CS102: Big Data

Tools and Techniques, Discoveries and Pitfalls

Spring 2017

Ethan Chan, Lisa Wang

Lecture 1: Course Overview

Teaching Team

Instructors:

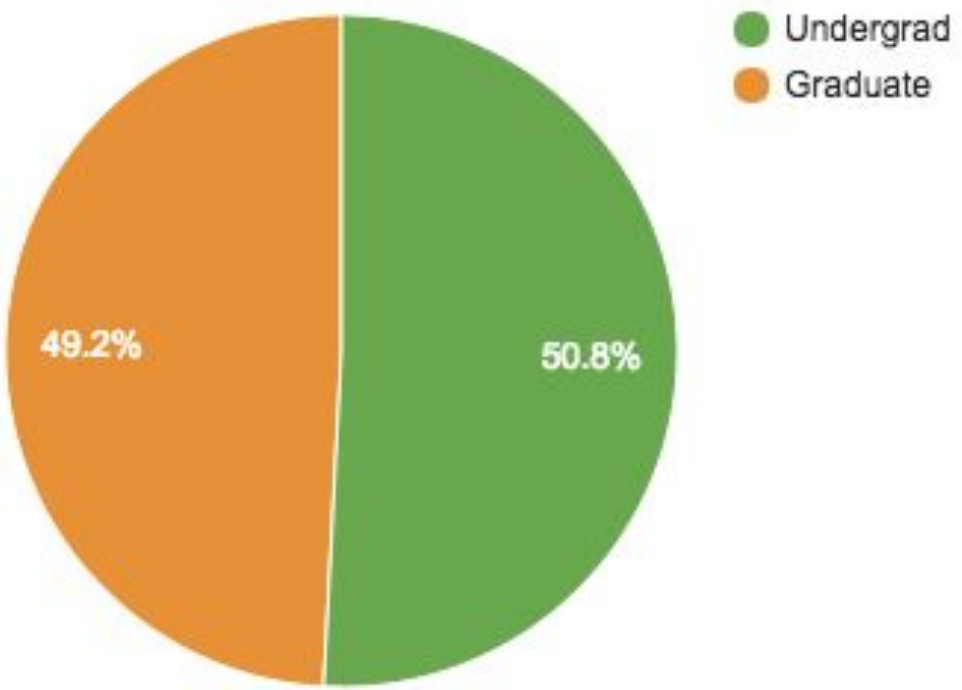
- Ethan Chan - *ethancys@stanford.edu*
- Lisa Wang - *lisa1010@stanford.edu*

Staff Mailing List:
cs102-spr1617-staff@lists.stanford.edu

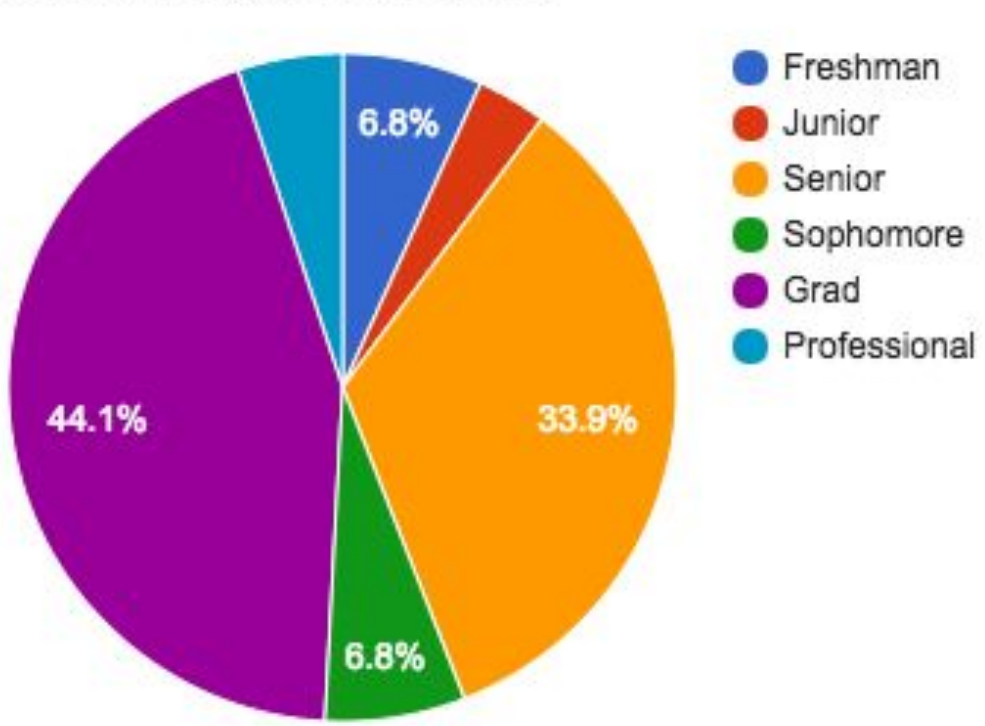
Who is this class for?

- Non-CS Majors
- Undergraduates or graduates
- Not afraid of numbers
- Not afraid of computer tools
- Took equivalent of one programming class at the level of CS106A
- Patient and tolerant..

