CPSC 340:
Machine Learning and Data Mining

Mike Gelbart
The University of British Columbia
2016-2017 Term 2

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

Big Data Phenomenon

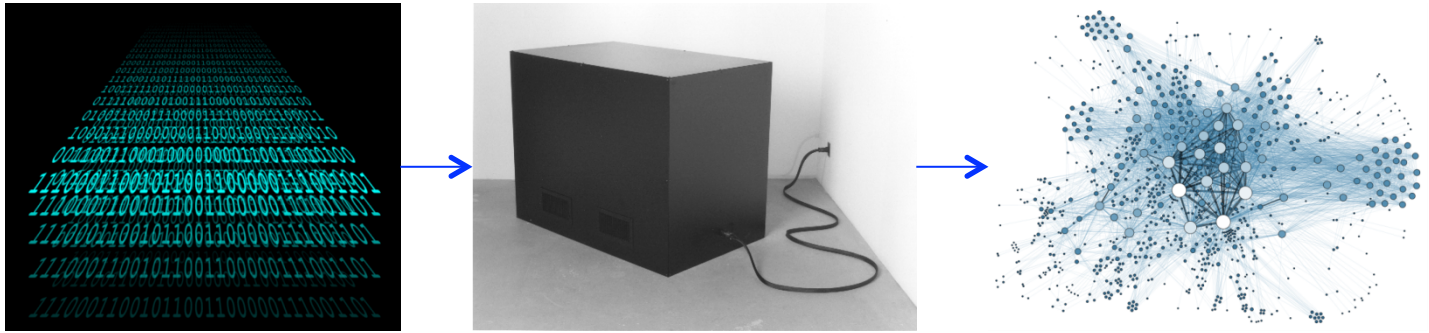
- We are collecting and storing data at an unprecedented rate.
- Examples:
  - YouTube, Facebook, MOOCs.
  - Credit cards transactions and Amazon purchases.
  - Transportation data (Google Maps, Waze, Uber)
  - Phone call records and speech recognition results.
  - Scientific experiments (biology, astronomy, …)
  - 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

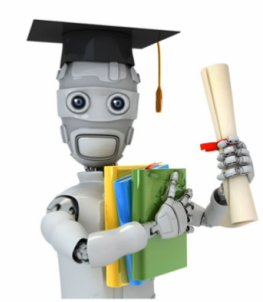
- Automatically extract useful knowledge from large datasets.

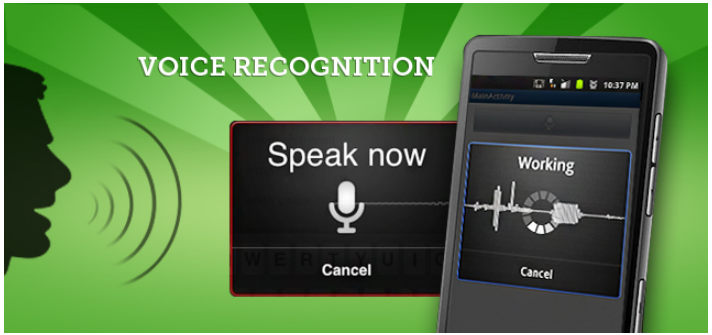


- Usually, to help with human decision making.

Machine Learning

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

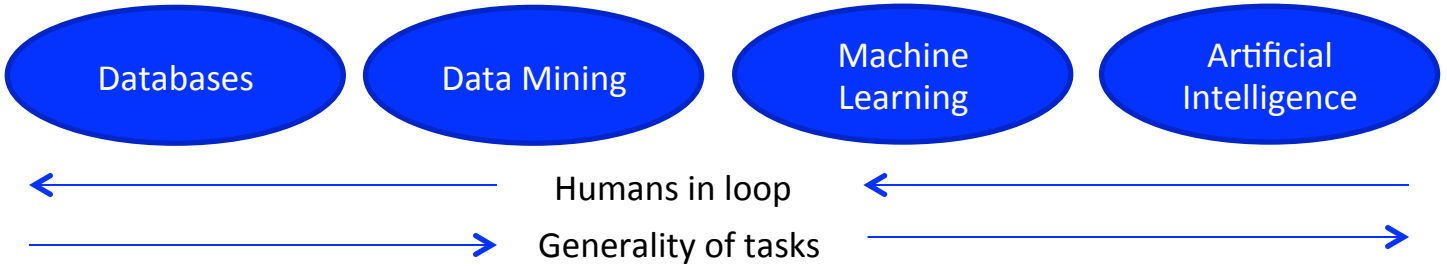




  - Sometimes, we want to automate something a human can do.
  - Sometimes, we want to do things a human can't do (like look at 1 TB of data).

