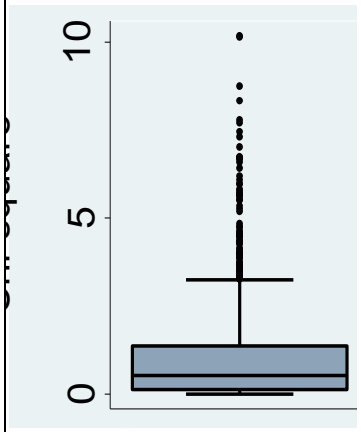
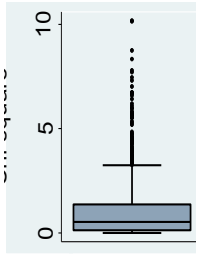
Biostat 202: Opportunities and Challenges of "Big Data"

Aaron Wolfe Scheffler,
Division of Biostatistics,
Department of Epidemiology and Biostatistics



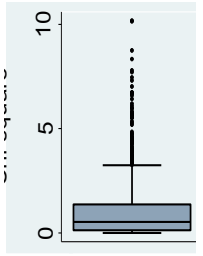
Biostat 202

Zoom lectures and labs

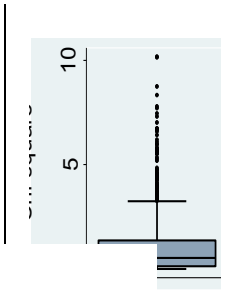


- Zoom lectures and labs will be recorded and posted for those who cannot attend
- Muted by default, please use chat to get my attention

What is “Big Data” today?



Big Data



Jasper Sitwell: Zola's algorithm is a program...for choosing Insight's targets.

Steve Rogers: What targets?

Jasper Sitwell: You! The TV anchor in Cairo, the Undersecretary of Defense, a high school valedictorian in Iowa city. Bruce Banner, Stephen Strange, anyone who's a threat to HYDRA. Now, or in the future.

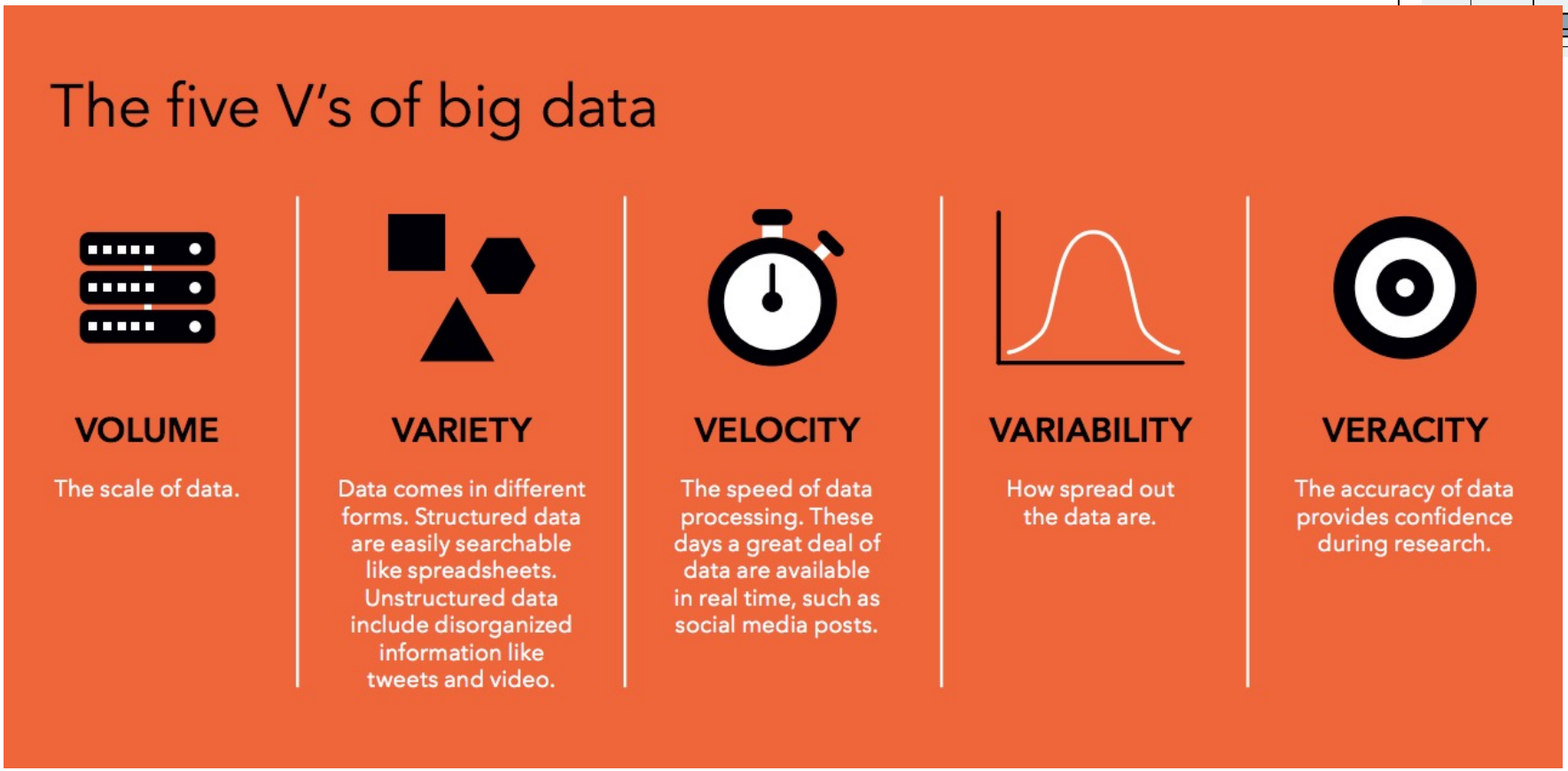
Steve Rogers The future? How can it know?

