introduction

Adam Okulicz-Kozaryn

adam.okulicz.kozaryn@gmail.com

this version: Thursday 5th September, 2019 20:32

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 see my goog scholar
- and i will tell you what data i have been using
- if you use same data then i can share my code with you!

- yourself? (see if others overlap: can collaborate!)
- research interests and data?
- software? eg SAS, SPSS, Stata, Python, R
- specific expectations for this class?

outline

why data management?

Stata v Sas v R

let's fire up stata

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

data revolution!!

- there are many jobs and there will be more that require programming
- there will be more jobs like that because we have more and more data: twitter, facebook, netflix etc
- these jobs will even ask you for a sample of your code
- data will be more important in any field
- 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

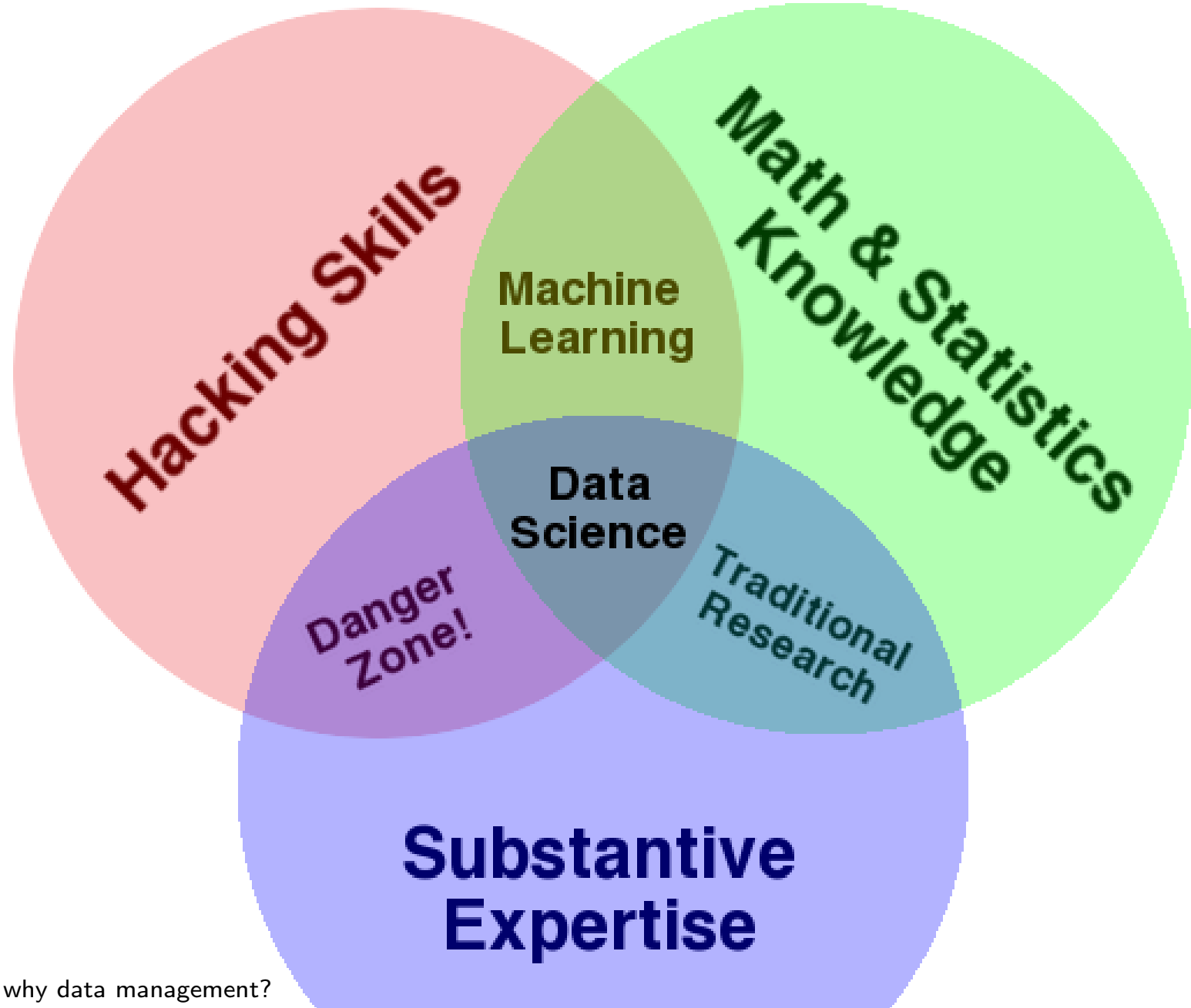
- in order to analyze data you need to manage it first
- GIGO (Garbage In Garbage Out)
  if data management fails, data analysis fails

- takes more time to prepare data than to analyze it
- start early with the right data!

