大数据分析与挖掘

2022秋，深高研院

培养目标

- 系统性学习大数据分析与挖掘的基本概念和原理，了解模型精度的评价方法。
- 深入地掌握互联网大规模数据挖掘与分布式处理方法，掌握各种回归、分类、聚类方法以及对其进行比较。
- 通过学习关联分析、社交网络分析以及文本挖掘，能够对实际数据进行处理、分析，并建立解释合理的模型。

课程基础

- **Algorithms**
  - Dynamic programming, basic data structures
- **Basic probability**
  - Moments(矩), typical distributions, Maximum-Likelihood Estimation (MLE), …
- **Programming**
  - Python or Java will be very useful

考核方式

- 考核成绩 = 平时成绩30% + 期末成绩70%
- 平时成绩 = 作业成绩80% + 考勤成绩20%

参考资料

- **Readings:** Book **Mining of Massive Datasets** with A. Rajaraman and J. Ullman

  Free online:

  http://www.mmds.org
- **Jure Leskovec, Anand Rajaraman, Jeff Ullman, 王斌[译]. 大数据：互联网大规模数据挖掘与分布式处理（第2版），人民邮电出版社，2015**
- **Reference material:**

  http://cs246.stanford.edu
  - Lecture slides
  - Homeworks, solutions
  - Readings

第一章：数据挖掘基本概念

主讲：陈爱国

大数据分析与挖掘

Outline

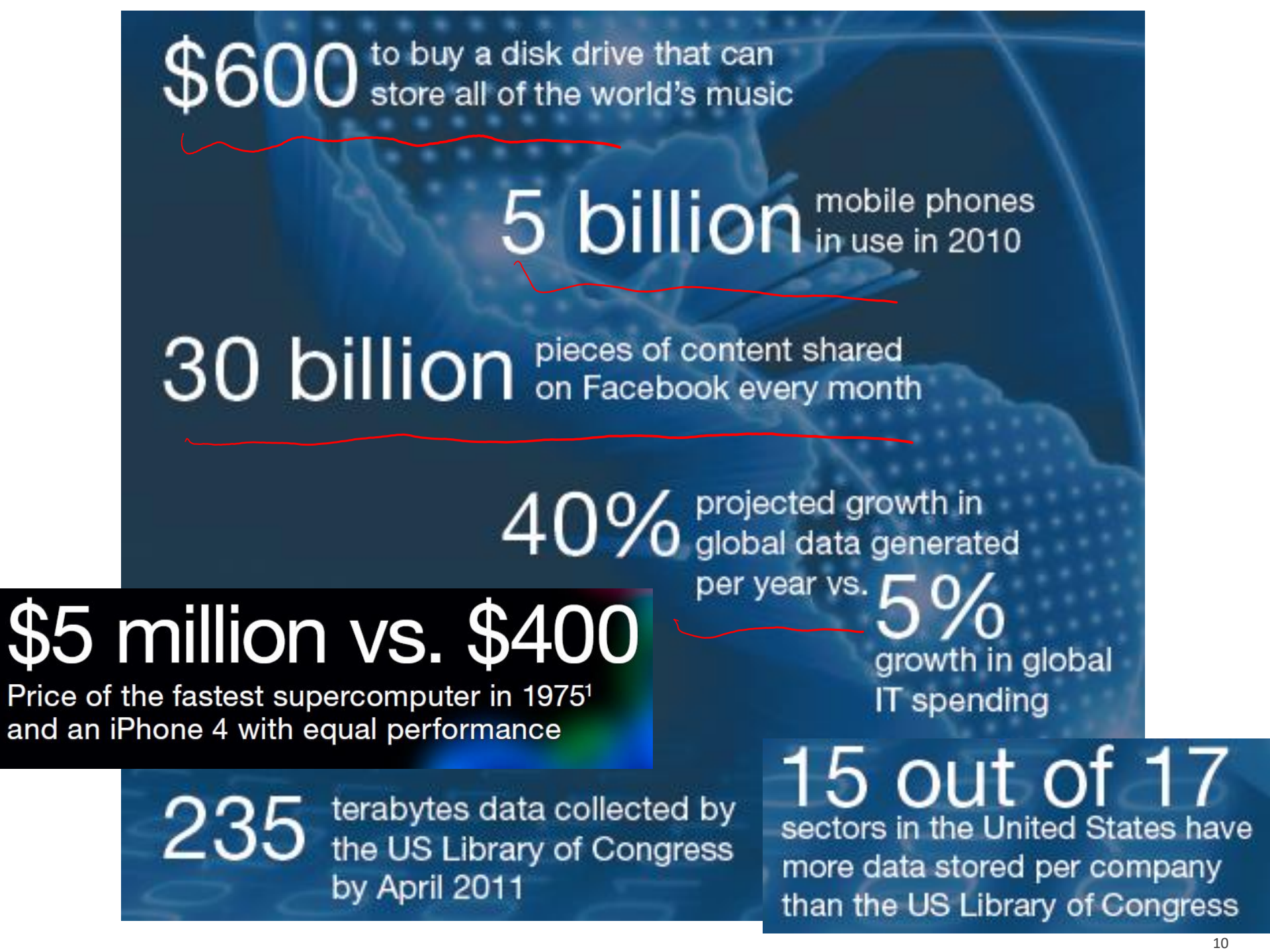
- 1.1 What is Data Mining?
  - 1.1.1 Modeling
  - 1.1.2 Statistical Modeling
  - 1.1.3 Machine Learning
  - 1.1.4 Computational Approaches to Modeling
  - 1.1.5 Summarization
  - 1.1.6 Feature Extraction
- 1.2 Statistical Limits on Data Mining
- 1.3 Things Useful to Know
- 1.4 Conclusion & Outline of the Book

