Overview of Big Data
Tools and Techniques, Discoveries and Pitfalls

CS102
Fall 2017
What Does “Big Data” Mean?

(1) Collecting large amounts of data
   Via computers, sensors, people, events ...

(2) Doing something with it
   Making decisions, confirming hypotheses,
   gaining insights, predicting future ...

“Data Science” = Going from (1) to (2)
Big Data is Here to Stay

- Ability to collect data will only increase
- Ability to analyze data will only improve

What would go away?
This Overview

- Promises of Big Data
  Applications and services

- Big Data tools and techniques
  Database management systems
  Data mining and machine learning

- Pitfalls of Big Data
  Correlation and causation
  Underfitting and overfitting
  Privacy and a few others

- Big Data systems and platforms
Promises of Big Data

(1) Collect large amounts of 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

Display Advertising Technology Landscape

1) Collect data
2) Do something with it
Sports

(1) Collect data

(2) Do something with it

“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

How big data gave the German football team a leg up

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

CNBC
Ocean Health

44,000 sensors, over 2 billion measurements
Physical, chemical, biological ...

(1) Collect *and curate* data

(2) Do something with it
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
Big 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 make inferences or 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")

- 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

- Spreadsheets
- Relational (SQL) database systems
- "NoSQL" / scalable systems
- Programming languages with big-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 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

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

Using data to make inferences or 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 the SAT
- 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 make inferences or 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 the SAT.
- 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.

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.

- Regression
- Classification
- Clustering
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

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812 - 1813.

Les nombres d'hommes présent sont représentés par les longueurs des gouttes alvéolées à raison d'une millième pour dix mille hommes, il y en a plus dix fois en hauteur des gouttes. Le rouge désigne les hommes qui restent en Russie, le noir ceux qui en rentrent. Les renforts qu'on a eus à recevoir de la carte ont été pris dans les travaux de M. Choder, de Pajot, de Troisièges de Chambray et le journal médical de Jacob, phototypie de l'Armée depuis le 25 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai ajouté que les corps du Prince Jérôme et du Marshal Davout, qui avaient été détachés pour Moskow et Mobilow en une région près Ostrok et Wilskop, avaient tousjours marche avec l'armée.

TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Les Courbes passent au goutte le Xiléum, gelé.
Basic Data Visualization

Don’t underestimate the power of basic visualizations

- Bar charts
- Pie charts
- Scatterplots
- Maps
Misleading Data Visualization

Relative Soft Drink Sales, 2010
(Sales proportional to can height)

Soft Drink Sales 2010
(millions of cases)
Data Collection and Preparation

The “dirty” secret of Big 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 Big Data project
Pitfalls of Big Data

(1) Collect large amounts of data

(2) Do something with it
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 SAT 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 SAT 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 \(\rightarrow\) A and C \(\rightarrow\) B
  - Homeless population and crime rate
    Confounding variable: unemployment
  - Forgetfulness and poor eyesight
    Confounding variable: age
  - Height and shoe size
  - Grades and SAT 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 $\rightarrow$ A and C $\rightarrow$ B

- Correlation is usually “easy” to test
- Causation is impossible to test
"Do you think all these film crews brought on global warming or did global warming bring on all these film crews?"
Correlation and Causation

“Do you think all these film crews brought on global warming or did warming bring on all these film crews?"

“I wish they didn’t turn on that seatbelt sign so much! Every time they do, it gets bumpy.”
Correlation and Causation

"Do you think all these film credits brought on global warming or did global warming bring on all these film credits?"

"I wish they didn’t turn on the seatbelt sign so much! Time they do, it gets boring.

"Excellent health statistics - smokers are less likely to die of age related illnesses."

Tobacco Industry Research Centre
Surprising Correlation #1

US crude oil imports from Norway correlates with Drivers killed in collision with railway train

- Railway train collisions
- US crude oil imports from Norway
Surprising Correlation #2

Worldwide non-commercial space launches correlates with Sociology doctorates awarded (US)

Sociology doctorates awarded (US) Worldwide non-commercial space launches

tylervigen.com
Surprising Correlation #3

Per capita cheese consumption correlates with Number of people who died by becoming tangled in their bedsheets

![Graph showing the correlation between per capita cheese consumption and deaths due to bedsheets tangling from 2000 to 2009. The graph indicates an increasing trend in both categories over the years.](https://tylervigen.com/graphs/bedsheet-tanglings)

Big Data Overview

CS102
“Spurious Correlations” Website

http://www.tylervigen.com/
Underfitting and Overfitting

Machine learning uses data to create a “model” and uses model to make inferences or 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 the SAT
- 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

- \( y_1 \) vs. \( x_1 \)
- \( y_2 \) vs. \( x_2 \)
- \( y_3 \) vs. \( x_3 \)
- \( y_4 \) vs. \( x_4 \)
Overfitting

Model is too specific
Big Data Pitfall: Google Translate

Automatically learns language translation from examples on the web

Anyone see a problem with this approach?
Big Data Scam: Soccer Match Prediction

- 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?
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

- Individual data collected covertly
  - Edward Snowden, “metadata” argument

- Individual data collected legally but used questionably
  - Individual “digital footprints” are enormous
  - Target stores pregnancy mailing
  - Engagement ring purchase broadcast on Facebook

- Individual data deduced from “anonymous” public data
  - Boston mayor’s health record
Languages, Systems, Platforms

- **Spreadsheets**
  Surprisingly versatile and powerful for data analysis tasks, but not truly big data

- **Programming languages with big-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, unstructured data
  - Key-value row stores (e.g., Cassandra, Dynamo)
  - Document databases (e.g., MongoDB, CouchDB)
  - Graph databases (e.g., Neo4J, Giraph)
Languages, Systems, Platforms

- Specialized languages on scalable systems
  - MapReduce / Hadoop
  - Spark generalized data flow

- Systems for data preparation

- Systems for data visualization
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
Big Data: How Big is Big?

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)
Size Isn’t Everything

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

Some applications actually do need twenty Spark servers in the cloud. More often a laptop with SQL, Python, or simple spreadsheets does the job.
What We’ll Cover

- Data Analysis & Visualization Using Spreadsheets
- Advanced Data Visualization Using Tableau
- Relational Databases and Basic SQL
- Python for Data Analysis & Visualization
- Machine Learning - Regression, Classification, Clustering
- Using Python for Machine Learning
- Advanced SQL
- Data Mining Algorithms
- Data Mining Using SQL and Python
- The R Language - Data Analysis, Visualization, and Machine Learning
- Social-Network Analysis

+ Guest speakers
+ Additional topics?
Overview of Big Data
Tools and Techniques, Discoveries and Pitfalls

Questions?