Applied Machine Learning

Syllabus and logistics

Reihaneh Rabbany

Remote Class

- **Live Lectures:** Tuesday & Thursdays, 1:05 pm - 2:25 (Montreal time)
  - Online Zoom meetings through Mycourses
  - Lectures will be **recorded** and automatically uploaded in Mycourses
- **Course Website:** http://www.reirab.com/comp551.html 

  - Syllabus, slides, deadlines, schedule, evaluation, etc.

Communications

- **Online Zoom Office Hours** **under zoom tab in Mycourses**
  - **Instructor:** Thursdays 3:00 pm - 4:00 pm
  - **TAs:** please check Mycourses's calendar
- **Course Email:** comp551mcgill@gmail.com 

- **Instructor Email**: rrabba@cs.mcgill.ca [for private communication "551 special" in title]
- **Course Slack:** will send invites soon

Prerequisites

- Strong linear algebra, probabilities, and Python programming is highly recommanded
- How can I refresh my background knowledge to follow the lectures better? a lot of excellent online materials, see which one you can follow easier, you can also refer to these reviews on probability and linear algebra.
- Tutorials to be scheduled for Math and Python

- Two quizzes on main concepts needed for lectures [with unlimitted attmpts allowed], due Jan 19th, released next week

Tutorials

Mid Jan.	Probability & Linear Algebra	
Late Jan.	Python	https://www.python.org/
Mid Feb.	Scikit-learn	https://scikit-learn.org/
Early Mar.	Pytorch	https://pytorch.org/

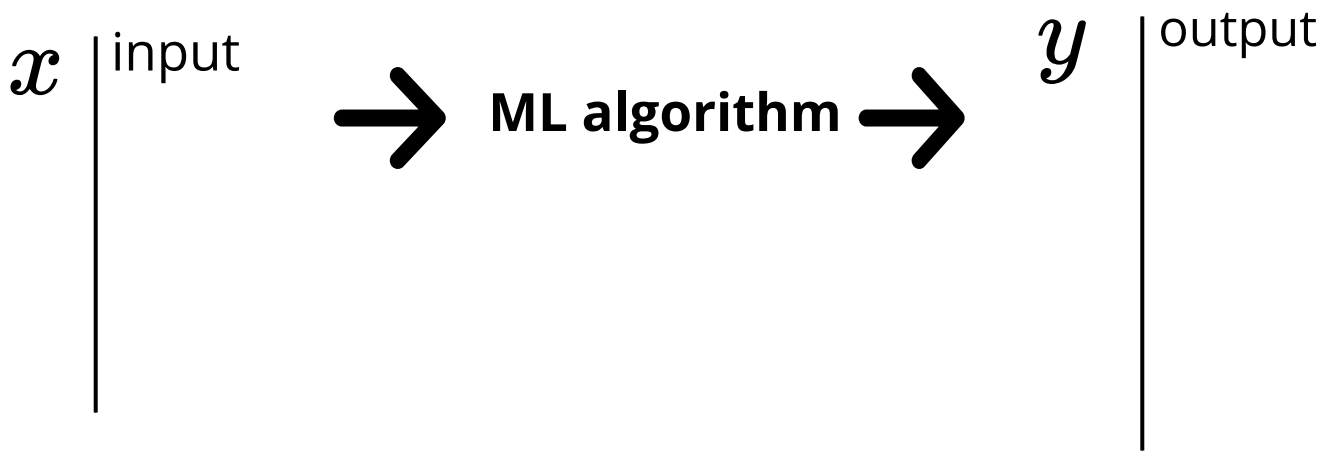
Pre-recorded, attend the corresponding TA's office hour for questions

About this course

Tentative Outline

- Introduction
- Nearest Neighbours
- Classification and regression trees
- Core concepts
- Maximum likelihood and Bayesian Reasoning
- Naive Bayes
- Linear regression
- Logistic and softmax regression
- Gradient descent methods
- Regularization
- Perceptrons & Multilayer Perceptrons
- Gradient computation and automatic differentiation
- Convolutional neural networks
- Linear support vector machines
- Bagging & Boosting
- Unsupervised learning
- Dimensionality reduction
- Learning with graphs

