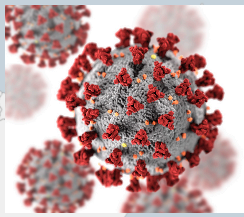
Overview of Data Science

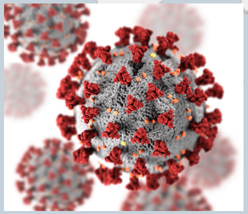
Promises and Pitfalls, Tools and Techniques

Professor Widom's Instructional Odyssey

www.professorwidom.org



COVID Edition



Stanford University

Google

Instabase

Google Cloud Platform

Association for Computing Machinery

Very Large Data Bases Endowment Inc.

amazon web services™

The "Instructional Odyssey"

STANFORD COMPUTER SCIENCE PROFESSOR OFFERS FREE IN-PERSON SHORT-COURSES AND WORKSHOPS, WORLDWIDE

Stanford University

Jennifer Widom, Stanford computer science professor, school of engineering dean, and MOOC pioneer, has traveled the globe offering free short-courses in data science, workshops in design thinking & collaborative problem-solving, and roundtables with women in technology. She began the odyssey full-time during her 2016-17 sabbatical year and has continued since then on an occasional basis. Navigate the website to learn more about Professor Widom and her motivation for the endeavor, to see details about the courses and workshops offered, to locate course materials, social media, and participant feedback, and to find out how to be an institutional host.

Locations (2016-2020)

The Second "COVID Edition"

Overview of Data Science

Promises and Pitfalls, Tools and Techniques

Stanford University

Google

Google Cloud Platform

Association for Computing Machinery

Very Large Data Bases Endowment Inc.

Instabase

amazon web services

Data is Everywhere

- Explosion in data-driven scientific discovery, business practices, medicine, education, politics, societal interventions, …
- And it's just the beginning
  - Ability to collect data across many domains will continue to accelerate
  - Data analysis techniques will continue to improve

"Data is the oil of the 21st century"

Data is Everywhere

- Explosion in data-driven scientific discovery, business practices, medicine, education, politics, societal interventions, …
- And it's just the beginning
  - Ability to collect data across many domains will continue to accelerate
  - Data analysis techniques will continue to improve

"Data is the ~~oil~~ **fuel** of the 21st century"

The Two Steps of Working with Data

(1) Collect data

Via computers, sensors, people, events …

(2) Do something with it

Make decisions, confirm hypotheses, gain insights, predict future …

"Data Science" = Going from (1) to (2)

This Session

- Promises of data science
  - Applications and services
- Data tools and techniques
  - Data analysis
  - Data mining and machine learning
  - Data visualization
- Pitfalls in data science
  - Correlation and causation
  - Underfitting and overfitting
  - Privacy and a few others
- Data systems and platforms

