CPSC 340
Machine Learning and Data Mining

Professors: Jiarui Ding and Jeff Clune
University of British Columbia, Spring 2024
https://www.students.cs.ubc.ca/~cs-340

Held on the traditional, ancestral, and
unceded territory of the Musqueam people

What is the most interesting thing happening in the world?

My answer

- ML / AI
- Culminating in AGI / HLAI?
- I'm on record as saying so…

Big Data Phenomenon

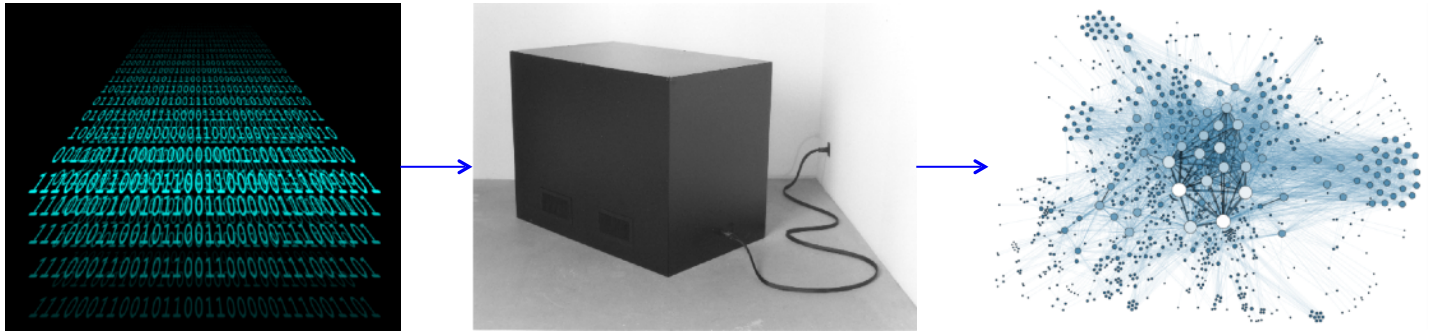
- We are collecting and storing data at an unprecedented rate.
- Examples:
  - YouTube, Facebook, MOOCs, news sites.
  - Credit cards transactions and Amazon purchases.
  - Transportation data (Google Maps, Waze, Uber)
  - Gene/protein sequencing/expression/structures.
  - Maps and satellite data.
  - Camera traps and conservation efforts.
  - 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

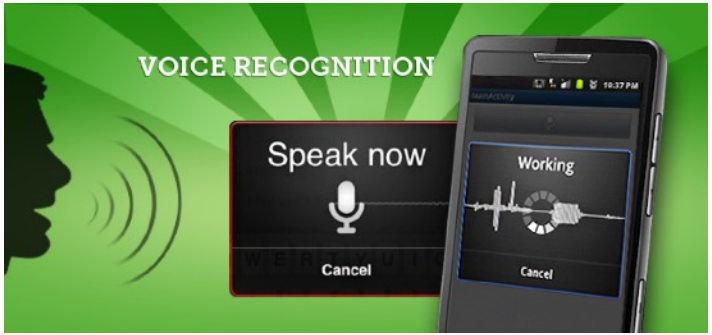
- Automatically extract useful knowledge from large datasets.



- Usually, to help with human decision making.

Machine Learning

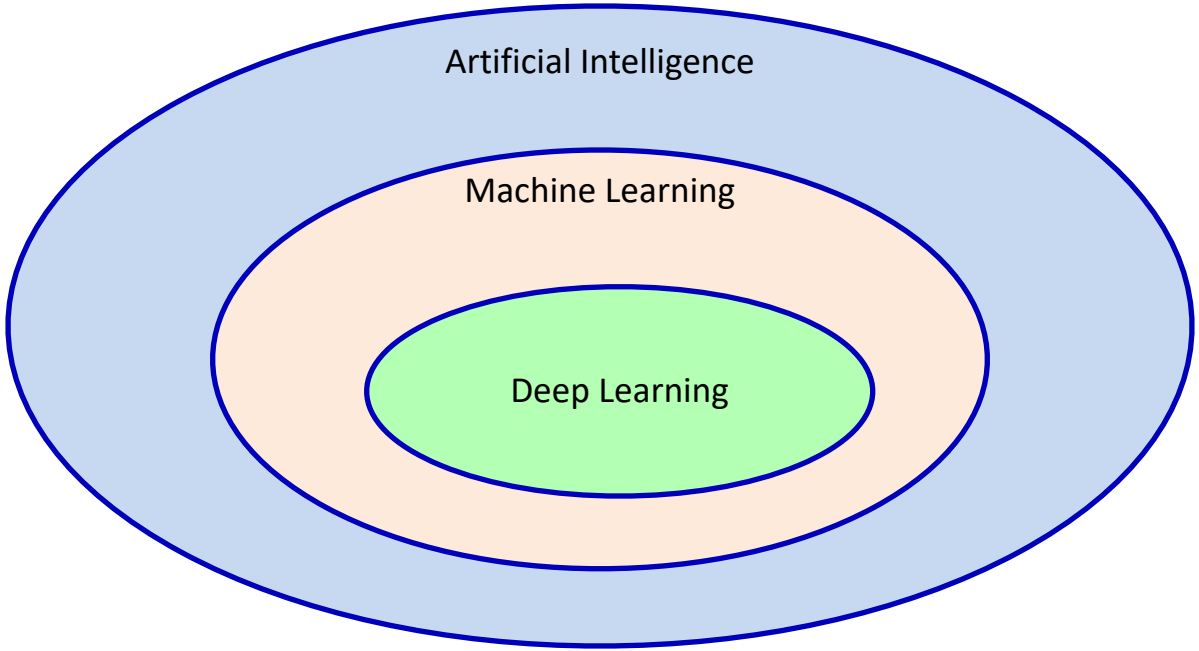
- Using computer to automatically detect patterns in data and use these to make predictions or decisions.