Data Mining vs. Machine Learning

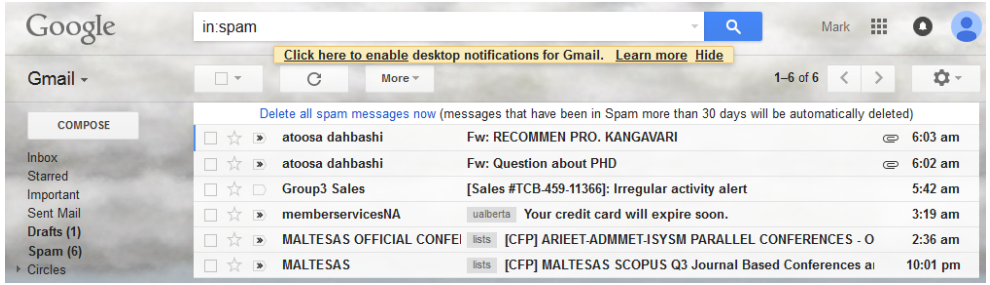
- DM and ML are very similar:
  - Data mining often viewed as closer to databases.
  - Machine learning often viewed as closer AI.



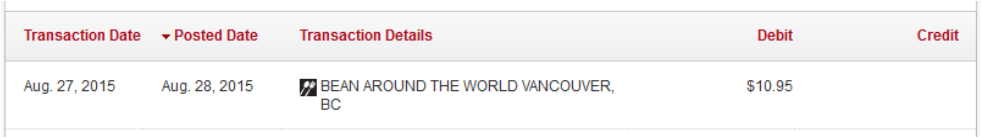
- Both are similar to statistics:
  - Less emphasis on ‘correct’ models and more focus on computation.

Applications

- Spam filtering:



- Credit card fraud detection:



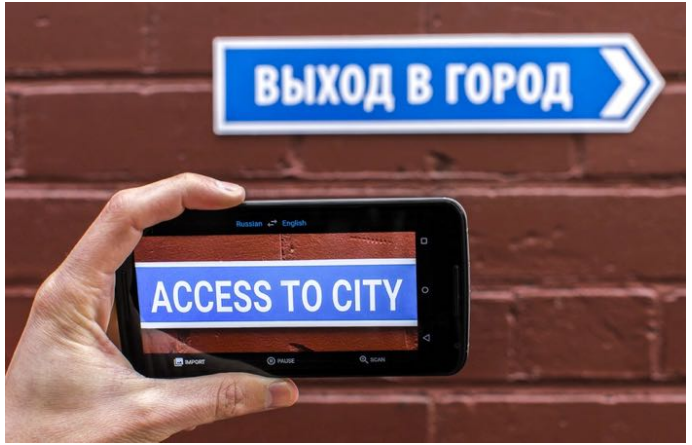
Transaction Date	Posted Date	Transaction Details	Debit	Credit
Aug. 27, 2015	Aug. 28, 2015	BEAN AROUND THE WORLD VANCOUVER, BC	$10.95	

- Product recommendation:

Applications

- Motion capture:

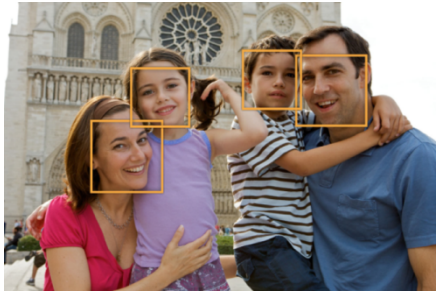
- Optical character recognition and machine translation:



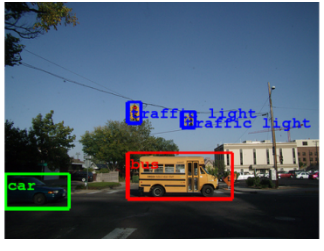
- Speech recognition:

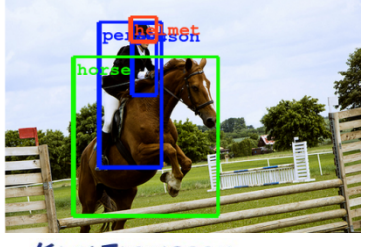
Applications

- Face detection:

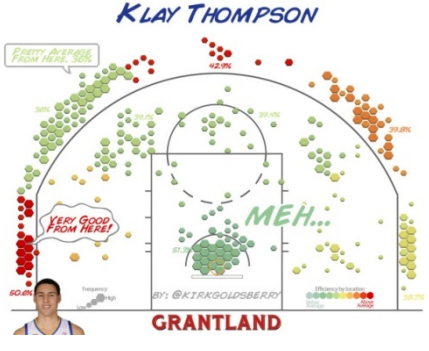


- Object detection:



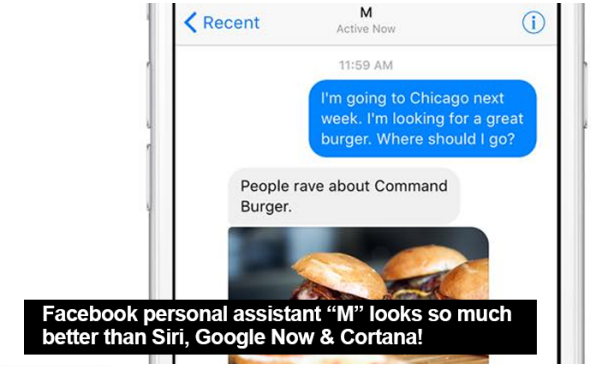


- Sports analytics:

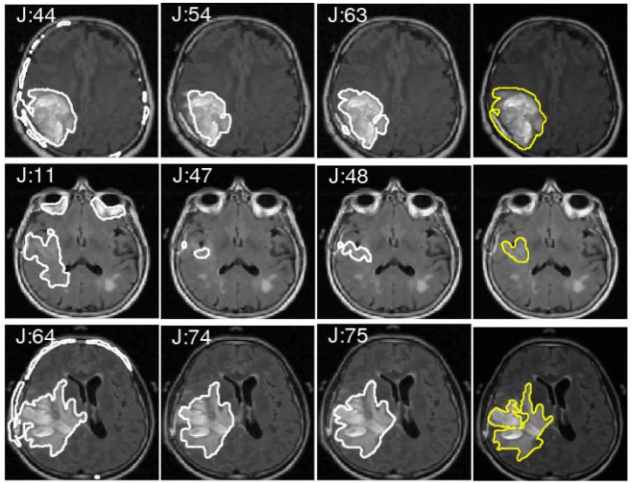


Applications

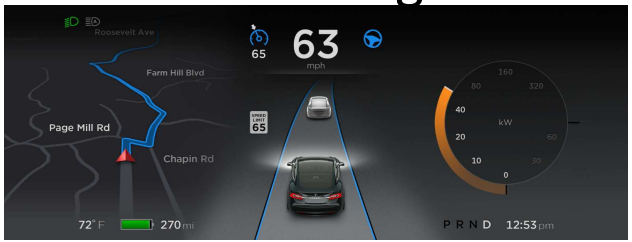
- Personal Assistants:



- Medical imaging:

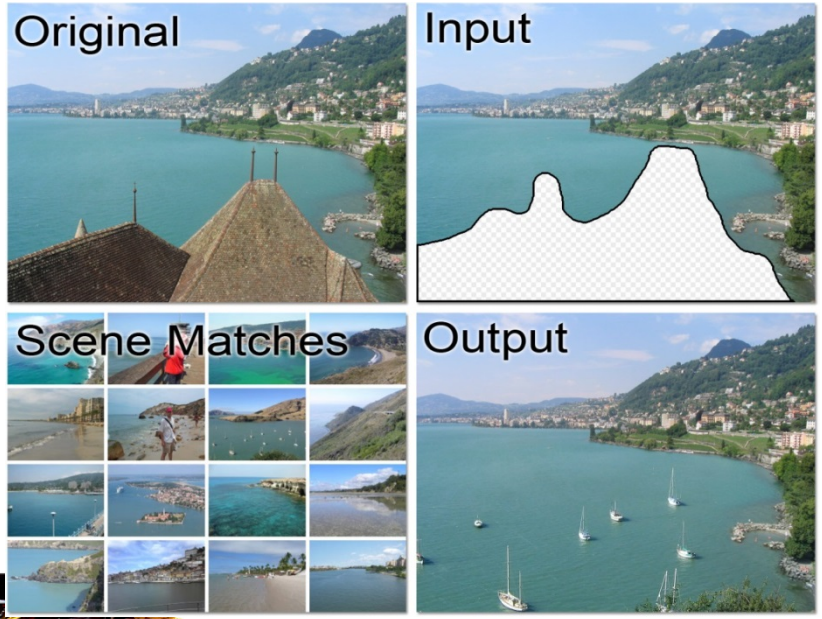


- Self-driving cars:



Applications

- Scene completion:



- Image annotation:

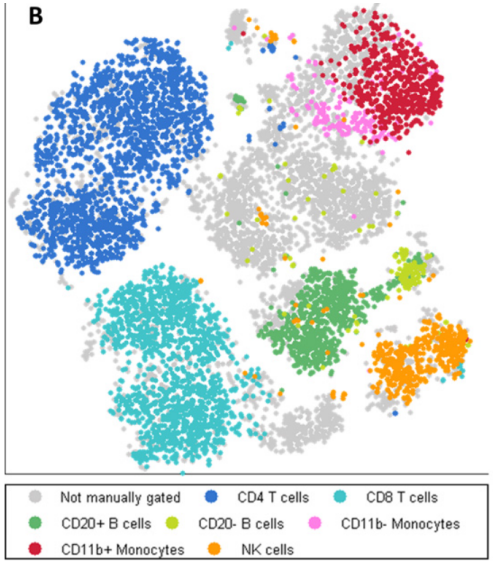
a cat is sitting on a toilet seat
logprob: -7.79

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

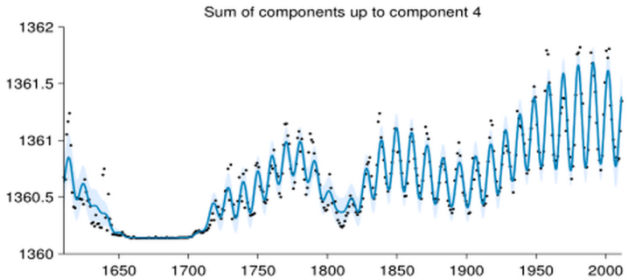
- Discovering new cancer subtypes:



- 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

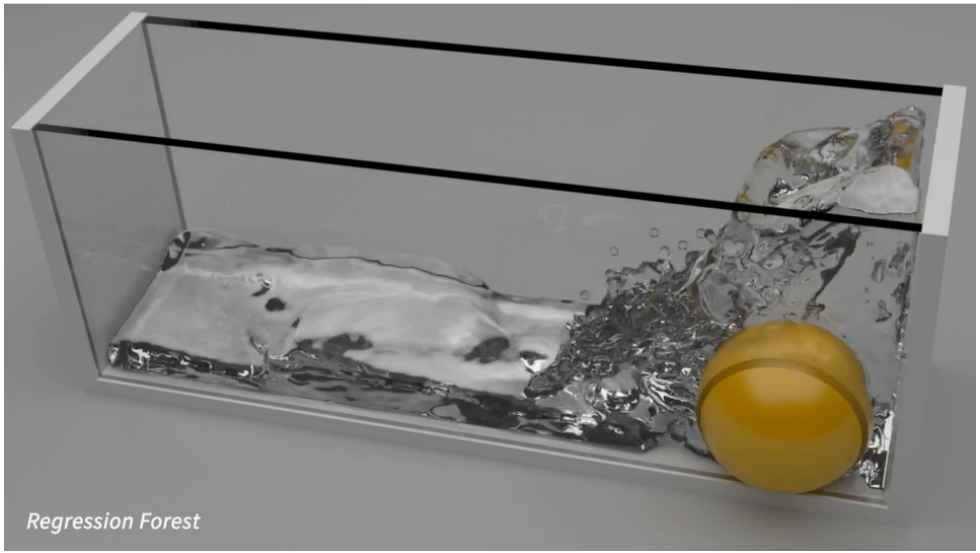
- Mimicking artistic styles and inceptionism:

Applications

- “Deep dream”:

Applications

- Fast physics-based animation:



- Mimicking art style in video.

Applications

- Beating human Go masters:

Summary

- There is a lot you can do with a bit of statistics and a lot data/ computation.
- It is important to know the limitations of what you are doing.
  - The future may not be like the past.
  - Associations do not imply causality.
- We are in exciting times.
  - Major recent progress in fields like speech recognition.
  - Things are changing a lot on the timescale of 3-5 years.
  - A bubble in ML investments?

About the course…

Outline

1) Intro to Machine Learning and Data Mining:
2) Course Administrivia
3) Course Overview

CPSC 340 vs. CPSC 540

- There is also a graduate ML course, CPSC 540:
  - More advanced material.
  - More focus on theory/implementation, less focus on applications.
  - More prerequisites and higher workload.
  - Offered in Term 2 (now)
- For almost all students, CPSC 340 is the right class to take:
  - CPSC 340 focuses on the most widely-used methods in practice.
  - CPSC 540 is intended as a continuation of CPSC 340.
  - You'll miss important topics if you skip CPSC 340.

Workload and difficulty

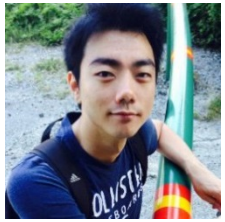
- For many people, this course is a LOT of work.
  - Some people spend tens of hours per assignment.
- Compared to typical CS classes, there is a lot more math:
  - Requires linear algebra, probability, and multivariate calculus.
  - Course is harder this year because of new calculus requirement.
- Compared to non-CS classes, there is a lot of programming:
  - This is not a class about running other people's software packages.
  - You are going to make/modify implementations of methods.

Resources

- Course homepage:
  - https://ubc-cs.github.io/cpsc340/
  - Please read all the info there
- Piazza for assignment/course questions:
  - https://piazza.com/ubc.ca/winterterm22016/cpsc340/home
- Optional weekly tutorials:
  - Start in second week of class (January 9).
  - In the first week you'll learn how to submit your homework.
  - You must be registered in a tutorial section to stay enrolled.

The Teaching Assistants (are outstanding)

- Tian Qi (Ricky) Chen

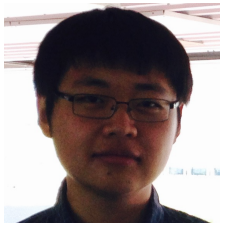


- Ritika Jain

- Issam Laradji

- Bita Nejat

- Wenyi Wang



Waiting List and Auditing

- The SSC currently lists this class as full
- 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!

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

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.).
  - Machine Learning: A Probabilistic Perspective (Murphy).
- List of related courses on the webpage, or you can use Google.

Course Website

- Please visit the course website: https://ubc-cs.github.io/cpsc340/
- The course website has links to other important places:
  - Piazza, for Q&A
  - UBC GitHub (github.ubc.ca), where you will submit all assignments
- The course website has info on:
  - Textbooks and other resources
  - Waitlist and registration
  - Auditing the course
  - Grading and exams
  - Much more!!!!

Python

- We will use Python for this course instead of Matlab.
- Please see the course website for:
  - Rationale for using Python
  - Info on different Python versions
  - Resources for learning Python
  - Installation instructions

UBC GitHub

- Some courses already run through github.com, e.g., CPSC 310
- We now have a GitHub Enterprise installation at github.ubc.ca
- Everything is on Canadian servers, which means we can store PII
  - You will receive your grades on github.ubc.ca
  - We don't need to use Connect!
- This is the 1st undergrad course at UBC to use this system
  - I have been piloting it with the Master of Data Science program since September
  - We're hoping to roll this out to more courses soon

Benefits of using GitHub

- No paper, lost or no-name assignments, missing staples, …
- You can work collaboratively with your partner from anywhere.
- Your TAs can mark collaboratively and from anywhere.
- Your TAs can see your work-in-progress anywhere/anytime.
- You will gain experience using git/GitHub, which are widely used in industry.