1.1 What is Data Mining?

- In the 1990's "data mining" was an exciting and popular new concept.
- Around 2010, people instead started to speak of "big data." ——from Data Modalities
- Today, the popular term is "data science." ——from systematic

1.1 What is Data Mining?

- However, during all this time, the concept remained the same :
  - use the most powerful hardware, the most powerful programming systems, and the most efficient algorithms to solve problems in science, commerce, healthcare, government, the humanities, and many other fields of human endeavor.


$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

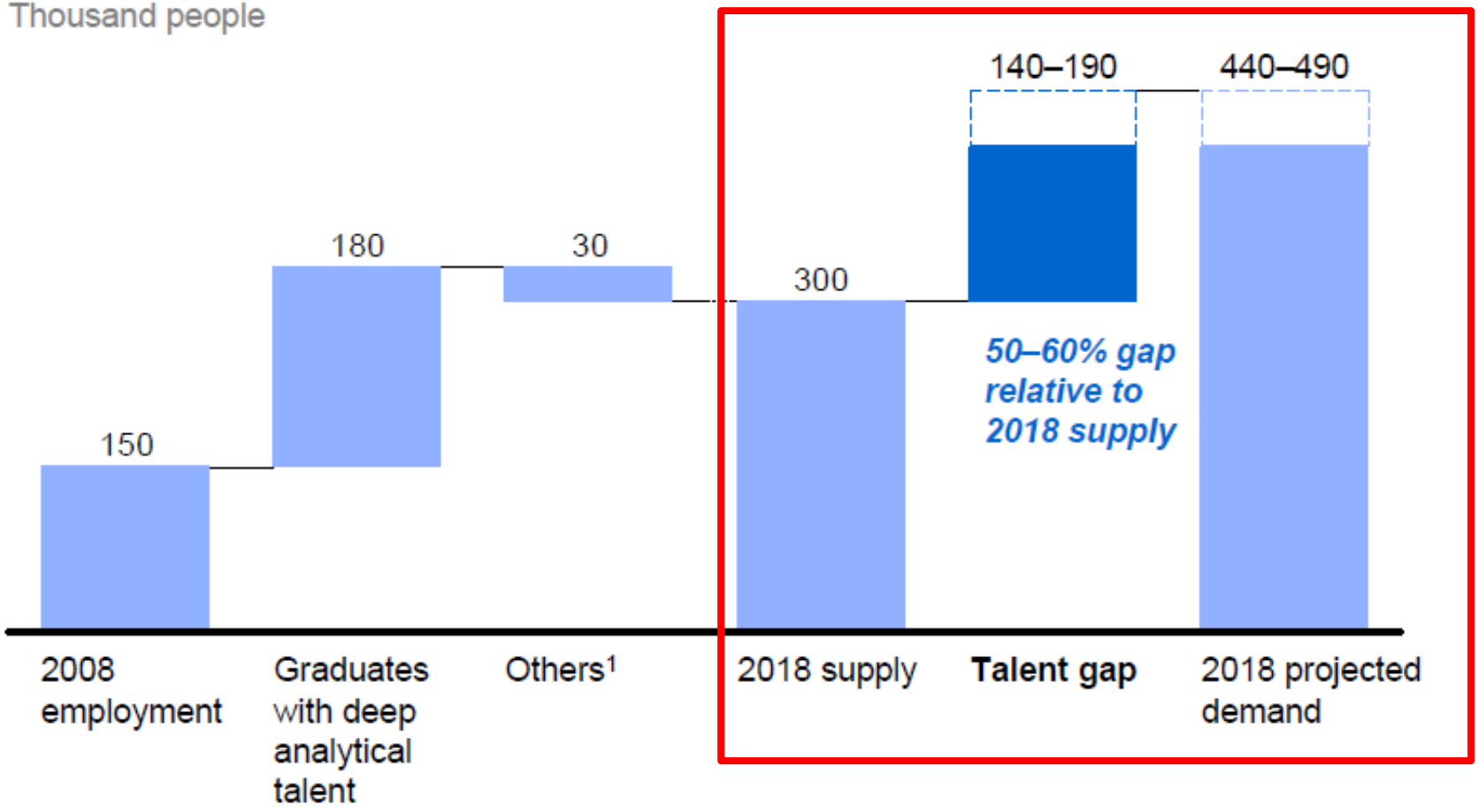
Data contains value and knowledge

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 This Course Is About

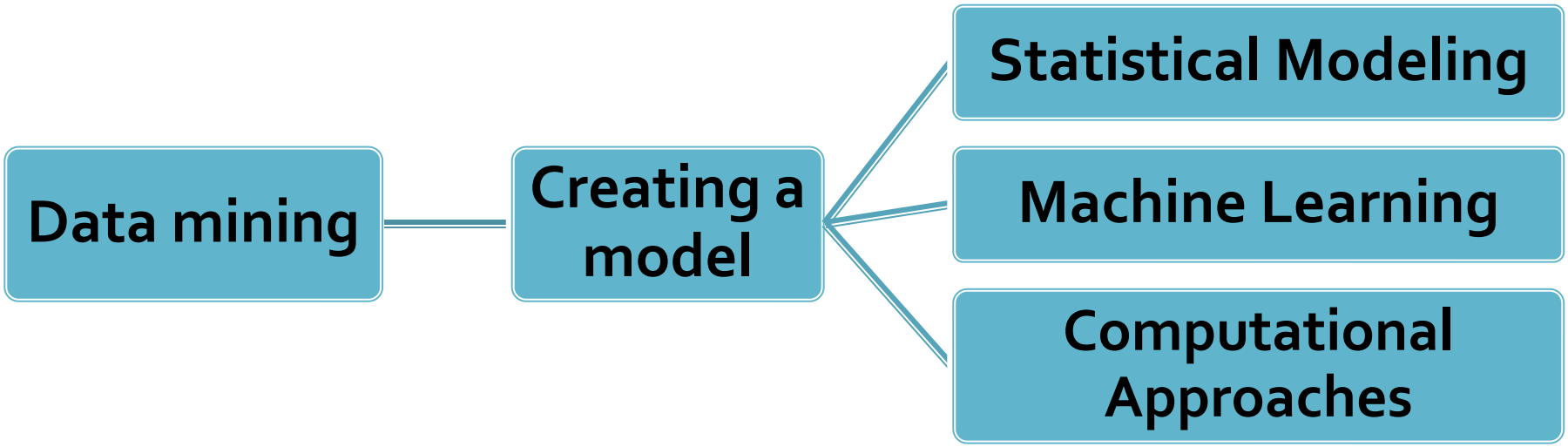
- ***Data mining*** **= extraction of actionable information from (usually) very large datasets, is the subject of extreme hype, fear, and interest**
- It's not all about machine learning
- But most of it is
- Emphasis in this course on algorithms that scale
  - Parallelization often essential

Data Mining——研究什么

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

Data Mining ≈ Big Data ≈ Predictive Analytics ≈ Data Science

1.1.1 Modeling



- To many, data mining is the process of creating a model from data
- Often by the process of machine learning, or other methods

1.1.1 Modeling

- However, more generally, the objective of data mining is an algorithm.
  - locality-sensitive hashing
  - a number of stream-mining algorithms
- Yet in many important applications, the hard part is creating the model, and once the model is available, the algorithm to use the model is straightforward.

What's the difference between model and algorithm?

1.1.2 Statistical Modeling

- Statisticians were the first to use the term “data mining.”
- Now, statisticians view data mining as the construction of a statistical model, that is, an underlying distribution from which the visible data is drawn.

1.1.3 Machine Learning

- There are some who regard data mining as synonymous with machine learning
- Machine-learning practitioners use the data as a training set, to train an algorithm of one of the many types used for machine-learning, such as Bayes nets, support-vector machines, decision trees, hidden Markov models, and a great variety of others
- Problems
  - be uncompetitive in situations where we can describe the goals of the mining more directly
  - not explainable

1.1.4 Computational Approaches to Modeling

- **Computer scientists** tend to look at data mining as an algorithmic problem
- In this case, a model of the data is simply the answer to a complex query about that data
  - For instance, given the set of numbers of Example 1.2, we might compute their average and standard deviation
- There are many different approaches to modeling data
  - the possibility of constructing a random process whereby the data could have been generated
  - Summarizing the data succinctly and approximately
  - Extracting the most prominent features of the data and ignoring the rest

(1) Summarization

e.g.

