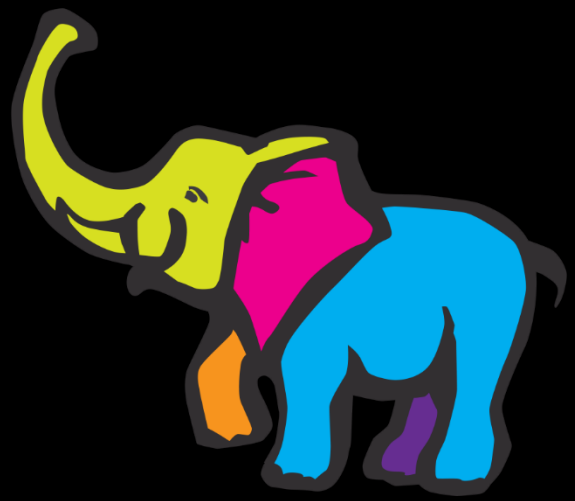
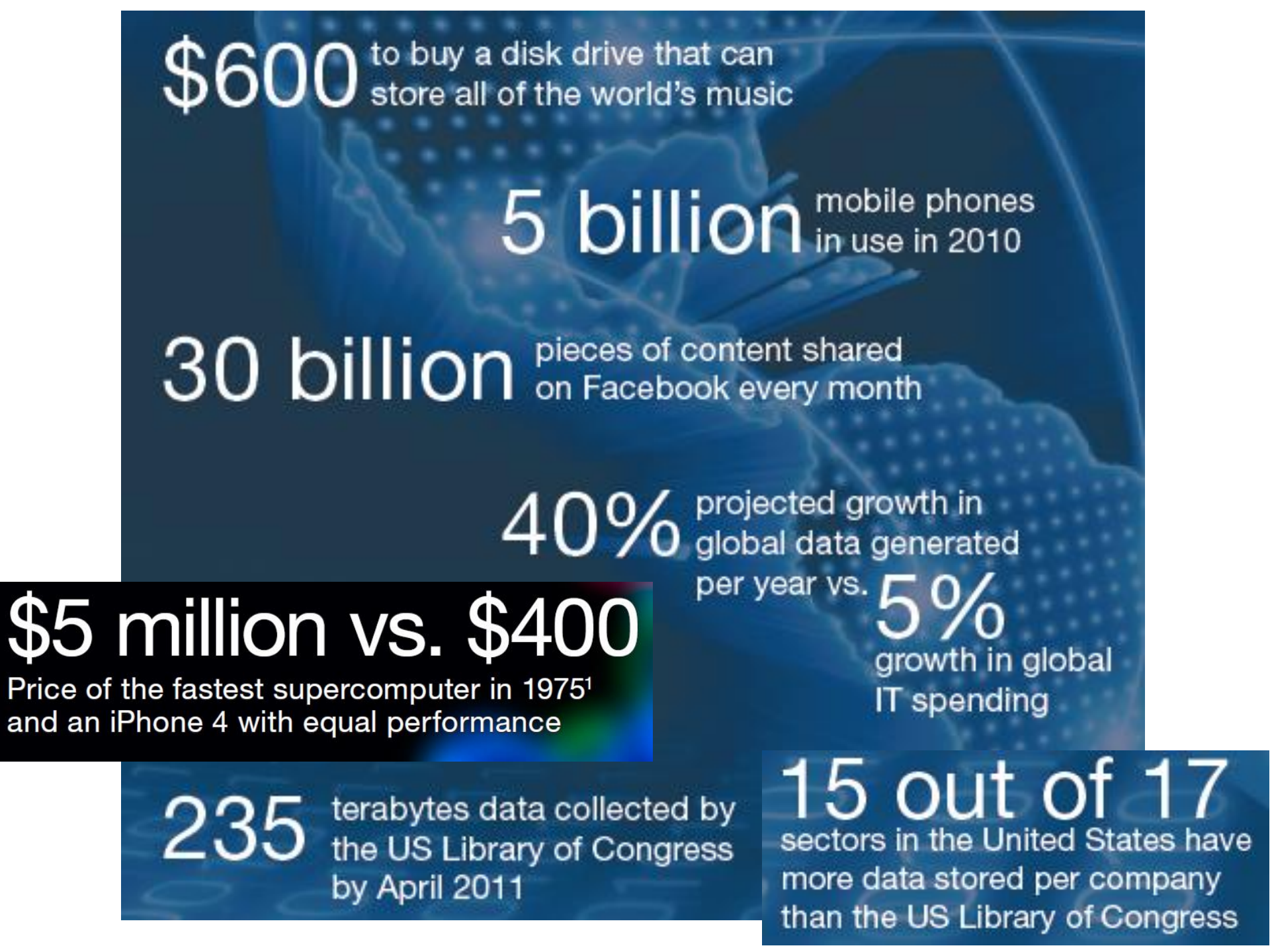
EECS4415: Big Data Systems