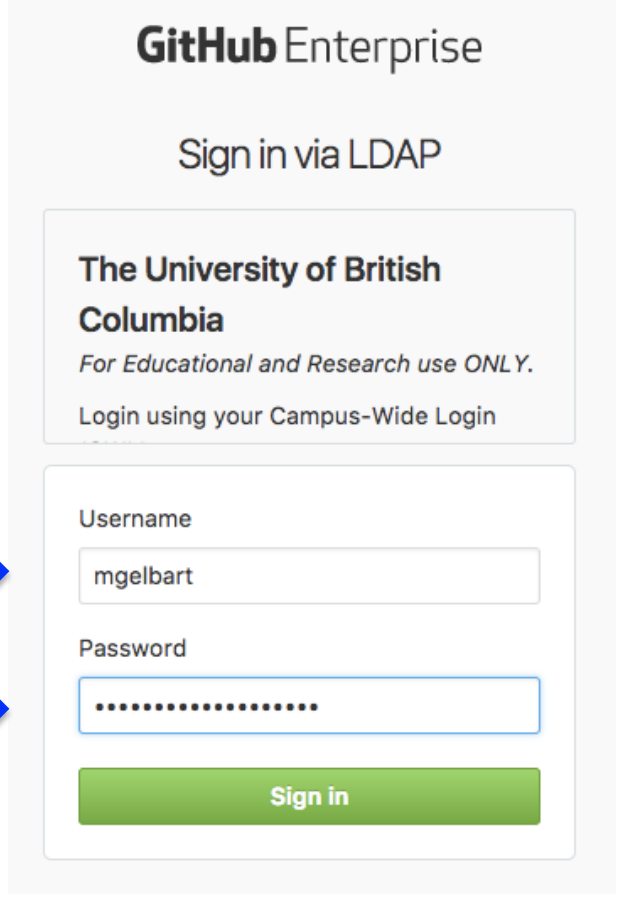
On your laptop/phone…

- Please go to github.ubc.ca and sign in with your CWL credentials.
- If you are enrolled in the course OR registered on the waitlist you should be able to log in successfully.
- If you registered but not able to log in, send me an email.

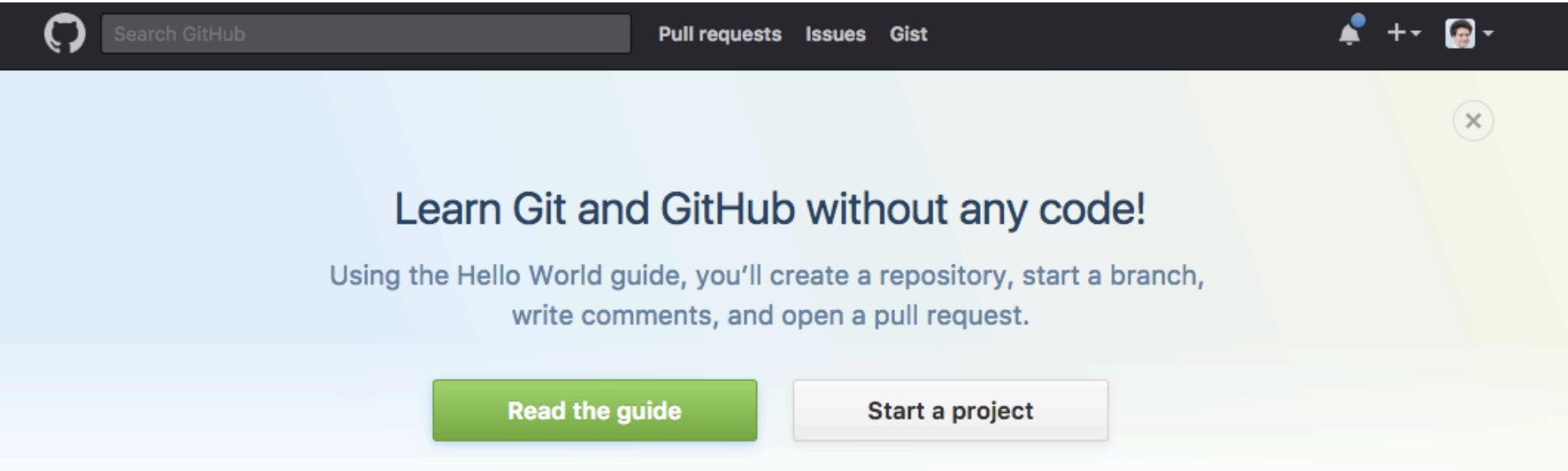
Step 1: Log in at github.ubc.ca

Your CWL username
(NOT student number)

Your CWL password



Step 2: You should see something like this



Step 3: You'll get you access soon

- Once you login you will have access to the cpsc340 organization
- This lives at github.ubc.ca/cpsc340
- From here you can access your homework repositories and other internal documents

Working in partners

- For Assignment 0 (due Jan 11) everyone must work individually
  - BTW, Assignment 0 is already posted; see the course timetable at https://ubc-cs.github.io/cpsc340/timetable/
- For all future assignments, you may work with a partner
- For this to happen, you must indicate your partnership **before** the assignment is **released**. Instructions are at the internal homepage, https://github.ubc.ca/cpsc340/home

Vote: do you want to see each other's work?

