Introduction to Machine Learning

Instructor: Dr. Alona Fyshe
alona@ualberta.ca
Fall 2022

A mask request, not a mandate

1. **Please wear a mask in class** and indoors. N95 or better, if possible.

2. Teamwork: Masks help protect you, fellow students, and everyone's friends and family.

3. It is not possible to tell just by looking at someone, if they are:
   - immunocompromised
   - have other medical vulnerabilities (e.g., diabetes, asthma)
   - asymptomatically infected
   - Have family members who cannot be vaccinated

* Infographic via Twitter

Alona Fyshe

A lawn

A fish

Me

- From Edmonton, AB
- Most recently lived in
  - Pittsburgh, PA
  - Victoria, BC (UVic)

About Me

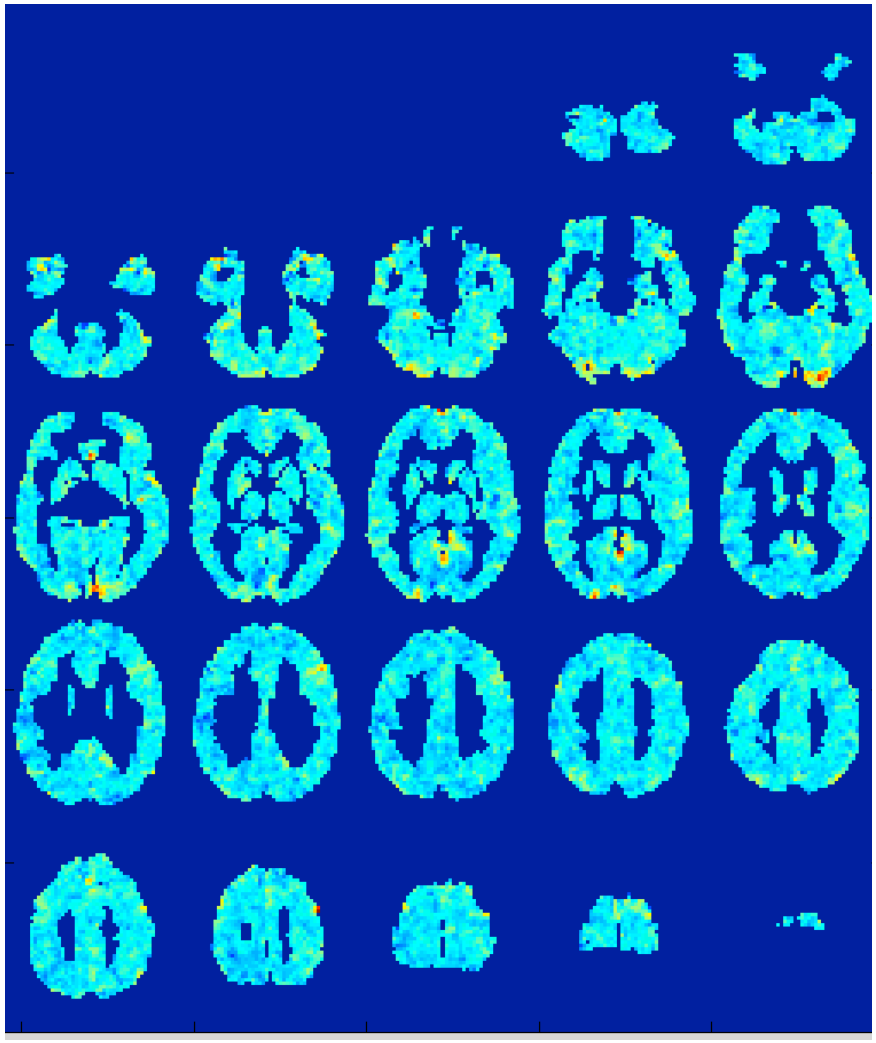
Alona Fyshe
Email: alona@ualberta.ca
Office: Ath 3-56 (*scheduled* office hours are virtual)
Office Hours:

Monday 3-4 PM MT

via gmeet: https://meet.google.com/yzo-ukgy-cea

Or by appointment (please don't hesitate to reach out, I'm happy to meet!).