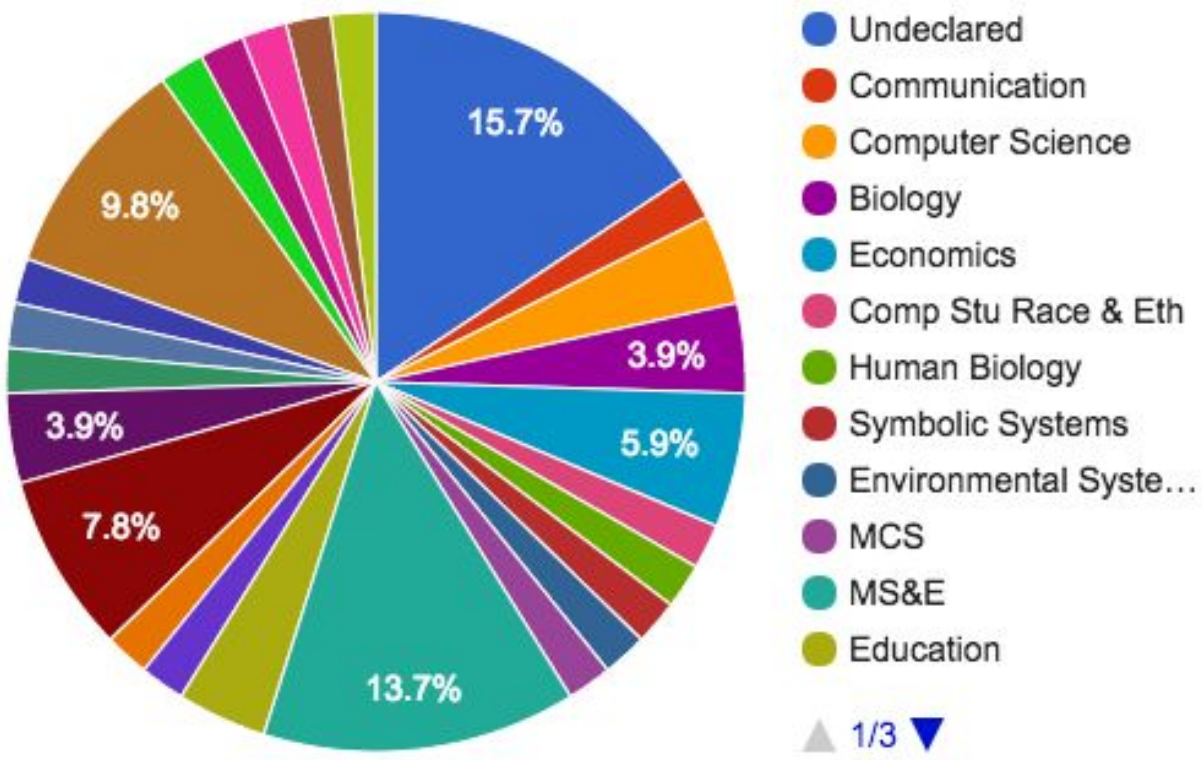
Breakdown by Student Type



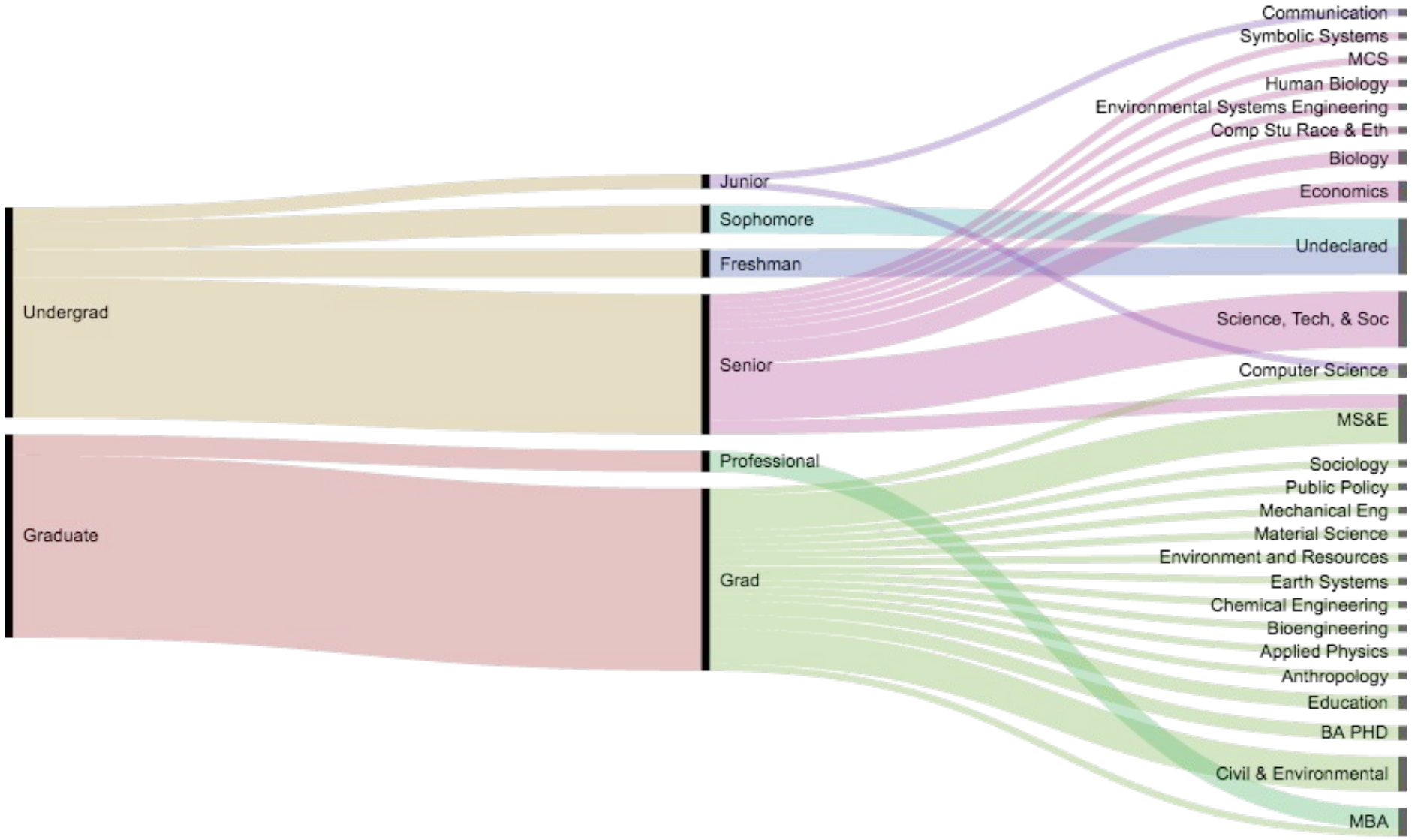
Breakdown by Student's Year



Breakdown by Major



Alluvial Chart



Hierarchical Bubble Chart

Today's Agenda

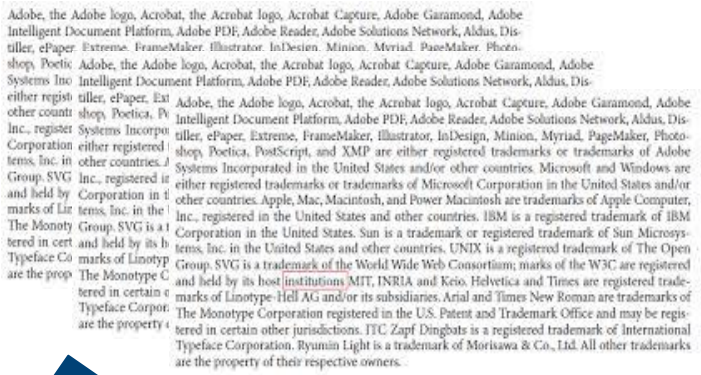
1. Defining Big Data
2. Tools and Techniques
3. Big Data Discoveries
4. Pitfalls
5. Data Visualization
6. Logistics

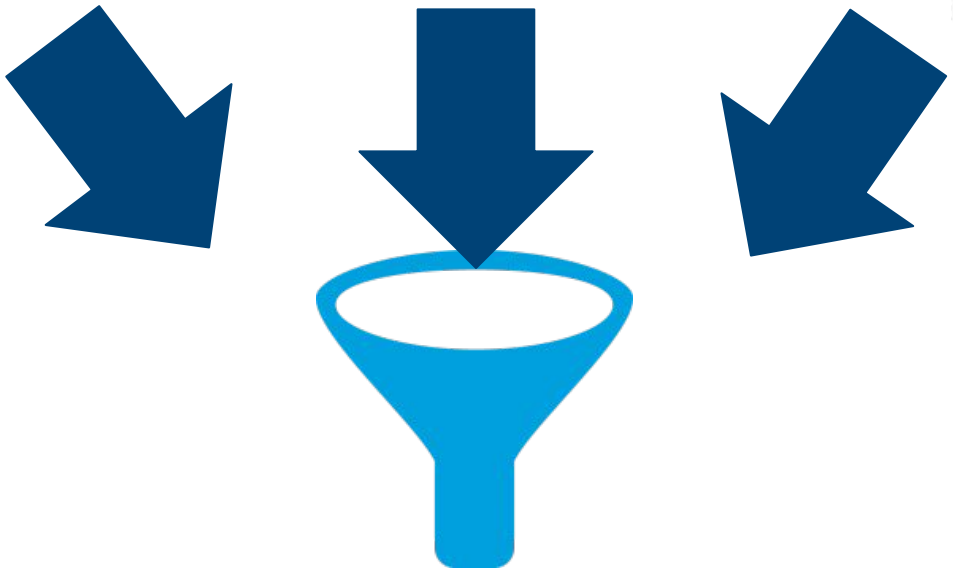
Defining Big Data

What is Big Data?

What does “Big Data” mean?

A) Collecting Large Amounts of Data





What does “Big Data” mean?

B) Making Sense of the Data

A) Collecting Large Amounts of Data

Big **V**olume

Need to collect huge amounts of data

Big **V**elocity

Need to collect incoming data in real time

Big **V**ariety

Need to collect many different types of data

B) Make sense of the data

What happened? (*descriptive*)

Why did it happen? (*diagnostic*)

What will happen? (*predictive*)

How can we make it happen? (*prescriptive*)

B) Make sense of the data

What happened? (*descriptive*)

”What is the demographic of this CS102 class?”

Why did it happen? (*diagnostic*)

”Why did the no. of enrollments increase ytd?”

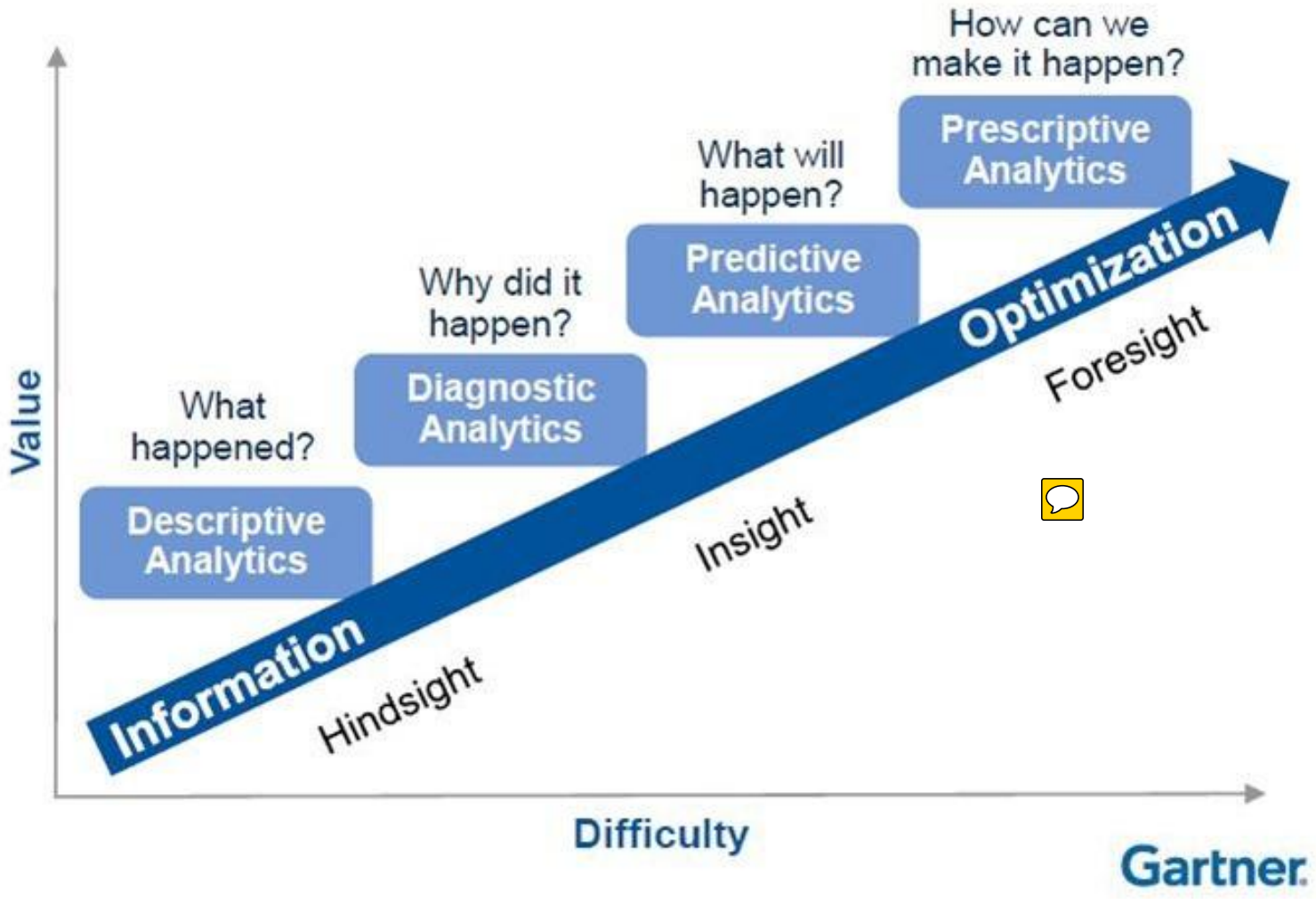
What will happen? (*predictive*)

”Who will score an A on this class?”

How can we make it happen? (*prescriptive*)

”Use slides or demos to teach the material?”

B) Make sense of the data



Tools and Techniques

What tools and techniques do you use to make sense of data?

Tools & Techniques You Will Learn

- Spreadsheets Data Analysis *(Google Sheets)*
- Relational Databases *(SQL)*
- Data Mining / Pattern Recognition *(Python, Pandas)*
- Machine Learning *(Python, Scikit-learn)*
  - Supervised, e.g. classification
  - Unsupervised, e.g. clustering
- Data Visualization *(Sheets, RawGraphs, Python, Tableau)*

Spreadsheet Data Analysis

E.g. Schoolkids data

Each row corresponds to a student and has attributes:

- Gender - The student's gender (boy or girl)
- Grade - The student's grade level (4, 5, or 6)
- Age - The student's age
- School - The student's school name
- Grades - A ranking between 1 (most important) and 4 (least important)
- Looks - A ranking between 1 and 4

Example question:

- Find the school with the highest girl-to-boy ratio

Relational Databases

- More flexible and scalable solution, industry standard for storing and manipulating data
- You will learn how to query databases with multiple tables (e.g. table for student, table for school)
- Example question:
  - Count how many 4th graders think Looks are more important than Grades for each school.

Data Mining

- Data mining refers to the *extraction of patterns and knowledge* from data, not the *extraction of data* itself.
- E.g. people who buy peanut butter often buy jelly as well. (Market Basket Analysis)

Supervised Machine Learning

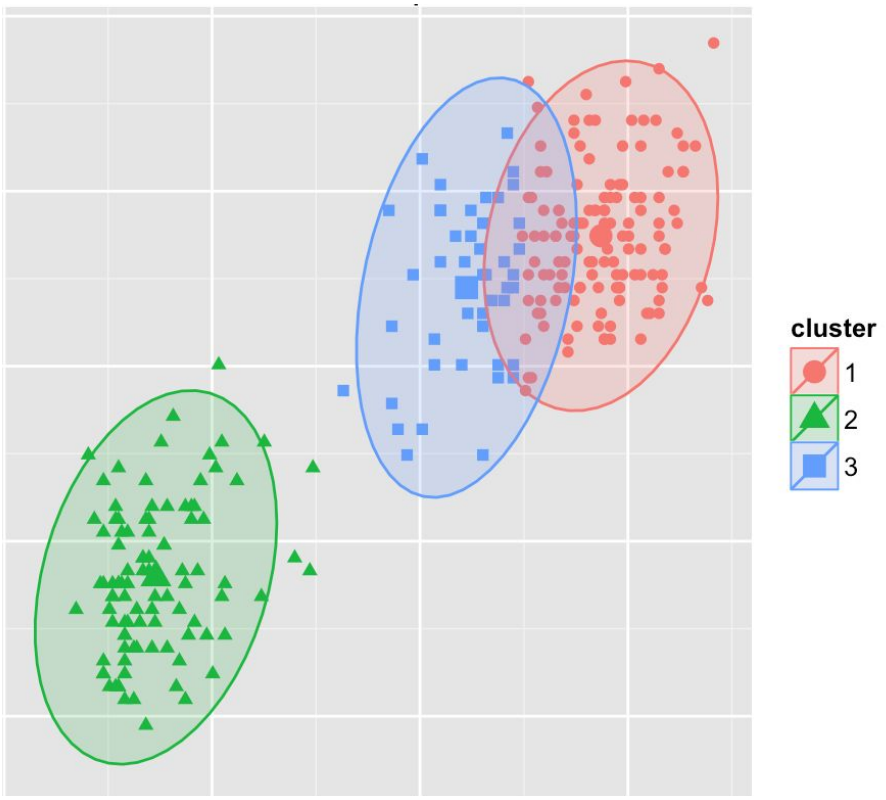
Given a labeled data set of (x,y) pairs, where x is the data and y is the label, learn a mapping from $x \rightarrow y$.

