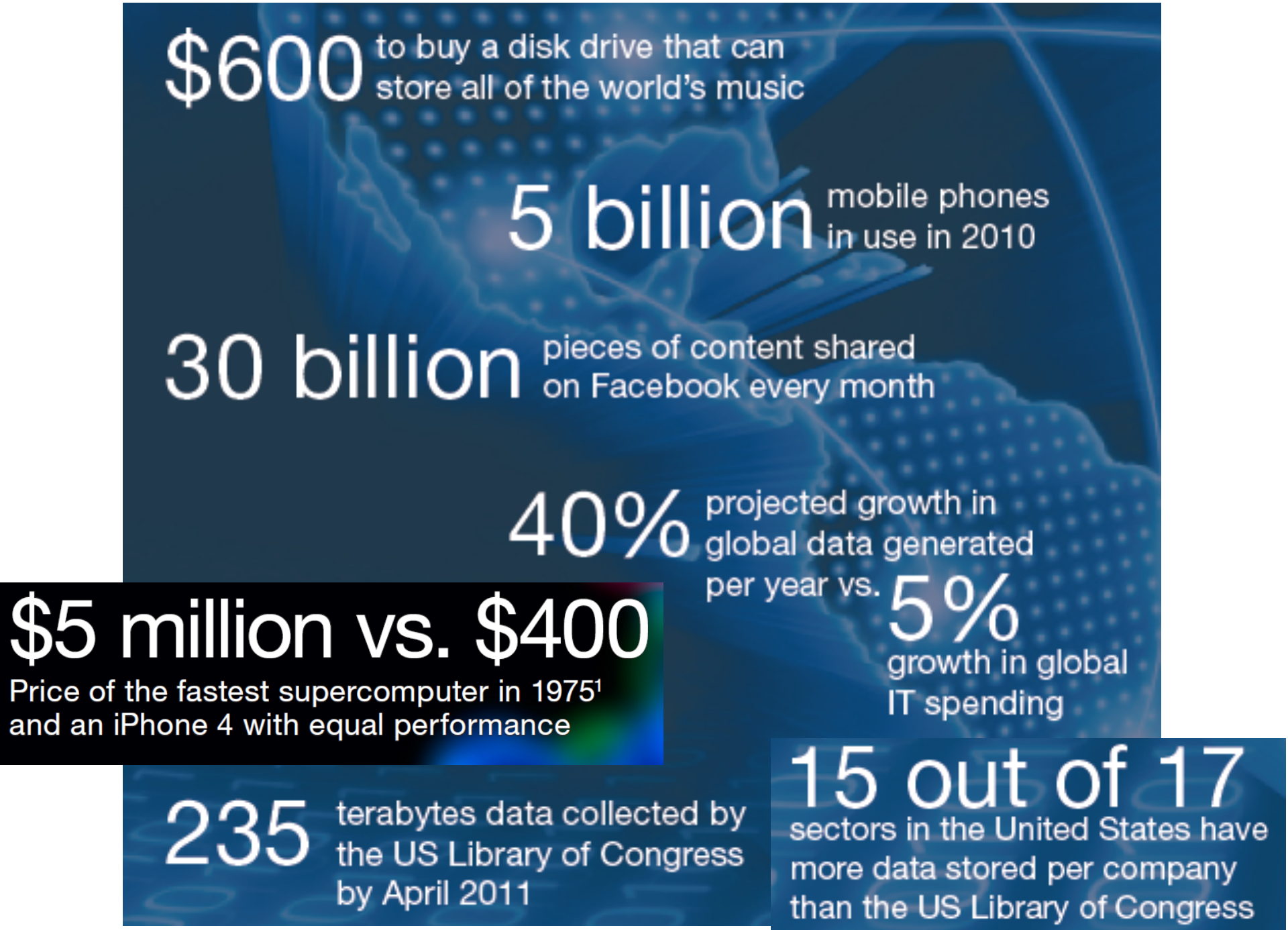
Big Data Analytics: Course Introduction

Jarek Szlichta

http://data.science.uoit.ca/

Acknowledgments: Mining of Massive Datasets (Jure Leskovec, Anand Rajaraman, Jeff Ullman)

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

The Era of Big Data

- Unprecedented growth in data being generated and its potential uses/value
  - Tweets, social networks (statuses, check-ins, shared content), blogs, click streams, various logs, …
  - *Facebook: > 1.2B active users, > 1B messages/day*
  - *Twitter: > 140M active users, > 500M tweets/day*
- Everyone is interested
  - Trade press and popular press: *"Big Data!"*
  - Enterprises, Web companies, online businesses, governments and public health researchers
  - Untapped value and countless new opportunities to understand, optimize, and/or compete

Facebook Country..

