Data analysis workflow

Before to become meaningful information and knowledge, data generally undergo repeated steps of being entered into tables or databases, extracted, transferred to other tables, edited, selected, transformed, summarized, and presented.

Workflow involves the entire process of data analysis including planning and documenting your work, cleaning data and creating new variables, producing and replicating analyses, presenting findings and archiving your work. Being able to reproduce the work presented or published is the cornerstone of any workflow. How projects are planned, work documented and results saved should anticipate the need to replicate

In the following section, three basic steps are detailed, constituting a data flow within which specific operations should be performed.

Steps	Activities
Preparing	• Data acquisition & extraction • Data examination • Data cleaning
Making sense	• Data transformation • Data exploration • Data analysis & interpretation
Presenting & sharing	• Data communication • Data preservation

Although there is a logical sequence to these steps, the dynamics of an effective workflow are flexible and highly dependent upon the specific projects. Some steps can occur in tandem and some can repeat more often than others. Preliminary results might uncover problems with variables that send you back to cleaning the data, thus starting the sequence again.

Within each of these steps, there are three tasks: planning/coordinating your work, organizing your materials, and documenting what you do.

A detailed workflow is available in annex.

A. Planning

Most of us spend too much time working and too little time planning. Before to start crunching numbers, draft a plan of what you want to do and assess your priorities. The general objective of planification is to maximize thinking about the data analysis and to minimize the time spent trying to interpret randomly selected data analyses.

A plan begin with broad considerations and goals for the entire project, anticipating the work that needs to be completed and thinking about how to complete these tasks most efficiently. Data analysis often involves side trips to deal with unavoidable problems and to explore unanticipated findings. A good plan keeps your work on track. Planning is invaluable, for instance in collaborative work where inadequate planning can lead to duplication of efforts, misunderstanding about access to data or working at cross purposes with one person undoing what someone else is doing. Also, the longer, the more diverse or the larger the project(s), the more important it is to plan.

Planning involves consideration such as:

- General goals and publishing plans: Broad objective of the research, outline of the end product, where will it be published, who is the audience?
- Scheduling: Timeline with target dates for completing key stages, meetings where to present preliminary or final results, intermediate reports, etc.
- Resources: Staff, computers, room, software, translation, etc.
- Size and duration: Size and duration have implications for how much details and structure is needed. A survey report takes more planning and organization than a brief or a series of articles.
- Dataset: What data and variables will be used? How will you collect and integrate data over a period of time? Anticipating the complexity of the dataset can prevent initial decisions that later cause problems. What types of missing data will be encountered, how will they be treated?

- Division of labour: Working in a group requires special considerations. Who is responsible for which tasks? Who coordinate data management? If multiple person have access to the data, how do you ensures that only one person can change it at a time? If the analysis begins while data collection continues, how do you make sure that people are working with the latest version of the data? Who handle back up and keeps the documentation up to date?
- Ownership: Data ownership needs to be clarified up front when other collaborators are involved. What agreements do team members have about joint ownership of the data?
- Analysis: what types of statistical analyses are anticipated? What software is needed, and is it locally available? Thinking about software helps you plan data formats, naming conventions and data structures.
- Documentation: What documentation is needed? Who will keep it? In what format? A plan for how to document the project makes it more likely that things will be documented.
- Backing up and archiving: Who is going to make regular backups of the files? Long term preservation should also be considered. If the research is funded, what requirements does the funding agency have for archiving the data? What sort of documentation do they expect and what data formats?

As the work progress, periodically reassess the plan by refining your goals and analytic strategy based on the output of the work already completed.

B. Organization

If you have good documentation about what you did, but you cannot find the files used to do the work, little is gained. Organization requires you to think systematically about what goes where, what to name it and how you will find it later. Careful organization helps you work faster: you can easily find things and avoid duplication of effort.

