Welcome to
CS4414 / CS9637 / CS9114
Introduction to Data Science I

Dr. Dan Lizotte (Comp. Sci., Epidemiology & Biostatistics)
“A data scientist is a statistician who lives in San Francisco”

“Data Science is Statistics on a Mac”

“A data scientist is better at statistics than any software engineer and better at software engineering than any statistician.”
Data
Data

1. As a count noun: an item of information; a datum; a set of data. Also *fig.*
2. As a mass noun.
   a. Related items of (chiefly numerical) information considered collectively, typically obtained by scientific work and used for reference, analysis, or calculation. Cf. *datum n. 1a.*
   b. *Computing.* Quantities, characters, or symbols on which operations are performed by a computer, considered collectively. Also (in non-technical contexts): information in digital form. Cf. *datum n. 1b.*

[OED]
Science
Science

A branch of study that deals with a connected body of demonstrated truths or with observed facts systematically classified and more or less comprehended by general laws, and incorporating trustworthy methods (now esp. those involving the **scientific method and which incorporate falsifiable hypotheses**) for the discovery of new truth in its own domain. [OED]
Science

1. Generate a hypothesis

2. Generate data through observation and/or experiment

3. Assess whether the data are consistent with the hypothesis or not
Data Science

1. Get some data
2. Try some methods
3. ...
• https://twitter.com/WesternGymBot
“Is there an association between student grades and weight room use?”

“Which student attributes are most associated with weight room use?”

“How busy will the weight room be in the next hour?”
"Is there an association between student grades and weight room use?"

"Which student attributes are most associated with weight room use?"

"How busy will the weight room be in the next hour?"
Confirmatory

"Is there an association between student grades and weight room use?"

Exploratory

"Which student attributes are most associated with weight room use?"

Predictive

"How busy will the weight room be in the next hour?"

Statistics

Machine Learning
Any dangers with starting from data?
Any dangers with starting from data?

• Spurious results because of random noise (variance)

• Spurious results because of missing information (bias)
Variance
Total revenue generated by arcades correlates with Computer science doctorates awarded in the US

Correlation: 0.98
Per capita consumption of chicken correlates with Total US crude oil imports

Correlation: 0.89
Bias
SCOOOP: With secret access to NYPD CCTV @IBM created software which tags people based on their skin tone + hair/clothing color. IBM gave NYPD access, then pitched them on a new AI product which identifies people on camera as "Black," "White," and "Asian":

[Link to article](theintercept.com/2018/09/06/nyp...)
Supervised Learning
Unsupervised Learning
DATA science

Statistic
Visualisation
Data Mining
Pattern recognition
Machine Learning
Neurocomputing
& Database & Data processing
ImageNet Challenge

- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.
ImageNet Challenge: 2014

AlexNet
- image
- conv-64
- conv-192
- conv-384
- conv-256
- conv-256
- FC-4096
- FC-4096
- FC-1000

I WAS WINNING IMAGENET

UNTIL A DEEPER MODEL CAME ALONG
MAP REDUCE

Distribute the documents across N computers

For each document, return a set of (word, frequency) pair

We get a big distributed list of sets of words frequency

Each Reduce function count the occurrences of one word.
Of transactions that included milk:
- 71% included bread
- 43% included eggs
- 29% included toilet paper
DATA science

- Statistics
- Visualisation
- Pattern recognition
- Machine Learning
- Neurocomputing
- Data Mining
- Database & Data processing
DATA science

Pattern recognition

Machine Learning

Neurocomputing

Data Mining

& Database & Data process

Visualisation

Statistic
Course Objective

• Introduce students to data science (DS) techniques, with a focus on application to substantive (i.e. "applied") scientific problems.

• Through group projects, students will gain experience in identifying which problems can be tackled by DS methods, and learn to identify which specific DS methods are applicable to a problem at hand.

• This course requires students to show substantial initiative in investigating methods that are applicable for their project. The lectures give an overview of important methods, but the lecture content alone is not sufficient to produce a high quality course project.
Logistics

- **READ. THE. WIKI.**
  http://www.csd.uwo.ca/~dlizotte/teaching/IDS/

- Instructor: Dan Lizotte – dlizotte at uwo dot ca – MC363
  TA: Nathan Phelps — nphelps3 at uwo dot ca

- Time: Tuesday from 2:30AM – 4:30PM, and on Thursday from 2:30PM – 3:30PM

- Place: Talbot College TC 205

- Communication: We will be using OWL for electronic communication.

- Question & Collaboration Hour: TBA
Materials

• **READ. THE. WIKI.**

• “Required” materials are materials that I expect you to consult if you have questions. Not required reading cover-to-cover.
Anticipated Topics and Schedule

- **Introduction to Data Science:** Definitions, Components, Relationships to Other Fields

- **Data Cleaning:** Working with structured data: selecting, filtering, joining, aggregating, Simple visualizations, “Face validity”

- **Supervised Machine Learning:** Regression, Classification. Linear Regression, SVMs, Trees, (Maybe also Reinforcement Learning and Sequential Decision Making)

- **(Re)-introduction to Statistics:** Data Summaries, Randomness, Sample Spaces and Events, Probability, Random Variables, Inference: Hypothesis testing, P-values, confidence Intervals Multivariate Statistics: conditional probability, correlation, independence

- **Evaluation:** Test set, cross-validation, bootstrap, confounding, causal inference

- **Unsupervised Machine Learning, Representations, and Feature Construction:** Clustering, Dimensionality reduction, Domain-specific Feature Development, Deep Learning, Images, Sounds, Text

- **Visualization

- **Your picks?"
Evaluation

• Midterm - 4414/9114: 35% 9637: 30%

• Brainstorming Session – 10%

• Project Proposal – 4414/9114: 15% 9637: 10%

• Report Draft – 5%

• Project Report – 35%

• Peer Review – 9637 only: 10%
Group Project (2 or 3)

- Project Proposal – 4414/9114: 15% 9637: 10%
  - Document detailing the plan for the project. See Project Guidelines on the wiki for detailed requirements.

- Report Draft – 5%
  - The purpose of the draft is to allow the instructor to provide feedback on the quality of the writing and the direction of the project.

- Project Report – 35%
  - Each student will prepare a research paper detailing a substantive problem, the data available, the applicable DS methods, and empirical results obtained on the problem.
Brainstorming - 10%

- Each group will prepare a presentation explaining an applied problem, as well as some potential data science methods that could be applied to the problem.

- The presentation should be no more than 10 minutes.

- We will then discuss the problem as a class, along with possible approaches for solving the problem using ML methods.

- Students are expected to be prepared to answer deep questions about the nature of their problem to ensure that they receive high quality feedback from the brainstorming session.

- See Project Guidelines on the Wiki for detailed requirements.
Brainstorming

• You must pick a brainstorming slot.

1. Find the wiki on OWL

2. Edit the schedule a replacing “SlotX” with your names.

• Pick one before Friday, 5 October at 5pm or we will pick one for you.
Peer Review

• Each research graduate (9637) student will be assigned three project reports to review

• Primary Purpose: Provide feedback to authors that they can make use of in their future careers, which gives them a better return on the investment they have made in their course project.

• Secondary Purpose: Give students a view of the variety of work that has been done in the course, and further develop reviewing skills.

  • Reviews from other students will not affect the grade of the author in any way.

  • See the wiki for more details.
Accessibility and Support, Missed Course Components

• Check the wiki.
Questions and Chat:
Why are you here?