Data Science: A Review

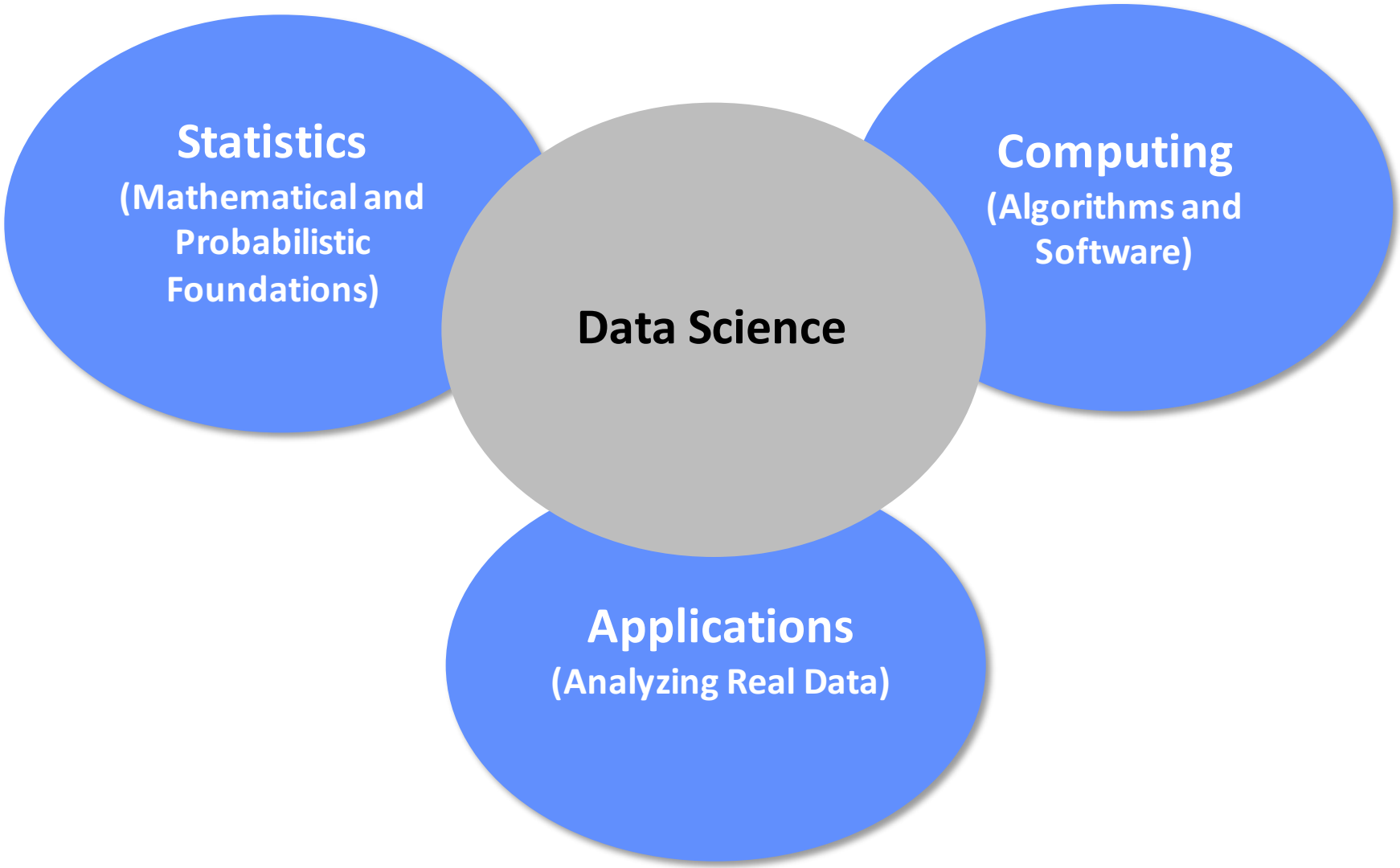
Stats 5, Winter 2018

Professor Padhraic Smyth
Departments of Computer Science and Statistics
University of California, Irvine

Schedule of Lectures

Date	Speaker	Department Or Organization	Topic
Jan 9	Padhraic Smyth	Computer Science	Introduction to Data Science
Jan 16	Padhraic Smyth	Computer Science	Machine Learning
Jan 23	Michael Carey	Computer Science	Databases and Data Management
Jan 30	Sameer Singh	Computer Science	Statistical Natural Language Processing
Feb 6	Zhaoxia Yu	Statistics	An Introduction to Cluster Analysis
Feb 13	Erik Sudderth	Computer Science	Computer Vision and Machine Learning
Feb 20	John Brock	Cylance, Inc	Data Science and CyberSecurity
Feb 27	Video Lecture (Kate Crawford)	Microsoft Research and NYU	Bias in Machine Learning
Mar 6	Matt Harding	Economics	Data Science in Economics and Finance
Mar 13	Padhraic Smyth	Computer Science	Review: Past and Future of Data Science

Components of Data Science



Core Data Science Skills

- Database systems
- Programming (Python, R, C, etc)
- Software engineering
- Algorithms
- Matrix-vector algebra and calculus
- Probability
- Machine learning
- Statistical modeling
- Communication and writing skills

What Classes will you take in the DS Major?

Statistics

Stats 120 ABC: Intro to Prob and Stats
Stats 68: Exploratory Data Analysis
Stats 110-112: Statistical Methods
CS 178: Machine Learning
(Stats 140: Multivariate Statistics)

Computing

ICS 46: Data Structures
IFMTX 43: Intro to Software Engineering
CS 122A: Intro to Data Management
CS 161: Design and Analysis of Algorithms
(CS 131: Parallel and Distributed Computing)
(CS 172: Neural Networks/Deep Learning)

Applications

Stats 170AB: Data Science Capstone Project
INF 143: Information Visualization
(INF 131: Human Computer Interaction)
(CS 121: Information Retrieval)
(CS 122B: Project in Databases/Web Applications)
(Summer intermships, e.g., junior year)

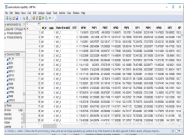
(Sample electives shown in parentheses)

Stats 170AB: Project in Data Science

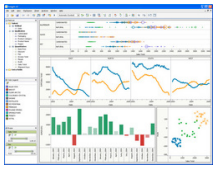
Unstructured Data

Extracted Data

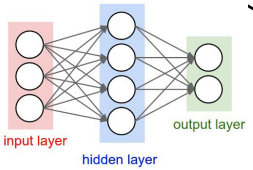
Transformed Data



Data for Modeling



Predictive Model

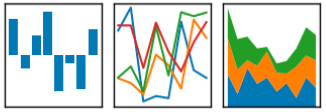


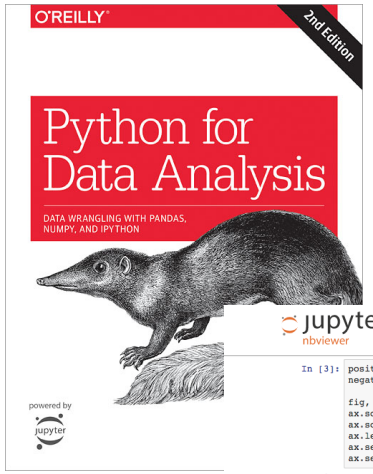
Predictions/ Decisions

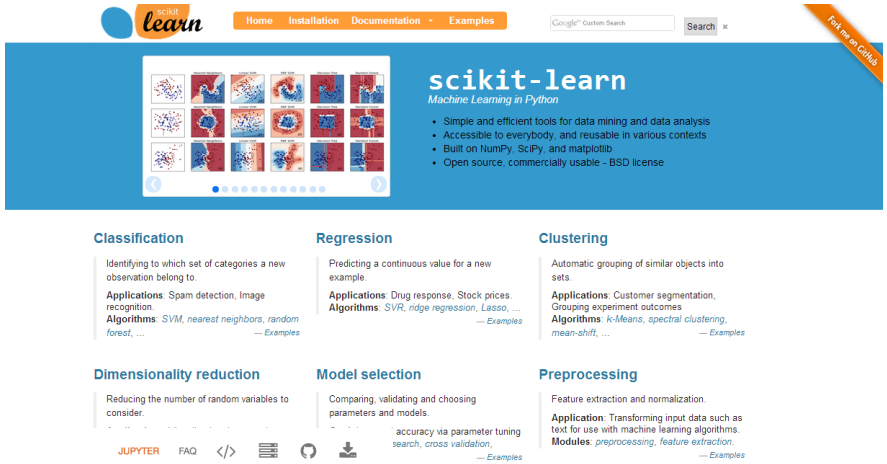
Stats 170AB: Project in Data Science

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

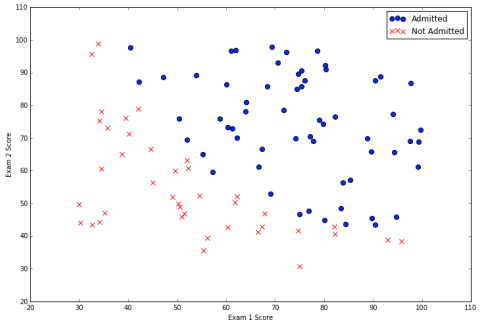






jupyter
nbviewer

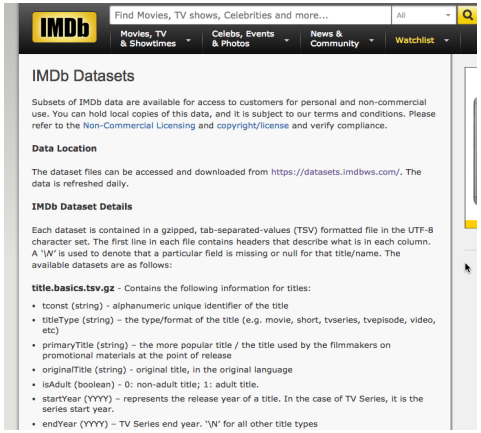
```
In [3]: positive = data[data['Admitted'].isin([1])]
        negative = data[data['Admitted'].isin([0])]

        fig, ax = plt.subplots(figsize=(12,8))
        ax.scatter(positive['Exam 1'], positive['Exam 2'], s=50, c='b', marker='o', label='Admitted')
        ax.scatter(negative['Exam 1'], negative['Exam 2'], s=50, c='r', marker='x', label='Not Admitted')
        ax.legend()
        ax.set_xlabel('Exam 1 Score')
        ax.set_ylabel('Exam 2 Score')

Out[3]: <matplotlib.text.Text at 0xd17d7b8>
```

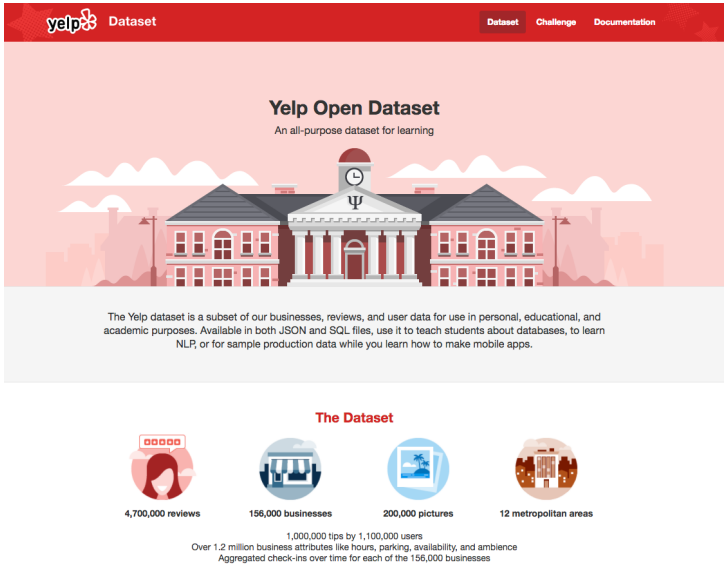


It looks like there is a clear decision boundary between the two classes. Now we need to implement logistic regression so we can train a model to predict the outcome. The equations implemented in the following code samples are detailed in "ex2.pdf" in the "exercises" folder.

Stats 170AB: Project in Data Science

twitter stream



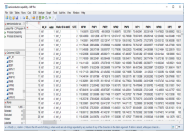


United States Census Bureau

Text from 4 million Wikipedia articles

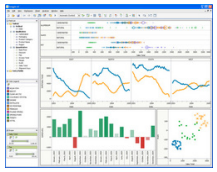
Data Science Skills: the Spectrum of Data Analysis

Raw Data

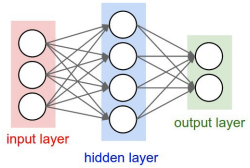


Data Wrangling

Data Management



Exploratory Data Analysis



Predictive Modeling

- Consumers
- External Business Customers
- Internal Business Customers
- Government
- Scientists

Databases, Algorithms, Software Engineering

Machine Learning, Statistics

Business, scientific, medical, and other domain knowledge

Sample Course of Study in the Major

Years 1 and 2: foundational courses in computer science, mathematics, statistics, including statistical computing

2015-16, First Year: 41 units

Fall	12	Winter	13	Spring	16
ICS 31	4	ICS 32	4	ICS 33	4
Math 2A	4	Math 2B	4	Math 2D	4
Writing 39A	4	Writing 39B	4	Stats 7	4
		Stats 5	1	Writing 39C	4

2016-17, Second Year: 46 units

Fall	16	Winter	14	Spring	16
ICS 6B	4	ICS 45C	4	Stats 68	4
Math 3A	4	ICS 51	6	Stats 120C	4
Stats 120A	4	Stats 120B	4	ICS 46	4
GE III	4			ICS 6D	4

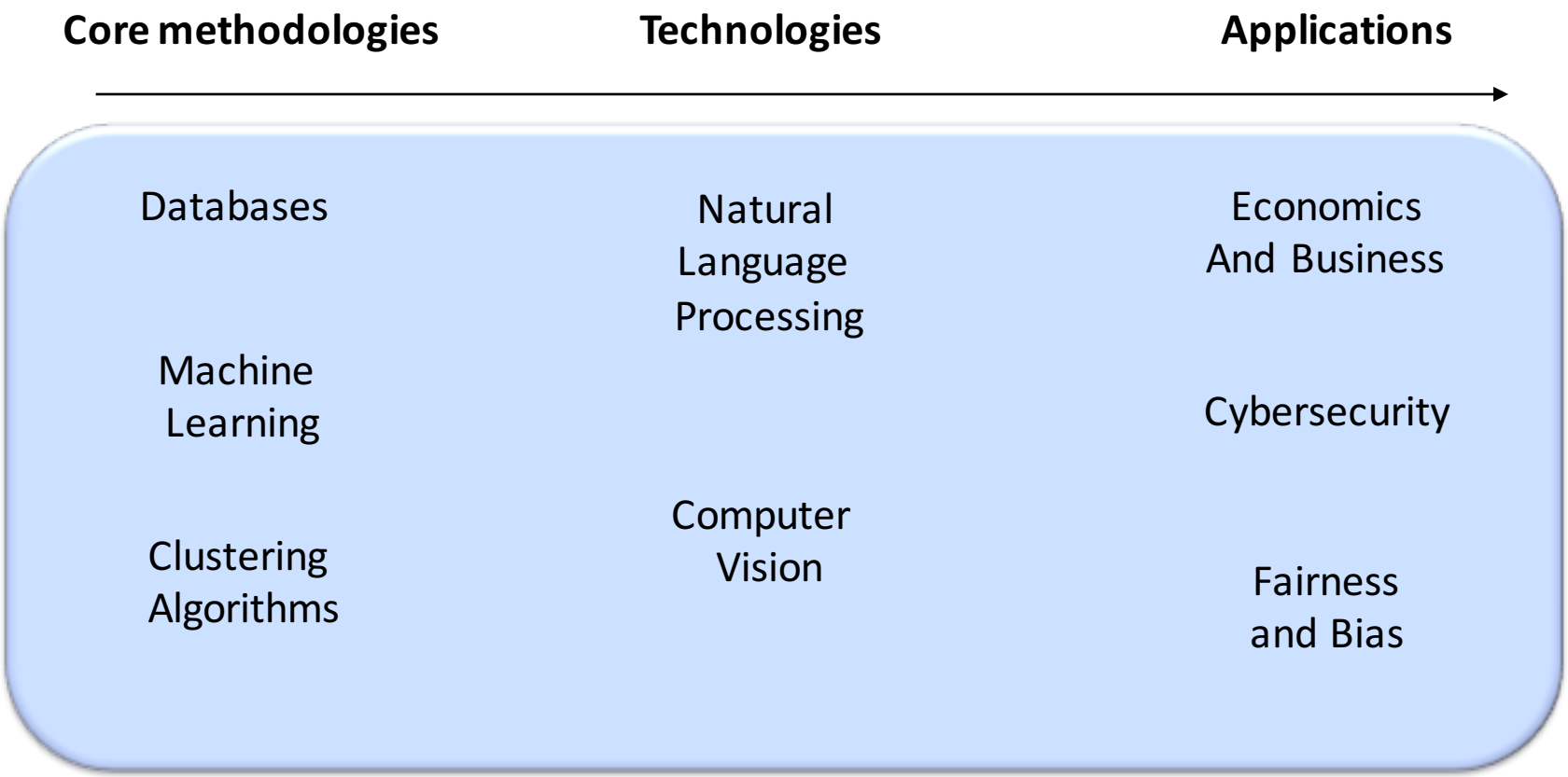
Years 3 and 4: more emphasis and specialization in data science topics such as machine learning, databases, visualization, advanced statistics

