Introduction

Advanced Search Techniques for Large Scale Data Analytics

Pavel Zezula and Jan Sedmidubsky
Masaryk University
http://disa.fi.muni.cz

Course

- **Teachers:**
  - prof. Ing. Pavel Zezula, CSc.
  - RNDr. Jan Sedmidubsky, Ph.D.

- **Inspiration:**
  - Inspired by the book of people from Stanford:
    - Jure Leskovec, Anand Rajaraman, Jeff Ullman: **Mining of Massive Datasets**. Cambridge University Press, 2^{nd} Edition, 476 pages, 2014.
    - Additional information: http://mmds.org/

Course Outline

- Introduction
- Block 1:
  - Support for Distributed Processing
  - Retrieval Evaluation
  - Clustering
  - Exercises on topics of Block 1
- Block 2:
  - Finding Frequent Item Sets
  - Finding Similar Items
  - Searching in Data Streams
  - Exercises on topics of Block 2
- Block 3:
  - Link Analysis
  - Search Applications
  - Seznam.cz – A Search Engine in Practice
  - Exercises on topics of Block 3

Course Outline – Block 1

- Block 1:
  - Support for Distributed Processing
    - Distributed file system
    - MapReduce, Algorithms using MapReduce
    - Cost model and performance evaluation
  - Retrieval Evaluation
    - Retrieval metrics
  - Clustering
    - K-means algorithms
    - Clustering in non-Euclidean spaces
    - Clustering for streams and parallelism

Course Outline – Block 2

- Block 2:
  - Finding Frequent Item Sets
    - Handling large datasets in main memory
    - Counting frequent items in a stream
  - Finding Similar Items
    - Applications of near-neighbor search
    - Shingling of documents
    - Similarity-preserving summaries of sets
    - Locality sensitive hashing
  - Searching in Data Streams
    - The stream data model
    - Filtering streams

Course Outline – Block 3

- Block 3:
  - Link Analysis
    - Page Rank
    - Topic sensitive
    - Link spam
  - Search Applications
    - Advertising on the web
    - Recommendation systems (collaborative filtering)
    - Mining social-network graphs
  - Seznam.cz
    - A Search Engine in Practice

What is Searching?

Search – The Goals

Goals:

1) We search to get **results**
2) We ask to find **answers**
3) We use filters so that the right staff **finds us**
4) We **browse** while wandering and way-finding in restricted space

- In reality, we move fluidly between modes of ***ask***, ***browse***, ***filter***, and ***search***

Search – The Traditional Way

- Defined by software
- Buy engine, then figure out what it is good for

- It often fails because
  - It is not easy to use
  - It is not able to handle needed content types

Search – Some Quantitative Facts

- 85% of all web traffic comes from search engines
- 450+ million searches/day are performed in North America alone
- 70%+ of all searches are done on Google sites

Search is the **most popular** application
(second to E-mail??)

Search – The Best First

- 60% of searchers NEVER go past 1st page of search results
- The top three results draw 80% of the attention
- The first few results inordinately influence query reformulation

Search – As an Interaction

- When we search, our next actions are reactions to the stimuli of a previous search
- What we find is changing what we seek
- In any case, search must be:

fast, ***simple***, and ***relevant***

Search – Basic Components

- Elements of global search:

  Users – goals, psychology, behavior

  Interface – interaction, affordances, language

  Engine – features, technology, algorithms

  Content – indexing structure, metadata

  Creators – tools, process, incentives

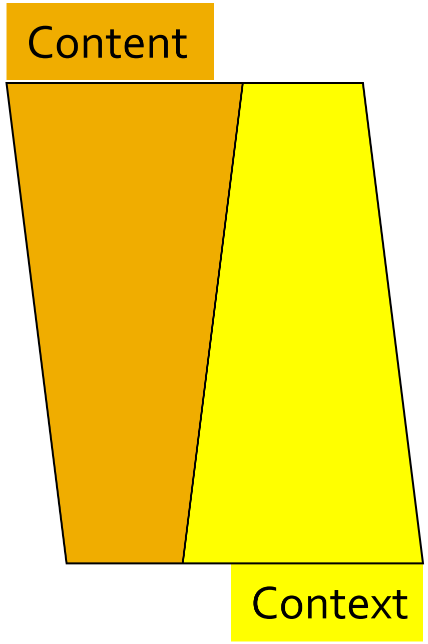
Search Changes Our Cognitive Habits

- Assuming information continually and instantaneously available on the web:
  1) We are increasingly handing off the job of remembering to search engines
  2) When we need answer, we do not think, we go immediately to a nearest Web connection
  3) When we expect information to be easily found again, we do not remember it well
  4) Our original memory of facts is changing to a memory of ways to find the facts

