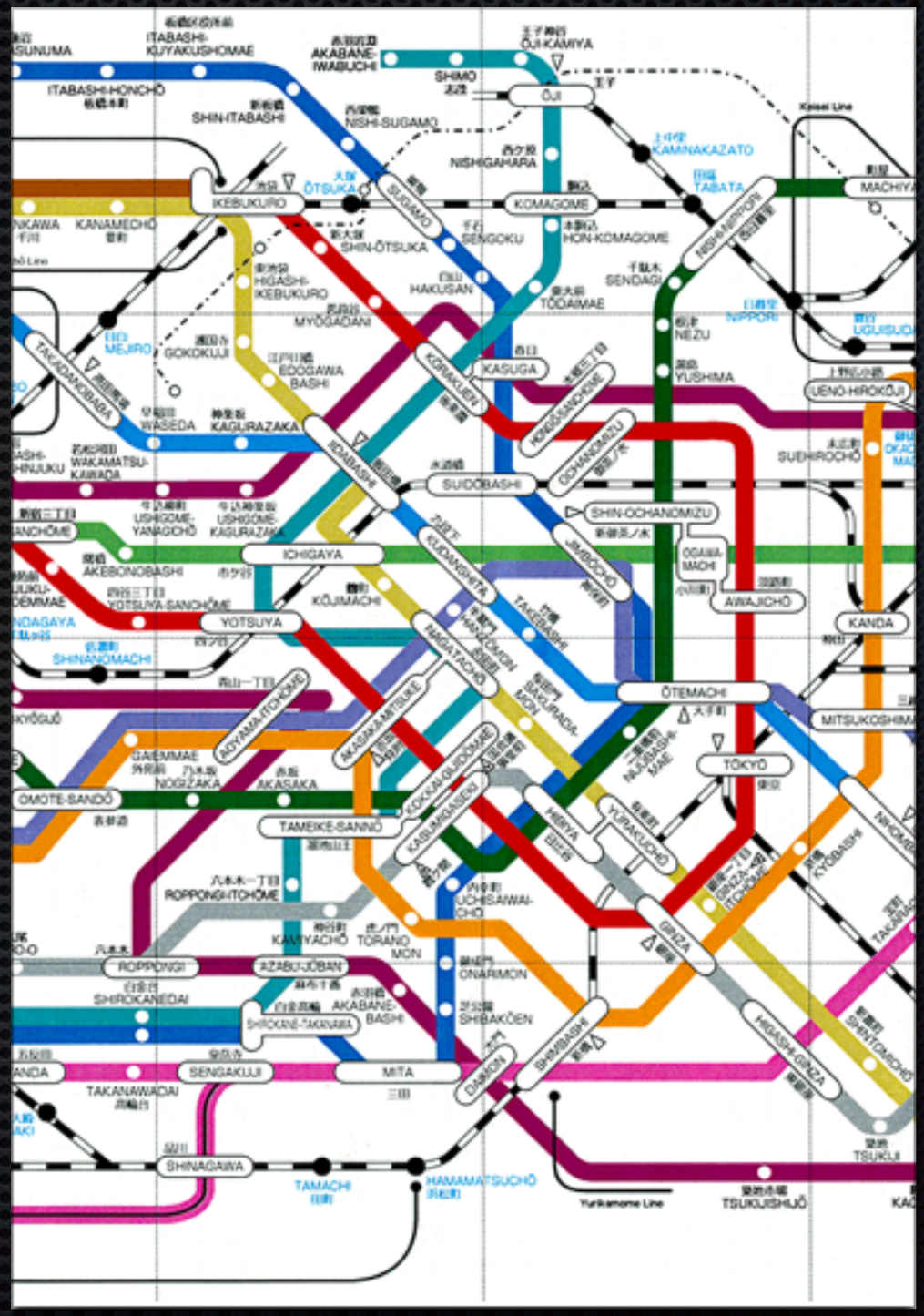
Data Workflows for Fun & Profit.

Hunter Owens (@hunter_owens),
Hack@UChicago Learnathon

Guiding Principle #1:

A Method to the Madness.



Getting Data

- Store raw data in $project/data if possible
- Use .gitignore to ignore this directory
- Put data notes in $project/README.md and/or $project/DataDictionary.md
- Use make/drake/luigi to script extraction build pipeline

Data Exploration Toolbelt

- Command Line Tools
- csvkit
- Pandas/IPython Notebook
  - Put in $project/notebooks
  - make sure each team member has separate directory

Guiding Principle #2

Take detailed notes of your data exploration. Make a mental map of the data.

