CSC321 Neural Networks and Machine Learning

Lecture 1

January 8, 2020

Welcome to CSC321!

Relevant Links

- Course Website: https://www.cs.toronto.edu/~lczhang/321/
- Piazza Message board:
  http://piazza.com/utm.utoronto.ca/spring2020/csc321
  - For all course related questions
- Markus (to be announced)

Introduction: Instructors

Pouria Fewzee (LEC0101)

- Prefers to be called "Pouria", but "Prof. Fewzee" is fine
- Email: pouria.fewzee [at] utoronto.ca
  - Logistic-related emails should go to Lisa
  - Please prefix email subject with 'CSC321'
- Office hours: Wednesday 12pm-2pm MN5107

Lisa Zhang (LEC0102) `*coordinator`

- Prefers to be called "Lisa", but "Prof. Zhang" is fine
- Email: lczhang [at] cs.toronto.edu
  - Please prefix email subject with 'CSC321'
  - Please do not email the surgeon with the same name
- Office hours: Monday 12pm-2pm DH3078

Introduction - You!

Survey: Why you are here

- Machine Learning is a growing field and I feel it is important to learn about it.

Survey: Why you are here

- Machine Learning is a growing field and I feel it is important to learn about it.
- It sounds really cool but also job opportunities

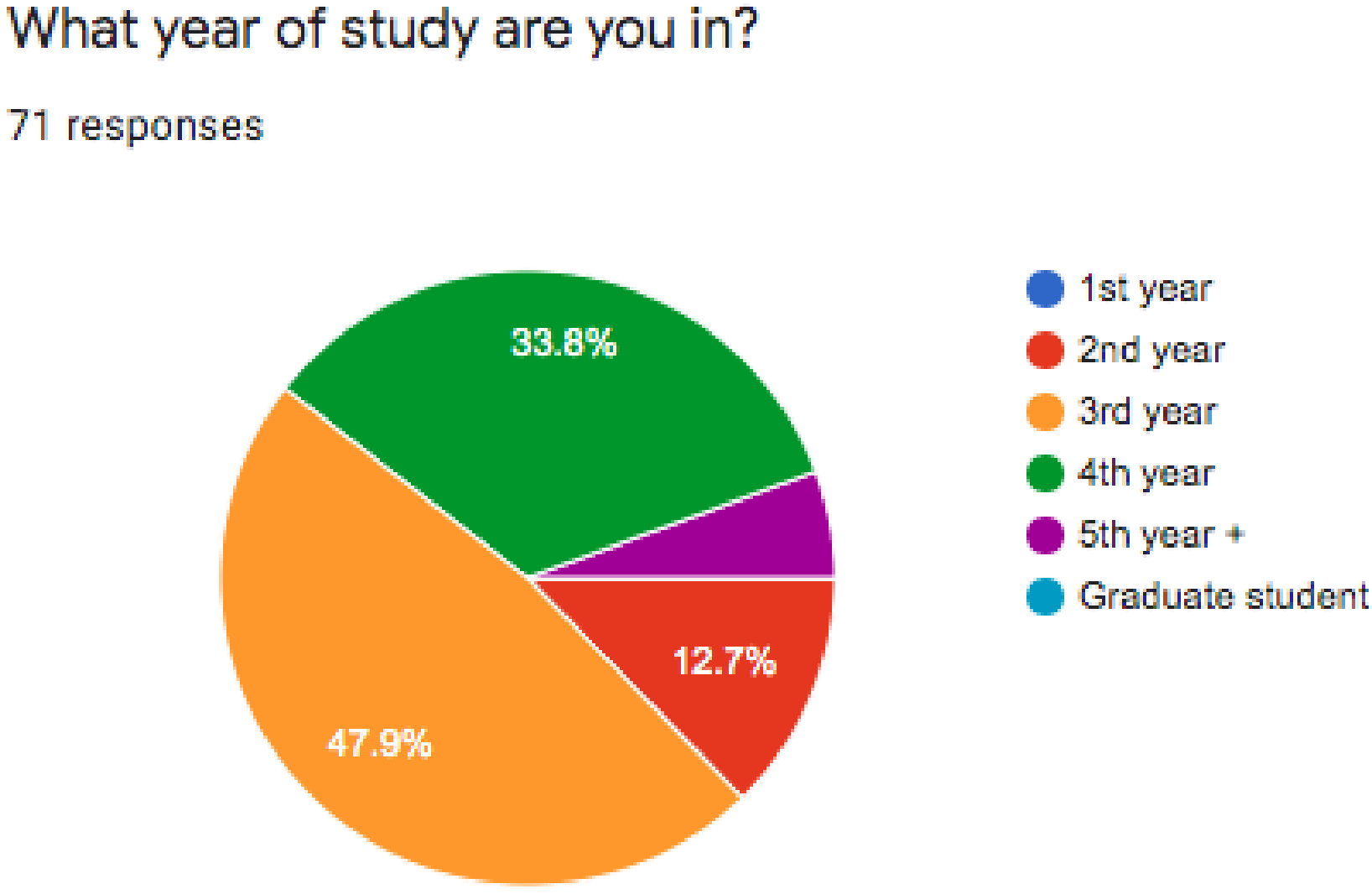
Survey: Why you are here

- Machine Learning is a growing field and I feel it is important to learn about it.
- It sounds really cool but also job opportunities
- I find how brains work fascinating, and Neural Networks are somewhat like digital brains.