Users and Their Intent

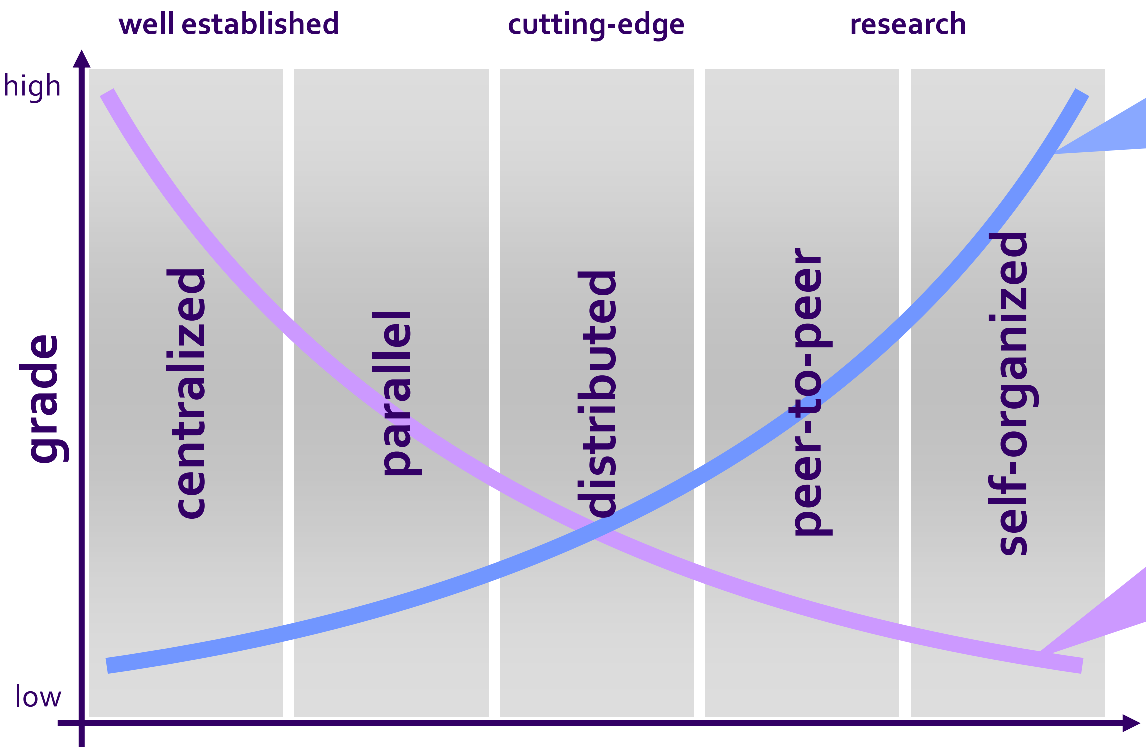
Search is **subjective** and also depends on **visual** and **emotional** attributes, e.g. *shocking*, *funny*, etc.

Prevalent strategy



- **Browser**
  - not clear end-goal; series of unrelated searches; jump across unrelated topics; expects surprises and random search hints
- **Surfer**
  - moderate clarity of end-goal; exploratory actions at the beginning; e.g. planning a holiday
- **Searcher**
  - very clear about what is searching for; completeness and clarity of results are important

Evolution of Search Engine Strategies



Scalability

data volume
number of users (queries)
variety of data types
multi-aspect queries

Determinism

exact match	‣ similarity
precise	‣ approximate
same answer	‣ variable
fixed infrastr.	‣ mapping

What is Data Mining?
Knowledge discovery from data

$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 contains value and knowledge

