CPSC 340:
Machine Learning and Data Mining

Alireza Shafaei

University of British Columbia, Summer 2020

http://cs340s20.shafaei.ca

Original version of slides by Mark Schmidt, modified by Mike Gelbart and Alireza Shafaei

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

Big Data Phenomenon

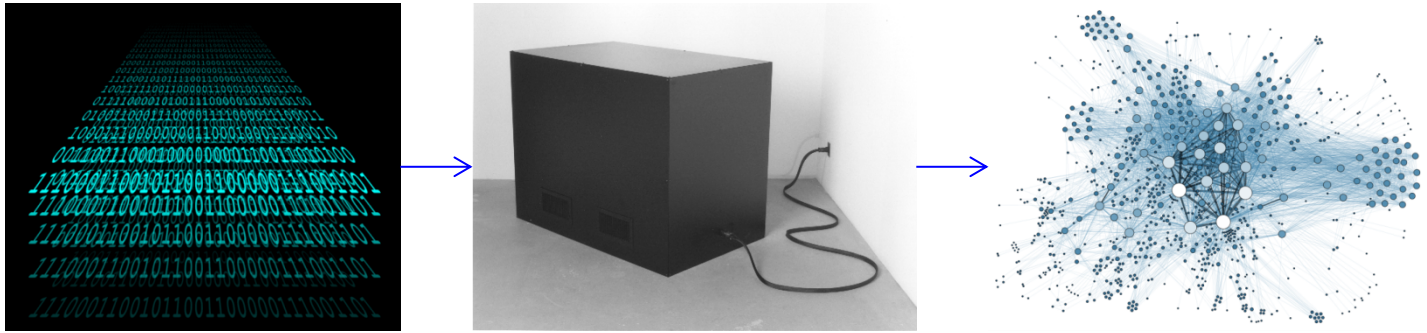
- We are collecting and storing data at an unprecedented rate.
- Examples:
  - YouTube, Facebook, MOOCs, news sites.
  - Credit cards transactions and Amazon purchases.
  - Transportation data (Google Maps, Waze, Uber)
  - Gene expression data and protein interaction assays.
  - Maps and satellite data.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

Big Data Phenomenon

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

Data Mining

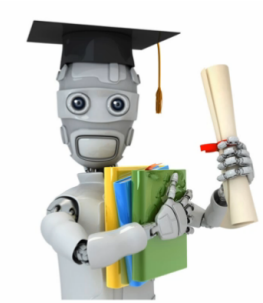
- Automatically extract useful knowledge from large datasets.

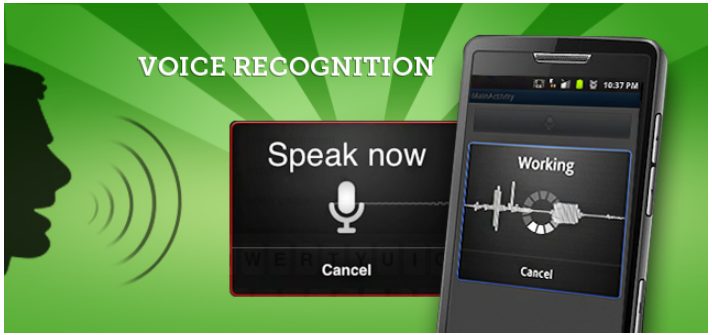


- Usually, to help with human decision making.

Machine Learning

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

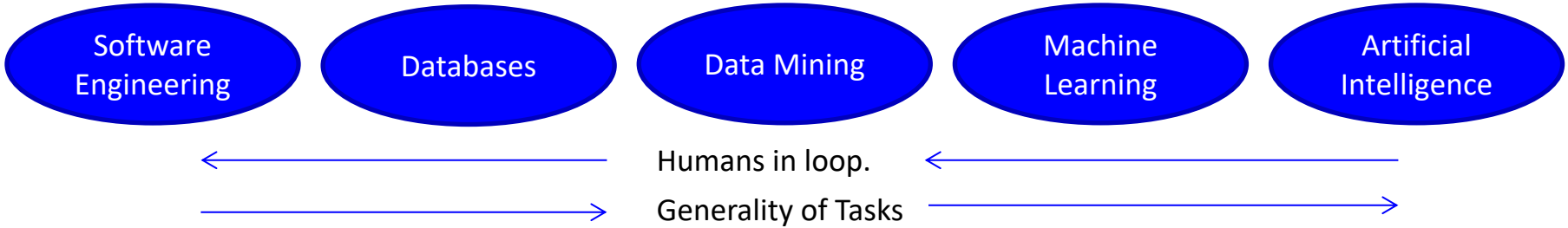




- Most useful when:
  - We want to automate something a human can do.
  - We want to do things a human can't do (look at 1 TB of data).