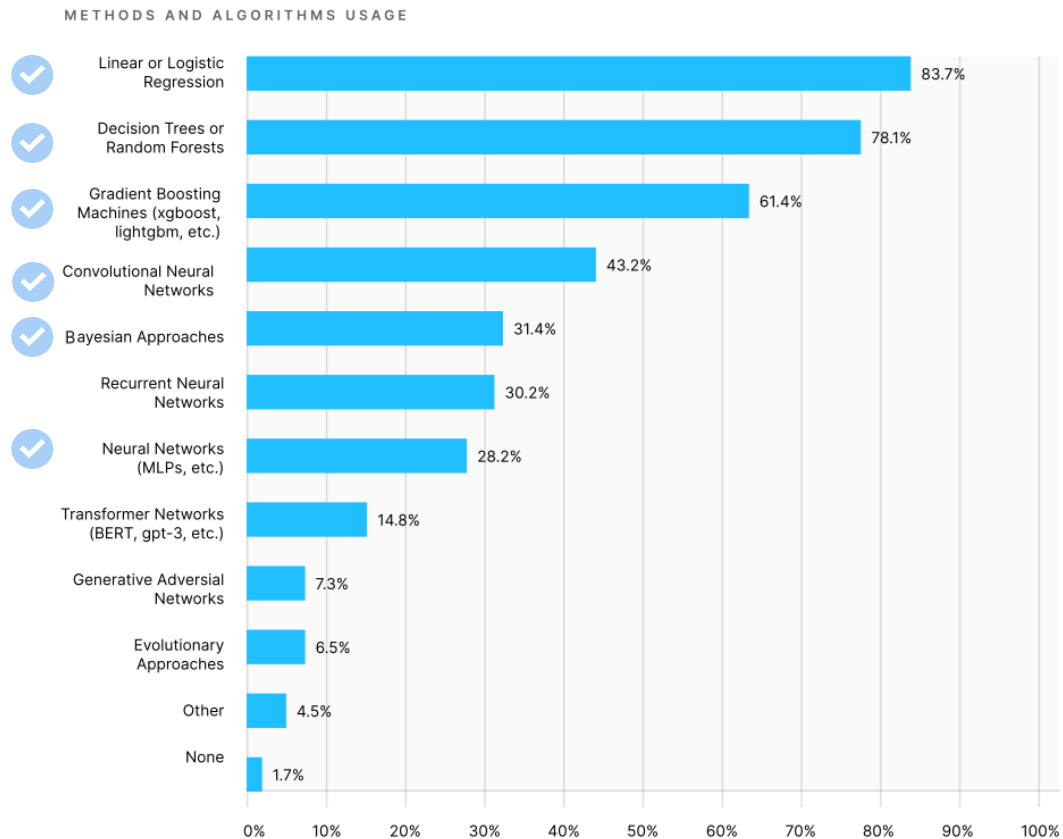
About this course



example

<tumorsize, texture, perimeter> = <18.2, 27.6, 117.5> → cancer = No

About this course



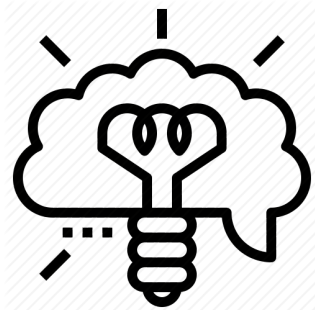
from 2020 Kaggle's survey on the state of Machine Learning and Data Science,
you can read the full version here

About this course

complementary components

Theory

Lectures
Weekly Practice Quizzes
Midterm Exam
Understand the theory behind learning algorithms



Application

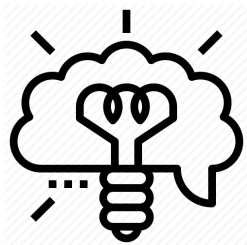
Codes in lectures
Mini-projects
Practice applying them in real-world

About this course:

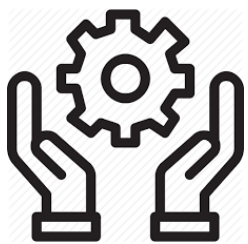
Evaluation and grading

Regular Practice Quizzes - **20%** {from last lecture - short}

Late Midterm exam - **30%**



Mini-projects - **50%** {group assignments}



About this course:

Evaluation and grading

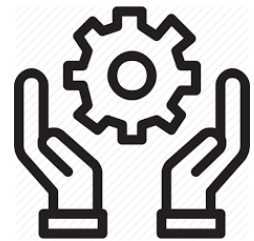
Regular Practice Quizzes - **20%**

{from last lecture - short}

- One per lecture to check the key concepts discussed in the last lecture
- Available until the start of the next lecture, timed
- From the two subsequent quizzes, the best one will be considered for your final grade
- The first 2 practice quizzes check the prerequisites for the course and have a different setting than the regular practice quizzes. In particular, unlimited attempts are allowed and due date is set to after add/drop instead of the next lecture

About this course:

Evaluation and grading



Mini-projects - **50%** {group assignments}

- Four programming assignments to be done in groups of three*, *no exception to this given the grading load on TAs
- Groups can stay the same between projects, you can also regroup when needed
- The goal is not to divide and conquer but to collaborate, do not wait for others to complete their tasks, help eachother do all the parts in the assignment
- All group members receive the same mark unless there are major complaints on not contributing, responding, etc. from group-mates, which will be resolved on a case-by-case basis. If a significant difficulty/conflict arises, please send an email to the course email, cc the group-TA and put 'Group-TA' in the title

Late submissions

All due dates are **11:59 pm** in Montreal unless stated otherwise. **No make-up quizzes** will be given. For mini-projects, 2^k% percent will be deducted per k days of delay.

If you experience barriers (including a covid related issue) to learning in this course, submitting the projects, etc., please do not hesitate to discuss them with me directly, and please make sure to put "551 special" in the header to make sure I see your email [for general course correspondence, please use the course email: comp551mcgill@gmail.com].

As a point of reference, you can reach the Office for Students with Disabilities at 514-398-6009

Code of Conduct

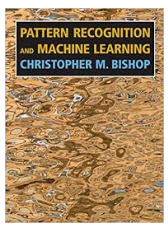
- Do not share or (re)post any of the course materials online. This includes: video lectures, codes, quizzes, zoom links, etc.
- Be respectful in the course forums and other communications
- Submit your own work for projects and quizzes

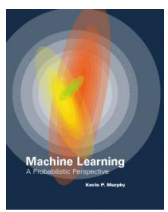
Academic Integrity

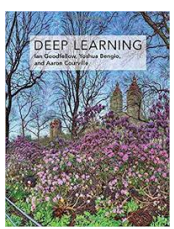
The ``McGill University values academic integrity. Therefore, all students must understand the meaning and consequences of cheating, plagiarism and other academic offenses under the Code of Student Conduct and Disciplinary Procedures'' (see McGill's webpage for more information). (Approved by Senate on 29 January 2003)

Relevant Textbooks

No required textbook but slides will cover chapters from the following books, all available online, which can be used as reference materials.







