Algorithms for Data Science
Course Unit 1 – Introduction

Slides based on those in: http://www.mmds.org

Spyros Kontogiannis & Christos Zaroliagis
$600 to buy a disk drive that can store all of the world’s music

5 billion mobile phones in use in 2010

30 billion pieces of content shared on Facebook every month

40% projected growth in global data generated per year vs.

5% growth in global IT spending

$5 million vs. $400 Price of the fastest supercomputer in 1975 and an iPhone 4 with equal performance

235 terabytes data collected by the US Library of Congress by April 2011

15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress
Data contains value and knowledge
What is Data Science?
Knowledge discovery from data

Interdisciplinary field

- Methods, processes, systems for extracting knowledge from data, similar to data mining
- Concept unifying statistics & data analysis, in order to understand & analyze certain phenomena with data
- Fourth paradigm of science (empirical, theoretical, computational, data-driven)
To extract the knowledge data needs to be

- Stored
- Managed
- And ANALYZED ← this class

Data Mining ≈ Big Data ≈ Predictive Analytics ≈ Data Science
Good news: Demand for Data Analysis

Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018

Supply and demand of deep analytical talent by 2018

<table>
<thead>
<tr>
<th>Year</th>
<th>Employment</th>
<th>Graduates with deep analytical talent</th>
<th>Others¹</th>
<th>2018 Supply</th>
<th>Talent Gap</th>
<th>2018 Projected Demand</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>150</td>
<td>180</td>
<td>30</td>
<td>300</td>
<td>140-190</td>
<td>440-490</td>
</tr>
</tbody>
</table>

¹ Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).

SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey Global Institute analysis
What is Data Analysis?

- Given lots of data
- Discover patterns and models that are
  - **Valid**: hold on new data with some certainty
  - **Useful**: should be possible to act on the item
  - **Unexpected**: non-obvious to the system
  - **Understandable**: humans should be able to interpret the pattern
Data Science Tasks

- **Descriptive methods**
  - Find human-interpretable patterns that describe the data
    - **Example:** Clustering

- **Predictive methods**
  - Use some variables to predict unknown or future values of other variables
    - **Example:** Recommender systems
Meaningfulness of Analytic Answers

- A risk with “Data mining” is that an analyst can “discover” patterns that are meaningless
- Statisticians call it **Bonferroni’s principle**
  - If expected number of patterns in random data $>>$ number of real instances, then almost anything found is bogus
  - Roughly, if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap
Example:

- We want to find (unrelated) people who **at least twice have stayed at the same hotel on the same day**
  - 10⁹ people being tracked
  - 1,000 days
  - Each person stays in a hotel 1% of time (1 day out of 100)
  - Hotels hold 100 people (so 10⁵ hotels)
  - **If everyone behaves randomly (i.e., no terrorists) will the data mining detect anything suspicious?**

- Prob(person P1 visits a hotel on a given day) = 1/100 = 0.01
- Prob(persons P1 & P2 visit a hotel on a given day) = (0.01)² = 0.0001 = 10⁻⁴
- Prob(P1 & P2 visit the same hotel on a given day) = 0.0001 / 10⁵ = 10⁻⁹
- Prob(P1 & P2 visit the same hotel on two given days) = (10⁻⁹)² = 10⁻¹₈
- Number of people pairs ≈ (10⁹)² /2 = 5 · 10¹⁷
- Number of pairs of days ≈ (10³)² /2 = 5 · 10⁵
- **Exp(# “suspicious” people pairs) = 5 · 10¹⁷ · 5 · 10⁵ · 10⁻¹₈ = 250000**
Expected number of “suspicious” pairs of people:

- 250,000
- … too many combinations to check – we need to have some additional evidence to find “suspicious” pairs of people in some more efficient way
What matters when dealing with data?

Challenges
- Usage
- Quality
- Context
- Streaming
- Scalability

Data Modalities
- Ontologies
- Structured
- Networks
- Text
- Multimedia
- Signals

Data Operators
- Collect
- Prepare
- Represent
- Model
- Reason
- Visualize
Data Science – Cultures

- **Data Science overlaps with**
  - **Databases**: Large-scale data, simple queries
  - **Machine learning**: Small data, Complex models
  - **CS Theory**: (Randomized) Algorithms

- **Different cultures**
  - To a DB person, data science is an extreme form of **analytic processing** – queries that examine large amounts of data
    - Result is the query answer
  - To a ML person, data science is the **inference of models**
    - Result is the parameters of the model

- **In this course we will do both!**
This Course

- Overlaps with machine learning, statistics, artificial intelligence, databases but more stress on
  - **Scalability** (big data)
  - **Algorithms**
  - **Computing architectures**
  - Automation for handling large data
What will we learn?

- We will learn to **mine different types of data**
  - Data is high dimensional
  - Data is a graph
  - Data is infinite/never-ending
  - Data is labeled
- We will learn to **use different models of computation**
  - MapReduce
  - Streams and online algorithms
  - Single machine in-memory
We will learn to **solve real-world problems**
- Recommender systems
- Market Basket Analysis
- Spam detection
- Duplicate document detection

We will learn **various “tools”**
- Linear algebra (SVD, Rec. Sys., Communities)
- Optimization (stochastic gradient descent)
- Dynamic programming (frequent itemsets)
- Hashing (LSH, Bloom filters)
# How It All Fits Together

<table>
<thead>
<tr>
<th>High dim. data</th>
<th>Graph data</th>
<th>Infinite data</th>
<th>Machine learning</th>
<th>Apps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locality sensitive hashing</td>
<td>PageRank, SimRank</td>
<td>Filtering data streams</td>
<td>SVM</td>
<td>Recommender systems</td>
</tr>
<tr>
<td>Clustering</td>
<td>Community Detection</td>
<td>Web advertising</td>
<td>Decision Trees</td>
<td>Association Rules</td>
</tr>
<tr>
<td>Dimensionality reduction</td>
<td>Spam Detection</td>
<td>Queries on streams</td>
<td>Perceptron, kNN</td>
<td>Duplicate document detection</td>
</tr>
</tbody>
</table>

- **Apps**
  - Recommender systems
  - Association Rules
  - Duplicate document detection

- **Machine learning**
  - SVM
  - Decision Trees
  - Perceptron, kNN

- **Infinite data**
  - Filtering data streams
  - Web advertising
  - Queries on streams

- **Graph data**
  - PageRank, SimRank
  - Community Detection
  - Spam Detection

- **High dim. data**
  - Locality sensitive hashing
  - Clustering
  - Dimensionality reduction
How do you want that data?
About the Course
Course Topics

- Advanced Frameworks for Large Data Sets (MapReduce)
- Complexity of Algorithms for Large Data Sets
- Finding Similar Items
- Locality-Sensitive Hashing
- Algorithms for Large Data Streams
- Algorithms for link analysis of large graphs (PageRank)
- Exploration of Frequent Itemsets
- Clustering
- Introduction to Social Networks
- Efficient Algorithms on Large Graphs
- Computational Advertising
- Recommendation Systems
- Dimensionality Reduction
- Large-Scale Machine Learning
Course Logistics

- Course website
  - Lecture slides
  - Bibliography

- Course Book: Mining of Massive Datasets (v3.0)
  by J. Leskovec, A. Rajaraman, and J. Ullman
  Free online: http://www.mmds.org

- Final mark: 50% presentation + 50% final exam