Duke SANFORD SCHOOL OF PUBLIC POLICY

Applied Big Data Science: Energy Data Analytics and Policy

Matthew Harding, Duke University

01/14/2015

“Without big data, you are blind and deaf in the middle of a freeway”

Geoffrey Moore, management consultant and theorist

“There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days.”

Eric Schmidt, of Google, said in 2010.

"Big data is at the foundation of all the megatrends that are happening today, from social to mobile to cloud to gaming."

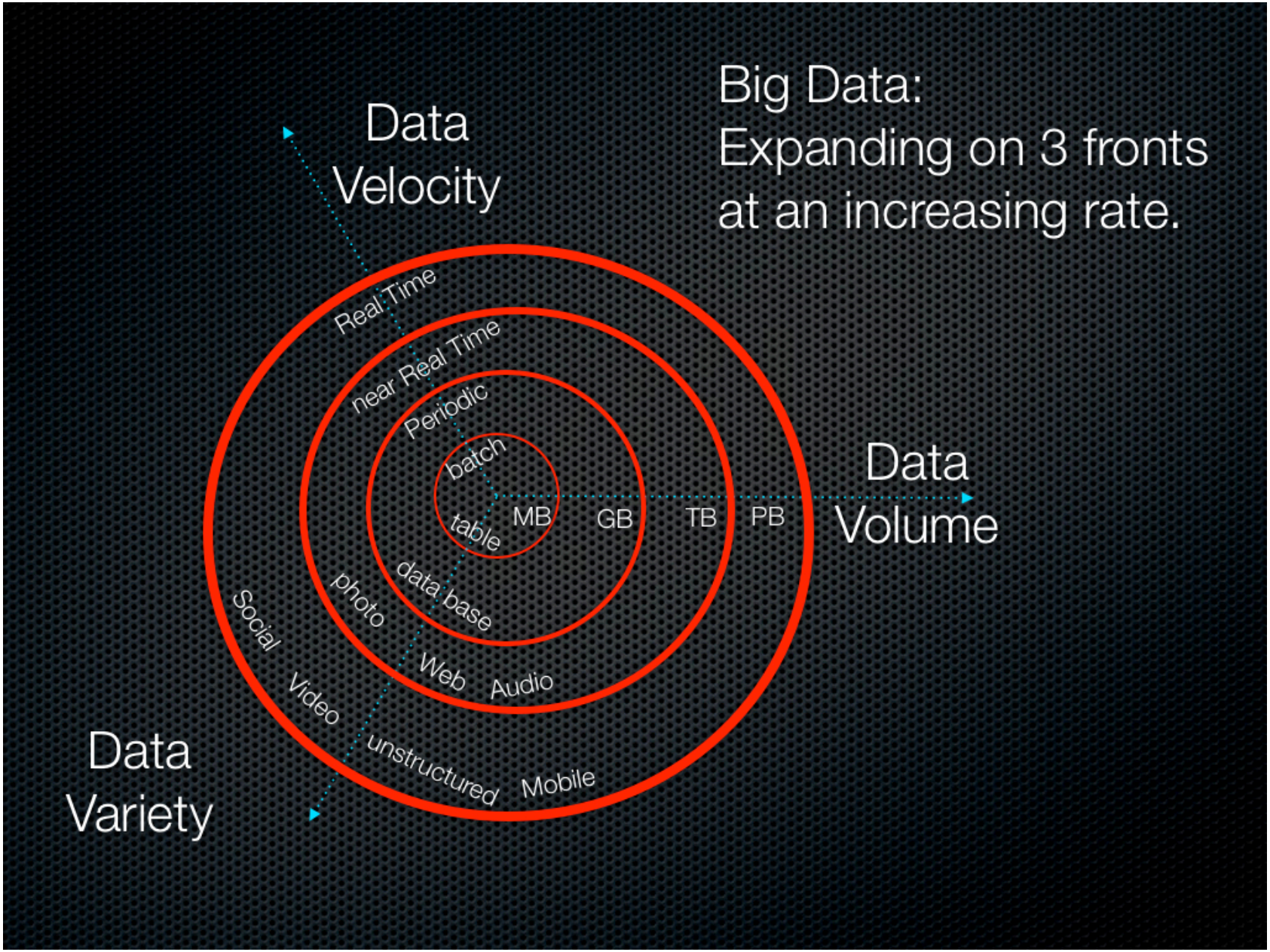
Chris Lynch, Vertica Systems

“I keep saying that the sexy job in the next 10 years will be statisticians, and I’m not kidding”

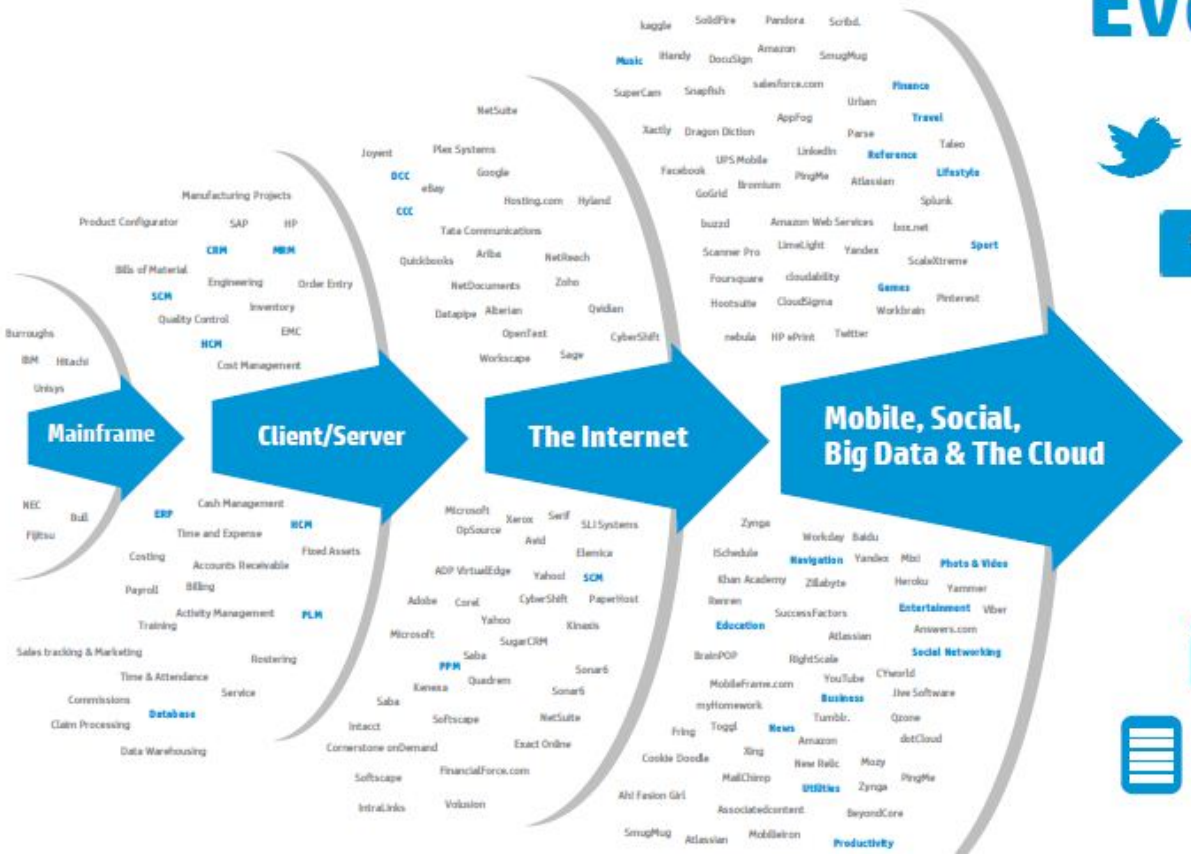
Hal Varian, Google

What is Big Data?