E.g. Movie review sentiment analysis.Given previous movie reviews and their labels (pos or neg), predict the sentiment of a new review.

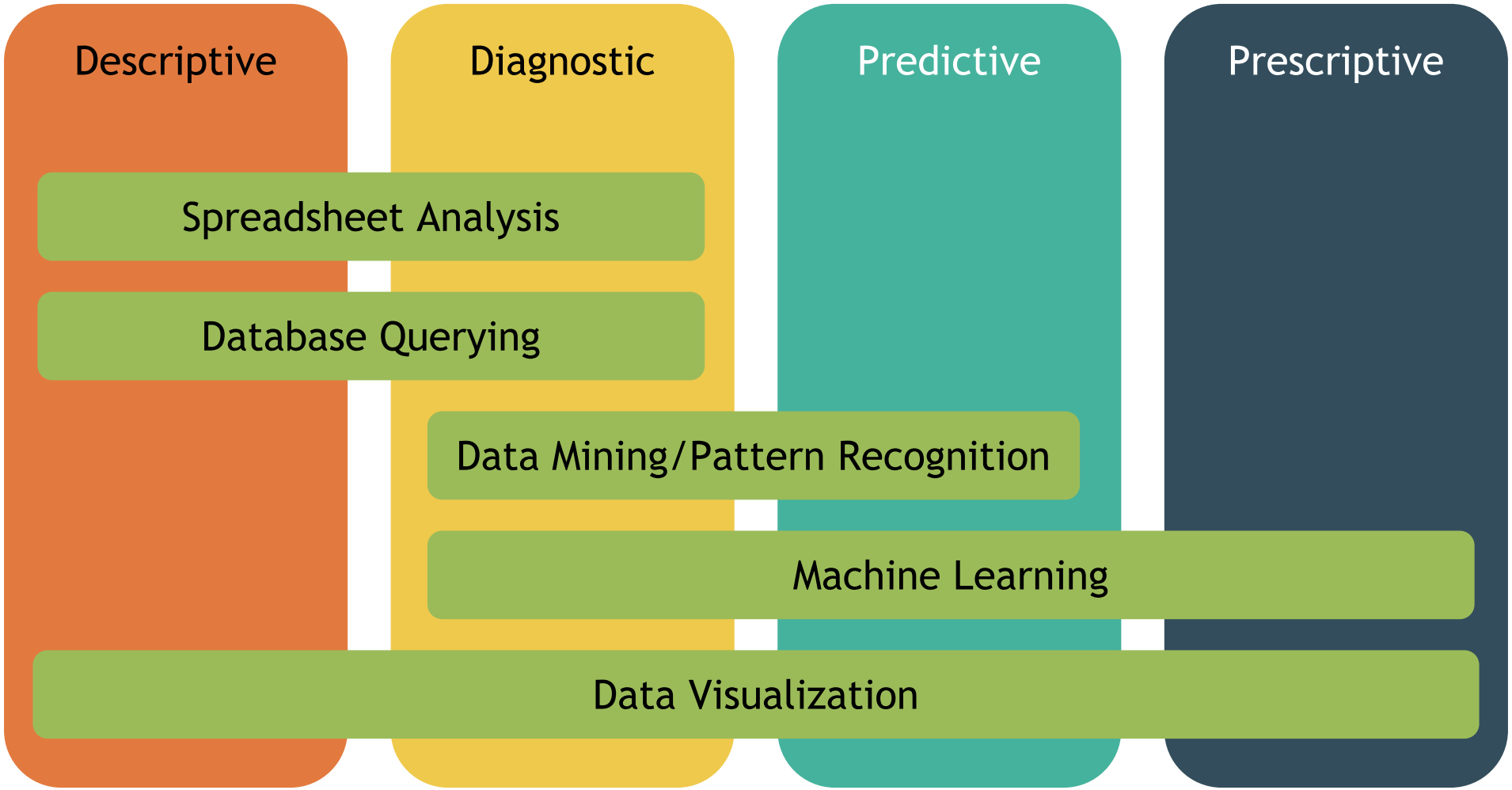
- Regression
- Classification

Unsupervised Machine Learning

- Clustering
- Example question:
  - Figure out which students are similar based on learning behavior



Tools & Techniques



Big Data Discoveries

Exploring Census Data

Data:

US Census Data

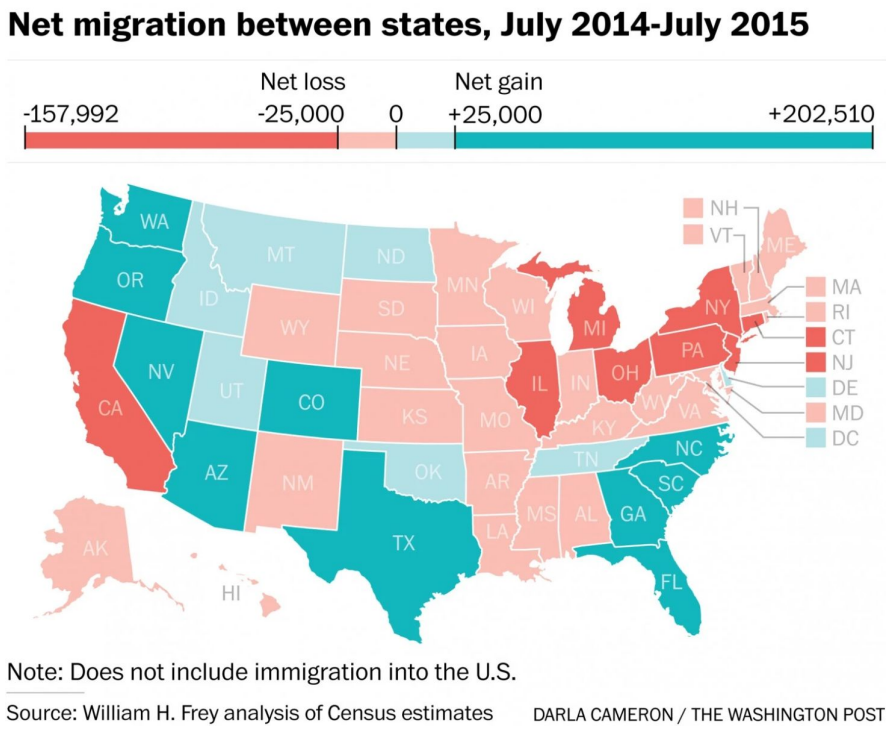
Task:

Answer queries like:

Which population group has the highest life expectancy?

Which cities are growing the fastest?

→ *Descriptive*



Healthcare: Medical Diagnosis

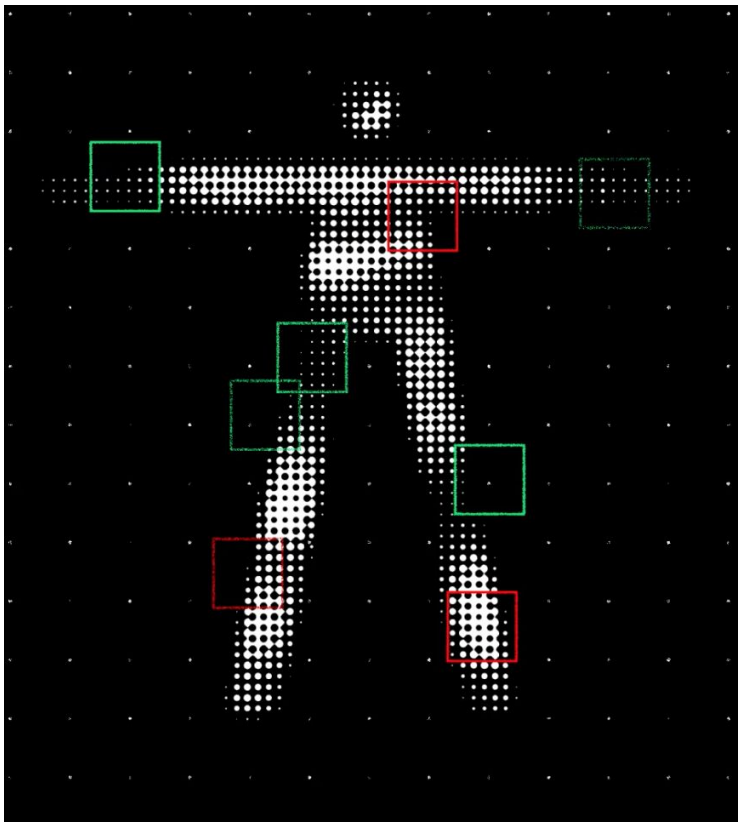
Data:

CT scans of patients' brains, labeled with "stroke" or "no stroke".

Task:

Given a new patient's CT scan, diagnose whether patient has stroke.

→ *Diagnostic*



Source: The New Yorker.
http://www.newyorker.com/magazine/2017/04/03/ai-versus-md

Sports Analytics

Data:

Game statistics: wins, losses, players, etc.

Task:

Predict which team will win the tournament

→ *Predictive*

Buy or Sell?

Data:

Stock performance over last 5 years, earning reports, news stories

Task:

Predict which stocks to sell or buy to maximize returns in 5 years

→ *Prescriptive*

→ *E.g. machine learning*

Poverty Mapping

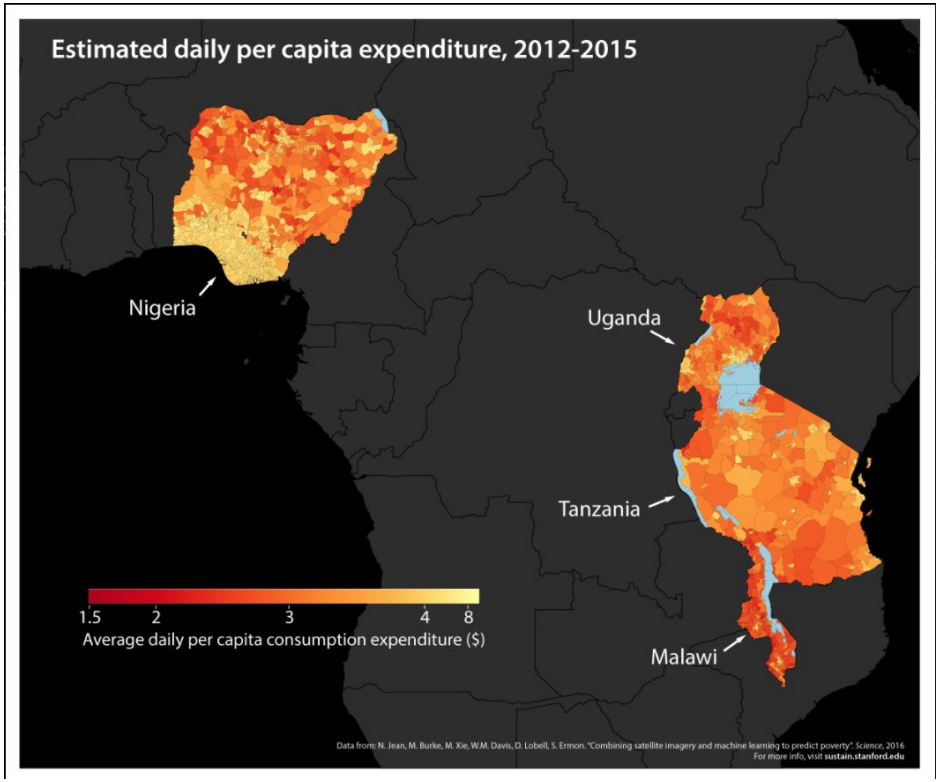
Data:

Satellite imagery of countries of interest

Task:

Map poverty for each region

Guest speaker Neal Jean will present his research project on poverty mapping!



