

intro

Adam Okulicz-Kozaryn

`adam.okulicz.kozaryn@gmail.com`

this version: Monday 16th January, 2023 18:20

outline

why data science?

stata v sas v r v python

python

[*] bonus—data sources [skip, can look at home]

introductions

- `https://theaok.github.io` and go to my goog sch
- and i will tell you what data i have been using
- if you use same data then i can share my code with you!
-
- you? (see if others overlap: collaborate!):
- research interests and data?

outline

why data science?

stata v sas v r v python

python

[*] bonus–data sources [skip, can look at home]

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 ?

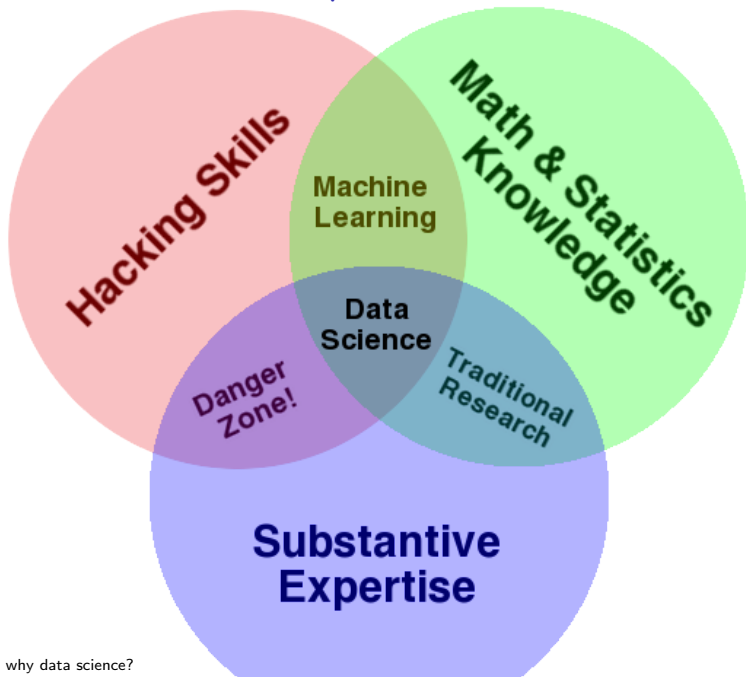
data management is fundamental

- ◇ in order to analyze data you need to manage it first
- ◇ GIGO (Garbage In Garbage Out)
if data management fails, data analysis fails
-
- ◇ 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!



outline

why data science?

stata v sas v r v python

python

[*] bonus–data sources [skip, can look at home]

a critical decision!

- it takes months to get productive with software
- it takes years to master software
- huge time investment
- in soc sci dat man the choice is: sas, stata, r, python
- there's more (Lisrel, HLM, etc) but the above are major
- excel and spss are junk that no one should use

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



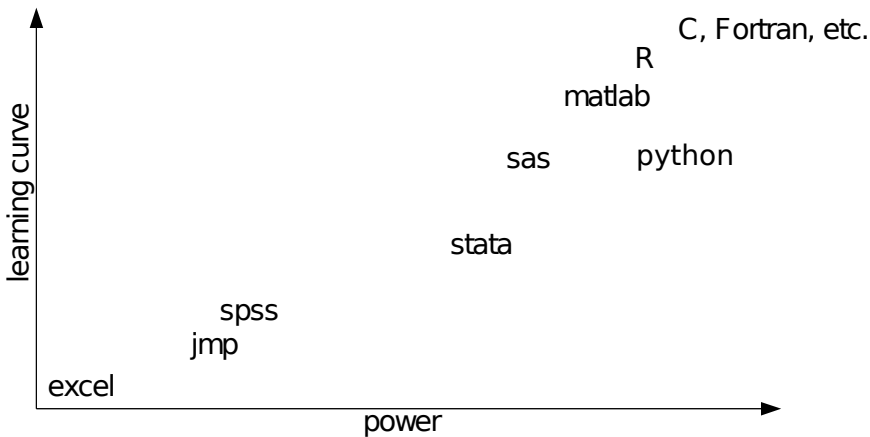
sas

spss



R

which one?



outline

why data science?

stata v sas v r v python

python

[*] bonus-data sources [skip, can look at home]



PYTHON!

YOU'RE FLYING!
HOW?



py is fun! and can accomplish so much more!

- why dat sci while we're already busy with soc/nat sci?
- so that you can do more soc sci and stop wasting time!
- don't say "no thanks! we are too busy"



No thanks!

We are
too busy

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/>

strategy/approach

- I won't bother you with dictionaries, tuples, etc
- I will introduce Python as for data science

but why another software for vis?

- stat software like Stata is behind
- R is clunky; and Py has better graphics

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

[*] bonus-data sources [skip, can look at home]

data sources

- ◇ <http://www.worldvaluessurvey.org/>
- ◇ <http://www.norc.uchicago.edu/GSS+Website/>
- ◇ <http://www.icpsr.umich.edu/icpsrweb/ICPSR/>
- ◇ <http://www.thearda.com/>
- ◇ <https://www.pippanorris.com/data>

more data sources

- ◇ <http://www.measureofamerica.org/>
- ◇ <http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/0,,contentMDK:20388241~menuPK:665266~pagePK:64165401~piPK:64165026~theSitePK:469382,00.html>
- ◇ <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://apps.facebook.com/usa_gnh/`
- ◇ `http://www.facebook.com/notes/facebook-data-team/relationships-and-happiness/304457453858`
- ◇ `http://www.springerlink.com/content/757723154j4w726k/fulltext.pdf`
- ◇ `http://www.wefeelfine.org/`

facebook data

- ◇ http://apps.facebook.com/usa_gnh/
- ◇ <http://www.facebook.com/notes/facebook-data-team/relationships-and-happiness/304457453858>
- ◇ <http://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919>
- ◇ <http://cyber.law.harvard.edu/node/4682>
- ◇ <http://www.thefacebookproject.com/resource/datasets.html>