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

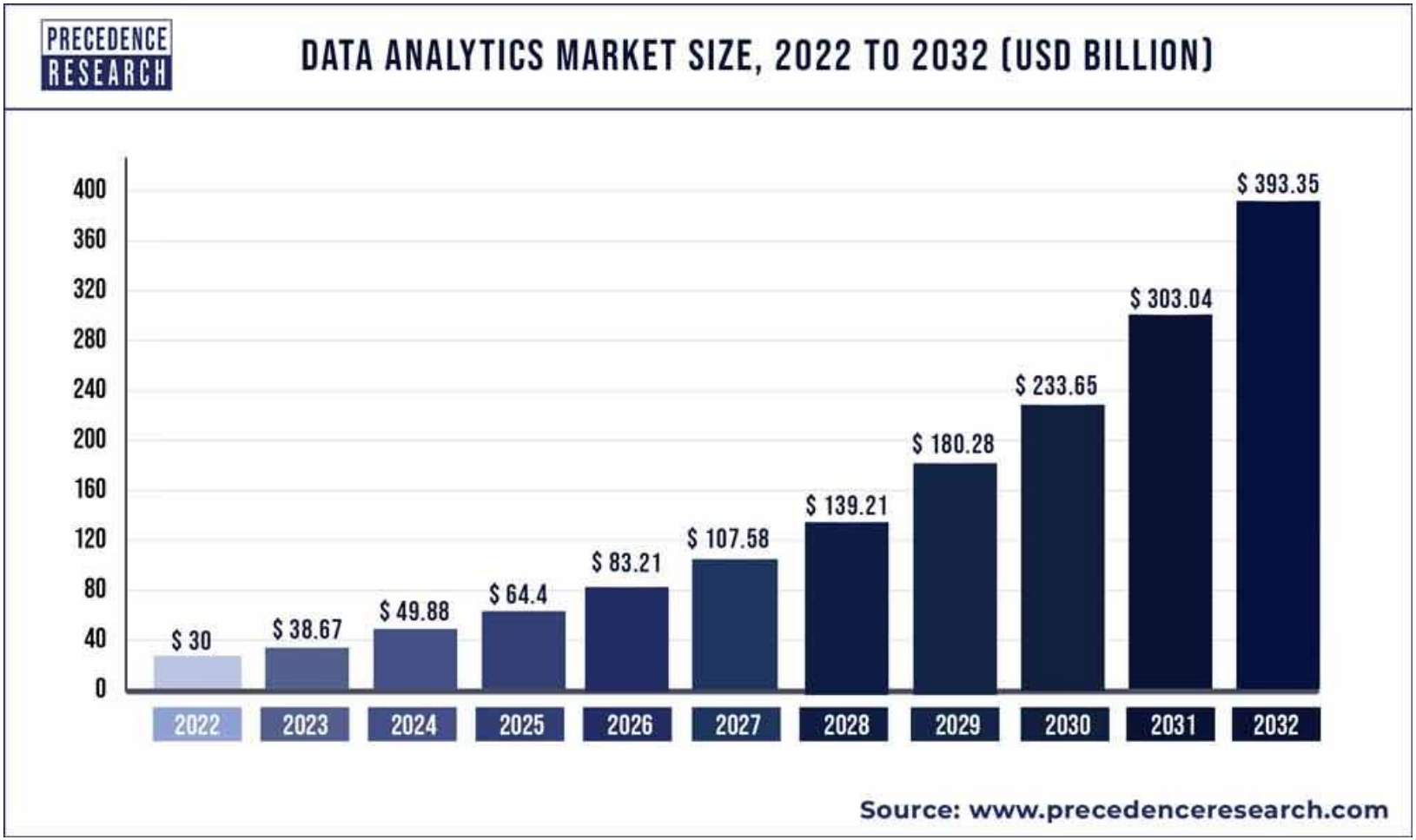
Data contains value and knowledge

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 Data Analytics Skills



Growing Data Analytics Market (2022 to 2032, USD billion)

