Causal Inference

Mixtape Session

Roadmap

- Hidden curriculum
  - Background
  - Empirical workflow
  - Hierarchical folder structure
  - Naming conventions
  - Version control
  - Soft skills

Empirical workflows require scarce time inputs

- All of us are living at the edge of our resource constraints, and our most scarce resources is *time*
- To do anything, we must sacrifice something else because all activities use time – including cleaning and arranging our data
- Just like running a marathon involves far far more time training than you ever spend running the marathon, doing empirical research involves far far more time doing tedious, repetitive tasks
- Since you do the tedious tasks repeatedly, they have the *most* potential for error which can be catastrophic ("anything that can happen will happen with enough trials")
- However long you think cleaning and organizing the data will take, **multiply it by 10** – prepare for it by managing your expectations

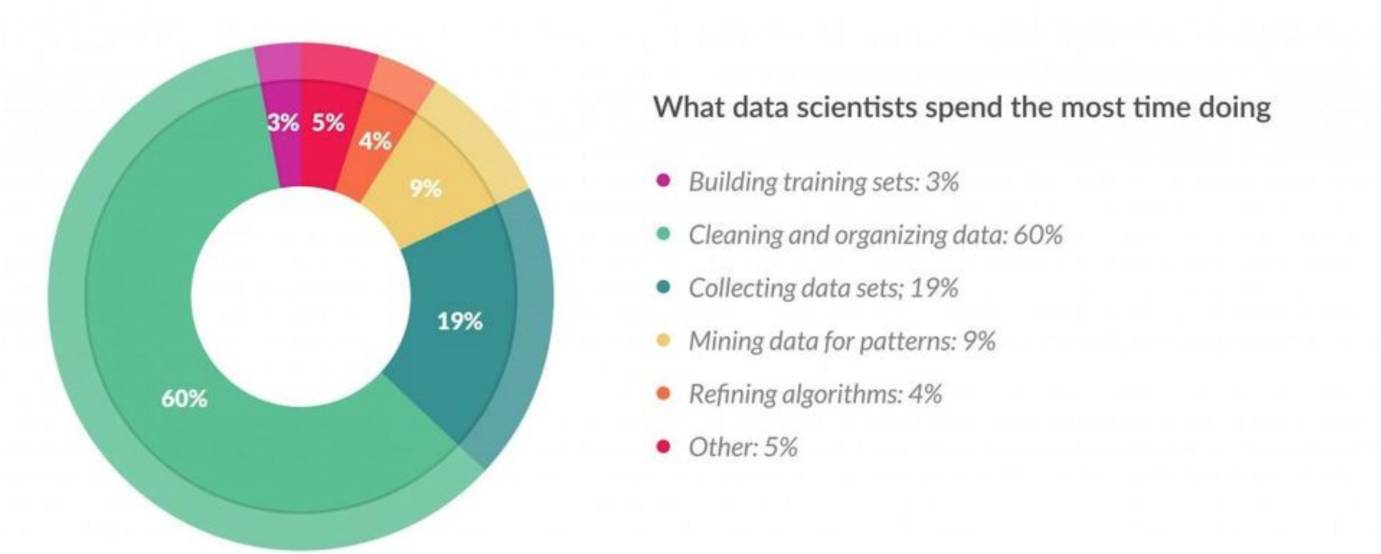

Figure: Image from Wenfei Xu at Columbia

Read the codebook

- Datasets often come with very large documents (either physical or digital) describing in detail the data production and arrangement
- You must become as much of an expert on the codebook as you are on your own research topic
- The codebook explains how to interpret the data you have acquired and it is not a step you can skip
- Set aside time to study it, and have it in a place where you can regularly return to it
- This goes for the `readme` that accompanies some datasets, too.