Data Storage Options:

(Protip: You probably don't have big data)

- Disk
- Network Drive
- S3/Azure Blob/etc
- HDFS (Hadoop File System)

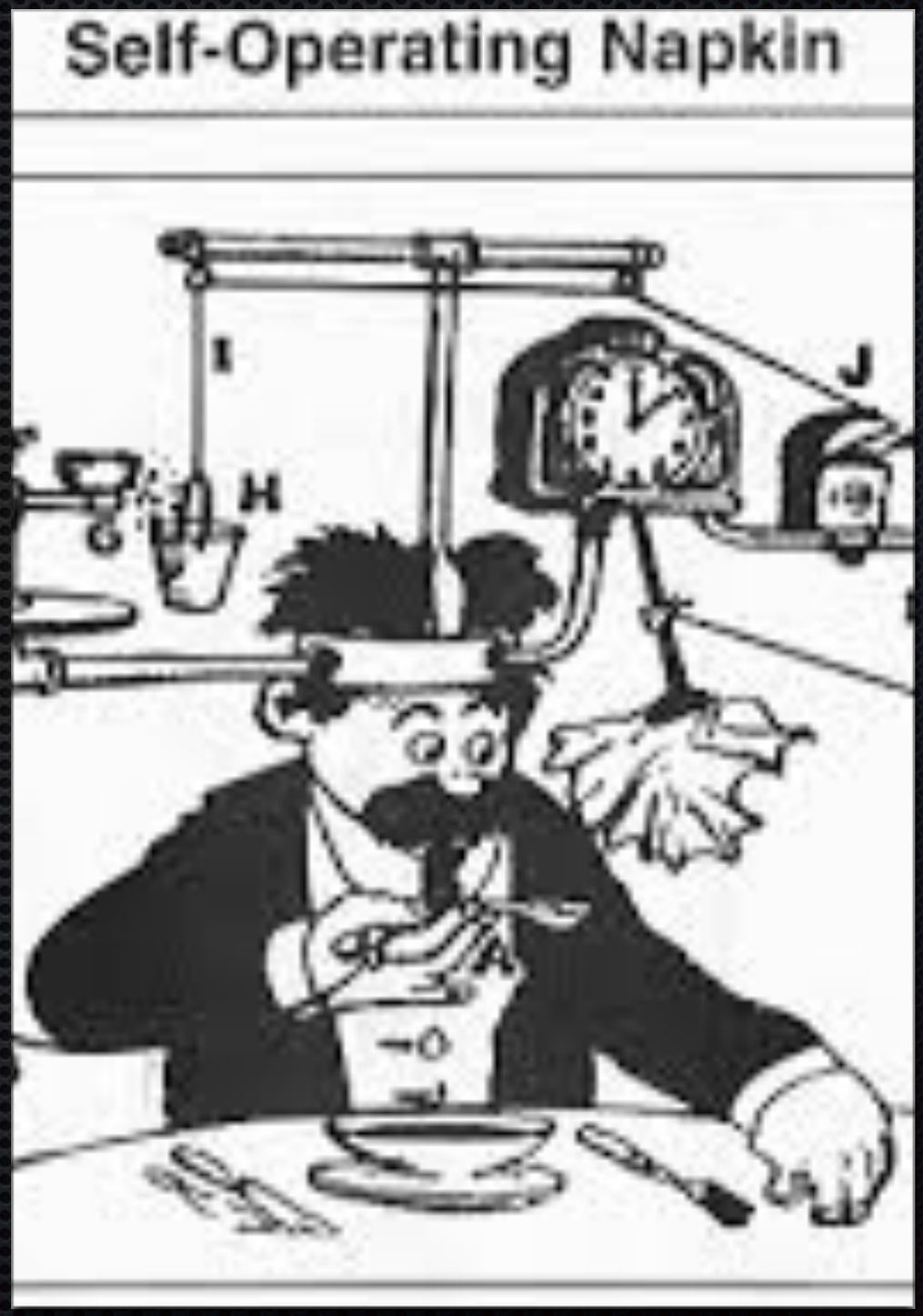
- Data Extraction is often a key step
  - Write where the data is from
  - Where it is extracted to
  - Store the entire raw dataset, if possible

Exploration Tools

- Command Line Tools - Use them
- Write Scripts (either .sh, .py or .r)
  - csvkit
  - pandas/ipython notebook
  - GDAL
  - Other, dataset specific tools (tilemill,QGIS for geographic data, etc)