(social) data science or CSS (comp soc sci)

- http://gking.harvard.edu/files/LazPenAda09.pdf
- a person good at computer programming (eg stata)
- a person who can manage, visualize and infer useful information from data
- need to like programming
- willing to learn programming beyond Stata, eg Python
- 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 $300; sas over $1k

which one?

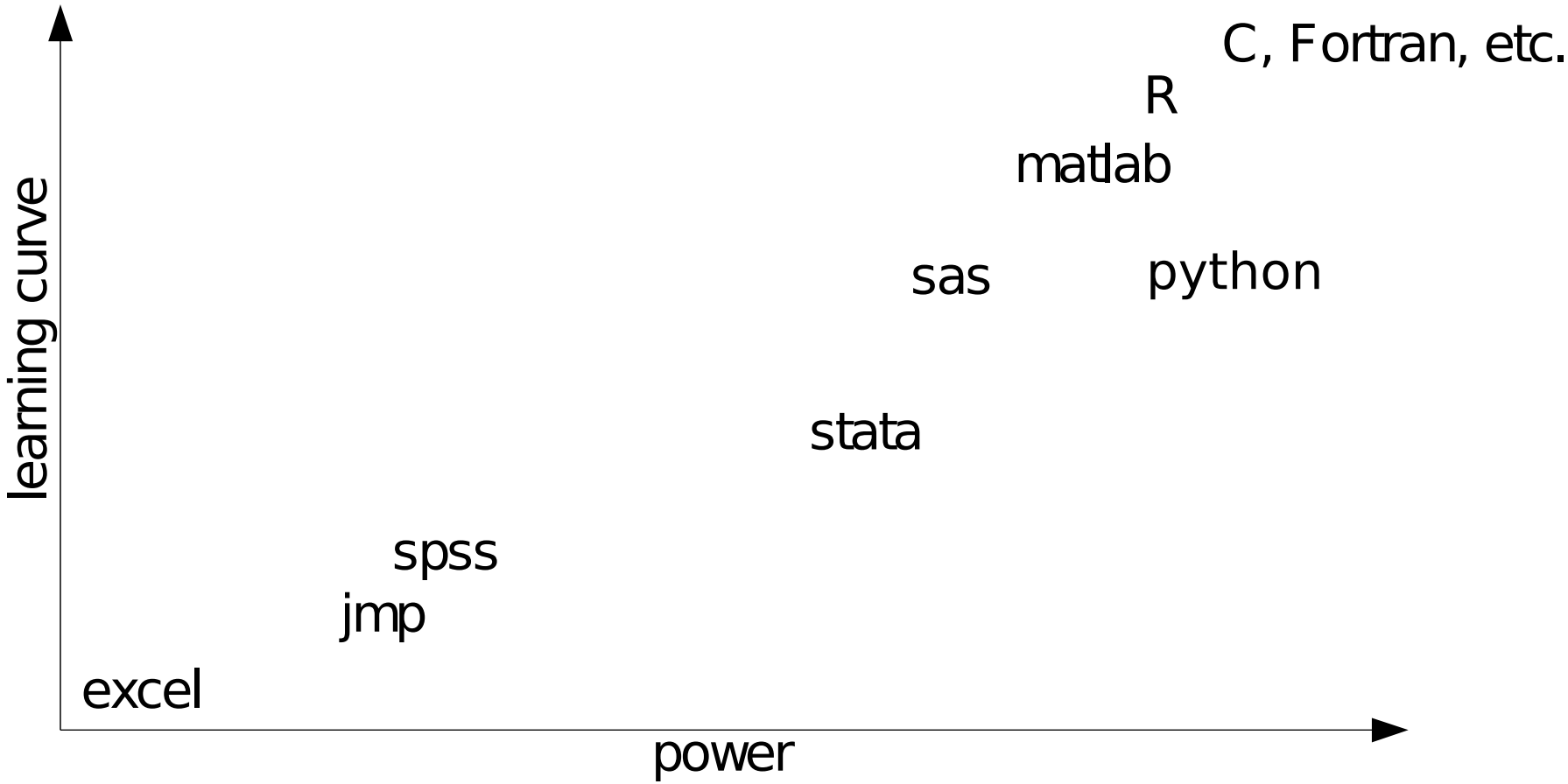
- "Stata is for people getting things done; R is for geeks showing off"
- "people using canned software like Stata don't know what they are doing"
- CONSIDER:
  - your supervisor/field usage
  - are you a geek?
  - do you need to do things that almost nobody does?
  - what's your style? (software is like cars or handbags)

sas

spss

R

which one?


