Data Mining

Yannis Velegrakis

Administrative Issues

- Schedule Wed & Thu @ 17:30
  - Lectures (in the class) – A212
  - Not online.

- Evaluation: Project
  - Topic: Will be announced later (not a free choice)
  - Development of a system and analysis of some dataset

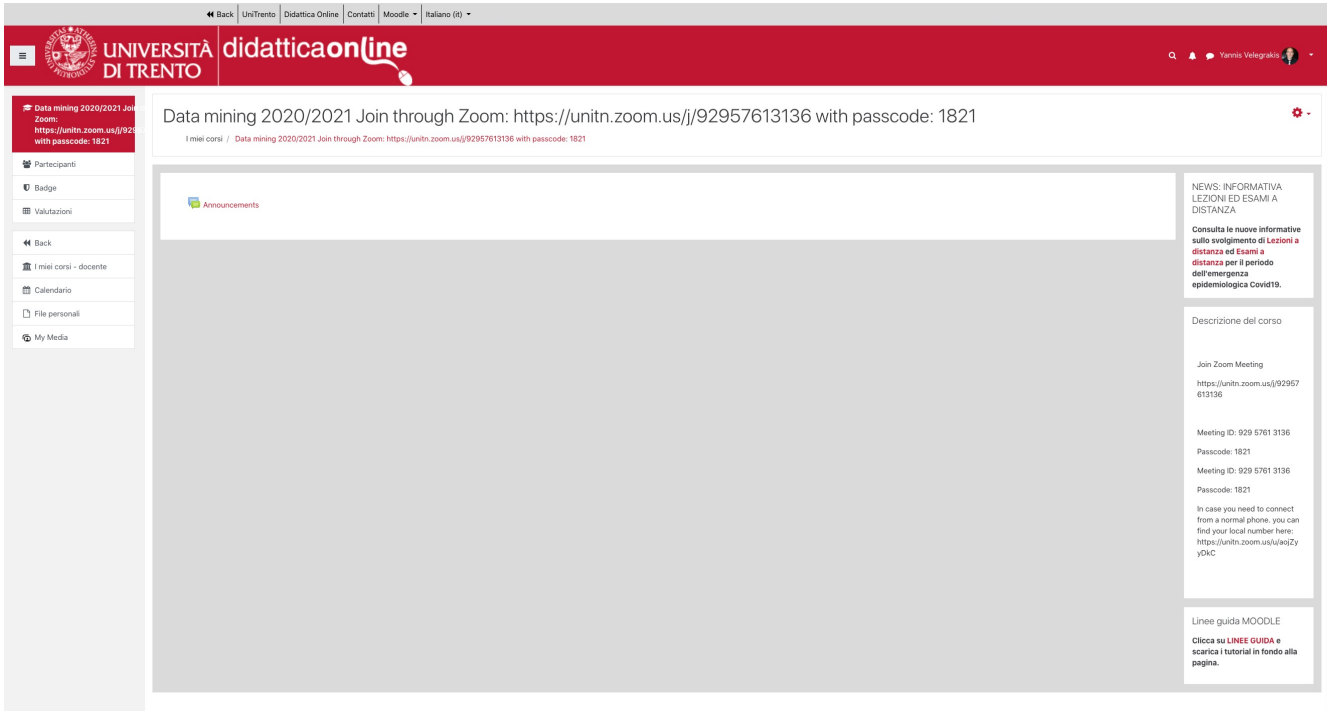
Course Syllabus

- https://www.esse3.unitn.it/Guide/PaginaRicercaInse.do
- Enter in the prof field: “Velegrakis”
- Select “Data Mining”

