

# CPSC 340: Machine Learning and Data Mining

The University of British Columbia

2017 Winter Term 2

Instructor: Mike Gelbart

# 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

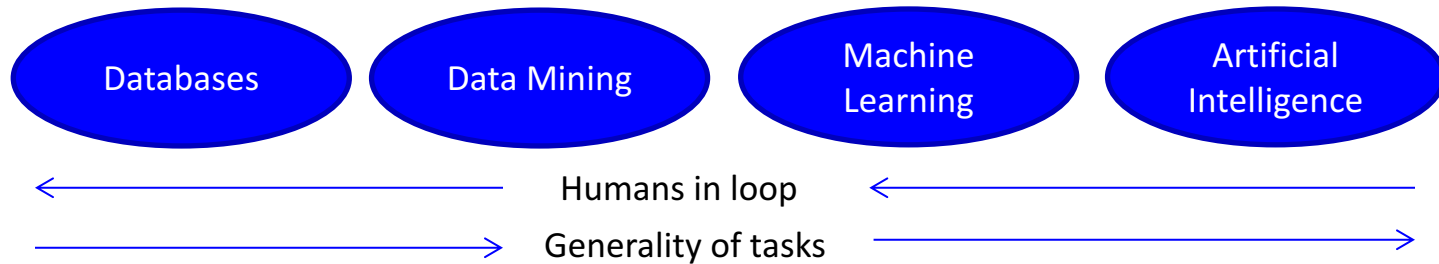
- What do you do with all this data?
  - Too much data to search through it manually.
- But there is valuable information in the data.
  - How can we use it for fun, profit, and/or the greater good?
- Data mining and machine learning are key tools we use to make sense of large datasets.

# 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), or contemplate something in 1000 dimensions.

# 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.



- If there is a difference, we'll be doing ML.
- Both are similar to statistics:
  - Less emphasis on 'correct' models and more focus on computation.

# About the course...

# Workload and difficulty

- For many people, this course is a LOT of work.
  - Some people spend **tens of hours per assignment**.
  - Your first assignment is due **next Wednesday** (in one week).
- Compared to typical CS classes, there is a **lot more math**:
  - Requires linear algebra, probability, and multivariate calculus.
- 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.
- Take this course to learn, not to get a certain grade.
  - Think of your desire grade,  $x$ . Now imagine you end up with  $x-10\%$  but you learned a ton, had a great time, and discovered an interest in ML. In this scenario, do you regret taking CPSC 340? If so, please consider dropping the course.

# Waitlist and prerequisites

- The waitlist is long. 130 students as of Dec 29.
- I have no power to move people off the waitlist.
  - This is done by the CS department administration, in a standard way for all CS courses, using priority queue.
  - If you're on the waitlist, you should still do the assignments.
  - If you're not serious about the course, please vacate your spot.
- If you do not meet the prerequisites...
  - If you do not have the prerequisites, **you will be automatically dropped from the course** unless you take action.
  - To appeal, follow instructions at <https://www.cs.ubc.ca/students/undergrad/courses-deadlines/prerequisites>



# Course Website

- Please visit the course website!!!
  - <https://github.ugrad.cs.ubc.ca/CPSC340-2017W-T2/home>
- The course website has info on:
  - Lecture schedule, notes, readings
  - TAs
  - Textbooks and other resources
  - Waitlist, registration, auditing
  - Grading and exams
  - Installing/learning the course software stack: Python 3, NumPy, git, GitHub
- **You are responsible for reading the information on the course website!!!!!!**

# UBC GitHub

- Some courses already run through github.com, e.g., CPSC 310
- We now have a GitHub Enterprise installation at github.ugrad.cs.ubc.ca
- Everything is on Canadian servers, which means we can store PII
  - You will receive your grades on github.ugrad.cs.ubc.ca
- This is still fairly new, hasn't yet been rolled out to all of CS or the whole university.
- To promote anonymity and privacy, this term we're using github.ugrad.cs.ubc.ca which is done through your CS ugrad accounts.
- Everyone in the course should be able to get an account at <https://www.cs.ubc.ca/getacct/>

# 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...

- If you don't already have an account, get one at <https://www.cs.ubc.ca/getacct/>
- Go to [github.ugrad.cs.ubc.ca](https://github.ugrad.cs.ubc.ca) and sign in with your ugrad 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.ugrad.cs.ubc.ca

