CS109a Data Science A

STAT 121a / AC209a / E 109A

Pavlos Protopapas
pavlos@seas.harvard.edu
Kevin Rader
rader@stat.harvard.edu
Weiwei Pan
weiweipan@g.harvard.edu

Outline

- What?
- Why?
- Who?
- How?

Outline

- What?
- Why?
- Who?
- How?

Employees' Choice Awards

Other Lists

Oddball Interview Questions
Best Jobs
Best Cities for Jobs

Trends

Additional Resources

Award FAQ

25 Best Jobs in America

Want a new job? Glassdoor is here to help, identifying the 25 Best Jobs in America for 2016. The jobs that make this list have the highest overall Glassdoor Job Score, determined by combining three key factors – number of job openings, salary and career opportunities rating. These jobs stand out across all three categories.

United States | 2016

1

Data Scientist

Job Openings	1,736
Median Base Salary	$116,840
Career Opportunity	4.1
Job Score	4.7

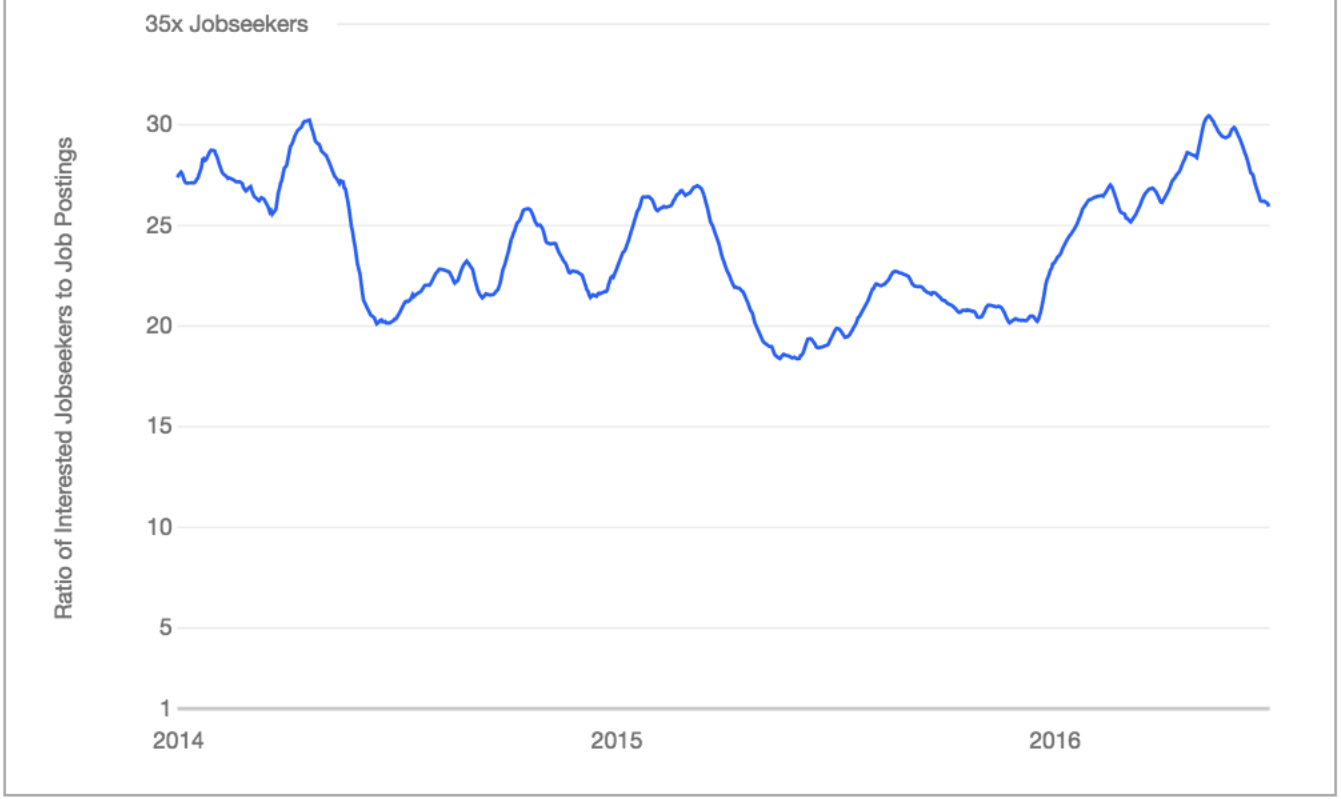
"Data Scientist" Job Trends

"Data Scientist" × | + Add Term | Find Trends

Scale: **Absolute** | Relative

Job Postings

Jobseeker Interest vs Job Postings



Find "Data Scientist" jobs

“By 2018, the US could face a shortage of up to 190,000 workers with analytical skills”

McKinsey Global Institute

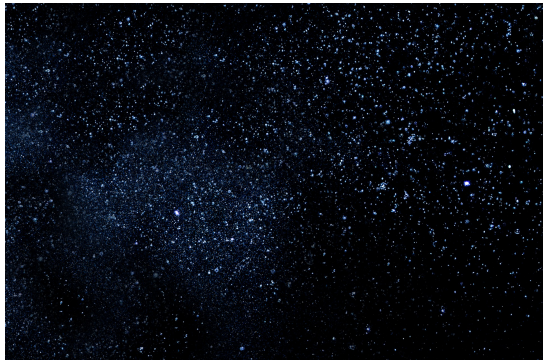
“The sexy job in the next 10 years will be statisticians.”

Hal Varian, Prof. Emeritus UC Berkeley Chief Economist, Google

Hal Varian Explains...

- The ability to take **data** – to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and **ubiquitous data**."

Long time ago (thousands of years): science was only empirical



People counted stars or crops and describe phenomena

Few hundred years: theoretical approach. Equations to describe general phenomena

1. $\nabla \cdot \mathbf{D} = \rho_V$
2. $\nabla \cdot \mathbf{B} = 0$
3. $\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$
4. $\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} + \mathbf{J}$

$$T^2 = \frac{4\pi^2}{GM} a^3$$

can be expressed as simply

$$T^2 = a^3$$

If expressed in the following units:

T Earth years

a Astronomical units AU (a = 1 AU for Earth)