Care in naming files and organizing their location is the key to keeping track of your files. The easiest approach in organizing project files is to start with a carefully designed directory structure. When files are created, place them in the appropriate directory. Like this, a file's location become an integral part of your documentation. Keep in mind that if you don't use your directory structure consistently, you will only make things worse. All the work related to a project should be contained within a single directory such as:

“Name”	Project name
Administration	Correspondence, budget, etc.
Documentation	Research log, codebooks, others
Posted	Completed text, datasets, log files
Readings	PDF files, articles, etc.
Work	Text and analysis that are being worked on

The folder “Work” and “Posted” are critical to the workflow. The directory “Work” holds work in progress. When a draft is ready to be shared, it can be moved to the folder “Posted”. Results are only shared after the associated files are posted. Once a file is posted, it is never changed.

When organizing your work:

- Start early: The more organized you are at the beginning of the project, the more you will be at the end. Organization is contagious.
- Simple but not too simple: More elaborate structures are not always better and can even make things harder for small projects.
- Consistency: If you use the same structure for all your projects, you will spend less time thinking about organization because you can take advantage of what you already know. If you organize different projects differently, you are bound to confuse yourself and spend time looking for things.
- Can you find it? Always keep in mind how you will find things. Software that search for files help, but work better if you plan your file naming and content so they work more effectively.
- Document your organization: You are more likely to stay organized if you document your procedures, specifically when collaborating, written procedures are essential.

C. Documentation

Documentation boils down to keeping track of what you have done and thought. It reminds you of decisions made, work completed and plans for future work. Unfortunately, it is more compelling to discover new things by analysing your data than it is to document where you stored a file or how you handled missing data. However, writing documentation can literally save weeks of work and frustration later. If you do not document your work, many of the advantages of planning and organizing are lost.

To ensure that you keep up with documentation, you need to include it as a regular part of your workflow, for instance by linking it to the completion of key steps in the project. You need both to document metadata (data about data) and process metadata (data about the process used to create, manipulate and analyse data). The ultimate criterion for whether something should be documented is whether it is necessary to replicate your findings. This includes:

- Data sources: If you are using secondary sources, keep track of where you got the data and which release of the data you are using.
- Data decisions: How were variables created and cases selected? Who did the work? When was it done? What coding decisions were made and why? For critical decisions, also document why you decided not to do something.
- Statistical analysis: What steps were taken in the statistical analysis, in what order, and what guided those analyses?
- Software version: Your choice of software can affect your results. This is particularly true with recent statistical packages using different algorithms, leading to different results.
- Storage: Where are the results archived? When you complete a project or put it aside to work on other projects, keep a record of where you are storing the files and other materials.
- Ideas and plan: Ideas for future research and list of tasks to be completed should be included in the documentation. What seems an obvious idea for future analysis today might be forgotten later.

Documentation includes:

- A research log that records what you. It chronicles the ideas underlying the project, the work you have done, the decisions made, and the reasoning behind each step in data construction and statistical analysis. It includes date when work was completed, who did the work, what files were used and where the materials are located. A good research log keeps your work on track by including your research plan, helps you deal with interruptions and facilitate replication.
- A codebook that documents the datasets you create and the variable they contain. It summarizes information on the variables of your dataset. It reflects the final decisions made in collecting and constructing variables, whereas the research log details the steps taken and computer programs used to implement these decisions.

If you have many datasets, you might want a registry of datasets. They will help you find a particular dataset and can help ensure that you are working with the latest data.

Key principles for writing good documentation include:

- Do it today: When things are fresh in your mind, you can write documentation faster and more accurately.
- Check it later: If you write documentation while doing the work, it is often easy to forget information that is obvious now but that should be recorded for future reference. Ideally, write your documentation soon after the work has been completed, then either have someone check the documentation of check it yourself at a later time.
- Know where to store documentation: If you cannot find it, it does not do you any good.
- Include full dates and names: When it comes to date, even the year is important. Include full names as initial might lose clarity in the future.

