

PRODUCTION AND OPERATIONS MANAGEMENT

**SPECIAL ISSUE ON PERSPECTIVES ON BIG DATA EDITED BY KALYAN SINGHAL,
QI FENG, RAM GANESHAN, NADA R. SANDERS, AND J. GEORGE SHANTHIKUMAR**

Kalyan Singhal, Qi Feng, Ram Ganeshan, Nada R. Sanders, J. George Shanthikumar
Introduction to the Special Issue on Perspectives on Big Data

Hau L. Lee
Big Data and the Innovation Cycle

Wallace J. Hopp, Jun Li, Guihua Wang
Big Data and the Precision Medicine Revolution

Marshall Fisher, Ananth Raman
Using Data and Big Data in Retailing

Qi Feng, J. George Shanthikumar
How Research in Production and Operations Management May Evolve in the Era of Big Data

Charles J. Corbett
How Sustainable Is Big Data?

Jayashankar M. Swaminathan
Big Data Analytics for Rapid, Impactful, Sustained, and Efficient (RISE) Humanitarian Operations

Sripad K. Devalkar, Sridhar Seshadri, Chitrabhanu Ghosh, Allen Mathias
Data Science Applications in Indian Agriculture

Maxime C. Cohen
Big Data and Service Operations

Samayita Guha, Subodha Kumar
Emergence of Big Data Research in Operations Management, Information Systems, and Healthcare:
Past Contributions and Future Roadmap

Information for Contributors

Submit your papers or possible publication in *Production and Operations Management* (POM) to an appropriate departmental editor or to the editor-in-chief electronically (preferably a pdf file) at the journal's Manuscript Central site. Use the following link: <https://mc.manuscriptcentral.com/poms>.

We follow a double-blind review process, and therefore please remove your name and any acknowledgment from the manuscript before submitting it.

Length

Although there is no page limit on initial (i.e., first-round) submissions, authors should strive to keep their papers at no longer than 38 pages double-spaced in a size 11 font. The page limit on the final version is 38 pages. The departmental editors can relax the page limit in exceptional cases. In addition, some of the proofs, lemmas, tables and other supporting material may be included in an online companion. There is no page limit for the online companion. A page limit is imposed because papers should focus on their main contribution and short papers are more likely to be read and cited.

Spacing and Formatting

The manuscript, including the abstract and references, should be double-spaced in size 11 font, and it should have one-inch margins on all four sides of the page. Each page of the manuscript should be numbered.

The first page of the manuscript should include the title of the paper and an abstract of 100-250 words. The title should be concise and descriptive, not exceeding 12 words. The abstract is a mini paper, and it should summarize important contents of the paper. It is not a roadmap of the paper, and it should not contain formulas, references, or abbreviations. Please list up to four key words at the end of the abstract.

Cite references in the text by enclosing the authors' names and the year of publication in parentheses. List references alphabetically in a double-spaced reference section at the end of the paper in the following style:

Hayes, R. H., G. P. Pisano. 1996. Manufacturing strategy: At the intersection of two paradigm shifts. *Production and Operations Management*, 5 (1), 25-41.

Submission of a paper to *Production and Operations Management* for possible publication implies that the author is certifying that the manuscript is not copyrighted; nor has it been accepted for publication or published by another refereed journal; nor is it being refereed elsewhere at the same time. If the paper or any version of it has appeared or will appear in a non-refereed publication, the details of such publication must be made known to the editor at the time of submission so that the suitability of the paper for publication in *Production and Operations Management* can be assessed.

Prior to submitting a paper, the authors should review the ethical guidelines of *Production and Operations Management* at <http://www.poms.org/Ethics-POMS-Website-total-document.pdf>. Upon submission, they will need to certify that the manuscript submission meets the POM Ethical Guidelines.

Once a manuscript is accepted for publication, the authors are asked to sign a copyright transfer agreement and to supply quality artwork for any figures.

PRODUCTION AND OPERATIONS MANAGEMENT

- 1639** Kalyan Singhal, Qi Feng, Ram Ganeshan, Nada R. Sanders, J. George Shanthikumar
Introduction to the Special Issue on Perspectives on Big Data
- 1642** Hau L. Lee
Big Data and the Innovation Cycle
- 1647** Wallace J. Hopp, Jun Li, Guihua Wang
Big Data and the Precision Medicine Revolution
- 1665** Marshall Fisher, Ananth Raman
Using Data and Big Data in Retailing
- 1670** Qi Feng, J. George Shanthikumar
How Research in Production and Operations Management May Evolve in the Era of Big Data
- 1685** Charles J. Corbett
How Sustainable Is Big Data?
- 1696** Jayashankar M. Swaminathan
Big Data Analytics for Rapid, Impactful, Sustained, and Efficient (RISE) Humanitarian Operations
- 1701** Sripad K. Devalkar, Sridhar Seshadri, Chitrabhanu Ghosh, Allen Mathias
Data Science Applications in Indian Agriculture
- 1709** Maxime C. Cohen
Big Data and Service Operations
- 1724** Samayita Guha, Subodha Kumar
Emergence of Big Data Research in Operations Management, Information Systems, and Healthcare: Past Contributions and Future Roadmap

Production and Operations Management

PRODUCTION AND OPERATIONS MANAGEMENT, (Print ISSN: 1059-1478; Online ISSN: 1937-5956), is published monthly on behalf of the Production and Operations Management Society by Wiley Subscription Services, Inc., a Wiley Company, 111 River St., Hoboken, NJ 07030-5774 USA. Periodicals Postage Paid at Hoboken, NJ and additional offices.

Postmaster: Send all address changes to *PRODUCTION AND OPERATIONS MANAGEMENT*, John Wiley & Sons Inc., C/O The Sheridan Press, PO Box 465, Hanover, PA 17331 USA.

Copyright and Copying (in any format)

Copyright © 2018 Production and Operations Management Society. All rights reserved. No part of this publication may be reproduced, stored or transmitted in any form or by any means without the prior permission in writing from the copyright holder. Authorization to copy items for internal and personal use is granted by the copyright holder for libraries and other users registered with their local Reproduction Rights Organisation (RRO), e.g. Copyright Clearance Center (CCC), 222 Rosewood Drive, Danvers, MA 01923, USA (www.copyright.com), provided the appropriate fee is paid directly to the RRO. This consent does not extend to other kinds of copying such as copying for general distribution, for advertising or promotional purposes, for republication, for creating new collective works or for resale. Permissions for such reuse can be obtained using the RightsLink "Request Permissions" link on Wiley Online Library. Special requests should be addressed to: permissions@wiley.com

Information for subscribers

Production and Operations Management is published in 12 issues per year. Institutional subscription prices for 2018 are:

Print & Online: US\$979 (US), US\$979 (Rest of World), €633 (Europe), £504 (UK). Prices are exclusive of tax. Asia-Pacific GST, Canadian GST/HST and European VAT will be applied at the appropriate rates. For more information on current tax rates, please go to www.wileyonlinelibrary.com/tax-vat. The price includes online access to the current and all online backfiles to January 1st 2014, where available. For other pricing options, including access information and terms and conditions, please visit www.wileyonlinelibrary.com/access

Delivery Terms and Legal Title

Where the subscription price includes print issues and delivery is to the recipient's address, delivery terms are Delivered at Place (DAP); the recipient is responsible for paying any import duty or taxes. Title to all issues transfers Free of Board (FOB) our shipping point, freight prepaid. We will endeavor to fulfill claims for missing or damaged copies within six months of publication, within our reasonable discretion and subject to availability.

Back issues: Single issues from current and recent volumes are available at the current single issue price from cs-journals@wiley.com. Earlier issues may be obtained from Periodicals Service Company, 351 Fairview Avenue – Ste 300, Hudson, NY 12534, USA. Tel: +1 518 822-9300, Fax: +1 518 822-9305, Email: psc@periodicals.com

Disclaimer

The Publisher, Production and Operations Management Society, and Editors cannot be held responsible for errors or any consequences arising from the use of information contained in this journal; the views and opinions expressed do not necessarily reflect those of the Publisher, Production and Operations Management Society, and Editors, neither does the publication of advertisements constitute any endorsement by the Publisher, Production and Operations Management Society, and Editors of the products advertised.

Publisher

Production and Operations Management is published by Wiley Periodicals, Inc., Commerce Place, 350 Main Street, Malden, MA 02148; Tel: (781) 388-8200; Fax: (781) 388-8210. Wiley Periodicals, Inc. is now part of John Wiley & Sons.

Journal Customer Services: For ordering information, claims and any enquiry concerning your journal subscription please go to www.wileycustomerhelp.com/ask or contact your nearest office.

Americas: Email: cs-journals@wiley.com; Tel: +1 781 388 8598 or +1 800 835 6770 (toll-free in the USA & Canada).

Europe, Middle East and Africa: Email: cs-journals@wiley.com; Tel: +44 (0) 1865 778315.

Asia Pacific: Email: cs-journals@wiley.com; Tel: +65 6511 8000.

Japan: For Japanese speaking support, Email: cs-japan@wiley.com.

Visit our Online Customer Help available in 7 languages at www.wileycustomerhelp.com/ask

Production Editor: Abigail Gutierrez (email: poms@wiley.com)

Advertising: Rachel Abbott (email: rabbott@wiley.com)

Wiley's Corporate Citizenship initiative seeks to address the environmental, social, economic, and ethical challenges faced in our business and which are important to our diverse stakeholder groups. Since launching the initiative, we have focused on sharing our content with those in need, enhancing community philanthropy, reducing our carbon impact, creating global guidelines and best practices for paper use, establishing a vendor code of ethics, and engaging our colleagues and other stakeholders in our efforts. Follow our progress at www.wiley.com/go/citizenship

View this journal online at wileyonlinelibrary.com/journal/pom

Wiley is a founding member of the UN-backed HINARI, AGORA, and OARE initiatives. They are now collectively known as Research4Life, making online scientific content available free or at nominal cost to researchers in developing countries. Please visit Wiley's Content Access – Corporate Citizenship site: <http://www.wiley.com/WileyCDA/Section/id-390082.html>

Printed in USA by The Sheridan Press

Abstracting and Indexing Services: The Journal is indexed by ISI, CrossRef, CSA ABI-INFORM, Current Abstracts, Current Contents, EBSCOhost, Inspec, OCLC, ProQuest, Science Citation Index, SCOPUS, Swets Information Services, Thomson Gale.

For submission instructions, subscription and all other information visit: [http://onlinelibrary.wiley.com/journal/10.1111/\(ISSN\)1937-5956](http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1937-5956)

PRODUCTION AND OPERATIONS MANAGEMENT

FOUNDER AND EDITOR-IN-CHIEF: PROFESSOR KALYAN SINGHAL, Merrick School of Business, University of Baltimore, 1420 N. Charles Street, Baltimore, MD 21201. Phone: (410) 837-4976. Fax: (410) 837-5722. ksinghal@ubalt.edu.

EXECUTIVE EDITOR: PROFESSOR JAYA SINGHAL, Merrick School of Business, University of Baltimore, 1420 N. Charles Street, Baltimore, MD 21201.

DEPUTY EDITOR: PROFESSOR SUBODHA KUMAR, Fox School of Business, Temple University, Philadelphia, PA 19122. subodha@temple.edu

MISSION STATEMENT: The mission of *Production and Operations Management* is to serve as the flagship research journal in operations management in manufacturing and services. The journal publishes scientific research into the problems, interests, and concerns of managers who manage product and process design, operations, and supply chains. It covers all topics in product and process design, operations, and supply chain management and welcomes papers using any research paradigm.

SUBMISSION: Manuscripts should be submitted electronically to the appropriate department editor or to the editor-in-chief. Department editors are listed below, along with their senior editors. The Statements of Objectives for the various departments are available on the website at www.poms.org and have been published in *Production and Operations Management*. The most recent statement is available in the issue identified by the date following each department title.

BEHAVIORAL OPERATIONS (1/2007)

Professor Elena Katok

University of Texas at Dallas
ekatok@utdallas.edu

Professor Mirko Kremer

Frankfurt School of Finance & Management, Germany
m.kremer@fs.de

- Elliot Bendoly, Ohio State University
- Kay-Yut Chen, University of Texas at Arlington
- Ernan E. Haruvy, University of Texas at Dallas
- Kyle Hyndman, University of Texas at Dallas
- Rogelio Oliva, Texas A&M University
- Kenneth Schultz, Air Force Institute of Technology
- Rachna Shah, University of Minnesota
- John Sterman, Massachusetts Institute of Technology
- Xuanming Su, The Wharton School
- Yaozhong Wu, National University of Singapore

DATA SCIENCE, STOCHASTIC AND OPTIMIZATION (6/2018)

Professor Qi (Annabelle) Feng

Purdue University
annabellefeng@purdue.edu

Professor Zuo-Jun (Max) Shen

University of California at Berkeley
maxshen@berkeley.edu

DISASTER MANAGEMENT (2/2015)

Professor Sushil K. Gupta

Florida International University
guptask@fiu.edu

Professor Martin K. Starr

Rollins College and Columbia University
mstarr@rollins.edu

- Nezih Altay, DePaul University
- Maria Besiou, Kühne Logistics University
- Nicole DeHoratius, The University of Chicago
- Mahyar Eftekhari, Arizona State University
- Paulo Gonçalves, Università della Svizzera Italiana
- Sameer Hasija, INSEAD
- Gyöngyi Kovács, HUMLOG Institute, Hanken School of Economics
- Alfonso J. Pedraza-Martinez, Indiana University
- Peter W. Robertson, University of Wollongong
- Tina Wakolbinger, WU (Vienna University of Economics and Business)

E-BUSINESS AND OPERATIONS (3/2006)

Professor Edward G. Anderson Jr.

University of Texas at Austin
edward.anderson@mcombs.utexas.edu

Professor Geoffrey G. Parker

Dartmouth College
geoffrey.g.parker@dartmouth.edu

- Jian Chen, Tsinghua University
- Amitava Dutta, George Mason University
- Richard Steinberg, London School of Economics
- Daewon Sun, University of Notre Dame
- Yinliang (Ricky) Tan, Tulane University

GLOBAL OPERATIONS STRATEGY (4/2013)

Professor M. Johnny Rungtusanatham

Ohio State University
rungtusanatham_1@fisher.osu.edu

Enno Siemsen

University of Wisconsin-Madison
siems017@umn.edu

- Gopesh Anand, University of Illinois-Urbana Champaign
- Aravind Chandrasekaran, Ohio State University
- Laurens G. Debo, University of Chicago
- Yan Dong, University of South Carolina
- Lawrence Fredendall, Clemson University
- Karan Girotra, INSEAD
- Murat Kristal, York University
- Manoj Malhotra, University of South Carolina
- Eve Rosenzweig, Emory University
- Fabrizio Salvador, Instituto de Empresa
- Gary D. Scudder, Vanderbilt University
- Bradley Staats, University of North Carolina
- Shawnee Vickery, Michigan State University
- Scott Webster, Arizona State University

HEALTHCARE OPERATIONS MANAGEMENT (2/2010)

Professor Sergei Savin

The Wharton School
savin@wharton.upenn.edu

- Tugba Cayirli, Özyegin University
- Rachel Chen, University of California at Davis
- Stephen Chick, INSEAD
- K. C. Diwas, Emory University
- Craig Froehle, University of Cincinnati
- Bruce Golden, University of Maryland
- Susan Lu, Purdue University
- Zhan Pang, City University of Hong Kong
- Martin Puterman, University of British Columbia
- Nicos Savva, London Business School
- Rachna Shah, University of Minnesota
- Steven Shechter, University of British Columbia
- Anita Tucker, Boston University
- Vedat Verter, McGill University
- Greg Zaric, University of Western Ontario

INDUSTRY STUDIES & PUBLIC POLICY (1/2012)

Professor Edward G. Anderson Jr.

University of Texas
Edward.Anderson@mcombs.utexas.edu

Professor Nitin R. Joglekar

Boston University
joglekar@bu.edu

- Saurabh Bansal, Pennsylvania State University
- W. C. Benton, Ohio State University
- Amy Cohn, University of Michigan
- Jane Davies, Cambridge University
- Pnina Feldman, Boston University
- Charles Fine, MIT
- John Gray, Ohio State University
- David Lane, Henley Business School
- Marvin Lieberman, UCLA
- Jeffrey Macher, Georgetown University
- Douglas Morrice, University of Texas
- Sriram Narayanan, Michigan State University
- Rogelio Oliva, Texas A&M University
- Kingshuk Sinha, University of Minnesota
- Bradley Staats, University of North Carolina
- Rohit Verma, Cornell University

MANAGEMENT OF TECHNOLOGY (12/2014)

Professor Cheryl Gaimon

Georgia Institute of Technology
cheryl.gaimon@dupree.gatech.edu

- Terrence August, University of California at San Diego
- Sulim Ba, University of Connecticut
- Elliot Bendoly, Ohio State University

- Janice Carrillo, University of Florida
- Raul Chao, University of Virginia
- Sarv Devaraj, University of Notre Dame
- Hong Guo, University of Notre Dame
- Manpreet Hora, Georgia Institute of Technology
- Moren Lévesque, York University
- S. Rajagopalan, University of Southern California
- Karthik Ramachandran, Georgia Institute of Technology
- Glen Schmidt, University of Utah
- Jeff Stratman, Texas Christian University
- D. J. Wu, Georgia Institute of Technology

MANUFACTURING OPERATIONS (1/2004)

Professor Panos Kouvelis

Washington University in St. Louis
kouvelis@olin.wustl.edu

Professor Chelliah Sriskandarajah

Texas A&M University
chelliah@mays.tamu.edu

- Anupam Agrawal, Texas A&M University
- Chester Chambers, Johns Hopkins University
- Zhi-Long Chen, University of Maryland
- Wen-Chyuan Chiang, University of Tulsa
- Jiri Chod, Boston College
- Lingxiu Dong, Washington University in St. Louis
- Neil Geismar, Texas A&M University
- Selçuk Karabati, Koç University, Turkey
- Michael Ketzenberg, Texas A&M University
- Chung Yee Lee, Hong Kong University of Science & Technology
- Cuihong Li, University of Connecticut
- Joseph Milner, University of Toronto, Canada
- Kamran Moinzadeh, University of Washington
- Chuck Munson, Washington State University
- Tava Lennon-Olsen, University of Auckland
- Anand Paul, University of Florida

NEW PRODUCT DEVELOPMENT, R&D, AND PROJECT MANAGEMENT (1/2004)

Professor Jurgen Mihm

INSEAD
jurgen.mihm@insead.edu

- Sreekumar Bhaskaran, Southern Methodist University
- Sanjiv Erat, University of California at San Diego
- Glen Schmidt, University of Utah
- Svenja Sommer, Purdue University
- Manuel Sosa, INSEAD
- Yi Xu, University of Maryland

POM-ACCOUNTING INTERFACE (1/2013)

Professor Anil Arya

Ohio State University
arya_4@fisher.osu.edu

- Mark Bagnoli, Purdue University
- Eva Labro, University of North Carolina-Chapel Hill
- Brian Mittendorf, Ohio State University
- Suresh Radhakrishnan, University of Texas at Dallas
- Shiva Sivaramakrishnan, Rice University

POM-ECONOMICS INTERFACE (3/2013)

Professor Haresh Gurnani

Wake Forest University
haresh@miami.edu

- Subramanian Balachander, Purdue University
- Bin Hu, University of North Carolina at Chapel Hill
- Ming Hu, University of Toronto
- Kinshuk Jerath, Columbia University
- Harish Krishnan, University of British Columbia
- Dimitri Kuksov, University of Texas at Dallas
- Guoming Lai, University of Texas
- Elie Ofek, Harvard University
- Amit Pazgal, Rice University
- Jeff Shulman, University of Washington
- Shubhanshu Singh, Johns Hopkins University
- David Soberman, University of Toronto
- Robert Swinney, Duke University
- Zhibin (Ben) Yang, University of Oregon
- Leon Zhu, University of Southern California

POM-FINANCE INTERFACE (2/2013)

Professor Sridhar Seshadri

Indian School of Business
Sridhar_Seshadri@isb.edu

- Rene Caldentey, New York University
- Nicola Secomandi, Carnegie Mellon University

POM-INFORMATION SYSTEMS INTERFACE (3/2014)

Professor Subodha Kumar

Temple University
subodha@temple.edu

Professor Asoo Vakharia

University of Florida
asoo.vakharia@warrington.ufl.edu

- Xue Bai, University of Connecticut
- Achal Bassamboo, Kellogg School of Management
- Gangshu Cai, Santa Clara University
- Yonghua Ji, University of Alberta
- Zhengrui (Jeffrey) Jiang, Iowa State University
- Amit Mehra, University of Texas at Dallas
- Sunil Mithas, University of Maryland
- Paul Pavlou, Temple University
- David Xiaosong Peng, University of Houston
- Liangfei Qiu, University of Florida
- Arvind Tripathi, University of Auckland
- Oliver Yao, Lehigh University
- Xiaohang Yue, University of Wisconsin-Milwaukee

POM-MARKETING INTERFACE (1/2013)

Professor Amiya K. Chakravarty

Northeastern University
akc@neu.edu

- Jian Chen, Tsinghua University
- Mabel C. Chou, National University of Singapore
- Tony Cui, University of Minnesota
- Laurens Debo, University of Chicago
- Hans Heese, North Carolina State University
- Arnd Huchzermeier, WHU – Koblenz, Germany
- Kinshuk Jerath, Columbia University
- Oded Koenigsberg, London Business School
- Noah Lim, University of Wisconsin – Madison
- Hosun Rhim, Korea University
- Alfred Taudes, Vienna University of Economics and Business

POM FORUM

Professor Kalyan Singhal

University of Baltimore
ksinghal@ubalt.edu

Professor Manmohan Sodhi

City University London
mohansodhi@gmail.com

POM PRACTICE

Professor Sean P. Willems

Boston University
willems@bu.edu

- Hongmin Li, Arizona State University
- Gilvan Souza, Indiana University

RETAIL OPERATIONS (4/2010)

Professor Felipe Caro

UCLA
fcaro@anderson.ucla.edu

Professor Nicole DeHoratius

University of Chicago
Nicole@DeHoratius.com

- Aydin Alptekinoglu, Pennsylvania State University
- Goker Aydin, Johns Hopkins University
- Srinagesh (Nagesh) Gavirneni, Cornell University
- Hans Heese, North Carolina State University
- Dorothee Honhon, University of Texas at Dallas
- Saravanan Kesavan, University of North Carolina
- Adam Mersereau, University of North Carolina
- Antonio Moreno, Harvard Business School

REVENUE MANAGEMENT (3/2006)

Professor Huseyin Topaloglu

Cornell University
ht88@cornell.edu

- Goker Aydin, Johns Hopkins University
- René Caldentey, New York University
- William L. Cooper, University of Minnesota
- Soulaymane Kachani, Columbia University
- Sumit Kunnumkal, Indian School of Business
- Costis Maglaras, Columbia University
- Georgia Perakis, MIT
- Robert L. Phillips, Nomis Solutions
- Paat Rusmevichientong, University of Southern California
- Kalyan Talluri, Imperial College, London
- Zizhuo Wang, University of Minnesota
- Dan Zhang, University of Colorado

SERVICE OPERATIONS MANAGEMENT (1/2004)

Professor Michael Pinedo

New York University
mpinedo@stern.nyu.edu

Professor Aleda V. Roth

Clemson University
aroth@clemson.edu

- Mor Armony, New York University
- Ryan Buell, Harvard Business School
- Maxime Cohen, New York University
- Milind Dawande, University of Texas at Dallas
- Craig Froehle, University of Cincinnati
- Wendell Gilland, University of North Carolina-Chapel Hill
- Harry Groenevelt, University of Rochester
- Gregory Heim, Texas A&M University
- Martin Lariviere, Northwestern University
- Larry Menor, University of Western Ontario
- Mahesh Nagarajan, University of British Columbia
- Suresh Nair, University of Connecticut
- Pedro Oliveira, Católica-Lisbon School of Business & Economics
- Rob Shumsky, Dartmouth College
- Anita Tucker, Boston University
- Guohua Wan, Shanghai Jiao Tong University
- Wenqiang Xia, New York University

SPECIAL RESPONSIBILITIES

Professor Wallace J. Hopp

University of Michigan
whopp@umich.edu

Professor Hau L. Lee

Stanford University
haulee@stanford.edu

Professor Suresh P. Sethi

University of Texas at Dallas
Sethi@utdallas.edu

Professor J. George Shanthikumar

Purdue University
shanthikumar@purdue.edu

Professor Sridhar Tayur

Carnegie Mellon University
stayur@andrew.cmu.edu

- Metin Cakanyildirim, University of Texas at Dallas
- Tao Li, Santa Clara University

SUSTAINABLE OPERATIONS (3/2006)

Professor Atalay Atasu

Georgia Institute of Technology
atalay.atasu@scheller.gatech.edu

- Vishal Agrawal, Georgetown University
- Tamer Boyaci, ESMT Berlin
- Nicole Darnall, Arizona State University
- Mark Ferguson, University of South Carolina
- Moritz Fleischmann, University of Mannheim
- Michael Galbreth, University of South Carolina
- Brian Jacobs, Michigan State University
- Robert D. Klassen, University of Western Ontario
- Suresh Muthulingam, Pennsylvania State University
- Gil Souza, Indiana University
- Vedat Verter, McGill University
- Yanchong (Karen) Zheng, MIT Sloan
- Eda Kemahlioglu Ziya, NCSU

SUPPLY CHAIN MANAGEMENT (1/2004)

Professor Albert Ha

Hong Kong University of Science and Technology
imayha@ust.hk

Professor M. Eric Johnson

Vanderbilt University
m.eric.johnson@owen.vanderbilt.edu

Professor Vinod Singhal

Georgia Institute of Technology
vinod.singhal@mgt.gatech.edu

Professor Jayashankar M. Swaminathan

University of North Carolina
msj@unc.edu

- Gad Allon, Northwestern University
- Metin Cakanyildirim, University of Texas
- Kyle Cattani, Indiana University
- Gregory A. DeCroix, University of Wisconsin
- Vinayak Deshpande, University of North Carolina
- Feryal Erhun, Stanford University
- Steve Gilbert, University of Texas
- Hans Heese, North Carolina State University
- Kevin Hendricks, Wilfrid Laurier University
- Manpreet Hora, Georgia Institute of Technology
- Arnd Huchzermeier, WHU, Otto-Beisheim Graduate School of Management
- Phil Kaminsky, University of California at Berkeley
- Harish Krishnan, University of British Columbia
- Guoming Lai, University of Texas at Austin
- Lauren Xiaoyuan Lu, University of North Carolina
- Manoj K. Malhotra, University of South Carolina
- Ann Maruchek, University of North Carolina
- Douglas Morrice, University of Texas at Austin
- Sriram Narayanan, Michigan State University
- Özalp Özer, University of Texas at Dallas
- Ali Parlakturk, University of North Carolina
- Kumar Rajaram, UCLA
- Saibal Ray, McGill University
- Milind Sohoni, Indian School of Business
- Greys Sobic, University of Southern California
- Jeff Stratman, Texas Christian University
- Doug Thomas, Penn State University
- Ulrich Thonemann, Universität zu Köln
- Brian Tomlin, Dartmouth College
- Andy A. Tsay, Santa Clara University
- Hongtao Zhang, Hong Kong University of Science and Technology
- Rachel Zhang, Hong Kong University of Science and Technology

TOPICS NOT COVERED ABOVE

Professor Vinod Singhal (Empirical Research)

Georgia Institute of Technology
vinod.singhal@mgt.gatech.edu

Professor Terry Taylor

University of California at Berkeley
taylor@haas.berkeley.edu

- Shannon Anderson, University of California at Davis
- Volodymyr Babich, Georgetown University
- Anant Balakrishnan, University of Texas at Austin
- Nagraj (Raju) Balakrishnan, Clemson University
- Kurt M. Bretthauer, Indiana University
- Xin Chen, University of Illinois at Urbana Champaign
- Thomas Choi, Arizona State University
- Tsan-Ming Choi, Hong Kong Polytechnic University
- Emre Demirezen, University of Florida
- Don Eisenstein, University of Chicago
- Jan C. Fransoo, Technische Universiteit Eindhoven
- Soumen Ghosh, Georgia Institute of Technology
- Pengfei Guo, The Hong Kong Polytechnic University
- Kevin Hendricks, Wilfrid Laurier University
- Qiaohai Joice Hu, Purdue University
- Woonghee Tim Huh, University of British Columbia
- Peter Kolesar, Columbia University
- K. Ravi Kumar, University of Southern California
- Guoming Lai, University of Texas at Austin
- Michael Lapré, Vanderbilt University
- Dengpan Liu, Tsinghua University
- Lauren Xiaoyuan Lu, University of North Carolina
- Ram Narasimhan, Michigan State University
- Roger Schmenner, Indiana University
- Bala Shetty, Texas A&M University
- Rick So, University of California, Irvine
- Mark Spearman, Factory Physics, Inc.
- Kwei Tang, National Chengchi University
- Giri Kumar Tayi, State University of New York at Albany
- Chris Voss, London Business School
- Peter Ward, Ohio State University
- Nan Yang, University of Miami
- Xiande Zhao, China Europe International Business School



EDITORIAL REVIEW BOARD

James Abbey, Texas A&M University
Jason Acimovic, Pennsylvania State University
Philipp Afeche, University of Toronto
Sal Agnihotri, Binghamton University
Naren Agrawal, Santa Clara University
Reza Ahmadi, University of California at Los Angeles
Hyun-Soo Ahn, University of Michigan
Arzum Akkas, Massachusetts Institute of Technology
Mehmet Altug, George Washington University
Joachim Arts, Eindhoven University of Technology
Diane Bailey, University of Texas at Austin
Burcu Balçik, Ozyegin University
Ravi Bapna, University of Connecticut
Amit Basu, Southern Methodist University
Sara L. Beckman, University of California at Berkeley
Gemma Berenguer, Purdue University
Fernando Bernstein, Duke University
Maria Besiou, Kuehne Logistics University
Jyoti Bhattacharjya, University of Sydney
Shantanu Bhattacharya, INSEAD
Martin Bichler, Technical University of Munich
Diane P. Bischak, University of Calgary
Eyal Biyalogorsky, Arison School of Business
Irineu Brito, Fatec-SJC
Ryan Buell, Harvard Business School
John Buzacott, York University
Gangshu Cai, Santa Clara University
Carri Chan, Columbia University
Aravind Chandrasekaran, Ohio State University
Liang Chao, Cheung Kong Graduate School of Business
Aadhaar Chaturvedi, University of Namur
Chien-Ming Chen, Nanyang Technological University
Pei-yu Chen, Carnegie Mellon University
Ying-Ju Chen, Hong Kong University of Science and Technology
Chun-Hung Chiu, Sun Yat-sen University
Soo-Haeng Cho, Carnegie Mellon University
Pui-Sze Chow, Hong Kong Polytechnic University
Geoffrey Chua, Nanyang Technological University
Eren Cil, University of Oregon
Christopher Craighead, Pennsylvania State University
Maqbool Dada, Johns Hopkins University
Tinglong Dai, Johns Hopkins University
Sriram Dasu, University of Southern California
Andrew Davis, Cornell University
Ozgun Caliskan Demirag, Penn State University
Sarang Deo, Indian School of Business
Preyas Desai, Duke University
Sripad Devalkar, Indian School of Business
Mike Dixon, Naval Postgraduate School
Ken Doerr, Naval Postgraduate School
Cheryl Druehl, George Mason University
Jean-Pierre Dubè, University of Chicago
Elizabeth Durango-Cohen, Illinois Institute of Technology
Rebecca Duray, University of Colorado at Colorado Springs
Kaushik Dutta, National University of Singapore
Cuneyt Eroglu, Northeastern University
Sinan Erzurumlu, Babson College
Gokce Esenduran, Ohio State University
Xiang Fang, University of Wisconsin-Milwaukee
Joy Field, Boston College
Adam Fleischhacker, University of Delaware
Lee O. Fleming, Harvard University
Barbara Flynn, Indiana University
Michael Fry, University of Cincinnati
Xianghua Gan, Polytechnic University of Hong Kong
Ram Ganeshan, College of William and Mary
Jean-Philippe Gayon, Grenoble Institute of Technology
Joseph Geunes, University of Florida
Jarrod Goentzel, Massachusetts Institute of Technology
Paulo Goes, University of Connecticut
Paulo Goncalves, University of Lugano
Erica Gralla, George Washington University
Walter Gutjahr, University of Vienna
Sean Handley, University of Notre Dame
Janet Hartley, Bowling Green State University
Xiuli (Sophia) He, University of North Carolina at Charlotte
Vernon Hsu, The Chinese University of Hong Kong
Tingliang Huang, University College London
Xiao Huang, Concordia University
Xiaowen Huang, Miami University
Baofeng Huo, Zhejiang University
Kyle Hyndman, Maastricht University

Rouba Ibrahim, University College London
Jonathan Jackson, Washington State University
Miguel Jaller, UC Davis
Ganesh Janakiraman, University of Texas at Dallas
Ramkumar Janakiraman, University of South Carolina
Stefanus Jasin, University of Michigan
Jayanth Jayaram, University of South Carolina
Vaidy Jayaraman, University of Miami
Houyuan Jiang, University of Cambridge
Li Jiang, Hong Kong Polytechnic University
David Johnston, Schulich School of Business
Moutaz Khouja, University of North Carolina, Charlotte
Jin Gyo Kim, Massachusetts Institute of Technology
Michael Kim, University of Toronto
Sang-Hyun Kim, Yale University
Kenneth J. Klassen, Brock University
Guangwen Kong, University of Minnesota Twin Cities
Xenophon Koufteros, Texas A&M University
Tim Kraft, University of Virginia
Ramayya Krishnan, Carnegie Mellon University
Trichy Krishnan, National University of Singapore
Thomas Kull, Arizona State University
Nathan Kunz, University of North Florida
Mumin Kurtulus, Vanderbilt University
Linda LaGanga, Mental Health Center of Denver
Deishin Lee, Harvard Business School
Hsiao-Hui Lee, University of Hong Kong
Adriana Leiras, PUC-Rio
Mingming Leng, Lingnan University
Meng Li, Rutgers University
Rong Li, Syracuse University
Shanling Li, McGill University
Wei Shi Lim, National University of Singapore
Yen-Ting Lin, University of San Diego
Kevin Linderman, University of Minnesota
Zhexin Liu, University of Michigan
Alan MacCormack, Harvard Business School
José A. Machuca, University of Sevilla
Alan Mackelprang, Georgia Southern University
B. Mahadevan, Indian Institute of Management, Bangalore
Pranab Majumder, Duke University
Vincent Mak, Cambridge Judge Business School
Debasish N. Mallick, University of St. Thomas
Suman Mallik, University of Kansas
Vidya Mani, Penn State University
Hirofumi Matsuo, University of Tsukuba
Ian McCarthy, Simon Fraser University
Richard Metters, Texas A&M University
Susan Meyer Goldstein, University of Minnesota
Anant Mishra, George Mason University
Sachin Modi, University of Toledo
Radha Mookerjee, University of Texas at Dallas
Liyang Mu, University of Delaware
Surender Munjal, Leeds University
Ana Muriel, University of Massachusetts
Nagesh Murthy, University of Oregon
Lin Nan, Purdue University
Anand Nair, Michigan State University
Barrie Nault, University of Calgary
Eitan Naveh, Technion-Israel Institute of Technology
David A. Nembhard, Penn State University
Sechan Oh, IBM
Jan Olhager, Lund University
Anton Ovchinnikov, University of Virginia
Rema Padman, Carnegie Mellon University
Mark Pagell, York University
Sofia Panagiotidou, University of Western Macedonia
Chris Parker, Penn State University
Jonathan Patrick, University of Ottawa
Valery Pavlov, University of Auckland
Olga Perdikaki, University of South Carolina
Vijay Pereira, University of Portsmouth
Sandun Perera, University of Michigan - Flint
Nicholas Petrucci, University of Illinois at Urbana Champaign
Çerag Pinçe, Kühne Logistics University
Madeleine Pullman, Portland State University
Hubert Pun, Western University
Xiangtong Qi, Hong Kong University of Science and Technology
Carrie Queenan, University of South Carolina
Elliot Rabinovich, Arizona State University
Tharanga Rajapakshe, University of Florida
Gal Raz, Ivey Business School

Guillaume Roels, University of California at Los Angeles
Dolores Romero-Morales, University of Oxford
Ying Rong, Shanghai Jiao Tong University
Jun Ru, University at Buffalo
Nils Rudi, University of Rochester
Jennifer Ryan, Rensselaer Polytechnic Institute
Reza Farzipoor Saen, Nottingham Trent University
Soroush Saghafian, Harvard University
Rajib L. Saha, Indian School of Business
Sibel Salman, Koc University
Scott Sampson, Brigham Young University
Nada R. Sanders, Lehigh University
Amar Sapra, University of Florida
Canan Savaskan, Northwestern University
Marc Schniederjans, University of Nebraska
Tobias Schoenherr, Michigan State University
Ralf Seifert, IMD
Jay Sethuraman, Columbia University
Andrew Shaefer, University of Pittsburgh
Wenjing Shen, Drexel University
Ruixia Shi, University of San Diego
Jeff Shockley, College of Charleston
Stephen Shum, City University of Hong Kong
Andreas Soteriou, University of Cyprus
Charles Sox, University of Alabama
Kathryn Stecke, University of Texas at Dallas
Xuemei Su, California State University at Long Beach
Haoying Sun, University of Kentucky
Aris Syntetos, Cardiff University
Srinivas (Sri) Talluri, Michigan State University
Burcu Tan, Tulane University
Tom Tan, Southern Methodist University
Jen Tang, Purdue University
Ou Tang, Linkopings Universitet
Yu Tang, University of Miami

Eylem Tekin, University of North Carolina
Ruud Teunter, University of Groningen
Sriram Thirumalai, University of Utah
Jordan Tong, University of Wisconsin-Madison
Danko Turcic, Washington University in St. Louis
Manoj Vanajakumari, Texas A&M University
Viswanath Venkatesh, University of Arkansas
Liana Victorino, University of Victoria
Veronica H. Villena, Pennsylvania State University
S. Vishwanathan, Nanyang Technological University
Stephan Wagner, ETH Zurich
Jingqi Wang, University of Hong Kong
Yimin Wang, Arizona State University
Yulan Wang, Hong Kong Polytechnic University
Don G. Wardell, University of Utah
Seunjin Whang, Stanford University
Yaozhong Wu, National University of Singapore
Zhou Xu, Hong Kong Polytechnic University
Mei Xue, Boston College
Houmin Yan, Chinese University of Hong Kong
Jian Yang, Rutgers Business School
Andy Yeung, Hong Kong Polytechnic University
Bo Zhang, IBM Research
Jun Zhang, University of Texas at Dallas
Michael Zhang, HKU of Science and Technology
Peter Zhang, Georgia State University
Rui Zhang, Xiamen University of Technology
Xuan Zhao, Wilfrid Laurier University
Xuying Zhao, University of Notre Dame
Jing (Jenny) Zhou, University of North Carolina, Charlotte
Xiang Zhu, University of Groningen
Yunxia (Peter) Zhu, Rider University
Eda Kemahliaglu Ziya, University of North Carolina, Chapel Hill
Serhan Ziya, University of North Carolina at Chapel Hill

Introduction to the Special Issue on Perspectives on Big Data

Kalyan Singhal

Merrick School of Business, University of Baltimore, ksinghal@ubalt.edu

Qi Feng

Krannert School of Management, Purdue University, annabellefeng@purdue.edu

Ram Ganeshan

Raymond A. Mason School of Business, The College of William and Mary, ram.ganeshan@mason.wm.edu

Nada R. Sanders

D'Amore-McKim School of Business, Northeastern University, n.sanders@northeastern.edu

J. George Shanthikumar

Krannert School of Management, Purdue University, shanthikumar@purdue.edu

Big data has the potential of offering valuable insights into the way organizations function, and it is changing the way organizations make decisions. Nine invited essays provide a wide range of perspectives on the role of big data in customer-driven supply chains, healthcare operations, retail operations, demand planning and manufacturing, environmental and social issues, humanitarian operations, agriculture supply chains, and service operations. Decision makers should have clean, valid, and reliable data, and they should have a thorough understanding of the contexts of applications. Big data shorten virtual distance to customers, and thus facilitate personalization of products and services. Successful implementation of big data applications requires sharing the data with appropriate stakeholders.

Key words: big data; data analytics; healthcare operations; operations management; service operations; supply-chain management

1. Introduction

Big data and associated technological developments (e.g., internet of things, industrial internet of things, cyber-physical systems) are dramatically changing the landscape of operations and supply-chain management. Researchers in our field, as in many others, are increasingly devoting significant effort to understand the emerging business models and changing business principles. Given many new phenomena, many unknowns have yet to be discovered, unknowns that will affect how the associated applications may evolve and how the related research should be shaped. We don't know whether big data is fundamentally changing the ways we conduct research, or whether we can just hit the new nails with the old hammers.

With these developments in mind, we invited some leading scholars in our field to share their thoughts on how big data is affecting research in POM. We have collected nine essays in this special issue. Each of these essays offers interesting views on how big data is reshaping the research across various domains. The authors provide examples of new business models enabled by big data.

2. Nine Essays

- *An evolutionary view:* Hau Lee adapts a three-stage framework for technology innovation to envision how big data may evolve and change the way we manage the supply chain. He highlights the shift toward managing the “long tail” and customer-driven supply chains.

- *Healthcare operations*: Wallace Hopp, Jun Li, and Guihua Wang explain the use of observational data from nontraditional sources to supplement the traditional experimental data from clinical studies and thus to transform the one-size-fit-all approach to precision medicine.
- *Retail operations*: Marshall Fisher and Ananth Raman share their knowledge of how data analytics lead to service innovations. They focus on the transformation from data to improved decisions concerning assortment, online dynamic pricing, online order fulfillment, and store closings.
- *Demand planning and manufacturing*: Qi Feng and George Shanthikumar focus on how big data is changing operational planning. They demonstrate how one should use prototype models in operations and production management.
- *An environmental and social perspective*: Charles Corbett describes examples of smart ways of using data to reduce the environmental footprint, to manage energy efficiency, and to plan weather-based service and production. He also stresses the challenges in using big data analysis, such as creating undesired social and ethical consequences.
- *Humanitarian operations*: Jayashankar Swaminathan discusses how descriptive, prescriptive and predictive analysis can enable rapid, impactful, sustained and efficient humanitarian decision making. He offers insights on how decision makers can use data to improve their identification of needy populations, optimize supplier bases and resolve production bottlenecks in the distribution of humanitarian relief products and services.
- *Agriculture supply chains*: Sripad Devalkar, Sridhar Seshadri, Chitrabhanu Ghosh, and Allen Mathias recommend innovative data collection applications and information dissemination applications. Using market price analytics as an example, they explain how algorithmic data analysis and decision making can help to improve the productivity of farmers.
- *Service operations*: Maxime Cohen highlights how the emergence of big data has led to the transformation from intuition-based decision making to evidence-based decision making. He also emphasizes the role of the marketplace in producing innovative offerings in financial services, transportation, hospitality, and online platforms.
- *An overview*: Samayita Guha and Subodha Kumar summarize existing research on various issues that arise with big data in research on information systems, operations, and healthcare.

3. Transformation from Data to Efficient Decisions

As a common theme, the essays in this issue identify opportunities and challenges in research using big data. Their authors emphasize the need for research and the value of research that enables the transformation from data to efficient decisions. For example, Devalkar, Seshadri, Ghosh, and Mathias write that while data access is provided to farmers, they lack ways of using the data to guide their planning. Guha and Kumar point to the need for research to evaluate the benefit of adopting technologies or systems that collect, process and analyze big data. Corbett argues that more data can lead to worse decisions if not aggregated and structured properly. Hopp, Li, and Wang point out that a pure machine learning approach, while shown to be effective for predicting medical outcomes, is not directly helpful to guide decisions for individual patients.

The authors particularly emphasized two aspects to ensure quality decision making processes. The first is the risk associated with the data itself. The veracity of big data requires careful design of data-acquisition and calibration strategies and of feature-extraction and selection strategies so that decision makers have clean, valid, and reliable inputs to use in making decisions. The second aspect lies in appropriate ways of using data, which requires a thorough understanding of the application contexts and a clever integration of data with structural knowledge.

4. Personalization

Many of the authors of the essays also recognize new research issues with the trend of personalization. Lee points out the shortened virtual distance to consumers and transformation toward customer-driven planning. Cohen discusses the mechanisms of real-time personalization and targeted promotions, and Feng and Shanthikumar devise prototype models to demonstrate data integration for personalized demand planning. Devalkar, Seshadri, Ghosh, and Mathias describe algorithmic analysis based on the diversity of individual farms. Fisher and Raman highlight the value of tracking the behaviors of individual retail customers, and Guha and Kumar point out its value for healthcare patients. Hopp, Li and Wang suggest that combination of data on patients' heterogeneous responses to treatment alternatives and data on individual patient preferences can enable medical decisions customized at the level of individual patients. Several authors also point out

the potential challenges of personalization that one should not ignore in one's research. These include *security and privacy* of personal data (Cohen; Guha and Kumar; Hopp, Li and Wang), *ethical* use of data (Corbett), and *fairness and discrimination* (Cohen; Swaminathan).

5. Data Sharing and Benefit Distribution

Another potential research avenue identified is related to data sharing and benefit distribution. Successful implementation of data analytics requires sharing the right data with appropriate stakeholders (Corbett). Feng and Shanthikumar discuss the need for data exchange and for a coordination mechanism in a manufacturing network to enhance overall capability, while avoiding a learning race. Devalkar, Seshadri, Ghosh, and Mathias stress the importance

of sharing the right data with the right entity to ensure fair distribution of the benefit generated by big data without creating unbalanced power structures.

6. System Thinking

Finally, operations management scholars have long advocated system thinking in our research. The emergence of big data requires us to enlarge the scope of the system when we perform analyses and develop models. In the process, we also generate new research questions about the system. Data analysis without a system view, however, can lead to pitfalls. Cohen elaborates on how analysis of big data without a thorough understanding of the system can lead to “machine bias” and “spurious correlations.” Corbett also points out the danger of “letting data availability drive priority” and advocates careful consideration of underlying phenomena.

Big Data and the Innovation Cycle

Hau L. Lee*

Graduate School of Business, Stanford University, Stanford, California 94305, USA, haulee@stanford.edu

Big data and the related methodologies could have great impacts on operations and supply chain management. Such impacts could be realized through innovations leveraging big data. The innovations can be described as first improving existing processes in operations through better tools and methods; second expanding the value propositions through expansive usage or incorporating data not available in the past; and third allowing companies to create new processes or business models to serve customers in new ways. This study describes this framework of the innovation cycle.

Key words: big data; innovation cycle; supply chain re-engineering; business model innovations

History: Received: October 2017; Accepted: December 2017 by Kalyan Singhal, after 1 revision.

1. Introduction

Big data and the related methodologies to make use of it: data analytics and machine learning have been viewed as digital technologies that could revolutionize operations and supply chain management in business and society at large. In the 2016 survey of over 1000 chief supply chain officers or similar senior executives, the SCM World found big data analytics at the top of the list of what these executives viewed as most disruptive to their supply chains (O'Marah and Chen 2016) (Figure 1).

Big data has great promises, and researchers have identified the research potentials (e.g., see Jin et al. 2015). At the same time, there have been skeptics who expressed the need to have proper steps to unleash such potentials. Otherwise, Blake-Plock (2017) indicated that big data could also be defined by its “smallness,” that is, limitations to getting its values. In a way, this seems to have some resemblance to the RFID saga in the early 2000s, when RFID was touted by the Economists (2003) as “The Best Thing Since the Bar-Code.” There were a lot of hypes, and Lee and Ozer (2007) used the term *credibility gap* to describe the problems that industry reports had regarding the value of RFID, that is, there seemed to be values given without much substantiation.

I like to offer some perspectives on how big data could provide great values in operations and supply chain management, and stimulate research in this area.

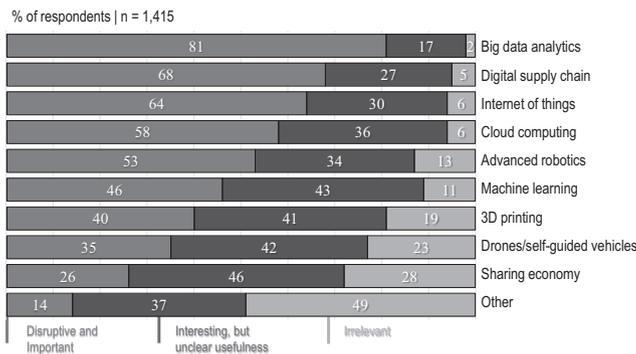
Will we have hypes again with big data? We have to be realistic and recognize that the use of big data and the associated development of tools to make use of it is a journey. This journey is a cycle that technological innovations often have to go through, and at every stage of the cycle, there are values and benefits,

as well as investments that we have to make in order to unleash the power and values. Understanding such a cycle could help to see how big data would not be a hype, and great impacts could be realized.

2. The 3-S Model of the Innovation Cycle

In Lee (2007), I described the 3-S model of technology evolution as a framework to see how RFID could be diffused. The model was based on the work of Malone and Rockart (1991). Such a model is equally applicable to think about the evolution of big data. Malone and Rockart (1991) used the introduction of automobile as a technological innovation to illustrate the cycle. Automobiles initially could replace horses or wagons as a means of transport. Increasing use of automobiles allowed people to travel more frequently and each trip could now cover more distance. With the subsequent development of highway systems, suburbs and shopping malls have been created, leading to a fundamental change in the structure of cities and people's work patterns. Malone and Rockart described another interesting evolution that occurred in the 12th century, when Dutch and Florence traders started using professional, traveling tradesmen to conduct trade, replacing the old means of barter trades. As this means of trade progressed, it was possible for regions that were geographically apart to engage in trades, and goods there were previously inaccessible became tradeable. Eventually, extensive trades required tradesmen providing insurance and loan services, and ultimately, whole new industries—insurance and financial services—were born.

The above examples illustrated that new technologies often evolve in three stages. The first one, which I called “Substitution,” is one when the new technology

Figure 1 Disruptive Technologies in Supply Chain [Color figure can be viewed at wileyonlinelibrary.com]

Source: O'Marah and Chen (2016).

is used in place of an existing one, to conduct a business activity. The second one, which I called “Scale,” is one when more items and more activities are used with the technology more frequently and extensively. The third is the “Structural Transformation” stage, when a new set of re-engineered activities can emerge with the new technology.

Electronic business has impacted supply chain management initially by improving the efficiency of very specific business processes. For example, digital sales channel has displaced some of the physical retail channel. Orders, invoices, and payments could be sent electronically instead of paper, phone, and fax. The value to the supply chain on digital communication and channels was clearcut. With millions and millions of people and businesses using the Internet as buyers and sellers, e-markets and the associated auctions, as well as P2P platforms such as Uber and AirBnB, have emerged. The deep impact of electronic business is fully realized when new products, new services, and ultimately new business models are created.

Let me illustrate the 3-S Model as an innovation cycle for big data, with some specific examples that I have come across.

3. The Substitution Stage of Big Data

The availability of big data can immediately allow new methods or processes to be developed to substitute existing ones for specific business activities. An obvious one is forecasting. Much deeper data analytics can now be used to replace previous forecasting methods, making full use of the availability of data. Such data were previously not easily accessible. For example, manufacturers and retailers used to not have their point-of-sales and inventory data in granular forms (time and geography), or that the data at other parts of the supply chain were not available.

Operations management researchers have long noted the value of information sharing in supply chains, and big data can certainly contribute to such sharing. Using more data can allow for more optimized inventory management, avoiding stockouts. Indeed, Economist (2017) reported on how Otto, a German e-commerce merchant, made use of big data and artificial intelligence (AI, which required extensive data for machine learning) to predict within 90% accuracy the 30-day demands for 200,000 of its items.

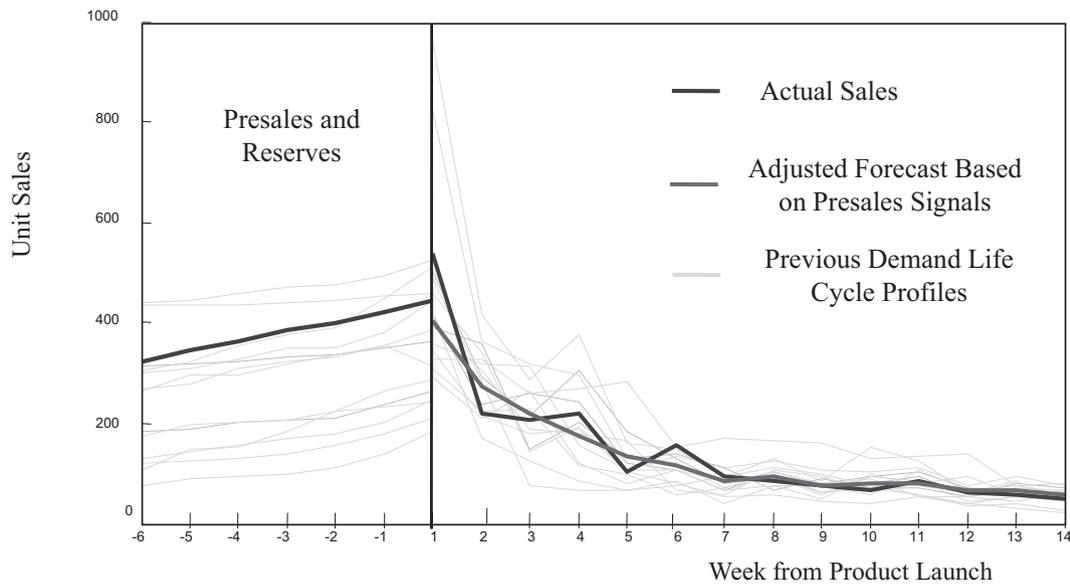
I observed another example of using big data to improve life cycle forecasting of new product demands that resembled the “accurate response” developed by Fisher and Raman (1996). An electronic game retailer had a hard time forecasting the very short life-time products of new computer games. It often had to guess what the life cycle demands could be, by picking among the life cycle sales of previous products, the one that perhaps resembled the closest product characteristics of the new product. The error rates were huge. The retailer created a database by collecting data of presales and reservations by potential customers prior to the launch of previous products, and then correlating such data with the actual subsequent life cycle sales. Such a database could then be used for a new forecasting method—making use of presales activities of the new product and the correlations to predict the subsequent life cycle sales. The following figure gave an example of how such a correlation has been able to produce an accurate forecast. This process is exactly what Fisher and Raman has dubbed “accurate response.” I think there are many such opportunities for using big data (Figure 2).

4. The Scale Stage of Big Data

Back in 2011, Gartner has identified the three Vs of big data: Volume, Velocity, and Variety (Sicular 2013). Rozados and Tjahjono (2014) gave a detailed account of the types of data that constituted the 3Vs. There, they described that most of the current usage of big data had been centered on core transactional data such as simple transactions, demand forecasts, and logistics activities. Certainly, as machine-generated data, Internet-of-things sensing, and social media data become readily available, there is an opportunity for us to use extensive data collected from outside of your business or even your industry. In Figure 3, I reproduce the 3V figure of Rozados and Tjahjono (2014).

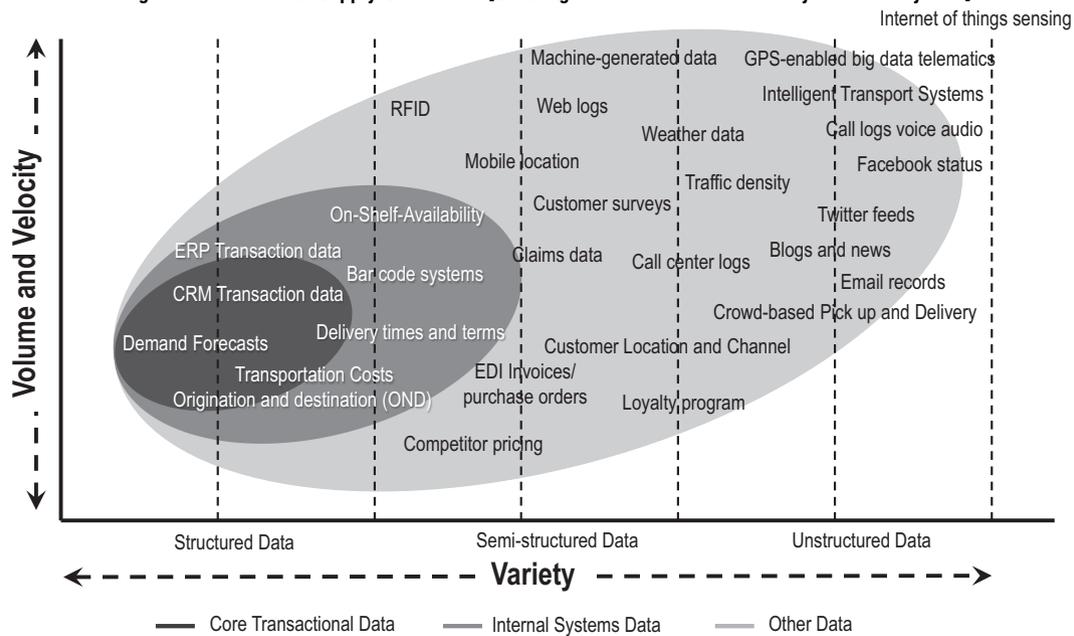
Of the three V's, Sicular (2013) claimed that the most interesting was variety, that is, data of different natures. Tableau (2017) wrote: “while all three Vs are growing, variety is becoming the single biggest driver of big data investments, as seen in the results of a recent survey by New Vantage Partners. This trend

Figure 2 Forecasting Improvement [Color figure can be viewed at wileyonlinelibrary.com]



Source: Data from Evant.