Year 3: sample program

Fall	Winter	Spring
Stats 110, Statistical Methods for Data Analysis I	Stats 111, Statistical Methods for Data Analysis II	Stats 112, Statistical Methods for Data Analysis III
CS 161, Design and Analysis of Algorithms	CS 178, Machine Learning and Data-Mining	CS 122A, Introduction to Data Management
In4matx 43, Introduction to Software Engineering	ICS 139W, Critical Writing on Information Technology	In4matx 143, Information Visualization
GE IV/VIII,	GE III/VII,	GE VI,

Year 4: two-quarter capstone "data-intensive" project, + statistics and CS electives

Topics from Lectures this Quarter



Final Assignment

- Write a ½ to 1 page short essay that takes any two of the topics from lectures 2 to 9, and describes how you think the two topics could "intersect" going forward,
  e.g.,
  - What aspects of each method could be combined to produce new ideas?
  - What new applications might be enabled by combining these methods?
  - What are the potential challenges in these areas?

- Possible combinations
  - Natural language and cybersecurity
  - Clustering algorithms and computer vision
  - Computer vision and fairness/bias

  - …feel free to pick any 2 topics that interest you

Final Assignment Instructions

- Put your name and student ID at the top of the page
- Submit as a PDF file
- Due to EEE dropbox by 9am on Monday March 19th (next week)
- Note: there is **no final exam** in this class

How is data measured and collected?

Data in Matrix Form

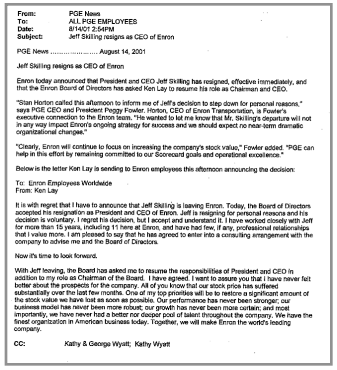
Measurements →

Entities ↓

ID	Income	Age	….	Monthly Debt	Good Risk?
18276	65,000	55	….	2200	Yes
72514	28,000	19	….	1500	No
28163	120,000	62	….	1800	Yes
17265	90,000	35	….	4500	No
…	…	…	….	…	…
…	…	…	….	…	…
61524	35,000	22	….	900	Yes

Text Collections

NYT
330,000 articles



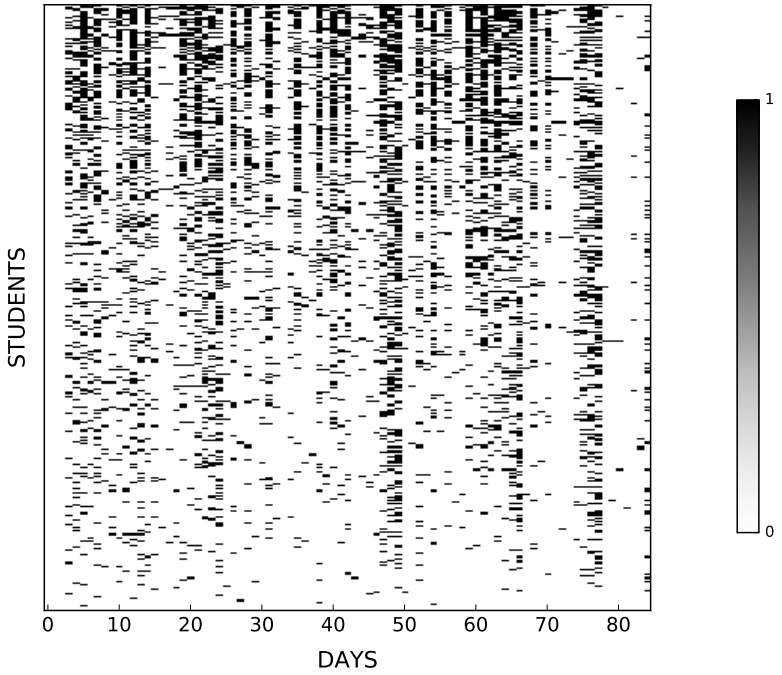
Enron
250,000 emails

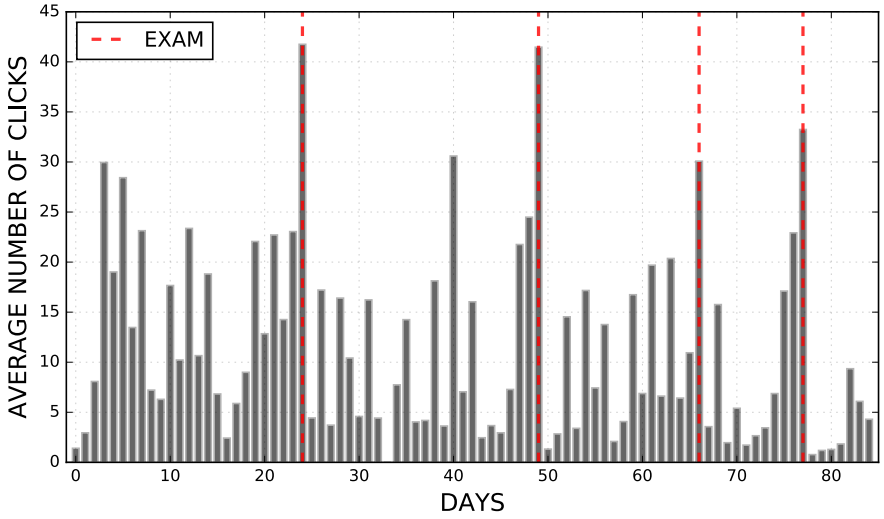
Pennsylvania Gazette
80,000 articles
1728-1800

NSF/ NIH
100,000 grants

16 million Medline articles

Examples of Student Clickstream Data





Yelp Dataset Challenge

Discover what insights lie hidden in our data

The challenge is a c
with us. Whether yo
from the local graph

The Challenge

We challenge students to use our data in innovative ways and break ground in research. Here are some examples of topics we find interesting, but remember these are only to get you thinking and we welcome novel approaches!

Photo Classification

Maybe you've heard of our ability to identify hot dogs (and other foods) in photos. Or how we can tell you if your photo will be beautiful or not. Can you do better?




Natural Language Processing & Sentiment Analysis

What's in a review? Is it positive or negative? Our reviews contain a lot of metadata that can be mined and used to infer meaning, business attributes, and sentiment.

Graph Mining

We recently launched our Local Graph but can you take the graph further? How do user's relationships define their usage
are the trend setters eating before it becomes popular?

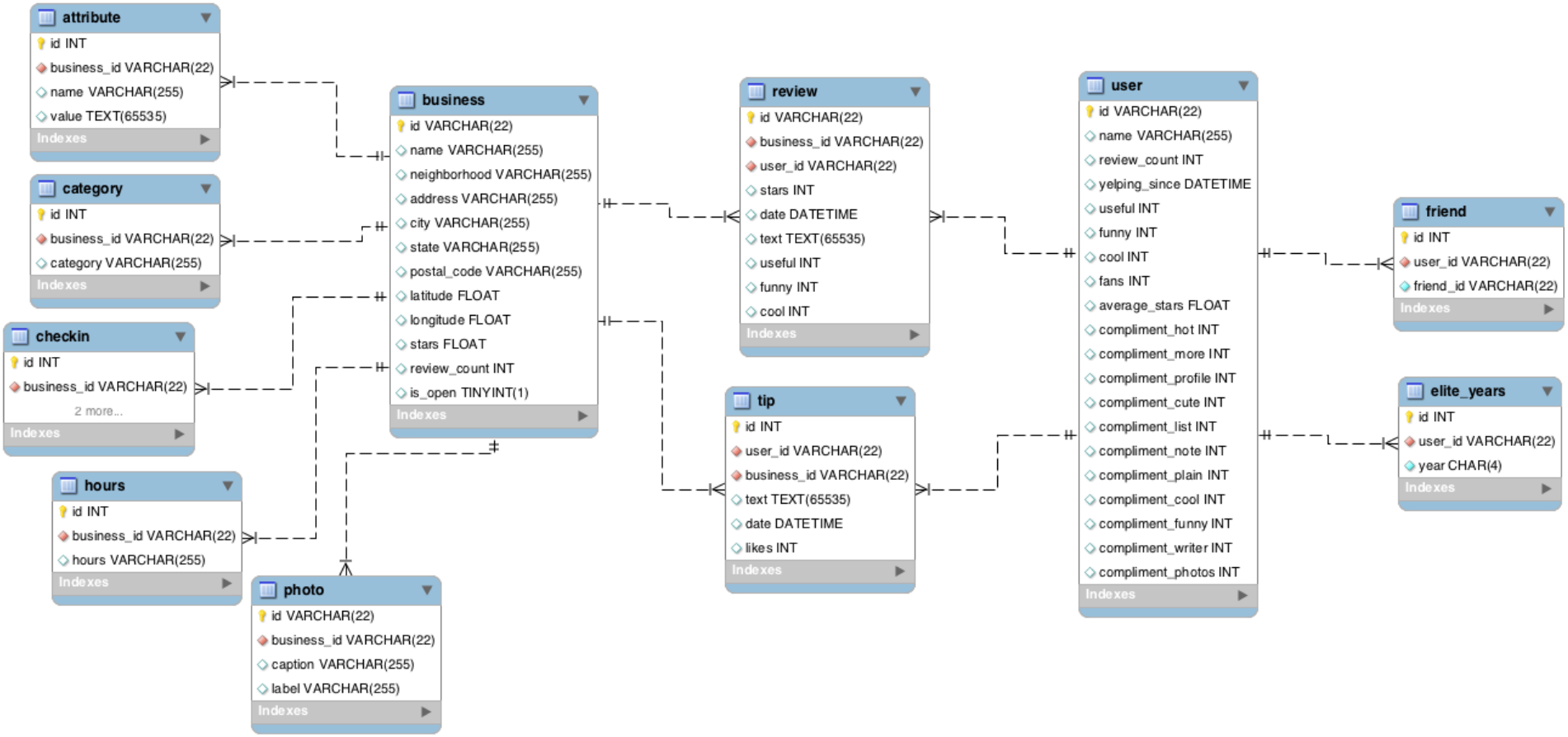
5.2 million reviews
174k businesses
11 metropolitan areas

Round 11

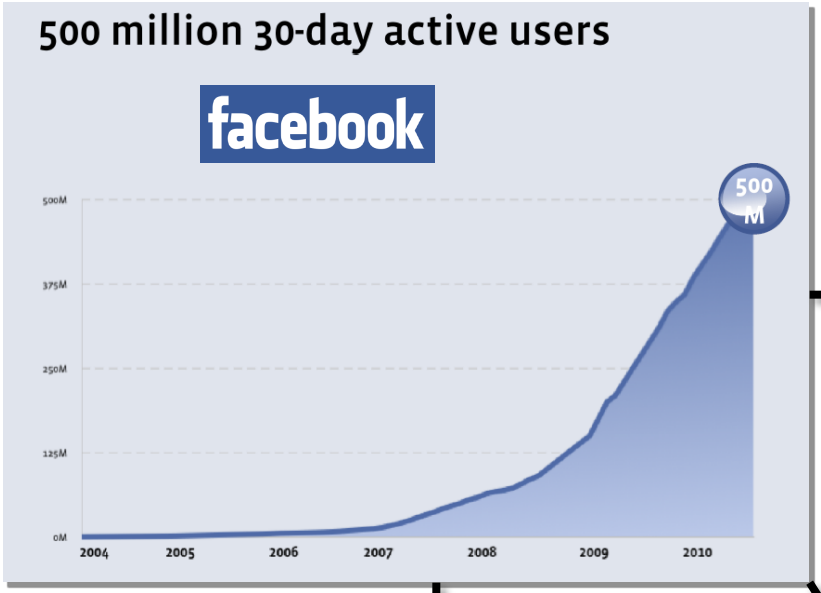
Our dataset has been updated for this iteration of the challenge - we're sure there are plenty of interesting insights waiting there for you. This set includes information about local businesses in 11 metropolitan areas across 4 countries. Round 11 has kicked off starting January 18, 2018 and will run through June 30, 2018.

