

CPSC 340: Machine Learning and Data Mining

Nam Hee Gordon Kim (nhgk@cs.ubc.ca)

Summer 2021

<https://www.cs.ubc.ca/~nhgk/courses/cpsc340s21>

REMINDER TO HIT RECORD

In This Lecture

1. Motivation (20 minutes)
2. Syllabus (30 minutes)

Coming Up Next

WHY MACHINE LEARNING?

Big Data Phenomenon

- We are **collecting and storing data** at an unprecedented rate.
- Examples:



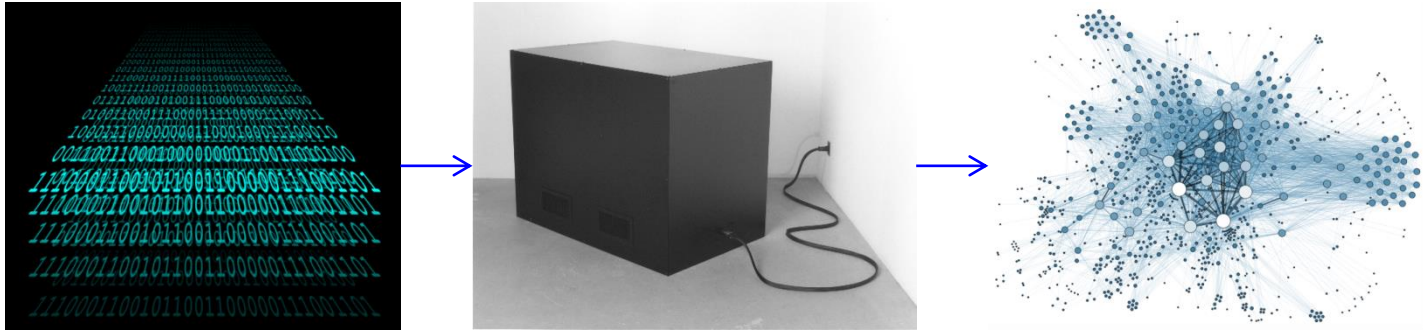
Q: What do you do with all this data?

Big Data Phenomenon

- There is valuable information in the data...
 - Too much data to search through it manually.
- Q: How can we use it for fun, profit, and/or the greater good?
- A: Machine learning and data mining!
 - key tools we use to make sense of large datasets.

What is Data Mining?

- Automatically **extract useful knowledge** from large datasets.



- Usually, to help with human decision making.

What is Machine Learning?

- Study of data-driven automations
- Detect patterns in data and use these to make predictions (or decisions)



Q: Why is ML useful?



Q: Does “machine learning” mean that a machine learns to do things?

“...learning refers to the process of extracting structure—statistical regularities—from input data, and encoding that structure into the parameters...”
[Zador 2019]

“Learning” is a Metaphor



- 1959: Arthur Samuel “teaches” computers to play checkers
- A program would iteratively optimize playing strategy, hence “learn to play”
 - Play, get feedback, remember, play better next time
- In computational view, learning is a special type of **optimization**
 - Optimize the automatic “behaviour” of a system to perform certain tasks
 - These tasks usually don’t have simple solutions



Do my shoes “learn” the shape of my feet??

About 4,610,000 results (0.39 seconds)

www.youtube.com › watch

This AI Learned To Stop Time! - YouTube



This AI Learned To Stop Time! ❤️ Check out Lambda here and sign up for their GPU Cloud: - The paper ...
Apr. 2, 2021 · Uploaded by Two Minute Papers

www.youtube.com › watch

This AI Learned To Animate Humanoids - YouTube



Check out Lambda here and sign up for their GPU Cloud: https://lambdalabs.com/papers The paper "Neural ...
Nov. 16, 2019 · Uploaded by Two Minute Papers

www.youtube.com › watch

This AI Learned to Summarize Videos - YouTube



Check out Linode here and get \$20 free credit on your account: https://www.linode.com/papers The paper ...
Apr. 18, 2020 · Uploaded by Two Minute Papers

www.youtube.com › watch

This AI Learned To See In The Dark - YouTube



The paper "Learning to See in the Dark" and its source code is available ...
May 29, 2018 · Uploaded by Two Minute Papers

www.sciencemag.org › news › 2018/05 › how-researcher...

How researchers are teaching AI to learn like a child | Science ...



Over time, artificial intelligence (AI) has shifted from algorithms that rely on programmed rules and logic ...
May 24, 2018 · Uploaded by Science Magazine

www.youtube.com › watch

This AI Learned To Create Dynamic Photos! - YouTube



Check out Weights & Biases and sign up for a free demo here: https://www.wandb.com/papers ❤️ Their report on ...
Jan. 26, 2021 · Uploaded by Two Minute Papers

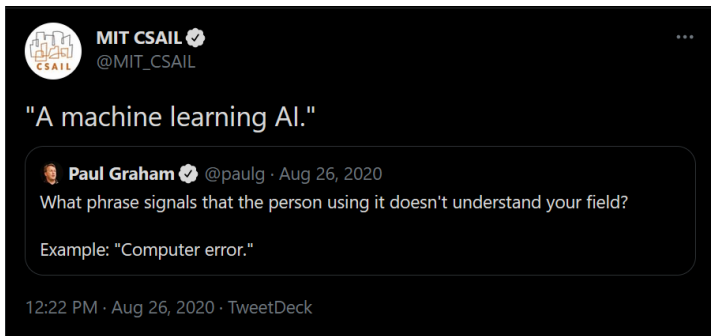
futurism.com › ai-learn-mistakes-openai

New Algorithm Lets AI Learn From Mistakes, Become a Little ...