- 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).

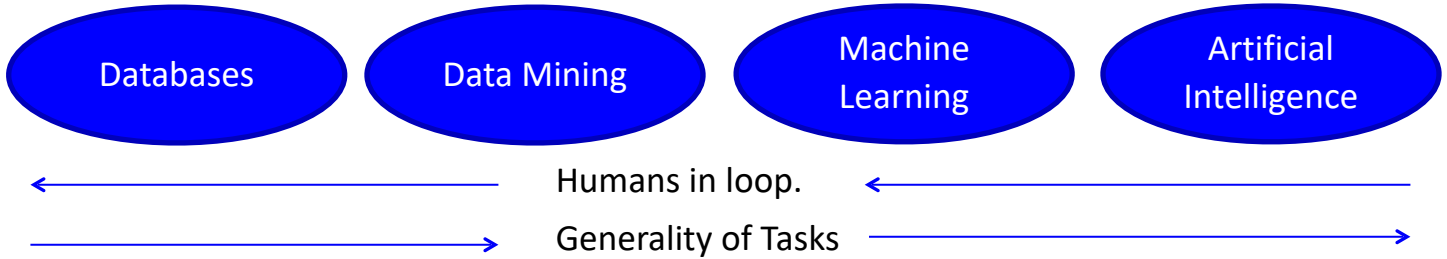
Deep Learning vs. Machine Learning vs. AI

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



Data Mining vs. Machine Learning

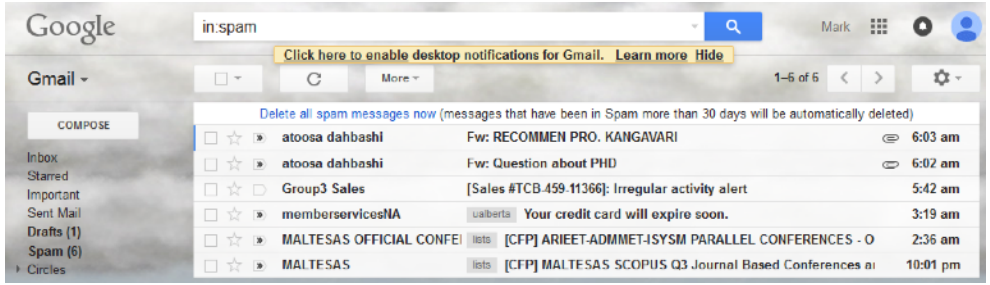
- Data mining and machine learning are very similar:
  - Data mining often viewed as closer to databases.
  - Machine learning often viewed as closer AI.



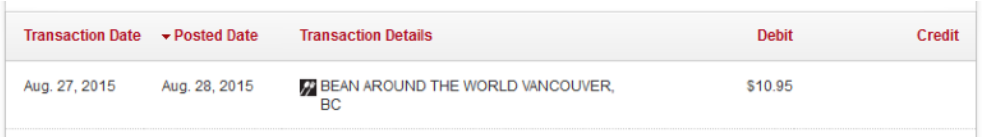
- Both are similar to statistics, but more emphasis on:
  - Large datasets and computation.
  - Predictions (instead of descriptions).
  - Flexible models (that work on many problems).

Applications

- Spam filtering:



- Credit card fraud detection:

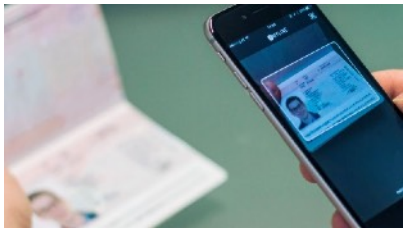


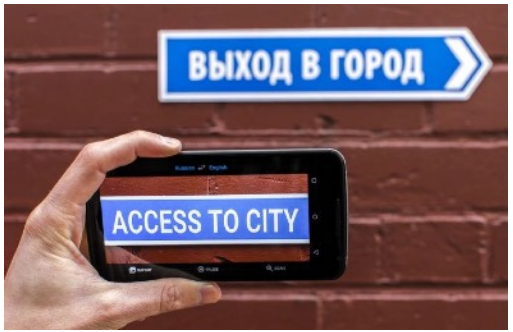
Transaction Date	Posted Date	Transaction Details	Debit	Credit
Aug. 27, 2015	Aug. 28, 2015	BEAN AROUND THE WORLD VANCOUVER, BC	$10.95	

- Product recommendation:

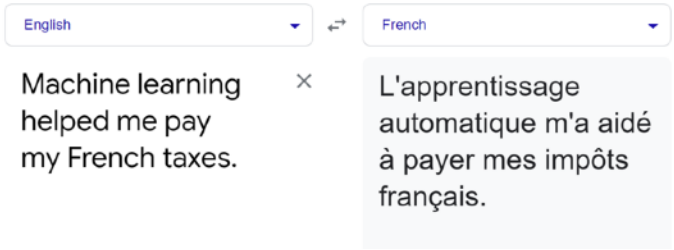
Applications

- Optical character recognition:

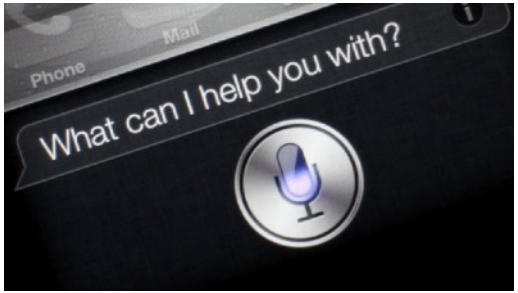




- Machine translation:

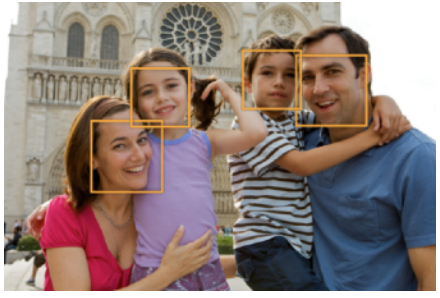


- Speech recognition:

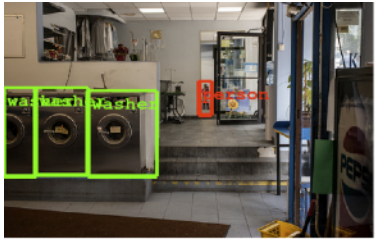


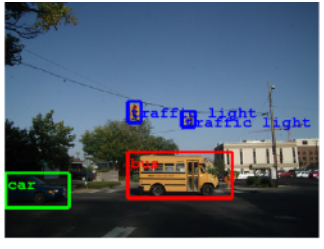
Applications

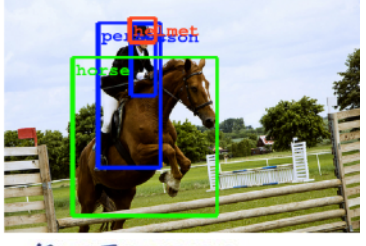
- Face detection/recognition:



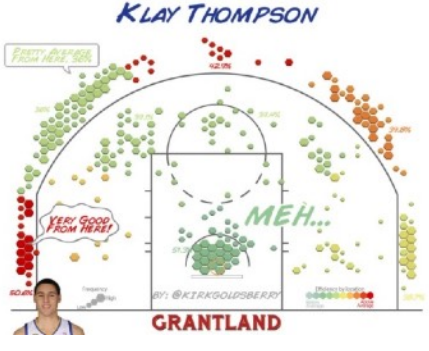
- Object detection:





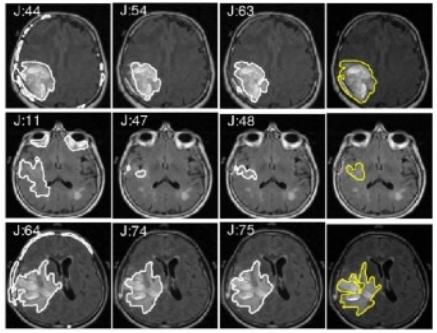


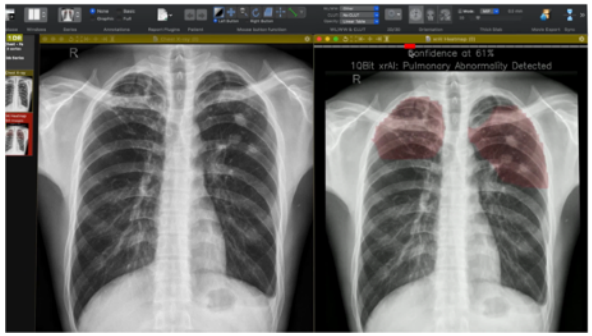
- Sports analytics:



Applications

- Medical imaging:



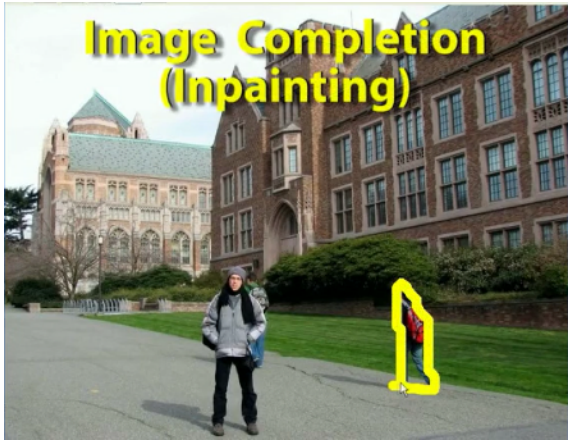


- Medical diagnostics:

- Self-driving cars:

Applications

- Image completion:



- Image annotation:

Applications

- Generating images from text:

vibrant portrait painting of Salvador Dalí with a robotic half face

a shiba inu wearing a beret and black turtleneck

a close up of a handpalm with leaves growing from it

an espresso machine that makes coffee from human souls, artstation

panda mad scientist mixing sparkling chemicals, artstation

a corgi's head depicted as an explosion of a nebula

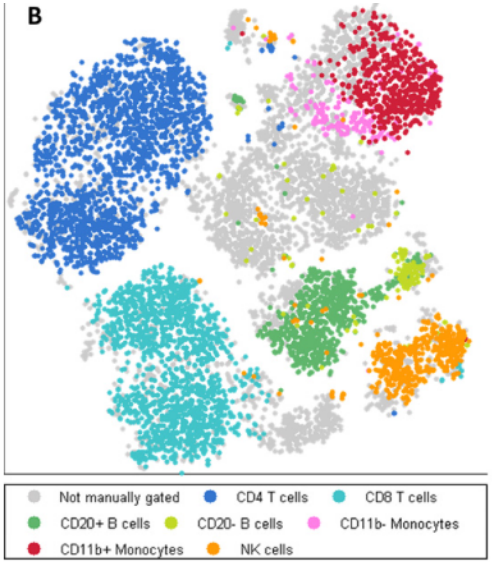
a dolphin in an astronaut suit on saturn, artstation

a propaganda poster depicting a cat dressed as french emperor

a teddy bear on a skateboard in times square

Applications

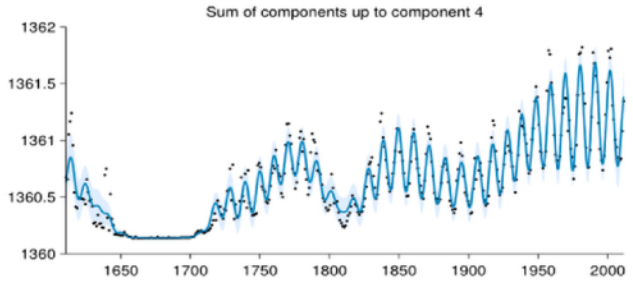
- Discovering new cancer subtypes:



- Automated Statistician:

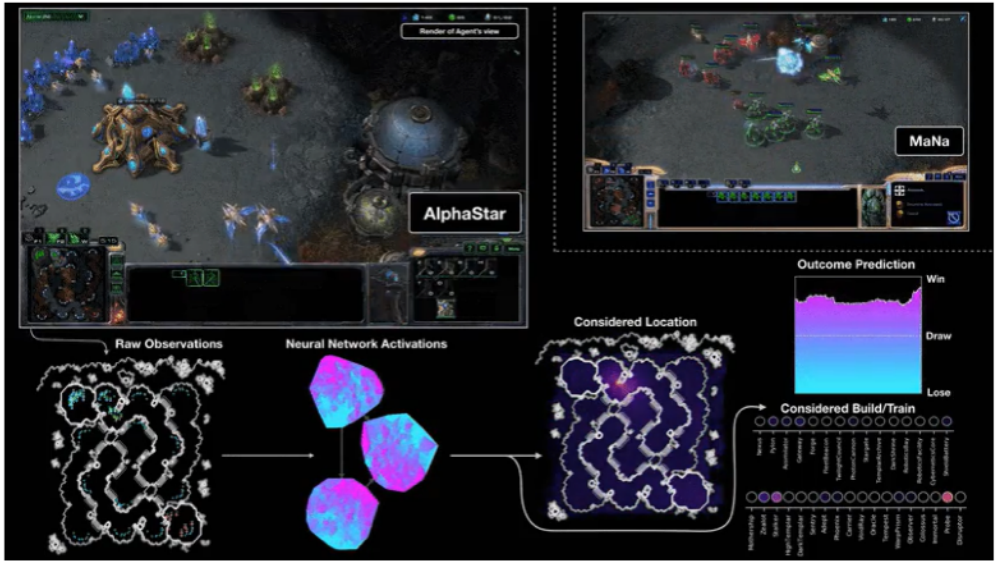
2.4 Component 4 : An approximately periodic function with a period of 10.8 years. This function applies until 1643 and from 1716 onwards

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

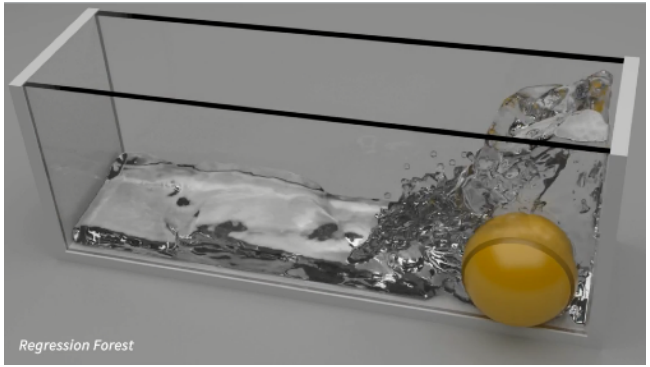
- Beating humans in Go and Starcraft:



Applications

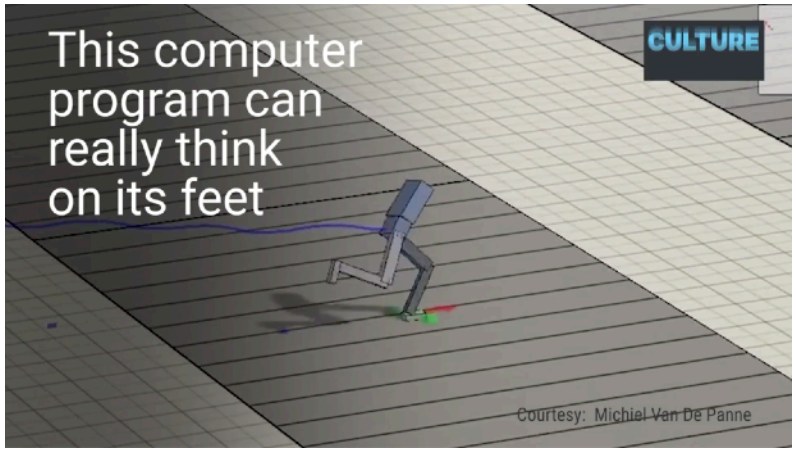
- Mimicking artistic styles (video).