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

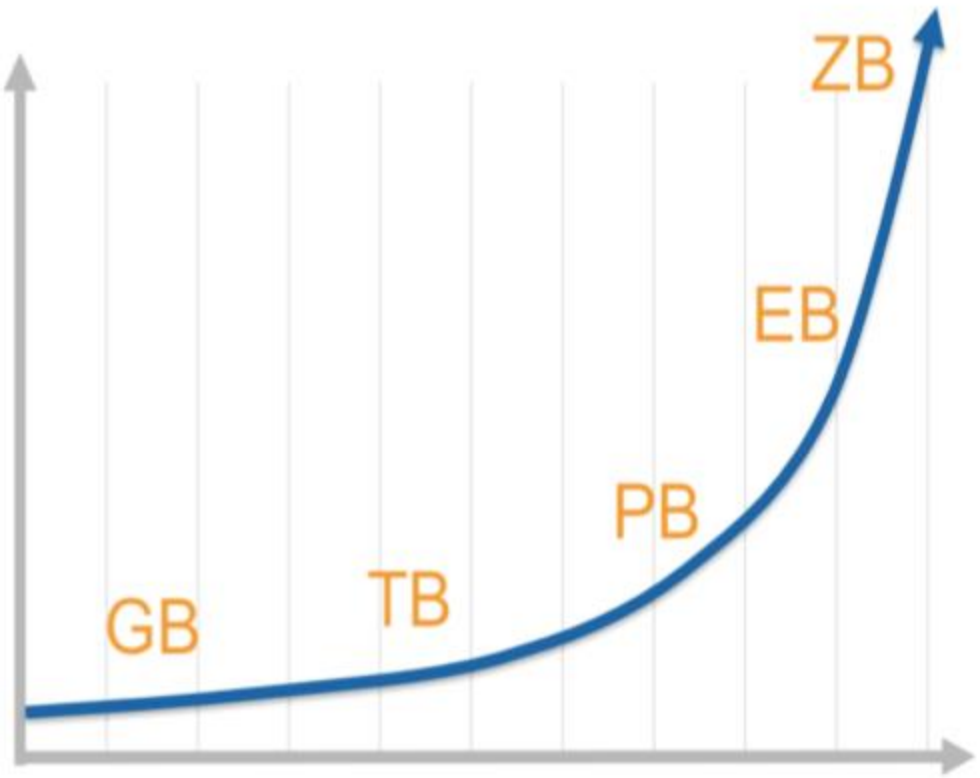
What is the purpose of big data systems?

To support analysis and knowledge discovery from very large amounts of 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

UNSTRUCTURED DATA GROWTH

SIGNIANT



Research from IDC shows that unstructured content accounts for 95% of all digital information, with estimates of compound annual growth at 65%.

By 2020, IDC predicts the volume of digital data will have reached 40,000 Exabytes (EB) or 40 Zettabytes (ZB).