I study language and the brain

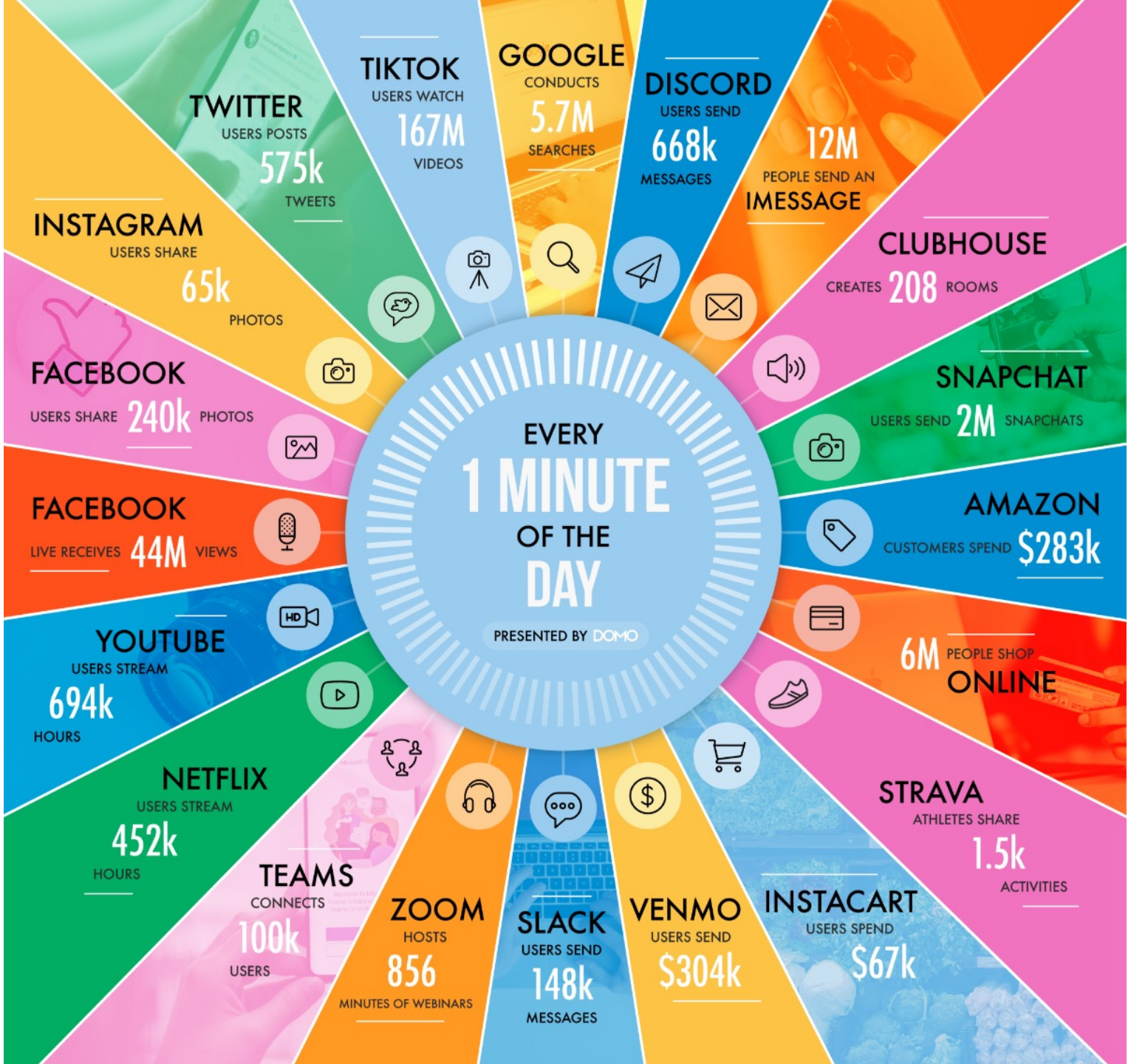


Why are you here?