Your turn!
You will get data, do something with it!

Stuck in Traffic?

LA Highway Traffic, as depicted in La La Land.

Data: Ride sharing trip histories in LA over the past year (with routes, speeds, wait times)

Come up with four data analysis tasks, one for each analytics type

Stuck in Traffic?

Data: Ride sharing trip histories in LA over the past year (with routes, speeds, wait times)

Come up with four data analysis tasks, one for each analytics type

Descriptive (what happened)

Diagnostic (Why did it happen)

Predictive (What will happen)

Prescriptive (How can you make it happen)

In Class Discussion

Descriptive

- What is the average wait times in Santa Monica?
- Which areas have the worst Traffic?

Diagnostic

- Why is there sudden congestion in the traffic?

Predictive 

- ETA's, predict when will riders arrive
- Predict ridership demand day to day
- Predict fare price

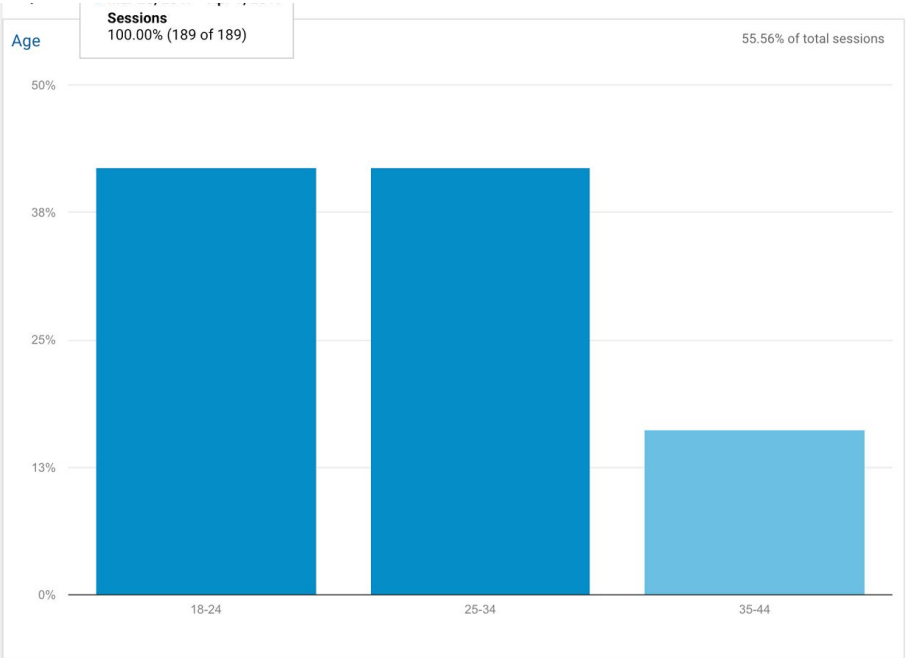
Prescriptive

- Where should I wait if I am a driver?
- How much bonus to give drivers to incentivize them?
- Finding quickest route

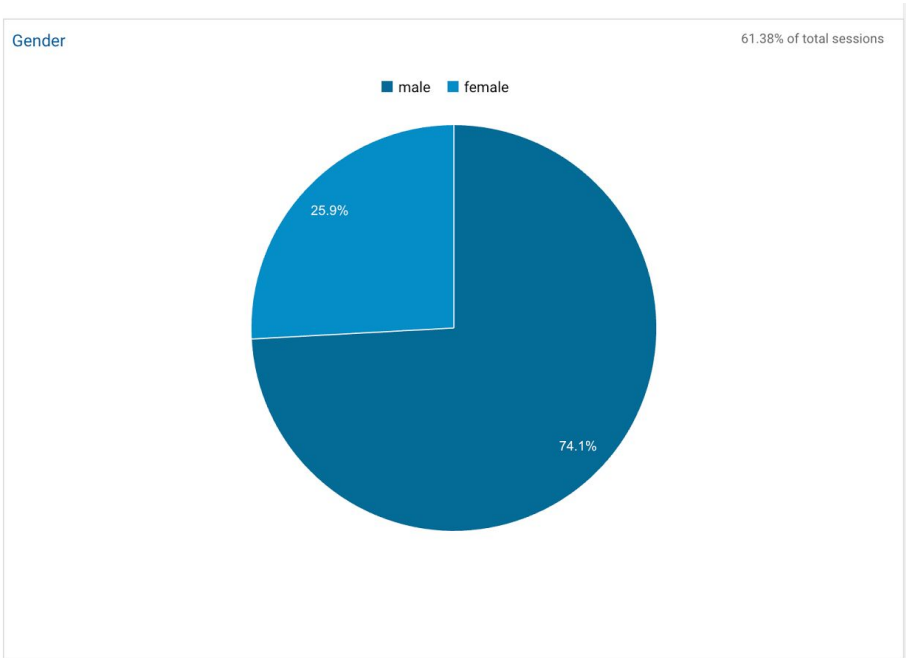
How many of you have visited the CS102 Website?

Google Analytics (Demographic)