1. China (1.339 billion)
2. India (1.218 billion)
3. Facebook (1.2 billion)

Data contains value and knowledge

Data Mining

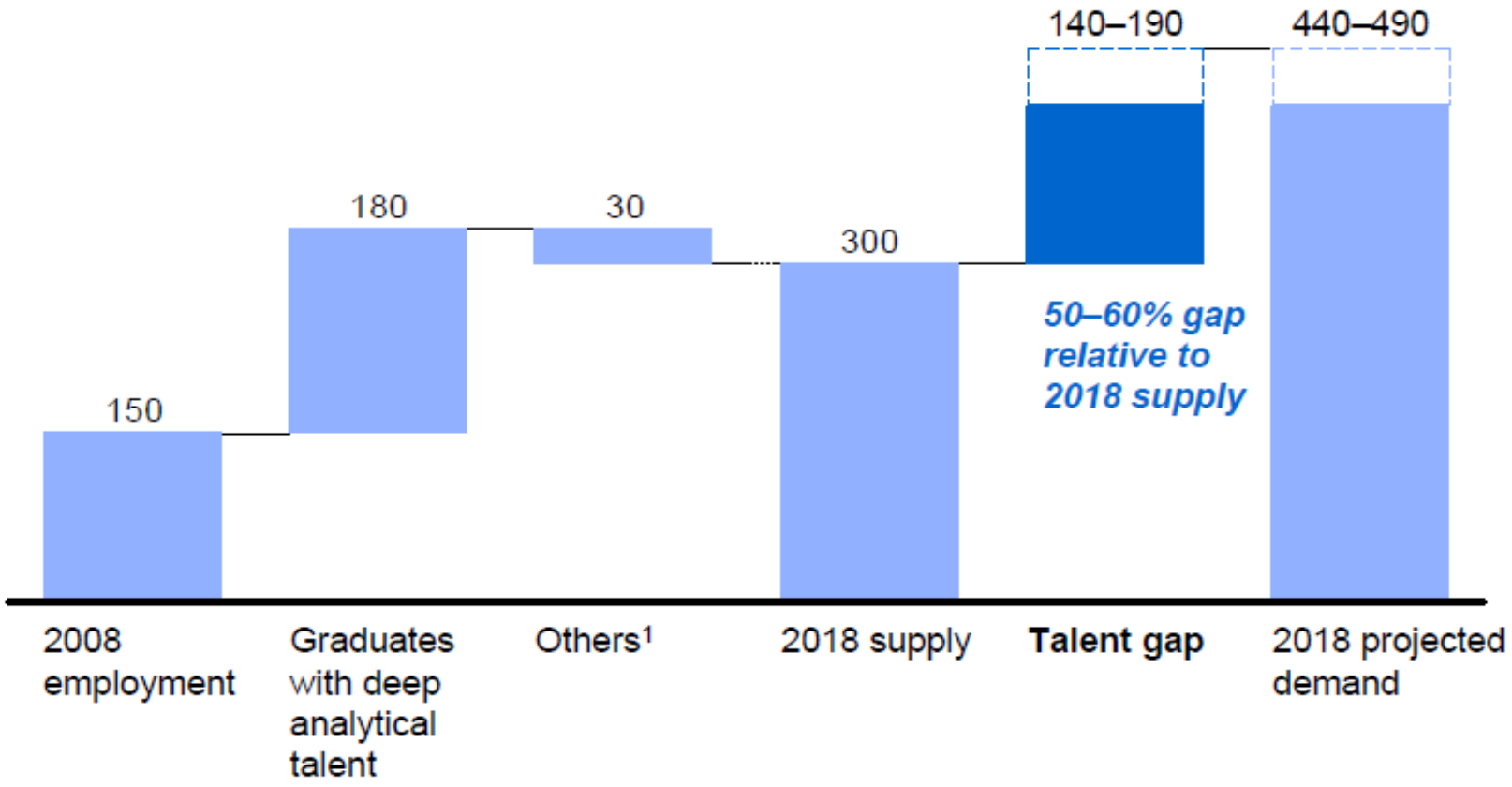
- **But to extract the knowledge data needs to be**
  - **Stored**
  - **Managed**
  - **And ANALYZED ← this class**

