CPSC 340:
Machine Learning and Data Mining

Mike Gelbart

University of British Columbia, 2020W Term 2

Big Data Phenomenon

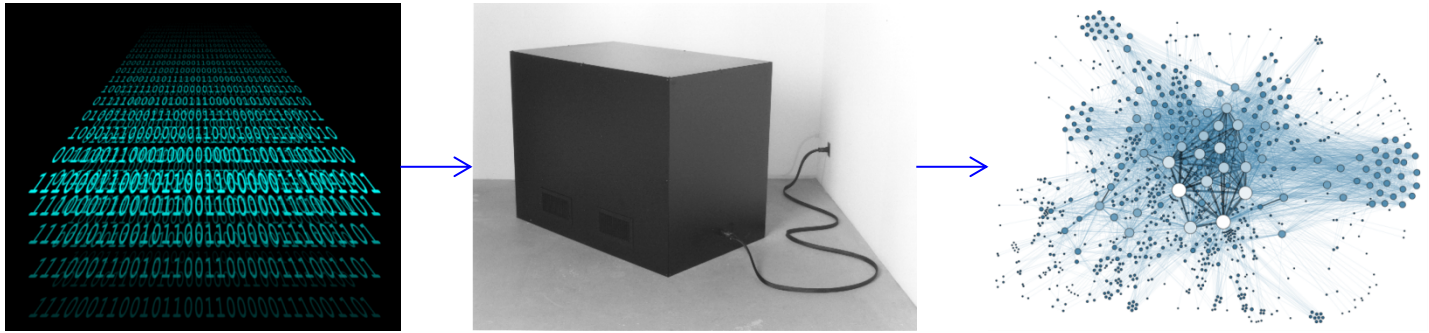
- We are collecting and storing data at an unprecedented rate.
- Examples:
  - YouTube, Facebook, MOOCs, news sites.
  - Credit cards transactions and Amazon purchases.
  - Transportation data (Google Maps, Waze, Uber)
  - Gene expression data and protein interaction assays.
  - Maps and satellite data.
  - Large hadron collider and surveying the sky.
  - Phone call records and speech recognition results.
  - Video game worlds and user actions.

Big Data Phenomenon

- What do you do with all this data?
  - Too much data to search through it manually.
- But there is valuable information in the data.
  - How can we use it for fun, profit, and/or the greater good?
- Data mining and machine learning are key tools we use to make sense of large datasets.

Data Mining

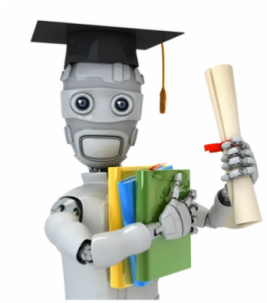
- Automatically extract useful knowledge from large datasets.

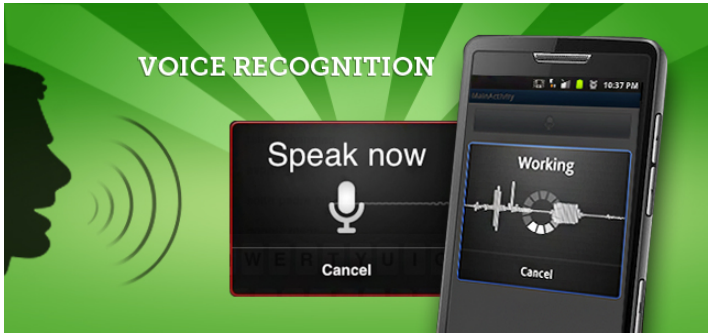


- Usually, to help with human decision making.

Machine Learning

- Using computer to automatically detect patterns in data and use these to make predictions or decisions.





- Most useful when:
  - We want to automate something a human can do.
  - We want to do things a human can't do (look at 1 TB of data).