Age Ranges



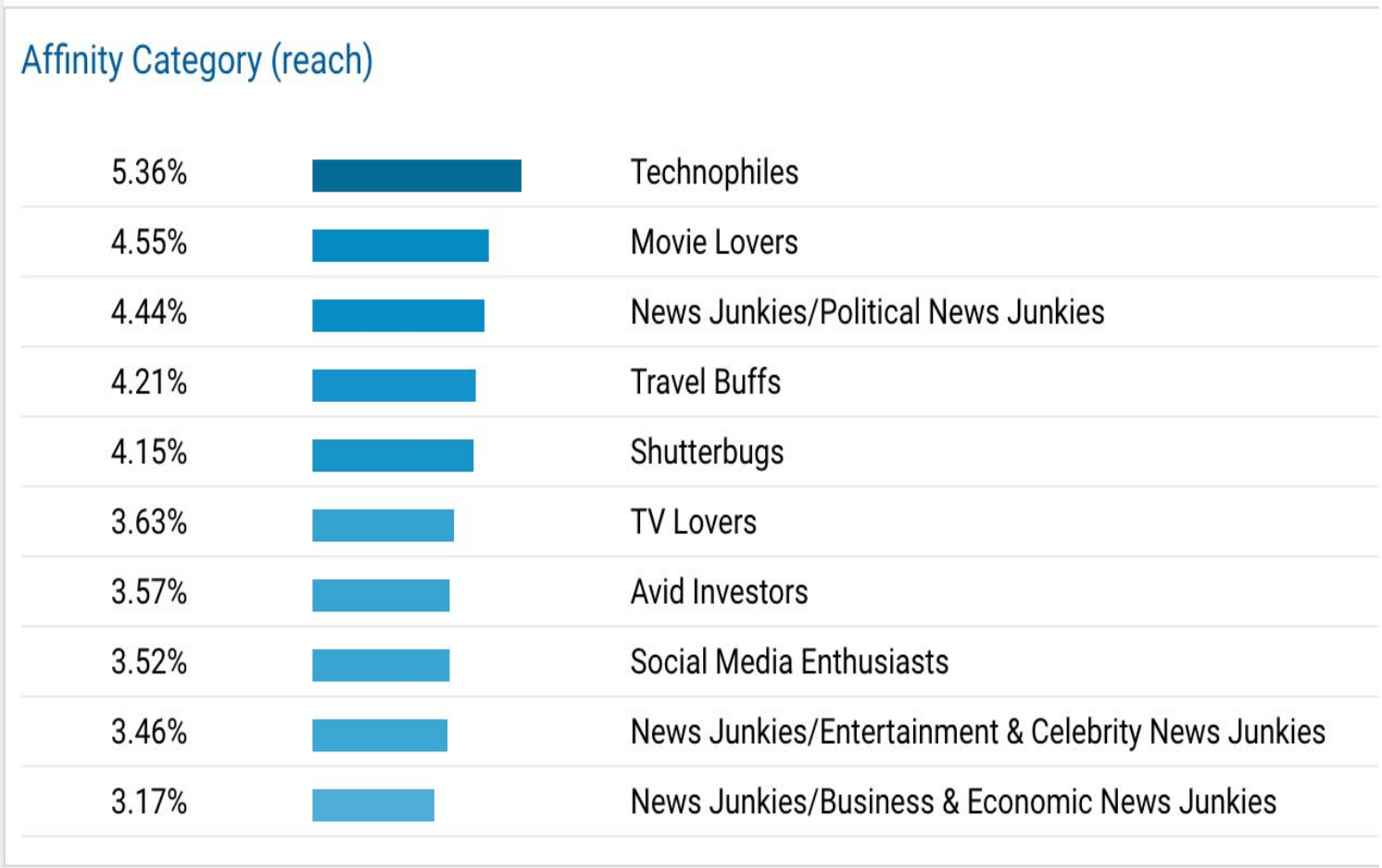
Gender



This report was generated on 4/2/17 at 12:53:15 PM - Refresh Report

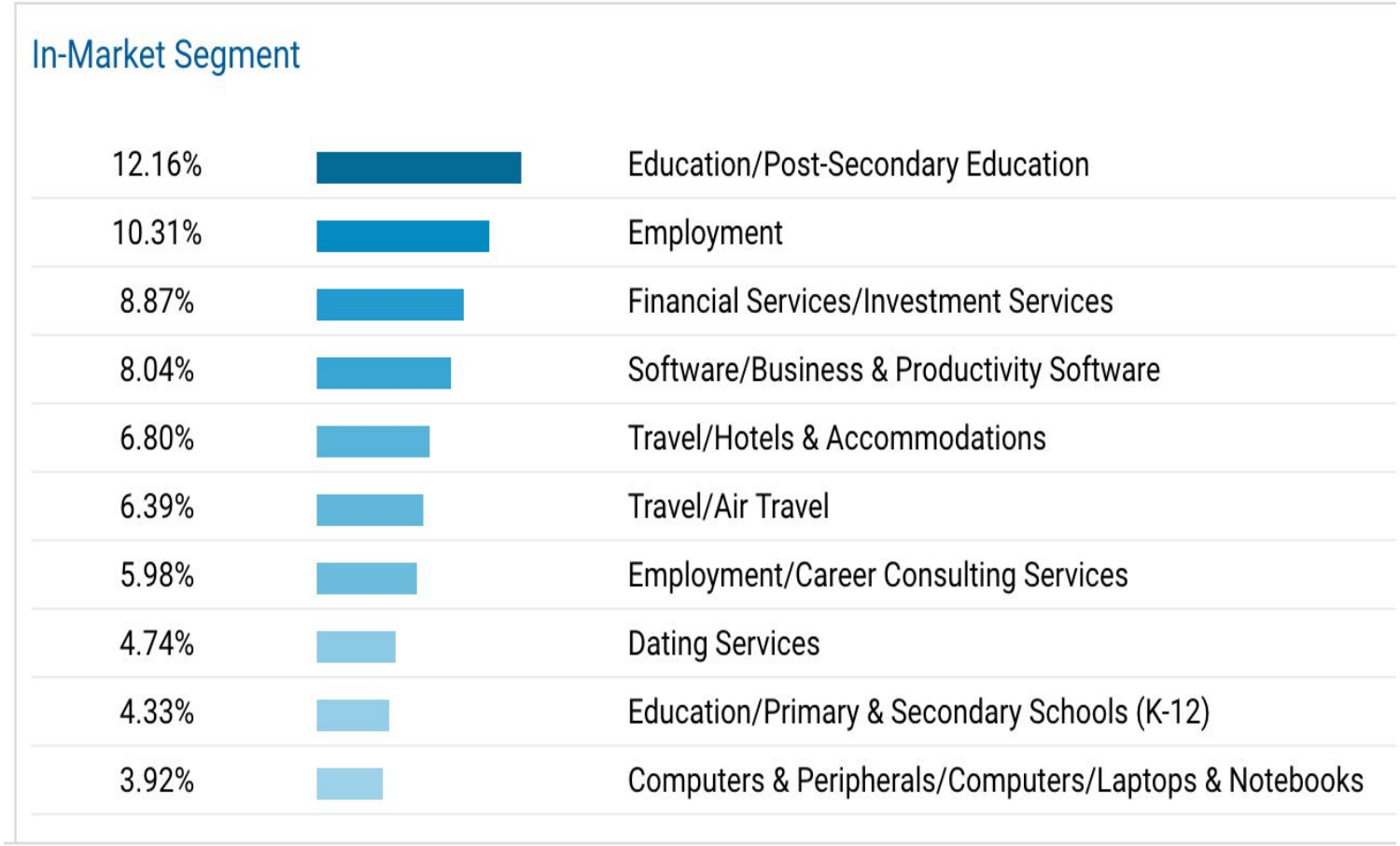
Google Analytics (Interests)

User profiles based on websites you visit



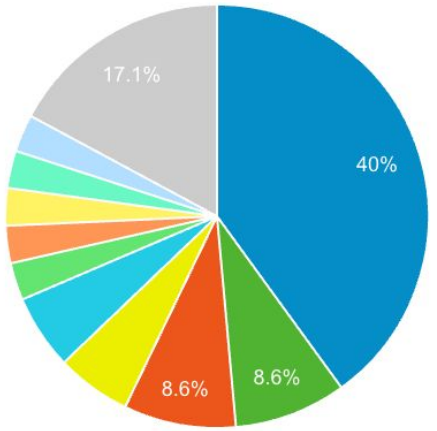
Google Analytics (Interests)

User profiles based on products you consume



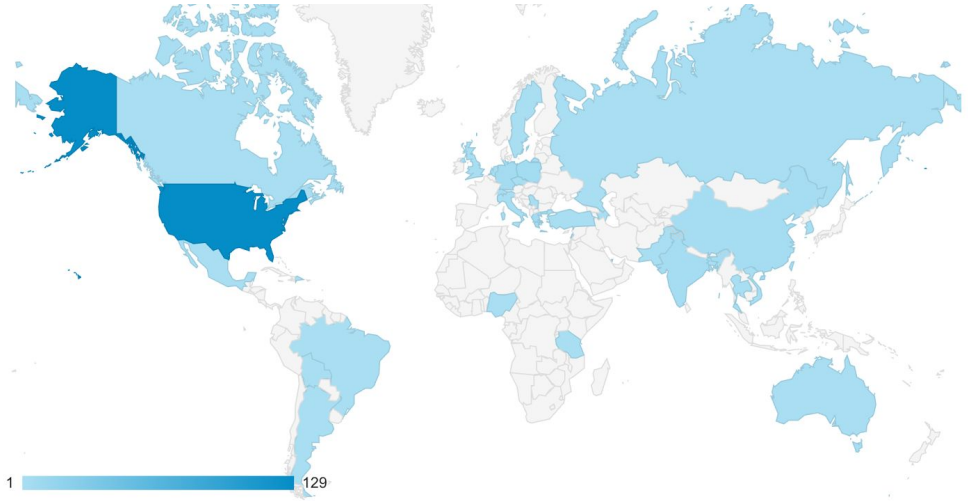
Google Analytics (Device Type)

Mobile Device Info	Sessions	Sessions	Contribution to total: Sessions
	35 % of Total: 18.52% (189)	35 % of Total: 18.52% (189)	
1. Apple iPhone	14	40.00%	
2. Apple iPad	3	8.57%	
3. Apple iPhone 6	3	8.57%	
4. Apple iPhone 5	2	5.71%	
5. Apple iPhone 6s	2	5.71%	
6. (not set)	1	2.86%	
7. Apple iPad Pro	1	2.86%	
8. Apple iPhone 5s	1	2.86%	
9. Apple iPhone 6s Plus	1	2.86%	
10. Apple iPhone 7	1	2.86%	



Google Analytics (Location)

Country Level



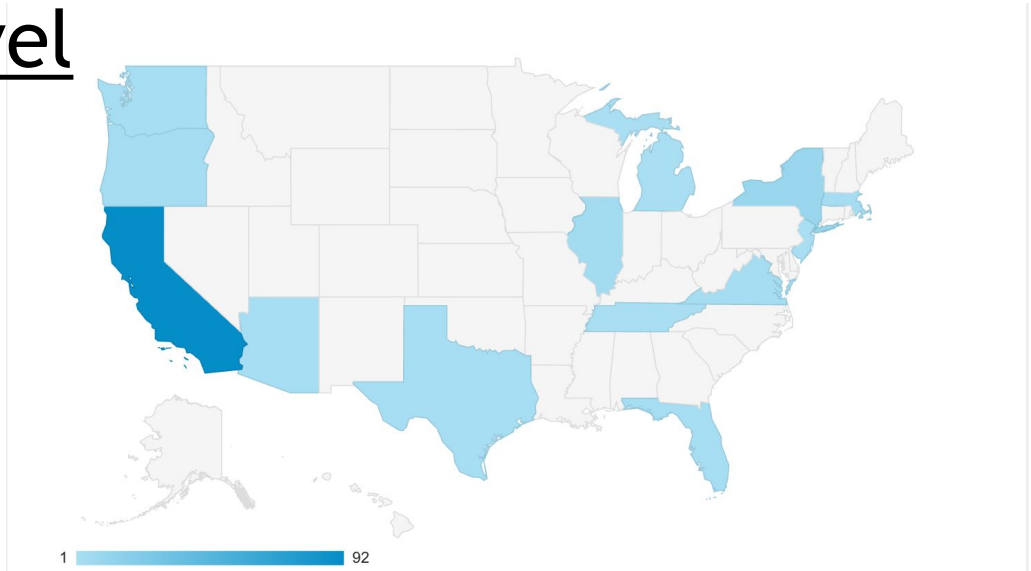
Primary Dimension: **Country** City Continent Sub Continent

Secondary dimension ▾ advanced

Country	Acquisition			Behavior			Conversions		
	Sessions ↓	% New Sessions	New Users	Bounce Rate	Pages / Session	Avg. Session Duration	Goal Conversion Rate	Goal Completions	Goal Value
	189 % of Total: 100.00% (189)	64.55% Avg for View: 64.02% (0.83%)	122 % of Total: 100.83% (121)	57.67% Avg for View: 57.67% (0.00%)	2.35 Avg for View: 2.35 (0.00%)	00:01:37 Avg for View: 00:01:37 (0.00%)	0.00% Avg for View: 0.00% (0.00%)	0 % of Total: 0.00% (0)	$0.00 % of Total: 0.00% ($0.00)
1. United States	129 (68.25%)	57.36%	74 (60.66%)	47.29%	2.81	00:02:11	0.00%	0 (0.00%)	$0.00 (0.00%
2. Greece	6 (3.17%)	16.67%	1 (0.82%)	100.00%	1.00	00:00:00	0.00%	0 (0.00%)	$0.00 (0.00%
3. Poland	6 (3.17%)	100.00%	6 (4.92%)	66.67%	1.50	00:02:03	0.00%	0 (0.00%)	$0.00 (0.00%
4. Germany	5 (2.65%)	100.00%	5 (4.10%)	100.00%	1.00	00:00:00	0.00%	0 (0.00%)	$0.00 (0.00%
5. China	4 (2.12%)	75.00%	3 (2.46%)	100.00%	1.00	00:00:00	0.00%	0 (0.00%)	$0.00 (0.00%
6. India	4 (2.12%)	100.00%	4 (3.28%)	100.00%	1.00	00:00:00	0.00%	0 (0.00%)	$0.00 (0.00%
7. Singapore	3 (1.59%)	100.00%	3 (2.46%)	66.67%	1.33	00:01:02	0.00%	0 (0.00%)	$0.00 (0.00%

Google Analytics (Location)

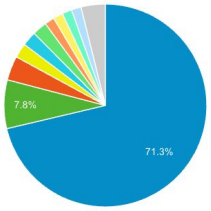
State Level



Primary Dimension: **Region** City Metro Other

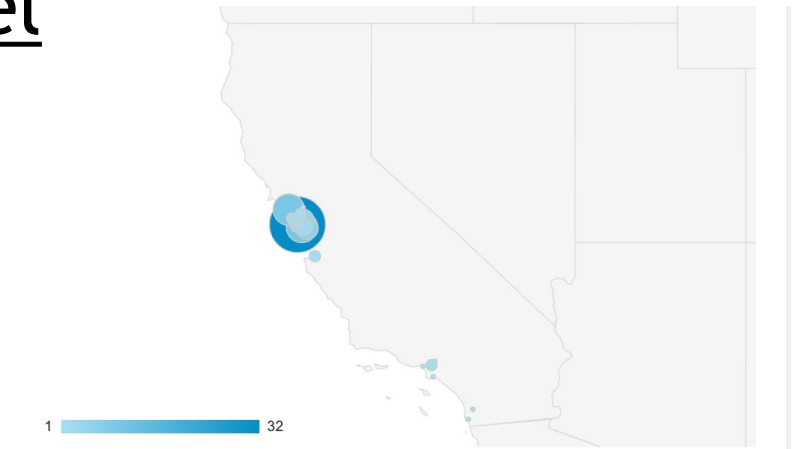
Secondary dimension | advanced

Region	Sessions	Sessions	Contribution to total: Sessions
	129 % of Total: 68.25% (189)	129 % of Total: 68.25% (189)	
1. California	92	71.32%	
2. New York	10	7.75%	
3. Illinois	5	3.88%	
4. Massachusetts	3	2.33%	
5. Oregon	3	2.33%	
6. Virginia	3	2.33%	
7. Arizona	2	1.55%	
8. District of Columbia	2	1.55%	
9. Florida	2	1.55%	
10. Texas	2	1.55%	