C, Fortran, etc.
R
matlab
sas
python
stata
spss
jmp
excel
learning curve
power

outline

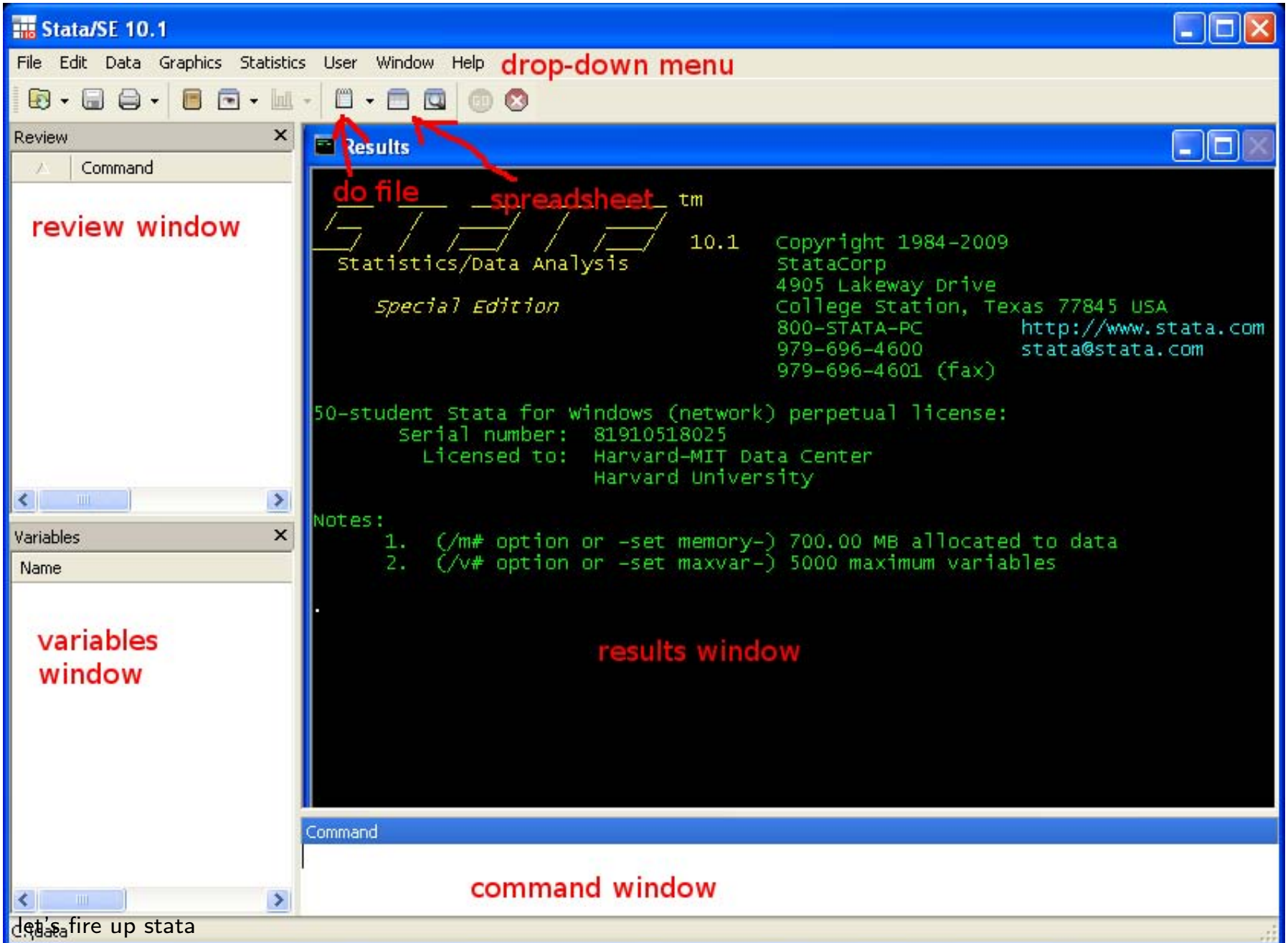
why data management?

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, you need it to understand it well
  - it takes a lot of time to understand it well, and hence
  - either manage data you are already familiar with
  - or data that you are ready to invest a lot of time into knowing
- again, the bottomline in this class is to manage data that is of great interest to you

outline

why data management?

Stata v Sas v R

let's fire up stata

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

data.gov

- `http://www.data.gov/`

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

LaTeX

- we may have a separate class, but some links follow:
- `http://people.hmdc.harvard.edu/~akozaryn/myweb/latex/`
- `http://www.ats.ucla.edu/stat/stata/latex/default.htm`
- i will post my do-files that output into LaTeX