Download Dataset

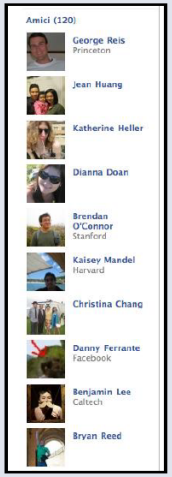
Yelp Challenge DataSet

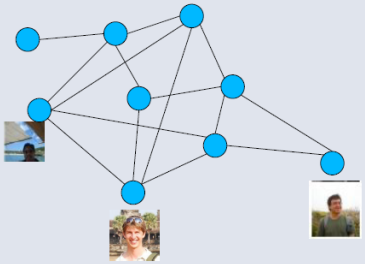


500 million 30-day active users



The Friendship graph





500M users each connect to an average of 130 other users = ~ 60 Billion Edges

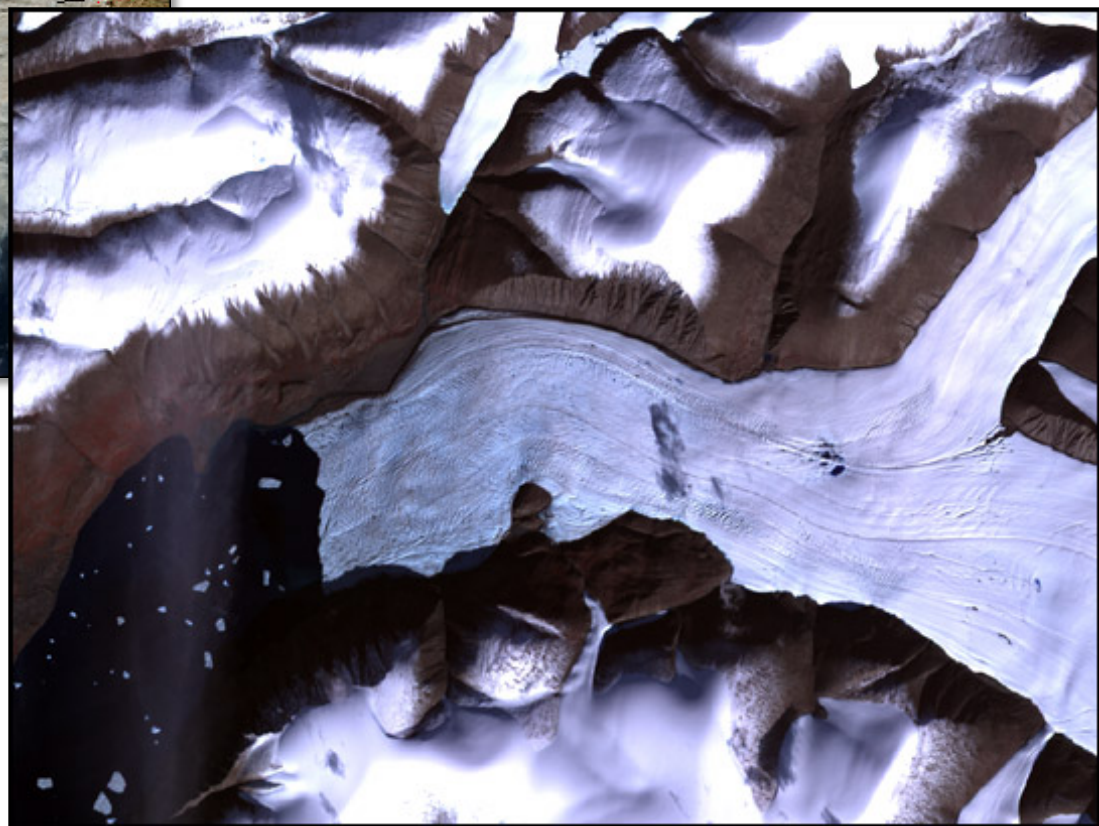
Over 30 billion pieces of content shared every month

Over **3 billion** photos uploaded each month

Graphics from Lars Backstrom, ESWC 2011

NASA's MODIS satellite

entire planet
250m resolution
37 spectral bands
every 2 days



Daily Report: At WWDC, Apple Expected to Expand Into Health and Home Monitoring

By THE NEW YORK TIMES JUNE 2, 2014 7:14 AM Comment

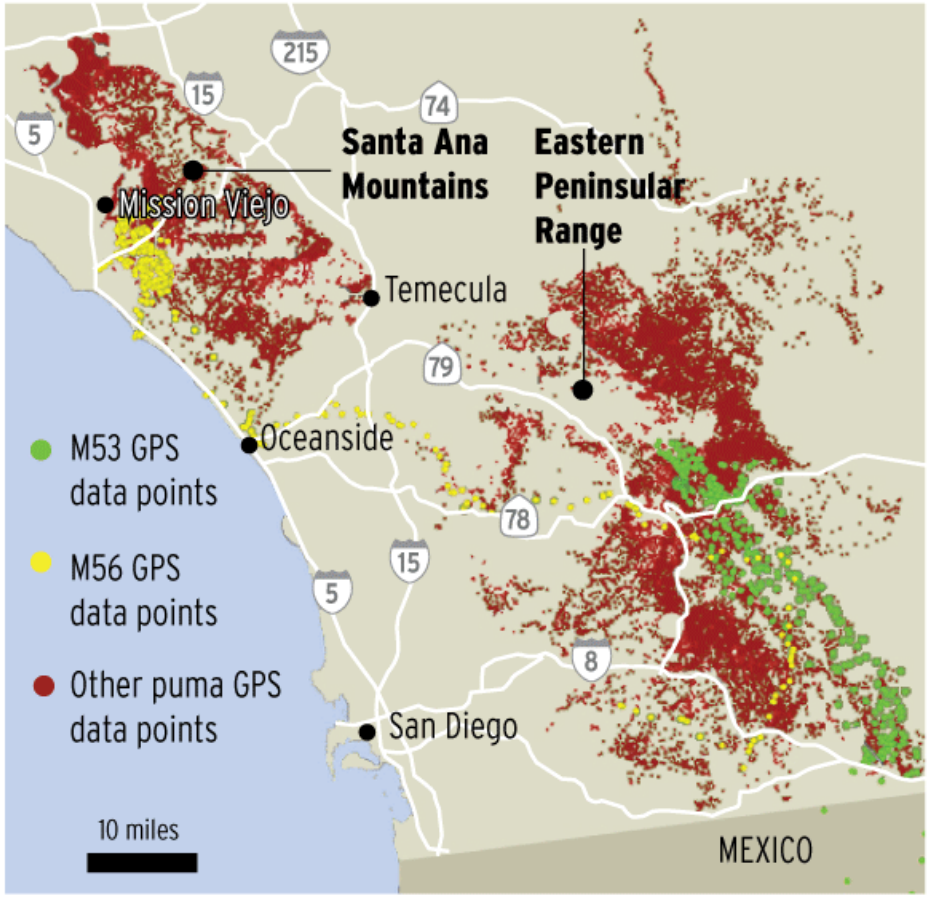
Apple is unlikely to introduce new devices this week, the things that most excite customers and investors these days. But the company is expected to dive deeper into two new areas: connected health and the so-called smart home, Brian X. Chen reports.

Daily Report

Along with operating system updates for mobile devices and desktop machines, Apple plans to introduce a new health-tracking app at its annual Worldwide Developers' Conference on Monday, according to a person briefed on the product, who spoke on the condition of anonymity because the plans were confidential. The app for mobile devices will track statistics for health or fitness, like a user's footsteps, heart rate and sleep activity.

Tracking pumas

From 2001 to 2013, scientists used GPS radio collars to track the pumas' movements in the Santa Ana Mountains and Eastern Peninsular Range in Orange and San Diego counties. Only one puma, M56, crossed between the mountains. Another, M53, moved out of the study area and into Mexico. The rest were hemmed in by highways and housing developments.

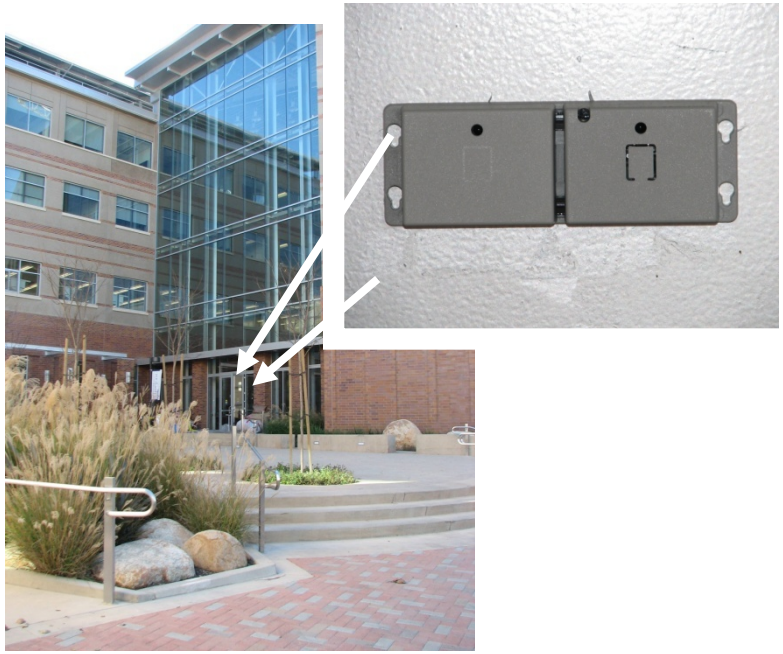


Source: UC Davis

STAFF GRAPHIC

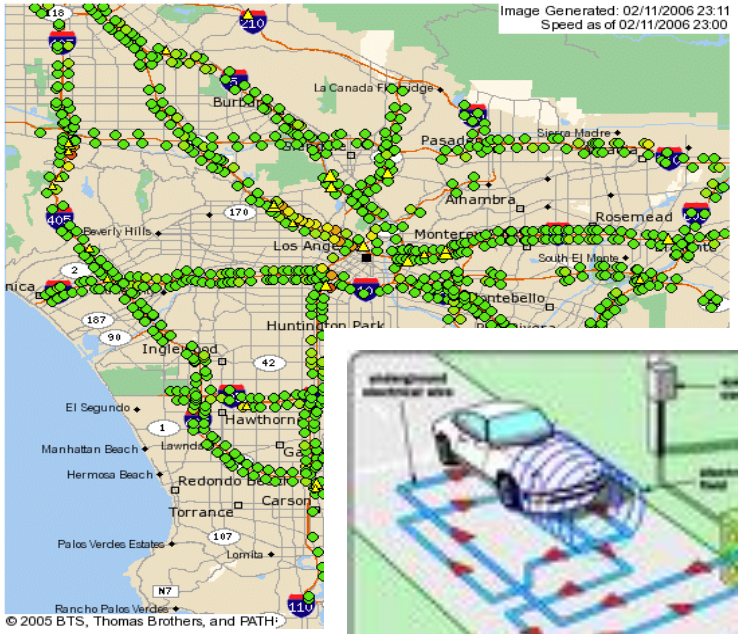
Graphic from Orange County Register

Sensors Measuring Human Activity

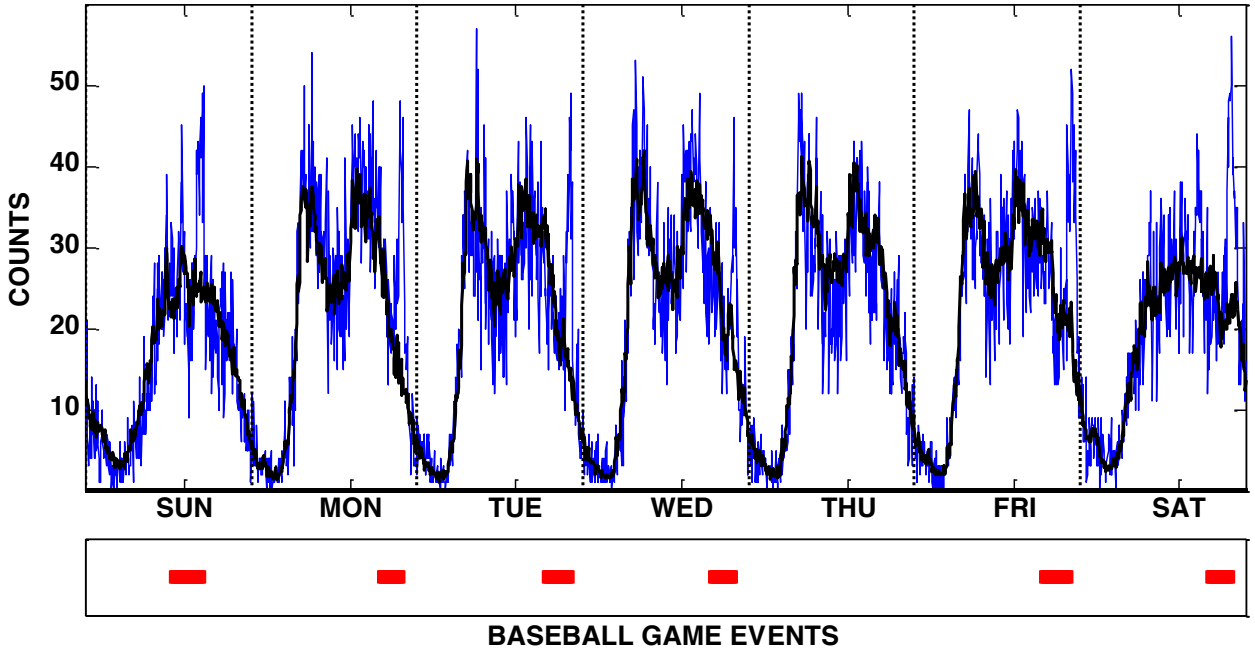


Optical people counter at a building entrance on campus

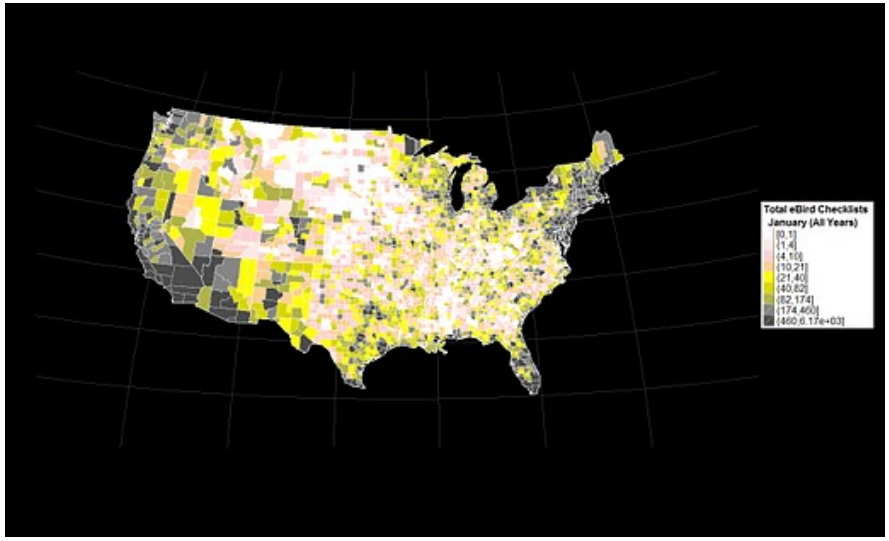
Loop sensors on Southern California freeways