[Bishop] Pattern Recognition and Machine Learning by Christopher Bishop (2007), available online

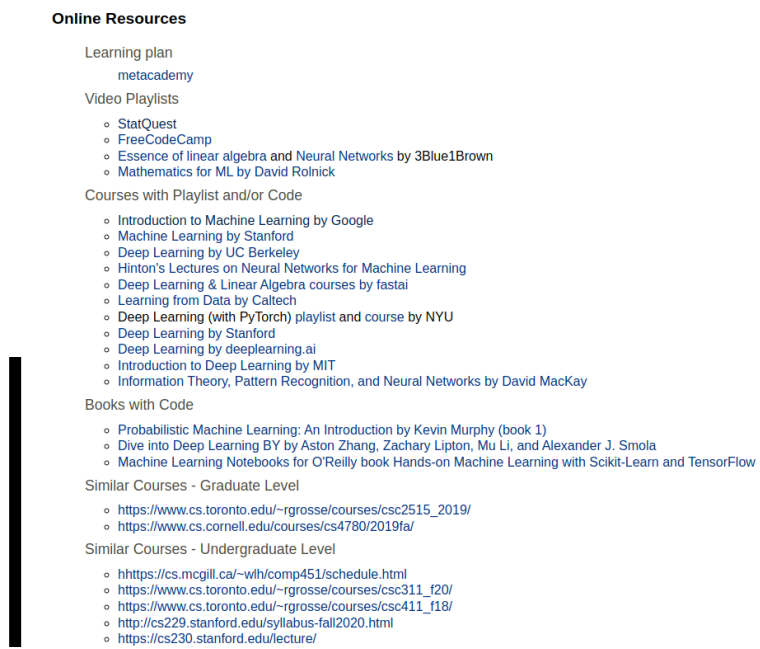
[Murphy] Machine Learning: A Probabilistic Perspective by Kevin Murphy (2012), available online through the library

[GBC] Deep Learning (2016) by Ian Goodfellow, Yoshua Bengio, and Aaron Courville, available online

Resources

Numerous great online resources at different levels, a selection is listed on the course website

Some may be more accessible than this course since they are designed for a different audience, but please note that this is a course designed for graduate students in computer science without ML background, with a heavy theory component.



Online Resources

Learning plan

- metacademy

Video Playlists

- StatQuest
- FreeCodeCamp
- Essence of linear algebra and Neural Networks by 3Blue1Brown
- Mathematics for ML by David Rolnick

Courses with Playlist and/or Code

- Introduction to Machine Learning by Google
- Machine Learning by Stanford
- Deep Learning by UC Berkeley
- Hinton's Lectures on Neural Networks for Machine Learning
- Deep Learning & Linear Algebra courses by fastai
- Learning from Data by Caltech
- Deep Learning (with PyTorch) playlist and course by NYU
- Deep Learning by Stanford
- Deep Learning by deeplearning.ai
- Introduction to Deep Learning by MIT
- Information Theory, Pattern Recognition, and Neural Networks by David MacKay

Books with Code

- Probabilistic Machine Learning: An Introduction by Kevin Murphy (book 1)
- Dive into Deep Learning BY by Aston Zhang, Zachary Lipton, Mu Li, and Alexander J. Smola
- Machine Learning Notebooks for O'Reilly book Hands-on Machine Learning with Scikit-Learn and TensorFlow

Similar Courses - Graduate Level

- https://www.cs.toronto.edu/~rgrosse/courses/csc2515_2019/
- https://www.cs.cornell.edu/courses/cs4780/2019fa/