Data Mining vs. Machine Learning

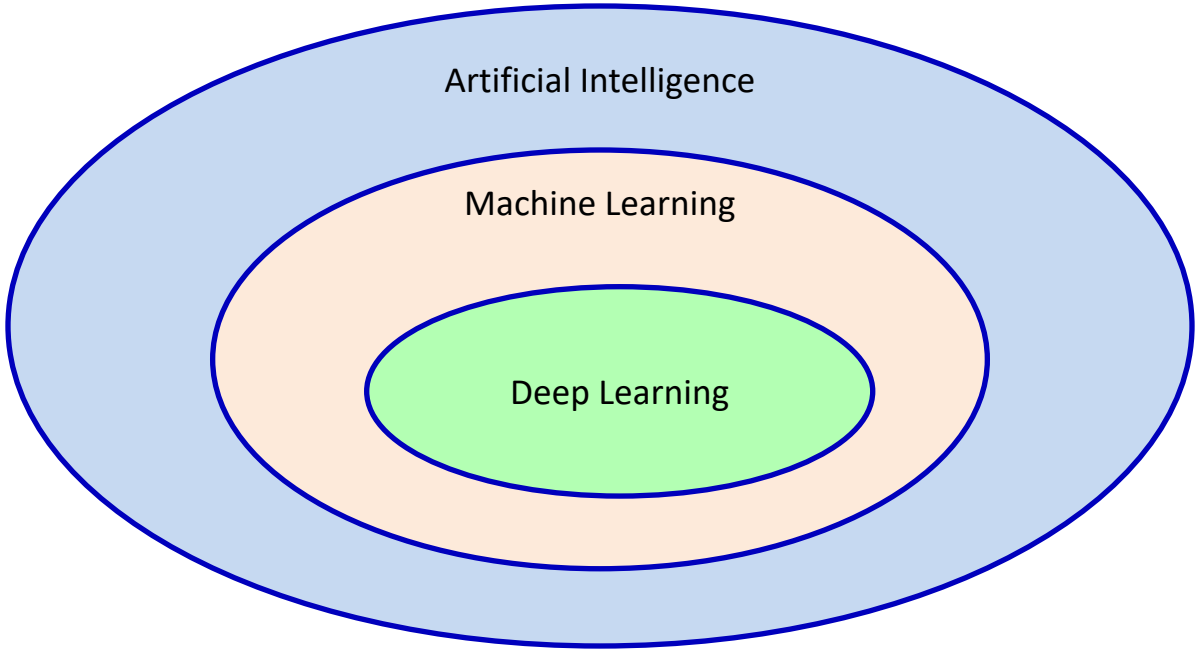
- Data mining and machine learning are very similar:
  - Data mining often viewed as closer to software engineering.
  - Machine learning often viewed as closer to AI.



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

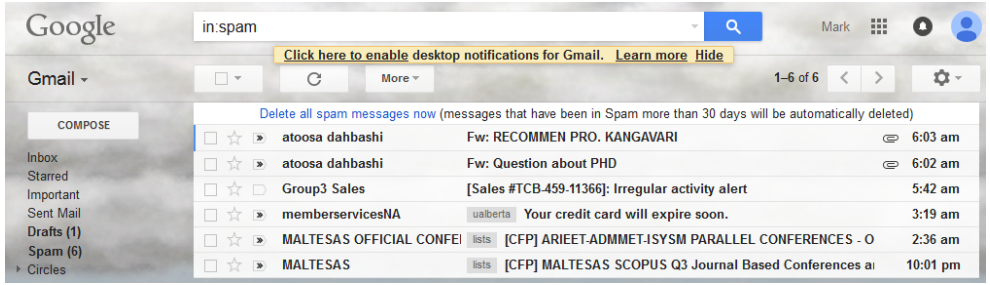
Deep Learning vs. Machine Learning vs. AI