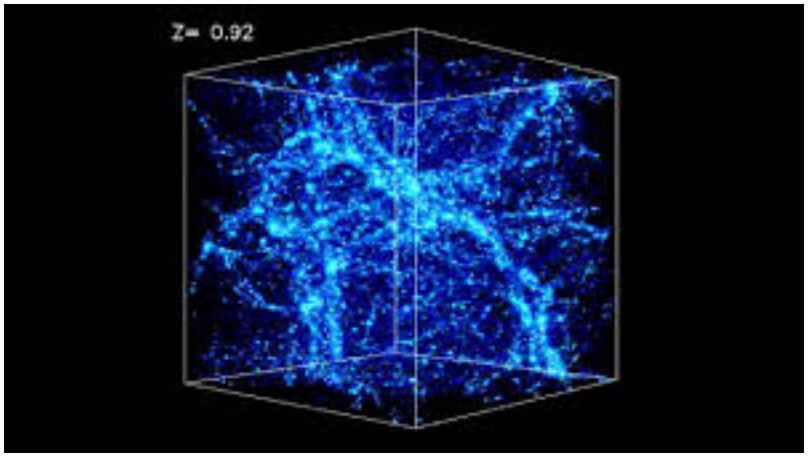
M Solar masses $M_\odot$

Then $\frac{4\pi^2}{G} = 1$

$$H(t)|\psi(t)\rangle = i\hbar \frac{\partial}{\partial t}|\psi(t)\rangle$$

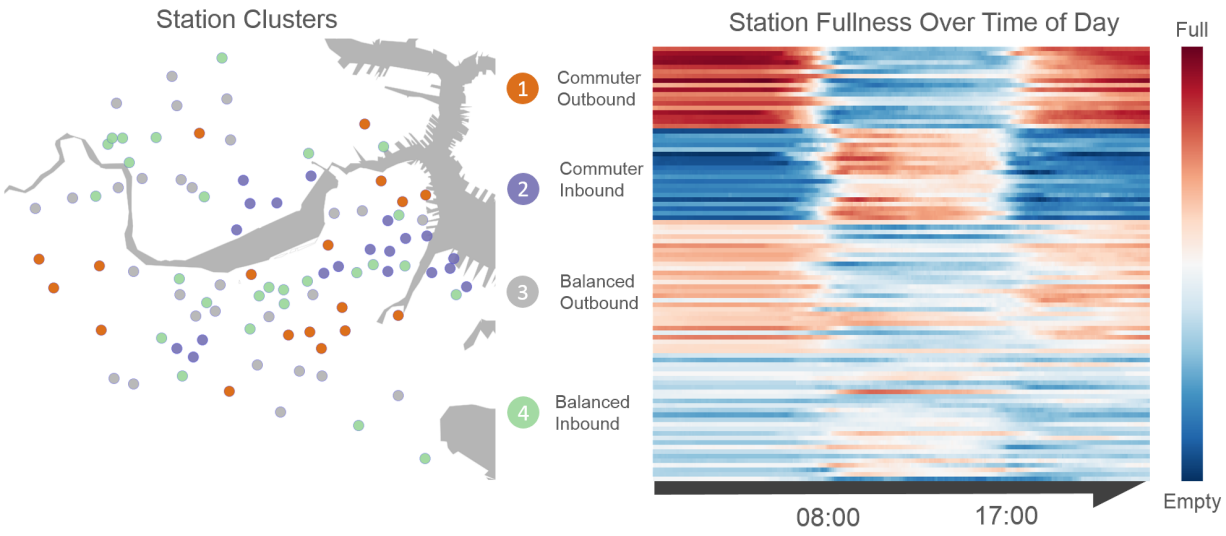
Last few decades:

Computational approach. Simulate complex phenomena



Today: **Data Science**

- Data are collected by sensors, instruments or simulators
- Processed by software (pipelines)
- Stored in computers in forms of databases
- Analyze data using statistics and machine learning

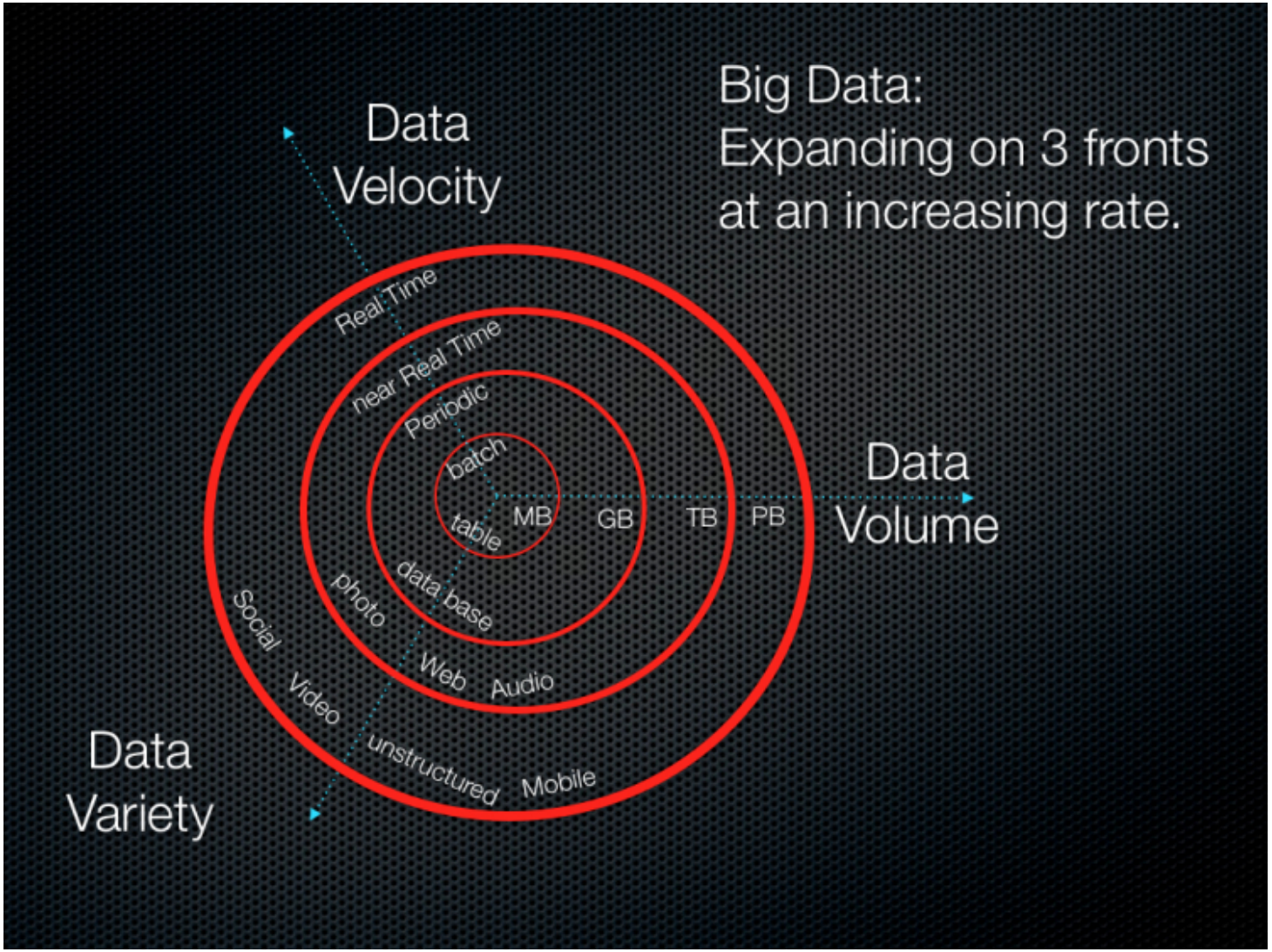


BIG DATA

Data … data … data …

“Between the dawn of civilization and 2003, we only created five exabytes of information; now we’re creating that amount every two days.”