- NSF/NIH: '[B]ig data' […] refers to large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available[.]
- Gartner, Inc (3V):'Big data' is high volume, high velocity and high variety information assets that demand cost-effective, innovative forms of information processing[.]


Big Data:
Expanding on 3 fronts
at an increasing rate.
Data
Velocity
Real Time
near Real Time
Periodic
batch
table
MB
GB
TB
PB
Data
Volume
data base
photo
Social
Web
Audio
Video
unstructured
Mobile
Data
Variety

A new style of IT emerging



Every 60 seconds

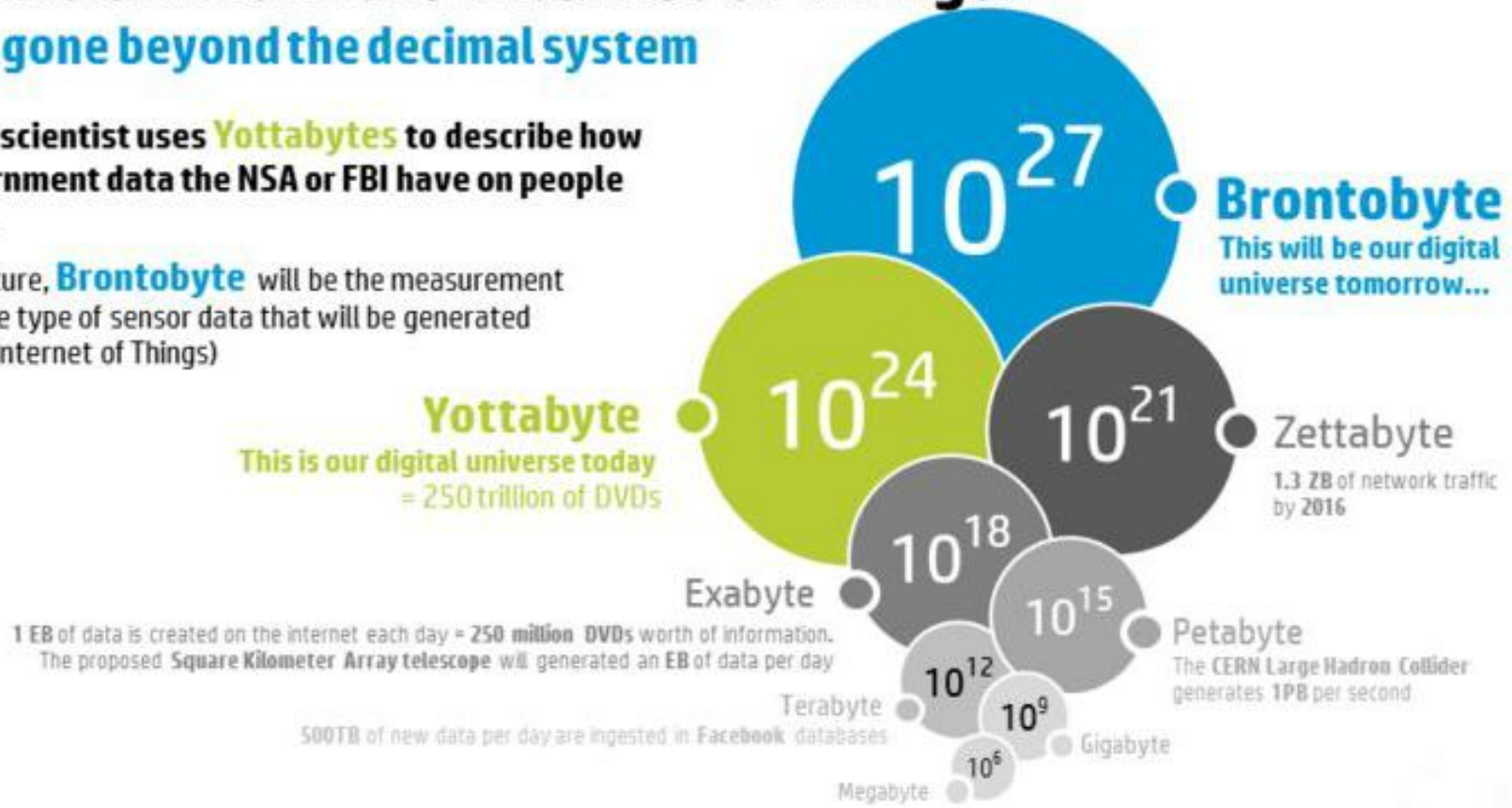
- **98,000+** tweets
- **695,000** status updates
- **11million** instant messages
- **698,445** Google searches
- **168 million+** emails sent
- **1,820TB** of data created
- **217** new mobile web users

Information from the Internet of Things:

We have gone beyond the decimal system

Today data scientist uses Yottabytes to describe how much government data the NSA or FBI have on people altogether.

In the near future, **Brontobyte** will be the measurement to describe the type of sensor data that will be generated from the IoT (Internet of Things)



What is special about Big Data?

- We already had volume and speed (even though it is accelerating
  - E.g. physics
- Variety makes a big difference
- What is special about the data today?

Types of Big Data

Web and social media
data includes clickstream and interaction data from social media such as Facebook, Twitter, LinkedIn, and blogs.

Machine-to-machine
data includes readings from sensors, meters, and other devices as part of the so-called "Internet of things."

Big transaction data
includes healthcare claims, telecommunications call detail records (CDRs), and utility billing records that are increasingly available in semi-structured and unstructured formats.

Biometric
data includes fingerprints, genetics, handwriting, retinal scans, and similar types of data.

Human-generated
data includes vast quantities of unstructured and semi-structured data such as call center agents' notes, voice recordings, email, paper documents, surveys, and electronic medical records.

Social data

- ‘Big data’ [is] the amassing of huge amounts of statistical information on social and economic trends and human behavior.

Deep Data


Transactions
Purchase History | Returns | Credit
Preferences
Subscriptions | Recommendations | Language | Personalization
Web Behaviors
Hit | Page View | Click Path | Session Duration
Social Behavior
Likes | Followers | Friends | Updates
Mobile Activity
SMS | GeoLocation | Badges | Gamification
Enterprise Data
CRM | Finance | ERP
Email Behavior
Open rates | Click through | Bounce | Spam complaints
Demographics
Age | Sex | Location | Profession

Granularity

- not only do these data sets document social phenomena, they do so at the granularity of individual people and their activities
- Makes lots of people unhappy
  - Ethics
  - Privacy
  - Bias

NYT 01/12/15

Prime Minister David Cameron, who said he would pursue banning encrypted messaging services if Britain's intelligence services were not given access to the communications.

