WEB大数据挖掘(一)

About the Course

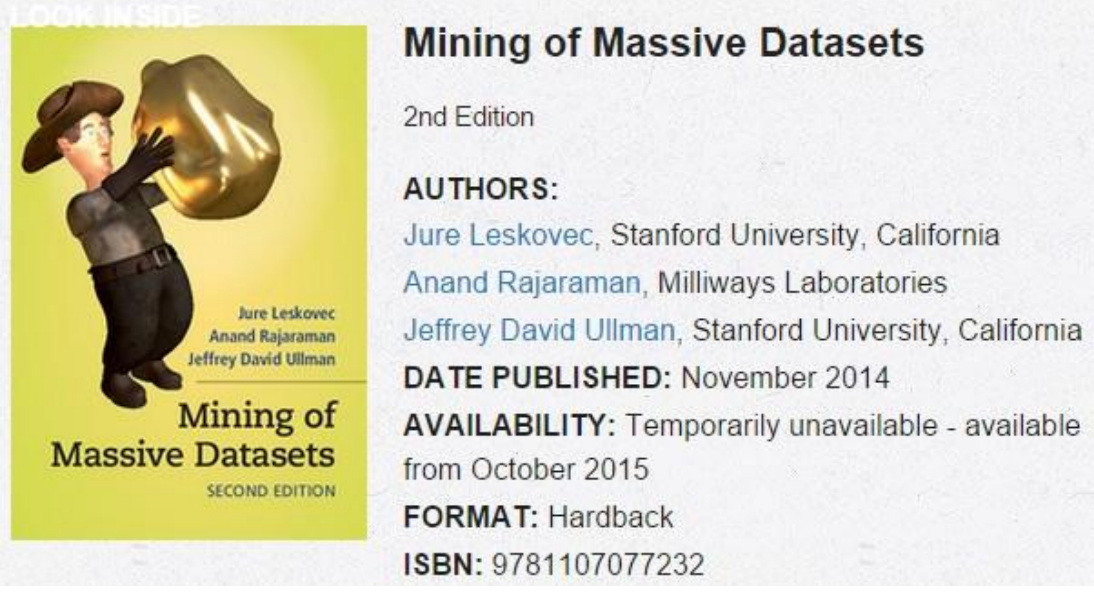
课程基本情况

- 参考书

  英文版: **Mining of Massive Datasets**

  参考资料网址: http://www.mmds.org/

  中文版：大数据:互联网大规模数据挖掘与分布式处理



基本情况

- 具体讲读内容

  参考书: 大数据挖掘

  参考资料网址: http://www.mmds.org/

Chapter	Title	Book		Slides	
	Preface and Table of Contents	PDF			
Chapter 1	Data Mining	PDF		PDF	PPT
Chapter 2	Map-Reduce and the New Software Stack	PDF		PDF	PPT
Chapter 3	Finding Similar Items	PDF		PDF	PPT
Chapter 4	Mining Data Streams	PDF	Part 1: Part 2:	PDF PDF	PPT PPT
Chapter 5	Link Analysis	PDF	Part 1: Part 2:	PDF PDF	PPT PPT
Chapter 6	Frequent Itemsets	PDF		PDF	PPT
Chapter 7	Clustering	PDF		PDF	PPT
Chapter 8	Advertising on the Web	PDF		PDF	PPT
Chapter 9	Recommendation Systems	PDF	Part 1: Part 2:	PDF PDF	PPT PPT
Chapter 10	Mining Social-Network Graphs	PDF	Part 1: Part 2:	PDF PDF	PPT PPT
Chapter 11	Dimensionality Reduction	PDF		PDF	PPT
Chapter 12	Large-Scale Machine Learning	PDF	Part 1: Part 2:	PDF PDF	PPT PPT
	Index	PDF			
	Errata	HTML			

课程成绩计算

- 平时成绩: **20%**
  - 课上(后)问答题
- **2**个大作业: **80%**
- 课件下载：学院网站
- 联系邮箱:

  教师：杨征路，邮箱：yangzl@nankai.edu.cn

  2个助教，问题发到： webbigdata@163.com

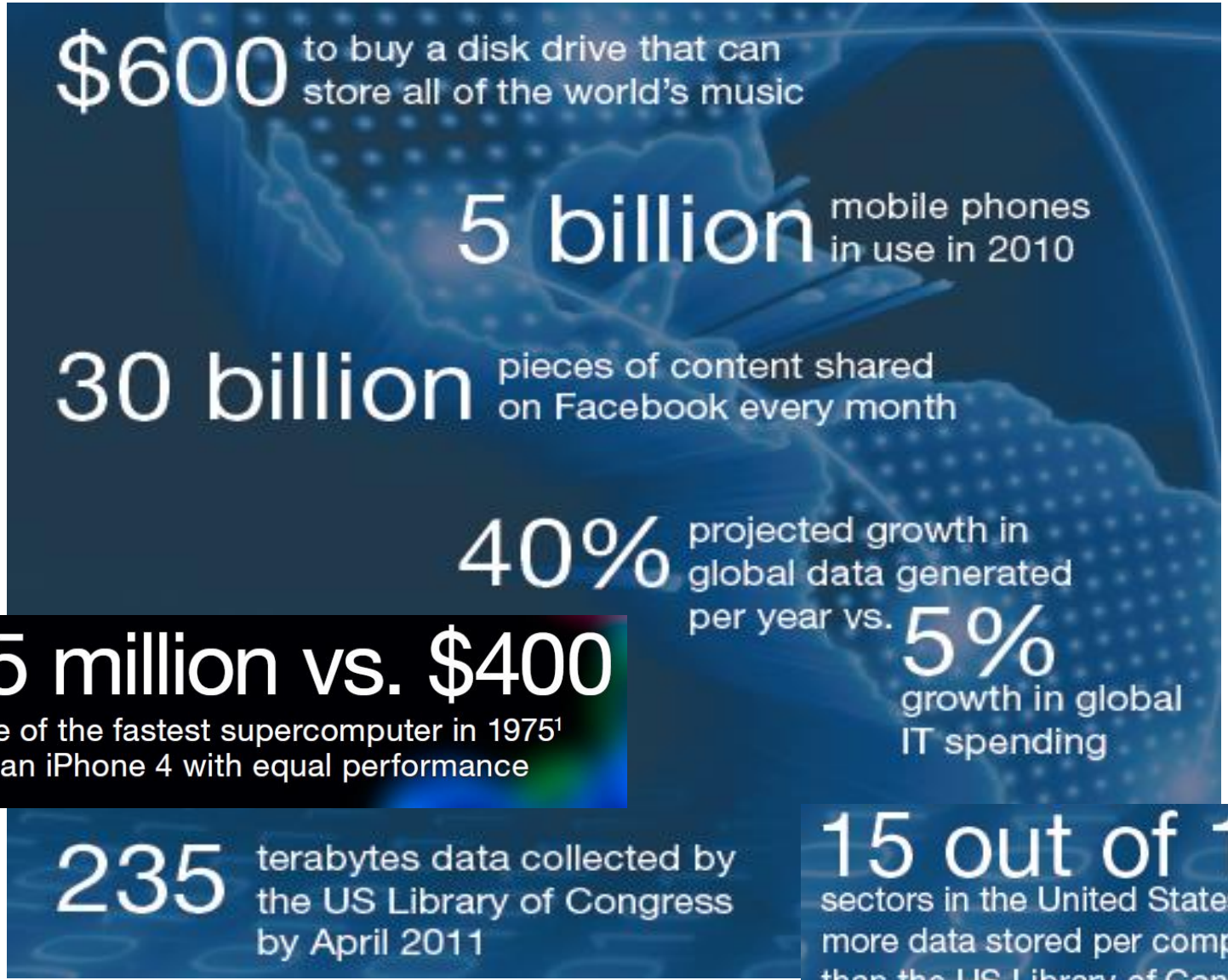
Prerequisites

- Algorithms
  - Dynamic programming, basic data structures
- Basic probability
  - Moments, typical distributions, MLE, …
- Programming
  - Your choice, but C++/Java will be very useful