- Also https://www.disi.unitn.it/it/bachelor-student

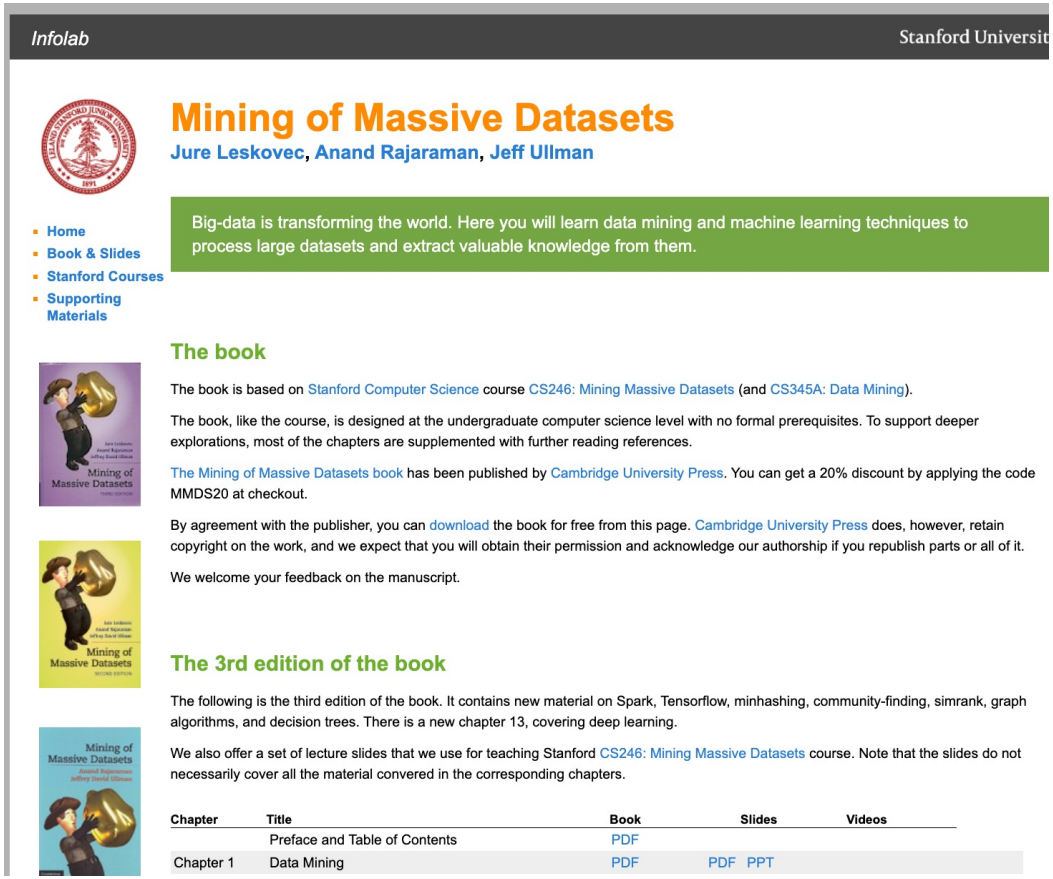
Administrative Issues

- Material Online (didattica Online)
  - https://didatticaonline.unitn.it/dol/course/view.php?id=34838



Administrative Issues

- Textbook
  - Mining Massive Datasets, Ullman et. al.
  - Available Online at http://www.mmds.org/



Infolab Stanford Universit

Mining of Massive Datasets

Jure Leskovec, Anand Rajaraman, Jeff Ullman

- Home
- Book & Slides
- Stanford Courses
- Supporting Materials

Big-data is transforming the world. Here you will learn data mining and machine learning techniques to process large datasets and extract valuable knowledge from them.

The book

The book is based on Stanford Computer Science course CS246: Mining Massive Datasets (and CS345A: Data Mining).

The book, like the course, is designed at the undergraduate computer science level with no formal prerequisites. To support deeper explorations, most of the chapters are supplemented with further reading references.

The Mining of Massive Datasets book has been published by Cambridge University Press. You can get a 20% discount by applying the code MMDS20 at checkout.

By agreement with the publisher, you can download the book for free from this page. Cambridge University Press does, however, retain copyright on the work, and we expect that you will obtain their permission and acknowledge our authorship if you republish parts or all of it.

We welcome your feedback on the manuscript.

The 3rd edition of the book

The following is the third edition of the book. It contains new material on Spark, Tensorflow, minhashing, community-finding, simrank, graph algorithms, and decision trees. There is a new chapter 13, covering deep learning.

We also offer a set of lecture slides that we use for teaching Stanford CS246: Mining Massive Datasets course. Note that the slides do not necessarily cover all the material convered in the corresponding chapters.

Chapter	Title	Book	Slides	Videos
	Preface and Table of Contents	PDF		
Chapter 1	Data Mining	PDF	PDF PPT	