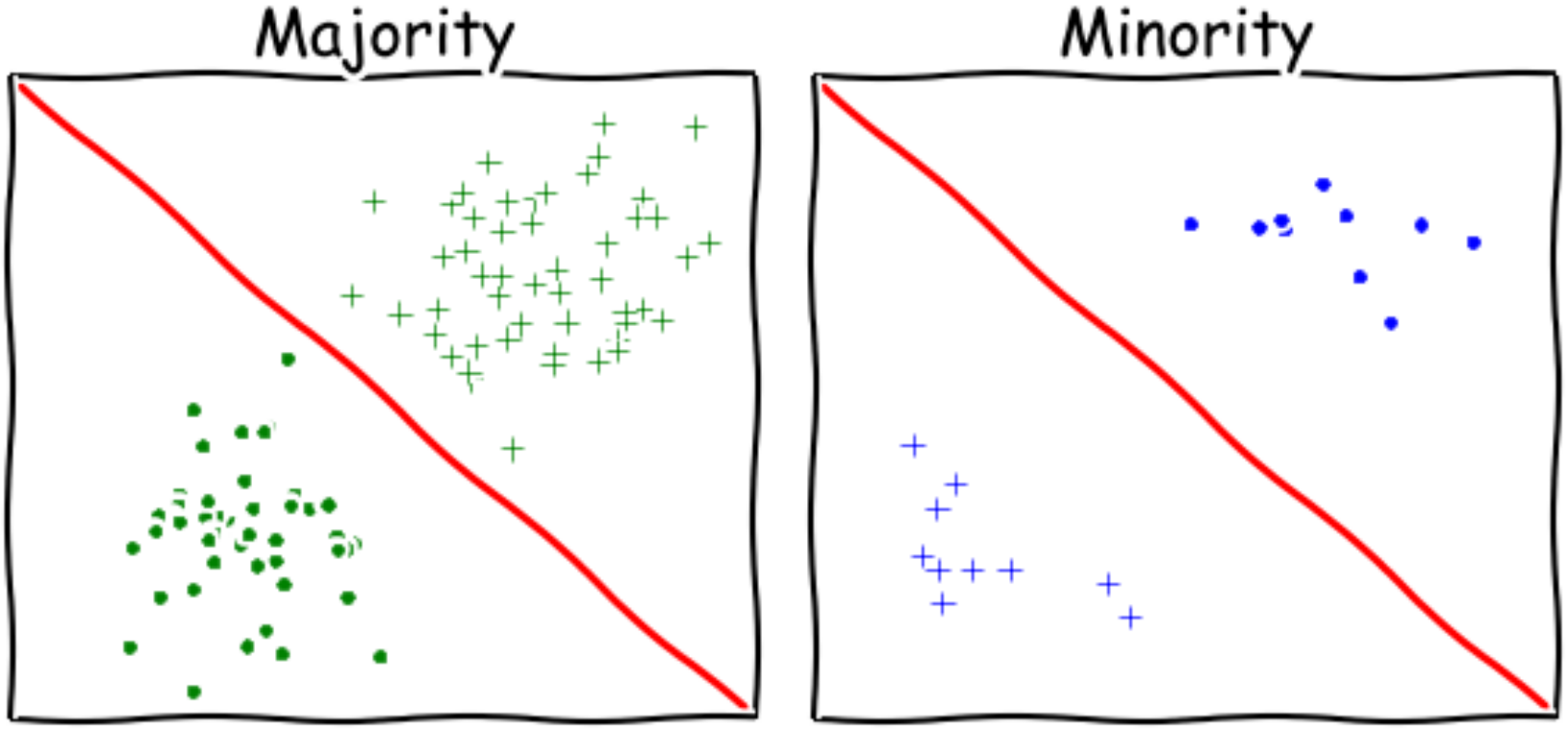
The statement comes as many European politicians are demanding that Internet companies like Google and Facebook provide greater information about people's online activities after several recent terrorist threats, including the attacks in Paris.

Mr. Cameron, who has started to campaign ahead of a national election in Britain in May, said his government, if elected, would ban encrypted online communication tools that could potentially be used by terrorists if the country's intelligence agencies were not given increased access. The reforms are part of new legislation that would force telecom operators and Internet services providers to store more data on people's online activities, including social network messages.

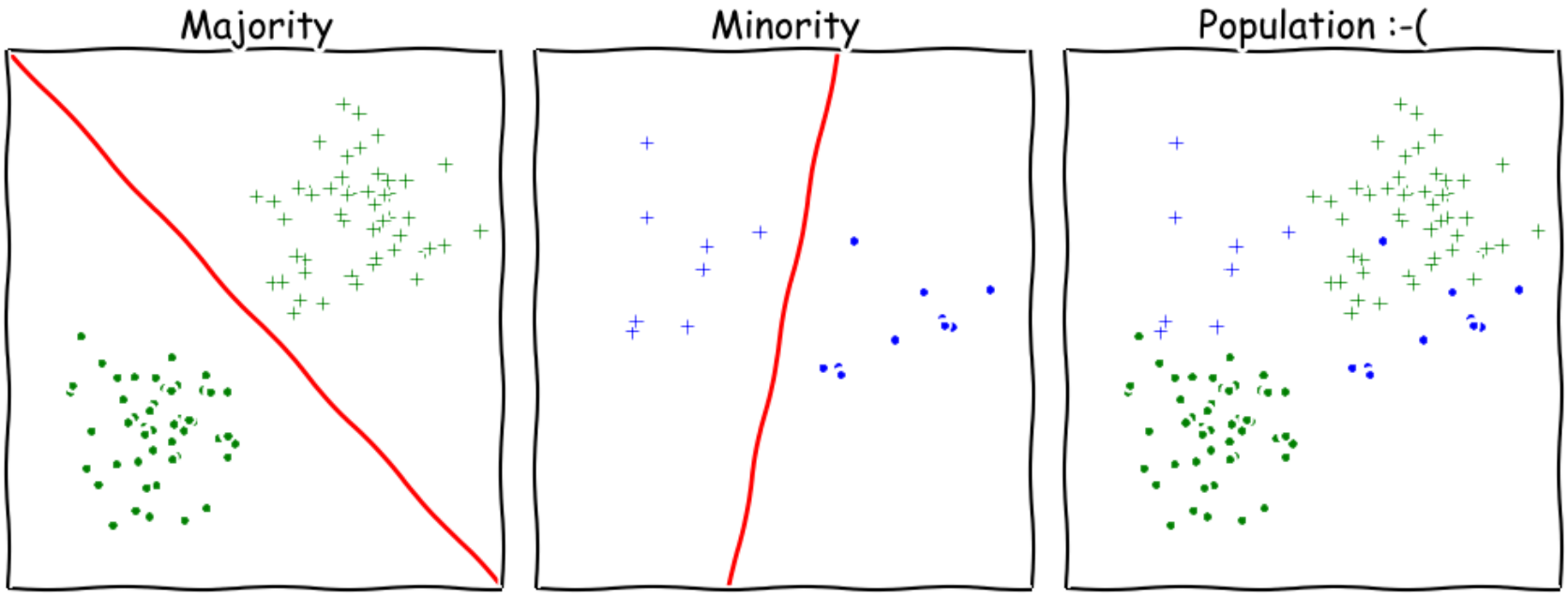
"Are we going to allow a means of communications which it simply isn't possible to read?" Mr. Cameron said at an event on Monday, in reference to services like WhatsApp, Snapchat and other encrypted online applications. "My answer to that question is: 'No, we must not.' "

WHY THE NEED FOR SOCIAL SCIENCE?