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

Types of Data Analysis

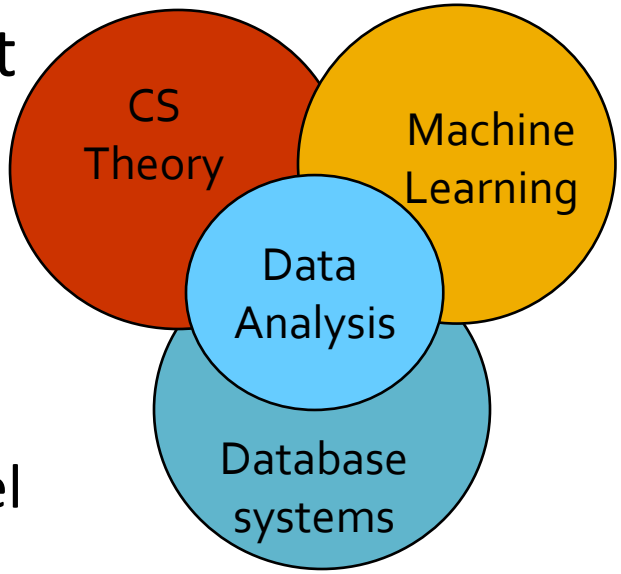
- **Descriptive methods**
  - Find human-interpretable patterns that describe the data
    - **Example:** Clustering (e.g., find communities of interest)
- **Predictive methods**
  - Use some variables to predict unknown or future values of other variables
    - **Example:** Recommendations (e.g., suggest new friends in a social network)

Meaningfulness of Data Analytics

- **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:**
  - **Database Systems:** Large data, simple queries
  - **Machine learning:** Large 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



This Class: EECS6414

- This class stresses more on
  - Data analysis of network data (graph model)
  - (Less on) Data analysis of high-dimensional data
  - Data visualization principles & examples

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/winter2024/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