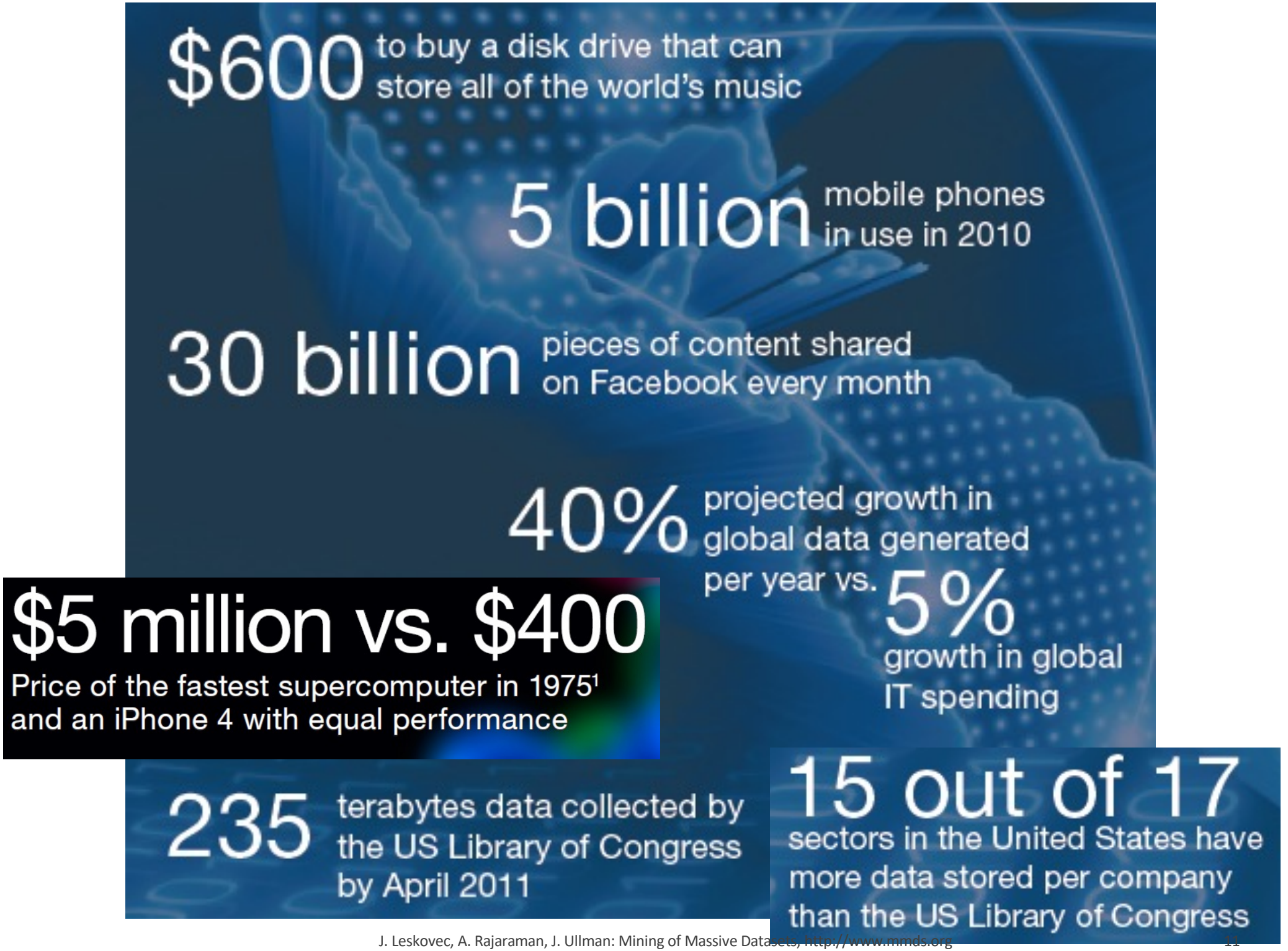
Questions?

Course Evaluation

- Project
- You get a real world problem. You get no instructions on what method to use. You have to select from those you have learned in the course.
- It is not a straight forward application of a method.
- You need to prove that your solution is correct and good.
  - Basically you have to convice us to “buy” your solution

What is Data Mining?
Knowledge discovery from data


1,000,000
VIDEOS
BLR
RS PUBLISH
2,340
POSTS
0,030
RE DOWNLOADED
00,000
ARE SENT
GOOGLE
4,500,000
VIDEOS
USERS
511,2
TWEETS
188,000,
EMAILS ARE
SKYPE
USERS
231,8
INSTAGRAM
2019
every
MINUTE
of the
DAY
PRESENTED BY DOMO

