CPSC 340 ~~and 532M~~: Machine Learning and Data Mining

Frank Wood

University of British Columbia, Fall 2020

https://www.cs.ubc.ca/~fwood/CS340/

Some images from this lecture are taken from Google Image Search, contact Mark Schmidt if you want the reference

Dr. *Frank* Wood

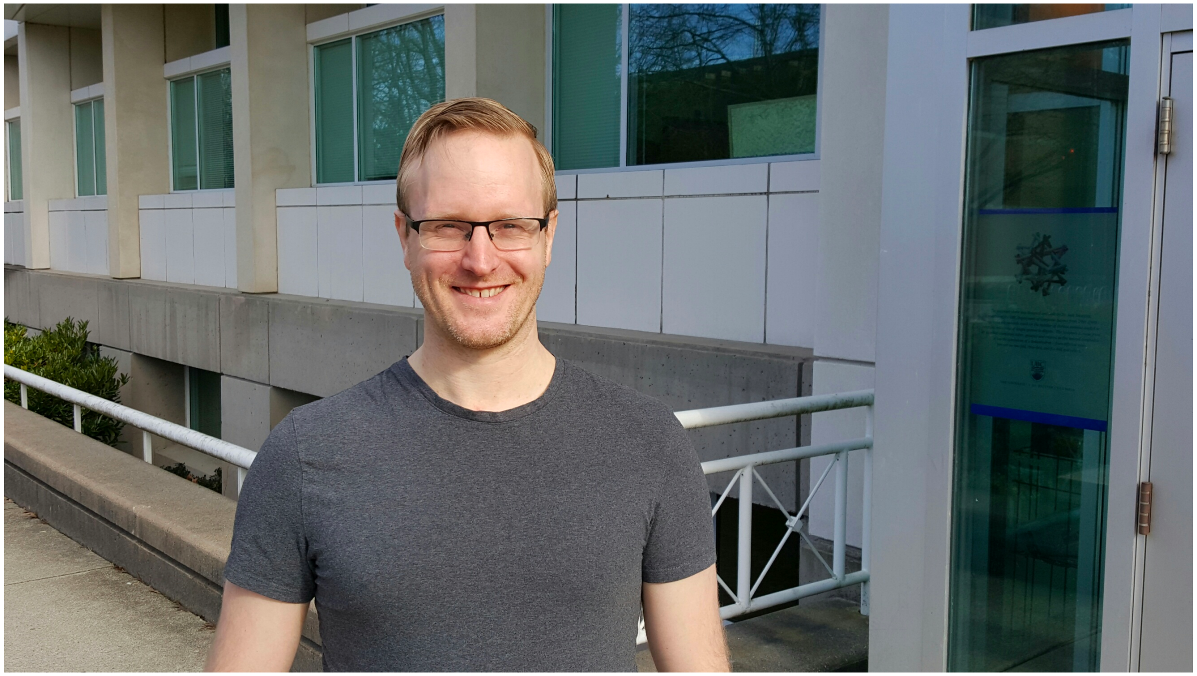
- Academic
  - BS, Computer science, Cornell
  - MS+PhD, CS/Computational Neuroscience, Brown
  - Postdoc, Gatsby Unit, UCL
  - Assistant Professor, Statistics, Columbia
  - Associate Professor, Engineering, Oxford
  - Associate Professor, Computer Science, UBC
  - CIFAR AI Chair affiliated to the Mila institute
- Principle **research** focus
  - Deep probabilistic programming, unsupervised machine learning, artificial intelligence
- Entrepreneur
  - Founder of UBC spin-out **Inverted AI** (bots for autonomous driving simulators)
  - CEO of first image search engine on web (sold in 1999 to AOL; $5M)
  - Investor/Founder Betacular (sold in 2017; undisclosed)
  - CEO of Interfolio (sold 2018; $110M)
- Social media
  - frankdonaldwood @twitter

PLAI Group

- Programming Languages for Artificial Intelligence
  - 2 Postdocs
  - 12 PhD students
  - 1 MS student

- Funding and (Projects)
  - DARPA (AutoML, LwLL)
  - Intel (Etalumis)
  - NSERC, CIFAR, Compute Canada, etc.

Credit



- Mark Schmidt: slides, homeworks, general course design, etc.
  - Michael Gelbart online lecture recordings, python homeworks, etc.