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

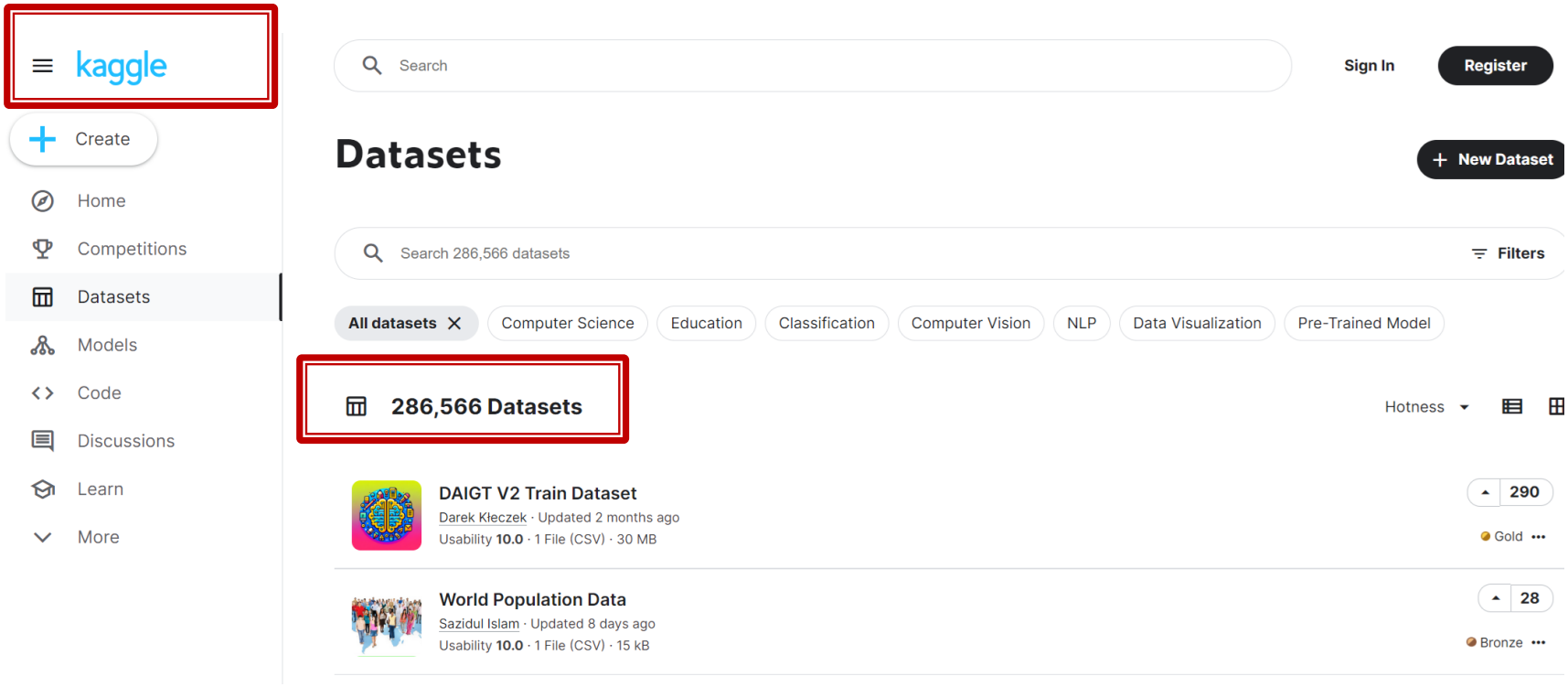
+ **team-work (up to 3 people)**

Elements of a DAV project

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

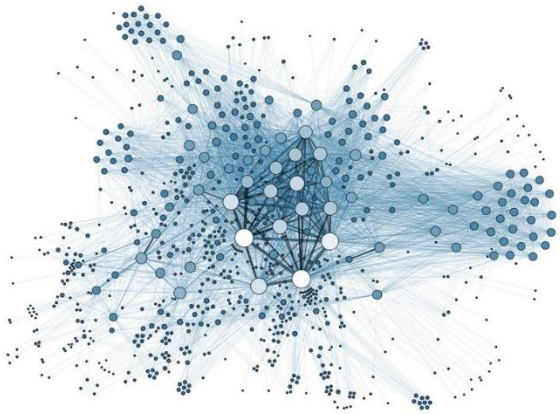
…but, more of an iterative process than a sequence

Open Data Initiatives

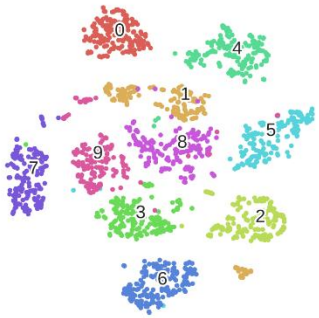


What Type of Data?

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)

Topics Covered

Network Analysis (~8 lectures)

Introduction to networks, basic graph theory, network measurements, network models, link analysis & link prediction, community detection, cascading behavior in networks, epidemic spreading models (if time allows)