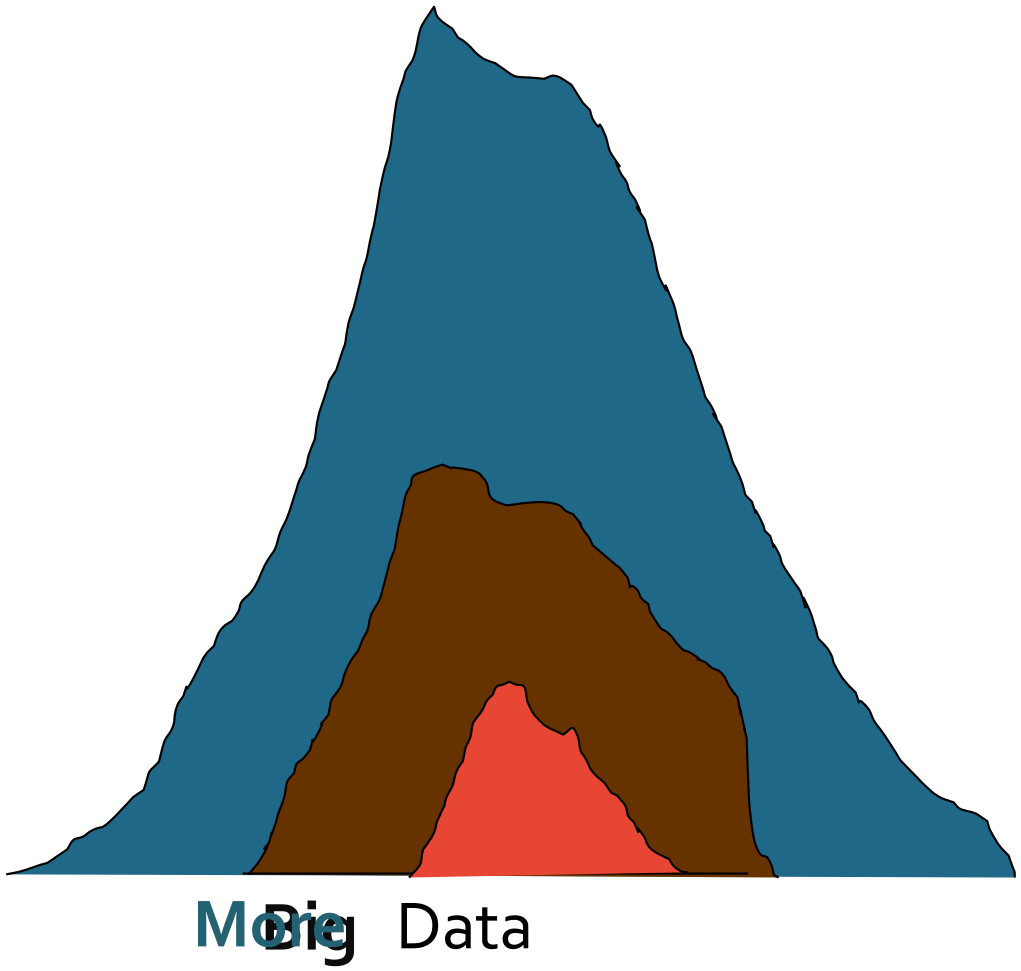

$600 to buy a disk drive that can store all of the world's music
5 billion mobile phones in use in 2010
30 billion pieces of content shared on Facebook every month
40% projected growth in global data generated per year vs. 5% growth in global IT spending
$5 million vs. $400
Price of the fastest supercomputer in 1975[1] 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 IS THE NEW GOLD"

Machine Generated Data

Internet Users

Employees / Data Engineers



Data contains value and knowledge

Data Mining

- But to extract the knowledge data needs to be
  - Stored
  - Managed
  - And ANALYZED ← this class

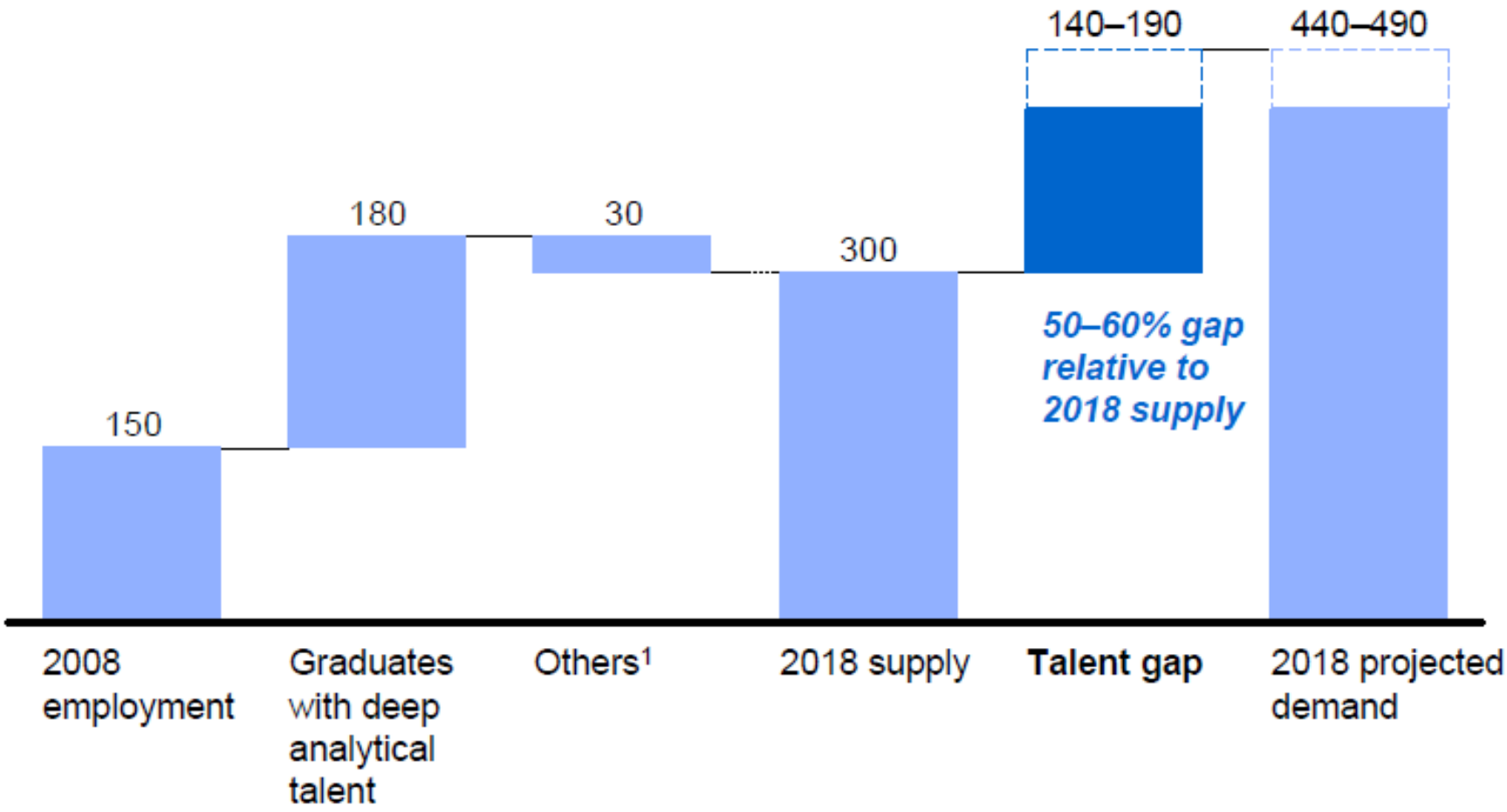
Data Mining ≈ Big Data ≈ Predictive Analytics ≈ Data Science

Or not?