- We provide some background, but the class will be fast paced

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

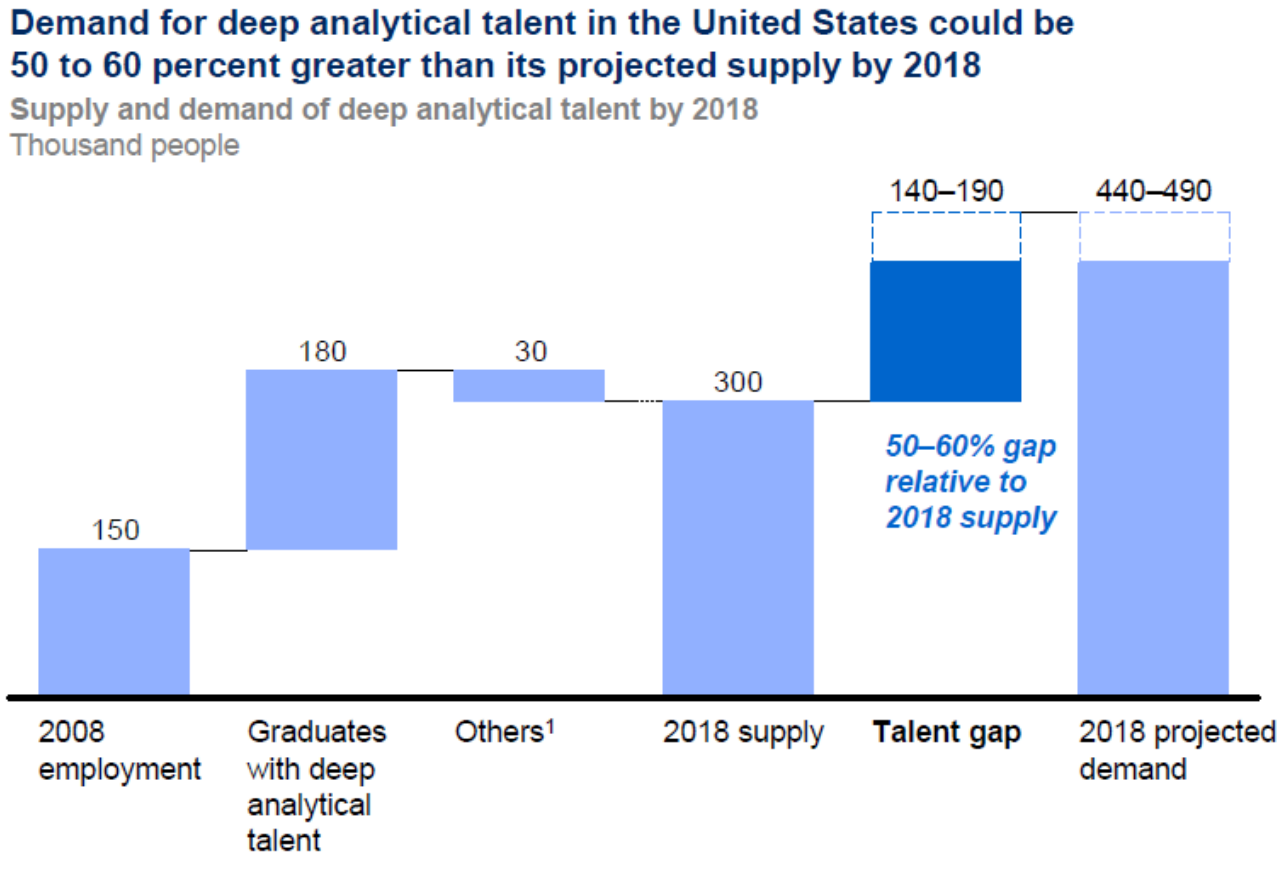
Data contains value and knowledge

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



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

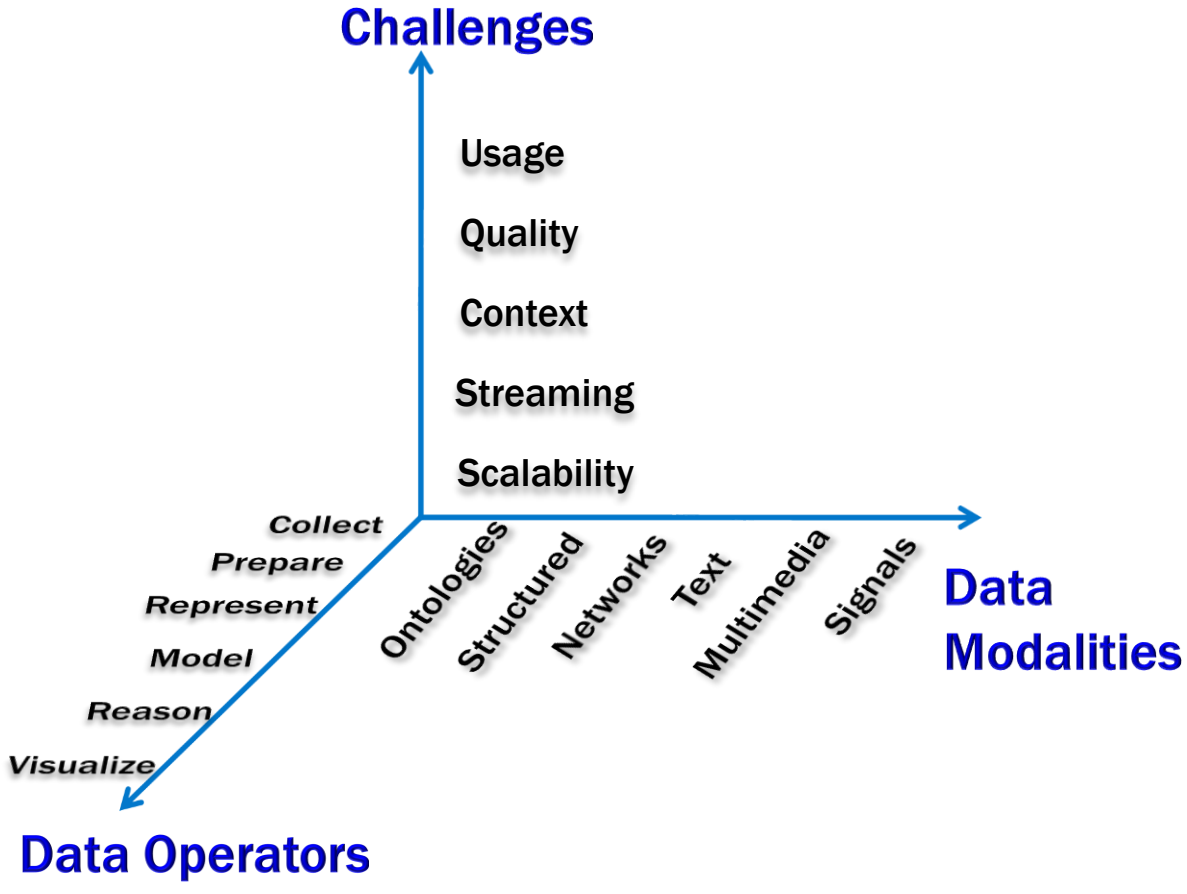
- **A risk with "Data mining" is that an analyst can "discover" patterns that are meaningless**
- Statisticians call it **Bonferroni's principle (邦弗朗尼原理)**:
  - Roughly, if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap

Meaningfulness of Analytic Answers