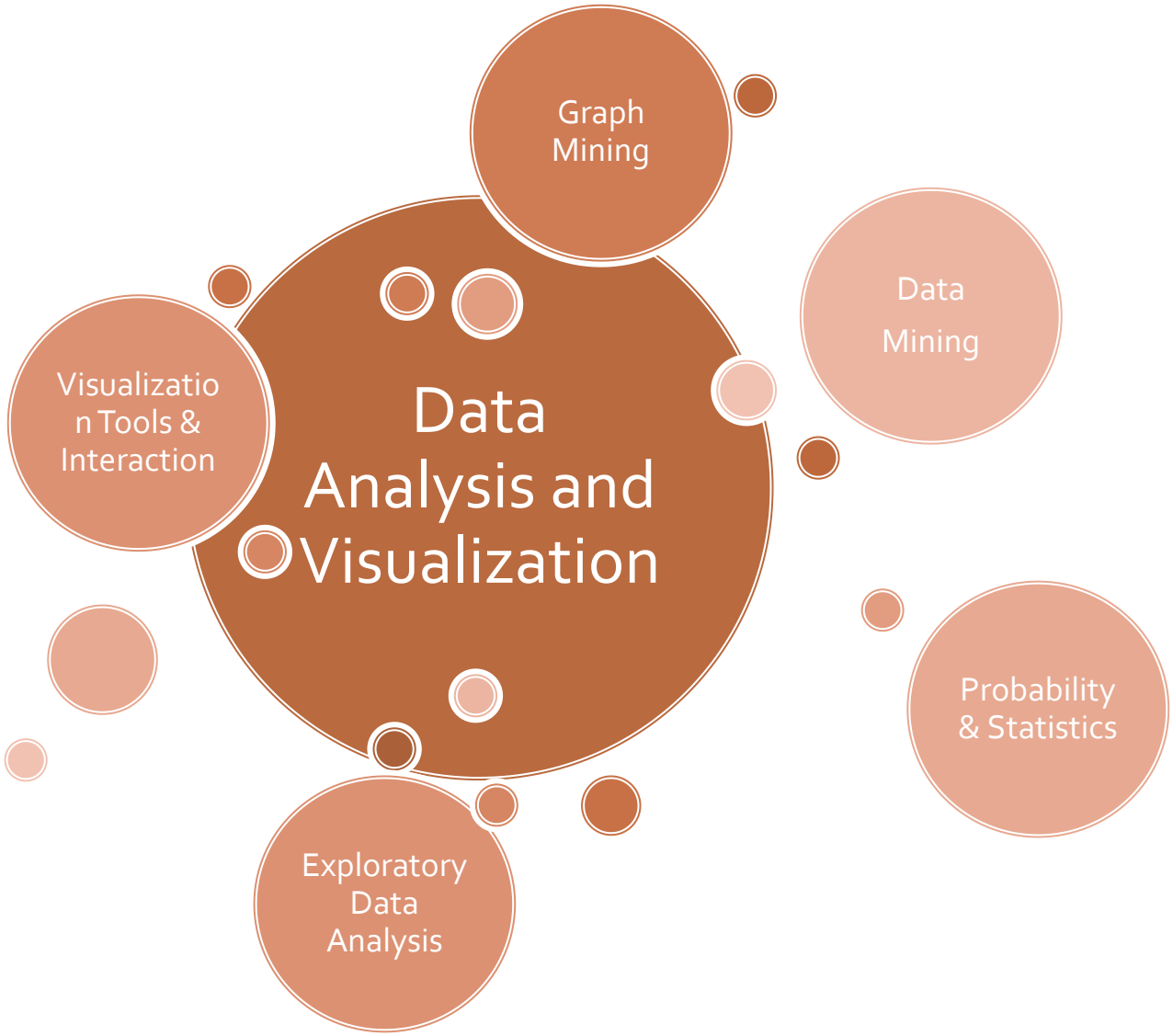
Data Visualization (~2 lectures)

Value of visualization, visual variables, cognition and perception, colors, pre-attentive vs attentive processing, visual metaphors, taxonomy of visualization, visualizations of qualitative and quantitative data

Team Project Presentations (~2 lectures)

Teams present their projects in class, share knowledge, get feedback.

Course Intellectual Content



Who Should Attend?

Current interest in DAV

You are currently working on an interesting DAV project

Continuous interest in DAV

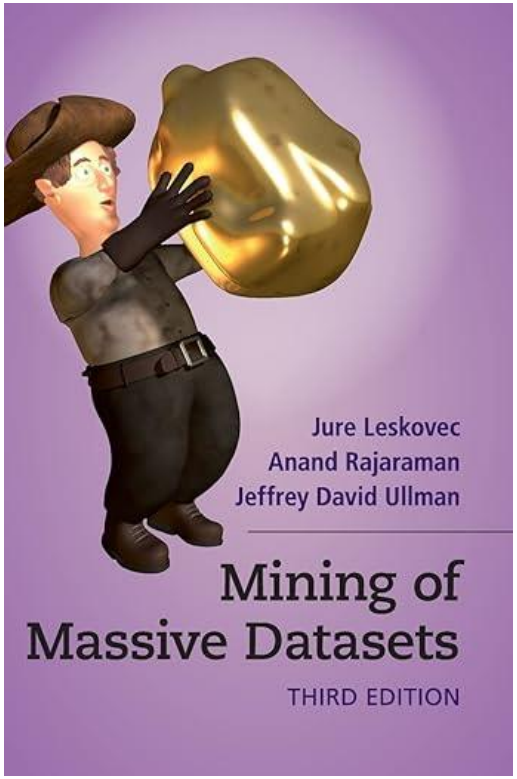
You worked on an interesting DAV project before (BSc thesis, MSc thesis, etc.) and would like to further expand it