SCHOLAR STRIKE CANADA

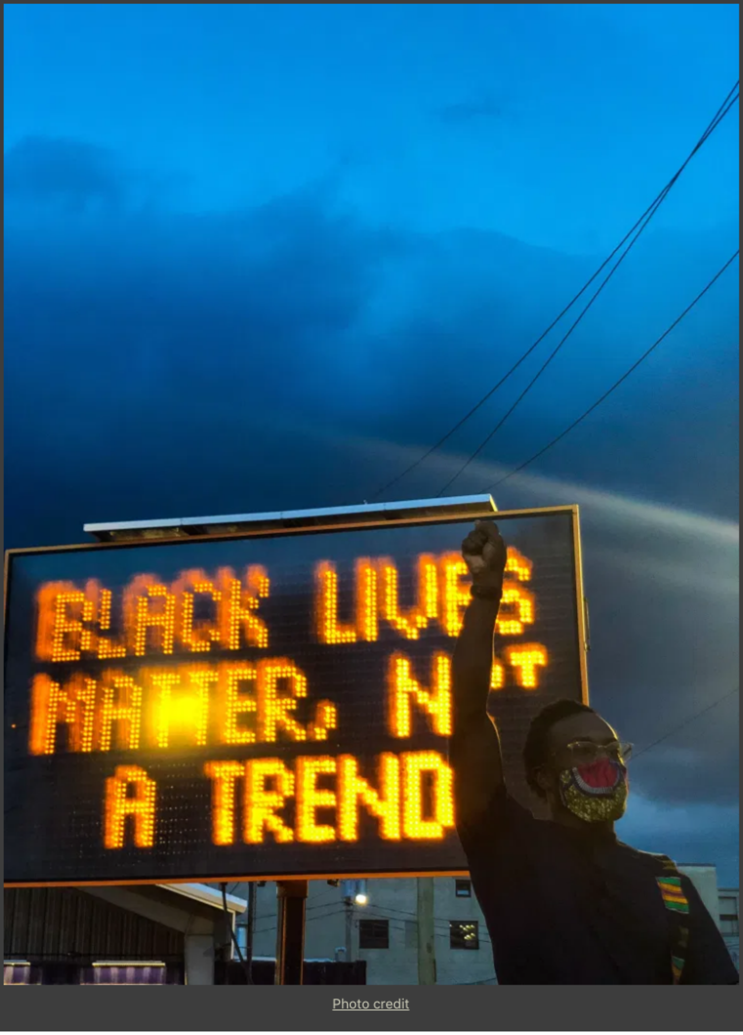
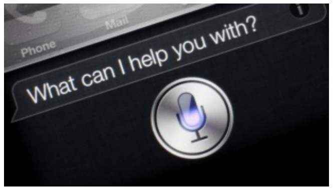


Photo credit

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.



What can I help you with?

Big Data Phenomenon

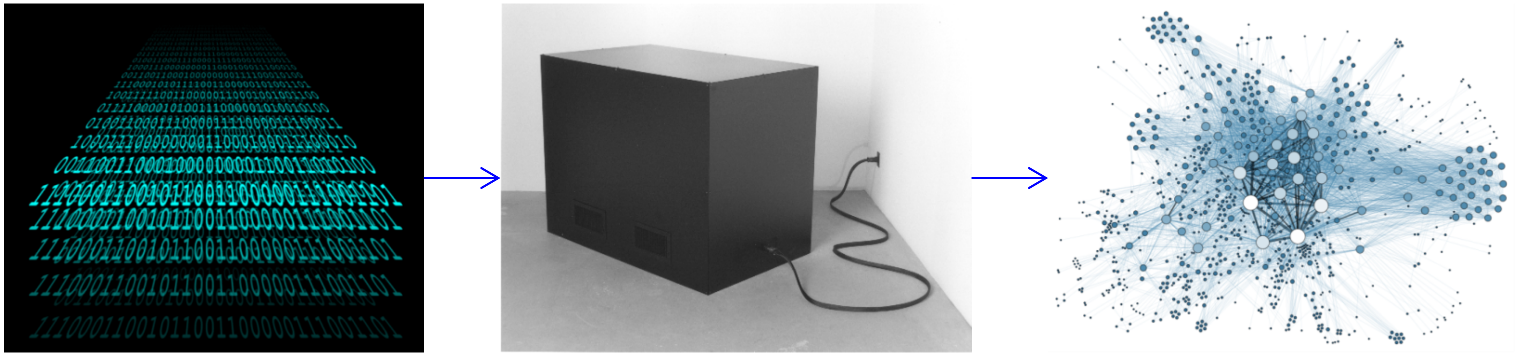
- What do you do with all this data?
  - Too much data to search through it manually.

- 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

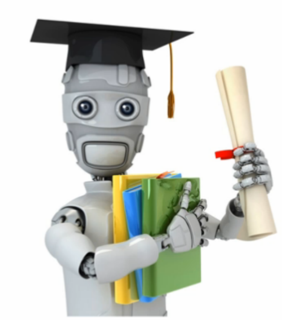
- Automatically extract useful knowledge from large datasets.

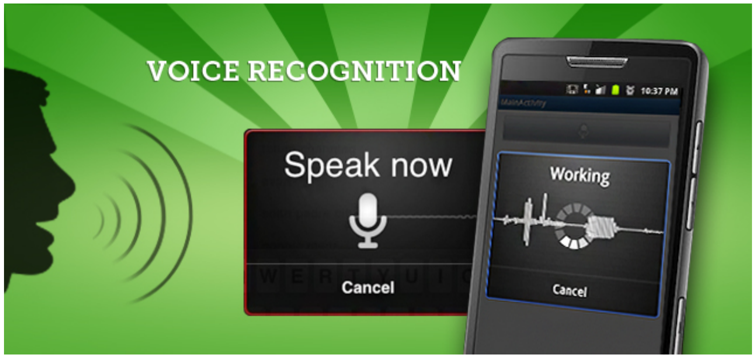


- Usually, to help with human decision making.

Machine Learning

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

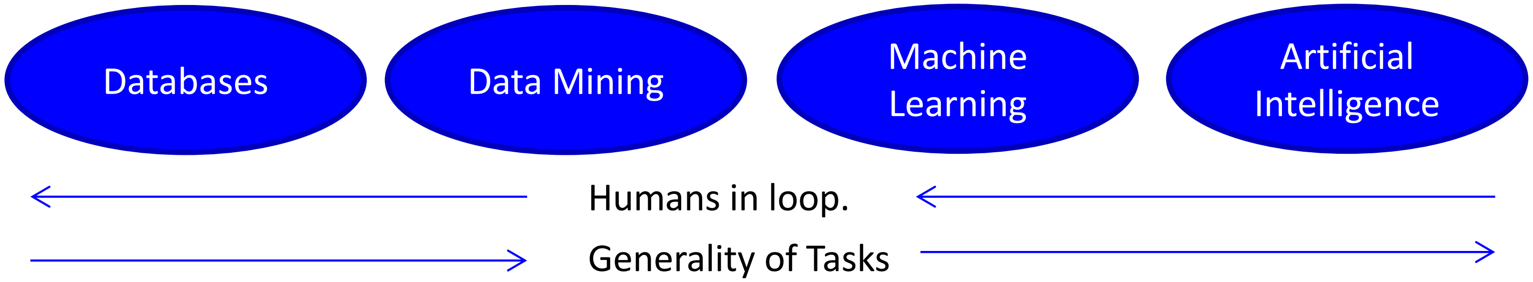




- Most useful when:
  - We want to automate something a human can do.
  - We want to do things a human can't do (look at 1 TB of data).

