Introduction

Max Turgeon

DATA 2010–Tools and Techniques in Data Science

Lecture schedule

We meet three times a week on Zoom:

- **Monday**, **Wednesday**, and **Friday** 1:30pm to 2:20pm

Lectures will be **recorded** and available on UM Learn.

Lab schedule

Once a week on Webex:

- **Monday** 4:30pm to 5:45pm

Assessments

- 5 labs (10% total)
- 3 assignments (20% total)
- 1 final project (40%)
- Final exam (30%)

Assignments

- Assignments will provide an opportunity to analyse data and practice the skills from the lectures.
    - They will require using a programming language: R or Python
    - We will review R in the first lab, and introduce Python as we go along

By the way, should I already know R and Python?

- You should have seen R in STAT 2150, but we will review together.
    - Concepts will be introduced as needed, and through examples.
    - See UM Learn for extra material on R.
- The goal at the end of the semester is to have the basics of both languages so that you can learn more on your own.
    - Both languages are used in research and industry!

Final project

- In teams of 2, you will have to choose a dataset from a pre-determined list, analyze it using the tools and techniques from this class, and summarize your findings.
- I will provide more details later, but you can start teaming up.

What is data science? i

- **Disclaimer**: Data science is not a well-defined field, so this is only my personal definition.
- Traditional statistics has focused on using statistical tests or models to understand data.
- Data science is concerned with the whole process of data analysis:
    - Store and retrieve data.
    - Visualizing data effectively.
    - Communicating results.

What is data science? ii

- It also blurs the line between traditional statistical models (e.g. linear regression) and machine learning (e.g. neural networks).
    - If a model does the job, who cares where it comes from!
- For these reasons, a good data scientist has good knowledge of both statistics and computer science.

What will we learn? i

- Here are the three main course objectives:
  - Become proficient in R/Python, to the level that you can analyse data using the tools from this class.
  - Be able to describe and analyze data through visualization and simple statistical procedures.
  - Be introduced to statistical thinking and be able to think critically about variation and biases.

What will we learn? ii

- We will cover the following tools and techniques:
  - Data cleaning and wrangling
  - Correlation, distributions, significance
  - Data Visualization
  - Scores and Rankings
  - Linear regression
  - Introduction to Machine Learning

What will we not learn?

- Cloud computing
- Big data technologies (e.g. Spark and Hadoop)
- Advanced statistical techniques.

But we have courses on all of them if you're interested!