The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. 2012. International Data Corporation (IDC). https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

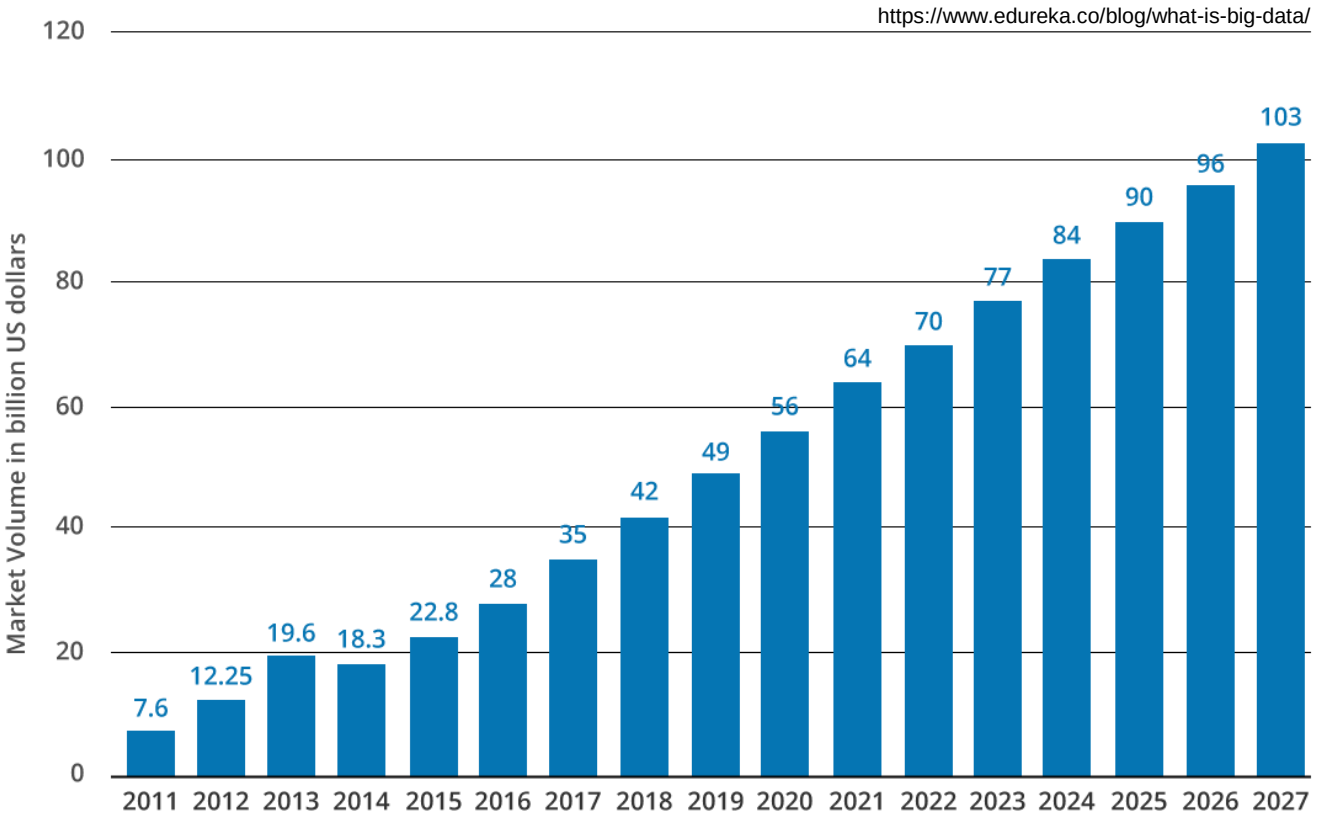
Data contains value and knowledge

Big Data Analytics

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

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

Adoption of Big Data Analytics



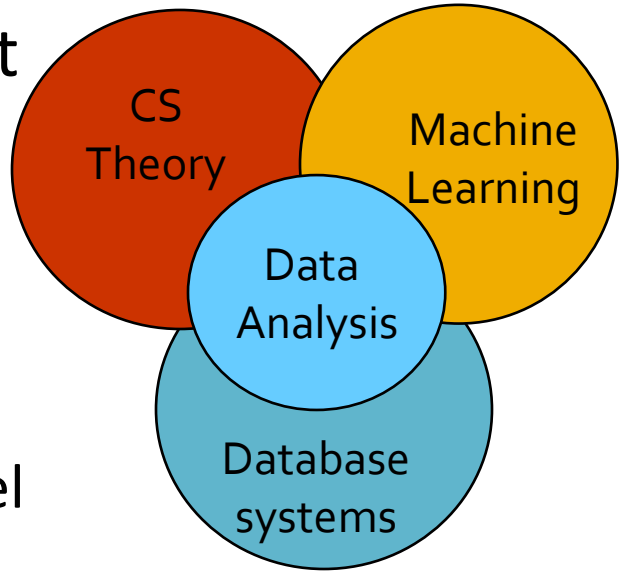
Growing market revenue of Big Data in billion U.S. dollars from the year 2011 to 2027