Jasper Sitwell: How could it not? The 21st century is a digital book. Zola taught HYDRA how to read it. Your bank records, medical histories, voting patterns, e-mails, phone calls, the damn SAT scores. Zola's algorithm evaluates peoples' past to predict their future.

Steve Rogers What then?

Jasper Sitwell: Then the Insight heli-carriers scratch people off the list.

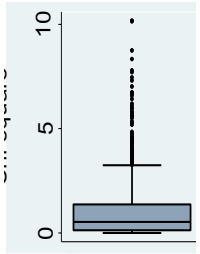
“Big Data”



Big Data

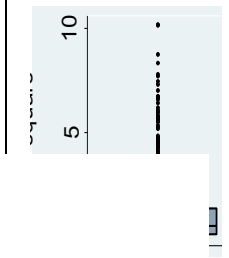
All Data

“Big Data” – the 5 Vs



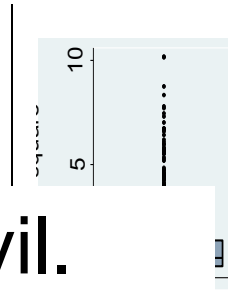
- **Volume**: Google flu trends, billions of data points
- **Variety**: electronic health records include diagnosis, imaging, genetics, etc…
- **Velocity**: real time model training, e.g. hypoglycemia prediction
- **Variability**: random variation, spurious correlations
- **Veracity**: harder to ensure data quality in big data

Thoughts on Big Data



1. Data mining can be used for good and evil.
2. Big data isn't the same as big information.
3. Having big data doesn't help if it was selected in a biased way.
4. Compared to causality, prediction is easy since you only have to understand associations and correlation.
5. If you are interested in causation you can't measure it all.
6. Modern machine learning methods are good at finding spurious correlations.
7. Randomization can be used with big data.

Thoughts on Big Data



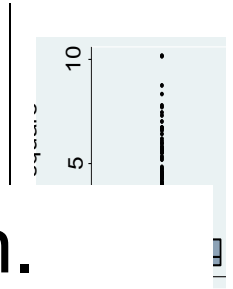
1. Data mining can be used for good and evil.

What 7 Creepy Patents Reveal About Facebook

By Sahil Chinoy

Illustrations by André Wee

Thoughts on Big Data



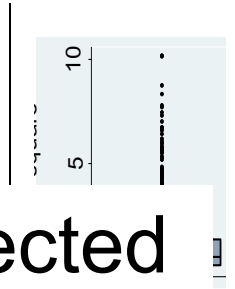
2. Big data isn't the same as big information.