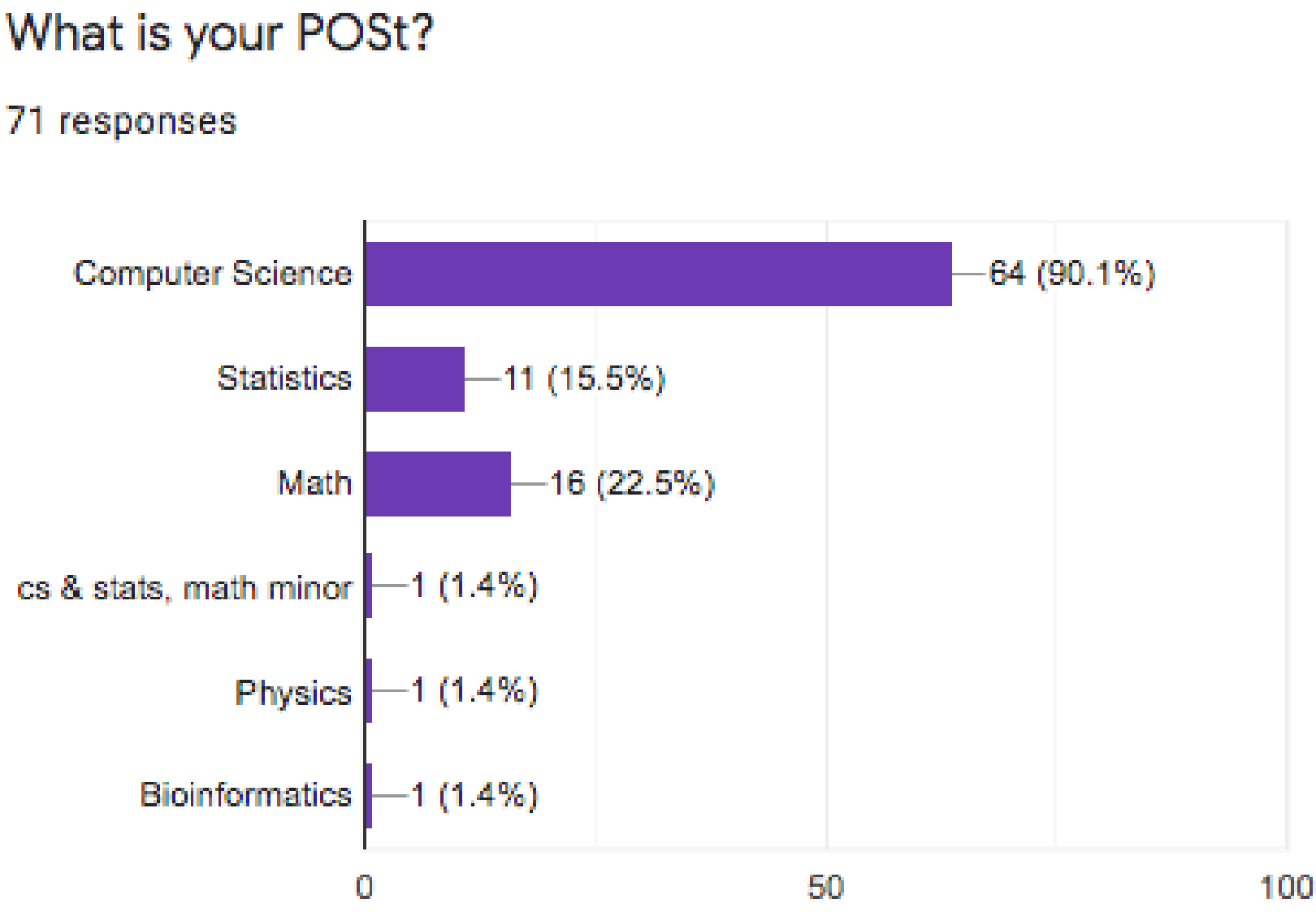
Survey: Why you are here

- Machine Learning is a growing field and I feel it is important to learn about it.
- It sounds really cool but also job opportunities
- I find how brains work fascinating, and Neural Networks are somewhat like digital brains.
- I don't understand anything