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

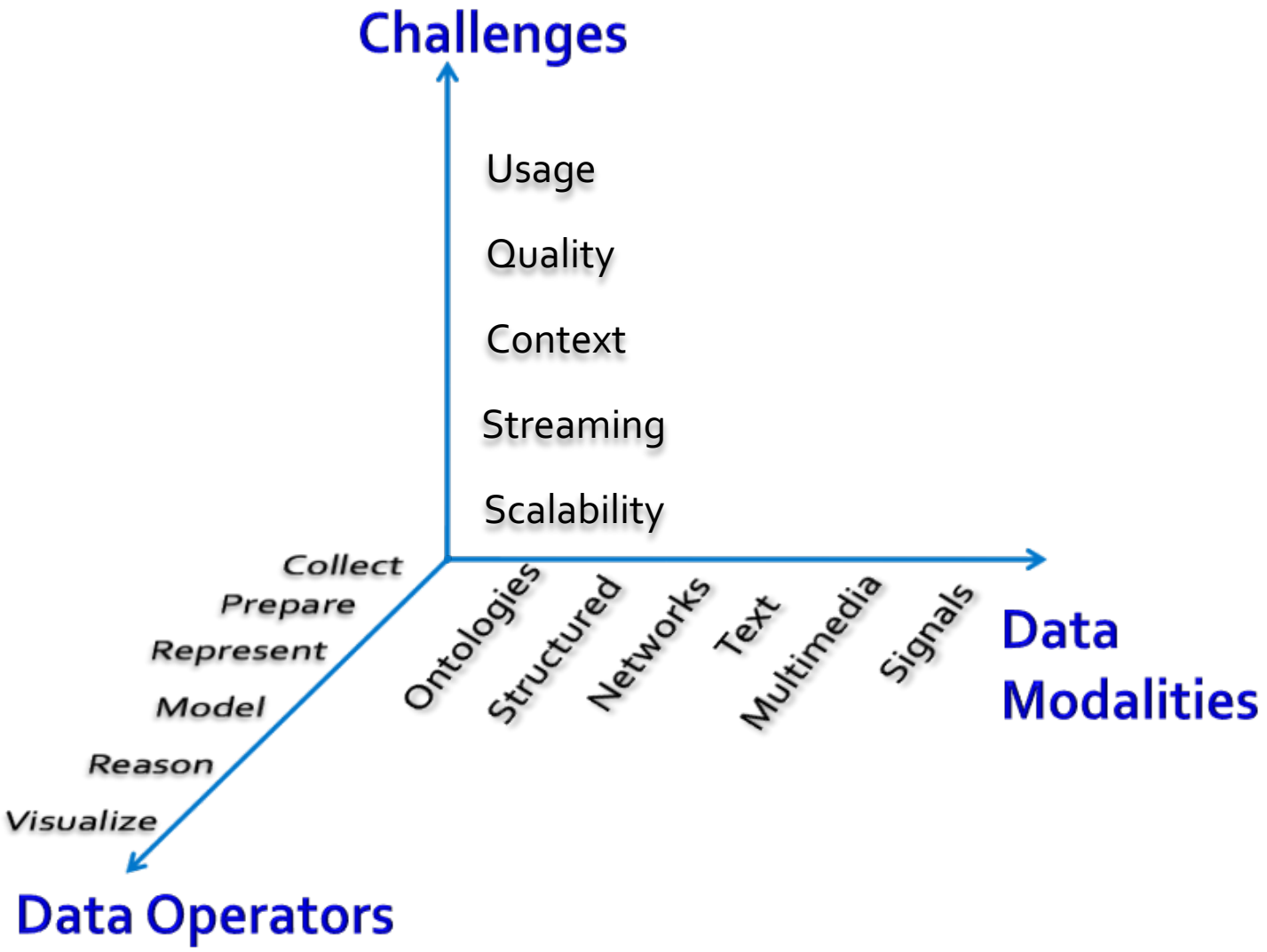
- **A risk with "Data mining" is that an analyst can "discover" patterns that are meaningless**
- Statisticians call it **Bonferroni's principle**:

Meaningfulness of Analytic Answers

Example:

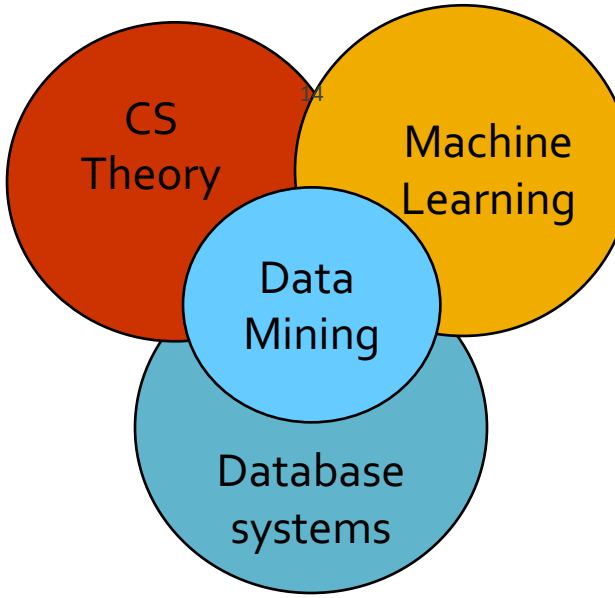
- We want to find (unrelated) people who **at least twice have stayed at the same hotel on the same day**
  - 10^9 people being tracked
  - 1,000 days
  - Each person stays in a hotel 1% of time (1 day out of 100)
  - Hotels hold 100 people (so 10^5 hotels)
  - **If everyone behaves randomly (i.e., no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



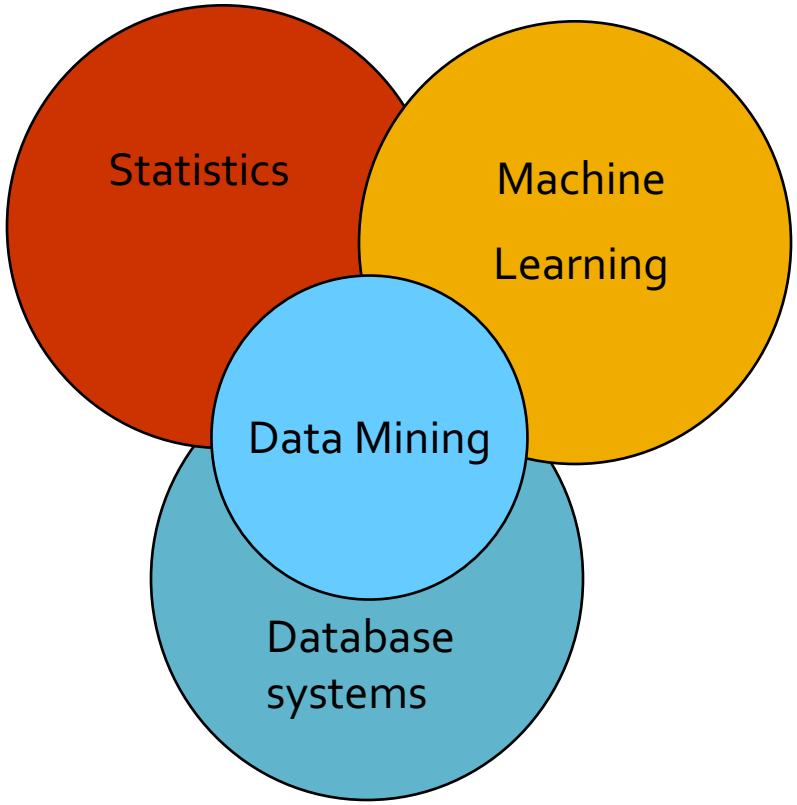
Data Mining: Cultures

- **Data mining overlaps with:**
  - **Databases:** Large-scale data, simple queries
  - **Machine learning:** Small data, Complex models
  - **CS Theory:** 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 all!**