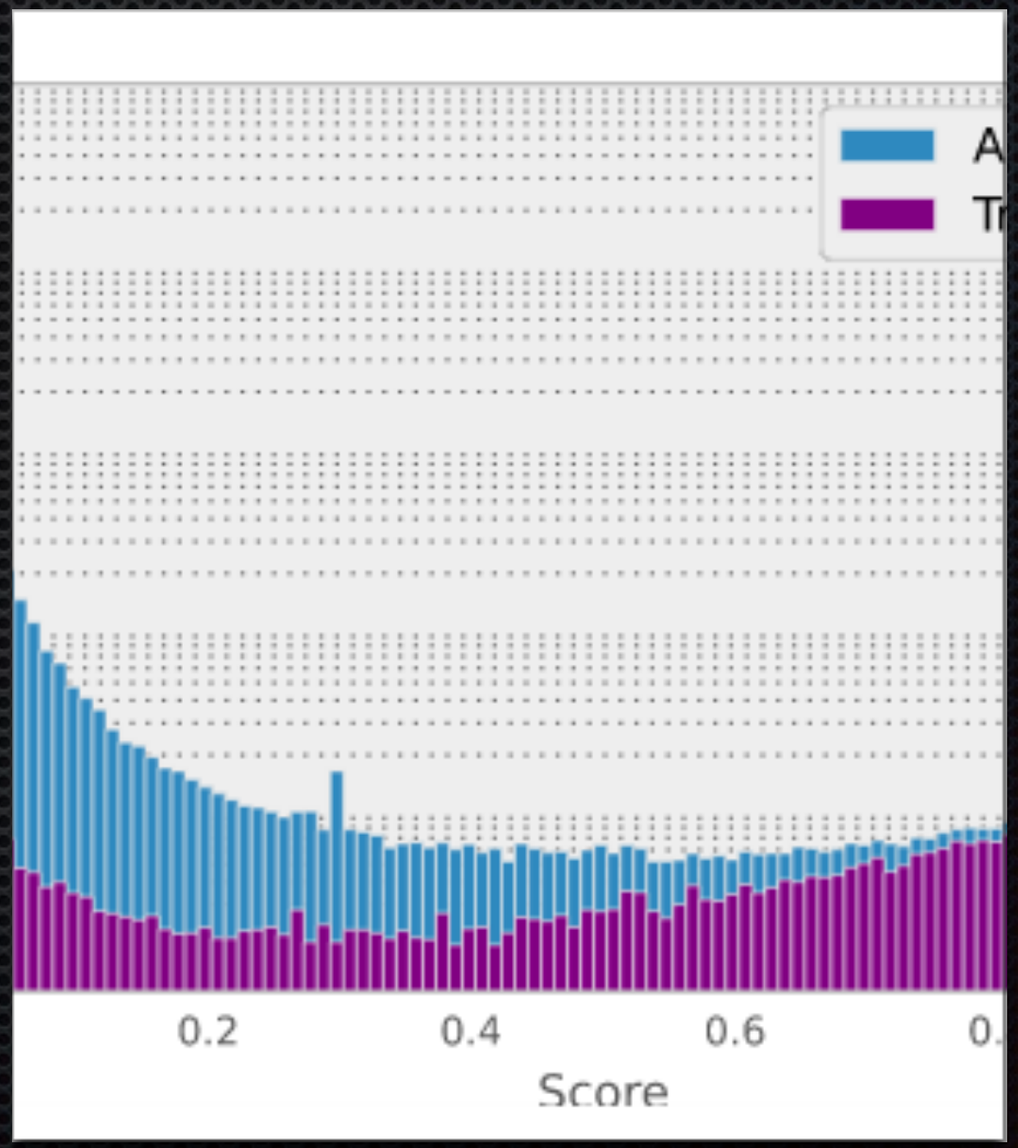
Guiding Principle #3

Automate all the things



Storing Model/Exploration Output

- Determine and find the place to store output for the long term
- Database, S3, etc.
- Store the output of all your models
- Ie, https://github.com/stripe/topmodel project
- Solving the how did I make that graph problem



Building Workflows

- Why set this up everytime?
- Use a template: Mine is http://github.com/dssg/project_template
- These things are opinionated- find one that works for you

Testing Assumptions

- Questions to ask after data cleaning, merging etc. (ie, tests)
  - Is the data of the correct size
  - Is it of the correct shape?
  - Fixing Data Horror Stories

Writing Data Tests

- Is hard
- Ensure that the data is meeting your goals
- Should be Complete, Correct and Connectable (thanks to Sasha Laundy for the phrase)

Thanks!

References

- https://www.youtube.com/watch?v=dOwmU-5ShJs
- https://18f.gsa.gov/2015/01/13/an-open-source-tool-for-easier-database-testing/
- https://github.com/dssg/project_template
- http://bost.ocks.org/mike/make/