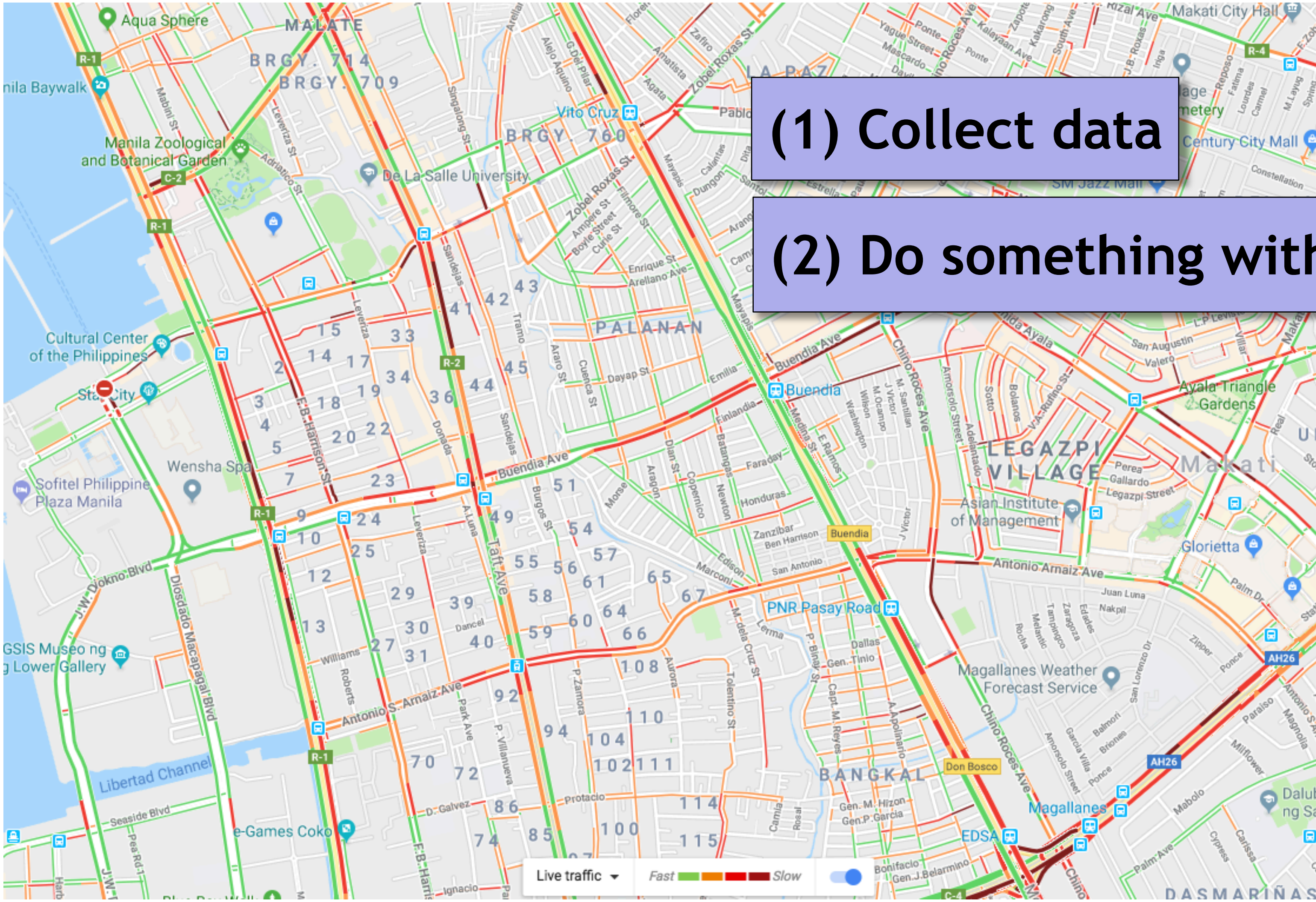
Promises of Data Science

(1) Collect data

(2) Do something with it

beneficial

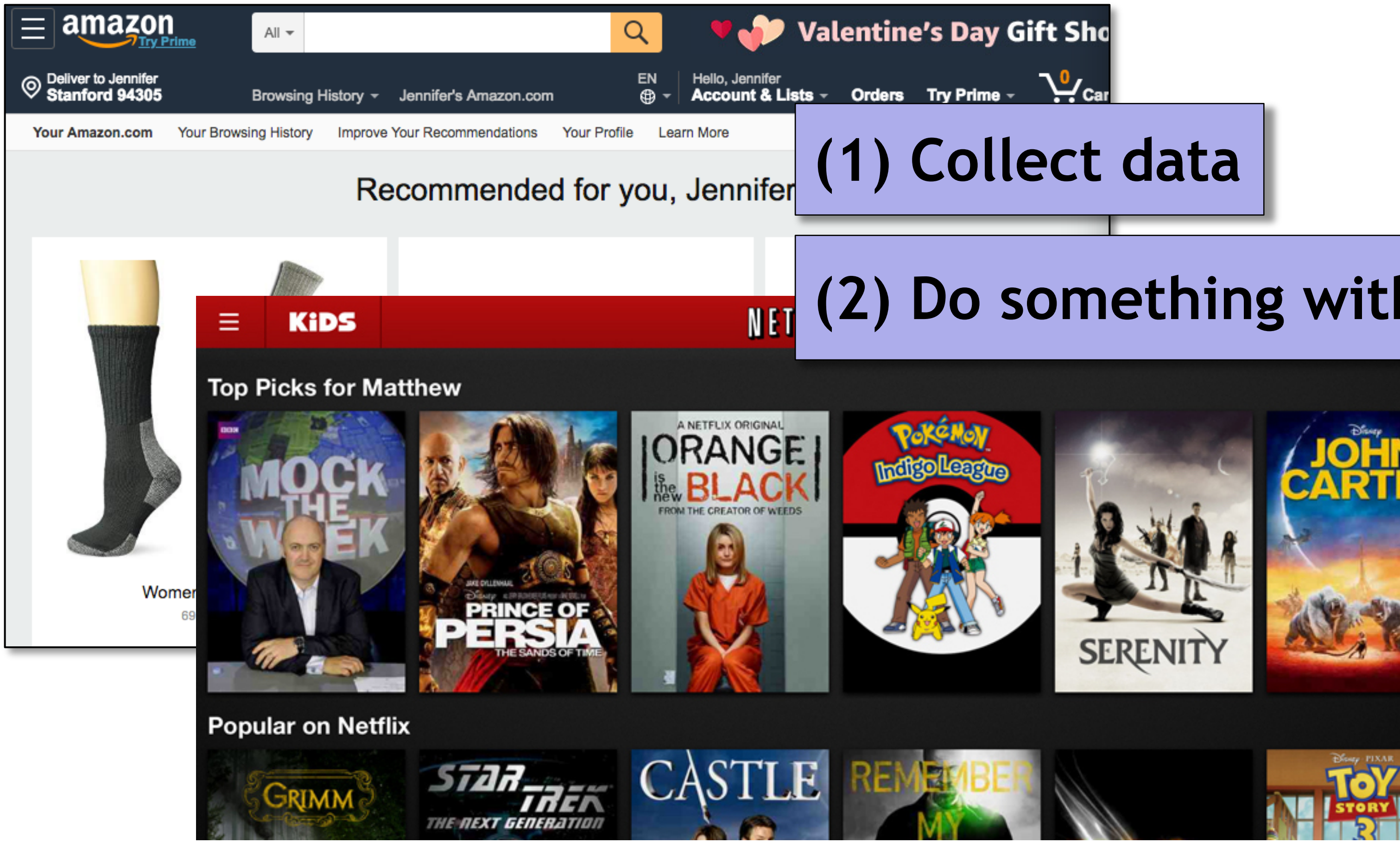
Traffic



(1) Collect data

(2) Do something with it

Recommender Systems



(1) Collect data

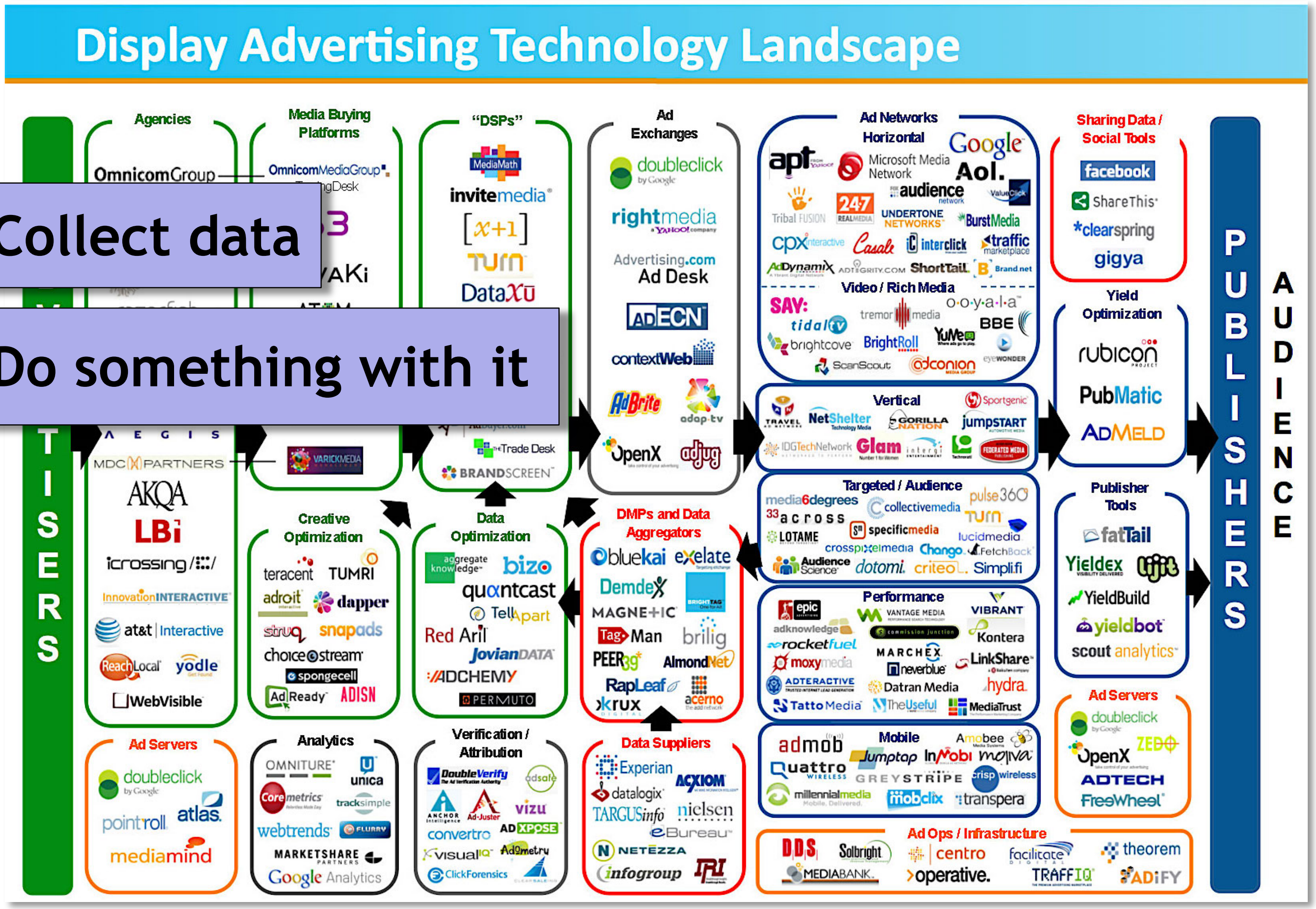
(2) Do something with it

+ music, news, friends, romantic partners, and many more!

Online Advertising

(1) Collect data

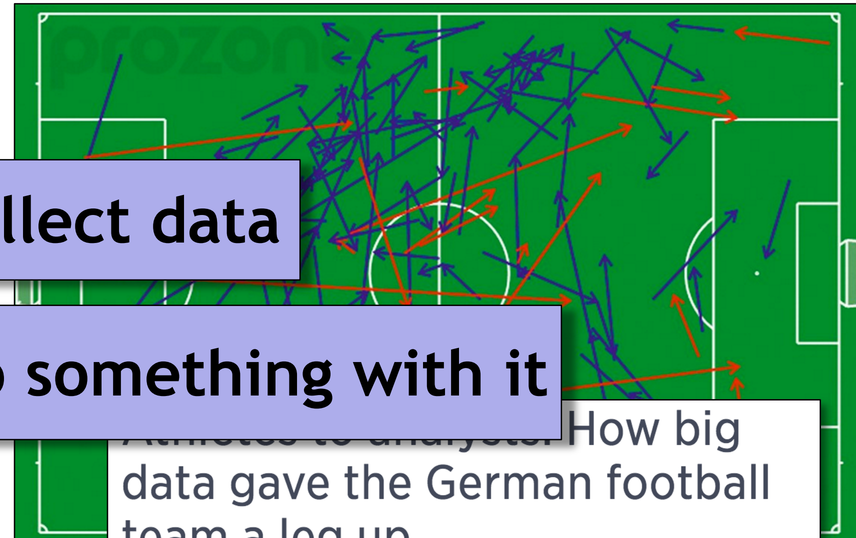
(2) Do something with it



Sports

"Remember, the other team is counting on Big Data insights based on previous games. So, kick the ball with your other foot."