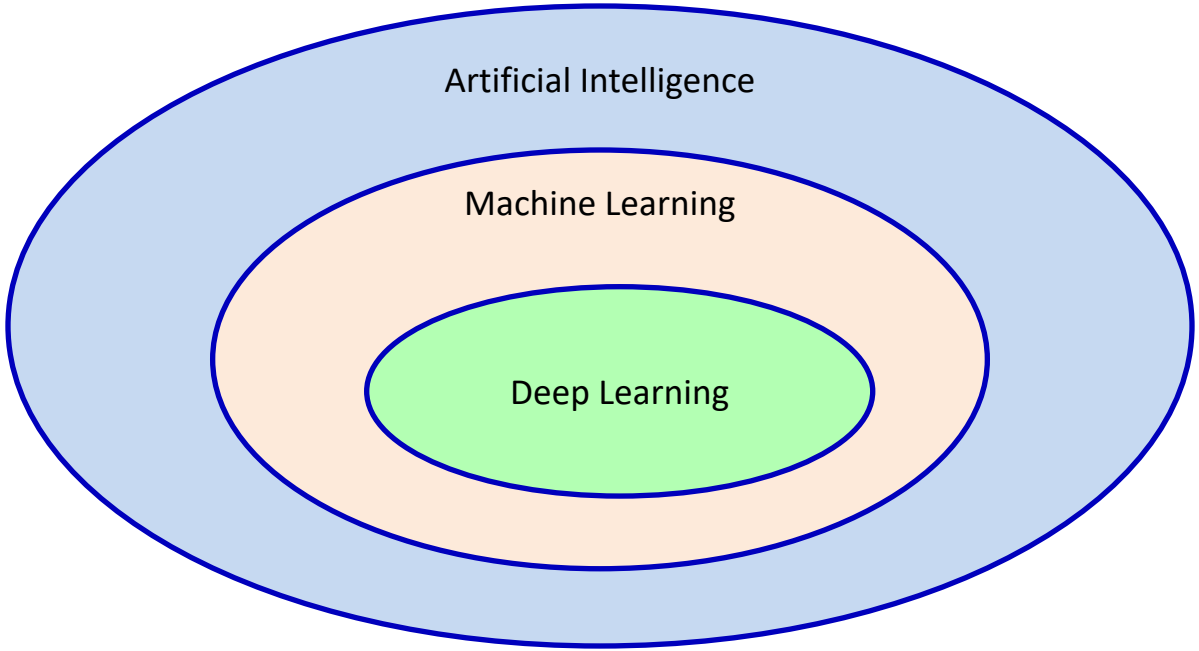
Data Mining vs. Machine Learning

- Data mining and machine learning are very similar:
  - Data mining often viewed as closer to databases.
  - Machine learning often viewed as closer AI.

- Both are similar to statistics, but more emphasis on:
  - Large datasets and computation.
  - Predictions (instead of descriptions).
  - Flexible models (that work on many problems).

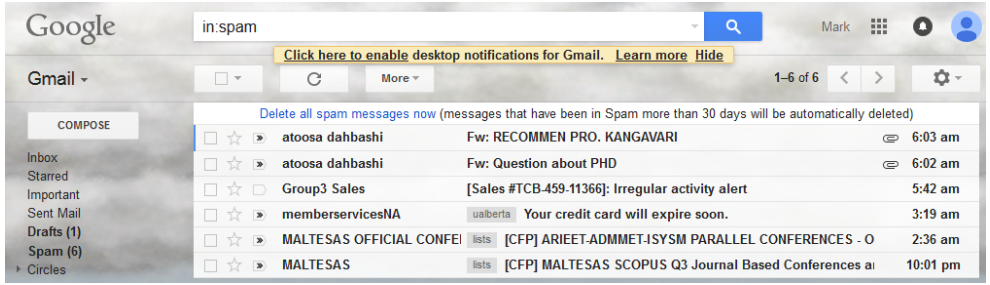
Deep Learning vs. Machine Learning vs. AI

- Traditional we've viewed ML as a subset of AI.
  - And "deep learning" as a subset of ML.

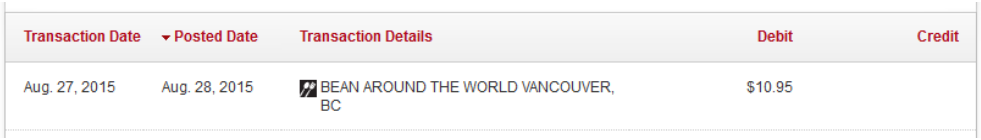


Applications

- Spam filtering:



- Credit card fraud detection:



