CPSC 440 and 540: Advanced Machine Learning

University of British Columbia, Jan-Apr 2023
www.cs.ubc.ca/~dsuth/440/22w2

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

Some images from this lecture are taken from Google Image Search.

Hi!

- (I’m not Mijung!)

- Danica Sutherland – cs.ubc.ca/~dsuth – ICICS X563 – she/her 🏳️‍⚧️
  - “Danica” (North Am. English pronunciation, not authentic Slavic one) / “Professor Sutherland” / “Dr. Sutherland” are all fine

  - New-ish at UBC (since January 2021)
    - 3 PhD students, 3 research MSc, 2 course MSc
    - Representation learning, kernel methods, statistical testing, learning theory
    - Grad course on statistical learning theory

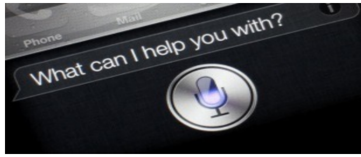
Where to find stuff

- Course website: cs.ubc.ca/~dsuth/440/
  - Schedule, slides, links to everything else
- Piazza
  - Discussion, announcements, etc
  - I mostly won’t send course-related emails
  - You’ll get faster, better responses here than if you email me
- Gradescope
  - Assignment handin (will have instructions)
- Canvas
  - Posting some files, etc
  - Should be able to access even if not registered yet – follow link from above

Next Topic: What is this course?

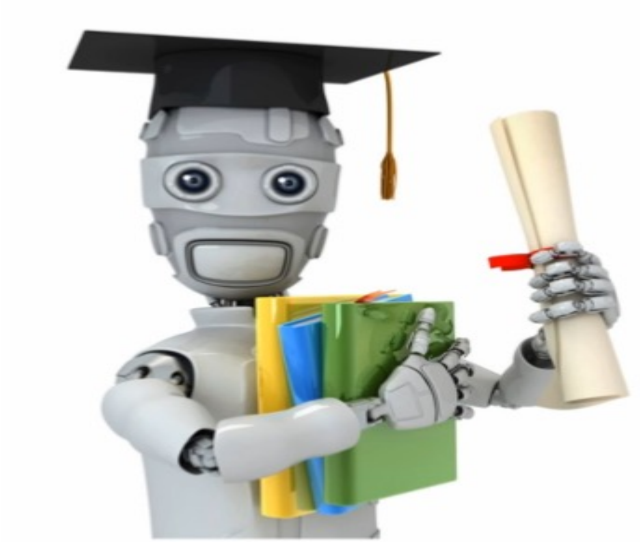
Big Data Phenomenon

- We are collecting and storing data at an unprecedented rate.
- Examples:
  - News articles and blog posts.
  - YouTube, Facebook, TikTok, the web at large….
  - Credit cards transactions and Amazon purchases.
  - 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.


Phone
Mail
What can I help you with?

Machine Learning

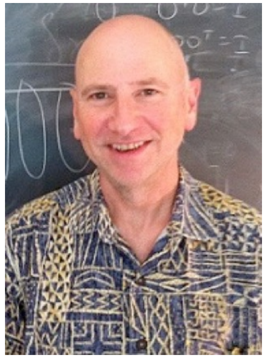
- What do you do with all this data?
  - Too much data to search through it manually.
- But there is valuable information in the data.
  - Can we use it for fun, profit, and/or the greater good?
- Machine learning: use computers to automatically detect patterns in data and make predictions or decisions.
- Most useful when:
  - Don’t have a human expert.
  - Humans can’t explain patterns.
  - Problem is too complicated.

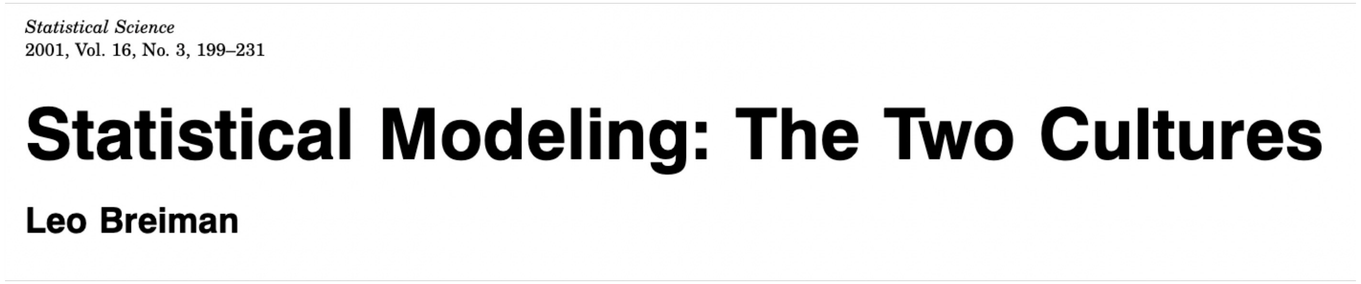


Machine Learning vs. Statistics

“If you’re analyzing data and proving theorems about it in [ESB], that’s statistics.
If you do it in [ICICS], that’s machine learning.”