How Big Data is Changing the World of Football



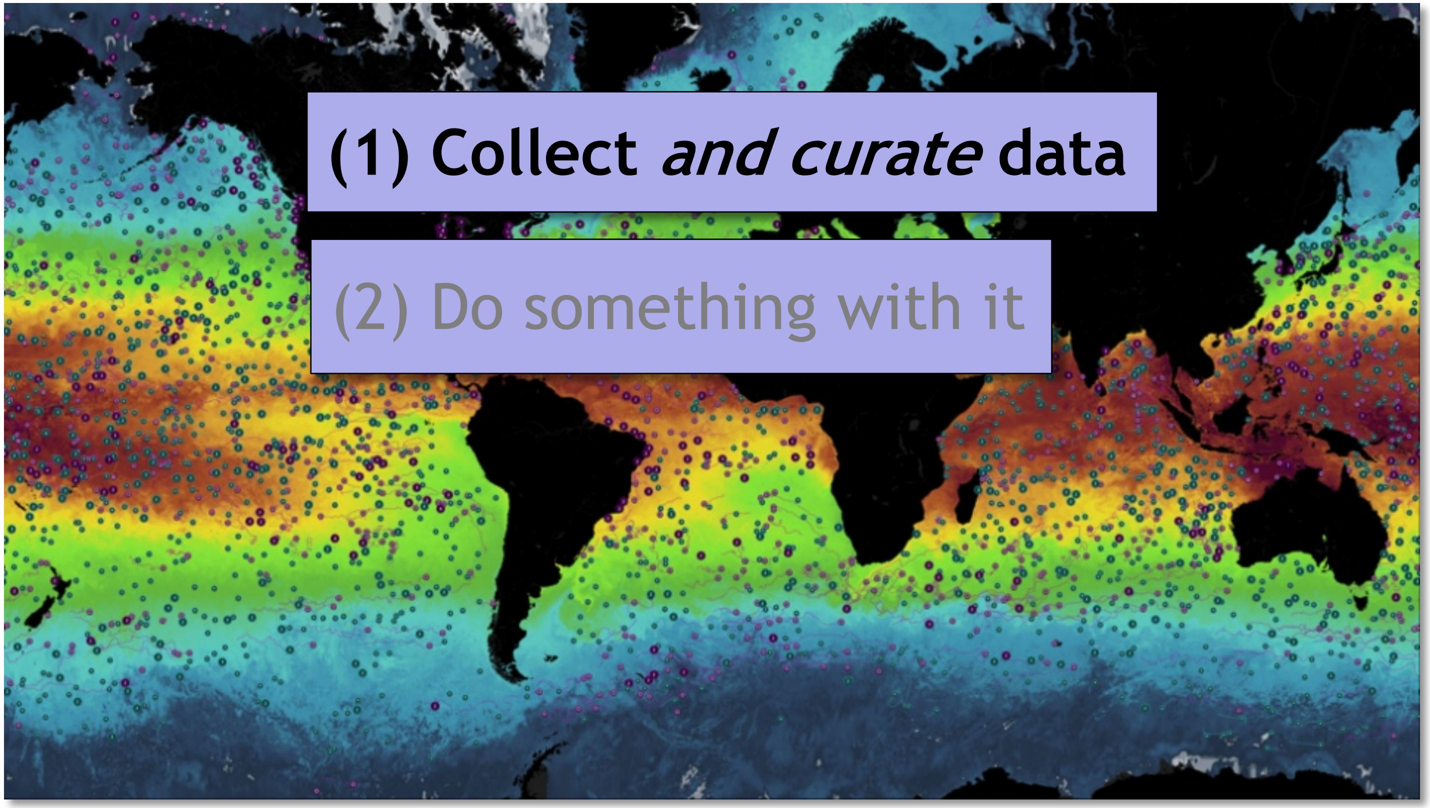
(1) Collect data

(2) Do something with it

How big data gave the German football team a leg up

Saheli Roy Choudhury | @sahelirc
Thursday, 7 Jul 2016 | 12:39 AM ET
CNBC

Ocean Health



44,000 sensors, over 2 billion measurements

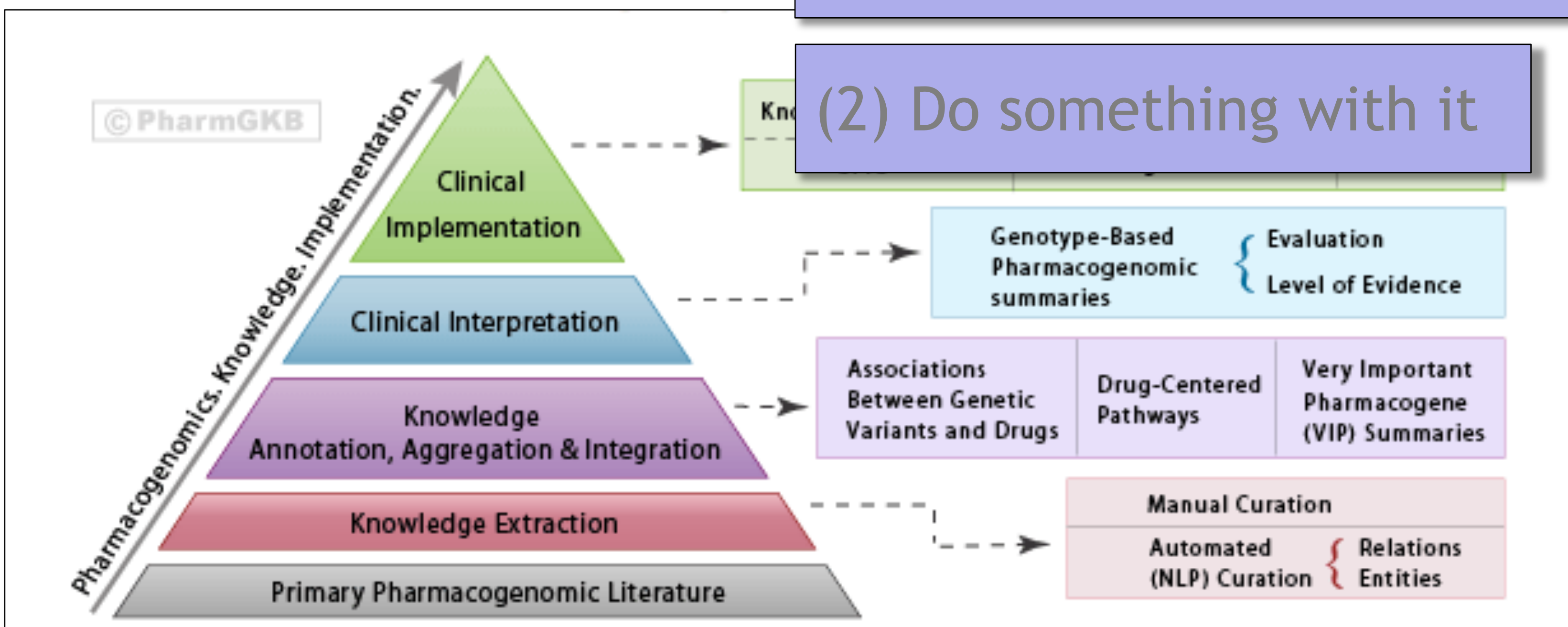
Physical, chemical, biological …

Genetics-Medicine Relationships

PharmGKB collects, curates, and disseminates knowledge about how human genetics affects response to medicines

(1) Collect *and curate* data

(2) Do something with it



And Many More

- Weather prediction
- Medical diagnosis
- Financial markets
- Resource management
- Computational social science
- Smart buildings and cities
-

The list goes on and on,
and it's still early days

Data Tools and Techniques

- Basic Data Manipulation and Analysis
  Performing well-defined computations or asking well-defined questions ("queries")
- Data Mining
  Looking for patterns in data
- Machine Learning
  Using data to build models and make predictions
- Data Visualization
  Graphical depiction of data
- Data Collection and Preparation

Basic Data Manipulation and Analysis

Performing well-defined computations or asking well-defined questions ("queries")

- Average January low temperature for each country over last 20 years
- Number of items over $100 bought by females between ages 20 and 30
- Frequency of specific medicine relieving specific symptoms
- The ten stocks whose price varied the most over the past year

Basic Data Manipulation and Analysis

Performing well-defined computations or asking well-defined questions ("queries")

- Ave... h cou...
- Nu... fe...
- Fre... specific symptoms
- The ten stocks whose price varied the most over the past year

- Spreadsheets
- Relational (SQL) database systems
- "NoSQL" / scalable systems
- Programming languages with data support (e.g., Python, R)

Data Mining

Looking for patterns in data

- Items X,Y,Z are bought together frequently
- People who like movie X also like movie Y
- Patients who respond well to medicines X and Y also respond well to medicine Z
- Students going to the same university are frequently online friends
- Wealthier people are moving from cities to suburbs

Data Mining

Looking for patterns in data

- Items X,Y,Z are bought together frequently
- People ... ovie Y
- Patien ... nes X and Y ... Z
- Studer ... y are frequently online friends
- Wealthier people are moving from cities to suburbs

- Frequent item-sets
- Association rules
- Specialized techniques for graphs, text, multimedia

Machine Learning

Using data to build models and make predictions

- Customers who are women over age 20 are likely to respond to an advertisement
- Students with good grades are predicted to do well on entrance exams
- The temperature of a city can be estimated as the average of its nearby cities, unless some of the cities are on the coast or in the mountains

Machine Learning

Using data to build models and make predictions

- Customers w... over age 20 are likely to res... rtisement
- Students wi... re predicted to do well on ent...