What is Big Data Analytics?

- **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 Analytics: Cultures

- **Data analysis overlaps with:**
  - **Databases:** 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: EECS4415

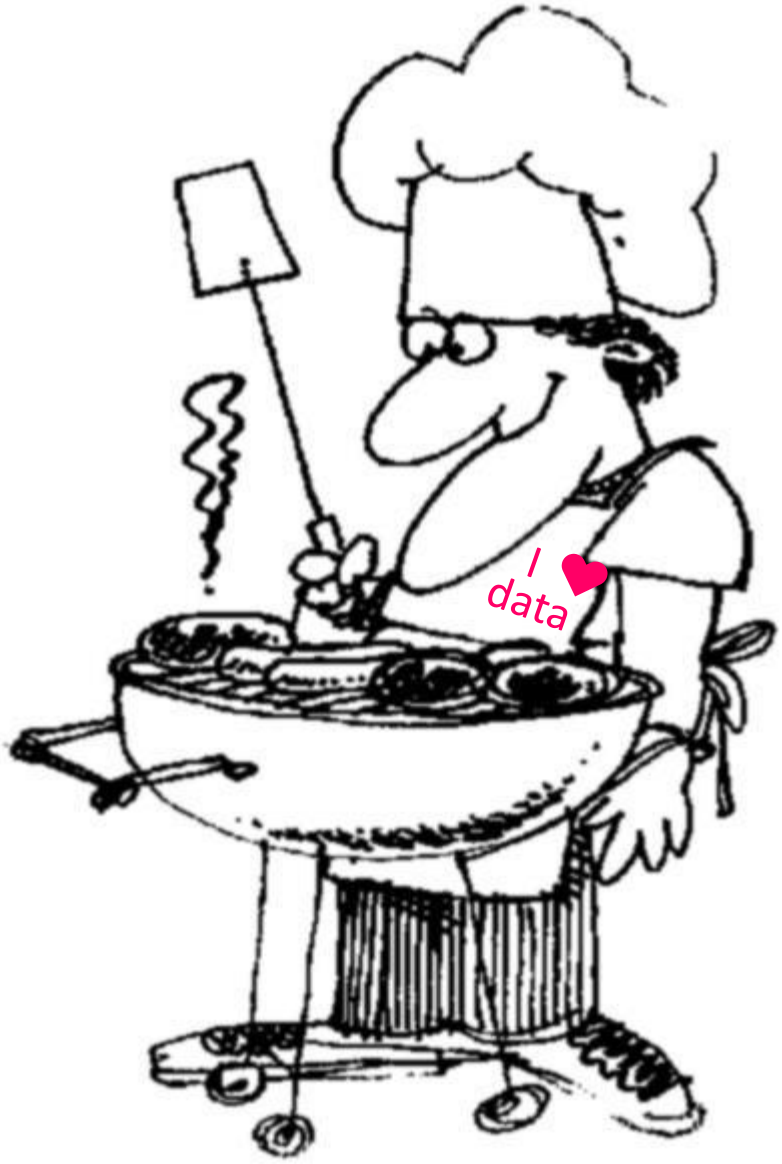
- **This class stresses more on**
  - **Storage systems**
  - **Distributed computing platforms**
  - **Algorithms, scalability**
  - Automation for handling **large data**

What will we learn?