Baboon study:

- 13 baboons studied.
- At least two electrical leads per animal.
- Up to 67 days of observation.
- Record electrical activity of uterine muscles 32 times per second
- Data ? Information ? Knowledge

About 4.8 billion data points (~4.5 GB)

Thoughts on Big Data



3. Having big data doesn't help if it was selected in a biased way.

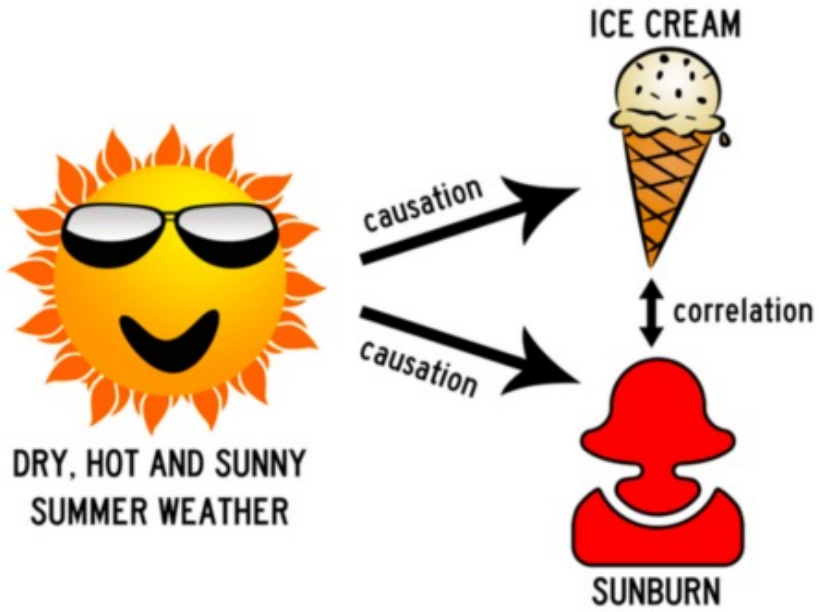
Table 1. The election of 1936.

	Roosevelt's percentage
The *Digest* prediction of the election result	43
Gallup's prediction of the election result	56
The election result	62

Note: Percentages are of the major-party vote. In the election, about 2% of the ballots went to minor-party candidates.
Source: George Gallup, *The Sophisticated Poll-Watcher's Guide*, 1972.