Figure 3 The 3Vs of Supply Chain Data [Color figure can be viewed at wileyonlinelibrary.com]



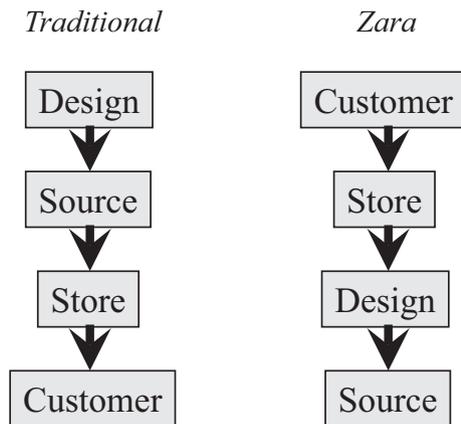
Source: Rozados and Tjahjono (2014).

will continue to grow as firms seek to integrate more sources and focus on the “long tail” of big data.”

The potential of using more data from sources not directly related to the product under focus, or even your own business, can be huge. DemandTec, a public price optimization software company which later has been acquired by IBM, had encountered a situation that serves as a simple illustration. They were analyzing data for D’Agostino, a New York-based supermarket chain. The bread department had been

using product sales data of all bread-related SKUs to make pricing and inventory decisions, while the meat department did the same with all meat products. By pooling data from the two seemingly unrelated product categories together, one could identify how pricing and inventory decisions of buns (from the bread department) and hot-dogs (from the meat department) should be made jointly, given the tight correlations in the demand elasticities of the two products.

Figure 4 Zara’s Supply Chain Model [Color figure can be viewed at wileyonlinelibrary.com]



Source: Zara.

Lazer et al. (2014) described the interesting challenge of predicting the arrival of the flu season, in which the Center of Disease Control usually used their scientific and statistically based method. Using big data, Google could also make use of how people searched for an extensive variety of words to make predictions. By combining the two methods, the prediction power could be improved even further.

Manenti (2017) gave the example of Transvoyant, which made use of one trillion events each day from sensors, satellites, radar, video cameras, and smart phones, coupled with machine learning, to produce highly accurate estimates of shipment arrival times. Such accurate estimates can help both shippers and shipping companies to be proactive with their operations, instead of being caught by surprise with either early or late arrivals of shipments. Similarly, Manenti (2017) reported the IBM Watson Supply Chain that used external data such as social media, newsfeeds, weather forecasts, and historical data to track and predict disruption and supplier evaluations.

5. The Structural Transformation Stage of Big Data

Ultimately, companies can make use of big data to re-engineer the business processes, leading to different paths of creating new products and serving customers, and eventually, potentially creating new business models.

One example is how the US Customs office could use big data to re-engineer its screening and border control process to improve supply chain security. Containers shipped from all over the world are currently undergoing sampling inspection by Customs at US ports for fear of illegal smuggling and terrorists’ use of the container to carry weapons of mass

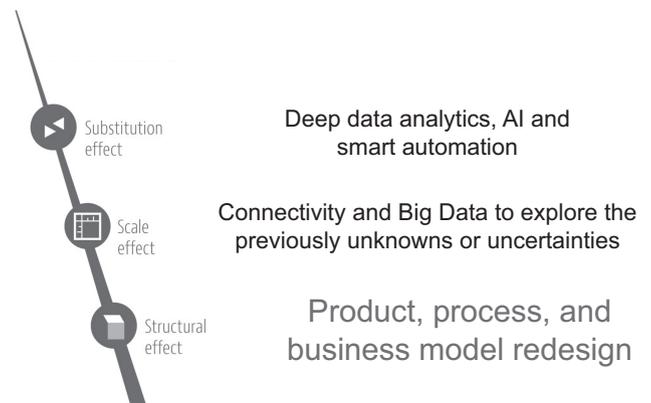
destruction. Inspection can lead to excessive delay and uncertainties in the lead time for shippers receiving such goods. When RFID was introduced, it was possible to create smart containers (containers armed with RFID-enabled electronic seals and possible sensors), and monitor a container after it was loaded and inspected at the port of origin, as well as in transit. The data captured during the journey could be used to determine if tightened or lessened inspection should be applied when going through customs. Thus, a new customs clearance process could be created. Pilots have shown that huge business values could be created (see Lee and Whang 2005). Today, the US Customs and Border Protection makes use of extensive data to change the customs clearance process—containers that pose a risk of terrorism are prescreened at the port of departure instead of arrival (US Customs and Border Protection Report 2011).

In 2008, a group of supply chain academics visited Zara in A Coruna, and I was struck by one of the pictures that the Zara executive showed us, reproduced here in Figure 4.

Zara’s supply chain model has basically restructured the supply chain process, and many case writers have described this in detail. The key is that customer information at the store formed the basis of input to their design, instead of professional designers using their sense of fashion trends and creativity. Such demand-sensing capability, coupled with quick response in production and distribution, form the foundation of Zara’s fast fashion success.

Here is an example of a company trying to develop the big data version of Zara. Li and Fung, the giant apparel sourcing and trading company, is building its digital supply chain, so that the new product generation process is no longer confined to the traditional one (Sourcing Journal, 2017). The company has tapped on-line platforms like Pinterest and a customized digital platform developed for its fashion

Figure 5 The 3-S Model of Big Data [Color figure can be viewed at wileyonlinelibrary.com]



business called WeDesign, to link designers across all of its verticals to create virtual samples for customers. Product designers will leverage data on fabric types, plastics, sensors, and most importantly, connectivity with customers. Real and direct customer needs are used to generate new products, identify winners, and then work with partners to produce the winners at scale.

Making use of data on items on web pages browsed by e-commerce shoppers, Sentient Technologies also created machine learning algorithms to do visual correlations of items, and delivered purchasing recommendations (Manenti 2017). Again, a new product generation process has been developed.

I believe there will be many more opportunities for big data to make similar disruptions to the supply chain processes of many industries.

6. Concluding Remarks

As indicated, I think the 3-S model of innovation can be a good framework for us to examine the impact of big data on operations and supply chain management. A summary of such impacts is shown in Figure 5.

Although I have described this model as a cycle, it does not mean that the innovations have to follow such a strict sequence. In fact, as we have seen, all three effects can happen at the same time. However, it is often the case that a new innovation requires many small-scale pilots to allow early users to gain familiarity as well as confidence, ascertaining the values that one can gain from the innovation. Such early usage had often been based on one particular business activity or one process of the supply chain. As time goes on, the scale and structural effects could make a much bigger impact. Hence, we must not forget the other two stages, and declare victory from just successful implementations at the substitution stage.

As researchers, we can also create methodologies to support, and conduct empirical validation for all three stages of the innovations. The good news is that data should be much more readily available for our research, with the advancement of big data.

References

- Blake-Plock, S. 2017. Where's the Value in Big Data. *Forbes*, April 14. Available at <https://www.forbes.com/sites/forbestechcouncil/2017/04/14/wheres-the-value-in-big-data/#78400ee430da> (accessed date April 14, 2017).
- Economist. 2003, February 6. The Best Thing since the Bar-Code.
- Economist. 2017, April 12. How Germany's Otto uses Artificial Intelligence.
- Fisher, M., A. Raman. 1996. Reducing the cost of demand uncertainty through accurate response to early sales. *Oper. Res.* **44** (1): 87–99.
- Jin, X., B. W. Wah, X. Cheng, Y. Wang. 2015. Significance and challenges of big data research. *Big Data Res.* **2**(2): 59–64.
- Lazer, D., R. Kennedy, G. King, A. V. Lee. 2014. The parable of Google Flu: Traps in big data analysis. *Science* **343**(14): 1203–1205.
- Lee, H. L. 2007. Peering through a glass darkly. *Int. Commer. Rev.* **7**(1): 61–68.
- Lee, H. L., O. Ozer. 2007. Unlocking the value of RFID. *Prod. Oper. Manag.* **16**(1): 40–64.
- Lee, H. L., S. Whang. 2005. Higher supply chain security at lower cost: Lessons from total quality management. *Int. J. Prod. Econ.* **96**: 289–300.
- Malone, T., J. Rockart. 1991. Computers, networks and the corporation. *Sci. Am.* **265**(3): 92–99.
- Manenti, P. 2017. Artificial intelligence and future supply chains. *SCM World Blog*, January 31. Available at <http://www.scmworld.com/artificial-intelligence-future-supply-chains/> (accessed date November 1, 2017).
- O'Marah, K., X. Chen. 2016. The Future of Supply Chain. *SCM World Report*.
- Rozados, I. V., B. Tjahjono. 2014. Big Data Analytics in Supply Chain Management: Trends and Related Research. 6th International Conference on Operations and Supply Chain Management, Bali. December.
- Sicular, S. 2013. Gartner's Big Data Definition Consists of Three Parts, Not to be Confused with Three 'V's. *Forbes* March 23.
- Sourcing Journal. 2017. Here's How Li & Fung Plans to Create the Supply Chain of the Future. March 30.
- Tableau. 2017. Top 10 Big Data Trends for 2017. Available at <https://www.tableau.com/learn/whitepapers/top-10-big-data-trends-2017?ref=lp&signin=34a8eb43de2cfc773726f6260c33bb5d®-delay=TRUE> (accessed date November 1, 2017).
- US Customs and Border Protection. 2011. Container Security Initiative In Summer. May.

Big Data and the Precision Medicine Revolution

Wallace J. Hopp*, Jun Li, Guihua Wang

Ross School of Business, University of Michigan, Ann Arbor, Michigan 48109, USA, whopp@umich.edu, junwli@umich.edu, guihuiw@umich.edu

The big data revolution is making vast amounts of information available in all sectors of the economy including health care. One important type of data that is particularly relevant to medicine is observational data from actual practice. In comparison to experimental data from clinical studies, observational data offers much larger sample sizes and much broader coverage of patient variables. Properly combining observational data with experimental data can facilitate precision medicine by enabling detection of heterogeneity in patient responses to treatments and tailoring of health care to the specific needs of individuals. However, because it is high-dimensional and uncontrolled, observational data presents unique methodological challenges. The modeling and analysis tools of the production and operations management field are well-suited to these challenges and hence POM scholars are critical to the realization of precision medicine with its many benefits to society.

Key words: big data; precision medicine; observational data; machine learning; causal inference

History: Received: April 2018; Accepted: April 2018 by Kalyan Singhal, after 1 revision.

1. Introduction

Two parallel revolutions are about to transform the health care industry in profound ways. The first is in the practice of medicine, where imprecise one-size-fits-all medicine is being replaced by individually tailored precision medicine. The second is the analysis of data, where big data techniques are making it possible to extract patterns from vast amounts of digital information to guide decision making in a wide range of sectors and applications. Together these revolutions will fundamentally alter the clinical practice of medicine, by changing the way patients are diagnosed, treatments are selected, and care is delivered. But they will also have ramifications for the health care industry far beyond the practice of medicine, including the way patients choose and interact with providers, the way providers make strategic and tactical decisions, and the way payers reimburse and incentivize both patients and providers.

The link between these two revolutions is data. The big data revolution is producing new ways to generate and analyze data and the precision medicine revolution is creating new ways to leverage it (see Kruse et al. 2016 for a broad review of the medical literature on using data in health care). In particular, big data methods are making it possible to go beyond experimental data generated by clinical studies and take advantage of observational data about patients in uncontrolled settings. Observational data can come from medical records, online reviews, mobile devices and many other sources. In contrast to experimental data, observational data is cheaper to obtain, has

much larger sample sizes and addresses a much wider range of variables. For example, less than 5% of adult cancer patients are part of a clinical trial (Unger et al. 2016). Furthermore, since only patients meeting certain criteria can enroll in trials, patients with uncommon tumor types, older patients, and patients with a poor performance status or comorbidities are frequently underrepresented. As a result, researchers are increasingly looking to observational data to study cancer treatments and to identify side effects not apparent in randomized clinical trials (Spigel 2010). By making virtually every act of health maintenance and patient treatment available for study, big data approaches are opening the possibility of true precision medicine that enables patients to receive care best matched to their specific health condition, individual characteristics and personal preferences.

However, although observational data has enormous potential, it presents problems not present with experimental data. In clinical experiments, subjects are selected to minimize bias and are organized into treatment and control groups, which make for crisp statistical conclusions. In contrast, observational data comes from actual patient experiences and can therefore suffer from bias, censoring, insufficient sample size and other problems. Therefore, in order to use observational data to guide health care decisions we must find ways to correct for the inherent flaws in the data. This is where the emerging tools of analytics and machine learning come in, and where the POM field can play an important role in bringing about the merger of big data and precision medicine to revolutionize health care.

To achieve its full potential, precision medicine must address all of these levels of health care decisions. It should help patients make good decisions about what type of care to seek. For example, a diabetic patient should be able to get information about what symptoms indicate he/she should seek diet advice on Dlife.com and what symptoms indicate a need for prompt medical intervention. Precision medicine should also help patients (and providers) choose the right provider for a particular patient. For example, a patient in need of heart surgery, should have access to rankings or outcome data to help him/her (perhaps with help from his/her cardiologist) choose a surgeon. Finally, precision medicine should help all of the actors involved in implementing a care path make the best decisions about tests, procedures, dosages, and other aspects of the health care process.

Although “precision medicine” has emerged as the most popular term for individually tailored health care, there are other terms in use. The older “personalized medicine” is regarded by many as interchangeable with personalized medicine, but by some as a literal call for individualized medicine (see National Research Council 2011). Since, in some cases, the most effective treatment may not be individually tailored, these people argue (without excuse for the pun) that “precision medicine” is the more precise term. “Pharmacogenomics” is a specific instance of precision medicine, which makes use of genetic information to tailor pharmacological treatments to patients. Finally, an overlapping term is “evidence-based medicine” which refers to the use of relevant data to guide medical decisions. However, while precision medicine is always evidence-based, not all evidence-based medicine is personalized. As we describe in Wang et al. (2018), many research studies focus on the response of an average patient to a particular treatment, and therefore imply “one-size-fits-all” protocols.

3. The Role of Big Data in Precision Medicine

Neither one-size-fits-all medicine, nor precision medicine, nor the stages in between them, are uniquely defined entities. Instead they represent a continuum of practices whose outcomes improve with the amount and type of data they leverage to guide decisions.

The continuum of one-size-fits-all medicine ranges from a pure trial-and-error process, in which the patient (or physician) selects randomly from a set of treatments in hopes of achieving an improvement, to evidence-based priority rules, in which the patient (physician) tries alternatives in decreasing order of their probability of success. By using success rate data to rank order the alternatives, we can increase the likelihood and speed of achieving a good outcome.

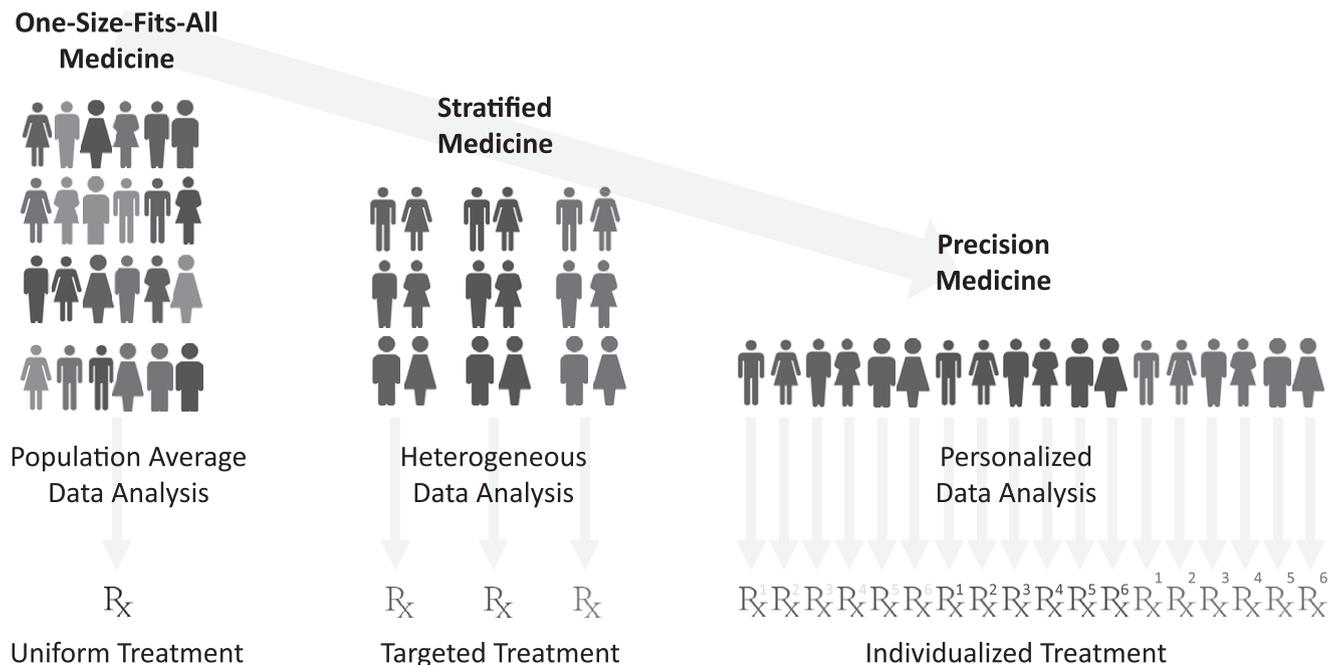
More accurate success rate data will produce a more efficient search process. However, regardless of whether success rates are raw or risk adjusted,¹ this type of data will lead to the same rank ordering of alternatives for all patients, and hence one-size-fits-all medicine. That is, all patients will start with the top-ranked option and work their way down the list in the same order. When patient responses to the various treatments are heterogeneous, this ensures inefficiency in the process, no matter how refined the success rate data.

Precision medicine can also involve trial-and-error, but the treatment alternatives and/or likelihoods of success are tailored to the individual patient. The most basic form of precision medicine would be a treatment that is determined by a single individual characteristic (e.g., a dosage based on the precise weight of the patient). As precision medicine takes into account more patient descriptors, such as comorbidities, genetic markers, preferences, lifestyle attributes, etc., the alternative set and outcome likelihoods can be refined. One reason pharmacogenomics has received so much attention within the precision medicine literature is that scientists expect many new indicators of patient responses to drugs to be discovered at the genetic level.

Between one-size-fits-all and precision medicine is stratified medicine, which divides patients into groups according to their response to a therapy. Since there may be other differences between patients in the same group, treatments will be rank ordered according to in-group success rates. Hence, we can think of this as “one-size-fits-some” medicine. An example of stratification is the classification of breast cancers into estrogen receptor (ER) positive and negative categories. ER positive cancers respond to hormone therapies, while ER negative cancers do not. Therefore, women in the two categories receive different chemotherapy protocols. Stratified medicine is also a continuum of practices because additional characteristics can be used to refine patient groupings. For example, breast cancers can also be classified as progesterone receptor (PR) positive or negative and as human epidermal growth factor receptor 2 (HER2) positive or negative. Because patients in these categories respond differently to drugs, chemotherapy protocols can be better targeted by using the PR and HER2 biomarkers. When indicators get specific enough to divide patients into groups of one, stratified medicine becomes precision medicine.

Figure 2 depicts the continuum from one-size-fits-all to precision medicine. For our purposes of linking the precision medicine revolution to the big data revolution, the most important aspect of this figure is the nature of the data analysis required to facilitate each type of medicine.

Figure 2 Progression toward Precision Medicine



One-size-fits-all medicine requires *population average effect analysis*. The simplest form of such analysis is the frequentist approach, which uses outcomes to calculate the fraction of the population for which each alternative is effective. But when health outcomes cannot be measured in binary success/failure terms (e.g., liver transplant patients are evaluated in graft survival time, a continuous metric), a more sophisticated approach is required. In these circumstances, a risk-adjusted outcome analysis can be used to rank alternatives. For example, to compare the effectiveness of different hospitals in treating a particular cardiovascular disease, we can evaluate each hospital by dividing the outcome (e.g., 5-year survival rate) by the expected outcome (i.e., the average 5-year survival rate that would be achieved by the full population of hospitals if they treated the same mix of patients as the individual hospital). The expected outcome requires a regression across all patients and hospitals to compute (see Glance et al. 2006). The resulting observed-to-expected, or O/E, ratio provides a single number metric with which to rank providers. However, a single rank ordering based on O/E ratio is only appropriate when the relative effectiveness of the providers does not depend on patient characteristics. When it does, rank order may be incorrect for some, or even all, patients.

Stratified medicine uses *heterogeneous effect analysis* to correct for the failure of average effect analysis to account for patient differences. This requires patient

characteristics along with outcomes. The more characteristics we can observe (e.g., age, gender, comorbidities, biomarkers, readings from wearable devices, etc.), the more likely we are to find characteristics that differentiate patients with regard to their responses. However, the more finely we stratify patients, the smaller our sample sizes become. In clinical studies, where data collection is expensive and difficult, it is often impossible to achieve large enough samples to permit subgroup analyses needed to detect heterogeneous patient responses.² As we discuss below, finding the right balance between difference distinction and statistical significance is a key analytical challenge in implementing stratified medicine.

Precision medicine uses *personalized effect analysis*, which often requires personalized data. Genome sequencing data is one form of personalized data. But so are preference (e.g., risk aversion) and lifestyle (e.g., diet and sleep) data. One could argue that all medicine makes use of at least some personalized data in the form of patient responses to provider inquiries (e.g., “How do you feel about the proposed course of action?”). But precision medicine is aimed at a much more objective, evidence-based incorporation of personalized data than these traditional subjective assessments. Some of these data are being collected through traditional channels (e.g., doctor appointments) and compiled in electronic medical records (EMRs) or electronic health records (EHRs). But more detailed personalized data will require new approaches, such as mining social media or using handheld devices.

A characteristic common to the data used for above analyses is multidimensionality. Because treatment alternatives are generally evaluated along multiple criteria, such as expected clinical outcome, risk of side effects, comfort and convenience of care, cost, and others, we need data with multiple dependent variables, as well as multiple independent variables, to support decision making. We can use big data to estimate the heterogeneity of patient responses along each dimension. But this will not be enough to facilitate patient treatment choices. We will also need some way to combine the results for the various dimensions. To do this at the personalized level, we will need individual preference data (in the form of linear weights or parameters for a nonlinear utility function). Furthermore, preferences on non-medical dimensions, such as travel distance, expense, familiarity, support services for family members, and many others, may be relevant to individual decisions. By combining heterogeneous outcome data and individual patient preference data we can achieve true precision medicine, in which choices are customized to the level of the individual patient.

Ultimately, the progression from one-size-fits-all to precision medicine will be driven by more and better sources of data, and the analytics techniques needed to use them to guide health care decisions. But collecting and analyzing these data present a wide array of challenges, as we describe below.

4. Challenges in Collecting and Analyzing Big Data

Precision medicine can make use of both experimental and observational data. But, although challenges remain in the collection and use of experimental data to support health care decisions, these are familiar problems clinical researchers have been dealing with for years. The new challenges presented by the incorporation of big data into health care decision making, and the ones most amenable to the skills of the POM community, are those associated with observational data. In this section, we first describe collection issues associated with this increasingly large-scale and personalized type of data. We then discuss estimation issues related to high dimensionality, bias, censoring and sample size issues that are common in uncontrolled observational data. Next, we highlight some modeling and optimization challenges that must be addressed to translate observational data into decision aids. Finally, because the ultimate goal is to enable all stakeholders in the health care system to make better decisions, we describe implementation challenges involved in presenting and disseminating information derived from big data analyses.

4.1. Data Collection

Health care observational data comes from many sources and in many formats. In this subsection, we describe common challenges that arise in the process of collecting large-scale observational data. These include technological barriers, incomplete or inaccurate data, privacy concerns and scientific challenges.

4.1.1. Technological Barriers. Advances in Information Technology (IT) have transformed health care data collection—from hand written notes, electronic health records, personal wearable devices, and many other sources. These advances also made it possible to capture both structured (e.g., administrative and claims data) and unstructured data (e.g., texts, images, audios, videos). While structured data are straightforward to analyze, and some text can be analyzed through textual analysis (see e.g., Wu 2016), technology for converting medical images, audios and videos to textual or structured data that can be easily understood by non-medical researchers is still lacking. As a result, for example, echocardiograms of mitral patients are routinely interpreted by a cardiologist or surgeon but are rarely available for use by a data analyst or policy maker.

There is a lack of coordination of health IT implementation across different organizations or even in different departments of the same organization. Variations in data collection platforms and storage methods make it difficult to convert data from different sources into the same format for analysis. The lack of coordination across different organizations is a major barrier to health information exchange when a patient switches to a different health care provider or payer. For example, when a patient changes insurance company, a new enrollee ID is usually created. Duplicate IDs make it hard to link data from multiple insurance companies to the same patient.

IT companies and researchers are continually seeking ways to translate complex and unstructured forms of information into usable data and hospitals and their IT providers are constantly working on better integration and interoperability. But both of these have a great deal of room for improvement. As a result, researchers seeking to use observational data to refine precision medicine must work with less-than-ideal data sets and are likely to do so for some time. But new research opportunities will continue to arise as these technological challenges are addressed.

4.1.2. Incomplete or Inaccurate Data. Missing values are common in observational data. These can occur due to unavailability of information (e.g., because a patient refused to disclose personal information or tests were not done), omissions by data administrators, accidental deletion, or data loss

during merge and conversion. Missing values can be categorized broadly into three types: missing completely at random (i.e., independent of both the outcome variable and other independent variables), missing at random (i.e., independent of either the outcome variable or other independent variables), and missing not at random (i.e., correlated with the outcome variable even after controlling for other independent variables). As we will discuss later, missing at random or completely at random may not be an important issue but missing not at random could cause biases in estimation. Techniques for dealing with missing value include imputation (i.e., replacing missing values with substituted values), deletion, interpolation and full analysis (see Little and Rubin 2014 for more details).

Worse than missing values are inaccurate values. Inaccurate values can be the result of failure to understand terminology, over-reliance on manual input, inconsistent coding practice, or change of documentation policies. Compared with missing values, inaccurate values are much more difficult to detect, especially when they are not obviously outliers. One approach to identifying inaccurate values is to compare different data (e.g., claims, clinical and administrative data) for the same information. Another is to verify the raw data used to calculate the values. For example, one might check a patient's height and weight if the patient's body mass index has a suspicious value.

4.1.3. Privacy Concerns. Privacy is becoming an increasingly serious concern for patients as individual sensors have made it possible to monitor health at an increasingly detailed level. Users may be unwilling to share their personalized data with the providers of these devices or their clients. This could hinder prospects for collecting vast amounts of personalized data to be used in precision medicine as well as developing new and more accurate devices for health monitoring. Encryption and de-identification are traditional approaches to addressing privacy concerns. Medicalchain is a recently developed blockchain approach to protecting patient information such as those from EMRs while sharing data across medical providers and treatment sites (Azaria et al. 2016).

Privacy is a concern to health care providers and payers, as well as to patients. Providers may be unwilling to share clinical data because there are laws and regulations that protect the privacy of patient information. They are also concerned that sharing a best practice with other providers may erode their competitive advantage in attracting patients. Payers may be unwilling to share their claims data because the data usually contain sensitive information about their cost structure and financial performance.

Finally, providers and payers may be unwilling to share information with each other, because they do not want to lose negotiating power. In theory, encryption and de-identification can be used to protect provider and payer data. But doing this while enabling integration of data across participants and platforms is difficult. Data science scholars (see, e.g., Li and Qin 2017, Miller and Tucker 2017) have proposed to use encryption and other means for protecting patient privacy when medical records are shared.

4.1.4. Scientific Challenges. Pharmacogenomics makes use of genetic information to predict patient responses to medication. Recent advances in genetic technology offer the prospect of being able to personalize medications to patients based on genetic tests. However, while an increasing number of biomarkers have been discovered, their influence on outcomes is complex. As a result, few biomarkers are currently used in clinical decision making or drug development and their ultimate utility is still a matter of debate.

Several scientific challenges make it difficult to collect useful and accurate pharmacogenomics data. First, there are potentially many yet-to-be-identified genes and biomarkers that affect the pharmacology of a drug. Second, even for the identified pharmacogenomics biomarkers, some of the clinical results regarding the genetic association between the biomarker and the pharmacology of a drug have been inconsistent (Lam 2013). Third, in addition to inherited genetic factors, many environmental, lifestyle and other factors could have important and complicated impacts on a patient's response to a drug. Researchers will continue to use clinical studies to link genetic and biomarker information to patient responses. As such information becomes part of patient records, it will also become available for use as observational data in uncontrolled studies.

4.2. Estimation

Analyzing large-scale observational data presents several estimation issues, which can be grouped into those related to "high dimensionality" and those associated with the "uncontrolled" nature of the data. Further complicating analysis is that treatments may have heterogeneous effects on different patients, and that this heterogeneity may differ among treatments. In this subsection, we first discuss issues related to high dimensionality, and then issues caused by the uncontrolled nature of the data. Finally, we discuss the challenges of addressing these two types of issues simultaneously in treatment effect analysis. We shall note that the issues and potential solutions we point out below are by no means exhaustive. While they represent some key challenges, the area offers a great potential for many more research opportunities.

4.2.1. High Dimensionality-Related Issues. A direct challenge of working with a large number of variables is that it is unclear which variables are important and how they affect the response variable. Finding the model, which could include non-linear terms and interactions between variables, with the highest explanatory power is both computationally complex and statistically challenging. The most important statistical challenges are the issues of over-fitting and multi-collinearity, and the problem of multiple testing.

Over-fitting is a common problem in statistical analysis of high-dimensional data, especially when the number of observations is small relative to the number of variables. To illustrate this point, consider a simple example with two observations and one variable. It is possible to perfectly fit the data with a straight-line model that has R^2 equal to one and mean-squared error equal to zero. As such, it has perfect in-sample performance, but will probably have very poor out-sample performance. If someone neglects the issue of over-fitting, he/she might erroneously pick such a model for prediction. Coefficient estimates and p -values in over-fitted models are misleading as well, as they often indicate that the effects of almost all variables are not statistically significant.

Multi-collinearity refers to a situation where some independent variables are highly correlated. An increase in the number of variables increases the probability that some variables are multi-collinear. This is partly because some variables in observational data, though labeled differently, describe similar or closely related attributes (e.g., height and weight) or the same attribute measured at different times (e.g., blood pressure measured throughout the hospital stay). When some variables suffer from multi-collinearity, the estimated effects of these variables will likely have large standard errors, which makes it appear that none of them is statistically relevant.

A systematic approach to addressing both over-fitting and multi-collinearity issues is stepwise variable selection. Forward stepwise selection starts with a null model (i.e., no variable) and recursively adds the remaining variable that has the highest predictive power as measured by R^2 or residual sum of square. Inverse to this, backward stepwise selection starts with the full model (i.e., with all variables) and recursively removes one existing variable that has the lowest predictive power. These stepwise selection processes result in a series of candidate models, whose performances are then compared for model selection. Note, however, this sequential variable selection is often order dependent and computationally inefficient, especially when the number of candidate variables is large.

A more sophisticated approach to addressing over-fitting and multi-collinearity is regularization, which

introduces a regularization factor to penalize the number of independent variables that enter the model. For example, the least absolute shrinkage and selection operator (LASSO) is a widely used regularization method that penalizes the number of variables in a regression model and shrinks the coefficients of insignificant or small-impact variables toward zero (Bastani and Bayati 2018, Bertsimas et al. 2016, Friedman et al. 2001). Under mild conditions, LASSO generates sparse models that identify the truly relevant variables. In contrast with stepwise variable selection, regularization methods can be applied even when the number of variables is larger than the number of observations. However, while methodologically intuitive, regularization methods raise several new questions, including: What is an appropriate amount of regularization? How to compare and select from different models that have similar explanatory powers?

These questions can be answered with cross validation, a technique commonly used in the machine learning literature to compare the performance of different models (Friedman et al. 2001). It uses one set of data (called the “training sample” or “in-sample”) to train a model and an independent set of data (called the “testing sample” or “out-sample”) to test the model performance, which is usually measured by mean-squared error (MSE) and coverage rate. Cross validation selects a model with the right number and type of variables that best predict outcomes in the testing sample. However, because (i) different training and testing samples might result in different models and (ii) there is a potential correlation between training and testing samples due to randomness, the model selection process is usually undertaken using multiple training and testing samples. A popular approach is k -fold cross validation, where the data is randomly divided into k parts. Each time, one of the k parts is held-out for testing and the remaining $k - 1$ parts are used for training. Out-sample MSE of a model is calculated as the average of the k mean-squared errors.

Besides cross validation, researchers have frequently used Akaike information criterion (AIC) and the Bayesian information criterion (BIC) for model selection as well. Both AIC and BIC balance the goodness-of-fit of the model and its complexity, while BIC penalizes complexity more strongly than AIC and thus leads to sparser models. If the objective is to make better predictions, AIC and cross validations are commonly used. If the objective is to select variables that are truly economically relevant, BIC is typically preferred (Wang et al. 2007).

Finally, multiple testing arises when one uses the same data to test a number of hypotheses. This issue is particularly common in health care settings, because multiple treatments are usually evaluated

across multiple metrics for multiple patient groups at multiple providers. For purposes of illustration, suppose the treatment and control groups are patients who received mitral valve repair and replacement, respectively. To guide future treatment decisions, we want to know whether mitral valve repair is superior to replacement for some patients and whether the outcome differences between providers are heterogeneous across different patients. We could partition patients into groups based on their gender and compare repair with replacement as well as different providers for each group in terms of mortality, complication, readmission, graft failure, etc. The more quality metrics used for comparison, the more likely that the treatment and control groups differ on at least one quality metric due to random errors. Similarly, the more patient groups and providers we consider, the more likely that mitral valve repair will be found to be superior for at least one patient group at one provider.

To address the issue of multiple testing, one needs to adjust the significance levels used for rejecting a null hypothesis. Bonferroni and Benjamini–Hochberg are two commonly used approaches partly due to their simplicity (see Bonferroni 1936, Benjamini and Hochberg 1995 for details). Note these tests rely on the assumption that different hypotheses are independent of each other, which might not be true in health care settings. For example, a patient who is more likely to have a complication is also more likely to have a readmission. A better and more complicated approach to finding adjusted significance levels is simulation. The null hypothesis is that the treatment does not have any effect on any observation for any metric. We can then randomly permute the treatment dummies among observations and test all hypotheses to obtain a set of p values. After that, we repeat the permutation N times and count the number of times at least one test is deemed significant based on the target significance level α . The adjusted significance level can be calculated as α/N .

4.2.2. Uncontrolled Data-Related Issues. The uncontrolled nature of observational data introduces a number of challenges beyond those presented by the high-dimensionality of the data. In randomized controlled experiments, because observations are randomly assigned to treatment and control groups, we can control the number of observations in each group and the treatment effect can be estimated using the average outcome difference between the treatment and control groups. In observational studies, because we cannot randomly assign treatments to individuals, the treatment group may have systematically different attributes than the control group, which may result in estimation biases. The sample size of the treatment

may also be insufficient to power the analysis, because unlike randomized experiments, the sample size of the treatment group in an observational study is not chosen by researchers for detecting a given effect at the desired level of significance. In addition to bias and sample size issues, censoring is common in health care observational data, because an event (e.g., mortality) may occur beyond the data collection period. Below we first review a list of common sources of biases, which include omitted variable, sample selection and missing value, and then discuss issues of sample size and censored data.

Omitted variable bias occurs because observational data may not include all variables of interest. While it is safe to omit a variable that is pure noise, omitting a variable that directly affects the outcome not only reduces the prediction accuracy of a model, but also creates biases in estimating the coefficients of all other variables that correlate with the omitted variable. For example, health-conscious patients may be more likely to seek treatments at a Center of Excellence (CoE). At the same time, these patients are healthier due to their lifestyles. However, we may not be able to observe from observational data whether a patient is health-conscious or not. Simply ignoring this unobservable variable will bias the estimate of the quality gap between CoEs and non-CoEs.

Sample selection bias occurs when patients described by observational data are not a representative sample of the population of interest. For example, some transplant centers decide whether to admit a patient to their waiting lists based on the acuity of the patient. A very sick patient may be denied access to waiting lists. Therefore, comparing centers based on samples of admitted patients can lead to biased conclusions, especially when different centers have different patient admission practices. Similarly, selection bias is the primary threat to using observational data in comparative treatment analysis. For example, early-stage cancer patients with better prognoses are more likely to receive milder forms of chemotherapy with fewer side effects, while advanced-stage cancer patients with poorer prognoses are more likely to receive stronger drugs with harsher side effects. Hence, both outcomes and side effects are strongly influenced by the mix of patients treated in a center. Giordano et al. (2008) demonstrated how selection bias can lead to different or even opposite conclusions to those from randomized clinical trials in cancer research.

Missing value bias is common in observational data that are collected over a long period of time and from different sources (e.g., multiple hospitals). If only a small portion of the values are missing, and the missing values happened completely at random, we can safely delete these observations without biasing the

results. However, if values are missing in a systematic way or for a specific group of observations (e.g., patients with certain comorbidities), simply ignoring the observations will create estimation biases.

Popular techniques for addressing these biases include instrumental variable, panel data, difference-in-differences, regression discontinuity, selection models, and others. The instrumental variable (IV) method explores the variation in an exogenous variable that correlates with the endogenous variable but does not directly correlate with the outcome variable. The IV approach extracts and uses only the exogenous variation in the variable of interest in the estimation, and obtains an unbiased estimate of its impact (see e.g., Bartel et al. 2016, Chan et al. 2016, Freeman et al. 2016, Ho et al. 2000, KC and Terwiesch 2011, 2012, Kim et al. 2014, Lee et al. 2017, Lu and Lu 2017, McClellan et al. 1994, Xu et al. 2018). For example, Hadley et al. (2010) showed that a properly chosen instrumental variable can correct the potential selection bias of observational data and provide consistent results as in randomized trials in the context of prostate cancer. Panel data models exploit longitudinal variations to take out unobserved time-invariant individual fixed effects (see e.g., Clark and Huckman 2012, Clark et al. 2013, KC 2013). The difference-in-differences method relies on the assumption that the treatment and control groups would have had parallel trends if there had been no treatment (see e.g., Song et al. 2015). It uses group fixed effects to capture time-invariant differences between the two groups and time fixed effects to capture the trends. Regression discontinuity exploits sharp changes in the variable of interest and compares outcomes of observations lying slightly above and below the threshold (see e.g., KC 2018). Other techniques involve direct modeling of the bias generating process. For example, the Heckman two-stage selection model (see Heckman 1976, 1977 for more details) corrects for biases caused by sample selection or non-random missing data. Which technique is most appropriate depends on the context and the availability of exogenous variables.

Even with big observational data, insufficient sample size may still be an issue, particularly as the stratifications in stratified medicine become finer. In experimental studies, researchers use power analysis to determine the sample size required detecting a certain treatment effect at a given significance level, sample size is a critical part of the experiment design. However, in observational studies, because researchers have no control over the number of patients who receive a given treatment, the treatment group may have a very small number of observations, which creates estimation difficulties. For example, suppose we want to compare the mortality rates of two hospitals

for a particular surgical procedure. One hospital treated 200 patients with 4 deaths and the other hospital treated 10 patients with no deaths. Which hospital is better? Obviously, the second hospital has a lower mortality rate, but we cannot conclude that it is better because the small sample size makes it likely that this rate was largely due to randomness.

The sample size issue described above is usually difficult to address without borrowing additional information from other sources. If we can construct a prior distribution regarding the outcome based on outside information, an empirical Bayes method can be used to estimate the effect of a treatment or compare multiple treatments. In the surgical example described above, such a prior could be the overall distribution of mortality rates across all hospitals, the relationship between a hospital's volume and its outcome, or the correlations between outcomes of similar procedures. Dimick et al. (2009) applied the empirical Bayes method that first estimates the effects of volume on outcomes from the data and then uses updates based on actual performance to predict performance of a specific hospital. Similar ideas might be suited to analyzing the effectiveness of more advanced chemotherapy treatment for patients of uncommon cancer types with a given chromosome gain or loss (e.g., a Wilms tumor patient with 16q chromosome loss). For example, one could construct a prior using data from patients of similar cancer types with the same chromosome change (e.g., all other kidney cancer patients with 16q loss) and then use the Bayes method to update this prior specifically for Wilms tumor patients.

Finally, censoring refers to the coding of a continuous variable into one number when the variable is larger or smaller than a specific value. There are three types of censoring: left censoring, interval censoring and right censoring, which correspond to different positions of the censored value. Right censoring is the most common in survival analysis, as an event of interest might happen beyond the study period. For example, suppose we use survival time as a quality metric to compare different hospitals for liver transplant surgery. The study period is all the years when data were collected, and the variable of interest is length of survival after transplant. However, we can only observe how long a patient survived only if the patient died before the last date of data collection.

One approach for dealing with censored data is to convert the variable of interest into binary indicators. For example, instead of analyzing how long a patient survived, we could analyze whether a patient survived more than 1 year after the transplant. However, the choice of the cut-off points may be arbitrary (e.g., 1 month vs. 1 year). Moreover, it usually fails to capture the long-term effect of the treatment. Another

approach is survival analysis, which explicitly models survival time as a continuous outcome variable. Parametric survival analysis assumes that the underlying distribution of survival time follows a certain known distribution (e.g., exponential) whereas nonparametric survival analysis such as the Kaplan–Meier method (see Wooldridge 2010 for more details) focuses on calculating and graphing survival probability as a function of time. A popular method for survival analysis is the Cox proportional hazards model, which is a semiparametric model that can be used to compare the survival time of two or more patient groups of interest.

4.2.3. Issues Related to Treatment Effect Analysis. To use big observational data to inform health care decisions we must simultaneously address issues related to both high-dimensionality and uncontrolled data in the context of treatment effect analysis. Estimating treatment effects is a different objective from predicting outcomes. Even though a more precise estimate of the treatment effect will improve the accuracy of outcome prediction, the model with the most predictive power does not necessarily capture the true treatment effect unbiasedly.

Classic machine learning methods referenced above are proven to be effective for predicting outcomes while addressing the high-dimensionality challenge. By themselves, these techniques are helpful in answering questions such as: How long will a patient with newly diagnosed stage III breast cancer survive? What is the probability that a patient with hypertension (pre-existing) may experience a major complication after a mitral valve surgery? Outcome prediction is particularly useful if a patient has already decided which health care path to follow or which treatment to receive. However, it is not directly useful for a patient who is trying to choose among alternative treatments, unless the causal effects of the treatments have been properly estimated. For example, a patient may ask questions like: How is my quality of life likely to differ if I receive a kidney transplant instead of continuing renal dialysis? What are the relative risks of complications if I choose to get a stent instead of a bypass graft?

While the effect of a treatment can be calculated by comparing predicted outcomes with and without the treatment, the results may be statistically misleading for two reasons. First, the best model for outcome prediction may not be the best for treatment effect estimation. For example, if age is an important factor affecting survival and race is an important factor affecting the effectiveness of a treatment, a model focusing on outcome prediction will include age as a key predictor whereas a model focusing on treatment effect estimation will instead include race as a key

predictor. Second, there may be endogeneity issues, which can bias the estimate of treatment effects. Because patients are not randomly assigned to observations, the treatment and control groups may have systematic differences (such as those cause treatment selections) that affect their outcomes. As a result, the effect estimated from a simple subtraction includes not only the true treatment effect but also the systematic difference between the treatment and control groups.

To modify classic machine learning methods for treatment effect analysis, we need to address two main issues. First, because we do not directly observe the treatment effect from data, we cannot use them as a dependent variable to train a model. If we instead use outcomes to train a model, an important variable that affects treatment effect but not outcome may be excluded during variable selection, whereas a less important variable that affects outcome but not treatment effect may be included in the final model. Second, because we do not know the treatment effects for observations in the testing sample, we cannot analyze the performance of a model by calculating its mean squared error or coverage rate or use cross validation for model selection. If we use outcomes in cross validation, we are likely to choose a model with the best outcome instead of the best treatment effect prediction.

One approach to addressing these issues is to use different penalization factors in LASSO to differentiate variables that affect outcomes from those that affect treatment effects (Imai and Ratkovic 2013). This approach allows for the possibility that some variables have a relatively small impact on outcomes but a large impact on treatment effects. This approach, however, requires institutional knowledge to distinguish between the two types of variables. An alternative approach first transforms the original outcomes and then applies the standard LASSO with the transformed outcomes and original independent variables for treatment effect analyses (Signorovitch 2007).

These approaches identify variables that significantly affect treatment effects when there are no pressing concerns of endogeneity. To address potential endogeneity issues associated with observational data, we need to integrate econometric methods such as the instrumental variable into machine learning models to obtain unbiased estimates of the treatment effects. For example, Li et al. (2017) combined LASSO and the instrumental variable technique to identify price competition in high-dimensional space. Approaches like this offer the best of both worlds by using machine learning to deal with big data and econometrics to establish causality.

Existing studies that analyze the causal effect of a treatment have focused on the average treatment effect. These studies implicitly assume that a

treatment has the same effect for all observations. However, it is possible that the same treatment has a positive effect on some patients but a negative or no effect on others. In cases when a treatment has a positive (or negative) effect on all patients, it is possible that the magnitude of the effect differs by patient. Recognizing the differences in patients' responses to the same treatment, researchers have called for stratified medicine that identifies patient groups with heterogeneous treatment effects.

One approach to identifying patient groups is to first interact the treatment dummy with patient characteristics and then include both the interaction terms and other variables in a regression model such as LASSO for variable selection. The main difficulty with this approach is that the number of interaction terms increases exponentially with the number of variables. As a result, this approach is computationally expensive and requires a large number of instruments (or other econometric techniques) to obtain unbiased estimates of all interaction terms.

An alternative approach is to partition observations into groups for subgroup analyses. The first challenge with this approach is that the number of ways to partition patients increases exponentially with the number of patient characteristics. It is not clear how many patient groups we should have and which characteristics should be used for partitioning. The second challenge is that there is a clear tradeoff between relevance and reliability. A finer partitioning of patients provides more relevant information but may not have enough statistical power due to a reduced sample size. A coarser partitioning of patients will have a larger sample size, but the information provided may not be as relevant.

These challenges can be addressed by tree-based methods, which are data-driven automatic processes that partition observations into groups such that observations in the same group have similar treatment effects and those across different groups have different treatment effects. A tree method usually starts at the root where all observations are in the same group (called "parent" node) and recursively partitions observations into "child" nodes using the variable that increases in-sample goodness of fit the most. It then treats each child node as a parent node and continues the partitioning until a stopping criterion is reached. Finally, it uses cross validation to select the tree that has the best out-sample performance.

While it is straightforward to calculate the treatment effect for a group in randomized controlled trials using average outcome differences between the treatment and control groups (Athey and Imbens 2016), we need to address potential endogeneity issues with observational data and use the

instrumental variable or another econometric method to obtain unbiased estimates of treatment effects. Furthermore, most existing tree-based methods are myopic in partitioning observations and may not achieve the best overall prediction accuracy. Bertsimas and Dunn (2017) proposed an optimal tree to address this issue, but this approach is computationally expensive if the dimension of the problem is high.

Finally, the main challenge of using big observational data to inform medical decisions at the individual patient level is that we cannot observe both the treatment and no-treatment conditions for a patient and no two patients are exactly alike. Even if we are able to perfectly match two patients based on observable characteristics, there is no guarantee that the two patients have the same unobservable characteristics. To address this main challenge and move practice closer to precision medicine, we can look to the inevitable increase in the amount of health care data and rise of real-time monitoring through use of wearable devices. In many health care settings, it is reasonable to assume that only a finite number of variables affect treatment effects, which means an increase in the number of observations will enable us to partition patients into finer groups without losing statistical power. Real-time monitoring by wearable devices allows us to collect observations for the same patient on a daily or even hourly basis. If a treatment is assigned to the same patient at random times or based on observable variables, these observations constitute trials for the same patient (see e.g., Klasnja et al. 2015). For example, a wearable device might remind the wearer to take deep breaths as an anxiety reducing therapy (see e.g., Sarker et al. 2017). With some randomization of the reminders and measurement of the physical responses, the device could optimize a breathing strategy for a given individual (see e.g., Walton et al. 2018). Finally, if the previously mentioned privacy concerns can be addressed, data from such devices could be pooled and used to determine the individual characteristics that make people most and least receptive to breathing therapy.

In summary, heterogeneous treatment analysis with high-dimensional observational data offers a wealth of research opportunities. New research in this area is emerging (see, e.g., Athey et al. 2017, Boruvka et al. 2017, Wang et al. 2017a,b) and has the potential to sharpen precision medicine protocols for a vast range of patients.

4.3. Modeling and Optimization

Once we have used big data to estimate patient responses to various health care alternatives, the problem becomes how to use the results to facilitate better decisions across the health care system. These decisions include choice of type of health care and

specific providers by patients, selection of specific treatments by patients and providers, strategic positioning and process improvement decisions by providers, reimbursement and incentive structure decisions by payers, and many others. We list some important decisions that present challenges amenable to the modeling and optimization skills of OM scholars below.

4.3.1. Individual Patient Level. In decentralized health care systems where patients can choose the type of treatment and the specific providers for the treatment, observational data can be used to understand how patients make choices and identify barriers that prevent patients from finding health care most suitable for them. When choosing a health care provider, patients consider not only outcome differences between providers but also travel distance, waiting time, insurance co-pay, etc. Patient choice models help understand not only the relative weights patients place on different factors but also how much outcomes would improve if one or more barriers are removed. For example, Wang et al. (2015) studied the impact of quality information, travel distance and insurance on patients' choice of cardiac surgeons for mitral valve surgeries and found that lack of quality information is the most important barrier preventing mitral patients from choosing the best care and reducing this barrier could improve mitral valve repair rate by 13%.

Although choice models using observational data can help understand what patients *are* doing, analytical models using OM techniques such as optimization are needed to tell patients and their providers what they *should* do. Compared with traditional models that rely on wide range of assumptions, data-driven analytical models estimate parameters based on observational data and use them as input for modeling and optimization. One application of data-driven analytical models is multi-criteria decision making, where multiple treatments (e.g., surgery, stents and statins for carotid disease) are available and a patient needs to consider multiple factors (e.g., recovery time from treatment, risk of complications and life expectancy) in choosing a treatment. Multi-criteria decision making is challenging because a given treatment may look better on one criterion but worse on others. OM techniques such as multi-objective optimization and the analytical network/hierarchical approach (Saaty 2013) can help patients choose the best treatment based on their individual preferences. These techniques can also help patients with sensitivity analyses that help them see how their choice depends on outcome estimates and/or their personal preferences.

In addition to multiple criteria, health care decisions often involve multiple periods. For example, patients with chronic diseases are typically faced with

a series of decisions to manage their condition. In some cases, such decisions are spread over time because treatment options change over time. For example, patients waiting for an organ transplant must decide whether to accept a currently available organ or to wait for a future and potentially better organ. Metastatic cancer patients can face similar decisions involving uncertain future options as they wait for improved chemotherapy options to become available. This type of multi-period decision making is challenging, because a patient's health status evolves over time. In the case of organ transplant, if a patient decides to reject an organ and stay on the waiting list, he/she might get a better offer in the future, but his/her health state might get worse. Problems like these can be addressed using familiar methods, such as linear programming and dynamic programming (Alagoz et al. 2009), but may also require new data analytics methods to update evolving options, risks, and patient characteristics to properly parameterize the models.

4.3.2. Health Care System Level. As data becomes more and more transparent, health care providers of all kinds will be faced with decisions of what services to offer. If patients can see which providers are substandard for which procedures, they will be able to selectively avoid them. Since hospitals will no longer be able to use the halo from their strengths to hide their weaknesses, it will become increasingly difficult to offer a full range of medical services. This will present hospitals and other medical providers with "invest or specialize" scenarios. They will either need to invest in process improvement to make their weakest services competitive or eliminate those services and compete on their strong services. Such decisions will be complicated by the fact that there are synergies between services (e.g., a strong transplant program for one type of organ may offer infrastructure and marketing advantages with which to build a transplant program for another type of organ). These complex strategic planning problems will require modeling support, which will need grounding in the types of data-driven evaluation of performance we have discussed in study.

Observational data can also help health systems improve outcomes through better matching of patients to facility and/or provider. As described earlier, the effect of a treatment may be heterogeneous for different patients and such heterogeneity may differ across treatments. Here, the "treatment" could be the actual medical treatment (e.g., surgery, stent, statin), but it could also be the type of facility (e.g., doctor's office, urgent care clinic, community hospital, research hospital) or the individual provider (e.g., specific surgeon). Heterogeneous outcomes imply

that a given facility or provider may be better suited to one type of patient than another and that the “best” facility or provider will be different for different patients. Using big data analysis to uncover such heterogeneity in outcomes can enable health systems to better guide patients (e.g., through physician referrals) to the types and sources of care that best meet their individual needs. By taking into account both clinical quality and operational efficiency, improvements in the matching process can help systems provide better health at a lower cost.

Modeling and optimization can also help health systems respond to pay-for-performance systems. Increased transparency of hospital performance data, made possible by more sophisticated big data analytics, will allow payers such as CMS to tie reimbursements more closely to the actual value hospitals provide to patients. An example of early efforts to do this are the hospital readmission reduction program (HRRP), which penalizes hospitals with excessive 30-day readmission rates and the hospital acquired condition reduction program (HACRP), which penalizes the worst quartile of hospitals with regard to hospital acquired infections. Because these programs base penalties on coarse evaluations of under-performing hospitals, they create incentives that may not align with the goal of promoting improvements in hospital performance. For example, Zhang et al. (2016) developed a game-theoretic model that captures the competition among hospitals induced by HRRP’s benchmarking mechanism and found that low-performing hospitals prefer paying penalties to reducing readmissions. Better performance evaluation will create opportunities for more targeted and more effective pay-for-performance mechanisms, which will require modeling and optimization support to exploit.

4.3.3. Societal Level. If a health care system were governed by a central planner who decides how to allocate limited treatment resources to different patients, heterogeneous and personalized data could be used directly to make better patient assignment decisions. But even this highly simplified version of the health care resource allocation problem is not entirely straightforward. While it is generally simple to solve an allocation problem in which both treatment effects and disease states are deterministic, the problem becomes much harder if we consider uncertainty in the effect of a treatment and the progression of a disease. To illustrate this point, consider two treatment types (Treatment 1 and 2) and two disease states (State 1 and 2). Treatment 1 is more effective and expensive than Treatment 2, but none of the treatments can cure the disease. Furthermore, the disease state evolves over time as a function of prior state and

treatment type. Because budget is limited in each period, the central planner must decide which patients to treat and the type of treatments to provide. This type of problem can be formulated using linear programming or dynamic program but is generally difficult to solve when the number of periods, treatment types or disease states under consideration is large. Zayas-Caban et al. (2017) proposed an asymptotically optimal heuristic based on a linear programming relaxation to the original stochastic formulation.

Since no country or health care system operates in pure central planning mode, analyses of such systems are used as guides for narrow decisions that do involve central planning, such as which procedures should be covered in medical plans, which patients should be prioritized in allocating scarce resources, and what guidelines and regulations should be adopted. Specific examples include allocating organs for transplant (Ata et al. 2017), optimizing colonoscopy screening for colorectal cancer prevention and surveillance (Erenay et al. 2014), allocating scarce healthcare resources in developing countries (Griffin et al. 2013), optimizing breast screening policies given heterogeneous adherence (Ayer et al. 2015).