Google Analytics (Location)

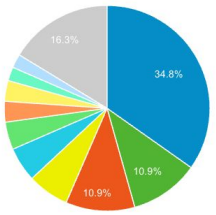
County Level



Primary Dimension: City Metro Other

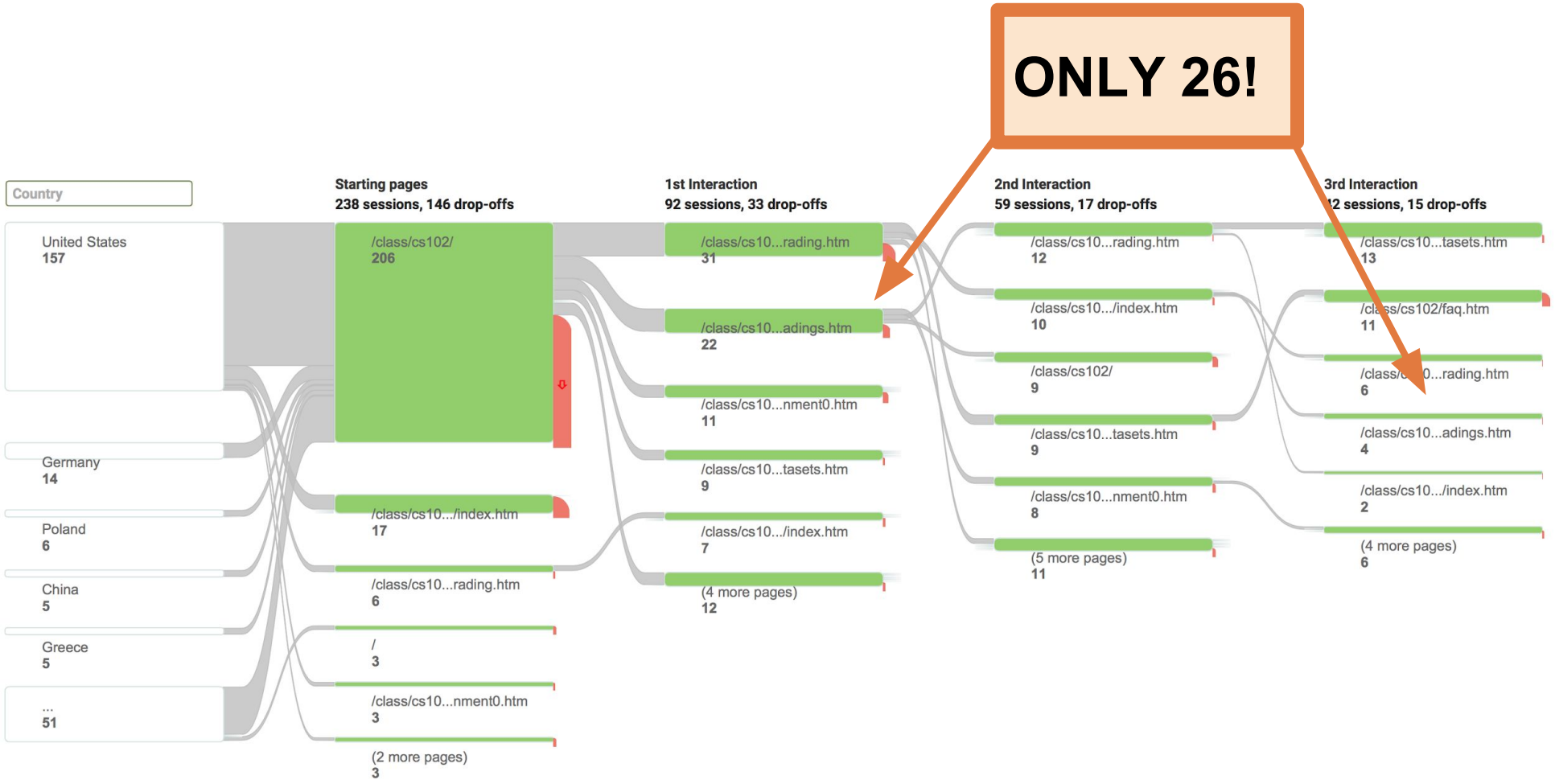
Secondary dimension | advanced

City	Sessions	Sessions	Contribution to total: Sessions
	92 % of Total: 48.68% (189)	92 % of Total: 48.68% (189)	
1. Stanford	32	34.78%	
2. San Francisco	10	10.87%	
3. Sunnyvale	10	10.87%	
4. San Jose	6	6.52%	
5. Fremont	5	5.43%	
6. Palo Alto	4	4.35%	
7. Mountain View	3	3.26%	
8. Santa Clara	3	3.26%	
9. Hayward	2	2.17%	
10. Los Angeles	2	2.17%	



How many of you have done the readings?

Google Analytics (User Activity)



Pitfalls

Privacy

Individual data collected covertly

- Edward Snowden, “metadata” argument

Data collected legally, used questionably

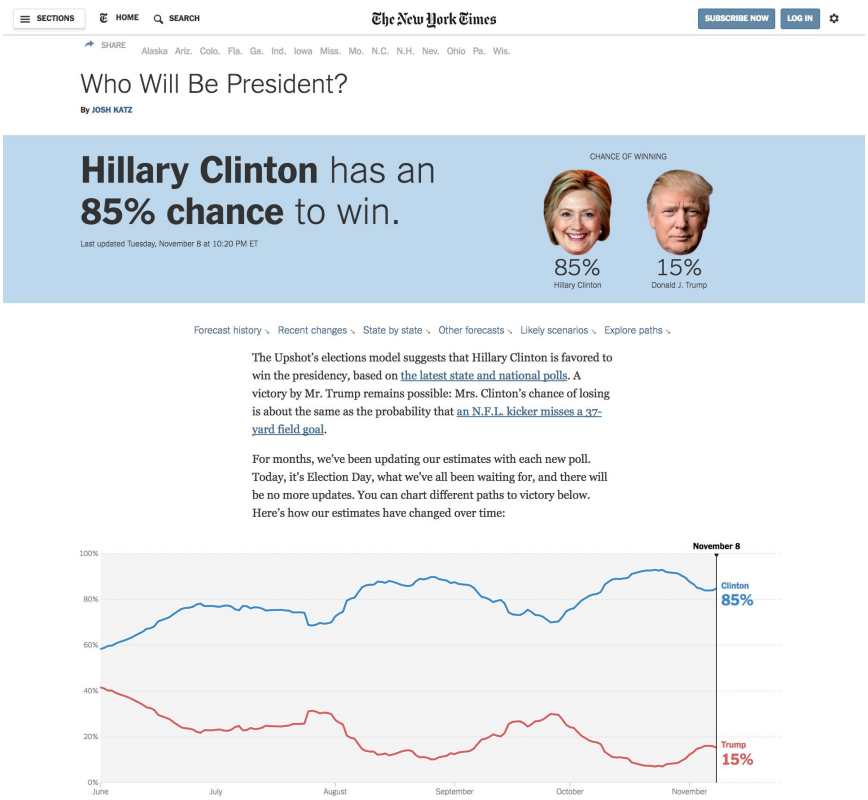
- Retailer Target’s pregnancy mailing scandal

Individiuals identified from “anonymous” data

- Boston mayor’s health record

Sampling Bias

Failure of election polls

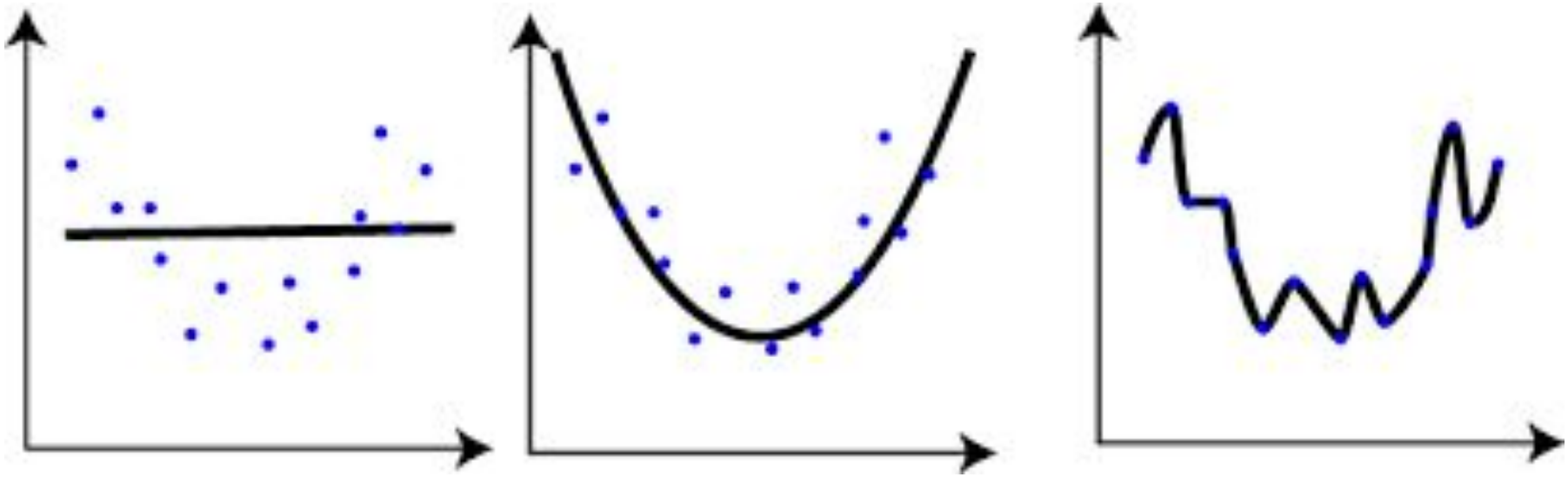


Adverserial Data

Microsoft's chatbot turning racist in <24hrs

Underfitting/Overfitting data

Model used for predictions too general or specific



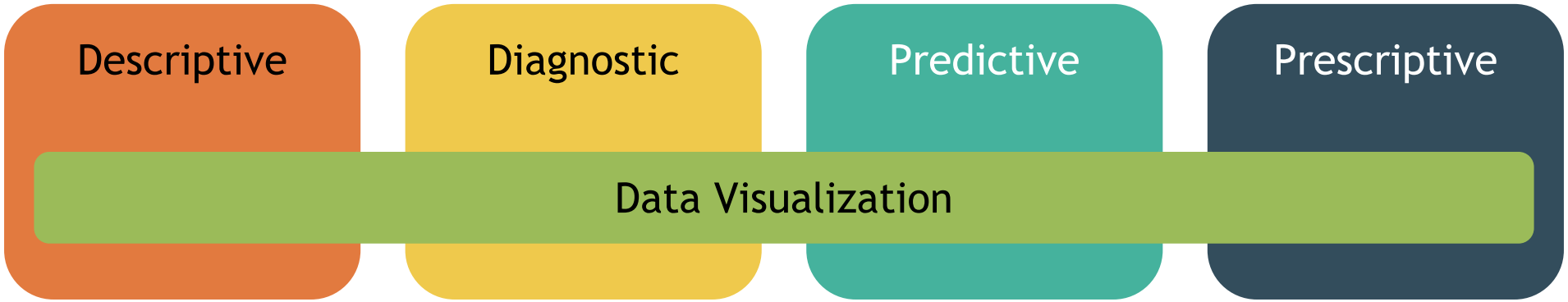
Underfitting "Just right" Overfitting

Correlation vs Causation

Discoveries Outweigh The Pitfalls

Data Visualizatiion

Data Visualization



- Visualization can be helpful at all stages of data analytics
- Serves two main purposes:
  - Summarize and explain results
  - Allow for exploration and discovery, e.g. to come up with hypotheses

