CMPUT 466
Machine Learning: Day 1

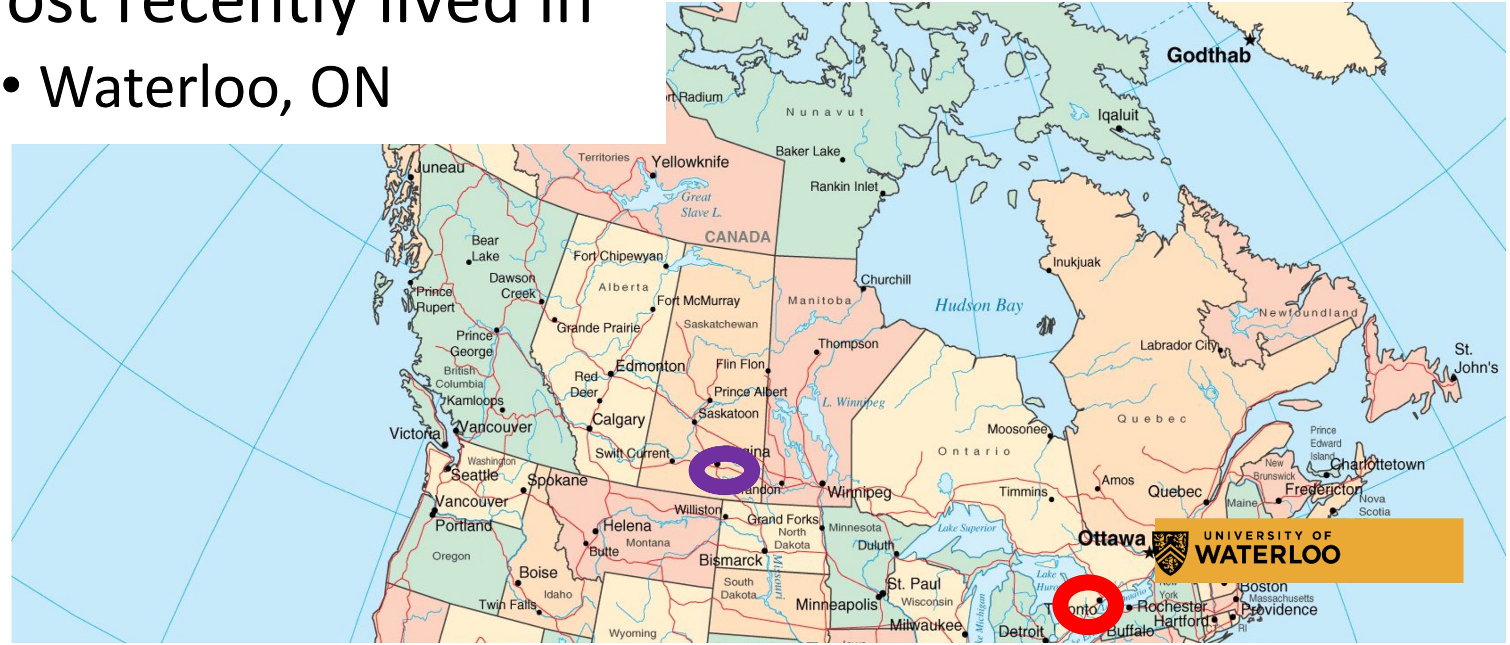
Professor: Bailey Kacsmar
kacsmar@ualberta.ca

Winter 2024

Many of these slides are derived from Alona Fyshe. Thanks!

Me

- From Saskatchewan
- Most recently lived in
  - Waterloo, ON



About and Info…

Bailey Kacsmar
Email: kacsmar@ualberta.ca
Office: Ath 3-17 (*scheduled* office hours - in person)
Office Hours:

Tuesdays 2-3 PM MT (starting next week)

Or by appointment.

I study privacy

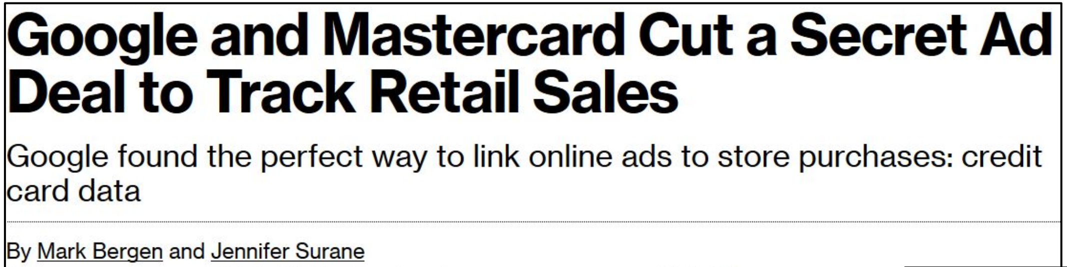
AI in Society

Entertainment Recommender Systems

Social Media Platforms

Fraud Detection for Financial Institutions

Data, Beyond the Abstraction

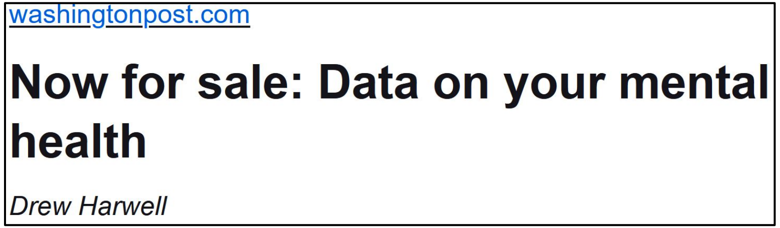


Google and Mastercard Cut a Secret Ad Deal to Track Retail Sales

Google found the perfect way to link online ads to store purchases: credit card data

