EECS6414: Data Analytics & Visualization
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 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 \(\Leftarrow\) emphasis on this class
  - Visualized \(\Leftarrow\) emphasis on this class

Data Analytics \(\approx\) Data Mining \(\approx\) Big Data \(\approx\) Predictive Analytics \(\approx\) Data Science
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
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)
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 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 stresses more on

- Data analysis of network data (graph model)
- Data analysis of high-dimensional data
- Data visualization principles & examples
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?

Challenges:
- Usage
- Quality
- Context
- Streaming
- Scalability

Data Modalities:
- Ontologies
- Structured
- Networks
- Text
- Multimedia
- Signals

Data Operators:
- Collect
- Prepare
- Represent
- Model
- Reason
- Visualize
How do you want that data?
EECS6414
About the Course
Logistics: Communication

- **Website**
  - [http://www.eecs.yorku.ca/~papaggel/courses/eecs6414/](http://www.eecs.yorku.ca/~papaggel/courses/eecs6414/)

- **Piazza Q&A website:**
  - Available from the website [https://piazza.com/yorku.ca/winter2020/eecs6414](https://piazza.com/yorku.ca/winter2020/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](mailto: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

Search datasets

Featured All

IMDB 5000 Movie Dataset
5000+ movie data scraped from IMDB website
chuansun76 · updated a year ago film

European Soccer Database
25k+ matches, players & teams attributes for European Professional Football
Hugo Mathien · updated 10 months ago association football, europe

Credit Card Fraud Detection
Anonymized credit card transactions labeled as fraudulent or genuine
Andrea · updated 10 months ago crime, finance

Human Resources Analytics
Why are our best and most experienced employees leaving prematurely?
ludoben · updated 9 months ago employment

~25,000 datasets
What Type of Data?

Text Data

Network Data

Multivariate Data
# (Tentative) Course Evaluation

<table>
<thead>
<tr>
<th>Milestone</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project proposal</td>
<td>10%</td>
</tr>
<tr>
<td>Project midterm report</td>
<td>20%</td>
</tr>
<tr>
<td>Project midterm in-class presentation</td>
<td>10%</td>
</tr>
<tr>
<td>Project final report</td>
<td>40%</td>
</tr>
<tr>
<td>Project final in-class presentation</td>
<td>20%</td>
</tr>
</tbody>
</table>

- project report in research paper format
- demo (if applicable)
... a number of lectures
## Topics Covered

<table>
<thead>
<tr>
<th>Topic</th>
<th>Lectures</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Network Analysis (~5 lectures)</strong></td>
<td></td>
</tr>
<tr>
<td>basic graph theory, network measurements,</td>
<td></td>
</tr>
<tr>
<td>network models, community detection, link</td>
<td></td>
</tr>
<tr>
<td>analysis &amp; link prediction</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>High-dimensional Data Analysis (~3 lectures)</strong></td>
<td></td>
</tr>
<tr>
<td>Frequent itemsets (a-priori algorithm, PCY</td>
<td></td>
</tr>
<tr>
<td>algorithm, SON algorithm); finding similar</td>
<td></td>
</tr>
<tr>
<td>items (shingling, min-hashing, locality-</td>
<td></td>
</tr>
<tr>
<td>sensitive hashing (LSH)); High dimensionality clustering (BFR algorithm, CURE algorithm)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Visualization (~2 lectures)</strong></td>
<td></td>
</tr>
<tr>
<td>Value of visualization, visual variables,</td>
<td></td>
</tr>
<tr>
<td>cognition and perception, colors, pre-</td>
<td></td>
</tr>
<tr>
<td>attentive vs attentive processing, visual</td>
<td></td>
</tr>
<tr>
<td>metaphors, taxonomy of visualization,</td>
<td></td>
</tr>
<tr>
<td>visualizations of qualitative and</td>
<td></td>
</tr>
<tr>
<td>quantitative data</td>
<td></td>
</tr>
</tbody>
</table>
Data Analysis and Visualization

- Graph Mining
- Data Mining
- Probability & Statistics
- Exploratory Data Analysis
- Visualization Tools & Interaction
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

<table>
<thead>
<tr>
<th>Item</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classes</td>
<td>Mon @ 16:00-19:00</td>
</tr>
<tr>
<td>Classroom</td>
<td>Accolade East Building 013 (ACE 013)</td>
</tr>
<tr>
<td>Course group</td>
<td>3</td>
</tr>
<tr>
<td>Credits</td>
<td>3</td>
</tr>
<tr>
<td>Website</td>
<td><a href="http://www.eecs.yorku.ca/~papaggel/courses/eecs6414/">http://www.eecs.yorku.ca/~papaggel/courses/eecs6414/</a></td>
</tr>
<tr>
<td>Office hour</td>
<td>Drop anytime by my office (LAS3050) or by appointment</td>
</tr>
</tbody>
</table>
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/