All real-world health systems are at least partially decentralized because patients have at least some choice about what treatment to get and where to get it. Such systems still require modeling and optimization to design and plan. For example, to help system managers evaluate the value of a new type of information (e.g., EMR information made possible by Medicalchain), we need to be able to predict how the new information will affect patient decision making and how this will impact patients and other stakeholders. An example of such a model is that of Wang et al. (2018), which analyzed the relative value of population-average and patient-specific quality information about cardiac surgeons. This involved modeling patient choice of surgeon as a function of outcome quality, travel distance and waiting time on the surgical schedule and combining this with a queueing model to estimate patient waiting time. The results suggested that societal benefits (i.e., sum of patient utility) from using patient-specific information about mitral valve surgery outcomes are comparable to those achievable by enabling the best surgeons to treat 10%–20% more patients under population-average information. Analyses like these can help identify areas where more detailed information can facilitate precision medicine to achieve the greatest societal benefits.

Modeling and optimization can also help understand and improve interactions between system managers and patients that govern how resources are used. For example, in organ transplant systems, UNOS plays the role of system manager responsible

for organ allocation. When an organ becomes available, patients with compatible blood type and antibody-antigen are sequenced according to their sickness, waiting time, proximity to donor, etc. The organ is then offered to patients in this sequence. UNOS quickly realized that patients who were more likely to receive an allocation were also more likely to reject the allocation. Hence, by using historical data to understand patients' decisions, UNOS developed the Liver Simulation Allocation Model to compare potential outcomes of alternative allocation policies. Several studies have made use of this simulation to incorporate patient choice into optimization models that improve the allocation process (Akan et al. 2012, Su and Zenios 2004, 2005, 2006). As more performance data becomes available as a result of the combined big data and precision medicine revolutions, analogous analyses of other scarce resources (e.g., elite surgeon time, experimental drugs, new imaging technologies, etc.) will become possible. POM scholars are well-equipped to provide the needed analytic innovations.

4.4. Implementation

Achieving precision medicine, as we have described it here, will require more than solving statistical and analytics problems or creating simulation and optimization models. We must also find ways to communicate complex information about treatment options and outcomes to patients and providers in ways they can use it. This will require better displays (e.g., customized rankings in web pages) and tools for dealing with multi-criteria decisions (e.g., aids to help users figure out weights for criteria or mechanisms for using partial weight information to choose between alternatives).

A number of observers have noted that public reporting of health care outcomes has been less effective than anticipated in altering patient and provider behavior (see e.g., DeVore et al. 2012, Ryan et al. 2012, Smith et al. 2012). Saghafian and Hopp (2018) concluded that the primary reason for this is that patients have acted as though they either do not have or do not understand the publically reported data. This implies that information contained in this data has not been communicated in a clear and usable way. If this is indeed the case, then there is considerable opportunity to advance progress toward precision medicine by better communicating the information being generated by the big data revolution. POM scholars can contribute to research into how patients perceive information about uncertain outcomes, how trust (e.g., of doctors, insurance companies, government) influences patient reception of information from various sources, and many other practical questions related to the use of information in real-world settings.

However, having and understanding the data needed to make effective decisions for precision health care is not enough. Both patients and providers must act upon it. If other factors distort treatment decisions, patient and societal benefits will fall short of their potential. Three issues that may prevent a provider from acting in patients' best interests are: (i) Some hospitals are profit-driven, so they may pick patients or prescribe medications based on how much revenue they could generate rather than how well the patients could be treated; (ii) Hospitals that are rated and reimbursed according to imperfectly risk-adjusted outcomes may intentionally avoid very sick patients who are likely to have bad outcomes (Dranove et al. 2003); (iii) Referral decisions by physicians may be influenced by non-clinical factors such as hospital affiliation, personal relationship, waiting time, etc. (Lu and Lu 2016). Analyses of the impact of big data on precision medicine, as well as designs of pay-for-performance systems, must take behaviors like this into account.

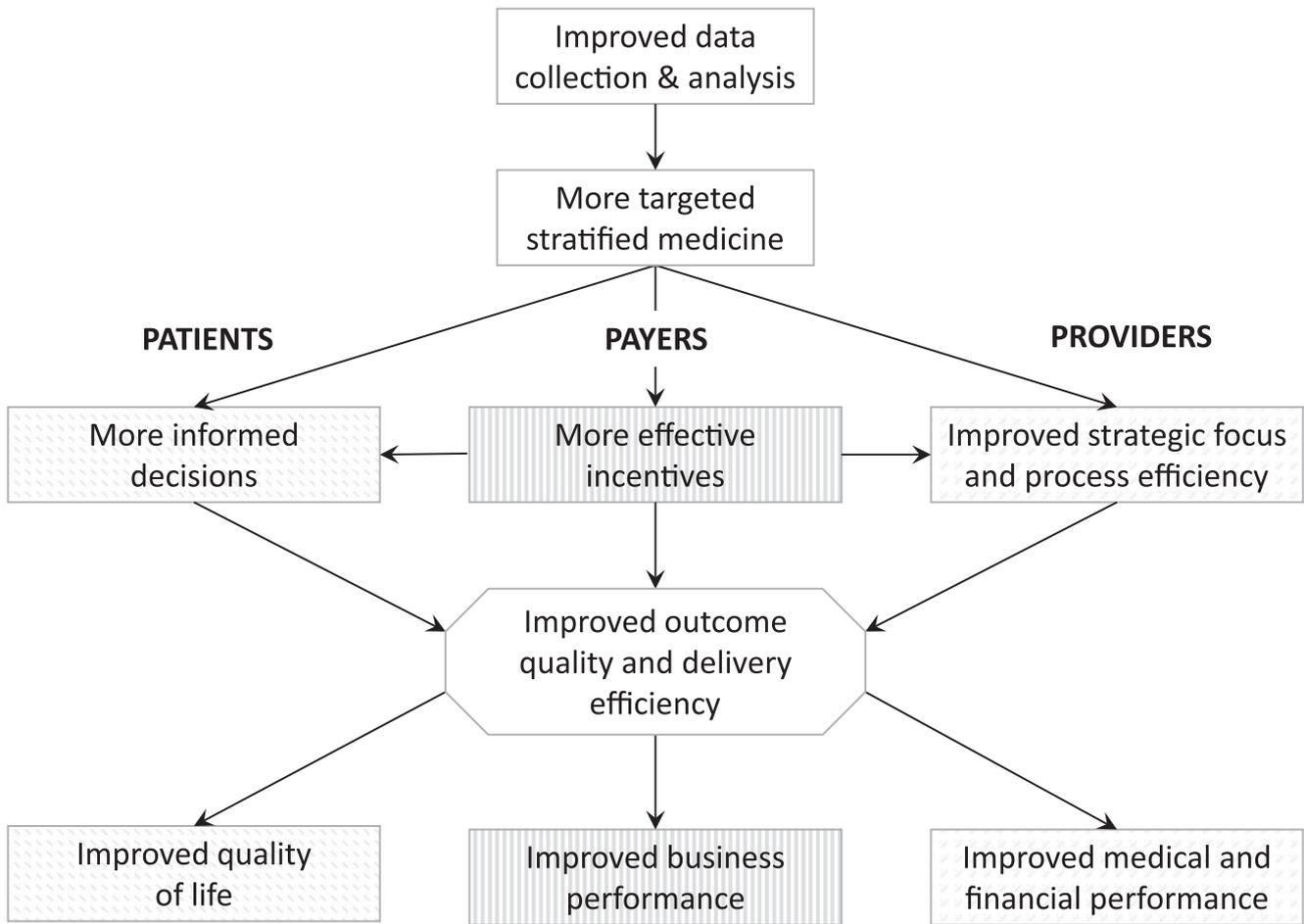
Finally, there are several important roles payers can play in the evolution of precision medicine. First, payers can help guide patients to the most suitable providers by structuring networks, co-pays and other policies to favor effective choices. Second, by combining the above patient incentives with reimbursements that favor high quality outcomes, payers can influence providers to focus on patients and procedures that fit their relative strengths. Third, payers can improve health care through pay-for-performance that incent providers to invest in process improvements that result in better patient outcomes. The ability of payers to carry these out will depend critically on the analytics that underlie performance evaluation.

5. Conclusion

As depicted in Figure 3, the improved data collection and analysis that will result from the trends and activities discussed here will lead to stratified medicine that is increasingly better targeted at individual patient needs. Patients will use the improved information to make more informed decisions about their own health care, while providers will use it to make better strategic planning and process improvement decisions, and payers will use it to design better incentives for both patients and providers. The combined effect will be improved patient outcomes and greater efficiency due to elimination of waste and errors caused by trial-and-error medicine.

For patients, the primary benefit of better decisions at all levels of the health care hierarchy will be better health. By providing evidence-based guidance to treatment options and provider choices that best suit an individual patient's needs, precision medicine will

Figure 3 Impact of Precision Medicine on Health Care Stakeholders



lead to significant improvements in patient outcomes. Big data will amplify this impact by providing increasingly detailed evidence that increases the power and specificity of precision medicine.

For example, a patient who suffers from severe depression today can expect it to take months, or even years, of consultation with psychiatrists, psychologists, therapists, yoga instructors, etc., to find a combination of medications and lifestyle changes that bring relief. In the coming world of precision medicine, the patient will receive treatments aligned with his/her individual characteristics and will see improvement much faster and will ultimately find a better health outcome.

But patients will benefit from more than better health because of precision medicine. By shifting the basis for medical decisions away from subjective judgement and toward statistical evidence, big data will move the locus of decision-making toward patients. Under trial-and-error medicine, in which providers choose a course of action largely on their personal experience, a patient has little choice but to accept a recommendation based on the provider’s judgment. Its not easy for a lay

person to argue with an experienced medical professional who says, “in my considered opinion you should do X.” But under precision medicine, where there is an explicit base of evidence to indicate a course of action, it will become possible for the patient to argue with a recommendation without disparaging the recommender. For example, suppose an orthopedist recommended a total knee replacement based on data showing an 80% chance of significant pain and mobility improvement and only a 5% chance of an outcome worse than the status quo. A particularly risk-averse patient could reasonably argue that the procedure is too risky for him/her. As data becomes more transparent, patients will become increasingly empowered to participate in more decisions regarding his/her health. The precision medicine revolution will ultimately bring about the end of paternalistic medicine.

In addition to improving health outcomes at the individual and societal levels, precision medicine powered by big data will also reduce costs. Brownlee (2008) estimates that between one-fifth and one-third of health care dollars are spent on unnecessary tests,

drugs and treatments (and that such overtreatment is responsible for as many as 30,000 deaths per year). By providing a clear evidence base for determining which interventions are clinically worthwhile, big-data-driven precision medicine will help patients and providers avoid unnecessary treatments and their associated costs. Furthermore, by forgoing treatments that are not well-suited to the patient, precision medicine will lead to fewer errors, complications, and follow-up corrections, all of which will reduce costs. Finally, precision medicine can and should indicate where less invasive measures (e.g., prevention) are the best course of action. In the right situations, these approaches can achieve better health at dramatically reduced costs.

The impact of the precision medicine revolution will impact provider behavior beyond interactions with patients. The detailed patient response data on which precision medicine is based will give clinicians and administrators granular feedback on their strengths and weaknesses. This will allow them to make strategic decisions about their focus, for example, choosing which surgical procedures to offer or which patient types to target. It will also help them focus their process improvement efforts on the patients and services where improvements are most needed.

Finally, precision medicine will open up a host of opportunities for payers to sharpen the incentives they provide to both patients and providers, with a goal of achieving better health outcomes at lower costs. When payers have detailed data on the best care options and choice of provider for a given patient, they can design pricing and co-payment schemes that incentivize patients to choose the most cost-effective alternatives. Similarly, they can design pay-for-performance reimbursement schemes that incentivize payers to seek out patients where they have a comparative quality advantage and to invest in targeted process improvements.

This impressive list of benefits will only be possible if we can resolve the data collection, estimation, modeling and optimization, and implementation issues discussed above. Since these are all challenges POM scholars are well-suited to address, the combined big data/precision medicine revolution is an area where our field can change the world for the better.

Acknowledgments

The authors are grateful to the Editor, Kalyan Singhal, for giving us the opportunity to write this essay. We are also indebted to our many colleagues and collaborators in the University of Michigan Hospital System and at the Mayo Clinic who have patiently helped us understand medical

challenges and have guided us to data sources with which to study them.

Notes

¹Risk-adjusted success rates correct for differences in patient characteristics by computing the average outcome from each alternative for a mix of patients that reflects the population. Note that the expected outcome for a “population average” patient may not reflect the outcome for any actual patient. Hence, as we note in Wang et al. (2018), population average data may not give appropriate rankings of alternatives when patient outcomes are heterogeneous.

²For example, Park et al. (2015) compared percutaneous coronary intervention (PCI) using stents with coronary-artery bypass grafting (CABG) in treating multi-vessel coronary artery disease. Although they found the rate of major adverse cardiovascular events was significantly higher among PCI patients than CABG patients, their ability to detect differences in outcomes among patient subgroups was limited by the size of the study. The study involved a total of 880 patients, which permitted analysis of only a small number of patient characteristics and restricted the statistical power of the analyses that were done.

References

- Akan, M., O. Alagoz, B. Ata, F. S. Erenay, A. Said. 2012. A broader view of designing the liver allocation system. *Oper. Res.* **60**(4): 757–770.
- Alagoz, O., A. J. Schaefer, M. S. Roberts. 2009. Optimizing organ allocation and acceptance. S. A. Shumaker, J. K. Ockene, K. A. Riekert, eds. *Handbook of Optimization in Medicine*. Springer, Boston, MA, 1–24.
- Ata, B., J. J. Freidewald, A. C. Randa. 2017. Organ transplantation. Working paper, University of Chicago, Chicago, IL.
- Athey, S., G. Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proc. Natl Acad. Sci.* **113**(27): 7353–7360.
- Athey, S., J. Tibshirani, S. Wager. 2017. Generalized random forests. Working paper, Graduate School of Business, Stanford University, Stanford, CA.
- Ayer, T., O. Alagoz, N. Stout, E. Burnside. 2015. Heterogeneity in women’s adherence and its role in optimal breast cancer screening policies. *Management Sci.* **62**(5): 1339–1362.
- Azaria, A., A. Ekblaw, T. Vieira, A. Lippman. 2016. MedRec: Using blockchain for medical data access and permission management. *IEEE Int. Conf. Open Big Data* **2016**: 25–30. Available at <https://ieeexplore.ieee.org/document/7573685/>
- Bartel, A. P., C. W. Chan, H. Kim. 2016. Should hospitals keep their patients longer? The role of inpatient and outpatient care in reducing readmissions. Working paper, Columbia Business School, New York, NY.
- Bastani, H., M. Bayati. 2018. Online decision-making with high-dimensional covariates. Working paper, Graduate School of Business, Stanford University, Stanford, CA.
- Benjamini, Y., Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**(1): 289–300.
- Bertsimas, D., J. Dunn. 2017. Optimal classification trees. *Mach. Learn.* **106**(7): 1039–1082.

- Bertsimas, D., A. O'Hair, S. Relyea, J. Silberholz. 2016. An analytics approach to designing combination chemotherapy regimens for cancer. *Management Sci.* **62**(5): 1511–1531.
- Bonferroni, C. 1936. Teoria Statistica Delle Classi e Calcolo Delle Probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze (Libreria Internazionale Seeber, Florence, Italy)* **8**: 3–62.
- Boruvka, A., D. Almirall, K. Witkiewitz, S. A. Murphy. 2017. Assessing time-varying causal effect moderation in mobile health. *J. Am. Stat. Assoc.* <https://doi.org/10.1080/01621459.2017.1305274>. [Epub ahead of print]
- Brownlee, S. 2008. *Overtreated: Why Too Much Medicine is Making us Sicker and Poorer*. Bloomsbury, New York.
- Chan, C. W., V. F. Farias, G. J. Escobar. 2016. The impact of delays on service times in the intensive care unit. *Management Sci.* **63**(7): 2049–2072.
- Clark, J. R., R. S. Huckman. 2012. Broadening focus: Spillovers, complementarities, and specialization in the hospital industry. *Management Sci.* **58**(4): 708–722.
- Clark, J. R., R. S. Huckman, B. R. Staats. 2013. Learning from customers: Individual and organizational effects in outsourced radiological services. *Organ. Sci.* **24**(5): 1539–1557.
- DeVore, A. D., B. G. Hammill, N. C. Hardy, Z. J. Eapen, E. D. Peterson, A. F. Hernandez. 2012. Has public reporting of hospital readmission rates affected patient outcomes? *J. Am. Coll. Cardiol.* **67**(8): 570–577.
- Dimick, J. B., D. O. Staiger, O. Baser, J. D. Birkmeyer. 2009. Composite measures for predicting surgical mortality in the hospital. *Health Aff.* **28**(4): 1189–1198.
- Dranove, D., D. Kessler, M. McClellan, M. Satterthwaite. 2003. Is more information better? The effects of “report cards” on health care providers. *J. Polit. Econ.* **111**(3): 555–588.
- Erenay, S., O. Alagoz, A. Said. 2014. Optimizing Colonoscopy Screening for colorectal cancer prevention and surveillance. *Manuf. Serv. Oper. Manag.* **16**(3): 381–400.
- Federal Drug Administration. 2018. Precision medicine. Available at <https://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/InVitroDiagnostics/PrecisionMedicine-MedicalDevices/default.htm> (accessed date February 23, 2018).
- Freeman, M., N. Savva, S. Scholtes. 2016. Gatekeepers at work: An empirical analysis of a maternity unit. *Management Sci.* **63**(10): 3147–3167.
- Friedman, J., T. Hastie, R. Tibshirani. 2001. *The Elements of Statistical Learning*, vol 1. Springer Series in Statistics. Springer, New York.
- Giordano, S. H., Y.-F. Kuo, Z. Duan, G. N. Hortobagyi, J. Freeman, J. S. Goodwin. 2008. Limits of observational data in determining outcomes from cancer therapy. *Cancer.* **112**(11): 2456–2466.
- Glace, L. G., A. Dick, T. M. Osler, Y. Li, D. B. Mukamel. 2006. Impact of changing the statistical methodology on hospital and surgeon ranking: The case of the New York State cardiac surgery report card. *Med. Care.* **44**(4): 311–319.
- Griffin, J., P. Keskinocak, J. Swann. 2013. Allocating scarce health-care resources in developing countries: A case for malaria prevention. B. Denton, ed. *Handbook of Healthcare Operations Management*. Springer, New York, NY, 511–532.
- Hadley, J., K. R. Yabroff, M. J. Barrett, D. F. Penson, C. S. Saigal, A. L. Potosky. 2010. Comparative effectiveness of prostate cancer treatments: Evaluating statistical adjustments for confounding in observational data. *J. Natl Cancer Inst.* **102**(23): 1780–1793.
- Heckman, J. J. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Ann. Econ. Soc. Meas.* **5**(4): 475–492.
- Heckman, J. J. 1977. Sample selection bias as a specification error. *Econometrica* **47**(1): 153–161.
- Ho, V., B. H. Hamilton, L. L. Roos. 2000. Multiple approaches to assessing the effects of delays for hip fracture patients in the United States and Canada. *Health Serv. Res.* **34**(7): 1499–1518.
- Imai, K., M. Ratkovic. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.* **7**(1): 443–470.
- KC, D. S. 2013. Does multitasking improve performance? Evidence from the emergency department. *Manuf. Serv. Oper. Manag.* **16**(2): 168–183.
- KC, D. S. 2018. Heuristic thinking in patient care. Working paper, Goizueta Business School, Emory University, Atlanta, GA.
- KC, D. S., C. Terwiesch. 2011. The effects of focus on performance: Evidence from California hospitals. *Management Sci.* **57**(11): 1897–1912.
- KC, D. S., C. Terwiesch. 2012. An econometric analysis of patient flows in the cardiac intensive care unit. *Manuf. Serv. Oper. Manag.* **14**(1): 50–65.
- Kim, S. H., C. W. Chan, M. Olivares, G. Escobar. 2014. ICU admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Sci.* **61**(1): 19–38.
- Klasnja, P., E. B. Hekler, S. Shiffman, A. Boruvka, D. Almirall, A. Tewari, S. A. Murphy. 2015. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychol.* **34**(5): 1220.
- Kruse, C., R. Goswamy, Y. Raval, S. Marawi. 2016. Challenges and opportunities of big data in health care: A systematic review. *JMIR Med. Inform.* **4**(4): e38.
- Lam, Y. W. 2013. Scientific challenges and implementation barriers to translation of pharmacogenomics in clinical practice. *ISRN Pharmacol.* **2013**: 1–17.
- Lee, D., J. Li, G. Wang, K. Croome, J. Burns, D. Perry, J. Nguyen, W. Hopp, B. Taner. 2017. Looking inward: Impact of operative time on patient outcomes. *Surgery* **162**(4): 937–949.
- Li, X., J. Qin. 2017. Anonymizing and sharing medical text records. *Inf. Syst. Res.* **28**(2): 332–352.
- Li, J., S. Netessine, S. Koulayev. 2017. Price to compete... with many: How to identify price competition in high-dimensional space. *Management Sci.* <https://doi.org/10.1287/mnsc.2017.2820>. [Epub ahead of print]
- Little, R. J., D. B. Rubin. 2014. *Statistical Analysis with Missing Data*, vol 333. John Wiley & Sons, New York.
- Lu, L. X., S. F. Lu. 2017. Distance, quality, or relationship? Inter-hospital transfer of heart attack patients. *Prod. Oper. Manag.* <https://doi.org/10.1111/poms.12711>. [Epub ahead of print]
- McClellan, M., B. J. McNeil, J. P. Newhouse. 1994. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *J. Am. Med. Assoc.* **272**(11): 859–866.
- Miller, A., C. Tucker. 2017. Privacy protection, personalized medicine, and genetic testing. *Management Sci.* <https://doi.org/10.1287/mnsc.2017.2858>. [Epub ahead of print]
- National Research Council. 2011. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. The National Academies Press, Washington, DC. <https://doi.org/10.17226/13284>.
- Park, S.-J., J.-M. Ahn, Y.-H. Kim, D.-W. Park, S.-C. Yun, J.-Y. Lee, S.-J. Kang, S.-W. Lee, C. W. Lee, S.-W. Park, S. J. Choo, C. H. Chung. 2015. Trial of everolimus-eluting stents or bypass surgery for coronary disease. *N. Engl. J. Med.* **372**: 1204–1212.

- Ryan, A. M., B. K. Nallamothu, J. B. Dimick. 2012. Medicare's public reporting initiative on hospital quality had modest or no impact on mortality from three key conditions. *Health Aff.* **31**(3): 585–592.
- Saaty, T. L. 2013. The modern science of multicriteria decision making and its practical applications: The AHP/ANP approach. *Oper. Res.* **61**(5): 1101–1118.
- Saghafian, S., W. Hopp. 2018. Can public reporting cure health-care? The role of quality transparency in improving patient-provider alignment. Working paper, Kennedy School of Government, Harvard University, Cambridge, MA.
- Sarker, H., K. Hovsepian, S. Chatterjee, I. Nahum-Shani, S. A. Murphy, B. Spring, E. Ertin, M. al'Absi, M. Nakajima, S. Kumar. 2017. From markers to interventions: The case of just-in-time stress intervention. J. Rehg, S. Murphy, S. Kumar, eds. *Mobile Health*. Springer, Cham, 411–433.
- Signorovitch, J. E. 2007. Identifying informative biological markers in high-dimensional genomic data and clinical trials. Doctoral dissertation, Harvard University, Cambridge, MA.
- Smith, M. A., A. Wright, C. Queram, G. C. Lamb. 2012. Public reporting helped drive quality improvement in outpatient diabetes care among Wisconsin physician groups. *Health Aff.* **31**(3): 570–577.
- Song, H., A. L. Tucker, K. L. Murrell. 2015. The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Sci.* **61**(12): 3032–3053.
- Spigel, D. R. 2010. The value of observational cohort studies for cancer drugs. *Biotechnol. Healthcare* **7**(2): 18–24.
- Su, X., S. Zenios. 2004. Patient choice in kidney allocation: The role of the queueing discipline. *Manuf. Serv. Oper. Manag.* **6**(4): 280–301.
- Su, X., S. A. Zenios. 2005. Patient choice in kidney allocation: A sequential stochastic assignment model. *Oper. Res.* **53**(3): 443–455.
- Su, X., S. A. Zenios. 2006. Recipient choice can address the efficiency-equity trade-off in kidney transplantation: A mechanism design model. *Management Sci.* **52**(11): 1647–1660.
- Unger, J. M., E. Cook, E. Tai, A. Bleyer. 2016. Role of clinical trial participation in cancer research: Barrier, evidence, and strategies. *Am. Soc. Clin. Oncol. Educ. Book* **35**: 185–198.
- Walton, A., B. Nahum-Shani, L. Crosby, P. Klasnja, S. Murphy. 2018. Optimizing digital integrated care via micro-randomized trials. *Clin. Pharmacol. Ther.* <https://doi.org/10.1002/cpt.1079>. [Epub ahead of print]
- Wang, H., R. Li, C. Tsai. 2007. Tuning parameter selector for the smoothly clipped absolute deviation method. *Biometrika* **94**(3): 553–568.
- Wang, G., J. Li, W. J. Hopp, F. L. Fazzalari, S. Bolling. 2015. Hospital quality and patient choice: An empirical analysis of mitral valve surgery. Working paper, Ross School of Business, University of Michigan, Ann Arbor, MI.
- Wang, G., J. Li, W. Hopp. 2017a. An instrumental variable tree approach for detecting heterogeneous treatment effects in healthcare and other observational studies. Working paper, Ross School of Business, University of Michigan, Ann Arbor, MI.
- Wang, G., J. Li, W. Hopp. 2017b. Personalized health care outcome analysis of cardiovascular surgical procedures. Working paper, Ross School of Business, University of Michigan, Ann Arbor, MI.
- Wang, G., J. Li, W. Hopp, F. Fazzalari, S. Bolling. 2018. Using patient-specific quality information to unlock hidden health care capabilities. *Manuf. Serv. Oper. Manag.* (e-pub ahead of print). <https://doi.org/10.1287/msom.2018.0709>.
- Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA.
- Wu, D. 2016. Shock spillover and financial response in supply chain networks: Evidence from firm-level data. Working paper, Ross School of Business, University of Michigan, Ann Arbor, MI.
- Xu, Y., M. Armony, A. Ghose. 2018. The Interplay between online reviews and physician demand: An empirical investigation. Working paper, Gies College of Business, University of Illinois, Urbana-Champaign, IL.
- Zayas-Caban, G., S. Jasin, G. Wang. 2017. An asymptotically optimal heuristic for general non-stationary finite-horizon restless multi-armed multi-action bandits. Working paper, Ross School of Business, University of Michigan, Ann Arbor, MI.
- Zhang, D. J., I. Gurvich, J. A. Van Mieghem, E. Park, R. S. Young, M. V. Williams. 2016. Hospital readmissions reduction program: An economic and operational analysis. *Management Sci.* **62**(11): 3351–3371.

Using Data and Big Data in Retailing

Marshall Fisher

The Wharton School, University of Pennsylvania, 3730 Walnut Street, 542 Jon M. Huntsman Hall, Philadelphia, Pennsylvania 19104, USA, fisher@wharton.upenn.edu

Ananth Raman*

Business Logistics, Harvard Business School, Morgan Hall 437, Boston, Massachusetts 02163, USA, araman@hbs.edu

In this essay, we examine how retailers can use data to make better decisions and hence, improve their performance. We have been studying retail operations for over two decades and have witnessed many, and been personally involved in a few, projects that delivered considerable value to retailers by better exploiting data. We highlight a few of these examples and also identify some other potential applications.

Key words: retailing; inventory; pricing; assortment; big data
History: Accepted: January 2018 by Kalyan Singhal.

1. Introduction

The two of us have been working on bringing data and analysis-based approaches to retailing since the mid-1990s. We did not have the clairvoyance to call it “Big Data” when we started but instead referred to it as “Rocket Science Retailing” in a *Harvard Business Review* article published in 2000 (Fisher et al. (2000)). In 1997, we were fortunate to receive a grant from the Alfred P. Sloan Foundation to examine and document how retailers were approaching the opportunities presented to them by the increasing availability of data and computer hardware and software to analyze the data. Over the next few decades, with the participation of over 50 retail partners, we studied and wrote about a variety of problems in retailing—ranging from inventory planning, labor scheduling, assortment planning, store execution (including inventory record inaccuracies and phantom stockouts), and incentives (within the organization and in the supply chain). Over time, we wrote many articles and case studies and summarized our findings in a book *The New Science of Retailing*, which we co-authored (Fisher and Raman 2010). Together, we direct the *Consortium for Operational Excellence in Retailing (COER)*, an industry-academia research group, and host the annual COER conference. By far, the biggest satisfaction we gained during the last 20 years was from mentoring and supervising over a dozen doctoral students directly and in teaching concepts in retail operations to thousands of undergraduates, MBA students, and executive participants.

In 1995, when we began to research how retailers could make better use of data, big data was not as big as it is today. The data available to retailers was mainly point-of-sale (POS) data, augmented by things like loyalty cards, that gave insight into the demographics of customers buying the retailer’s products. During the 15 years leading up to publishing *The New Science of Retailing*, we discovered that even only sales data, properly analyzed, can be extremely useful to a retailer in improving decision-making and adding a component of science to their art. But in 1995, retailing was much more art than science and the use of even sales data was limited. Over time this has changed, and many exciting analytic tools have emerged, some described below. Moreover, data truly has gotten bigger, as more and better data have emerged. Now we have much new data, including ever-growing data from e-tailing, in-store video, tracking of customers within store via mobile phones, and smart fitting rooms. Most e-tailers now scrape the websites of their competitors to obtain competitive intelligence, like what prices they are charging, that day. The Chinese Internet retailer Yihaodian created a remarkable technology in which a “hot spot” can be created in a parking lot or other open space that creates a display on a customer smart phone that shows product on shelves as though they were walking around a store. And of course, we all await the opening of the technologically rich Amazon ‘Go Store.’

These days, when we mention in a social setting that we work with retailers, we are often met with sympathy. “Oh, that’s too bad. Retailing’s in trouble, isn’t it?” And certainly, we are confronted by a steady stream of retail bad news about physical store

retailers closing stores, seeing their stocks plunge, and not infrequently declaring bankruptcy.

Among the many examples of retail bad news is JC Penny. While commenting on the turnaround that incoming CEO Ron Johnson attempted unsuccessfully at iconic US retailer JC Penney, investor Bill Ackman summarized Johnson's challenge as, "Build the fastest growing retailer (0–7 million square feet in 15 months) while winding down the 110-year-old iconic retailer at the same time . . . that has a different customer base and driving pricing experience. . . by the way, we need those cash flows to fund the growth." Most retailers in the United States and in many other developed markets can relate to this challenge—they have to find ways to launch and scale-up innovations successfully, while also executing and unwinding their current business model. In other words, they have to "explore" new business models and "exploit" their mature business concurrently (O'Reilly and Tushman 2011).

Much of the physical store retail malaise is attributed to the steady advance of e-tailing, led by Amazon. One would think in a few years physical stores would be a thing of the past. But to paraphrase Mark Twain, reports of the stores' death have been greatly exaggerated. An August 17, 2017 U.S. Department of Commerce report lists Q2 2017 E-Commerce Sales at 8.9% of total retail. One year prior, that number was 8% and the year before that 7%. So e-Commerce's share of retail revenue is growing linearly at the rate of about 1%. If that were to continue (and most growth rates do not continue unabated, they slow down), in 40 years e-Commerce would still be less than half of retailing. So stores will be with us for a long time.

Fisher et al. (2017a,b) point out that mature retailers can continue to prosper, even when they run out of room for new stores, if they focus on growing sales in their existing stores using data to optimize the drivers of sales, including assortment, inventory, price, and store staffing levels. Examples of retailers discussed in the article that are doing this well, and prospering thereby, include Home Depot and McDonald's.

Data—and Big Data—can help retailers not only to explore new opportunities and innovations but also to execute their current business models and wind down parts of the business that are not working. This essay provides some guidelines on why Big Data represents a significant opportunity in retailing, and identifies a few specific application areas that retailers could consider.

Before we discuss how data can enable retailers to improve their performance, it is helpful to understand that retailing—both brick-and-mortar retailing and e-commerce—is a high fixed-cost, low net margin business. Net margins in retailing are commonly

around 2–3% of sales and significantly lower than gross margin, labor cost, the cost of real estate, and inventory levels. This means even small increases in revenue have a big impact on profit. Consider, for example, the impact of increasing sales by 1% for a retailer with gross margins of 50% and net margin of 2%. The incremental sales will increase gross margin dollars by 0.5% of sales; assuming fixed costs stay unchanged, this will flow to the bottom line and increase net profits by 25%. Similarly, as has been shown in prior literature (Gaur et al. 2014), even small errors in valuing a retailer's inventory level can have very significant impact on a retailer's market capitalization. These examples help to illustrate why even small changes in retail operations can have very significant impact on a retailer's net profits and market capitalization. They also serve to explain why US publicly listed retailers on average have delivered stock-market returns that exceed the market average even while facing a significantly higher risk of bankruptcy than other listed firms.

2. How Data (and Big Data) Can Help Retailers Innovate while Executing Better

We describe below how retailers can use—and are using—data about their customers and their business operations more creatively to explore and exploit better. We had discussed many illustrative examples a few years ago in our book that was published in 2010; so, we will highlight some more recent examples here.

2.1. Innovation

Data and analytics provide retailers with tremendous opportunity to innovate in their operations and their business models. When we started studying retailing a few decades ago, our primary focus had been on the vast amounts of point-of-sale (POS) data that retailers were gathering. Today, the POS data have been augmented in numerous ways. In many ecommerce and brick-and-mortar contexts, retailers can track not only what is being sold at different locations and times (through POS systems) but also *who* is buying these items. Moreover, technology potentially allows retailers to observe what a customer browses or tries on in the fitting room before purchase. It is easy to visualize how ecommerce retailers can track their customers' browsing behavior. Brick-and-mortar retailers have similar options too. For example, one retailer that we are familiar with used RFID tags on clothes to track which products were being tried on by customers in fitting rooms. A supermarket chain (Larson et al. 2005) was able to use data obtained through RFID

tags to track the path of shopping carts and hence, customers, within the store. Video analysis and face recognition technology potentially allow retailers to track when an individual customer enters a store and also identify the social networks that exist among customers by knowing which customers tend to enter a store together.

Leveraging these technologies will require retailers to think creatively about offering new services and developing new business models. Some retailers—like Amazon.com (Amazon)—seem especially adept at developing such services. Today, most products are reviewed extensively at Amazon, and the company uses its vast data to identify for any consumers “items that are frequently bought together” and “customers who bought this item also bought.” It is reasonable to expect that when “Amazon Go”—Amazon’s future store concept with very little or no store staff—is rolled out, it will likely use cell phones and video recognition technology to track customers and RFID tags to track products.

It is hard to generalize about retailing broadly, given its size and diversity. It is our sense though that while retailers have been adept at using technology incrementally, they have found it hard to reinvent their business models based on the new technology. The video retailer Blockbuster is often held up as a canonical example. Why did Blockbuster not move more aggressively in reinventing its business model? Similarly, Borders—an excellent bookstore that we have studied closely on multiple occasions—failed to capitalize on the opportunity for selling books on the Internet. In Borders case, the company also saw the opportunity for doing so, and its annual report had discussed the possibility in 1990 of distributing floppy disks that customers could use to order books from home—well before the launch of Amazon.com!

2.2. Execution

Data and Big data can also help retailers execute their current business model better. In addition to enhancing the retailer’s market capitalization, superior execution can provide the cash flow needed to invest in innovation projects. In recent years, we have seen at our annual COER conferences how data can be used to tackle new kinds of problems in retailing; in the following paragraphs, we explain how data can be used to improve assortment, pricing, and store closing, and to quantify the impact of stockouts on lost sales and profits and of delivery time on revenue. Feng Qi and Shanthikumar (2017) describe elsewhere in this issue how the enhanced availability of data is influencing academic research in many areas of operations management. Their article, which includes many applications in retailing, complements our study very well.

2.2.1. Optimizing Store Assortments. Periodically, retailers update the assortment of products carried in the various categories in their stores, dropping some products and adding others, in response to changing demand patterns and to new products that have entered the market. The hardest part of this process is knowing what a potential new product will sell if added to the assortment in a store, and how much of those sales will be incremental and how much will cannibalize the sales of existing products. Fisher and Vaidyanathan (2012, 2014) describe a way to do this by first identifying the attributes of the products in a category, then using sales of existing products to estimate the demand for attributes, and finally, estimate the demand for a potential new product from the demand for its constituent attributes. Examples of attributes include size and price/value for tires, and product, material, primary gemstone, and price/value for jewelry. The parameters in a demand model, attribute level demand shares in a store for attributes and the probability a customer will substitute to another attribute level if their first choice is not in the assortment, are chosen to minimize the deviation between actual sales and demand model predictions for existing products. Implementations of this approach have produced revenue increases of 3% to 6%.

2.2.2. Online Dynamic Pricing. A type of new data that is enabled by e-commerce is the ability to track a competitor’s prices, and indeed, many e-tailers will use software to ‘scrape’ competitor web sites daily to download the assortment of products they carry, their prices and whether or not they are in stock. Fisher et al. (2017a) describes a project with an e-tailer who was collecting prices daily from several competitors and wondering how to respond in their own pricing. They formulate a model that estimates demand for each product sold by the e-tailer based on their prices and competitor prices, an experiment with randomized prices to estimate the parameters in the model, and a best-response pricing algorithm that takes into account consumer choice behavior, competitors’ actions, and supply parameters (procurement costs, margin target, and manufacturer price restrictions). Initial implementation of the algorithm produced an 11% revenue increase in one category and a 19% increase in a second, while maintaining margin above a retailer-specified target.

2.2.3. Online Order Fulfillment Speed. Many e-tailers and omni-channel retailers are making substantial investments to shorten customer order fulfillment time. While it is plausible that customers will be happier getting their order fulfilled more quickly, there is little to no evidence on whether the increased

revenue from one of these initiatives is sufficient to justify the cost. Gallino et al. (2014) describes a retailer that added a second Distribution Center (DC) that reduced order fulfillment time for a segment of their customers from 5–9 days to 2–4 days. A difference-in-difference analysis of revenue before and after this event, while controlling for other demand drivers, showed that this resulted in a 4% revenue increase, an amount sufficient to more than justify the cost of the second DC. The demand model was also used to show that given the retailers' current volume, a third DC would not be cost justified. This thus affords a retailer a technique for determining the optimal investment in order fulfillment speed.

2.2.4. Store Closing and Liquidation. Retailers periodically need to identify those of their stores that are not performing well and close them. Moreover, entire retail chains need to be liquidated from time to time. In either case, retailers usually find ways to liquidate, as profitably as possible, the inventory in the store so as to recover as much cash as possible for their investors and lenders; more importantly, as a consequence of retailers becoming more efficient at liquidation, lenders (e.g., big banks in the United States) have become more comfortable using inventory as a collateral for loans (see Foley et al. 2015).

One of our former students, Nathan Craig, studied the store liquidation problem extensively, and identified ways in which the process of liquidation could be made more efficient. He showed that the performance of a store during liquidation differs substantially from its performance prior to liquidation. It is common for sales to increase by a factor of 5 or 6 during the liquidation period (relative to the same period the prior year) at certain stores but hardly increase at other stores in the same chain. Moreover, to the best of our knowledge, the difference in performance cannot be predicted in advance.

Professional liquidator firms (like our research partner Gordon Brothers) understand the importance of tracking store performance during liquidation and making adjustments in price, inventory, and duration of liquidation accordingly. Research described in Craig and Raman (2016) showed that data-based forecasting and optimization could have improved the net recovery on cost (i.e., the profit obtained during a liquidation, stated as a percentage of the cost value of liquidated inventory) by two to five percentage points in the cases we examined. Interestingly, the experts' decisions differed in some systematic ways from the method that we proposed in our study. On markdowns, for example, experts typically take a smaller-than-optimal markdown early in the liquidation and

take too steep a markdown toward the end of the liquidation.

2.2.5. Estimating the Impact of B2B Service Level on Demand. Most manufacturers know that the orders they receive from retailers (and other B2B customers) is a function of the service levels they provide. Most manufacturers are unable to quantify the relationship in large part because they lack a methodology to identify the relationship in a retailer's operating data.

Craig et al. (2016) uses data from a field experiment at Hugo Boss to quantify the relationship in that context and to present a method that could be used more widely. Their method examines the relationship between historical fill rates and order quantity, and finds that a 1% increase in historical fill rate is associated with a statistically significant 11% increase in current retailer demand, controlling for other factors that might affect retailer demand.

3. Data and Big Data in Retailing: The Potential is Huge, the Pilots are Very Promising, but the Scale-up has Been Challenging

The potential for using data and big data to improve retail operations is huge. Numerous pilots have shown that retailers can improve their performance substantially. Moreover, as we have argued above, the retail business is such that even small improvements—such as, increases in sales (or reductions in lost sales), reductions in bad inventory, or reductions in labor costs—could improve profits substantially.

References

- Craig, N. C., A. Raman. 2016. Improving store liquidation. *Manuf. Serv. Oper. Manag.* 18(1): 89–103.
- Craig, N., N. DeHoratius, A. Raman. 2016. The impact of supplier inventory service level on retailer demand. *Manuf. Serv. Oper. Manag.* 18(4): 461–474.
- Feng, Q., J. G. Shanthikumar. 2018. How research in production and operations management may evolve in the era of big data. *Prod. Oper. Manag.* 27(9): 1670–1684. <https://doi.org/10.1111/poms.12836>.
- Fisher, M., A. Raman. 2010. *The New Science of Retailing: How Analytics are Transforming Supply Chains and Improving Performance*. Harvard Business School Press.
- Fisher, M., R. Vaidyanathan. 2012. Which Products Should You Stock? *Harvard Business Review*, November 2012.
- Fisher, M., R. Vaidyanathan. 2014. An algorithm and demand estimation procedure for retail assortment optimization. *Management Sci.* 60(10): 2401–2415.
- Fisher, M., A. Raman, A. McClelland. 2000. Rocket science retailing is coming: Are you ready. *Harvard Business Review*, July 2000.

- Fisher, M., S. Gallino, J. Li. 2017a. Competition-based dynamic pricing in online retailing: A methodology validated with field experiments. *Management Sci.* Available at <https://doi.org/10.1287/mnsc.2017.2753> (accessed date January 4, 2017).
- Fisher, M., V. Gaur, H. Kleinberger. 2017b. Curing the Addiction to Growth. *Harvard Business Review*, January–February 2017.
- Foley, C. F., A. Raman, N. C. Craig. 2015. “Inventory-Based Lending Industry Note.” Harvard Business School Background Note 612-057, January 2012. (Revised June 2015.)
- Gallino, S., J. Xu, M. Fisher. 2014. The value of rapid delivery in online retailing. Working paper, Wharton School Operations, Information and Decisions.
- Gaur, V., S. Kesavan, A. Raman. 2014. Retail inventory: Managing the canary in the coal mine! *Ca. Manag. Rev.* 56(2): 55–76.
- Larson, J. S., E. T. Bradlow, P. Fader. 2005. An exploratory look at in store supermarket shopping paths. *Int. J. Res. Market.* 22(4): 395–414.
- O’Reilly III, C. A., M. L. Tushman. 2011. Organizational ambidexterity in action: How managers explore and exploit. *Ca. Manag. Rev.* 53(4): 5–21.

How Research in Production and Operations Management May Evolve in the Era of Big Data

Qi Feng, J. George Shanthikumar*

Krannert School of Management, Purdue University, Rawls 4024, 100 S. Grant Street, West Lafayette, Indiana 47907, USA,
annabellefeng@purdue.edu, shanthikumar@purdue.edu

We are living in an era in which data is generated in huge volume with high velocity and variety. Big Data and technology are reshaping our life and business. Our research inevitably needs to catch up with these changes. In this short essay, we focus on two aspects of supply chain management, namely, demand management and manufacturing. We feel that, while rapidly growing research on these two areas is contributed by scholars in computer science and engineering, the developments made by production and operations management society have been insufficient. We believe that our field has the expertise and talent to push for advancements in the theory and practice of demand management and manufacturing (of course, among many other areas) along unique dimensions. We summarize some relevant concepts emerged with Big Data and present several prototype models to demonstrate how these concepts can lead to rethinking of our research. Our intention is to generate interests and guide directions for new research in production and operations management in the era of Big Data.

Key words: Big Data; demand learning and planning; manufacturing; individualization; mass customization

History: Received: November 2017; Accepted: December 2017 by Kalyan Singhal, with no revision.

1. Introduction

Technology has dramatically changed our life and business in many ways. It is now possible to access, store, and process a massive amount of data, which we would not have imagined not so long ago. Traditionally, firms make decisions based on data that are directly coming from their day-to-day operations and transactions or that are known to the industries they are operating in and interacting with. Nowadays, such data is available in much larger scopes and in much greater details. Moreover, a huge amount of external sources of data has become accessible.

Many new frameworks and concepts have appeared and have stimulated tremendous attentions on Big Data, as well as associated technologies and business models, in industry and academia. As an immediate implication of Big Data, there are now a large amount of variables (co-factors, co-variates, and contexts) available to describe and predict, which aid improved decision making. The advancements in data mining and machine learning have significantly improved the accuracy of the descriptions and predictions derived from data. All these have intrigued researchers from computer science, statistics, information and communication technologies, and manufacturing science and technology to devote to the development of new technologies and management

principles that would lead to transformations of today's supply chains.

In this short essay, we would like to focus on two specific areas, namely, (i) demand learning and planning and (ii) manufacturing. These two areas have gained rapidly increasing interests from scholars in many different fields, while they seem to be, in our opinion, under-researched by Production and Operations Management community in recent years. We believe, with the fourth industrial revolution and the rapid development of internet-based business models, POM researchers can make unique contributions, which cannot be neglected by other fields, to these two areas.

An important trend enabled by Big Data and the associated technological developments, which is closely related to POM research, is *personalization* or *individualization*. Firms now possess a tremendous amount of data on individual customers, allowing for identifying their unique characteristics. It is now possible for firms to become innovative in creating distinct values to individual customers in an economical way. Values can be generated through personalized sales process, personalized product design, personalized production, and personalized services. For example, smart devices are now recording and communicating diverse data which can help firms to identify opportunities of innovative service using learning algorithms; with the advances in laser sintering

technology, the end customers may now create their own designs and have their products “printed.” We are truly entering an age of *mass customization*. As our discussion unfolds, it will become evident that many of the emerging concepts find their genetic roots to personalization or individualization.

While we discuss many concepts that are becoming popular, our focus is not to describe or speculate the industrial trends. Instead, we would like to make it explicit that research on issues or phenomena emerged with Big Data should not be a mere inclusion of data into our studies. We need to rethink our modeling and analysis to *integrate* the available data. Such an integration inevitably requires fundamental changes in the way we develop methodologies and applications. As Simchi-Levi (2014) states, “connecting the realms of data, models, and decisions will require our community to move out of our comfort zone and address intellectual challenges that demand bringing together statistics, computational science, and operations research techniques.” We present several prototype models to demonstrate how the thinking and modeling of POM issues may need to evolve in order to integrate the data and to incorporate the emerging concepts. The models presented below are only examples, which can be elaborated and enhanced in many different directions. Our intention is to entice contributions from POM researchers to areas that are traditionally within the core of POM but may have been neglected recently by some of us.

2. Demand Learning and Planning

Accurate demand forecast has always been a major puzzle in supply chain management. An industrial survey conducted recently (Chase 2013) suggests that the top focus area to use analytics and Big Data is improving demand forecasting and planning. Despite the presence of massive data, firms find it challenging to convert the data into actionable information for improved decision making. In this section, we review the research development in demand forecasting, and discuss how one may extract useful features from Big Data to improve predictive accuracy and facilitate prescriptive solutions.

2.1. From Aggregate to Individualized Demand Forecasts

A vast literature has been devoted to develop and analyze forecasting methods based on time series, including exponential smoothing (Gardner 1985), autoregressive models (Makridakis and Winkler 1983), diffusion models (Parker 1994), and Bayesian models (West and Harrison 1997). As a commonality among these methods, the inputs used are the (recent) history of demands, and their relationship to the

output, i.e., the future demand, is described through some functional form. A significant amount of effort in this area is devoted to compare and combine various forecasting methods (see, e.g., Clemen 1989, Petropoulos et al. 2014).

Improving demand forecasts using features (co-factors, co-variables, contexts, etc.) has been advocated in the very long past. Statistical techniques played a crucial part in the distant past, while statistical and machine learning methods are now used to learn, extract and select these features. For instance, when developing forecasting models for specific industries, domain knowledge can be used to refine the functional form used to fit the data. Use of domain knowledge in developing forecasts can be found in, for example, the study on semiconductor companies by Çakanyildirim and Roundy (2002) and in that on military distribution systems by Gardner (1990). In addition, information of the economic, social and climate environment in which the systems are operating can be helpful to improve forecasts. The demographic information is often used in retail demand forecast (e.g., Feng et al. 2013, Hoch et al. 1995) and weather index is often used in electricity consumption forecast (e.g., Carbonneau et al. 2008).

With the emergence and rapid growth of the Internet, decision makers are now swamped with a massive amount of accessible information, which enriches the data describing the environment of the operations. Real-time data on economic activities is also provided by many companies including Google, MasterCard, Federal Express, UPS, Intuit, among others. Such data is shown to help improving the demand forecasts. Goel et al. (2010) use web search data in linear or log-linear time series models to forecast the revenue generated from movie, video games and music; Choi and Varian (2012) use Google Trend data to predict the demand for auto parts; Yang et al. (2014) use web traffic data to predict hotel demands; Huang and Van Mieghem (2014) use online clickstream data to forecast the off-line demand for a door manufacturer; Cui et al. (2017) use Facebook data to predict daily demand for an online retailer.

Though theoretically more information leads to better forecasts, the challenge, however, comes from dealing with the increased number of variables and their ambiguous relationships. It is unlikely that the commonly used linear, loglinear, exponential or quadratic functional forms would fit all the variables. Identifying a functional form with good fit generally requires a large sample generated by repeated events. Though the amount of the data available is huge, the data collected, however, are often sparse and non-repetitive. Take the web traffic data as an example, the search on a trendy topic exhibits high frequencies only for a short period of time. Recently,

semi-nonparametric approach like quantile regression (see, e.g., Feng and Shanthikumar 2018b, Lu and Shen 2015) and nonparametric approach like regression tree (see, e.g., Ferreira et al. 2016), neural networks (see, e.g., Srinivasan et al. 1995) and support vector machines (see, e.g., Carbonneau et al. 2008) have been introduced. Many studies suggest machine learning approaches generally outperform the traditional time series analysis in forecast accuracy, though the improvement may not be always statistically significant.

The research on demand forecast, including the aforementioned studies, heavily focuses on describing the aggregate demands. Features carefully identified from Big Data can be used to reduce the uncertainties and improve the predictions at the aggregate level. More importantly, a great detail of individual customers' information is becoming available in Big Data, and it is now possible to predict the demands at the individual level. Individualized or personalized learning can facilitate demand planning and shaping at a much finer granularity than before. For the purpose of demonstration, we will present some simple models to obtain individualized forecasts and then discuss how firms may use such forecasts for planning. Our intention here is to put seeds for more indepth research in this area.

2.2. Individualized Demand Learning Models: A Retail Example

Nowadays, retail stores can collect a tremendous amount of information about their consumers. In the case of an online retailer, a consumer's entire purchasing history is recorded in his registered account. In the case of a brick-and-mortar store, an individual consumer can be identified through his membership information or payment methods. Such information allows the firm to understand and analyze the unique characteristics of a repeat shopper, which makes the prediction of individual's future preference and behavior possible.

2.2.1. A Model for Online Retailer. Take a typical consumer of an online store. At the time of registration, say, t_s , the consumer provides his basic information, which is denoted by an attribute vector, \mathbf{x}_0 . This information may contain the consumer's name, address, payment information, etc. Over time, the consumer visits the store website and leaves a history of browsing and purchasing. Suppose the consumer's n th visit of the store website happens at time a_n (with $a_1 = t_s$). The clickstream during this arrival is recorded in a vector, \mathbf{c}_n , which specifies the pages or products browsed and the sequence. If the consumer makes a purchase, then the order information is also recorded. We use S_n to denote the set of products

bought and $q_{n:i}$ the quantity of product $i \in S_n$ purchased. We would also like to include the entire product offering and the entire price list at the time of this consumer's arrival, as the available choice set may very well influence the consumer's purchasing decision. Specifically, the retail store knows the price $p_{n:j}$ and available quantity $y_{n:j}$ of each product $j \in J_n$ offered at that time, where J_n is the set of entire product offering. We also define $\mathbf{q}_n = (q_{n:1}, q_{n:2}, \dots, q_{n:|S_n|})$, $\mathbf{p}_n = (p_{n:1}, p_{n:2}, \dots, p_{n:|J_n|})$ and $\mathbf{y}_n = (y_{n:1}, y_{n:2}, \dots, y_{n:|J_n|})$. If the consumer visits the website without a purchase, then $S_n = \emptyset$ and only the clickstream data \mathbf{c}_n are recorded. We use $\mathbf{z}_n = (\mathbf{c}_n, S_n, \mathbf{q}_n, J_n, \mathbf{p}_n, \mathbf{y}_n)$ to denote the information collected on the consumer's n th visit. With a slight abuse of notation, we use $n(t)$ to denote the number of visits the consumer makes up to time t .

The entire history of one consumer, i.e., $\{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_{n(t)}\}$, can contain a massive amount of variables. In addition, the retailer may be closely monitoring the economic and social trends as well as the offerings of its online competitors, which may also play a role in influencing the consumer's decision. For ease of exposition, we do not explicitly include such environmental data, as the same methodology can be applied to the model involving such data.

In any case, fully describing a consumer with all the available data requires dealing with a large number of variables. To characterize and predict the behavior of a consumer, one would extract and select critical features of this consumer, denoted by \mathbf{x}_t , from the raw data \mathbf{x}_0 and $\{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_{n(t)}\}$. Feature extraction and selection help to improve the efficiency of data storage and processing. The commonly used methods include filters, search algorithms, regularizations, fuzzy neurons, Bayesian neural networks, support vector machines, etc. Guyon et al. (2006) provide an excellent summary of basic feature extraction and selection strategies as well as recent advancements in this area. Additional updates on feature selection techniques can be found in Chandrashekar and Sahin (2014) and Miao and Niu (2016). It is important to keep in mind that the extraction and selection of features depend heavily on the goal. Even from the same data set, the groups of features identified can be very different for different predictive variables.

In our context of a retail consumer, an elaborate model for the consumer's visit and purchase pattern may assume that $\{(\mathbf{a}_n, \mathbf{z}_n), n = 1, 2, \dots\}$ form a Markov renewal process. Depending on the application context, the state space of this Markov renewal process can be reduced by applying feature extraction and selection techniques mentioned before on \mathbf{z}_n to obtain a vector \mathbf{w}_n with a manageable size.

Now suppose that the store has m registered customers at time t and the transformed data for the j th

consumer is $\{(a_n^{(j)}, \mathbf{w}_n^{(j)}), n = 1, 2, \dots, n^{(j)}(t)\}$ with extracted set of features $\mathbf{x}_t^{(j)}$. We can estimate or predict the next arrival time of a consumer and his choice of products based on the products and prices to be offered. For example, if we choose to use the maximum likelihood estimator, the log-likelihood of the next arrival time for all m consumers can be computed as

$$\sum_{j=1}^m \sum_{k=1}^{n^{(j)}(t)-1} \log (f_A^{(j)}(a_{k+1}^{(j)} - a_k^{(j)}, \mathbf{w}_k^{(j)}, \mathbf{x}_t^{(j)})) - \sum_{j=1}^m \log (\bar{F}_A^{(j)}(t - a_{n^{(j)}(t)}^{(j)}, \mathbf{w}_{n^{(j)}(t)}^{(j)}, \mathbf{x}_t^{(j)})),$$

where $f_A^{(j)}(\cdot, \mathbf{w}_k^{(j)}, \mathbf{x}_t^{(j)})$ is the density function of the next arrival of consumer j after the k th arrival given the consumer's clickstream and purchase information $\mathbf{w}_k^{(j)}$ and the extracted features $\mathbf{x}_t^{(j)}$ at time t , and $\bar{F}_A^{(j)}$ is the corresponding survival function.