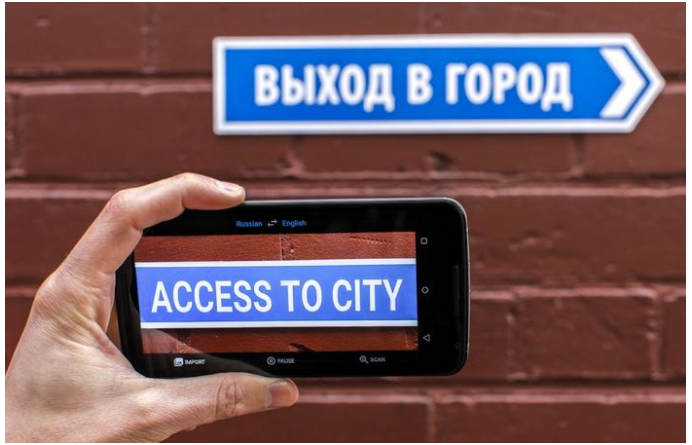
Transaction Date	Posted Date	Transaction Details	Debit	Credit
Aug. 27, 2015	Aug. 28, 2015	BEAN AROUND THE WORLD VANCOUVER, BC	$10.95	

- Product recommendation:

Applications

- Motion capture:

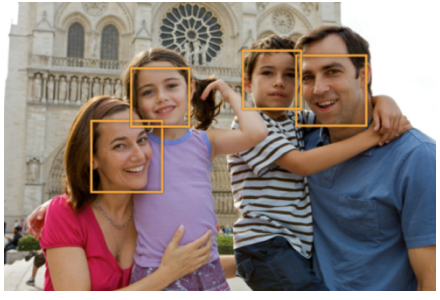
- Optical character recognition and machine translation:

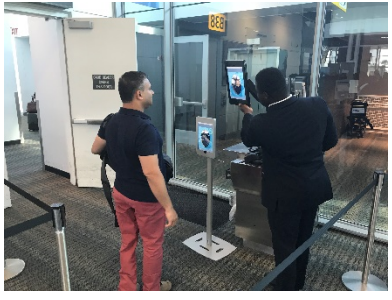


- Speech recognition:

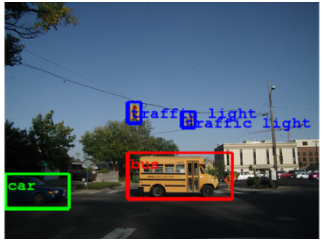
Applications

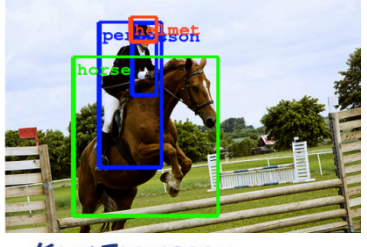
- Face detection/recognition:



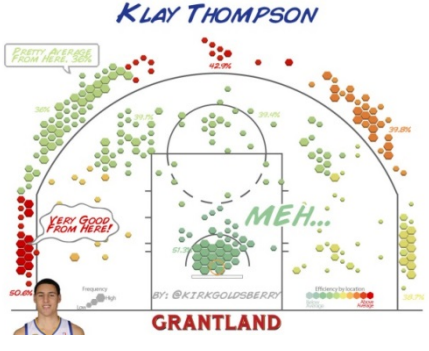


- Object detection:



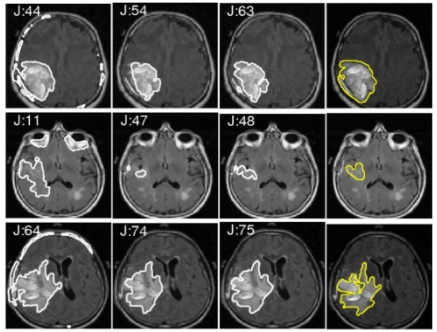


- Sports analytics:



Applications

- Medical imaging:

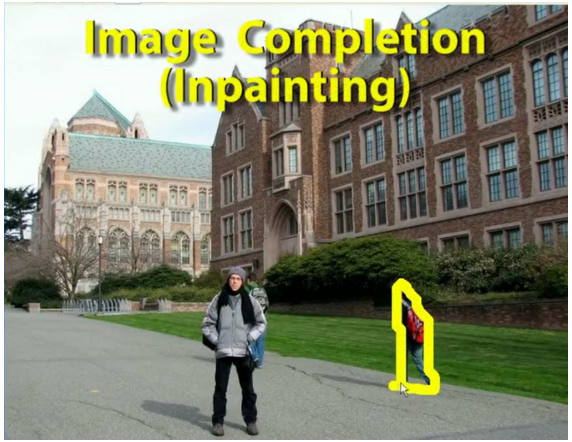


- Medical diagnostics:

- Self-driving cars:

Applications

- Image completion:



- Image annotation:

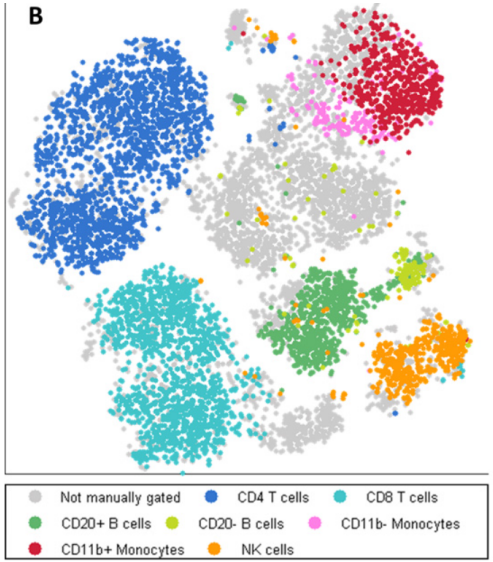
a cat is sitting on a toilet seat
logprob: -7.79

a display case filled with lots of different types of donuts
logprob: -7.78

a group of people sitting at a table with wine glasses
logprob: -6.71

Applications

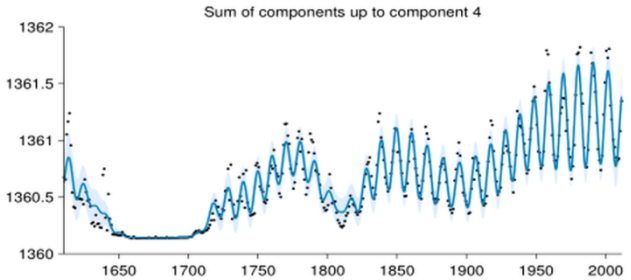
- Discovering new cancer subtypes:



- Automated Statistician:

2.4 Component 4 : An approximately periodic function with a period of 10.8 years. This function applies until 1643 and from 1716 onwards

This component is approximately periodic with a period of 10.8 years. Across periods the shape of this function varies smoothly with a typical lengthscale of 36.9 years. The shape of this function within each period is very smooth and resembles a sinusoid. This component applies until 1643 and from 1716 onwards.

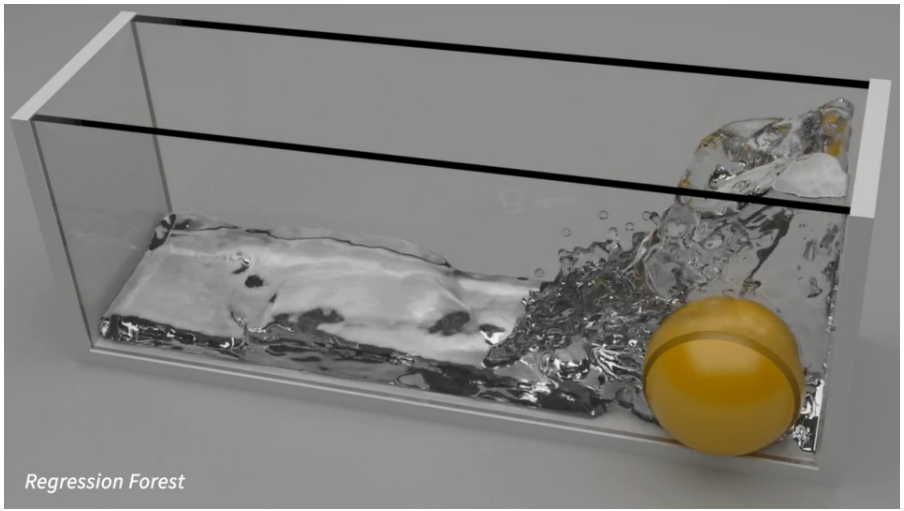


Applications

- Mimicking artistic styles:

Applications

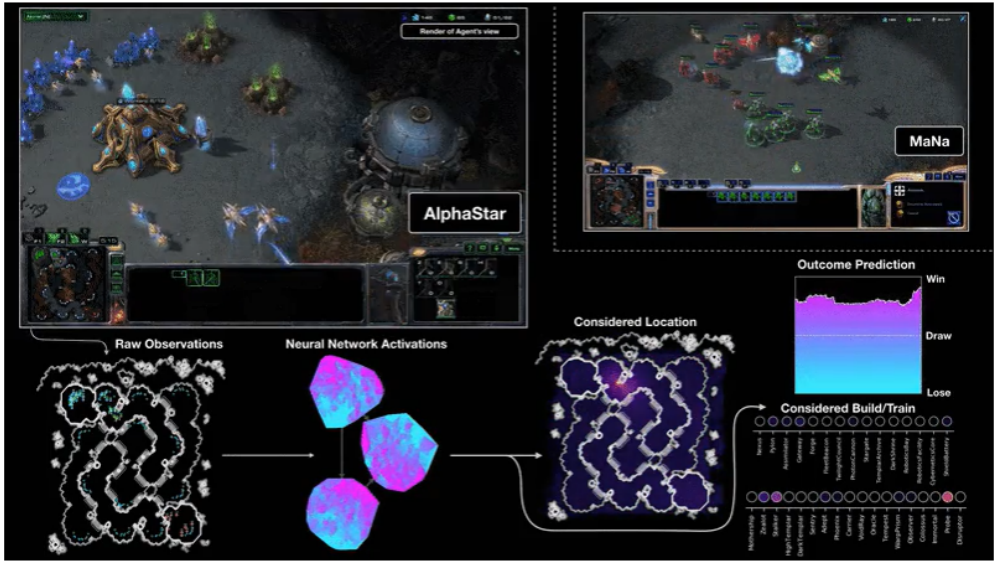
- Fast physics-based animation:



- Mimicking art style in video.
- Recent work on generating text/music/voice/poetry/dance.

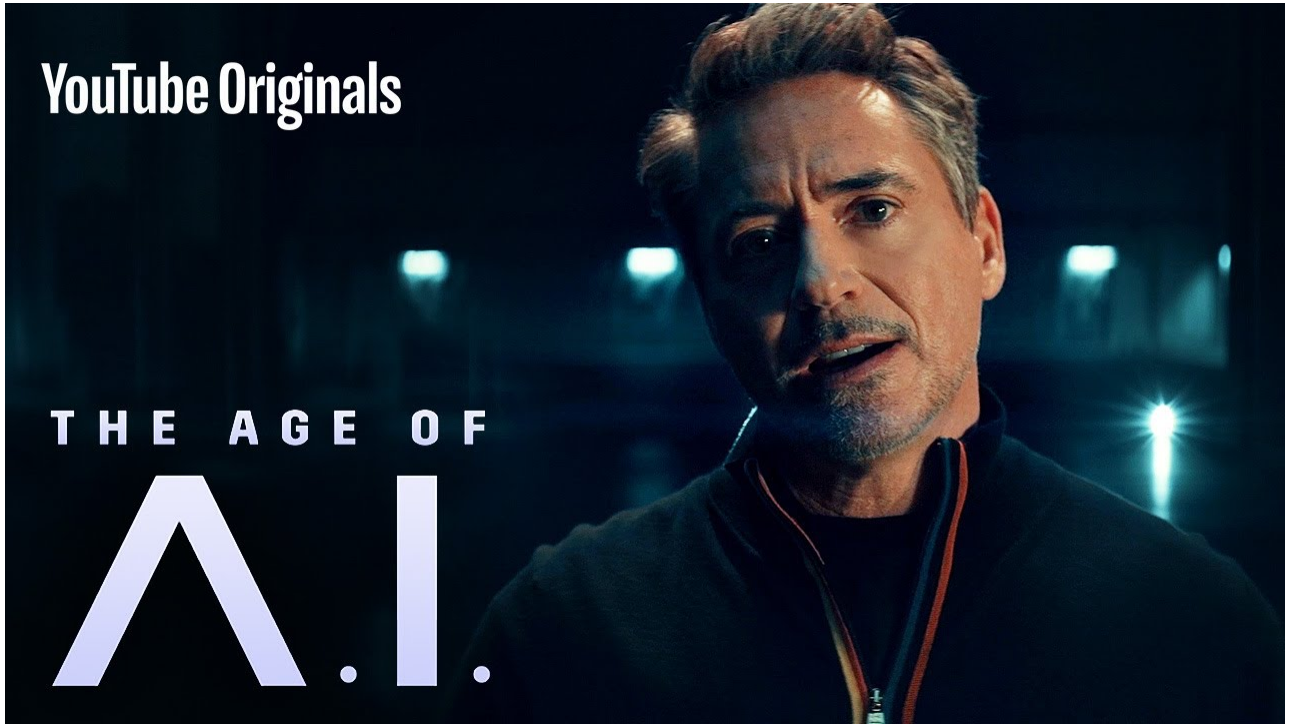
Applications

- Beating humans in Go and Starcraft:



Applications

- “Age of AI” YouTube series:



- Summary:
  - There is a lot you can do with a bit of statistics and a lot data/computation.