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

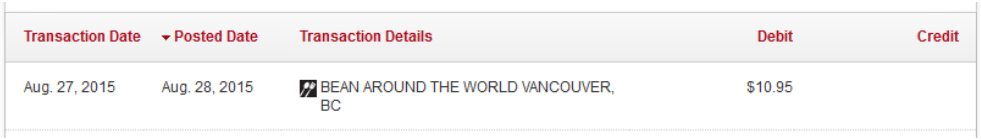


Applications

- Spam filtering:



- Credit card fraud detection:



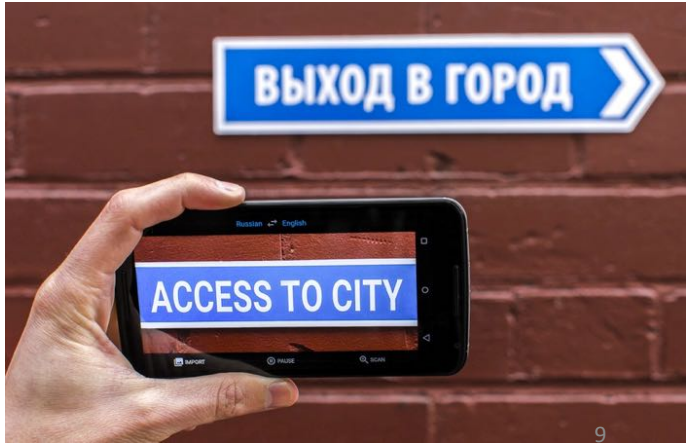
Transaction Date	Posted Date	Transaction Details	Debit	Credit
Aug. 27, 2015	Aug. 28, 2015	BEAN AROUND THE WORLD VANCOUVER, BC	$10.95	

- Product recommendation:

Applications

- Motion capture:

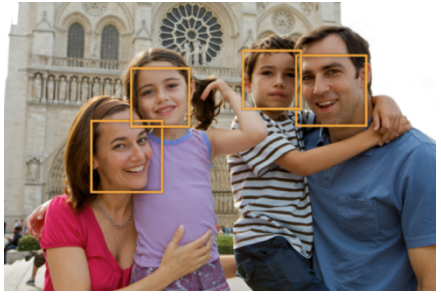
- Optical character recognition and machine translation:



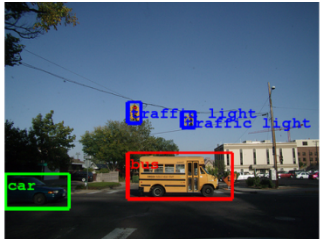
- Speech recognition:

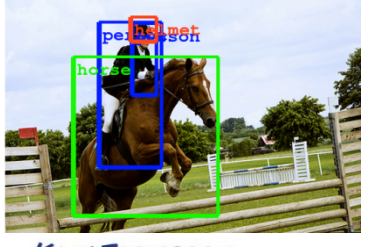
Applications

- Face detection:

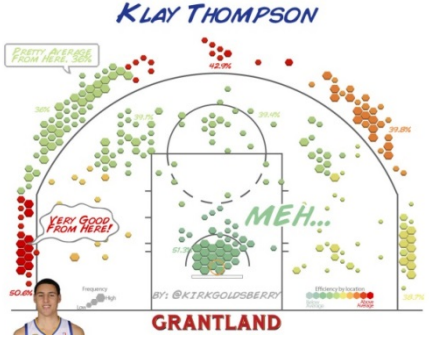


- Object detection:



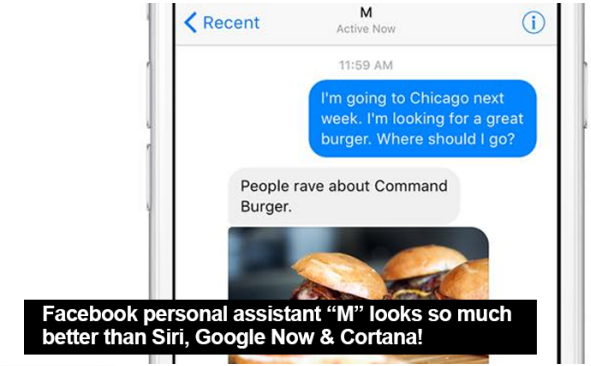


- Sports analytics:

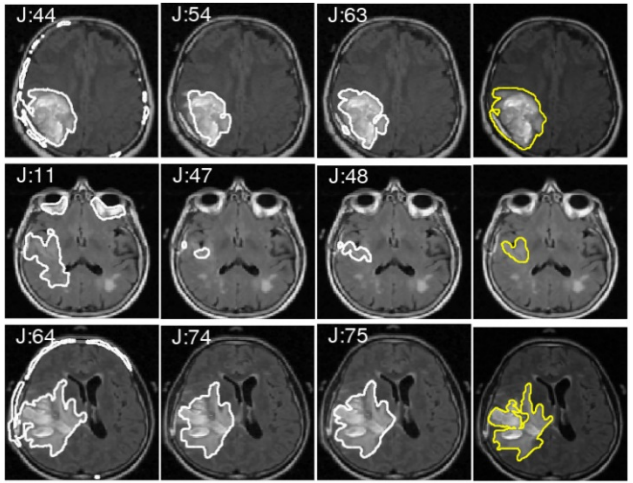


Applications

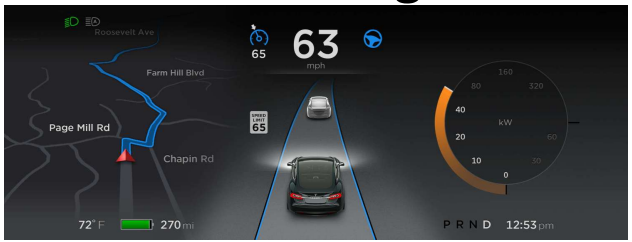
- Personal Assistants:



- Medical imaging:

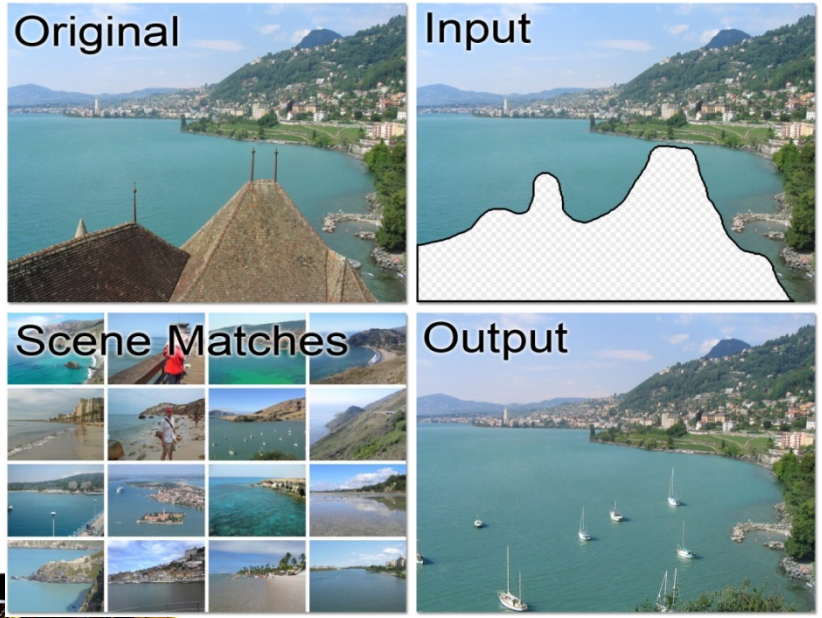


- Self-driving cars:



Applications

- Scene completion:



- Image annotation:

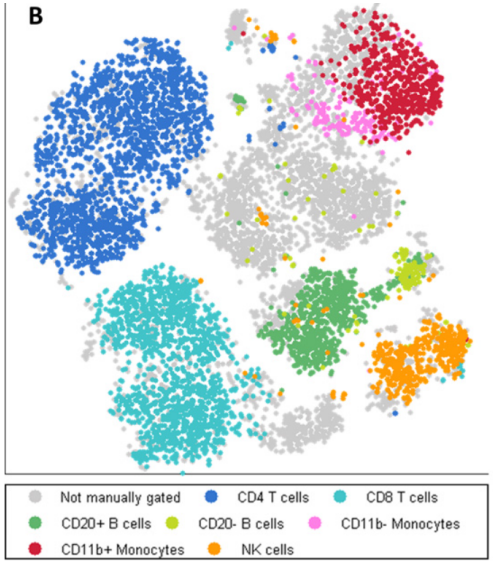
a cat is sitting on a toilet seat
logprob: -7.79