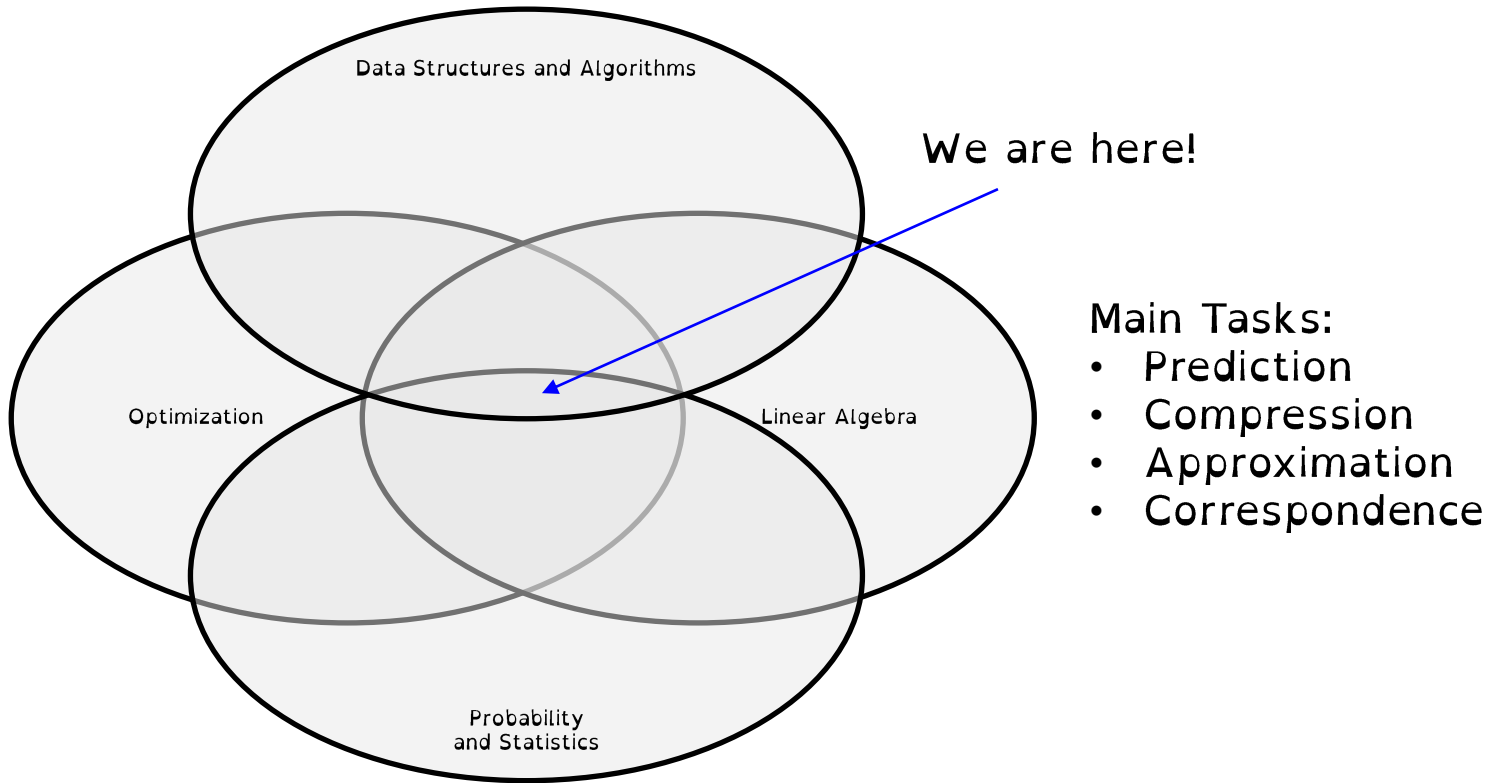


OpenAI's latest algorithm lets AI learn from its mistakes by re-framing past failures. This method helps AI to ...
Mar. 2, 2018 · Uploaded by OpenAI

- An intelligent system can use learned solutions
- Learned solutions are optimized with data
- Does that mean the AI learns to do something? (I don't know)