- "YES" means that, after solutions are posted, you will be able to see the submitted work of all your classmates, and they will be able to see yours. Benefit: learn from each other.

- "NO" means that only your partner, and no other classmates, will be able to see your work at any time. Benefit: privacy.

- Either way, the instructor and TAs can see your work at all times.

Assignments

- 6 Assignments worth 25% of final grade:
  - Written portion and Python programming.
  - Submitted via github.ubc.ca
  - You can have up to 3 total "late days":
    - For example, if assignment is due on Monday at 9am:
      - Handing it in before Monday 9am uses 0 late days.
      - Handing it in before Tuesday 9am uses 1 late days.
      - Handing it in Tuesday 9:05am uses 2 late days.
  - You do not need to notify me or anyone to use a late day.
  - You will get a mark of 0 on an assignment if you run out of late days.
- You can work alone or in partners (groups of 2).
  - You don't need to have the same partner for every assignment.
- Acknowledge all sources, including webpages and other students.

Midterm and Final

- Midterm details:
  - 30% of final grade
  - Closed book, two-page double-sided ‘cheat sheet’.
- No ‘tricks’ or ‘surprises’:
  - Given a list of things you need to know how to do.
  - Mostly minor variants on assignment questions.
- If you miss the exam, see me with doctor’s note or relevant documentation.
- Final will follow same format:
  - 45% of final grade.
  - Cumulative.

Lecture Style and Lecture Slides

- The course we will cover a lot of topics:
  - Some topics will not be covered in much depth.
    - But we'll go into depth on a few key recurring issues.
    - To keep things sane, I'll give you a list of topics to know for the midterm/final.
  - It can be better to know many methods than learning only a few in detail:
    - I'll explain why when we discuss the "best" machine learning algorithm.
    - Some class time will be devoted to important ideas that you won't be tested on.
- All class material will be available online or on Piazza.
  - I'll try to post topics/readings before each class.
  - After class, I'll post annotated/updated slides.
  - Do not record without permission.
- In early October, we'll do an unofficial instructor evaluation:
  - Will let me adapt lecture/assignment/tutorial style.

Code of Conduct

- Do not post offensive or disrespectful content on Piazza or GitHub.
- If you have a problem or complaint, let me know immediately. Maybe we can fix it!
- Do not distribute any course materials without permission.
- Do not record lectures (audio or video) without permission.
- If you commit to working with a partner, do your part of the work.
- Think about how/when to ask for help (in person, email, Piazza, etc.)
  - Don't ask for help after being stuck for only 10 seconds. Make a reasonable effort to solve your problem.
  - Read all the instructions before asking for help. Don't ask questions whose answers are on the course website or homework instructions or on the first page of a Google search. These questions take time away from other instructional activities.
  - On the other hand, don't ask for help only after 10 hours of painful debugging. Don't be shy!

Lectures

- All materials posted online.
- Please ask questions. If your question is not relevant (or I don't know the answer), I'll deflect it for later (so don't hesitate to ask).
- If you raise your hand and I don't call on you by name, please say your name before asking your question. This will help me learn your names.
- We will try out an online system for pacing the course. Please go to http://128.189.230.5:6169/test/. Participation is anonymous.
  - You must be connected to the "ubcsecure" wifi for this to work.

About Me

- Please call me Mike. I do not believe academics (or physicians) are particularly more deserving of a special title than anyone else.
- I'm originally from Vancouver but studied in the U.S. for 8 years.
- My PhD thesis was on automatically tuning ML algorithms.
- I co-designed and teach in UBC's Master of Data Science program.
- Outside of UBC I advise at Vanedge Capital and some ML startups.
- More about me on my website: http://www.cs.ubc.ca/~mgelbart/
- I really enjoy teaching and am glad to be on this journey with you for the next 14 weeks!

Outline

1) Intro to Machine Learning and Data Mining:
2) Course Administrivia
3) Course Overview

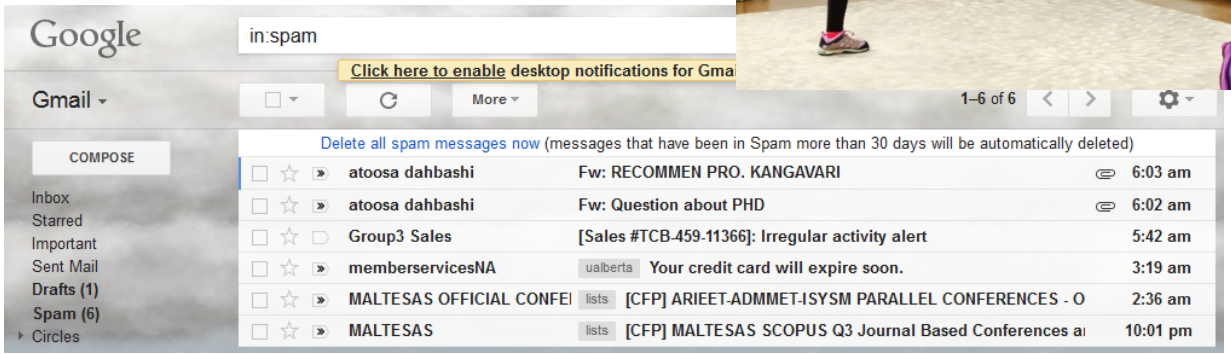
- All the remaining slides are “extra”.
- We may go through them briefly, if time permits.