Good news: Demand for Data Mining

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
Thousand people



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 Mining?

- **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

Machine Learning

- Good when we do not know what we are looking for.
- Not good if you know characteristics of the data you are looking

- Supervised vs unsupervised.

Data Mining

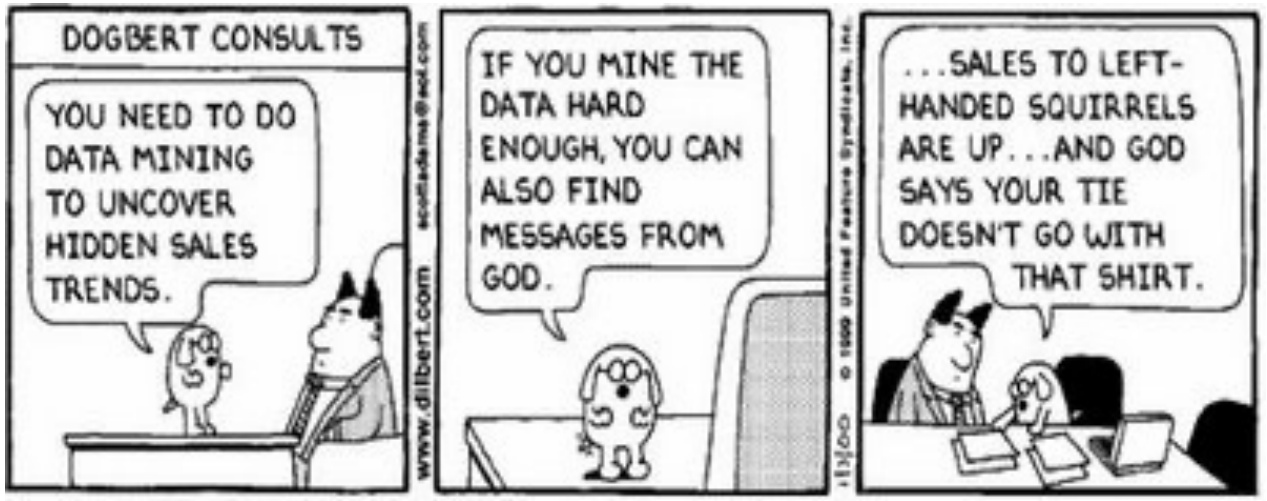
- Algorithmic
  - 1. Summarization
    - e.g. Page Rank
  - 2. Feature Extraction
    - The Cholera case
    - Frequent Itemsers
    - Similar Items

Data Mining 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**:
  - Roughly, if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap



Bonferroni by example

- We need to find evil-doers
  - they periodically gather in a hotel to plan an attack
- 1,000,000,000 = 10^9 people
- Everyone goes to a hotel once every 100 days
- A hotel holds 100 people

- In 100 days almost everyone has been to a hotel. So, 10^9 people in 100 days,
- hence 10^9 /100 = 10,000,000 = 10^7 people in a hotel per day
- 10^7 / (100 people per hotel) = 100,000 hotels we have
- We examine the hotel records for 1000 days
- We try to find two people that in 2 different days they met in different hotels.
- Suppose there are no evil-doers
- Probability 2 people on the same day to visit a hotel on the same day is 1/100 * 1/100 = 10^{-4}
- Chance that it is the same hotel is V * 10^{-4} = 10^{-4} We
  - Probability they both select the 1st hotel = $\frac{1}{10^5} * \frac{1}{10^5} = \frac{1}{10^{10}}$
  - Probability they both select the 2nd hotel = $\frac{1}{10^5} * \frac{1}{10^5} = \frac{1}{10^{10}}$
  - Probability they both select the 3rd hotel = $\frac{1}{10^5} * \frac{1}{10^5} = \frac{1}{10^{10}}$
  - …
  - Probability they both select the 10^5 th hotel = $\frac{1}{10^5} * \frac{1}{10^5} = \frac{1}{10^{10}}$
  - ---
  - Probability they choose the same hotel = prob they meet in the 1st + prob they meet in the 2nd etc = $\frac{1}{10^{10}} + \frac{1}{10^{10}} + \ldots + \frac{1}{10^{10}} = 10^5 * \frac{1}{10^{10}} = \frac{1}{10^5}$

- Probability they meet = prob decide to go on same day and choose the same hotel = $10^{-4} * \frac{1}{10^5} = 10^{-9}$