Similar Courses - Undergraduate Level

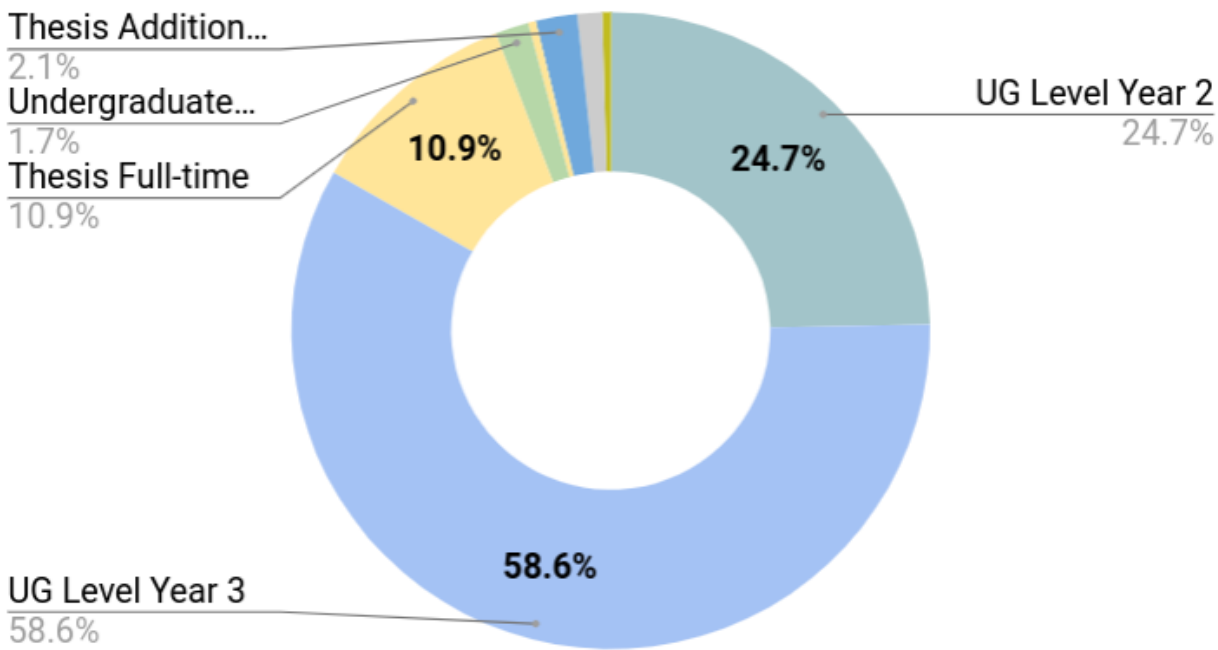
- hhttps://cs.mcgill.ca/~wlh/comp451/schedule.html
- https://www.cs.toronto.edu/~rgrosse/courses/csc311_f20/
- https://www.cs.toronto.edu/~rgrosse/courses/csc411_f18/
- http://cs229.stanford.edu/syllabus-fall2020.html
- https://cs230.stanford.edu/lecture/
- Cheatsheets: https://stanford.edu/~shervine/teaching/

Who is in this class? You

250 registered

Mostly undergraduates year 3

Mostly with Computer background



Who is in this class? Me

Reihaneh Rabbany

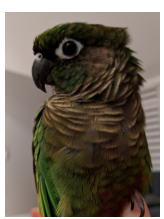
Canada CIFAR AI Chair and core member at Mila

Assistant Professor in the School of Computer Science

http://www.reirab.com/

Had CMPUT 551 Winter 2009 with Enrl of 9!

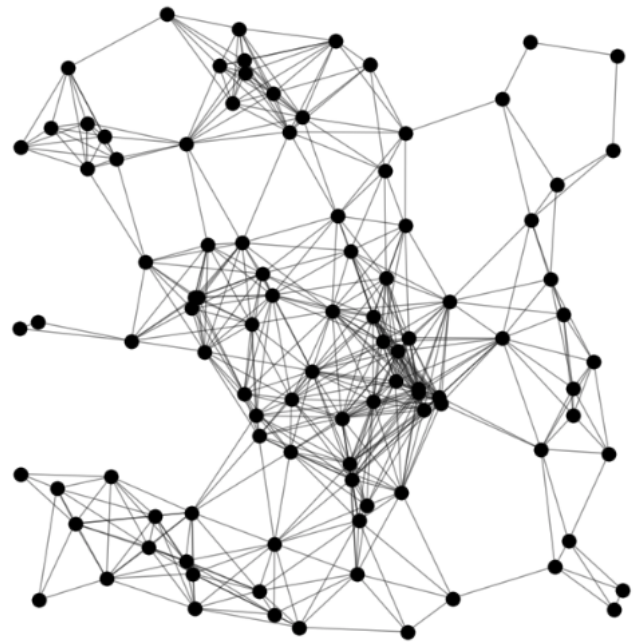
We might have some background noise



Who is in this class? Me

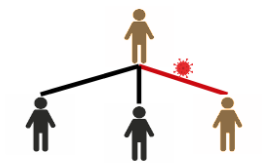
My research is on Network science, data mining and machine learning, with a focus on analyzing real-world interconnected data, and social good applications.

- Physics (complex systems)
- Sociology (social networks)
- Mathematics (graph theory)
- Data Mining (graph mining)
- Machine Learning (relational learning, graph neural networks)



Who is in this class? Me

AI4Good applications where mining connections is the key



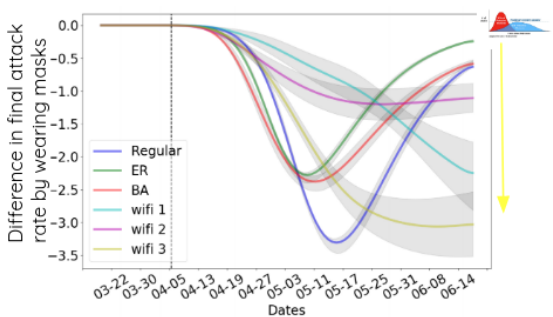
- Interconnected populations
  - Modelling covid-19 by incorporating contact graphs and flight networks
- Interconnected crime
  - Detecting organized human trafficking in online escort markets
- Interconnected discussions
  - Detecting coordinated groups in online political discourse

Questions?