(Self-)Selection and Bias



Heterogeneity



"Information is the oil of the 21st century, and analytics is the combustion engine."

Peter Sondergaard, Gartner Research

"Big data is not about the data"

Gary King, Harvard University, making the point that while data is plentiful and easy to collect, the

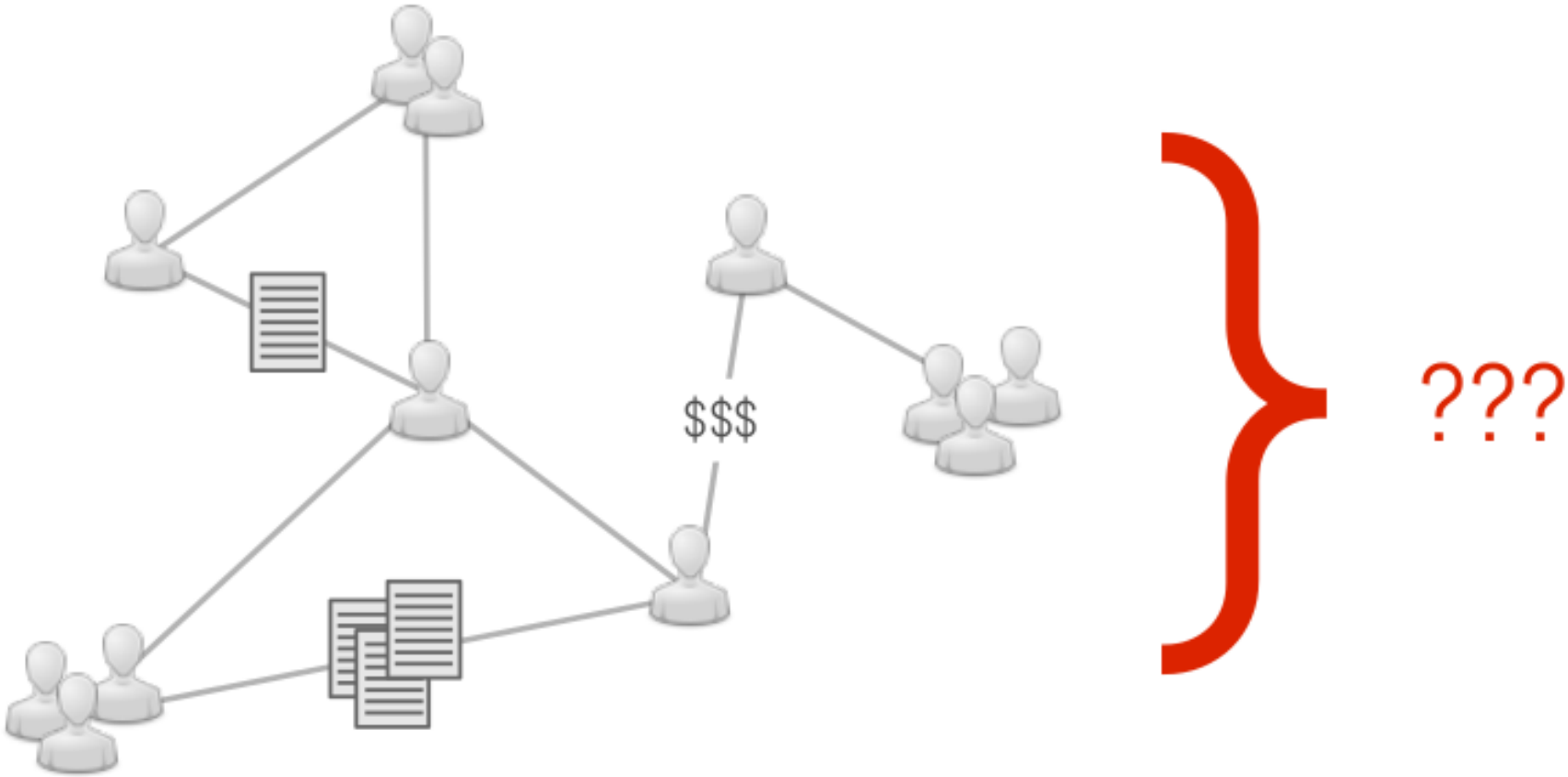
real value is in the analytics.

Differences in modeling approaches

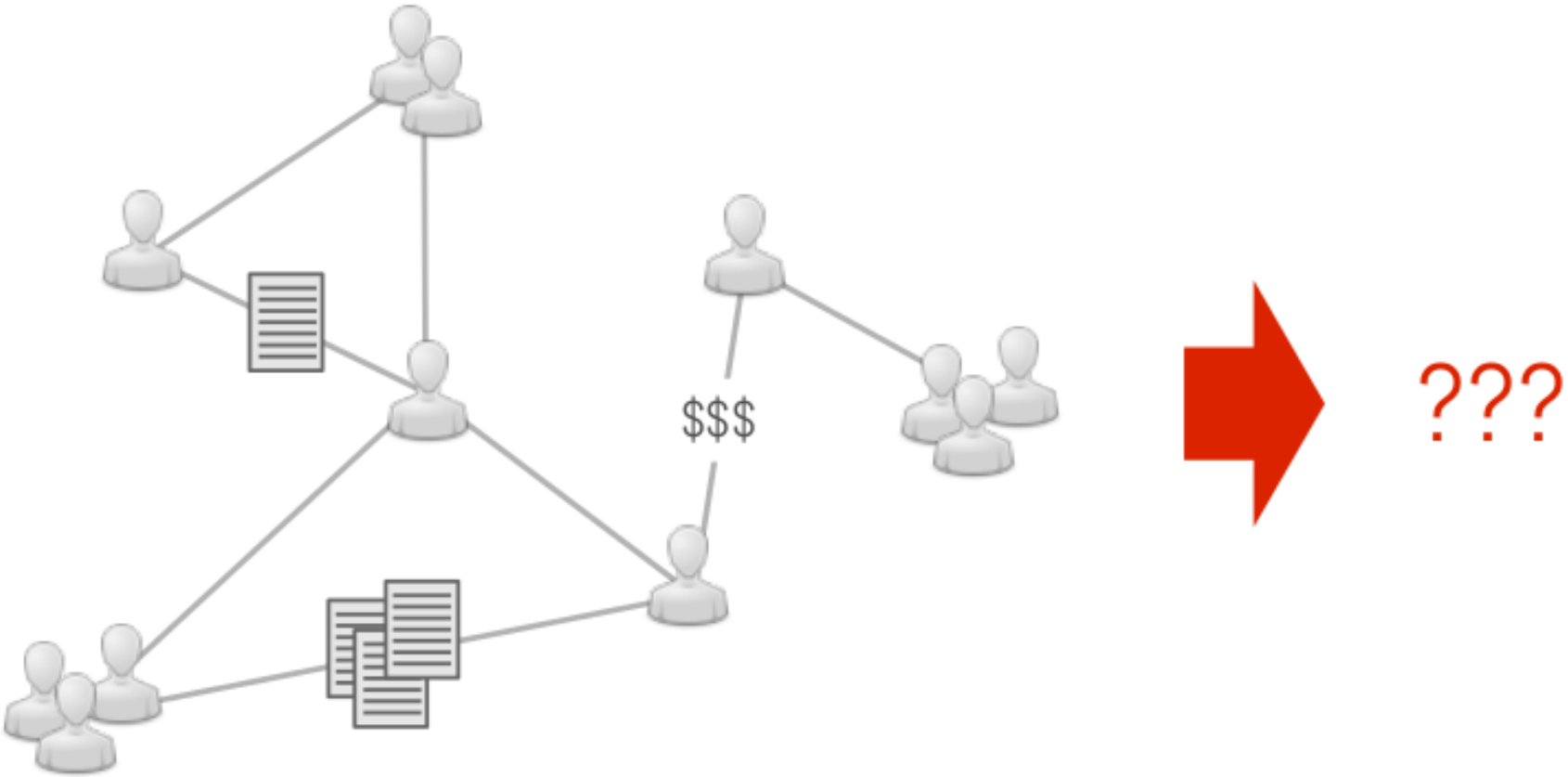
- "[C]omputer scientists may be interested in finding the needle in the haystack — such as […] the right web page to display from a search — but social scientists are more commonly interested in characterizing the haystack."