Applications



- Fast physics-based animation:

- Character animation (video):



- Recent work on generating text/music/voice/poetry/dance.

ChatBots

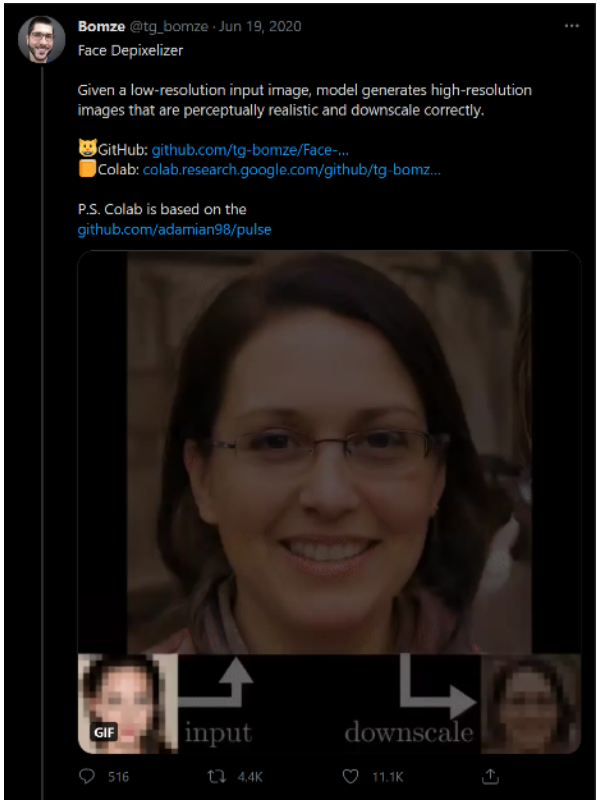
- e.g. ChatGPT

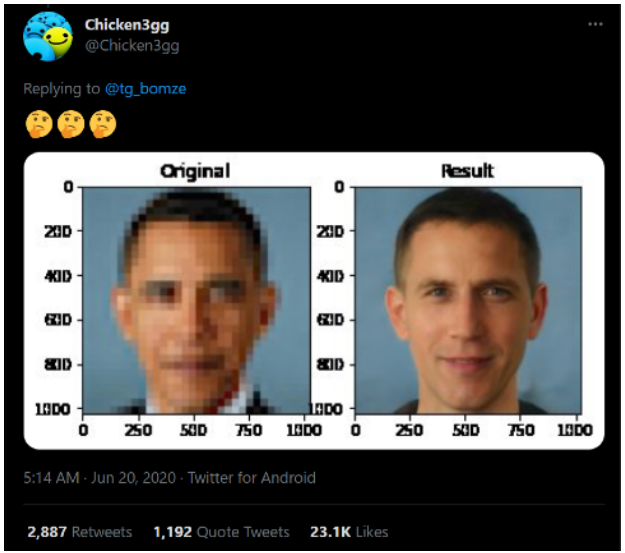
- Summary:
  - There is a lot you can do with a bit of statistics and a lot data/computation.

- We are in exciting times.
  - Major recent progress in many fields:
    - Speech recognition, computer vision, natural language processing, image generation.
  - Things are changing a lot on the timescale of 3-5 years.
  - AI and AI products (e.g. ChatGPT) are “household names”
  - A bubble in ML/AI investments/startups

- But it is important to know the limitations of what you are doing.
  - A huge number of people applying ML are just “overfitting”.
    - Their methods do not work when they are released “into the wild”.

Failures of Machine Learning





Failures of Machine Learning

Amazon reportedly scraps internal AI recruiting tool that was biased against women

The secret program penalized applications that contained the word "women's"

By James Vincent | Oct 10, 2018, 7:09am EDT

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By James Vincent | Mar 24, 2016, 6:43am EDT

Via The Guardian | Source TayandYou (Twitter)

Uber self-driving car kills pedestrian in first fatal autonomous crash

by Matt McFarland @mattmcfarland

March 19, 2018: 1:40 PM ET

Get Educated, Be Cautious

- ML/AI hubris is not healthy and can be very harmful
- Learn how things work "under the hood", listen for cautionary tales, have a healthy dose of skepticism, proceed cautiously!

Next Topic: Course Syllabus

These slides ARE the syllabus

Lectures

- All slides will be posted online (before lecture, and final version after).
- Please ask questions: you probably have similar questions to others.
  - I may deflect to the next lecture or Piazza for certain questions.
- 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.
- Is it wrong to have only have shallow knowledge?
  - In this field, it's better to know many methods than to know 5 in detail.
    - Different problems need different solutions.

Bonus Slides

- We 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 are useful if you want to take 440 or work in this area.

- Wel will use this colour of background on bonus slides.