Data Mining vs. Machine Learning

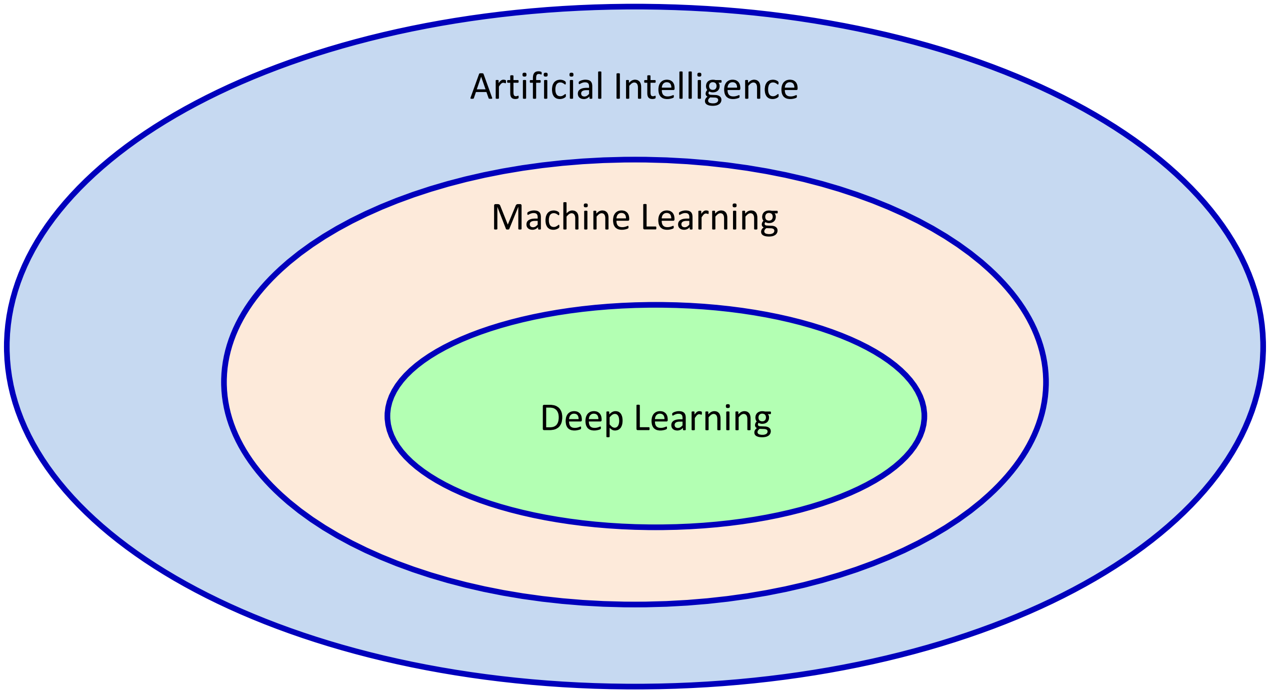
- Data mining and machine learning are very similar:
  - Data mining often viewed as closer to databases.
  - Machine learning often viewed as closer AI.



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

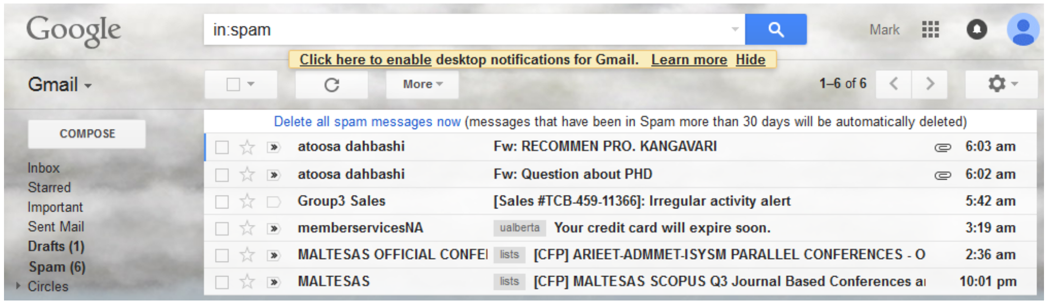
Deep Learning vs. Machine Learning vs. AI

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



Applications

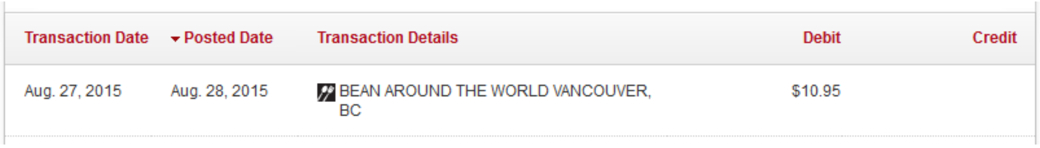
- Spam filtering:
- Credit card fraud detection:
- Product recommendation:



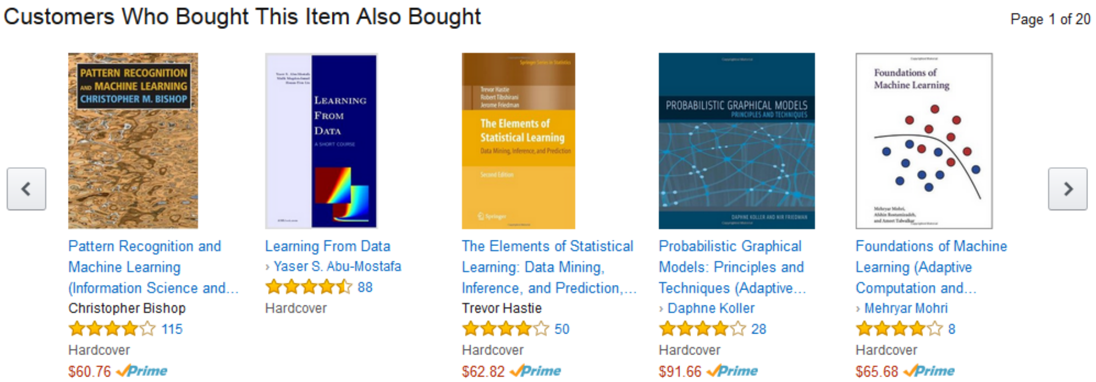
Click here to enable desktop notifications for Gmail. Learn more Hide

Delete all spam messages now (messages that have been in Spam more than 30 days will be automatically deleted)

atoosa dahbashi		Fw: RECOMMEN PRO. KANGAVARI	6:03 am
atoosa dahbashi		Fw: Question about PHD	6:02 am
Group3 Sales		[Sales #TCB-459-11366]: Irregular activity alert	5:42 am
memberservicesNA	ualberta	Your credit card will expire soon.	3:19 am
MALTESAS OFFICIAL CONFEI	lists	[CFP] ARIEET-ADMMET-ISYSM PARALLEL CONFERENCES - O	2:36 am
MALTESAS	lists	[CFP] MALTESAS SCOPUS Q3 Journal Based Conferences aı	10:01 pm



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



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 ✓Prime

Learning From Data
› Yaser S. Abu-Mostafa
★★★★½ 88
Hardcover

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

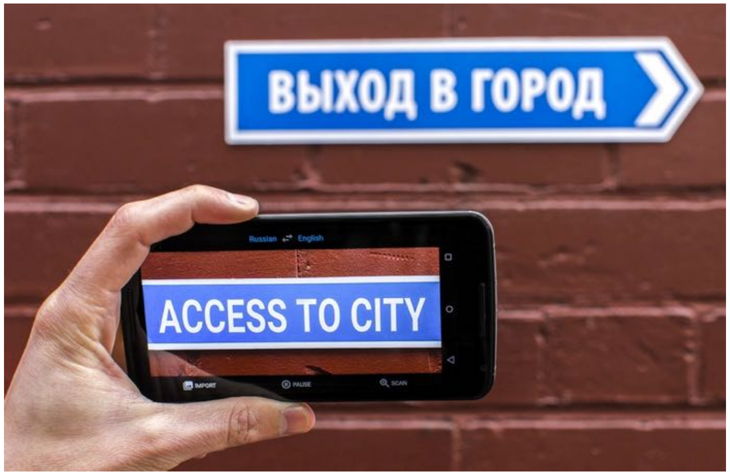
Probabilistic Graphical Models: Principles and Techniques (Adaptive...
› Daphne Koller
★★★★☆ 28
Hardcover
$91.66 ✓Prime

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

Applications

- Motion capture:

- Optical character recognition and machine translation:

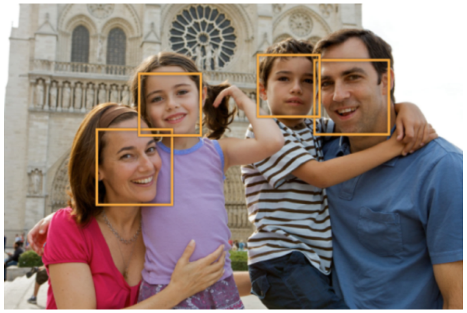


- Speech recognition:

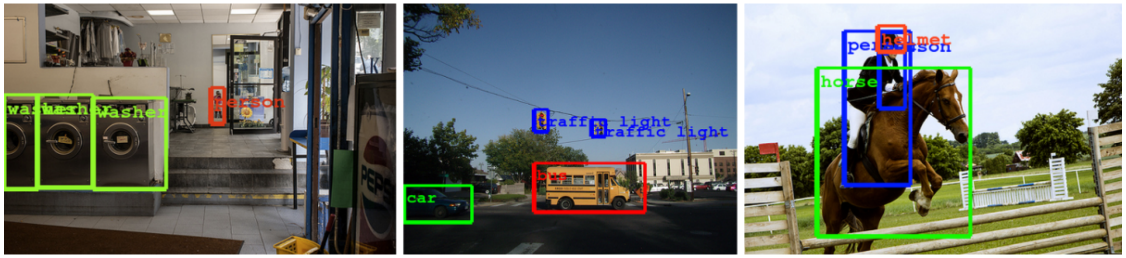
What can I help you with?

