Welcome to CS4437 / CS9637
Introduction to Data Science

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

- Statistic
- Visualisation
- Data Mining
- Pattern recognit.
- Machine Learning
- Neurocomputing & 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.”
DATA science

Pattern recognition
Machine learning
Neurocomputing

Statistic
Visualisation
Data mining
& Database & Data processing
Supervised Learning
Unsupervised Learning
Reinforcement Learning

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

\[
Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_t(s_t, a_t) \cdot \left( R_{t+1} + \gamma \max_a Q_t(s_{t+1}, a) - Q_t(s_t, a_t) \right)
\]

Diagonal Line Node

Face Node

Cat Node
Map Reduce

1M documents -> Function (ie Map) -> Big Set of Word

5M documents

Function (ie Map) for each document, return a set of (word, frequency) pair

Big Set of Word

We get a big distributed list of sets of words frequency

Function (ie Reduce) for each word

Each Reduce function count the occurrences of one word.
DATA science

- Statistic
- Visualisation
- Pattern recognition
- Machine Learning
- Neurocomputing
- Data Mining
- Database and Data process
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
- Databases
- Data processing
DATA science

Statistic

Visualization

Data Mining

Pattern recognition

Machine Learning

Neurocomputing

& Database & Data processing
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/cs4437/

• Instructor: Dan Lizotte – dlizotte at uwo dot ca – MC363

• Time: Tuesday from 11:30AM – 1:30PM, and on Thursday from 3:30PM – 4:30PM

• Place: Middlesex College MC320

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

- READ. THE. WIKI.
Anticipated Topics and Schedule

- READ. THE. WIKI.
Evaluation

• Daily Quizzes – 5%

• Brainstorming Session – 35%

• Project Proposal – 4437: 15% 9637: 10%

• Report Draft – 5%

• Project Report – 40%

• 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 11 Feb.

• The lowest quiz mark will be dropped.

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

- Project Proposal – 4437: 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 – 40%
  - 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 - 35%

- 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 between **20 and 30 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. Create account on the Wiki

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

• Pick one before Friday, 29 January at 5pm or I will pick one for you and you won’t like it.
Peer Review

• Each 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?