Course Outline

- Next class discusses data exploration, cleaning, and preprocessing.
- After that, the remaining lectures focus on the six topics:
  1) Supervised Learning.
  2) Unsupervised learning.
  3) Linear prediction.
  4) Latent-factor models.
  5) Deep learning.
  6) Density estimation.

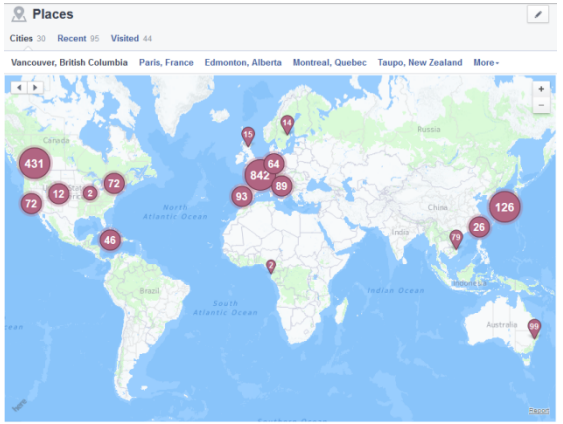
Supervised Learning

- Classification:
  - Given an object, assign it to predefined ‘classes’.
- Examples:
  - Spam filtering.
  - Body part recognition.



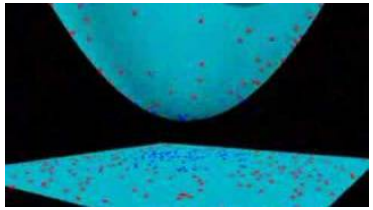
Unsupervised Learning

- Clustering:
  - Find groups of `similar' items in data.
- Examples:
  - Are there subtypes of tumors?
  - Are there high-crime hotspots?
- Outlier detection:
  - Finding data that doesn't belong.
- Association rules:
  - Finding items frequently 'bought together'.

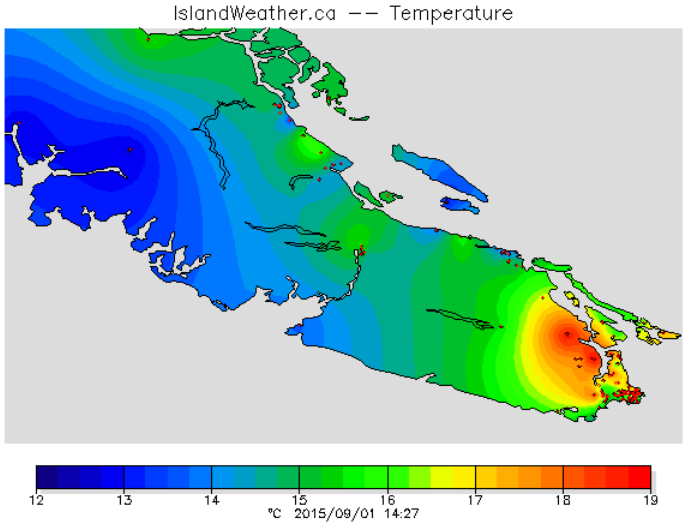


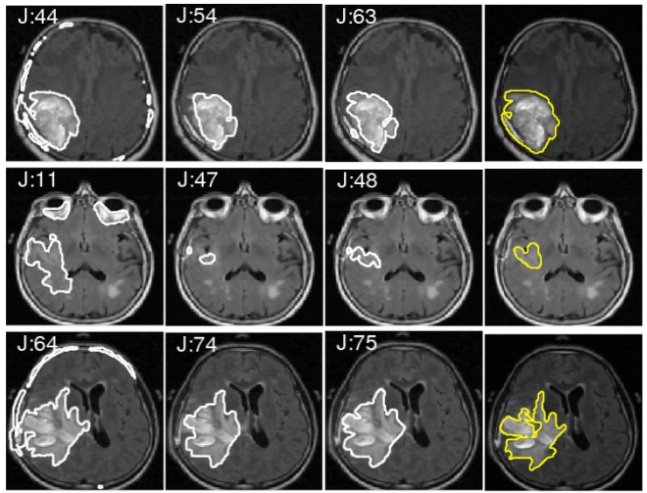
Transaction Date	Posted Date	Transaction Details	Debit	Credit
Aug. 27, 2015	Aug. 28, 2015	BEAN AROUND THE WORLD VANCOUVER, BC	$10.95	

Linear Prediction

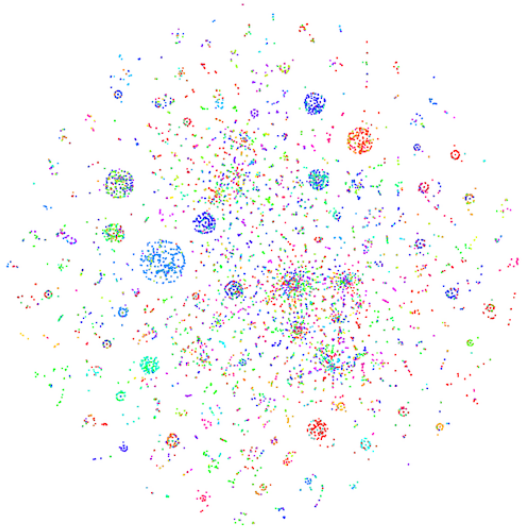


- Regression:
  - Predicting continuous-valued outputs.
- Working with very high-dimensional data.