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

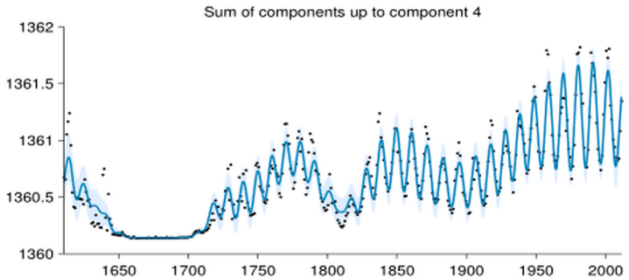
- Discovering new cancer subtypes:



- 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.

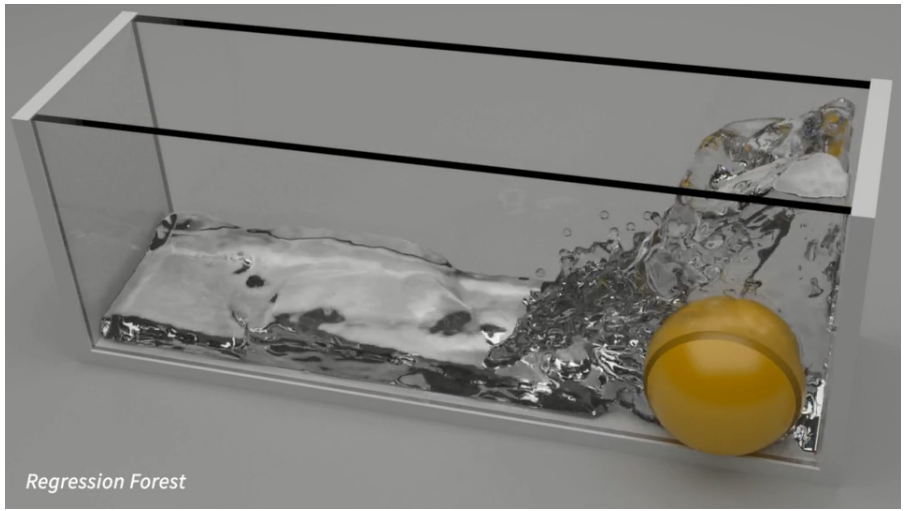


Applications

- Mimicking artistic styles:

Applications

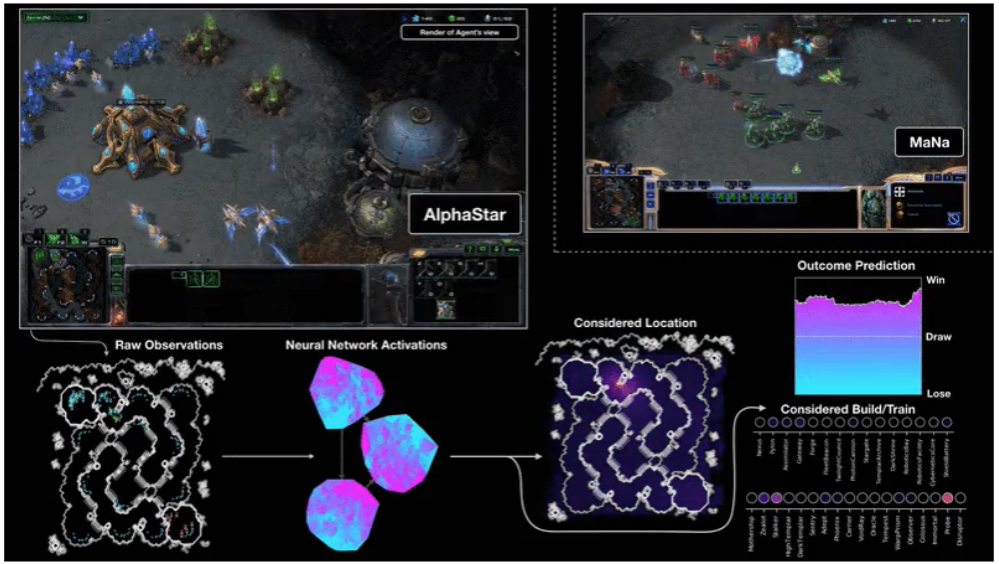
- Fast physics-based animation:



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

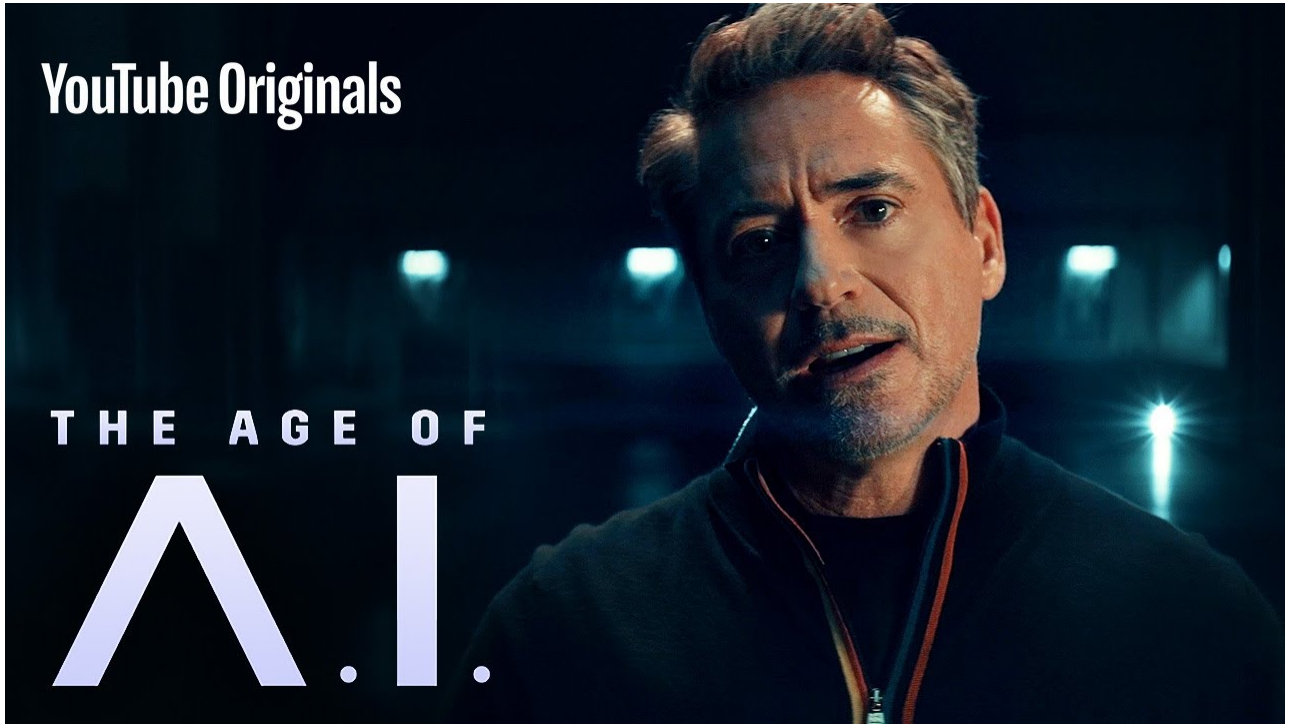
Applications

- Beating humans in Go and Starcraft:



Applications

- “Age of AI” YouTube series:



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

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

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

The Gap

- In this course you need to
  - Fill the knowledge gap.
    - Learn about the field and the standard practices.
    - Learn the existing techniques and understand how/why/when the ideas apply.
  - Fill the skill gap.
    - Internalize the standard practices and correctly execute the strategies.
    - Learn to implement and run experiments on real-world problems.
    - Develop the skill to pick the right approach to your problems.
- The lectures will cover the knowledge gap at a broad level. But the most important component will be the homework assignments.

(pause)

Credits

- Course designed by Mark Schmidt.
- With contributions from Mike Gelbart.