Pointer 34°04'23.06" N 118°15'02.64" W elev 465 ft

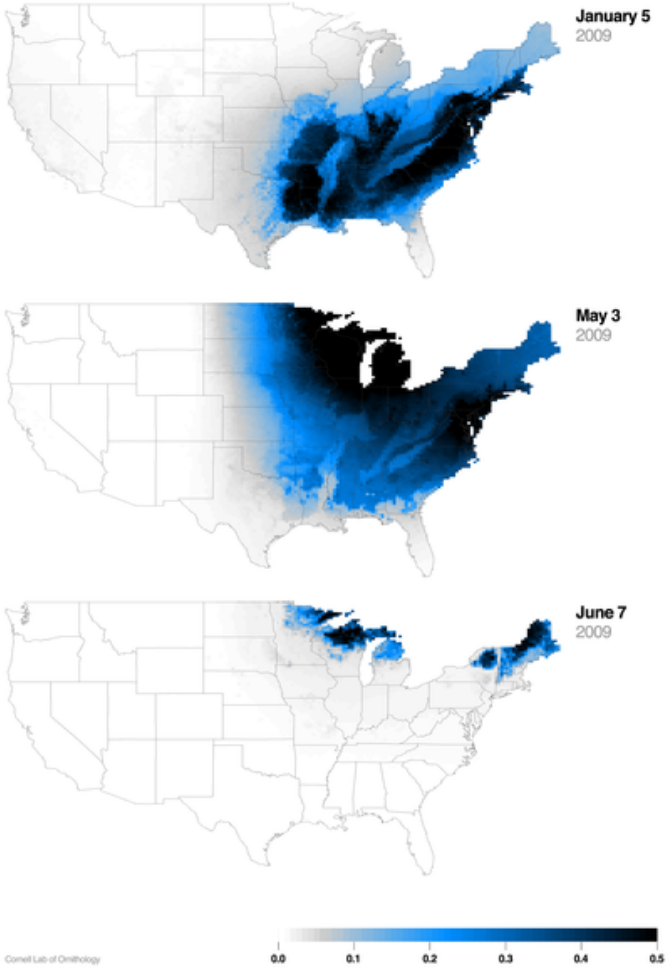

COUNTS
50
40
30
20
10
SUN
MON
TUE
WED
THU
FRI
SAT
BASEBALL GAME EVENTS

Ebird.org



Over 1.5 million submissions per month

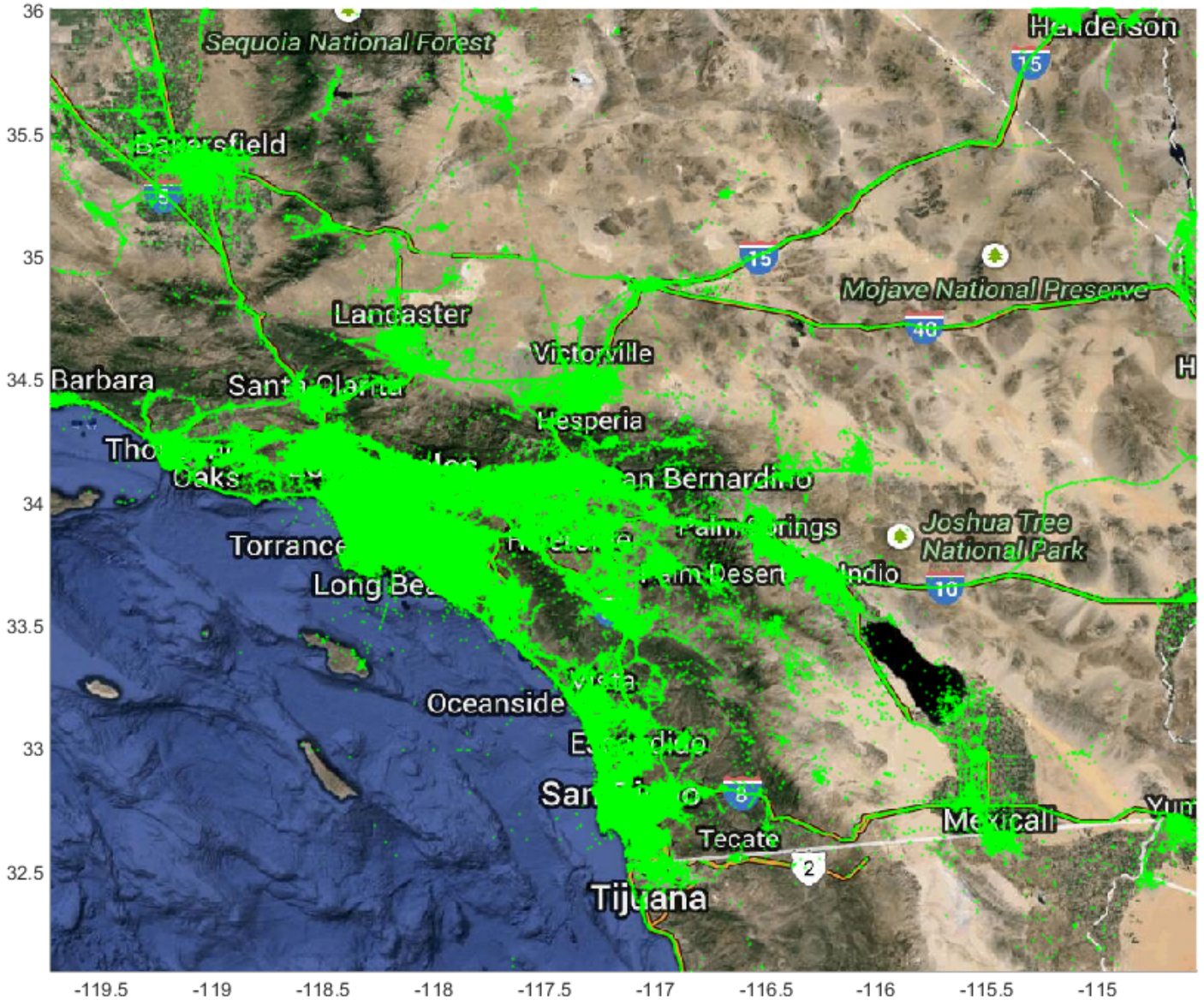
White-throated sparrow distribution



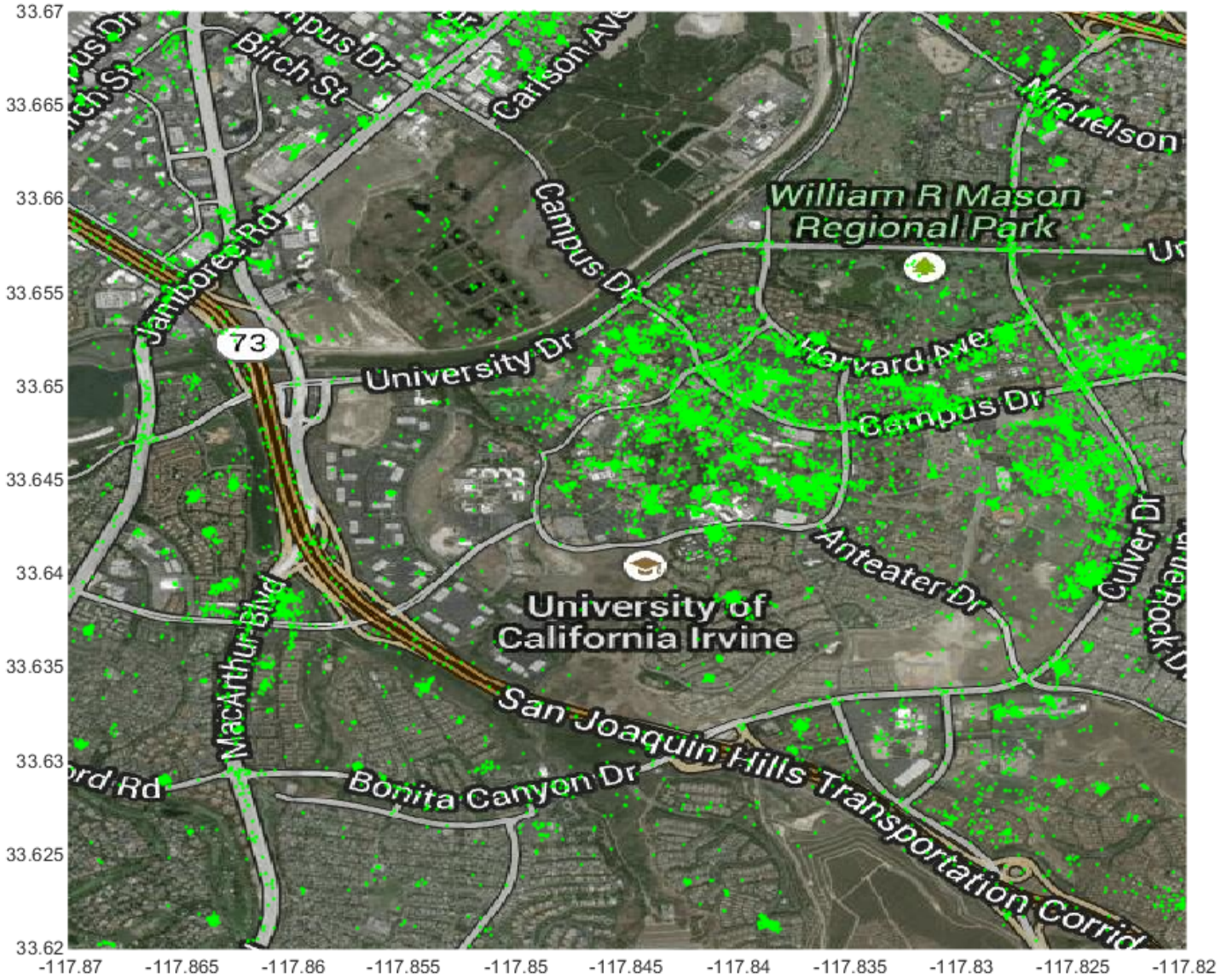
From Wood et al, PLOS Biology, 2011

What are potential issues with data collection?

Geolocated Tweets in Southern California



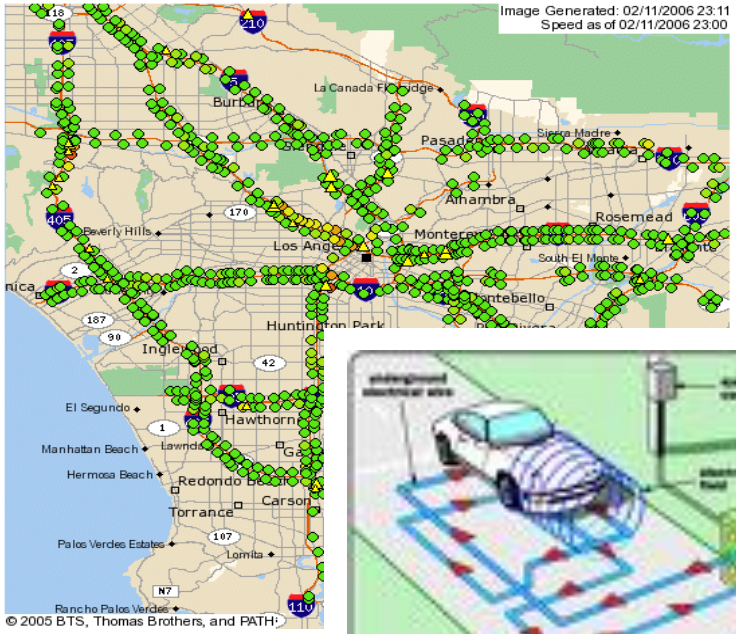
Geolocated Tweets around UC Irvine



Sensors Measuring Human Activity

Optical people counter at a building entrance on campus

Loop sensors on Southern California freeways

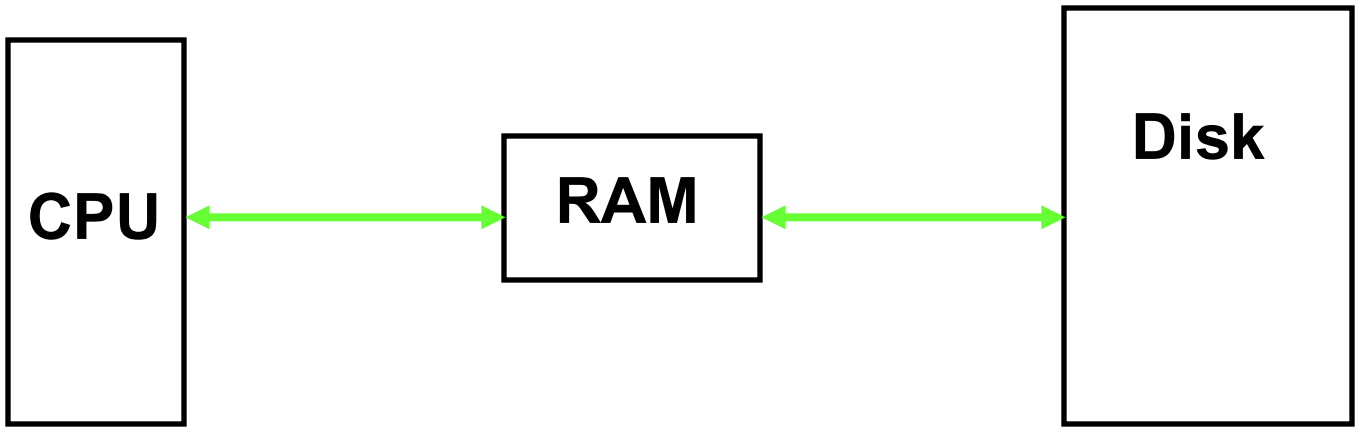


Typical Challenges with "Large Data"

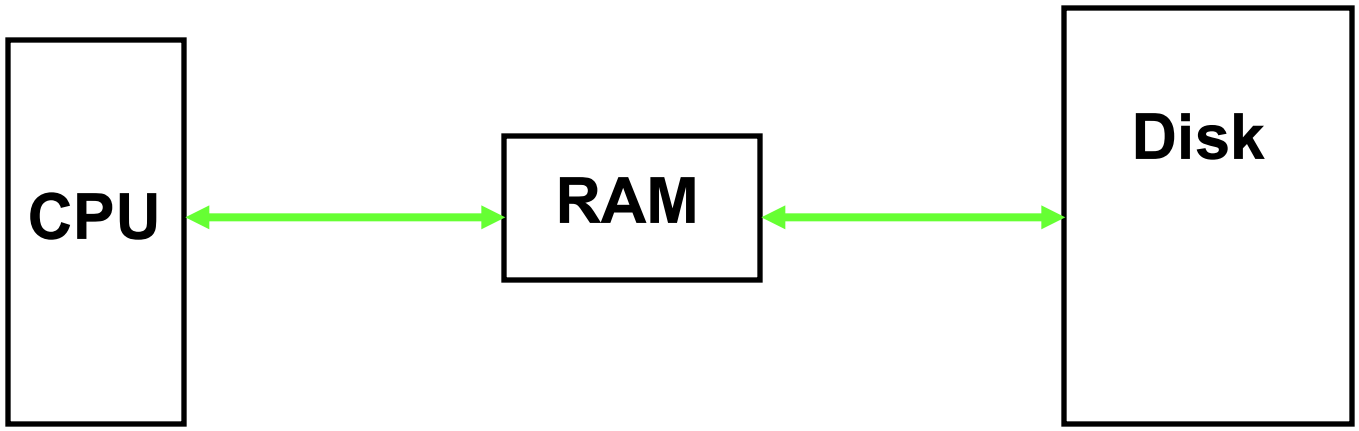
- Observational/secondary
  - Collected for some other purposes, e.g., from social media
- Noisy, Biased
  - Measurement mechanisms are often unclear, subject to whims of data owners
- Size
  - Size brings complexity: in data management, in interactive analysis, etc
- Complex and Multisource
  - e.g., text data, location data, demographic data: poses challenge in analysis
- Non-Stationary
  - Changing over time: trends, seasonality, etc

Why is data management and organization important?

Computer Architecture 101



How Far Away are the Data?

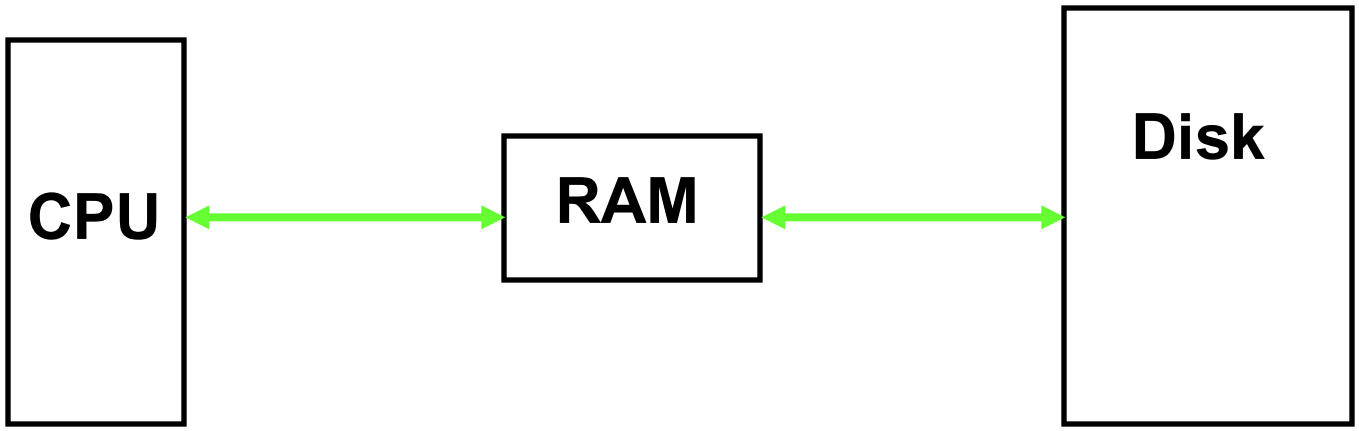


10^{-8} seconds

10^{-3} seconds

Random Access Times

How Far Away are the Data?



