Introduction to Data Science
SKKU University, Summer 2015

Prof. Jevin D. West
University of Washington

Lecture 5 – July 3, 2015

Week in Review

- Introduction to Data Science
- Data Science Opportunities
- Version Control Systems (Github)
- Introduction to R

Next Week

- Management in Data Science
- Experimental Design
- Cloud Computing
- Probability Distributions
- Central Limit Theorem
- Visualizing Data in R (ggplot)
- Data Science Examples

Schedule

** The schedule will update regularly so make sure to check it regularly*

- Week 1
  - Introduction to R (and python)
  - Version Control Systems
  - Opportunities in Data Science
  - Data Ingestion
  - **Assignments:** Quiz #1, Assign. #1, Group Project (Data Set Identified)
- Week 2
  - Cloud Computing
  - Experimental Design
  - Basics in probability and statistics
  - **Assignments:** Quiz #2, Assign. #2, Group Project (Question, Preliminary Stats)
- Week3
  - Basics in machine learning
  - Network analysis
  - Information visualization
  - Data Ethics
  - **Assignments:** Quiz #3, Assign. #3, Group Project (Final Paper, Final Presentation)

Agenda

9:30 – 10:00	Black Box
10:00 – 10:30	R Functions
10:30 – 10:40	Break
10:40 – 11:30	Empirical Frameworks
11:30 – 12:00	Questions

Logistics

- Attendance
- Repositories (individual and team)
- Class materials can be found in this repository:
    - https://github.com/jevinw/SKKU_DataScience_2015

Schedule

- DUE Today: Assignment #1
- DUE Today: Script and figures from last class
- DUE Today: Data set identified (repository)
- **DUE Tuesday, July 7**: Question Identified with a 1-page description and answers to the following:
  - Why are you investigating this question?
  - How are you going to try and answer your question?
  - What are the limitations of your question?
  - Who else has answered this question? How will you build upon other work done with this question?
  - Provide references
- Wednesday, July 8: Quiz #2
- DUE Friday, July 10: Assignment #2

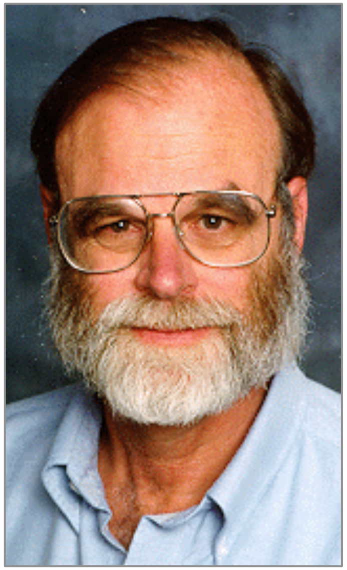
Empirical Frameworks

* The following slides on empirical frameworks were developed by Prof. Joshua Blumenstock

Readings

- Thomas Davenport (2006). "Competing on Analytics", Harvard Business Review, Jan. 2006, Vol. 84 Issue 1, pp. 99-107
- The Fourth Paradigm, *Jim Gray on eScience: A Transformed Scientific* Method, pgs xvii – xxxi

Science Paradigms



Jim Gray

Empirical

Theoretical

Computational

Data Exploration

- The Fourth Paradigm, *Jim Gray on eScience: A Transformed Scientific* Method, pgs xvii – xxxi

Empirical Frameworks

- ***Empirical*** (Merriam-Webster):
  1. originating in or based on observation or experience
  2. ~~relying on experience or observation alone often without due regard for system and theory~~
  3. capable of being verified or disproved by observation or experiment

Empirical frameworks and DS

- You have a question, a theory, or a decision
  - Note that this is not a foregone conclusion!
- *How to answer, test, decide based on data?*
  - This is your empirical framework
- Key components
  - What data will you use?
  - What empirical methods?
  - How will you communicate results?