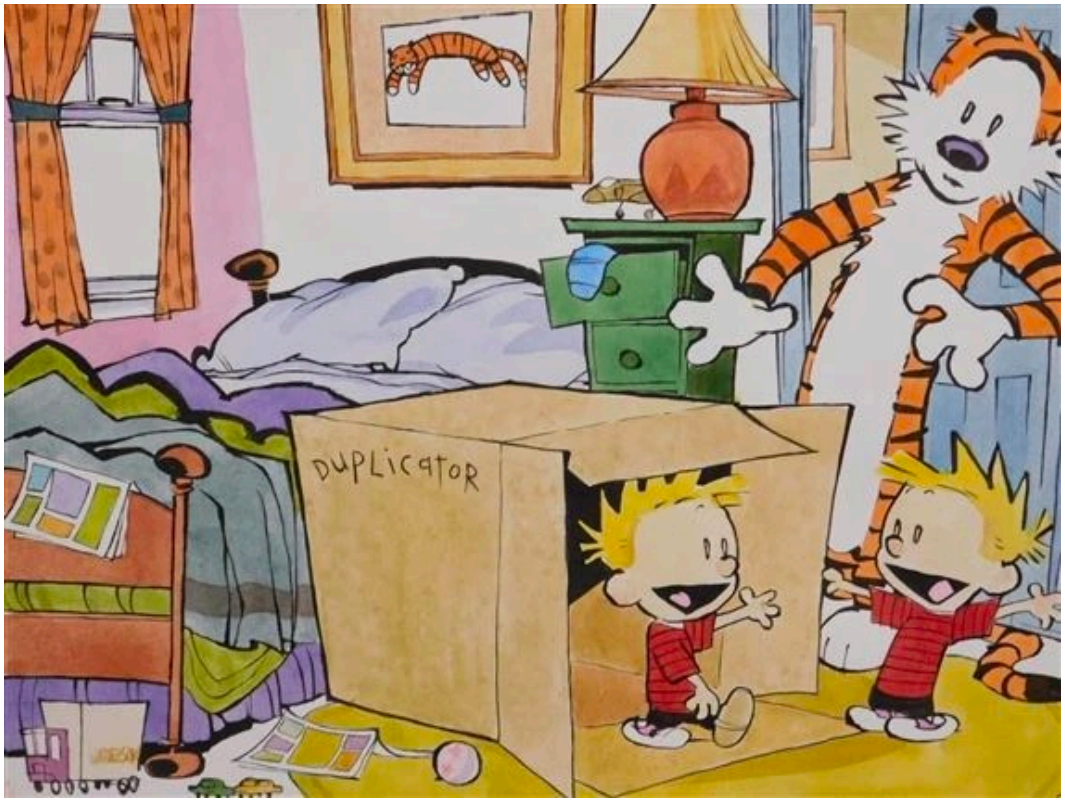
Thoughts on Big Data

4. Compared to causality, prediction is easy since you only have to understand associations and correlation.



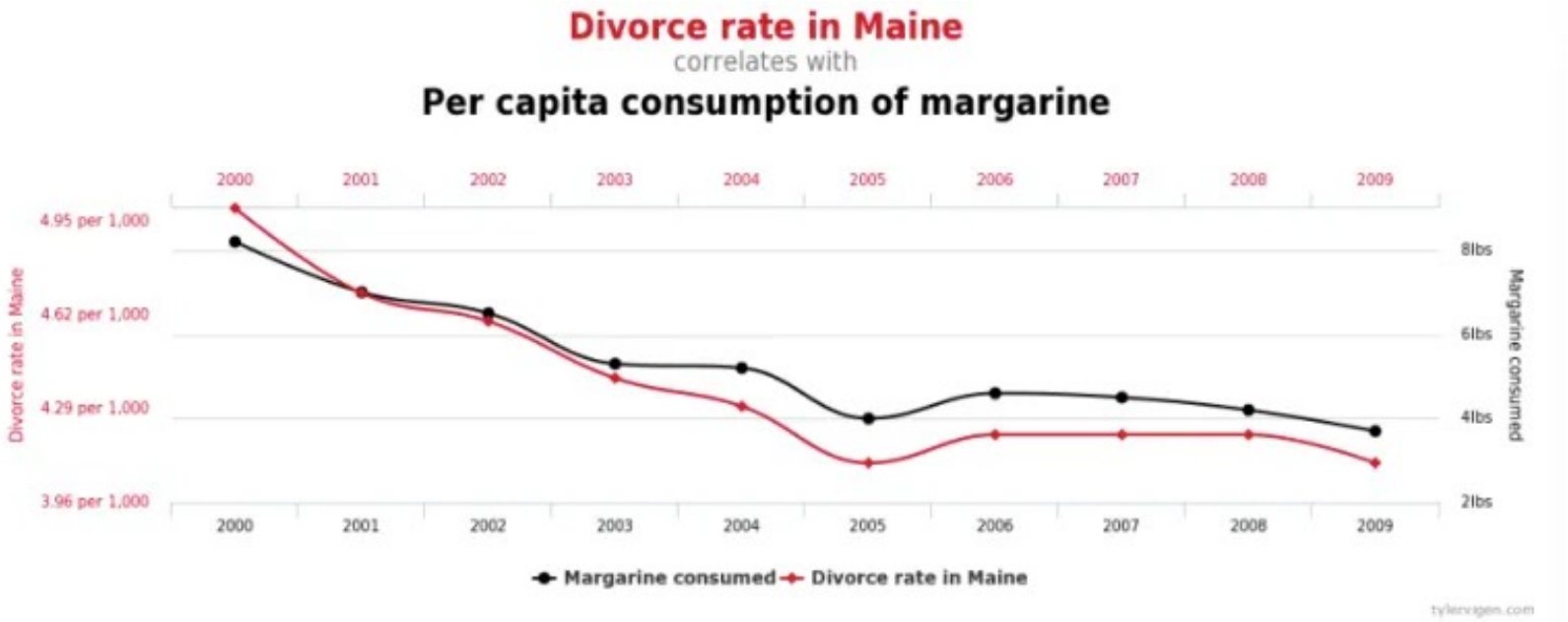
Thoughts on Big Data

5. If you are interested in causation you can't measure it all. You can at most, measure half of it.



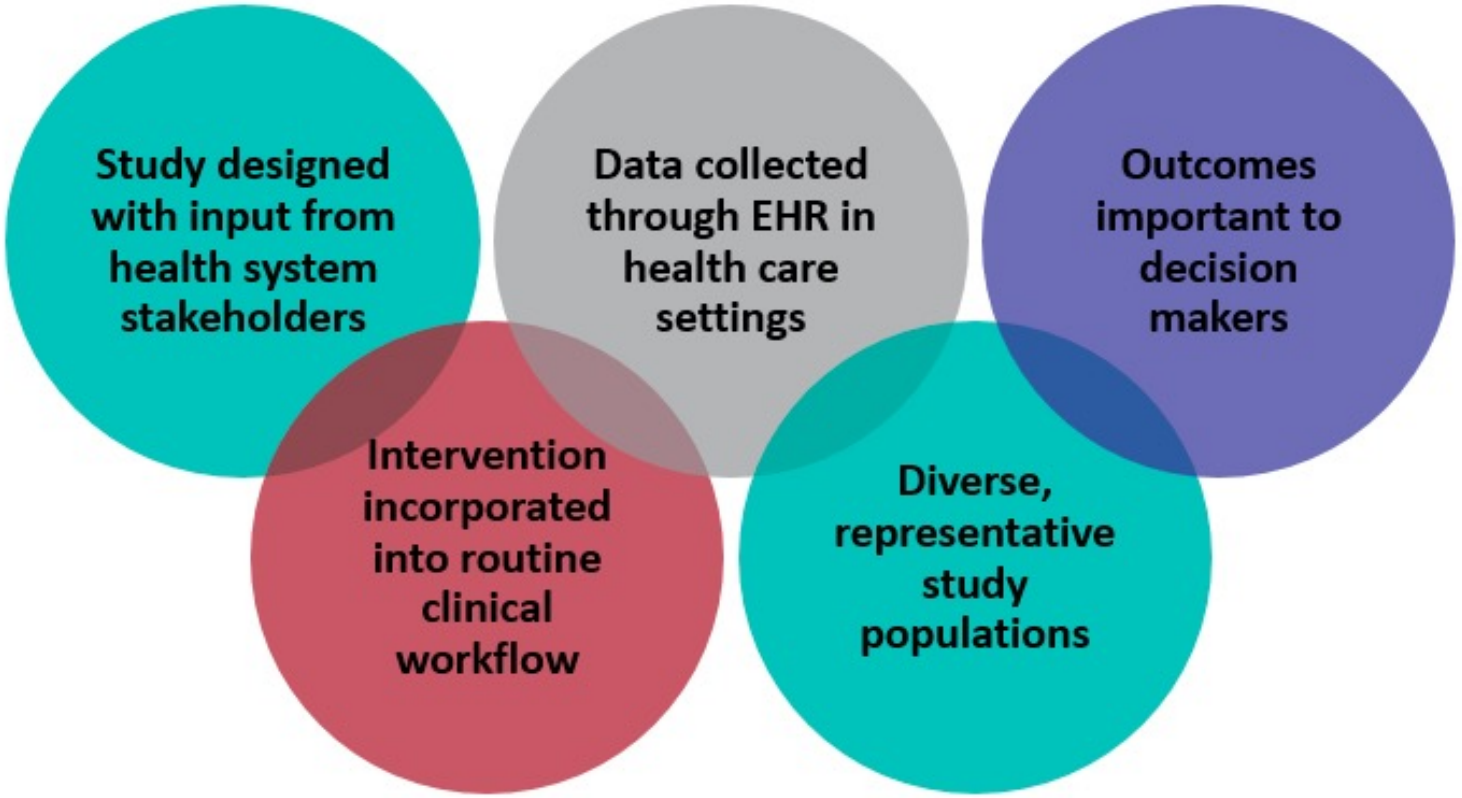
Thoughts on Big Data

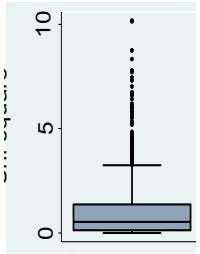
6. Modern machine learning methods are just as good or even better than the more usual statistical methods at finding spurious correlations. Solution? validation.



Thoughts on Big Data

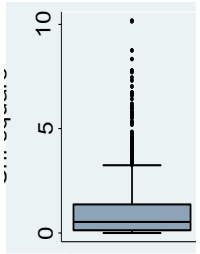
7. Randomization can be used with big data.





Questions related to content?