Example:

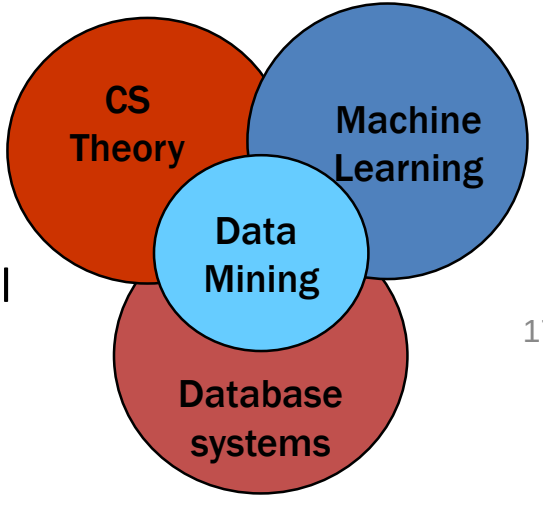
- We want to find (unrelated) people who **at least twice have stayed at the same hotel on the same day**
  - 10^9 people being tracked
  - 1,000 days
  - Each person stays in a hotel 1% of time (1 day out of 100)
  - Hotels hold 100 people
  - **If everyone behaves randomly (i.e., no terrorists) will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



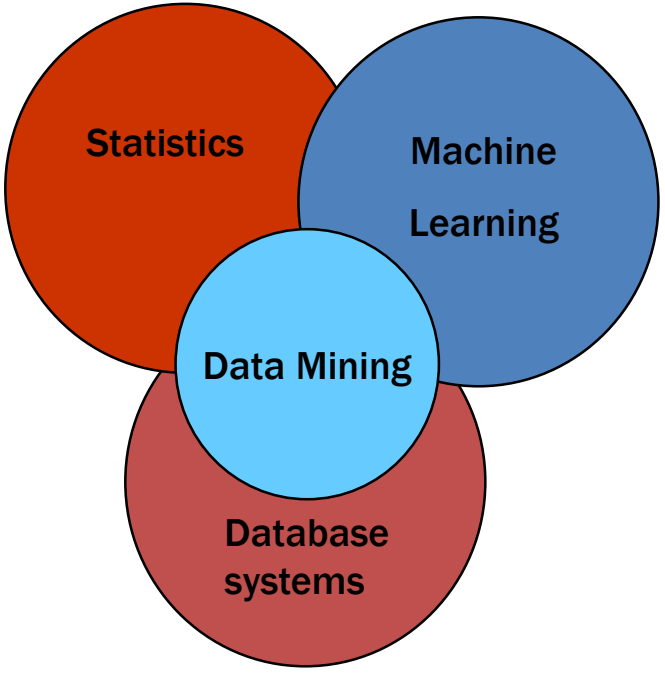
Data Mining: Cultures

- **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 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