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

What will we learn?

- Hands-on experience working with systems and tools for storing and processing big data:
  - MapReduce/Hadoop
  - Hive/BigQuery
  - Apache Spark
  - OpenRefine
  - …



How do you want that data?

EECS4415
About the Course

Logistics: Communication

- **Website**
  - http://www.eecs.yorku.ca/~papaggel/courses/eecs4415/
- **Piazza Q&A website:**
  - Available from the website
    https://piazza.com/yorku.ca/summer2021/eecs4415
  - You need to register with your ***yorku.ca*** email

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

Prerequisites

- **Course Prerequisites**
  - **EECS-3421: Introduction to Database Systems**
  - **EECS-3101: Design and Analysis of Algorithms**
  - **General prerequisites**
- **No single topic in the course is too hard by itself**
- **But we will cover and touch upon many topics and this is what makes the course hard**
  - **Good background in:**
    - Database Systems
    - Algorithms
  - **Programming:**
    - You should be able to write non-trivial programs (in Python)

Topics Covered

Component I

Data-driven Organizations, Data Ingestions, Data Quality, Data Lakes, Data Cleaning

Component II

Computing Platforms, Storage Systems, Distributed Processing Systems (for general-purpose batch data, structured data, graph data, streaming data), Data processing methods (Aggregation, grouping, filtering)

Component III

Serving data, Exploratory Data Analytics, Advanced Topics on Big Data Mining

Coursework

Work	Weight	Comment
Weakly Readings	10%	1% each
3 Assignments	50%	10%, 20%, 20%, respectively
Final Exam	40%	Final exam grade must be > 40%

Project

You need to:

- identify a problem
- find data
- design a big data architecture
- prepare data for analysis
- process data
- uncover insights
- communicate critical findings
- create a data-driven solution

+ **team-work**

Elements of a Big Data project

Need for data collection
Need for data storage
Need for data analysis
Need for data visualization (optionally)

…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

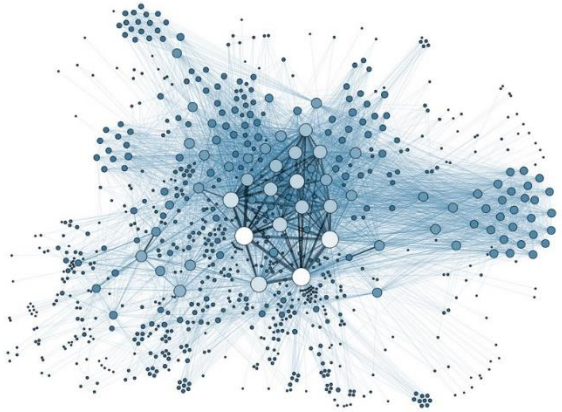
Human Resources Analytics

Why are our best and most experienced employees leaving prematurely?

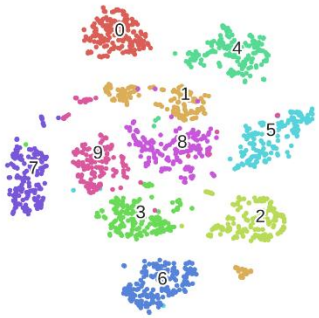
ludoben · updated 9 months ago · employment

What Type of Data?