At the completion of each stage of the work, take the necessary time to review what was done, to check that the procedures were documented and to confirm that the materials are archived.

D. Execution

The following section details the three basic steps and the tasks required within each to get the most out of your data: Preparing, making sense, presenting and sharing your data.

1. Preparing the data

Preparing the data is very likely to be the most time and labour intensive aspect of your project.

Data acquisition and extraction describes gaining access to data, either by generating fresh data (e.g. field assessment or survey, phone interviews, etc.), or through acquiring data from others. Data can be obtained in different format depending on the source: Website, HTML format, PDF, (good luck), XLS, XML, Csv, Spreadsheets, etc.

In the extraction stage, data is converted from whatever input format has been acquired into a form that can be used for further processing and analysis. This often involves loading data into a database system (Excel, google spreadsheets, MySQL, etc.).

Data examination: You need to know your source material to tell good stories about data. Examination activities entail to:

- Learn about the numbers and metrics. Data is only as good as its source. Figure out where they came from, if it is reliable and how they were estimated, see if they even make sense.
- Examine completeness and usability of the data. *Does it fit purpose? Is it all there or do you need more? Is the size and shape consistent with your expectations? Does it have all the categories you were expecting? Does it cover the time period you wanted? Are all the fields or variables included? Does it contain the expected number of records?*
- Determine if you need additional datasets or information to complete or refine your message.
- Determine physical properties, architecture and hierarchy within your data. Identify data types (categorical nominal or ordinal, quantitative interval or ratio).

Data cleaning: Data collected during assessment, either secondary or primary, needs always to be checked for quality. Care should be taken to ensure data is as accurate and consistent (i.e. spellings, to allow aggregation) as possible. Inconsistencies can wreak havoc when trying to perform analysis on the data, so they have to be addressed before starting any analysis.

Data cleaning is the process of detecting, correcting, replacing, modifying or removing incomplete, incorrect, irrelevant, corrupt or inaccurate records from a record set, table, or database. It involves repeated cycles of screening, diagnosing, and treatment.

Screening involves systematically looking for suspect features in assessment questionnaires, databases, or analysis datasets (in small assessments, with the analysts closely involved at all stages, there may be little or no distinction between a database and an analysis dataset).

The diagnostic (identifying the conditions of the suspect data) and treatment (deleting, editing or leaving the data as it is) phases of data cleaning require insight into the sources and types of errors at all stages of the assessment.

2. Making sense of the data

Data transformation, exploration and analysis activities are not happening in a linear way, but are rather following an iterative process that potentially can go on endlessly. This back and forth tasks imply shifting focus and activities as new questions arise and new insight are discovered.

Data transformation (or re-expression) refers to the application of new calculation or mathematical operation to one or several observations (data points) in a data set, creating new values. It implies filtering and editing the data to keep it consistent and relevant for analysis.

There are an infinite number of possible transformations. The most common are:

- Parsing (split up) any variables, such as extracting month from a date value.

- Merging variables to form new ones, i.e. creating a whole name out of title, forename, and surname.
- Converting qualitative data/free-text into coded values or keywords.
- Deriving new values out of others, such as gender from title, overweight from the values for height and weight. This includes simplifying individual variables (collapsing 6 possible values to a smaller number such as index) and deriving compound variables (e.g. “socioeconomic status” derived from education and occupation.).
- Standardizing or normalizing data (ratio, %, average, etc.).
- Creating calculations for use in analysis, such as percentage proportions.
- Removing redundant data for which you have no planned use (Do a backup first!).

Data transformation should be used with care and never unless there is a clear reason as it can alter the fundamental nature of the data (i.e. such as changing the measurement scale from interval or ratio to ordinal), complicating analysis and interpretation.