Big Data


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

THE BIG V'S OF BIG DATA

Turning Information Overload Into Big Sales

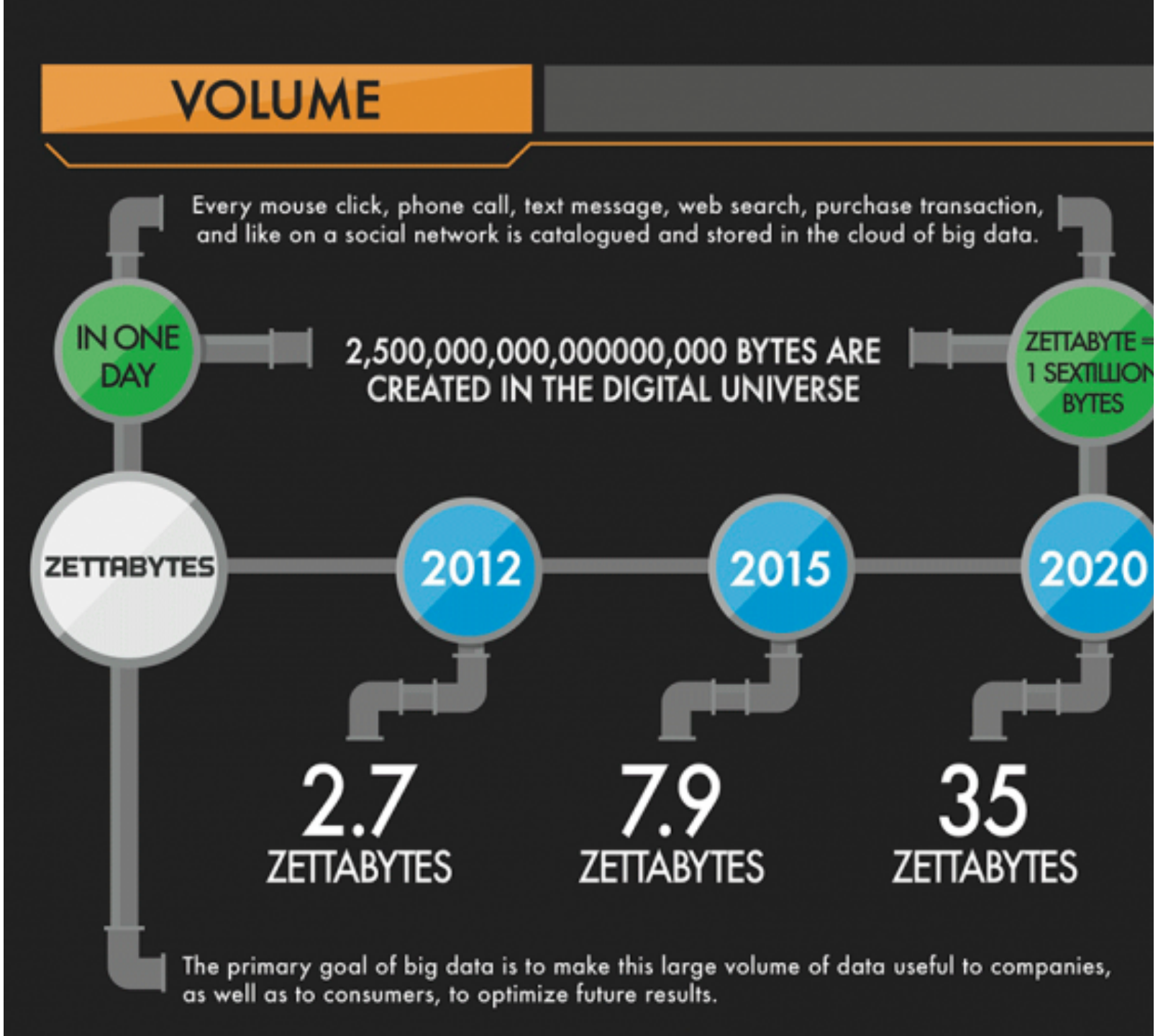
In the emerging market of Big Data, three "V" words have often been used to describe the issues at hand with information overload in our digital world.

THE EXISTING V'S

Big data has brought both great opportunity and change to the technological industry. Data scientists traditionally look at the existing V's, the ones that have classically been utilized to understand key variables of any data set.

VOLUME

Every mouse click, phone call, text message, web search, purchase transaction, and like on a social network is catalogued and stored in the cloud of big data.


VOLUME
Every mouse click, phone call, text message, web search, purchase transaction, and like on a social network is catalogued and stored in the cloud of big data.
IN ONE DAY
2,500,000,000,000000,000 BYTES ARE CREATED IN THE DIGITAL UNIVERSE
ZETTABYTE = 1 SEXTILLION BYTES
ZETTABYTES
2012
2015
2020
2.7 ZETTABYTES
7.9 ZETTABYTES
35 ZETTABYTES
The primary goal of big data is to make this large volume of data useful to companies, as well as to consumers, to optimize future results.

ZETTABYTES ZETTABYTES ZETTABYTES

The primary goal of big data is to make this large volume of data useful to companies, as well as to consumers, to optimize future results.

VARIETY

In today's multi-faceted Internet culture, the great volume of data is also extremely varied in its form. So many variables can be thrown at a company that the true value of information can often be lost in the sea of data.

PURCHASE TRANSACTIONS

WEBSITE TRAFFIC

REWARDS PROGRAMS

QUARTERLY BUSINESS REPORTS

TWITTER

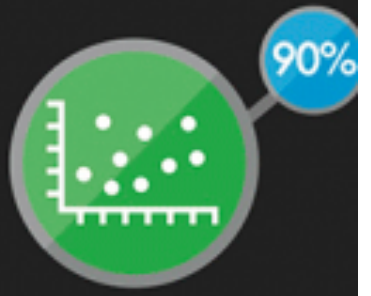
FACEBOOK

BLOG CONTENT

VELOCITY

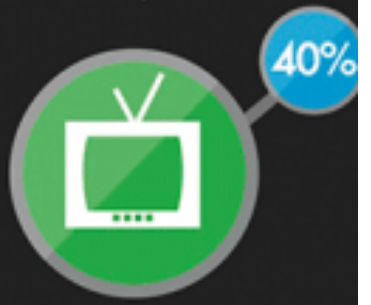
Information is being created at a faster pace than ever before. The varied channels of big data are each increasing their output of content, daily.

USERS GENERATE 2.7 BILLION LIKES ON FACEBOOK PER DAY



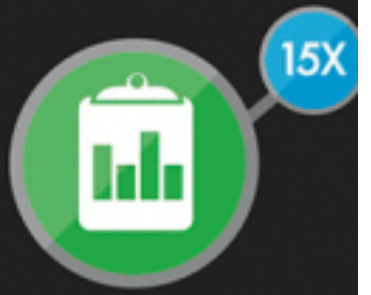
of the data in the world today has been created in the last two years alone

NEW TWEETS ARE CREATED BY ACTIVE USERS EACH DAY



40% of tweets are related to television and are beginning to be implemented in TV ratings

OF VIDEO IS UPLOADED TO YOUTUBE EVERY MINUTE



In 7 years, 15x the amount of data that exists today will be created every single year

The 10 Vs by Kirk Borne

- **Volume**: = lots of data (which I have labeled a "Tonnabytes", to suggest that the actual numerical scale at which the data volume becomes challenging in a particular setting is domain-specific, but we all agree that we are now dealing with a "ton of bytes").
- **Variety**: = complexity, thousands or more features per data item, the curse of dimensionality, combinatorial explosion, many data types, and many data formats.
- **Velocity**: = high rate of data and information flowing into and out of our systems, real-time, incoming!