Assuming a proportional hazard rate model for all individual consumers or for a clustered subset of consumers, a parametrized hazard rate function, and a parametrized function that maps (\mathbf{w}, \mathbf{x}) to the proportion, one may compute the maximum likelihood estimates for the parameters. In addition, we may estimate the transition probability for the consumer's purchases. That is, $p(\mathbf{w}, \mathbf{w}_k; \mathbf{x}_t) \equiv P\{\mathbf{W}_{k+1} = \mathbf{w} | \mathbf{W}_k = \mathbf{w}_k, \mathbf{x}_t\}$. Hence for a consumer with feature vector \mathbf{x}_t , whose last arrival before t is at time τ with click and purchase data \mathbf{w} , we can obtain the density function of his arrival time after time t as

$$\frac{f_A(s + (t - \tau), \mathbf{w}, \mathbf{x}_t)}{\bar{F}_A(t - \tau, \mathbf{w}, \mathbf{x}_t)}, \quad s \geq 0,$$

and his next visit and purchase probability as $p(\hat{\mathbf{w}}, \mathbf{w}; \mathbf{x}_t)$.

If one prefers a non-parametric approach, techniques such as Kernel smoothing regression (Warnd and Jones 1995) could be used. Furthermore, by appropriately using the full data set and state space reduction, it is also possible to estimate lost sales and consumer substitutions in the event of stockout. For example, we may find that a consumer, who has regularly bought a specific product (i.e., a product with high purchase probability) that is out of stock, purchases a similar item that he rarely picks (i.e., a product with low purchase probability). With the above model, such an event can be identified, which can then be used to estimate the lost sales on a product and the resulting substitution behaviors. In other words, the data-integrated personalized demand model can lead to a new way of learning and planning under demand censoring.

2.2.2. A Model for In-Store Purchase. Compared with online purchase, an important feature of in-store shopping lies in its periodic pattern. For example, a particular consumer may mostly shop during week-ends, while occasionally picking up items during the middle of a week. In this case, it is natural to model a weekly purchase demand pattern.

Consider the situation where the store is able to track the purchases of a consumer starting from week t_s . The record of a consumer reveals whether this consumer has arrived ($a_n = 1$) or not ($a_n = 0$) during week $t_s + n$. In the event of the consumer arrival, the set of items purchased S_n and the quantity purchased \mathbf{q}_n are also recorded. Like in the case of the online store, the retailer knows the store offering J_n as well as the corresponding prices and inventories $(\mathbf{p}_n, \mathbf{y}_n)$.

Based on our discussion of the previous model, the retailer is able to extract feature \mathbf{x} to characterize this consumer. Let $p_A(1, S, \mathbf{q}, J, \mathbf{p}, \mathbf{y}, \mathbf{x})$ be the probability that the consumer with features \mathbf{x} will visit the store in the next week given that this consumer has arrived in the current week, purchased the set of products S in amounts \mathbf{q} , and the store has offered the set of products J with list prices \mathbf{p} and inventories \mathbf{y} at this consumer's arrival. We also define $p_A(0, J, \mathbf{p}, \mathbf{y})$ to be the probability of a visit by the consumer with features \mathbf{x} in the next week given that this consumer has not visited in the current week. In the same token, we can further estimate the exact visit day and purchase quantities of a consumer conditioning on the event that he will visit in the next week. Specifically, given that the consumer will come in the next week, let $p_{day}(t), t = 1, 2, \dots, 7$, be the probability that the consumer will arrive on the j th day of the next week and $D_j(p_j), j \in J$, be the quantity of product j to be purchased by the consumer when the price for product j is set at p_j .

We would also like to point out that personalized consumer characterizations can help the retailer to better predict the behavior of a new consumer who does not have a long purchase history, or to better forecast the selling of a new product which is on shelf for only a short period of time. Techniques like clustering (Jain and Dubes 1988), support vector machines and Kernel smoothing (Warnd and Jones 1995) can be used to match the features of the new consumer to those with longer purchasing histories for a specific purpose (e.g., adoption of a certain product). Recent development in this area has suggested that with the massive amount of individualized data, segmentation of consumers based on Big Data can lead to categorizations that are very different from those generated in the past—Consumers who are seemingly alike (e.g., with similar demographic features and visit patterns) may behave very differently in

their choices of specific products (see, e.g., Zhang et al. 2017).

2.3. Demand Shaping with Individualized Learning

Certainly, the individualized forecasts can be consolidated to generate the aggregate forecast. In spirit, this is similar to the approach used in many utility-based consumer models, where the firm's demand function is derived by aggregating consumers' choices based on their respective utility functions (see the examples discussed in Talluri and Van Ryzin 2004).

This, however, seems to go against the commonly held notion in supply chain management that aggregations along, say, products, regions and time, help to improve forecast accuracy when individual demands are not positively correlated. This notion has shown to reduce supply-demand mismatch for decisions with long lead time. In contrast, based on individualized short-term forecast, retailers now can make daily adjustments on shipment schedule, merchandising, advertisement, and pricing. Moreover, individualized demand forecasts have shown to enable firms to use personalized pricing (e.g., Liu and Zhang 2006) and personalized marketing (e.g., Bauer et al. 2005, Shankar et al. 2016) to increase sales and revenue. Following the discussions in section 2.2, the demand models derived can be directly used to predict how an individual consumer would react to changes in price, product offering or even web page design. The retail firm thus can use these models to decide personalized product information or promotion deals offered to specific consumers. For example, using the prediction of each individual consumer's store visit day and purchase pattern in section 2.2.2, the retailer can now try to nudge the consumers to come on a specific day of a week by sending coupons valid only for that day. This way, the retailer can smooth out the store traffic and the demands for certain products throughout the week to improve staff scheduling and shelf space management. Furthermore, when the stock of an item is running low, the retailer may choose to entice a small identified group of consumers from whom a great revenue can be generated. When excess inventory is found for a newly introduced product, the retailer may give promotions only to selected consumers who are likely to generate continuous future values.

More importantly, the benefit of aggregated demand forecast comes under the premises of the demand-driven planning framework, an approach adopted by most firms. Under this framework, demand forecasts are generated, marketing and sales plans are derived, and then resource and production planning are conducted based on the marketing parameters. With the knowledge of individual

consumers, firms are now ready to shift from the marketing-lead paradigm to a full integration of sales and operations planning. Both the environmental features and individual customer features are likely to change rapidly. It is crucial that when adapting to observed changes in demand patterns, the firm integrates the the marketing strategies with resource and inventory decisions. On the implementation front, the new planning paradigm requires an integration of predictive tools with prescriptive solutions, which invokes the rethinking of forecasting and decision making models to facilitate a direct transformation of a massive amount of consumer data into efficient decisions.

2.4. Preference Learning and Planning

Learning about customer preferences on product functionalities or service offerings is an integral part of demand learning. Customers' behaviors and responses have always been important input for the successful designs of products and services (Bloch 1995). Many researchers in marketing have examined how various factors may influence consumer preferences.

With the wide use of search engines, consumer reviews, blogs and so on, researchers have developed various techniques for opinion mining and sentiment analysis (see, e.g., Pang and Lee 2008) to extract useful information to understand the trend of consumer preferences, which may guide firms in their product design and production processes. Moreover, with the emergence of Internet of Things (IoT), an increasing number of sensors and communication devices are built into the products. The engineers can now collect data to monitor and understand how customers are using the products in the field, for example, whether certain functionalities are never used and whether the products are likely to break down under certain way of use. Such data, together with analytics, can provide accurate predictions of customer preferences on design features and help firms to improve the design and process.

Consumers' need for uniqueness has been documented in marketing studies (e.g., Tian et al. 2001). Personalized designs are becoming popular in areas like apparel, medicine, tourist trip and cars. Digital shoppers, in particular, are increasingly demanding personalized service and experience (see, e.g., Amato-Mccoy 2017). Big Data and IoT allow firms to better understand consumers' self-perceptions of uniqueness and the role their purchasing decisions play in expressing personal identity.

Before ending this subsection, we would like to emphasize that, though we focus much of our discussion in retail contexts, the concept of personalization is not limited to consumer interfaces. Industries have

moved toward personalized demand learning and planning even in business-to-business interactions (see, e.g., Balci and Cetin 2010). Moreover, as we discuss personalization, we mostly take the angle of a firm's planning to satisfy the customers. With the growing development of technologies for online marketplaces, there soon will come a time where individuals, with guaranteed anonymity and security, can "sell" their personal data of product exploration and purchasing behavior. This, in effect, can create competition among firms to supply specific products that the customers are looking to have (see, e.g., Wolf 2017) or accurately advise the customers on the timing and quantity availability. Such a trend would lead to a completely different landscape for demand learning and planning.

2.5. A Newsvendor Demonstration of Big Data Enabled Forecasting and Planning

In this subsection, we use the newsvendor model as the workhorse to show how Big Data can change the demand modeling and operational planning, as well as the possible evolution of research. The development below is intended to provide a concrete demonstration of the concepts and ideas discussed previously in a simple setting and to generate interests in further research along these directions. More detailed analysis and results on data-integrated newsvendor models can be found in Feng and Shanthikumar (2017).

The newsvendor model has served as a basis to explain phenomena, explore strategies and develop understandings in operations management. In the classical newsvendor model, the demand is specified in an aggregate way as a random variable D and the product (newspaper) is generic with no design variations. The newsvendor, when making a quantity decision y , obtains a profit of

$$\Psi(y, D) = s \min\{y, D\} - cy,$$

where s is the per-unit revenue and c is the per-unit cost. The goal is to maximize the expected profit defined as

$$\phi(y) = \mathbb{E}[\Psi(y, D)] = s \int_0^y \bar{F}_D(x) dx - cy,$$

where \bar{F}_D is the survival function of random variable D .

In what follows, we introduce the design parameters and the market or environmental features into the newsvendor model separately in sections 2.5.1 and 2.5.2 to demonstrate how one may use available data to enhance the aggregate forecast and planning. Then in section 2.5.3, we show how to reconcile all the available data and enable personalized newsvendor

planning. Finally, we include the aspect of supply shortage in section 2.5.4 and show how a personalized newsvendor can appropriately ration the limited stock by enticing the "right" customers to purchase.

2.5.1. A Design Focused Newsvendor. There are many extensions of the newsvendor model. The most studied one is the pricing newsvendor model, where the demand is price-dependent. That is, D is a stochastic function of the price s chosen. To make some connection to our discussion in section 2.4, one may also consider design parameters of the product, denoted by a vector \mathbf{w} . Due to the complexity of forecasting the joint demand distribution for a large number of design alternatives and the challenge of dealing with the issues in traditional manufacturing (e.g., costly change over, efficiency reduction with reduced batch size), it is natural to limit the number of alternative designs. Suppose we can choose only one design and $D(s, \mathbf{w})$ is the aggregate demand. Using the techniques discussed previously, the newsvendor can learn through experiment, expert opinion or customer input to identify a model describing $D(s, \mathbf{w})$. With this knowledge, the newsvendor's profit becomes

$$\Psi(s, y, D(s, \mathbf{w}), \mathbf{w}) = (s - \delta(\mathbf{w})) \min\{D(s, \mathbf{w}), y\} - c(\mathbf{w})y,$$

where $\delta(\mathbf{w})$ and $c(\mathbf{w})$, respectively, represent the costs of selling and producing the product with design variation \mathbf{w} compared with the generic product. The problem then becomes one of choosing the optimal quantity decision y , price decision s and design decision \mathbf{w} to maximize the expected profit expressed as

$$\phi(s, y, \mathbf{w}) = (s - \delta(\mathbf{w})) \int_0^y \bar{F}_D(x; s, \mathbf{w}) dx - c(\mathbf{w})y.$$

2.5.2. An Information Gathering Newsvendor.

Echoing the discussion in sections 2.1 and 2.2, the newsvendor may use available environmental information including web search data, industry trend, and competitors' moves to improve the demand forecast. Suppose V is the set of features that are selected based on all the available data. These features V take values $\mathbf{V} \in \{\mathbf{v} : \mathbf{v} \in \mathcal{V}\}$ for some state space $\mathcal{V} \subseteq \mathbb{R}^{|V|}$. Statistical and machine learning approaches discussed in section 2.2 can provide the predictive characterization of the demand $D(s, \mathbf{v})$. The newsvendor's profit is now defined as

$$\Psi(s, y, D(s, \mathbf{V})) = s \min\{D(s, \mathbf{V}), y\} - cy,$$

and the expected profit as

$$\begin{aligned} \phi(s, y; \mathbf{v}) &= \mathbb{E}[\Psi(s, y, D(s, \mathbf{V})) | \mathbf{V} = \mathbf{v}] \\ &= s \int_0^x \bar{F}_D(x; s, \mathbf{v}) dx - cy. \end{aligned}$$

This model could be seen as a version of the Big Data newsvendor (see, e.g., Ban and Rudin 2017), as it requires a predictive characterization of the demand model based on the massive data available. Next, we will add an important layer to this model by integrating the design parameters described in section 2.5.1, which enables personalization.

2.5.3. A Big Data Newsvendor with Personalized Design. As we demonstrate in section 2.2, Big Data with statistical and machine learning algorithms allows for identification of features that accurately predict an individual’s behavior towards a certain purchase (e.g., purchasing a specific item at a certain price or choosing one item over others). Moreover, with the emerging technologies advancing smart manufacturing (see more discussions in section 3), customized products can now be produced efficiently. Firms are able to use the knowledge learned from individual customers to offer products that are catered toward unique preferences.

Suppose that the newsvendor has identified such characterizations of m high-value customers. These high-value customers may have generated a large revenue in the past and have shown significant loyalty. Also, they highly value uniqueness or personal identity in the sense that they may exhibit strong individual preferences on particular functionalities of the product. Specifically, for each customer j in this set, the newsvendor has a fairly accurate prediction of the probability $p_j(s, \mathbf{v}, \mathbf{w})$ that customer j would purchase a product at price s when the environmental features take the values \mathbf{v} and the design parameters are \mathbf{w} .

The newsvendor would like to offer personalized designs and prices targeting the high-value customer. In the meanwhile, outside of the group of high-value customers, general customers form an aggregate demand $D_{m+1}(s, \mathbf{v}, \mathbf{w})$, which the newsvendor may learn using the framework discussed in section 2.2. Suppose that the newsvendor orders $y(\geq m)$ units of the generic product (\mathbf{w}_{m+1}), which takes a significant lead time. Upon receiving the units, the newsvendor can approach each high-value customer to offer a personalized variation $\mathbf{w}_j, j = 1, 2, \dots, m$. If a high-value customer accepts the offer, a unit is modified at a cost $\delta(\mathbf{w}_j)$ and sold to that customer. Otherwise, that unit would be made available to general customers in $D_{m+1}(s_{m+1}, \mathbf{v}, \mathbf{w}_{m+1})$. Thus, the Big Data newsvendor with personalized design makes a profit of

$$\begin{aligned} s_{m+1} \min & \left\{ D_{m+1}(s_{m+1}, \mathbf{v}, \mathbf{w}_{m+1}), y - \sum_{j=1}^m D_j(s_j, \mathbf{v}, \mathbf{w}_j) \right\} \\ & + \sum_{j=1}^m (s_j - \delta(\mathbf{w}_j)) D_j(s_j, \mathbf{v}, \mathbf{w}_j) - cy, \end{aligned}$$

for $y \geq m$. Here $D_j(s_j, \mathbf{V}, \mathbf{w}_j)$ is the personalized demand with $\mathbb{E}[D_j(s_j, \mathbf{V}, \mathbf{w}_j) | \mathbf{V} = \mathbf{v}] = p_j(s_j, \mathbf{v}, \mathbf{w}_j), j = 1, 2, \dots, m$. Note that we have assumed that a high-value customer, if not purchasing the personalized product, will not purchase the generic product.

While we consider m high-value customers with unique preferences, the above model can be easily recast to consider m classified groups of high-value customers. Many clustering methods can be used to categorize the customers in terms of their preferences in purchasing the product considered. Refinements along several other directions are also possible. For example, in the above model, we have assumed that the high-value customers exhibit strong personal preferences on the design and they are unlikely to choose a generic product regardless of the price differences between the products. In other words, high-value customers do not substitute. The consideration of substitution would require one to predict the high-value customers’ demands when only the generic product is offered. Specifically, suppose $\hat{D}_j(s_{m+1}, \mathbf{v}, \mathbf{w}_{m+1})$ is the demand for the generic product (\mathbf{w}_{m+1}) from high-value customer j . We must have $D_j(s_j, \mathbf{v}, \mathbf{w}_j) + \hat{D}_j(s_{m+1}, \mathbf{v}, \mathbf{w}_{m+1}) \leq 1$, as the customer will purchase at most one unit. Then, the above profit function is modified to

$$\begin{aligned} s_{m+1} \min & \left\{ D_{m+1}(s_{m+1}, \mathbf{v}, \mathbf{w}_{m+1}) + \sum_{j=1}^m \hat{D}_j(s_{m+1}, \mathbf{v}, \mathbf{w}_{m+1}), \right. \\ & \left. y - \sum_{j=1}^m D_j(s_j, \mathbf{v}, \mathbf{w}_j) \right\} \\ & + \sum_{j=1}^m (s_j - \delta(\mathbf{w}_j)) D_j(s_j, \mathbf{v}, \mathbf{w}_j) - cy. \end{aligned}$$

In addition, the Big Data newsvendor with personalization may potentially offer individualized prices to every customer (including the general group). We will discuss this possibility in the context of limited supply in the next subsection.

2.5.4. A Big Data Newsvendor with Limited Supply. Now suppose that there are only one design of the product and the design parameters are defined by the supplier, who offers a limited amount to the newsvendor. Specifically, let $D(s)$ be the aggregate demand the newsvendor is facing when a price s is

posted. Without supplier limitation, the newsvendor’s optimal pricing and quantity decisions are

$$(s^*, y^*) = \arg \max \left\{ s \int_0^y \bar{F}_D(x; s) dx - cy : \underline{s} \leq s \leq \bar{s}, y \geq 0 \right\},$$

where \underline{s} and \bar{s} are lower and upper bounds of a feasible price. Supply shortage suggests that the supplier can only offer a quantity $y < y^*$. If the newsvendor chooses to receive all y units from the supplier, the classical newsvendor would choose the following optimal price

$$\hat{s}(y) = \arg \max \left\{ s \int_0^y \bar{F}_D(x; s) dx : \underline{s} \leq s \leq \bar{s} \right\}.$$

However, the newsvendor would like to set a higher price $s_0 > \hat{s}(y)$, while at the same time trying to sell all y units. With aggregate planning, such a policy is certainly suboptimal. A Big Data newsvendor with personalization may actually achieve a better profit by posting a price s_0 than by posting a price $\hat{s}(y)$, yet selling all y units. Specifically, the newsvendor would attempt to obtain personalized features \mathbf{x} that characterize the individual customers. Let \mathbf{x}_j be the feature vector obtained for customer j (e.g., through past purchasing behavior), and $p(s, \mathbf{v}, \mathbf{x}_j)$ be the predicted probability that customer j would purchase the product at price s when the environmental feature vector is \mathbf{v} . In cases where such characterization is available for all customers (Zhang et al. 2017), the Big Data newsvendor can potentially offer personalized prices to each customer. Such offers can be implemented by setting a common list price s_0 and providing selected customers with individualized coupons which give discounts over the list price.

If the individualized coupons are offered to the selected customers at the same time, the individualized demand is still uncertain to the newsvendor, as customers’ purchases are predicted in probabilities. A potential strategy that the newsvendor can adopt is to identify a set $R \subseteq \{1, 2, \dots, m\}$ of customers who are price sensitive but are insensitive to the stockout of this particular product. Coupons (i.e., price discounts) that can only be redeemed at the end of the selling season are offered to these customers. Let $\bar{R} = \{1, 2, \dots, m\} \setminus R$. Then the Big Data newsvendor’s profit becomes

$$\max \left\{ \begin{array}{l} s_0 \mathbb{E}[\min\{D_T(\bar{R}, s_0), y\}] \\ + \mathbb{E} \left[\frac{(y - D_T(\bar{R}, s_0))^+}{\max\{(y - D_T(\bar{R}, s_0))^+, D_T(R, s)\}} \cdot \sum_{j \in R} s_j D_j(s_j, \mathbf{v}, \mathbf{x}_j) \right] \end{array} \right\},$$

where

$$D_T(\bar{R}, s_0) = \sum_{j \in \bar{R}} D_j(s_0, \mathbf{v}, \mathbf{x}_j) \text{ and}$$

$$D_T(R, s) = \sum_{j \in R} D_j(s_j, \mathbf{v}, \mathbf{x}_j)$$

are, respectively, the aggregate demand generated by customers without coupon and that by customers with coupons. Because the customers with coupons must wait until the end of the selling season for redemptions, there is a chance that some of these customers find the product sold out.

In the above formulation, the customers seeking to redeem their coupons are randomly assigned the available inventory in the event of stockout. In view of this, one may, of course, extend the above model by incorporating strategic customers’ behaviors. Specifically, a customer with coupon may choose between purchasing at the full price and waiting to use the coupon by trading off his desire of possessing the product and his sensitivity to price reduction. It is not hard to see that a Big Data newsvendor with personalization is poised to reap a much higher profit than the classical or even the information gathering newsvendor is.

With today’s technologies, it is possible to further improve the implementation of personalized offers. Specifically, the newsvendor can choose customers in sequence, assign a price, observe the outcome and then choose the next customer (provided that the newsvendor chooses a price-only strategy without further interactions with individual customers; see Feng and Shanthikumar 2018a). Such a sequential selling process is now viable with the wide adoption of smart mobile devices and the internet. The newsvendor can push the individualized offers directly to customers through text messaging or email, and can have the customers committing purchases through direct mobile or online payment. With the inventions of technologies like Blockchain (Pilkington 2016), such transactions can be implemented in trusted and even anonymous ways.

To describe the sequential selling decisions, a dynamic programming formulation is needed for the Big Data newsvendor. Specifically, let $v(i, \bar{R})$ be the newsvendor’s optimal profit when there are i items left and \bar{R} is the set of customers who have not yet received an offer. Then, the optimality equation is

$$\begin{aligned} v(i, \bar{R}) = \max \{ & p_j(s, \mathbf{v}, \mathbf{x}_j)(s + v(i - 1, \bar{R} - j) \\ & - v(i, \bar{R} - j)) + v(i, \bar{R} - j) : j \in \bar{R}, \underline{s} \leq s \leq s_0 \}, \\ & i = 1, 2, \dots, y, \quad \bar{R} \subset \{1, 2, \dots, n\}. \end{aligned}$$

The dynamic model provides an efficient way to sell the limited supply y with the most profit through sequential price discrimination.

Realizing that the newsvendor is utilizing Big Data in demand learning and planning, the supplier may react accordingly. It would be interesting to investigate how personalization strategies adopted by the newsvendor may affect the supplier's decisions (e.g., contract price, capacity, or even directly personalizing the sales to the newsvendor). Moreover, while personalization is seen as a direction to build the core competence by many firms, myopically implemented strategies can backfire. Anecdotal evidence from online retailing has suggested that an aggressive push of personalized recommendation may lead to reduced sales of the recommended products. Also, consumers who take personalized offers may tend to reduce their visits to the store website, leading to reduced co-selling of other products. The consideration of Big Data with personalization opens up many new avenues for research in supply chain interactions.

3. Manufacturing

In manufacturing, the internet-triggered revolution, termed Industry 4.0, is expected to change the landscape of the industries (Tomlin 2017). The most important directions along this are represented by Cyber-Physical Systems (CPSs) and Industrial Internet of Things (IIoT). CPSs offer integrations of computation, networking, and physical processes (Khaitan and McCalley 2015). Moreover, the virtual and physical worlds are growing closer together to form the Internet of Things (Monostori 2014).

In today's manufacturing environment, sensors, processors, actuators, and communication devices are often in place to collect and process real time data about the machines, the processes, and the products. CPS would enable the communications between machines, between machines and products, and between machines and humans. Machines may become self-controlled on certain tasks, as well as interacting with humans through interfaces. Via learning and interaction over networks, both factories and products become *smart*. Areas including robotic systems, smart grid, autonomous automobile, and medical monitoring are among the first adopters of these concepts. With the advancement in automation using sensors and networked machines, an extreme amount of data is continuously generated.

In the classical automation pyramid, the communication hierarchy moves from the sensor level up to the device level, control level, plant level and eventually the enterprise level, and the exchanges often occur only between adjacent levels. The CPS-based automation, in contrast, is breaking the communication boundaries of the hierarchical levels, transforming the pyramid into a network. In other words, the

vision is an evolution toward autonomous and social production systems.

In the last decade, the related research has rapidly emerged and grown in computer science, statistics, information and communication technologies, and manufacturing science and technology. Many of the studies there have intricate connections to our research. For example, an enabler for automation is the design of learning architectures and learning algorithms that are adaptive in response to significant events while remaining stable when facing irrelevant events. Moreover, the changes in information communication and learning structure entail the need for deviations from the traditional hierarchical planning approaches.

3.1. Contributions Needed from Production and Operations Management

The focus of the computer science and engineering literature is on developing *predictive* informatics tools to manage Big Data. Adaptation and integration of these predictive tools into prescriptive decisions are where we believe POM researchers can contribute. We identify three important aspects that are closely related to our expertise. They are individualized production, integration and coordination over distributed networks, and connecting production and services.

3.1.1. Individualized Production. The idea of individualized production is not new. Considerable attentions were given to this topic around early 1990's from researchers in manufacturing and service operations and in marketing. Techniques for mass customization, including modularization, setup reduction, and postponement, have been taught in operations management classes.

In the smart manufacturing environment, mass customization is leading its way through. Techniques like additive manufacturing (e.g., 3D printing) and reconfigurable manufacturing have enabled quick reaction to changing customer requirements with cost-effective production. The new information communication, processing, and analytics tools now reconcile knowledge of manufacturing process and consumer applications, making customized products also intelligent. All these have expanded the landscape of mass customization.

Recently, a few researchers have paid attention to new operational phenomena induced by individualized production techniques. For example, Song and Zhang (2016) study management of inventory and logistics systems when 3D printing are used to produce spare parts. Dong et al. (2017) analyze assortment decisions after adopting 3D printing. Additional research on understanding the process, variety, scheduling and inventory management would make contributions to shape the trend in industry.

3.1.2. Integration and Coordination over Distributed Networks. With the increasing complexity of products and processes, integration has always been the focal point in the development of Industry 4.0. Practitioners and academics identify the challenges in *horizontal integration* through value network, *vertical integration* of networked manufacturing, and *end-to-end digital integration* of engineering across the entire value chain systems (Kagermann et al. 2013).

Data availability and visibility throughout the entire network is a prerequisite to build and integrate the analytics capabilities (Srinivasan and Swink 2017). This requires extensive real time information sharing across the boundaries of legally independent organizations, such that these organizations can share competence and exploit increased values and opportunities, eventually forming virtual corporations (Davidow and Malone 1992). For small and medium enterprises (SME), who each have limited resources and capabilities, collaborative development and manufacturing environments have become particularly important for their survival and success. In cloud-based design and manufacturing systems (Wu et al. 2015), resources are highly distributed over the network. To sustain collaborations, the business model has shifted from offering superior products to collectively offering superior manufacturing capability (Brettel et al. 2014).

Two aspects associated with this shifting business model are particularly relevant to POM research, namely, *information exchange* and *incentive coordination*. While we all know that information sharing enables collaboration efficiency, long-standing and significant challenges are yet to be addressed. One obstacle hindering seamless collaborations is the lack of trust when data is transmitted across organizational boundaries. This is especially the situation when a collaborator may also compete on the market. Even otherwise, firms concern the implied risks through data sharing on security of proprietary information and protection of intellectual property. Studies on SMEs suggest that, while information sharing can trigger innovation, asymmetric learning can lead to learning races (see, e.g., Bounckena and Kraus 2013). These issues, resulting in high coordination costs, have led to the failures of many collaborations. For example, more often than not, the data from sensors on machines are collected by the equipment supplier, while the data from sensors on products are collected by the manufacturer. Due to the concern of leaking proprietary information, equipment suppliers are reluctant to share the real time machine status, which may very well affect the process productivity and quality. Though there has been a large literature in operations management on coordination through appropriate contracting, little on

collaboration with data exchange mechanism is studied.

On the coordination front, the goal of a collaborative network is to balance and consolidate limited resources across different organizations to expand the overall capability. This requires optimization of resource allocation and capability investment in a distributed network. Understanding the distributions of resource, information, responsibility, and decision authority requires rethinking of the coordination models studied in supply chain management. Approaches like evolutionary games (see, e.g., Friedman 1991), which have been extensively used in economics, computer science, social science, and physics, can be introduced to understand the dynamic adaptation and evolution of new technology in collaborated networks.

3.1.3. Connecting Production and Services. An envisioned smart manufacturing system involves machines that are self-aware, self-learning and self-maintained. Machines would be able to actively suggest task arrangements and adjust operational parameters to maximize productivity and quality. Moreover, machines could assess their own health and use information from their peers for maintenance or adjustment to avoid potential issues (Lee et al. 2014). With these capabilities, real time data from outside of the organization (e.g., suppliers or end customers) can be directly adopted to enhance the manufacturing processes in order to eventually improve the experience of the end customers. Though self-learning machines are yet far from wide implementation, it is recognized in industries that service is becoming an increasing part of production. The concept of manufacturing servitization (Vandermerwe and Rada 1989), advocating customer focus, emphasizes the combination of products, services, support, and knowledge in developing innovated value-added service to enhance core competence.

With an increasing amount of sensors and communication devices placed on smart products, production engineers can now gather real time information on how end customers are using the products in the field. Such data allows manufacturing facilities to quickly adjust the production plan and quality control, or even process design and product design, to better service the end customer.

3.2. Example: Collaborative Learning in Machine Maintenance

In this subsection, we use the planning of machine maintenance as an example to illustrate how some of the aspects discussed in the previous subsection can lead to management approaches that are different from the classical models.

Big Data learning from IIoT (e.g., sensors) for machine maintenance and root cause analysis has gained momentum recently. According to a McKinsey report (Dilda et al. 2017), predictive maintenance typically reduces machine down time by 30%–50% and increases machine life by 20%–40%. In general the learning for such predictive maintenance uses sensor measurements and machine status. The learning about maintenance is generally conducted by the machine tool suppliers, while the learning on process and quality improvement is generally performed by the product manufacturer. Data sharing between the supplier and the manufacturer is often stopped by the concerns of leaking proprietary information. Also, the machine tool suppliers often worry that unexpected fluctuations in sensor collected data may lead to immediate complaints from the manufacturers, hurting the perceived quality and brand image. In many situations, however, the process information is useful for the equipment supplier to understand the factors influencing machine performance, and the machine status information is valuable for the manufacturer to guide process improvements and production schedules. Due to the lack of data exchange, it is very typical that the machine tool supplier, who offers service contracts and takes the responsibility for maintaining the equipment, would trigger a maintenance solely based on the machine sensor measurements. While this approach of maintenance planning may work well in automatic transfer lines (see, e.g., Buzacott and Shanthikumar 1992) that produce the same parts in repetitive cycles, it can be suboptimal for multi-product manufacturing facilities.

Let us focus on a manufacturing facility (e.g., a job shop or a flexible manufacturing system) that produces multiple product types. To effectively and efficiently manage the line productivity, a predictive algorithm to identify maintenance timing is needed. Certainly, the machine tools have their intrinsic aging processes (see, e.g., Cinlar et al. 1989). The levels of stress, wear and energy use placed on the machine tools are generally different when producing different products. These factors play a crucial role in affecting machine deteriorations. Therefore, a pure time-based aging model cannot accurately predict the health of the machine tools. Instead, one needs to understand the product mix and its implication on the aging processes. In addition, it may be appropriate to perform a maintenance before the production of a specific product that induces high stress, wear or energy requirements on the machine tool, but to postpone maintenance to after the production of a product that is less demanding. Take an example from the semiconductor manufacturing process (Nurani et al. 1996). The chamber of an etcher may get dirty with tiny particles and may need to be cleaned. The

particles in the chambers can potentially cause yield loss to chips with a small line width and line spacing, while they may not affect the yield for less sensitive chips (i.e., those with a wider line width and line spacing). Hence learning for predictive maintenance has to be carried out in conjunction with the production schedule.

In the reminder of this subsection, we first present a basic model of stochastic machine maintenance. Then, we add layers to this model to demonstrate how collaborative learning with Big Data and IIoT can lead to rethinking of the classical planning approach.

3.2.1. Sensor Data Triggered Machine Maintenance. Suppose the sensor identifiable states of the machine are $\{1, 2, \dots, m\}$ with 1 denoting the healthiest or new state and m representing the failure state. Careful monitoring of the machine using sensors has revealed an aggregate transition probability matrix $\mathbf{P} = (p_{ij})$, $i, j \in \{1, 2, \dots, m\}$ (see details in Sloan and Shanthikumar 2000). The knowledge \mathbf{P} is aggregate because it is derived based on the historical sensor data. The actual production schedule involving product types is not integrated into the learning of \mathbf{P} .

Now suppose we are producing n different product types at a required ratio of $\gamma_1 : \gamma_2 : \dots : \gamma_n$ with $\sum_{j=1}^n \gamma_j = 1$. This ratio reflects the mix of the product demands that the firm anticipates from its customers. Product type j brings a profit of r_j and the aggregate yield of product j is y_j .

At the beginning of period t , the machine status $X_t \in \{1, 2, \dots, m\}$ is observed. If the machine fails, i.e., $X_t = m$, repair is performed at a cost c_R to bring the machine state back to new, i.e., $X_{t+1} = 1$. If the machine is working properly, i.e., $X_t < m$, a decision needs to be made in terms of whether or not to schedule a preventive maintenance in the current period. If so, the machine state is restored to $X_{t+1} = 1$ at the end of the current period and a cost of $c_M (< c_R)$ is charged for maintenance. Note that we have assumed that repair or maintenance takes exactly one period. Consideration of a general repair time can be easily incorporated and does not change the main insights of the model.

If production continues in period t , the machine state changes according to the transition probability matrix \mathbf{P} . Due to possible yield losses, the firm has to carefully decide the input mix so that the numbers of final products passing quality assurance follow roughly the ratios $\gamma_1 : \gamma_2 : \dots : \gamma_n$. To achieve this, product j is chosen to be processed at random with probability

$$\beta_j = \frac{\gamma_j}{y_j} \frac{1}{\sum_{k=1}^n \frac{\gamma_k}{y_k}}.$$

Note that we seek policies that meet the required output ratio in expectation. We will use this requirement in the rest of this section. However, one may use actual counts on successful production (which enlarges the state space) and dynamically adjust the input mix to make the output mix as close as possible to the required ratio.

Let $v_{t:C}(i)$ and $v_{t:M}(i)$, respectively, denote the optimal profits for machine state i when production continues and when maintenance is scheduled. Then, the optimal profit function at the beginning of period t , denoted by $v_t(i)$, is simply the larger one of $v_{t:C}(i)$ and $v_{t:M}(i)$. That is,

$$v_t(i) = \max\{v_{t:M}(i), v_{t:C}(i)\}, \quad i = 1, 2, \dots, m - 1,$$

where

$$\begin{aligned} v_{t:M}(i) &= -c_M + \alpha v_{t+1}(1), \quad i = 1, 2, \dots, m - 1, \\ v_{t:C}(i) &= r + \sum_{j=1}^m p_{ij} v_{t+1}(j), \quad i = 1, 2, \dots, m - 1, \\ r &= \sum_{j=1}^n \beta_j y_j r_j, \end{aligned}$$

and α is the discount factor. Furthermore,

$$v_t(m) = -c_R + \alpha v_{t+1}(1).$$

The above equations constitute the formulation of the Markov decision model for machine maintenance.

Naturally, the machine in production is likely to stay in a healthier state if its initial state is healthier. This monotone relationship can be reflected by assuming that the transition probability matrix \mathbf{P} is stochastically monotone. Under this assumption, it is easy to show (see, e.g., Ross 1983) that there exists a threshold ℓ such that the machine should be subject to maintenance if the sensor reported machine state is ℓ or larger. This control is monitored by the machine tool supplier, who triggers a maintenance based on sensor data.

3.2.2. Sensor Data Triggered Machine Maintenance with Yield Knowledge. A little reflection of the above model reveals that the yields of different products can be different at different states of the machine. To predict the yield of product j produced under machine state i , denoted by y_{ij} , the machine tool supplier and the manufacturer need to collaborate on data sharing and learning. This knowledge certainly improves the machine maintenance decision.

To ensure the output ratios $\gamma_1 : \gamma_2 : \dots : \gamma_n$, the production input mix must satisfy

$$\beta_{ij} = \frac{\gamma_j}{y_{ij} \sum_{k=1}^n \gamma_k / y_{ik}}, \quad i = 1, 2, \dots, m.$$

In other words, product j is selected for processing with probability β_{ij} when the machine state is i . Also, the revenue generated by the machine becomes state-dependent, i.e.,

$$r(i) = \sum_{j=1}^n \beta_{ij} y_{ij} r_j.$$

Now the problem formulated in section 3.2.1 can be modified to

$$\begin{aligned} v_t(i) &= \max\{v_{t:M}(i), v_{t:C}(i)\}, \\ v_t(m) &= -c_R + \alpha v_{t+1}(1), \\ v_{t:M}(i) &= -c_M + \alpha v_{t+1}(1), \\ v_{t:C}(i) &= r(i) + \sum_{j=1}^m p_{ij} v_{t+1}(j). \end{aligned}$$

It is easy to show that a threshold maintenance policy continues to be optimal provided that the revenue $r(i)$ generated is increasing in the healthiness of the machine (i.e., decreasing in i). Under such collaborative learning on machine state dependent yields, the machine tool supplier can trigger maintenance calls.

3.2.3. Integrated Production and Maintenance Scheduling. When the manufacturer collaborates with the machine tool supplier on learning about machine state, the production scheduling can be coordinated with the machine state. This, of course, must be under the premises that the machine status data is made available by the supplier.

In addition to jointly determining whether or not production process should be interrupted for maintenance, the manufacturer also needs to decide the production input mix $\beta_{t;x_{ij}}$, $j = 1, 2, \dots, n$, in period t when the machine state is x_t . Certainly the input mix decision is only made if no maintenance activity is scheduled, which we denote by an indicator function $I_{t:C}(x_t)$. Specifically, $I_{t:C}(x_t) = 1$ if we choose to continue production at time t when the machine state is $x_t (\neq m)$, and $I_{t:C}(x_t) = 0$ if we choose to schedule a maintenance at time t . Naturally, $I_{t:C}(m) = 0$. Thus, the dynamic production and maintenance scheduling problem becomes

$$\begin{aligned} \max \left\{ \begin{array}{l} \mathbb{E} \left[\sum_{t=1}^T \alpha^t \left(\begin{array}{l} I_{t:C}(X_t) \sum_{j=1}^n \beta_{t;X_t} y_{X_t,j} r_j \\ -(1 - I_{t:C}(X_t)) c_M - I_{\{X_t=m\}} (c_R - c_M) \end{array} \right) \right] : \\ \beta_{t;ij}, \quad j = 1, 2, \dots, n, \quad i = 1, 2, \dots, m; \\ I_{t:C}(i), \quad i = 1, 2, \dots, m; \quad I_{t:C}(m) = 0 \end{array} \right\} \\ \text{s.t. } \mathbb{E} \left[\sum_{t=1}^T I_{t:C}(X_t) \beta_{t;X_t} y_{X_t,j} \right] \\ = \gamma_j \mathbb{E} \left[\sum_{t=1}^T I_{t:C}(X_t) \sum_{k=1}^n \beta_{t;X_t,k} y_{X_t,k} \right], \quad j = 1, 2, \dots, n. \end{aligned}$$

Here the state transitions are governed by:

$$\begin{aligned} P\{X_{t+1} = \ell | X_t = i, I_{t:C}(i) = 1\} &= p_{i\ell}, \\ i &= 1, 2, \dots, m-1; \ell = 1, 2, \dots, m, \\ P\{X_{t+1} = 1 | X_t = i, I_{t:C}(i) = 0\} &= 1, \quad i = 1, 2, \dots, m-1, \\ P\{X_{t+1} = 1 | X_t = m\} &= 1. \end{aligned}$$

With IIoT and IoT development platforms (like GE's Predix, Amazon's AWS, Microsoft's Azure and others), many machine tool suppliers have opted to maintain tools at the shop floors. Such servitization is becoming increasingly popular, making it possible to jointly optimize the production and maintenance schedules. The formulation above attempts to identify the best policy for the entire production system. In reality, collaborative learning and decision making are facilitated through bilateral contracts between the machine tool supplier and the manufacturer. Though contracts and coordination have been extensively researched by the production and operations management researchers, little work has considered appropriate contract terms involving information exchange. This is an area that needs new development to guide practice. For example, in many industries, the machine tool suppliers charge manufactures based on machine up time. It is worth investigating whether such a payment scheme would facilitate long-term coordination and information exchange.

3.2.4. Integrated Production and Maintenance Scheduling with Collaborative Learning. Further collaboration between the machine tool supplier and the manufacturer can enable the learning of the machine deterioration based on the products processed. Matching the data from machine sensors and production schedules, it is possible to understand how the machine state transition varies with the processing of different products. Let $\mathbf{P}^{(j)}$ be the transition probability matrix of the machine status when product j is produced. Now the problem formulation in section 3.2.3 has a different state transition (though the expressions of the objective and product mix constraint remain the same):

$$\begin{aligned} P\{X_{t+1} = \ell | X_t = i, \beta_{t:ij}, j = 1, 2, \dots, n, I_{t:C}(i) = 1\} \\ = \sum_{j=1}^n \beta_{t:ij} p_{i\ell}^{(j)}, \quad i = 1, 2, \dots, m-1; \ell = 1, 2, \dots, m, \\ P\{X_{t+1} = \ell | X_t = i, I_{t:C}(i) = 0\} &= 1, \quad i = 1, 2, \dots, m-1, \\ P\{X_{t+1} = 1 | X_t = m\} &= 1. \end{aligned}$$

As we have mentioned before, one major obstacle in collaborative learning and planning lies in the concern of intellectual property violation in the data exchange process. Researchers from engineering have

given tremendous emphasis on the design of secured data exchange technologies and protocols. From the production and operations management perspective, an important question is to identify the right data in the right format (e.g., the right level of aggregation or the right surrogate measures) so that collaboration can be achieved, while ill incentives are not created.

4. Remarks

We have briefly discussed some thoughts on how Big Data may change the POM research in demand planning and manufacturing. There are of course many other important and interesting topics (e.g., crowd sourcing, blockchain, sharing economy, lock boxes or drones, among many others) left out in our discussions. Even on topics we have covered, their impact on supply chain may go much beyond the scope of demand planning and manufacturing. For example, an issue that has gathered many debates is regarding whether to set up local 3D printing facilities or to build centralized foundries for consumer and industrial products. Such a decision is unlikely to be dictated solely by the traditional logistics network design parameters like the costs of production, transportation, and inventory. Instead, it will be heavily driven by customers' need for personalization with customized designs.

While Big Data enables extractions of useful features for better prediction, researchers should be mindful of the *veracity* associated with Big Data. On the one hand, unexpected trend in the data may contain early warnings of pattern changes. Timely detection of such changed patterns can generate great business value. On the other hand, biases, noise and abnormality in data post the biggest challenges in data and analytics based decision making. For example, Mukherjee and Sinha (2017) suggest that noisy signals contained unstructured user reports may induce under-reactions in medical device recall decisions.

We are at the very beginning of developing models for supply chain management that incorporate the real effects of Big Data. These effects are felt throughout the supply chain via the recognition of features and personalization. Going forward we have to pay attention to the way the supply chain will be managed as a consequence of the available information. Using Big Data is not merely to expand our models with additional features. Moreover, not every phenomenon or issue must be studied through models that are data driven or data integrated. For example, when a new policy or technology is to be introduced, one may need to understand its potential impact using data-supported theories and data-supported models. Also, models formulated with the awareness of potentially

available data can still provide understandings for general contexts without having to postulate the real data as inputs. In our view, what is important for our research in the era of Big Data is not merely the ability to collect data and carry out regressions. The way we think about operational processes and decision making needs to be transformed fundamentally.

Acknowledgments

Qi Feng's research is partly supported by National Natural Science Foundation of China (NSFC-71471107 and 71431004).

References

- Amato-McCoy, D. M. 2017. Study: Online shoppers want a personalized experience. Available at <https://www.chainstoreage.com/real-estate/german-discount-grocery-chain-heads-east/> (accessed on January 12, 2018).
- Balci, G., I. B. Cetin. 2010. Market segmentation in container shipping services: A qualitative study. *Manage. Res. Rev.* **40**(10): 1100–1116. <https://doi.org/10.1108/MRR-01-2017-0012>.
- Ban, G.-Y., C. Rudin. 2017. The big data newsvendor: Practical insights from machine learning. Working paper.
- Bauer, H. H., S. J. Barnes, T. Reichardt, M. M. Neumann. 2005. Driving consumer acceptance of mobile marketing: A theoretical framework and empirical study. *J. Electron. Commer. Res.* **6**(3): 181–191.
- Bloch, P. H. 1995. Seeking the ideal form: Product design and consumer response. *J. Market.* **59**(3): 16–29.
- Bounckena, R. B., S. Kraus. 2013. Innovation in knowledge-intensive industries: The double-edged sword of cooperation. *J. Bus. Res.* **66**(10): 2060–2070.
- Brettel, M., N. Friederichsen, M. Keller, M. Rosenberg. 2014. How virtualization, decentralization and network building change the manufacturing landscape: An industry 4.0 perspective. *Int. J. Inf. Commun. Eng.* **8**(1): 37–44.
- Buzacott, J. A., J. G. Shanthikumar. 1992. *Stochastic Models of Manufacturing Systems*. Prentice Hall, Englewood Cliffs, NJ.
- Çakanyildirim, M., R. O. Roundy. 2002. SeDFAM: Semiconductor demand forecast accuracy model. *IEE Trans.* **34**(5): 449–465.
- Carbonneau, R., K. Laframboise, R. Vahidov. 2008. Application of machine learning techniques for supply chain demand forecasting. *Eur. J. Oper. Res.* **184**: 1140–1154.
- Chandrashekar, G., F. Sahin. 2014. A survey on feature selection methods. *Comput. Electr. Eng.* **40**: 16–28.
- Chase, C. W. 2013. Using big data to enhance demand-driven forecasting and planning. *J. Bus. Forecast.* **32** (Summer): 27–32.
- Choi, H., H. Varian. 2012. Predicting the present with google trends. *Econ. Record* **88**(June): 2–9.
- Cinlar, E., M. Shaked, J. G. Shanthikumar. 1989. On lifetimes influenced by a common environment. *Stoch. Process. Appl.* **33**: 347–359.
- Clemen, R. T. 1989. Combining forecasts: A review and annotated bibliography. *Int. J. Forecast.* **5**(4): 559–583.
- Cui, R., S. Gallino, A. Moreno, D. J. Zhang. 2017. The operational value of social media information. *Prod. Oper. Manag.* <https://doi.org/10.1111/poms.12707>.
- Davidow, W., M. Malone. 1992. *The Virtual Corporation*. Harper Collins, New York.
- Dilda, V., L. Mori, O. Noterdaeme, C. Schmitz. 2017. Manufacturing: Analytics unleashes productivity and profitability. Available at <https://www.mckinsey.com/business-functions/operations/our-insights/manufacturing-analytics-unleashes-productivity-and-profitability> (accessed on January 12, 2018).
- Dong, L., D. Shi, F. Zhang. 2017. 3D printing vs. traditional flexible technology: Implications for operations strategy. Working paper.
- Feng, Q., J. G. Shanthikumar. 2017. Data integrated operations management. Working paper.
- Feng, Q., J. G. Shanthikumar. 2018a. Posted pricing versus bargaining in sequential selling process. *Oper. Res.* **66**(1): 92–103. <https://doi.org/10.1287/opre.2017.1651>.
- Feng, Q., J. G. Shanthikumar. 2018b. Supply and demand functions in inventory models. *Oper. Res.* **66**(1): 77–91. <https://doi.org/10.1287/opre.2017.1648>.
- Feng, Q., S. Luo, D. Zhang. 2013. Integrating dynamic pricing and replenishment decisions under supply capacity uncertainty. *Manuf. Serv. Oper. Manag.* **16**(1): 149–160.
- Ferreira, K. J., B. H. A. Lee, D. Simchi-Levi. 2016. Analytics for an online retailer: Demand forecasting and price optimization. *Manuf. Serv. Oper. Manag.* **18**(1): 69–88.
- Friedman, D. 1991. Evolutionary games in economics. *Econometrica* **59**(3): 637–666.
- Gardner, E. S. 1985. Exponential smoothing: The state of the art. *J. Forecast.* **4**(1): 1–28.
- Gardner, E. S. 1990. Evaluating forecast performance in an inventory control system. *Management Sci.* **36**(4): 490–499.
- Goel, S., J. M. Hofman, S. Lahaie, D. M. Pennock, D. J. Watts. 2010. Predicting consumer behavior with web search. *Proc. Natl Acad. Sci.* **107**(41): 17486–17490.
- Guyon, I., S. Gunn, M. Nikravesh, L. A. Zadeh. 2006. *Feature Extraction: Foundations and Applications*. Springer-Verlag, Berlin, Heidelberg.
- Hoch, S. J., B.-D. Kim, A. L. Montgomery, P. E. Rossi. 1995. Determinants of store-level price elasticity. *J. Mark. Res.* **32**(1): 17–29.
- Huang, T., J. A. Van Mieghem. 2014. Clickstream data and inventory management: Model and empirical analysis. *Prod. Oper. Manag.* **23**(3): 333–347.
- Jain, A. K., R. C. Dubes. 1988. *Algorithms for Clustering Data*. Prentice-Hall, Upper Saddle River, NJ.
- Kagermann, H., W. Wahlster, J. Helbig. 2013. Securing the future of German manufacturing industry: Recommendations for implementing the strategic initiative industrie 4.0. (April).
- Khaitan, S. K., J. D. McCalley. 2015. Design techniques and applications of cyberphysical systems: A survey. *IEEE Syst. J.* **9**(2): 350–365.
- Lee, J., B. Bagheri, H.-A. Kao. 2014. Service innovation and smart analytics for Industry 4.0 and big data environment. *Procedia CIRP* **16**: 3–8.
- Liu, Y., Z. J. Zhang. 2006. Research note—The benefits of personalized pricing in a channel. *Market. Sci.* **25**(1): 97–105.
- Lu, M., S. Shen. 2015. Not-for-profit surgery block allocation with cross-subsidization. Working paper.
- Makridakis, S., R. L. Winkler. 1983. Averages of forecasts: Some empirical results. *Management Sci.* **29**(9): 987–996.
- Miao, J., L. Niu. 2016. A survey on feature selection. *Procedia Comput. Sci.* **96**: 919–926.
- Monostori, L. 2014. Cyber-physical production systems: Roots, expectations and r&d challenges. *Procedia CIRP* **17**: 9–13.
- Mukherjee, U. K., K. K. Sinha. 2017. Product recall decisions in medical device supply chains: A big data analytic approach to evaluating judgment bias. *Prod. Oper. Manag.* <https://doi.org/10.1111/poms.12696>.
- Nurani, R. K., R. Akella, J. G. Shanthikumar. 1996. Optimal monitoring and control of deteriorating processes under two repair options. Working paper.

- Pang, B., L. Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retrieval* 2: 1–135.
- Parker, P. M. 1994. Aggregate diffusion forecasting models in marketing: A critical review. *Int. J. Forecast.* 10: 353–380.
- Petropoulosa, F., S. Makridakisb, V. Assimakopoulosc, K. Nikolopoulosd. 2014. ‘Horses for Courses’ in demand forecasting. *Eur. J. Oper. Res.* 237(1): 152–163.
- Pilkington, M. 2016. Blockchain technology: Principles and applications. F. X. Olleros, M. Zhegu, eds. *Research Handbook on Digital Transformations*. Edward Elgar, Northampton, 225–253.
- Ross, S. M. 1983. *Introduction to Stochastic Dynamic Programming*. Academic Press, New York, NY.
- Shankar, V., M. Kleijnen, S. Ramanathan, R. Rizley, S. Hollande, S. Morrissey. 2016. Mobile shopper marketing: Key issues, current insights, and future research avenues. *J. Interactive Market.* 34: 37–48.
- Simchi-Levi, D. 2014. OM forumOM research: From problem-driven to data-driven research. *Manuf. Serv. Oper. Manag.* 16(1): 2–10.
- Sloan, T. W., J. G. Shanthikumar. 2000. Combined production and maintenance scheduling for multiple-product single machine production system. *Prod. Oper. Manag.* 9: 379–399.
- Song, J.-S., Y. Zhang. 2016. Stock or print? Impact of 3D printing on spare parts logistics. Working paper.
- Srinivasan, R., M. Swink. 2017. An investigation of visibility and flexibility as complements to supply chain analytics: An organizational information processing theory perspective. *Prod. Oper. Manag.* (Forthcoming) <https://doi.org/10.1111/poms.12746>.
- Srinivasan, D., C. S. Chang, A. C. Liew. 1995. Demand forecasting using fuzzy neural computation, with special emphasis on weekend and public holiday forecasting. *IEEE Trans. Power Syst.* 10(4): 1897–1903.
- Talluri, K. T., G. J. Van Ryzin. 2004. *The Theory and Practice of Revenue Management*. Springer, Berlin.
- Tian, K. T., W. O. Bearden, G. L. Hunter. 2001. Smart manufacturing: Past research, present findings, and future directions. *J. Consum. Res.* 28(1): 50–66.
- Tomlin, B. 2017. Industry 4.0. Presentation at MOSM Supply Chain Management Special Interest Group Conference.
- Vandermerwe, S., J. Rada. 1989. Servitization of business: adding value by adding services. *Eur. Manage. J.* 6(4): 314–324.
- Warnd, M. P., M. C. Jones. 1995. *Kernel Smoothing*. Chapman & Hall/CRC, Boca Raton, FL.
- West, M., J. Harrison. 1997. *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York.
- Wolf, C. 2017. Car dealers compete in an online bidding war. Available at <https://www.autocheatsheet.com/new-car/car-dealers-compete-online.html> (accessed on January 12, 2018).
- Wu, D., D. W. Rosen, L. Wang, D. Schaefer. 2015. Cloud-based design and manufacturing: A new paradigm in digital manufacturing and design innovation. *Comput. Aided Des.* 59: 1–14.
- Yang, Y., B. Pan, H. Song. 2014. Predicting hotel demand using destination marketing organization’s web traffic data. *J. Travel Res.* 53(4): 433–447.
- Zhang, N., K. Kannan, J. G. Shanthikumar. 2017. A recommender system to nudge customers in a capacity constrained supply chain. Working paper.

How Sustainable Is Big Data?

Charles J. Corbett

UCLA Anderson School of Management, 110 Westwood Plaza, Box 951481, Los Angeles, California 90095-1481, USA,
 charles.corbett@anderson.ucla.edu

The rapid growth of “big data” provides tremendous opportunities for making better decisions, where “better” can be defined using any combination of economic, environmental, or social metrics. This essay provides a few examples of how the use of big data can precipitate more sustainable decision-making. However, as with any technology, the use of big data on a large scale will have some undesirable consequences. Some of these are foreseeable, while others are entirely unpredictable. This essay highlights some of the sustainability-related challenges posed by the use of big data. It does not intend to suggest that the advent of big data is an undesirable development. However, it is not too early to start asking what the unwanted repercussions of the big data revolution might be.

Key words: big data; sustainability; energy; operations; life-cycle assessment

History: Received: December 2017; Accepted: December 2017 by Kalyan Singhal, with no revision.

1. Introduction

Big data is here to stay, but what are some of the environmental and social consequences of the big data revolution? How sustainable is big data? The advent of big data provides revolutionary new opportunities for increased understanding of the environmental and social impacts of supply chains, with the concomitant potential for improvement along those dimensions. Big data also gives rise to both known and unknown environmental and social challenges. The purpose of this essay is to highlight some of those challenges. My intention is not to argue that big data is a phenomenon to be resisted. However, any technological breakthrough, if adopted on a sufficiently wide scale, will have far-reaching externalities, both positive and negative.

I use the term *big data* according to the emerging consensus (e.g., Etzion and Aragon-Correa 2016, p. 148), which holds that big data is not necessarily “big” but rather that it is differentiated from “traditional data” by any of the “4Vs”: *volume*, *variety*, *velocity*, and *veracity*. Goes (2014) argues that the promise of big data is the ability to exploit various combinations of these 4Vs.