Data exploration: Most analysts don’t just wrangle data for fun. They have a story to tell or a problem to solve. Often you will start with a question in mind, defined in your analysis plan. This could be anything from: ‘Do people feel secure in this area?’ to ‘what copying mechanisms are being used by the population? A question is a good starting point for inquiring your data – it makes you focused and helps you to detect interesting patterns in the data. Understanding for whom your question is interesting will also help you to define the audience you need to work for, and will help you to shape your story.

What if you start without a question? You’re just exploring to get a “feel” for the data. You explore your data by searching for unanticipated trends and anomalies in order to gain understanding and ideas. If you find something that looks interesting in your data, you can start examining it as if this was the question you had in mind. Sometimes patterns in data can be explained by investigating what causes the patterns. This is often a story worth telling.

Whether you began with a question or not, you should always keep your eyes open for unexpected patterns/results, or anything that surprises you. Often, the most interesting stories aren’t the ones you were looking for. Use inductive/deductive reasoning to find stories in your data. Look for similarities (comparing) and differences (contrasting) to detect patterns, trends and relationships in the data.

The exploration process involve a high level of interaction with the data, using manipulation such as sorting, filtering, restricting, appending, adding or removing variable, aggregating or disaggregating, re-expressing, highlighting, zooming and panning, re-scaling, annotating, re-visualizing and grouping. When manipulating the data, look especially for:

Comparisons and proportions

- Range and distribution: Discovering the range of values and the shape of their distribution within each variable and across combinations of variables.
- Ranking: Learning about the order of data in terms of general magnitude, identifying the big, medium, and small values.
- Context: Judging values against the context of averages, standard deviations, targets, and forecasts.
- Part of the whole: Determining the composition, distribution, and relevance of the data's categories and subcategories.

Relationship and connections

- Associations: Can we identify any important connections between different combinations of variables or values?
- Exceptions: Can we identify any significant values that sit outside of the norm, such as outliers that change the dynamics of a given variable's range?
- Correlations: Is there evidence of strong or weak correlations between variable combinations?
- Clusters and gaps: Where is there evidence of data being "bunched"? Where are there gaps in values and data points?

Patterns and trends

- Direction: Are values changing in an upward, downward, or flat motion?
- Rate of change: How steep or flat do pattern changes occur? Do we see a consistent, linear pattern, or is it much more exponential in shape?
- Fluctuation: Do we see evidence of consistent patterns or is there significant fluctuation? Maybe there is a certain rhythm, such as seasonality, or perhaps patterns are more random.
- Intersections: Do we observe any important intersections or overlaps between variables, crossover points that indicate a significant change in relationship?
- Importance and significance: Can we determine if the patterns we see are meaningful signals of practical value or simply represent the noise within the data?

Data Analysis and interpretation: Data rarely speak for themselves. Data analysis and interpretation can be defined as what we do to make sense of information, an activity which always demands experience and human thinking. The focus of analysis is always related to the research objectives and main questions of your study.

While data exploration involves the uncovering of patterns, trends and relationship in data, the analysis and interpretation phases entail explaining those patterns and trends through contextualisation and use of experience.

Comparison is at the core of the analysis process. It is used to identify how things are alike and how they are different. This requires thinking about specific attributes or characteristics of the data (male vs. female, rural vs. urban, etc.) and uses these characteristics as the basis for identifying meaningful similarities (correspondence) or differences (contrast).

Interpretation is the process of attaching meaning to data, determining why a particular pattern or trend is there and what the reason behind it is. Deriving this entails outlining alternative explanations and suggesting which conclusions can or cannot be drawn. This calls for identifying *why* data indicates a particular condition for one group and not another and *why* people behave as they do. During interpretation, evidence supporting the claims must be weighed and findings contextualized before they acquire real meaning and practical value to decision makers.

The most interesting output of analysis is certainly the exploration of possible futures. Anticipating, developing scenarios or forecasting are activities requiring strong analytical skills and the ability to project alternative ways in which a situation might evolve, speculate about an uncertain future and envisaging different possible outcomes for a given initial situation.