- We are in exciting times.
  - Major recent progress in fields like speech recognition and computer vision.
  - Things are changing a lot on the timescale of 3-5 years.
  - NeurIPS conference sold out in ~11 minutes last year.
  - A bubble in ML investments (most "AI" companies are just doing ML).

- But it is important to know the limitations of what you are doing.
  - "The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data." – John Tukey
  - A huge number of people applying ML are just "overfitting".
    - Or don't understand the assumptions needed for them to work.
    - Their methods do not work when they are released "into the wild".

(pause)

Reasons NOT to take this class

- Compared to typical CS classes, there is a lot more math:
  - Requires linear algebra, probability, and multivariate calculus (at once).
  - "I think the prerequisites for this course should require that students have obtained at least 75% (or around there) in the required math courses. As someone who did not excel at math, I felt severely under prepared and struggled immensely in this course, especially seeing that I have taken CPSC courses in the past with similar math requirements, but were not nearly as math heavy as CPSC340."
- If you've only taken a few math courses (or have low math grades), this course may well be a bad experience.
- It's better to improve your math, then take this course later.
  - A good reference covering the relevant math is here (Chapters 1-3 and 5-6).

Reasons NOT to take this class

- This is not a class on "how to use scikit-learn or TensorFlow or PyTorch".
  - You will need to implement things from scratch, and modify existing code.

- Instead, this is a 300-level computer science course:
  - You are expected to be able to quickly understand and write code.
  - You are expected to be able to analyze algorithms in big-O notation.

- If you only have limited programming experience,
  this course may well be a bad experience.
- It's better to get programming experience, then take this course later.
  - Take CPSC 310 and/or 320 instead, then take this course later.

Programming Language: Python

- 3 most-used languages in these areas: Python, Matlab, and R.

- We will be using Python which is a free high-level language.
  - Expected to be able to learn a programming language on your own.

- No, you cannot use Matlab/R/TensorFlow/Julia/etc.
  - Assignments have prepared code: we won't translate to many languages.

Reasons NOT to take this class

- Do NOT take this course expecting a high grade with low effort.

- Many people find the assignments very long and very difficult.
  - You will need to put time and effort into learning new/difficult skills.
  - If you aren't strong at math and CS, they may take all of your time.

- Class averages have only been high because of graduate students.
  - NOT because this is an "easy" course, for most people it's not.

CPSC 330 vs. CPSC 340

- There is also a less mathematical ML course, CPSC 330:
  - "Applied Machine Learning", designed and taught by me.
  - 330 emphasizes "when to use" tools, 340 emphasizes "how they work".
  - Fewer prerequisites:
    - 330 spends more time on data processing, applications and has no math.
      - More "learning by doing" and less discussion of fundamental principles.
    - 330 spends more time on data cleaning, communicating results, etc.
      - More emphasis on the entire "pipeline" of data of analysis.
    - 330 cannot be used as a prereq for CPSC 440.
  - You can take both for credit and I fact I highly encourage this!!

CPSC 340 vs. CPSC 440

- There is also a more-advanced ML course, CPSC 440:
  - Starts where this course ends.
  - More focus on theory/implementation, less focus on applications.
  - More prerequisites and higher workload.

- For almost all students, CPSC 340 is the better class to take:
  - CPSC 330/340 focus on the most widely-used methods in practice.
    - It covers much more material than standard ML classes like Coursera.
  - CPSC 440 focuses on less widely-used methods and research topics.
    - It is intended as a continuation of CPSC 340.
    - You'll miss important topics if you skip CPSC 340.

Essential Links

- Canvas course: https://canvas.ubc.ca/courses/58981
  - Contains link to everything else you need

- You should sign up for Ed:
  - https://edstem.org/us/join/GSsHth
  - Can be used to ask questions about lectures/assignments/exams.
  - May occasionally be used for course announcements.
  - Most questions should be “public” and not “private”,
    I will switch viewability of generally-relevant questions to “public”.

Assignments

- There will be 6 Assignments worth 30% of final grade (5% each):
  - Usually a combination of math, programming, and short answer.