Text Data

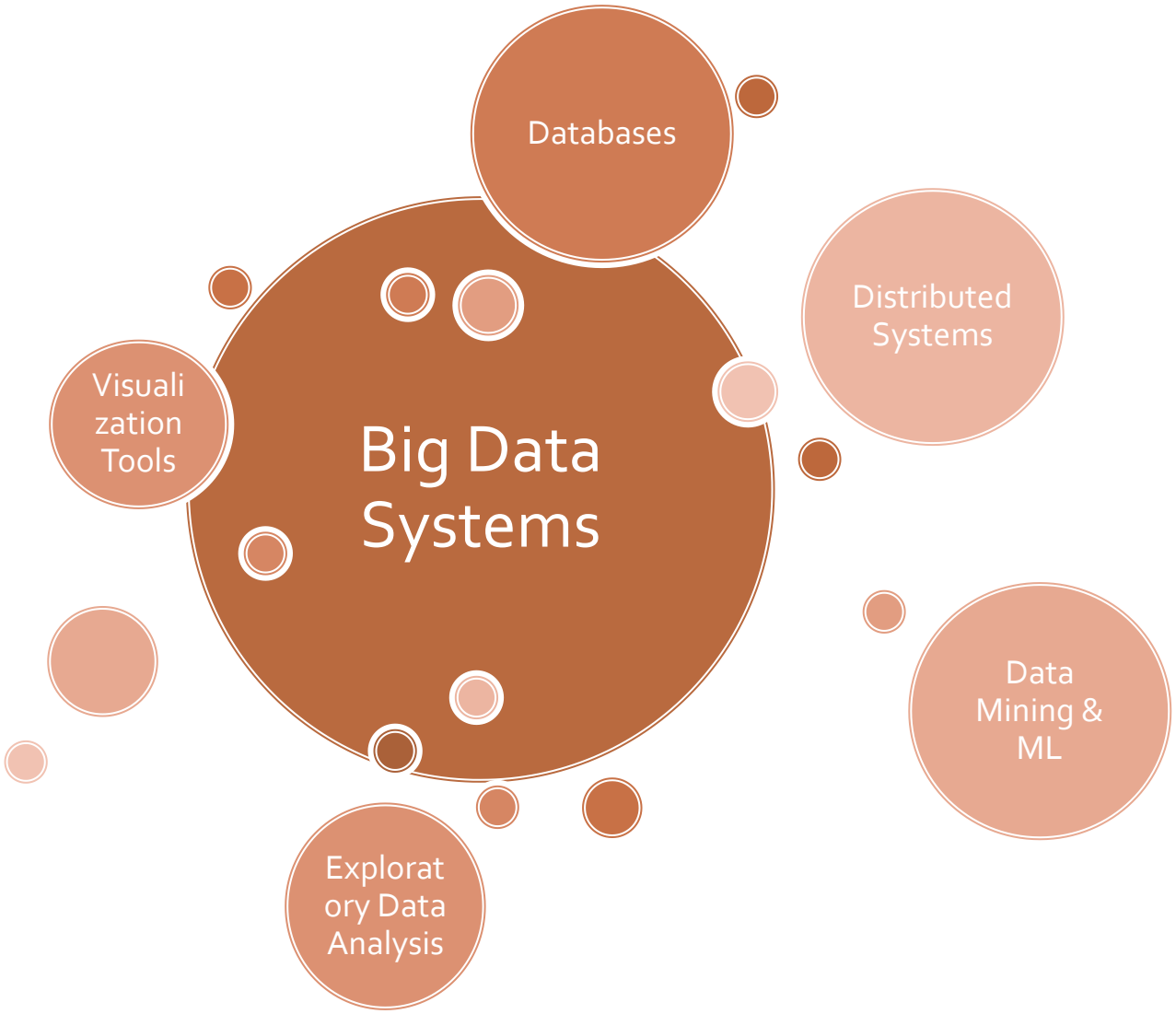


Network Data



Multivariate Data

Course Intellectual Content



Who Should Attend?

