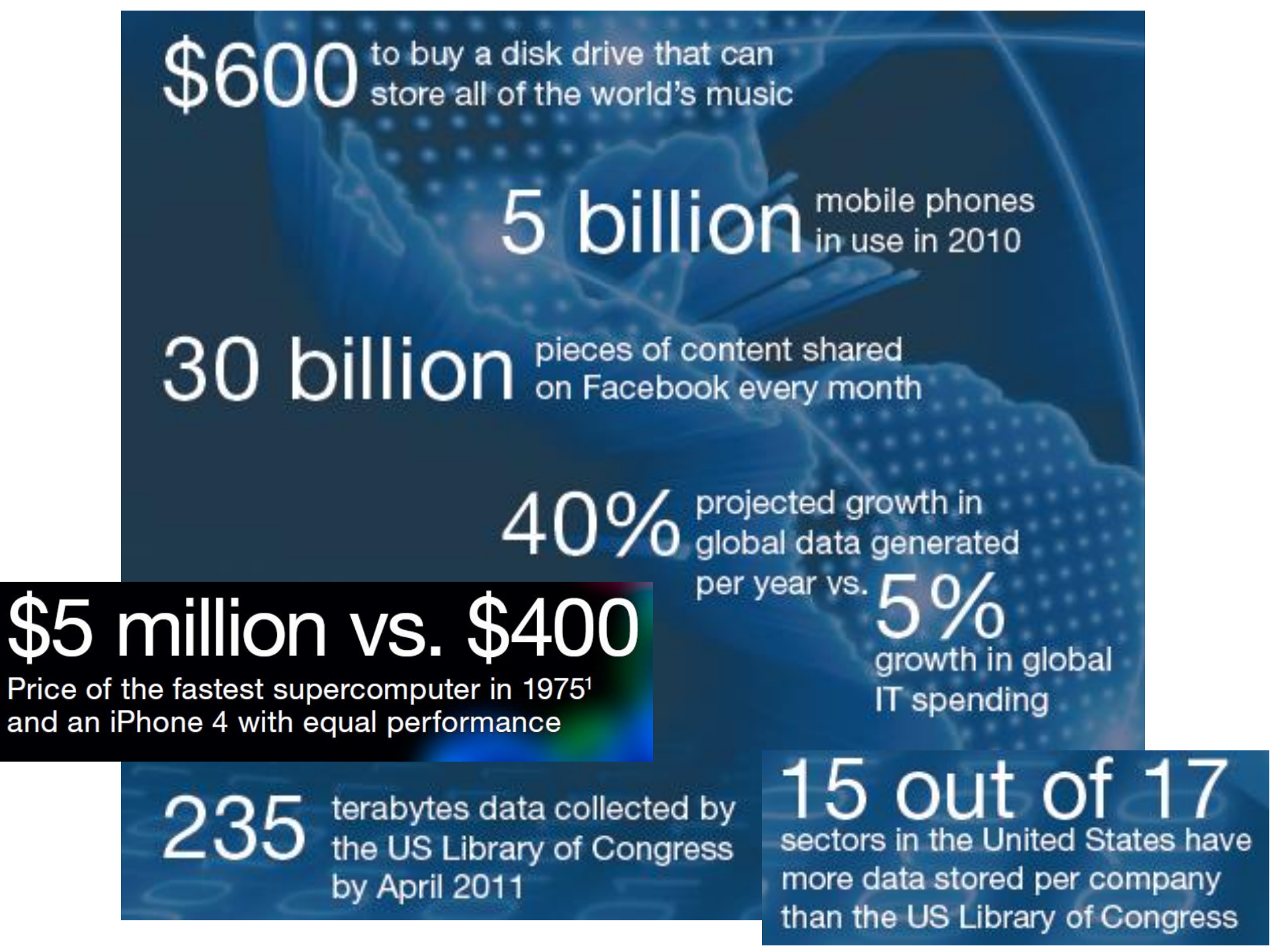
EECS6414: Data Analytics & Visualization

