Welcome to CS4437 / CS9637 / CS9114
Introduction to Data Science

Dr. Dan Lizotte (Comp. Sci., Epidemiology & Biostatistics)

IF YOU ARE NOT REGISTERED,
please have a seat and we will discuss as a group.
DATA
science

Pattern recognition

Machine Learning

Neurocomputing

Statistic

Visualisation

Data Mining

& Database & Data process
“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.”
“Choosing and using methods from computer science and statistics to understand something about the world.”

–Dan
DATA science

- Pattern recognition
- Machine Learning
- Neurocomputing
- Statistic
- Visualisation
- Data Mining
- Database & Data processing
Supervised Learning

![Supervised Learning Diagram](image)
Unsupervised Learning
Reinforcement Learning

Figure 1: Screen shots from five Atari 2600 Games: (Left-to-right) Pong, Breakout, Space Invaders, Seaquest, Beam Rider

\[
Q_{t+1}(s_t, a_t) = \underbrace{Q_t(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha_t(s_t, a_t)}_{\text{learning rate}} \cdot \left( \underbrace{R_{t+1}}_{\text{reward}} + \gamma \underbrace{\max_a Q_t(s_{t+1}, a)}_{\text{estimate of optimal future value}} - \underbrace{Q_t(s_t, a_t)}_{\text{old value}} \right)
\]
DATA Science

- Statistic
- Visualisation
- Pattern recognition
- Machine Learning
- Neurocomputing

& Database & Data processing
**MAP REDUCE**

1. Distribute the documents across N computers.
2. For each document, return a set of (word, frequency) pair.
3. We get a big distributed list of sets of words frequency.
4. 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

Neurocomputing

Statistic
Visualisation
Data Mining
Pattern recognition
Machine Learning
### 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

Image
- conv-64
- maxpool
- conv-128
- maxpool
- conv-256
- maxpool
- conv-256
- maxpool
- conv-256
- maxpool
- conv-512
- maxpool
- conv-512
- maxpool
- conv-512
- maxpool
- FC-4096
- FC-4096
- FC-1000
- softmax
Course Objective

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

• Through individual 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: Brent Davis — bdavis56 at uwo dot ca

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

• Place: Middlesex College MC105-B

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

• Question & Collaboration Hour: TBA, Thursdays after class
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**
Evaluation

• Daily Quizzes – 5%
• Midterm - 35%
• Brainstorming Session – 5%
• Project Proposal – 4414/9114: 15% 9637: 10%
• Report Draft – 5%
• Project Report – 35%
• Peer Review – 9637 only: 5%
Daily Quizzes

• Very short quiz at the beginning of class covering the previous day's materials

• The final quiz will be on 31 Oct.

• The lowest quiz mark will be dropped.

• Quiz marks will only be excused for medical reasons.
Individual* Project

- 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 - 5%

• Each student 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.

• Student is 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. Request an account on the Wiki

2. Edit the schedule at the bottom of the main page, replacing “SlotX” with your name.

• Pick one before Friday, 6 October at 5pm or I will pick one for you and you won’t like it.
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?