- PageRank
- Clustering

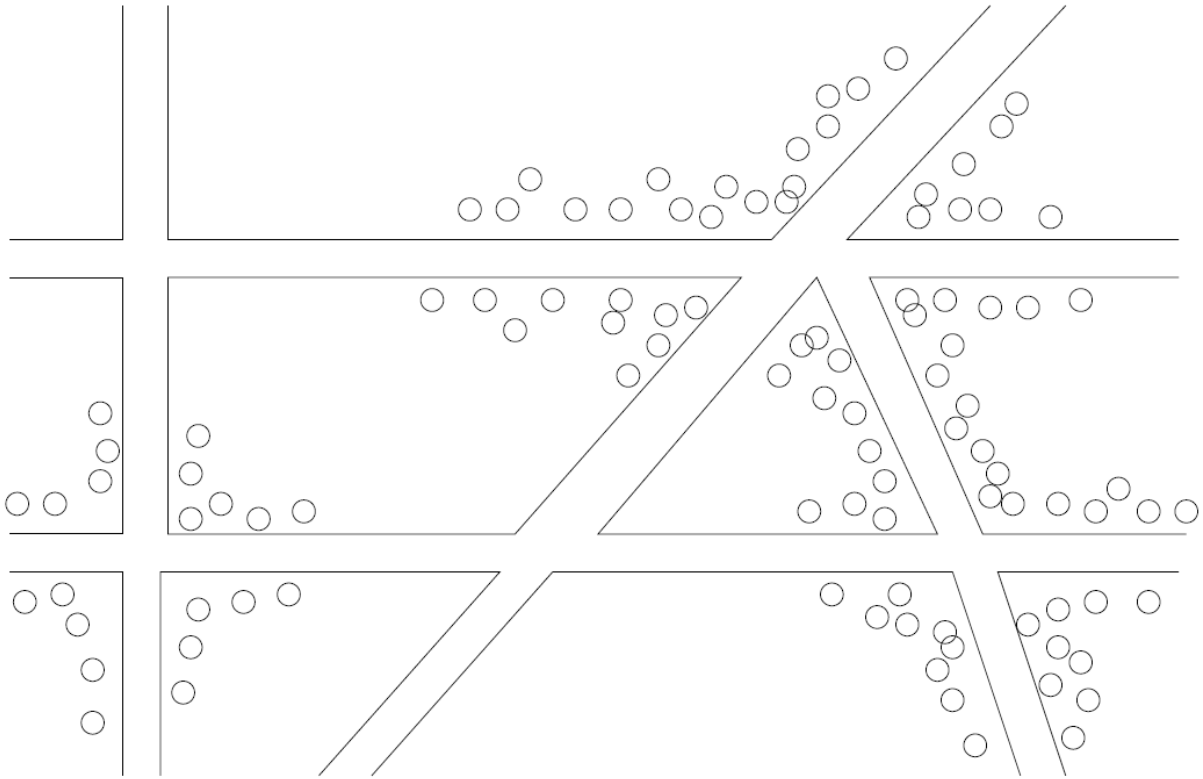


Figure 1.1: Plotting cholera cases on a map of London

(2) Feature Extraction

- The typical feature-based model looks for the most extreme examples of a phenomenon and represents the data by these examples
- Some of the important kinds of feature extraction from large-scale data that we shall study are:
  - Frequent Itemsets
  - Similar Items: collaborative filtering

Outline

- 1.1 What is Data Mining?
- 1.2 Statistical Limits on Data Mining
  - 1.2.1 Total Information Awareness
  - 1.2.2 Bonferroni's Principle
- 1.3 Things Useful to Know
- 1.4 Conclusion & Outline of the Book

1.2.1 Total Information Awareness

- Total Information Awareness (TIA), also been called Terrorism Information Awareness
- a massive U.S. data mining project focused on scanning travel, financial and other data from public and private sources
- with the goal of detecting and preventing transnational threats to national security
- one particular technical problem:
  - if you look in your data for too many things at the same time, you will see things that look interesting, but are in fact simply statistical artifacts and have no significance.

1.2.2 Bonferroni's Principle——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

1.2.2 Bonferroni's Principle——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?**

1.2.2 Bonferroni's Principle——Meaningfulness of Analytic Answers

Example:

- The probability of any two people both deciding to visit a hotel on any given day is .0001 （10^{-4}）
- The chance that they will visit the same hotel is this probability divided by 10^5, the number of hotels. The result is 10^{-9}
- The chance that they will visit the same hotel on two different given days is the square of this number, 10^{-18}.

1.2.2 Bonferroni's Principle——Meaningfulness of Analytic Answers

Example:

- Now, we must consider how many events will indicate evil-doing.
- An "event" in this sense is a pair of people and a pair of days, such that the two people were at the same hotel on each of the two days.
  - For large n $\binom{n}{2}$ is about $n^2/2$
  - The number of pairs of people $\binom{10^9}{2} = 5 \times 10^{17}$
  - The number of pairs of days $\binom{1000}{2} = 5 \times 10^5$
  - The number of events $5 \times 10^{17} \times 5 \times 10^5 \times 10^{-18} = 250,000$

1.2.2 Bonferroni's Principle——Meaningfulness of Analytic Answers

Example:

- **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

1.2.2 Bonferroni's Principle——Meaningfulness of Analytic Answers

Bonferroni校正（扩展学习）：

- 如果在同一数据集上同时检验n个独立的假设，那么用于每一假设的统计显著水平，应为仅检验一个假设时的显著水平的1/n。
- 显著性水平是估计总体参数落在某一区间内，可能犯错误的概率，用α表示。统计显著， 是指零假设为真的情况下拒绝零假设所要承担的风险水平，又叫概率水平，或者显著水平。