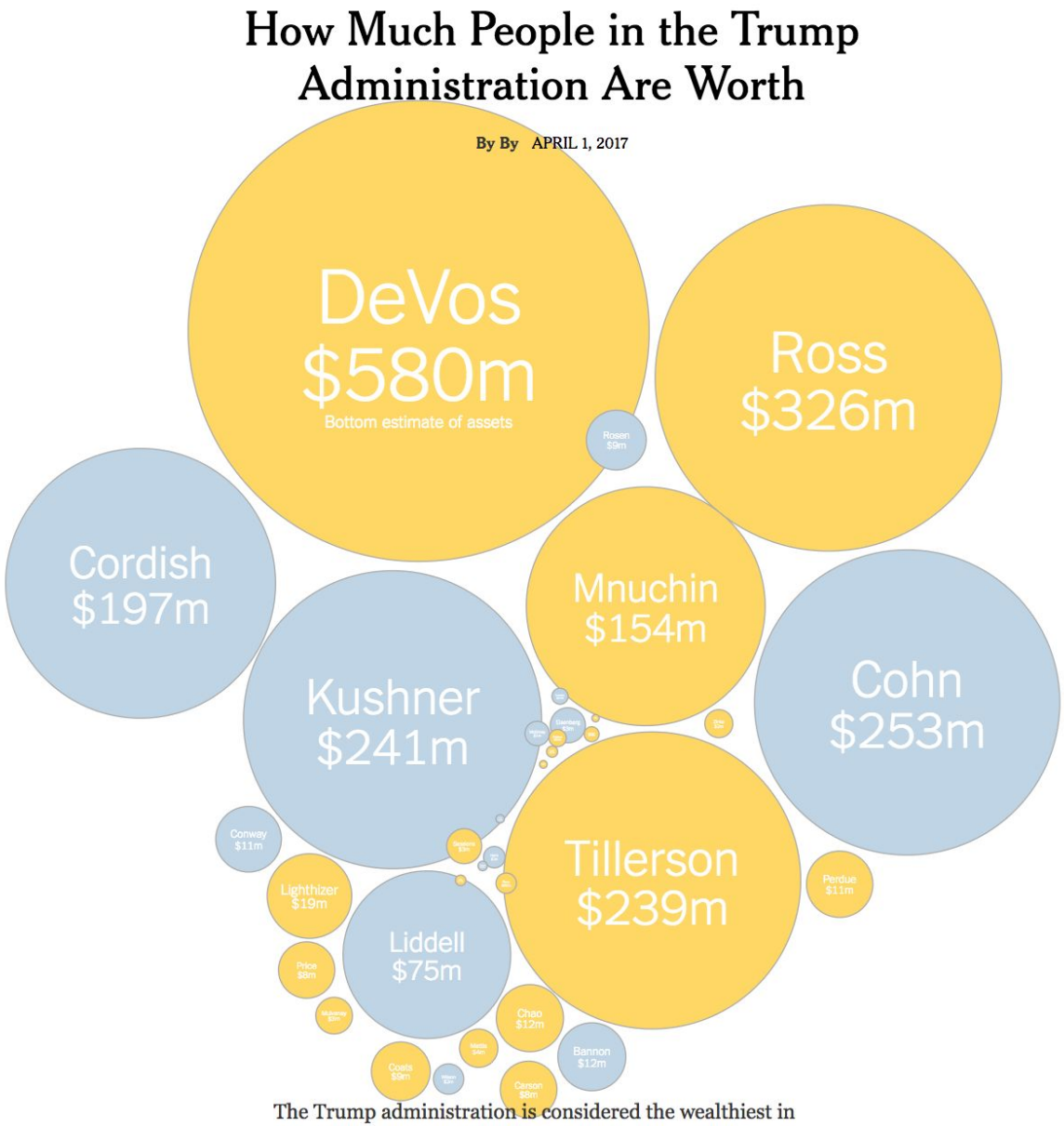
What makes a visualization good?

Good Visualizations

- Displays the data concisely and accurately
- Easy and fast to understand
- Facilitates comparison of data points
- Serves a clear purpose, e.g. description or exploration

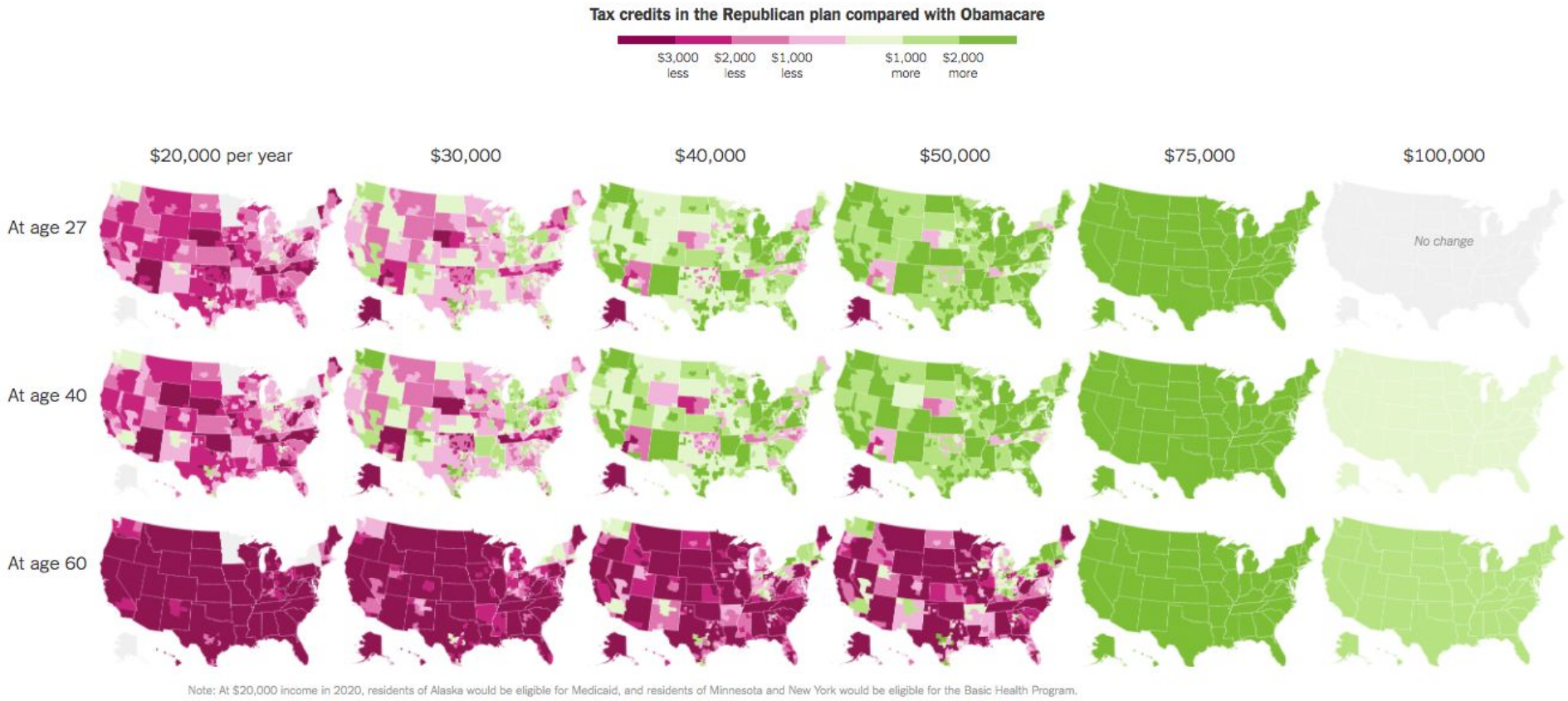
Loosely adopted from *Visual Display of Quantitative Information*.

Data Visualization Gallery

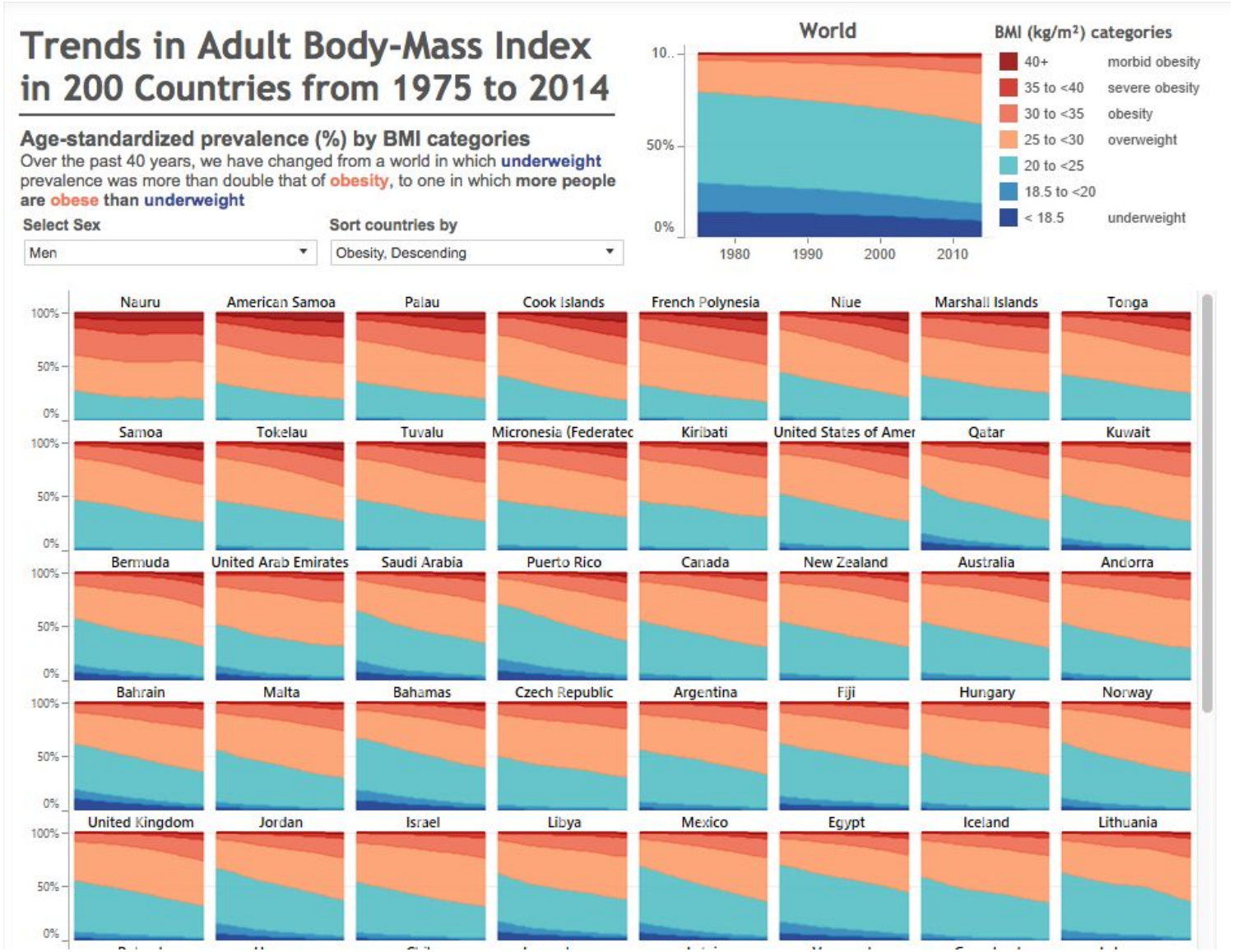


Source: New York Times.
https://www.nytimes.com/interactive/2017/04/01/us/politics/how-much-people-in-the-trump-administration-are-worth-financial-disclosure.html

