Lab #1: Data Wrangling and Exploration

Everything Data

CompSci 216 Spring 2015

DUKE
COMPUTER SCIENCE

Announcements (Wed. Jan. 14)

- Welcome our UTAs!

Billy Wan

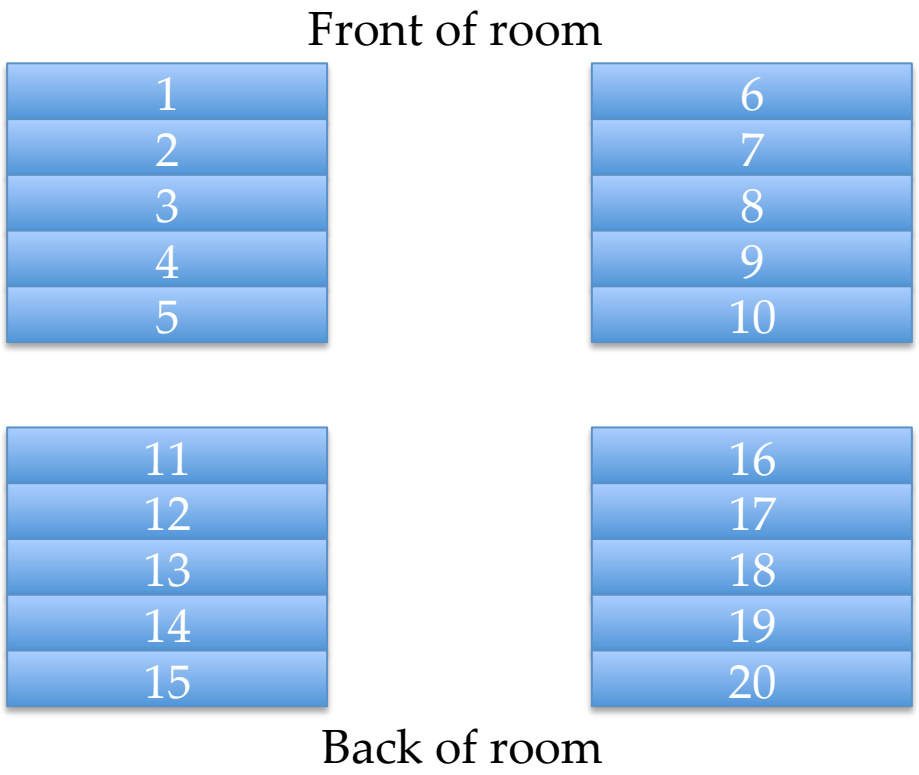
Eric Wu

Kevin Wu

- Most of our office hours have been posted on the website

Seat assignment

See course website (under "Schedule") for team assignment



Format of this lab

- Discuss last homework
  - Continue working on your setup if needed
- Work on new exercises
  - Raise your hand to get your answer checked by the course staff
  - Win challenge to get little prizes!
- Summarize lessons learned (10 min.)

Lesson learned: reality check

- Data is messy
- Reality is messy
- Garbage in, garbage out

☞ *One reason why you need data wrangling*

Image credit: http://www.practicalpolymath.com/wp-content/uploads/2012/10/eliminate_the_paper_mess.jpg

Lesson learned: "abstraction"

- More structure and semantics ➠ more powerful analysis, and
- Different data models ➠ different questions and processing methods

☞ *Another reason you need data wrangling*

- Examples in this lab
  - Character strings: regular expressions
  - Dates: meaningful range conditions, subtraction
  - Tables: filtering, grouping, counting, sorting…

Picture of Prof. Carlo Sequin, UC Berkeley. Image credit: http://scgp.stonybrook.edu/archives/4169

Lesson learned: UI or not UI

- Exploratory analysis goes a long way
  - You can get fair amount of insight without a single line of code

- Interactivity is nice
- Easy to learn, difficult to master
  - Tool-specific recipes vs. universal primitives

Image credit: http://www.24tee.com/movie_related_shirts/lunch_is_for_wimps_tee_shirt

What's next

- No class next Monday (MLK Day)
- No homework until next Wednesday (lecture on relational data processing)
  - Resolve any remaining setup issues
  - Try to gain some efficiency with VM and shell