Essential Links

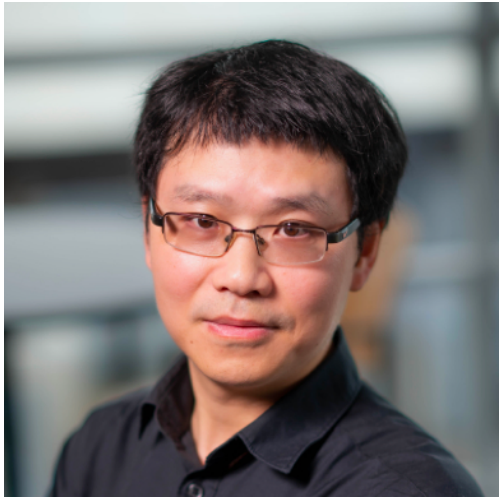
- Please bookmark the course webpage:
  - https://www.students.cs.ubc.ca/~cs-340/
  - Contains lecture slides, assignments, optional readings, additional notes.

- Required: sign up for Piazza:
  - Enter the class number and term on Piazza's website
  - Can be used to ask questions about lectures/assignments/exams.
  - **Will be used for course announcements.**
  - Most questions should be "public" and not "private",
    we will switch viewability of generally-relevant questions to "public".

- Use Piazza instead of e-mail for questions:
  - Please direct administrative questions and questions regarding assignment/exam accommodations to cpsc340-admin@cs.ubc.ca

Different Sections of CPSC 340

- Prof. Ding and I are jointly teaching both sections of 340 this term.
  - Both sections have the same webpage, assignments, and exams.
- You are free to attend the lectures of the other section.
  - But do not take a seat if you are not registered and people are standing.
- We are planning to alternate classes.
  - I will give some lectures, Prof. Ding the others
- The AM and PM lectures will cover the same set of topics
- Not that we are research faculty
  - We will do our best, but may not be as effective as teaching-only faculty
  - You may wish to look up our research
    - JeffClune.com
    - https://www.cs.ubc.ca/~jiaruid/



Prof. Ding

Brief Intro to Me: Professional

MICHIGAN

UNIVERSITY OF WYOMING

UBER AI Labs

OpenAI

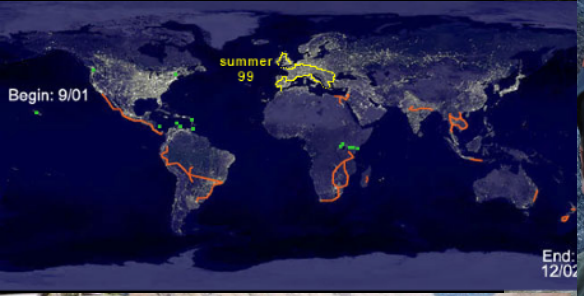
UBC THE UNIVERSITY OF BRITISH COLUMBIA

VECTOR INSTITUTE

Google DeepMind

COLLEGE FOOTBALL PLAYOFF
NATIONAL
CHAMPIONSHIP
JANUARY 8, 2024 | HOUSTON
WASHINGTON VS. MICHIGAN

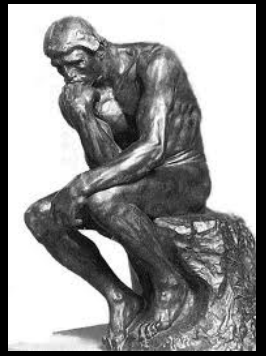
GO
BLUE
MICHIGAN

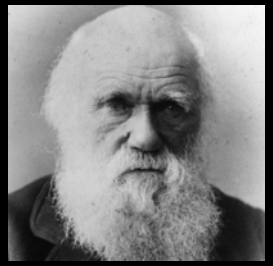

summer
99
Begin: 9/01
End:
12/02

CONGRATULATI
YOU ARE NOW AT
AMSL.
AFRICA'S HIGHEST POINT
WELCOME

Background & Research Areas

- Deep learning
- Deep reinforcement learning
- Robotics
- Neuroevolution
- Past
  - Computational evolutionary biology





Interested in Learning More?

- JeffClune.com
  - videos of talks I have given
  - Research papers
  - etc.
- @jeffclune

COVID / Sickness

- Masks are fully welcome
- Please do not come to class if you have any symptoms
- Please do not come to class if you have any symptoms
- Please do not come to class if you have any symptoms

Videos

- Lectures are in-person (no live streaming)
  - We will try to record and post, but no guarantee that works
  - Do not share videos publicly

- Videos from a previous offering are available here:
  - https://www.youtube.com/playlist?list=PLWmXHcz_53Q02ZLeAxigki1JZFfCO6M-b
  - From Mike Gelbart, who is an excellent teacher
  - Material is usually almost identical, particularly for the “testable” concepts
    - But not guaranteed to be, so it is your responsibility to check slides

Assignments

- There will be 6 Assignments worth 30% of final grade (for 340):
  - Usually a combination of math, programming, and very-short answer.

- Assignment 1 will be on webpage soon, and is due next Friday.
  - Submission instructions will posted on webpage/Piazza.
  - The assignment should give you an idea of expected background.
  - Make sure to submit repeatedly before the deadline and check your submission

- Start early, work often!!!! There is a lot there!
  - Don't wait to see you if get off the waiting list to start.
  - You should be able to do the first few questions already.

Working in Teams for Assignments

- Assignment 1 must be done individually.

- Assignments 2-6 can optionally be done in pairs.
  - You do not need to use the same pairing for all assignments.
  - But the midterm/final will ultimately be individual.