– Larry Wasserman
(who said it with Baker and Gates, CMU’s equivalents)





Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

ML vs. Statistics

- ML places more emphasis on:
  1. Computation and large datasets.
  2. Predictions, rather than descriptions.
  3. Non-asymptotic performance.
  4. Models that work across domains.
- The field is growing very fast:
  – Influence of $$$ too:
    AI report estimates US $93.5 billion
    of private corp. investment in 2021

NUMBER of AI PUBLICATIONS by FIELD of STUDY (EXCLUDING OTHER AI), 2010–21

Source: Center for Security and Emerging Technology, 2021 | Chart: 2022 AI Index Report

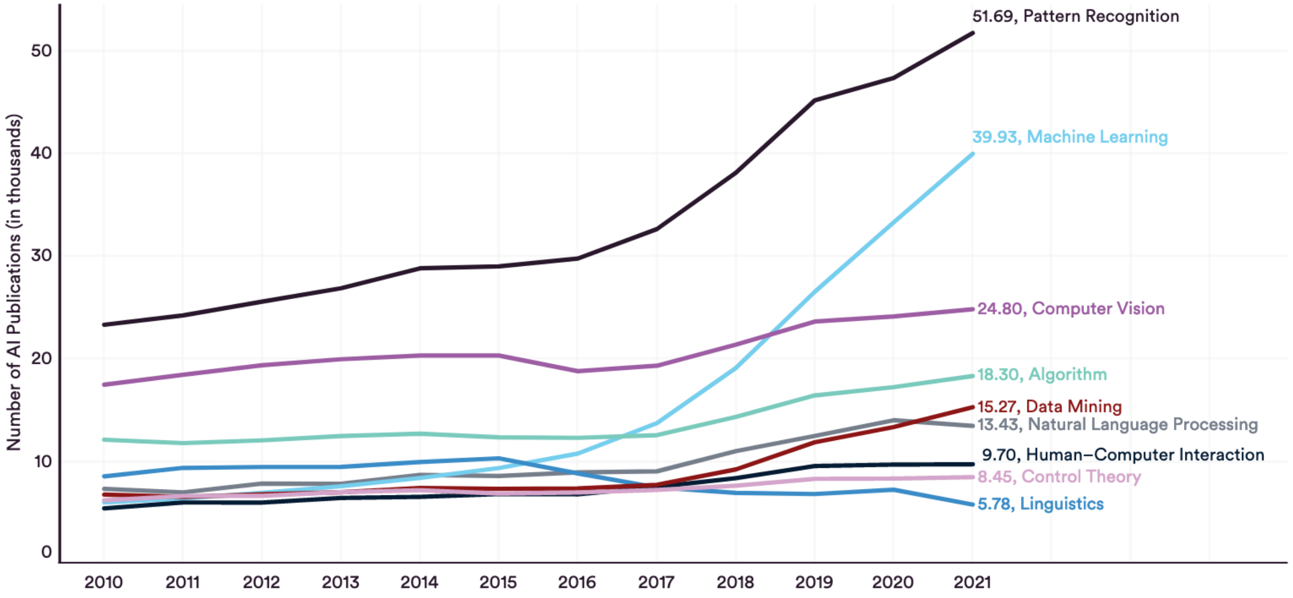


Figure 1.1.3

ATTENDANCE at LARGE AI CONFERENCES, 2010–21

Source: Conference Data, 2021 | Chart: 2022 AI Index Report

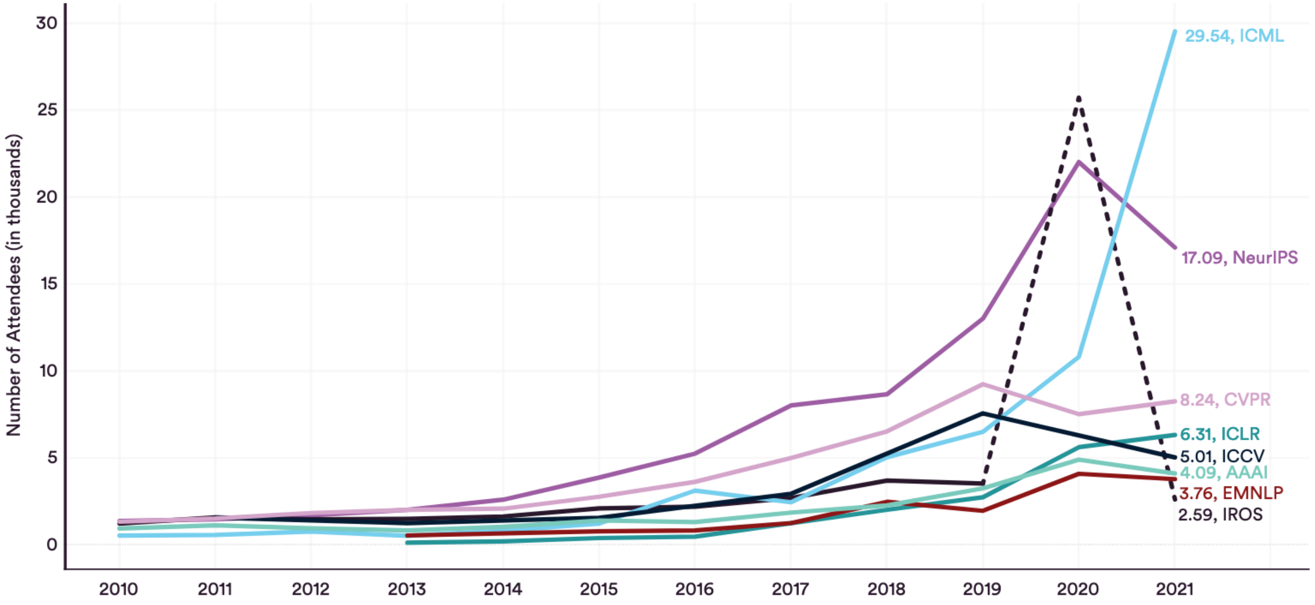
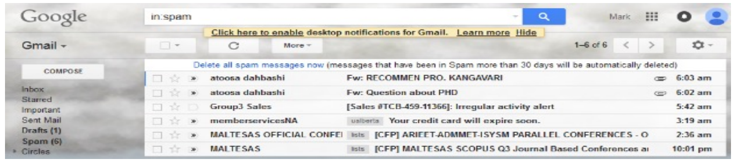


Figure 1.2.2

Applications

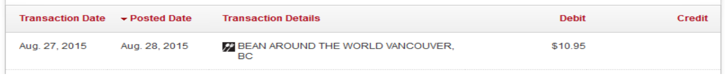
- Spam filtering.