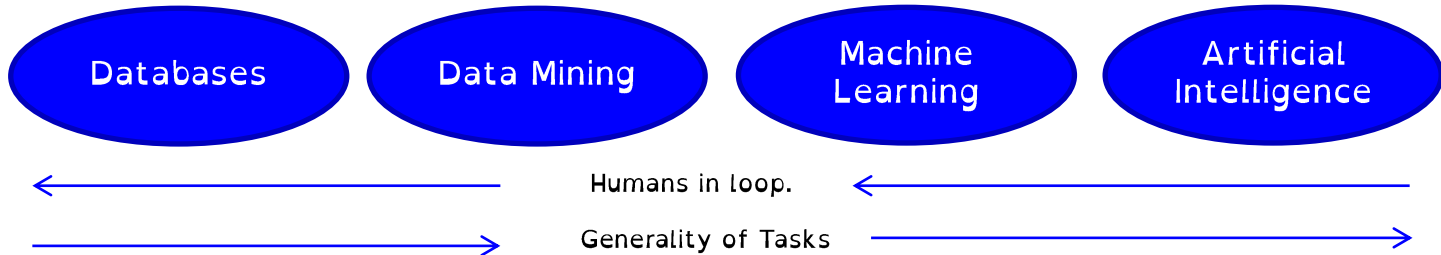


CPSC 340's View of ML



Data Mining vs. Machine Learning

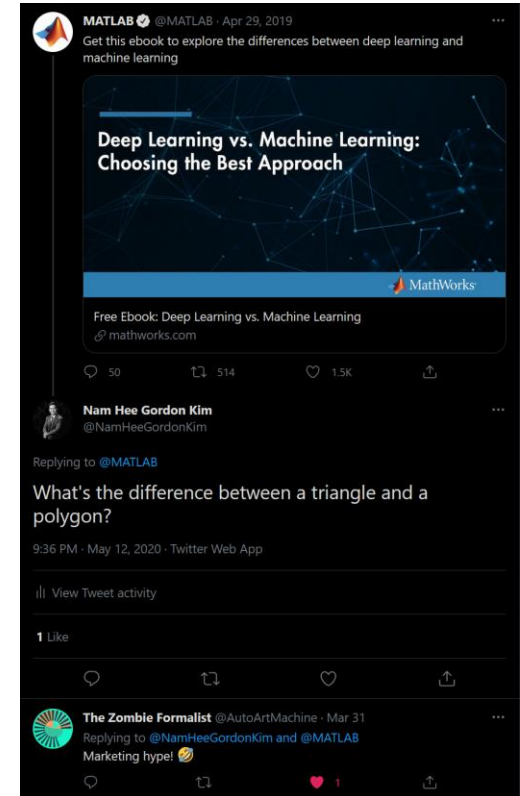
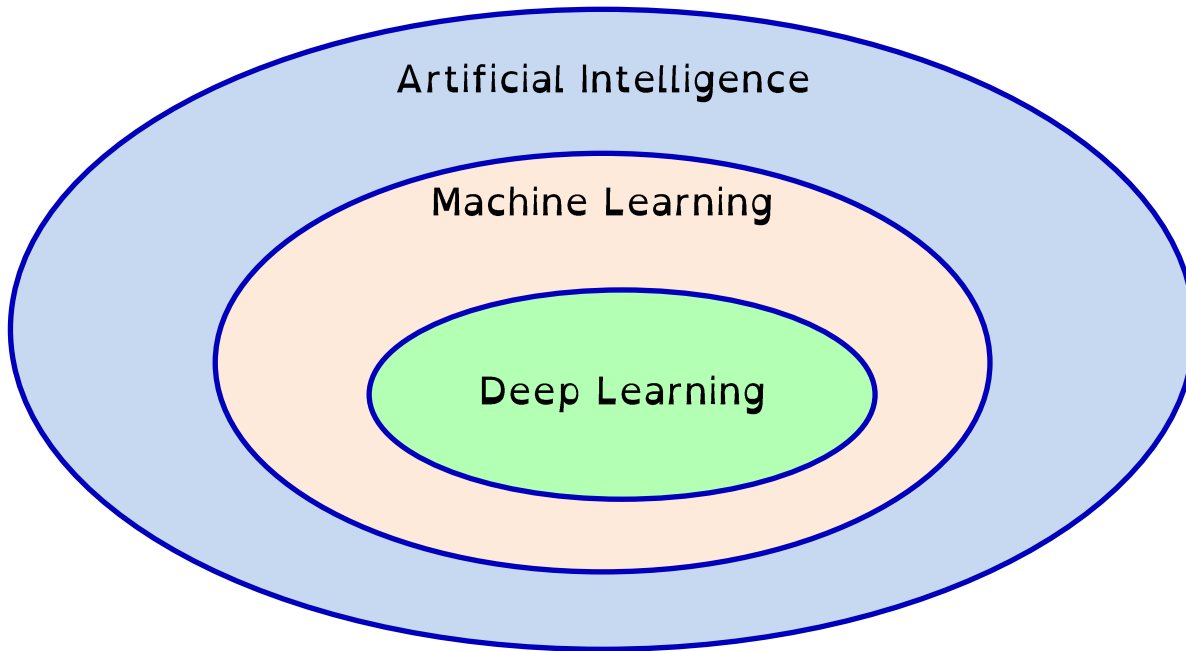
- Data mining and machine learning are very similar:
 - Data mining often viewed as closer to databases.
 - Machine learning often viewed as closer AI.



Q: How are these different from statistics?

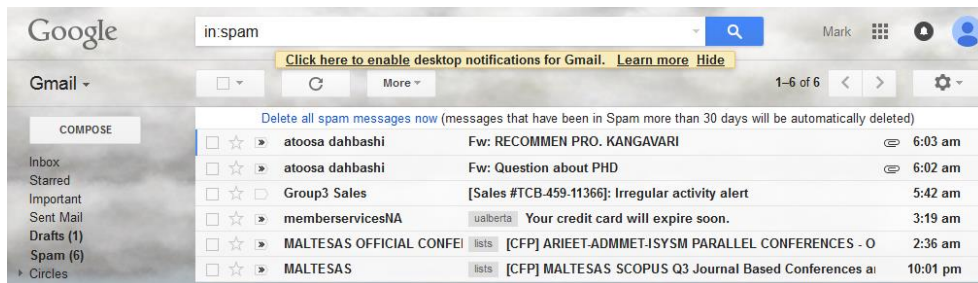
Deep Learning vs. Machine Learning vs. AI

- Traditionally we've viewed ML as a subset of AI.
 - And “deep learning” as a subset of ML.