Bonferroni by example

- Probability they meet = prob decide to go on same day and choose the same hotel = $10^{-4} * \frac{1}{10^5} = 10^{-9}$
- Probability they meet at the same hotel in two different days = $10^{-9} * 10^{-9} = 10^{-18}$
- Pairs of two people meeting on the same hotel in two different days:
- Number of pairs of people is $\binom{10^9}{2} \cong \frac{10^{9^2}}{2} = 5*10^{17}$
- Number of pairs of days is $\binom{1000}{2} \cong \frac{1000^2}{2} = 5*10^5$
- The number of cases of 2 random people visiting the same hotel in two different days is:
  - $5*10^{17} * 5*10^5 * 10^{-18}$ = 250,000

- Assuming that there are 10 evil doers… then you need to do 250,000 checks to catch them. … too many combinations to check – we need to have some additional evidence to find "suspicious" pairs of people in some more efficient way

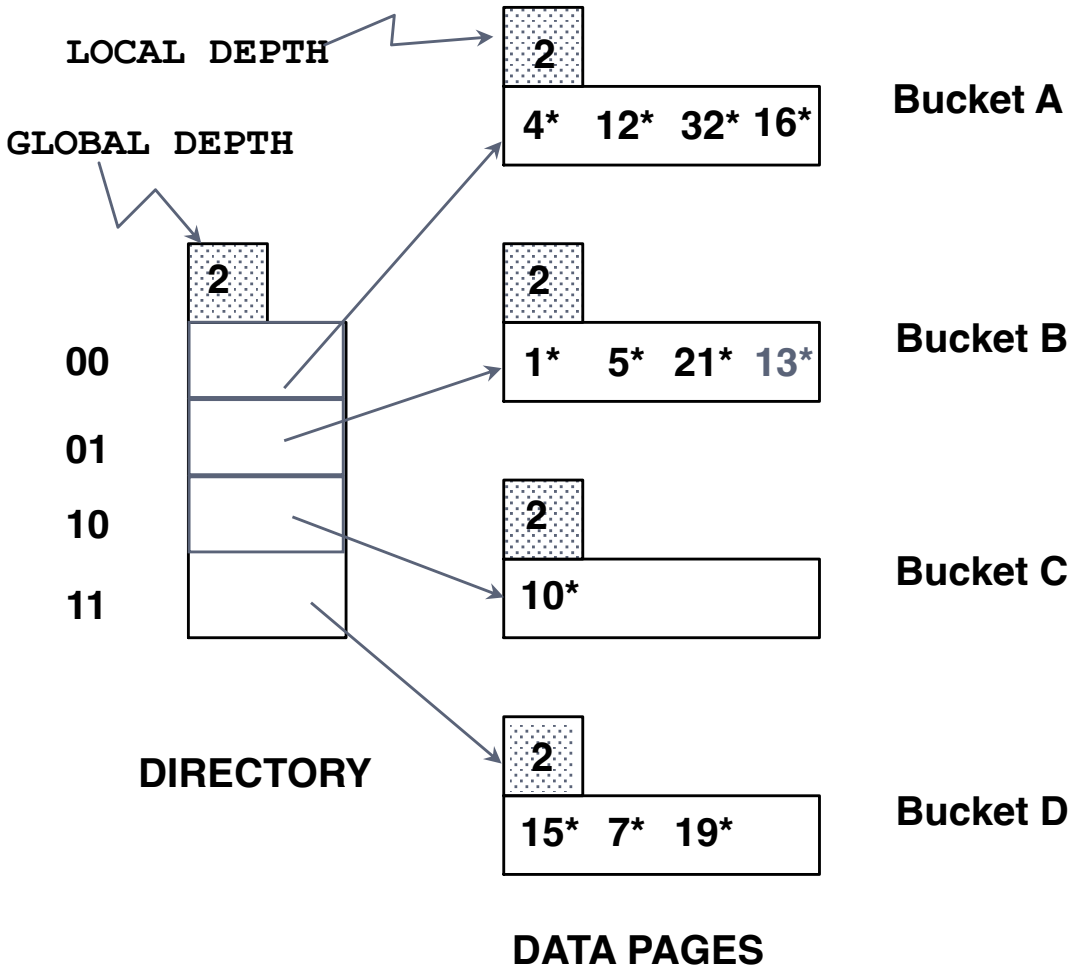
TF-IDF

- Collection of N documents
  - Need to find documents related to COVID
    - (or decide if a document is related to COVID)
- Frequency of term word i in document j:

$$TF_{ij} = \frac{f_{ij}}{max_k f_{kj}}$$

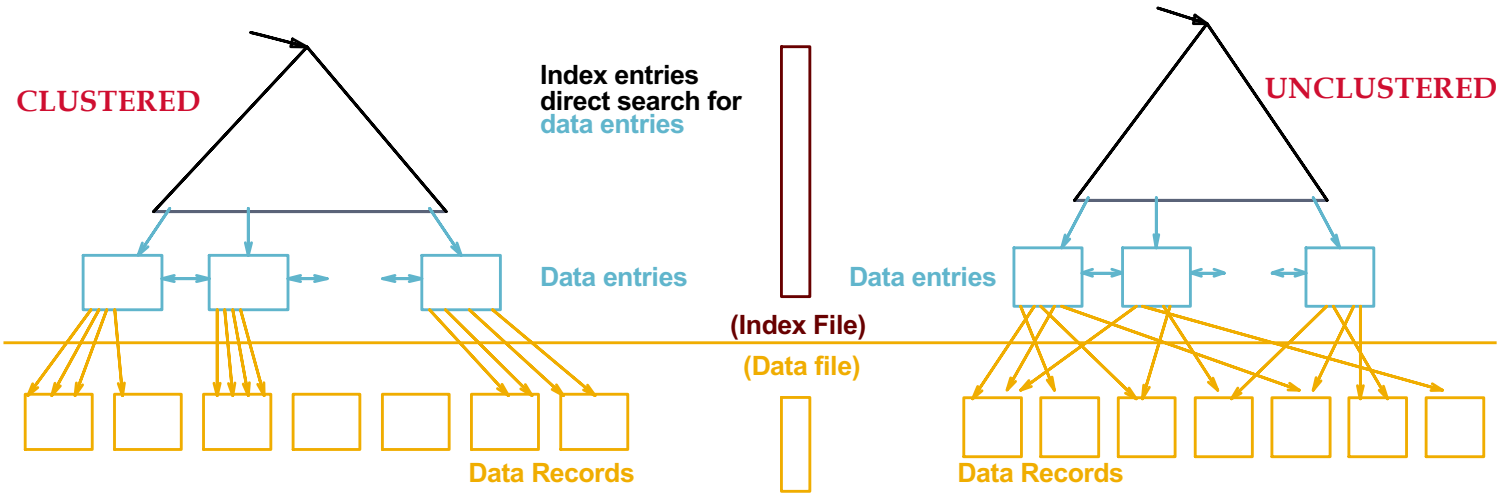
- Inverse Term Frequency
- $IDF_i = log_2\left(\frac{N}{n_i}\right)$
- $TF_{ij} * IDF_i$