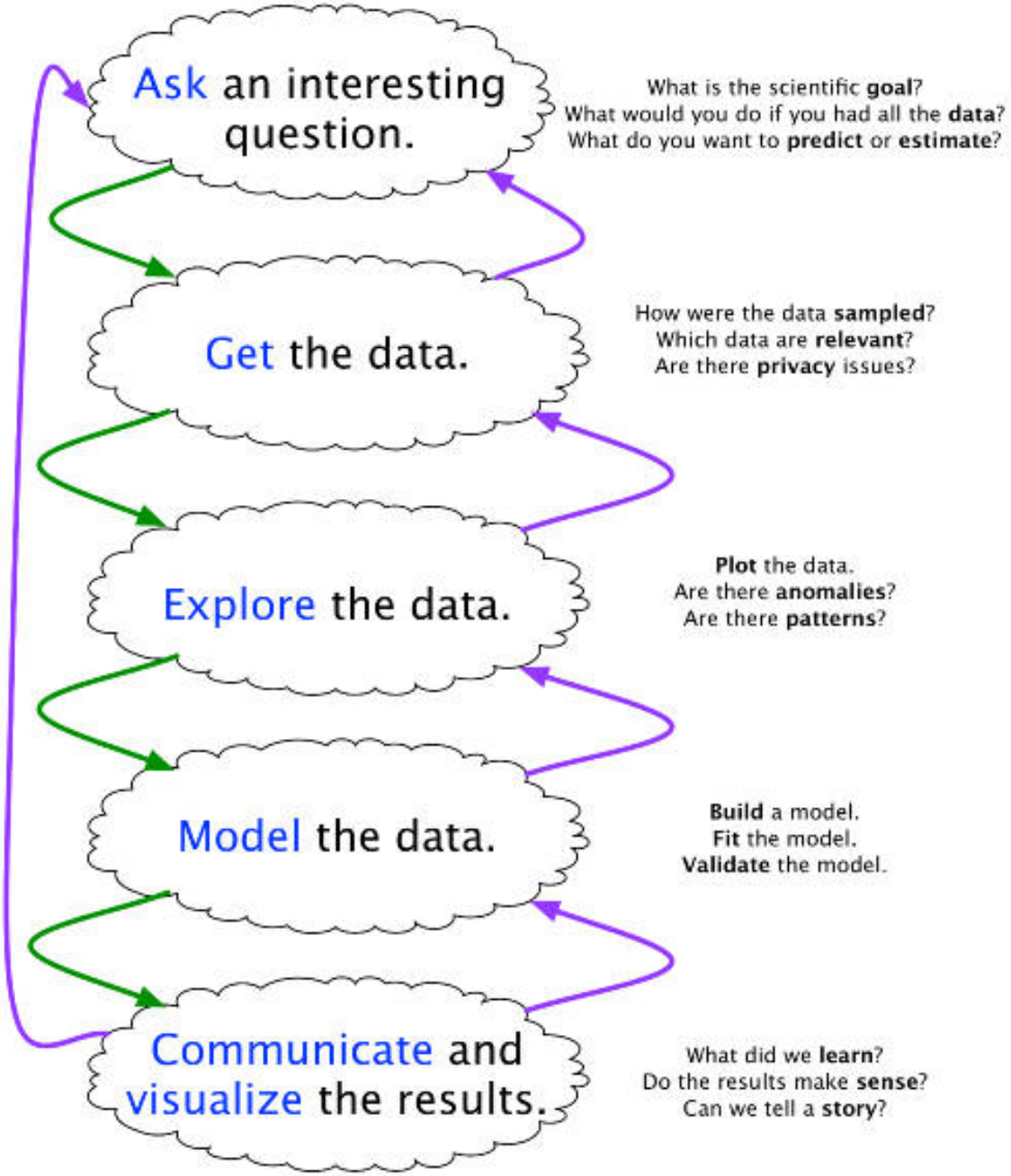
The 10 Vs Kirk Borne

- **Veracity**: = necessary and sufficient data to test many different hypotheses, vast training samples for rich micro-scale model-building and model validation, micro-grained "truth" about every object in your data collection, thereby empowering "whole-population analytics".
- **Validity**: = data quality, governance, master data management (MDM) on massive, diverse, distributed, heterogeneous, "unclean" data collections.
- **Value**: = the all-important V, characterizing the business value, ROI, and potential of big data to transform your organization from top to bottom (including the bottom line).
- **Variability**: = dynamic, evolving, spatiotemporal data, time series, seasonal, and any other type of non-static behavior in your data sources, customers, objects of study, etc.

The 10 Vs Kirk Borne

- **Venue**: = distributed, heterogeneous data from multiple platforms, from different owners' systems, with different access and formatting requirements, private vs. public cloud.
- **Vocabulary**: = schema, data models, semantics, ontologies, taxonomies, and other content- and context-based metadata that describe the data's structure, syntax, content, and provenance.
- **Vagueness**: = confusion over the meaning of big data (Is it Hadoop? Is it something that we've always had? What's new about it? What are the tools? Which tools should I use? etc.)
  Venkat Krishnamurthy (Director of Product Management at YarcData.

The Data Science Process



Outline

- What?
- Why?
- Who?
- How?