Coming Up Next

APPLICATIONS OF MACHINE LEARNING



Spam filtering

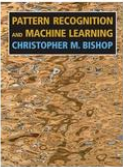
Transaction Date	Posted Date	Transaction Details	Debit	Credit
Aug. 27, 2015	Aug. 28, 2015	BEAN AROUND THE WORLD VANCOUVER, BC	\$10.95	

Fraud detection


Customers Who Bought This Item Also Bought

Page 1 of 20


<



Pattern Recognition and Machine Learning
(Information Science and...)
Christopher Bishop
★★★★☆ 115
Hardcover
\$60.76



Learning From Data
Yaser S. Abu-Mostafa
★★★★☆ 88
Hardcover



The Elements of Statistical Learning: Data Mining, Inference, and Prediction...
Trevor Hastie
★★★★☆ 50
Hardcover
\$62.82



Probabilistic Graphical Models: Principles and Techniques (Adaptive...
Daphne Koller
★★★★☆ 28
Hardcover
\$91.66



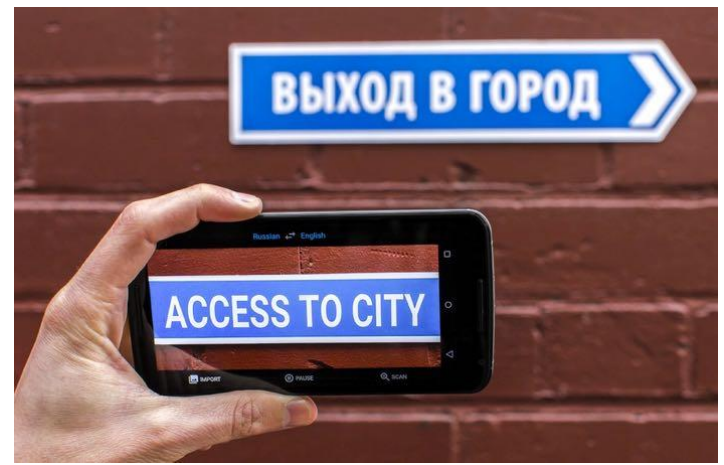
Foundations of Machine Learning (Adaptive Computation and...
Mehryar Mohri
★★★★☆ 8
Hardcover
\$65.68