Applications

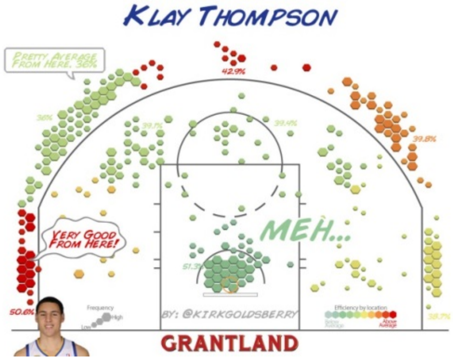
- Face detection:



- Object detection:

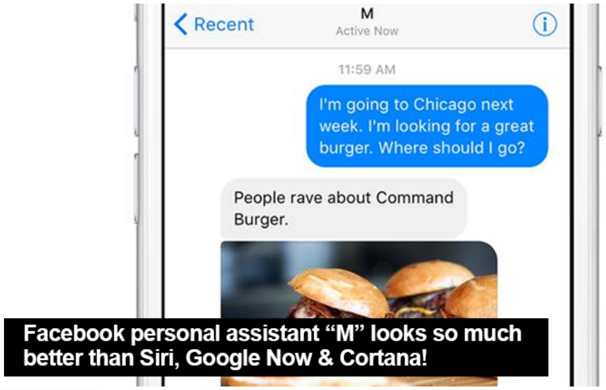


- Sports analytics:



Applications

- Personal Assistants:



Recent

M

Active Now

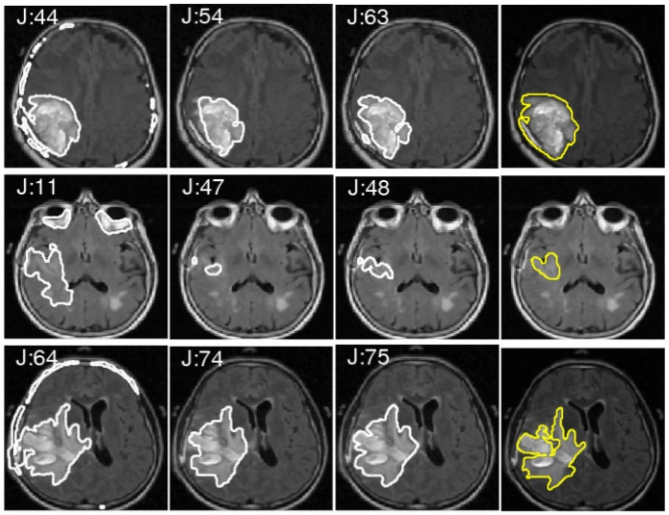
11:59 AM

I'm going to Chicago next week. I'm looking for a great burger. Where should I go?

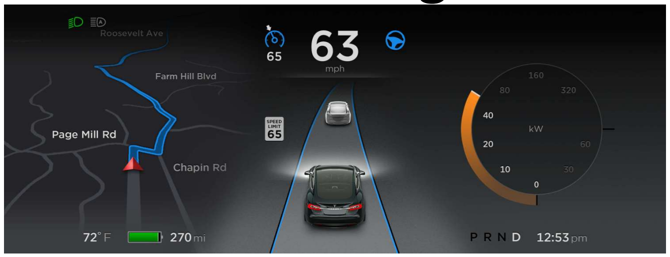
People rave about Command Burger.

Facebook personal assistant "M" looks so much better than Siri, Google Now & Cortana!

- Medical imaging:

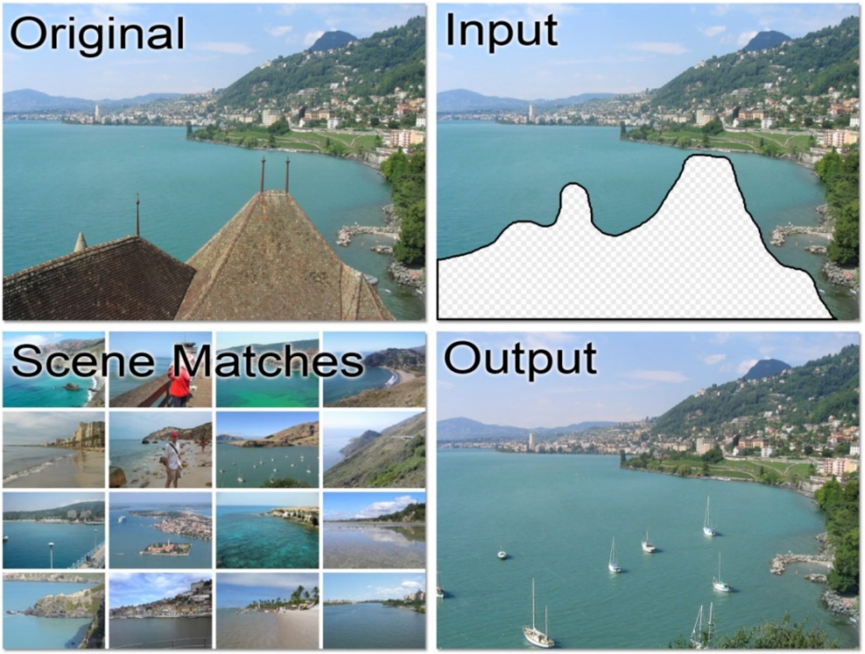


- Self-driving cars:



Applications

- Scene completion:



- Image annotation:

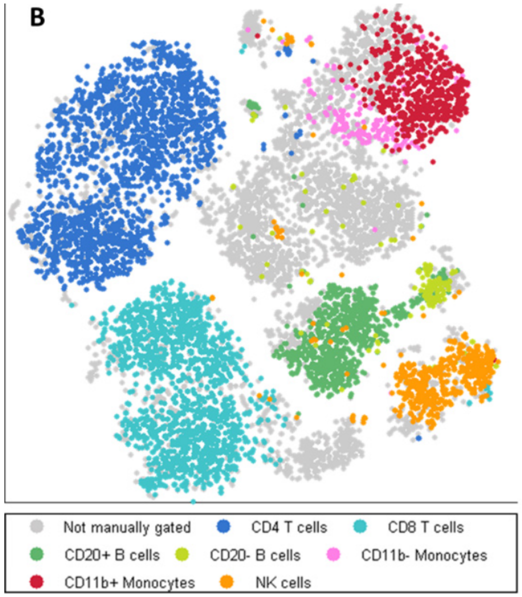
a cat is sitting on a toilet seat
`logprob: -7.79`