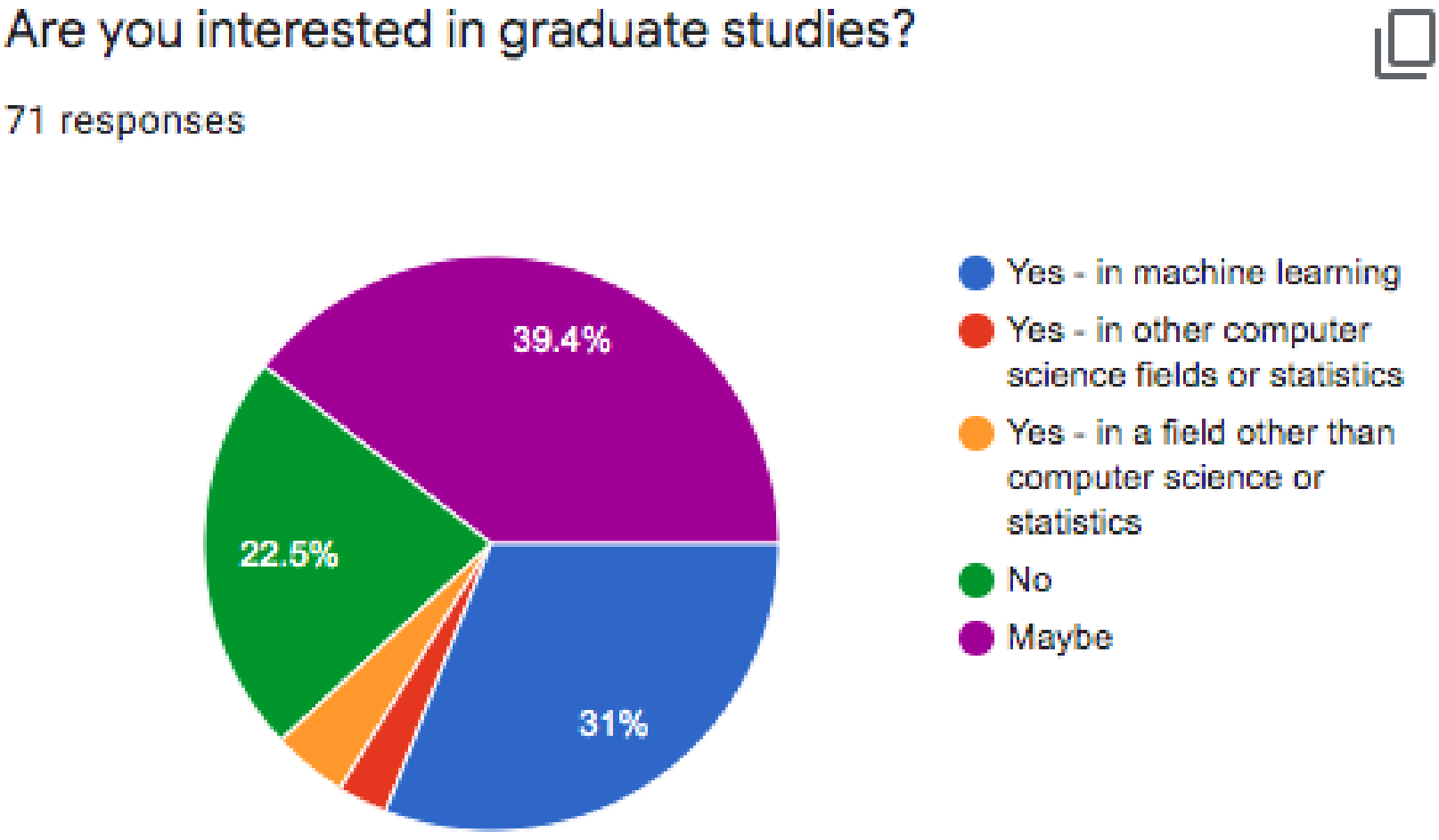
Survey: Demographic



Survey: POSt


What is your POSt?
71 responses
Computer Science
64 (90.1%)
Statistics
11 (15.5%)
Math
16 (22.5%)
cs & stats, math minor
1 (1.4%)
Physics
1 (1.4%)
Bioinformatics
1 (1.4%)
0
50
100