Data is everywhere all the time

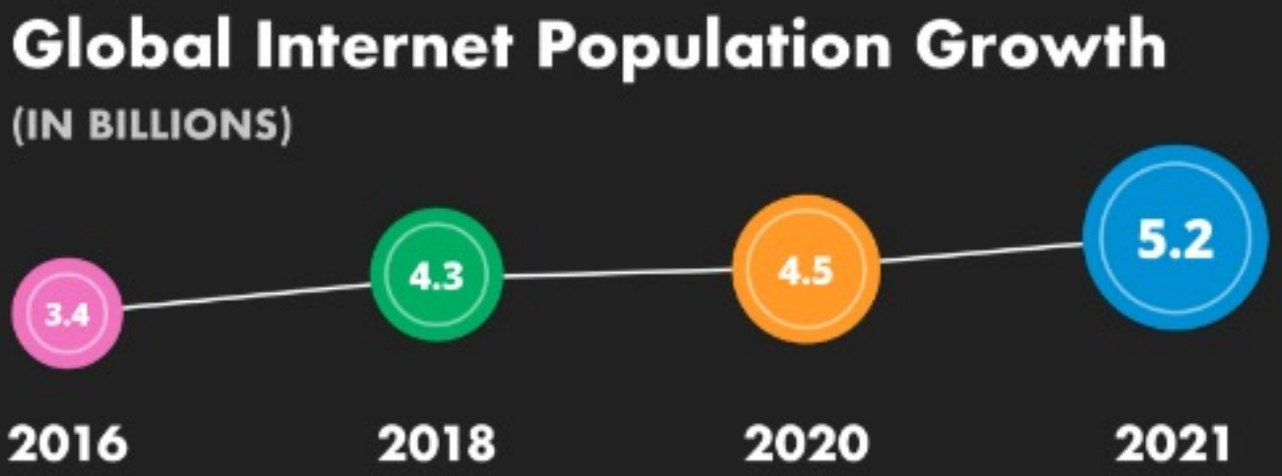
Amount of data generated every MINUTE!

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



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

As of July 2021, the internet reaches 65% of the world's population and now represents 5.17 billion people—a 10% increase from January 2021. Of this total, 92.6 percent accessed the internet via mobile devices. According to Statista, the total amount of data consumed globally in 2021 was 79 zettabytes, an annual number projected to grow to over 180 zettabytes by 2025.



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

Deluge of Data

- 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

- There's a lot of data out there!

Machine Learning

- 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**.
  - With great power comes great responsibility

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?
- The same idea applies to all kinds of content sources
  - The Social Dilemma: https://www.youtube.com/watch?v=uaaC57tcci0

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:

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

ML is generally divided into:

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

ML is generally divided into:

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 as easy as they were in the previous case. 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	**?**

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…

What would it take for you to believe AI is intelligent?

- What's better than a Turing test?

ML in the news

- What interesting examples of ML have you seen lately?

Fun Demo: DALL·E

- https://openai.com/dall-e-2/

- From openai website for DALLE 1

TEXT PROMPT an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED IMAGES

Edit prompt or view more images↓

TEXT PROMPT an armchair in the shape of an avocado. . . .