Define *sustainability* loosely as making decisions while simultaneously taking into account economic, environmental, and social considerations. When sustainability is defined this way, it becomes clear that sustainability is inherently intertwined with big data. When we seek to measure the environmental and social impact of our decisions, an explosion in both the volume and the variety of data naturally results. Environmental impacts could be on global climate,

health of watersheds, human health, biodiversity, etc., while social impacts could affect workers, consumers, communities, societies, or value chain actors.

In the past, we may have received periodic updates on climatic conditions, or sporadic insights into the treatment of workers at vendor facilities. Now, however, the real-time monitoring of such phenomena at ever greater granularity results in a much greater *volume* and *velocity* of data. Now that such data can include anything from temperatures to satellite images to social media posts, the *variety* is widening too. The *veracity* of data also varies widely, depending on factors such as whether weather data are observed or extrapolated, or whether worker conditions are self-reported or independently verified.

Although the main purpose of this essay is to highlight sustainability-related challenges associated with big data, I do not want it to sound negative. So, I will first provide a few (unrepresentative and unscientifically selected) examples of the exciting benefits and opportunities that big data already provides or promises.

2. Examples of Big Data and Sustainable Operations

This section offers a few examples of how big data is, or can be, used to enhance our understanding of the impacts supply chains have on environmental and social conditions, and vice versa. These examples are not intended to be comprehensive or representative. Instead, they are provided to illustrate the wide range of (potential) applications of big data to sustainable operations, and to serve as a counterweight to the

subsequent section's emphasis on the emerging sustainability-related challenges associated with big data.

Many of the examples of firms successfully reducing their environmental footprint involve harnessing large amounts of timely data. The importance of performance measurement as a tool to improve environmental performance has long been known, as the examples in Corbett and Van Wassenhove (1993) illustrate. One good example is the energy management program at Walt Disney World (Allen 2005), which was initiated in the late 1990s and already involved collecting hourly information on the consumption of electricity, water, and other resources at a highly localized level. By gathering and sharing this information with the appropriate stakeholders (e.g., the maintenance crew or executive managers), Walt Disney World has been able to reduce its annual electricity usage by some 100 million kWh, while earning a 53% internal rate of return for their efforts. Monthly report cards provide historical and benchmarking information; showing the managers of Epcot how they performed relative to the Animal Kingdom helps to generate healthy competition among them. Conversely, real-time monitoring of a building's HVAC system allows a repair crew to be notified immediately if a control for a door to an air-conditioned space malfunctions in the evening, rather than only finding out when the next utility bill arrives a month or two later. The velocity with which the data is collected, processed, and shared, is critical to ensuring the data has the intended impact on energy consumption.

In a different setting, Marr (2017) points out how Caterpillar's Marine Division uses shipboard sensors to monitor a wide range of systems. The resulting data provide new insights into optimal operating practices; in one instance, a customer discovered that running more generators at lower power is more efficient than running fewer generators at maximum output. Big data help firms design better materials (National Academy of Sciences 2014), such as new photovoltaic materials with higher efficiency (p. 17) or GM's new thermoelectric materials for higher fuel efficiency (p. 24).

It is well understood that a changing climate will have a wide range of consequences for all kinds of organizations and supply chains. The exact effects of climate change on the conditions in any given location are still not well understood, but the combination of increasingly comprehensive historical data and fine-grained climate simulation models allows for more tailored predictions of how different regions will be affected. Some activities, such as wine growing and ski tourism, depend heavily on highly local microclimates, so forecasts must be available at a much more detailed spatial scale than was previously possible.

For instance, Jones et al. (2005) combine data on the ratings of wines from regions around the world with a widely used climate simulation model (HadCM3) to predict how the wines from each region will be affected, positively or negatively, by changes in local climatic conditions from 2000 to 2049. Ski tourism, which depends on the thickness and the persistence of snowpack on specific slopes, provides a similar example. Sun et al. (2016) used statistical downscaling to predict that various mountain locations in Southern California will be snow-free several weeks earlier by mid-century than is currently the case. Investors in winter sports facilities would do well to consider this kind of detailed forecast. In some instances, the activities of the supply chain itself cause changes in weather patterns: Thornton et al. (2017) combine data on shipping emissions and on 1.5 billion lightning strikes on a $10 \times 10 \text{ km}^2$ grid to find that the number of lightning strikes was elevated by 20%–100% along polluted shipping lanes. These are just a few examples; there are numerous similar studies.

This kind of data gathering is no longer constrained to earthbound monitoring stations, as various satellite-based systems provide more data with more detail and at higher frequencies. This trend will continue to accelerate because of the constant effort to miniaturize satellites. Woellert et al. (2011) outline a range of opportunities being opened by the use of CubeSats. These are small satellites that have a mass of around 1 kg and a volume of 10 cm^3 . CubeSats provide much more fine-grained monitoring of atmospheric conditions. These cheap satellites will also allow near real-time tracking of animal populations, and they can help disaster relief agencies allocate resources by providing images of earthquake damage. Satellite imagery is already being used to detect illegal logging after the fact, but Lynch et al. (2013) argue that daily observations are required if we are to take preventive action against illegal logging, instead of just observing it from a distance. Laurance and Balmford (2013) propose that satellite data could help prevent inappropriate road-building. This would help prevent ecological disasters long before they occur, because even a single road through a forest can wreak environmental damage far out of proportion to the physical footprint of the road itself. All these kinds of analyses will require a staggering amount of data. And, the attendant databases will rapidly become massive, because many studies of this kind require longitudinal data with very fine spatial and temporal resolution.

At the opposite end of the spectrum, firms are becoming more and more interested in the conditions experienced by workers in their supply chains. Traditionally, much work in this area depended on sending

auditors, third-party or otherwise, to assess the extent to which factories implemented the environmental and social practices expected of them. With the advent of smartphones, workers in factories around the world can now directly and anonymously report any conditions or practices they are exposed to. For instance, Walmart uses such worker-generated data collected through LaborVoices (de Felice 2015, p. 553). LaborLink is a similar effort. Firms must navigate between great opportunities and serious challenges to make the best decisions about how to use this data, since it comes from individual workers, in real time, covering a range of dimensions, and with unknown accuracy.

It would be easy to provide countless other examples of how big data can enhance sustainability, whether by allowing firms to make better decisions about the operation of their supply chains, or by allowing regulators to exert tighter control over those supply chains. The opportunities are boundless and exciting. Nevertheless, the big data revolution is also rapidly generating sustainability-related challenges of its own, which is the subject of the next section.

3. Sustainability Challenges Related to Big Data

The advent of big data presents unbounded opportunities to improve decision-making and to ensure more sustainable outcomes. However, like any other technological advance, it also brings challenges that will become increasingly acute as the use of big data becomes more prevalent. Some of these challenges may seem far-fetched today, but recall that the internal combustion engine was once considered an environmental breakthrough (Kirsch 2000). It is unlikely that advocates of fossil fuel-powered vehicles in the early 1900s could have had an inkling of the dramatic effects this technology would have on global air quality and climate over the course of the subsequent century. As mentioned earlier, the intention of this essay is not to argue that the rise of big data is an undesirable trend; the intent is to stress that we should be cognizant of some of its concomitant downsides. Below, I will review some of these risks associated with using or managing big data.

3.1. Social and Ethical Consequences of Using Big Data

When initially deciding where to roll out its Prime Free Same-Day Delivery service, Amazon aimed to serve as many people as it could, using its data to identify ZIP codes that contained a high concentration of Amazon Prime members. Maps of various cities produced by *Bloomberg BusinessWeek* (Ingold and Soper 2016), show the areas within the city limits that

initially received same-day delivery. In the case of Boston, it shows that the Roxbury neighborhood was excluded, while all surrounding neighborhoods did receive same-day delivery. The extent to which Roxbury was an anomaly is highlighted when one considers how far beyond the city limits the same-day delivery area stretched. The population of Roxbury is 59% Black. The *Bloomberg BusinessWeek* analysis found similar (though not quite as striking) effects in other major cities.

Without a doubt, Amazon did not set out to distinguish neighborhoods based on racial or ethnic composition. In the *Bloomberg BusinessWeek* article, an Amazon spokesperson stated, “Demographics play no role in it. Zero.” However, the result of their analysis was undesirable enough that Amazon rapidly backtracked. They added same-day service to initially excluded regions in Boston, New York, and Chicago. Even though race played no direct role in Amazon’s analysis, the algorithm they used led to “apparent discrimination,” as defined in Galhotra et al. (2017).

Fairness in algorithms is now the subject of significant research efforts, as also mentioned in Cohen (2018). For instance, Calders and Verwer (2010) define a *discrimination score* to measure the strength of group discrimination, and Galhotra et al. (2017) highlight some limitations of that score and generalize it to settings with more complex inputs. They also propose ways of testing software for fairness, something that Amazon presumably wishes it had done before rolling out the early phase of its same-day delivery program.

A different way in which virtually all of us have been negatively affected by big data is through the various hacks that have occurred over the years, exposing our personal data to unauthorized parties. Equifax revealed a particularly large data breach in September 2017. The company disclosed that sensitive personal information for some 146 million customers were stolen, including financial records and social security numbers. Many other large organizations around the world have been the subject of similar hacks.

The vast majority of individuals whose data was compromised are not directly financially affected, beyond perhaps the cost associated with additional identity protection services. Lai et al. (2012) observe that there appears to be relatively little research on the consequences for victims of the subsequent potential identity theft. They mention various studies that estimate the total costs to consumers on the order of \$50 billion in 2008–2009, or \$6000 per victim (p. 353). The horror stories reported by victims such as Amy Krebs are more salient (Shin 2014). Cohen (2018) mentions that the breach at Ashley Madison, a Canadian online dating site that specializes in

extramarital affairs, had severe consequences for families affected, and may have led to several unconfirmed suicides. The social and emotional costs incurred by individuals following data breaches can be substantial, in turn causing reputational damage and increased regulatory scrutiny of the firms that were hacked. Even firms that were not hacked can suffer consequences when a competitor is breached: Experian issued a warning about the risks it may face due to the “increased legislative and regulatory activity” that followed the Equifax breach (*Financial Times*, 2017).

Discussions of various ethical aspects of big data are emerging. Zwitter (2014) observes that global big data is shifting the power balance between the various stakeholders. They argue that major consequences can follow from many small actions taken by many individuals online (such as re-tweets or Facebook likes), and these consequences require a different perspective on what constitutes ethical behavior. Even offline actions, such as an individual parking his or her car in front of his or her own house, can be used to predict information such as demographics and voting behavior, as Gebru et al. (2017) describe in their application of deep learning to images from Google Street View. They also note that this raises important ethical concerns (p. 5). Richards and King (2013) highlight three paradoxes related to the ethics of big data. First, they note that although big data supposedly enhances transparency, much of the process by which the data are collected and analyzed is invisible to the public. Second, they pinpoint that in many cases, even though big data is about large numbers of individuals, its purpose is often precisely to identify specific individuals. And third, they also caution that big data will fundamentally change existing power structures.

A particular ethical quandary associated with big data is how it is used to make technology as addictive as possible. As Alter (2017a,b) notes, the people who design games, websites, and other interactive experiences run endless tests on millions of users to collect massive amounts of data on which features (fonts, sounds, swipes, incentives, etc.) maximize user engagement and keep users coming back time and again. In Alter’s (2017a) words: “As an experience evolves, it becomes an irresistible, weaponized version of the experience it once was. In 2004, Facebook was fun; in 2016, it’s addictive.” Tristan Harris, a former Google Design Ethicist, lists 10 ways in which product designers hijack our psychological vulnerabilities in order to keep our attention focused on their creations; among others, he lays out how our phone and the collection of apps that reside on it are like carrying slot machines in our pockets (Harris 2016). He argues that designers should use the data at their disposal to protect us from our addictive tendencies,

rather than exploit them; we should protect our time just as we protect our privacy.

To summarize, the use of big data is generating a wide range of ethical challenges. Some are more obvious than others, but they must be addressed if society is to reap the many positive benefits that big data has to offer.

3.2. Big Data May Not be the Right Data

How do we prioritize all these data? We have ever-growing amounts of increasingly fine-grained data on variables related to climate, and we should use that data by all means. But, when world leaders adopted 17 sustainable development goals (SDGs) at a United Nations summit in September 2015, climate change mitigation was Goal 13 of those 17.¹ It may be harder to measure progress on some of the other SDGs, such as “zero hunger” (Goal 2) or “peace, justice, and strong institutions” (Goal 16), but that difficulty does not mean those areas should be neglected.

Even within each goal, it is essential to first define the objective, and only then try to determine appropriate indicators, rather than the reverse. Hák et al. (2016) point out the danger of letting data availability drive priorities: “Operationalisation of the targets through indicators would be methodologically incorrect and might lead to distortions in the development of the policy agenda (such an approach might cause the false interpretation that only what can be measured is important to our lives)” (p. 568). This lack of data is not an idle threat; Sachs (2012) identified data shortcomings as “one of the biggest drawbacks” (p. 2210) of the Millennium Development Goals (the predecessor to the SDGs). With all the excitement about the vast amount of data becoming available for analysis, we must always ask what is not being captured.

Moreover, even when data appears to exist, it may not be correct. Veracity is always a concern. Firms and nations increasingly formulate quantitative targets related to sustainability: For example, firms set “science-based targets” and nations work together under the Paris agreement to reduce greenhouse gas emissions. The success of this kind of initiative inevitably hinges on the credibility of the associated emissions data. Ongoing debates, such as the discussion on China’s CO₂ emissions, indicate that there is as much as 10% uncertainty about the magnitude of those emissions over the period 2000–2013, which could be a decisive factor in whether China’s cumulative emissions will be consistent with a 2°C warming target (Korsbakken et al. 2016, Liu et al. 2015). At the firm level, Melville and Whisnant (2014) document various kinds of errors two firms made in their greenhouse gas emissions reporting. Blanco et al. (2016) point out that firms’ reports of supply chain emissions are even more difficult to interpret. Melville et al.

(2017) find that firms that paid attention to the accuracy of their carbon emissions data also achieved lower emissions. In the context of big data, these errors would suggest that improving the veracity of sustainability-related data may be beneficial in itself.

Finally, even when the data are correct in a narrow technical sense, it can easily be abused. With data arriving ever more rapidly, it is easy to fall into the trap of churning out analyses and rankings without due consideration of the underlying phenomena or the impact of those rankings. There are multitudes of rankings of countries, states, or firms on all kinds of environmental or social metrics. These rankings may sometimes have some informative value, but it is rare that a device as simple as a ranking can capture the many nuances involved in social or environmental concerns. Delmas et al. (2013) provide an illustration of how combining multiple rankings of firms' corporate social responsibility performance can lead to substantially better insights than any single ranking by itself. Many readers of this essay will find the concerns about business school rankings to be all too familiar. These ranking methods are dissected in Bachrach et al. (2017), who argue that the fundamental flaws in the methodologies used to rank business schools negatively impact those schools' ability to meet their social obligations. This danger of the misuse of big data is well examined in Lazer et al. (2014). They use the large error in predictions made by Google Flu Trends as an example of "big data hubris," which is the "often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis" (p. 1203). They close with the reminder that, despite the enormous opportunities provided by big data, much information is also contained in "small data" (p. 1205) that is not, or cannot be, contained in big data.

In short, despite the excitement associated with emerging big data related to environmental and social indicators, "small data" is still a critical component of environmental and social progress too.

3.3. Big Data May Not Mean Better Decisions

Part of the implicit premise underlying the excitement around big data is the assumption that more data will lead to better decisions. There is a vast literature on biases and heuristics (e.g., Kahneman 2011). Muthulingam et al. (2013) provide one example in a sustainability-related context: they document that managers who are faced with well-structured information about energy-efficiency initiatives will disproportionately choose items that appear closer to the top of the list, even when other initiatives further down are economically and environmentally superior. In the face of more data, these biases are likely to persist, or possibly become even more acute, due to the additional

cognitive burden associated with big data. The concern that more data might lead to worse decisions is not new, as illustrated by a well-known quote from Herbert Simon (1971, pp. 40–41): "In an information-rich world, the wealth of information means a dearth of something else: a scarcity of whatever it is that information consumes. What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it."

Even when there is no obvious bias at work, individuals may not respond to big data in the intended manner. One well-documented instance is the "rebound effect" in energy conservation, in which individuals who adopt energy-efficient technologies sometimes partly or fully negate these energy savings by consuming more energy elsewhere. Asensio and Delmas (2016, p. 207) find mixed evidence on this effect using real-time energy consumption data from 118 households over 9 months, yielding 374 million observations. The households in the treatment group were given feedback on their energy use, either in terms of cost or environmental health. The health treatment generally led to more durable energy conservation, but the households that received the cost treatment, after an initial reduction, ended up increasing their energy use related to heating and cooling. Even more surprisingly, both treatment groups experienced an increase in refrigerator energy use, which the authors attribute to unintuitive design of the temperature controls in the appliances concerned, making it hard for the participants to know which way to adjust the knob. This goes to show that no amount of data can compensate for the simple issue of poor product design. (Although, without the large dataset produced during this study, the product design issue would not have become clear.)

Discussions about big data often focus on the increasing *volume* of data, but the increasing *variety* of data poses growing challenges for decision-making. Decisions that have environmental consequences often involve data about a range of impacts, such as neurotoxicity, carcinogenicity, biotoxicity, global warming potential, land use effects, and various social indicators that may also be relevant. Multi-criteria decision methods can help to inform policies or decisions that balance such a range of environmental and social indicators in an appropriate manner. The use of these methods is growing, but still relatively nascent (Linkov and Moberg 2012, Ch. 2), and these methods are not without their drawbacks.

Consider the case of Alternatives Analysis, an approach intended to identify safer chemicals while avoiding inadvertent substitution of toxic chemicals

with even more undesirable substances. This approach is a response to chemical policies changing from a risk management focus to a prevention focus. Data on the human health and ecosystem health impacts of a chemical become more reliable as we gain more experience with that chemical. However, negative “experience” is precisely what policies such as the California Safer Consumer Products program and the European Union’s Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH) program seek to avoid. By the time we have big data on the ecological or human health impacts of a chemical, we are already well past the time for prevention and well into the time for risk management and mitigation. Malloy et al. (2016) describe some of the challenges involved with making early assessments about safer alternatives when the data on risks and impacts is severely incomplete and multidimensional. Linkov and Moberg (2012) provide an in-depth discussion of how multi-criteria decision analysis can be used in a variety of environmental decision contexts. A recent workshop we held at UCLA with some 15 participants involved in various aspects of alternatives analysis revealed that there is a great need for more systematic approaches to making these kinds of decisions, and also that application of existing multi-criteria decision methods to these questions is far from straightforward. In each of these settings, the main challenge is in how to deal with the *small* data, rather than the big data.

To summarize, big data can clearly help inform environmental and social decisions in some situations, but in other contexts, we must focus on making better decisions with little or no data rather than waiting for more information to arrive.

3.4. Big Data Can Change the Manufacturing Landscape

Big data is not only a technological revolution in itself. It will also facilitate other potentially large shifts, including in the physical world. One such potential consequence is that the spread of big data may foster wider application of mass customization, which in turn would likely be linked to a broader use of 3D printing or additive manufacturing. Two other articles in this special issue, by Feng and Shanthikumar (2018) and Guha and Kumar (2018), also point to this potential effect.

This raises the question whether additive manufacturing is more or less sustainable than conventional production. Several recent articles, including a special issue of the *Journal of Industrial Ecology*, investigate this question. The overall takeaway is that the answer is not obvious. Summarizing existing studies, Kellens et al. (2017) observe that the specific energy involved in additive manufacturing is “1 to 2 orders of

magnitude” (p. S63) higher than for conventional manufacturing. However, they also note that additive manufacturing can lead to environmental benefits if a larger portion of a damaged part can be reused in a repair process, or if parts are redesigned appropriately. This need for appropriate redesign was also found by Mami et al. (2017) in the manufacture of aircraft doorstops. In addition, Walachowicz et al. (2017) find that additive manufacturing has lower impact along various dimensions than conventional processes for the repair of gas turbine burners. In the case of injection molding, Huang et al. (2017) estimate that additive manufacturing uses slightly less energy than conventional manufacturing. In the manufacture of eyeglasses, Cerdas et al. (2017) find that the comparison between additive and conventional manufacturing depends heavily on the material used. They caution that one of the potential benefits of 3D printing is to allow more distributed manufacturing, but that such a more dispersed production system is harder to regulate. Taking a life-cycle perspective, Huang et al. (2017) predict major savings due to the weight reduction in aircraft components that 3D printing allows.

The point here is not to argue that 3D printing is the only innovation that is likely to be accelerated by the big data revolution, nor that 3D printing is more or less sustainable than conventional manufacturing. The point is to highlight that the consequences of big data on sustainability are likely to reach well beyond big data itself and even into the physical world. The example of CubeSats mentioned earlier is another instance of this linkage. The big data that CubeSats will enable will lead to high demand for such satellites, which in turn will have real environmental consequences: launch is clearly energy-intensive, but many satellites will eventually become space junk, a very different environmental challenge.

3.5. Managing and Storing Big Data

There is no question that in many instances, better data can help reduce energy or material consumption, but managing that data still requires physical processes, which consume energy. A widely cited analysis by Gartner (2007) claims that information and communication technology (ICT) accounted for about 2% of global CO₂ emissions in 2007, and this amount was comparable to the emissions associated with aviation. Although Malmodin et al. (2010) point out that the comparison is distorted, they confirm that the 2% estimate is about right. (They estimate the ICT portion of CO₂e emissions as 1.3%.) In other words, the energy consumption of ICT is not huge, but it is already significant, and it is growing.

Predicting how the impact of ICT will evolve over time involves balancing two counteracting factors. On

the one hand, storing and transmitting data is becoming more efficient over time. Aslan et al. (2018) estimate that electricity intensity of data transmission has decreased by about 50% per year since 2000. Operators of some of the largest data centers increasingly rely on renewable energy. Apple's data centers operate on 100% renewable energy (Apple 2017, p. 41), and Google will reach 100% renewable energy for its operations in 2017 (Google 2016, p. 9); Facebook is similarly committed to powering its operations with 100% renewable energy (Facebook 2017). On the other hand, our data footprint is growing dramatically, and with it the energy required by data centers, data networks, and connected devices. The International Energy Agency projects that data center electricity use will increase by 3% by 2020, despite a tripling in workload; its forecast for data networks ranges anywhere from a 70% increase to a 15% decrease by 2021 (IEA 2017, p. 103). An encouraging sign is that, as Khuntia et al. (2018) find, firms that invest in green IT not only achieve lower IT equipment energy consumption, but also earn higher profits.

Whether the energy used for storing and transmitting data is renewable or not, that energy has to be generated somehow, and even renewable energy comes with (significant) costs, in the form of land use, material use, noise and visual pollution, and more. A rapid increase in the energy demands associated with big data is therefore a concern which we need to confront. How can we translate figures such as "2% of global CO₂ emissions" to numbers on a scale that apply to individuals or companies? For instance, what are the greenhouse gas emissions associated with storing 1 TB of data? Estimates vary widely. An analysis by the Natural Resources Defense Council (2012) compares various technologies and data hosting scenarios, and the result was a range of 0.6 kg CO₂e per year (i.e., using best practices, public cloud storage) to 15.9 kg CO₂e per year (worst case scenario, on-premise with virtualization) (see figure 2 in Natural Resources Defense Council 2012). The high end of this range translates to 15.9 metric tons of CO₂e per year per TB of data, which would be comparable to the annual emissions of several passenger vehicles. This seems unreasonably high. A report by Google (2011, p. 6) estimates that keeping email on a locally hosted server at a small business causes 103 kg of CO₂ per year per user. Because this still seems high, I asked an expert in this area, Professor Eric Masanet at the McCormick School of Engineering at Northwestern University. He provided some very helpful estimates that seem more believable (Masanet 2017).

Taking data from Shehabi et al. (2016), a Lawrence Berkeley National Laboratory report he coauthored, he estimates US data centers stored about 300 million TB of data in 2016 (figure 12, p. 14), which

consumed around 8.3 billion kWh (figure 16, p. 17), hence 27.7 kWh per TB of data. This is the energy directly used for storage by the IT equipment, which we must multiply by the power usage effectiveness (PUE, or *total data center facility energy use* ÷ *IT equipment energy use*) to obtain the total consumption of the data centers, which includes the energy consumption of associated technology such as cooling and lighting. According to figure 21 in Shehabi et al. (2016, p. 25), the total data center energy usage is about 72 billion kWh in 2016, of which 43 billion kWh was for IT equipment, yielding a PUE of $72 \div 43 = 1.67$. The total energy consumption of data centers is then $1.67 \times 27.7 \text{ kWh} = 46.33 \text{ kWh}$ per TB of data per year.

Converting this to CO₂e emissions is not straightforward, since energy consumption varies widely across data centers. Using an EPA (2017) estimate for the emission factor (the US national weighted average CO₂ marginal emission rate) of 0.744 kg of CO₂ per kWh, the carbon emissions associated with data storage would be 0.744×46.33 , or approximately 35 kg of CO₂ per TB per year. Of course, this is a very rough estimate at best, and with changes in storage technology and energy mix, this footprint per TB will likely improve. At the same time, with the continuation of big data and the advent of augmented reality and virtual reality, our data footprint will likely increase.

The numbers so far have only focused on storage, but *transmitting* data also consumes energy. Coroama et al. (2013) estimate that a videoconference transmission between Switzerland and Japan in 2009 accounted for 200 kWh per TB, substantially above the 46.33 kWh per TB estimated above for storage. Weber et al. (2010) find that online delivery of music to consumers generally causes a lower carbon footprint than physical CDs, but Mayers et al. (2015) predict that downloading large games over the Internet will cause higher carbon emissions than physical delivery via Blu-Ray discs. Altogether, these (and other) studies illustrate that, while there is considerable uncertainty about the energy impacts of big data storage and transmission, they are large enough to take seriously.

What makes all these figures more disconcerting is that the vast majority of data stored can safely be considered as waste. This is sometimes referred to as "dark data," or "the information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes." Again, reliable estimates of how much dark data exists are hard to come by, but the numbers are quite astounding. IBM estimates that 90% of all data stored is never used (Johnson 2015). Paul Rogers, chief development officer at General Electric, stated in

Wharton (2014) that “[o]nly about one-half of 1% of the world’s data is being analyzed” (p. 2), which means that the other 99.5% is dark data. Cohen (2018) cites another source that mentions the same figure of 0.5%. A Veritas survey suggests that 54% of data is dark, while 32% is ROT (*redundant, obsolete, trivial*), and only 14% is business-critical. This survey also estimates that there will be almost \$900 billion in avoidable costs of data storage by 2020 (Hernandez 2015). The prevalence of dark data and the costs of data storage suggest that a sizable proportion of the energy use associated with big data is avoidable, and we must begin to consider the waste hidden in big data in the same way we think about physical waste. It is not just energy that is wasted: Shehabi et al. (2016, p. 28) estimate that data centers in the US were responsible for the consumption of 626 billion liters of water in 2014.

We need to start using tools such as value stream mapping for data flows, just as we already do to identify waste in physical flows. This would help uncover the vast amounts of unnecessary and obsolete copies of data currently being stored because we have not yet started treating it as actual waste with a real cost. The ISO/IEC 38505 series of standards for governance of data provide a helpful organizing framework for how to collect, store, report, decide, distribute, and dispose of data. The concepts of the circular economy that are already being applied to supply chains (e.g., Agrawal et al. 2018) might also help curtail the unbounded growth of data being transmitted or sitting in storage. As data become too big, it will become excessively costly to distribute it, so it will need to be analyzed locally at data repositories. Space missions may provide inspiration on how to do this: the explosion of data collected during a mission far exceeds the severe constraints on what can be transmitted back to earth, which means that local data reduction will be necessary (National Academy of Sciences, 2014; p. 12).

This is a good time to recall the analogy of the internal combustion engine. When that technology was in its infancy, people were certainly concerned about the visible air pollution it created, but the thought that something as small as a car could change our climate must have been inconceivable. Now, several billion cars later, we know better. The point I wish to make here is that, even though the environmental benefits of big data are often large and obvious, it is not too early to start measuring and minimizing the environmental costs, as some firms already do by investing in renewable energy to power their data centers. This applies not only to big data but also to other complementary technologies that have surfaced earlier in this essay, such as additive manufacturing and CubeSats.

4. Conclusions

We have only begun to explore the potential of big data to improve decision-making in many areas of life. These applications permeate the field of sustainability, broadly defined. I have described a few examples, and other articles in this special issue provide more. Devalkar et al. (2018) outline how big data can help agriculture in India. Swaminathan (2018) describes opportunities for using big data to assist in humanitarian operations. These opportunities are exciting and profoundly promising, and we should pursue them accordingly.

However, while doing so, we should not lose track of the fact that rushing to collect and exploit ever “bigger” data will inevitably have undesirable side effects. Some of these side effects have already surfaced, but others may arise in unexpected areas. We should not dampen the excitement deservedly attached to big data, but we should also be vigilant about potential side effects. I have catalogued some of these potential unwanted byproducts. Some of them may turn out to be irrelevant in the long term, but some others not mentioned here will surely emerge.

One often hears the slogan *Big data is the new oil*, but like all analogies, it only goes so far. It is true that fossil fuel was a critical driver of growth and change in the global economy during the 20th century (*The Economist*, 2017), but as Thorp (2012) points out, information is inherently renewable, unlike fossil fuel. Thorp does draw a different parallel, arguing that oil has also been the cause of untold environmental and social devastation. He observes that “data spills” have already occurred, and he asks when we might encounter “dangerous data drilling practices” or suffer the long-term effects of “data pollution.” In her discussion of ethical issues in big data, Martin (2015) also refers to the surveillance that results from the systematic way in which individual data are collected as pollution. She draws a number of parallels between traditional supply chains and information supply chains, with implications for how data should be managed throughout the supply chain to minimize the negative aspects of the growth of big data. For instance, just as manufacturers are concerned about ethical sourcing, firms in the big data industry should ensure that the data they rely on are obtained ethically.

One could argue that, in its earliest days, the fossil fuel revolution was mostly beneficial and relatively harmless. Its disastrous side effects were the result of the sheer breadth and depth of the penetration of fossil fuel-based products into every aspect of human life. Moreover, the collateral inertia associated with the vast investments made over the years have created the pronounced path dependency that has

caused so much difficulty as we try to migrate away from fossil fuels.

We are now making important decisions about big data—decisions about issues such as technology platforms, governance mechanisms, ownership structures, and access rights. All these decisions could have pivotal implications for what options will be available to us later, when the costs of big data start coming into focus. In order to ensure we use big data in a sustainable way, we must always be on the alert for potential repercussions, even repercussions that seem far-fetched to us now. The big data revolution has opened a vast uncharted frontier, and we must not only explore this frontier with enthusiasm, but also with caution.

Acknowledgments

I am grateful to the Editor, Kal Singhal, for providing this opportunity; to Eric Masanet for his quick and detailed response to my questions about energy use of data centers and for other suggestions, and to Christian Blanco, Suresh Muthulingam and Vincente LeCornu for helpful comments on an earlier version of this essay. As usual, all errors are my own.

Notes

¹See <http://www.un.org/sustainabledevelopment/development-agenda/> (accessed date November 15, 2017).

²<https://www.gartner.com/it-glossary/dark-data> (accessed date October 27, 2017).

References

- Agrawal, V. V., A. Atasu, L. N. Van Wassenhove. 2018. New opportunities for operations management research in sustainability. *Manuf. Serv. Oper. Manag.* (e-pub ahead of print). <https://doi.org/10.1287/msom.2017.0699>.
- Allen, P. 2005. How Disney saves energy and operating costs. HPAC Engineering. Available at <http://www.hpac.com/building-controls/how-disney-saves-energy-and-operating-costs> (accessed date December 2, 2017).
- Alter, A. 2017a. Tech bigwigs know how addictive their products are. Why don't the rest of us? *Wired*, March 24, 2017. Available at <https://www.wired.com/2017/03/irresistible-the-rise-of-addictive-technology-and-the-business-of-keeping-us-hooked/> (accessed date December 13, 2017).
- Alter, A. 2017b. *Irresistible: The Rise of Addictive Technology and the Business of Keeping Us Hooked*. Penguin Press, New York, NY.
- Apple. 2017. Environmental responsibility report. 2017 Progress Report, Covering Fiscal Year 2016. Available at https://images.apple.com/environment/pdf/Apple_Environmental_Responsibility_Report_2017.pdf (accessed date December 13, 2017).
- Asensio, O. I., M. A. Delmas. 2016. The dynamics of behavior change: Evidence from energy conservation. *J. Econ. Behav. Organ.* **126**: 196–212.
- Aslan, J., K. Mayers, J. G. Koomey, C. France. 2018. Electricity intensity of internet data transmission: Untangling the estimates. *J. Ind. Ecol.* **22**(4): 785–798. <https://doi.org/10.1111/jiec.12630>.
- Bachrach, D. G., E. Bendoly, D. Beu Ammeter, R. Blackburn, K. G. Brown, G. Burke, T. Callahan, K. Y. Chen, V. H. Day, A. E. Ellstrand, O. H. Erikson. 2017. On academic rankings, unacceptable methods, and the social obligations of business schools. *Decis. Sci.* **48**(3): 561–585.
- Blanco, C., F. Caro, C. J. Corbett. 2016. The state of supply chain carbon footprinting: analysis of CDP disclosures by US firms. *J. Clean. Prod.* **135**: 1189–1197.
- Calders, T., S. Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Min. Knowl. Disc.* **21**(2): 277–292.
- Cerdas, F., M. Juraschek, S. Thiede, C. Herrmann. 2017. Life cycle assessment of 3D printed products in a distributed manufacturing system. *J. Ind. Ecol.* **21**(S1): S80–S93.
- Cohen, M. C. 2018. Big data and service operations. *Prod. Oper. Manag.* **27**(9): 1709–1723. <https://doi.org/10.1111/poms.12832>.
- Corbett, C. J., L. N. Van Wassenhove. 1993. The green fee: Internalizing and operationalizing environmental issues. *Calif. Manage. Rev.* **36**(1): 116–135.
- Coroama, V. C., L. M. Hilty, E. Heiri, F. M. Horn. 2013. The direct energy demand of internet data flows. *J. Ind. Ecol.* **17**(5): 680–688.
- Delmas, M. A., D. Etzion, N. Nairn-Birch. 2013. Triangulating environmental performance: What do corporate social responsibility ratings really capture? *Acad. Manage. Perspect.* **27**(3): 255–267.
- Devalkar, S. K., S. Seshadri, C. Ghosh, A. Mathias. 2018. Data science applications in Indian agriculture. *Prod. Oper. Manag.* **27**(9): 1701–1708. <https://doi.org/10.1111/poms.12834>.
- EPA. 2017. Greenhouse gases equivalencies calculator – Calculations and references. Available at <https://www.epa.gov/energy/greenhouse-gases-equivalencies-calculator-calculations-and-references> (accessed date November 16, 2017).
- Etzion, D., J. A. Aragon-Correa. 2016. Big data, management, and sustainability: Strategic opportunities ahead. *Organ. Environ.* **29**(2): 147–155.
- Facebook. 2017. Available at <https://sustainability.fb.com/clean-and-renewable-energy/> (accessed date December 13, 2017).
- de Felice, D. 2015. Business and human rights indicators to measure the corporate responsibility to respect challenges and opportunities. *Hum. Rights Q.* **37**: 511–555.
- Feng, Q., J. G. Shanthikumar. 2018. How research in production and operations management may evolve in the era of big data. *Prod. Oper. Manag.* **27**(9): 1670–1684. <https://doi.org/10.1111/poms.12836>.
- Financial Times. 2017. Experian warns of increased scrutiny after Equifax hack. November 15, 2017. Available at <https://www.ft.com/content/ec01484c-c9e8-11e7-ab18-7a9fb7d6163e> (accessed date November 15, 2017).
- Galhotra, S., Y. Brun, A. Meliou. 2017. Fairness testing: Testing software for discrimination. Proceedings of 2017 11th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering. Paderborn, Germany, September 4–8, 2017 (ESEC/FSE'17), pp. 13. Available at <https://doi.org/10.1145/3106237.3106277>.
- Gartner. 2007. Gartner estimates ICT industry accounts for 2 percent of global CO₂ emissions. April 26, 2007. Available at <https://www.gartner.com/newsroom/id/503867> (accessed date November 16, 2017).
- Gebru, T., J. Krause, Y. Wang, D. Chen, J. Deng, E. L. Aiden, L. Fei-Fei. 2017. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proc. Natl Acad. Sci.* **114**: 12108–13113.

- Goes, P. B. 2014. Big data and is research. *MIS Q.* **38**(3): iii–viii.
- Google. 2011. Google's green computing: Efficiency at scale. Available at <https://static.googleusercontent.com/media/www.google.com/en//green/pdfs/google-green-computing.pdf> (accessed date October 30, 2017).
- Google. 2016. Environmental report. Available at <https://static.googleusercontent.com/media/www.google.com/en//green/pdf/google-2016-environmental-report.pdf> (accessed date December 13, 2017).
- Guha, S., S. Kumar. 2018. Emergence of big data research in operations management, information systems, and healthcare: Past contributions and future roadmap. *Prod. Oper. Manag.* **27**(9): 1724–1735. <https://doi.org/10.1111/poms.12833>.
- Hák, T., S. Janoušková, B. Moldan. 2016. Sustainable development goals: A need for relevant indicators. *Ecol. Ind.* **60**: 565–573.
- Harris, T. 2016. How technology is hijacking your mind – From a Magician and Google Design Ethicist. *Medium*, May 18, 2016. Available at <https://journal.thriveglobal.com/how-technology-hijacks-peoples-minds-from-a-magician-and-google-s-design-ethicist-56d62ef5edf3> (accessed date December 13, 2017).
- Hernandez, P. 2015. Enterprises are hoarding 'dark' data: Veritas. Datamation. Available at <https://www.datamation.com/storage/enterprises-are-hoarding-dark-data-veritas.html> (accessed date October 30, 2017).
- Huang, R., M. E. Riddle, D. Graziano, S. Das, S. Nimbalkar, J. Cresko, E. Masanet. 2017. Environmental and economic implications of distributed additive manufacturing: The case of injection mold tooling. *J. Ind. Ecol.* **21**(S1): S130–S143.
- IEA. 2017. Digitalization & energy. Available at <http://www.iea.org/publications/freepublications/publication/DigitalizationandEnergy3.pdf> (accessed date December 13, 2017).
- Ingold, D., S. Soper. 2016. Amazon doesn't consider the race of its customers. Should it? *Bloomberg*, April 21, 2016. Available at <http://www.bloomberg.com/graphics/2016-amazon-same-day> (accessed date October 27, 2017).
- Johnson, H. 2015. Digging up dark data: What puts IBM at the forefront of insight economy. *Silicon Angle*. Available at <https://siliconangle.com/blog/2015/10/30/ibm-is-at-the-forefront-of-insight-economy-ibminsight/> (accessed date October 27, 2017).
- Jones, G. V., M. A. White, O. R. Cooper, K. Storchmann. 2005. Climate change and global wine quality. *Clim. Change.* **73**(3): 319–343.
- Kahneman, D. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.
- Kellens, K., M. Baumers, T. G. Gutowski, W. Flanagan, R. Lifset, J. R. Dufflou. 2017. Environmental dimensions of additive manufacturing: Mapping application domains and their environmental implications. *J. Ind. Ecol.* **21**(S1): S49–S68.
- Khuntia, J., T. J. V. Saldanha, S. Mithas, V. Sambamurthy. 2018. Information technology and sustainability: Evidence from an emerging economy. *Prod. Oper. Manag.* **27**(4): 756–773. <https://doi.org/10.1111/poms.12822>.
- Kirsch, D. A. 2000. *The Electric Vehicle and the Burden of History*. Rutgers University Press, New Brunswick, NJ; London.
- Korsbakken, J. I., G. P. Peters, R. M. Andrew. 2016. Uncertainties around reductions in China's coal use and CO₂ emissions. *Nat. Clim. Chang.* **6**(7): 687–690.
- Lai, F., D. Li, C. T. Hsieh. 2012. Fighting identity theft: The coping perspective. *Decis. Support Syst.* **52**(2): 353–363.
- Laurance, W. F., A. Balmford. 2013. Land use: A global map for road building. *Nature* **495**(7441): 308–309.
- Lazer, D., R. Kennedy, G. King, A. Vespignani. 2014. The parable of Google Flu: Traps in big data analysis. *Science* **343**(6176): 1203–1205.
- Linkov, I., E. Moberg. 2012. *Multi-Criteria Decision Analysis: Environmental Applications and Case Studies*. CRC Press, Boca Raton, FL.
- Liu, Z., D. Guan, W. Wei, S. J. Davis, P. Ciaia, J. Bai, S. Peng, Q. Zhang, K. Hubacek, G. Marland, R. J. Andres. 2015. Reduced carbon emission estimates from fossil fuel combustion and cement production in China. *Nature* **524**(7565): 335–338.
- Lynch, J., M. Maslin, H. Balzter, M. Sweeting. 2013. Choose satellites to monitor deforestation. *Nature* **496**: 293–294.
- Malloy, T. F., V. M. Zaunbrecher, C. Batteate, A. Blake, W. F. Carroll, C. J. Corbett, S. F. Hansen, R. Lempert, I. Linkov, R. McFadden, K. D. Moran. 2016. Advancing alternative analysis: Integration of decision science. *Environ. Health Perspect.* **125**(6): 066001-1–066001-12.
- Malmodin, J., Å. Moberg, D. Lundén, G. Finnveden, N. Lövehagen. 2010. Greenhouse gas emissions and operational electricity use in the ICT and entertainment & media sectors. *J. Ind. Ecol.* **14**(5): 770–790.
- Mami, F., J. P. Revéret, S. Fallaha, M. Margni. 2017. Evaluating eco-efficiency of 3D printing in the aeronautic industry. *J. Ind. Ecol.* **21**(S1): S37–S48.
- Marr, B. 2017. IoT and big data at caterpillar: How predictive maintenance saves millions of dollars. *Forbes*, February 7, 2017. Available at <https://www.forbes.com/sites/bernardmarr/2017/02/07/iot-and-big-data-at-caterpillar-how-predictive-maintenance-saves-millions-of-dollars/#203abf737240> (accessed date December 11, 2017).
- Martin, K. E. 2015. Ethical issues in the big data industry. *MIS Q. Exec.* **14**(2): 67–85.
- Masanet, E. 2017. Private communication.
- Mayers, K., J. Koomey, R. Hall, M. Bauer, C. France, A. Webb. 2015. The carbon footprint of games distribution. *J. Ind. Ecol.* **19**(3): 402–415.
- Melville, N. P., R. Whisnant. 2014. Energy and carbon management systems. *J. Ind. Ecol.* **18**(6): 920–930.
- Melville, N. P., T. J. Saldanha, D. E. Rush. 2017. Systems enabling low-carbon operations: The salience of accuracy. *J. Clean. Prod.* **166**: 1074–1083.
- Muthulingam, S., C. J. Corbett, S. Benartzi, B. Oppenheim. 2013. Energy efficiency in small and medium-sized manufacturing firms: Order effects and the adoption of process improvement recommendations. *Manuf. Serv. Oper. Manag.* **15**(4): 596–615.
- National Academy of Sciences. 2014. *Big Data in Materials Research and Development: Summary of a Workshop*. The National Academies Press, Washington, DC.
- Natural Resources Defense Council. 2012. The carbon emissions of server computing for small- to medium-sized organizations: A performance study of on-premise vs. the cloud. Available at https://www.nrdc.org/sites/default/files/NRDC_WSP_Cloud_Computing_White_Paper.pdf (accessed date October 27, 2017).
- Richards, N. M., J. H. King. 2013. Three paradoxes of big data. *Stanf. Law Rev.* (Online) **66**(41): 41–46.
- Sachs, J. D. 2012. From millennium development goals to sustainable development goals. *Lancet* **379**(9832): 2206–2211.
- Shehabi, A., S. Smith, D. Sartor, R. Brown, M. Herrlin, J. Koomey, E. Masanet, N. Horner, I. Azevedo, W. Lintner. 2016. *United States Data Center Energy Usage Report*. Lawrence Berkeley National Laboratory, Berkeley, CA. LBNL-1005775.
- Shin, L. 2014. 'Someone had taken over my life': An identity theft victim's story. *Forbes*. November 18, 2014. Available at <https://www.forbes.com/sites/laurashin/2014/11/>

- 18/someone-had-taken-over-my-life-an-identity-theft-victims-story/#3e6b4a3f25be (accessed date November 15, 2017).
- Simon, H. A. 1971. Designing organizations for an information-rich world. M. Greenberger, ed. *Computers, Communication, and the Public Interest*. The Johns Hopkins Press, Baltimore, MD, 37–72.
- Sun, F., A. Hall, M. Schwartz, D. B. Walton, N. Berg. 2016. Twenty-first-century snowfall and snowpack changes over the southern California Mountains. *J. Clim.* **29**(1): 91–110.
- Swaminathan, J. M. 2018. Big data analytics for rapid, impactful, sustained, and efficient (RISE) humanitarian operations. *Prod. Oper. Manag.* **27**(9): 1696–1700. <https://doi.org/10.1111/poms.12840>.
- The Economist. 2017. Fuel of the future: Data is giving rise to a new economy. May 6, 2017. Available at <https://www.economist.com/news/briefing/21721634-how-it-shaping-up-data-giving-rise-new-economy> (accessed date December 3, 2017).
- Thornton, J. A., K. S. Virts, R. H. Holzworth, T. P. Mitchell. 2017. Lightning enhancement over major oceanic shipping lanes. *Geophys. Res. Lett.* **44**(17): 9102–9111.
- Thorp, J. 2012. Big data is not the new oil. *Harvard Business Review*, November 30, 2012. <https://hbr.org/2012/11/data-humans-and-the-new-oil> (accessed date December 3, 2017).
- Walachowicz, F., I. Bernsdorf, U. Papenfuss, C. Zeller, A. Graichen, V. Navrotsky, N. Rajvanshi, C. Kiener. 2017. Comparative energy, resource and recycling lifecycle analysis of the industrial repair process of gas turbine burners using conventional machining and additive manufacturing. *J. Ind. Ecol.* **21**(S1): S203–S215.
- Weber, C. L., J. G. Koomey, H. S. Matthews. 2010. The energy and climate change implications of different music delivery methods. *J. Ind. Ecol.* **14**(5): 754–769.
- Wharton. 2014. What big data can mean for sustainability. September 12, 2014. <http://knowledge.wharton.upenn.edu/article/what-big-data-means-for-sustainability/> (accessed date December 3, 2017).
- Woellert, K., P. Ehrenfreund, A. J. Ricco, H. Hertzfeld. 2011. Cube-sats: Cost-effective science and technology platforms for emerging and developing nations. *Adv. Space Res.* **47**(4): 663–684.
- Zwitter, A. 2014. Big data ethics. *Big Data Soc.* **1**(2): 1–6.

Big Data Analytics for Rapid, Impactful, Sustained, and Efficient (RISE) Humanitarian Operations

Jayashankar M. Swaminathan*

Kenan-Flagler Business School, University of North Carolina, Chapel Hill, North Carolina 27599, USA, msj@unc.edu

There has been a significant increase in the scale and scope of humanitarian efforts over the last decade. Humanitarian operations need to be—rapid, impactful, sustained, and efficient (RISE). Big data offers many opportunities to enable RISE humanitarian operations. In this study, we introduce the role of big data in humanitarian settings and discuss data streams which could be utilized to develop descriptive, prescriptive, and predictive models to significantly impact the lives of people in need.

Key words: big data; humanitarian operations; analytics

History: Received: November 2017; Accepted: November 2017 by Kalyan Singhal, after 1 revision.

1. Introduction

Humanitarian efforts are increasing on a daily basis both in terms of scale and scope. This past year has been terrible in terms of devastations and losses during hurricanes and earthquake in North America. Hurricanes Harvey and Irma are expected to lead to losses of more than \$150 billion US dollars due to damages and lost productivity (Dillow 2017). In addition, more than 200 lives have been lost and millions of people have suffered from power outages and shortage of basic necessities for an extended period of time in the United States and the Caribbean. In the same year, a 7.1 earthquake rattled Mexico City killing more than 150 people and leaving thousands struggling to get their lives back to normalcy (Buchanan et al. 2017). Based on the Intergovernmental Panel on Climate Change, NASA predicts that global warming could possibly lead to increase in natural calamities such as drought, intensity of storms, hurricanes, monsoons, and mid-latitude storms in the upcoming years. Simultaneously, the geo-political, social, and economic tensions have increased the need for humanitarian operations globally; such impacts have been experienced due to the crisis in Middle East, refugees in Europe, the systemic needs related to drought, hunger, disease, and poverty in the developing world, and the increased frequency of random acts of terrorism. According to the Global Humanitarian Assistance Report, 164.2 million people across 46 countries needed some form of humanitarian assistance in 2016 and 65.6 million people were displaced from their homes, the highest number witnessed thus far. At the same time, the international humanitarian aid increased to all time high of \$27.3 billion US

dollars from \$16.1 billion US dollars in 2012. Despite that increase, common belief is that funding is not sufficient to meet the growing humanitarian needs. Therefore, humanitarian organizations will continue to operate under capacity constraints and will need to innovate their operations to make them more efficient and responsive.

There are many areas in which humanitarian operations can improve. Humanitarian operations are often blamed for being slow or unresponsive. For example, the most recent relief efforts for Puerto Rico have been criticized for slow response. These organizations also face challenges in being able to sustain a policy or best practice for an extended period of time because of constant turnover in personnel. They are often blamed for being inefficient in how they utilize resources (Vanrooyen 2013). Some of the reasons that contribute to their inefficiency include operating environment such as infrastructure deficiencies in the last mile, socio-political tensions, uncertainty in funding, randomness of events and presence of multiple agencies and stake holders. However, it is critical that humanitarian operations show high level of performance so that every dollar that is routed in these activities is utilized to have the maximum impact on the people in need. Twenty-one donor governments and 16 agencies have pledged at the World Humanitarian Summit in 2016 to find at least one billion USD in savings by working more efficiently over the next 5 years (Rowling 2016).

We believe the best performing humanitarian operations need to have the following characteristics—they need to be **Rapid**, they have to be **Impactful** in terms of saving human lives, should be effective in terms of providing **Sustained** benefits and they

should be highly **Efficient**. We coin **RISE** as an acronym that succinctly describes the characteristics of successful humanitarian operations and it stands for **Rapid, Impactful, Sustained, and Efficient**.

One of the major opportunities for improving humanitarian operations lies in how data and information are leveraged to develop above competencies. Traditionally, humanitarian operations have suffered from lack of consistent data and information (Starr and Van Wassenhove 2014). In these settings, information comes from a diverse set of stakeholders and a common information technology is not readily deployable in remote parts of the world. However, the Big Data wave that is sweeping through all business environments is starting to have an impact in humanitarian operations as well. For example, after the 2010 Haiti Earthquake, population displacement was studied for a period of 341 days using data from mobile phone and SIM card tracking using FlowMinder. The data analysis allowed researchers to predict refugee locations 3 months out with 85% accuracy. This analysis facilitated the identification of cholera outbreak areas (Lu et al. 2012). Similarly, during the Typhoon Pablo in 2012, the first official crisis map was created using social media data that gave situation reports on housing, infrastructure, crop damage, and population displacement using metadata from Twitter. The map became influential in guiding both UN and Philippines government agencies (Meier 2012).

Big Data is defined as large volume of structured and unstructured data. The three V's of Big Data are Volume, Variety, and Velocity (McCafee and Brynjolfsson 2012). Big Data Analytics examines large amounts of data to uncover hidden patterns and correlations which can then be utilized to develop intelligence around the operating environment to make better decisions. Our goal in this article is to lay out a framework and present examples around how Big Data Analytics could enable RISE humanitarian operations.

2. Humanitarian Operations—Planning and Execution

Planning and Execution are critical aspects of humanitarian operations that deal with emergencies (like hurricanes) and systemic needs (hunger). All humanitarian operations have activities during preparedness phase (before) and disaster phase (during). Emergencies also need additional focus on the recovery phase (after). Planning and execution decisions revolve around **Where, When, How, and What**. We will take the UNICEF RUTF supply chain for the Horn of Africa (Kenya, Ethiopia, and Somalia) as an example. RUTF (ready to use therapeutic food) also called

Plumpy' Nut is a packaged protein supplement that can be given to malnourished children under the age of 5 years. The supplement was found to be very effective; therefore, the demand for RUTF skyrocketed, and UNICEF supply chain became over stretched (Swaminathan 2010). UNICEF supply chain showed many inefficiencies due to long lead times, high transportation costs, product shortages, funding uncertainties, severe production capacity constraints, and government regulations (So and Swaminathan 2009). Our analysis using forecasted demand data from the region found that it was important to determine where inventory should be prepositioned (in Kenya or in Dubai). The decision greatly influenced the speed and efficiency of distribution of RUTF. The amount of prepositioned inventory also needed to be appropriately computed and operationalized (Swaminathan et al. 2012). Given that the amount of funding and timing showed a lot of uncertainty, when funding was obtained, and how inventory was procured and allocated, dramatically influenced the overall performance (Natarajan and Swaminathan 2014, 2017). Finally, understanding the major roadblocks to execution and addressing those for a sustained solution had a great impact on the overall performance. In the UNICEF example, solving the production bottleneck in France was critical. UNICEF was able to successfully diversify its global supply base and bring in more local suppliers into the network. Along with the other changes that were incorporated, UNICEF RUTF supply chain came closer to being a RISE humanitarian operations and estimated that an additional one million malnourished children were fed RUTF over the next 5 years (Komrska et al. 2013). There are a number of other studies that have developed robust optimization models and analyzed humanitarian settings along many dimensions. While not an exhaustive list, these areas include humanitarian transportation planning (Gralla et al. 2016), vehicle procurement and allocation (Eftekar et al. 2014), equity and fairness in delivery (McCoy and Lee 2014), funding processes and stock-outs (Gallien et al. 2017), post-disaster debris operation (Lorca et al. 2017), capacity planning (Deo et al. 2013), efficiency drivers in global health (Berenguer et al. 2016), and decentralized decision-making (Deo and Sohoni 2015). In a humanitarian setting, the following types of questions need to be answered.

Where

- Where is the affected population? Where did it originate? Where is it moving to?
- Where is supply going to be stored? Where is the supply coming from? Where will the distribution points be located?

- c. Where is the location of source of disruption (e.g., hurricane)? Where is it coming from? Where is moving to?
- d. Where are the debris concentrated after the event?

When

- a. When is the landfall or damage likely to occur?
- b. When is the right time to alert the affected population to minimize damages as well as unwanted stress?
- c. When should delivery vehicles be dispatched to the affected area?
- d. When should supply be reordered to avoid stock-outs or long delays?
- e. When should debris collection start?

How

- a. How should critical resources be allocated to the affected population?
- b. How much of the resources should be prepositioned?
- c. How many suppliers or providers should be in the network?
- d. How to transport much needed suppliers and personnel in the affected areas?
- e. How should the affected population be routed?

What

- a. What types of calamities are likely to happen in the upcoming future?
- b. What policies and procedure could help in planning and execution?
- c. What are the needs of the affected population? What are reasons for the distress or movement?
- d. What needs are most urgent? What additional resources are needed?

3. Big Data Analytics

Big Data Analytics can help organizations in obtaining better answers to the above types of questions and in this process enable them to make sound real-time

decisions during and after the event as well as help them plan and prepare before the event (see Figure 1). Descriptive analytics (that describes the situation) could be used for describing the current crisis state, identifying needs and key drivers as well as advocating policies. Prescriptive analytics (that prescribes solutions) can be utilized in alert and dispatch, prepositioning of supplies, routing, supplier selection, scheduling, allocation, and capacity management. Predictive analytics (that predicts the future state) could be utilized for developing forecasts around societal needs, surge capacity needs in an emergency, supply planning, and financial needs. Four types of data streams that could be utilized to develop such models are social media data, SMS data, weather data, and enterprise data.