Thanks to Jure Leskovec, Anand Rajaraman, Jeff Ullman Stanford University - **http://www.mmds.org**

What is Data Analytics? Knowledge discovery from data

what is data analysis?

a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making


$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

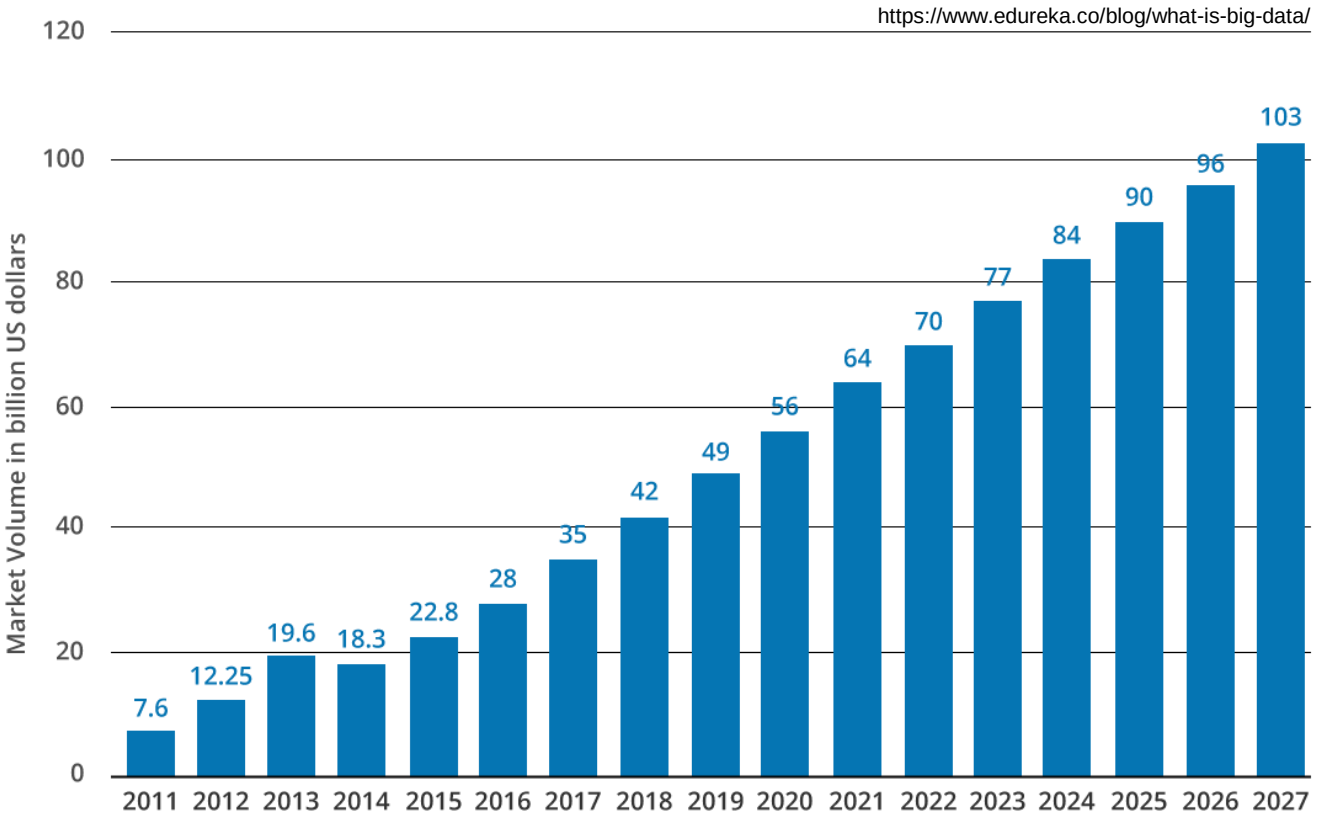
Data contains value and knowledge

Data Analytics

- But to extract the knowledge data needs to be
  - Stored
  - Managed
  - Analyzed ← emphasis on this class
  - Visualized ← emphasis on this class