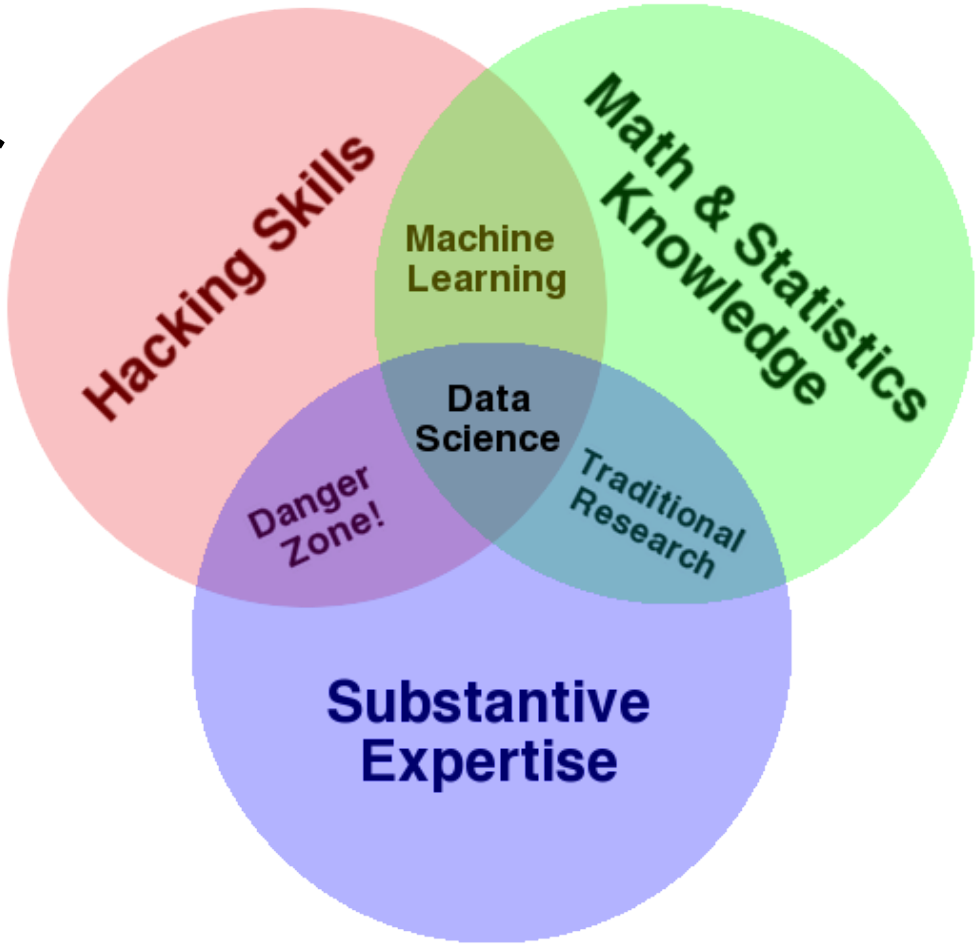
Data Science

To gain insights into data through computation, statistics, and visualization

Hanspeter Pfister, Joe Blitzstein, Verena Kaynig

Data Science

Computer Science



Statistics

Domain Science

Drew Conway

Data Scientist

Data Scientist Is

- “A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician.”

Josh Blumenstock

“Data Scientist = statistician + programmer + coach + storyteller + artist”

Shlomo Aragm

Data Scientist

- A new key player in organizations: the "data scientist." It's a high-ranking professional with the training and curiosity to make discoveries in the world of big data.

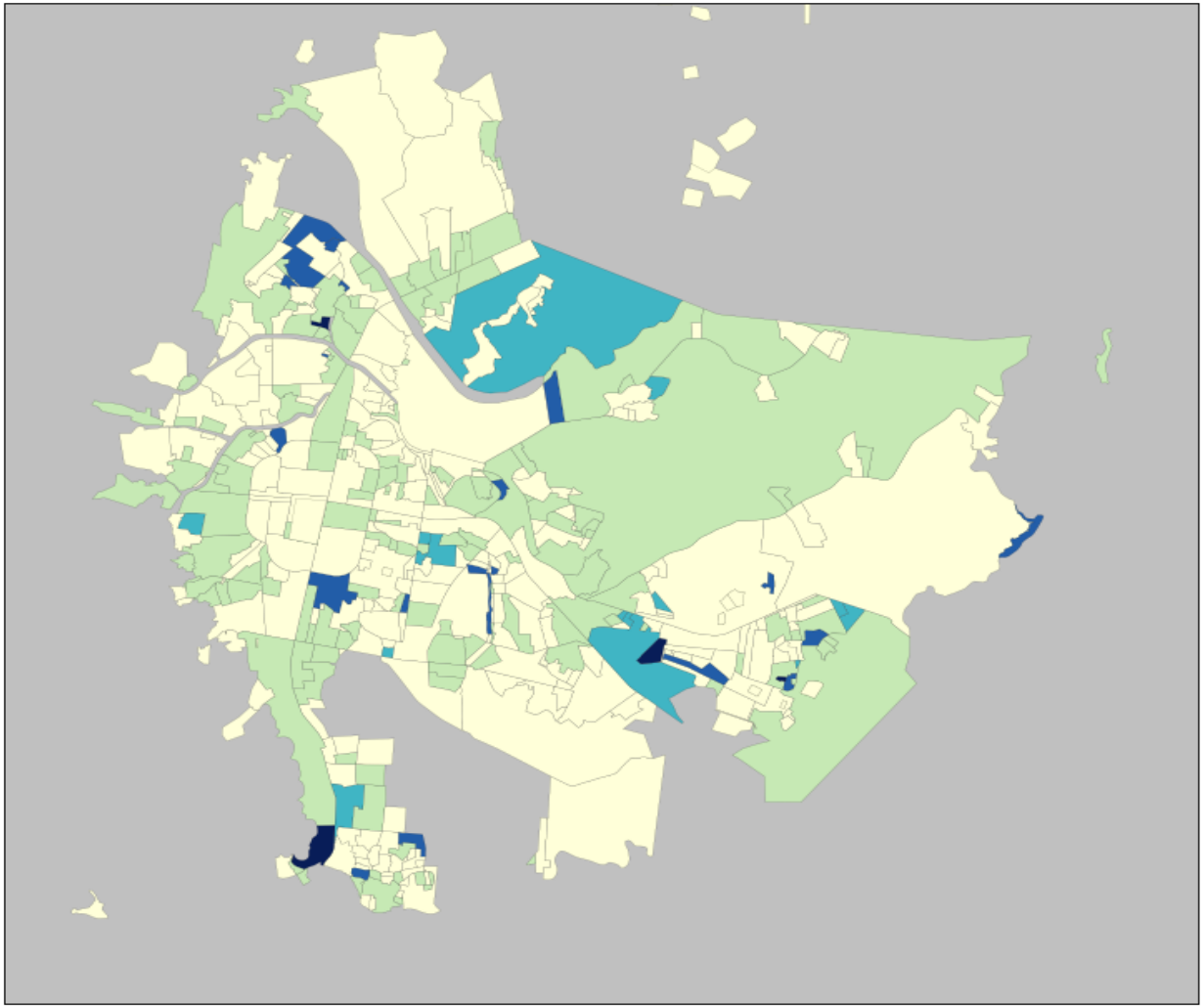
- The title has been around for only a few years.
  - (It was coined in 2008 by D.J. Patil, and Jeff Hammerbacher, then the respective leads of data and analytics efforts at LinkedIn and Facebook.)

- Nate Silver said, "I think data-scientist is a sexed up term for a statistician....Statistics is a branch of science. Data scientist is slightly redundant in some way and people shouldn't berate the term statistician."

WHO CARES

Can you solve these kind of problems?

Violence-Type Clusters



Neighborhood Type

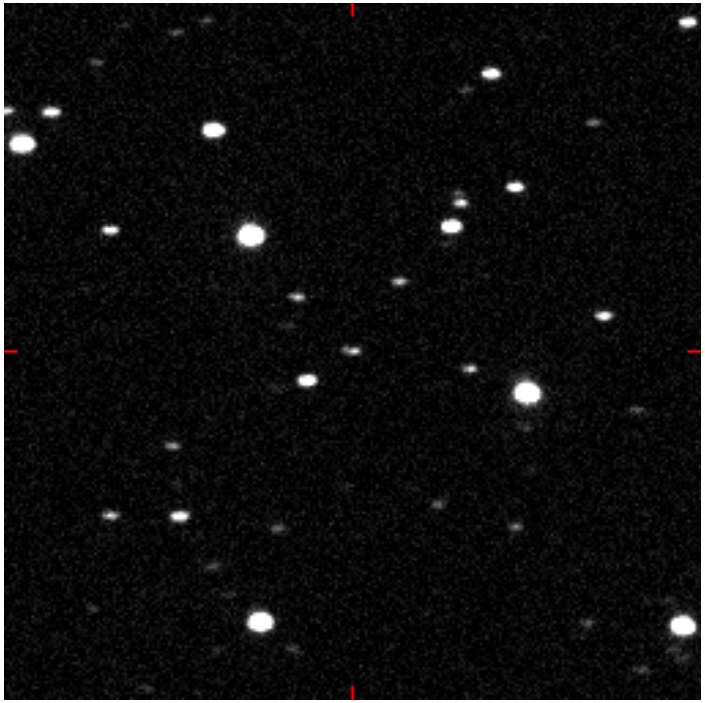
- non-violent
- violent—high rate
- violent—high absolute
- violent—both

violent neighborhoods are concentrated in four out of eight regions of San Pedro Sula