3.1. Social Media Data

The availability of data from social media such as Twitter has opened up several opportunities to improve humanitarian emergency response. Descriptive analytics from the data feed during an emergency could help create the emergency crisis map in rapid time and inform about areas of acute needs as well as movement of distressed population. This could help with rapid response into areas that need the most help. Furthermore, such data feed could also be used to predict the future movement of the affected population as well as surges in demand for certain types of products or services. A detailed analysis of these data after the event could inform humanitarian operations about the quality of response during the disaster as well as better ways to prepare for future events of a similar type. This could be in terms of deciding where to stock inventory, when and how many supply vehicles should be dispatched and also make a case for funding needs with the donors. Simulation using social media data could provide solid underpinning for a request for increased funding. Analysis of information diffusion in the social network could present new insights on the speed and efficacy of messages relayed in the social network (Yoo et al. 2016). Furthermore, analyzing the population movement data in any given region of interest could provide valuable input for

Figure 1 Big Data Analytics and Rapid, Impactful, Sustained, and Efficient Humanitarian Operations

Descriptive	Prescriptive	Predictive
Create Crisis Maps Identify Needs Advocate Policies Identify Success Drivers	Alert and Dispatch Preposition Supplies Supplier Selection Routing Scheduling/Allocation Capacity Management	Societal Needs Planning Supply Planning Surge Capacity Planning Shelter Capacity Planning

ground operations related to supply planning, positioning, and vehicle routing. Finally, social media data is coming from the public directly and sometimes may contain random or useless information even during emergency. There is an opportunity to develop advanced filtering models so that social media data are leveraged in real-time decision-making.

3.2. SMS Data

Big Data Analytics can also be adapted successfully for SMS-based mobile communications. For example, a number of areas in the United States have started using cell phone SMS to text subscribers about warnings and alerts. Timely and accurate alerts can save lives particularly during emergencies. Predictive analytics models can be developed to determine when, where, and to whom these alerts should be broadcasted in order to maximize the efficacy of the alerts. The usage of mobile alerts is gaining momentum in the case of sustained humanitarian response as well. For example, frequent reporting of inventory at the warehouse for food and drugs can reduce shortages. Analytics on these data could provide more nuances on the demand patterns which in turn could be used to plan for the correct amount and location of supplies. Mobile phone alerts have also shown to improve antiretroviral treatment adherence in patients. In such situations, there is a great opportunity to analyze what kinds of alerts and what levels of granularity lead to the best response from the patient.

3.3. Weather Data

Most regions have highly sophisticated systems to track weather patterns. This type of real-time data is useful in improving the speed of response, so that the affected population can be alerted early and evacuations can be planned better. It also has a lot of information for designing humanitarian efforts for the future. For example, by analyzing the data related to the weather changes along with population movement, one could develop robust prescriptive models around how shelter capacity should be planned as well as how the affected population should be routed to these locations. So, rather than trying to reach a shelter on their own, an affected person can be assigned a shelter and directed to go there. Prepositioning of inventory at the right locations based on weather data could improve response dramatically as reflected by the actions of firms such as Wal-Mart and Home Depot that have made it a routine process after successful implementation during hurricane Katrina. Finally, the weather pattern data could be utilized to develop predictive models around the needs of the population in the medium to long term. For example,

the drought cycles in certain regions of Africa follow a typical time pattern. A predictive model around the chances of famine in those regions could then inform the needs and funding requirements for food supplements.

3.4. Enterprise Data

Most large humanitarian organizations such as UNICEF have information systems that collect a large amount of data about their operations. Analytics on such data can be useful to develop robust policies and guide the operational decisions well. For example, in systemic and emergent humanitarian needs, analyzing the demand and prepositioning inventory accordingly has shown to improve the operational performance. Furthermore, the analysis of long-term data could provide guidelines for surge capacity needed under different environments as well as predict long-term patterns for social needs across the globe due to changing demographics and socioeconomic conditions.

As the Big Data Analytics models and techniques develop further, there will be greater opportunities to leverage these data streams in more effective ways, particularly, given that the accuracy of data coming out of the different sources may not have the same level of fidelity in a humanitarian setting. While data are available in abundance in the developed world, there are still geographical areas around the globe where cell phone service is limited, leave alone social media data. In those situations, models with incomplete or missing data need to be developed. Also the presence of multiple decentralized organizations with varied degree of information technology competencies and objectives limits their ability to effectively synthesize the different data streams to coordinate decision-making.

4. Concluding Remarks

Big data has enabled new opportunities in the value creation process including product design and innovation (Lee 2018), manufacturing and supply chain (Feng and Shanthikumar 2018), service operations (Cohen 2018), and retailing (Fisher and Raman 2018). It is also likely to impact sustainability (Corbett 2018), agriculture (Devalkar et al. 2018), and healthcare (Guha and Kumar 2018). In our opinion, humanitarian organizations are also well positioned to benefit from this phenomenon. Operations Management researchers will have opportunity to study newer topics and develop robust models and insights that could guide humanitarian operations and make them more *Responsive, Impactful, Sustained, and Efficient*.

Acknowledgments

The author wishes to thank Gemma Berenguer, Anand Bhatia, Mahyar Eftekar, and Jarrod Goentzel for their comments on an earlier version of this study.

References

- Berenguer, G., A. V. Iyer, P. Yadav. 2016. Disentangling the efficiency drivers in country-level global health programs: An empirical study. *J. Oper. Manag.* **45**: 30–43.
- Buchanan, L., J. C. Lee, S. Pechanha, K. K. R. Lai. 2017. Mexico City before and after the earthquake. *New York Times*, September 23, 2017.
- Cohen, M. C. 2018. Big data and service operations. *Prod. Oper. Manag.* **27**(9): 1709–1723. <http://doi.org/10.1111/poms.12832>.
- Corbett, C. J. 2018. How sustainable is big data? *Prod. Oper. Manag.* **27**(9): 1685–1695. <http://doi.org/10.1111/poms.12837>.
- Deo, S., M. Sohoni. 2015. Optimal decentralization of early infant diagnosis of HIV in resource-limited settings. *Manuf. Serv. Oper. Manag.* **17**(2): 191–207.
- Deo, S., S. Iravani, T. Jiang, K. Smilowitz, S. Samuelson. 2013. Improving health outcomes through capacity allocation in a community based chronic care model. *Oper. Res.* **61**(6): 1277–1294.
- Devalkar, S. K., S. Seshadri, C. Ghosh, A. Mathias. 2018. Data science applications in indian agriculture. *Prod. Oper. Manag.* **27**(9): 1701–1708. <http://doi.org/10.1111/poms.12834>.
- Dillow, C. 2017. The hidden costs of hurricanes. *Fortune*, September 22, 2017.
- Eftekar, M., A. Masini, A. Robotis, L. Van Wassenhove. 2014. Vehicle procurement policy for humanitarian development programs. *Prod. Oper. Manag.* **23**(6): 951–964.
- Feng, Q., J. G. Shanthikumar. 2018. How research in production and operations management may evolve in the era of big data. *Prod. Oper. Manag.* **27**(9): 1670–1684. <http://doi.org/10.1111/poms.12836>.
- Fisher, M., A. Raman. 2018. Using data and big data in retailing. *Prod. Oper. Manag.* **27**(9): 1665–1669. <http://doi.org/10.1111/poms.12846>.
- Gallien, J., I. Rashkova, R. Atun, P. Yadav. 2017. National drug stockout risks and global fund disbursement process for procurement. *Prod. Oper. Manag.* **26**(6): 997–1014.
- Gralla, E., J. Goentzel, C. Fine. 2016. Problem formulation and solutions mechanisms: A behavioral study of humanitarian transportation planning. *Prod. Oper. Manag.* **25**(1): 22–35.
- Guha, S., S. Kumar. 2018. Emergence of big data research in operations management, information systems and healthcare: Past contributions and future roadmap. *Prod. Oper. Manag.* **27**(9): 1724–1735. <http://doi.org/10.1111/poms.12833>.
- Komrska, J., L. Kopczak, J. M. Swaminathan. 2013. When supply chains save lives. *Supply Chain Manage. Rev.* **January–February**: 42–49.
- Lee, H. L. 2018. Big data and the innovation cycle. *Prod. Oper. Manag.* **27**(9): 1642–1646. <http://doi.org/10.1111/poms.12845>.
- Lorca, A., M. Celik, O. Ergun, P. Keskinocak. 2017. An optimization based decision support tool for post-disaster debris operations. *Prod. Oper. Manag.* **26**(6): 1076–1091.
- Lu, X., L. Bengtsson, P. Holme. 2012. Predictability of population displacement after 2010 Haiti Earthquakes. *Proc. Natl Acad. Sci.* **109**(29): 11576–11581.
- McCafee, A., E. Brynjolfsson. 2012. Big data: The management revolution. *Harvard Business Review*, October 1–9, 2012.
- McCoy, J., H. L. Lee. 2014. Using fairness models to improve equity in health delivery fleet management. *Prod. Oper. Manag.* **23**(6): 965–977.
- Meier, P. 2012. How UN used social media in response to typhoon Pablo. Available at <http://www.irevolutions.org> (accessed date December 12, 2012).
- Natarajan, K., J. M. Swaminathan. 2014. Inventory management in humanitarian operations: Impact of amount, schedule, and uncertainty in funding. *Manuf. Serv. Oper. Manag.* **16**(4): 595–603.
- Natarajan, K., J. M. Swaminathan. 2017. Multi-Treatment Inventory Allocation in Humanitarian Health Settings under Funding Constraints. *Prod. Oper. Manag.* **26**(6): 1015–1034.
- Rowling, M. 2016. Aid efficiency bargain could save \$1 billion per year. *Reuters*, May 23, 2016.
- So, A., J. M. Swaminathan. 2009. The nutrition articulation project: A supply chain analysis of ready-to-use therapeutic foods to the horn of Africa. UNICEF Technical Report.
- Starr, M., L. Van Wassenhove. 2014. Introduction to the special issue on humanitarian operations and crisis management. *Prod. Oper. Manag.*, **23**(6), 925–937.
- Swaminathan, J. M. 2010. Case study: Getting food to disaster victims. *Financial Times*, October 13, 2010.
- Swaminathan, J. M., W. Gilland, V. Mani, C. M. Vickery, A. So. 2012. UNICEF employs prepositioning strategy to improve treatment of severely malnourished children. Working paper, Kenan-Flagler Business School, University of North Carolina, Chapel Hill.
- Vanrooyen, M. 2013. Effective aid. *Harvard International Review*, September 30, 2013.
- Yoo, E., W. Rand, M. Eftekar, E. Rabinovich. 2016. Evaluating information diffusion speed and its determinants in social networks during humanitarian crisis. *J. Oper. Manag.* **45**: 123–133.

Data Science Applications in Indian Agriculture

Sripad K. Devalkar, Sridhar Seshadri*

Indian School of Business, Gachibowli, Hyderabad 500032, India, sripad_devalkar@isb.edu, sridhar_seshadri@isb.edu

Chitrabhanu Ghosh

Independent, Bengaluru 560103, India, chitrabhanu_ghosh_2016@cba.isb.edu

Allen Mathias

Microsoft India R&D Private Limited, Gachibowli, Hyderabad 500032, India, allen_mathias_2016@cba.isb.edu

Agricultural supply chains in the developing world face the daunting task of feeding a growing population in the coming decades. Along with the provision of food, sustaining livelihoods, enhancing nutrition and the ability to cope with rapid changes in the environment and marketplaces are equally important to millions of small farmers. Data science can help in many ways. In this article, we outline the beginnings of data science applications in Indian agriculture. We cover various initiatives such as data collection, visualization and information dissemination, and applications of algorithmic data analysis techniques for decision support. We describe one application under development that provides timely price information to farmers, traders, and policy makers.

Key words: data science; agricultural operations; India; price analytics

History: Received: October 2017; Accepted: November 2017 by Kalyan Singhal, after 1 revision.

1. Introduction

The food output of the world has to increase by 50% by 2050 to cater to the needs of a growing world population (FAO 2017). It is also a fact that a significant portion of the demand will come from developing countries in Asia and Africa. And the burden of increasing the output will most likely fall on large producers such as China, India, and the United States. In this context, there has been a call to modernize every aspect of the sector (The Economist 2016). Several initiatives, both in the public and the private sectors, are underway to explore the use of analytics and data science in agriculture. The Operations Management profession is also heeding the call to arms (INFORMS 2016).

This essay is about data science applications in India, where the application context is very different. The Indian agricultural sector is characterized by small landholdings, with an average landholding size of one hectare and more than 120 million farmers that are classified as marginal or small farmers. While the agricultural sector's contribution to India's GDP has shrunk from close to 52% in the 1950s to about 14% in 2011, over 50% of India's labor force is still employed in this sector (Government of India 2016). Despite the fragmented holdings, India is the second largest agricultural producer in the world (based on FAO data available at <http://www.fao.org/faostat/en/#data>).

Along with being tasked with the provision of food, sustaining livelihoods, enhancing nutrition and being prepared to cope with rapid changes in the environment and marketplaces are equally important to millions of small farmers.

In developed countries, the infrastructure for supporting agriculture is well developed. Information about inputs and markets for produce is easily accessible for most players in the agricultural supply chain. As a result, most of the focus has been on data science applications for improving the production process (see Wolfert et al. 2017, and articles alluded to above). In contrast, initiatives in the Indian context have tried to address more basic aspects of the agricultural supply chain, such as, strengthening the infrastructure that supports agricultural activity, increasing access to, and information about input and output markets and the cultivation process to various degrees. These interventions are aimed at improving:

- The infrastructure that supports agricultural activity, including connectivity and communications support, provision of financial assistance, and crop insurance.
- Access to, and information about inputs used in agricultural activity. These include technology to measure soil health, suitability of different fertilizers, pesticides, seeds research, information about the weather, etc.

- Production process, including use of sensors to collect data about crop growth on a continuous basis and provide targeted advice as well as inputs, use of automated harvesters for higher harvesting yields, etc.
- Access to, and information about output markets to enable farmers to identify markets to sell in, including data about prices and availability of various commodities in different markets in a neighbourhood.

In the rest of the article, we first provide an overview of various existing data science applications. We use the data value chain framework of Miller and Mork (2013) to classify them into the following broad life cycle stages: (i) data collection, cleaning and archiving, (ii) data visualization and information dissemination, (iii) application of algorithmic data analysis techniques to derive insights and (iv) using insights from data analysis for decision making. We then describe an application under development that spans all the different stages.

2. Existing Data Science Applications in the Indian Agriculture Sector

2.1. Data Collection Applications

The application of data science in the Indian agriculture sector traces back to the early periods of post-independence India (after 1947) when the government set up the National Sample Survey Office (NSSO) with the aim of creating a database of information related to the key sectors of the economy. Given the complexity of the Indian economy, the cost and time required for frequent census of the different sectors was enormous. The cost had to be mitigated by the clever use of data collection and estimation techniques. The Indian Statistical Institute (ISI) was entrusted with the development of a sampling strategy to collect data using nationally representative samples and develop methods to use the sample data to make inferences (Rao 2010). While the NSSO survey methodology continues to be used for collecting data about various aspects of the Indian economy, a well-developed, decentralized Agriculture Statistics System exists today to collect and compile various agricultural statistics such as land holding, land and crop use, crop production forecasts, etcetera (Srivastava 2012). These methods have been used to estimate agricultural production and perhaps, indirectly contributed to the implementation of the “Green Revolution” agricultural reforms in the 1960s. The Green Revolution enabled India to transform itself from a food-grain deficient to a food-grain surplus country. According to Government of India (2015) the total production of food grains has increased from

50.8 million tonnes in 1950–1951 to over 257 million tonnes in 2012–2013, more than enough to meet the current population’s consumption requirement of 160 million tonnes.

While the existing, predominantly manual, process continues to be used for collecting data related to agriculture, there have been various initiatives to use technology to collect data related to the physical, financial and legal infrastructure that supports agricultural activity in the country. Many state governments have undertaken projects such as digitization of land records for accurate recording of land ownership and creation of soil health cards to capture information about soil quality (see, e.g., Meebhoomi 2017, a portal where land owners in the state of Andhra Pradesh can access their land records electronically). The data collected can enable targeted advice to farmers about better farming practices suited for their landholdings. Together with other programs, such as, improving financial assistance and loan availability to backward classes and instituting crop insurance support, these steps help create an environment that is conducive to realizing higher profits from agricultural activity (see Chhatre et al. 2016 for a discussion on factors constraining farmers in India). The challenge continues in being able to collect and store the data on a continuous basis in a costless manner. Future efforts to improve data collection will have to focus on developing processes and technology for automated collection and dissemination of information to make sure these initiatives can be easily scaled.

2.2. Information Dissemination Applications

The dissemination of information is equally critical to development efforts. Therefore, alongside the data gathering applications described above, various schemes are underway to disseminate the information gathered to different stakeholders in the agricultural supply chain. State governments have invested in developing satellite based communication networks to provide information and advice to farmers (Indian Space Research Organization 2000). This network can be used to broadcast systematic and targeted information, e.g., recommendations on fertilizer use and crop rotation to farmers by combining the digitized land records and soil health information for individual farmers. In addition to targeted information, the government has also developed portals to disseminate technical information; e.g., *Mkisan* is a government run information portal for farmers (Ministry of Agriculture and Farmers Welfare 2017b). While the information provided through these portals is not specific to an individual farmer, they nevertheless have to take into account the wide diversity of agricultural operations in India. Significant diversity in agro-climatic conditions and soil quality, combined with

differential access to irrigation, varied rainfall patterns and diversity in language mean that information disseminated through such portals is state/district specific. For example, the Agrinet portal maintained by the Andhra Pradesh state government provides information on farming practices, availability of inputs, and advice that is specific to farmers in the state, in the local language, Telugu (Agriculture Department 2017). One may even envisage the concept of a one-to-one mapping of virtual and real farms in the future! Such a map can create a virtual reality where farmers can experiment with different ideas and obtain advice without waiting for the rains and harvest to take place. The idea of virtualization has already been implemented in manufacturing, thus, extension to agriculture is not far in the future.

A bigger challenge in translating the productivity gains from initiatives on the input side to higher profits for farmers is the lack of adequate access to markets and reliable information about markets. Farmers in India, especially small farmers, face significant hurdles in accessing markets to sell their output—they often do not have accurate information about prices in the output market. As a result, farmers are heavily dependent on middlemen, who influence what farmers grow, what they pay for services and the price of the output, thus, depressing the returns that farmers get from cultivation. This in turn makes it less affordable for farmers to invest in technologies that can help improve farm productivity. In recent years, many private organizations have tried to address this issue by providing close to real-time market information to farmers (e.g., RML Agtech Pvt. Ltd., formerly known as Reuters Market Light). Some agri-business firms have also invested in procurement infrastructure to provide viable channels for farmers to sell directly to the firms rather than be dependent on middlemen (e.g., ITC e-Choupal). The government has also taken steps to provide market information through public portals and mobile applications through the Agmarket portal and AgriMarket App (Ministry of Agriculture and Farmers Welfare 2017a), and more recently enable access to markets through the Electronic National Agriculture Market platform (NAM 2017). One interesting area for study would be the unbundling, fragmenting, and servicification of the activities in the agriculture supply chain as and when “data science as a decision-making tool” gets accepted and used widely (see Bhagwati 1984, a prescient paper that discusses the possibility of fragmentation and outsourcing of service processes.)

2.3. Algorithmic Data Analysis and Decision Support Applications

In the United States and other developed countries, many private companies have developed

technological and data analytics solutions targeted towards providing decision support systems to farmers to improve productivity (see Sparapani 2017 for examples). Unlike developed countries in the West, there has not been any large scale adoption of such technologies to-date in India. Part of the reason for this is the nature of landholding in India, where the average landholding size for farmers is very small, of the order of one hectare, making investments in such technology unviable at the level of a single farmer. Also, because the agricultural sector is a major employer, capital intensive technologies might meet resistance due to reduction in the usage of labour. Thus, a careful analysis of the types of innovations that benefit small farms is necessary. At this moment, the ideas are scattered and available only as anecdotes. Also, Indian agriculture is primarily rain-fed. Therefore, soil conservation, water conservation, proper crop rotation, and practice of the art, and craft handed out over generations combined with modern science and data is very much necessary (Chaudhary and Singh 2016).

Some organizations, both private and government, are foraying into using big data analytics techniques to provide decision support to small farmers (Ramesh 2015). For example, ICRISAT is piloting a service which uses data intensive climate simulation combined with crop growth models to provide advice to peanut farmers in the district of Anantapur in Andhra Pradesh. The decision support provided by ICRISAT includes advice on when to prepare the land for sowing contingent on current rainfall in the growing season, which variety of seed to sow given the weather and soil conditions specific to the farmer, what inputs to use and such (ICRISAT 2017).

Unlike initiatives similar to the green revolution where the goal was to improve productivity and ensure overall food sufficiency—the current context needs applications that are concomitantly geared towards increasing returns to employment in the agricultural sector, provide stable income and drive economic growth. A necessary factor for improving profitability from farming is to connect farmers to markets & markets to farms, provide visibility of market prices, and enable farmers to identify crops that have the potential to generate more profits. Past examples of success show that individual entrepreneurs have worked with farmer-communities to grow crops that are not necessarily consumed in the local markets but have high demand in other markets to increase supply chain profitability (these include age-old examples of tea and coffee plantation in India, to a recent example of how Gherkin cultivation has improved livelihood for farmers in the state of Karnataka (Basu 2014)). Use of data science techniques, spanning the different stages of the lifecycle described

earlier, can play a significant role in coordinating agricultural supply chains and enabling farmers to discover such profitable opportunities.

To the best of our knowledge, while existing applications in India provide access to information, they barely provide analytics or decision support tools based on the data. Herein lies the opportunity for Operations Management researchers to contribute to a truly worthwhile cause. We describe below a potential application of data analytics that collects available information from output markets and renders it useful for various players in the agricultural supply chain such as, farmers, traders, and policy makers.

3. A Specific Example—Market Price Analytics Application for Agricultural Commodities

For many farmers in India, price discovery and price realization for their produce is simultaneous, i.e., farmers know what price they will get for their produce only after it is taken to the market, after having incurred significant costs of travel and time. As a result, farmers often settle for lower prices than what they possibly can get. Given the small scale of operations many farmers are dependent on intermediaries for marketing their produce, leading to double-marginalization and hence lower prices. As a consequence, the net income of most farmers in India is around 25,000 INR (little less than 400 USD) per year, leaving most of them below the poverty line (Government of India 2014).

Ensuring easy access to price information should help address the farmers' information disadvantage regarding market prices. Combined with information visibility, decision support tools to help the farmers decide when, and to which market, to take their produce will help improve the returns to farming. Similarly, while intermediaries/traders possibly have better information about market prices, a decision support tool that allows traders to improve profits from marketing agricultural produce will improve overall welfare. Additionally, data analysis algorithms based on price and volumes traded across multiple markets can help policy makers identify anomalies in prices and volumes traded, differences in patterns across years and markets, et cetera and take steps to correct any imbalances. All these different applications will have the effect of improving overall supply chain surplus, with some of these gains accruing to the farmers.

The market price analytics application we are currently developing (expected to be available by May 2018) is aimed at providing all of the above benefits to the various stakeholders in the agricultural supply chain. Below, we describe the data used by the

application, the process of data collection and cleaning, and key use cases describing how farmers, traders and policy makers can use this application.

3.1. Data Source and Data Collection Process

The Government of India has mandated that every regulated agricultural commodity market in India publish details of daily transactions. These data are collated by the states and aggregate data at the daily level is made publicly available through the Agmarket portal (Ministry of Agriculture and Farmers Welfare 2017a). We rely on this portal for our data. Data from each physical agricultural market (known as a "mandi") across India are updated on the portal on a daily basis. We are able to "scrape" data from the portal every day, after the closing time of the current day and before the opening time of the next day. Currently, the application is focused on the state of Punjab, which has over 450 mandis. In the future, the scope of the application will be expanded to include all mandis in India. The data collected consists of the fields shown in Table 1. After obtaining the data, a number of data cleaning steps are performed to handle missing or incomplete information and inconsistencies in data about prices and arrival.

While the application is currently envisaged to provide information about prices and arrivals and decision support with regards to marketing and procurement for farmers and traders, the application has the potential to add value to other players in the agricultural eco-system by integrating other data sources and providing an interface for providers of ancillary services such as logistics, banking and insurance, etc. In addition, the physical agricultural market data in the application can be complemented with data such as weather forecasts and spot and futures prices from various commodity exchanges. Appendix A provides a schematic of how this application can integrate with

Table 1 Description of Data Collected by the Application from agmarket.gov.in

Field	Description
Date	Date of reporting
Market (Mandi) name	Name/Address of the Market (Mandi)
Commodity name	Uniquely identifiable commodity name
Variety	The variety/variant (grade) of the commodity
Group	Food group of the commodity (cereal, pulse, etc.)
Minimum selling price	Minimum price at which the commodity was sold on the day
Maximum selling price	Maximum price at which the commodity was sold on the day
Modal selling price	Modal price at which the commodity was sold on the day
Arrivals	Volume of commodity transacted (in 100s of Kg) on the day

various data sources, and the steps in the data collection and cleaning process.

3.2. Price and Arrivals Information Dissemination

With over 500 million active mobile subscribers and growing, India is on course to become one of the largest mobile telephone markets in the world. The high level of mobile phone penetration, combined with modest incomes and limited literacy levels imply that the Indian farmer is of short of means, is more likely to have access to a feature phone without any internet access, and can only read and write to a limited extent in the native language. To this end, the application provides services via an SMS-based system that the farmer can interact with using any simple feature phone. The SMS-based system is available in the local language of the state where the farmer resides, and the farmer can interact with the system using fixed-form text. The farmer can query the application to enquire about the price of any particular crop in the mandis closest to him/her. The mandis closest to the farmer are identified based on the farmer's location either through GPS tracking of the phone or pre-determined location of the farmer specified at the time of registration for the service. In addition to prices, the farmer can also query for the arrivals at the mandis closest to him/her, prices and/or quantity for a given commodity at a particular mandi in the immediate past week. The application also allows users such as traders and farmers with access to smartphones and internet access to make more sophisticated queries through the smartphone app and/or web portal of the application. Large buyers who wish to transact at one or more mandis and buy from one or more farmer (or, alternately large farmers who wish to sell at one or more mandis) can use the application to identify mandi(s) with the best price to fulfil their buying (or, sales) needs. Appendix B provides a schematic on how different users interact with the application through SMS/web-based portal/smartphone app.

3.3. Algorithmic Data Analysis and Decision Support

In addition to information about historical prices and arrivals, the application provides localized information to help farmers and traders make decisions on when and where to transact for the produce. Automated forecasting and advanced algorithmic data analysis embedded in the application can generate price and arrival predictions for different commodities in the upcoming week at a particular mandi or group of mandis. These predictions can be combined with optimization techniques to help buyers (sellers) to identify the combination of mandis that can provide the quantity required (to be sold) at the best

price, source (or, sell) multiple commodities through multiple mandis within a stipulated price and distance range.

The application can also provide an aggregate view across all mandis to policy makers and regulators. The application's database has price and quantity information for multiple commodities across all mandis for multiple years. This data can be mined to identify patterns and help in making effective, data-driven policy interventions. Some of the important questions that this application helps answer at a macro-level, include: information about aggregate arrivals of a particular commodity across the mandis of a state/district, identify patterns in arrivals over multiple years and identify deviations from the past, predict estimated arrivals/prices of particular commodities in the upcoming season at major mandis, using algorithmic techniques on historical data from previous years and combine it with other information such as weather and yield forecasts for the current season, track deviation between supply and demand, monitor deviations from the minimum support price guaranteed for some commodities by the government, use the data to create region-specific price indices for food, create dynamic models of supply-demand matching on a spatial network, etcetera. We do not provide specific use cases but expect to allow integration with the data to enable such applications.

In addition to the data collected from external sources mentioned before, an additional source of data, namely the queries and searches performed by various users of the application, will become available to be mined and generate insights. For example, based on queries by farmers, the application will be able to analyze patterns and predict likely arrivals of a commodity at a given mandi on a given day or week. Similarly, based on queries by the traders the application will be able to predict expected demand for various commodities at a given mandi, and based on the estimated supply and demand, forecast prices for major commodities that are likely to prevail in mandis within a specific geographic region. These and other possible insights from mining the social media data on the platform will be useful for members of the agricultural supply chain to make better decisions and improve profitability. One can imagine applications wherein bulletin boards can be provided to help clear specific commodities that are highly seasonal or of high quality. We leave it to the readers' imagination to speculate about other possible benefits that the application will bring about.

The obvious question is whether the cost of developing and maintaining this application at scale is going to be prohibitive, rendering this a mere theoretical exercise and a small scale proof of concept. The surprising answer is no. The costs of developing and

maintaining this application—hardware costs for data storage, software development, manpower costs for maintaining the application, incremental costs of data acquisition from multiple sources, cost for users to access the application—can be estimated in a fairly straightforward manner (see Appendix C for an estimate of the development and maintenance costs of the application). Given the potential impact of the information, the low cost, less than USD 3500 per month, reflects the progress data science has made in the last decade.

4. Data Science and Social Change

The role of big data in transforming businesses and society is becoming more and more apparent with the convergence of data availability, technology, and democratization of access (see Fisher and Raman 2018 and Swaminathan 2018, in this volume for two examples of how big data can help improve retail and humanitarian operations). To play a truly transformational role, data science applications have to additionally take into account local contextual factors.

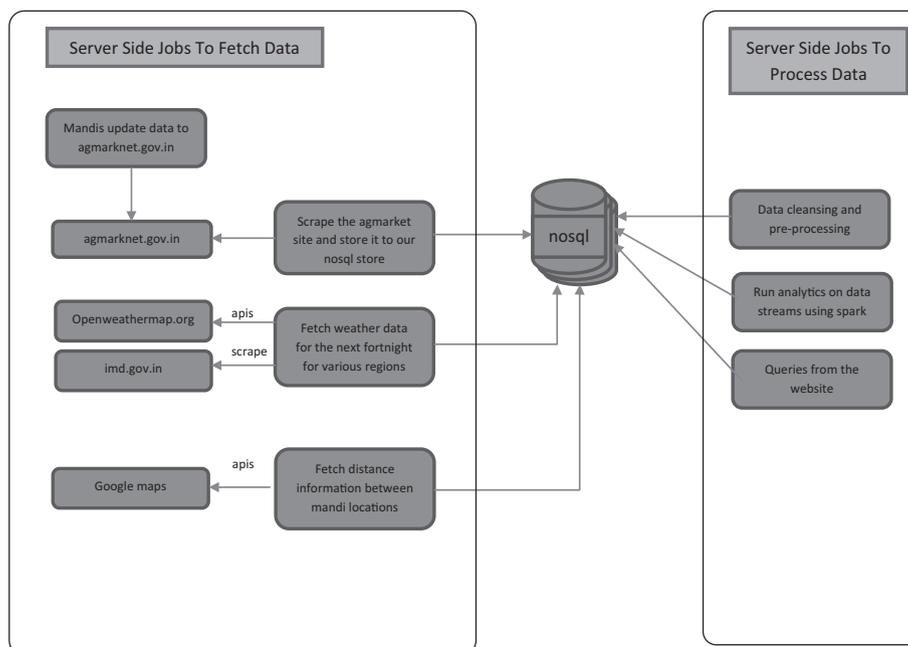
For example, a key feature of the Indian agricultural sector, and those in other developing countries, is that in addition to being a major supplier of food to the world, the sector is also a major employer. Thus, while we discussed the technological and informational aspects related to various data science applications, we should note that the ability of these

applications to address the challenges facing the Indian agricultural sector also depends on having the right combination of user-skills and market processes. To this end, there is a need to equip the users with the right skills to make the best use of the data science applications. We also cannot end this essay without mentioning the need for good governance and regulation. In the absence of safeguards and provisions of recourse, it is likely that these applications can lead to asymmetric distribution of benefits and leave the most vulnerable players worse-off (see Helbing et al. 2017 for a discussion on the pitfalls of control of data and information being centralized in the hands of a few, large enterprises and/or governments). Thus, with careful design, we foresee data science as being truly transformational in the techno-social-political-economic sense in rural India.

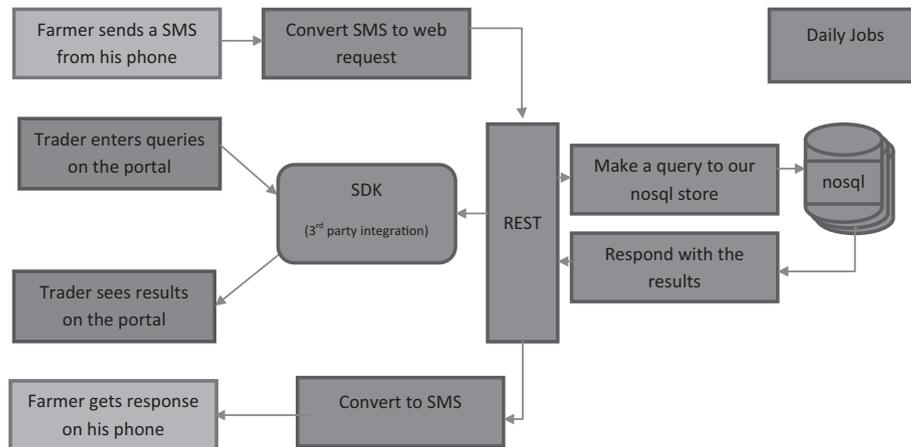
Acknowledgments

The authors thank Prasanna Tantri and Sai Harsha Katuri at the Centre for Analytical Finance, and Ashwini Chhatre, at the Indian School of Business for their help in collecting the price and arrivals data from various mandis that provided the starting point for the development of the application discussed in this essay. The authors also thank Lal Mani Pandey, Bhimasankaram Pochiraju and Deepak Agrawal at the Applied Statistics and Computing Lab at the Indian School of Business for their support. Two anonymous reviewers helped with their comments. We are very grateful to the editor-in-chief, Kalyan Singhal, for the opportunity to write this essay.

Appendix A. Data Collection and Cleaning Process [Color figure can be viewed at wileyonlinelibrary.com]



Appendix B. Process Flow for the Interaction between Users and the Application [Color figure can be viewed at wileyonlinelibrary.com]



Appendix C. Application Development and Maintenance Costs

Item	Cost per month (in INR)	Reference
Application development phase		
Computation Resources	71,000	For dedicated computing resources on m4 instance https://aws.amazon.com/ec2/dedicated-hosts/pricing/
Data Storage	20,000	Estimated 10TB initial requirement https://aws.amazon.com/govcloud-us/pricing/s3/
API access	32,500	https://openweathermap.org/price
Backend development	3 × 20,000	3 developers
UX development	3 × 20,000	2 developers
Office space	50,000	
Total monthly expenses	293,500	We estimate development to take ~6 months
Maintenance phase		
Computation, Storage, API access	123,500	
Development team	2 × 20,000	2 developers in the long run to maintain the application
Office space	50,000	
Total monthly expenses	213,500	

Note. One USD = 65 INR approximately.

References

- Agriculture Department. 2017. E-Vyavasayam (“E-Agriculture”). Available at <http://www.apagrisnet.gov.in/> (accessed date December 5, 2017).
- Basu, S. 2014. Go Gherkin. Available at <http://www.thehindu.com/features/metroplus/society/go-gherkin/article6291788.ece> (accessed date November 10, 2017).
- Bhagwati, J. N. 1984. Splintering and disembodiment of services and developing nations. *World Economy* 7(2): 133–144.
- Chaudhary, V., S. Singh. 2016. What is the future of agriculture in India? Available at <https://thewire.in/52228/what-is-the-future-of-agriculture-in-india/> (accessed date November 10, 2017).
- Chhatre, A., S. K. Devalkar, S. Seshadri. 2016. Crop diversification and risk management in Indian agriculture. *Decision* 43(2): 167–179.
- FAO, IFAD, UNICEF, WFP and WHO. 2017. The state of food security and nutrition in the world 2017: Building resilience for peace and food security. Report, Food and Agriculture Organization, Rome, Italy.
- Fisher, M., A. Raman. 2017. Using data and big data in retailing. *Prod. Oper. Manag.* 27(9): 1665–1669. <https://doi.org/10.1111/poms.12846>.
- Government of India. 2014. Key indicators of situation of agricultural households in India, NSS 70th Round, Report. Ministry of Statistics and Programme Implementation, New Delhi, India.
- Government of India. 2015. Agricultural statistics at a glance, Report. Ministry of Agriculture, Department of Agriculture and Cooperation, New Delhi, India.
- Government of India. 2016. State of Indian agriculture. Report, Ministry of Agriculture and Farmers Welfare, Department of Agriculture, Cooperation and Farmers Welfare, New Delhi, India.
- Helbing, D., B. S. Frey, G. Gigerenzer, E. Hafen, M. Hagner, Y. Hofstetter, van den Hoven J., R. V. Zicari, A. Zwitter. 2017. Will democracy survive big data and artificial intelligence. *Scientific American*, February 25. Available at <https://www>

- scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/ (accessed date December 5, 2017).
- ICRISAT. 2017. Integrated assessment of climate change impacts on groundnut productivity in Anantapur, India. Available at <http://spatial-tools.icrisat.ac.in/Geographical%20Information%20System.aspx> (accessed date December 5, 2017).
- Indian Space Research Organization. 2000. ISRO And Andhra Pradesh Sign MOU on Satellite Based Communication Network. Available at <https://www.isro.gov.in/update/08-jul-2000/isro-and-andhra-pradesh-sign-mou-satellite-based-communication-network> (accessed date December 5, 2017).
- INFORMS. 2016. Feeding the world through analytics. Available at <http://pubsonline.informs.org/editorscut/agribusiness/industry> (accessed date December 5, 2017).
- Meebhoomi. 2017. Meebhoomi. Available at <http://meebhoomi.aip.gov.in/Home.aspx> (accessed date December 5, 2017).
- Miller, H. G., P. Mork. 2013. From data to decisions: a value chain for big data. *IT Prof.* 15(1): 57–59.
- Ministry of Agriculture and Farmers Welfare. 2017a. AgMarknet - Connecting farmer to markets. Available at <http://agmarknet.gov.in/> (accessed date December 5, 2017).
- Ministry of Agriculture and Farmers Welfare. 2017b. MKisan. Available at <http://mkisan.gov.in/> (accessed date December 5, 2017).
- NAM. 2017. National Agriculture Market. Available at <http://enam.gov.in> (accessed date December 5, 2017).
- Ramesh, R. 2015. 10 technological innovations that are revolutionizing Indian agriculture. Available at <http://www.thealternative.in/business/10-technological-innovations-revolutionizing-indian-agriculture/> (accessed date November 10, 2017).
- Rao, T. J. 2010. Official statistics in India: the past and the present. *J. Off. Stat.* 26(2): 215–231.
- Sparapani, T. 2017. How big data and tech will improve agriculture, from farm to table. Available at <https://www.forbes.com/sites/timsparapani/2017/03/23/how-big-data-and-tech-will-improve-agriculture-from-farm-to-table/#46d1b63f5989> (accessed date November 10, 2017).
- Srivastava, A. K. 2012. Agricultural statistics system in India. U. C. Sud, H. Chandra, K. Aditya, eds. *E-Book on Techniques of Estimation and Forecasting of Crop Production in India*. Indian Agricultural Statistics Research Institute, New Delhi, 120–126.
- Swaminathan, J. M. 2018. Big data analytics for rapid, impactful, sustained, and efficient (RISE) humanitarian operations. *Prod. Oper. Manag.* 27(9): 1696–1700. <https://doi.org/10.1111/poms.12840>.
- The Economist. 2016. Technology quarterly. The future of agriculture. Available at <http://www.economist.com/technology-quarterly/2016-06-09/factory-fresh> (accessed date December 5, 2017).
- Wolfert, S., L. Ge, C. Verdouw, M. J. Bogaardt. 2017. Big data in smart farming – A review. *Agric. Syst.* 153: 69–80.

Big Data and Service Operations

Maxime C. Cohen*

NYU Stern School of Business, New York City, New York 10012, USA, maxcohen@nyu.edu

This study discusses how the tremendous volume of available data collected by firms has been transforming the service industry. The focus is primarily on services in the following sectors: finance/banking, transportation and hospitality, and online platforms (e.g., subscription services, online advertising, and online dating). We report anecdotal evidence borrowed from various collaborations and discussions with executives and data analysts who work in management consulting or finance, or for technology/startup companies. Our main goals are (i) to present an overview of how big data is shaping the service industry, (ii) to describe several mechanisms used in the service industry that leverage the potential information hidden in big data, and (iii) to point out some of the pitfalls and risks incurred. On one hand, collecting and storing large amounts of data on customers and on past transactions can help firms improve the quality of their services. For example, firms can now customize their services to unprecedented levels of granularity, which enables the firms to offer targeted personalized offers (sometimes, even in real-time). On the other hand, collecting this data may allow some firms to utilize the data against their customers by charging them higher prices. Furthermore, data-driven algorithms may often be biased toward illicit discrimination. The availability of data on sensitive personal information may also attract hackers and gives rise to important cybersecurity concerns (e.g., information leakage, fraud, and identity theft).

Key words: big data; service operations

History: Accepted: December 2017 by Kalyan Singhal.

1. Introduction

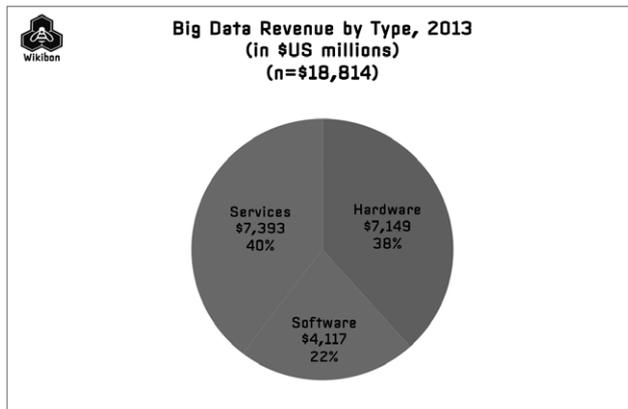
This study presents a broad overview of how the service industry has been affected by the presence of big data (i.e., granular data collected and stored at unprecedented levels of variety, velocity, and volume). Today, most firms collect and store large amounts of data on their customers (e.g., personal information, demographics, social networks, geolocation, past searches, and previous purchases), and keep a record of all their past transactions. Several experts claim that we are only at the beginning of a revolution, and that in the near future every company will base most operational decisions on data. Such a practice is often referred to as *data-driven decision making* or *data-driven algorithms*. It has become surprisingly simple to find impressive statistics on the massive size of this phenomenon. For instance, more data were created within the past 2 years than in the entire previous history of the human race. By 2020, it is predicted that 1.7 megabytes of new information will be created every second for every person on the planet. For example, Walmart handles more than 1 million customer transactions every hour. These data are imported into databases, which are estimated to contain more than 2.5 petabytes of data.¹ A recent study by Wikibon.org reported that big data will be a \$50 billion business by 2017.² According to the same study, the big data market as measured by vendor

revenue derived from sales of related hardware, software, and services reached \$18.6 billion in 2013. Broken down by type, the revenue generated from big data services made it to the first place with 40% of the total market (\$7.39 billion), as can be seen in Figure 1.

The vast majority of companies across different industries are aware of the potential of utilizing big data to increase their bottom line. It is now common to find a data science team in most organizations (e.g., Walmart, Staples, Zara, the New York Yankees, Marriott, American Airlines, Spotify, and Disney, to name a few). Even small startup companies often include one co-founder with expertise in data science and analytics. A 2012 survey indicated that 54% of financial services firms have appointed a chief data officer (Bean 2016). Companies are well aware of the benefits of collecting and storing past data. Interestingly, research by MGI and McKinsey's Business Technology Office revealed that firms are facing a steep shortage of talent capable of managing and analyzing big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with strong analytical skills.

Most firms have invested heavily in collecting and storing data. However, it was reported that less than 0.5% of all collected data has ever been analyzed or used.³ Thus, finding innovative and efficient ways to utilize existing data represents a major challenge for firms to unlock the benefit of big data. In the service

Figure 1 The Big Data Market as Measured by Vendor Revenue in 2013 (source: wikibon.org) [Color figure can be viewed at wileyonlinelibrary.com]



industry, this task is of primary importance, as historical data can help service providers learn their customers' behavior and preferences to enhance service quality and therefore, increase profit and customer satisfaction. Given the competitive landscape of the service industry, it is crucial for firms to know their existing customers (by learning their preferences from the data) in order to offer better customized services. The data on existing customers can also be used to attract similar new customers and increase the firm's market share.

In the digital era of today's economy, almost all types of services have a digital presence and are data oriented. Interestingly, firms interact with their customers via several channels. For example, users can connect and interact with a service company (e.g., an airline, hotel, or restaurant) via their Internet website, smartphone application, phone customer service, chatbot, Twitter account, etc. It is important to aggregate the multiple sources of data and to understand how a typical user utilizes the different channels offered by the firm. The challenge then is to transform the data into strategic levers that help the firm improve the current customer experience, and thus, the firm's bottom line. Big data has widely impacted management practices. According to McAfee et al. (2012): "Data-driven decisions are better decisions—it's as simple as that. Using big data enables managers to decide on the basis of evidence rather than intuition. For that reason it has the potential to revolutionize management." A similar claim also applies to service operations, which should strive to unlock the powerful information hidden in big data.

In this study, we discuss how the presence of big data has transformed services. In particular, section 2 focuses on financial services (credit cards, online payments, peer-to-peer lending, and trading services). Section 3 discusses the transportation and hospitality

sectors, with an emphasis on ride-hailing and marketplaces, as well as traditional firms, such as airlines and hotels. Section 4 considers services delivered by online platforms. In section 5, we present some of the main pitfalls and risks due to the abundance of big data. Finally, our conclusions are reported in section 6.

For each topic and industry, we convey some of the recent changes induced by the presence of large datasets. We discuss in greater detail companies with which the author collaborated through research, consulting, and/or informal discussions. The ideas presented in this study are by no means exhaustive and portray only a partial representation of the actual situation which is evolving every day. Our goal is to make the reader aware of the impact of big data on services, as well as to discuss some of the marketing and operational tools commonly used to improve the decision-making process by exploiting historical data. The focus of this study is less on the technical aspects and more on the high-level picture. Before presenting concrete examples from the aforementioned sectors, we briefly discuss several tools and mechanisms the service industry uses to analyze and exploit big data.

1.1. Tools and Mechanisms

Companies seek to use their data to identify trends, detect patterns, and know better their customers' habits and preferences. As big data holds the potential to describe customers with a high accuracy, many firms use their data to refine their definition of the Ideal Customer Profile (ICP). Furthermore, one can use past digital information to define target audiences (age, location, education, income, interests, etc.). By observing users' habits and online search results, the firm can identify the type of customer who is attracted by certain products. Exploiting big data provides insights into how to lower the Customer Acquisition Cost (CAC), increase the Customer Lifetime Value (CLTV), reduce customer churn, and manage several other customer-driven metrics.

The process of collecting, storing, and analyzing data is based on several steps, and each step can be performed, using specific tools and software. These steps include data collection, data storage and management (e.g., Hadoop and Cloudera), data cleaning (e.g., OpenRefine and DataCleaner), data mining and analysis (e.g., IBM SPSS Modeler and BigML), data visualization (e.g., Tableau), and data integration (e.g., Talend). In this study, we focus on the data mining and analysis step. Next, we describe mechanisms (or methods) commonly used in the service industry that leverage the potential information hidden in big data.

- **Real-time personalization:** Companies aim to send the right offer to the right user at the right time. Big data can be used to effectively

personalize offers, prices, displayed assortments, and the order of online searches. Personalization can be performed dynamically, as well as at the user level. Depending on the context, personalization can utilize customers' attributes, such as geo-localization, demographic information, time of the day, etc. Today, differentiating pricing strategies at the customer-product-time level and optimizing prices using big data have become ubiquitous. Several academic papers were written on this topic (see, e.g., Amatriain 2013, Golrezaei et al. 2014).

- **Targeted promotions and campaigns:** Sending promotions to existing or new customers can be expensive and often results in low conversion rates. Using large historical data sets can help guide the design of future promotion campaigns. For example, firms need to decide which types of customers to target, and what are the most important features (e.g., geo-localization, demographics, or past behavior). Search engine optimization and targeted marketing campaigns (email and mobile offers) are two domains where big data is having a significant impact. Designing targeted promotions has been extensively studied in the academic literature (see, e.g., Andrews et al. 2015, Arora et al. 2008, Fong et al. 2015).
- **Pattern and trend identification:** This process, sometimes also called lead scoring, involves the use of historical data to identify existing trends and predict future outcomes (see, e.g., Thomas et al. 2017 and the references therein). Companies try to identify shopping trends (e.g., by using Google Trends), so they can highlight their top-selling products and eliminate the underperforming ones (Choi and Varian 2012). Other examples include identifying demographic trends, which allows businesses to better cater to specific groups. Advanced statistical and machine learning algorithms are routinely used to search for new patterns and correlations in the data. Subsequently, such patterns are used to inform future operational decisions. One common technique is *association rule learning*, which aims to discover interesting hidden relationships among several attributes in large databases. One application is market basket analysis, in which a retailer can determine which products are frequently purchased together and then use this information for marketing purposes.
- **Customer segmentation:** This consists of clustering customer data to create customized marketing campaigns that cater to several specific groups of users. Examples include recommendation engines

that create value for customers by reducing their search and evaluation costs (e.g., Amazon and Netflix). Firms try to segment users based on several attributes in order to leverage the data on existing users to make decisions for similar new customers (see, e.g., Bobadilla et al. 2013, Zhou et al. 2010).

- **Predictive modeling and analytics:** Predictive analytics is a commonly used statistical technique to algorithmically predict future outcomes based on historical, current, and even social data (Cui et al. 2017). In practice, predictive analytics can be applied to a wide spectrum of disciplines—from predicting the failure of an aircraft engine based on the data stream from sensors to predicting customers' next actions based on their previous purchases and their social media engagement. One popular application is fraud analytics, where the data are used to detect fraudulent activities.
- **Text, audio, and video analytics:** Text analytics (or text mining) refers to techniques that extract information from textual data, obtained from online reviews/forums, emails, survey responses, social network feeds, and call center logs. Text analytics involves statistical analysis, computational linguistics, and machine learning. Audio or speech analytics analyzes and extracts information from unstructured audio data. Currently, call centers are one of the primary application areas of audio analytics. Recently, video analytics has also received a lot of attention.

In the next three sections, we report several concrete examples of services from different sectors that use at least one of the mechanisms above.

2. Financial Services

In this section, we discuss how big data has been shaping some of the financial services. We first comment on the impact of credit cards applications, which is a multi-billion dollar industry. Second, we discuss the impact on online payment platforms that have emerged in recent years. Third, we consider the market of matching borrowers with lenders (peer-to-peer lending), which is a growing industry powered by large datasets. Fourth, we briefly discuss the impact of big data on trading and investment services. More details on operations in financial services can be found in Xu et al. (2017) and Pinedo and Xu (2017).

2.1. Credit Cards

Banks and financial institutions are constantly seeking to improve their credit card departments. Credit card

services yield a comfortable margin and have become a very lucrative business. For example, in 2014, American Express made a net interest income of approximately \$5.8 billion.⁴ From the bank's perspective, credit cards generate revenue by charging a membership fee, a small fee per transaction to merchants (called the interchange fee, which typically ranges between 1% and 6% of the transaction amount), international fees on transactions in a foreign currency, and exchange rate loading fees. In many cases, banks also charge a high penalty fee for late or overdue payments, a fee for exceeding the credit limit (called the over-limit fee), as well as relatively high interest rates on outstanding balances. From the user's perspective, having a credit card is almost a necessity (at least, in the United States). The variety of credit card options has exploded in the last two decades. It is now common for companies to partner with a bank in order to offer their own credit card (e.g., Amazon, American Airlines, Macy's, Costco, Best Buy, Toys "R" Us, and the recent example of the Uber VISA card). The number of credit cards in circulation and the number of active users have also exploded. In 2015, the number of credit cards from the four primary credit card networks (VISA, MasterCard, American Express, and Discover) was 636 million.⁵ In addition, Americans made 33.8 billion credit card transactions in 2015, for a total value of \$3.16 trillion.⁶ As of November 2016, a typical American held an average of 2.35 credit cards.⁷

Picking the right credit card (either for personal or business purposes) can save users thousands of dollars each year on interest, helps book travel for free, as well as earns significant rewards (e.g., receiving cash back, collecting airfare miles, and accessing airport lounges). It is also important for users to use their credit cards wisely, as it affects their credit score, which is crucial for loan applications. With the unprecedented volumes of data on users and past transactions, credit card companies seek to send targeted offers to users to increase their profits. In particular, companies use data-driven models to predict the default probability of users, as well as their propensity to spend. Such companies often use their data in order to guide decisions, such as (i) which individuals to offer a credit card, (ii) the credit limit, (iii) when and by how much to increase the credit limit for existing users, and (iv) which benefits or rewards to offer in order to increase the conversion rate.

Consider the concrete problem faced by a credit card company that needs to decide whether to issue credit cards to a specific set of users. The first goal is to assess their risk of default. Note that this task can be challenging as the historical data is available only for individuals who were given a credit card in the past. Consequently, for users who were never issued

a credit card, it is not possible to know if they would have defaulted. If the population of people applying is "similar" to the population who were issued cards in the past, one can use the past information to infer future decisions. But very often, this is not quite the case. In particular, there are a large number of applications from customers who are completely unbanked, with no credit history (e.g., new immigrants and students who get their first job). The presence of large datasets can help identify similar users (by using clustering techniques) and leverage some detailed information on similar users.

An additional business decision related to credit cards is to decide whether to give a credit limit increase to existing customers. Existing users may request a credit limit increase every several months. Then, the credit card company needs to decide whether or not to approve such a request. On one hand, the company wants to increase the spending power of users, as it may potentially enhance the firm's profits. On the other hand, this can lead to a higher risk of defaulting, and such decisions are subject to strict laws and regulations. More precisely, credit card companies cannot grant infinite credit limit increases as the companies must keep a portion of their capital on hand to cover the credit they are issuing. Consequently, with the credit limit constraint, the firm wants to increase credit limits for individuals who are the most likely to spend, while minimizing the default risk. Solving this optimization problem calibrated with historical data is definitely not a simple task.

In addition to credit card companies, banks face similar data-driven decision-making problems. Consider, for example, a bank that needs to decide which financial product to suggest next to a particular client. Banks always try to persuade customers to acquire new products (e.g., a credit card, a savings account, a mortgage, or a private wealth account). However, advertising a new product can often be expensive. Therefore, several banks put great efforts into carefully selecting whom to advertise to, and which product(s). Similarly, banks need to decide which price to offer for long-term savings accounts. Many individuals have at least one savings account. When the term ends (e.g., every year) and the account rolls over, the user has to decide whether to renew, and this decision depends on the rate offered by the bank. Today, banks often rely on past data to solve this problem. Several related works can be found, see, e.g., Bharath et al. (2009).

Interestingly, some credit card companies sell (anonymous and aggregated) customers' data to other businesses, such as retailers that would like to garner better insights into consumer spending habits. The data can be aggregated by ZIP code, which

informs retailers what areas are more likely to make purchases. Alternatively, credit card companies can sell data to advertisers that can use this information to target specific users with ads. Similarly, companies, such as creditkarma, sell aggregate information to credit card companies and to advertisers. Creditkarma is a personal financial management platform that offers free credit scores (with weekly monitoring), tax preparation, tools for identifying and disputing credit report errors, and tools for simulating credit score evolution. The revenue stream of such (free to consumers) platforms is typically covered by targeted advertisements for financial products (e.g., credit card recommendations based on users' credit profiles).

2.2. Online Payments

Online payment systems can be based either on fiat currencies or on virtual currencies. Traditionally, small payments in the United States, as well as in many other countries, were made through checks. Launched in December 1998, PayPal quickly became the leader in accepting and receiving online payments. It appears to be the de facto online payment solution for online customers, freelancers, and small business owners. For example, many transactions on eBay are performed via PayPal. Similarly, a large number of websites that aim to request donations use PayPal. PayPal's shares increased by nearly 80% between January and October 2017 with \$85.8 billion in market capital (October 26, 2017). In October 2017, PayPal launched a new product called PayPal for Marketplaces. This new modular system is designed for businesses that operate online marketplaces (e.g., ride-sharing and room rental platforms, crowdfunding portals, peer-to-peer e-commerce sites, and online gaming). In recent years, technological advances have opened the door for a number of competitors to challenge PayPal by offering cheaper fees, faster transactions, and enhanced security. Over the last several years, new payment systems proliferated. One can count more than a dozen of alternatives. Examples include Stripe, Due, Apple Pay, Google Wallet, Payer, Square, Alipay, Amazon Pay, Skrill, WePay, and Venmo; a more exhaustive list can be found online.⁸ Such systems may allow users to complete payments via email or by using their smartphones. Consider, for example, the service offered by Venmo (which was acquired by Braintree in 2012, which was itself acquired by PayPal in 2013). Venmo is designed for quick and small payments (after verifying their identity, users can send up to \$2999.99 during each seven-day period). Although the number of users is not publicly reported, the dollar amount in transactions is quite impressive. Venmo handled \$17.6 billion in transactions in 2016, \$6.8 billion in

transactions in Q1 of 2017,⁹ and more than \$8.0 billion in transactions in Q2 of 2017.¹⁰ The main competitive edge of Venmo lies in its social dimension. In particular, a popular use case is when friends conveniently split bills (e.g., for meals, movies, rent, or trips). Venmo allows users to log in to the platform using Facebook, thus providing access to social network data to the provider. When a user completes a payment transaction, the transaction details (the payer, receiver, amount, and specifics of the expense) are shared on the user's "feed," so that other friends can also see it.¹¹ In addition, Venmo encourages social interactions on the platform through likes and comments on each transaction. Consequently, the richness of the data available to a platform like Venmo is striking. The platform has access to the network of friends, to the types of relationships and mutual interests people have (e.g., going to watch a movie on a weekend), and to all the pairwise money transactions. Monetizing this data is a challenge but if done properly, can lead to very high profits.