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

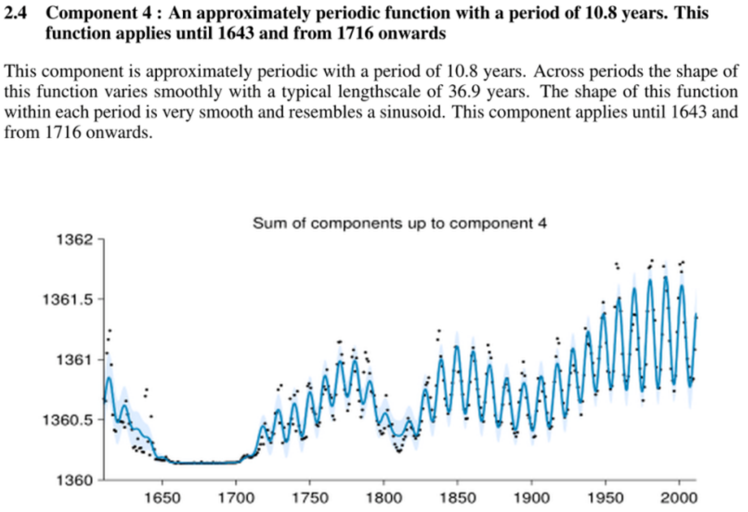
a group of people sitting at a table with wine glasses
`logprob: -6.71`

Applications

- Discovering new cancer subtypes:



- Automated Statistician:

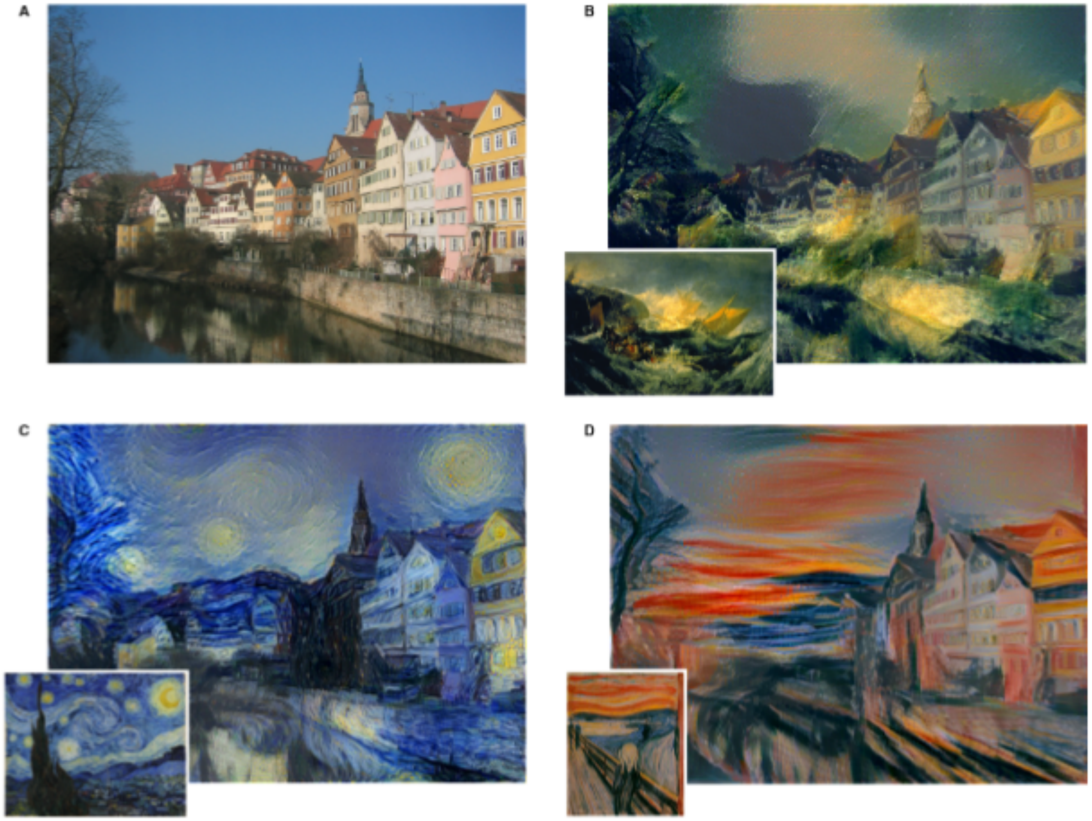


2.4 Component 4 : An approximately periodic function with a period of 10.8 years. This function applies until 1643 and from 1716 onwards

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.

