

introduction

Adam Okulicz-Kozaryn

`adam.okulicz.kozaryn@gmail.com`

this version: Tuesday 5th September, 2023 14:38

outline

why data management?

Python v R v Stata

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

introductions (see if others overlap: collaborate!)

- `https://theaok.github.io`
- your research interests and data?

outline

why data management?

Python v R v Stata

[*] bonus—soc sci 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 ?

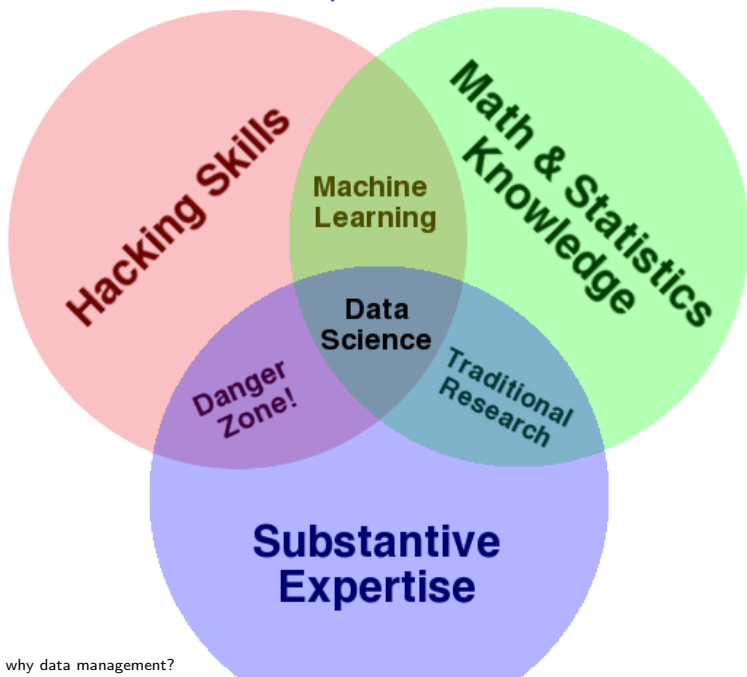
data management is fundamental

- ◇ 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

data science

- see venn diag next p

already have stat/math and subst, need hacking!



outline

why data management?

Python v R v Stata

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

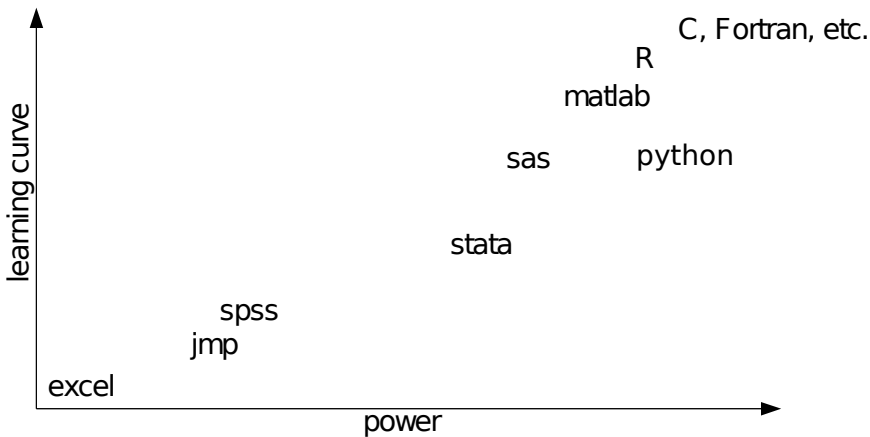
a critical decision!

- takes months to get productive with software
- takes years to master
- excel and spss are junk that no one should use
- sas: a dinosaur (still, often industry standard)

which one?

- stata: user friendly, fast, very concise code
- r: user unfriendly, slow; weird code! for math people
- py: clean and fun; for IT people

which one?



outline

why data management?

Python v R v Stata

[*] bonus–soc sci 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>