Data Mining

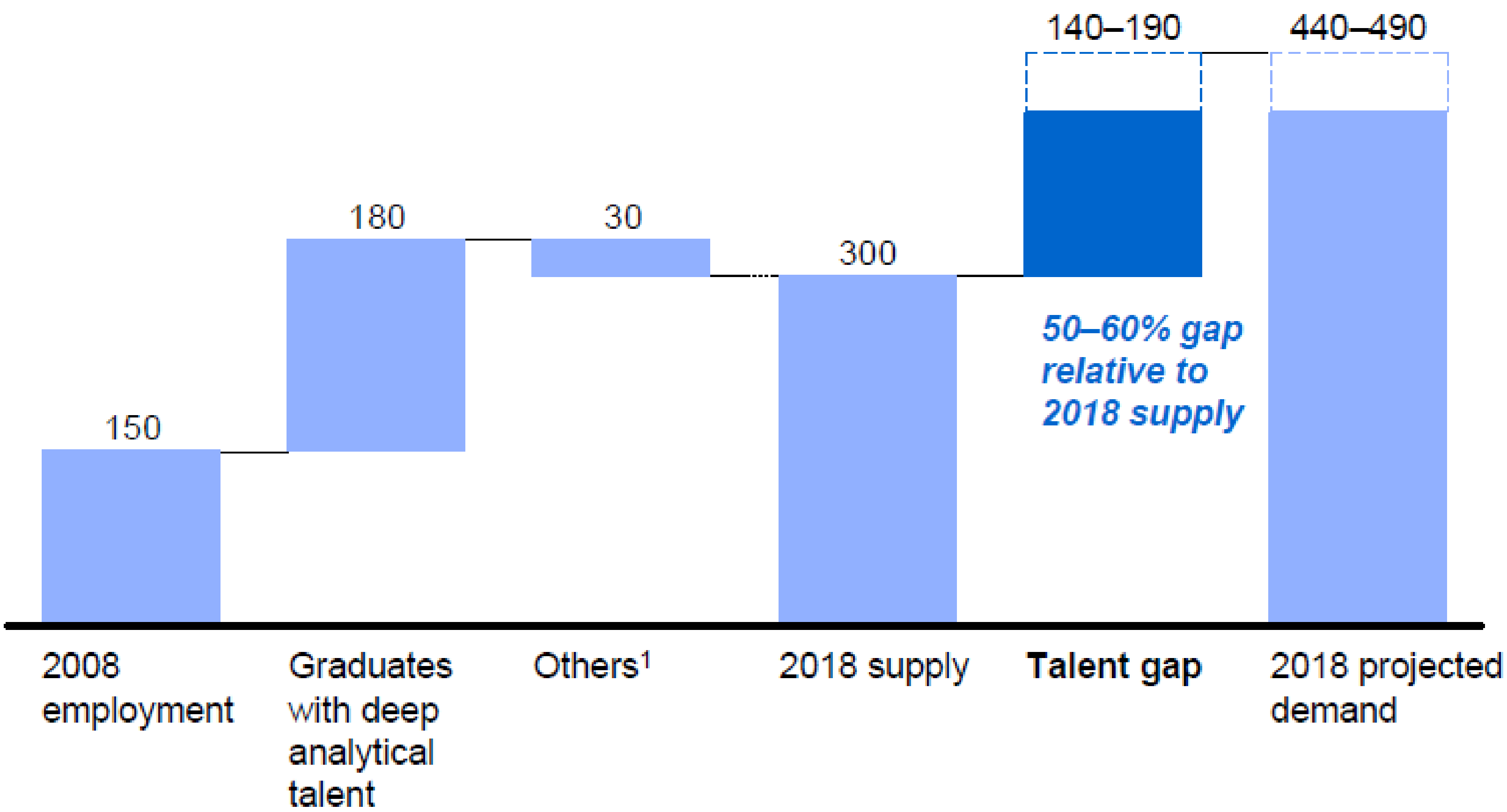
- But to extract the knowledge data needs to be
  - Stored
  - Managed
  - And ANALYZED ← this class

Data Mining ≈ Big Data ≈ Predictive Analytics ≈ Data Science

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?