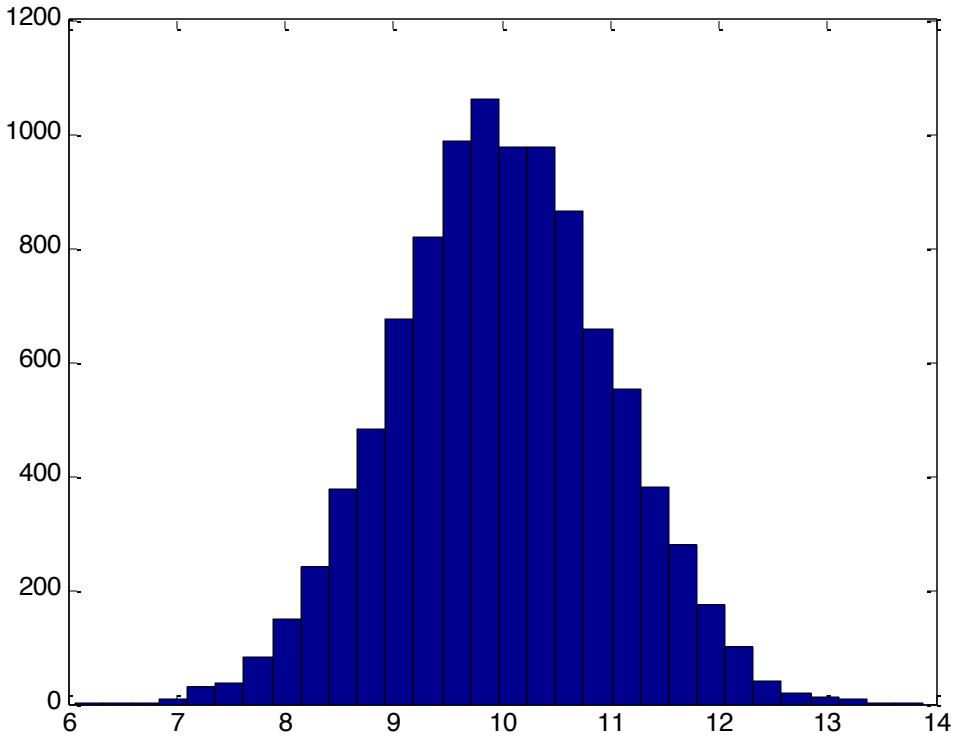
Effective Distances

Data Engineering at Web Scale

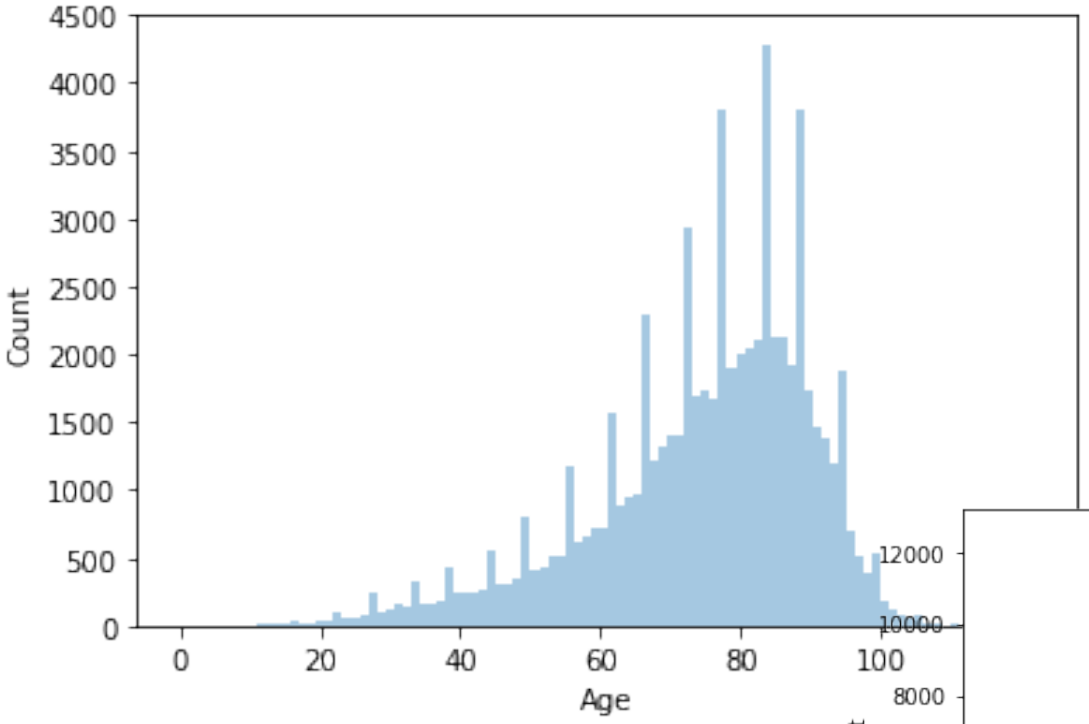
Why is it important to explore and understand data before analysis?

Histogram of Unimodal Data

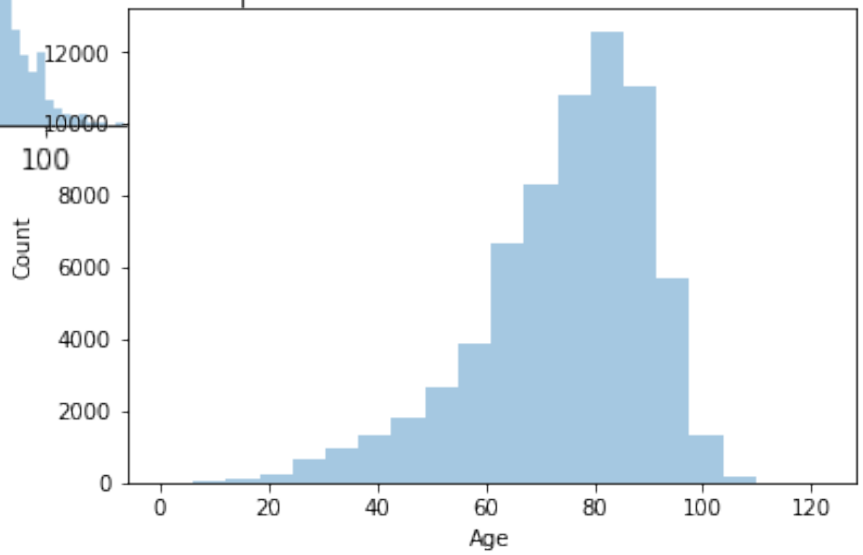
1000 data points simulated from a Normal distribution, mean 10, variance 1, 30 bins



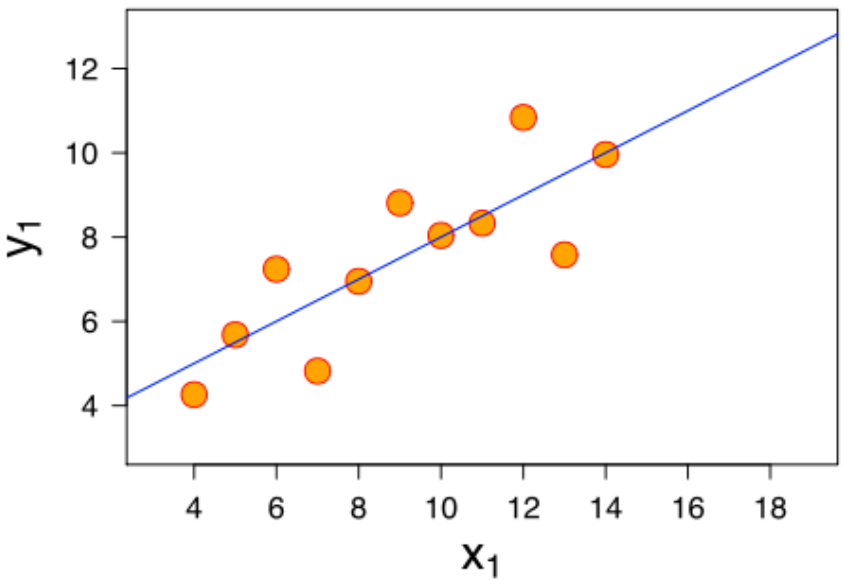
Histogram of Age at Death of 68,000 individuals



Notice anything unusual?



Summary Statistics



Summary Statistics of the Data:

$N = 11$

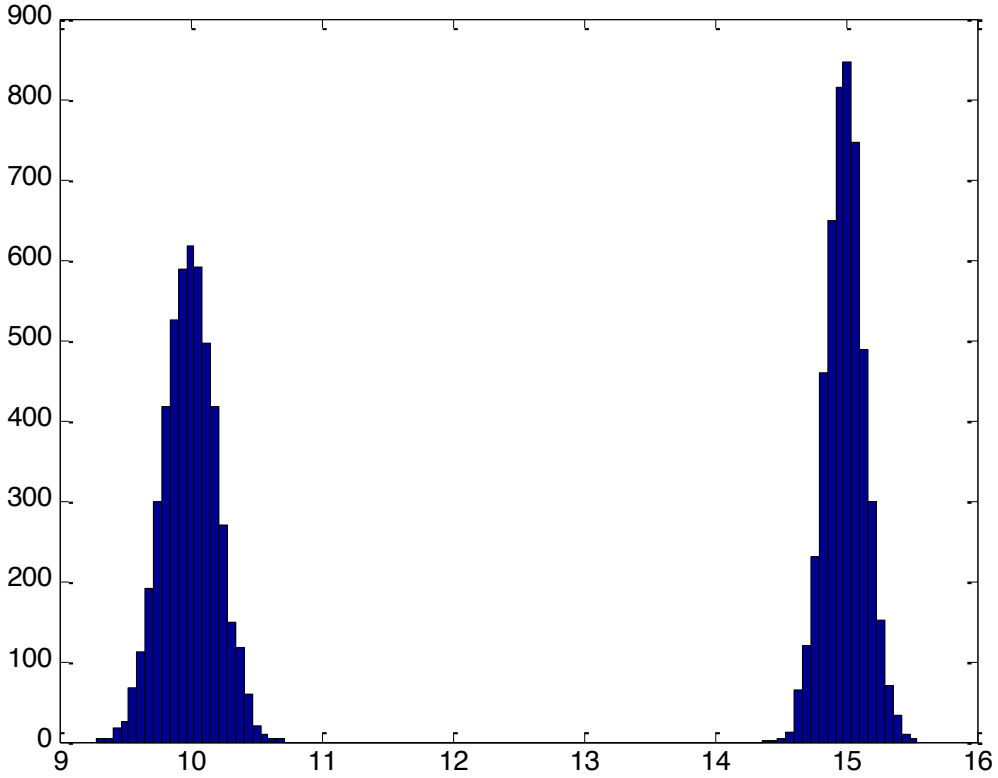
Mean of $X = 9.0$

Mean of $Y = 7.5$

Intercept = 3

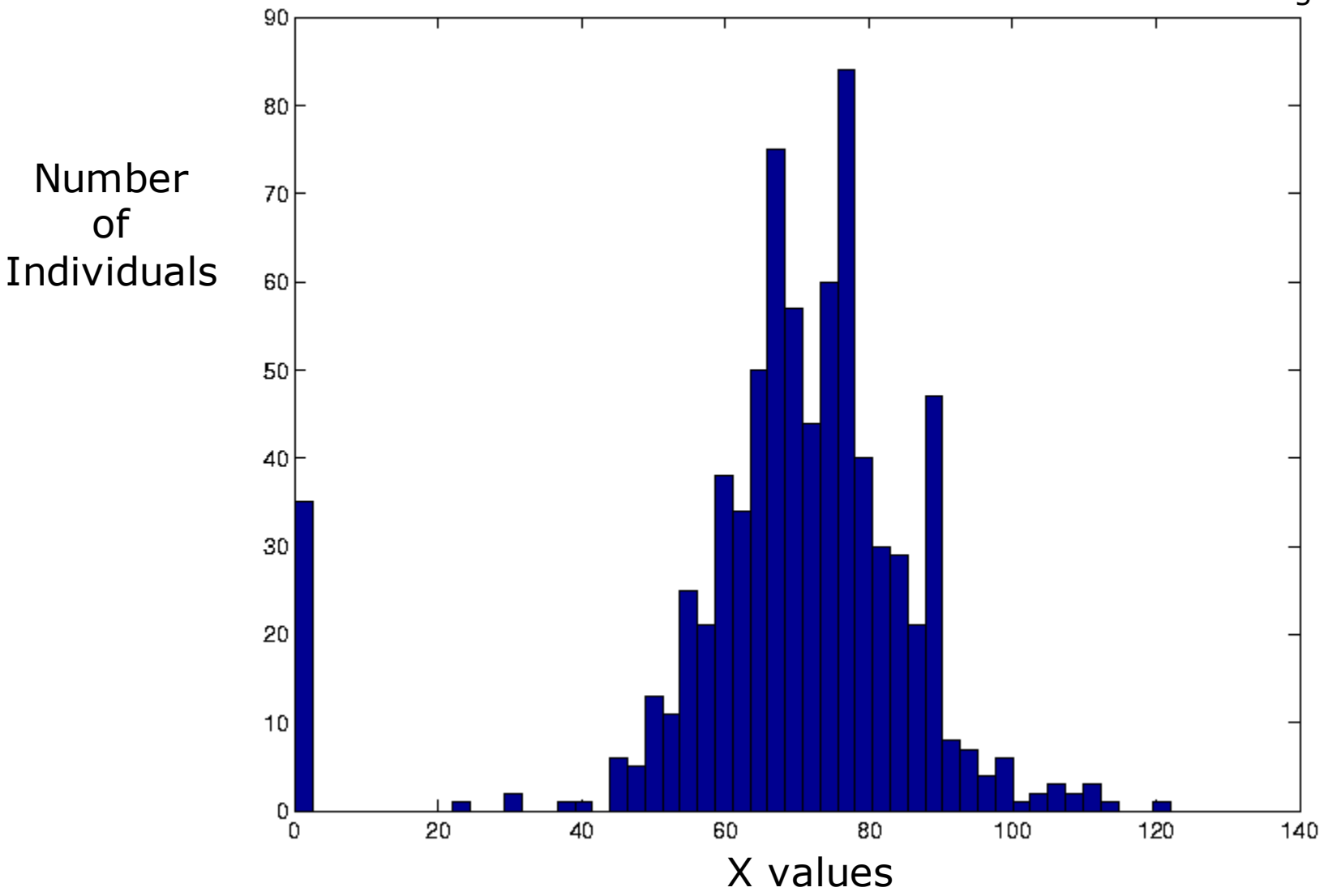
Slope = 0.5

What will the mean or median tell us about this data?