AI-GENERATED IMAGES

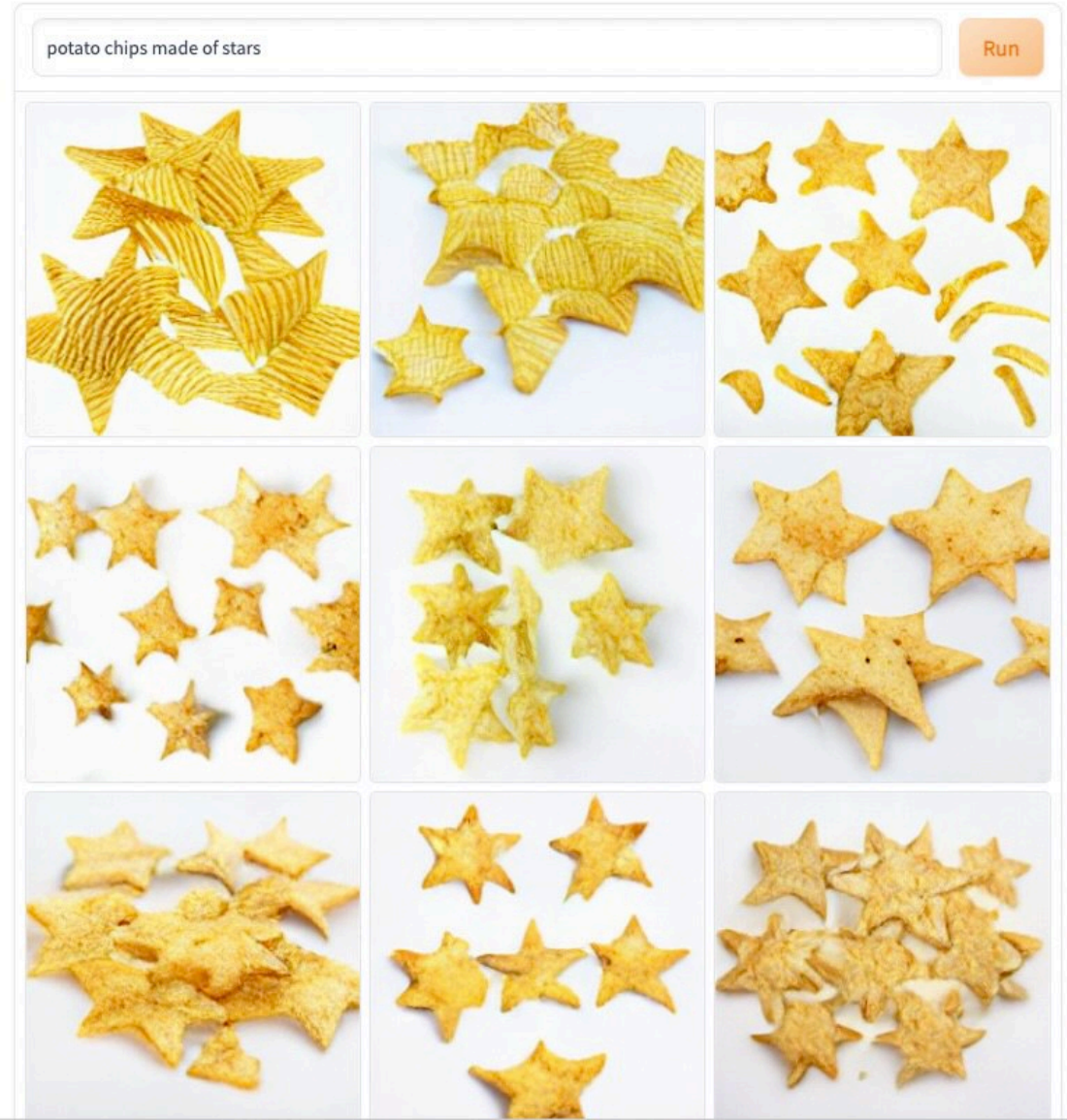
Edit prompt or view more images↓

TEXT PROMPT a store front that has the word 'openai' written on it. . . .

AI-GENERATED IMAGES

Edit prompt or view more images↓

Fun Demo: **DALL·E** (mini)



a really big house with balconies

Run

a purple tree with a swing

Run

a smelly eraser

Run

Other 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

Details...

- Eclass website (registered students should be able to view automatically)
- Labs are Monday
  - 5-7:50 pm, CCIS 1-140 or online
  - TAs will (likely) cycle coverage, most weeks open office hours
- Office hours: Monday 3-4 PM MT
  - via gmeet https://meet.google.com/yzo-ukgy-cea
  - Or by appointment (please reach out via email, I'm happy to meet!).
  - I have the most flexible schedule on Mondays and Fridays, so try and suggest a time on those days.

Where can I find information?

- Eclass
- Syllabus

https://docs.google.com/document/d/1dnH5v-rOQjzm2muIRBIRRJaB2ulS-Q8Z6P1wOEjq7Zs/edit?usp=sharing

- Schedule

https://docs.google.com/spreadsheets/d/1w9JPpNdMjedLoZnoj1ebegC2O2-z5TzAFdUdRWpvRGk/edit?usp=sharing

- Videos from Winter 2021 (similar but not identical content)

  https://youtube.com/playlist?list=PLkSkB8HpnMo7nnkww67yoBECV-SOo20J5

Machine Learning

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

Real Talk

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

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 and unsupervised learning
  - supervised = ?
  - unsupervised = ?
- 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

Assessment	Weight	Date
Assignments (3)	3*15%=45%	See schedule
Midterm	20%	**Oct 20**
Final	35%	****TENTATIVELY Dec 20** (will announce when the final exam schedule is released)
Project Final report (**566 students only**) Teams of 2-3 students	15% (Students will receive the average of the 3 highest marks from {project report, As1, As2, As3} I.e., the mark on the project can replace a lower assignment mark.)	Proposal: **Sept. 22** (required but not graded, for feedback purposes only) Final presentations: **Dec 1, 6** Final report: **Dec. 8**

Project

- 566 students only
  - Groups of 2-3
- The project is **optional**
  - 566 students will receive the average of the 3 highest marks from {project report, As1, As2, As3}
  - I.e., the mark on the project can replace a lower assignment mark.
  - Students who choose not to do the project will receive the average of the 3 assignments

Project

- Worth 15% of your mark (same as an assignment)
  - Proposal: Sept 22
    - **Required** but not graded, for guidance purposes
  - 5% Final Presentation (in-class): Dec. 1, 6
  - 10% Final report & code: Dec. 8
- Students self-select groups, use the forums to find a group.
- If you are wondering a project will be easier than an assignment, it will not.
  - But it may be more fun!

Plagiarism

- I have **zero tolerance** for 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/

Arguments I have had

- 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**

Arguments I have had

- 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 next week!

- Monday is a holiday, FYI ☺