2.3. Peer-to-Peer Lending

In today's economy, borrowing and lending often occur online, especially when the borrowing party does not have a high enough credit score. Borrowing and lending (peer-to-peer) can take place between two individuals without involving a banking institution. Peer-to-peer lending refers to the practice of lending money to individuals or businesses through online services that match lenders with borrowers. The borrower must be able to provide sufficient information about his or her creditworthiness on the online platform in order for the lender to be able to assess the credit risk. The interest rates can be set by lenders who compete for the lowest rate on the reverse auction model or fixed by the intermediary firm based on an analysis on the borrower's credit.

For many years, private individuals have already been offered the option to secure mortgages on their homes through websites such as LendingTree.com. This platform is an online lending exchange that matches individuals with given credit rating scores (FICO) with established banking institutions that compete for business. Such transactions are made possible thanks to reliable credit scores provided by three credit rating agencies (Equifax, Experian, and TransUnion) that collect extensive financial information on individuals. At times, the borrower may be a small company (e.g., a startup) that would have difficulties to obtain financing from a banking institution. Such a company may also resort to crowdfunding to obtain financing in the form of a loan or an equity stake in the company.

LendingClub is the world's largest peer-to-peer lending platform.¹² More precisely, it is an online

lending platform that offers loan trading on a secondary market and enables borrowers to create unsecured personal loans between \$1000 and \$40,000. The company claims that \$28.75 billion in loans have been originated through its platform up to June 30, 2017.¹³ Each loan displayed on the website includes information about the borrower (e.g., FICO score, credit history, debt-to-income ratio, and personal income), the amount of the loan, the loan's grade, and the loan's purpose. Investors earn money from the interest, whereas the platform charges borrowers an origination fee and investors a service fee. The amount of historical data available to the platform is enormous. This data (which is partially made publicly available) is transforming the lending industry and has incentivized several investment firms to enter this market.

2.4. Investment and Trading Services

Investment and trading services, which often occur through online platforms, are other prominent financial services that have been affected by big data. Users can easily create accounts allowing them to trade securities, stocks, bonds, and exchange-traded funds (ETFs). Such platforms have access to unique datasets. For example, the platform can have access to how often users connect to the platform and which types of financial products users monitor. Subsequently, the platform can sell such aggregate information to advertisers. In addition, the platform can send targeted offers to its users, such as free webinars, referral promotions, and online workshops on different topics. An interesting example is the platform eToro, which is a social trading and multi-asset brokerage company.¹⁴ Users can trade currencies, commodities, indices, and contract for difference (CFD) stocks online. The unique characteristic of eToro's platform is that users can decide to follow investors by replicating the same portfolio investment. The slogan on their website reads as follows: "Join the Social Trading revolution! Connect with other traders, discuss trading strategies, and use our patented CopyTrader technology to automatically copy their trading portfolio performance." Their data is very rich as it includes performance data and risk measures for each user at each point in time. In addition, eToro has access to fascinating data on social connections and influences among the different users. Each user is rated with a risk factor, the gain or loss in percentage during the last 12 months, and the variation in gain or loss during the last trading day. One can also access historical statistics on the performance and the current portfolio (i.e., open trades). Finding effective ways to monetize such data is an interesting challenge, and several companies are working on this type of problem.

3. Transportation and Hospitality

In this section, we discuss some of the recent disruptions in the transportation and hospitality industries that were partially driven by the presence of big data. The world is clearly moving toward personalization in these sectors, implemented through a large-scale decentralized system. Most service providers aim to constantly improve their customer service by collecting large relevant datasets on users. We first consider transportation services (with a focus on ride-hailing platforms), and then consider hospitality services (with a discussion on hotels and online marketplaces, such as Airbnb).

3.1. Transportation

On-demand ride-hailing platforms have changed the way people commute and travel for short distances. Several well-known players in this market are Uber, Lyft, Didi Chuxing, Grab, Ola, Via, Gett, and Juno, to name a few. In October 2016, it was reported that Uber had 40 million monthly riders worldwide.¹⁵ Today, using this type of transportation service has become the norm in most major cities (e.g., Uber now operates in more than 600 cities around the world). During the first few years, growth was moderate, but within the last 2 years, this industry has expanded rapidly. For example, it took Uber 6 years to complete their first billion rides (from 2009 to 2015) but only an additional 6 months to reach their two-billionth ride.¹⁶ This means that during the first 6 months of 2016, the company was providing an average of 5.5 million rides a day (or 230,000 an hour). This type of statistics illustrates the impressive scale and growth of the ride-hailing industry. More importantly, ride-hailing platforms collect a massive amount of granular data at a very large scale. Each transaction (in this case, ride request) comes with a stream of information: rider ID, drop-off/pick-up times and locations, number of passengers, day of the week, price offered, weather conditions, waiting time, type of car requested, and much more. For example, services like Uber allow riders to split the fare with friends. Thus, this provides information on users' social networks. Platforms can collect and store information on geo-localization, willingness to pay, willingness to wait, as well as many other features related to their customers. Using this rich data on every single user remains challenging but has unprecedented potential in terms of increasing the service personalization and the long-term profits. For example, if the platform knows that some users do not mind waiting for rides, whereas others do, this information could be potentially used in the "dispatch algorithm" (i.e., deciding in real-time which drivers are assigned to which ride requests). These platforms always try to

find new ways to exploit the data in order to improve service, retention, and revenue. A recent example is related to geo-localization. When a user requests a ride and does not accept it (e.g., the price was too high), the application can notify the user a few minutes later that the price quote is now lower (while knowing the rider's exact position). The firm can also use the data to decide how to send targeted promotions to its users. By leveraging available fine-grained data, promotion and referral campaigns can now be customized to a very high degree of granularity. Since such platforms often operate in an on-demand supply mode, they also collect a vast amount of data on workers/drivers (vehicle type, sometimes demographics information, geo-localization, work hours, sensitivity to promotions, etc.). As a result, the platforms can refine their algorithms and increase the efficiency of their operations by using this data to create better incentives for both riders and drivers. This topic is a very active research area in the operations management community (see, e.g., Bimpikis et al. 2016, Chen and Hu 2016, Chen and Sheldon 2016, Cohen and Zhang 2017, Hu and Zhou 2017, Tang et al. 2017, Taylor 2017).

Recently, taxi services have dedicated great efforts to modernize their operations to better fit into today's economy. For example, in several cities, taxi rides can now be directly ordered from a smartphone application, and the payment (including the tip) can be completed either via the application or in person. One such company based in the United States is Curb.¹⁷ On their website, one can read: "Curb is the #1 taxi app in the United States that connects you to fast, convenient and safe rides in 65 cities (50,000 Cabs—100,000 Drivers)." Other similar examples include Ace Taxi in Cleveland, Ohio, Talixo in Germany, and taxi.eu which operates in 100 European cities. These companies offer taxi rides by using an online platform, and therefore, can easily collect data on previous transactions. Historical data can help taxi companies improve operational decisions, such as dispatching drivers across the city, predicting demand in real-time, and sending targeted offers. Most optimization and prediction algorithms used by such platforms are data-driven and are tuned very often in order to dynamically capture the high-paced changes observed in the data.

Interestingly, several platforms go beyond just passively collecting data. In particular, several firms design and run a multitude of micro-experiments with the goal of generating high-quality data. A typical platform can decide to run a series of carefully designed experiments (often called *A/B tests*), in order to validate intuition and gain important novel knowledge on users' behavior. For example, are users more likely to react to promotions sent in the morning or in

the evening? To answer such a question, the firm can design a small experiment and randomly send promotions to two samples of users. Then, by testing the statistical significance of the results, the platform can gain important knowledge that will be valuable moving forward. Today, Microsoft and several other leading companies, including Amazon, Booking.com, Facebook, and Google, each conduct more than 10,000 online controlled experiments annually, with many tests engaging millions of users (Kohavi and Thomke 2017). Startups companies and firms without digital roots, such as Walmart, Hertz and Singapore Airlines, also run this type of test regularly, albeit at a smaller scale. For more details on this topic, we refer the reader to the recent article by Kohavi and Thomke (2017) and to the paper by Kohavi et al. (2013).

Airline companies have also been greatly affected by big data. The pricing and scheduling decisions for flights are often controlled by data-driven algorithms. Airline companies collect rich datasets on customer transactions (customers can now easily be identified via frequent flyer numbers). Firms subsequently use this data to customize price offerings and to enhance customer loyalty. The demand prediction for each leg is also performed by using large datasets which include previous performance, weather conditions, as well as many additional factors. Today, airlines make the majority of their operational decisions (scheduling, pricing, inventory, staffing, etc.) based (at least partially) on historical data. From the customer perspective, things have also evolved significantly. The increased level of competition and the explosion of reservation systems (e.g., Kayak, Orbitz, and Expedia) allow consumers to easily compare the different alternatives. Some of these reservation systems even offer advice on whether to book now or to wait for a potential price decrease. These reservation systems have access to very fine-grained data about users: geographical location, IP address, browser used, mobile or desktop platform, past searches, previous reservations, number of clicks, and so on. The systems can sell aggregate information to advertisers and use some of this information to discriminate searches and prices among users (see more details on this topic in section 5.3).

Big data have also affected the transportation industry from a completely different angle. Manufacturers and in particular, large aeronautics companies, such as Boeing and Airbus, now routinely use data obtained from sensors to manage the maintenance of their aircrafts, and to create after-sales personalized services (e.g., proactive maintenance and special monitoring). More precisely, they place hundreds (if not thousands) of different sensors to collect information in a very fine-grained fashion. Those sensors are very sophisticated and can often record

measurements as fast as every second. They are located in different parts of the aircraft and typically measure the temperature, the humidity level, the vibrations, as well as various physical and mechanical properties. The data collected from these sensors gives rise to very large time series (such datasets are often hard to store and to visualize). Subsequently, the firm's goal is to analyze these datasets in order to improve the current maintenance strategy. The ultimate goal is to send specific aircraft parts for maintenance before a critical issue occurs, but at the same time not too early. This practice involves very high costs and risks, and thus, the potential impact of properly analyzing such data is tremendous. Firms that offer this type of service to airlines have a unique competitive advantage. A very similar story (albeit, at a smaller scale) is present in the automobile industry, which has also started to use a large number of data sensors.

3.2. Hospitality

Today, hotels collect and store as much information as possible on their customers. Nevertheless, it can be challenging to efficiently exploit the data as individuals often use different email addresses (e.g., business versus personal). One common technique to ease the data collection process is to offer a loyalty card to customers. Very often, the loyalty program can be shared among several hotel partners (e.g., the Starwood Preferred Guest network includes Sheraton, Le Meridien, Westin, W Hotels, and several other hotel chains). This allows the firm to identify each user by a unique identifier. After each stay, hotels can record the special requests and preferences (e.g., vegetarian or booking room dining services). During the next stay, the hotel can then accommodate the customer's needs in a more effective and personalized fashion.

Most hotels also use historical data to design and send targeted promotional offers. For example, hotels can collect data on which types of customers are likely to accept which type of promotional offer, and at what time of the year. Then, by leveraging the data from past offers, the hotels can decide the specifics of the next campaign. For example, hotels in the Intercontinental Hotels Group use price optimization tools (see Koushik et al. 2012). Another example is the Carlson Rezidor Hotel Group that uses data-driven methods to maximize revenue (see Pekgün et al. 2013). This type of practice is particularly relevant to hotels in Las Vegas (or other casino resorts), which use data-driven predictive algorithms to infer the spending capital of each potential customer. Such practices are often taken very seriously as they can drive a relatively large portion of the hotel's revenue. This topic has been extensively studied in the academic literature (see, e.g., Kilby et al. 2005, Lucas and

Brewer 2001). In addition, a large number of patents were issued on this topic during the last two decades. Having access to larger datasets can only make this lucrative practice more exciting. It is now common for hotel groups to hire a team of analysts who are dedicated to improving their data-driven algorithms. In addition, with the explosion of online travel search websites (e.g., Kayak, Orbitz, and Expedia), hotels need to decide at each point in time the portion of reservations to assign to those channels, as well as the pricing policy. Managing operational decisions for cruises has also been affected by the presence of large datasets (see, e.g., Gibson 2006). Today, several cruise companies have a department fully dedicated to developing and improving the management and use of their data.

It does not seem reasonable to end this section without mentioning online marketplaces such as Airbnb, Homeaway, Homestay, VRBO, Vacasa, and Flipkey, to name a few. These online platforms allow people to lease or rent short-term lodging, including vacation and apartment rentals. These companies typically do not own any lodging but receive a percentage service fee (commission) for every booking from guests and hosts. As of October 2017, Airbnb has more than 3 million listings in 65,000 cities and 191 countries.¹⁸ The price is typically set by the host and can depend on the time of the year, the day of the week, the number of similar listings, the amenities available, the number of nights, etc. The amount of data gathered by such platforms is impressive, as they can collect data on both the users and the properties. This data allows the platform to deliver a better service by recommending prices to the host, and improving the ranking of the different options for each browsing user (based on previous preferences). It also leads to more transparency for the industry (users can write reviews of good and bad experiences so that other users' knowledge increases considerably before booking). Such platforms often hire senior data analysts who are constantly working on exploiting historical data to improve future tactical and operational decisions.

4. Online Platforms

In this section, we focus on services which are offered via online platforms. These services are also greatly affected by the presence of big data. Today, many firms interact with their customers via online platforms. This is true for transportation services (ride-hailing), as discussed in section 3.1. Other examples include health and medical services (e.g., Zocdoc), dating services (e.g., match.com), recruiting services (e.g., CornerJob), restaurant reservations (e.g., OpenTable), food delivery services (e.g., Grubhub and

Slice), delivery and home services (e.g., TaskRabbit and Handy), and subscription services (e.g., Stitch Fix and Blue Apron). Services delivered by online platforms have transformed a big part of the service industry. These companies can collect vast amounts of fine-grained data on customers' habits and preferences with the goal of improving service customization. The ultimate objective is to offer the optimal product for each customer, and vary the prices dynamically to maximize long-term revenue and retention. Learning user preferences (for food, clothing style, dating affinities, etc.) can be performed by analyzing survey data, questionnaires, online reviews, data from social networks (friends, pictures, and interests) and matching or clustering users with other carefully selected similar users. This can be accomplished only by leveraging and analyzing the large amounts of historical data. For example, to use the service offered by Stitch Fix, customers fill out a survey about their style, pay a \$20 up-front styling fee, and then receive five clothing items tailored to their taste. Stitch Fix runs relatively large data science operations that leverage the data from in-depth surveys to increase the accuracy in styling choices for customers. This type of firm owns a valuable growing dataset of detailed customer preferences and product trends and is constantly working on refining its data-driven algorithms. As mentioned before, such datasets can be sold to online advertisers. Online services such as music (e.g., Spotify) and movies (e.g., Netflix) also dedicate significant efforts to collecting and exploiting large amounts of data. One of the basic challenges is to accurately learn users' preferences from past usage in order to provide effective recommendations. These companies are actively hiring top data scientists to develop efficient algorithms that operate in real-time at a very large scale.

Apart from these relatively new examples, one can find similar data-driven practices by companies like Amazon, Facebook, YouTube and many others. Online retailers (e.g., Amazon) know when a user recently visited their website and browsed a specific product. Then, such a user may have a high valuation for being shown an ad for this specific product. This common practice is known as remarketing or retargeting, and can generate significant revenue for retailers and for advertisers. Facebook has access to endless data on its users, and strives to exploit this data to optimize advertising content. In particular, Facebook can track users by using cookies which allow fine-grained targeting for online advertising.

The media and entertainment industry has also been greatly impacted by big data and analytics. Entertainment theme parks (e.g., Disney) use state-of-the-art machine learning techniques to improve their user experience and to increase the profits generated

by their parks and by their consumers' derived products (e.g., costumes, mugs, hats, and stuffed toys). In most attraction parks, users can download a smartphone application that allows them to navigate through the park. This application provides the firm access to users' geographic locations in real-time, and thus, allows Disney to better estimate wait times for the different park attractions. This also allows the park restaurants to send targeted offers to specific users, depending on their geographic locations and other attributes.

Traditional restaurants try to use previous data on consumers to improve their service quality and the customer experience. Restaurants often track their customers by requesting phone numbers during reservations. They then record dietary restrictions and preferences in order to customize the service for future visits. At a higher level, restaurants collect data on previous dishes (costs, statistics about people ordering and reordering, online reviews, etc.) and on prices with the goal to constantly improve their offerings and their profits. It even seems that some restaurant owners practice *A/B* tests by varying the menus and prices in order to learn customers' preferences.

Finally, we conclude this section by discussing a recent tool for customer service. Several decades ago, many providers (e.g., banks, telecom companies, and hospitals) opened call centers with the goal of addressing customer complaints and concerns (see, e.g., Aksin et al. 2007, Koçağa et al. 2015). The data obtained from call centers is very large and is often used to learn the problems customers most frequently encounter, and how the company can efficiently address those concerns in a timely manner. Recently, artificial intelligence brought *chatbots* to replace or complement these services (e.g., ReplyYes, Interactbot, and Twyla). A chatbot is a computer program which conducts a conversation via auditory or textual methods to address customers' concerns by appropriately querying the relevant databases (chatbots are often used as part of instant messaging platforms making the interaction with customers convenient). Companies such as Domino's, Pizza Hut, Disney, KLM, Starbucks, Sephora, Staples, Whole Foods, and many others use chatbots to increase customer engagement, promote their products and services, and provide their customers a convenient way to communicate and/or to place orders. Some chatbots use sophisticated natural language processing tools in order to provide a higher quality of service by efficiently scanning databases for matching keywords and word patterns. The presence of large accessible data on previous customers' transactions and complaints has allowed firms in this sector to build enormous datasets. These data sets are routinely used by programs based on artificial intelligence to improve customer service, as well as for information

acquisition. For example, a firm can directly ask the chatbot questions about its suppliers, its pending orders, and other matters that can be answered directly from the data.

5. Pitfalls and Risks

The big data revolution raises a number of ethical issues related to privacy, confidentiality, transparency, and identity. Big data brings the natural requirement for issuing laws and regulations on the ways firms can use data, as well as the potential development of big data ethics (Richards and King 2014). This will help protect values such as privacy, confidentiality, and identity theft, as well as avoid illegal discrimination. In this section, we discuss some of the main pitfalls incurred by big data. First, we report examples of recent data leakages that had serious consequences. Second, we highlight several practical challenges related to data accessibility and aggregation. Finally, we consider a serious problem coined *machine bias* that corresponds to the phenomenon in which data-driven algorithms often lead to unfair decisions based on illicit discrimination (e.g., race or gender).

5.1. Data Leakage and Identity Theft

It was reported that 1093 data breaches took place in 2016.¹⁹ One can also count several dozens of large substantial data breaches in the United States in 2017.²⁰ Two notable examples are Deloitte and Equifax. Deloitte was the victim of a cybersecurity attack in 2017 that went unnoticed for several months. The hackers compromised the firm's global email server

through an administrator's account that gave them privileged, unrestricted access. Emails to and from Deloitte's 244,000 staff were stored (some emails had attachments with sensitive security and design details). In addition, the hackers had potential access to usernames, passwords, Internet Protocol (IP) addresses, architectural diagrams for businesses, and health information.

Equifax is one of the three main credit monitoring agencies in the United States that provides credit reports. In September 2017, Equifax announced that the personal data of 143 million US customers had been accessed or stolen in a massive hack. The breach is thought to have revealed the names, Social Security numbers, dates of birth, addresses, and driver's numbers of almost half the US population (44%). Also compromised were the credit card numbers of 209,000 consumers and the personal identifying information of 182,000 users. In addition, the company admitted in October 2017 that the data of some 694,000 British customers was also compromised, some of whom had their financial information and passwords stolen, including partial credit card information. The company's share price plummeted 35% the week after the breach was disclosed (see Figure 2).

This type of data breach can allow hackers to apply for credit cards and loans by stealing the identity of the hacked users. The total losses from identity theft in 2014 amounted to \$15.4 billion, with an out-of-pocket loss average of \$2895 for the victims.²¹ A survey by Gallup News reported that 27% of US adults claim that they were affected by stolen credit card information between October 2015 and October 2016 (up from 22% between October 2014 and October

Figure 2 The Equifax Inc. Stock Price on the NYSE between August 4, 2017, and October 6, 2017 (*source: investopedia.com*) [Color figure can be viewed at wileyonlinelibrary.com]



2015). This increase is partially driven by the presence of larger datasets.

Two additional recent massive hacks were Yahoo and Ashley Madison. Yahoo disclosed in 2017 that all of its 3 billion email users were likely compromised in a 2013 breach, breaking a potential record for the largest ever potential data breach. Yahoo took action and invalidated unencrypted security questions and answers so they could not be used to access an account. Ashley Madison is a Canadian online dating service specializing in extramarital affairs, marketed to people who are married or in relationships. In July 2015, a group of hackers stole user data, by copying personal information, and threatened to release users' names and personal information if the website did not immediately shut down. In August 2015, the hackers leaked more than 25 gigabytes of company data, including user details (containing several thousand corporate emails). This breach received extensive media coverage and had heavy consequences for families, with a few unconfirmed suicides linked to the data breach.

Cybersecurity and privacy are evidently important issues in financial services. Several major operational risk events have happened over the years. A significant breach occurred at JPMorgan Chase in 2014 when information was stolen regarding 83 million accounts.²² The hackers gained access to the names, addresses, phone numbers, and email addresses of account holders. The bank did not reveal the total cost of the data breach but the bank announced it would spend \$250 million a year to improve its cybersecurity. Another event involved the bitcoin exchange Mt. Gox which experienced several security breaches between 2010 and 2014 resulting in losses of approximately \$500 million, forcing it to shut down in 2014. Other types of operational risk events that financial institutions have to deal with involve crimes in the form of insider trading, rogue trading (i.e., an employee authorized to make trades on behalf of an employer who makes unauthorized trades), and money laundering. Banks and the US Securities and Exchange Commission (SEC) have put several anomaly detection mechanisms in place to detect such events. These mechanisms are typically based on machine learning techniques, such as Neural Networks, Bayesian Belief Networks, and Support Vector Machines. The presence of big data and online transactions clearly accentuates the risks of rogue trading and money laundering.

Many companies that deal with digital information have several data and cyber analysts who are responsible for detecting fraud (e.g., fraud accounts, advertising fraud, and payment fraud). For instance, it is common to develop data-driven models for cleaning up illicit content from websites (e.g., reviews),

primarily based on text and language processing techniques. A second example is models that are based on a set of rules that classify fraudulent and legitimate massive registrations to detect fake accounts. In such models, it is very important to avoid false positives as much as possible. Several startup companies (e.g., Shift Technology) work on developing methods based on artificial intelligence to detect patterns and flag fraudulent insurance claims. This type of algorithm can be trained on hundreds of millions of past insurance claims. Finally, online platforms need to also deal with detecting fraudulent ads (such models are often based on building a dictionary of past fraudulent ads and detecting similarities). It has also become the norm to use multi-factor authentication in order to securely log into many online services. Payment companies such as PayPal invest significant efforts in reducing fraud by developing state-of-the-art algorithms. Such algorithms can be tuned on very large datasets and are typically predictive models that can vary depending on various critical user features. For example, the most valued customers (e.g., monthly users who are high spenders) may face an easier verification process. Such problems are challenging due to the scale of the data and the dynamic aspect.

In the last few years, the field of fraud analytics has exploded. Using data analytics to handle fraud allows organizations to keep control over every transaction. Fraud analytics also identifies hidden patterns and trends to control the exposure of sensitive data across the organization. In addition, it can constantly scan internal systems by trying to detect anomalies in several targeted statistics (e.g., by calculating statistical parameters to find out if or when values exceed averages of standard deviations or by classifying outliers) and learn from previous breaches. A typical method is to use predictive models to compute fraud propensity scores for each transaction in real-time. Then, the firm needs to set the threshold values that dictate the boundary detection for different types of anomalies. Adapting the threshold values dynamically for different users is crucial in order to avoid errors. For more details on data-driven techniques to handle fraud analytics, see, e.g., Bolton and Hand (2002), Delamaire et al. (2009), and Bhattacharyya et al. (2011).

5.2. Data Accessibility, Aggregation, and Acquisitions

Banks and financial institutions have data on millions of customers (and up to hundreds of millions for the largest banks). As a result, the data often needs to be stored in a distributed fashion. Today, most sophisticated banks know how to collect all the relevant data and store it securely. However, one of the main challenges is to make this data easily available to all the relevant users within the institution. Note that these

users may range from advanced individuals who can deal with systems and software, such as Hadoop Distributed File System (HDFS), Spark, and Map Reduce, directly to others who have limited coding skills. Many data scientists currently work on such projects with the goal of making the existing datasets accessible to the largest possible number of employees within the institution. This, of course, should be done while being aware of access controls (i.e., sensitive data should be accessible only to a very few users), as well as security concerns. This raises a trade-off between accessibility and security. Several factors can impede having the data easily accessible to a large number of users. One such factor is acquisitions. If an institution acquires another institution (e.g., Bank of America acquired Merrill Lynch in 2008, and Capital One acquired ING Direct USA in 2011), unless a serious effort is invested in integrating the data properly, it will be forever difficult to treat all customers in the same fashion. In addition, there are several regulatory concerns that banks should cope with. One example is the European Union (EU) General Data Protection Regulation (GDPR),²³ which is arguably the most important change in data privacy regulation in the past 20 years. This regulation, which is set to take effect in 2018, dictates, for instance, that algorithmic decisions related to users must be easily explained (the so-called right for explanation) at a later stage.

An additional challenge is to develop efficient ways to use all the data that comes from different channels. Today, customers interact with firms using different modes: offline when customers go to brick-and-mortar locations, online (either via the Internet website or the smartphone application), and through social networks (e.g., many customers report issues with service quality using Twitter). Combining all these interactions and merging the different observations generated by the same user is crucial. Several companies are putting a great effort into this data aggregation endeavor, as ultimately this will give rise to richer and more valuable data. A similar challenge is the one of using private datasets from the firm together with publicly available data (e.g., the NYC Taxi and Limousine Commission database or data from Google Trend). Leveraging the strength of both types of data can allow a better understanding of the market and consumers.

5.3. Machine Bias and Discrimination

In 2017, ProPublica, an American nonprofit organization that produces investigative journalism, launched the Documenting Hate project to systematically track hate crimes and bias incidents. It has also used machine learning and natural language processing techniques to monitor and collect news stories about hate crimes and bias incidents. Some of the findings

provide evidence for a *machine bias*. As ProPublica puts it: “There’s software used across the country to predict future criminals. And it’s biased against blacks.”²⁴ The organization provided statistical analyses to support this claim and affirmed that basing such decisions on data can be problematic. Recidivism in crimes is only one of many examples. Similar issues arise when insurance companies using data analytics appear to have a bias toward minority neighborhoods that often pay higher car insurance premiums relative to other areas with the same risk. Other typical examples include credit scoring and loan assessment, as decisions in these areas may have ethical and/or legal implications. Consequently, this type of issue raises a flag about being careful when using past data to generalize future insights. The data and algorithms can be biased, and this is not acceptable. This important topic is highlighted in O’Neil (2017) recent book *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. It is shown that in many cases, decision making via mathematical models driven by big data can reinforce discrimination and lead to unfair outcomes. Many researchers in statistics and computer science are working on this problem by trying to ensure the fairness of data-driven algorithms (see, e.g., Calders and Verwer 2010, Hardt et al. 2016). They also address and quantify the bias introduced by data collection, and the issue induced by the fact that many predictor variables can be correlated with the outcome (which we are trying to predict) and some protected attributes (e.g., race or gender). This type of correlation is often referred to as *spurious correlations*, and should be carefully controlled for.

At the end of 2016, an investigation by ProPublica revealed that Facebook not only allows advertisers to target users by specific attributes (such as age, interests, and likes), but Facebook may also let advertisers eliminate users based on race.²⁵ Facebook released an official statement to defend itself and claims to strictly avoid such practices. As we discussed, the era of big data allows advertisers to target users to an exceptional degree of specificity. However, it is important to train those algorithms to understand which types of attributes are acceptable and which are not. Embedding such knowledge in data-driven algorithms is a clear necessity. The features should be tested for discrimination, fairness, and additional desired requirements depending on the context.

Apart from bias and discrimination, the presence of big data together with digitization allow firms to quickly react. For example, prices on online platforms can vary several times a day and can differ for different users. It is not surprising to observe that two users in the same city are offered different price points for

the same product via the same website at the same time. Firms use past data on users' behavior in order to refine their pricing strategies. This gives rise to new issues where users can receive a higher price depending on how often they look at the website, their past searches, the day of the week, the device they are using (mobile versus computer), whether they are using an ad blocker, their geo-localization, etc. Consequently, firms can significantly improve their prediction and profits. At the same time, users can be hurt as they will often get charged a higher price, and this can raise fairness issues. In other words, the fine-grained personalization induced by big data can be perceived as a disadvantage for buyers. A team of researchers at Northeastern University examined 16 popular e-commerce sites (10 general retailers and six hotel and car rental sites) to measure two specific forms of personalization: (i) price discrimination, in which a product's price is user customized, and (ii) price steering, in which the order of the search results is user customized. They found evidence of personalization on four retailer sites and five travel sites, including cases where sites altered prices by hundreds of dollars for the same product. Overall, travel sites showed price inconsistencies in a higher percentage of cases, relative to the control samples (Hannak et al. 2013). It was also claimed that websites such as Expedia and Hotels.com steered a subset of users toward more expensive hotels. It is worth mentioning that in some cases, using big data can actually reduce discrimination. The recent work in Cui et al. (2016) provides evidence for discrimination by hosts against guests of certain races in the marketplace Airbnb. The authors also showed that a review posted on a guest's page significantly reduces discrimination, suggesting that sharing-economy platforms can alleviate discrimination by providing more information and incentivizing peer reviews.

As mentioned in section 3.1, there is a growing recent trend to design experiments that can produce valuable data. Carefully controlled experiments not only attempt to depict the shape of the demand-price curve but also track how this curve changes hour to hour. For example, in some contexts, online purchases may peak during weekday office hours; therefore, retailers are commonly advised to raise prices in the morning and lower them in the early evening. The different deals can vary according to the location, the browsing history, and even the operating system used by the potential buyer. A well-known example is Orbitz which has supposedly targeted Mac users with more expensive search results. Those findings raise the following question: Could the Internet, whose transparency was supposed to empower consumers, be doing the opposite? To alleviate the negative effects of these practices, several tools have emerged

to help customers track price changes and detect the best available offers. Examples of such tools are camelcamelcamel.com (a free Amazon price tracker), honey (a free deal-finding browser add-on), and InvisibleHand (a free automatic price-tracker). Those tools offer price history charts and price drop alerts and may also allow users to search for coupon codes whenever they check out online. Such companies generate revenue by earning commissions when users find a sale or a coupon. In summary, the presence of big data allows firms to better price discriminate customers. On one hand, big data can generate higher profits for firms that efficiently exploit historical data. On the other hand, big data can be perceived as unfair by some customers and thus reduce the market share of businesses that use these methods. Finding the right trade-off between these two conflicting effects can be quite challenging.

6. Conclusion

In this study, we discussed how the large amounts of data collected by firms have transformed the service industry. We focused our discussion on services in the following sectors: finance/banking, transportation and hospitality, and online platforms. We presented an overview of how big data has shaped the service industry, discussed several mechanisms that leverage the potential information hidden in big data, and pointed out some of the pitfalls and risks incurred. We conveyed that firms can now collect unprecedented levels of granular data on customers and on transactions. Firms are also developing quantitative data-driven tools to improve operational decisions, such as prices and quality of service. It is clear that having access to large amounts of data can help enhance the service quality by tailoring the offerings to the users' needs.

Combining the power of big data analytics with high-speed computing (which is becoming affordable) allows for real-time service personalization at a very large scale (e.g., online recommendation systems for movies). However, this personalization benefit seems to come at a price. Firms that have access to this rich data can utilize it to price discriminate against customers. In addition, data-driven algorithms can include a machine bias that accentuates illicit discrimination. This raises several legal issues which need to be carefully addressed by governments. Furthermore, the availability of data on sensitive personal information attracts hackers. The number of breaches has increased and is now a major concern for most firms.

Interestingly, there is growing interest in cross-disciplinary services, where many companies try to exploit the interactions between different types of services. For example, Amazon operates in multiple

spaces (retail, cloud computing, media streaming, and food delivery services). Airbnb is entering the dining reservation market, and IKEA acquired TaskRabbit, among many other examples.

It seems that having access to big data on different types of services can allow firms to exploit the multi-dimensionality of their users' interactions in order to reach a more comprehensive picture and to enhance the service quality, as well as the long-term profits.

In summary, it is clear that big data has been transforming the way firms interact with customers in the service industry. It is also clear that this transformation is only in its infancy. What is less clear is the extent of the long-term impact of such a disruption. Although big data certainly brings several advantages, some drawbacks are in order. One of the major challenges for firms is to carefully exploit and unlock the power of big data while preserving fairness, trust, and consumers' happiness. Identifying the fine line involved in this trade-off seems to be subtle and may require data scientists, marketers, psychologists, lawyers and regulators to work together.

Acknowledgment

The author thanks Tarek Abdallah, Daniel Guetta, Ilan Lobel, Mike Pinedo, Olivier Rubel, and Luis Voloch for their valuable feedback, which has helped improve this paper.

Notes

¹https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/big-data-meets-big-data-analytics-105777.pdf.

²http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017.

³<https://www.technologyreview.com/s/514346/the-data-made-me-do-it/>.

⁴https://materials.proxyvote.com/Approved/025816/20150313/AR_239749/HTML2/american_express-ar2014_0022.htm.

⁵Sources: visa.com, mastercard.com, americanexpress.com, discover.com, 2015.

⁶<https://www.federalreserve.gov/newsevents/press/other/2016-payments-study-20161222.pdf>.

⁷<http://www.experian.com/live-credit-smart/state-of-credit-2016.html>.

⁸https://en.wikipedia.org/wiki/List_of_online_payment_service_providers.

⁹<http://www.businessinsider.com/venmos-monetization-will-be-worth-watching-2017-1>.

¹⁰<https://www.recode.net/2017/7/26/16044528/venmo-8-billion-transaction-volume-growth-rate-chart>.

¹¹Users can decide to opt for a private mode, where not all the details of the transactions are revealed. However, it was reported that many users keep the default public setting, as they do not bother change the privacy settings.

¹²<https://www.economist.com/blogs/schumpeter/2013/01/lending-club>.

¹³<https://www.lendingclub.com/info/statistics.action>.

¹⁴<https://www.eto.com/>.

¹⁵<http://fortune.com/2016/10/20/uber-app-riders/>.

¹⁶<https://techcrunch.com/2016/07/18/uber-has-completed-2-billion-rides/>.

¹⁷<https://gocurb.com/>.

¹⁸<https://www.airbnb.com/about/about-us>.

¹⁹<http://www.idtheftcenter.org/2016databreaches.html>.

²⁰<https://www.techworld.com/security/uks-most-important-data-breaches-3604586/>.

²¹<https://www.bjs.gov/content/pub/pdf/vit14.pdf>.

²²<https://www.reuters.com/article/us-jpmorgan-cybersecurity/jpmorgan-hack-exposed-data-of-83-million-among-biggest-breaches-in-history-idU.S.KCNOHR23T20141003>.

²³<http://www.eugdpr.org/>.

²⁴<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

²⁵This alarming issue was the topic of extensive media coverage, see, e.g., <http://fortune.com/2016/10/28/facebook-ad-propublica-race/>.

References

- Aksin, Z., M. Armony, V. Mehrotra. 2007. The modern call center: A multi-disciplinary perspective on operations management research. *Prod. Oper. Manag.* 16(6): 665–688.
- Amatriain, X. 2013. Big & personal: Data and models behind netflix recommendations. Proceedings of the 2nd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, ACM, New York, pp. 1–6.
- Andrews, M., X. Luo, Z. Fang, A. Ghose. 2015. Mobile ad effectiveness: Hyper-contextual targeting with crowdedness. *Market. Sci.* 35(2): 218–233.
- Arora, N., X. Dreze, A. Ghose, J. D. Hess, R. Iyengar, B. Jing, Y. Joshi, V. Kumar, N. Lurie, S. Neslin, S. Sajeesh, M. Su, N. Syam, J. Thomas, Z. J. Zhang. 2008. Putting one-to-one marketing to work: Personalization, customization, and choice. *Market. Lett.* 19(3–4): 305.
- Bean, R. 2016. Just using big data isn't enough anymore (online). *Harvard Business Review*. Available at <https://hbr.org/2016/02/just-using-big-data-isnt-enough-anymore> (accessed date July 22, 2017).
- Bharath, S. T., S. Dahiya, A. Saunders, A. Srinivasan. 2009. Lending relationships and loan contract terms. *Rev. Financ. Stud.* 24(4): 1141–1203.
- Bhattacharyya, S., S. Jha, K. Tharakunnel, J. C. Westland. 2011. Data mining for credit card fraud: A comparative study. *Decis. Support Syst.* 50(3): 602–613.
- Bimpikis, K., O. Candogan, S. Daniela. 2016. Spatial pricing in ride-sharing networks. Working paper.
- Bobadilla, J., F. Ortega, A. Hernando, A. Gutiérrez. 2013. Recommender systems survey. *Knowl.-Based Syst.* 46: 109–132.
- Bolton, R. J., D. J. Hand. 2002. Statistical fraud detection: A review. *Stat. Sci.* 17: 235–249.
- Calders, T., S. Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Min. Knowl. Disc.* 21(2): 277–292.
- Chen, Y., M. Hu. 2016. Pricing and matching with forward-looking buyers and sellers. Working paper.
- Chen, M. K., M. Sheldon. 2016. Dynamic pricing in a labor market: Surge pricing and flexible work on the uber platform.

- Proceedings of the 2016 ACM Conference on Economics and Computation, ACM, pp. 455.
- Choi, H., H. Varian. 2012. Predicting the present with Google trends. *Econ. Record* 88(s1): 2–9.
- Cohen, M. C., R. P. Zhang. 2017. Coopetition and profit sharing for ride-sharing platforms. Working paper.
- Cui, R., J. Li, D. J. Zhang. 2016. Discrimination with incomplete information in the sharing economy: Field evidence from airbnb. Working paper.
- Cui, R., S. Gallino, A. Moreno, D. J. Zhang. 2017. The operational value of social media information. *Prod. Oper. Manag.* (forthcoming). <https://doi.org/10.1111/poms.12707>.
- Delamaire, L., H. Abdou, J. Pointon. 2009. Credit card fraud and detection techniques: A review. *Bank Bank Syst.* 4(2): 57–68.
- Fong, N. M., Z. Fang, X. Luo. 2015. Geo-conquesting: Competitive locational targeting of mobile promotions. *J. Mark. Res.* 52(5): 726–735.
- Gibson, P. 2006. *Cruise Operations Management*, Elsevier, Oxford, UK.
- Golrezaei, N., H. Nazerzadeh, P. Rusmevichientong. 2014. Real-time optimization of personalized assortments. *Management Sci.* 60(6): 1532–1551.
- Hannak, A., P. Sapiezynski, A. Molavi Kakhki, B. Krishnamurthy, D. Lazer, A. Mislove, C. Wilson. 2013. Measuring personalization of web search. Proceedings of the 22nd International Conference on World Wide Web, ACM, New York, pp. 527–538.
- Hardt, M., E. Price, N. Srebro. 2016. Equality of opportunity in supervised learning. D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, R. Garnett, eds. *Advances in Neural Information Processing Systems Conference*. pp. 3315–3323.
- Hu, M., Y. Zhou. 2017. Price, wage and fixed commission in on-demand matching. Working paper.
- Kilby, J., J. Fox, A. F. Lucas. 2005. *Casino Operations Management*. John Wiley & Sons, New York, NY.
- Koçağa, Y. L., M. Armony, A. R. Ward. 2015. Staffing call centers with uncertain arrival rates and co-sourcing. *Prod. Oper. Manag.* 24(7): 1101–1117.
- Kohavi, R., S. Thomke. 2017. The surprising power of online experiments. *Harv. Bus. Rev.* 95(5): 74–81.
- Kohavi, R., A. Deng, B. Frasca, T. Walker, Y. Xu, N. Pohlmann. 2013. Online controlled experiments at large scale. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, pp. 1168–1176.
- Koushik, D., J. A. Higbie, C. Eister. 2012. Retail price optimization at intercontinental hotels group. *Interfaces* 42(1): 45–57.
- Lucas, A. F., K. P. Brewer. 2001. Managing the slot operations of a hotel casino in the las vegas locals' market. *J. Hospital. Tourism Res.* 25(3): 289–301.
- McAfee, A., E. Brynjolfsson, T. H. Davenport. 2012. Big data: the management revolution. *Harv. Bus. Rev.* 90(10): 60–68.
- O'Neil, C. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books, New York.
- Pekgün, P., R. P. Menich, S. Acharya, P. G. Finch, F. Deschamps, K. Mallery, J. V. Sistine, K. Christianson, J. Fuller. 2013. Carlson rezidor hotel group maximizes revenue through improved demand management and price optimization. *Interfaces* 43(1): 21–36.
- Pinedo, M., Y. Xu. 2017. Operations in financial services: Processes, technologies, and risks. *Foundations Trends Technol. Inf. Oper. Manag.* 11(3): 223–242. <http://dx.doi.org/10.1561/02000000048>.
- Richards, N. M., J. H. King. 2014. Big data ethics. Working paper.
- Tang, C. S., J. Bai, K. C. So, X. M. Chen, H. Wang. 2017. Coordinating supply and demand on an on-demand platform: Price, wage, and payout ratio. Working paper.
- Taylor, T. 2017. On-demand service platforms. Working paper.
- Thomas, L., J. Crook, D. Edelman. 2017. *Credit Scoring and Its Applications*. SIAM, Philadelphia.
- Xu, Y., M. Pinedo, M. Xue. 2017. Operational risk in financial services: A review and new research opportunities. *Prod. Oper. Manag.* 26(3): 426–445.
- Zhou, R., S. Khemmarat, L. Gao. 2010. The impact of youtube recommendation system on video views. *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, ACM, New York, pp. 404–410.

Emergence of Big Data Research in Operations Management, Information Systems, and Healthcare: Past Contributions and Future Roadmap

Samayita Guha*, Subodha Kumar

Fox School of Business, Temple University, Philadelphia, Pennsylvania 19122, USA, samayita.guha@temple.edu, subodha@temple.edu

In this day, in the age of big data, consumers leave an easily traceable digital footprint whenever they visit a website online. Firms are interested in capturing the digital footprints of their consumers to *understand and predict* consumer behavior. This study deals with how big data analytics has been used in the domains of information systems, operations management, and healthcare. We also discuss the future potential of big data applications in these domains (especially in the areas of cloud computing, Internet of Things and smart city, predictive manufacturing and 3-D printing, and smart healthcare) and the associated challenges. In this study, we present a framework for applications of big data in these domains with the goal of providing some interesting directions for future research.

Key words: big data; information systems; operations management; healthcare

History: Received: October 2017; Accepted: November 2017 by Ram Ganeshan and Nadia Sanders, after 1 revision.

Data is the new science. Big Data holds the answers
 —Pat Gelsinger (Gelsinger 2012)

1. Introduction

This is an era where we are generating data at an exponential rate. Large quantities of data representing our digital footprint are generated whenever we interact over social media and chat applications, use online shopping portals, or even when we use such ubiquitous applications as Google Search or Google Maps (Marr 2017a). Aside from data generated by us as users, an enormous amount of data comes from “smart” devices, that is, devices with sensors that collect data from the physical world and convert them into a digital form (Hashem et al. 2016, Riggins and Wamba 2015). This ever-growing stream of data generation is made possible by the advancements in computing and mobile technology and the increasing accessibility of the Internet. For example, according to a report by the United States Census Bureau, in 2015, 78% of U.S. households had a desktop or laptop, 75% had a handheld computer such as a smartphone, and 77% had a broadband Internet connection (Ryan and Lewis 2017). All of these devices, when connected to the Internet, have the ability to generate data in large quantities for those who know how to aggregate it.

It is these data—texts, reviews, ratings, news, images, videos, audio, email, chat communications, search history, etc.—that form the foundation of big data. Big data is characterized by four dimensions: *Volume, Velocity, Variety, and Veracity* (Dykes 2017, McAfee et al. 2012, Zikopoulos and Eaton 2011). Since the data is in unstructured form, a few years ago, it was almost impossible to analyze the data in this form and get meaningful insights. However, today with betterment of analytics tools and technology, not only can we obtain valuable information from the data but also use the insights to predict future trends (Chen et al. 2012). Most of the analytics involve artificial intelligence and machine learning (Marr 2017b). The computers are trained to identify patterns from the data and they can spot patterns much more reliably and efficiently than humans. Advanced analytics tools can produce millions of these results in a very short time. A report by Rubinson Partners, a marketing and research firm, shows that advertisers can boost their *Return on Advertisement Spending* (ROAS) by up to 16× using aggregated big data which give them information about the right time of advertising to the consumer (Rubinson 2017).

As a result, there is tremendous curiosity about the application of big data among corporate houses. Anyone who wants to have or maintain leverage over their competitors today is encouraged to gather data and analyze them using big data analytics. However, there is still a lack of knowledge about how to implement big data analytics in many companies. In this

article, we investigate how several disciplines, specifically Information systems, operations and supply chain management, and healthcare, have applied big data in their domain. We also explore future research avenues for big data in these areas.

2. Information Systems

There was a time in academic research when data were collected solely for testing hypotheses to confirm our belief about certain phenomena or behaviors. However, when we use the Internet today, we leave a digital footprint that can be easily traced, collected, and utilized by big data analytics to *understand and predict* consumer behavior. Today it is even possible to store and analyze such massive data at an inexpensive rate. These analytics technologies can deliver new knowledge on their own without active human intervention (Dhar 2013), and as such can be very valuable.

Information systems (IS) has been an interdisciplinary domain conducting research at the intersection of computer technology and data from the business world (Agarwal and Dhar 2014). A majority of the existing research in the IS domain focuses on understanding and implementing processes that increase the efficiency of business operations. Since IS researchers were accustomed to handling huge volume of data, they started with an early advantage as far as research in big data is concerned, when compared to other business disciplines (Goes 2014). IS has contributed to the field of work surrounding big data in many ways, including surrounding issues of data integrity, data security and cybersecurity, social media, e-commerce, and web/mobile advertising. We briefly discuss the recent work in each of these areas.

Data integrity is critical to big data. To semantically integrate heterogeneous databases, it is essential to identify what entities in a data source map to the same entities in some other data sources so that data have a uniform and common structure across all heterogeneous databases (Kong et al. 2016). This process is called *entity reconciliation* (Enríquez et al. 2017, Zhao and Ram 2005). Entity reconciliation is of paramount importance to the process of data integration and management in the big data environment. Researchers have studied entity reconciliation from various perspectives. For example, Li et al. (2011) propose a context-based entity description (CED) for entity reconciliation where objects can be compared with the CED to ascertain their corresponding entities. Some researchers have also studied rule-based frameworks for entity reconciliation (Li et al. 2015).

Data security is another topic in big data where several research studies have been conducted (e.g., Chen and Zhang 2014, Demchenko et al. 2013, Katal et al.

2013). Some studies suggest the use of *real-time security analysis* as a measure for risk prevention (Lafuente 2015), whereas some others investigate *privacy-preserving data mining* (PPDM) operations (Xu et al. 2014). PPDM is a method of preserving data in such a way that applying data mining algorithms on the data do not disclose any sensitive information about the data. Big data analytics and optimization can be used as an answer against advanced cybersecurity threats (Ji et al. 2016). Since big data covers massive breadth of information sources and enormous depth of data, specifying and detecting risks become very precise (Hurst et al. 2014, Sagioglu and Sinanc 2013).

Some work at the interface of IS-Marketing research has also touched on the topic of big data. For example, data from social media have been analyzed to comprehend behavior and predict events (Ruths and Pfeffer 2014, Xu et al. 2017). In this direction, Qiu and Kumar (2017) study the performance of prediction markets through a randomized field experiment and find that an increase in audience size and a higher level of online endorsement lead to more precise predictions. Moreover, they also suggest integrating social media in predicting market because social effects and reputational concerns improve the participants' prediction accuracy. The results from this study recommend that the predictions will be more refined by targeting people of intermediate abilities. Another area of social media research where big data has contributed is text analysis and sentiment mining (Mallipeddi et al. 2017, Salehan and Kim 2016). In this area, Kumar et al. (2018a) study the importance of management responses to online consumer reviews. The results show that organizations who chose to respond to consumer comments and reviews experienced a surge in the total number of check-ins. Findings from this study also confirm that the spillover effect of online management response on neighboring organizations depends on whether the focal organization and the neighboring organizations are direct competitor of each other. Furthermore, Millham and Thakur (2016) examine the pitfalls of applying big data techniques to social media data. In this direction, Kumar et al. (2018b) propose a novel hierarchical supervised-learning approach to increase the likelihood of detecting anomalies in online reviews by analyzing several user features and then characterizing their collective behavior in a unified manner. The dishonest online reviews are difficult to detect because of complex interactions between several user characteristics, such as review velocity, volume, and variety. Kumar et al. (2018b) model user characteristics and interactions among them as univariate and multivariate distributions. They then stack these distributions using several supervised-learning techniques, such as Logistic Regression, Support

Vector Machine, and k-Nearest Neighbors yielding robust meta-classifiers.

Big data analytics has also been studied from the point of view of strategic decision-making in e-commerce (Akter and Wamba 2016) and digital marketing (Fulgoni 2013, Minelli et al. 2012). Some of the growing areas of research in e-commerce include the advertising strategy of online firms and their use of recommender systems (Ghoshal et al. 2014, 2015, Liu et al. 2012). For example, Liu et al. (2012) study the advertising game between two electronic retailers subject to a given level of information technology (IT) capacity. They reach the conclusion that if IT capacity constraints of the firms are not included in advertisement decisions, then it may result in wastage of advertisement expenditure. Based on their results, they present implementable insights for policy makers regarding how to control wasteful advertising. Ghoshal et al. (2015) find that recommendation systems impact the prices of products in both personalizing and non-personalizing firms.

Furthermore, web and mobile advertising has been an interesting area of research since the arrival of dot-com firms (Dawande et al. 2003, 2005, Fan et al. 2007, Kumar and Sethi 2009, Kumar et al. 2006). Dutta et al. (2017) and Kumar (2015) summarize the use and future trends of data analytics and optimization in web and mobile advertising. Mookerjee et al. (2016) develop a model predicting visitor's click on web advertisements. They then discuss an approach to manage Internet ads so that both click-rate and revenue earned from clicks are increased. The above group of scholars has also developed a decision-model that maximizes the advertising firm's revenue subject to a click-through rate constraint (Mookerjee et al. 2012, 2016). Another study uses the real-world data to validate new optimization methods for mobile advertising (Mookerjee et al. 2014).

IS scholars have also studied big data as a service, for example, a platform combining big data and analytics in cloud computing (Assunção et al. 2015, Demirkan and Delen 2013, Zheng et al. 2013). For instance, the Big-Data-as-a-Service (BDaaS) has been explored to yield user-friendly application programming interfaces (APIs) so that the users can easily access the service-generated big data analytic tools and corresponding results (Zheng et al. 2013). Cloud computing plays a vital role in the use and adaption of big data analytics because infrastructure requirement and cost of resources can be adjusted according to actual demand (Assunção et al. 2015).

Some studies have also been conducted on IT governance from the perspective of big data (Hashem et al. 2015, Tallon 2013) and deception detection (Fuller et al. 2013, Rubin and Lukoianova 2015). In

the IT governance domain, Tallon (2013) suggests that good data governance practices maintain a balance between value creation and risk exposure. Implementing such practices help firm earn a competitive leverage from their use of big data and application of big data analytics.

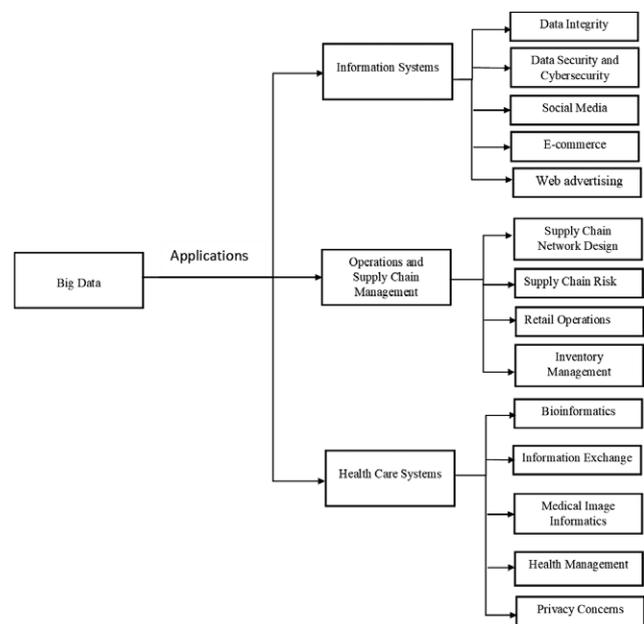
Figure 1 summarizes the above discussion. This figure also includes the contributions of big data in Operations and Supply Chain Management, and Healthcare (discussed in the following sections).

3. Operations and Supply Chain Management

With the betterment of *enterprise resource planning* (ERP) software, it is easier to capture data at different levels of operations. Firms want to analyze these data to develop more efficient processes. Hence, big data and big data analytics are being used by operations and supply chain academia as well as the industry to get insights from existing data in order to make better and informed decisions (Muhtaroglu et al. 2013, Wamba et al. 2015). The key areas in this domain where big data has left an impact are supply chain network design, risk management, inventory management, and retail operations.

Big data analytics has been used to align sourcing strategies with the organizational goals (Romano and Formentini 2012) and to evaluate the performance of suppliers (Chai and Ngai 2015, Choi 2013). Supply chain network design can itself account for a massive amount of data and hence is a favorite area for

Figure 1 Summary of Big Data Research in Operations Management, Information Systems, and Healthcare



applying big data analytics. Researchers have studied supply chain network design where the demand is uncertain (Benyoucef et al. 2013, Bouzembrak et al. 2012, Soleimani et al. 2014) as well as where the demand is certain (Jindal and Sangwan 2014, Tiwari et al. 2012). Firms can use analytics to ascertain the cost, quality, and time-to-market parameters of products to gain leverage over competitors (Bloch 2011, Luchs and Swan 2011, Srinivasan et al. 2012).

Big data analytics has also been applied to maximize production (Noyes et al., 2014) and minimize the material waste (Sharma and Agrawal 2012). Noyes et al. (2014) recommend that changes in existing manufacturing processes, incorporating automation, and simplification of methods and raw materials, will result in increasing the speed and throughput of in-process analytics during polysaccharide manufacturing processes. Moreover, Sharma and Agrawal (2012) implemented fuzzy analytic hierarchy process to solve production control policy selection problem. Inventory challenges, such as cost, demand, and supply fluctuations have also been studied using big data analytics (Babai et al. 2009, Hayya et al. 2006). In this direction, Babai et al. (2009) discuss a new dynamic inventory control method where forecasts and uncertainties related to forecast are exogenous and known at each period.

Big data has also been increasingly used in retailing. In the last decade, retailing has been one of the key areas of research for the OM researchers, especially with the growth of multi-channel retailing (Mehra et al. 2018). Big data analytics has also been applied to retail operations by firms to reduce cost and to market themselves better than the competition (Dutta et al. 2017, Janakiraman et al. 2013, Kumar et al. 2017). For instance, big data techniques are now being heavily used in recommender systems that reduce consumer search efforts (Dutta et al. 2017). Kumar et al. (2017) study how the presence of brick-and-mortar stores impacts consumers' online purchase decision. Furthermore, Janakiraman et al. (2013) study product returns in multi-channel retailing taking into consideration consumers' channel preference and choice.