Outline

- 1.1 What is Data Mining?
- 1.2 Statistical Limits on Data Mining
  - 1.2.1 Total Information Awareness
  - 1.2.2 Bonferroni's Principle
- 1.3 Things Useful to Know
- 1.4 Conclusion & Outline of the Book

1.3 Things Useful to Know

1. The TF.IDF measure of word importance.

2. Hash functions and their use.

3. Secondary storage (disk) and its effect on running time of algorithms.

4. The base e of natural logarithms and identities involving that constant.

5. Power laws.

1.3.1 Importance of Words in Documents

- In several applications of data mining, we shall be faced with the problem of categorizing documents (sequences of words) by their topics
  - Topics： identified by finding the special words that characterize documents about that topic
- classification often starts by looking at documents, and finding the significant words in those documents
- stop words are often removed from documents before any attempt to classify them
- the indicators of the topic are relatively rare words. However, not all rare words are equally useful as indicators

1.3.1 Importance of Words in Documents

- The formal measure of how concentrated into relatively few documents are the occurrences of a given word is called TF.IDF
  - TF，Term Frequency times 词项频率
  - IDF，Inverse Document Frequency 逆文档频率

1.3.1 Importance of Words in Documents

- TF，Term Frequency times 词项频率

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

- 其中，f_{ij}是词项i在文档j中的词项频率，TF_{ij}是f_{ij}归一化的结果

1.3.1 Importance of Words in Documents

- IDF，Inverse Document Frequency 逆文档频率

$$IDF_i = \log_2(N/n_i)$$

- 其中N是集合中的总文档数量，n_i是出现词项i的文档数量
- 词项i在文档j中的得分被定义为：

$$TF_{ij} \times IDF_i$$

- 得分最高词，通常是刻画文档主题的最佳词

1.3.2 Hash Functions

- The hash functions that make hash tables feasible are also essential components in a number of data-mining algorithms, where the hash table takes an unfamiliar form.
- a hash function h takes a hash-key value as an argument and produces a bucket number as a result

1.3.3 Indexes

- An index is a data structure that makes it efficient to retrieve objects given the value of one or more elements of those objects
- The most common situation is one where the objects are records, and the index is on one of the fields of that record
- There are many ways to implement indexes, a hash table is one simple way

1.3.6 Power Laws

- The Matthew Effect
  - Often, the existence of power laws with values of the exponent higher than 1 are explained by the Matthew effect.
  - Many phenomena exhibit this behavior, where getting a high value of some property causes that very property to increase.

