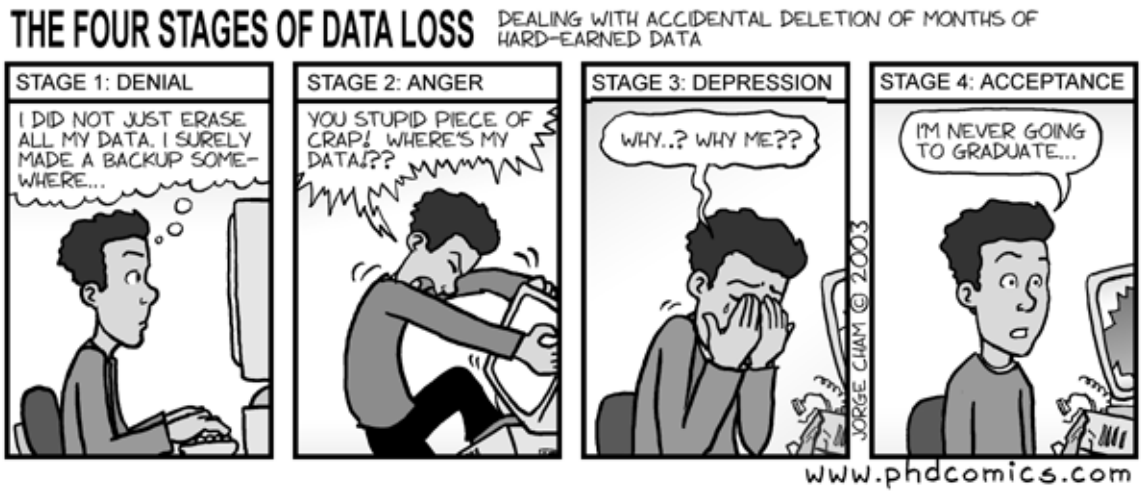
Josh Clinton
Vanderbilt University

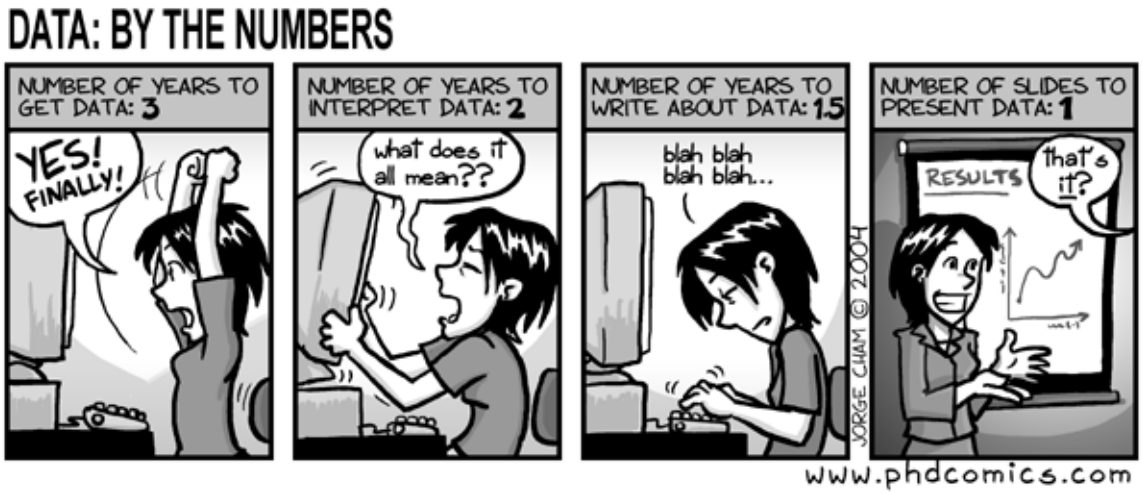
Advice on Computing Issues:

Motivation: don't let this happen to you…



How we deal with data and files is one of the most essential aspects of the job. It is also one on which there is no formal instruction and an area that is almost entirely behind the scenes. We are in the business of asking and answering questions, but to do so requires an inordinate amount of care and attention to aspects that have absolutely nothing to do with political science and which people almost never observe (unless they request your data).

The following is an entirely apt description of the inputs and outputs:



Here is some advice gleaned from my past experiences and mistakes:

Your goal should also be motivated by the fear that in five years when you are working on more interesting projects, teaching new classes, and serving on committees you are asked for your data and code for the project. You want to help out your "Future Self" as much as possible by making good choices now.