4. Healthcare Systems

Healthcare systems in the United States have been rapidly adopting electronic health records (EHRs) and Healthcare Information Exchanges (HIEs) that are contributing to the accumulation of massive quantities of heterogeneous medical data from various sections of the healthcare industry—payers, providers, and pharmaceuticals (Demirezen et al. 2016, Rajapakshe et al. 2018). These data can be analyzed in order to derive insights that can improve quality of healthcare

(Groves et al. 2016). However, the analyses and practical applications of such data become a challenge because of its enormity and complexity. Since big data can deal with massive data volume and variety at high velocity, it has the potential to create significant value in healthcare by improving outcomes while lowering costs (Roski et al. 2014). It has been shown to improve the quality of care, make operational processes more efficient, predict and plan responses to disease epidemics, and optimize healthcare spending at all levels (Nambiar et al. 2013). Here, we explore how big data analytics has revolutionized the healthcare industry.

4.1. Bioinformatics

One of the subsections of the healthcare industry where big data has contributed the most is biomedical research. With the emergence and enhancement of parallel computing and cloud computing—two of the most important infrastructural pillars of big data analytics—and with the extensive use of EHRs and HIEs, the cost and effort of capturing and exploring biomedical data are decreasing.

In bioinformatics, big data contributes in yielding infrastructure for computing and data processing, including error detection techniques. Cloud-based analytics tools, such as Hadoop and MapReduce, are extensively used in the biomedical domain (Taylor 2010). Parallel computing models, such as CloudBurst (Schatz 2009), Conrail (Schatz et al. 2010), and Crossbow (Gurtowski et al. 2012), are making the genome mapping process easier. CloudBurst improves the performance of the genome mapping process as well as reduces the time required for mapping significantly (Schatz 2009). DistMap, a scalable, integrated workflow on a Hadoop cluster, supports nine different mapping tools (Pandey and Schlötterer 2013). SeqWare (D O'Connor et al. 2010), based on Apache HBase database (George 2011), is used for accessing large-scale whole-genome datasets, whereas Hydra (based on Hadoop-distributed computing framework) is used for processing large peptide and spectra databases (Lewis et al. 2012). Tools such as SAMQA (Robinson et al. 2011), ART (Huang et al. 2011), and CloudRS (Chen et al. 2013a) help in identifying errors in sequencing data. Furthermore, *Genome Analysis Toolkit* (GATK) (McKenna et al. 2010, Van der Auwera et al. 2013), BlueSNP (Huang et al. 2012), and Myrna (Langmead et al. 2010) are toolkits and packages that aid researchers in analyzing genomic data.

4.2. Healthcare Information Exchange

Clinical informatics focuses on the application of IT in the healthcare domain. It includes activity-based research, analysis of the relationship between a patient's main diagnosis (MD) and underlying cause

of death (UCD), and storage of data from EHRs and HIEs (Luo et al. 2016). Big data's main contributions have been to the manner in which EHR and HIE data are stored. The clinical real-time stream data are stored using NoSQL database, Hadoop, and HBase database because of their high-performance characteristics (Dutta et al. 2011, Jin et al. 2011, Mazurek 2014). Some research work has also studied and proposed several interactive methods of sharing medical data from multiple platforms (Chen et al. 2013b).

Healthcare Information Exchanges are used for efficient information sharing among heterogeneous healthcare entities, thus increasing the quality of care provided. Janakiraman et al. (2017) study the use of HIEs in emergency departments (EDs) and find that the benefits of HIEs increase with more information on patients, doctors, and prior interaction between them. Yaraghi et al. (2014) model HIE as a multi-sided platform. Users evaluate the self-service technologies of the model based on both user-specific and network-specific factors. Another body of research studies whether healthcare reforming models leads to better patient-centric outcomes (Youn et al. 2016).

Big data techniques have enabled the availability and analyses of a massive volume of clinical data. Insights derived from this data analysis can help medical professionals in identifying disease symptoms and predicting the cause and occurrence of diseases much better, eventually resulting in an overall improved quality of care (Genta and Sonnenberg 2014, McGregor 2013, Wang and Krishnan 2014). Since the size and complexity of data are enormous and often involve integrating clinical data from various platforms to understand the bigger picture, data security is often compromised during analysis of clinical data. Big data techniques can address this issue (Schultz 2013). Researchers have proposed several models and frameworks to efficiently protect the privacy of the data as well as effectively deal with concurrent analyses of datasets (Lin et al. 2015, Sobhy et al. 2012).

4.3. Medical Image Informatics

With the dawn of improved imaging technology, EHRs are often accompanied with high quality medical images. Studying the clinical data along with the analysis of such images will lead to better diagnoses, as well as more accurate prediction of diseases in future (Ghani et al. 2014). Medical image informatics focuses on processing images for meaningful insights using big data tools and technologies. Similarly, picture archiving and communication systems (PACS) have been critically advantageous for the medical community, since these medical images can be used for improved decision regarding treatment of patients and predicting re-admission (Ghani et al. 2014). Silva

et al. (2012) discuss how to integrate data in PACS when the digital imaging and communications in medicine (DICOM) object repository and database system of PACS are transferred to the cloud. Since analyzing large quantities of high quality clinical images using big data analytics generates rich, spatially oriented information at the cellular and sub-cellular levels, systems such as Hadoop-GIS (Wang et al. 2011), that is, cost-effective parallel systems, are being developed to aid in managing advanced spatial queries.

4.4. Health Management

Recent studies have also used big data techniques to analyze the contents of social media as a means for contagious disease surveillance, as well as for monitoring the occurrence of diseases throughout the world (Hay et al. 2013, Young et al. 2014). Big data analytics tools are used on social media communications to detect depression-related emotional patterns, and thus identify individuals suffering from depression from among the users (Nambisan et al. 2015). Health IT infrastructures, such as the US Veterans Health Administration's (VHA), have facilitated improved quality of care by providing structured clinical data from EHRs as well as unstructured data such as physician's notes (Kupersmith et al. 2007).

4.5. Privacy Concerns

In coming times, there is a massive potential of HIEs becoming public utility intermediaries that many interested markets can access to derive information (De Brantes et al. 2007). However, a major hurdle that adaptation of HIEs faces is privacy concern among consumers. A section of researchers is building HIE frameworks incorporating privacy and security principles. For example, Pickard and Swan (2014) have created a health information sharing framework, which increases sharing of health information, built on trust, motivation, and informed consent. Trust is necessary for dealing with access control issues, motivation maps the willingness to share, and informed consent enforces the legal requirement to keep the information safe. In another study, Anderson and Agarwal (2011) find that type of the requesting stakeholder and how the information will be used are two important factors that affect the privacy concern of an individual while providing access to one's health information. Numerous states in the United States have enacted laws that incentivize HIE efforts and address the concerns of patients regarding sharing of health information. In another study, Adjerid et al. (2015) observe whether various forms of privacy regulation policies facilitate or decrease HIE efforts. They find that although privacy regulation alone negatively affects HIE efforts, when combined with incentives,

privacy regulation with patient consent requirement positively impacts HIE efforts.

5. Way Ahead: Potential Applications and Challenges

In this section, we discuss the potential of big data applications in Information Systems, Operations/Supply Chain, and Healthcare domains. Figure 2 summarizes the key areas of future research.

5.1. Internet of Things (IoT) and Smart City

The Internet of Things creates a world of interconnected sensory devices containing sensors that can collect and store information from their respective real-world surroundings (Hashem et al. 2016, Riggins and Wamba 2015). According to Business Insider, the number of IoT devices will be 75 billion by the year 2020 (Danova 2013). These devices can be sensors, databases, Bluetooth devices, global positioning system (GPS), and radio-frequency identification (RFID) tags (O'Leary 2013). These devices collect massive amount of data, and if we delve down deep into this information using big data analytic tools and techniques, we may be able to derive useful insights. The applications of IoT and big data analytics combined have the potential to bring path-breaking changes to various industries and academic research. However, at the same time, since these subjects are still very new, there are uncertainties among scholars about how to implement them, and how best to extract the business value from these concepts (Riggins and Wamba 2015).

One of the domains where the coupling of big data techniques and IoT has made significant progress is the concept of a smart city, that is, where each component of urban surrounding consists of devices that are connected to a network (Hashem et al. 2015). These devices can collect data from their surroundings and share among themselves. These data can be used to monitor and manage the city in a refined dynamic manner, to improve the standard of living, and to also support the sustainability of the smart city (Kitchin 2014). IoT concepts enable information sharing across various devices, thus aiding in the creation big data caches. Furthermore, big data analytics are used to conduct real-time analysis of smart city components. Kitchin (2014) mentions that urban governance decisions and future policies regarding city life are based on these analyses. Some sub-areas under smart city where the bulk of research is being conducted are energy grids (Chourabi et al. 2012), smart environments (Atzori et al. 2010, Nam and Pardo 2011, Tiwari et al. 2011), waste management (Neirotti et al. 2014, Washburn et al. 2009), smart healthcare (Nam

and Pardo 2011, Washburn et al. 2009), and public security (Neirotti et al. 2014, Washburn et al. 2009). An emerging field surrounding smart city research is an area where big data has the potential to make a lot of contribution in the coming days.

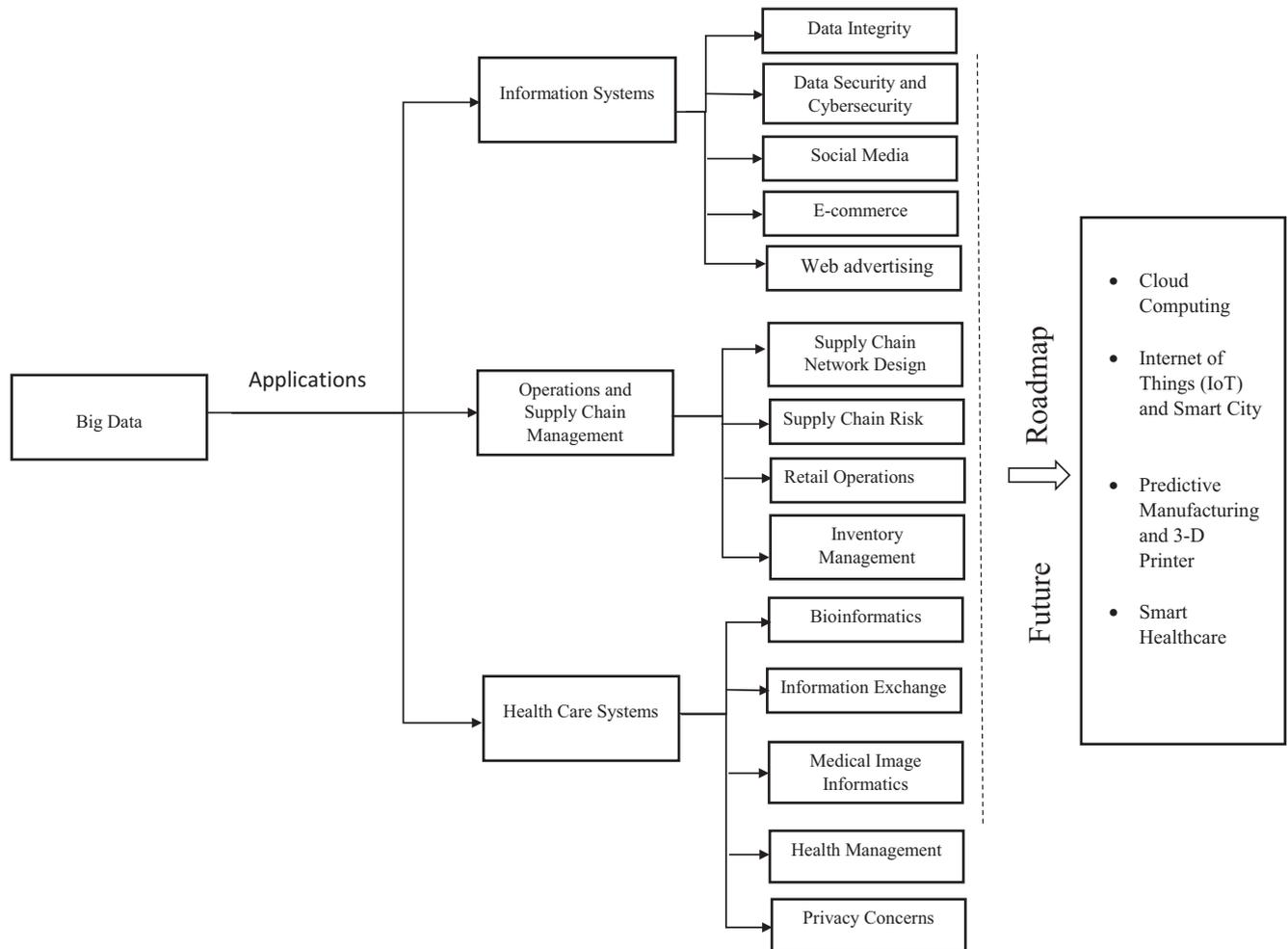
5.2. Predictive Manufacturing and 3-D Printer

Predictive manufacturing is based on cyber physical systems (CPS). CPS consists of devices that communicate with each other, as well as with the physical world, with the help of sensors and actuators (Alur 2015). CPS technology is becoming increasingly popular among manufacturers in the United States and Europe as it allows them to gain an edge in international manufacturing dynamics (Wright 2014). CPS technology can also be used to improve the design of products, to track its production and in-service performance, and to enhance productivity and efficiency of the manufacturers. General Electric (GE) and Rolls Royce have embedded sensors on their jet engines that capture data during flight and post-flight, and maintenance decisions can then be made based on these logged data (Dai et al. 2012).

Massive amounts of data are being collected from manufacturing plants through RFID and CPS technologies (Lee et al. 2013). As more advancement is made in big data analytics, these data about production equipment and operations can be processed better. Security of CPS and predictive manufacturing is another potential area where big data techniques can be applied for better security outcomes. Furthermore, additive manufacturing processes, also known as 3-D printing, are used to build three-dimensional objects by depositing materials layer-by-layer (Campbell et al. 2011, Conner et al. 2014). 3-D printing is a path-breaking technology that, in coming future, will make the existing models of manufacturing for certain products obsolete (Waller and Fawcett 2013). Hence, it is profoundly important that we study the applications of big data analytics to additive manufacturing in order to derive insights.

5.3. Smart Healthcare

Smart Healthcare is an extension of IoT ideas in the healthcare industry; that is, IoT devices equipped with RFID, Wireless Sensor Network (WSN), and advanced mobile technologies are being used to monitor patients and biomedical devices (Catarinucci et al. 2015). In the smart healthcare architecture, IoT-supporting devices are being used for seamless and constant data collection, and big data technology on the cloud is being used for storing, analyzing, and sharing this information (Muhammad et al. 2017). The nexus of IoT and big data analytics hosted on

Figure 2 Possible Future Research Directions for Big Data Applications in Operations Management, Information Systems, and Healthcare

cloud technology will not only help in more accurate detection and treatment of illnesses, but will also provide quality healthcare at a reduced cost (Varshney and Chang 2016). Moreover, smart healthcare enables to bring specialized healthcare to people who have restricted movement, or who are in remote areas where there is a dearth of specialized doctors (Muhammad et al. 2017).

Recently, the use of wearable devices has seen a rapid growth, and the number of such units shipped annually is expected to reach 148 million by 2019 (Danova 2015). Olshansky et al. (2016) discuss how data captured by wearable devices can be transmitted to health data aggregation services, such as Human API (humanapi.co) and Welltok (welltok.com), who can transform the data into measures of risk. These measures can be used to observe health trends as well as to detect and prevent diseases. Some promising topics of research in the smart healthcare domain where big data can play an important role are smart and connected health (Carroll 2016, Harwood et al.

2014, Leroy et al. 2014), and privacy issues in the smart healthcare framework (Ding et al. 2016).

6. Fading Boundaries

In this article, we explored the application of big data in three different domains—information systems, operations and supply chain, and healthcare. But, the line between these disciplines are blurring with each passing day. Several new avenues of research are becoming popular that are common to at least two of these domains. One such topic is use of ERP platforms in healthcare that is common to all the three fields.

Healthcare organizations accumulate massive amounts of information from various departments and then different entities in healthcare management rely on to carry out their services. An automated integrated system, such as an ERP system to manage the information coming from different services and processes, will enable healthcare organizations to improve efficiency of service and quality of care

(Handayani et al. 2013). The motivations underlying the adoption of ERP system in healthcare management are technological, managerial, clinical, and financial (Poba-Nzaou et al. 2014). An ERP system integrates various business units of healthcare organization, such as finance, operation and supply chain management, and human resource, and provides easy access within each unit. It can also address the disparity in healthcare quality between urban and rural settings. ERP provides connectivity among all healthcare centers and hence information can also be accessed from rural centers (Padhy et al. 2012). Benefits from implementing ERP can be classified into four categories—patients' satisfaction, stakeholders' satisfaction, operations efficiency, and strategic and performance management (Chiarini et al. 2017). However, ERP systems are costly to acquire and involve hidden costs even after successful implementation such as integration testing and staff members training costs (Gupta 2000, Wailgum 2008). Till date, majority of research work involving ERP in healthcare domain has revolved around implementation of ERP systems (Mucheleka and Halonen 2015). One potential research avenue is to conduct empirical studies to quantify the benefits from implementation of such systems.

7. Closing Thoughts

We generate data whenever we use the Internet. Aside from the data generated by us, several interconnected smart devices collect data, that is, devices with sensors collect data from their surrounding real world. With this tremendous quantity of data generated each day, big data and big data analytics are very much in demand in several industries as well as among scholars. In this study, we discussed the contributions of big data in information systems, operations and supply chain management, and healthcare domains. At the end, we talked about four sub-areas of these domains—cloud computing, Internet of things (IoT) and smart city, predictive manufacturing and 3-D printer, and smart healthcare—where big data techniques can lead to significant improvements. We also discussed the corresponding challenges and future research opportunities in the field, noting numerous areas for growth and exploration.

References

- Adjerid, I., A. Acquisti, R. Telang, R. Padman, J. Adler-Milstein. 2015. The impact of privacy regulation and technology incentives: The case of health information exchanges. *Management Sci.* 62(4): 1042–1063.
- Agarwal, R., V. Dhar. 2014. Big data, data science, and analytics: The opportunity and challenge for IS research. *Inf. Syst. Res.* 25(3): 443–448.
- Akter, S., S. F. Wamba. 2016. Big data analytics in e-commerce: A systematic review and agenda for future research. *Electron. Mark.* 26(2): 173–194.
- Alur, R. 2015. *Principles of Cyber-Physical Systems*. MIT Press, Cambridge, MA.
- Anderson, C. L., R. Agarwal. 2011. The digitization of healthcare: Boundary risks, emotion, and consumer willingness to disclose personal health information. *Inf. Syst. Res.* 22(3): 469–490.
- Assunção, M. D., R. N. Calheiros, S. Bianchi, M. A. Netto, R. Buyya. 2015. Big data computing and clouds: Trends and future directions. *J. Parallel Distrib. Comput.* 79: 3–15.
- Atzori, L., A. Iera, G. Morabito. 2010. The Internet of things: A survey. *Comput. Netw.* 54(15): 2787–2805.
- Babai, M. Z., A. A. Syntetos, Y. Dallery, K. Nikolopoulos. 2009. Dynamic re-order point inventory control with lead-time uncertainty: Analysis and empirical investigation. *Int. J. Prod. Res.* 47(9): 2461–2483.
- Benyoucef, L., X. Xie, G. A. Tanonkou. 2013. Supply chain network design with unreliable suppliers: A Lagrangian relaxation-based approach. *Int. J. Prod. Res.* 51(21): 6435–6454.
- Bloch, P. H. 2011. Product design and marketing: Reflections after fifteen years. *J. Prod. Innov. Manag.* 28(3): 378–380.
- Bouzembrak, Y., H. Allaoui, G. Goncalves, H. Bouchriha, M. Baklouti. 2012. A possibilistic linear programming model for supply chain network design under uncertainty. *IMA J. Manag. Math.* 24(2): 209–229.
- Campbell, T., C. Williams, O. Ivanova, B. Garrett. 2011. *Could 3D Printing Change the World: Technologies, Potential, and Implications of Additive Manufacturing*. Atlantic Council, Washington, DC. Available at <http://www.cbpp.uaa.alaska.edu/afef/Additive%20MFG%20.pdf> (accessed date October 10, 2017).
- Carroll, N. 2016. Key success factors for smart and connected health software solutions. *Computer* 49(11): 22–28.
- Catarinucci, L., D. De Donno, L. Mainetti, L. Palano, L. Patrono, M. L. Stefanizzi, L. Tarricone. 2015. An IoT-aware architecture for smart healthcare systems. *IEEE Internet Things J.* 2(6): 515–526.
- Chai, J., E. W. Ngai. 2015. Multi-perspective strategic supplier selection in uncertain environments. *Int. J. Prod. Econ.* 166: 215–225.
- Chen, C. P., C. Y. Zhang. 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Inf. Sci.* 275: 314–347.
- Chen, H., R. H. Chiang, V. C. Storey. 2012. Business intelligence and analytics: From big data to big impact. *MIS Q.* 36(4): 1165–1188.
- Chen, C. C., Y. J. Chang, W. C. Chung, D. T. Lee, J. M. Ho. 2013a. CloudRS: An error correction algorithm of high-throughput sequencing data based on scalable framework. J. Han, ed. 2013 *IEEE International Conference on Big Data*. Silicon Valley, CA, pp. 717–722. doi: 10.1109/BigData.2013.6691642.
- Chen, J., F. Qian, W. Yan, B. Shen. 2013b. Translational biomedical informatics in the cloud: Present and future. *Biomed. Res. Int.* 2013: 1–8.
- Chiarini, A., E. Vagnoni, L. Chiarini. 2017. ERP implementation in public healthcare, achievable benefits and encountered criticalities: An investigation from Italy. *Int. J. Serv. Oper. Manag. (Online)* 29(1): 1–17.
- Choi, T. M. 2013. Optimal apparel supplier selection with forecast updates under carbon emission taxation scheme. *Comput. Oper. Res.* 40(11): 2646–2655.
- Chourabi, H., T. Nam, S. Walker, J. R. Gil-Garcia, S. Mellouli, K. Nahon, T. A. Pardo, H. J. Scholl. 2012. Understanding smart cities: An integrative framework. R. H. Sprague Jr., ed. *45th*

- Hawaii International Conference on System Sciences*. Maui, HI, pp. 2289–2297. doi: 10.1109/HICSS.2012.615.
- Conner, B. P., G. P. Manogharan, A. N. Martof, L. M. Rodomsky, C. M. Rodomsky, D. C. Jordan, J. W. Limperos. 2014. Making sense of 3-D printing: Creating a map of additive manufacturing products and services. *Addit. Manuf.* 1: 64–76.
- D O'Connor, B., B. Merriman, S. F. Nelson. 2010. SeqWare query engine: Storing and searching sequence data in the cloud. *BMC Bioinformatics* 11(12): S2.
- Dai, X., K. Sasloglou, R. Atkinson, J. Strong, I. Panella, L. Yun Cai, H. Mingding, W. Ang Chee, I. Glover, J. Mitchell, P. Dutta, W. Schiffers. 2012. Wireless communication networks for gas turbine engine testing. *Int. J. Distrib. Sens. Netw.* 8(3), 212876.
- Danova, T. 2013. Morgan Stanley: 75 billion devices will be connected to the Internet of Things by 2020. *Business Insider*. Available at <http://www.businessinsider.com/75-billion-devices-will-be-connected-to-the-internet-by-2020-2013-10> (accessed date October 4, 2017).
- Danova, T. 2015. The wearables report: Growth trends, consumer attitudes, and why smartwatches will dominate. *Business Insider*. Available at <http://www.businessinsider.com/the-wearable-computing-market-report-2014-10> (accessed date October 4, 2017)
- Dawande, M., S. Kumar, C. Sriskandarajah. 2003. Performance bounds of algorithms for scheduling advertisements on a web page. *J. Sched.* 6(4): 373–394.
- Dawande, M., S. Kumar, C. Sriskandarajah. 2005. Scheduling web advertisements: A note on the minspace problem. *J. Sched.* 8(1): 97–106.
- De Brantes, F., D. W. Emery, J. M. Overhage, J. Glaser, J. Marchibroda. 2007. The potential of HIEs as infomediaries. *J. Healthc. Inform. Manag.* 21(1): 69–75.
- Demchenko, Y., C. Ngo, C. de Laat, P. Membrey, D. Gordijenko. 2013. Big security for big data: Addressing security challenges for the big data infrastructure. W. Jonker, M. Petkovic, eds. *Workshop on Secure Data Management*. Springer, Cham, 76–94.
- Demirezen, E. M., S. Kumar, A. Sen. 2016. Sustainability of healthcare information exchanges: A game-theoretic approach. *Inf. Syst. Res.* 27(2): 240–258.
- Demirkan, H., D. Delen. 2013. Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. *Decis. Support Syst.* 55(1): 412–421.
- Dhar, V. 2013. Data science and prediction. *Commun. ACM* 56(12): 64–73.
- Ding, D., M. Conti, A. Solanas. 2016. A smart health application and its related privacy issues. A. Mohan, A. S. Uluagac, M. Conti, eds. *2016 Smart City Security and Privacy Workshop (SCSP-W)*. IEEE, Vienna, pp. 1–5. doi: 10.1109/SCSPW.2016.7509558.
- Dutta, H., A. Kamil, M. Pooleery, S. Sethumadhavan, J. Demme. 2011. Distributed storage of large-scale multidimensional electroencephalogram data using Hadoop and HBase. S. Fiore and G. Aloisio, eds. *Grid and Cloud Database Management*. Springer, Berlin and Heidelberg, 331–347.
- Dutta, K., A. Ghoshal, S. Kumar. 2017. The interdependence of data analytics and operations management. M. K. Starr, S. K. Gupta, eds. *Routledge Companion for Production and Operations Management (POM)*. Taylor & Francis, New York, 291–308.
- Dykes, B. 2017. Big data: Forget volume and variety, focus on velocity. *Forbes*, June 28. Available at <https://www.forbes.com/sites/brentdykes/2017/06/28/big-data-forget-volume-and-variety-focus-on-velocity/#1ed5e4236f7d> (accessed date October 6, 2017).
- Enríquez, J. G., F. J. Domínguez-Mayo, M. J. Escalona, M. Ross, G. Staples. 2017. Entity reconciliation in big data sources: A systematic mapping study. *Expert Syst. Appl.* 80: 14–27.
- Fan, M., S. Kumar, A. B. Whinston. 2007. Selling or advertising: Strategies for providing digital media online. *J. Manage. Inf. Syst.* 24(3): 143–166.
- Fulgoni, G. 2013. Big data: Friend or foe of digital advertising? *J. Advertising Res.* 53(4): 372–376.
- Fuller, C. M., D. P. Biros, J. Burgoon, J. Nunamaker. 2013. An examination and validation of linguistic constructs for studying high-stakes deception. *Group Decis. Negot.* 22(1): 117–134.
- Gelsing, P. 2012. Data: The new science. *DellEMC*, June 25. Available at <https://blog.dellemc.com/en-us/data-the-new-science/> (accessed date October 6, 2017).
- Genta, R. M., A. Sonnenberg. 2014. Big data in gastroenterology research. *Nat. Rev. Gastroenterol. Hepatol.* 11(6): 386–390.
- George, L. 2011. *HBase: The Definitive Guide: Random Access to Your Planet-Size Data*. O'Reilly Media Inc, Sebastopol, CA.
- Ghani, K. R., K. Zheng, J. T. Wei, C. P. Friedman. 2014. Harnessing big data for healthcare and research: Are urologists ready? *Eur. Urol.* 66(6): 975–977.
- Ghoshal, A., S. Kumar, V. S. Mookerjee. 2014. Locking effects of recommender systems: A competitive analysis. Proceedings of the 24th Annual Workshop on Information Technologies and Systems (WITS), December 17–19, 2014, Auckland, New Zealand.
- Ghoshal, A., S. Kumar, V. Mookerjee. 2015. Impact of recommender system on competition between personalizing and non-personalizing firms. *J. Manage. Inf. Syst.* 31(4): 243–277.
- Goes, P. B. 2014. Editor's comments: Big data and IS research. *MIS Q.* 38(3): iii–viii.
- Groves, P., B. Kayyali, D. Knott, S. V. Kuiken. 2016. The “big data” revolution in healthcare: Accelerating value and innovation. *McKinsey Q.* Available at http://www.pharmatalents.es/assets/files/Big_Data_Revolution.pdf (accessed date October 4, 2017).
- Gupta, A. 2000. Enterprise resource planning: The emerging organizational value systems. *Ind. Manage. Data Syst.* 100(3): 114–118.
- Gurtowski, J., M. C. Schatz, B. Langmead. 2012. Genotyping in the cloud with crossbow. *Curr. Protoc. Bioinformatics* 39: 15.3.1–15.3.15.
- Handayani, P. W., M. Z. Rahman, A. N. Hidayanto. 2013. Information technology assessment on hospital information system implementation: Case study a teaching hospital. *Int. J. Eng. Technol.* 5(2): 631–634.
- Harwood, J., J. J. Dooley, A. J. Scott, R. Joiner. 2014. Constantly connected—The effects of smart-devices on mental health. *Comput. Hum. Behav.* 34: 267–272.
- Hashem, I. A. T., I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, S. U. Khan. 2015. The rise of “big data” on cloud computing: Review and open research issues. *Inf. Syst.* 47: 98–115.
- Hashem, I. A. T., V. Chang, N. B. Anuar, K. Adewole, I. Yaqoob, A. Gani, E. Ahmed, H. Chiroma. 2016. The role of big data in smart city. *Int. J. Inf. Manage.* 36(5): 748–758.
- Hay, S. I., D. B. George, C. L. Moyes, J. S. Brownstein. 2013. Big data opportunities for global infectious disease surveillance. *PLoS Med.* 10(4): e1001413.
- Hayya, J. C., J. G. Kim, S. M. Disney, T. P. Harrison, D. Chatfield. 2006. Estimation in supply chain inventory management. *Int. J. Prod. Res.* 44(7): 1313–1330. <https://doi.org/10.1080/00207540500338039>.
- Huang, W., L. Li, J. R. Myers, G. T. Marth. 2011. ART: A next-generation sequencing read simulator. *Bioinformatics* 28(4): 593–594.
- Huang, H., S. Tata, R. J. Prill. 2012. BlueSNP: R package for highly scalable genome-wide association studies using Hadoop clusters. *Bioinformatics* 29(1): 135–136.

- Hurst, W., M. Merabti, P. Fergus. 2014. Big data analysis techniques for cyber-threat detection in critical infrastructures. F. Xhafa, S. Wang, H.-H. Hsu, D. Taniar, eds. *28th International Conference on Advanced Information Networking and Applications Workshops*. Victoria, BC, pp. 916–921. doi: 10.1109/WAINA.2014.141.
- Janakiraman, R., Y. Liu, R. Bezawada, S. Kumar. 2013. A structural model of consumers' perception of channel fit and consumer channel choice: Evidence from a multichannel retailer. Conference on Information Systems and Technology (CIST), October 5–6, Minneapolis, MN.
- Janakiraman, R., E. Park, E. Demirezen, S. Kumar. 2017. The effects of health information exchange access on healthcare quality and efficiency: An empirical investigation. Available at <https://ssrn.com/abstract=2915190> (accessed date October 9, 2017).
- Ji, Y., S. Kumar, V. Mookerjee. 2016. When being hot is not cool: Monitoring hot lists for information security. *Inf. Syst. Res.* 27(4): 897–918.
- Jin, Y., T. Deyu, Z. Yi. 2011. A distributed storage model for EHR based on HBase. W. Susheng, H. Sun, K. M. Sim, L. Li, T. VM Rao, M. Zhang, eds. *2011 4th International Conference on Information Management, Innovation Management and Industrial Engineering*. Shenzhen, China, pp. 369–372. doi: 10.1109/ICIII.2011.234.
- Jindal, A., K. S. Sangwan. 2014. Closed loop supply chain network design and optimisation using fuzzy mixed integer linear programming model. *Int. J. Prod. Res.* 52(14): 4156–4173.
- Katal, A., M. Wazid, R. H. Goudar. 2013. Big data: Issues, challenges, tools and good practices. M. Parashar, A. Zomaya, eds. *Sixth International Conference on Contemporary Computing (IC3)*. IEEE, Noida, pp. 404–409. doi: 10.1109/IC3.2013.6612229.
- Kitchin, R. 2014. The real-time city? Big data and smart urbanism. *GeoJournal* 79(1): 1–14.
- Kong, C., M. Gao, C. Xu, W. Qian, A. Zhou. 2016. Entity matching across multiple heterogeneous data sources. S. Navathe, W. Wu, S. Shekhar, X. Du, X. Wang, H. Xiong, eds. *International Conference on Database Systems for Advanced Applications*. Springer, Cham, 133–146.
- Kumar, S. 2015. *Optimization Issues in Web and Mobile Advertising: Past and Future Trends*. Springer, New York.
- Kumar, S., S. P. Sethi. 2009. Dynamic pricing and advertising for web content providers. *Eur. J. Oper. Res.* 197(3): 924–944.
- Kumar, S., V. S. Jacob, C. Sriskandarajah. 2006. Scheduling advertisements on a web page to maximize revenue. *Eur. J. Oper. Res.* 173(3): 1067–1089.
- Kumar, A., A. Mehra, S. Kumar. 2017. How do stores drive online sales? The less obvious effects of stores on revenues of a multi-channel retailer. Conference on Information Systems and Technology (CIST). Houston, Texas.
- Kumar, N., L. Qiu, S. Kumar. 2018a. Exit, voice, and response in digital platforms: An empirical investigation of online management response strategies. *Inf. Syst. Res.* (e-pub ahead of print). <https://doi.org/10.1287/isre.2017.0749>.
- Kumar, N., D. Venugopal, L. Qiu, S. Kumar. 2018b. Detecting review manipulation on online platforms with hierarchical supervised learning. *J. Manage. Inf. Syst.* 35(1): 350–380. <https://doi.org/10.1287/isre.2017.0749>.
- Kupersmith, J., J. Francis, E. Kerr, S. Krein, L. Pogach, R. M. Kolodner, J. B. Perlin. 2007. Advancing evidence-based care for diabetes: Lessons from the Veterans Health Administration. *Health Aff.* 26(2): w156–w168.
- Lafuente, G. 2015. The big data security challenge. *Netw. Secur.* 2015(1): 12–14.
- Langmead, B., K. D. Hansen, J. T. Leek. 2010. Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol.* 11(8): R83.
- Lee, J., E. Lapira, B. Bagheri, H. A. Kao. 2013. Recent advances and trends in predictive manufacturing systems in big data environment. *Manuf. Lett.* 1(1): 38–41.
- Leroy, G., H. Chen, T. C. Rindfleisch. 2014. Smart and connected health. *IEEE Intell. Syst.* 29(3): 2–5.
- Lewis, S., A. Csordas, S. Killcoyne, H. Hermjakob, M. R. Hoopmann, R. L. Moritz, E. Deutsch, J. Boyle. 2012. Hydra: A scalable proteomic search engine which utilizes the Hadoop distributed computing framework. *BMC Bioinformatics* 13(1): 324.
- Li, L., J. Li, H. Wang, H. Gao. 2011. Context-based entity description rule for entity resolution. M. Lease, F. Cacheda, eds. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. ACM, Glasgow, Scotland, 1725–1730.
- Li, L., J. Li, H. Gao. 2015. Rule-based method for entity resolution. *IEEE Trans. Knowl. Data Eng.* 27(1): 250–263.
- Lin, W., W. Dou, Z. Zhou, C. Liu. 2015. A cloud-based framework for home-diagnosis service over big medical data. *J. Syst. Softw.* 102: 192–206.
- Liu, D., S. Kumar, V. S. Mookerjee. 2012. Advertising strategies in electronic retailing: A differential games approach. *Inf. Syst. Res.* 23(3-part-2): 903–917.
- Luchs, M., K. S. Swan. 2011. Perspective: The emergence of product design as a field of marketing inquiry. *J. Prod. Innov. Manag.* 28(3): 327–345.
- Luo, J., M. Wu, D. Gopukumar, Y. Zhao. 2016. Big data application in biomedical research and healthcare: A literature review. *Biomed. Inform. Insights* 8: 1.
- Mallipeddi, R. R., R. Janakiraman, S. Kumar, S. Gupta. 2017. The effects of social media tone on engagement: Evidence from Indian general election 2014. Conference on Information Systems and Technology (CIST). Available at SSRN: <https://ssrn.com/abstract=2980481> (accessed date October 9, 2017).
- Marr, B. 2017a. The complete beginner's guide to big data in 2017. *Forbes*, March 14. Available at <https://www.forbes.com/sites/bernardmarr/2017/03/14/the-complete-beginners-guide-to-big-data-in-2017/#799e5d117365> (accessed date October 1, 2017).
- Marr, B. 2017b. New tool uses machine learning and artificial intelligence to improve IT operations. *Forbes*, September 26. Available at <https://www.forbes.com/sites/bernardmarr/2017/09/26/new-tool-uses-machine-learning-and-artificial-intelligence-to-improve-it-operations/#648fa05a3789> (accessed date October 4, 2017).
- Mazurek, M. 2014. Applying NoSQL databases for operationalizing clinical data mining models. S. Kozielski, D. Mrozek, P. Kasprowski, B. Malysiak-Mrozek, D. Kostrzewa, eds. *BDAS 2014*. Springer, Cham, 527–536.
- McAfee, A., E. Brynjolfsson, T. H. Davenport. 2012. Big data: The management revolution. *Harv. Bus. Rev.* 90(10): 60–68.
- McGregor, C. 2013. Big data in neonatal intensive care. *Computer* 46(6): 54–59.
- McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytzsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo. 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20(9): 1297–1303.
- Mehra, A., S. Kumar, J. S. Raju. 2018. Competitive strategies for brick-and-mortar stores to counter “showrooming”. *Management Sci.* 64(7): 3076–3090. <https://doi.org/10.1287/mnsc.2017.2764>.

- Millham, R., S. Thakur. 2016. *Social media and big data*. G. S. Tomar, N. S. Chaudhari, R. S. Bhadoria, G. C. Deka, eds. *The Human Element of Big Data: Issues, Analytics, and Performance*. CRC Press, Boca Raton, FL, pp. 179–194.
- Minelli, M., M. Chambers, A. Dhiraj. 2012. *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*. John Wiley & Sons, New York.
- Mookerjee, R., S. Kumar, V. S. Mookerjee. 2012. To show or not show: Using user profiling to manage Internet advertisement campaigns at Chitika. *Interfaces* 42(5): 449–464.
- Mookerjee, R., S. Kumar, V. S. Mookerjee, C. Sriskandarajah. 2014. Demand-supply optimization in mobile advertising. Proceedings of the 24th Annual Workshop on Information Technologies and Systems (WITS), Poster Session, December 17–19, 2014. Auckland, New Zealand.
- Mookerjee, R., S. Kumar, V. S. Mookerjee. 2016. Optimizing performance-based Internet advertisement campaigns. *Oper. Res.* 65(1): 38–54.
- Mucheleka, M., R. Halonen. 2015. ERP in healthcare. *2015 ICEIS 17th International Conference on Enterprise Information Systems*. Barcelona, Spain, April 27–30, 2015. pp. 162–171.
- Muhammad, G., S. M. M. Rahman, A. Alelaiwi, A. Alamri. 2017. Smart health solution integrating IoT and cloud: A case study of voice pathology monitoring. *IEEE Commun. Mag.* 55(1): 69–73.
- Muhtaroglu, F. C. P., S. Demir, M. Obali, C. Girgin. 2013. Business model canvas perspective on big data applications. J. Han, ed. *2013 IEEE International Conference on Big Data*. Silicon Valley, CA, pp. 32–37. doi: 10.1109/BigData.2013.6691684.
- Nam, T., T. A. Pardo. 2011. Conceptualizing smart city with dimensions of technology, people, and institutions. J. Bertot, K. Nahon, eds. *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*. ACM, College Park, MD, pp. 282–291.
- Nambiar, R., R. Bhardwaj, A. Sethi, R. Varghese. 2013. A look at challenges and opportunities of big data analytics in healthcare. J. Han, ed. *2013 IEEE International Conference on Big Data*. Silicon Valley, CA, pp. 17–22. doi: 10.1109/BigData.2013.6691753.
- Nambisan, P., Z. Luo, A. Kapoor, T. B. Patrick, R. A. Cisler. 2015. Social media, big data, and public health informatics: Ruminating behavior of depression revealed through twitter. T. X. Bui, R. H. Sprague Jr., eds. *2015 48th Hawaii International Conference on System Sciences*. Kauai, HI, 2015, pp. 2906–2913. doi: 10.1109/HICSS.2015.351.
- Neirotti, P., A. De Marco, A. C. Cagliano, G. Mangano, F. Scorrano. 2014. Current trends in Smart City initiatives: Some stylised facts. *Cities* 38: 25–36.
- Noyes, A., R. Godavarti, N. Titchener-Hooker, J. Coffman, T. Mukhopadhyay. 2014. Quantitative high throughput analytics to support polysaccharide production process development. *Vaccine* 32(24): 2819–2828.
- O'Leary, D. E. 2013. Big data, the “internet of things” and the “internet of signs”. *Intell. Syst. Account. Financ. Manag.* 20(1): 53–65.
- Olshansky, S. J., B. A. Carnes, Y. C. Yang, N. Miller, J. Anderson, H. Beltran-Sanchez, K. Ricanek. 2016. The future of smart health. *IEEE Comput.* 49(11): 14–21.
- Padhy, R. P., M. R. Patra, S. C. Satapathy. 2012. Design and implementation of a cloud based rural healthcare information system model. *Univers J. Appl. Comput. Sci. Technol.* 2(1): 149–157.
- Pandey, R. V., C. Schlötterer. 2013. DistMap: A toolkit for distributed short read mapping on a Hadoop cluster. *PLoS ONE* 8(8): e72614.
- Pickard, K. T., M. Swan. 2014. Big desire to share big health data: A shift in consumer attitudes toward personal health information. *2014 AAAI Spring Symposium Series*, pp. 2168–7161.
- Poba-Nzaou, P., S. Uwizeyemungu, L. Raymond, G. Paré. 2014. Motivations underlying the adoption of ERP systems in healthcare organizations: Insights from online stories. *Inf. Syst. Front.* 16(4): 591–605.
- Qiu, L., S. Kumar. 2017. Understanding voluntary knowledge provision and content contribution through a social-media-based prediction market: A field experiment. *Inf. Syst. Res.* 28(3): 529–546.
- Rajapakshe, T., S. Kumar, A. Sen, C. Sriskandarajah. 2018. Sustainability planning for healthcare information exchanges with supplier rebate program. Working Paper.
- Riggins, F. J., S. F. Wamba. 2015. Research directions on the adoption, usage, and impact of the internet of things through the use of big data analytics. T. X. Bui, R. H. Sprague Jr., eds. *2015 48th Hawaii International Conference on System Sciences*. Kauai, HI, 2015, pp. 1531–1540. doi: 10.1109/HICSS.2015.186.
- Robinson, T., S. Killcoyne, R. Bressler, J. Boyle. 2011. SAMQA: Error classification and validation of high-throughput sequenced read data. *BMC Genom.* 12(1): 419.
- Romano, P., M. Formentini. 2012. Designing and implementing open book accounting in buyer-supplier dyads: A framework for supplier selection and motivation. *Int. J. Prod. Econ.* 137(1): 68–83.
- Roski, J., G. W. Bo-Linn, T. A. Andrews. 2014. Creating value in healthcare through big data: Opportunities and policy implications. *Health Aff.* 33(7): 1115–1122.
- Rubin, V. L., T. Lukoianova. 2015. Truth and deception at the rhetorical structure level. *J. Assoc. Inf. Sci. Technol.* 66(5): 905–917.
- Rubinson, J. 2017. The next sea change in marketing is coming fast. Are you ready? Available at <http://blog.joelrubinson.net/2017/08/the-next-sea-change-in-marketing-is-coming-fast-are-you-ready/> (accessed date October 8, 2017).
- Ruths, D., J. Pfeffer. 2014. Social media for large studies of behavior. *Science* 346(6213): 1063–1064.
- Ryan, C., J. Lewis. 2017. Computer and Internet use in the United States: 2015. Available at <https://www.census.gov/content/dam/Census/library/publications/2017/acs/acs-37.pdf> (accessed date October 14, 2017).
- Sagioglu, S., D. Sinanc. 2013. Big data: A review. D. W. Barnett, J. Zubairi, D. W. Barnett, J. Zubairi, eds. *2013 International Conference on Collaboration Technologies and Systems (CTS)*. San Diego, CA, 2013, pp. 42–47. <https://doi.org/10.1109/cts.2013.6567202>.
- Salehan, M., D. J. Kim. 2016. Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. *Decis. Support Syst.* 81: 30–40.
- Schatz, M. C. 2009. CloudBurst: Highly sensitive read mapping with MapReduce. *Bioinformatics* 25(11): 1363–1369.
- Schatz, M., A. L. Delcher, S. L. Salzberg. 2010. Assembly of large genomes using cloud computing. *Genome Res.* 20(9): 1165–1173.
- Schultz, T. 2013. Turning healthcare challenges into big data opportunities: A use-case review across the pharmaceutical development lifecycle. *Bull. Assoc. Inf. Sci. Technol.* 39(5): 34–40.
- Sharma, S., N. Agrawal. 2012. Application of fuzzy techniques in a multistage manufacturing system. *Int. J. Adv. Manuf. Technol.* 60(1): 397–407.
- Silva, L. A. B., C. Costa, J. L. Oliveira. 2012. A PACS archive architecture supported on cloud services. *Int. J. Comput. Assist. Radiol. Surg.* 7(3): 349–358.
- Sobhy, D., Y. El-Sonbaty, M. A. Elnasr. 2012. MedCloud: Healthcare cloud computing system. C. A. Shoniregun, ed. *2012*

- International Conference for Internet Technology and Secured Transactions*. London, pp. 161–166.
- Soleimani, H., M. Seyyed-Esfahani, G. Kannan. 2014. Incorporating risk measures in closed-loop supply chain network design. *Int. J. Prod. Res.* **52**(6): 1843–1867.
- Srinivasan, R., G. L. Lilien, A. Rangaswamy, G. M. Pingitore, D. Seldin. 2012. The total product design concept and an application to the auto market. *J. Prod. Innov. Manag.* **29** (S1): 3–20.
- Tallon, P. P. 2013. Corporate governance of big data: Perspectives on value, risk, and cost. *Computer* **46**(6): 32–38.
- Taylor, R. C. 2010. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. *BMC Bioinformatics* **11**(12): S1.
- Tiwari, R., R. Cervero, L. Schipper. 2011. Driving CO₂ reduction by integrating transport and urban design strategies. *Cities* **28** (5): 394–405.
- Tiwari, A., P. C. Chang, M. K. Tiwari. 2012. A highly optimised tolerance-based approach for multi-stage, multi-product supply chain network design. *Int. J. Prod. Res.* **50**(19): 5430–5444.
- Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, E. Banks. 2013. From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**: 11.10.11–11.10.13.
- Varshney, U., C. K. Chang. 2016. Smart health and well-being. *Computer* **49**(11): 11–13.
- Wailgum, T. 2008. ERP Definition and solutions. *White Paper*, November 15, 2017.
- Waller, M. A., S. E. Fawcett. 2013. Click here for a data scientist: Big data, predictive analytics, and theory development in the era of a maker movement supply chain. *J. Bus. Log.* **34**(4): 249–252.
- Wamba, S. F., S. Akter, A. Edwards, G. Chopin, D. Gnanzou. 2015. How “big data” can make big impact: Findings from a systematic review and a longitudinal case study. *Int. J. Prod. Econ.* **165**: 234–246.
- Wang, W., E. Krishnan. 2014. Big data and clinicians: A review on the state of the science. *JMIR Med. Inform.* **2**(1).
- Wang, F., A. Aji, Q. Liu, J. Saltz. 2011. Hadoop-GIS: A high performance spatial query system for analytical medical imaging with MapReduce. Technical report, Emory University. Available at <http://www3.cs.stonybrook.edu/~fuswang/papers/CCI-TR-2011-3.pdf> (accessed date October 4, 2017).
- Washburn, D., U. Sindhu, S. Balaouras, R. A. Dines, N. Hayes, L. E. Nelson. 2009. Helping CIOs understand “smart city” initiatives. *Growth* **17**(2): 1–17.
- Wright, P. 2014. Cyber-physical product manufacturing. *Manuf. Lett.* **2**(2): 49–53.
- Xu, L., C. Jiang, J. Wang, J. Yuan, Y. Ren. 2014. Information security in big data: Privacy and data mining. *IEEE Access* **2**: 1149–1176.
- Xu, Z., H. Zhang, C. Hu, Y. Liu, J. Xuan, L. Mei. 2017. Crowdsourcing-based timeline description of urban emergency events using social media. *Int. J. Ad Hoc Ubiquitous Comput.* **25**(1–2): 41–51.
- Yaraghi, N., A. Y. Du, R. Sharman, R. D. Gopal, R. Ramesh. 2014. Health information exchange as a multisided platform: Adoption, usage, and practice involvement in service co-production. *Inf. Syst. Res.* **26**(1): 1–18.
- Youn, S., G. R. Heim, S. Kumar, C. Sriskandarajah. 2016. Hospital quality, medical charge variation, and patient care efficiency: Implications for bundled payment reform models. *Healthcare Conference: Patient-Centric HealthCare Management in the Age of Analytics*. Available at <https://ssrn.com/abstract=2876358> or <https://doi.org/10.2139/ssrn.2876358> (accessed date October 4, 2017).
- Young, S. D., C. Rivers, B. Lewis. 2014. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Prev. Med.* **63**: 112–115.
- Zhao, H., S. Ram. 2005. Entity identification for heterogeneous database integration—A multiple classifier system approach and empirical evaluation. *Inf. Syst.* **30**(2): 119–132.
- Zheng, Z., J. Zhu, M. R. Lyu. 2013. Service-generated big data and big data-as-a-service: An Overview. S. Sakr, ed. *IEEE International Congress on Big Data*. Santa Clara, CA, pp. 403–410. <https://doi.org/10.1109/BigData.Congress.2013.60>.
- Zikopoulos, P., C. Eaton. 2011. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw-Hill Osborne Media, New York.

Discover
**Business &
Management**
with Wiley

For Business & Management researchers, students, and faculty, Wiley's varied list of distinctive journals, books, and online resources provides the highest level of scholarship that spans the breadth of the discipline.

Business Ethics

Conflict Resolution

Corporate Governance

Creativity & Innovation Management

Consumer Behavior

Decision Sciences

Human Resource Management

Industrial & Labor Relations

International Management

Leadership & Teams

Management Science

Marketing Management

Non-Profit Organizations

Operational Research

Organizational Behavior

Organizational Development

Organization & Management Theory

Production Operations Management

Project Management

Public Administration

Sales

Small Business & Entrepreneurship

Strategic Management

Training & Development

Discover all that Wiley has to offer in your field
wileyonlinelibrary.com/subject/business

WILEY

New from TED and Wiley

TEDStudies

Instructor materials created by **WILEY**

We are proud to present in collaboration with TED a curated series of talks plus specially commissioned educational support materials. Two new courses will be released each month.

Invite guest speakers into your classroom with the following courses:

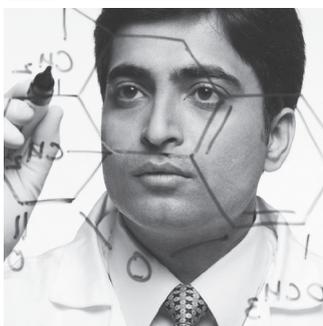
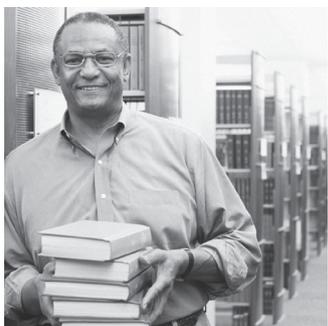
- Psychology: Understanding Happiness
- Statistics: Visualizing Data
- Government and Politics: Cyber-Influence and Power
- Religion: Understanding Islam
- Linguistics: Exploring the Evolution of Language
- Urban Design: The Ecology of Cities
- Marine Biology: Deep Oceans
- Media & Journalism: Covering World News
- Design+Engineering: Ingenuity in the Developing World
- Education: Creative Problem-Solving
- Environmental Studies: Climate Change
- Physics: The Edge of Knowledge
- Medicine: Rethinking Cancer
- Visual Arts: Mastering Tech-Artistry
- Evolution: Excavating Origins
- Design Thinking: Reimagine the Designer
- Applied Mathematics: Exploring the Geometry of Form
- Sustainable Consumption: Reworking the Western Diet
- Management: Leading Wisely
- Neuroscience: Mapping and Manipulating the Brain
- Computer Science: Exploring Robotics

www.wiley.com/go/tedstudies



WILEY

12-50283



Decision Sciences Journal of Innovative Education

"Their support is strong, continuous, enthusiastic, prompt, and supports the long-term growth of the journal and the institute."

**Chetan S. Sankar,
Editor**

Institute of Food Technologists

"By combining excellent customer service and a strategic view of the publishing market, they have enabled our scientific society to meet our goals..."

**Jerry Bowman,
Vice President of
Communication**

Veterinary Clinical Pathology

"I continue to be impressed with both the day-to-day management, including careful attention to quality issues, and long-term planning. We look forward to continuing our relationship for many years..."

**Karen M. Young,
Editor-in-Chief**

WILEY

exchanges.wiley.com/societies

PRODUCTION AND OPERATIONS MANAGEMENT SOCIETY

Officers and Board for 2018

President

J. George Shanthikumar
Purdue University, West Lafayette,
IN, USA

President—Elect

Nada Sanders
Northeastern University, Boston,
USA

Past Presidents

Manoj Malhotra
University of South Carolina,
Columbia, SC, USA

Asoo J. Vakharia
University of Florida, USA

Vice President—Finance

Shailesh Kulkarni
University of North Texas, Texas,
USA

Vice President—Education

Bala Shetty
Texas A&M University,
College Station, Texas, USA

Vice President—Meetings

Jerry Burke
Georgia Southern University,
Statesboro, Georgia, USA

Vice President—Member Activities

Mark Ferguson
University of South Carolina, USA

Vice President—Publications

Chelliah Sriskandarajah
Texas A&M University,
College Station, Texas, USA

Vice President—Industry

Russell Allgor
Chief Scientist, Worldwide Operations
and Amazon Logistics, USA

Vice President—Colleges

Xiuli He
University of North Carolina,
Charlotte, North Carolina, USA

Vice President—Communications

Henrique Correa
Rollins College, USA

Secretary

Haoying Sun
University of Kentucky, USA

Regional Vice President—Americas

Hugo Yoshijaki
University of São Paulo (USP), Brazil

Regional Vice President—Europe

Amit Eynan
University of Richmond, Richmond,
Virginia, USA

Regional Vice President—Africa and Middle East

Dino Petraolo
SVP Competitive Capabilities
International, Johannesburg,
South Africa

Regional Vice President—Australasia

Janny M.Y. Leung
The Chinese University of Hong
Kong, Hong Kong, China

Board Members

Saif Benjaafar
University of Minnesota, USA

Annabelle Feng
Purdue University, USA

Norma Harrison
Macquarie University, Sydney,
Australia

Jack Kanet
University of Dayton, Dayton, Ohio,
USA

Gal Raz
Western University, Ontario, Canada

Founder & Editor-in-Chief

Production and Operations Management

Kalyan Singhal
University of Baltimore, USA

Director of Strategic Planning

Martin K. Starr
Rollins College, USA

Executive Director

Sushil K. Gupta
Florida International University, USA

Associate Executive Director for Global Initiatives and Outreach

Nagesh Murthy
University of Oregon, USA

Associate Executive Director of POMS

Information Technology Services

Subodha Kumar
Temple University, USA

The mission of *Production and Operations Management* is to serve as the flagship research journal in operations management in manufacturing and services.

The journal publishes scientific research into the problems, interests, and concerns of managers who manage product and process design, operations, and supply chains. It covers all topics in product and process design, operations, and supply chain management and welcomes papers using any research paradigm.

For additional information, please visit our website: www.poms.org.