By Mark Bergen and Jennifer Surane

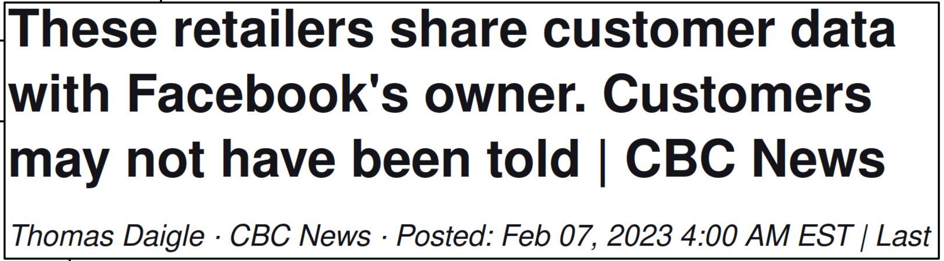
August 30, 2018, 3:43 PM EDT *Updated on August 31, 2018, 12:40 PM EDT*



washingtonpost.com

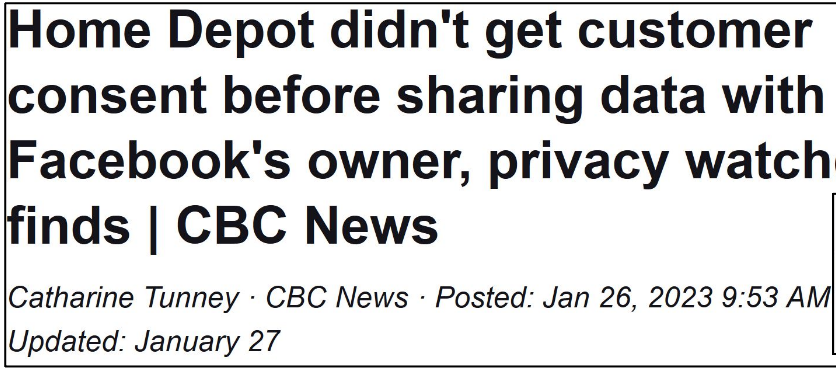
Now for sale: Data on your mental health

Drew Harwell



These retailers share customer data with Facebook's owner. Customers may not have been told | CBC News

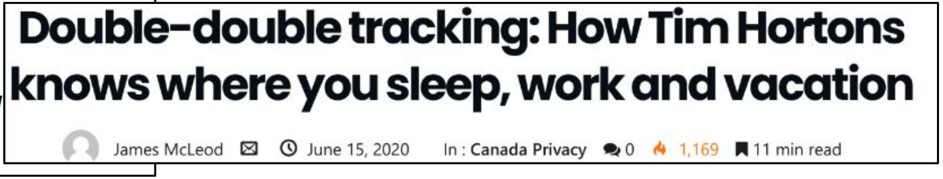
Thomas Daigle · CBC News · Posted: Feb 07, 2023 4:00 AM EST | Last



Home Depot didn't get customer consent before sharing data with Facebook's owner, privacy watchdog finds | CBC News

Catharine Tunney · CBC News · Posted: Jan 26, 2023 9:53 AM

Updated: January 27



Double-double tracking: How Tim Hortons knows where you sleep, work and vacation

James McLeod ✉ June 15, 2020 In : Canada Privacy 0 1,169 11 min read

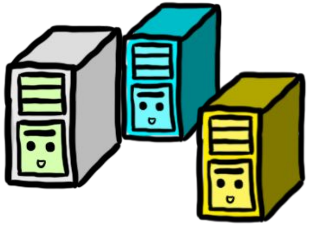
The Use of Health Data

P20986: “It depends. I think it can be beneficial **under certain circumstances**, but I would be hesitant having any healthcare data shared outside my practitioners. However, I recognize how it can improve goods/services, but there **has to be a lot of protection** in place **anytime data is shared**"

P94865: “**Repugnant**, especially in light of for profit health systems attempting to maximize profitability from patient interactions"

B. Kacsmar, K. Tilbury, M. Mazmudar, and F. Kerschbaum. "Caring about Sharing: User Perceptions of Multiparty Data Sharing." In 31st USENIX Security. 2022.

Data and Abstraction



A company wants to analyze data

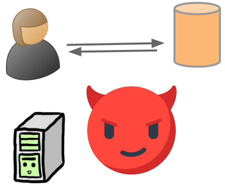
→

But the data has privacy implications for the data subjects

→

Researchers develop technical solutions

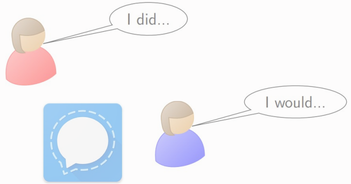
Technical Privacy



Technical Privacy

Conceptual Privacy

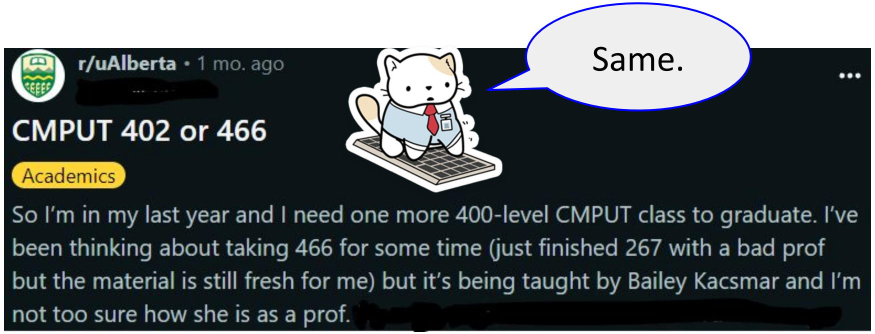
Legal Privacy



Usable Privacy

Define, **what** is being protected, from **who**, and under what **conditions** this protection will hold.

