# IRS 990 e-File Data (8) – Data Wrangling and Export to Stata

---

# IRS 990 e-File Data (7) – Generate Codebook

---

# IRS 990 e-File Data (6) – Download IRS 990, 990EZ, and 990PF Filings and Associated Schedules into MongoDB

---

# IRS 990 e-File Data (part 5)

## – Download IRS 990 Filings and Associated Schedules into MongoDB

---

## IRS 990 e-File Data (part 4) – Download IRS 990 Filings into MongoDB

---

## IRS 990 e-File Data (part 3) – Load Index Files into PANDAS

---

## IRS 990 e-File Data (part 2)

# – Load Index Data and Insert into MongoDB

---

# Tutorials for Sarbanes-Oxley Paper Data

Dan Neely (from University of Milwaukee-Wisconsin) and I just had the following article published at the *Journal of Business Ethics*:

Saxton, G. D., & Neely, D. G. (2018). The Relationship Between Sarbanes–Oxley Policies and Donor Advisories in Nonprofit Organizations. *Journal of Business Ethics.*

This page contains tutorials on how to download the IRS 990 e-file data that was used for the control variables in our study.

## Tutorials

- IRS 990 e-File Data (part 1) – Set up AWS CLI credentials and grab index files
- IRS 990 e-File Data (part 2) – Load Index Data and Insert into MongoDB
- IRS 990 e-File Data (part 3) – Load Index Files into PANDAS
- IRS 990 e-File Data (part 4) – Download IRS 990 Filings into MongoDB
- IRS 990 e-File Data (part 5) – Download IRS 990 Filings

and Schedules into MongoDB
- IRS 990 e-File Data (part 6) – Download IRS 990, 990PF, and 990EZ Filings into MongoDB
- IRS 990 e-File Data (part 7) – Generate Data Codebook
- IRS 990 e-File Data (part 8) – Data Wrangling and Export to Stata

I hope you have found this helpful. If so, please spread the word, and happy coding!

---

# Analyzing Big Data with Python PANDAS

This is a series of iPython notebooks for analyzing Big Data – specifically Twitter data – using Python's powerful PANDAS (Python Data Analysis) library. Through these tutorials I'll walk you through how to analyze your raw social media data using a typical social science approach.

The target audience is those who are interested in covering key steps involved in taking a social media dataset and moving it through the stages needed to deliver a valuable research product. I'll show you how to import your data, aggregate tweets by organization and by time, how to analyze hashtags, how to create new variables, how to produce a summary statistics table for publication, how to analyze audience reaction (e.g., # of retweets) and, finally, how to run a logistic regression to test your hypotheses. Collectively, these tutorials cover essential steps needed to move from the data collection to the research product stage.

**Prerequisites**

I've put these tutorials in a GitHub repository called PANDAS. For these tutorials I am assuming you have already downloaded some data and are now ready to begin examining it. In the first notebook I will show you how to set up your ipython working environment and import the Twitter data we have downloaded. If you are new to Python, you may wish to go through a series of tutorials I have created in order.

If you want to skip the data download and just use the sample data, but don't yet have Python set up on your computer, you may wish to go through the tutorial "Setting up Your Computer to Use My Python Code".

Also note that we are using the iPython notebook interactive computing framework for running the code in this tutorial. If you're unfamiliar with this see this tutorial "Four Ways to Run your Code".

For a more general set of PANDAS notebook tutorials, I'd recommend this cookbook by Julia Evans. I also have a growing list of "recipes" that contains frequently used PANDAS commands.

As you may know from my other tutorials, I am a big fan of the free Anaconda version of Python 2.7. It contains all of the prerequisites you need and will save you a lot of headaches getting your system set up.