Data Analytics ≈ Data Mining ≈ Big Data ≈ Predictive Analytics ≈ Data Science

Demand for Big Data Skills



Growing market revenue of Big Data in billion U.S. dollars from the year 2011 to 2027

Objective of Data Analysis

- **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 Analysis Tasks

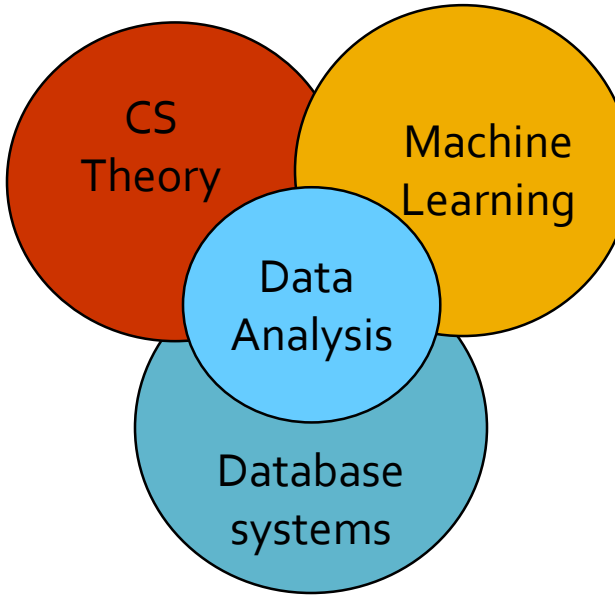
- **Descriptive methods**
  - Find human-interpretable patterns that describe the data
    - **Example:** Clustering
- **Predictive methods**
  - Use some variables to predict unknown or future values of other variables
    - **Example:** Recommender systems

Meaningfulness of Analytic Answers

- **A risk with "Data analysis" 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

