# Analysis of Big Data

## Lectures

Wed, 12:00 - 13:00 @ IA2050

Fri, 10:00-12:00 @ IA2160

## Instructor

### Qiang Sun

Email: qiang.sun@utoronto.ca

Office hour: TBD

In your email, please include "$YOUR_NAME_$STUDENT_ID:" in your subject line. For example, your email subject line should read as "Jason Roy_123: regarding problem 1 in homework 1."

## Teaching Assistant

TBD

## Description

The amount of data in our world has been exploding, and analyzing large data sets is becoming a central problem in our society. Large data sets include data with high-dimensional features and massive sample size. This course introduces the statistical principles and computational tools for analyzing big data: the process of acquiring and processing large datasets to find hidden patterns and gain better understanding and prediction, and of communicating the obtained results for maximal impact. Topics include optimization algorithms, inferential analysis, predictive analysis, and exploratory analysis.

This is an advanced learning and optimization course for fourth year undergraduate students.

## Prerequisite

STAC58, STAC67, and CSCC11.

## Computing

Python Programming Language.

## Evaluation

Participation: 10%. For more details, see here.

Assignments: 30%.

Midterm Exam: 30%, in class midterm on theory.

Final Project: 30%, final project.

## Reference Textbooks

The main reference will be scribe notes. We will develop a set of scribe notes along the way. Please do not distribute them without my permission.

Other than the scribe notes, some other references are:

1. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R.*
2. Hastie et al., *Elements of Statistical Learning.*
3. Aston Zhang, et al. *Dive into Deep Learning.*

## Topics

We will in general cover supervised learning, generative models, representation learning, and geometric methods (with a focus on graphical models). The following are potential topics and might change subject to the instructor's view on the progress of the course:

- Fundamental conceptions and principals.
- Supervised learning: Regression and classification.
- High dimensional statistics
- Introduction to deep learning
- Expressivity of neural networks
- Optimization: Adaptive (stochastic) gradient descent, back propagation, the muon optimizer
- Neural architecture designs: CNN and transformer
- From Bert to ChatGPT
- Generative models: linear, GAN, VAE, and diffusion models
- Transfer learning: Fine tuning, prompt tuning, causal transfer learning, lifelong learning
- Dependence learning: Gaussian graphical models, random dot graphs, and GNNs
- Ensemble learning: Bagging, stacking, boosting, and meta learning.

## Assignment Policy:

We will only take .rmd, .tex and .pdf (along with related materials), python notebook for assignments. Associated and reproducible code, if any, must be attached to the .rmd file. Late submissions will NOT be accepted.

If plagiarisms are found, they will be reported. Both (or multiple) copies of the assignments will be given zero grades.

## Exam details and dates

Mid Term (In-class exam): TBD

Final Project (TBD)