This Class: CSCI 4030

- **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**



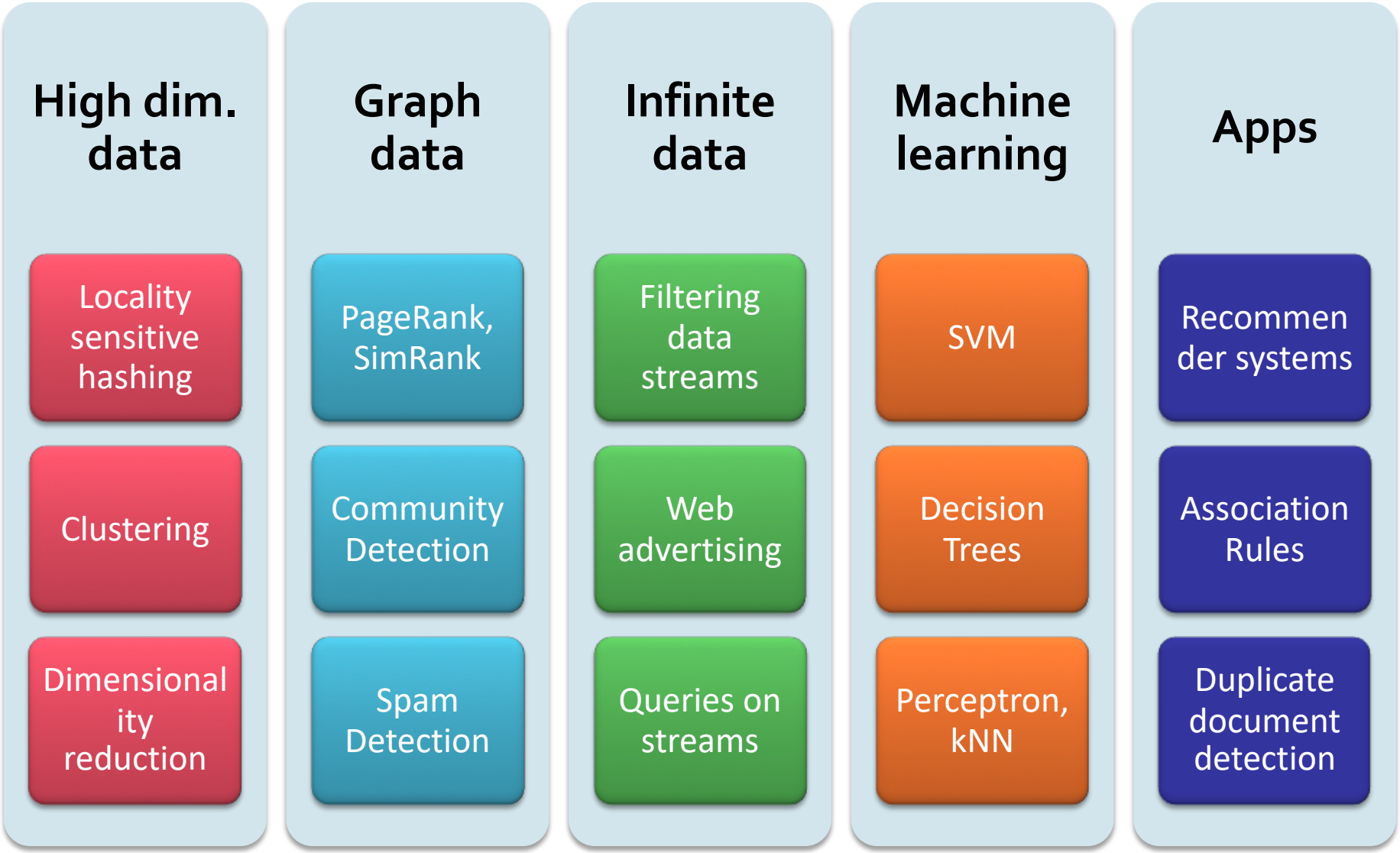
What will we learn?