Data Analytics: Cultures

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



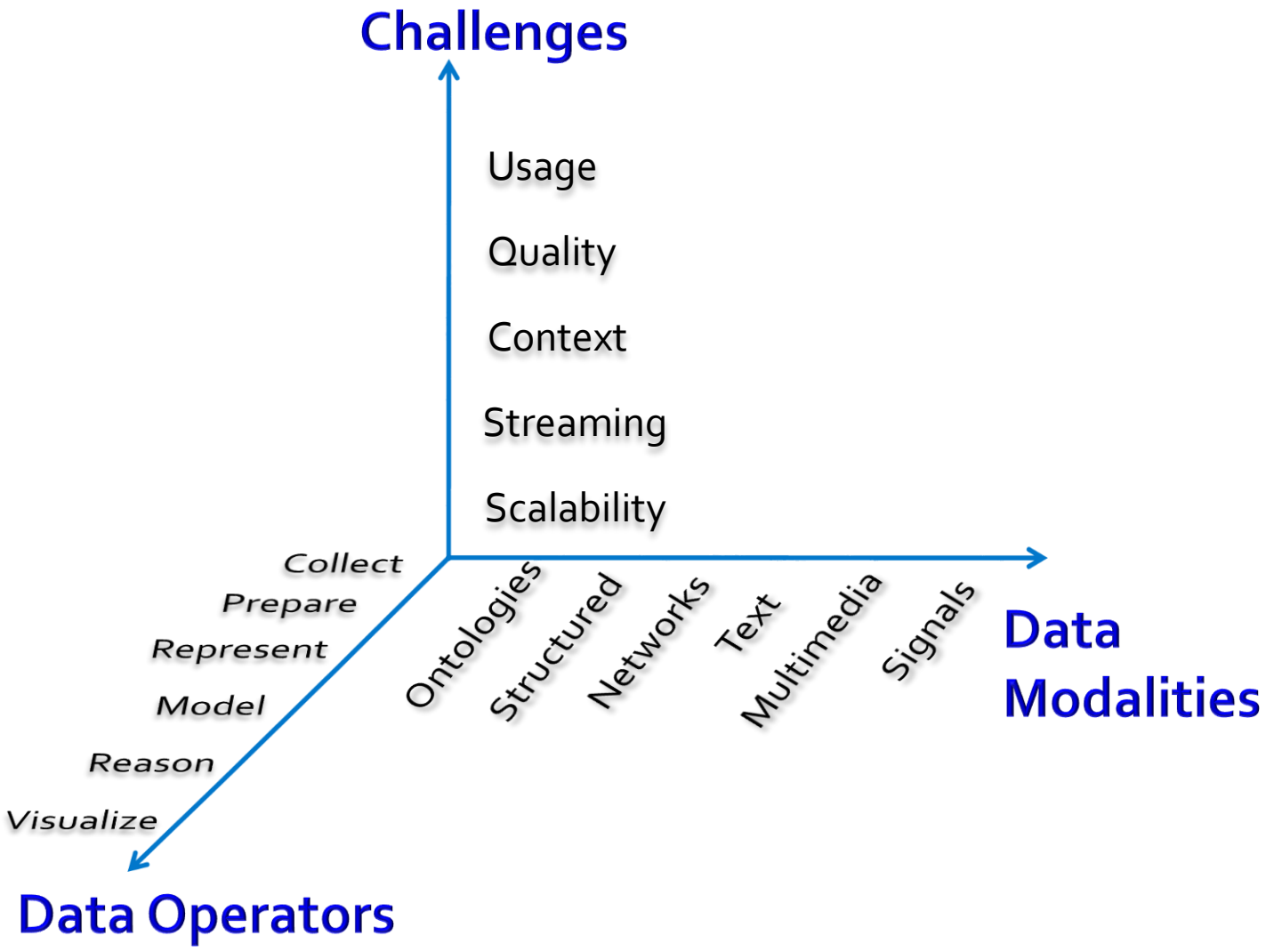
This Class: EECS6414

- This class stresses more on
  - Computing architectures
  - Algorithms for handling large data
  - Visualization

What will we discuss?

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

What matters when dealing with data?



How do you want that data?

EECS6414
About the Course

Logistics: Communication

- **Website**
  - http://www.eecs.yorku.ca/~papaggel/courses/eecs6414/
- **Piazza Q&A website:**
  - Available from the website
    https://piazza.com/yorku.ca/winter2019/eecs6414
  - You need to register with your ***yorku.ca*** email

    Please participate and help each other!
- **e-mail for personal issues:**
  - papaggel@eecs.yorku.ca

Project-focused course

What Does it Mean?

No final exam, no assignments

But, you need to:

identify a problem
find data
prepare data for analysis
create visualizations for data exploration
uncover insights
communicate critical findings
create data-driven solutions

+ **team-work (up to 3 people)**

Elements of a DAV project

Need for data collection
Need for data storage
Need for data analysis
Need for data visualization