Primary types of frameworks

1. Experimental
   - You are able to affect the environment
2. Observational / Non-experimental
   - You have no/limited control over the environment
3. Middle ground: Quasi-Experimental
   - You look for something resembling an experiment

- Any of the above can be *causal* or *descriptive*

Experimental

- You can affect the environment
- Common scenarios:
  - You can offer subjects incentives or promotions
  - You can assign different treatments
- Examples
  - AT&T: What causes people to churn?
  - Kaiser: How to reduce patient recidivism?
  - IMT 589: How to get people to read?

Experimental

PROS

- Well-defined counterfactual
- Causal inference simpler
- Greater statistical power

CONS

- Difficult to implement
- Can cause confusion
- Can create inequity
- May be unethical

Observational

- You have no or limited control over the environment
- Common scenarios
  - Want to know the effect of something in the past
  - You want to segment customers
- Examples
  - AT&T: What causes people to churn?
  - Kaiser: How to reduce patient recidivism?
  - IMT 589: How to get people to read?

Observational

PROS

- Easy to implement
- Does not interfere with normal operations

CONS

- Weak counterfactual
- Correlation vs. causality
- Limited control

Quasi-Experimental

- Idea: Look for something resembling an experimental intervention
- Common Scenarios
  - Natural experiments
  - Policy experiments
- Examples:
  - Weather patterns and air pollution
    - Schlenker & Walker (2012): "Airports, Air Pollution, and Contemporaneous Health"
  - College scholarships and lifetime earnings
    - Alex Solis (2012): "Credit access and college enrollment"

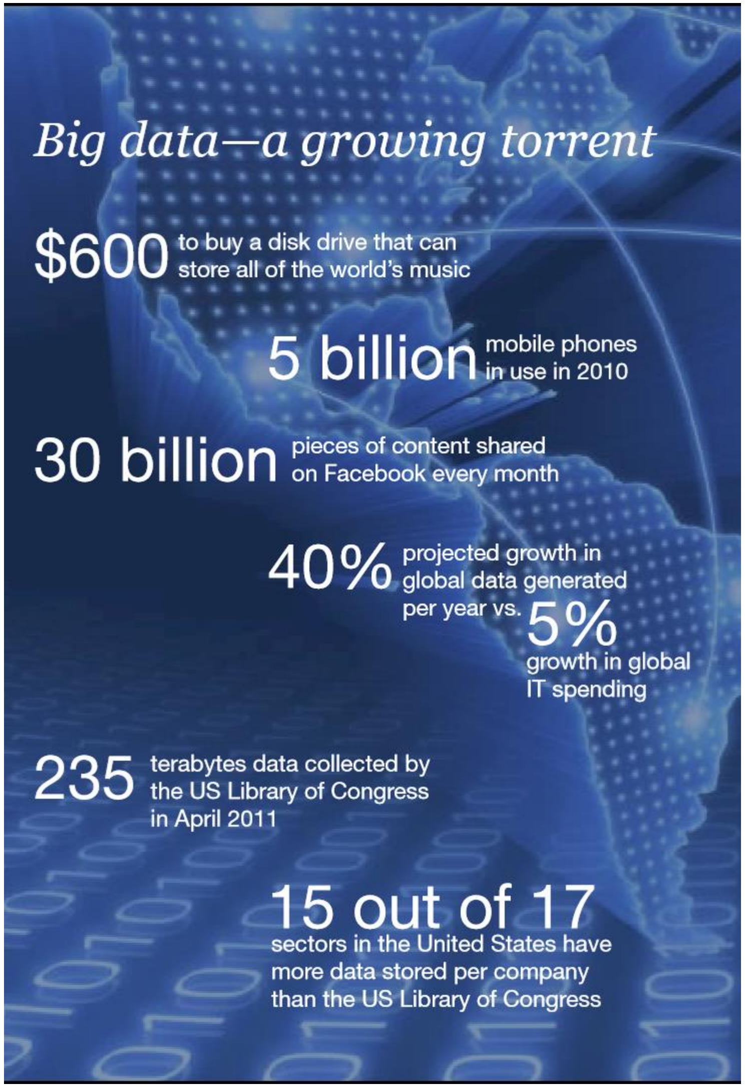
Questions?

