

intro

Adam Okulicz-Kozaryn

`adam.okulicz.kozaryn@gmail.com`

this version: Sunday 16th April, 2023 14:05

outline

why data science?

stata v sas v r v python

python: fun and useful

[*] bonus—data sources [skim quickly; look at home more]

introductions

- <https://theaok.github.io> and my goog sch
- what data and vis i have been using
 - eg py map in livability paper, scatterplots
- if your res is similar, i can help more
 -
- you? (see if others overlap: collaborate!):
 - research interests
 - software
 - and data?

outline

why data science?

stata v sas v r v python

python: fun and useful

[*] bonus–data sources [skim quickly; look at home more]

data revolution!!

- ◇ most jobs/tasks require or benefit from programming
- ◇ qualitative data (pictures, text, etc.) are just rich quantitative data and can be analyzed like quantitative!
- everything can be quantified or not? any examples of non-quantifiable things ?

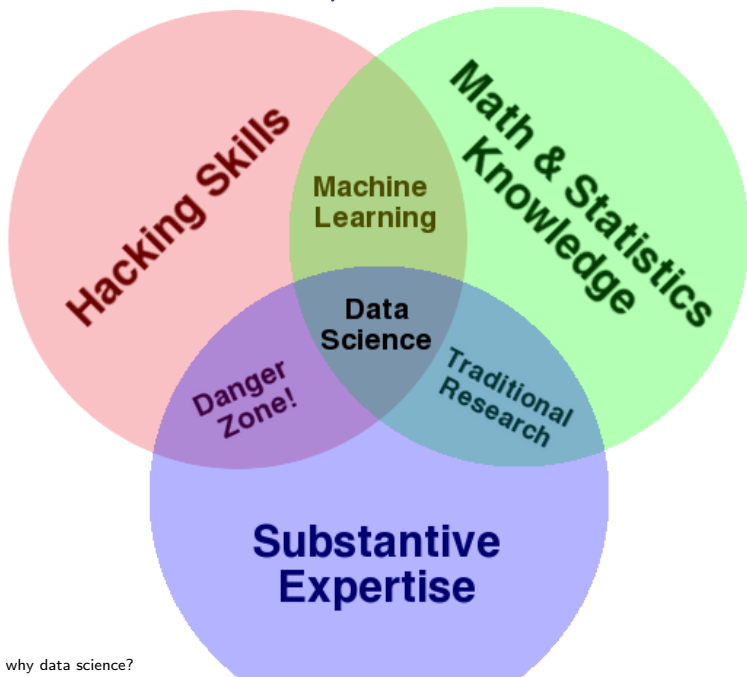
data management is fundamental

- ◇ takes more time to prepare data than to analyze it
- ◇ start early with the right data!
 - sth you're passionate about
 - sth that will advance your career/think beyond school

(social) data science or comp soc sci

- see venn diag next p
- <http://gking.harvard.edu/files/LazPenAda09.pdf>
- <http://radar.oreilly.com/2010/06/what-is-data-science.html>
- <http://tdwi.org/Articles/2011/01/05/Rise-of-Data-Science.aspx?Page=1>
- <http://www.quora.com/Educational-Resources/How-do-I-become-a-data-scientist>

already have stat/math and subst, need hacking!



dat sci skills

- <https://www.kdnuggets.com/2023/04/top-19-skills-need-know-2023-data-scientist.html>

outline

why data science?

stata v sas v r v python

python: fun and useful

[*] bonus–data sources [skim quickly; look at home more]

which one?

- stata: powerful, no need to learn any other soft; sufficient for vast majority of projects
- r: most powerful stat soft (py seems to be taking over)
- stata: user friendly, fast, very concise code
- r: user unfriendly, slow; weird code!
- py: somewhere in between
- all: great user community: listserv, websites, etc.
- sas: a dinosaur (still, often industry standard), very verbose
- r,py: free, stata: around \$200; sas over \$1k



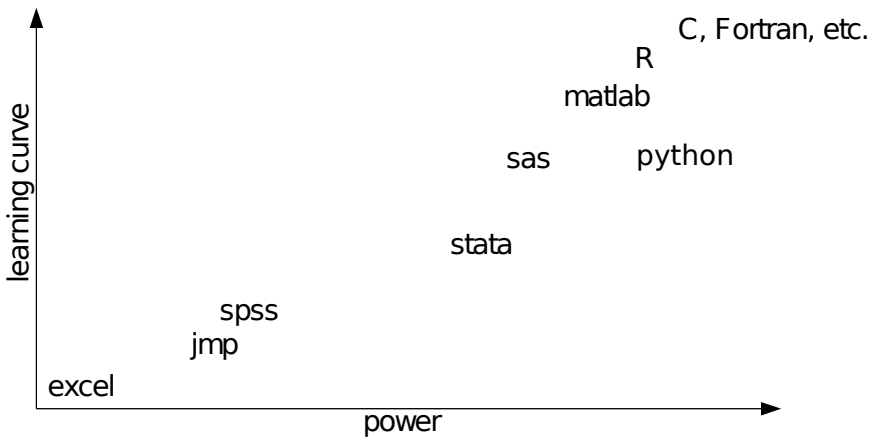
sas

spss



R

which one?



outline

why data science?

stata v sas v r v python

python: fun and useful

[*] bonus—data sources [skim quickly; look at home more]



PYTHON!

YOU'RE FLYING!
HOW?





No thanks!

We are
too busy

general references/tutorials

- we'll skip a lot of stuff, for more see:
- com sci, quick, readable! <http://code.google.com/edu/languages/google-python-class/>
- looks fun <https://automatetheboringstuff.com/#toc>
- eg email/text <https://automatetheboringstuff.com/chapter16/>
- manipulate images <https://automatetheboringstuff.com/chapter17/>
- general, complete, lengthy
<http://www.diveintopython3.net/>
- for soc sci, comprehensive
<http://nealcaren.github.io/python-tutorials/>

Python for dat sci

- automating os, like BASH scripts (os, system)
- data management (pandas)
- text processing (re, nltk)
- statistics (pandas, scipy, statsmodels, scikit-learn)
- API (urllib, rauth)
- gis (pysal)
- web scrapping (beautifulsoup, scrapy, html5lib)
- beautiful graphs, I'd say THE BEST (pandas, matplotlib)

outline

why data science?

stata v sas v r v python

python: fun and useful

[*] bonus–data sources [skim quickly; look at home more]

data sources

- ◇ <http://icpsr.umich.edu>
- ◇ <http://www.worldvaluessurvey.org/>
- ◇ <https://gss.norc.org/>
- ◇ <http://www.thearda.com/>
- ◇ <https://www.pippanorris.com/data>

more data sources

- ◇ `http://www.measureofamerica.org/`
- ◇ `http://usa.ipums.org/usa/`
- ◇ `https://international.ipums.org/international/`

“non-traditional” data

- ◇ <http://dvn.iq.harvard.edu/dvn/dv/patent>
- ◇ http://www.trustlet.org/wiki/Trust_network_datasets

happiness data

- ◇ `http://www.bmj.com/content/337/bmj.a2338.full`
- ◇ `http://www.wefeelfine.org/`