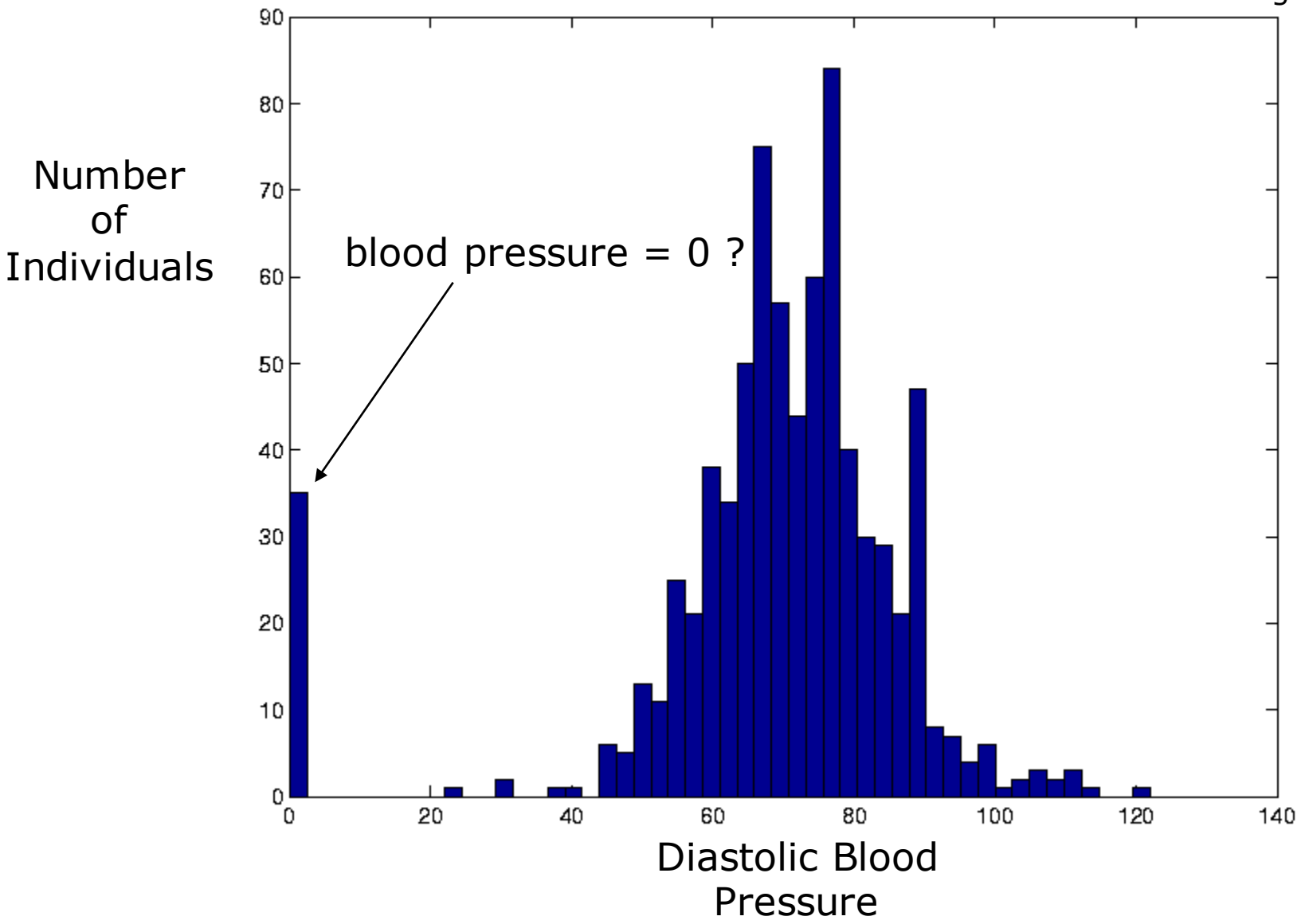
Histogram with Outliers

Pima Indians Diabetes Data,
From UC Irvine Machine Learning Reposito



Histogram with Outliers

Pima Indians Diabetes Data,
From UC Irvine Machine Learning Reposito



Matrix of Scatter Plots with Color Overlays

Iris classification data set, 3 classes

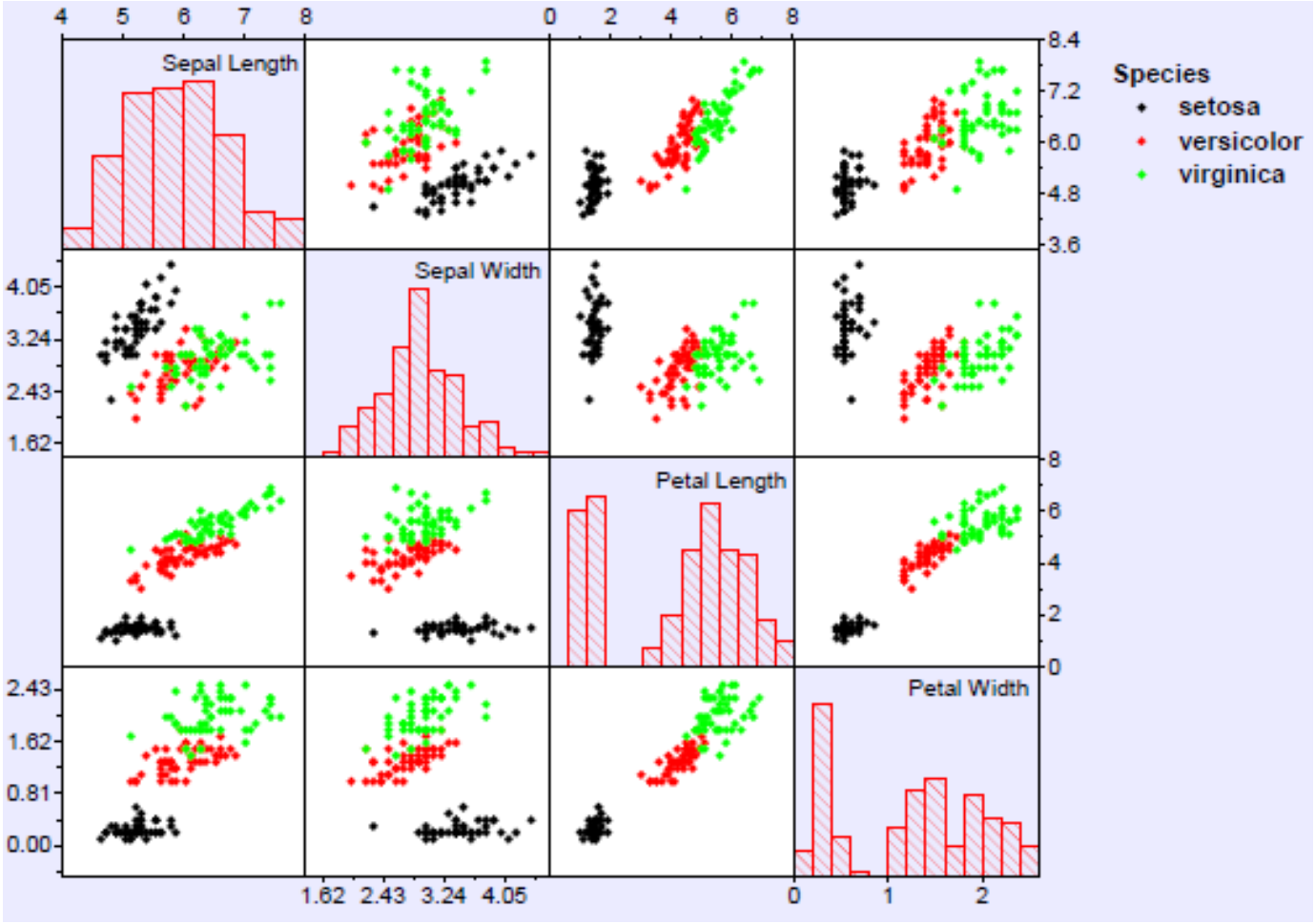


Figure from www.originlab.com

Linear Correlation Coefficient

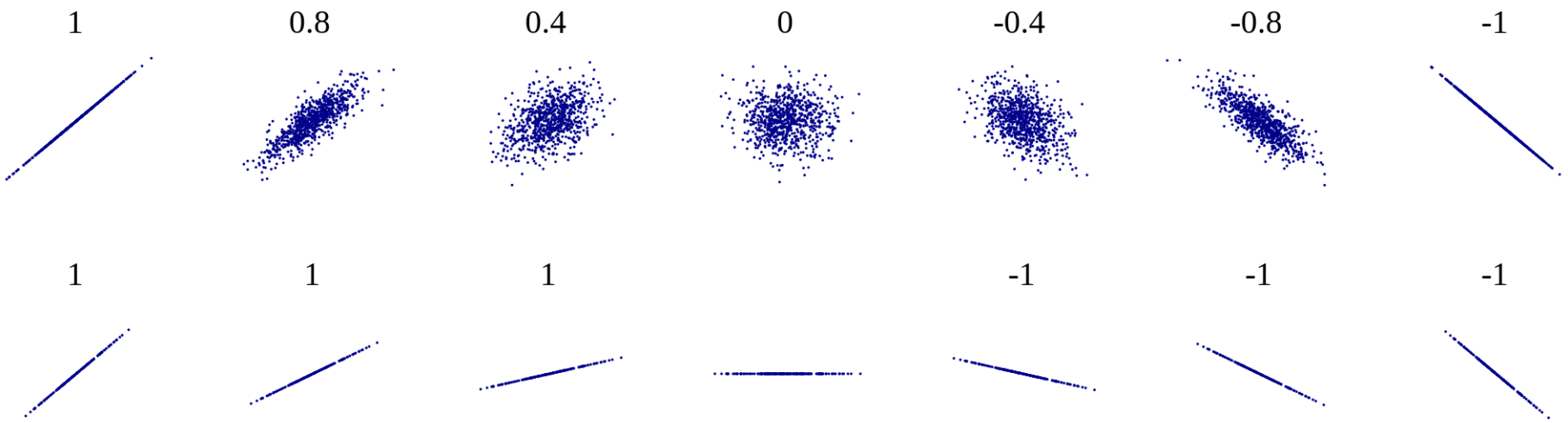
- Measures the degree of linear dependence of two variables
- Linear correlation coefficient is defined as:

$$\rho(X,Y) = \frac{\sum_{i=1}^{n}(x(i)-\bar{x})(y(i)-\bar{y})}{\left(\sum_{i=1}^{n}(x(i)-\bar{x})^2 \sum_{i=1}^{n}(y(i)-\bar{y})^2\right)^{\frac{1}{2}}}$$