Survey: Graduate Studies

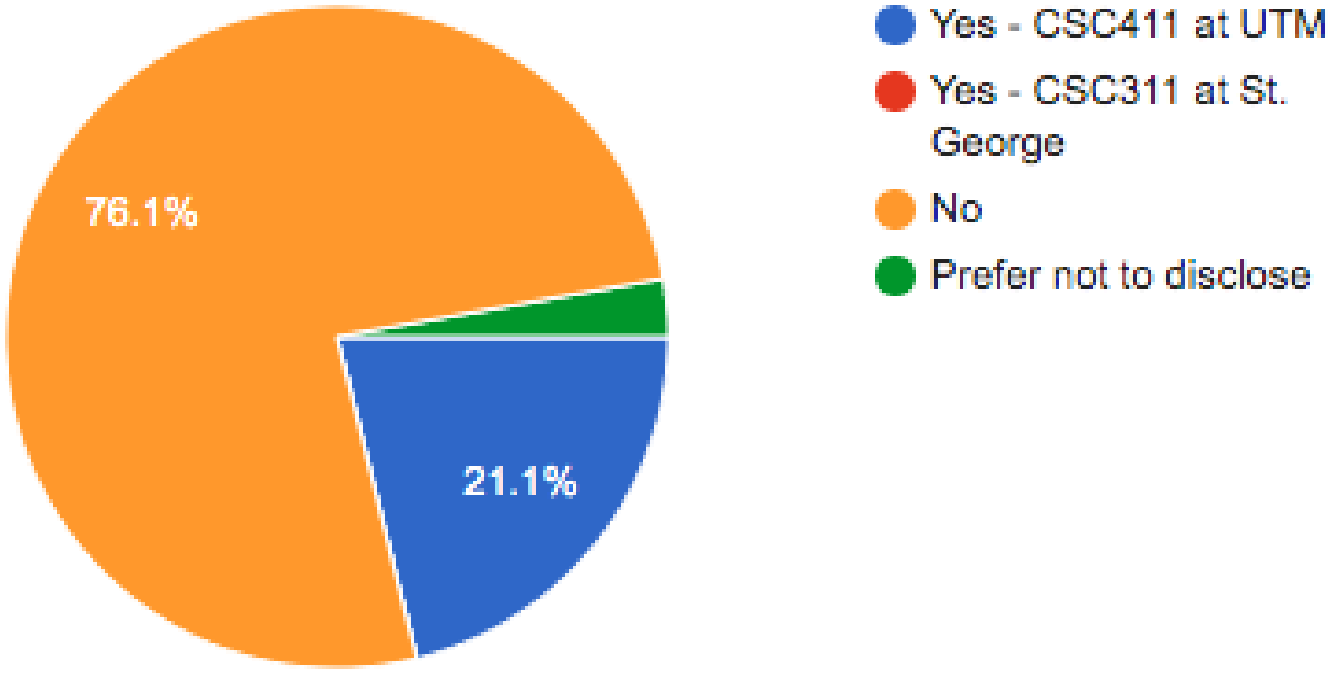


Consider taking CSC413 downtown instead if you've taken CSC411 and want to do graduate studies.

Survey: CSC411

This course does not have CSC411 Machine Learning as a pre-requisite, so the two courses will overlap. Have you already taken CSC411 (or CSC311)?

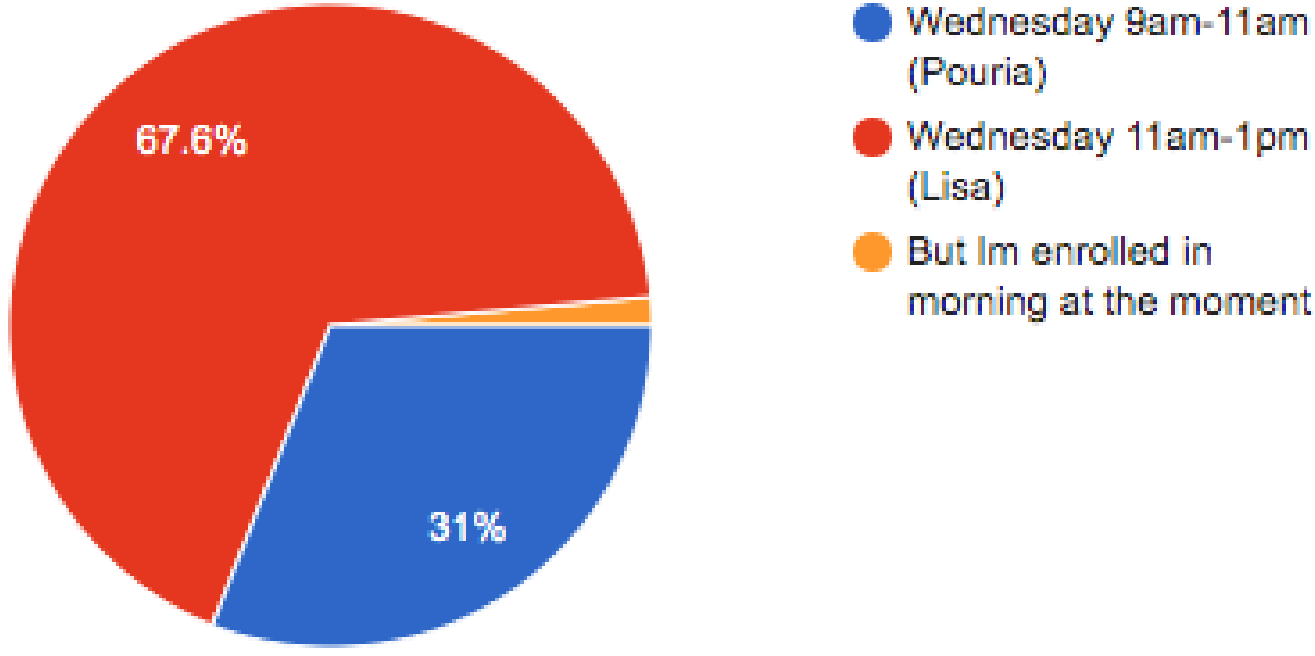
71 responses



Survey: Section

Which lecture session do you intend to attend?