…but, more of an iterative process than a sequence

Open Data Initiatives

1,028 featured datasets

www.kaggle.com

Sort by

Featured All

Search datasets

714

IMDB 5000 Movie Dataset

5000+ movie data scraped from IMDB website

chuansun76 · updated a year ago · film

656

European Soccer Database

25k+ matches, players & teams attributes for European Professional Football

Hugo Mathien · updated 10 months ago · association football, europe

617

Credit Card Fraud Detection

Anonymized credit card transactions labeled as fraudulent or genuine

Andrea · updated 10 months ago · crime, finance

539

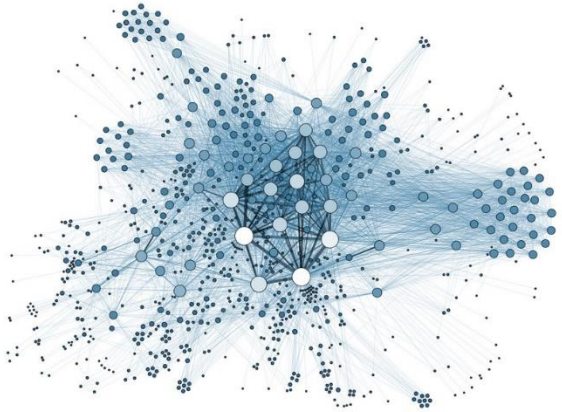
Human Resources Analytics

Why are our best and most experienced employees leaving prematurely?

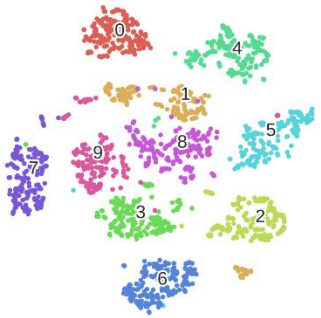
ludoben · updated 9 months ago · employment

What Type of Data?

Text Data



Network Data



Multivariate Data

(Tentative) Course Evaluation

Milestone	Weight
Project proposal	10%
Project midterm report	20%
Project midterm in-class presentation	10%
Project final report	40%
Project final in-class presentation	20%

+ project report in research paper format
+ demo (if applicable)

… a number of lectures

Topics Covered

Data Analysis

Finding Similar Items, Frequent Itemsets, Mining Data Streams, Clustering, Dimensionality Reduction