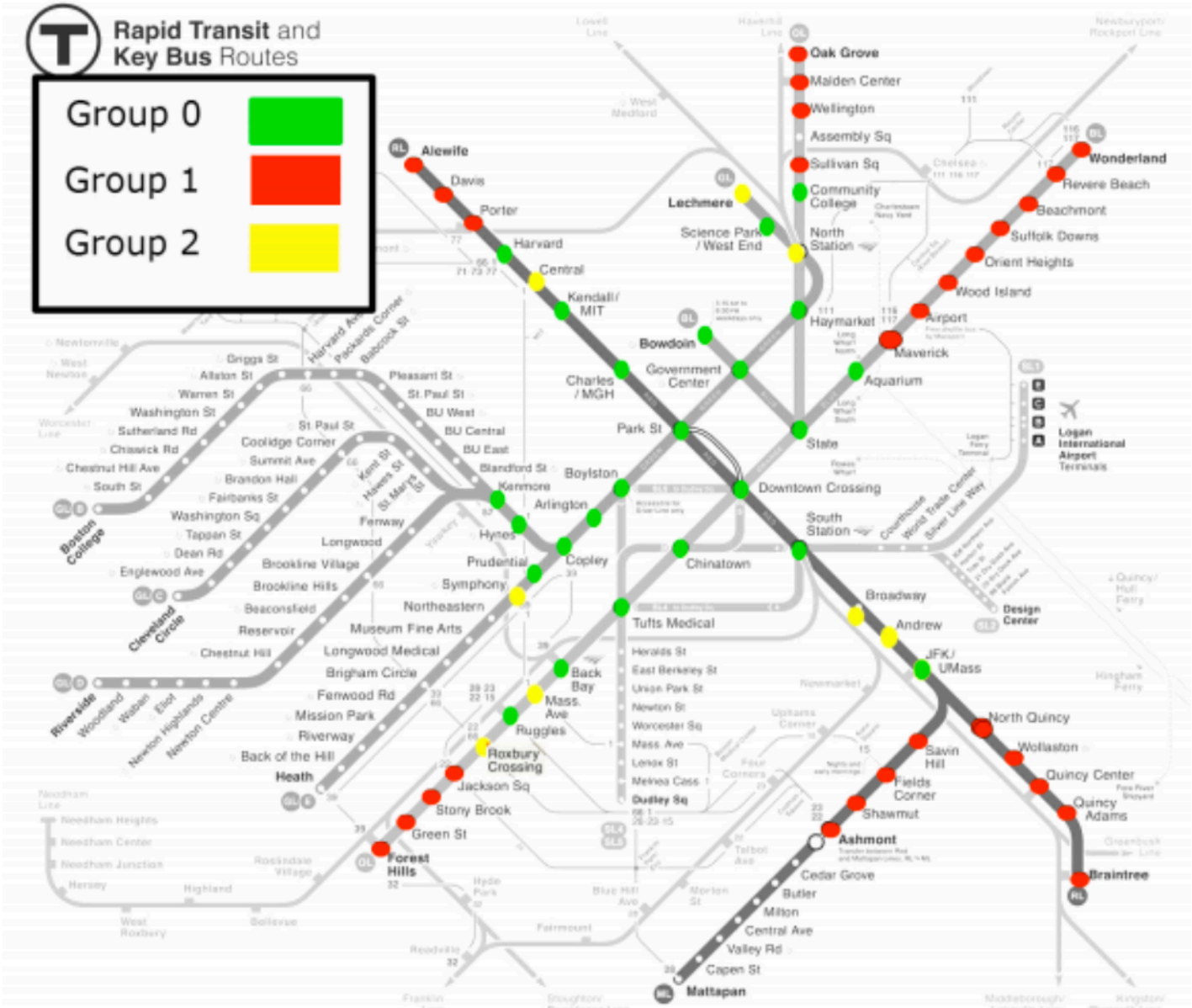
We Can Predict Future Homicides

Discovering Near Earth Objects

Capstone project (ME thesis 2016-2017)

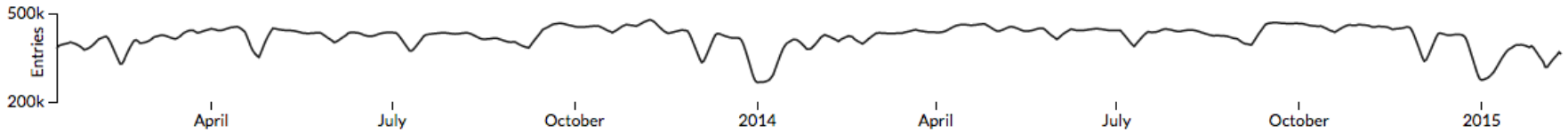


Correlations



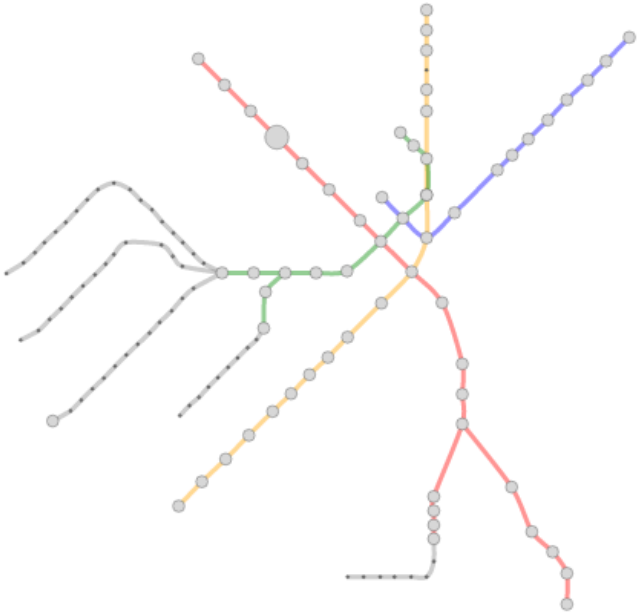
Ridership Timeline

*Days in Selection: **770***



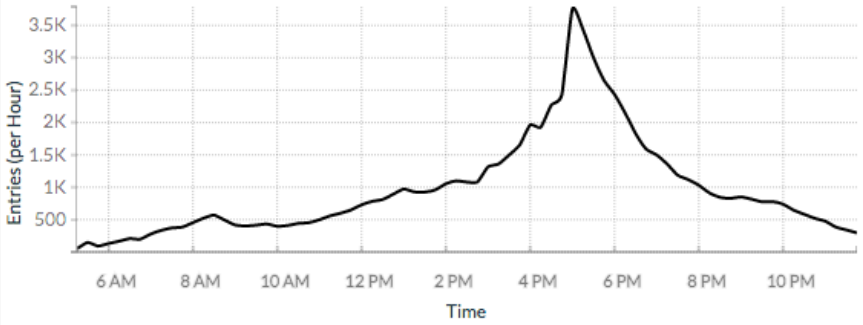
Station Selection

*Currently Selected: **Harvard***



Station Average Entries per Hour

*Currently Selected: **Harvard***



Condition Controls

Weekday | Weekend

None Trace 0-2 in. 2-4 in. 4-8 in. 8-15 in. 15+ in.