Why are you here?



r/uAlberta • 1 mo. ago

[redacted]

CMPUT 402 or 466

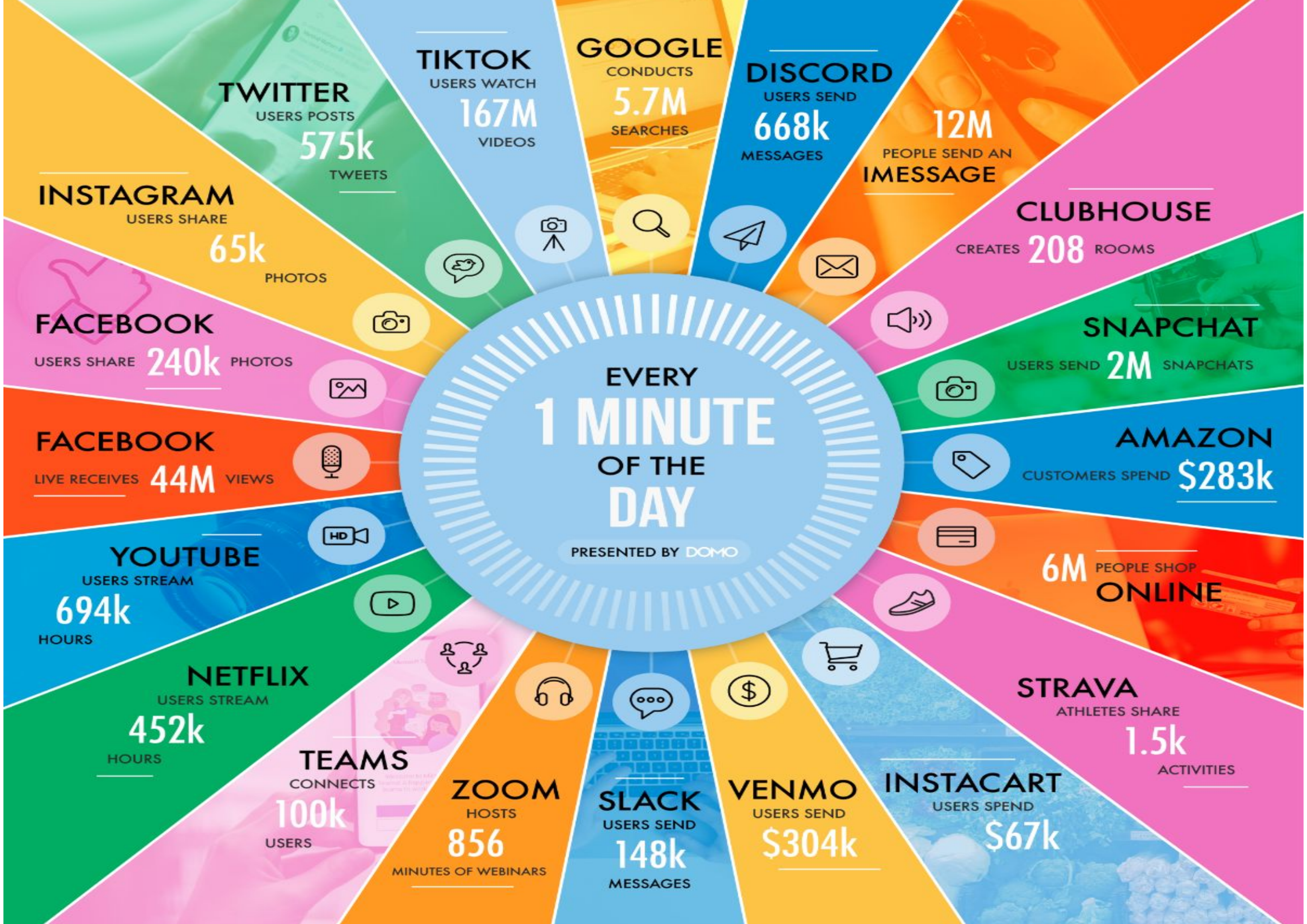
Academics

So I'm in my last year and I need one more 400-level CMPUT class to graduate. I've been thinking about taking 466 for some time (just finished 267 with a bad prof but the material is still fresh for me) but it's being taught by Bailey Kacsmar and I'm not too sure how she is as a prof. [redacted]

Why are you here?

Why is this course here?

Data is everywhere all the time



https://www.domo.com/learn/data-never-sleeps-9

Detrimental (?) Deluge of Data

- Youtube: 500 **hours** of video uploaded every **minute**

- Recommending videos becomes very important!
  - What happens if recommendations are bad?
  - What if they are good?

- Certain videos keep you on Youtube longer… but is that a good thing?
  - What sort of videos keep you watching?

Data is everywhere all the time

No team of staff is large enough to handle the data manually (clearly)

That’s where Machine Learning comes in

Machine Learning

- This semester we will be discussing how to use **computer models** to **explore and understand data**, and **generalize to new data**

Machine Learning

- This semester we will be discussing how to use **computer models** to **explore and understand data**, and **generalize to new data**
- This semester we will learn very powerful techniques for leveraging that data

Machine Learning

- This semester we will be discussing how to use **computer models** to **explore and understand data**, and **generalize to new data**
- This semester we will learn very powerful techniques for leveraging that data
- **Remember**: behind a lot of data (including most of the data you see online) there are **people**.