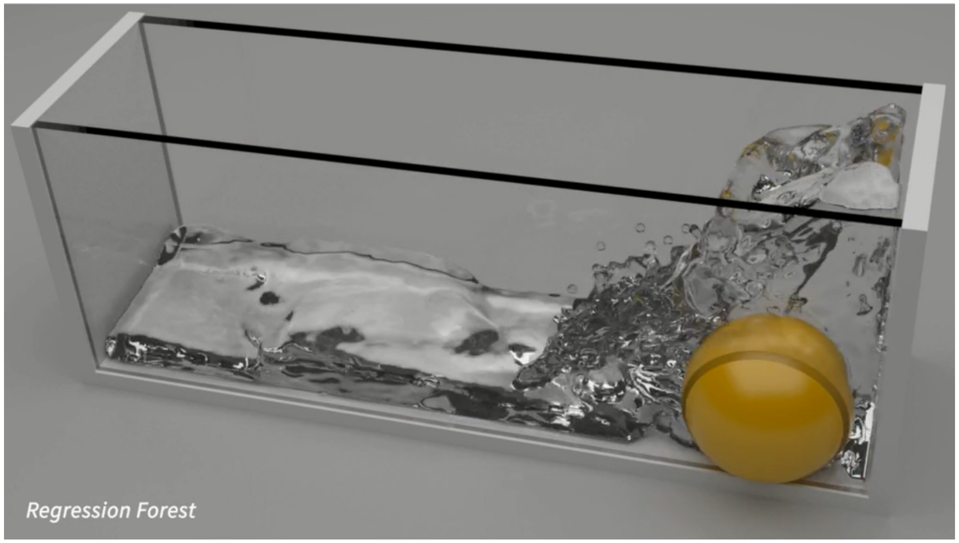
Applications

- Mimicking artistic styles:



Applications

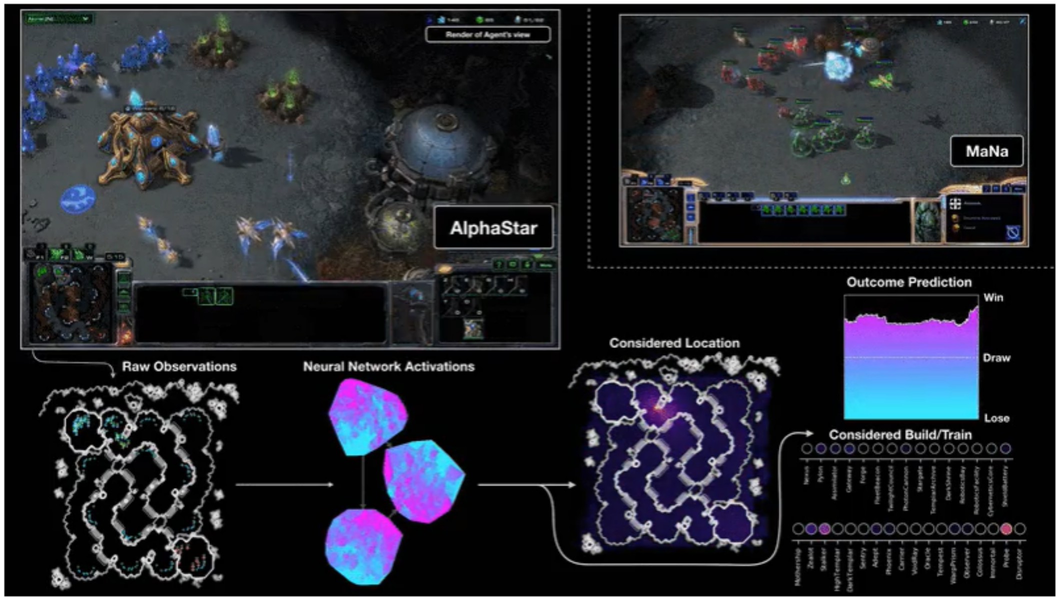
- Fast physics-based animation:



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

Applications

- Beating humans in Go and Starcraft:



- 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 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 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.

Programming Language: Python

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

- No, you cannot use Matlab/R/TensorFlow/Julia/etc.
  - Assignments have prepared code: we won’t translate to many languages.
  - TAs will not grade answers in other 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.
- 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.”

Different Sections of 340 ~~and 532M~~

- I am teaching both sections of 340 this term.
  - Both sections have the same webpage, assignments, and exams.

- You are free-ish to attend the lectures of the other section.
  - However, don't enter the zoom meeting until slightly after the start of the class time; the UBC zoom meeting participant limit would be hit if everyone shows up to one lecture.

- Lectures will cover roughly the same set of topics.
  - You will only be tested on material that appears in both sections.
  - They will be recorded and made available in the cloud through Canvas.

- Next term it will be taught by Mike Gelbart (probably in Python).
  - Mark and I are research faculty.
  - Mike Gelbart is teaching faculty: consider this as an option if you want a better experience.

Multiple Sections of 340 and 532M

- Another section of 340 is taught MWF at 12 by Mike Gelbart.
  - It has more seats available than this section does.
  - Both sections have the same GitHub webpage, assignments, and exams.

- Lectures won’t be identical but will cover roughly the same set of topics.
  - You will only be tested on material that appears in both sections.

- Differences between Mike and I:
  - Mike is a teaching faculty, I am a research faculty.
  - Mike does more in-class Python demos, I do more traditional lecturing.

- Neither of us mind if you attend the other person’s lectures.
  - However, don’t take a seat if you aren’t registered and people are standing.

- If you are a CS undergrad, consider switching to the other section (smaller class that is not full).
  - This will let more people on the waiting lists register.

CPSC 340 vs. 532M

- One section of CPSC 340 is also cross-listed as CPSC 532M.
  - For graduate students who want/need graduate credit.
- Students in CPSC 532M must do a small research project.
  - Literature survey on an ML topic not covered in class.
  - Must be done in groups of 2-3.
  - More details later.
- Grading will be slightly different:

Number	Assignments	Midterm	Final Exam	Survey
340	30	20	50	0
532M	25	15	40	20

CPSC 330 vs. CPSC 340

- There is also a less-advanced ML course, CPSC 330:
  - Fewer prerequisites (and probably lower workload).
  - You can take both for credit (if you do this then take 330 first).
  - 330 emphasizes "when to use" tools, 340 emphasizes "how they work".
  - 330 is more like the Coursera course and other online courses.