Hash Functions

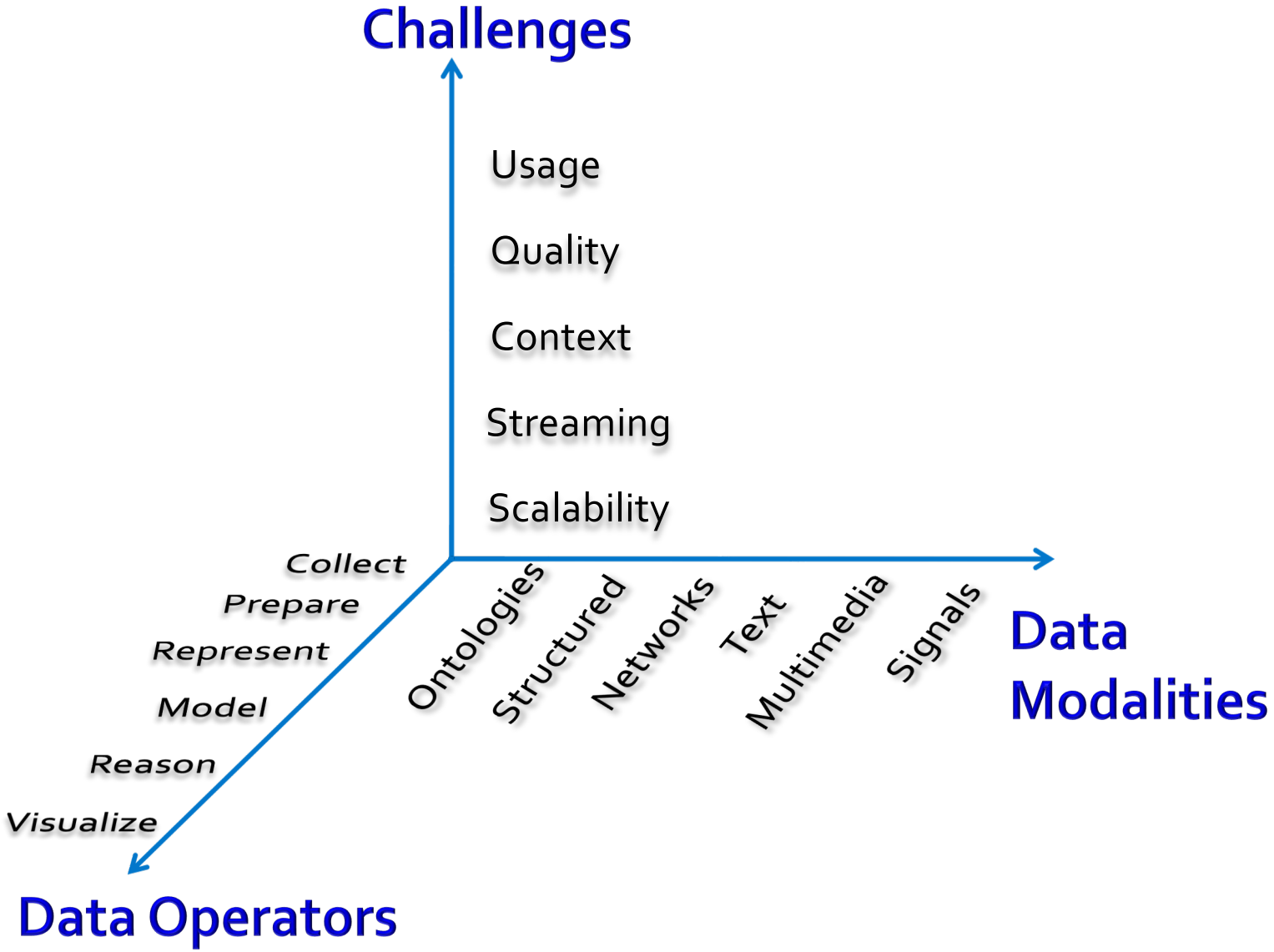


Hash function = last 2 bits in its binary representation

Clustered vs. Unclustered Index

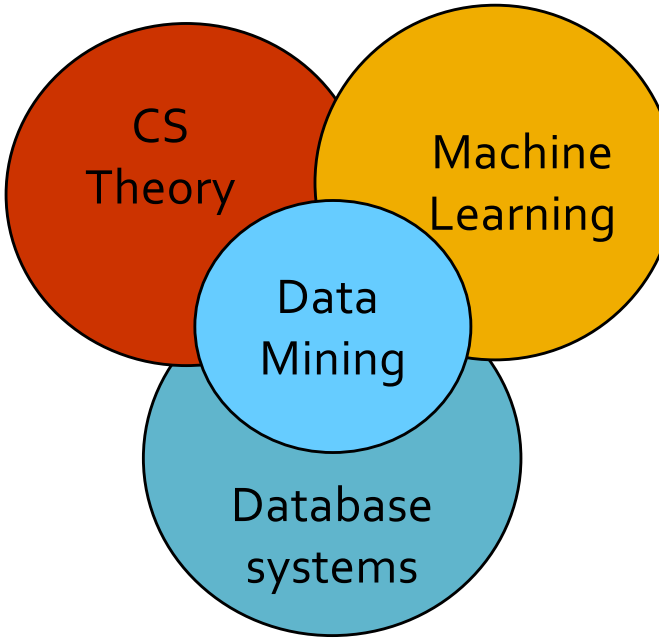


What matters when dealing with data?



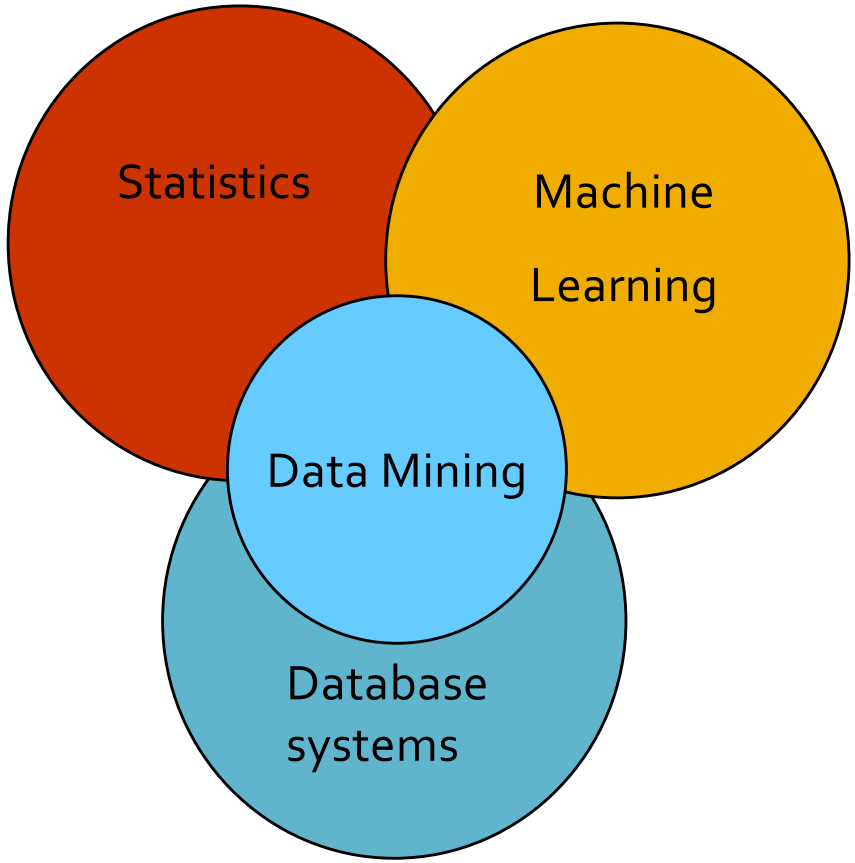
Data Mining: Cultures

- **Data mining 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 mining is an extreme form of **analytic processing** – queries that examine large amounts of data
    - Result is the query answer
  - To a ML person, data-mining is the **inference of models**
    - Result is the parameters of the model
- **In this class we will do both!**



This Class

- **This class 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 (high level. See Big Data Course for details)
  - Streams and online algorithms
  - Single machine in-memory

What will we learn?

- **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