>

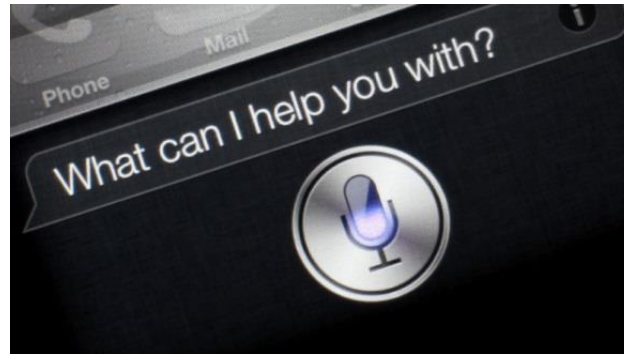
Product recommendation



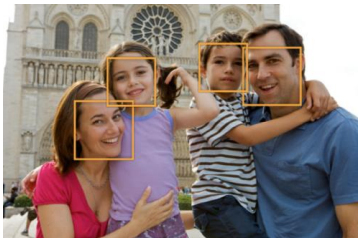
Motion capture



Character recognition and translation



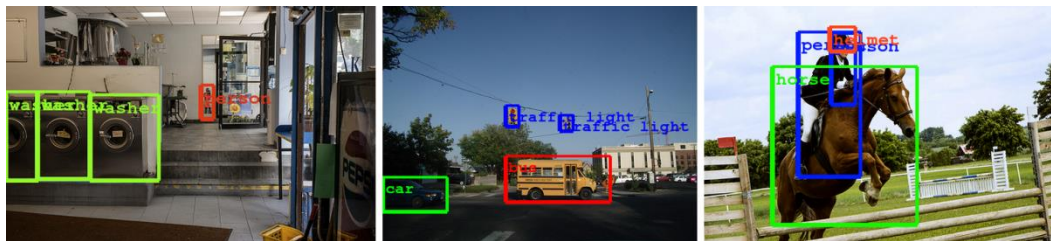
Speech processing and smart interface



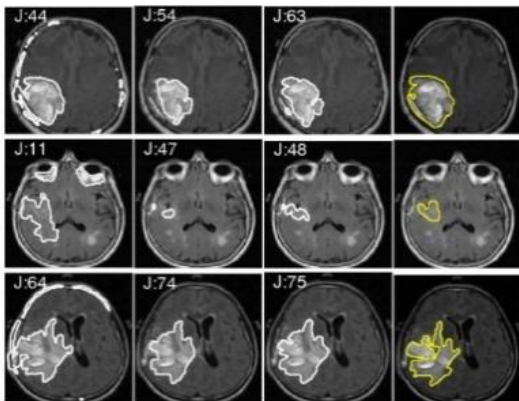
Face detection and recognition



Sports analytics



Object detection and recognition



Medical imaging



Personalized healthcare



Autonomous driving



Image editing



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

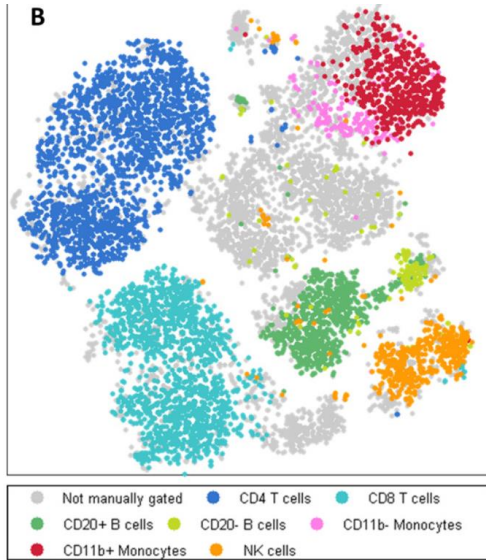


a display case filled with lots of different types of donuts
logprob: -7.78



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

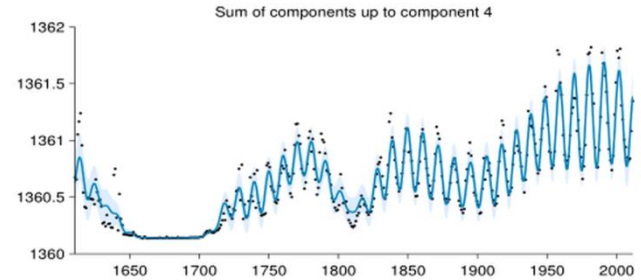
Image captioning



Finding new cancer types

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

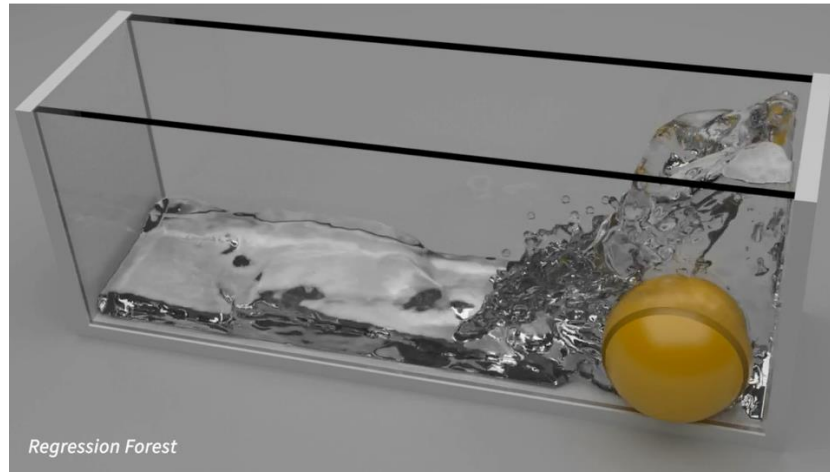
This component is approximately periodic with a period of 10.8 years. Across periods the shape of this function varies smoothly with a typical lengthscale of 36.9 years. The shape of this function within each period is very smooth and resembles a sinusoid. This component applies until 1643 and from 1716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



Automated statistician



Mimicking art styles



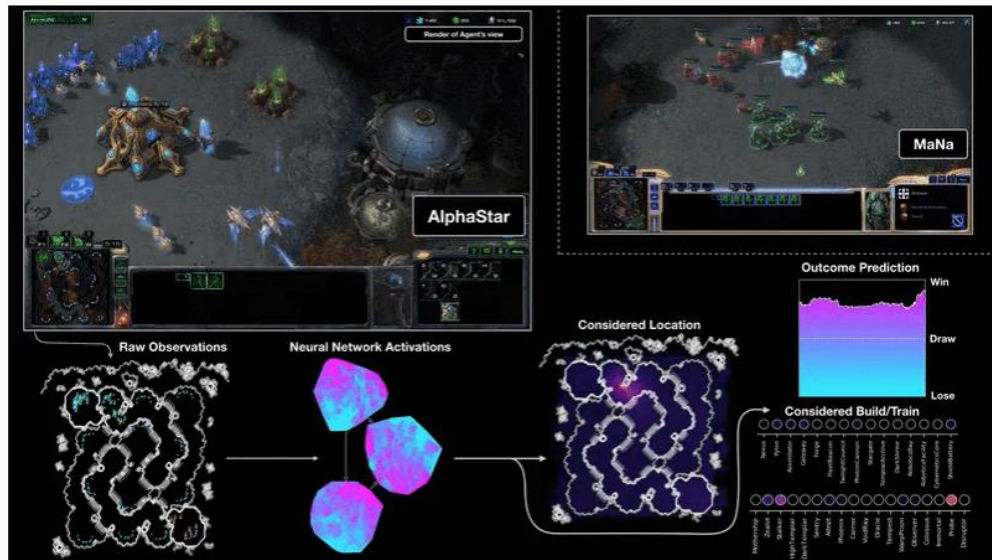
Accelerating physics simulations



Character animation



Playing Go

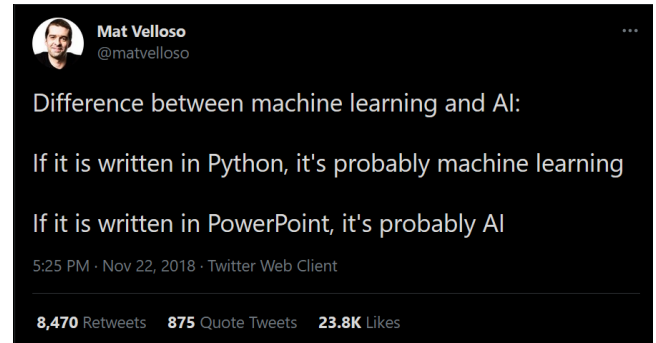


Playing StarCraft II



Playing Dota 2

- A bit of statistics + a lot data/computation = LOTS of interesting things
- We are in exciting times!
 - Things are changing a lot on the timescale of 3-5 years.
 - A bubble in ML investments (most “AI” companies are just doing ML).
- 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”.