- Credit card fraud detection.
- Product recommendation.
- Motion capture.
- Machine translation.
- Speech recognition.
- Face detection.
- Object detection.
- Sports analytics.
- Cancer subtype discovery.



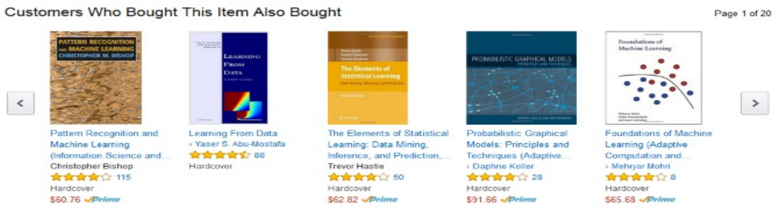
Click here to enable desktop notifications for Gmail. Learn more Hide

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 CONFE	lists [CFP] ARIEET-ADMMET-ISYSM PARALLEL CONFERENCES - O	2:36 am
MALTESAS	lists [CFP] MALTESAS SCOPUS Q3 Journal Based Conferences ai	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 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

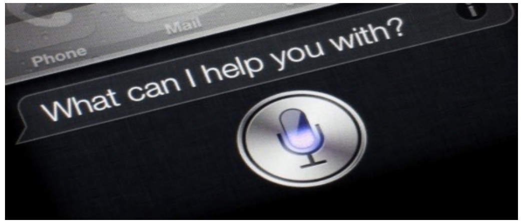
English – detected

French

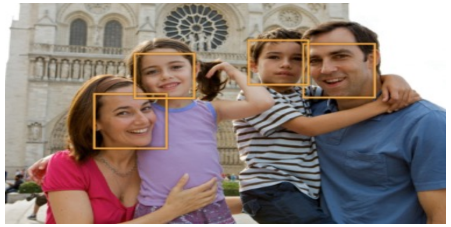
Data mining.

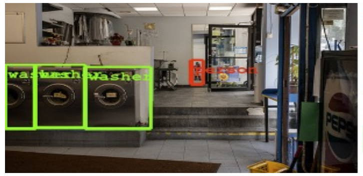
L'exploration de données .

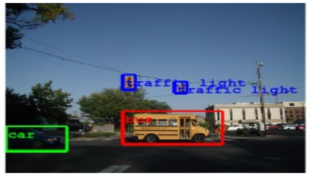
Open in Google Translate

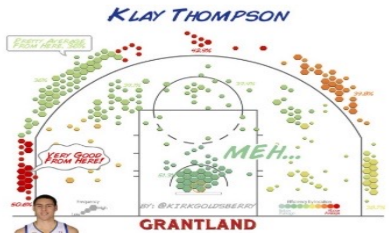


What can I help you with?