> - Regression
> - Classification
> - Clustering

- Roughly: Basic data analysis and data mining give answers from the available data, while machine learning uses the available data to make predictions about missing or future data

Data Visualization

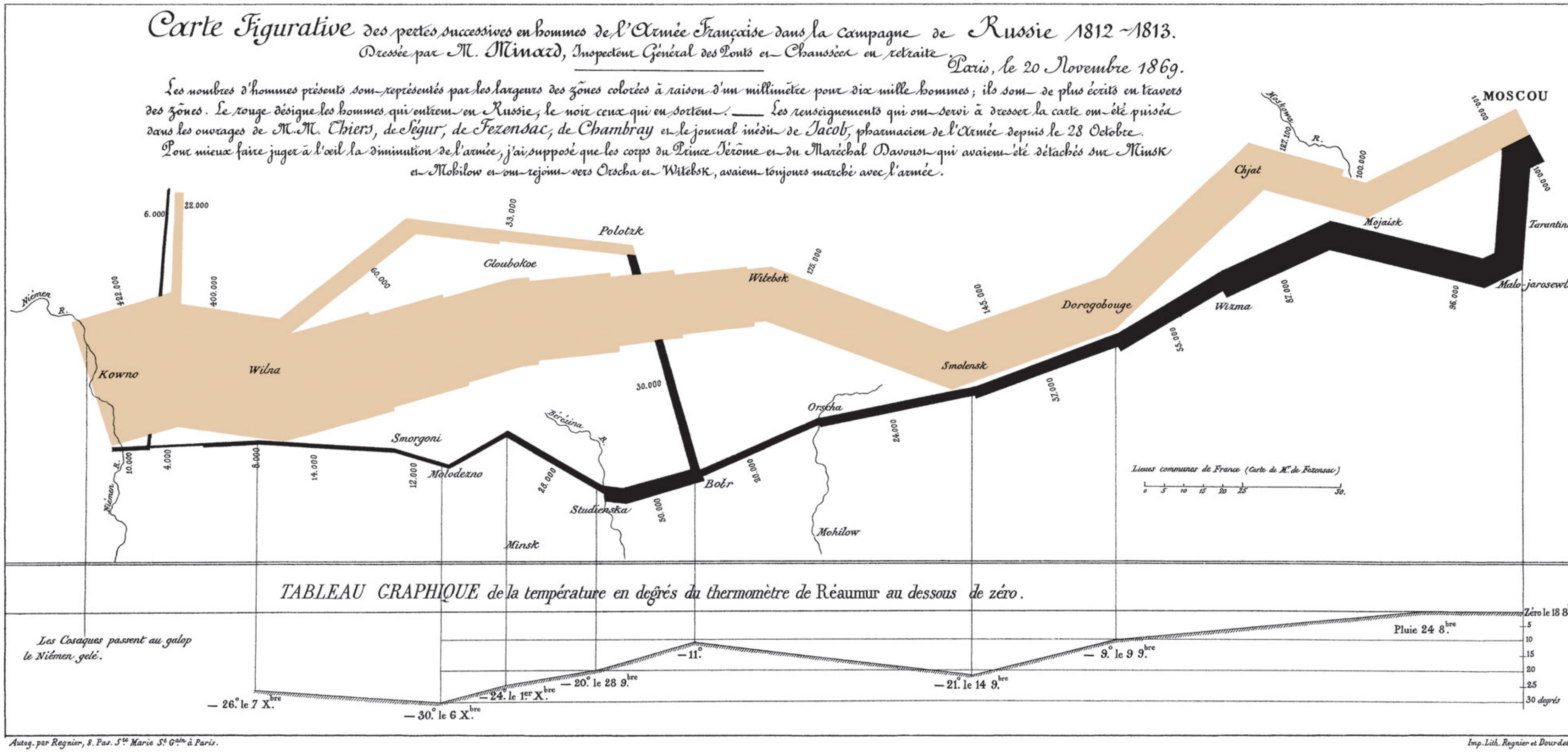
"A picture is worth a thousand words"

Data Visualization

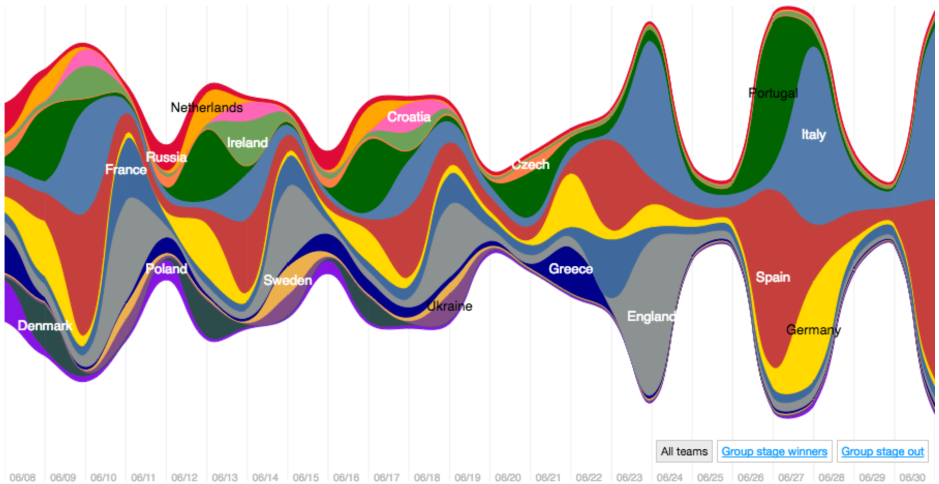
"A picture is worth a ~~thousand words~~"
trillion data points