So…What is Machine Learning?

Origins of Machine Learning

- Has its roots in AI
- Draws ideas from: Statistics, computing science, psychology, neuroscience…

Types of Machine Learning

ML is generally divided into:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Types of Machine Learning

1) Supervised Learning, Predictive tasks

[Use some attributes to predict unknown or future values of other attributes.]

- Classification
- Regression
- e.g. distinguish spam from non-spam email

Types of Machine Learning

2) Unsupervised Learning, Descriptive tasks

[Find human-interpretable patterns that describe the data.]

- Recommender Systems
- Clustering
- Self-supervised learning
- GANs (Generative Adversarial Nets)
- e.g. cluster emails by topic (school, friends, etc), generate an email that doesn’t look like spam

Types of Machine Learning

3) Reinforcement Learning

[Learn to ”interact” with an “environment” to work towards a “goal”]

- Model based and model-free variants
- E.g. learn to play Atari games

Supervised Learning

- Given a collection of records (***training set***)
  - Each record contains a set of ***attributes/features***, one of which is the ***class***.
  - e.g. for emails:
    - the features could be the words of the email
    - class is spam/not spam
- Find ("learn") a ***model*** to predict the class attribute as a ***function*** of the features.
  - e.g. predict spam/not spam based on the words of the email
- Goal: Accurately assign class **previously unseen** records.
  - [Generalize]

Learning

We can think of at least three different problems being involved in learning:

- Memory – memorize the data exactly
- Averaging – learn a simple summary of the data
- Generalization – learn to use the data in a more complex way that (hopefully!) works better for new examples

Example problem

(Adapted from Leslie Kaelbling's example in MIT courseware)

- Imagine I'm trying predict whether my neighbor is going to drive into work, so I can ask for a ride.

- Whether she drives into work seems to depend on the following attributes of the day:
  - temperature
  - expected precipitation
  - day of the week
  - what she's wearing

Memory

- Okay. Let's say we observe our neighbor on three days:

Temp	**Precip**	**Day**	**Clothes**	
25	None	Sat	Casual	**Walk**
-5	Snow	Mon	Casual	**Drive**
15	Snow	Mon	Casual	**Walk**

Memory

- Now, we find ourselves on a snowy “-5” degree Monday, and the neighbor is wearing casual clothes.
- Do you think she's going to drive?

Temp	**Precip**	**Day**	**Clothes**	
25	None	Sat	Casual	**Walk**
-5	Snow	Mon	Casual	**Drive**
15	Snow	Mon	Casual	**Walk**
-5	**Snow**	**Mon**	**Casual**	

Memory

- Standard answer in this case is "yes".
  - This day is just like one of the ones we've seen before, and so it seems like a good bet to predict "yes."
- This is the most rudimentary form of learning, which is just to memorize the things you've seen before.

Temp	Precip	Day	Clothes	
25	None	Sat	Casual	**Walk**
-5	Snow	Mon	Casual	**Drive**
15	Snow	Mon	Casual	**Walk**
-5	Snow	Mon	Casual	**Drive**

Noisy Data

Things aren’t always that easy. What if you get this set of noisy data?

Temp	**Precip**	**Day**	**Clothes**	
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Drive**
25	None	Sat	Casual	**Drive**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	**None**	**Sat**	**Casual**	**?**

We have certainly seen this case before, but the problem is that it has had different answers. Our neighbor is not entirely reliable.

Averaging

One strategy would be to predict the majority outcome.

Temp	Precip	Day	Clothes	
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Drive**
25	None	Sat	Casual	**Drive**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**

Generalization

- Dealing with previously unseen cases
- **Will she walk or drive?**

Temp	**Precip**	**Day**	**Clothes**	
22	None	Fri	Casual	**Walk**
3	None	Sun	Casual	**Walk**
10	Rain	Wed	Casual	**Walk**
30	None	Mon	Casual	**Drive**
20	None	Sat	Formal	**Drive**
25	None	Sat	Casual	**Drive**
-5	Snow	Mon	Casual	**Drive**
27	None	Tue	Casual	**Drive**
24	**Rain**	**Mon**	**Casual**	**?**

Generalization

- Dealing with previously unseen cases
- **Will she walk or drive?**

Temp	Precip	Day	Clothes	
22	None	Fri	Casual	**Walk**
3	None	Sun	Casual	**Walk**
10	Rain	Wed	Casual	**Walk**
30	None	Mon	Casual	**Drive**
20	None	Sat	Formal	**Drive**
25	None	Sat	Casual	**Drive**
-5	Snow	Mon	Casual	**Drive**
27	None	Tue	Casual	**Drive**
24	**Rain**	**Mon**	**Casual**	**?**

We might plausibly make any of the following arguments:

- **She's going to walk because it's raining today and the only other time it rained, she walked.**
- **She's going to drive because she has always driven on Mondays…**