(King and Hopkins)

Exploration

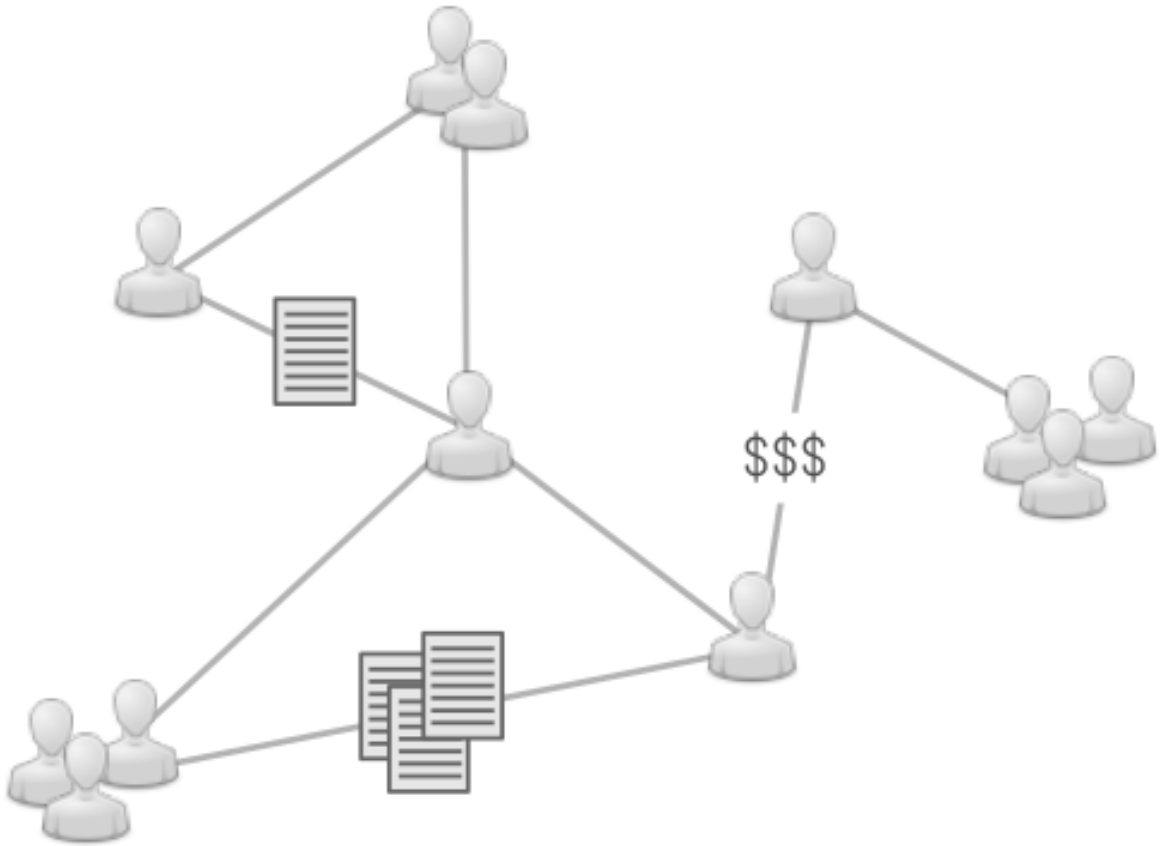


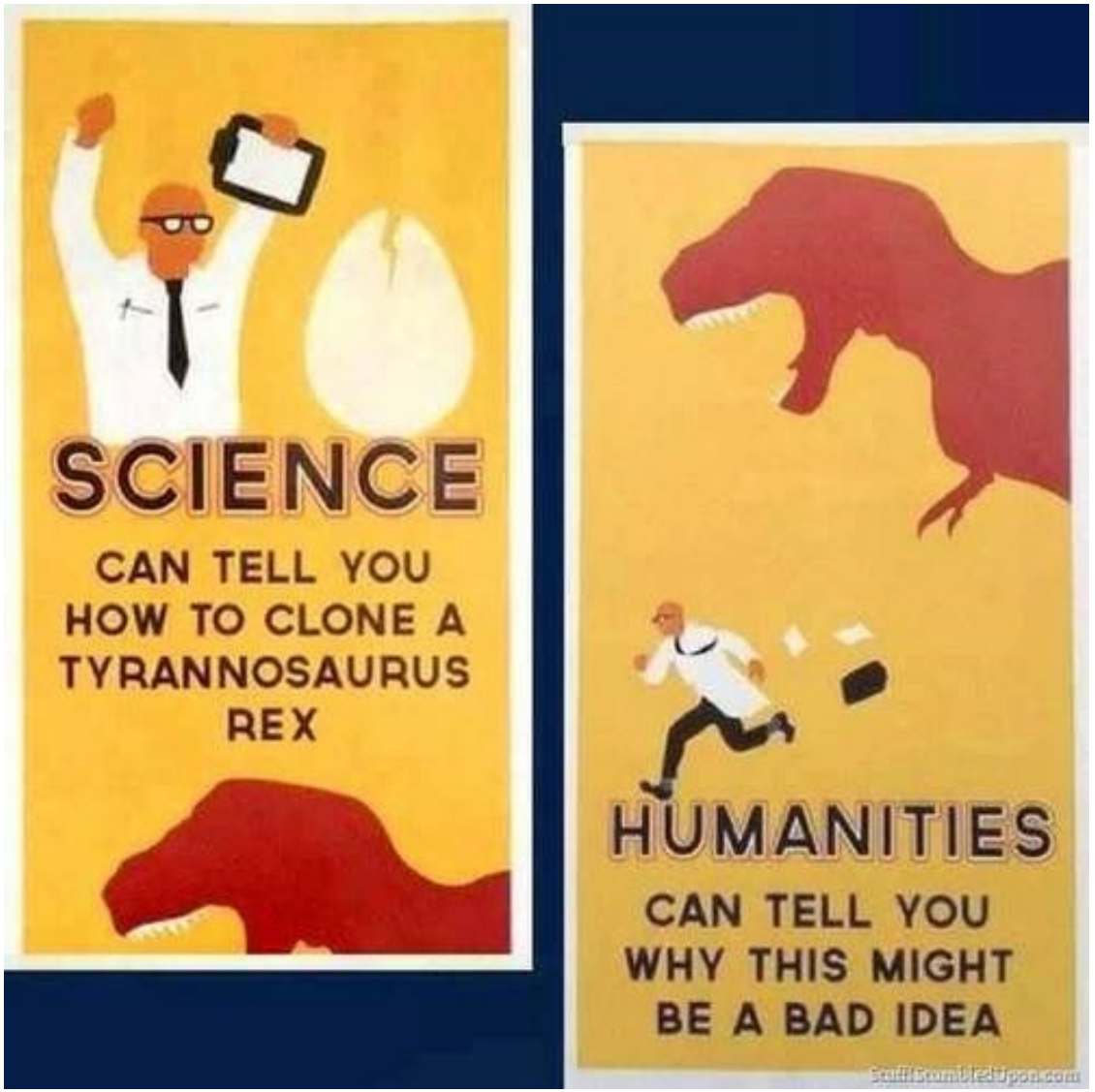
Prediction



Explanation

???




SCIENCE
CAN TELL YOU HOW TO CLONE A TYRANNOSAURUS REX
HUMANITIES
CAN TELL YOU WHY THIS MIGHT BE A BAD IDEA

Warning!

- We need to be careful not to abuse economic theory to rationalize our prior biased beliefs
- Models can often be set up so as to rationalize our preconceptions
- There is always an element of interpretation
- Need to be honest about what the data tells us
- Be aware of our scientific limitations

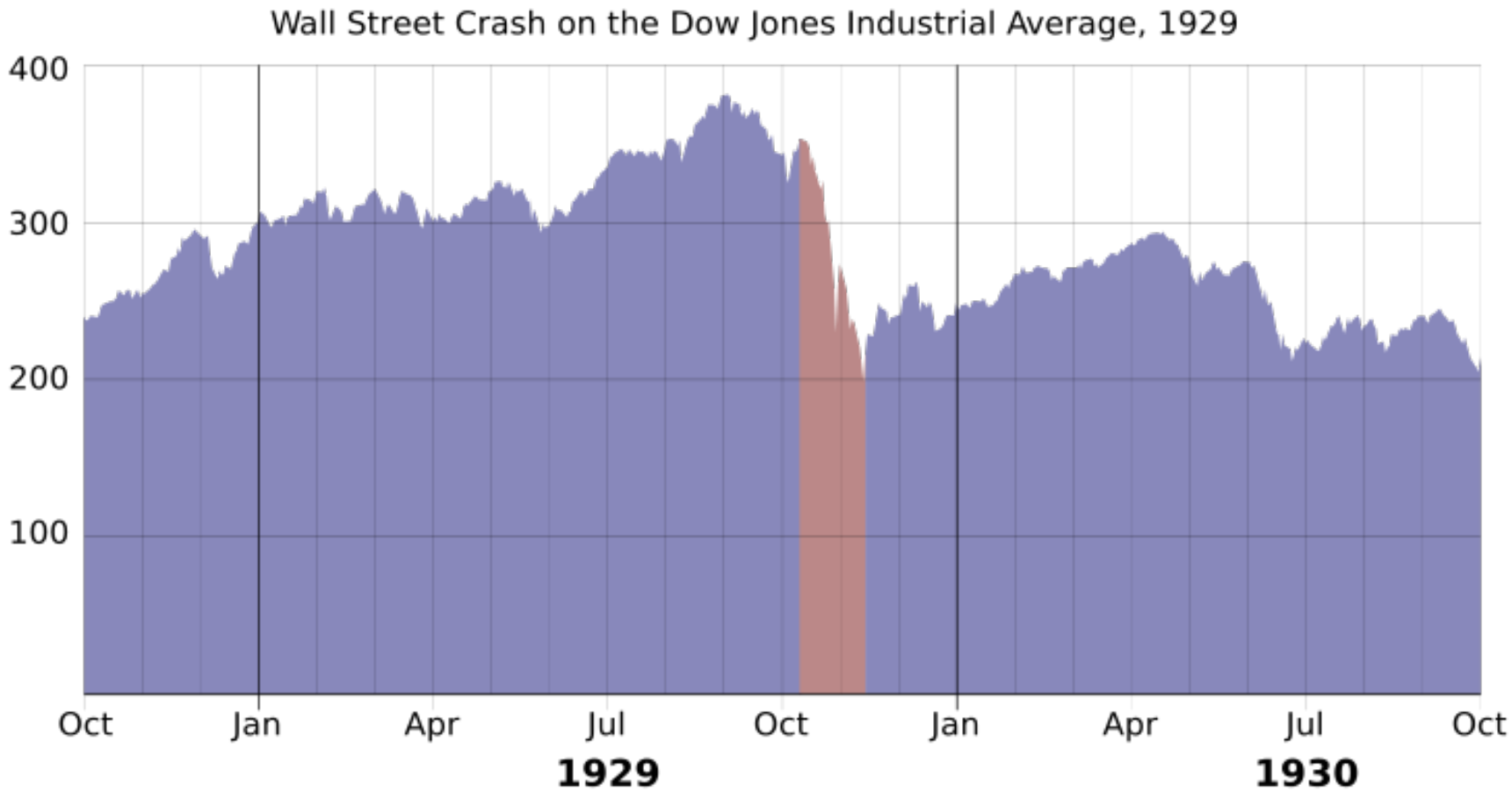
Wall Street 1920s (Dow Jones Industrial Average)

There is a danger in trying to rationalize patterns in the data.

Charles Dice

Evidence on the new "electrical economy"

Wall Street 1920s



Why are we (sometimes/often) wrong?

- Limited experience
  - Missing variables, unobserved heterogeneity
- Simplified models
  - Representative agent versus heterogeneous agents
  - Rational vs Irrational
- World is random and unstable
  - Structural breaks
- *We only observe one of many possible histories*
  - *Typically non-experimental data*
  - *Increasingly common to have access to large scale experimental data*