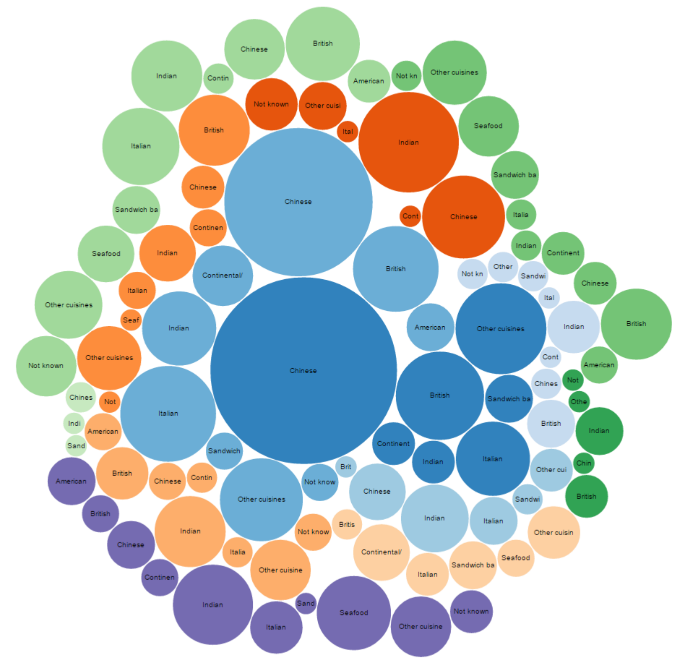
Early Data Visualization

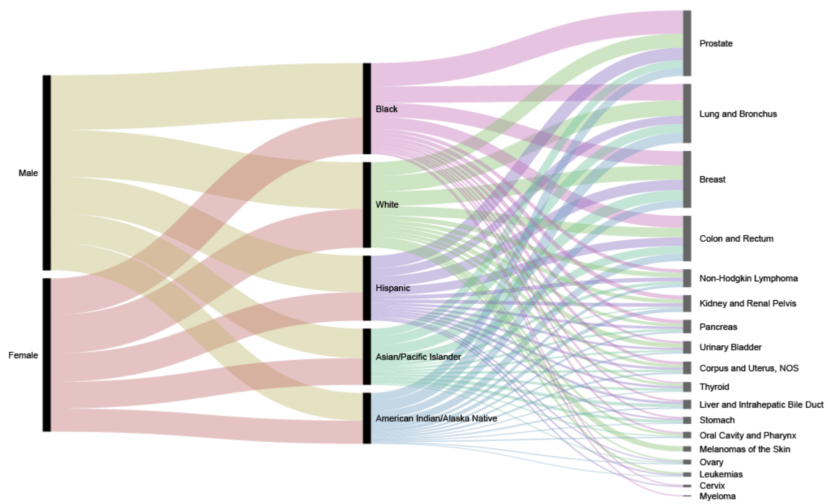
Napoleon's Army



Modern Data Visualization



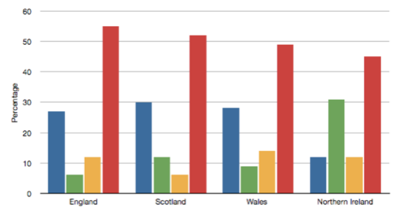


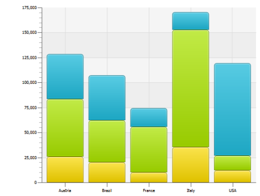


Basic Data Visualization

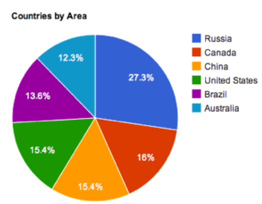
Don’t underestimate the power of basic visualizations

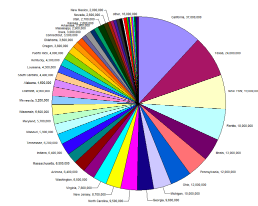
- Bar charts



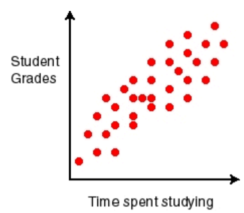


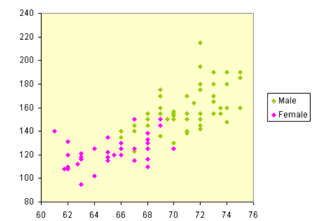
- Pie charts



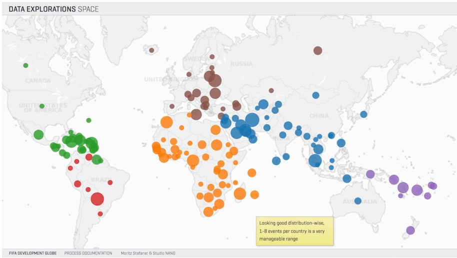


- Scatterplots

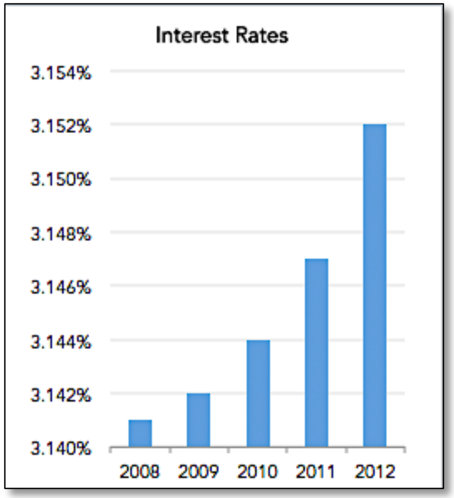


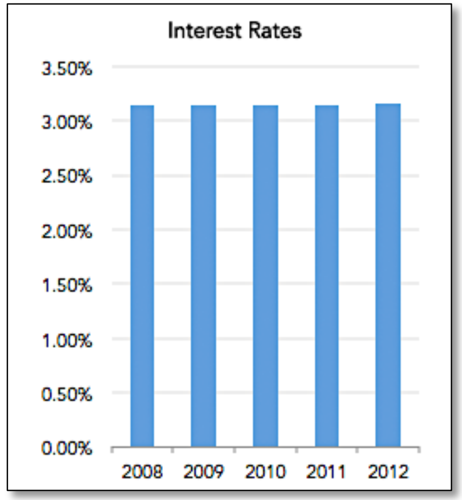


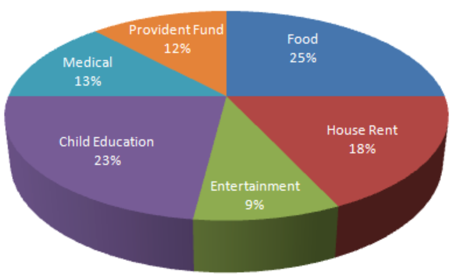
- Maps

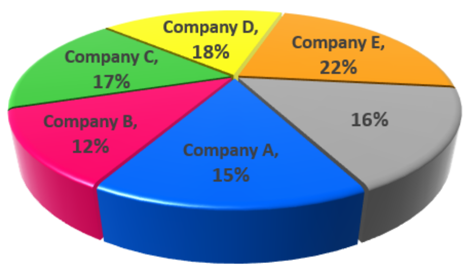


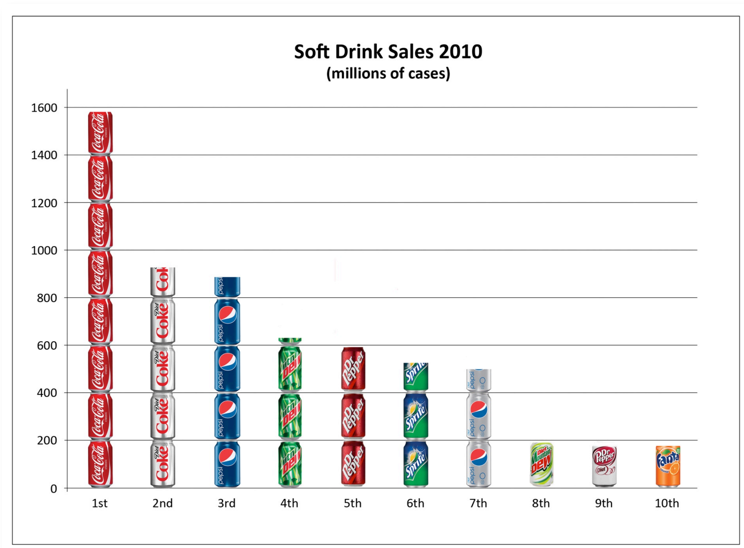
Misleading Data Visualization











Data Collection and Preparation

The "dirty" secret of working with data

- Extracting data from difficult sources
- Filling in missing values
- Removing suspicious data
- Making formats, encoding, and units consistent
- De-duplicating and matching

Data preparation often consumes 80% or more of the effort in a data-driven project

Pitfalls of Data Science

(1) Collect data

(2) Do something with it

correct

Correlation and Causation

Data analysis, data mining, and machine learning can reveal relationships between data values

Correlation – Values track each other

- Height and Shoe Size
- Grades and Entrance Exam Scores

Causation – One value directly influences another

- Education Level → Starting Salary
- Temperature → Cold Drink Sales

Correlation and Causation

"Correlation does not imply causation"

Correlation – Values track each other

- Height and Shoe Size
- Grades and Entrance Exam Scores

Causation – One value directly influences another

- Education Level → Starting Salary
- Temperature → Cold Drink Sales

Correlation and Causation

"Correlation does not imply causation"

- Correlation can be result of causation from a hidden "confounding variable"
- A and B are correlated because there's a hidden C such that C → A and C → B
  - ❖ Homeless population and crime rate
    Confounding variable: unemployment
  - ❖ Forgetfulness and poor eyesight
    Confounding variable: age
  - ❖ Height and shoe size
  - ❖ Grades and entrance exam scores