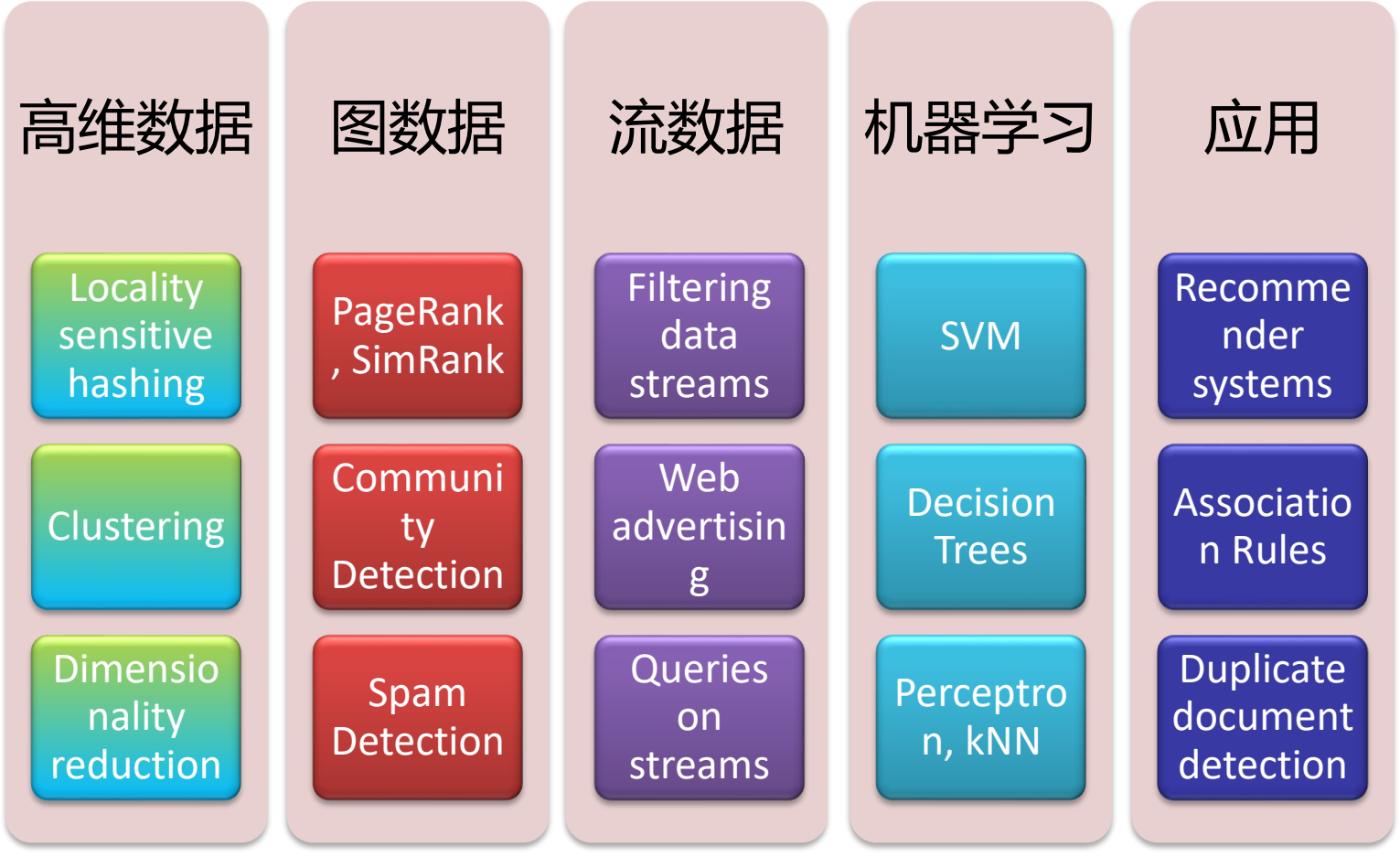
What will we learn?

- 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:
  - 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



How do you want that data?