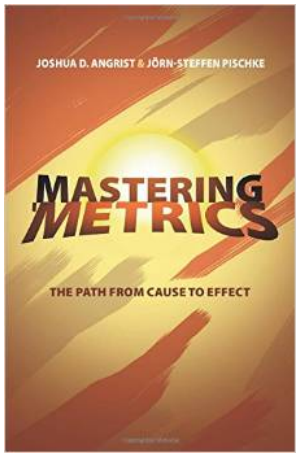
This class

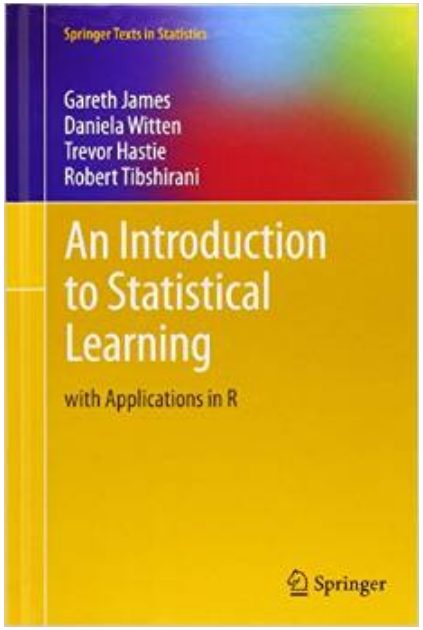
- **Professor: Matt Harding**
  Office: Rubenstein 196
  Office Hours: TuTh 9:30AM-10:00AM
  Email: matthew.harding@duke.edu
- TA: Danton Noriega-Goodwin
  Office: SSRI 230J
  Office Hours: TBA
  Email: danton.noriega@duke.edu

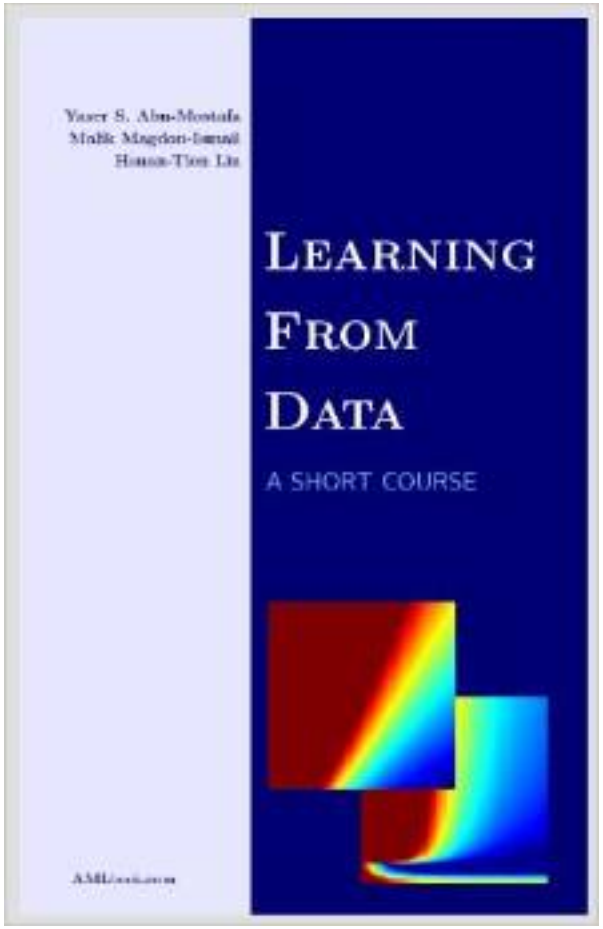
Pre-requisites

- introductory probability and statistics; basic familiarity with scientific programming. A number of different statistics courses will satisfy the prerequisites. Familiarity with basic concepts such as random variables, sampling, conditional expectations, mean regression, linear models, multivariate models, and binary variables is required.

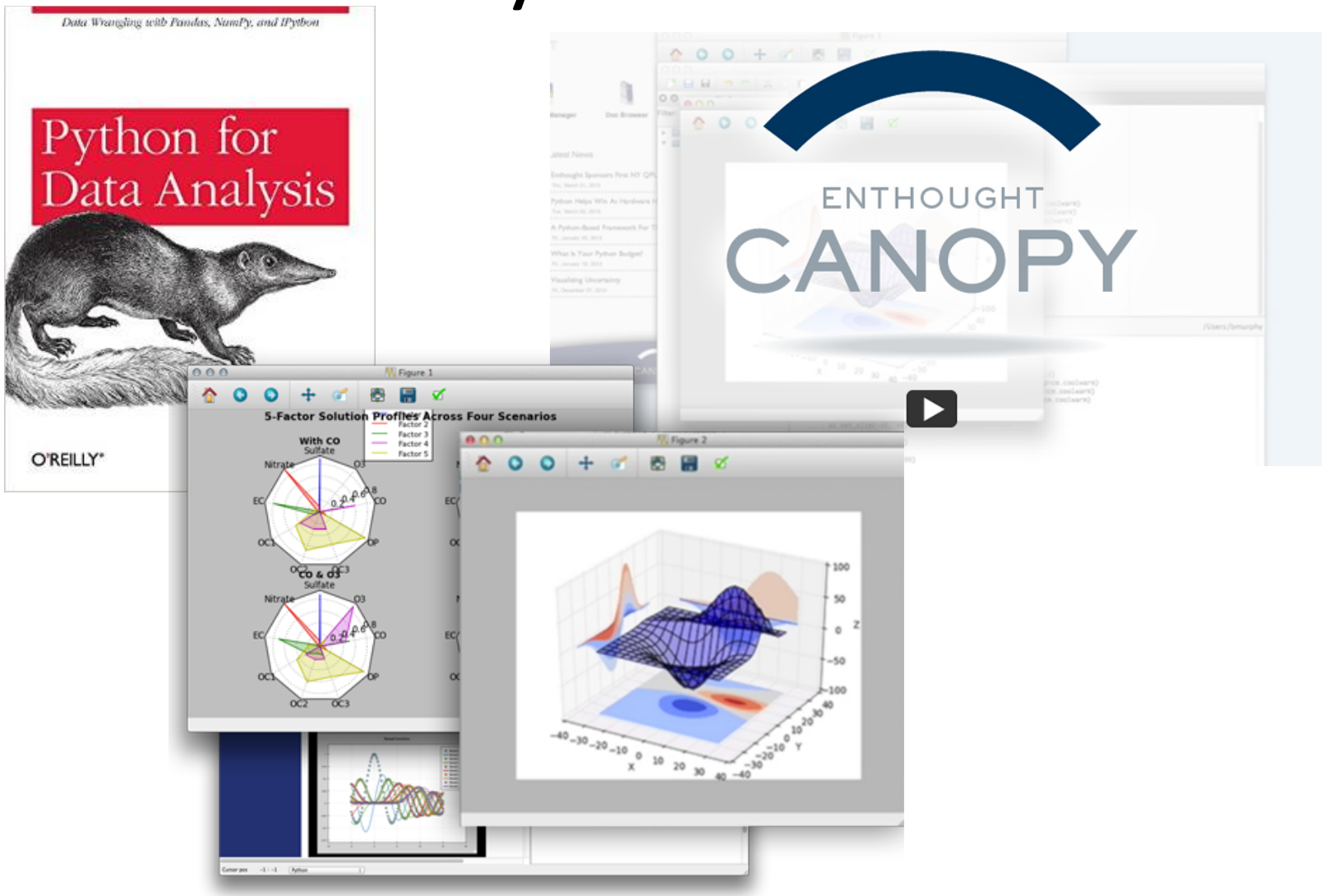
Books


JOSHUA D. ANGRIST & JÖRN-STEFFEN PISCHKE
MASTERING METRICS
THE PATH FROM CAUSE TO EFFECT


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer


Yaser S. Abu-Mostafa
Malik Magdon-Ismail
Hsuan-Tien Lin
LEARNING FROM DATA
A SHORT COURSE

Python



https://www.kevinsheppard.com/Python_for_Econometrics

Course grades

- Team projects: 40%
- Final project: 45%
- Class participation: 15%

Attendance

- Class attendance is mandatory. In fact the success of this class depends to a large extent on everyone's engagement in class teamwork. Moreover, class attendance will factor into the grade. In addition to traditional lectures, you will work in teams during class time to accomplish specific goal. If you do not attend class you will jeopardize your team's progress. If you have to miss class for valid reasons such as medical issues you are required to provide a written note.