Correlation and Causation

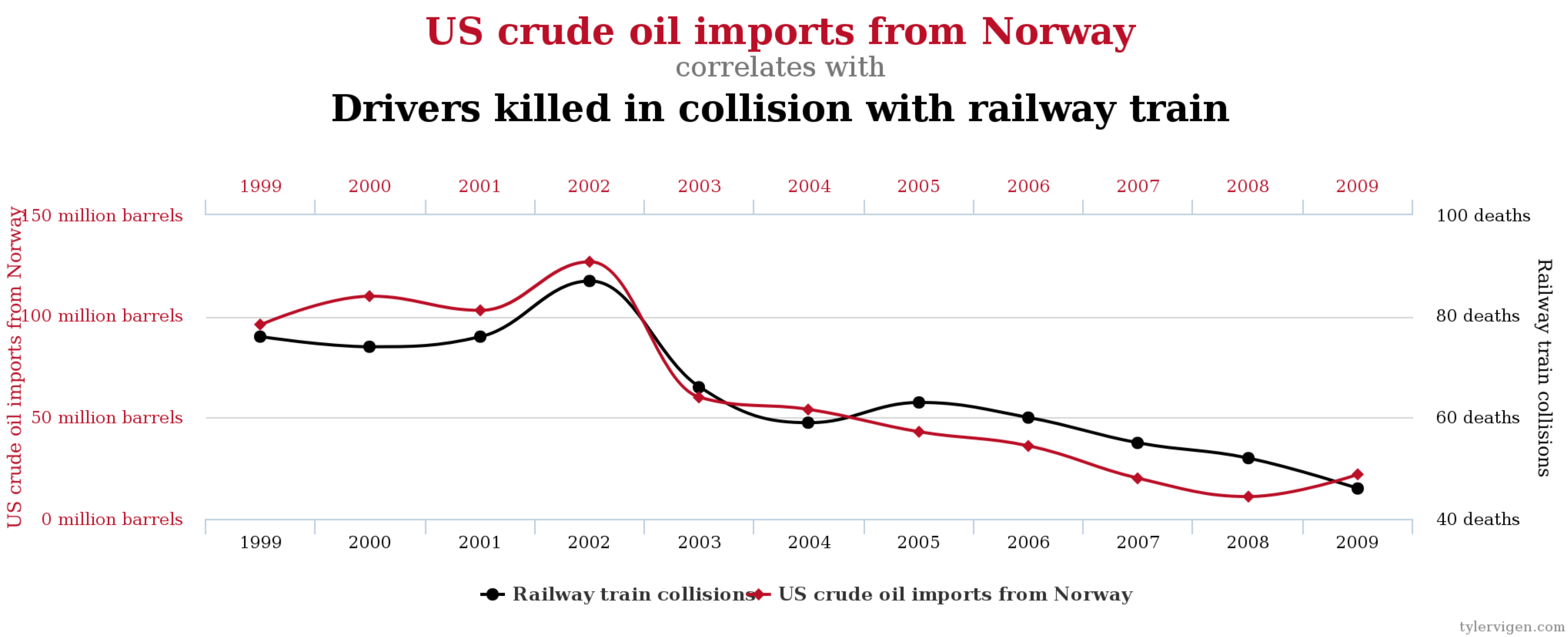
"Correlation does not imply causation"

- Correlation can be result of causation from a hidden "confounding variable"
- A and B are correlated because there's a hidden C such that C → A and C → B

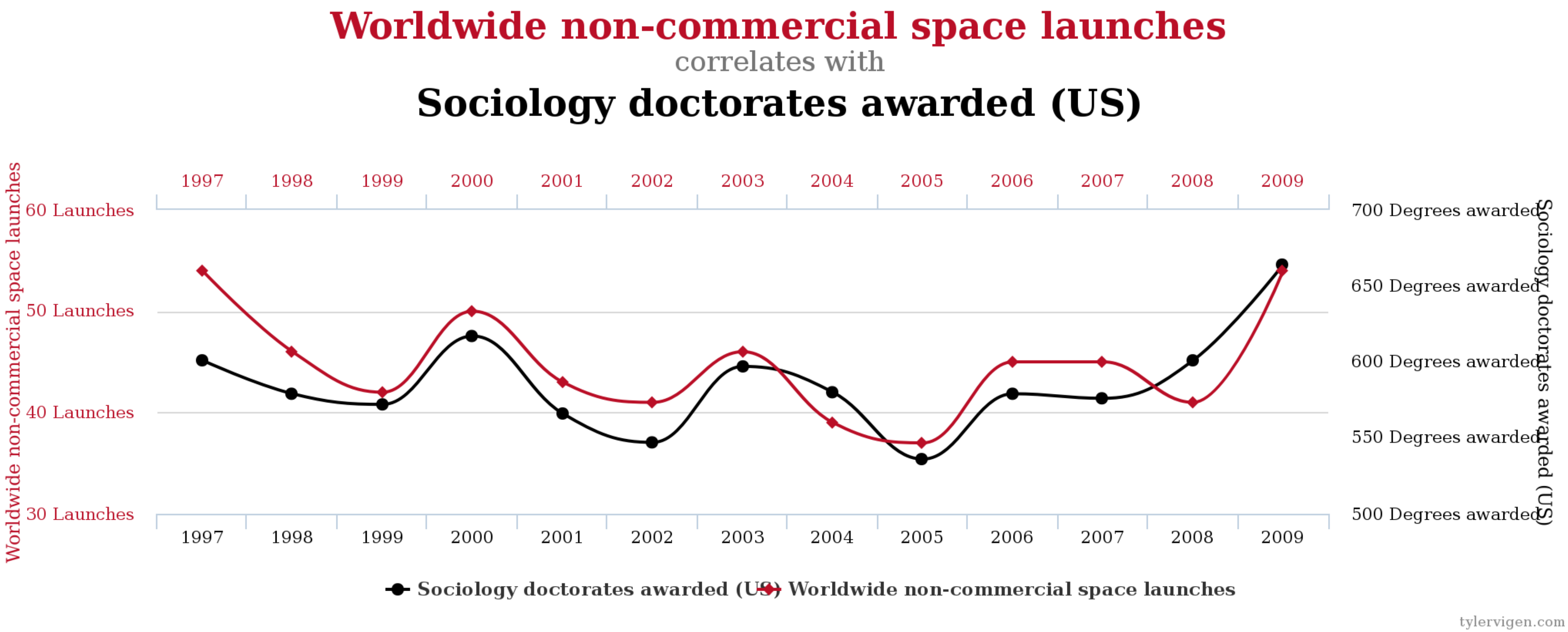
- Correlation is usually "easy" to test
- Causation is typically impossible to test