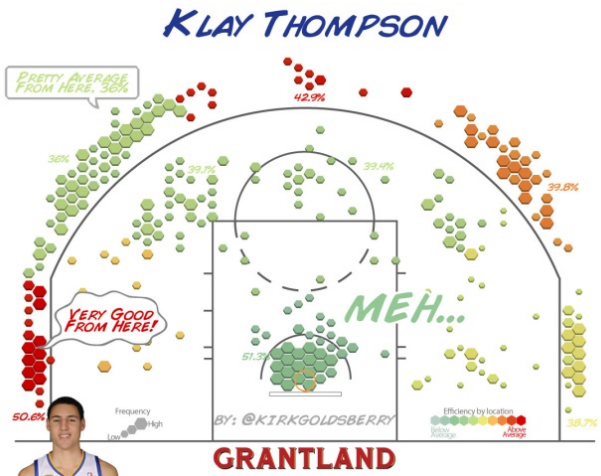




Latent-Factor Models

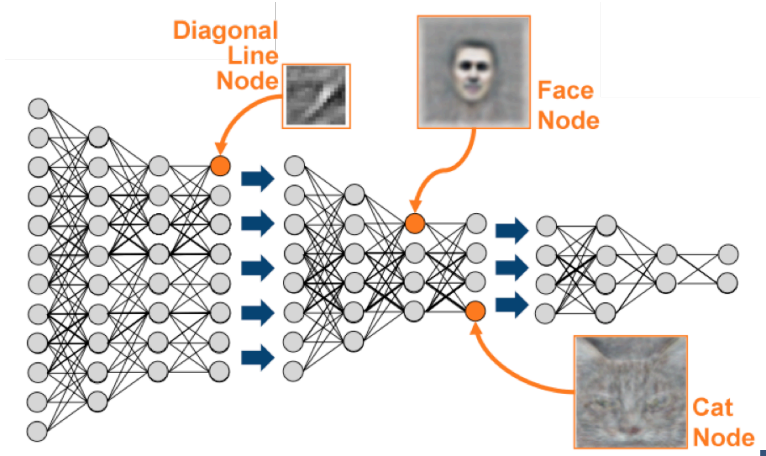


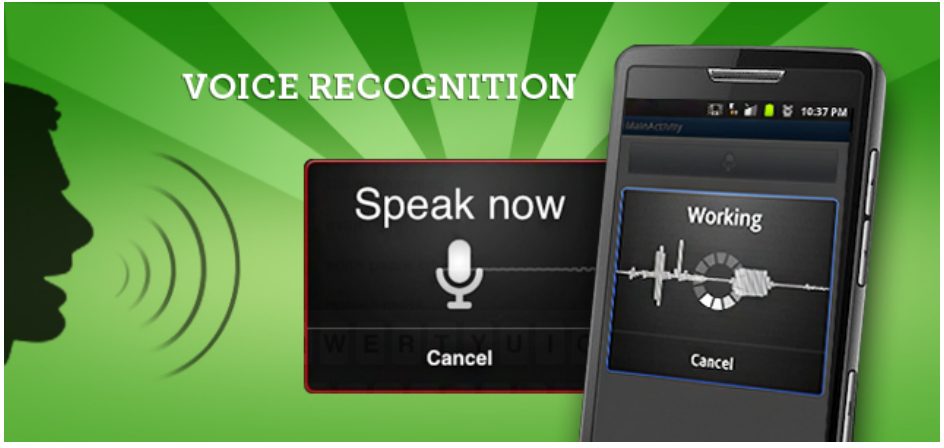
- Principal component analysis and friends:
  - Low-dimensional representations.
  - Decomposing objects into “parts”.
  - Visualizing high-dimensional data.
- Collaborative filtering:
  - Predicting user ratings of items.

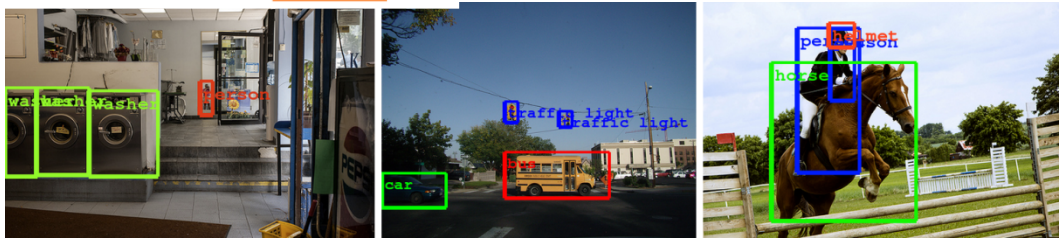


Deep Learning

- Neural networks: Brain-inspired ML when you have a lot of data/ computation but don't know what is relevant.







Markov Chains

- Markov chains: a model of sequence data.