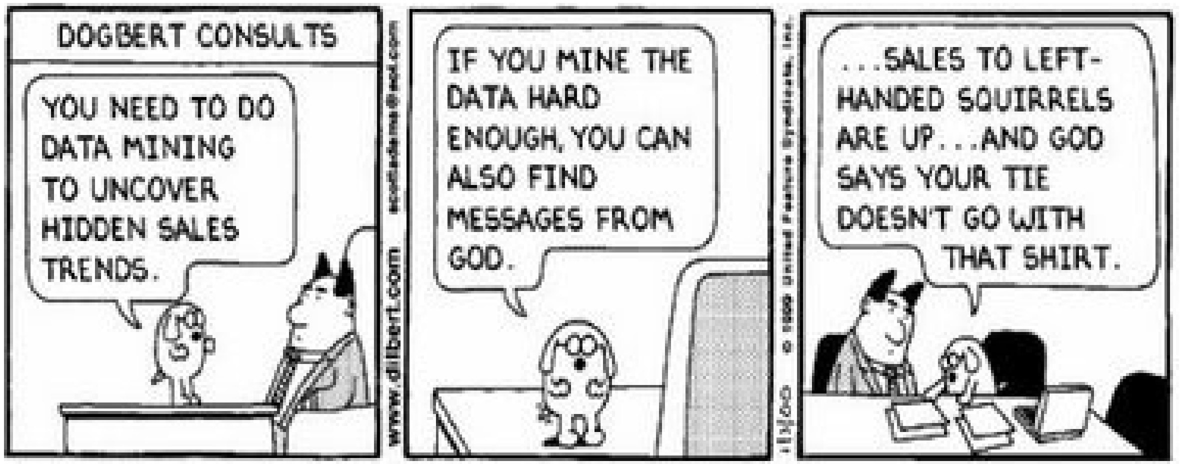
- **Given lots of data**
- **Discover patterns and models that are:**
  - **Valid:** hold on new data with some certainty
  - **Useful:** should be possible to act on the item
  - **Unexpected:** non-obvious to the system
  - **Understandable:** humans should be able to interpret the pattern

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

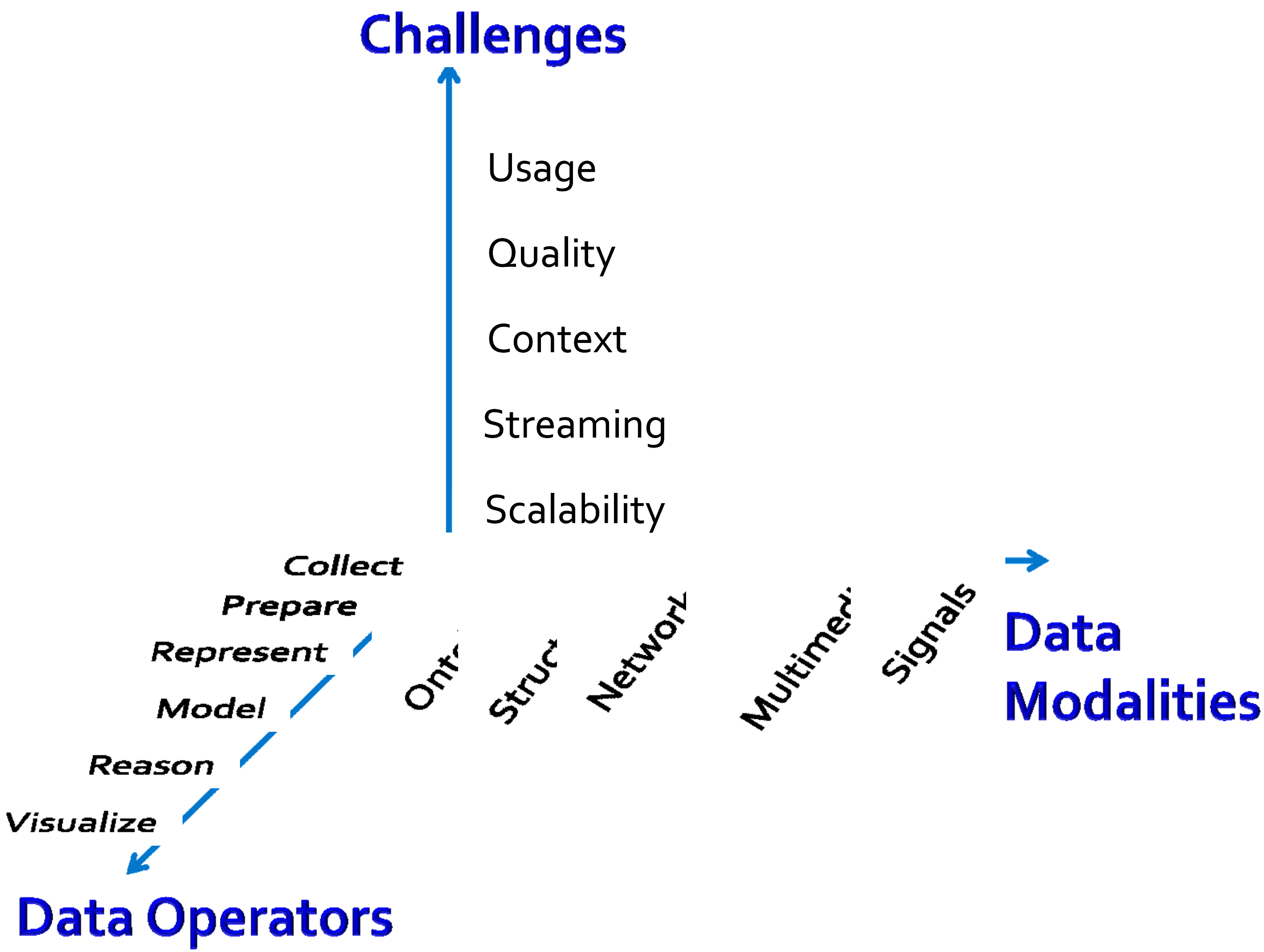


Meaningfulness of Analytic Answers

Example:

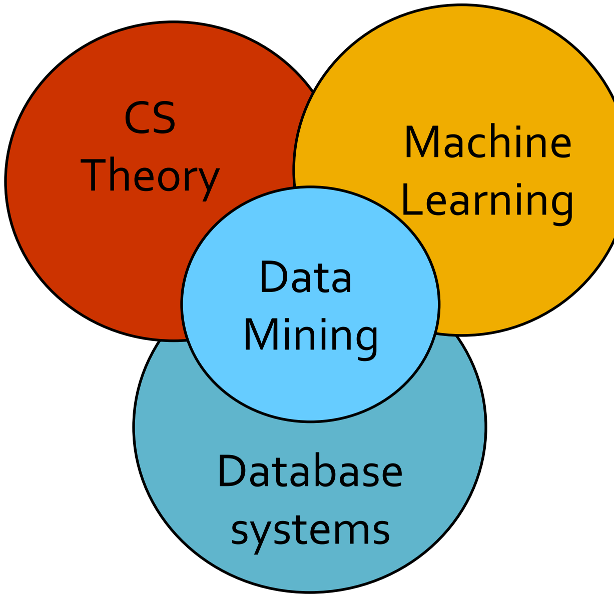
- We want to find (unrelated) people who **at least twice have stayed at the same hotel on the same day**
  - 10^9 people being tracked
  - 1,000 days
  - Each person stays in a hotel 1% of time (1 day out of 100)
  - Hotels hold 100 people (so 10^5 hotels)
  - **If everyone behaves randomly (i.e., no terrorists) will the data mining detect anything suspicious?**
- **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?



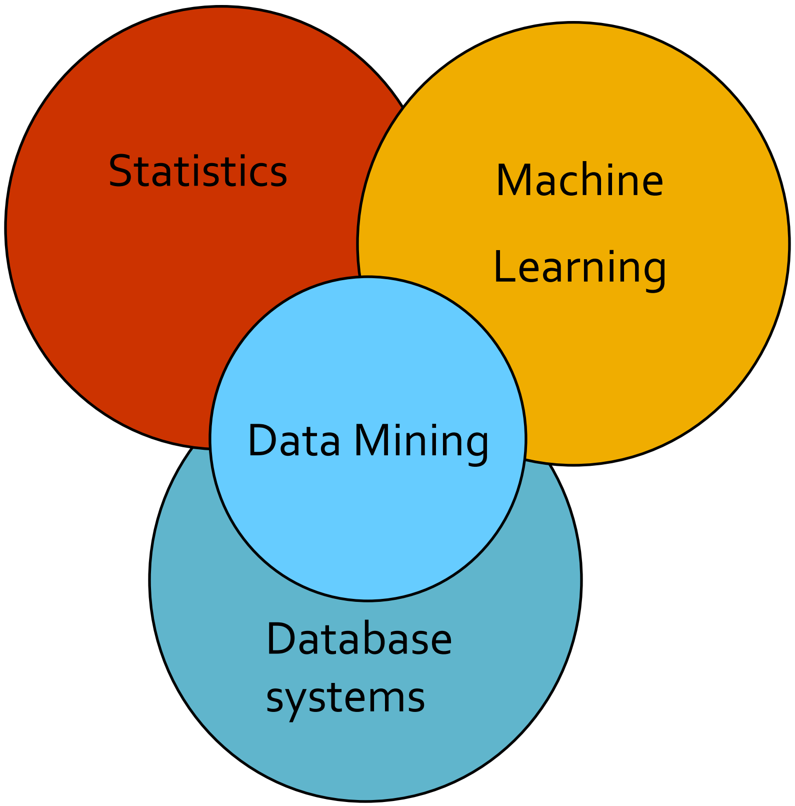
Data Mining: Cultures

- **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 Course

- **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

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

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

How do you want that data?