None Drizzle Rain

How do I know I am a data scientist

- know how to deal with data
- use many data sources
- understand how the data were collected (sampling is essential)
- understand what is important
- use statistical models (not just hacking around in Excel)
- understand correlations (e.g., states that trend similarly)
- think like a Bayesian and a frequentist
- have good communication skills (What does a 60% probability even mean? How can we visualize, validate, and understand the conclusions?)

kaggle Competitions Datasets Kernels Forums Jobs

11 active competitions Sort By Prize

Active All Entered Hosted Main Site All Eval Metrics

Predicting Red Hat Business Value

Classify customer potential

19 days to go · Featured

1,515 teams
1,347 kernels
$50,000

Bosch Production Line Performance

Reduce manufacturing failures

2 months to go · Featured

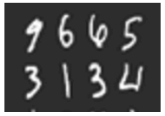
177 teams
$30,000

TalkingData Mobile User Demographics

Get to know millions of mobile device users

5 days to go · Featured

1,714 teams
2,684 kernels
$25,000



Digit Recognizer

Classify handwritten digits using the famous MNIST data

1,041 teams
5,816 kernels

Outline

- What?
- Why?
- Who?
- How?

Pavlos Protopapas

Scientific Director

Institute for Applied Computational Science (IACS)

pavlos@seas.harvard.edu

Expert in astrostatistics and particularly time series and astrophysics

Kevin Rader

Senior Preceptor

Stat Department

rader@stat.harvard.edu

sports analytics and applied biostatistical research specializing in survival data and large complex surveys

Weiwei Pan

Preceptor

Institute for Applied Computational Science (IACS)

weiweipan@g.harvard.edu

Machine learning, theoretical machine learning with applications

Harikrishna Narasimhan

Postdoc

Institute for Applied Computational Science (IACS)

hnarasimhan@g.harvard.edu

Eleni Kaxiras

eleni@seas.harvard.edu

Head TF

CS109a Staff

Xiaowen(Crystal) Chang
Siv Lu
Abhishek Malali
Qing Zhao
Joseph Song
Dylan Tan
Chae Christine Hwang
Yunhan Xu
Michael Farrell
Steve Klosterman

Zelong Qiu
Zona Kostic
Yoon Kim
Zeerak Ahmed
Reinier Maat Jacob
Kela Roberts

Jonathan Seitz
David Wihl

Outline

- What?
- Why?
- Who?
- How?

Main Ideas

- ***Data munging/scraping/sampling/cleaning*** in order to get an informative, manageable data set;
- ***Data management*** in order to be able to access data quickly and reliably during subsequent analysis;
- ***Exploratory data analysis*** to generate hypotheses and intuition about the data;
- ***Prediction/Inference*** based on statistical tools such as regression, classification, and clustering
- ***Communication*** of results through visualization, stories, and interpretable summaries

Topics

- Data Collection & Web Scraping Exploratory Data Analysis & Visualization
- K-Nearest Neighbors
- Linear regression (simple and multiple)
- Linear Model Regularization: Ridge & Lasso
- Principle Components & High Dimensionality
- Logistic Regression
- Bayesian Thinking
- Decision Trees, Random Forest, Boosting
- Support Vector Machine
- Experimental Design and Testing

How will you learn

- Lectures: Basic ideas and concepts with examples
- Labs: Examples and interactive learning
- Reading: 10-15 min before class
- Homework: Hands on and working with classmates
- Midterms: Preparation
- Projects:
- Office hours: TA and instructors

Is there a difference between CS109a and AC209a and STAT121a

CS109a and STAT121a are identical, each may satisfy a core requirement for different degree programs.

AC 209a is a graduate level course and involves completing extra course work which are more difficult or open-ended

- additional questions on HWs
- midterms

All students will have access to the extra problems on assignments and the project. If you're not enrolled in the graduate version your work will be credited (TBD)

Homework

- Eight homework

  Release: Wednesday morning 9:00am

  Due: Tuesday 11:59pm

  [see policies below]

- Each homework has:
  - One simple question
  - One real-world and/or open ended question
  - One challenge question (mandatory for AC209 students only extra credit for all others)

Homework schedule

HW #	Release	Due	Focus
HW0a	9/1	9/6	Python basics and libraries
HW1	9/7	9/13	Scraping and data exploration
HW2	9/14	9/20	KNN and linear regression
HW3	9/21	9/27	Cross validation and model selection
HW4	9/28	10/4	Model selection and dimensionality reduction
HW5	10/19	10/25	Logistic regression
HW6	10/26	11/1	LDA and Bayesian inference
HW7	11/2	11/8	Random Forest
HW8	11/9	11/15	SVM

Homework learning outcomes

Learn how:

- Scrape and wrangle messy data
- Apply sophisticated statistical analysis
- **Visualize and communicate results**
- **Interpret results**

Labs

- Guide+direct+solve one of the homework problems
- Introduce tools & skills
- Not mandatory but strongly recommended
- Two "**identical"** sessions

  Thursday 4:00pm-5:30pm @ red couch area (outside lecture hall)

  Friday 10:00-11:30am @ red couch area

Labs

- Video will be available only to DCE students
- Some material will also be posted for everyone
- First one this Thursday/Friday