**Chapters:**

At the GitHub site you'll find the following chapters in the tutorial set:

Chapter 1 – Import Data, Select Cases and Variables, Save DataFrame.ipynb
Chapter 2 – Aggregating and Analyzing Data by Twitter Account.ipynb
Chapter 3 – Analyzing Twitter Data by Time Period.ipynb
Chapter 4 – Analyzing Hashtags.ipynb

Chapter 5 – Generating New Variables.ipynb
Chapter 6 – Producing a Summary Statistics Table for Publication.ipynb
Chapter 7 – Analyzing Audience Reaction on Twitter.ipynb
Chapter 8 – Running, Interpreting, and Outputting Logistic Regression.ipynb

I hope you find these tutorials helpful; please acknowledge the source in your own research papers if you've found them useful:

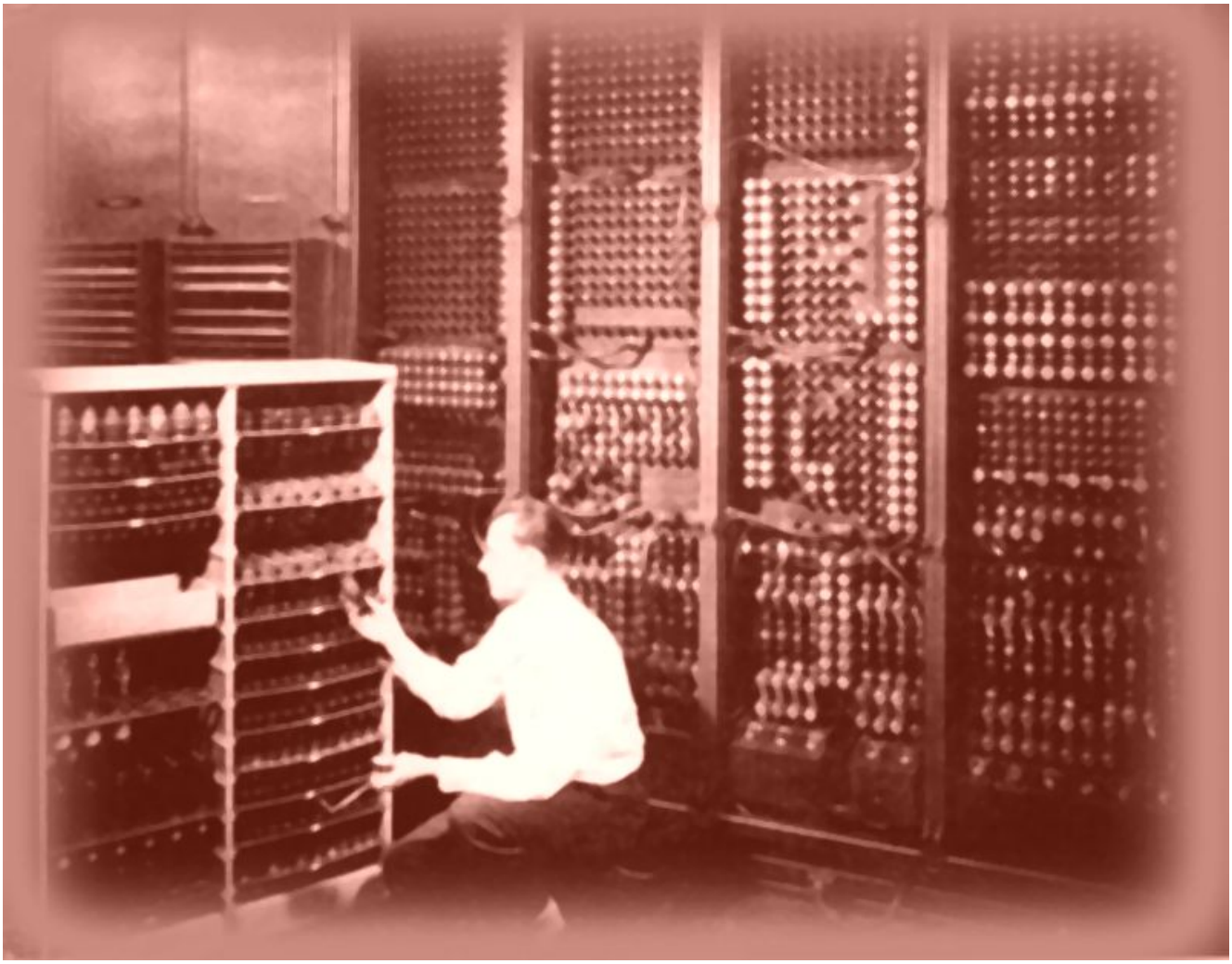
Saxton, Gregory D. (2015). *Analyzing Big Data with Python.* Buffalo, NY: http://social-metrics.org