Correlation and Causation

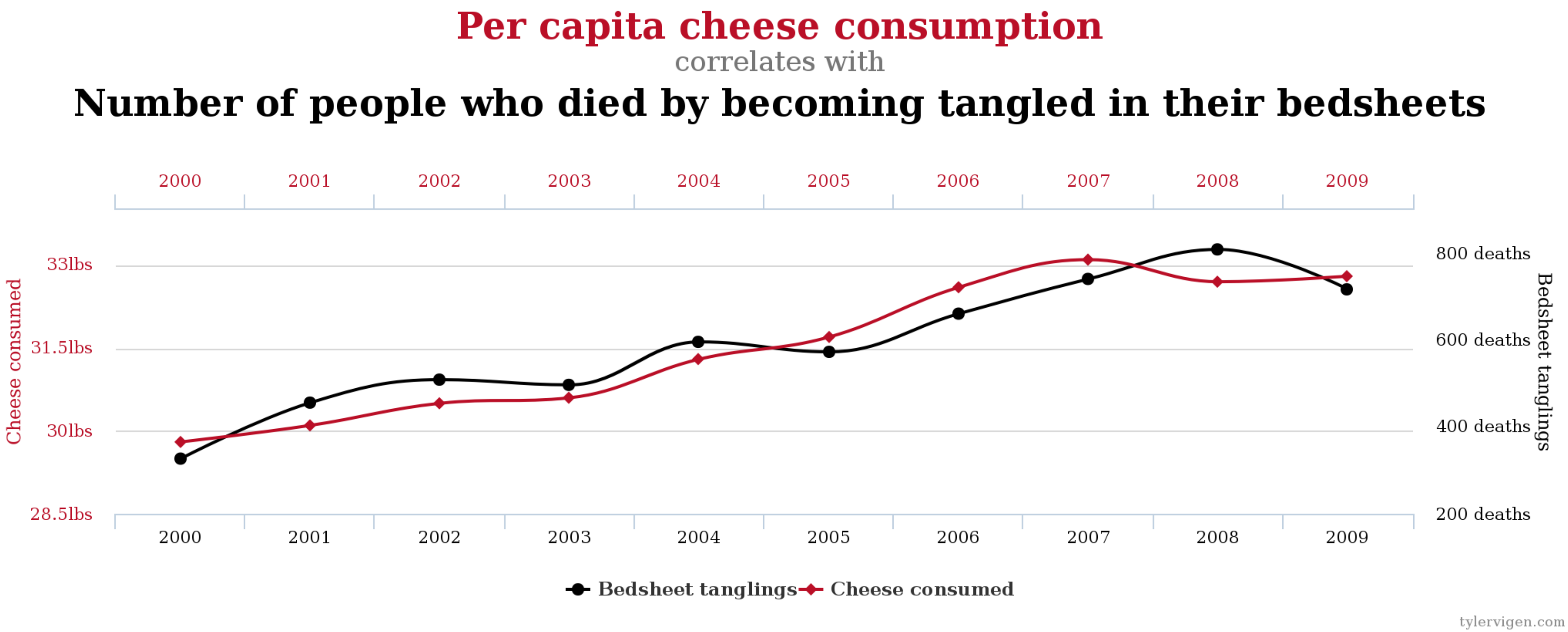
Surprising Correlation #1



Surprising Correlation #2



Surprising Correlation #3



“Spurious Correlations” Website

http://www.tylervigen.com/

Underfitting and Overfitting

Machine learning uses data to create a "model" and uses model to make predictions

- Customers who are women over age 20 are likely to respond to an advertisement
- Students with good grades are predicted to do well on entrance exams
- The temperature of a city can be estimated as the average of its nearby cities, unless some of the cities are on the coast or in the mountains

Underfitting

Model used for predictions is **too simplistic**

- 60% of men and 70% of women responded to an advertisement, therefore all future ads should go to women
- If a furniture item has four legs and a flat top it is a dining room table
- The temperature of a city can be estimated as the average of its nearby cities, unless some of the cities are on the coast or in the mountains

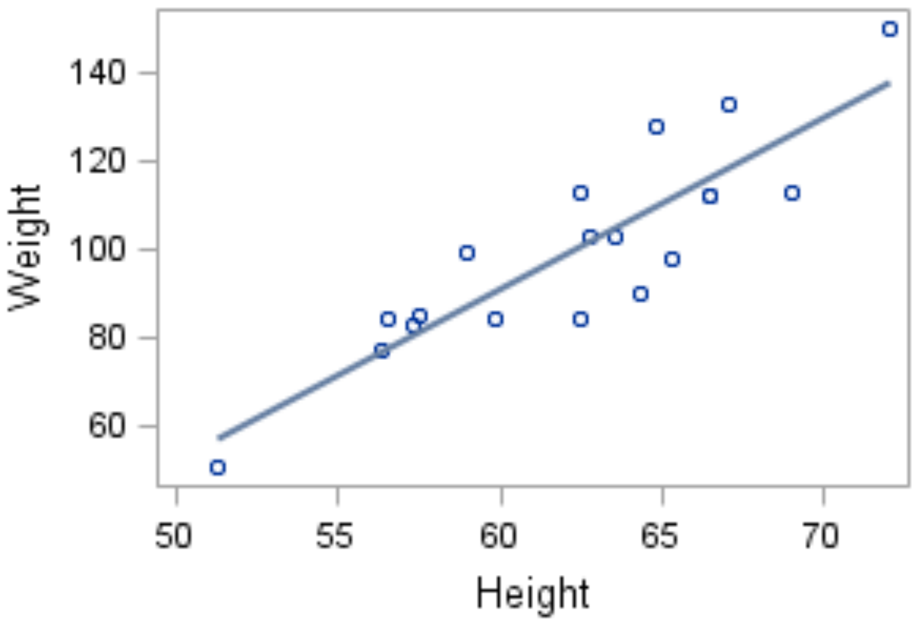
Overfitting

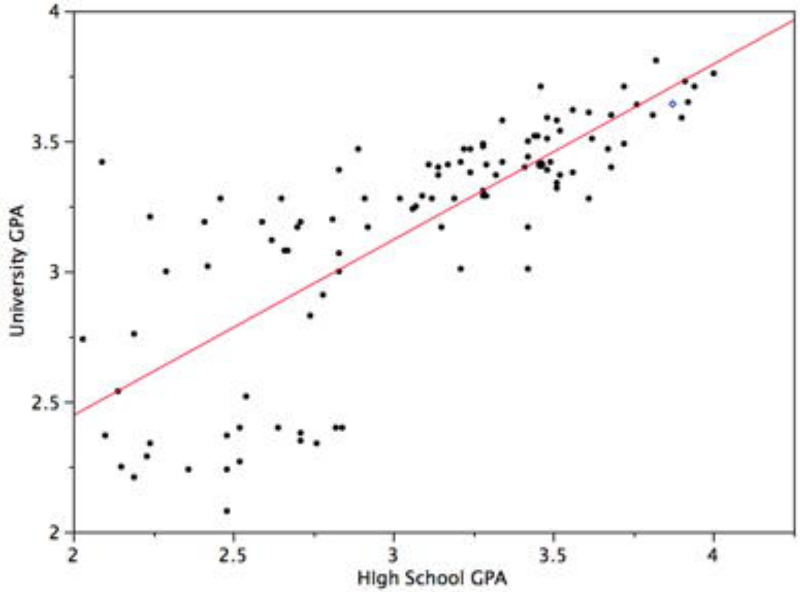
Model used for predictions is **too specific**

- The best targets for an advertisement are married women between 25 and 27 years with short black hair, one child, and one pet dog
- If a furniture item has four 100 cm legs with decoration and a flat polished wooden top with rounded edges then it is a dining room table

Regression

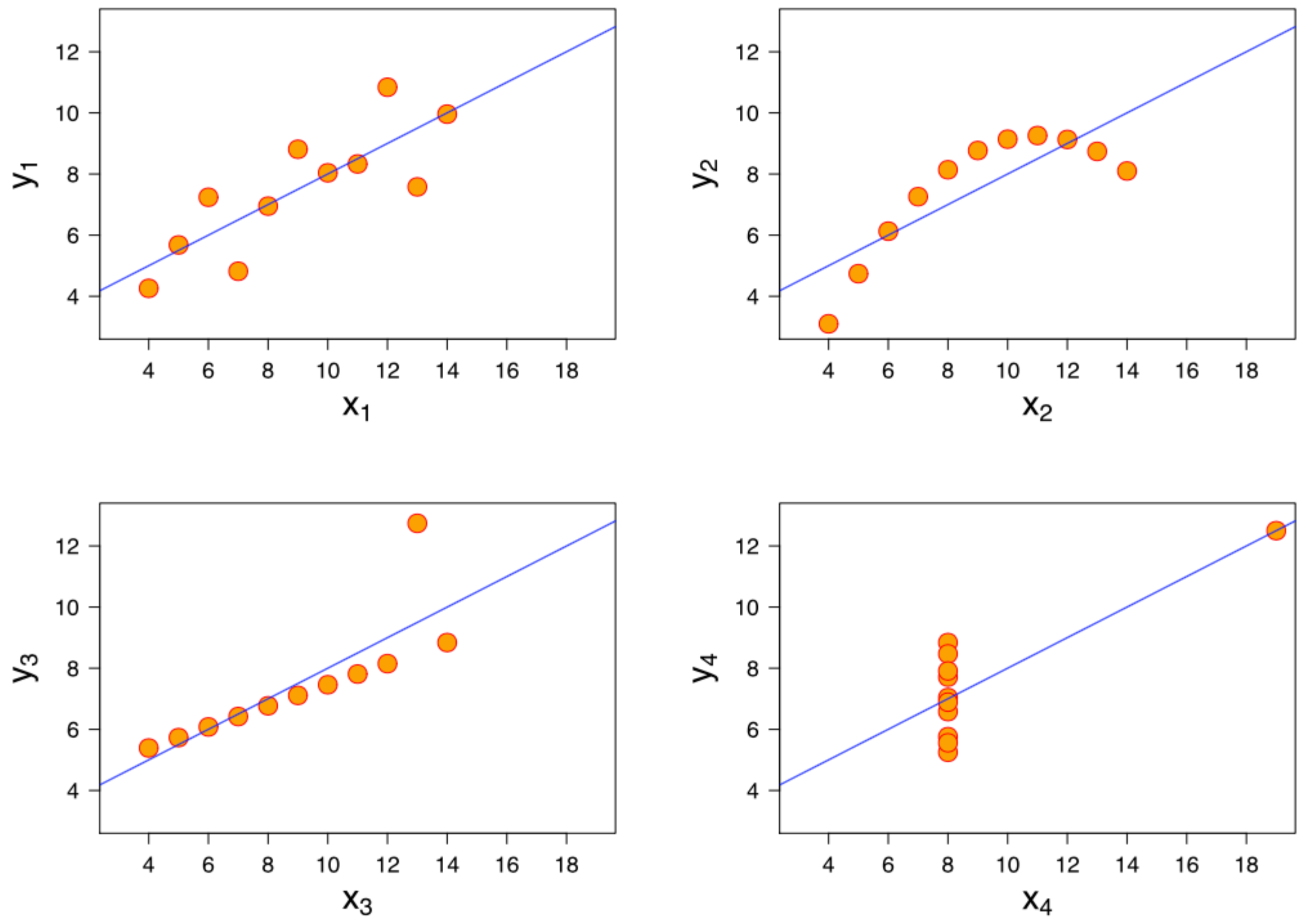
- Fit a line or curve to a set of points (model)
- Use model to predict values for new points