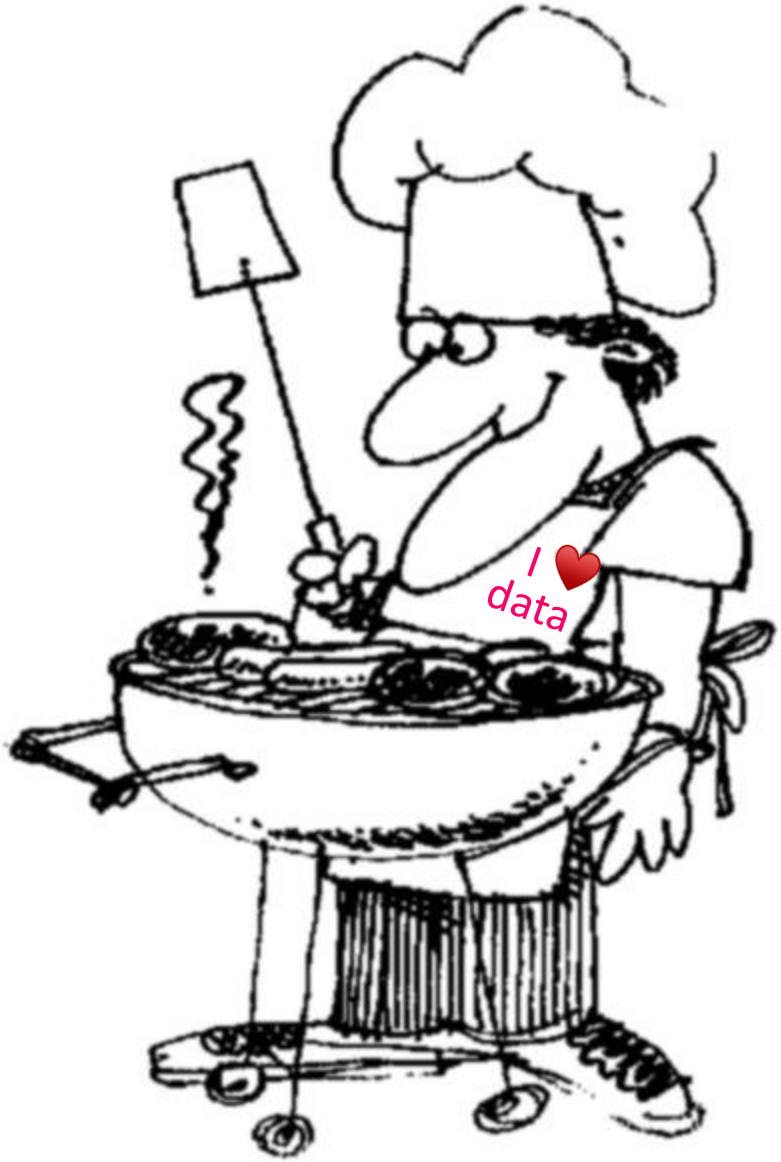
- **We will learn to mine different types of data:**
  - Data is high dimensional
  - Data is a graph
  - Data is infinite/never-ending
  - Data is labeled
- **We will learn to use different models of computation:**
  - Streams and online algorithms
  - …
  - Single machine in-memory

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)
  - Optimization (stochastic gradient descent)
  - Dynamic programming (frequent itemsets)
  - Hashing (LSH, Bloom filters)

How It All Fits Together





How do you want that data?

About the Course

CSCI4030 Course Staff

- **TAs:**
  - **We have great TAs!**
    - **Spencer Bryson:** spencer.bryson@ontariotechu.net
    - **Bahare Askari:** Bahare.AskariFiroozjayi@ontariotechu.net
- **Office hours:**
  - See course website for office hours

Course Logistics

- **Course website:**
  http://data.science.uoit.ca
  - Lecture slides (at least 30min before the lecture)
- **Readings:** Book **Mining of Massive Datasets**
  Authors: Jure Leskovec, A. Rajaraman and J.Ullman
  Free PDF: See Blackboard

Logistics: Communication

- **Blackboard:**
  - https://uoit.blackboard.com
    - Posting Labs and Project / Midterm / Final / ..
- **Slack for course communication**
- **For e-mailing use:**
  - jarek@uoit.ca

- **We will post course announcements to the website (make sure you check it regularly)**

Auditors are welcome to sit-in & audit the class

Work for the Course

- **Labs & Project:** 30% (10% + 20%)
  - Research project on Big Data (midterm report + final report)
  - Labs will start in the week of 20th of Jan!
- **Midterm I:** 20 %
- **Participation & Presentation:** 10% (5% + 5%)
- **Final Midterm:** 40%

- It's going to be fun and hard work. ☺
- We have big-data graduate positions open!

Equipment

Prerequisites

- **STAT 2010 - STATS & PROB FOR PHYSICAL SCI.**
  - Basic statistics and probability
- **CSCI 3030 - DATABASE SYSTEMS & CONCEPTS**
  - Fundamentals of database systems
- **We provide background, but the class makes an assumption about prerequisites**

Actions

- **Read Chapter 1: Data Mining!**
  - **Includes some prerequisites!**