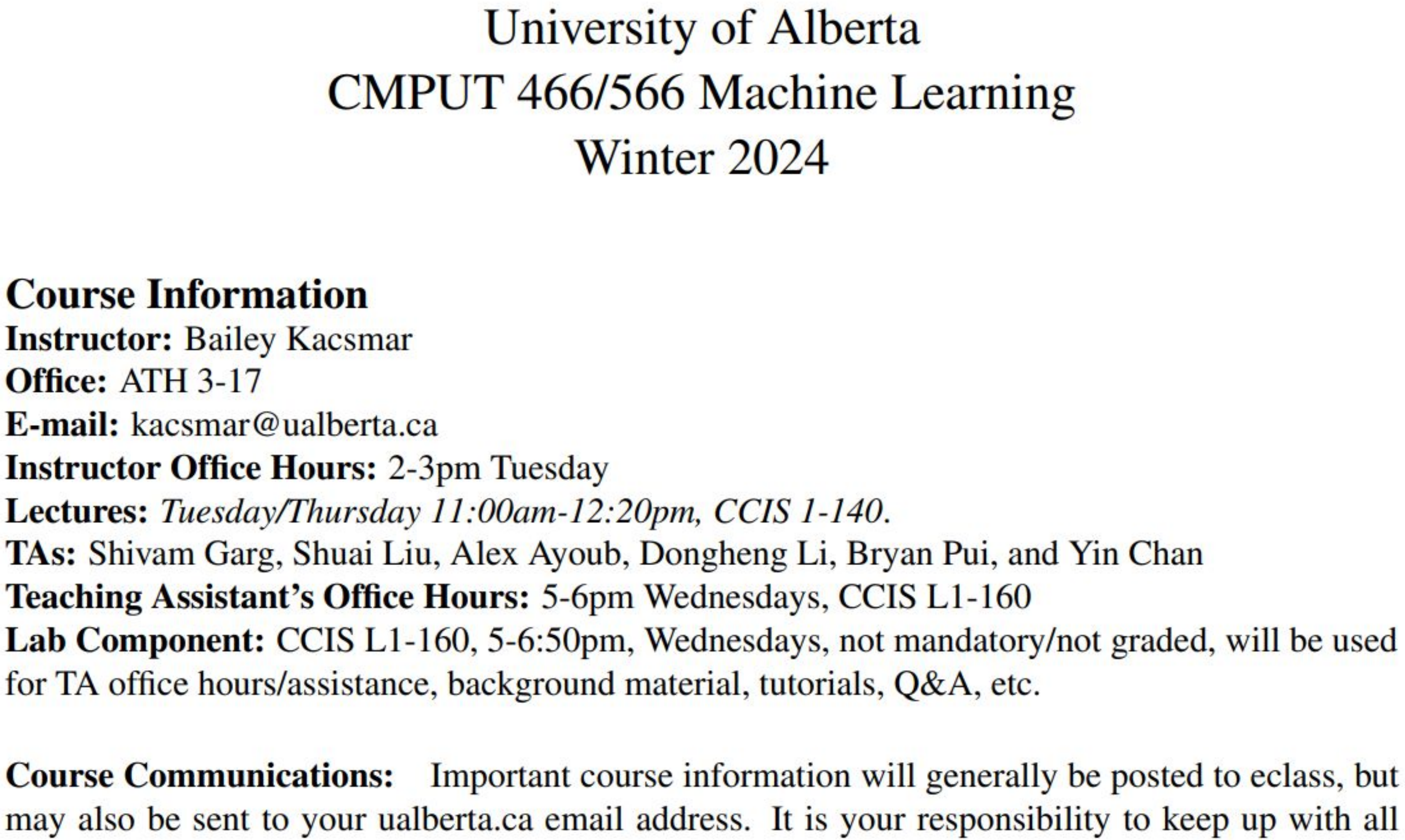
Some Class Logistics

Details/Info Dump…

- Eclass website (registered students should be able to view automatically)
- Labs are Wednesdays
  - 5-7:50 pm, CCIS L1-160
  - TAs will (likely) cycle coverage, most weeks open office hours
- Office hours: Tuesdays 2-3 PM MT
  - Or by appointment (reach out via email), can be online or in-person.
  - I have the most flexible schedule on Mondays and Wednesdays, so try and suggest a time on those days.

Where can I find information?

- Eclass (has everything)
- Syllabus
- Schedule (in the syllabus)