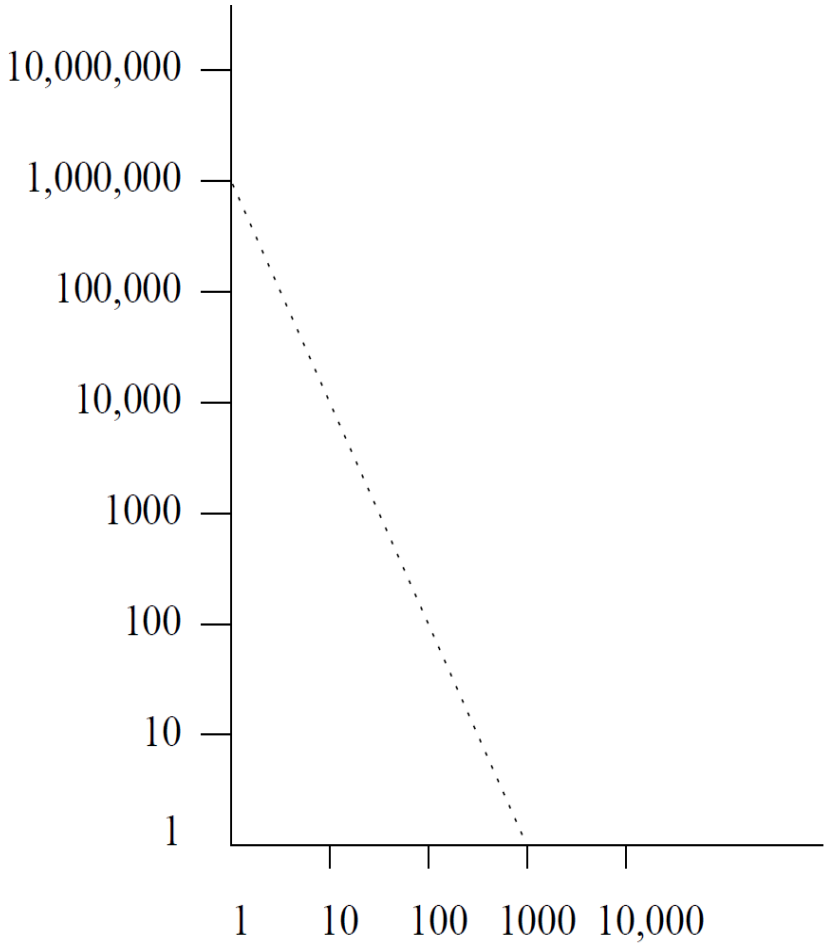
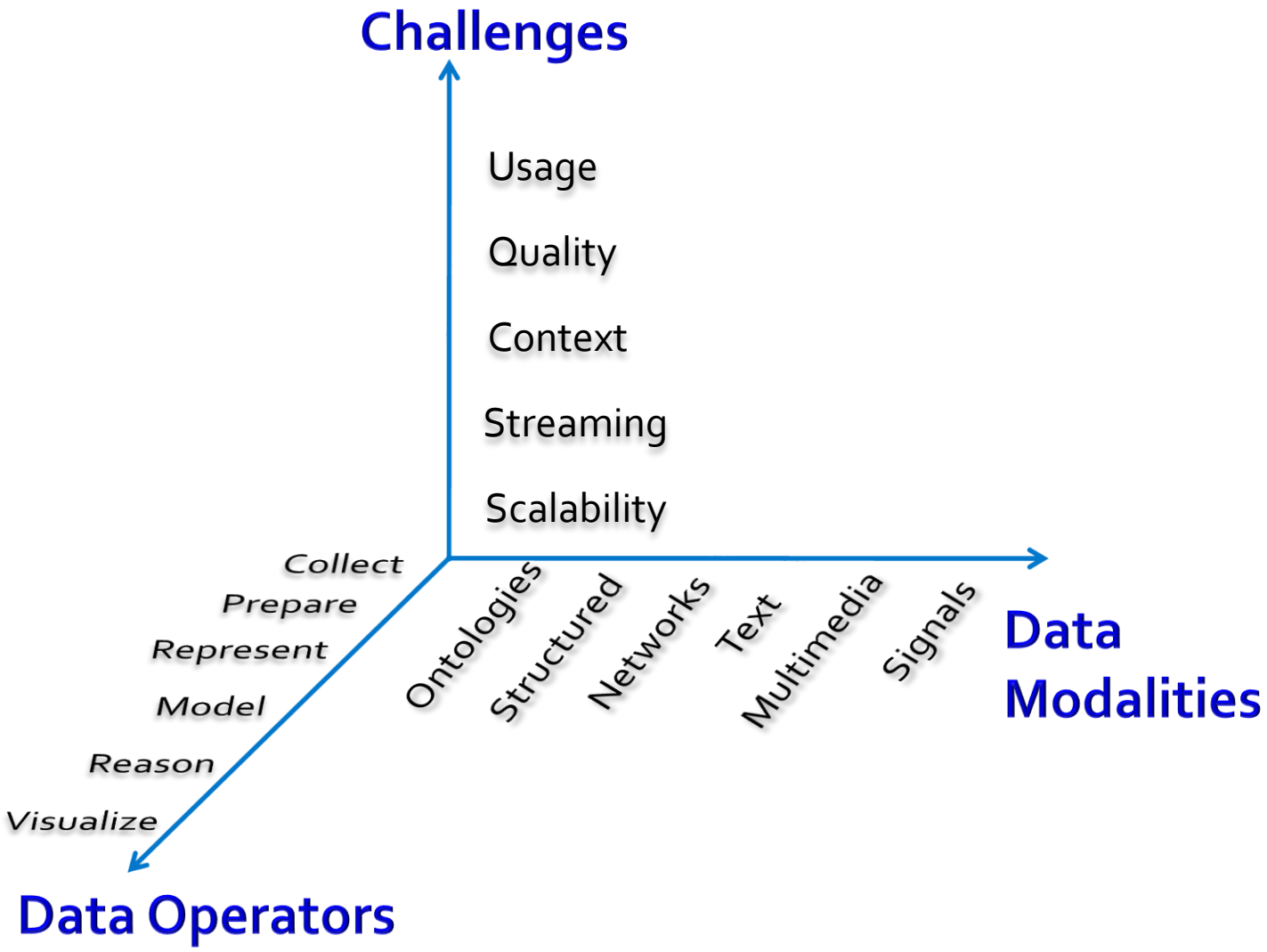


Figure 1.3: A power law with a slope of −2

Outline

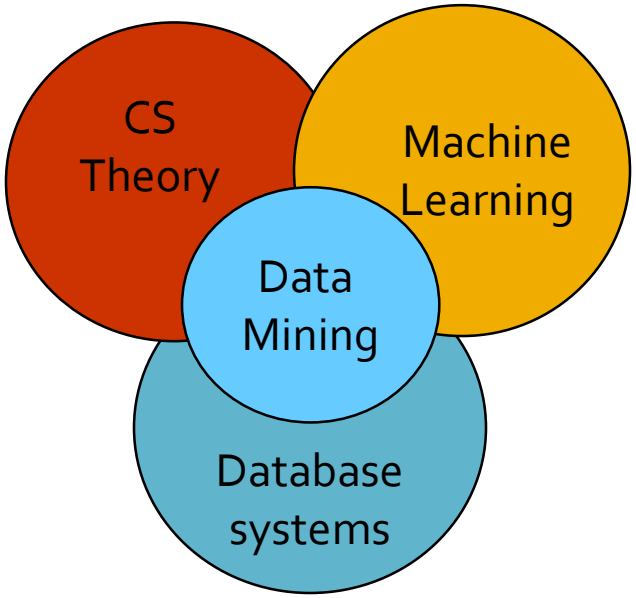
- 1.1 What is Data Mining?
- 1.2 Statistical Limits on Data Mining
  - 1.2.1 Total Information Awareness
  - 1.2.2 Bonferroni's Principle
- 1.3 Things Useful to Know
- 1.4 Conclusion & Outline of the Book

What matters when dealing with data?