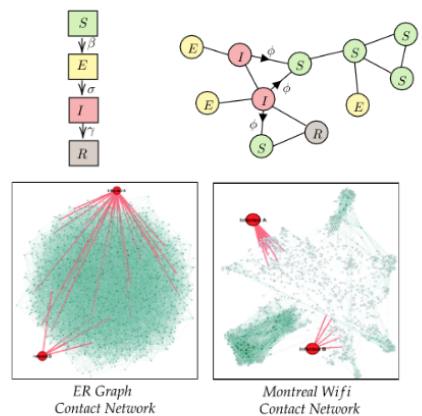
Application Highlights - COVID

structure of the contact networks significantly changes the epidemic curves, effectiveness of different NPIs, and outcomes of re-opening strategies

Classic compartment based models assume a random structure



A significantly delayed peak is the result of random mixing assumption, which also predicts wrongly that the final attack rate stays similar with or without NPIs

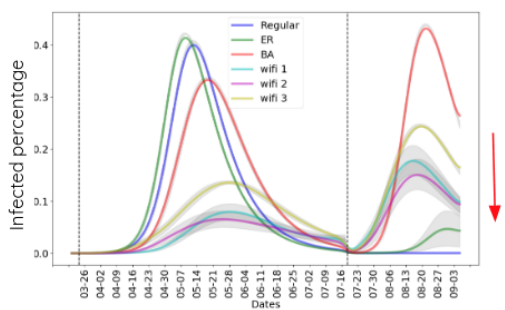


Leung A, Ding X, Huang S, Rabbany R. Contact Graph Epidemic Modelling of COVID-19 for Transmission and Intervention Strategies. arXiv preprint arXiv:2010.03081. 2020 Oct 6.

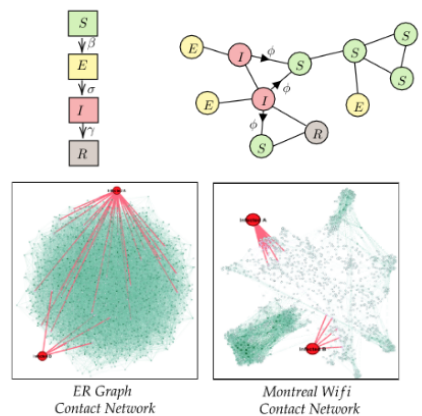
Application Highlights - COVID

structure of the contact networks significantly changes the epidemic curves, effectiveness of different NPIs, and outcomes of re-opening strategies

Classic compartment based models assume a random structure



ER graph significantly underestimates the second peak after reopening public places, i.e. allowing back hubs.

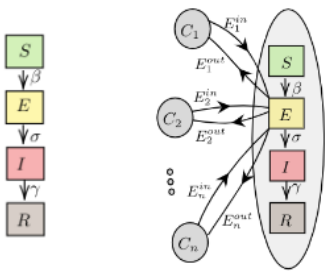


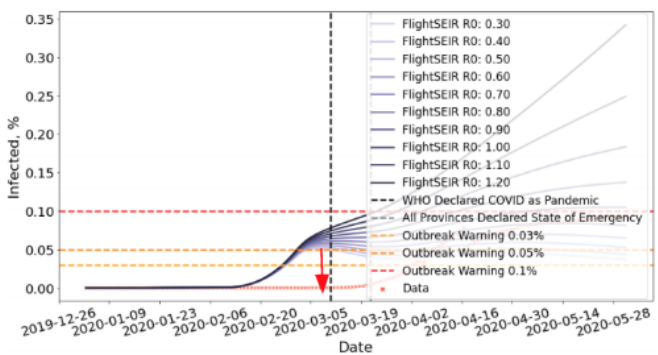
Leung A, Ding X, Huang S, Rabbany R. Contact Graph Epidemic Modelling of COVID-19 for Transmission and Intervention Strategies. arXiv preprint arXiv:2010.03081. 2020 Oct 6.

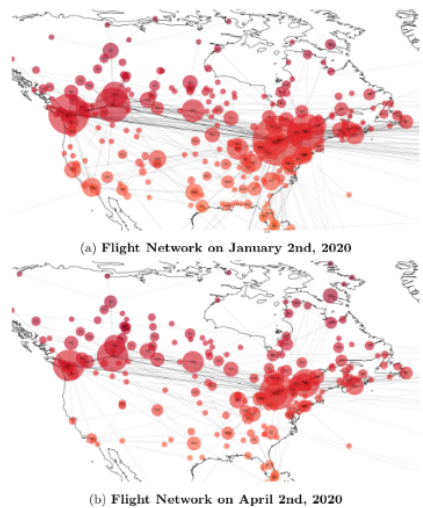
Application Highlights - COVID

Incorporating between population movements allows early detection of outbreaks, more accurate estimation of the reproduction number, and evaluation of the impact of travel restrictions and the implications of lifting them

Classic compartment based models assume a closed population







(a) **Flight Network on January 2nd, 2020**

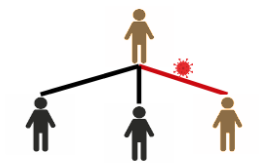
(b) **Flight Network on April 2nd, 2020**

Figure 1: **Flight network before and after imposing travel restrictions**

Ding X, Huang S, Leung A, Rabbany R. Incorporating Dynamic Flight Network in SEIR to Model Mobility between Populations. arXiv preprint arXiv:2010.01408. 2020 Oct 3.