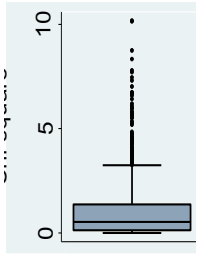
Use the **chat** to let me know, I will attend in that order.



Course outline

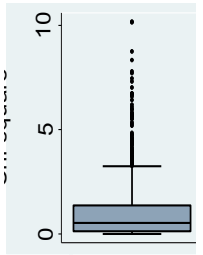
- Introduction
- Getting data (large databases, electronic health records)
- Managing data
- Algorithms (predictive, clustering)
- Predictor/variable selection
- Causal inference and big data
- Describing data with tables and figures
- Case studies

Is this course for you?



- High-level survey course
- Fast, comprehensive, but not deep
- No programming, graphical computer interface
- Not statistics!
- Goal: teach basic fluency in language and methods of machine learning/data mining in big data. Hold your own at a SF dinner party!

Course Delivery

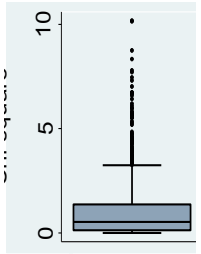


- As you may have noticed, we are remote!
- **Online format**
  - Lecture (2x/week): 1 hour recorded (watch before)
  - Summary/QA (2/week): M/Th 1:00 – 1:45 PM (live Zoom)
  - Lab: Th 2:00 – 4:00 PM (MH-1400)
  - See CLE for links: https://courses.ucsf.edu/course/view.php?id=8594

Attendance is not required but I urge you to log in!

All class is pre-recorded or recorded via Zoom!

Course Grading

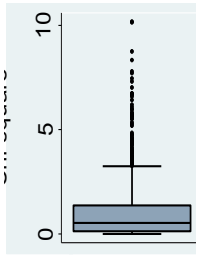


- Assignments (55%)
- Project (45%) – four data options
  - 3 “components” (0/1/2 grading – 15%)
  - Final project write-up (30%)

Turn assignments in on time. *We move fast.*

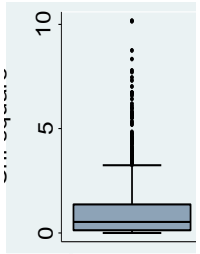
To ease grading, copy and paste answers into a separate document – do not respond inline.

TICR Professional Conduct Statement



- I will maintain the highest standards of academic honesty
- I will neither give nor receive aid in examinations or assignments *unless such cooperation is expressly permitted by the instructor* (Our policy: mutual collaboration on labs and assignments is fine. The project should be your own work)
- I will conduct research in an unbiased manner, report results truthfully, and credit ideas developed and work done by others
- I will not use answer keys from prior years
- I will write answers in my own words, and, when collaboration is permitted, acknowledge collaborators when answers are jointly formulated.

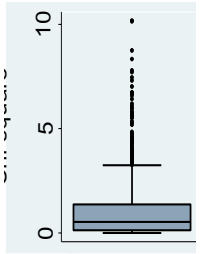
Course due dates



- Assignments are due Thursdays starting 7/29.
- Project components due on 8/8, 8/15 and 8/22.
- Final project due Thursday 9/2.

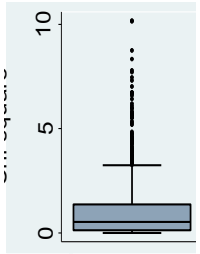
All materials submitted via CLE.

Course software:



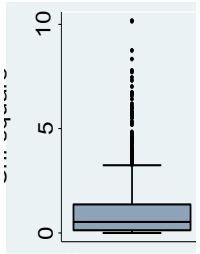
- Orange Data Mining
- A key aspect of the labs will be hands-on use of the software.
- Allows access to many of the features of big data analytics without the need to learn programming.

Optional reading (posted on CLE):



- *An Introduction to Statistical Learning*. eBook free via http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf
  Or within UCSF network from Springer directly.

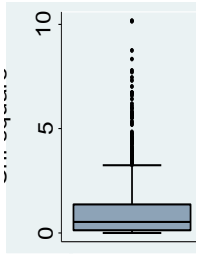
Questions?



Class forum – let other people see your questions!

Or contact me: aaron.scheffler@ucsf.edu

Next up, lab



Take a 15 minute break and we will return and go through lab.