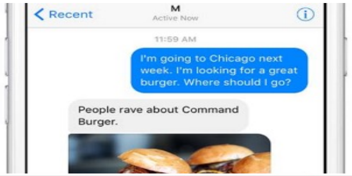






Applications

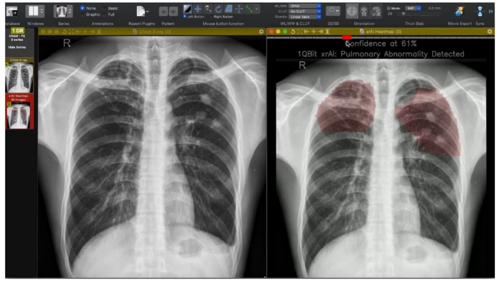
- Gene localization/functions/editing.
- Personal Assistants.
- Medical imaging.
- Self-driving cars.
- Scene completion.
- Image search and annotation.
- Artistic rendering.
- Physical simulations.
- Image colourization.
- Source separation
- Game-playing.

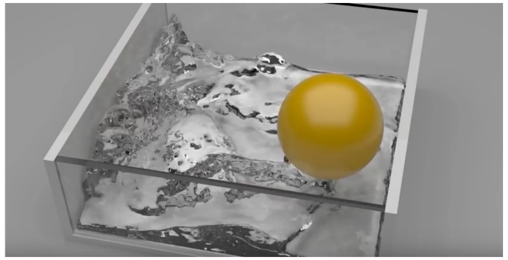


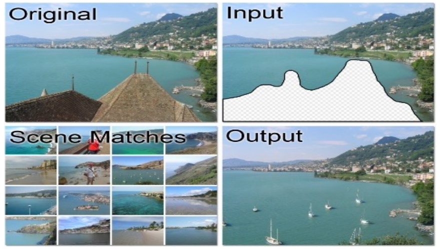
11:59 AM

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

People rave about Command Burger.







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

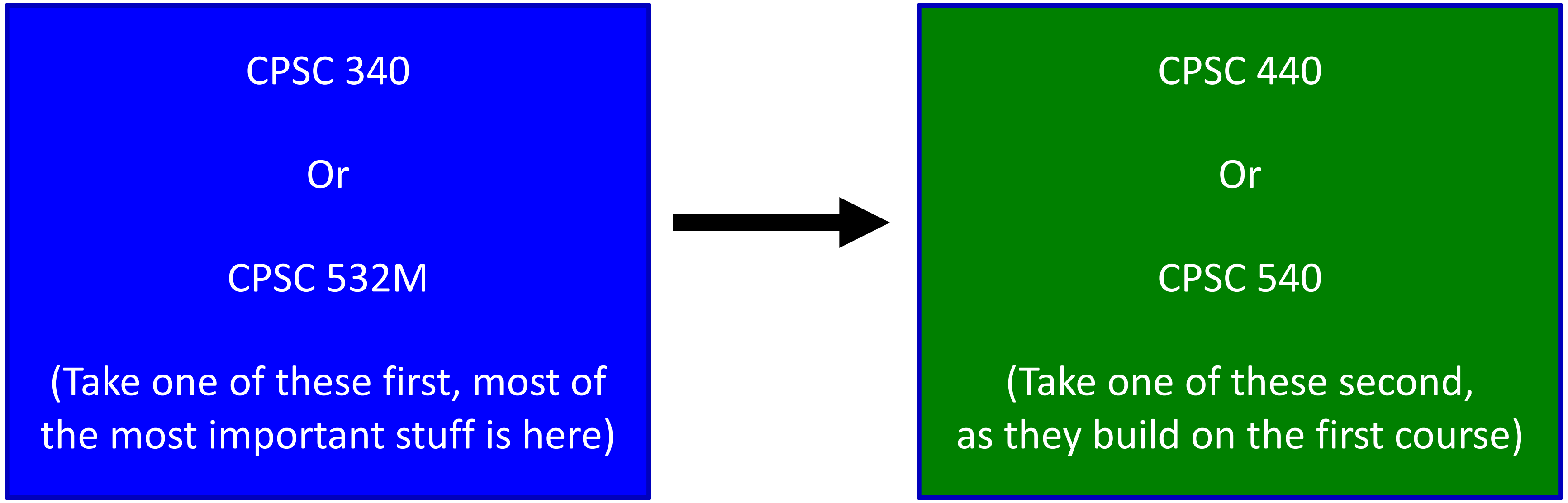
Youngsters, May 1912

A graphical representation of AlphaStar's processing. The system sees whole map from the top down and predicts what

Next Topic: Course Registration

CPSC 340 and CPSC 440 (and 532M and 540)

- There are two "core" ML classes: CPSC 340 and CPSC 440.
  - You can take CPSC 340 for grad credit as CPSC 532M (usually).
  - You can take CPSC 440 for grad credit as CPSC 540.
  - Structured as one full-year course: 440 starts where 340 ends.