3. Presenting and sharing

Communicating using data requires as much attention as the calculations and manipulations that went into your analysis. This requires balancing respect for the data, honesty about the limitations of your analysis and care for the end user and the decisions he might have to take based on your analysis.

Communicating your data involves:

Identify your message and plan: Not everything in the data should always be presented or communicated. Only those numbers which are new, important or interesting will ultimately be of interest to decision makers. This entails defining clearly the message and the way the data will be used to support the main storyline. Building a data narrative and using storytelling techniques maximise the potential insight within the data. Effective statistical narratives are generally simple, seamless, informative, true, contextual, familiar, concrete, personal, emotional, actionable and sequential.

Knowing your audience will influence how the data will be presented. Thinking about what your audience cares about allow for expressing your results in those terms (e.g. funds for donors, resources for program officers, etc.). In addition, knowing the decision making process and determining how data can best support it helps shaping the message and to provide decision makers with just enough information to allow choosing between options.

As mentioned earlier, sharing documentation is important for giving credit where credits is due, leaving a trail for you to remember (i.e. one or two years later), creating a trail for others to follow (i.e. one or two years later) and checking your work ((i.e. one or two years later). Information is required both on the data and on the methods by which they were collected and analysed, so that any competent analyst can recreate the work.

Communicate uncertainty: Limitations in analysis will emerge from the interpretation phase and should be reported, either in written form (be explicit and honest about limitations) in the final report or in verbal presentations (be prepared to discuss limitations).

Writing about numbers is an essential skill for data analysts who are due to present and communicate trends, patterns and numerical concepts. Even for works that are not inherently quantitative, one or two numeric facts can help convey the importance or context of your topic. Numbers can be incorporated in several different ways: a few carefully chosen facts in an executive summary or a presentation, a table in the analytic section of your report, a chart of trends in the slides for a speech, etc. In each of these contexts, the numbers support other aspects of the written work. The results of the calculations are used to answer some real-world question, generate interest in the topic or provide evidence for supporting a claim or a conclusion.

In many ways, writing about numbers is similar to other kinds of analytical writing: it should be clear, concise, and written in a logical order. It should start by stating an idea or proposition, then provide evidence to support that thesis. It should include examples that the expected audience can relate to and descriptive language that enhances their understanding of how the evidence relates to the question. It should be written at a level of detail that is consistent with its expected use. It should set the context and define terms the audience might not be expected to know, but do so in ways that distract as little as possible from the main thrust of the work.

Data visualization: Tables and graphs are used regularly to communicate quantitative information, and they have become common place since the advent of spread sheet software in the 1970s. When they are well-designed and used properly, they are fantastic tools. But unfortunately, this is not always the case and they often fall far short of their potential. Poor design choices can hide important data and clutter valuable information behind the noise.

Designing appropriate charts demands as much attention as ensuring the validity of the collected data. Understanding design principles, appropriate use of colour, when to use which type or chart or another is key to communicating data effectively.

Data preservation: Once data has been communicated, it must be preserved and archived for future use and access. You want to prevent the loss of files that you are actively working on, to maintain files from completed work that you might need later, and to preserve critical data and analyses for future use. Files can be lost because they are deleted, corrupted, your computer suffer hardware failure, or media or formats become obsolete.

Planning for protection implies balancing ease of use and degree of protection. Three levels of protection exists:

Short term protection focuses on making sure that the files you are using today will be there tomorrow. It involves continuously making duplicate copies of your files, referred to a mirroring. This protects against the failure of your hard drive, a computer virus or accidental deletion. This concern documents, log files, datasets, and other document that are part of ongoing analysis.

Mid-term protection involves protecting files that you have finished working on but that you might want to use in the years to come. This is a backup, where you are not concerned with preserving the files beyond your own interest in them. This concerns posted files that have been completed and will no longer change.