Coming Up Next

FAILURES OF MACHINE LEARNING


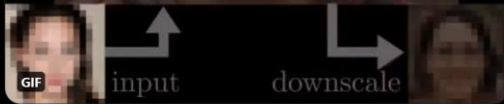
Bomze @tg_bomze · Jun 19, 2020

Face Depixelizer

Given a low-resolution input image, model generates high-resolution images that are perceptually realistic and downscale correctly.

🐱 GitHub: github.com/tg-bomze/Face-...
 📄 Colab: colab.research.google.com/github/tg-bomz...


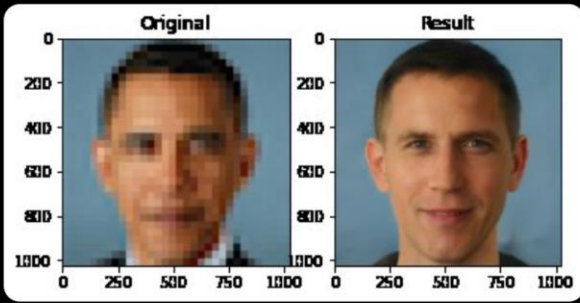
P.S. Colab is based on the github.com/adamian98/pulse

516 4.4K 11.1K

Chicken3gg @Chicken3gg

Replying to @tg_bomze

5:14 AM · Jun 20, 2020 · Twitter for Android

2,887 Retweets 1,192 Quote Tweets 23.1K Likes

Racial bias

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

Gender bias

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



Self-driving failure



Sherif Elsayed-Ali ∞
@sherifea

No



We don't need to go back to the office to be creative, we need AI
Remote working is making us more efficient, but is seriously damaging
innovation. AI could change that
[wired.co.uk](https://www.wired.co.uk)

12:05 AM · Feb 25, 2021 · Twitter Web App

2 Retweets 10 Likes



Philippe Beaudoin @PhilBeaudoin · Feb 25

Replying to @sherifea

AI. The shortest wrong answer to all the world's problems.



Bottom line: current trend of ML/AI worship is **not healthy**.
Be a machine learning skeptic!

Coming Up Next:

ADMINISTRATIVE STUFF

Lectures

- Lectures: MWF 9:30am-11:50am (140 minutes)
 - Part 1: 9:30am-10:25am (55 minutes)
 - break and self-study (30 minutes)
 - Part 2: 10:55am-11:50am (55 minutes)
- All slides will be posted online (before lecture, and final version after).
 - I'll also post Mike's demos on the webpage.
- Lectures will be recorded and posted on Canvas.
- **Please ask questions:** you probably have similar questions to others.
 - I may deflect to the next lecture or Piazza for certain questions.
- **Be warned:** 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.
- **Isn't it wrong to have only have shallow knowledge?**
 - In this field, it's **better to know many methods** than to know 5 in detail.
 - This is called the "no free lunch" theorem: different problems need different solutions.

Essential Links

- Please **bookmark** the course webpage:
 - <https://cs.ubc.ca/~nhgk/courses/cpsc340s21/>
 - Contains lecture slides, assignments, optional readings, additional notes.
- You should sign up for Piazza:
 - <https://piazza.com/ubc.ca/summer2021/cpsc340911>
 - 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”.
- Use Piazza instead of email for questions

Instructor

- **First name: Nam Hee (IPA: nɒmhi:)**
- **Last name: Kim**
- **English name that I never use: Gordon**
- **B.Sc., Rice University (2012-2016)**
- **B.Sc., UBC (2016-2018)**
- **M.Sc., UBC (2019-2021)**
- **Ph.D., Aalto University (2021-)**



Special Thanks



Mark Schmidt



Mike Gelbart

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).
 - “MATH 307, STAT 305, STAT 306, CPSC 320, CPSC 406 in One Breath”
- If you’ve only taken a few math courses (or have low math grades), **this course will ruin your life for the next 1.5 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).

Table of Contents
Part I: Mathematical Foundations
1 Introduction and Motivation
2 Linear Algebra
3 Analytic Geometry
4 Matrix Decompositions
5 Vector Calculus
6 Probability and Distribution
7 Continuous Optimization
Part II: Central Machine Learning Problems
8 When Models Meet Data
9 Linear Regression
10 Dimensionality Reduction with Principal Component Analysis
11 Density Estimation with Gaussian Mixture Models
12 Classification with Support Vector Machines

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 1.5 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.

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.
 - TAs shouldn't have to know many languages to grade.

Anecdote

- I took this course in 2017W1 and did well.