CPSC 340

Or

CPSC 532M

(Take one of these first, most of the most important stuff is here)

CPSC 440

Or

CPSC 540

(Take one of these second, as they build on the first course)

  - We have also have CPSC 330 (focuses on using machine learning methods), and some other undergrad courses on the way

CPSC 340 and CPSC 440 (and 532M and 540)

- CPSC 340/532M (the other course):
  - Introductory course on data mining and ML.
  - Emphasis on applications and core ideas of ML.
    - And how to implement methods, not just how to use them.
  - Most useful techniques that you can apply to your research/work.
- CPSC 440/540 (this course):
  - Sequel course covering topics that do not appear in 340.
  - More "stuff that requires more time/background" than "advanced ML".
    - Assumes strong background on fundamental ML concepts.
    - Assumes stronger math/CS background.

You should take CPSC 340/532M first!

- If you can only take one class, take the other class (CPSC 340/532M).
  - 340/532M covers the most useful methods and ideas.
  - If you take 440/540 first, you'll miss half the story and a lot will seem random.
    - This is not an intro course and there is not a lot of review in 440/540.
      - So 440/540 is missing a lot important topics.
  - 440/540 is NOT an "advanced" version of 340/532M.
    - It just covers the methods that require more advanced math/CS background to use.

- It is much better to do CPSC 340/532M first:
  - Many people have taken 340/532M *after* 440/540 (not recommended).
  - A few people took 440/540 then 340/532M then *sat in on 440/540 again*, (REALLY not recommended).
  - In industry/research, BIG mistakes are usually related to the (340) fundamentals!
    - Not because of an "advanced" machine learning error.

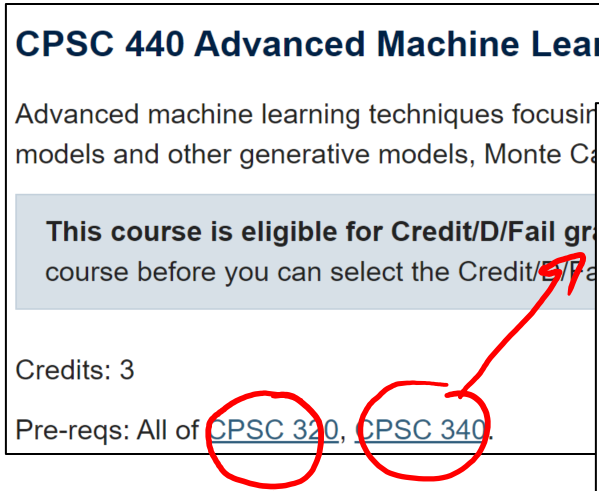
CPSC 340 and CPSC 440 (and 532M and 540)

- I'm assume you already know the below "typical" topics (covered in 340):
  - Calculus in matrix notation, including derivation of normal equations for least squares.
  - IID assumption, complexity vs. generalization trade-off, ensemble methods, and cross-validation.
  - Probabilistic classifiers, maximum likelihood, and MAP estimation.
  - Radial basis functions, how to show a function is convex.
  - Stochastic gradient descent, softmax loss, and L1-regularization.
  - PCA and collaborative filtering.

- **You will get lost very quickly if you don't know this material**.
  - You should already be able to write code implementing (almost) all of the above ideas.

- CPSC 440/540 Course Outline:
  - Density estimation, Bayesian methods, graphical models, mixtures and latent variables.

Prerequisites (A lot of Math and CS)



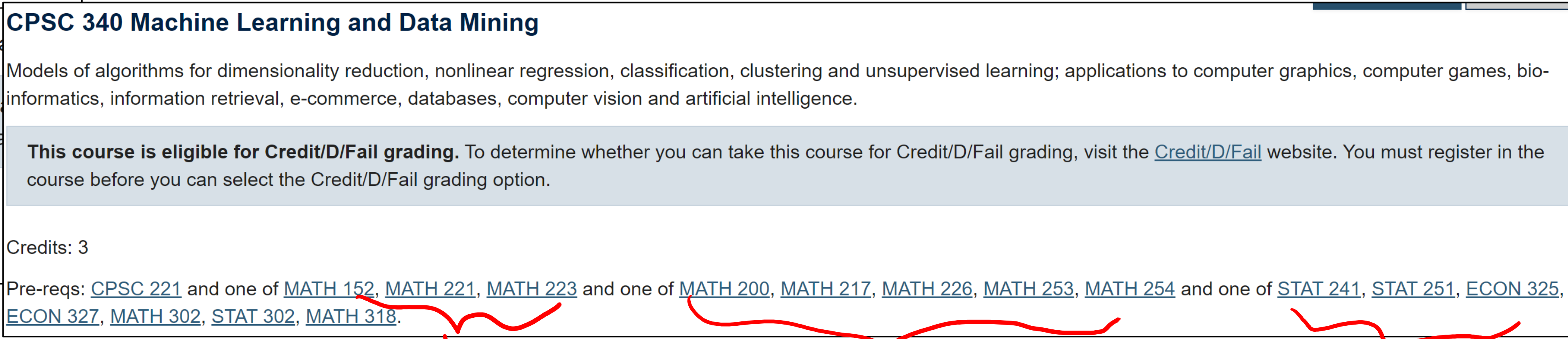
CPSC 440 Advanced Machine Learning

Advanced machine learning techniques focusin
models and other generative models, Monte Ca

This course is eligible for Credit/D/Fail gr
course before you can select the Credit/D/Fa

Credits: 3

Pre-reqs: All of CPSC 320, CPSC 340.