71 responses

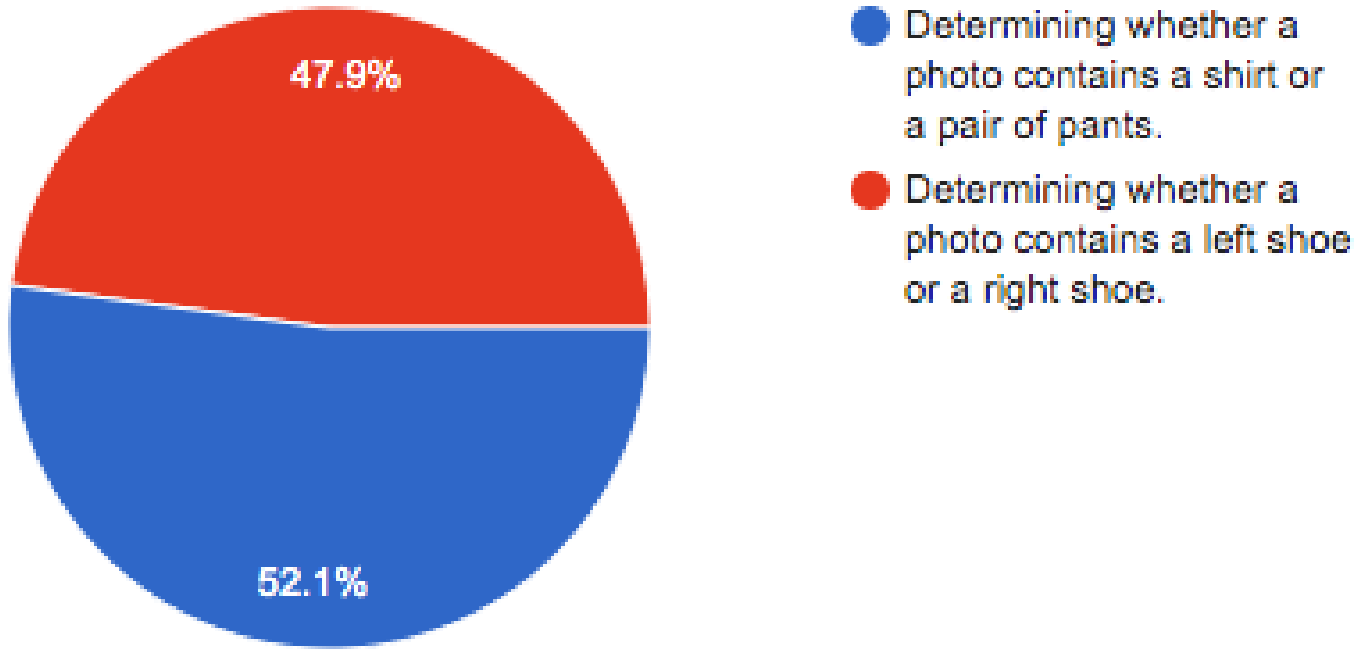


Students who are actually registered in this section have seating priority.

Survey: Project 3

Which machine learning task sounds more interesting?

71 responses



Survey – What you wanted us to know

- Very excited for this course! :)
- Excited for interesting machine learning projects.

Survey – What you wanted us to know

- Very excited for this course! :)
- Excited for interesting machine learning projects.
- Could you let us know about the intended difficulty of the course?
- Make it harder

Survey – What you wanted us to know

- Very excited for this course! :)
- Excited for interesting machine learning projects.
- Could you let us know about the intended difficulty of the course?
- Make it harder
- my linear algebra is very rusty

Survey – What you wanted us to know

- Very excited for this course! :)
- Excited for interesting machine learning projects.
- Could you let us know about the intended difficulty of the course?
- Make it harder
- my linear algebra is very rusty
- Looking forward to build some awesome project to put on my resume and get a job in related field.
- Please introduce us material that would prepare us for a potential career in this field.

Survey – What you wanted us to know

- Very excited for this course! :)
- Excited for interesting machine learning projects.
- Could you let us know about the intended difficulty of the course?
- Make it harder
- my linear algebra is very rusty
- Looking forward to build some awesome project to put on my resume and get a job in related field.
- Please introduce us material that would prepare us for a potential career in this field.
- I'm 6'1 but I haven't measured myself in a few years so I might be taller.

What is the difference between...