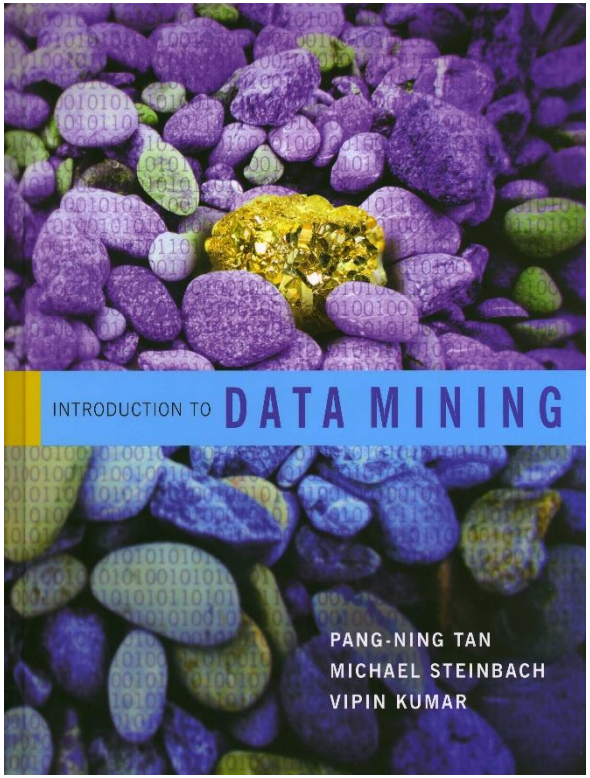
Potential interest in DAV

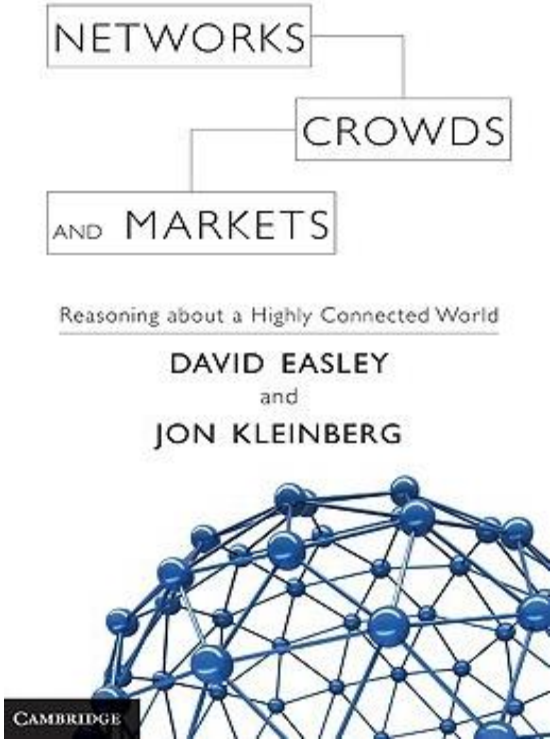
You are interested to work on a DAV project and looking for inspirations