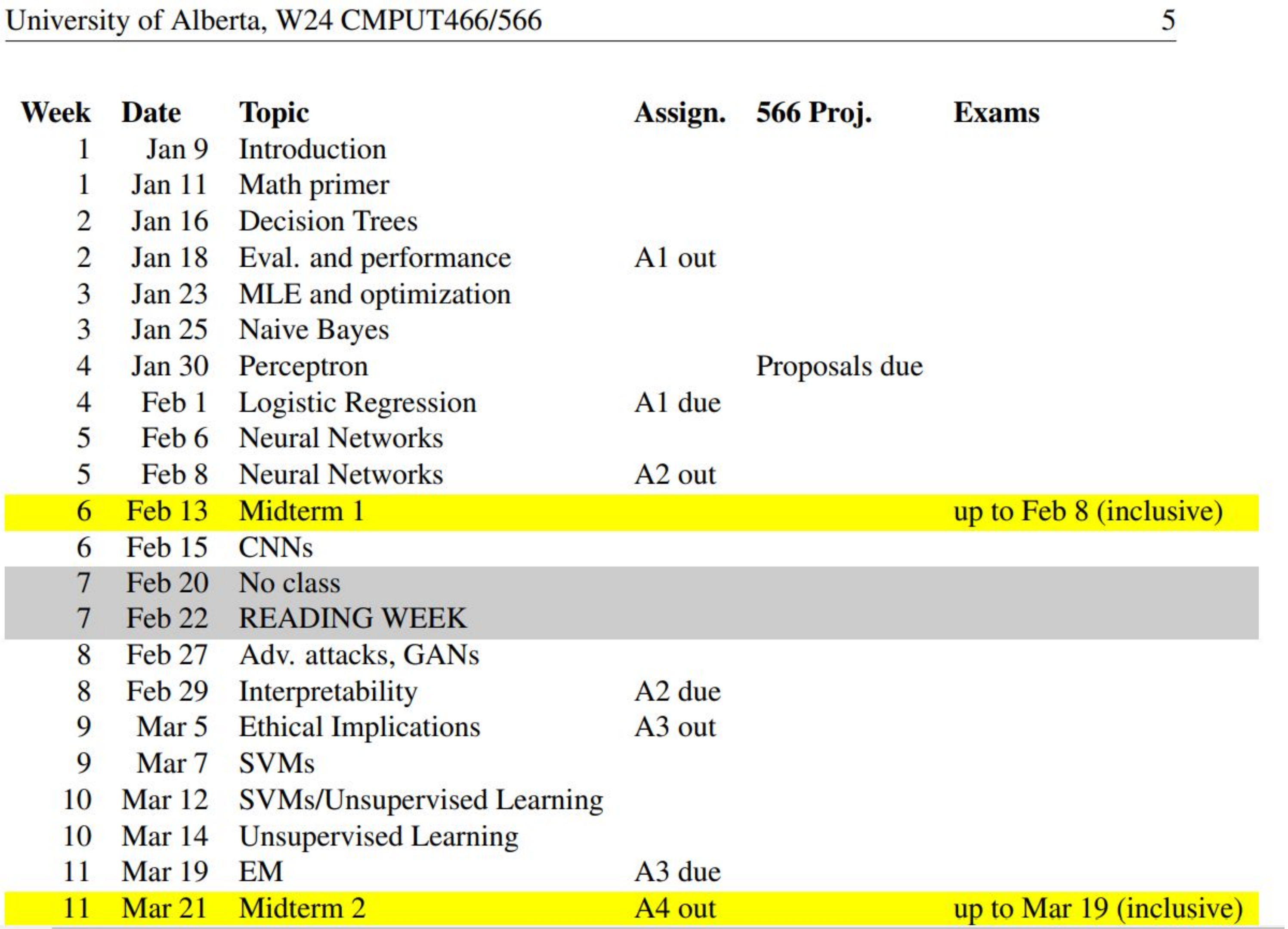


University of Alberta
CMPUT 466/566 Machine Learning
Winter 2024

Course Information

Instructor: Bailey Kacsmar
Office: ATH 3-17
E-mail: kacsmar@ualberta.ca
Instructor Office Hours: 2-3pm Tuesday
Lectures: *Tuesday/Thursday 11:00am-12:20pm, CCIS 1-140*.
TAs: Shivam Garg, Shuai Liu, Alex Ayoub, Dongheng Li, Bryan Pui, and Yin Chan
Teaching Assistant's Office Hours: 5-6pm Wednesdays, CCIS L1-160
Lab Component: CCIS L1-160, 5-6:50pm, Wednesdays, not mandatory/not graded, will be used for TA office hours/assistance, background material, tutorials, Q&A, etc.

Course Communications: Important course information will generally be posted to eclass, but may also be sent to your ualberta.ca email address. It is your responsibility to keep up with all



Week	Date	Topic	Assign.	566 Proj.	Exams
1	Jan 9	Introduction			
1	Jan 11	Math primer			
2	Jan 16	Decision Trees			
2	Jan 18	Eval. and performance	A1 out		
3	Jan 23	MLE and optimization			
3	Jan 25	Naive Bayes			
4	Jan 30	Perceptron		Proposals due	
4	Feb 1	Logistic Regression	A1 due		
5	Feb 6	Neural Networks			
5	Feb 8	Neural Networks	A2 out		
6	Feb 13	Midterm 1			up to Feb 8 (inclusive)
6	Feb 15	CNNs			
7	Feb 20	No class			
7	Feb 22	READING WEEK			
8	Feb 27	Adv. attacks, GANs			
8	Feb 29	Interpretability	A2 due		
9	Mar 5	Ethical Implications	A3 out		
9	Mar 7	SVMs			
10	Mar 12	SVMs/Unsupervised Learning			
10	Mar 14	Unsupervised Learning			
11	Mar 19	EM	A3 due		
11	Mar 21	Midterm 2	A4 out		up to Mar 19 (inclusive)

A Course on Machine Learning

- Family of algorithms/techniques to formalize the process of finding patterns in data

Important Note on Course Topic

- Machine Learning is not magic

Important Note on Course Topic

- Machine Learning is not magic
- Machine Learning is CS, math and statistics
-

Important Note on Course Topic

- Machine Learning is not magic
- Machine Learning is CS, math and statistics
- There is a lot of math…a lot.
-

Important Note on Course Topic

- Machine Learning is not magic
- Machine Learning is CS, math and statistics
- There is a lot of math…a lot.
- We will be doing math every day
-

Important Note on Course Topic

- Machine Learning is not magic
- Machine Learning is CS, math and statistics
- There is a lot of math…a lot.
- We will be doing math every day
- Some days we will also do programming
-

Important Note on Course Topic

- Machine Learning is not magic
- Machine Learning is CS, math and statistics
- There is a lot of math…a lot.
- We will be doing math every day
- Some days we will also do programming
- Then we will do statistics
-

Important Note on Course Topic

- Machine Learning is not magic
- Machine Learning is CS, math and statistics
- There is a lot of math…a lot.
- We will be doing math every day
- Some days we will also do programming
- Then we will do statistics
- Then we will do more math
-

Important Note on Course Topic

- Machine Learning is not magic
- Machine Learning is CS, math and statistics
- There is a lot of math…a lot.
- We will be doing math every day
- Some days we will also do programming
- Then we will do statistics
- Then we will do more math
- There will be math…and statistics
-

Important Note on Course Topic

- Machine Learning is not magic
- Machine Learning is CS, math and statistics
- There is a lot of math…a lot.
- We will be doing math every day
- Some days we will also do programming
- Then we will do statistics
- Then we will do more math
- There will be math…and statistics
- 2 months from now: we are still doing math
- …