- Assignments are due every 2 weeks (with one exception)
- a1 is already available, and is due next Wednesday night (Jan 20).
  - Submission instructions are posted on course webpage.
  - The assignment should give you an idea of expected background.
  - Make sure to submit before the deadline and check your submission.

- Start early, CPSC 340 assignments are notoriously very long.
  - You won't be able to finish if you start the day before the deadline.

Working in Teams for Assignments

- Assignments can optionally be done in pairs.
  - Teams are formed on Gradescope during the submission process, see submission instructions.

Getting Help

- Many students find the assignments long and difficult.
- But there are many sources of help:
  - TA office hours and instructor office hours.
    - Starting this week!
    - See office hours calendar, linked to from Canvas homepage.
  - Ed (for general questions).
  - Weekly tutorials (optional).
    - Starting next week.
    - Will go through provided code, review background material, review big concepts, and/or do exercises.
  - Other students.
  - The web (almost all topics are covered in many places).

Midterm and Final

- Midterm worth 19% and a (cumulative) final worth 50%
  - Open-book.
  - No need to pass the final to pass the course (but recommended).

- Midterm is scheduled for Friday Feb 26 in class.

- Format is TBD. Currently thinking of a multiple choice quiz on Canvas.

Lecture Format

- This course will be a flipped classroom.
  - Before each class, you are expected to watch the lecture video.
  - Starting Wednesday, in 2 days!
  - These lecture videos were recorded when I taught the class 3 years ago.
  - During class, we will work on practice problems and activities.
  - Relevant parts of the classes will be recorded and recordings will be available on the Canvas Zoom tab.

- Hopefully this approach will enhance your learning.
  - Many topics are too difficult to grasp in real-time.
  - I tried this approach last term and I think it was effective.

Lecture Format

For each lecture there will be 2 sets of slides available:

1. The exact slides used in the lecture videos (links in video descriptions)
2. The activities that we do in our synchronous time together (posted on course website)

Likewise there will be 2 recordings:

1. The recorded lectures on YouTube
2. Recordings of the synchronous time, found on the Canvas Zoom area

Lecture Format

- Be warned that the course we will move fast and cover a lot of topics:
  - Big ideas will be covered slowly and carefully.
  - But a bunch of other topics won't be covered in a lot of detail.

Tutorials

- Tutorials are optional.
  - You don’t have to be registered in a tutorial section.
  - If registered, try to attend your section, but we’re flexible.

- Tutorials will be run by our excellent TAs.
  - Extra concepts
  - Going through assignments

Upcoming deadlines

- Syllabus Quiz due Friday Jan 15 at 11:55pm
- Assignment 1 due Wed Jan 20 at 11:55pm

Bonus Slides

- The lectures include a lot of “bonus slides”.
  - May mention advanced variations of methods from lecture.
  - May overview big topics that we don’t have time for.
  - May go over technical details that would derail class.

- You are not expected to learn the material on these slides.
  - But they’re useful if you want to take 440 or work in this area.

- I’ll use this colour of background on bonus slides.

Code of Conduct

- Do not post offensive or disrespectful content.
- If you have a problem or complaint, let me know (maybe we can fix it).
- Do not distribute any course materials without permission.
- Do not record lectures without permission.

- Think about how/when to ask for help:
  - Don't ask for help after being stuck for 10 seconds. Make a reasonable effort to solve your problem (check instructions, Ed, and Google).
  - But don't wait until the 10th hour of debugging before asking for help.
    - If you do, the assignments could take all of your time.

Cheating and Plagiarism

- Read about UBC's policy on "academic misconduct" (cheating):
  - http://www.calendar.ubc.ca/Vancouver/index.cfm?tree=3,54,111,959

- When submitting assignments, acknowledge all sources:
  - Put "I had help from Sally on this question" on your submission.
  - Put "I got this from another course's answer key" on your submission.
  - Put "I copied this from the Coursera website" on your submission.
  - Otherwise, this is plagiarism (course material/textbooks are ok with me).

- At Canadian schools, this is taken very seriously.
  - Automatic grade of zero on the assignment.
  - Could receive 0 in course, be expelled from UBC, or have degree revoked.

Course Outline

- Next class discusses “exploratory data analysis”.

- After that, the remaining lectures focus on five topics:
  1) Supervised Learning.
  2) Unsupervised learning.
  3) Linear prediction.
  4) Latent-factor models.
  5) Deep learning.

- “What is Machine Learning?” (overview of many class topics)