Deadlines

- You will be clearly informed of all deadlines such as submitting assignments. Late submissions are not accepted. If you cannot make a deadline because of a valid reason such as an illness, you are required to provide an appropriate written note from an academic advisor or medical professional as soon as possible. You will then be offered the choice to reweigh the course grade, for example by adding more weight to Final project if you missed the deadline for submitting a required assignment. Again, please keep in mind that this may affect your team's performance. No make-up assignments are offered. You cannot miss submitting more than two (2) assignments. Please remember that extracurricular activities or job interviews do not represent valid reasons no matter how exciting you think they are!

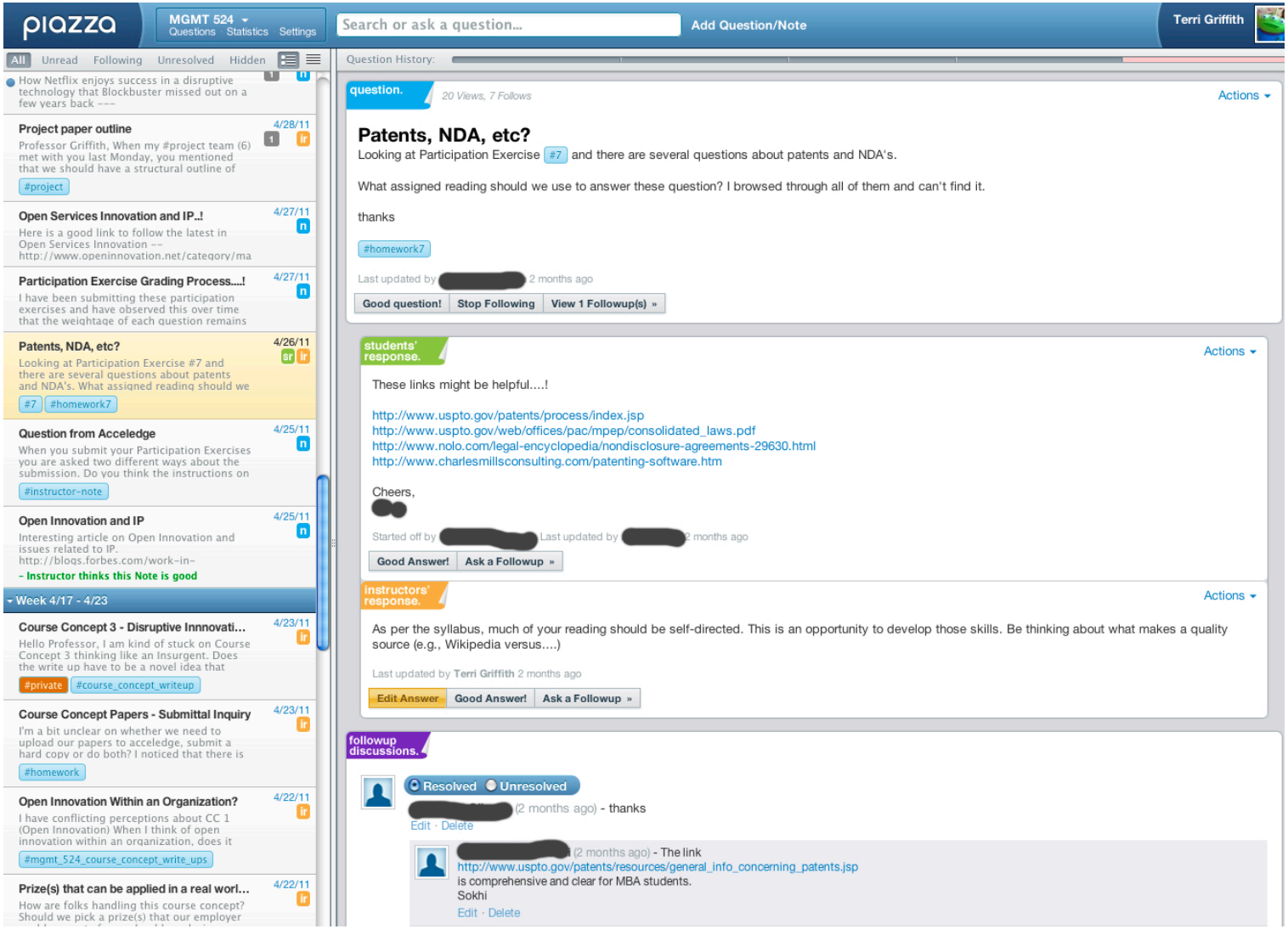
Re-grades

- If you think there was a mistake or oversight in grading an assignment you can contact the TA within 7 days of receiving the graded assignment. You need to explain in writing which precise issue that you have identified and why you want the assignment regarded. You should keep in mind that your entire assignment will however be re-evaluated, possibly by myself, so your overall grade may go up or down as a result. An additional review may well conclude with an even stricter interpretation of the grading policy than the first one. The second decision is then final.

Honor System

- This class is subject to the Duke Honor System: http://integrity.duke.edu/ugrad/
- You should be familiar with what is expected of you. Ignorance is not an excuse. Instances of academic misconduct will be reported, and subject to disciplinary action by the Office of Student Conduct: http://studentaffairs.duke.edu/conduct/undergraduate-disciplinary-system
- Please note if found in violation of current policies you will receive a failing grade for this class in addition to any other disciplinary action taken by the University.

https://piazza.com/duke/spring2015/pubpol590/home



Task management: www.workboard.com

The Big Cats

[1] The Jaguars

[2] The Tigers

[3] The Pumas

[4] The Leopards

The Not-so-big Cats

[5] The Ragdolls

[6] The Sand Cats

[7] The Sphynxes

[8] The Himalayans

Data: CER Smart Metering Project

- The *Commission for Energy Regulation* initiated the Smart Metering Project in 2007 with the purpose of undertaking trials to assess the performance of Smart Meters, their impact on consumers' energy consumption and the economic case for a wider national rollout

http://www.ucd.ie/issda/data/commissionforenergyregulationcer/

Tasks for next week

1. Data request form – complete and submit to get the data
2. Review materials on the CER website for the electricity pilot:
   - What is the pilot about?
   - What data was collected?
   - How was the pilot implemented?
   - What surveys were conducted?
   - Implementation issues and analysis challenges?
   - Any relevant literature on this program?

(Light) Reading

- Moritz Hardt (2014) "How big data is unfair: Understanding sources of unfairness in data driven decision making" https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de
- Hannah Wallach (2014) "Big Data, Machine Learning, and the Social Sciences: Fairness, Accountability, and Transparency" https://medium.com/@hannawallach/big-data-machine-learning-and-the-social-sciences-927a8e20460d