Network Analysis

Link Analysis, Mining Social-Network Graphs, Recommendation Algorithms

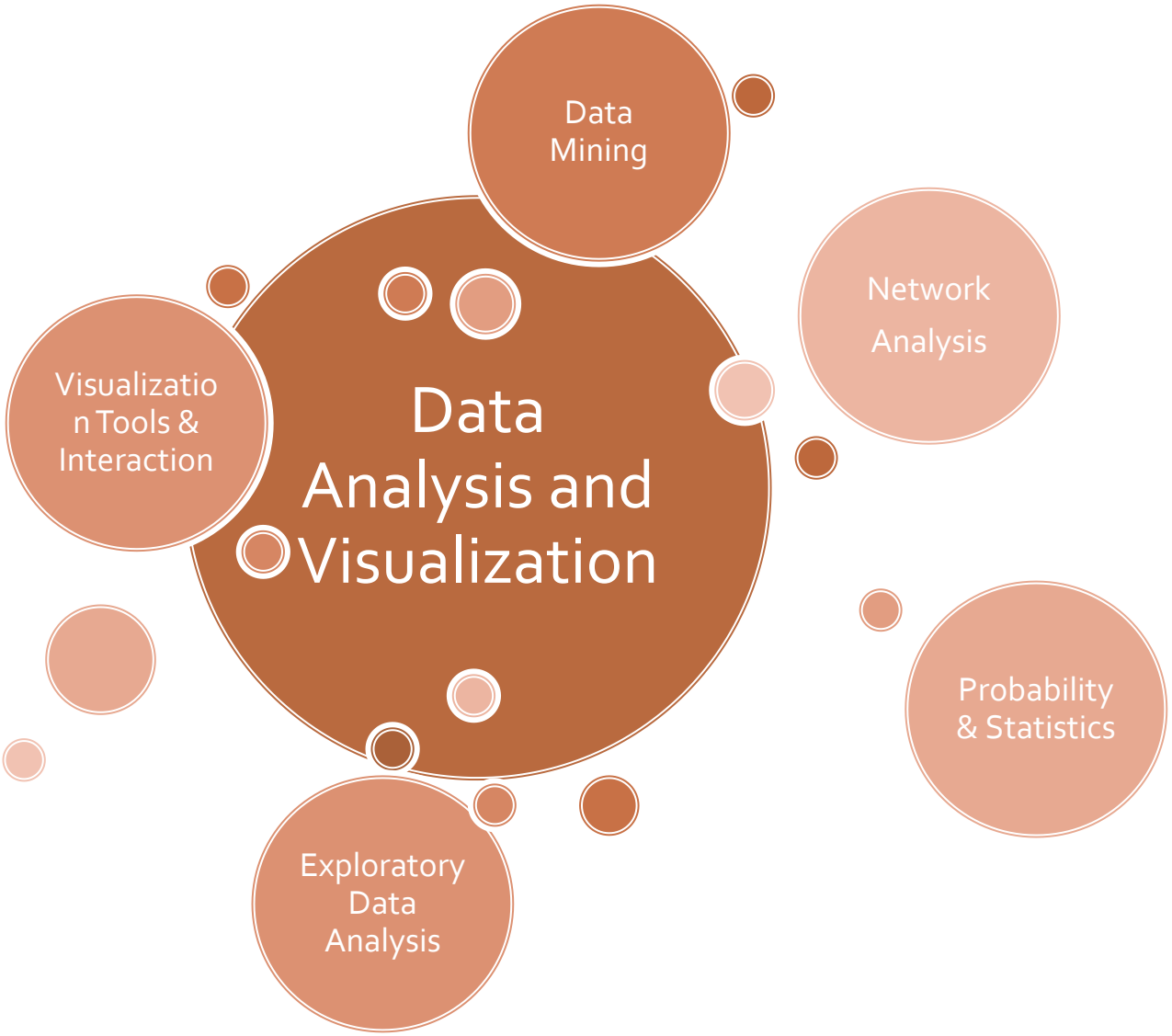
Visualization

Value of Visualization, Exploratory Data Analysis, Multidimensional Data, Networks

Data Analysis and Visualization Tools

OpenRefine, Apache Hadoop/MapReduce, Apache Spark, Google BigTable, Tableau, D3, Google Embedding Projector

Course Intellectual Content



Who Should Attend?

Current interest in DAV

You are currently working on an interesting DAV project

Continuous interest in DAV

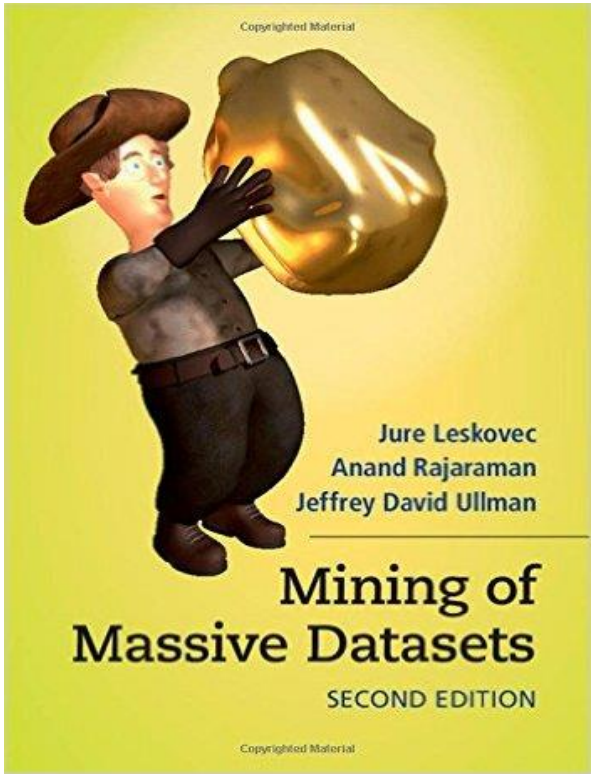
You worked on an interesting DAV project before (BSc thesis, MSc thesis, etc.) and would like to further expand it