CPSC 340 Machine Learning and Data Mining

Models of algorithms for dimensionality reduction, nonlinear regression, classification, clustering and unsupervised learning; applications to computer graphics, computer games, bio-informatics, information retrieval, e-commerce, databases, computer vision and artificial intelligence.

This course is eligible for Credit/D/Fail grading. To determine whether you can take this course for Credit/D/Fail grading, visit the Credit/D/Fail website. You must register in the course before you can select the Credit/D/Fail grading option.

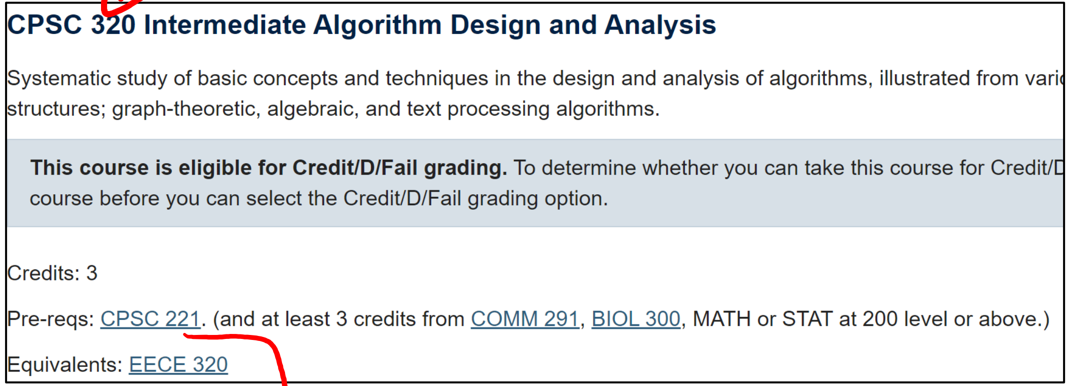
Credits: 3

Pre-reqs: CPSC 221 and one of MATH 152, MATH 221, MATH 223 and one of MATH 200, MATH 217, MATH 226, MATH 253, MATH 254 and one of STAT 241, STAT 251, ECON 325, ECON 327, MATH 302, STAT 302, MATH 318.

linear algebra

multivariate calculus

probability meets calculus



CPSC 320 Intermediate Algorithm Design and Analysis

Systematic study of basic concepts and techniques in the design and analysis of algorithms, illustrated from vari
structures; graph-theoretic, algebraic, and text processing algorithms.

This course is eligible for Credit/D/Fail grading. To determine whether you can take this course for Credit/D
course before you can select the Credit/D/Fail grading option.

Credits: 3

Pre-reqs: CPSC 221. (and at least 3 credits from COMM 291, BIOL 300, MATH or STAT at 200 level or above.)

Equivalents: EECE 320

Basic Algorithms and data Structures

Examples of CS concepts you should know:

- writing/debugging complex programs, binary trees, hash functions, graphs, big-O, randomized algorithms, dynamic programming, NP-completeness.

Examples of math concepts you should know:

- matrix algebra, norms, gradients, random variables, expectations, minimizing quadratic functions, random vectors.

Auditing

- Auditing is an excellent option:
  - Pass/fail on transcript rather than grade.
  - Do 1 assignment, or write a 2-page report on one technique from class, or attend > 90% of classes.
  - But please do this officially:
    - http://students.ubc.ca/enrolment/courses/academic-planning/audit
  - I strongly expect we'll have auditor space, but I'll sign forms next week

Next Topic: Lectures

Lectures

- Slides will be posted online (before lecture, and final marked-up 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 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.
    - If you know why something is important, and the core ideas, you can fill in details later.

Lecture recordings

- I’m planning to record lectures on Panopto (link on Canvas/Piazza)
- Please come to class anyway

- Intended to help you review things you didn’t get the first time, or as a last resort if you’re sick/etc
- But it’s *really* easy to fall behind remotely, and this course moves fast

Warning regarding teaching quality

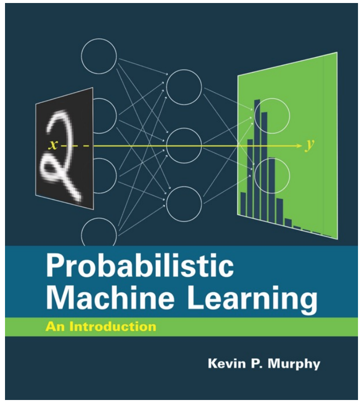
- This is only the third time that CPSC 440 has been offered.
  - The first time by me (or anyone who isn't Mark).
  - And I really didn't have as much time to prepare as I wanted.
    - It will still rely a lot on old CPSC 540 material, which is not always "undergrad friendly".
  - So the course isn't as "put together" as you might like.
    - Won't be as smooth as a course that has been offered many times.
  - I'm also not a teaching faculty member.
    - Relatively new faculty, and I run a larger-than-median lab.
    - I try but may not be as available/good as some the full-time teachers.