**GitHub** Enterprise

Sign in via LDAP

**The University of British Columbia**  
*For Educational and Research use ONLY.*  
Login using your Campus-Wide Login

Username

Password

**Sign in**

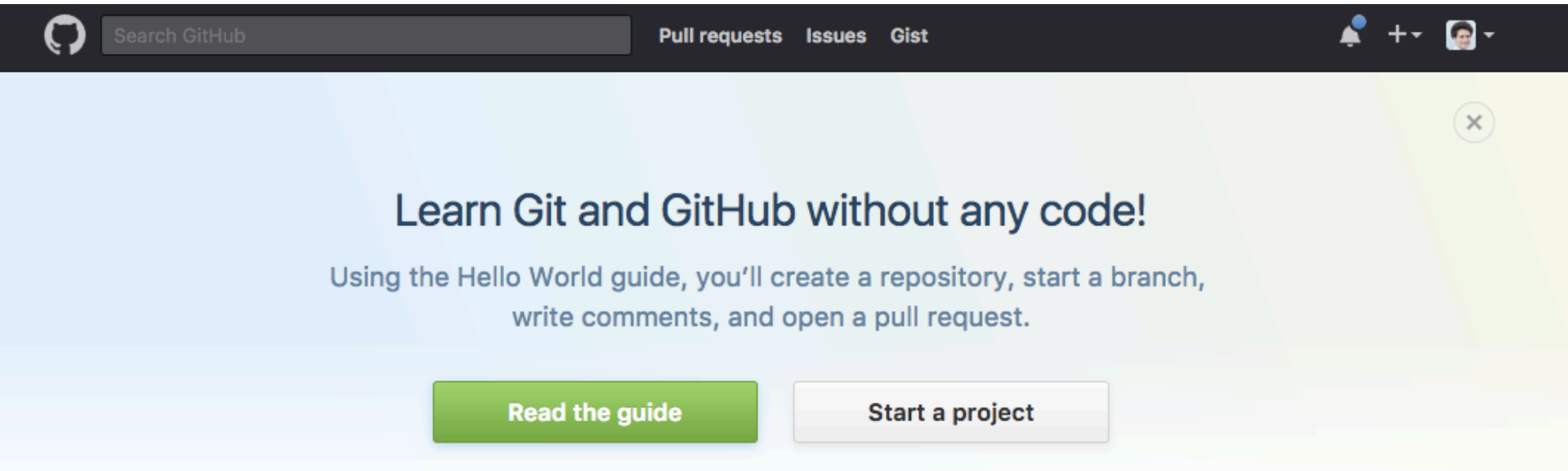
Your CS ugrad id  
(not student number,  
not CWL)



Your CS ugrad password



# Step 2: You should see something like this



The screenshot shows the GitHub homepage with a dark navigation bar at the top. On the left is the GitHub logo and a search bar containing the text "Search GitHub". To the right of the search bar are links for "Pull requests", "Issues", and "Gist". Further right are icons for notifications, a plus sign, and a user profile picture. The main content area features a light blue and yellow gradient background with a large heading "Learn Git and GitHub without any code!". Below the heading is a sub-heading: "Using the Hello World guide, you'll create a repository, start a branch, write comments, and open a pull request." At the bottom of the banner are two buttons: a green "Read the guide" button and a white "Start a project" button. A close button (an 'x' in a circle) is located in the top right corner of the banner area.

Search GitHub

Pull requests Issues Gist

Learn Git and GitHub without any code!

Using the Hello World guide, you'll create a repository, start a branch, write comments, and open a pull request.

Read the guide

Start a project

## Step 3: Access the course website

- Once you login you will have access to the course organization
- This lives at <https://github.ugrad.cs.ubc.ca/CPSC340-2017W-T2/>
- From here you can access your homework repositories and other internal documents

# Homework submission

- There's an entire document about homework submission procedures on the site – please read it!
- The short version is: you'll have a repository automatically created for you to work in.
  - For now I've posted a .zip file of assignment 0 just so that you can get started right away
  - I plan to create those repositories tonight, I had to wait for you to all log in initially (which you hopefully just did)



# Working in partners

- For Assignment 0 everyone must work individually
- 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 on the course page.
  - Last term people often forgot and I ended up having to fix this manually.
  - I will not do this again – you are responsible for remembering.
  - If you plan to work with the same partner for the whole term, I strongly recommend setting yourselves as partners for all assignments right now, so that you don't forget later.
  - The system is admittedly not great, we are working on improving it.

# Code of Conduct

- Do not post offensive or disrespectful content on Piazza or GitHub.
- If you have a concern, 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 share 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.
  - On the other hand, don't ask for help only after 10 hours of painful debugging. Don't be shy!
  - Perhaps 10-30 minutes of effort is a good guideline before asking for help.

# Lecture recordings

- As you can see, there's a camera at the back of the room!
  - This means I have to use a microphone, which I don't particularly like.
  - But hopefully it's worth it.
- I plan to make these lecture recordings publicly available on YouTube.
  - You will have access to lecture recordings.
  - In future terms I may switch to more of a discussion-based lecture model
- Because of copyright concerns, I'm removing most of the images from the slides (sorry).
  - You can Google for last term's slides if you want to see them.

# Lecture recordings (continued)

- If you're concerned about your head (or laptop screen!) appearing in the video, don't sit near the aisle in front of the podium.
- PLEASE don't be shy to ask questions
  - Students asking questions is standard for posted lecture videos.
  - I'm not sure if the microphones will pick them up anyway; I'll repeat all questions.
  - If you are really concerned about something you said, email me and I can try to edit it out.
- I will try to learn your names. If this bothers you, send me an email.

# 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 few months!

# The two URLs you need to write down

- Get a CS ugrad account: <https://www.cs.ubc.ca/getacct/>
- Course website: <https://github.ugrad.cs.ubc.ca/CPSC340-2017W-T2/home>