Important Note on Course Topic

- Machine Learning is not magic
- Machine Learning is **CS**, **math** and **statistics**
- There is a lot of **math**…a lot.
- We will be doing **math** every day
- Some days we will also do programming
- Then we will do **statistics**
- Then we will do more **math**
- There will be **math**…and **statistics**
- 2 months from now: we are still doing **math**
- **Math** doesn't even look like a real word anymore

I think…this course might be **math**

Intro to Machine Learning:

- This class involves:
  - **Python programming**. If you are rusty, or have not used Python*, there are many online tutorials to help you get started. https://wiki.python.org/moin/BeginnersGuide/Programmers
  - **Linear Algebra****. Here is a set of short videos for review: http://www.cs.cmu.edu/~zkolter/course/linalg/
  - **Statistics and Probability****. If you need some statistics review, here are some videos: https://www.youtube.com/playlist?list=PLRCdqbn4-qwoRTW3OpaB8-GnQwr6ta756

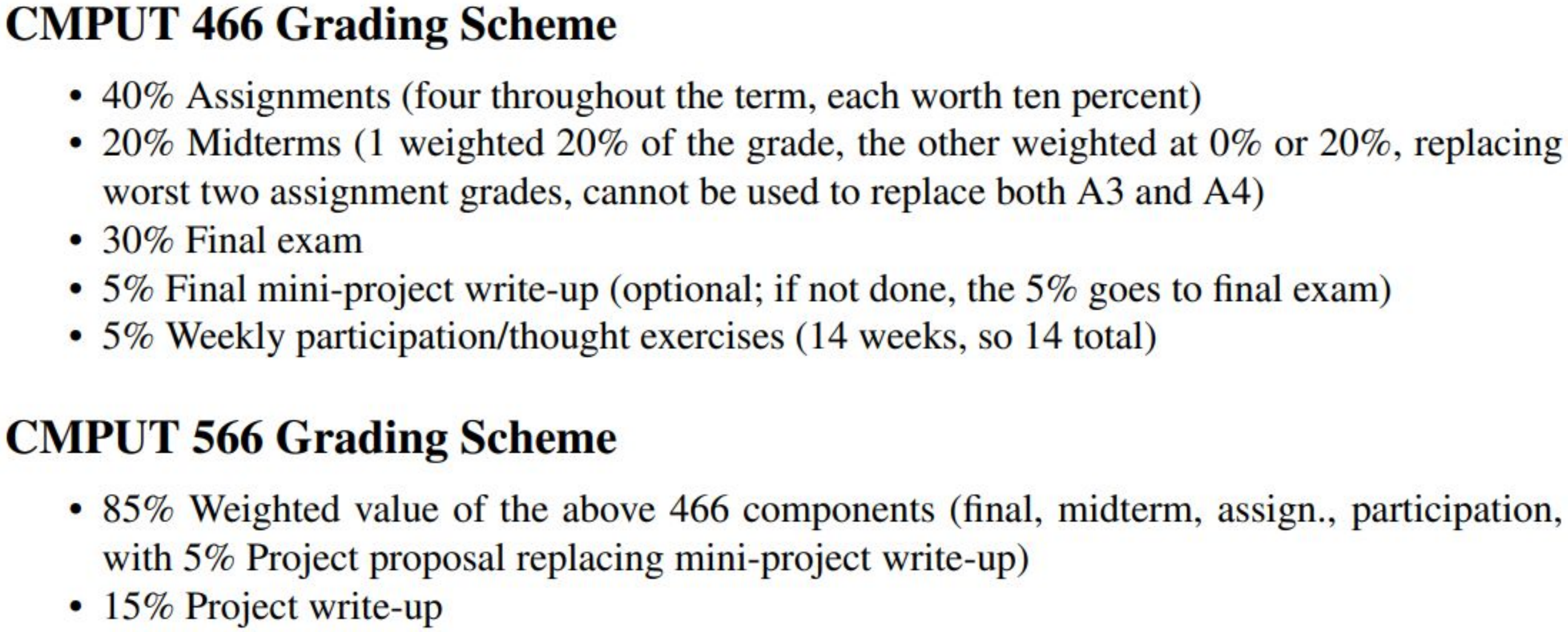
*You need to have strong programming skills for this class, but no necessarily Python experience

** We will go through some of this in class

What Will We Cover?

- Supervised Learning
  - Decision Trees
  - SVMs
  - Linear/logistic regression
  - Neural networks / deep learning
- Unsupervised learning
  - Clustering algorithms
  - Expectation Maximization (EM)
  - Neural networks / deep learning
- Reinforcement Learning

Grade Evaluation



CMPUT 466 Grading Scheme

- 40% Assignments (four throughout the term, each worth ten percent)
- 20% Midterms (1 weighted 20% of the grade, the other weighted at 0% or 20%, replacing worst two assignment grades, cannot be used to replace both A3 and A4)
- 30% Final exam
- 5% Final mini-project write-up (optional; if not done, the 5% goes to final exam)
- 5% Weekly participation/thought exercises (14 weeks, so 14 total)

CMPUT 566 Grading Scheme

- 85% Weighted value of the above 466 components (final, midterm, assign., participation, with 5% Project proposal replacing mini-project write-up)
- 15% Project write-up

Assignments

- Four assignments
- Each include some written (proofs, contemplation, etc) and programming
- You can replace your two worst assignment grades with one of the midterms (but not both A3 and A4)

Assignments

- Four assignments
- Each include some written (proofs, contemplation, etc) and programming
- You can replace your two worst assignment grades with one of the midterms (but not both A3 and A4)

Midterms/Final

- Various question styles
- Will assume you did assignments and prompts
- Best midterm grade of the two, 20%
- Second best midterm grade, can replace worst two assignment grades
- Final exam is 30% if you do the project, 35% if you do not (466 only)
- Final exam is 30% for 566

Participation/Thought Exercises

- One per week
- Worth 5% total, so ~0.35% each…minor to miss, but adds up.
- Does not mean attendance required
- Indicated in slide, easily spotted
- Helps you, more work for us

Participation/Thought Exercises

- One per week
- Worth 5% total, so ~0.38% each…minor to miss, but adds up.
- Does not mean attendance required
- Indicated in slide, easily spotted
- Helps you, more work for us

Participation Prompt

Act.