CAVEATS:

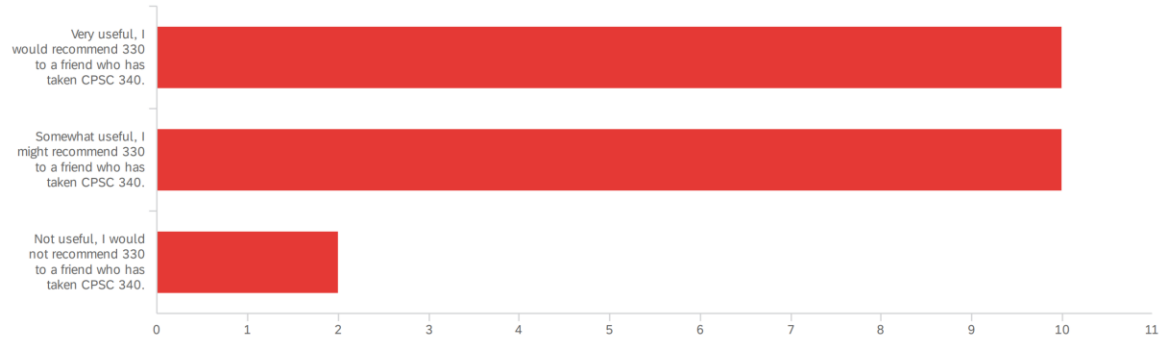
- It was the **ONLY** course I was taking
- I did research projects involving numerical computing (MATLAB)
- I was CMAJ CPSC/STAT and had taken these courses:
 - MATH 303, STAT 305, STAT 306, MATH 307, CPSC 320, CPSC 310

CPSC 330 vs. CPSC 340

- There is also a **less-advanced ML course, CPSC 330**:
 - “Applied Machine Learning”, taught by Mike Gelbart.
 - 330 emphasizes “**when to use**” tools, 340 emphasizes “**how they work**”.
 - 330 is more like the Coursera course and online courses.
 - **Fewer prerequisites**:
 - 330 spends more time on **low-level coding details** and has **basically no equations**.
 - More “learning by doing” and less discussion of fundamental principles.
 - 330 spends more time on **data cleaning, communicating results**, and so on.
 - More emphasis on the entire “pipeline” of data of analysis.
 - 330 **cannot be used as a prereq** for the more-advanced CPSC 440.
 - You **can take both** for credit (better to take 330 first or at same time).

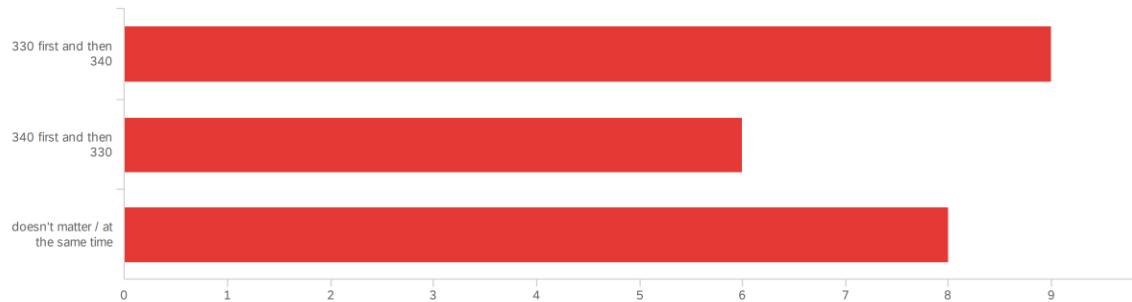
Q4 - Please rate how useful CPSC 330 was to you as someone who has taken CPSC 340

340.



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

ultimately takes both courses?



CPSC 340 vs. CPSC 440/540

- There is also a **more-advanced ML course, CPSC 440/540**:
 - 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 440/540 focuses on less widely-used methods and research topics.
 - It is intended as a continuation of CPSC 340 with even more math.
 - You'll miss important topics if you skip CPSC 340.

Textbooks

- No required textbook.
- I'll post relevant sections out of these books as optional readings:
 - Artificial Intelligence: A Modern Approach (Russell & 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 available online through UBC Library.
- List of related courses on the webpage, or you can use Google.

Excellent TAs



Austin Beauchamp



Peyman Gholami



Lironne Kurzman



Farnoosh Hashemi



Gabriel Huang



Shahriar Shayesteh



Mohamad Amin Mohamadi



Ali Seyfi



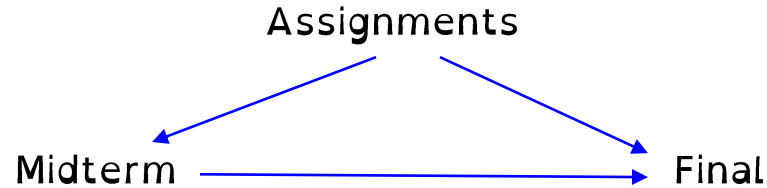
Frank Yu



Yuxin Tian

Grading Breakdown

- Assignments: 30%
- Midterm: 20%
- Final: 50%



Doing well on assignments will help you do well on midterm.
Doing well on assignments and midterm will help you do well on final.

Assignments

- There will be **6 Assignments** worth 30% of final grade
 - Usually a combination of math, programming, and very-short answer.
- **Assignment 1 is on webpage**, and is due **next Monday at 9:25am PST**.
 - Submission instructions will posted on webpage/Piazza.
 - The assignment should **give you an idea of expected background**.
 - Make sure to submit before the deadline and check your submission.