Who is in this class? Me

AI4Good applications where mining connections is the key



- Interconnected populations
  - Modelling covid-19 by incorporating contact graphs and flight networks
- **Interconnected crime**
  - Detecting organized human trafficking in online escort markets
- Interconnected discussions
  - Detecting coordinated groups in online political discourse

Scope and Scale: Modern Slavery

$150bn yearly profit & 21 million victims worldwide [UN Office on Drugs and Crimes]

- 71% women & girls, 29% Men & boys
- 54% sexual exploitation, 38% forced labour, 8% other including organ trafficking
- A majority of victims being advertised online

Trafficking in persons in Canada

- Ninety-three per cent of sex trafficking victims are Canadian citizens
- Number of police-reported incidents of human trafficking on the rise
- Victims of human trafficking most often young women, 1 in 4 are underage

Human Trafficking in Canada

2017

Quebec: **12 million dollars**
for 2017-2022

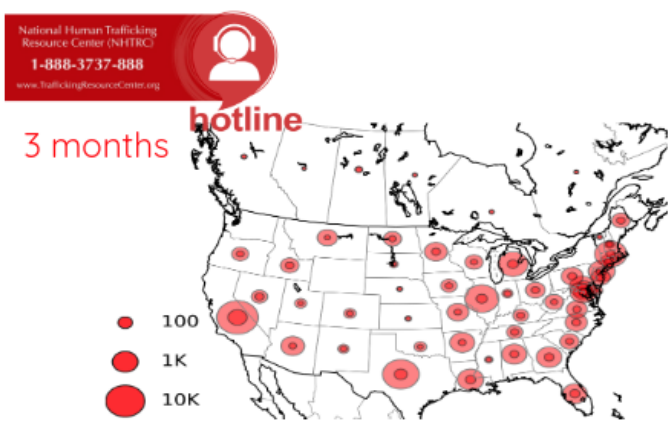
2020

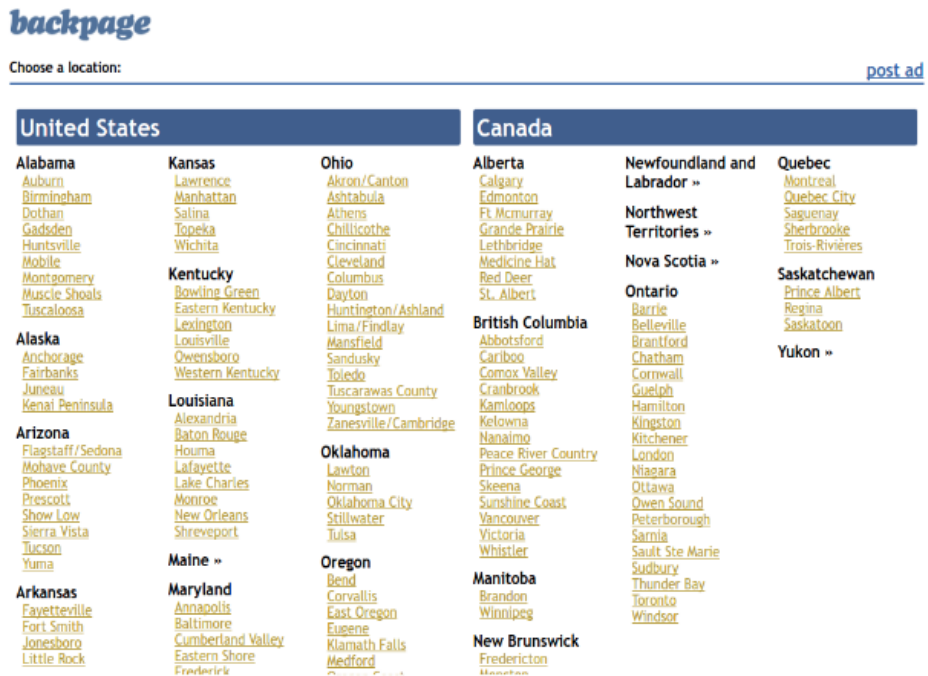
https://www.publicsafety.gc.ca/cnt/cntrng-crm/hmn-trffcknq/index-en.aspx
National Strategy To Combat Human Trafficking 2019-2024
https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/2019-ntnl-strtgy-hmnn-trffc/index-en.aspx

Online Escort Markets

Majority of sex-trafficking victims are advertised online

⇒ Use AI to monitor online “Escort” Markets



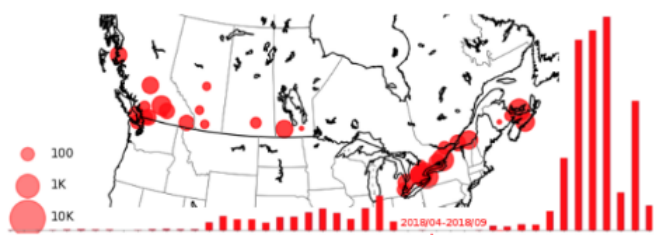


AI for Combating Human Trafficking (CTH)