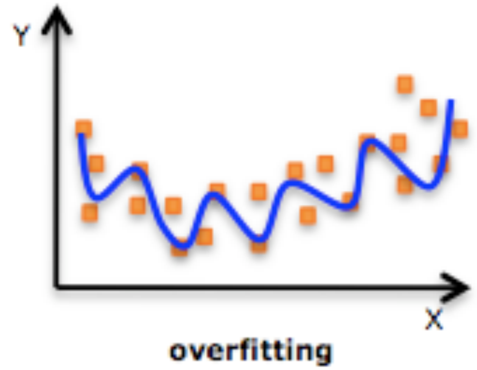
Underfitting

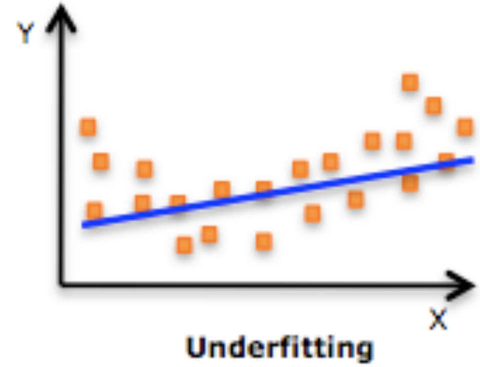
Model is **too simplistic**

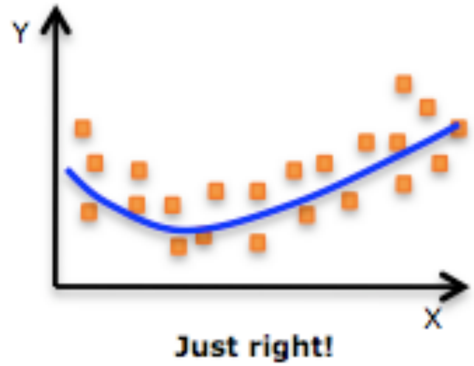


Overfitting

Model is **too specific**







Soccer Match Prediction Scam

- Friday: receive email from "Psychic Sally" predicting which teams will be the winners in the weekend's five soccer matches. She's right about all of them!
- Same thing the following weekend: five games, all winners predicted correctly
- And the following one: five more correct
- Fourth Friday: Sally offers to give you her predictions for the coming weekend's games, for a fee

Should you do it?

Soccer Match Prediction Scam

How many contacts must Sally start with on week one to ensure she has 100 potential buyers by week four, i.e., 100 people who received 15 correct predicted winners?

(Assume no draws)

Data Privacy

Of significant concern in some sectors

- Individual data collected covertly
  - Edward Snowden, “metadata” argument
- Individual data collected legally but used questionably
  - Individual “information trails” are enormous
  - Target stores pregnancy mailing
- Individual data deduced from “anonymous” public data
  - Governor of Massachusetts health record

Languages, Systems, Platforms

- Spreadsheets

  Surprisingly versatile and powerful for data analysis tasks, provided data is not *too* large

- Programming languages with data support
  - R Language – powerful statistical features
  - Python – general-purpose language with R-like add-ons (Pandas, SciPy, scikit-learn)

Languages, Systems, Platforms

- Relational Database Management Systems
  - Also called RDBMS, SQL Systems
  - Long-standing solution for reliability, efficiency, powerful query processing
  - Works for all but truly extreme data sizes, or highly unstructured data
- "NoSQL" Systems
  - Distributed/scalable processing
  - Some specifically target unstructured data (documents, graphs)

Languages, Systems, Platforms

- Specialized languages on scalable systems
  - MapReduce / Hadoop
  - Spark generalized data flow
- Systems for data preparation
- Systems for data visualization

Languages, Systems, Platforms

- Data processing in the cloud
  - Amazon Web Services, Google Cloud, Microsoft Azure
  - Data storage
  - Data processing: SQL, Hadoop, Spark
  - Machine learning libraries
  - Integration with visualization systems

How Much Data is There?

Complete works of William Shakespeare
5 megabytes

Average individual
50 gigabytes (10,000 Shakespeares)

USA Library of Congress
10 terabytes (2 million Shakespeares)

Uploaded to Facebook daily
1 petabyte (200 million Shakespeares)

Produced by humanity daily
2.5 exabytes (500 trillion Shakespeares)

"Big Data"

Some domains produce vast quantities of data, and some analyses require "big data" to be effective

- Most tools and techniques apply to data of all sizes
- Big insights can come from small/medium data

> Sometimes twenty Spark servers in the cloud are required.
> More often a laptop with SQL, Python, or simple spreadsheets does the job.

Rest of the Short-Course

- Data Analysis Using Spreadsheets
- Data Visualization Using Spreadsheets
- Advanced Data Visualization Using Tableau
- Relational Databases and Basic SQL
- Using Python for Data Analysis

If extra time (last day) decide among…

- Machine Learning – Regression and Classification
- Network Analysis
- Advanced SQL
- The R Language

Getting Started with Google Sheets

professorwidom.org > Courses > Course Materials

Step 1: Create a Google account: accounts.google.com

Step 2: Copy data files

a) Follow link for module in "Getting Started" document
b) Select all files (control/command-a), control/right-click, "Make a copy"
c) Go to "My Drive", rename/reorganize files if desired

Step 3: Open files using Google Sheets: double-click

Step 4 (optional): Enable working offline; see "Getting Started" document for details

Getting Started with Tableau Public

professorwidom.org > Courses > Course Materials

Step 1: Install Tableau Public: public.tableau.com

Step 2: Copy data files

a) Follow link in "Getting Started" document
b) Download each file: double-click to open, select File > Download as > Comma-separated values
c) Rename/relocate files if desired

Step 3: Open Tableau Public, click "Text File", select downloaded file. (If the data loads but looks strange, try clicking "Excel" instead of "Text File".)

Getting Started with Instabase

professorwidom.org > Courses > Course Materials

Step 1: Get an Instabase account: follow link in "Getting Started" document, use provided token

Step 2: Manage workspaces and repositories

Step 3: Copy notebooks and data files

a) Follow link for module in "Getting Started" document
b) Select all files, Actions menu > Copy
c) Select repository, click Instabase Drive, Copy

Step 4: Open notebook: right-click, Open With > Jupyter

Overview of Data Science

Promises and Pitfalls, Tools and Techniques

Questions?

Stanford University

Google

Instabase

Google Cloud Platform

Association for Computing Machinery

Very Large Data Bases Endowment Inc.

amazon web services™