Homework Assignments

Post Date	Due Date	Files	Notes/Links
Mon May 10	Mon May 17	a1.pdf · a1.zip (contains the LaTeX template + code)	Setting up Python

- **Start early!** 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 Q1-Q4 already.

Working in Teams for Assignments

- Assignment 1 must be done individually.
- *Assignments 2-6 can optionally be done in pairs.*
 - No need to have the same partner for all assignments.

Notes on Working Together

- Divide-and-conquer may be counter-productive!
 - Very common case (happened to me):
 - Member 1 does everything on time,
Member 2 waits until last minute
Member 1 freaks out and does Member 2's work
 - Hard to be on the same page
 - What if Member 1's portion shows up on exam?
 - Member 2 has to learn Member 1's portion from scratch right before exam (defeats the purpose of the assignment!)
- Most students using a **scrum-style workflow** struggled more than those who **pair-programmed!**

Notes on Working Together

- Consider a “synchronized sprint” model of teamwork
 - Members do Q1 independently
 - Members compare answers and discuss*
 - Then Members do Q2 independently
 - Members compare answers and discuss*
 - Etc.
- Ideally, a stronger student should lead a weaker student, while keeping workload even.

*use office hours for any questions that emerge!

Late Submissions

- By default, we will allow **no late submissions!**
 - Assignments are due immediately before Monday lectures (9:25am PST).
 - Summer terms have to move extra fast.
 - If you really need late submission, talk to me.
 - Will grant extra time in exceptional cases (e.g. family emergency)
- We'll release solutions to assignments on Wednesdays

Assignment Issues

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

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

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.
 - Times will be posted on the course webpage and Canvas.
 - **Piazza** (for general questions).
 - **Weekly tutorials** (optional).
 - Starting TODAY.
 - Will go through Python setup, look at provided code, provide tips for homework problems.
 - **Other students**
 - **The web** (almost all topics are covered in many places).

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.
 - We have actually given 0 to people before.

Code of Conduct

- Do not post offensive or disrespectful content on Piazza.
- 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, Piazza, 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.
- There will be no post-course grade changes based on grade thresholds:
 - 48% will not be rounded to 50%, and 70% will not be rounded to 72%, and so on.

Notes on Question-Asking

- Your questions are extremely important (to you and me)
- Student questions are **canon**
 - I will take note of good questions asked in and out of class
 - I might turn some of these into exam questions
- Piazza is **canon**
 - Similarly, I might sample questions from Piazza to build exams.
- When asking, start with lecture/slide/assignment/question numbers
 - E.g. “Lecture 17’s slide 43 says that L1 loss is robust to outliers. Can you clarify what robust means in this context?”
 - E.g. “In A2 Question 5, do we need to submit the learning curve plot?”
 - If posting on Piazza, include screenshots

How to Ask Good Questions

- Provide analogy whenever possible
 - E.g. “My impression is that gradient descent is similar to hill-climbing, using gradients as some kind of search heuristic. Is this correct?”
- Ask about the position of the topic in relation to other topics in the course
 - E.g. “How does PCA with $k < d$ relate to vector quantization? They both seem to perform compression.”
 - E.g. “Aren’t latent factor models just special cases of unsupervised learning?”
- Use active recall (I recall that...)
 - E.g. “I recall that convex functions can have multiple global minima. Does that mean there can be multiple linear models that perform equally well in testing?”

Midterm and Final

- Midterm worth 20% and a (cumulative) final worth 50%
 - Midterm: open-book, on Canvas AND Gradescope, 90 minutes.
 - Final: open-book, on Canvas AND Gradescope, 150 minutes.
 - Can take any time during the day, but must complete in one seating
 - No need to pass the final to pass the course (but recommended).
- Midterm is tentatively scheduled for Monday, May 31st.
 - Let us know if you have a conflict that cannot be resolved.
- Final date is TBA. Don't make travel plans before Friday, June 25th.
- 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.

I Will Make Your Lives Easier

- If I find something very exam-worthy in my slides, I will highlight the title **green**.
 - Can be based on student questions
 - Can be based on my opinion
 - Disclaimer: may or may not be actually on the exam
- I will provide practice exams
 - so you get familiar with the format.
- In lecture slides, I will include **lots of questions** to help you review for exams.

Videos from Previous Offering

- Videos of Mike's January 2018 offering of the course:
 - https://www.youtube.com/playlist?list=PLWmXHcz_53Q02ZLeAxigki1JZFfCO6M-b
- You may find these useful:
 - Material is almost identical, but now you can rewind (or fast-forward).
 - Mike is a more experienced teacher than I am.

Bonus Slides

- I will 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 540 or work in this area.
- I’ll use this colour of background on bonus slides.

Course Outline

- Next lecture 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)

30-minute Break and Self-Study

- Sign up on Piazza
- Check out Assignment 1 (you can do Q1-Q4!)
- Check out recommended readings

Stop Calling Everything AI, Machine-Learning Pioneer Says

Michael I. Jordan explains why today's artificial-intelligence systems aren't actually intelligent

Artificial Intelligence—The Revolution
Hasn't Happened Yet

by Michael I. Jordan

Published on Jul 01, 2019

Machine Learning: The Great Stagnation

The bureaucrats are running the asylum



Mark Saroufim Dec 4, 2020 · 14 min read ★