- Artificial Intelligence
- Machine Learning
- Deep Learning

Discuss with your neighbour

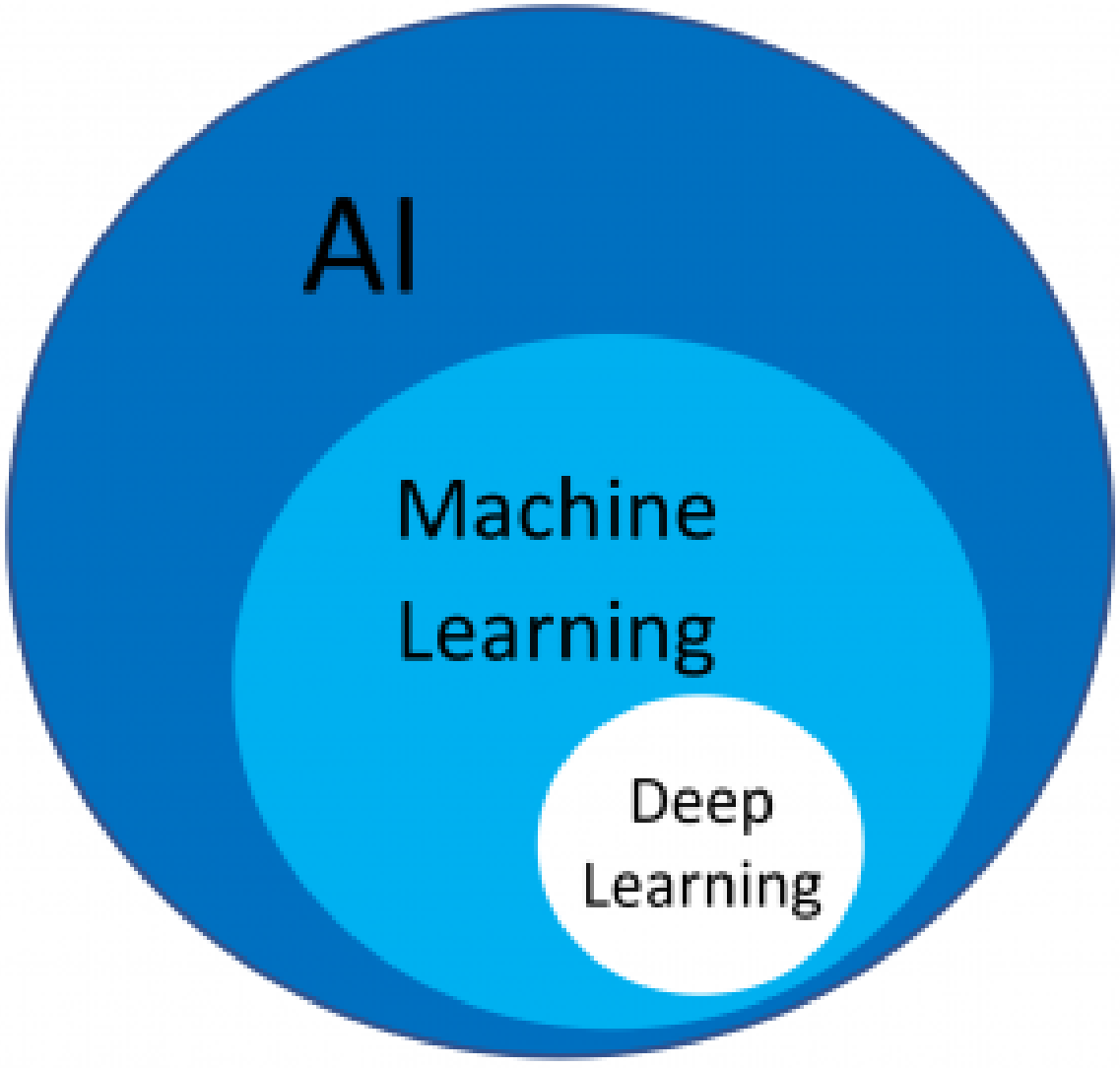
AI vs ML vs DL

Artificial Intelligence: Create intelligent machines that work and act like humans. (CSC384)

Machine Learning: Find an algorithm that automatically learns from example data. (CSC411/CSC311)

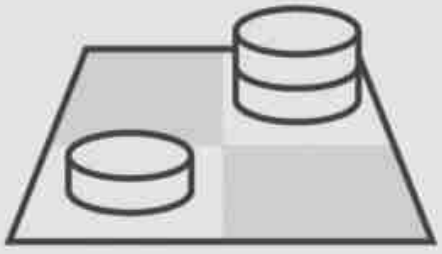
Deep Learning: Using deep neural networks to automatically learn from example data. (CSC321/CSC413)

Relationship



ARTIFICIAL INTELLIGENCE

Artificial Intelligence captures the imagination of the world.



MACHINE LEARNING

Machine learning starts to gain traction.

DEEP LEARNING

Deep learning catapults the industry.

Turing Test Devised
1950

ELIZA
1964 - 1966

Edward Shortliffe writes MYCIN, an Expert or Rule based System, to classify blood disease
1970s

IBM Deep Blue defeats Grand Master Garry Kasparov in chess
1996

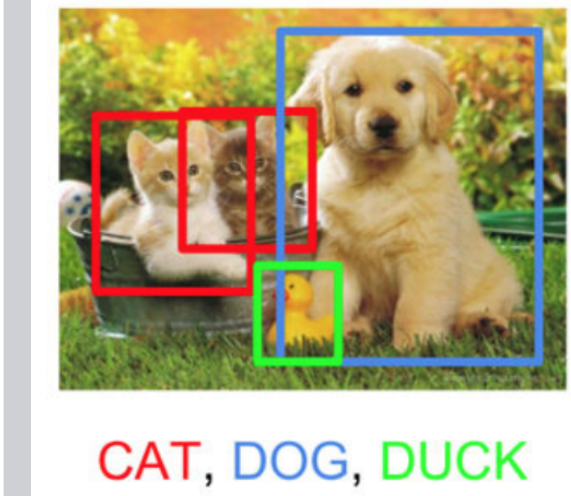
ImageNet Feeds Deep Learning
2009

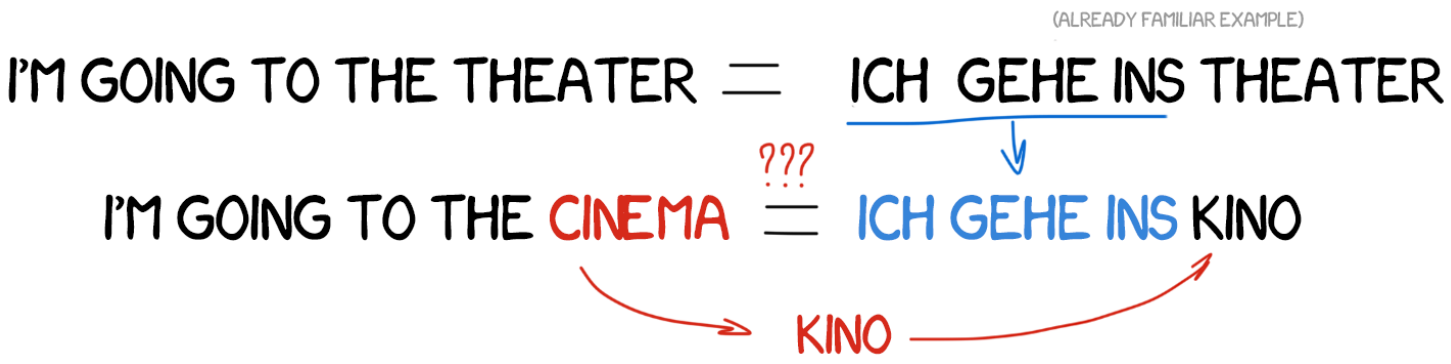
AlphaGo defeats Go champion Lee Sedol
2016

1950s 1960s 1970s 1980s 1990s 2000s 2010s

Why machine learning?