Week One: Data, Data, Everywhere…

Go through your day. For a portion of your day, write down every time your using your computer/device that you access something where you have to send data. Don’t forget the cookies.

How often was it? Of the times you went to do something using your computer how many resulted in information about you/from you going somewhere?

What types of data do you think were sent? Any surprise you?

Project

- 466 students optional (and different write-up)
- 566 students, not optional
- Both: Formulate a task into a machine learning problem. Implementation of the task.
- 566 Related work/literature very important
- Details in syllabus

Very Important Section

Plagiarism

- If you copy answers for **any part** an assignment/ project/ exam, your case will be sent to the Faculty of Science ethics committee
- For your project reports, you must cite all sources, including tables/figures.
- If you copy paste from other resources, you must make it clear that the material is verbatim from another source.
  - e.g. **using quotation marks** or indenting the text
  - https://owl.english.purdue.edu/owl/resource/747/03/

DO NOT DO THIS (please)

- I googled the question and copied the answer I found
  - This is not “research”
- I let someone copy my assignment
  - You have also violated the code of conduct
- I copied their code and changed the variable names
- **I copied text from a published research paper or webpage, but I cited the paper**
- **I copied text from a published research paper or webpage, but I changed some of the words and cited the paper**

Academic Integrity

- I don’t decide cheating cases, they go to the academic integrity committee.
- A helpful handout:

  https://www.ualberta.ca/science/media-library/studentservices/studentforms/forms-cabinet-2018/donotdo-it2018.pdf

All UofA students are responsible for following the Code of Student Behaviour

“For these freedoms* to exist, it is essential to maintain an atmosphere in which the safety, the security, and the inherent dignity of each member of the community are recognized…”

In summary, while you’re a student at the University (**even off-campus/online**):

- No cheating/plagiarism/sharing of confidential or course materials
- No disrupting class
- No discrimination or harassment
- No abuse, threats, or violence

https://www.ualberta.ca/governance/media-library/documents/resources/policies-standards-and-codes-of-conduct/cosb-updated-july-1-2020.pdf

What to do when the code of student behaviour is broken?

Report to the department (all profs and TAs are mandatory reporters if there's a crime)

- Professors (either in your class or others you trust) or TAs
- CS EDI Anonymous Form https://forms.gle/XhZGjKGdLWYNLixM8

Specific offices on campus to reach out to (who are not mandatory reporters)

- Office of the Dean of Students (this is their job)
- Office of Safe Disclosure & Human Rights
- Helping Individuals at Risk Program
- Office of the Student Ombuds
- Student Legal Services

How to help if you witness something you feel is wrong?

1. **Distract**: Diffuse or otherwise attempt to distract the student who broke the code so their focus is off the target
2. **Delegate**: Find someone with authority (TA, Prof, etc.) and ask to intervene.
3. **Document**: Watch and witness, write down or film the behaviour
4. **Direct**: Call the behaviour out. This is an important aspect of building a culture without tolerance for bad behaviour. Though only use this as a last resort if the situation is or might become violent or dangerous.
5. **Delay**: Comfort the targeted student, acknowledge the behaviour was wrong, and be a friend.

https://www.standup-international.com/ca/en/our-training/bystander

What to do if your behaviour is called out?

1. Be open to learning! You’re a student after all, and you may not have known the behaviour was harmful.
2. Apologize and endeavour not to do it again.
3. Drop the topic in the moment.
4. Educate yourself separately on the issue, don’t expect the person who called you out to educate you, that’s not their responsibility.

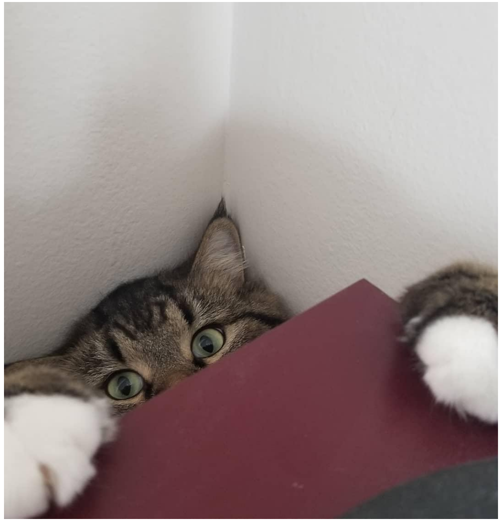
What to do if it escalates and you feel you are wrongly sanctioned?

- Office of the Student Ombuds
- Student Legal Services
- University Appeal Bodies

Thank you!

Questions? Concerns? Please contact cs.edi@ualberta.ca

Thanks, see you Thursday!



Some fun demos

https://quickdraw.withgoogle.com
https://magenta.tensorflow.org/assets/sketch_rnn_demo/index.html
http://playground.tensorflow.org/
http://nlp.stanford.edu:8080/sentiment/rntnDemo.html

https://huggingface.co/spaces/dalle-mini/dalle-mini