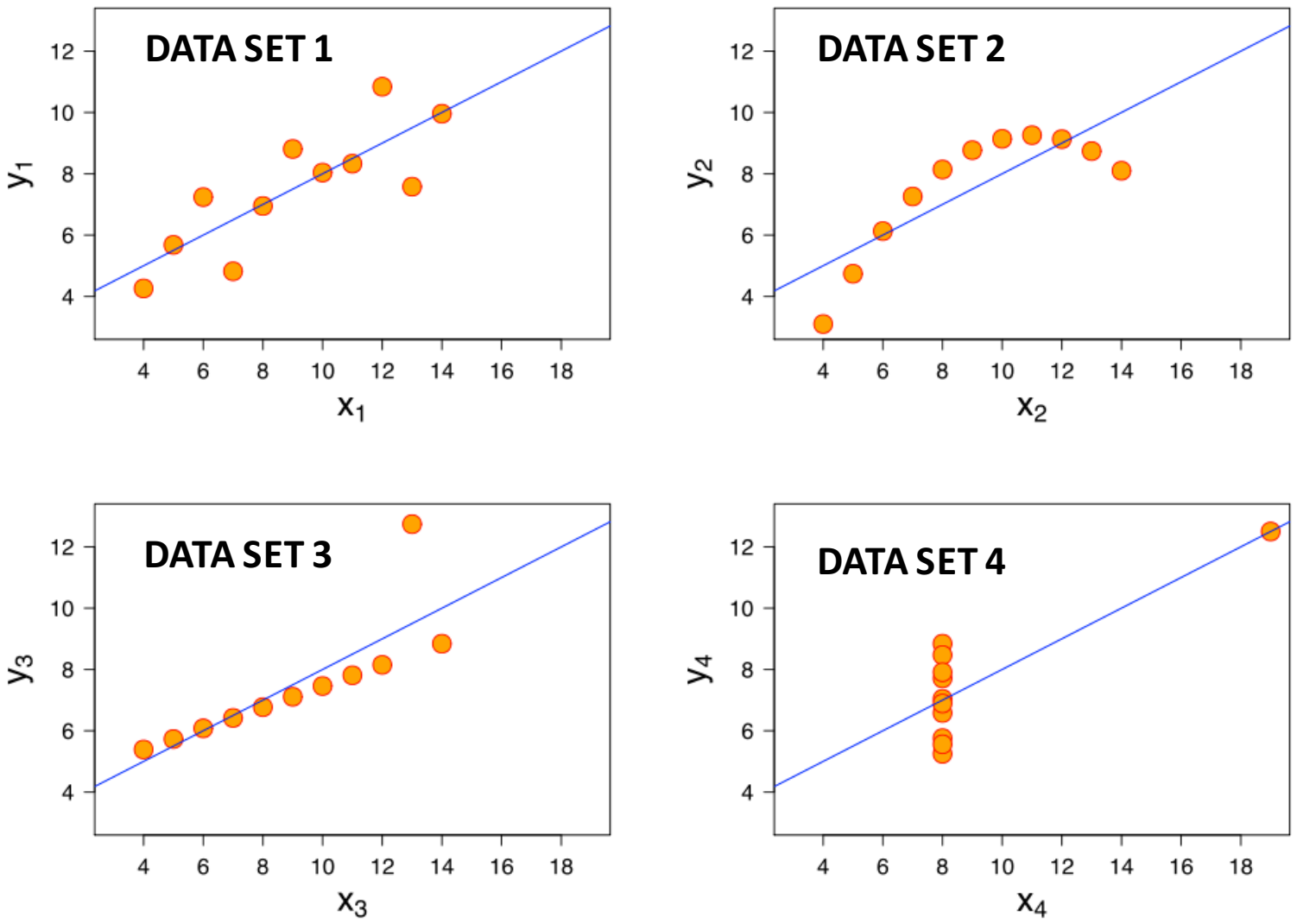
- Ranges between -1 and +1

- Note: lack of linear correlation does not imply lack of dependence

Examples of X-Y plots and linear correlation values

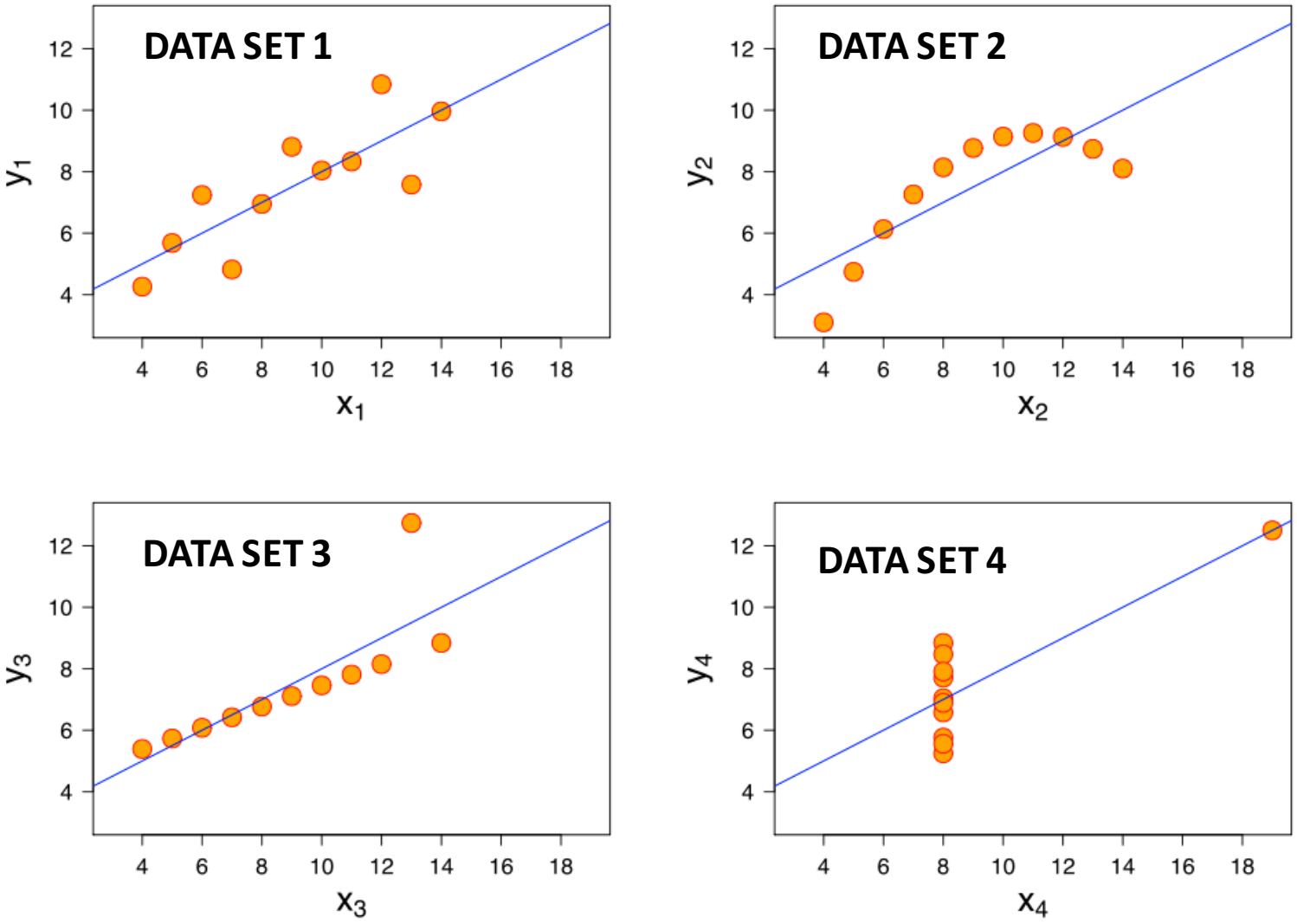


Example: 4 Data Sets, Y versus X



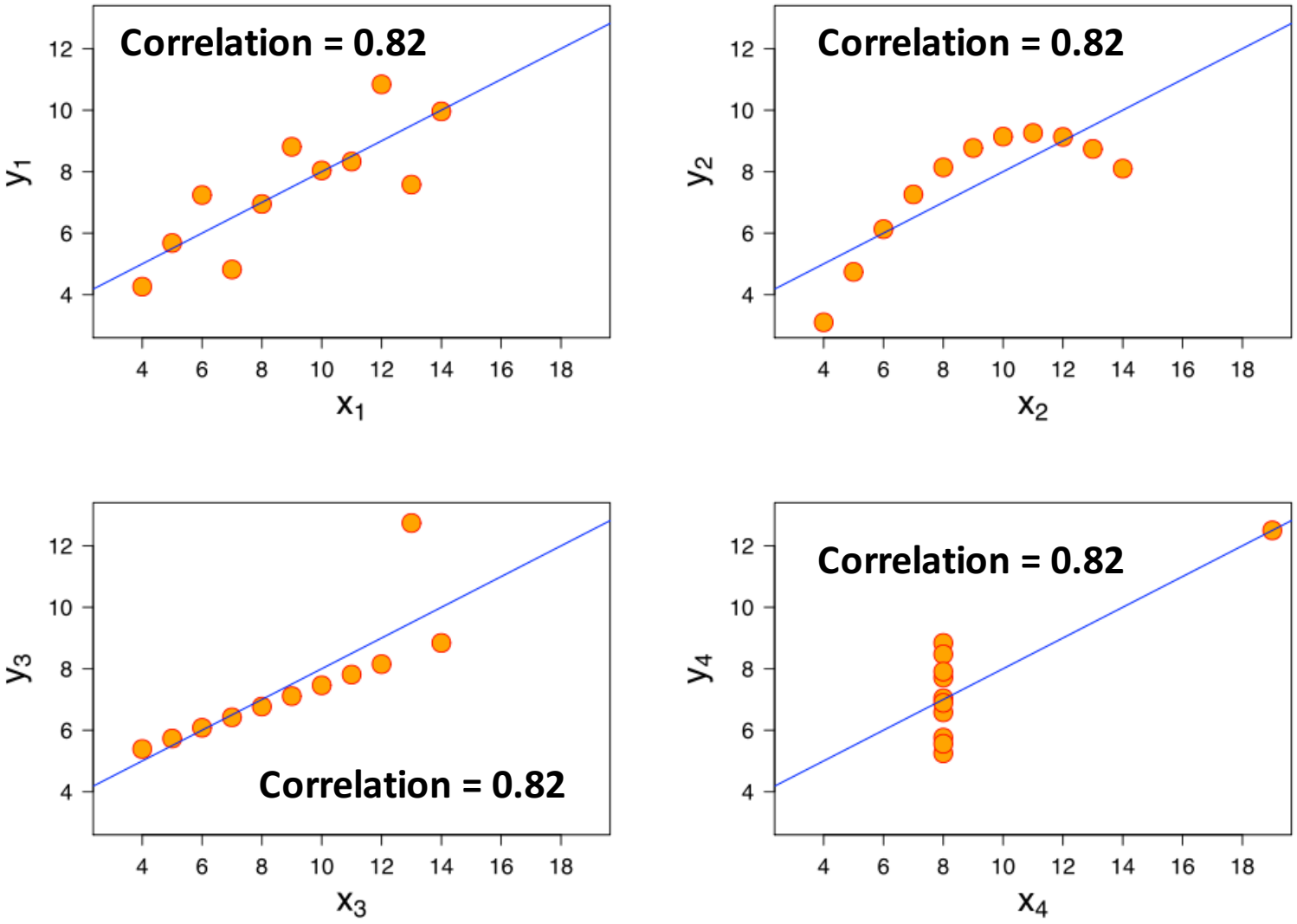
Anscombe, Francis (1973), *Graphs in Statistical Analysis*, The American Statistician, pp. 195-199.

Guess the Linear Correlation Values for each Data Set



Anscombe, Francis (1973), *Graphs in Statistical Analysis*, The American Statistician, pp. 195-199.

Actual Correlation Values



Anscombe, Francis (1973), *Graphs in Statistical Analysis*, The American Statistician, pp. 195-199.

Summary Statistics for each Data Set

Summary Statistics of Data Set 1
N = 11
Mean of X = 9.0
Mean of Y = 7.5
Intercept = 3
Slope = 0.5
Correlation = 0.82

Summary Statistics of Data Set 2
N = 11
Mean of X = 9.0
Mean of Y = 7.5
Intercept = 3
Slope = 0.5
Correlation = 0.82

Summary Statistics of Data Set 3
N = 11
Mean of X = 9.0
Mean of Y = 7.5
Intercept = 3
Slope = 0.5
Correlation = 0.82

Summary Statistics of Data Set 4
N = 11
Mean of X = 9.0
Mean of Y = 7.5
Intercept = 3
Slope = 0.5
Correlation = 0.82

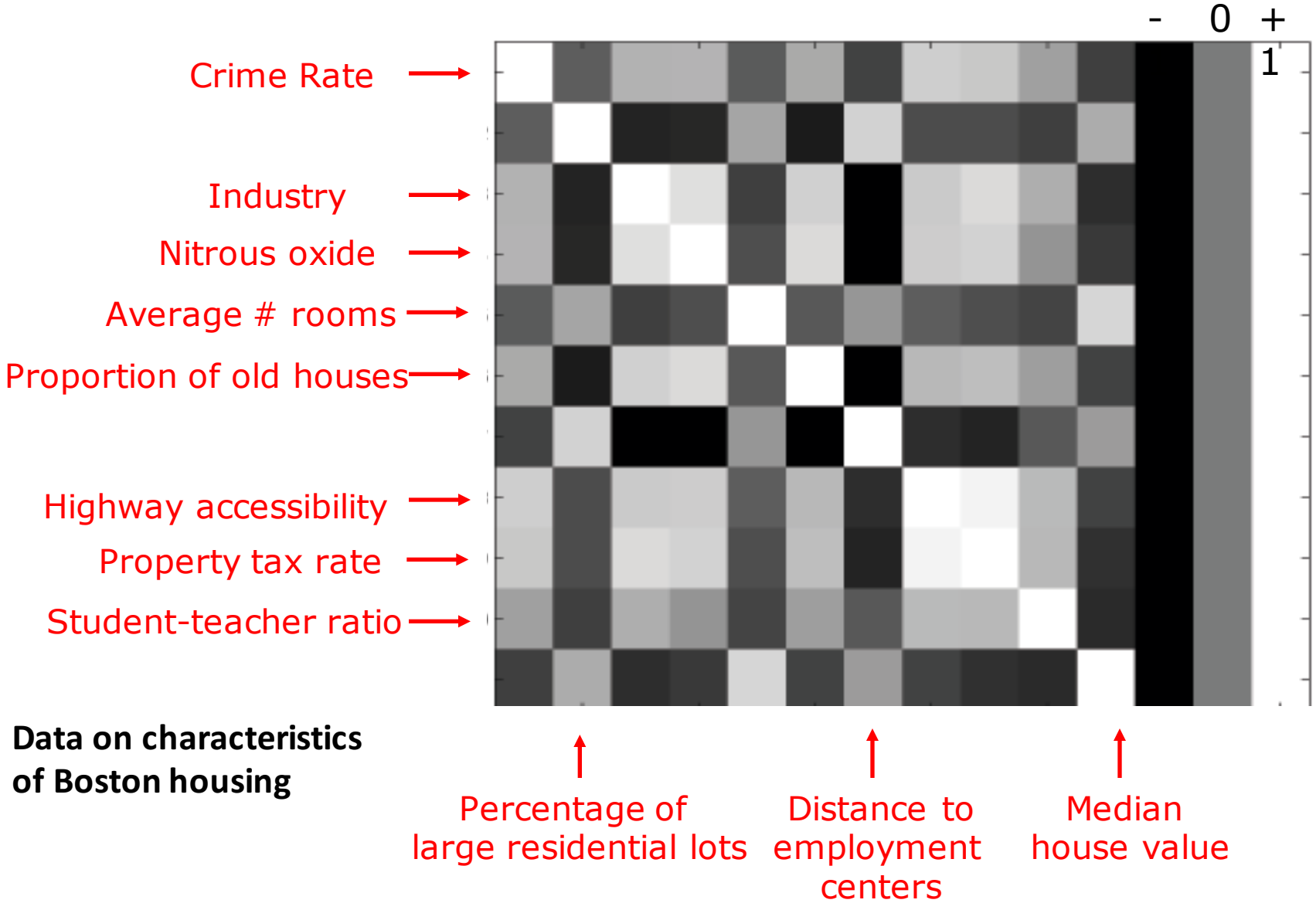
Anscombe, Francis (1973), *Graphs in Statistical Analysis*, The American Statistician, pp. 195-199.

Data Set on Housing Prices in Boston

(widely used data set in research on prediction models)

1	CRIM	per capita crime rate by town
2	ZN	proportion of residential land zoned for lots over 25,000 ft^2
3	INDUS	proportion of non-retail business acres per town
4	NOX	Nitrogen oxide concentration (parts per 10 million)
5	RM	average number of rooms per dwelling
6	AGE	proportion of owner-occupied units built prior to 1940
7	DIS	weighted distances to five Boston employment centres
8	RAD	index of accessibility to radial highways
9	TAX	full-value property-tax rate per $10,000
10	PTRATIO	pupil-teacher ratio by town
11	MEDV	Median value of owner-occupied homes in $1000's

Matrix of Pairwise Linear Correlations



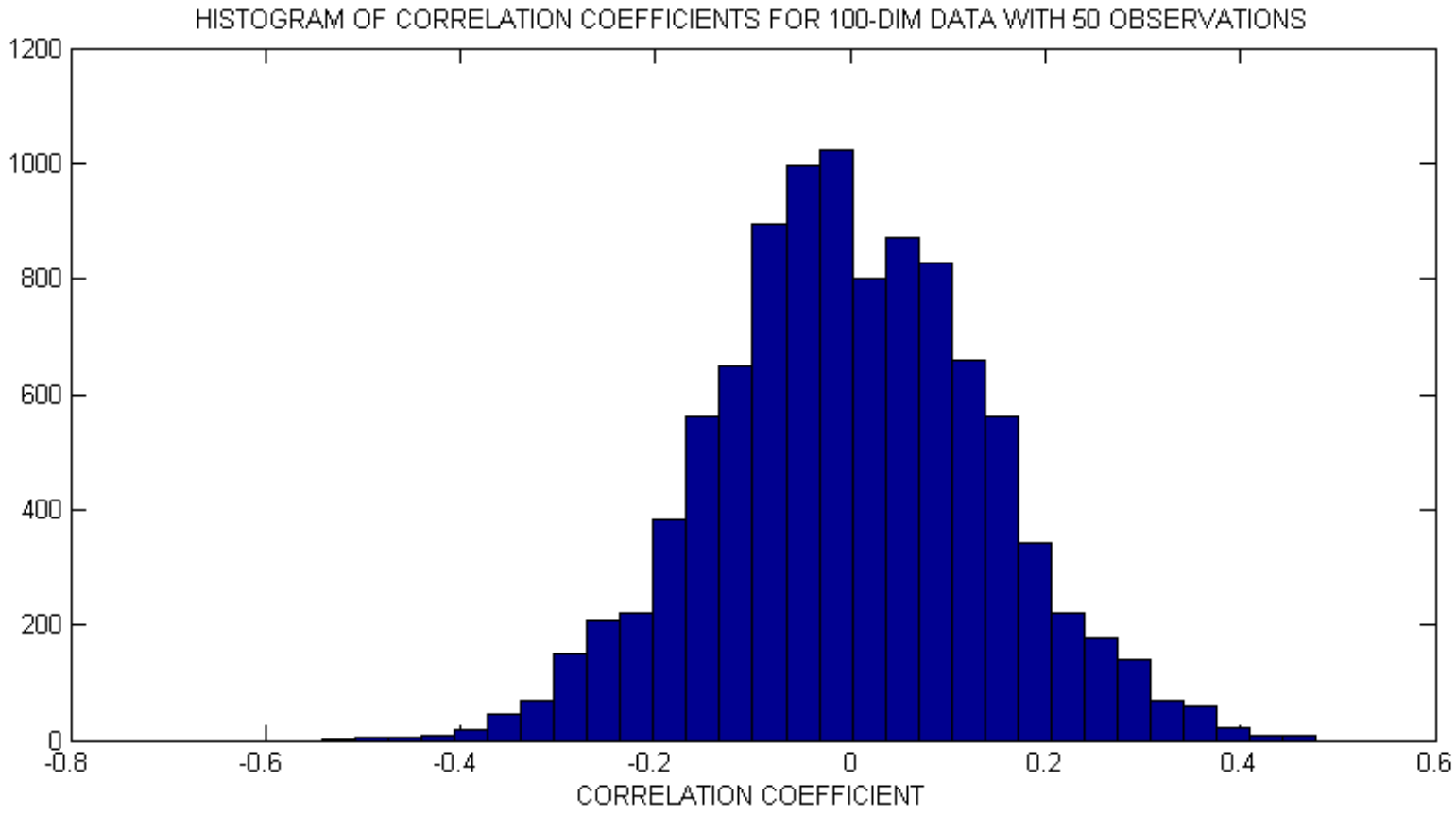
Data on characteristics of Boston housing

Human judgement is important in data analysis

Example: a data set with

- 100 independent variables
- Simulate 50 data vectors
- Compute the correlation of all pairs of variables from the data
- This gives us 50*49/2 correlation values

What do you think these correlation values will look like if we plot them as a histogram?



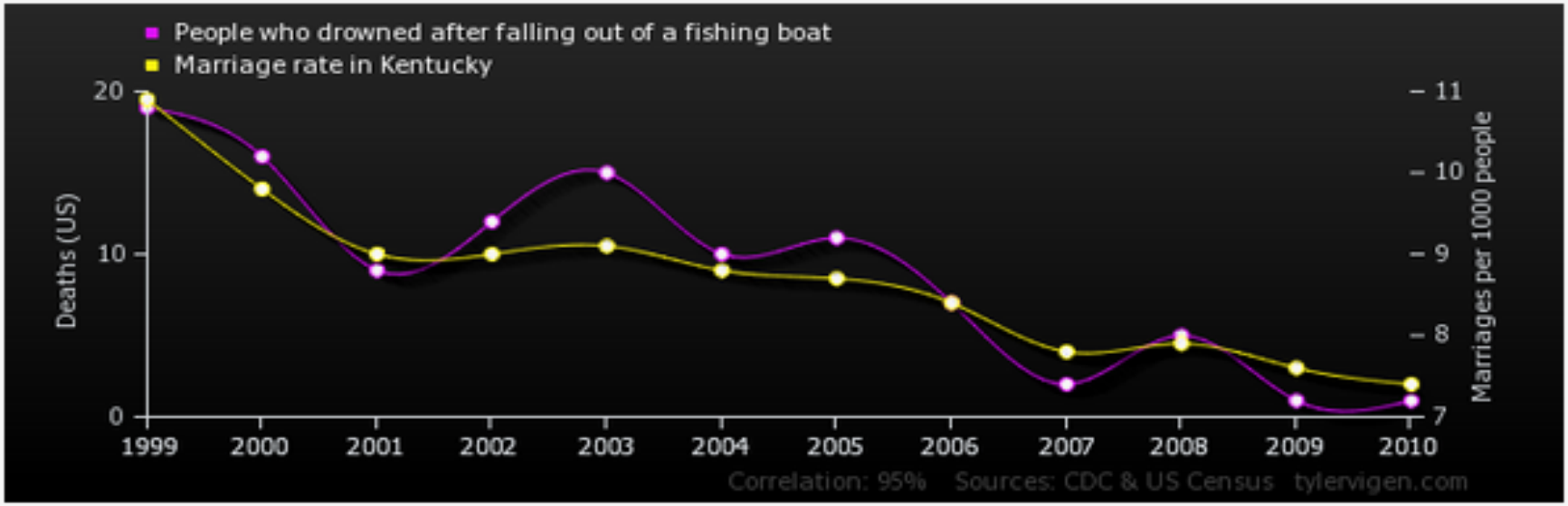
Conclusion: even if data are entirely random (no dependence) there is a very high probability some variables will appear dependent just by chance.