- Don't expect a high grade without a high effort.
  - We cover a lot of material and the assignments are LONG.
  - 340 is not an "easy" class; 440 is harder.

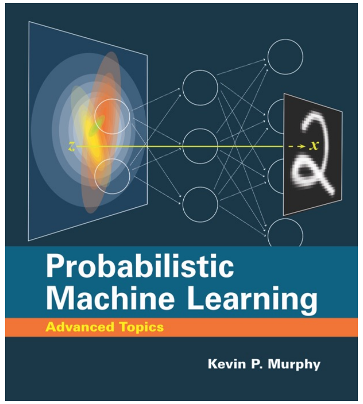
- If these things are going to bother you, it might be better to take this course later and/or take a different course this term.

Textbook and Other Optional Reading

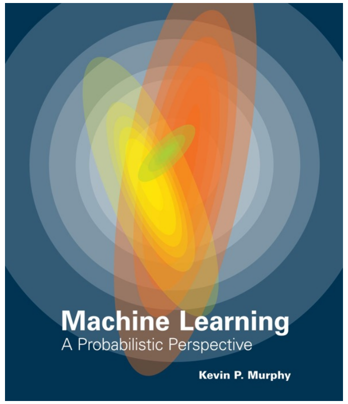
- No textbook covers all course topics.
- The closest is Kevin Murphy’s “Probabilistic Machine Learning”.
  - But we’re using a very different order.



+



or



- For each lecture:
  - I’ll give relevant sections from these books.
  - I’ll give other related online material.
- There is a list of related courses on the webpage.

bonus!

Textbook and Other Optional Reading

- Other good machine learning textbooks:
  - All of Statistics (Wasserman).
  - Elements of Statistical Learning (Hastie et al.).
  - Pattern Recognition and Machine Learning (Bishop).

- Good textbook covering needed mathematical background:
  - Mathematics for Machine Learning (Deisenroth, Faisal, Ong).
  - Available online.

- Good textbooks on specialized topics from this course:
  - Probabilistic Graphical Models (Koller and Friedman).
  - Deep Learning (Goodfellow et al.).
  - Bayesian Data Analysis (Gelman).

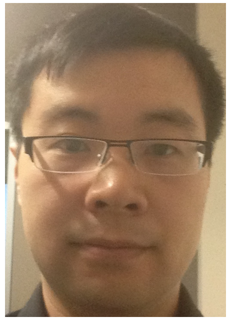
bonus!

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 you may find them interesting or useful in the future.
- I'll use a different colour of background on bonus slides.
  - I often include "post-lecture" bonus slides after the "Summary" slide.

Getting Help

- Use Piazza for course-related questions:
  - Link from Canvas and the course site.
  - Private posts asking about general information will be made public without asking.
- Weekly or almost-weekly Tutorials:
  - Run by TAs covering related material, mainly to help with assignments.
  - After class Monday/Wednesday and 4-5 on Friday; optional, starting next week.
- Instructor and TA office-hours:
  - Schedule TBA (starting end of this week).
- Teaching Assistants:
  - Wonho Bae
  - Ali Behrouz
  - Wu Lin
  - Alan Milligan
  - Justin Yang



Next Topic: Grades and Coursework

50%: Assignments

- 50% of your course grade will be based on 4 assignments:
  - Written answers, math, and Python programming.
  - No, you can't do the assignments in Julia, R, or Matlab.

  - First assignment due Friday of next week **at noon**.
    - NO LATE DAYS WILL BE CONSIDERED FOR THE FIRST ASSIGNMENT (even if you "didn't know you would be registered").
    - Subsequent assignments due every ~3-4 weeks, with more flexibility.
    - Gradescope handin; instructions soon.

- Start early, the assignments are a lot of work:
  - Previous students estimated that each assignments takes 6-25 hours:
    - This was heavily correlated with satisfying prerequistes.
    - Please look through the assignment in previous offerings to see length/difficulty.

- Doing assignments in groups:
  - Assignment 1 should be done on your own.
  - Assignments 2-4 can be done in groups of 1 to 2 (hand in one assignment for the group).

50%: Choose Your Own Adventure

- 50% of your grade will be based on a combination of the following:
  1. Final exam.
  2. Course project.
  3. Example lecture and assignment.

- You do not have to do all three of the above.
  - 440 students: 50% based on the maximum grade across these 3 options.
    - Have to do at least 1 out of the 3.
  - 540 students: 25% based on the maximum, 25% based on 2nd-best.
    - Have to do at least 2 out of the 3.