Market is moving to alternative sites, getting more complicated

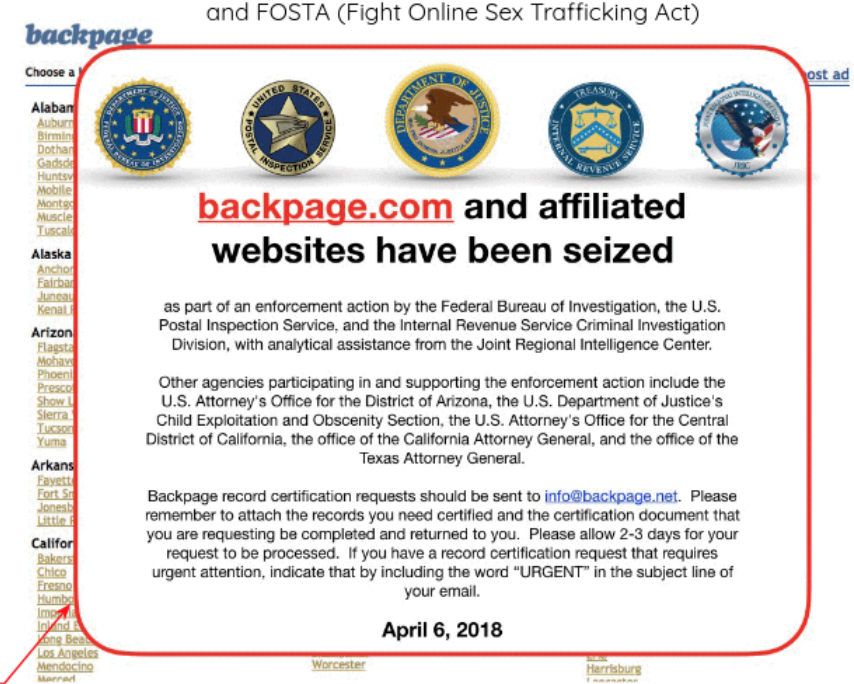
Canadian Human Trafficking Hotline call 1-833-900-1010

Launched on May 2019



Sample of 20K ads from two of the new hubs

April 11, 2018: SESTA(Stop Enabling Sex Traffickers Act) and FOSTA (Fight Online Sex Trafficking Act)


backpage

backpage.com and affiliated websites have been seized

as part of an enforcement action by the Federal Bureau of Investigation, the U.S. Postal Inspection Service, and the Internal Revenue Service Criminal Investigation Division, with analytical assistance from the Joint Regional Intelligence Center.

Other agencies participating in and supporting the enforcement action include the U.S. Attorney's Office for the District of Arizona, the U.S. Department of Justice's Child Exploitation and Obscenity Section, the U.S. Attorney's Office for the Central District of California, the office of the California Attorney General, and the office of the Texas Attorney General.

Backpage record certification requests should be sent to info@backpage.net. Please remember to attach the records you need certified and the certification document that you are requesting be completed and returned to you. Please allow 2-3 days for your request to be processed. If you have a record certification request that requires urgent attention, indicate that by including the word "URGENT" in the subject line of your email.

April 6, 2018

Trafficking is Organized

An average pimp has control 4 to 6 victims many of whom report of having “no input into the wording used in the advertisements”

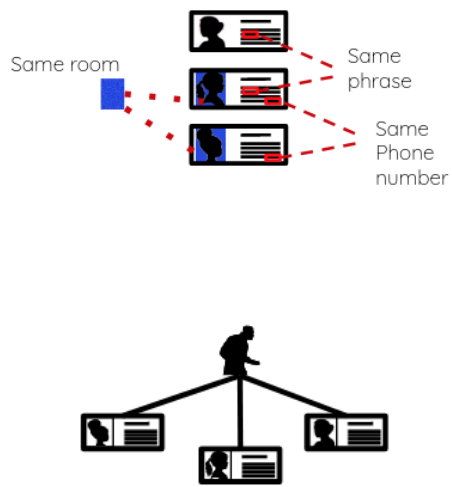
Example Trafficking Cases using Online ads

2014

February 2009 and October 2011

2018

June 2014 and July 2015



Related CHT Projects

Finding & Characterizing Micro Clusters

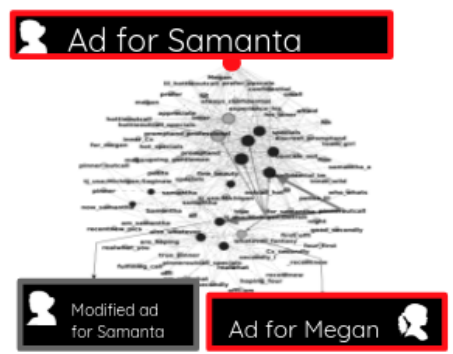
NER from Noisy Adversarial Data

Crawling and Storing data in Graph Databases

Background Matching

Active Learning for Connection Inference

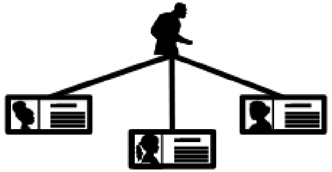
Hotels-50K: A Global Hotel Recognition Dataset

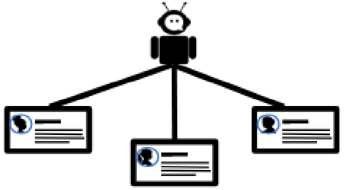


Rabbany, Reihaneh, David Bayani, and Artur Dubrawski. "Active search of connections for case building and combating human trafficking." Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018.

Detecting Micro-clusters of Text

- Organized Trafficking in Escort Ads
- Identifying Twitter Bots





Constant | Slot | Insertion | Deletion | Substitution

T_1		**sismo**	**richter**		**km al sureste de puerto escondido oax lat lon pf km**
#1		sismo	richter		km al sureste de puerto escondido oax lat lon pf km
Omit	21 Identical	Tweets as #1 ...			
#23	sismologicomx	sismo	magnitud	loc	km al sureste de puerto escondido oax lat lon pf km

TABLE IX: INFOSHIELD *is language-independent:* Spanish template from Twitter dataset.