Alireza Shafaei

- Please call me *Alireza.*
- *5th year Ph.D. Student of Mark Schmidt and Jim Little.*
- *My research is on theory and application of deep neural networks.*
- First time teaching a full course @ UBC.
- I’m passionate about teaching an am thrilled to have this opportunity to walk you through the most fascinating field of science (in my opinion).

The Situation

- Teaching online
  - I can’t get direct feedback (to take in all the excitement on your faces).
  - There’s no best way to run the course.
  - The exam is going to be difficult to design/take.
- We will
  - Record the lectures and share them with you.
  - Make sure there are enough office hours and resources available to you.
  - Listen to your suggestions.

Platforms

- We will use Piazza for course-related questions and HW solutions.
  - You can sign up through Canvas.
- We will use Gradescope for HW submission and marking.
  - You can sign up through Canvas. (?)
- The main lectures will be delivered through Zoom.
  - The links will be shared in Canvas.
  - The lectures will be recorded.
- The tutorials/office hours will be through Collaborate Ultra.
  - Accessible through Canvas.

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.

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 2 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 2 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

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

- We will be using Python which is a free and fast high-level language.

- 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.

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."

CPSC 330 vs. CPSC 340

- There is also a less-advanced ML course, CPSC 330:
  - Taught by Mike Gelbart.
  - 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.

CPSC 340 Grading

- Grading will be slightly different from the previous offerings:

Assignments	Midterm	Final Exam
30	30	40

Essential Links

- Please bookmark the course webpage:
  - http://cs340s20.shafaei.ca/
  - Contains lecture slides, assignments, recordings, optional readings, additional notes.

- You should sign up for Piazza:
  - Accessible through Canvas, or
  - http://piazza.com/ubc.ca/summer2020/cpsc340.
  - Can be used to ask questions about lectures/assignments/exams.
  - **I do not watch piazza**. Come to office hours to ask questions directly of me. **TAs handle Piazza**.
  - Most questions should be “public” and not “private”,

- Use Piazza instead of e-mail for questions:
  - I can take a long time to respond e-mails.

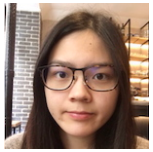
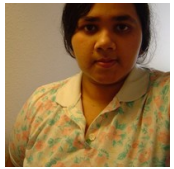
Textbooks

- 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 reserved in the ICICS reading room (if you can get in).
- List of related courses on the webpage, or you can use Google.

TA Cheat Sheet

- Ke (Mark) Ma 
- Lironne Kurzman 
- Ming Zhang 
- Ramya Rao Basava 
- Shahriar Shayesteh 
- Farnoosh Javadi
- Egor Peshkov 

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 is on webpage, and is due Friday.
  - The next five assignments will be due Sundays at midnight.
  - 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.
  - Using Gradescope you submit PDFs of your assignment (after editing it in latex to include code, figures etc). You can also specify your partner in Gradescope.
  - We expect you won't need to have the same partner for all assignments.

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 7 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 50% penalty:
  - Missing names or student IDs on assignments.
  - Corrupted submission.
  - 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.

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, check the calendar regularly.
  - Piazza (for general questions).
  - Weekly tutorials (optional).
    - Starting this week.
    - Will go through provided code, review background material, review big concepts, and/or do exercises.
  - Other students (connect to the other students).
  - The web (almost all topics are covered in many places).

Midterm and Final

- Midterm worth 30% and a (cumulative) final worth 40%
  - More info will be discussed later.

- Midterm is tentatively scheduled for 9:00 am June 1st in class.
  - Details coming later.

- I don't control when the final is.
  - 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.

Lectures

- All slides will be posted online (before lecture, and final version after).
- All lectures *will be recorded* and shared with you, check the homepage for details.
- 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 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.

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 me know (maybe we can fix it).
- **Do not distribute any course materials without permission**.
- Do not record lectures without permission.
- Check out https://keeplearning.ubc.ca for tips on online learning.

- 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.

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.

To Do

- Sign up on Piazza through Canvas.
- Sign up on Gradescope through Canvas (or let me know if you can’t).
- Bookmark the course webpage.
- Get started on the first homework assignment.

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)

Photo Mark 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