Long term protection (archiving) seeks to maintain the information in perpetuity. Archiving is extremely difficult, requiring constant concern with migrating files to new media and formats and the availability of documentation that is accessible to anyone interested in the files. It is not necessary to archive all the data products generated. Consider which data will be the most useful for future data users and which data version would be the most difficult to reproduce. This implies datasets, codebooks, articles, final reports, research logs, data sources, etc. Consider also legal and policy considerations that can affect the use and reuse of your data. It is important to provide the most comprehensive access possible with the fewest barriers or restrictions. Consider your own organization policy regarding data sharing, copyright issues, and the manner on which you licence your data (i.e. the Creative Commons Zero Licence provides for very broad access).

Key rules for copying files include:

- Make at least two copies on top of the original of all files.
- Store the copies at different locations
- Verify that copies are exact duplicates

E. References

Few 2009, Statistical Narrative. Telling Compelling Stories with Numbers

Jonathan G. Koomey. 2006. Best practices for understanding quantitative data

Jonathan G. Koomey. 2004. Practical problem solving: A beginner's guide

Data visualization: a successful design process by Andy Kirk, 2012

Data management guide for public participation in scientific research, 2013. http://www.dataone.org/sites/all/documents/DataONE-PPSR-DataManagementGuide.pdf

The Chicago guide to writing about numbers by Jane E. Miller, 2004

Now You See It, Stephen Few, 2008

The Wall Street Journal: Guide to Information Graphics, Dona M. Wong, 2010

Show Me the Numbers, Stephen Few, 2006

Data Visualization: a successful design process, Andy Kirk, 2012

Juiceanalytics, http://www.juiceanalytics.com/

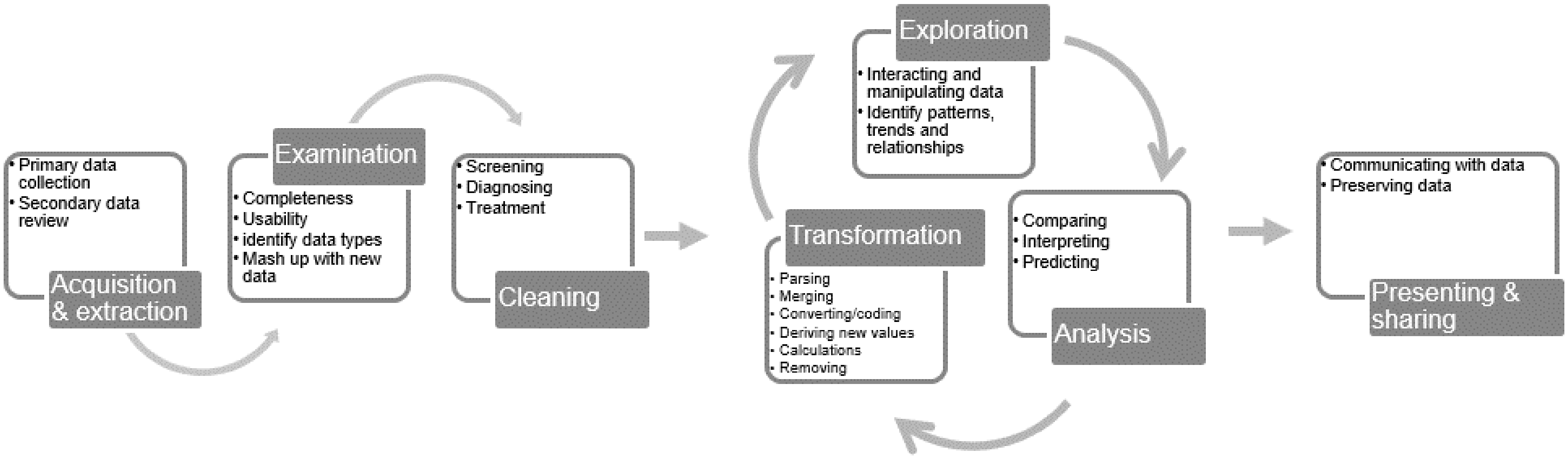
The Functional Art: An introduction to information graphics and visualization, Alberto Cairo, 2012

http://oak.ucc.nau.edu/rh232/courses/EPS625/Handouts/Data%20Transformation%20Handout.pdf