- All the various permutations of partners are allowed:
  - Partners can be from different sections of 340.
  - One of you can be in 340 and one can be in 540
  - Partnering with an auditor is ok.

Programming Language: Python

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

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

Late Policy for Assignments

- Assignments will be due just before midnight on the due date.

- If you can't make it, you can use "late classes":
  - For example, if assignment is due on a Friday:
    - Handing it in Monday is 1 late class.
    - Handing it in Wednesday is 2 late classes.

  - There is no penalty for using "late classes",
    but you will get a mark of 0 on an assignment if you:
    - Use more than 2 late classes on the assignment.
    - Use more than 4 late classes across all assignments.

  - If you are working in a pair, you both must have late classes remaining.

- We will try to put up grades within 10 days of final late class.
  - You see the solutions by coming to TA/instructor office hours.
  - We are not releasing solutions anymore (too much cheating and time spent taking them down).

Midterm and Final

- Midterm worth 20% and a (cumulative) final worth 50%
  - Closed-book (no electronic devices)
  - One doubled-sided ‘cheat sheet’ for midterm, two doubled-sided pages for final.
  - No need to pass the final to pass the course (but recommended).

- Midterm Room, Date, Time TBD:
  - Let us know if you have a conflict that cannot be resolved.
- We don’t control when the final is, don’t make travel plans before last possible final slot.
  - If it’s scheduled early, we may restrict the number “late classes” for the last assignment.

- There will be two types of questions:
  - ‘Technical’ questions requiring things like pseudo-code or derivations.
    - Similar to assignment questions, and will only be related topics covered in assignments.
  - ‘Conceptual’ questions testing understanding of key concepts.
    - All lecture slide material except “bonus slides” is fair game here.
- Final: same style as midterm, but cumulative for the whole course

Tentative Schedule

- Tentative schedule
  - Will likely change! Your responsibility to check back often
  - Let’s go look

Reasons NOT to take this class

- Compared to typical CS classes, there is a lot more math:
  - Requires linear algebra, probability, and multivariate calculus (and combining them).
  - "the prerequisites for this course should require that students have obtained at least 75% (or around there) in the required math courses … 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"
- If you've only taken a few math courses (or have low math grades), this course will ruin your life for the next 4 months.
- 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 will ruin your life for the next 4 months.
- It’s better to get programming experience, then take this course later.
  - Take CPSC 310 and/or 320 instead, then take this course later.

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.

- From "Rate My Professors":
  - "Lectures were dull, dry, and glossed over the material skipping over the theoretical details. Ironically, assignments were detail-heavy and LONG. Doesn't seem to care about students because some of us have 4 other classes and well, if they're all like this course, my girlfriend would have broken up with me two months ago."

CPSC 330 vs. CPSC 340

- There is a less-advanced ML course, CPSC 330:
  - "Applied Machine Learning".
  - 330 emphasizes "when to use" tools, 340 emphasizes "how they work".
    - 330 is more like the Coursera course and online courses.
  - Fewer prerequisites and more emphasis "learning by doing".
    - 330 spends more time on low-level coding details and has basically no equations.
    - 330 spends more time on data cleaning, communicating results, and so on.
  - You can take both for credit (better to take 330 first or at same time).

CPSC 340 vs. CPSC 440

- There is also a more-advanced ML course, CPSC 440:
  - "Advanced machine Learning": starts where this course ends.
  - More focus on theory/implementation, less focus on applications.
  - More prerequisites and higher workload (330 cannot be used as a prereq).

- For almost all students, take CPSC 340 first:
  - CPSC 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 will miss important topics if you skip CPSC 340.
      - "I am familiar with ML from my research/company so I can skip CPSC 340" – someone who is wrong.

Waiting List and Auditing

- Right now only CS students can register directly.
  - All other students need to sign up for the waiting list to enroll.

- We're going to start registering people from the waiting list.
  - Being on the waiting list is the only way to get registered:
    - https://www.cs.ubc.ca/students/undergrad/courses/waitlists
  - You might be registered without being notified, be sure to check!
    - They might also ask to submit a prereq form, let me know if you have issues.

- If the room is full, we may not have seats for auditors.
  - If there is space, I'll describe (light) auditing requirements then.

Getting Help

- There are many sources of help:
  - TA office hours and instructor office hours.
    - Starting in the second week of class.
    - Times on the course webpage, location based on who teaches Friday
  - Piazza (for general questions).
  - Weekly tutorials (optional).
    - Starting in second week of class.
    - Will go through provided code, review background material, review big concepts, and/or do exercises.
  - Other students (ask your neighbor for their e-mail).
  - The web (almost all topics are covered in many places, but don't use it to cheat on homeworks!).

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 had help from GPT 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.
  - We have actually given 0 to people before.

Generative AI / ChatGPT

- See answer on FAQ
  - (let’s go read it now)

Code of Conduct

- Do not post offensive or disrespectful content on Piazza.
- If you have a problem or complaint, let me know.
  - We will try to fix it
- Do not distribute any course materials without permission.
  - For example, do not post your solutions to the internet, nor exam questions
- Do not share 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, Piazza, and Google).
  - But do not wait until the 10th hour of debugging before asking for help.
    - If you do, the assignments could take all of your time.

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)

This is the end of the lecture.
(Future lectures will end on a “Summary” slide.)

The slides after the “Summary” slide are typically “bonus” material related to the topics of the lecture.