- Empirical Frameworks
  - Experimental
  - Observational
  - Quasi-experimental

Empirical frameworks and final projects

- Setting up a solid framework is critical!
- Start with a single, well-defined, intriguing and non-obvious question
  - How did Twitter behavior change in response to the crisis in Syria?
  - Is the sentiment of Yelp! reviews correlated with global and local economic trends?
  - This is not easy!
- What sub-questions you must answer along the way?
- Plan out your analysis by listing every step and every figure/table you will produce *in advance*

“Data Science” and “Big Data”



Big data—a growing torrent

$600 to buy a disk drive that can store all of the world’s music

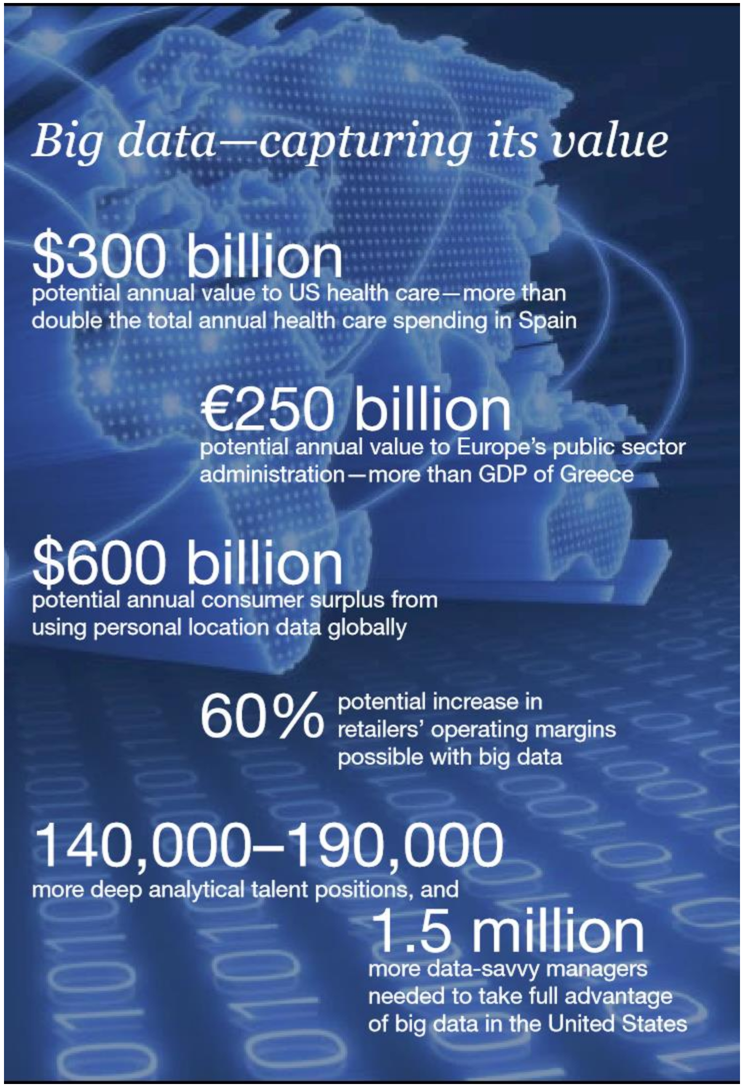
5 billion mobile phones in use in 2010

30 billion pieces of content shared on Facebook every month

40% projected growth in global data generated per year vs. 5% growth in global IT spending

235 terabytes data collected by the US Library of Congress in April 2011

15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress



Big data—capturing its value

$300 billion potential annual value to US health care—more than double the total annual health care spending in Spain

€250 billion potential annual value to Europe’s public sector administration—more than GDP of Greece

$600 billion potential annual consumer surplus from using personal location data globally

60% potential increase in retailers’ operating margins possible with big data

140,000–190,000 more deep analytical talent positions, and

1.5 million more data-savvy managers needed to take full advantage of big data in the United States

Data Science is about asking good questions

Homework

Big Data