For many problems, it is difficult to program the correct behavior by hand. Machine learning approach: program an algorithm to automatically learn from data.





Types of Machine Learning Problems

- Supervised Learning
  - Regression
  - Classification
- Unsupervised Learning
- Reinforcement Learning
- (. . . and more)

Supervised Learning Task

Supervised Learning: learning a function that maps an input to an output *based on example input-output pairs*.

Examples:

- Age prediction given a headshot:
  - Input: headshot image
  - Output: person's height
- Sentiment classification given a tweet:
  - Input: tweet text
  - Output: whether the tweet is happy or sad

If we can collect *labeled* data, or data for which both (input, output) are known, then we can use supervised learning techniques.

Supervised Learning Task

- **Regression**: when the output is a continuous value
  - e.g. height prediction
- **Classification**: when the output is a categorical value
  - e.g. sentiment classification

Unsupervised Learning

Unsupervised Learning: learning the structure of some (unlabelled) data

Examples:

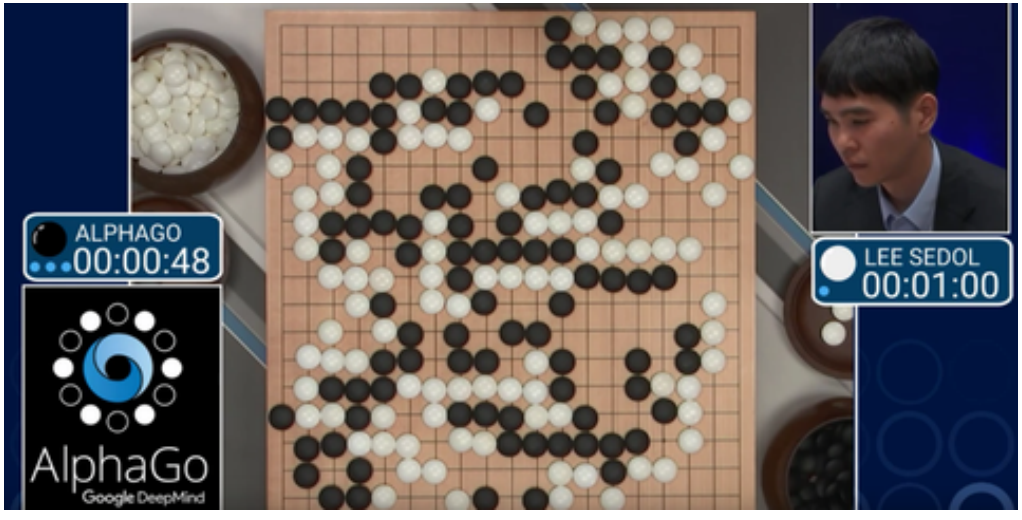
- clustering
- generating new images
- style transfer

Reinforcement Learning

Reinforcement Learning: learning what actions to take to optimize long-term reward.

Example:

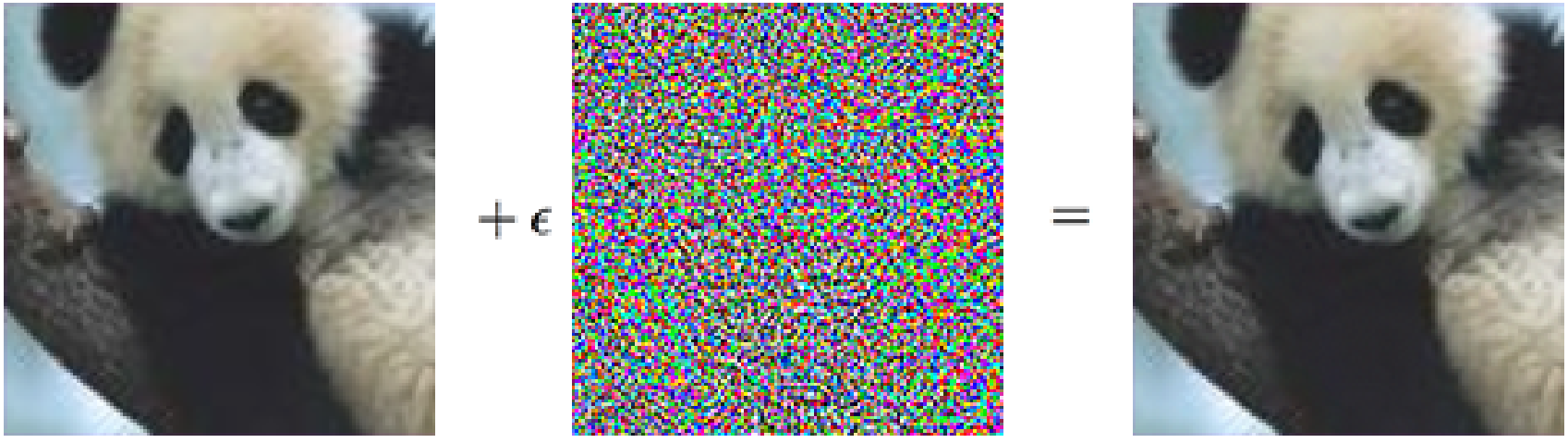
- playing a video game
- playing a game like go