This is sometimes referred to as “data fishing”

People who drowned after falling out of a fishing boat

correlates with

Marriage rate in Kentucky



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
People who drowned after falling out of a fishing boat *Deaths (US) (CDC)*	19	16	9	12	15	10	11	7	2	5	1	1
Marriage rate in Kentucky *Marriages per 1000 people (US Census)*	10.9	9.8	9	9	9.1	8.8	8.7	8.4	7.8	7.9	7.6	7.4

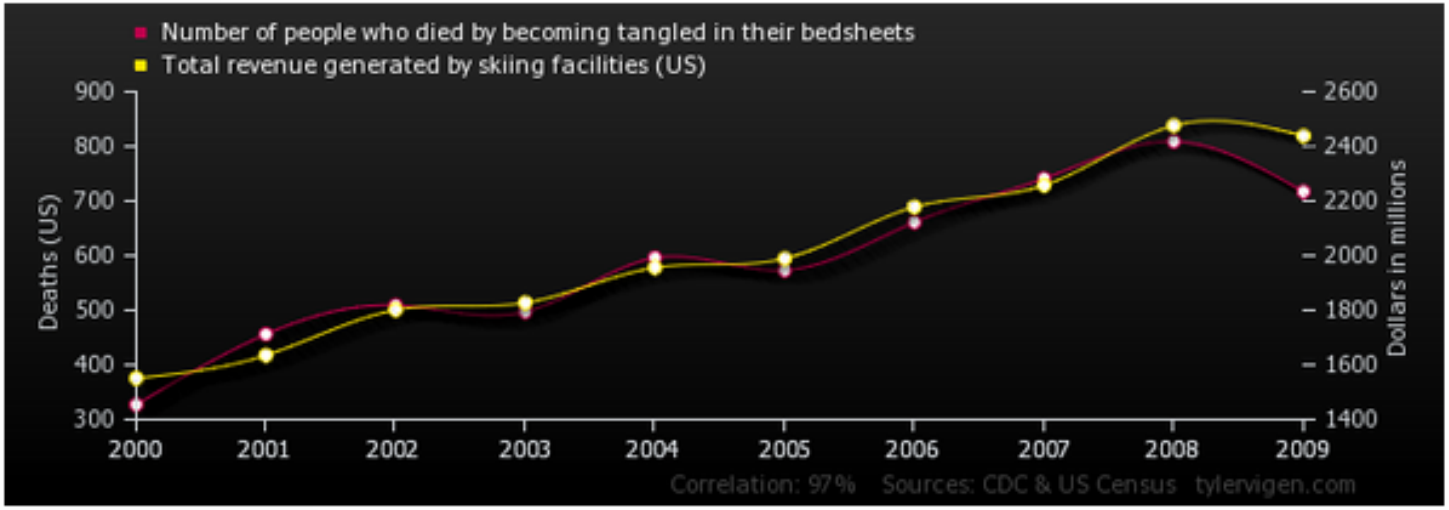
Correlation: 0.952407

Graphics from http://www.tylervigen.com/

Number of people who died by becoming tangled in their bedsheets

correlates with

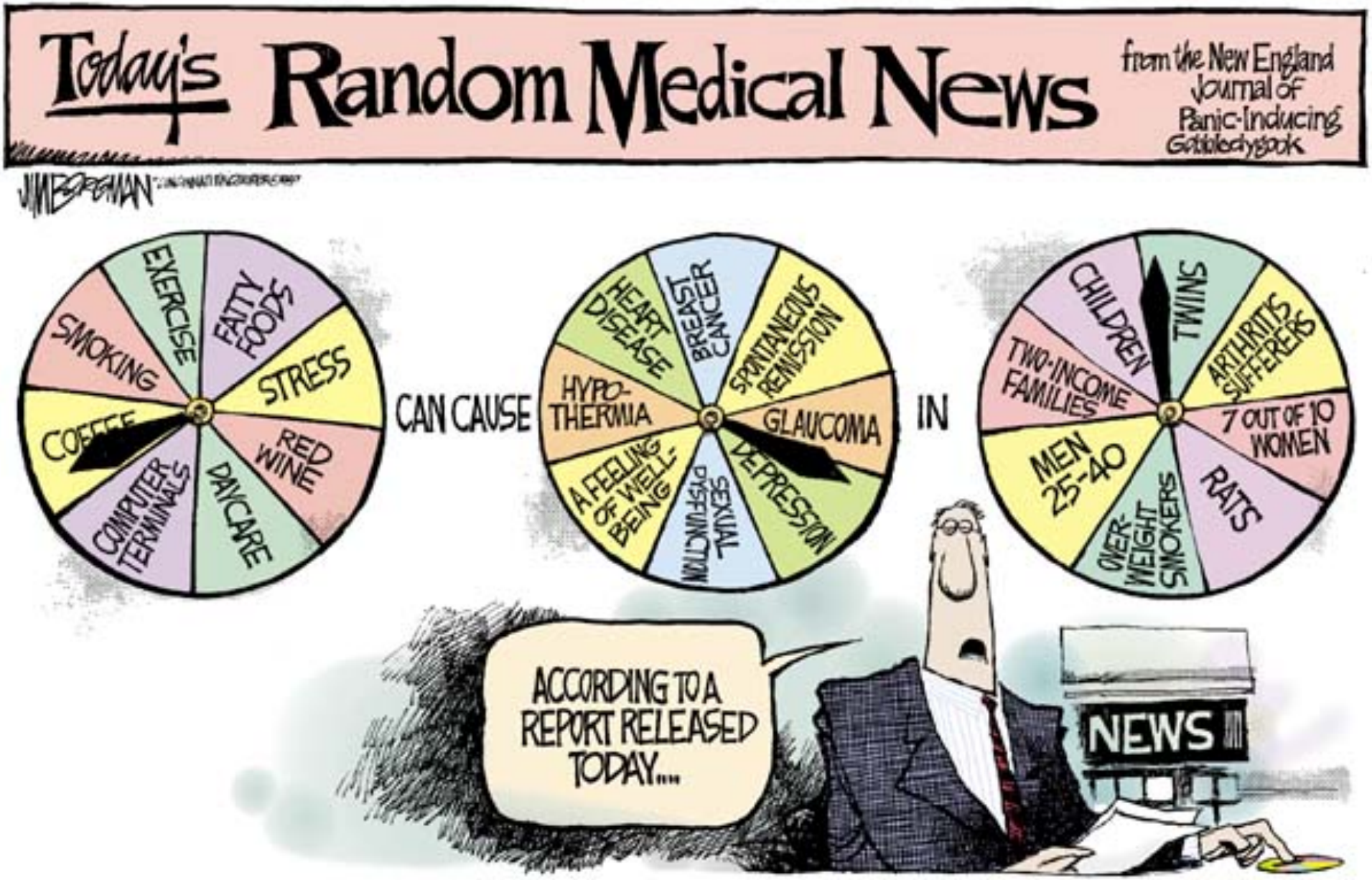
Total revenue generated by skiing facilities (US)



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Number of people who died by becoming tangled in their bedsheets *Deaths (US) (CDC)*	327	456	509	497	596	573	661	741	809	717
Total revenue generated by skiing facilities (US) *Dollars in millions (US Census)*	1,551	1,635	1,801	1,827	1,956	1,989	2,178	2,257	2,476	2,438

Correlation: 0.969724

Graphics from http://www.tylervigen.com/


Today's Random Medical News
from the New England Journal of Panic-Inducing Gobbledygook
JIM BORGMAN
SMOKING
EXERCISE
FATTY FOODS
STRESS
COFFEE
RED WINE
COMPUTER TERMINALS
DAYCARE
CAN CAUSE
HEART DISEASE
BREAST CANCER
SPONTANEOUS REMISSION
HYPO-THERMIA
GLAUCOMA
A FEELING OF WELL-BEING
SEXUAL DYSFUNCTION
DEPRESSION
IN
CHILDREN
TWINS
ARTHRITIS SUFFERERS
TWO-INCOME FAMILIES
7 OUT OF 10 WOMEN
MEN 25-40
RATS
OVER-WEIGHT SMOKERS
ACCORDING TO A REPORT RELEASED TODAY...
NEWS

Another Example: Automated Essay Grading

From Inside Higher Ed, April 2012

Report on a major study comparing automated essay-grading software with trained human readers, on 22,000 high-school essays.

"The differences, across a number of different brands of automated essay scoring software (AES) and essay types, were minute. "

Why is automated essay grading of interest?

Human graders: 20 to 30 essays an hour

Automated: millions per hour

Human Interpretation of Automated Essay Grading

From New Statesman and New York Times, April 2012

Les Perelman, MIT, experimented with different essays to test the Educational Testing Service (ETS)'s automated eRater program

All of his essays received a perfect score

Human Interpretation of Automated Essay Grading

From New Statesman and New York Times, April 2012

SAT prompt:

"*The rising cost of a college education is the fault of students who demand that colleges offer students luxuries unheard of by earlier generations of college students -- single dorm rooms, private bathrooms, gourmet meals, etc.*"

Discuss the extent to which you agree or disagree with this opinion. Support your views with specific reasons and examples from your own experience, observations, or reading.

Portions of a Perfect-Scoring Essay

Teaching assistants are paid an excessive amount of money. The average teaching assistant makes six times as much money as college presidents. In addition, they often receive a plethora of extra benefits such as private jets, vacations in the south seas, a staring roles in motion pictures.

Portions of a Perfect-Scoring Essay

In Heart of Darkness, Mr. Kurtz is a teaching assistant because of his connections, and he ruins all the universities that employ him. Finally, teaching assistants are able to exercise mind control over the rest of the university community. The last reason to write this way is the most important. Once you have it down, you can use it for practically anything. Does God exist? Well, you can say yes and give three reasons, or no and give three different reasons. It doesn't really matter.

What are the legal and ethical aspects of data analysis?

Who Owns Your Data?

Collection of Individual-Level Data

1960's

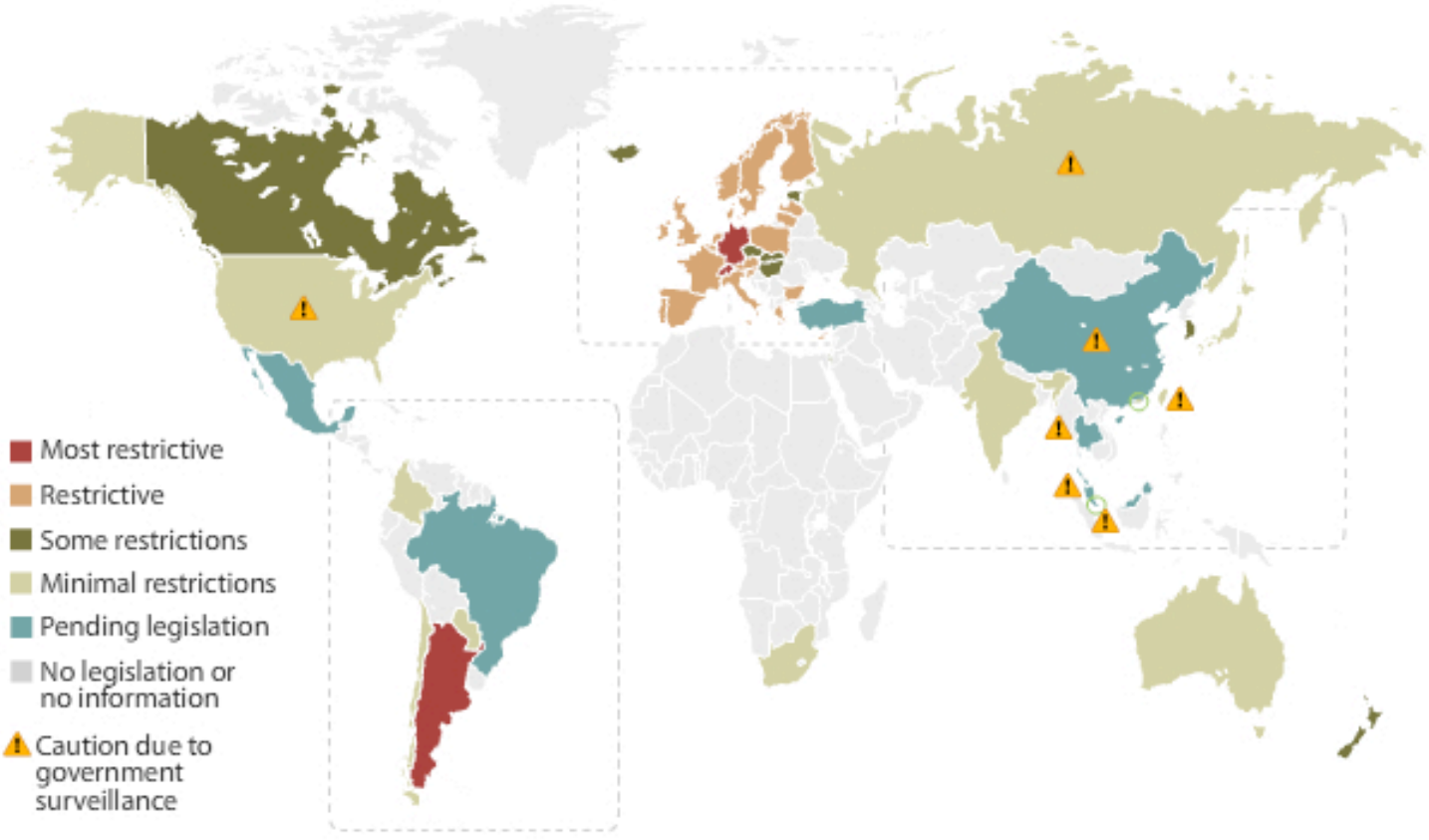
1980's

2000's

2020's

BIG DATA:
SEIZING OPPORTUNITIES,
PRESERVING VALUES

Executi



Source: US Department of Commerce and country specific legislation

Source: Forrester Research, Inc.

The Future of Data Science

What types of new data might we collect?

What new analysis techniques might be developed?

What new application areas might emerge?

What are the societal implications of data science?

Final Assignment

- Write a ½ to 1 page short essay that takes any two of the topics from lectures 2 to 9, and describes how you think the two topics could "intersect" going forward,
  e.g.,
  - What aspects of each method could be combined to produce new ideas?
  - What new applications might be enabled by combining these methods?
  - What are the potential challenges in these areas?

- Possible combinations
  - Natural language and cybersecurity
  - Clustering algorithms and computer vision
  - Computer vision and fairness/bias

  - …feel free to pick any 2 topics that interest you

Final Assignment Instructions

- Put your name and student ID at the top of the page
- Submit as a PDF file
- Due to EEE dropbox by 9am on Monday March 19th (next week)
- Note: there is **no final exam** in this class