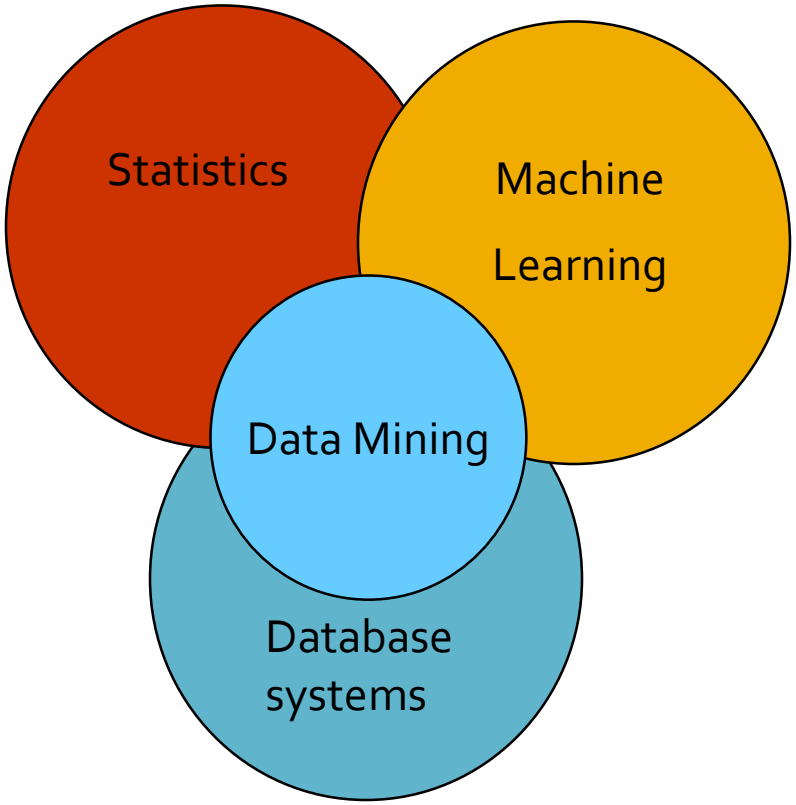
Data Mining: Cultures

- **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**



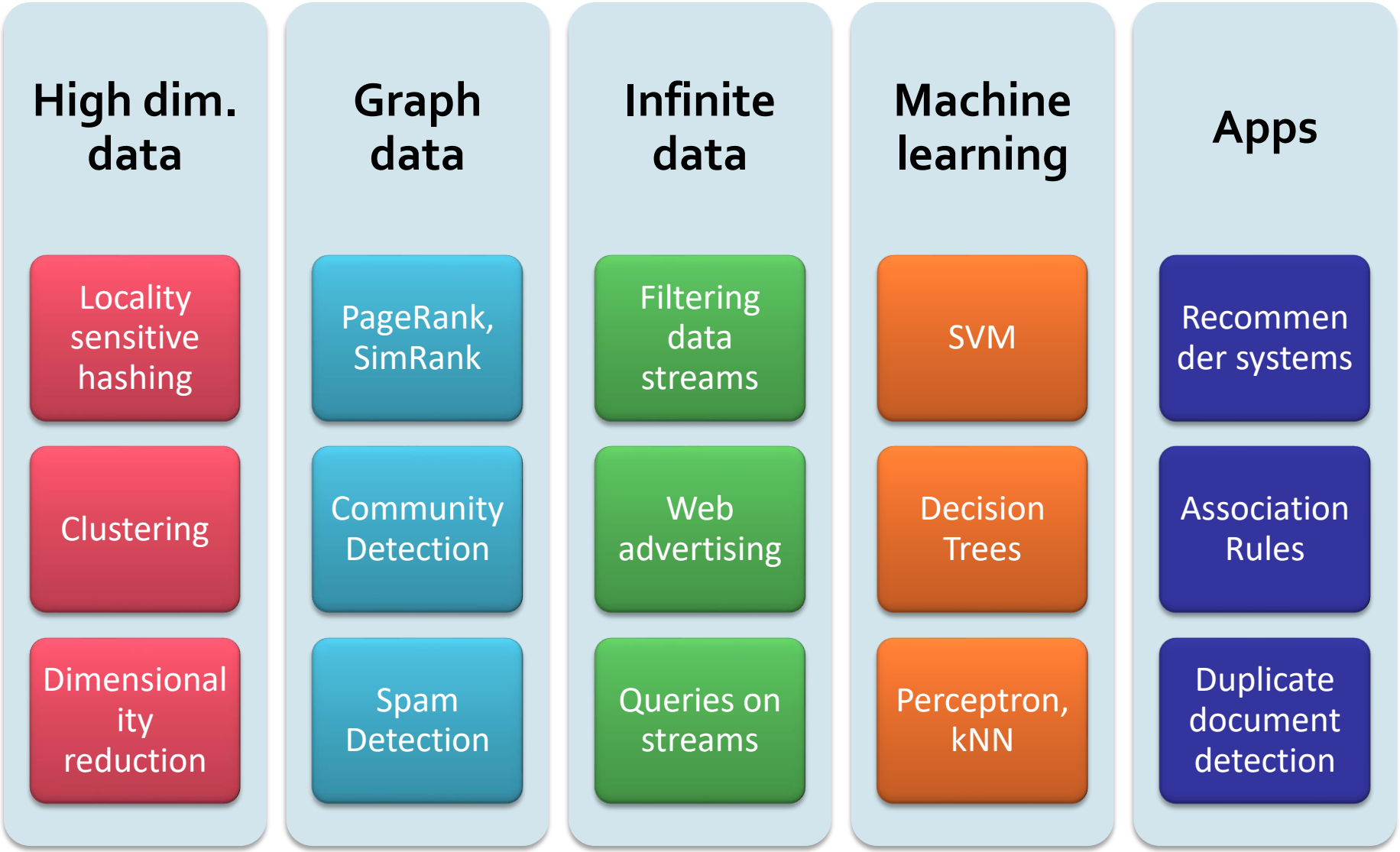
What will we learn?