Final Exam

- Final exam details:
  - Scheduled by UBC, date currently unknown.
    - If you plan to take it, don’t make travel plans before April 27th.
  - Closed book; three pages of double-sided “cheat sheets”.

- There will be two types of questions:
  - “Technical” questions requiring things like pseudo-code or derivations.
    - On topics covered in assignments (similar to assignment questions).
  - “Conceptual” questions testing understanding of key concepts.
    - All lecture slide material except “bonus slides” is fair game here.

Course Project and Example Lecture/Assignment

- Course projects and example lecture/assignment:
  - More details coming later in the term.
  - Should be done in groups of 2-3.
  - Project scope will be smaller than projects in most classes.
  - You can use any programming language.
  - Will (probably) be due on the last day of exams.
  - You can use different groups for different projects/assignments.
    - Mixing and matching 440 and 540 students is also ok.

Late Assignment Policy

- You can submit one of Assignments 2-4 up to one week late.

- Example:
  - Assignment 3 is due on a Friday at midnight.
  - You can hand it in anytime before the following Friday.
    - You do not need to "declare" this, there is no penalty and no questions asked.
    - But then your group cannot hand in Assignment 4 late.

- For groups, can hand in late if at least one hasn't handed in late already.

- No other extensions except in *very* exceptional circumstances.
  - Academic concessions process (e.g. medical issues, death in family, etc).
  - **Not** "I had a lot of other assignments" or "there was a conference deadline."
  - You can submit more than once; do something preliminary by the deadline.
  - The assignments are out for a long window; don't wait to the last minute.

Cheating and Plagiarism

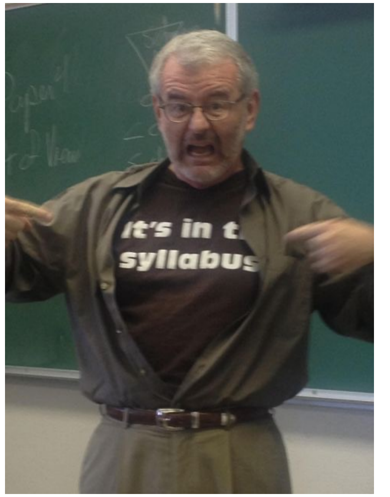
- Read about UBC’s policy on “academic misconduct” (cheating):
  - http://www.calendar.ubc.ca/Vancouver/index.cfm?tree=3,54,111,959

- When submitting assignments, acknowledge all sources:
  - Course material / textbooks are fine without mentioning.
  - Write “I had help from Lucy on this question.”
  - Write “I got this from another course’s answer key.”
  - Write “I copied this from the Coursera website.”
  - Write “ChatGPT wrote this answer for me.”
  - Otherwise, this is plagiarism.

- At Canadian schools, this is taken very seriously.
  - Could receive 0 in course, be expelled from UBC, or have degree revoked.

Summary

- Machine learning:
  - Automatically detecting patterns in data to help make predictions and/or decisions.
- CPSC 440:
  - Advanced/difficult 2nd or 3rd+ course on this topic.
  - Also called CPSC 540 for grad students.
  - "Sequel" class to CPSC 340 (not an "advanced" version of it).
- Course admin:
  - These slides are the syllabus!
  - Check here before asking course admin question so I don't do this:



- Next time: estimating COVID-19 prevalence.

- UBC provides resources to support student learning and to maintain healthy lifestyles but recognizes that sometimes crises arise and so there are additional resources to access including those for survivors of sexual violence. UBC values respect for the person and ideas of all members of the academic community. Harassment and discrimination are not tolerated nor is suppression of academic freedom. UBC provides appropriate accommodation for students with disabilities and for religious and cultural observances. UBC values academic honesty and students are expected to acknowledge the ideas generated by others and to uphold the highest academic standards in all of their actions. Details of the policies and how to access support are available here: https://senate.ubc.ca/policies-resources-support-student-success

bonus!

CPSC 340/532M and CPSC 440/540

- Quotes from people who probably should have taken CPSC 340:
  - "I did Coursera [or other online class] and have have done well in Kaggle competitions."
    - Neither of these cover calculus in matrix notation or MLE and MAP estimation.
  - "I've used SVMs, PCA, and L1-regularization in my work."
    - Sure, but do you know how to implement them from scratch?
  - "I've seen most of the 340 topics before."
    - Sure, but at what level of detail and do you know how to implement them from scratch?
  - "I want to apply machine learning in my research."
    - Great! Take 340 to learn how the most useful tools work and also **what can go wrong**.
  - "I took a machine learning course at my old school."
    - 340 is more broad/advanced than at most schools (talk to me if unsure).
  - "I've already learned about deep learning, so can I skip the basic stuff?"
    - When something goes wrong, you are going to want to understand the fundamentals.
  - " I took CPSC 540 with you. I wish I would have taken CPSC 340 first."
    - From a really-smart person who was working in a machine learning research job at the time.