Photo I took in the UK on the way home from the “Optimization and Big Data” workshop:

Less-inspirational quote: “Without data you're just another person with an opinion.” W.E. Deming

Bonus Slide: “Machine Learning” vs. “Data Mining”

- Machine learning and data mining have many similarities (as do other fields like statistics and signal processing), and the similarity is increasing due to the 'arXiv' effect (people from both fields can now easily read each other's papers and are using standard notation).
- However, as a subjective answer I would say that the focuses are different. Data mining is broader in scope and includes things like how to organize data, models that simply look up answers or are based on counting (KNN and naive Bayes are also often covered in data mining, and in data mining there is a greater focus on interpretable models), and tasks like information visualization. Machine learning is more narrow, focusing largely on the modeling aspect, generalization error, and using methods that rely on numerical optimization or high-dimensional integration (that may not necessarily be interpretable).
- Another subjective comment would be that data mining often focuses on tools that help professionals analyze their data, while machine learning often focuses on automating data analysis. For example, here is a recent very-interesting project by some machine learning folks from Cambridge and MIT:
  - http://www.automaticstatistician.com

Textbooks

- No required textbook.

- I'll post relevant sections out of these books as optional readings:
  - Artificial Intelligence: A Modern Approach (Rusell & Norvig).
  - Introduction to Data Mining (Tan et al.).
  - The Elements of Statistical Learning (Hastie et al.).
  - Mining Massive Datasets (Leskovec et al.)
  - Machine Learning: A Probabilistic Perspective (Murphy).
- Most of these are on reserve in the ICICS reading room.
- List of related courses on the webpage, or you can use Google.

Assignment Issues

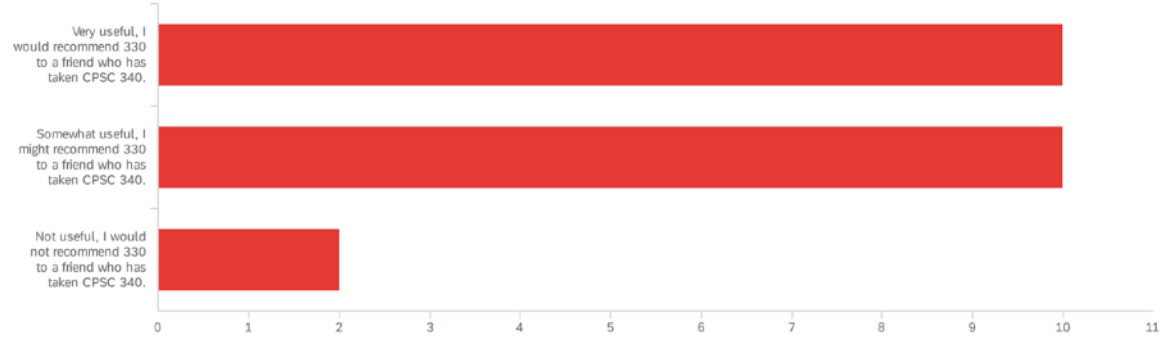
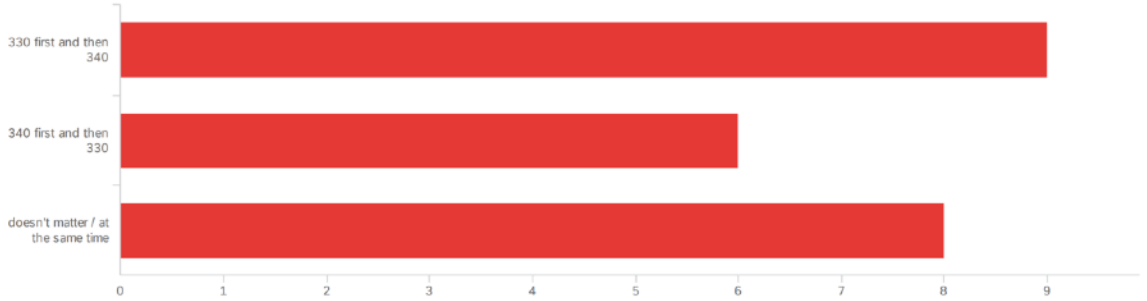
- No extensions will be considered beyond the late days.
  - Also, since you can submit more than once, you have no excuse not to submit something preliminary by the deadline.

- Further, due to grouchiness, these issues are a 50% penalty:
  - Missing names or student IDs on assignments.
  - Submitting the wrong assignment or corrupted files.
  - Not including answers in the correct location in the .pdf file.

Example Tentative Course Schedule from Prior Year

- First class: Sep 7
- Assignment 1 due: Sep 16 (Friday of week 2)
- Drop deadline: Sept 19 (Monday of week 3)
- Assignment 2 due: Sep 30 (Friday of week 4)
- Assignment 3 due: Oct 14 (Friday of week 6)
- Midterm: Oct 24 (Monday of week 8)
- Assignment 4 due: Nov 7 (Monday of week 10)
- Assignment 5 due: Nov 23 (Wednesday of week 12)
- Assignment 6 due: Dec 7 (Wednesday of week 14, last day of class)
- Final: random day decided by UBC, sometime between Dec 11 and 22

Q4 - Please rate how useful CPSC 330 was to you as someone who has taken CPSC 340.



Q5 - Which order of the courses do you think makes more sense for a student who ultimately takes both courses?