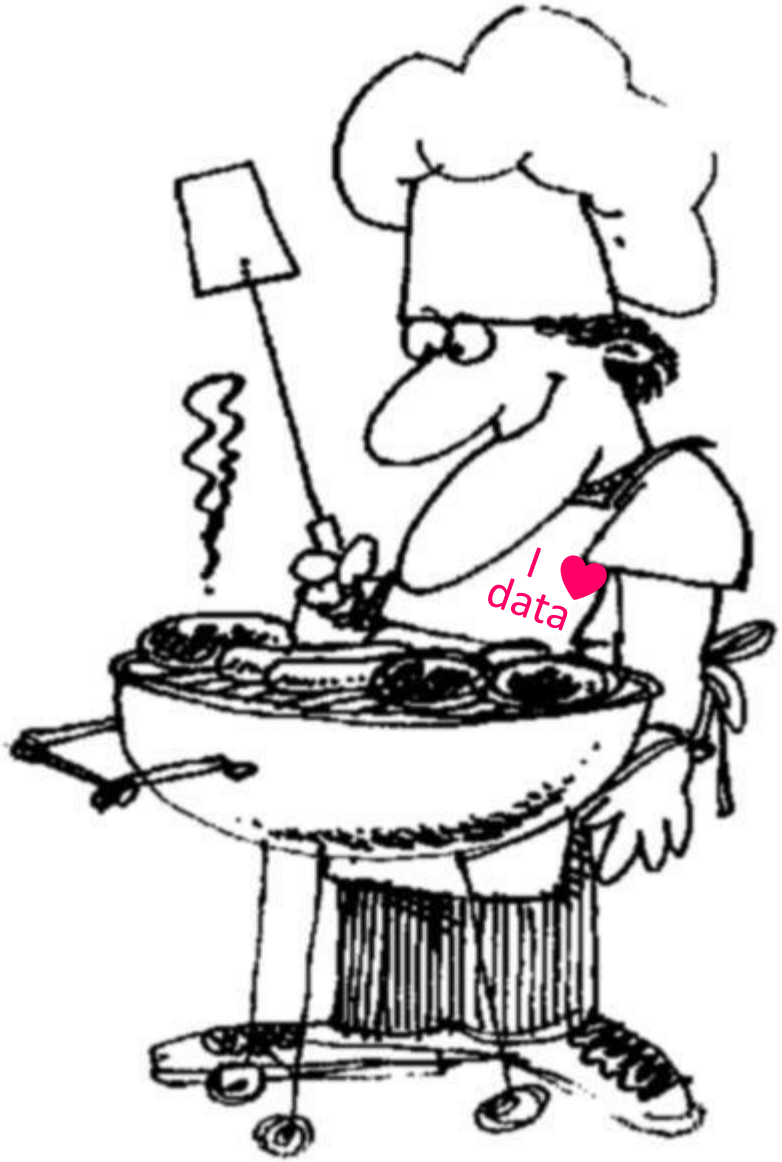
- **We will learn to mine different types of data:**
  - Data is high dimensional
  - Data is a graph
  - Data is infinite/never-ending
  - Data is labeled
- **We will learn to use different models of computation:**
  - MapReduce
  - Streams and online algorithms
  - Single machine in-memory

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?