introduction

Adam Okulicz-Kozaryn

`adam.okulicz.kozaryn@gmail.com`

this version: Thursday 20th January, 2022   16:16

outline

why data management?

stata v sas v r

let's fire up stata

[*] 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 management?

stata v sas v r

let's fire up stata

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

data revolution!!

$\diamond$ most jobs/tasks require or benefit from programming

$\diamond$ 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
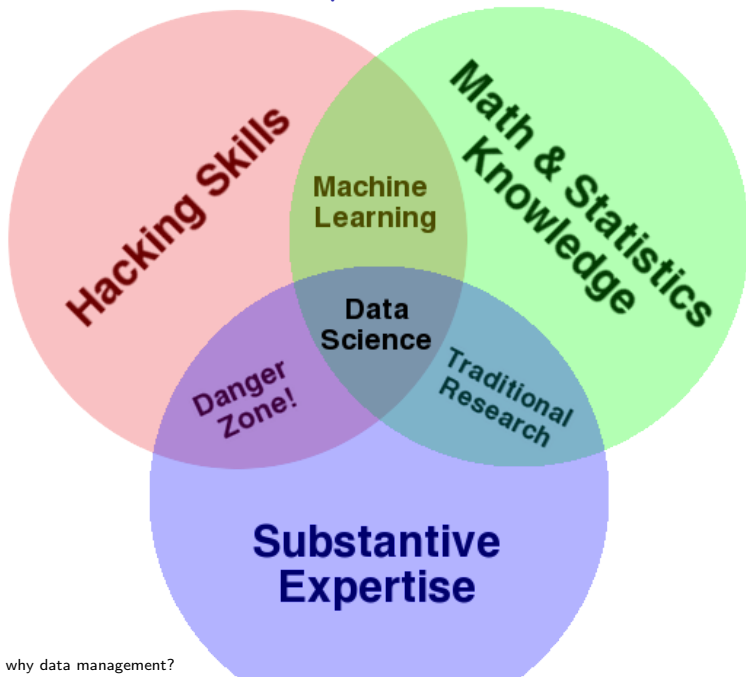 non-quantifiable things ?

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 management?

stata v sas v r

let's fire up stata

[*] 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
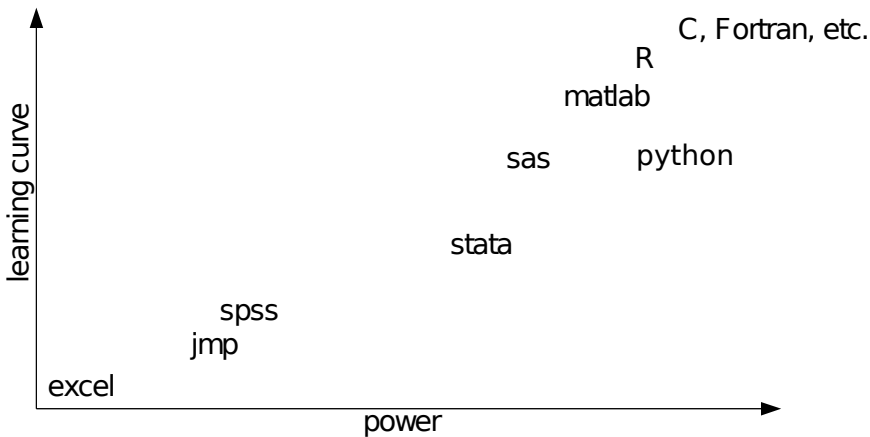- r,py: free, stata: around $200; sas over $1k

spss

sas

R

which one?

<u>outline</u>
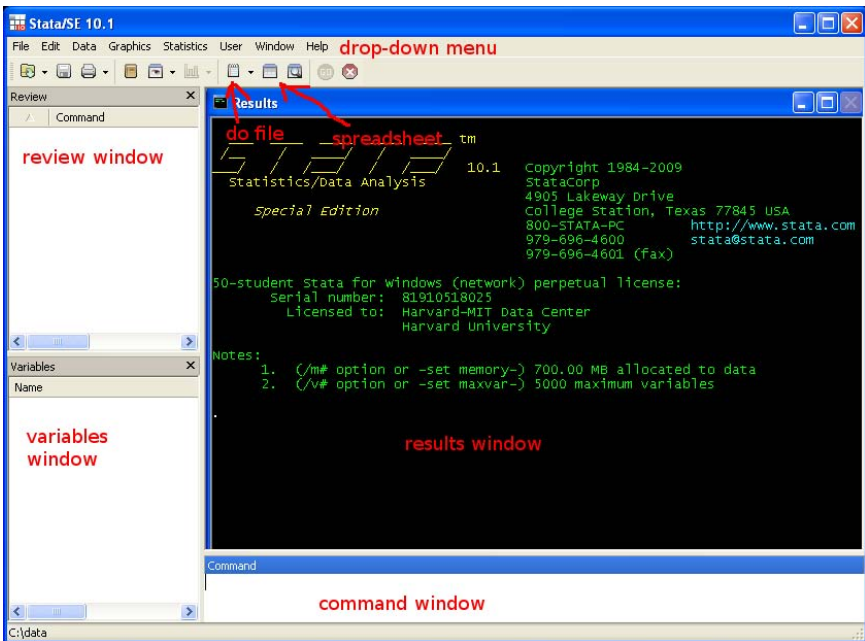
why data management?

stata v sas v r

let's fire up stata

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

stata interface; and do intro.do

looking

- exploring your data is critical!
- we will learn many commands to understand data
- most basic: `d` `sum` `tab`
- remember this:
○ to manage data well, need it to understand it well
○ takes a lot of time to understand it well, so:
○ either manage data you already know
○ or data you're ready to invest ton of time into knowing
- the bottomline is only to deal with data that you're really
  into!

outline

why data management?

stata v sas v r

let's fire up stata

[*] 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/

◇ http:
  //ksghome.harvard.edu/~pnorris/Data/Data.htm

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

more data

- http://www.stateoftheusa.org/blog.php
- http://www.stateoftheusa.org/content/
  health-measures-for-the-develo.php
- http://www.stateoftheusa.org/content/
  fbi-report-violent-crime-down.php
- http://www.stateoftheusa.org/content/
  economy-seen-as-prompting-cohabitation.php
- http://stateoftheusa.org/content/
  measuring-economic-well-being.php
- http://www.stateoftheusa.org/content/
  report-hispanics-outlive-other-american.php