Midterm

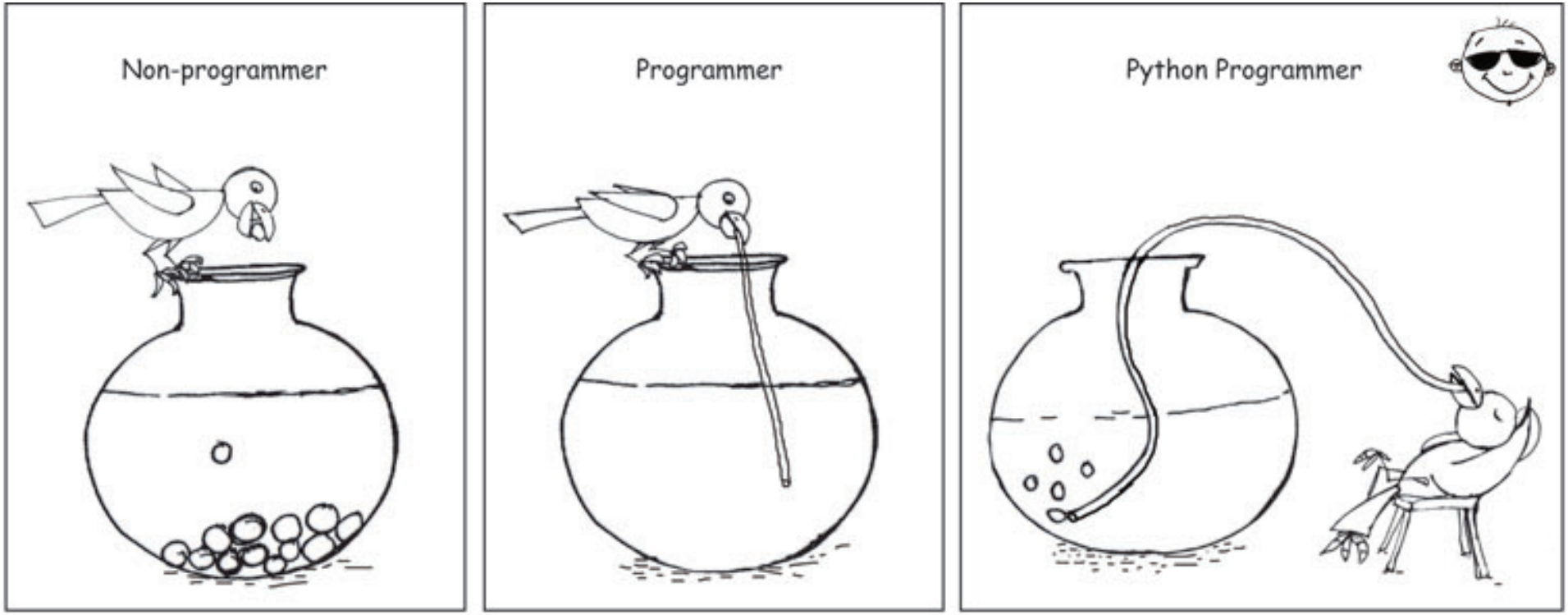
Each topic will be "touched" multiple times

- Lectures (more than once)
- Labs, Homework (sometimes multiple times)
- Midterm

Midterm 1: Take home (24 hour), Thursday, October 13 @ 9:00am

Midterm 2: Monday, November 21 @1-2:30pm

Programming



Python 2.7

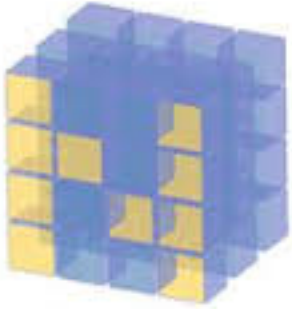
SciPy

BeautifulSoup

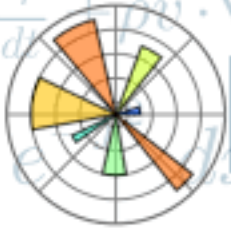
jupyter

IP[y]: IPython
Interactive Computing

scikits learn
machine learning in Pyt



NumPy



matplotlib

Projects

- **WE** will provide projects
- Work in teams (3-4 students per team)
- You work through the semester on 5 milestones before the actual final 3-4 weeks dedicated to projects
- One TF per project. TF will monitor the progress, guide and give ideas

Projects Schedule

Projects will be released: 9/12

M1: Form a team and choose projects (due 9/20)

Teams assign to projects: 9/23

M2: Literature review and reading (due 10/4)

M3: Data collection and data exploration (due 11/1)

M4: Base line model (due 11/15)

M5: Proposed work (due 11/15)

Projects due: 12/14

Poster+webpage

Predicting Demand for Ride-sourcing Services

Online ride-sourcing services such as Uber are gaining enormous popularity across the globe. A challenge in running these services efficiently is to accurately estimate the customer demand for a ride at a given location and time, and to predict surges in demand. With better demand estimates, the service provider can then allocate more drivers to locations where the demand is likely to go up. Can one use techniques from machine learning to analyze historical data of customer rides, and predict future demand for rides at a given location and time?

Milestones:

1. Project Selection: Form teams of 2 or 3 and select a project from the provided list.

2. Literature Study: Go through the following resources for background on the project and write a half to1 page summary for each one:

· Gonzales, E.J., Yang, J., Morgul, E.F., Ozbay, K., 2014. "Modeling Taxi Demand with GPS Data from Taxis and Transit", MNTRC Report 12-16, Mineta National Transit Research Consortium. URL:http://transweb.sjsu.edu/PDFs/research/1141-modeling-taxi-demand-gps-transit-data.pdf

· A previous project on this topic: https://github.com/ajsmith007/UberDemandPrediction

3. Data Exploration and Cleaning: The primary source of data for this project is a publicly available repository of around 4.5 million Uber rides in New York City from April-September 2014:

- http://data.beta.nyc/dataset/uber-trip-data-foiled-apr-sep-2014
- In addition, one can also make use of weather data recorded during this period: http://www.ncdc.noaa.gov/cdo-web/datasets
- and data about trips made by local taxi cabs outside the Uber network: http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

....

Projects Proposals

If you have a killer project that you absolutely feel you must do please send as an email in the next 3-4 days

Collaboration Policy

- You can discuss homework with other students (and with the instructor and TAs, of course)
- Work you turn in must be your own
- Indicate on your problem sets the names of the students with whom you worked
- Midterms are individual work
- Projects are 3-4 students team effort
- Harvard Honor Code

Grades

Homework 40%

Quizzes/Readings 10%

Midterm I 15%

Midterm II 15%

Project 20%

Piazza

- The purpose of piazza is for discussion amongst students
- Help each other
  - Please do not post solutions to homework problems
- Course announcements will be on canvas
- For questions regarding course it is best to send email to help-line
- We will monitor piazza but mainly for students

JupyterHub

Provide a Jupyter Notebook server

Jupyter Notebook is started for each student on the system when they log in

No need for installations

After midterm we will switch to anaconda

Instructions on how to login to be posted on canvas and demonstrated during lab

Prerequisites

Programming:

knowledge at the level of CS 50 or above,

Statistics:

knowledge at the level of Stat 100 or above

(Stat 110 recommended).

Is this course for me ???

Office hours

Monday and Thursday: 7:00-8:30pm (location TBD)
Tuesday: 4:00-5:30pm

Common CS office hours
Tuesday 6:00-8:00pm

Other office hours

- Pavlos: Northwest Building B155, Thursday 1-2:00pm
- Kevin: SC 614, Tuesday 1:00-2:00pm
- Weiwei: Thursday 5:30-7:00pm

Policies

- HWs due on Tuesdays, 11:59 pm EST
  4 late days for HW (no questions asked)
- Cannot submit HW later than 2 days
- Holistic grading (0-5) with a lot of comments on notebooks via vocareum

email help-lines for the course.

Send all questions regarding grading to:

cs109A+grading@gmail.com.

All questions regarding course materials and requests for extra assistance should be sent to:

cs109A+help@gmail.com.

Include your name and affiliation in the email

Next

- HW0 is mandatory: needs to be submitted but it will not be graded - (solutions will be provided)
  - **HW0a is a placement test for DCE students (passing grade TBD)**
  - HW0b very good introduction to python, libraries etc
- Read syllabus carefully
- **Keep an eye on canvas for the next few days**
  - Instructions for preparing for tomorrow's lab
  - Follow instructions on jupyterhub that will be posted on canvas
  - Office hours locations

How to have fun?