GUI is evil! Never, ever, ever, ever, ever use it unless you copy and paste the code. It is not replicable, the GUI changes when the software updates, and you may mistakenly press the wrong button.

What should you do? Here you go:

1) Subscribe to *DropBox* (or some other remote file system): www.dropbox.com

Everything you are currently working on should be remotely located so you can access it from any computer. This ensures that you will always be working on the most recent copy (i.e., you won't have a proliferation of different files that you will then try to reconcile which is the most recent and you will also never face the unhappy situation where you have edited two different versions of the same paper and now you have to merge the edits together.

This is great for collaboration because you can also share the folder and give multiple people access to the folder.

All work should be done on DropBox.

2) Embrace the Folder System:

Folders are your friend. Within DropBox, you should have a folder for each project you are working on. Whether you label them according to the coauthors (e.g., "ClintonLewis") or the project (e.g., "AgencyIdealPoint") is secondary, but be consistent in your labeling. Don't use titles like"MPSA2012" because in 3 years you will have no idea what you presented at Midwest in 2012.

Make subfolders. Within every Folder, I have several subfolders. The most vital include:

1) *Paper* (containing the most current draft (and perhaps a folder within *Paper* called *Older Drafts* if I ever want to go back to older versions),
2) *Data* (obvious, but sometimes subfolders are also good),
3) *Code* (containing the code I used),
4) *Presentations* (folder consisting of the most current presentation (with the possibility of a folder of older presentations as well).

2.1) The Data Folder:

This folder should contain all the data required to produce every table in the paper, appendix and presentations. Data that was originally collected with the intention of being

used but was not should end up in an *Old Data* folder (don't destroy data – you never know when it may come in handy).

You should have a file list that describes what each file is and what it contains (e.g., where the data was collected from, when, etc.). This is your road map to the project data. You do not want to be in the situation where you have a bunch of excel spreadsheets containing data that you have no idea what they mean.

I like to have the complete data process: whereby I go from the original data set all the way to the data used in the paper. At the end of the process, you should have a dataset that contains only the actual data you used in the paper, presentation, or presentation. Including extraneous variables only confuses things (but have them available).

If the analyses are done at different levels, sometimes I will have data files such as "Chapter1Data" etc to help locate the role that the day play in the analysis.

2.2) The Code Folder:

Everything you do should be replicable by a first-year. Well-documented code is essential to this effort. Comments are your friend. Use them!

At the start of every Code should be something like:

```
#       Josh Clinton
#       Vanderbilt University
#       January 2012
#
#       Project: The Fox News Effect Reconsidered
#       Reading data on Fox News coverage from Excel *.csv file and reading it into R
#       Version: R .14.0
```

Your name, affiliation and date tell about you for those who request the code. The project tells what the file is being used for, a short description helps remind you what the code does, and the version reminds you what version the code works with in case future updates makes changes to the commands.

You should have code that takes the project from the initial data all the way to the tables and figures that you discuss. I prefer several scripts to do so to keep the scripts short and manageable (scrolling through pages of code is bad). I have one script to get the data and create the file I am going to analyze with the variables I need. I have another script to process the data – recode, create, label and manipulate the data. I then have at least one more script to analyze the data. Sometimes I have scripts by section (e.g., Section1Analysis), while other times I organize it by the end-product (e.g., PaperAnalysis, AppendixAnalysis, PresentationAnalysis).
When conducting analysis, you will take make false steps. I keep prior analysis because it is good to see how alternative specifications may affect things. However, I keep such in an

Old Code sub-folder because I do not want a multitude of analyses to confuse or clutter the code I really care about.

You should comment along the way. What are you doing at each step? The goal is to remind your Future Self of the decisions you are making and the reasons why.

I tell the code to save the figures and Tables to the appropriate *Paper* folder.

3) Folder Migration

When I am no longer working on a project either because it is published or else it is unpublishable, I move it to the *Published Papers* or *Unpublished Papers* folder on my desktop. (And I make sure to have a back-up system on my desktop!)

I have debated the efficacy of a clean folder – i.e., a folder containing just the paper, data, and code required to reproduce the figures and tables without any extraneous data for each paper that I could zip and make available, but I have not yet done this myself.

In general:

Replicability is becoming increasingly important. QJPS, AJPS both require it for publication.

Dataverse (at Harvard) collects data from other projects. (http://dvn.iq.harvard.edu/dvn/)

Sharing is influencing. Science is iterative.

Do it now so your Future Self doesn't hate you.