Constant | Slot | Insertion | Deletion | Substitution

T_1	**the**	**mostpopular**	**most popular**	**stories on pr daily this week from**	*	**are**	*
#1	the		most popular	stories on pr daily this week from	instagram to mr t and perhaps even your grocers produce httpicokbfwdfis		
...							
#14	the		most popular	stories on pr daily this week from	new cover photo rules on facebook and a battle of the soci httpicoouetyugbku		
#15	the	mostpopular		stories on pr daily this week from	whimsical words to hillarys texts here	are	this weeks mos httpteoymwflapn
...							
#27	the	mostpopular		stories on pr daily this week from	understanding sopa to dating a pr professional here	are	the httpteoplace

TABLE X: INFOSHIELD *detects slots:* template from Twitter dataset.

Constant | Slot | Insertion | Deletion | Substitution

T_1	**not shown for victim's safety**							
#1	(empty)		time		(empty)		(empty)	
#2	personal description		time		(empty)		cost	
#3	(empty)		time		(empty)		cost	
#4	personal description		(empty)		preferences		cost	
...18 similar ads								

TABLE XI: *Slots contain user-specific information:* template from HT dataset.

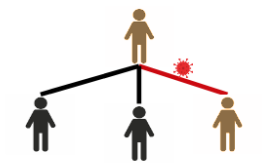
MARINUS ANALYTICS

Carnegie Mellon University

Meng-Chieh Lee, Catalina Vajiac, Aayushi Kulshrestha, Sacha Levy, Namyong Park, Cara Jones, Reihaneh Rabbany, and Christos Faloutsos,
InfoShield: Generalizable Information-Theoretic Human-Trafficking Detection, IEEE International Conference on Data Engineering (ICDE) 2021

Who is in this class? Me

AI4Good applications where mining connections is the key



- Interconnected populations
  - Modelling covid-19 by incorporating contact graphs and flight networks
- Interconnected crime
  - Detecting organized human trafficking in online escort markets
- **Interconnected discussions**
  - Detecting coordinated groups in online political discourse

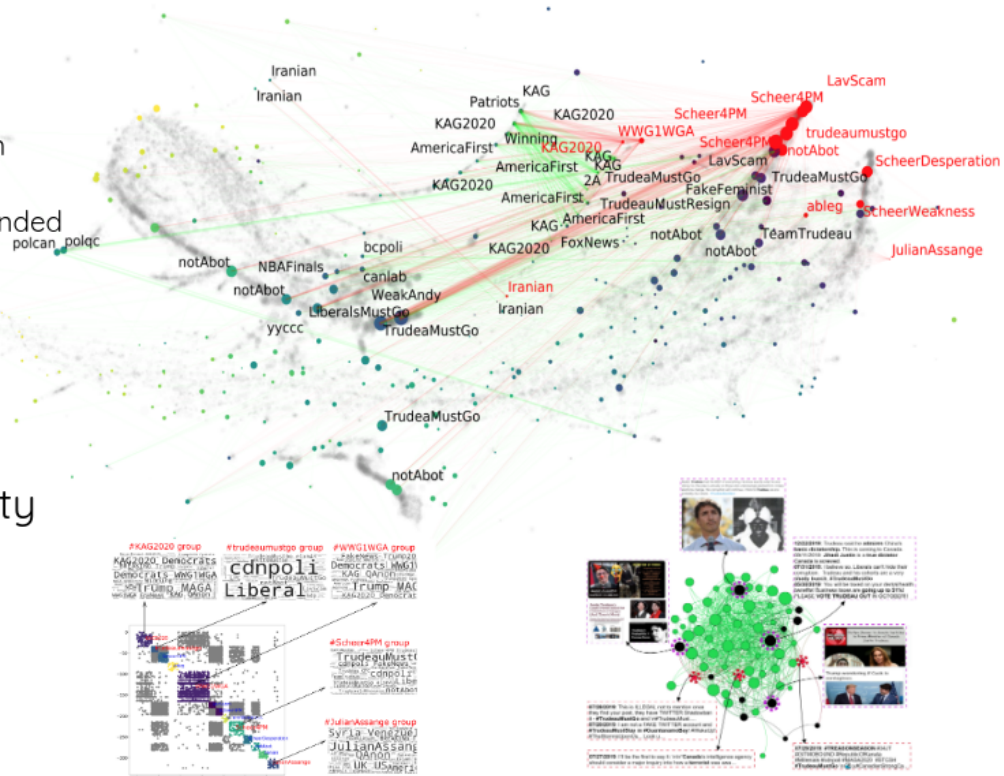
Being [Politically] Smart in the Age of Social Media

Spotting Coordinated Groups in Social Media

- 60 thousand user accounts with 3.4 million followership connections, and 1.3 million unique hashtags engaged in twitter around the 2019 Canadian Federal Elections
- users in coordinated groups 4x more likely to get suspended
- hashtags which characterize their creed linked to misinformation campaigns

Understanding Troll behavior

Studying Polarization and Polarizing Activity



Wang J, Levy S, Wang R, Kulshrestha A, Rabbany R. SCG: Spotting Coordinated Groups in Social Media. arXiv preprint arXiv:1910.07130. 2019 Oct 16.