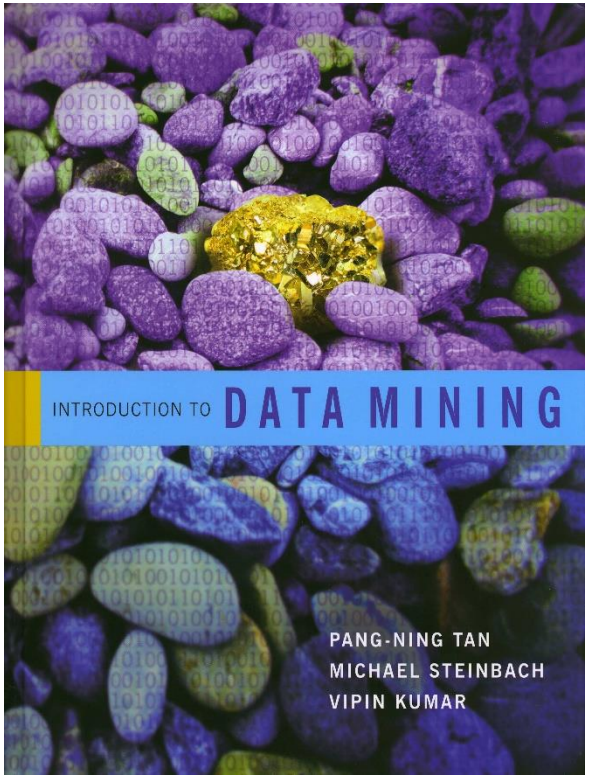
Potential interest in DAV

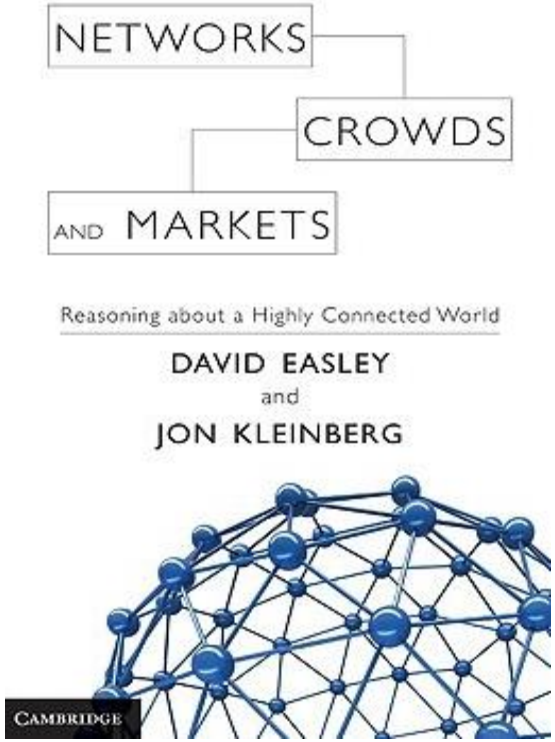
You are interested to work on a DAV project and looking for inspirations