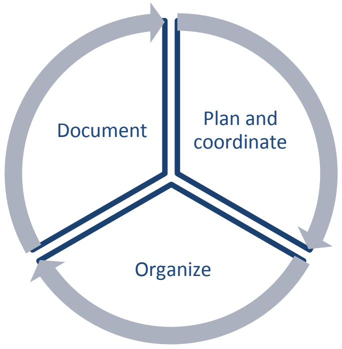
Jason W. Osborne, Notes on the use of trasnformation, http://pareonline.net/getvn.asp?v=8&n=6

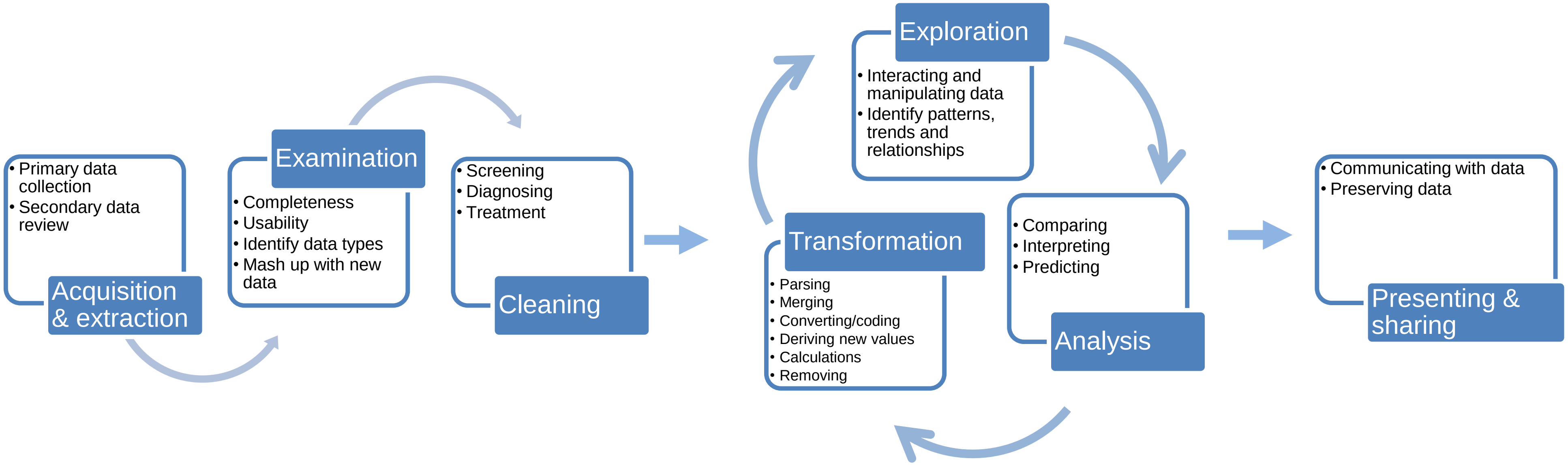
J. Scott Long, The workflow of data analysis using Stata, 2009

Annex - The data analysis workflow (Kirk 2012, ACAPS 2013):



And within each of the above steps:




Acquisition & extraction
• Primary data collection
• Secondary data review
Examination
• Completeness
• Usability
• Identify data types
• Mash up with new data
Cleaning
• Screening
• Diagnosing
• Treatment
Exploration
• Interacting and manipulating data
• Identify patterns, trends and relationships
Analysis
• Comparing
• Interpreting
• Predicting
Transformation
• Parsing
• Merging
• Converting/coding
• Deriving new values
• Calculations
• Removing
Presenting & sharing
• Communicating with data
• Preserving data