Also, please share and spread the word to help build a vibrant community of PANDAS users.

Happy coding!

---

# Do I Need to Learn Programming to Download Big Data?



You want to download and analyze “Big Data” — such as messages or network data from Twitter or Facebook or Instagram. But you’ve never done it before, and you’re wondering, “Do I need to learn computer programming?” Here are some decision rules, laid out in the form of brief case studies.

## One-Shot Download with Limited Analysis

Let’s say you have one organization you’re interested in studying on Twitter and want to download all of its tweets. You are doing only basic analyses in a spreadsheet like Excel. In this case, if you have a PC, you can likely get away with something like NodeXL — an add-on to Excel. ***VERDICT: COMPUTER PROGRAMMING LIKELY NOT NECESSARY***

## One-Shot Download with Analysis in Other Software

Let's start with the same data needs as above: a one-shot download from one (or several) organizations on Twitter. You wish to undertake extensive analyses of the data but can rely on some other software to handle the heavy lifting – maybe a qualitative analysis tool such as ATLAS or statistical software such as SAS, R, or Stata. Each of those tools has its own programming capabilities, so if you're proficient in one of those tools – and your data-gathering needs are relatively straightforward – you *might* be able to get away with not learning programming. ***VERDICT: COMPUTER PROGRAMMING MAY BE UNNECESSARY***

## Anything Else

In almost any other situation, I would recommend learning a programming language. Why is this necessary? For one case, let's say you wish to download tweets for a given hashtag over the course of an event. In this case you'll want to use a database – even a simple database like SQLite – to avert duplicates from being downloaded. The programming language, meanwhile, helps you download the tweets and "talk" to the database. In short, if you are downloading tweets *more than once* for the same sample of organizations, you should probably jump to learning a programming language. Similarly, if you have any need at all for *manipulating the data* you download – merging, annotating, reformulating, adding new variables, collapsing by time or organization, etc. – then a programming language becomes highly desirable. Finally, if you have any interest in or need of medium- to advanced-level *analysis of the data*, then a programming language is similarly highly desirable. ***VERDICT: PICK A PROGRAMMING LANGUAGE AND LEARN IT***

## Conclusion

Not *everyone* needs to learn a programming language to accomplish their social media data downloading objectives. If your needs fall into one of the simple cases noted above then you may wish to skip it and focus on other things. On the other hand, if you are going to be doing data downloads again in the future, or if you have anything beyond basic downloading needs, or if you want to tap into sophisticated data manipulation and data analysis capabilities, then you should seriously consider learning to program.

Learning a programming language is a challenge. Of that there is little doubt. Yet the payoff in improved productivity alone can be substantial. Add to that the powerful analytical and data visualization capabilities that open up to the researcher who is skilled in a programming language. Lastly, leaving aside the buzzword "Big Data," programming opens up a world of new data found on websites, social media platforms, and online data repositories. I would thus go so far as to say that any researcher interested in social media is doing themselves a great disservice by not learning some programming. For this very reason, one of my goals on this site is to provide guidance to those who are interested in getting up and running on Python for conducting academic and social media research. If you are a beginner, I'd recommend you work through the tutorials listed here in order.