- From a former 340 student:
  - "I took Andrew Ng's Coursera course and had a lot of fun and so I would recommend it. But before you spend any time, the Coursera course (I feel) covers only a subset of the concepts covered in this class and wouldn't be an efficient way of gaining understanding of the course material."

CPSC 340 vs. CPSC 540

- There is also a more-advanced ML course, CPSC 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 540 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

- Please bookmark the course webpage:
  - https://www.cs.ubc.ca/~fwood/CS340/
  - Contains lecture slides, assignments, optional readings, additional notes.

- You should sign up for Piazza:
  - Can be used to ask questions about lectures/assignments/exams.
  - May occasionally be used for course announcements.
  - I **do not watch piazza**; do not message me directly. **TAs handle Piazza**.

- **I do not read or answer my own email.**

Textbooks

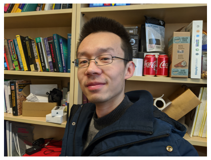
- No required textbook.

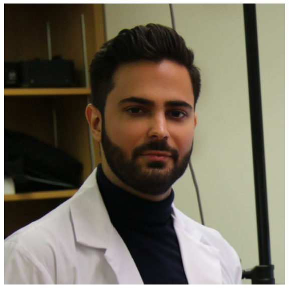
- I'll post relevant sections out of these books as optional readings:
  - Artificial Intelligence: A Modern Approach (Rusell & Norvig).
  - Introduction to Data Mining (Tan et al.).
  - The Elements of Statistical Learning (Hastie et al.).
  - Mining Massive Datasets (Leskovec et al.)
  - Machine Learning: A Probabilistic Perspective (Murphy).

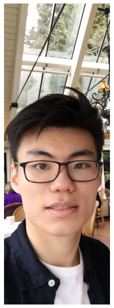
- Most of these are on reserve in the ICICS reading room.
- List of related courses on the webpage, or you can use Google.

TA Cheat Sheet

- Ivy Qiuhan
- Mark Ma
- Peyman Gholami
- Amit Kadan
- Larry Liu
- Shahriar Shayesteh*
- Erik Ryhorchuk
- Yancey Yang
- Yuxin Zhang







Assignments

- There will be 6 Assignments worth 40% of final grade:
  - Usually a combination of math, programming, and very-short answer.
  - Because of Covid-19 grading will be very terse and generally lenient, the product of:
    - Complete or not (0 for incomplete, 1 for complete)
    - Cheating or not (0 for cheating, 1 for not cheating)
  - You are responsible for learning – be honest with yourself – do the work!
  - Answers will be distributed via Piazza and you can compare your answers with the distributed answers!

- Assignment 1 is on the webpage, 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 before the deadline and check your submission.

- Start early, there is a lot there.
  - Don't wait to see you if get off the waiting list to start.

Working in Teams for Assignments

- Assignment 1 must be done individually.

- Assignments 2-6 can optionally be done in pairs.
  - You will use Gradescope to submit PDFs of your assignment (after editing it in latex to include code, figures etc). You specify your partner in gradescope using their partner's csID. Both partners need to submit a copy of the assignment
  - You don't need to have the same partner for all assignments.

- All the various permutations of partners are allowed:
  - Partnering with an auditor is ok.

Late “Class” Policy for Assignments

- Assignments will be due at 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.

- We’ll release solutions to assignments after 2 “late classes”.
  - We’ll try to put grades up within 10 days of this.

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 100% penalty:
  - Missing names or student IDs on assignments.
  - Not using your CS ID and/or associated @ugrad.cs.ubc.ca email alias (as listed in https://www.cs.ubc.ca/getacct) to identify yourself on Gradescope
  - Corrupted submission files or not using a .zip file.
  - 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 room is full, we will not have seats for auditors.

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.
    - Schedule is posted on the course webpage.
  - Piazza (for general questions).
  - Weekly tutorials (optional – attend any you wish to; multiple OK).
    - 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).

Midterm and Final

- Midterm worth 20% and a (cumulative) final worth 40%
  - Take-home.
  - No need to pass the final to pass the course (but recommended).

- Both the midterm and final will consist of two different kinds of questions and activities
  - Short written questions
  - Kaggle-competitions

Lectures

- All slides will be posted online (before lecture, and final version after).
- All lectures will be recorded. Zoom cloud recordings will be available through the canvas course website.

- 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.

- 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.

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 teaching faculty.

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.

Code of Conduct

- Do not post offensive or disrespectful content on Piazza.
- If you have a problem or complaint, let us know (maybe we can fix it).
- Do not distribute any course materials without permission.
- Do distribute recorded 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:
  - Because of Covid-19 effectively the course mark will effectively be pass/fail.
  - There will be three default marks: A+=100% A=90% D=50%.
  - I retain the right to assign any other mark for any other reason.

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)

Privacy Notice

Please note that this course uses Gradescope to facilitate the grading of exams and/or assignments. Your CS ID and/or associated @ugrad.cs.ubc.ca email alias (as listed in https://www.cs.ubc.ca/getacct) must be used to identify you on Gradescope. As Gradescope is hosted outside Canada, we want to remind you to keep your @ugrad alias private, just as you would any other account information. If you choose not to keep your @ugrad alias confidential, please note that UBC will proceed on the assumption that you do not object to Gradescope potentially identifying you personally, and that you are consenting to the storage of personal information on Gradescope servers outside Canada.