Current interest in Data Science

You are interested in the general area of data science

Interest in Big Data Technologies

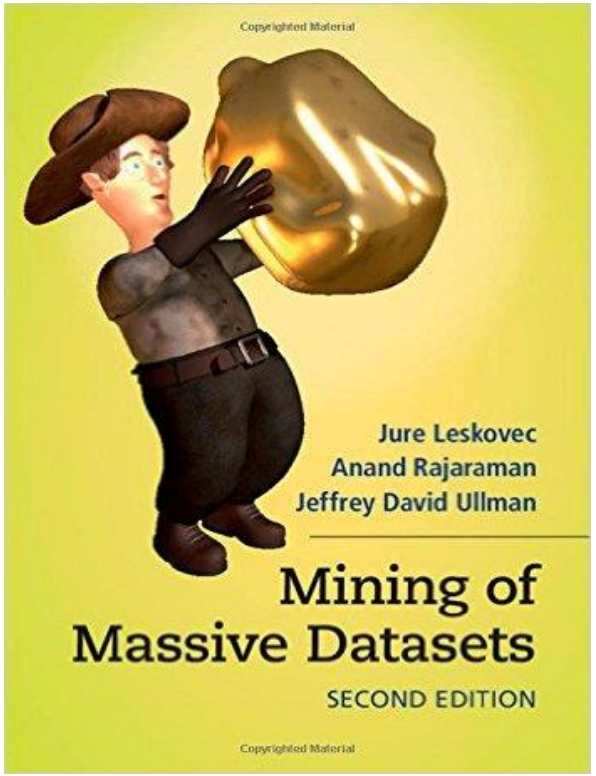
You are interested in big data systems and engineering

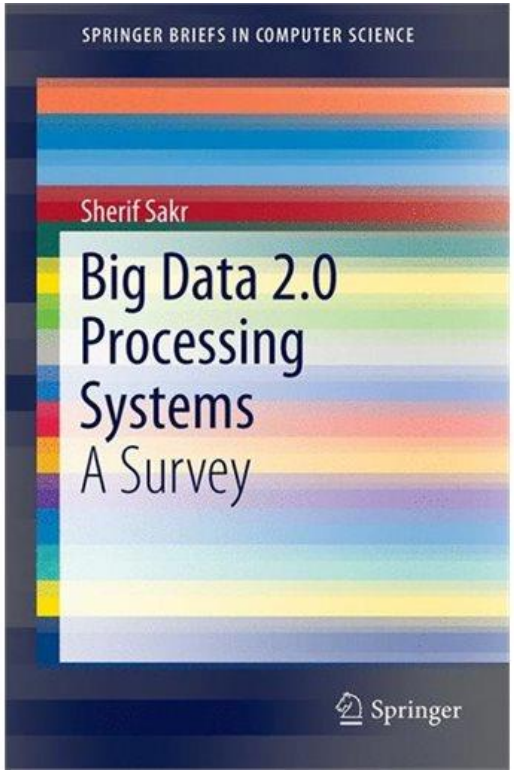
Interest in Big Data Analytics

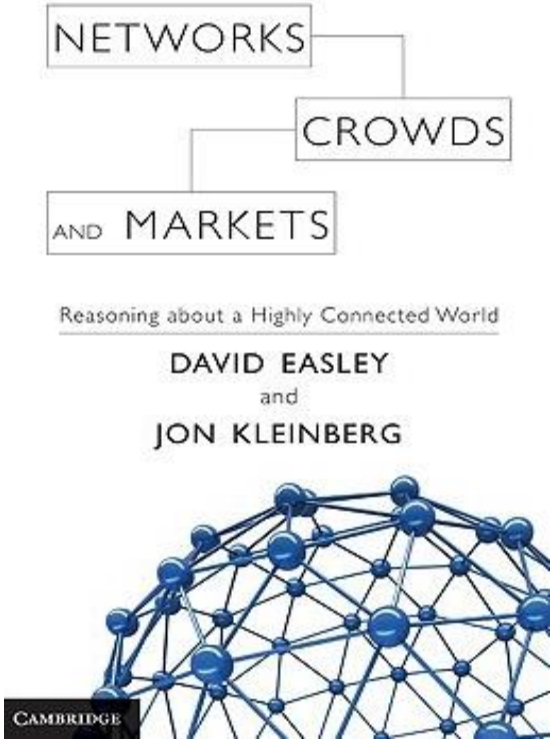
You are interested in finding interesting patterns and insights in large amounts of data

"Suggested" Textbooks

Data Analytics







+ tools for data analytics

Logistics

Item	Comment
Classes	Tue and Thu @ 16:00-19:00
Classroom	VH 3006 (Vari Hall)
Credits	3
Website	http://www.eecs.yorku.ca/~papaggel/courses/eecs4415/
Office hour	Tue, 13:00-14:00 (Online)

Welcome!

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