"Suggested" Textbooks 1/2

Data Analytics



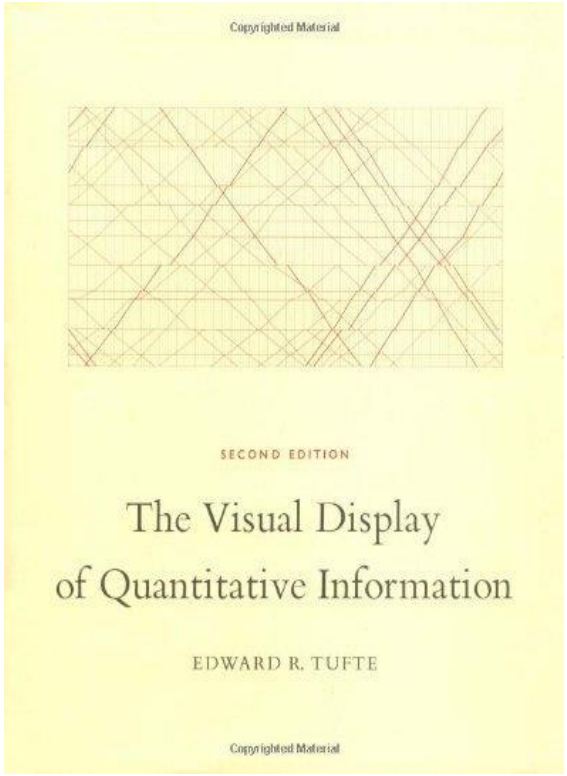


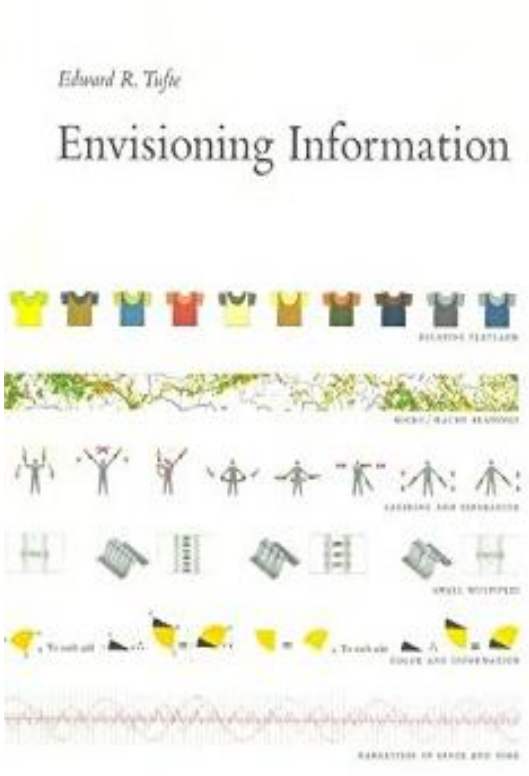


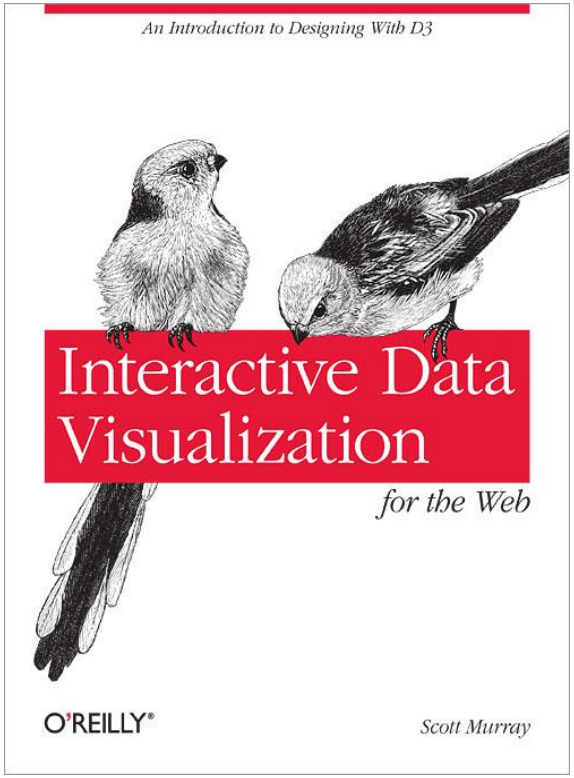
+ tools for data analytics

"Suggested" Textbooks 2/2

Data Visualization







+ tools for visualization of high-dimensional data

Logistics

Item	Comment
Classes	Mon @ 16:00-19:00
Classroom	Calumet College 109 (CC109)
Course group	3
Credits	3
Website	http://www.eecs.yorku.ca/~papaggel/courses/eecs6414/
Office hour	Drop anytime by my office (LAS3050) or by appointment

Background

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

It's going to be fun and hard work. ☺

Welcome!

Contact:
Manos Papagelis, LAS 3050
papaggel@eecs.yorku.ca
www.eecs.yorku.ca/~papaggel/