"Suggested" Textbooks 1/2

Data Analytics



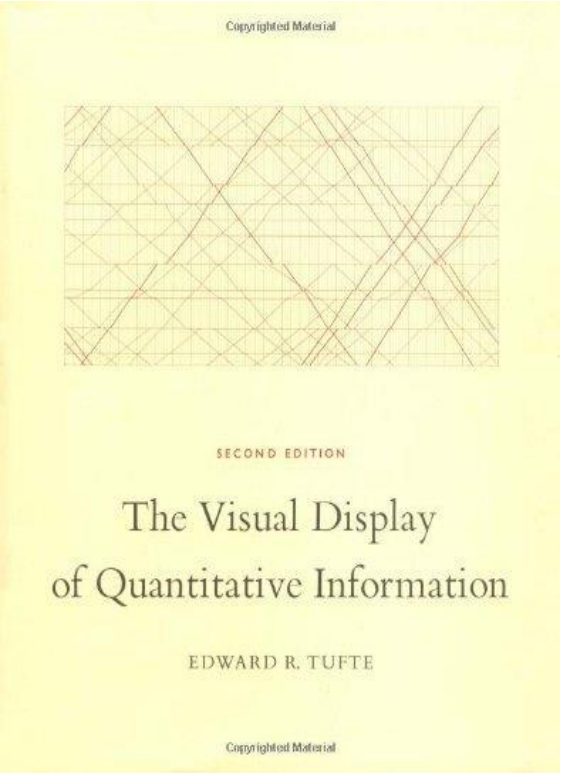


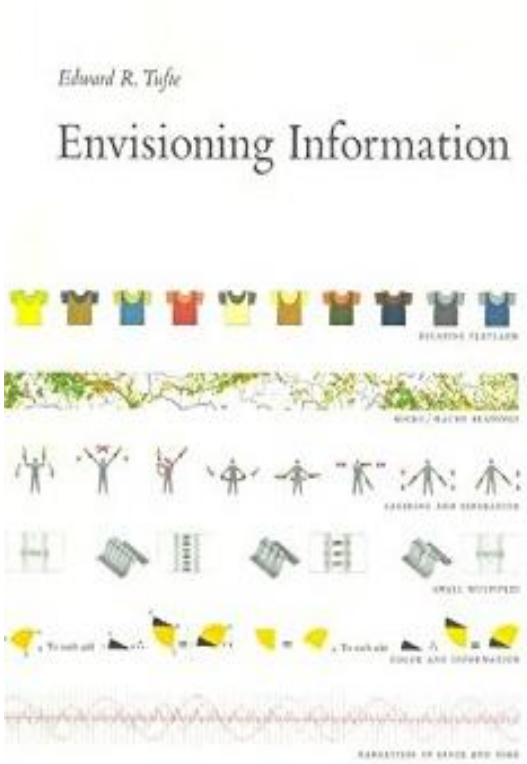


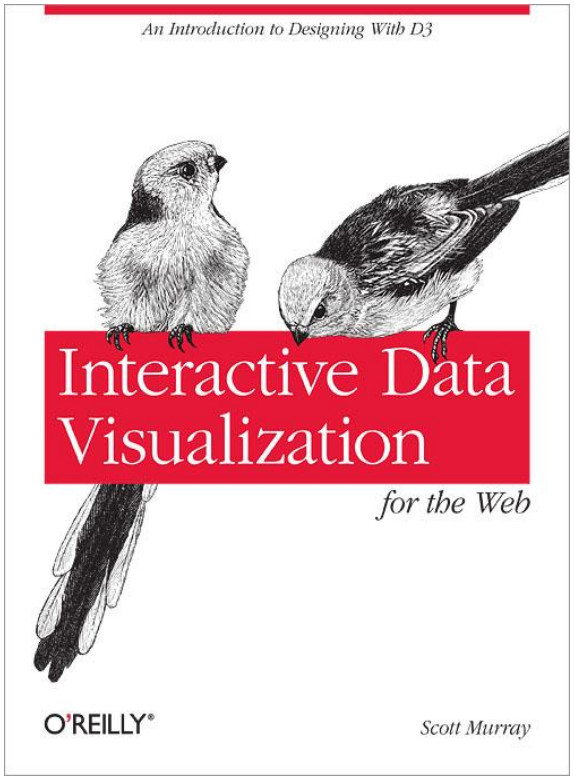
+ tools for data analytics

"Suggested" Textbooks 2/2

Data Visualization







+ tools for visualization of high-dimensional data

Logistics

Item	Comment
Classes	Wed @ 13:00-16:00
Classroom	BRG 213
Course group	3
Credits	3
Website	http://www.eecs.yorku.ca/~papaggel/courses/eecs6414/
Office hour	By appointment with teams

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 <u>fun</u> and <u>hard</u> work. ☺

Welcome!

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