Who Wins and Who Loses Under Republicans' Health Care Plan

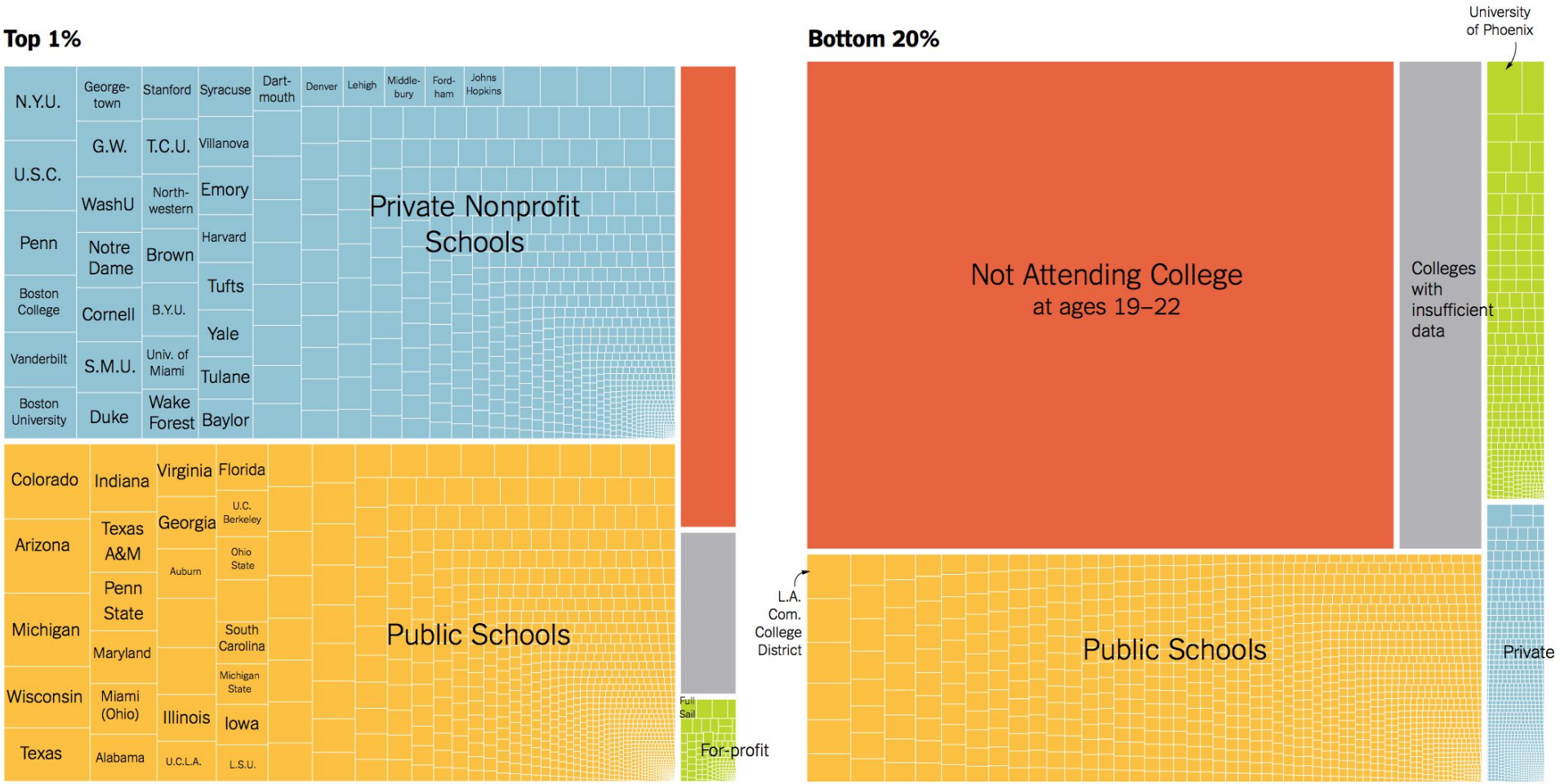


By Kevin Quealy and Margot Sanger-Katz. Source: New York Times.



Source: Tableau.
https://public.tableau.com/en-us/s/gallery/four-decades-prevalence-adult-bmi

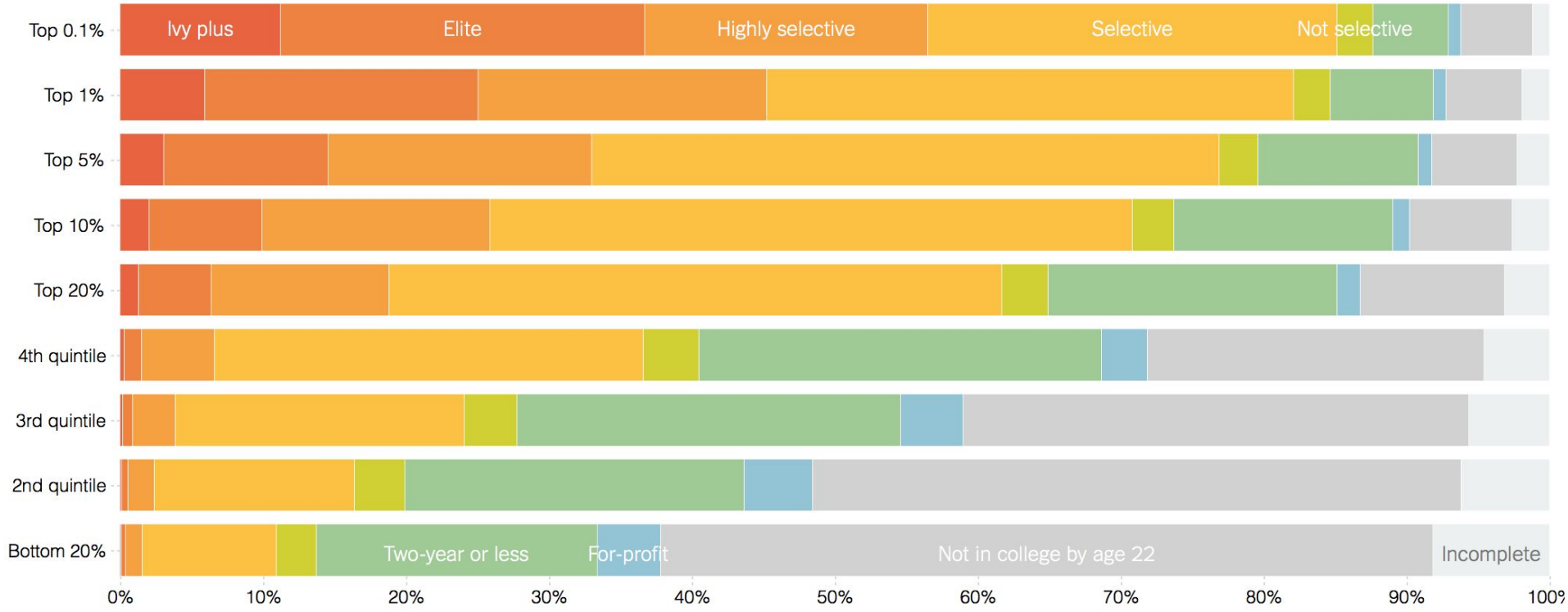
Where the top 1% and the bottom 20% go to college



Source: New York Times.
https://www.nytimes.com/interactive/2017/01/18/upshot/some-colleges-have-more-students-from-the-top-1-percent-than-the-bottom-60.html

Where today's 25-year-olds went to college, grouped by their parents' income

About four in 10 students from the top 0.1 percent attend an Ivy League or elite university, roughly equivalent to the share of students from poor families who attend any two- or four-year college.



Source: New York Times.
https://www.nytimes.com/interactive/2017/01/18/upshot/some-colleges-have-more-students-from-the-top-1-percent-than-the-bottom-60.html

Course Objectives Summary

- Explore big data through case studies and guest speakers
- Learn data analysis techniques through databases, data mining and machine learning
- Learn data analysis tools including Spreadsheets, SQL and Python
- Learn data visualization techniques and tools
- Apply techniques to different application areas

Logistics

Course Website

https://cs102.stanford.edu

- **Syllabus** *with lecture topics, readings, materials, assignments and due dates*
- Link to course calendar
- Datasets
- FAQ
- Piazza link
- Canvas link

Office Hours

- Refer to calendar on course website.
- You can also add this calendar to your own Google calendar.

- Regular OH times (Starts this week!):
  - Lisa: Wed, 7.30 - 9pm, Huang Basement
  - Ethan: Thu, 3 - 4.30pm, Lathrop Tech Lounge

Piazza

https://piazza.com/class/iz1v14otfga59p

- Q&A platform where anyone can ask class-related questions and post answers
- You can opt to post anonymously, or privately
- Please enroll if you haven't already

Canvas

https://web.stanford.edu/group/canvas/discovery.html

- You will use Canvas to submit assignments

Course Requirements

Point Distribution:

Item	Points
Assignment 0	1
Assignment 1	10
Assignment 2	10
Assignment 3	10
Assignment 4	20
Midterm	10
Final Project Proposal	1
Final Project	15
Final Exam	20
Class Attendance*	3
Total	**100**

Attendance Policy

- Goal: Maximize what you get out of CS102!
- Difference to other CS courses: Small class size, in-class activities and discussions.

- Mandatory attendance
- Up to 3 absences (excused or unexcused) allowed

Waitlist Policy

- Thanks for coming today!
- If a spot opens up for you, axess will enroll you automatically.
- If you can’t take the class anymore, we kindly ask you to drop it.

Assignment 0: Jupyter Notebook Setup

- We will use Jupyter Notebook for assignments and in class demos (from week 2)
- Goal of Assignment 0: Making sure that everyone can run Jupyter Notebook
- Setup instructions on course website
- If you have difficulties, please find us at OH or post on piazza.
- **Due date: Sun, Apr 9**

Questions?

Market Basket Analysis in practice.