Deep Learning Caveats: Interpretability

Figure 1: from https://xkcd.com/1838/

Deep Learning Caveats: Adversarial Examples



"panda"

57.7% confidence

"gibbon"

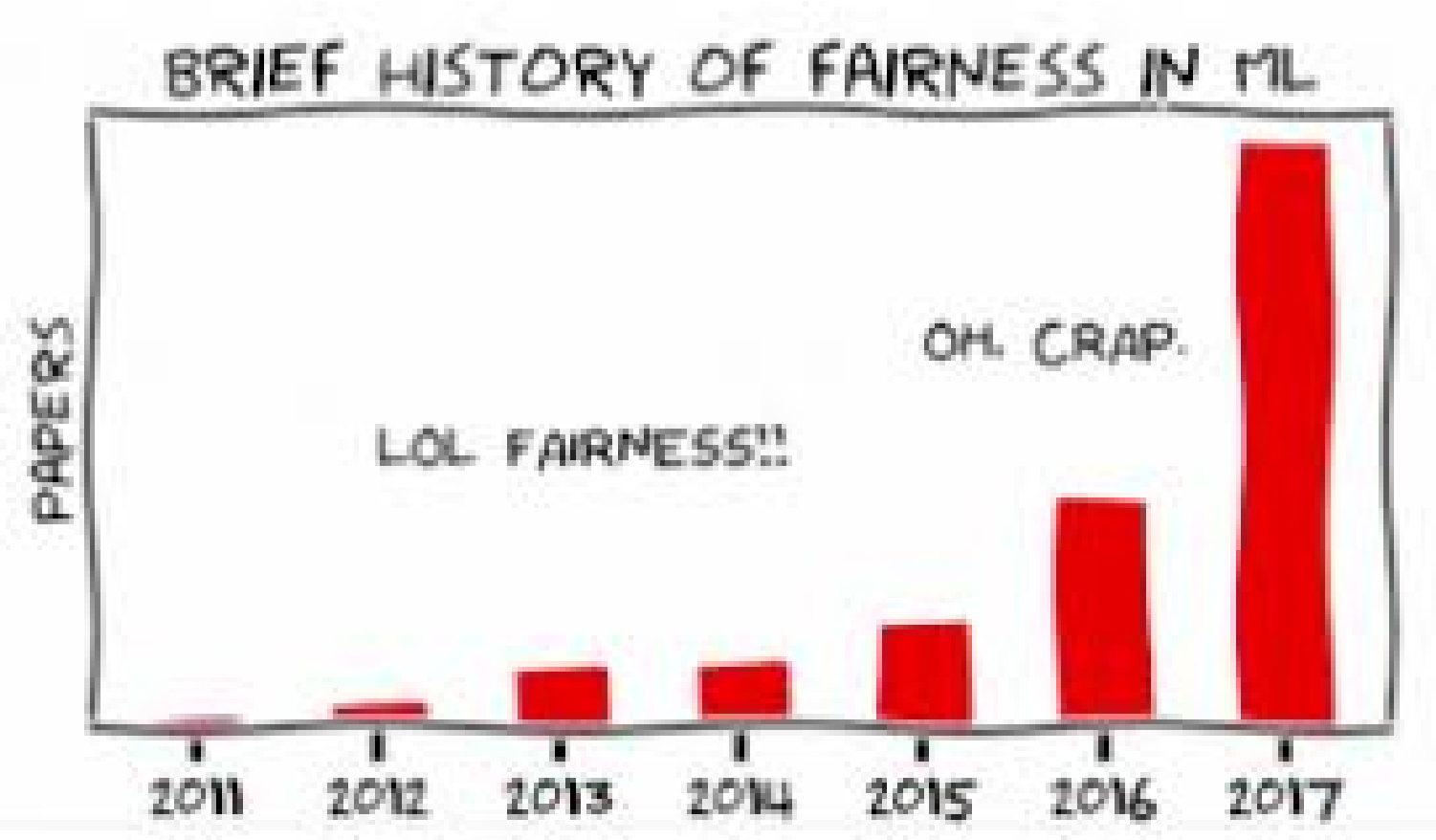
99.3% confidence

Deep Learning Caveats: Fairness

The U.S. military built an AI tool to find suitable combat personnel but had to shut it down because it was discriminating against women

News

Fairness in Machine Learning



Course Coverage

- We will focus mostly on neural networks and deep learning
- Mostly supervised learning (2/3 of the course)
- Some unsupervised learning (1/3 of the course)
- A tiny bit of reinforcement learning (weather-permitting)

Course Coverage

- We will focus mostly on neural networks and deep learning
- Mostly supervised learning (2/3 of the course)
- Some unsupervised learning (1/3 of the course)
- A tiny bit of reinforcement learning (weather-permitting)

Shameless plug: There will be a Machine Learning Reading Group this term! We'll either present supplemental material following CSC321 or following a Reinforcement Learning lecture series.

Pre-requisites

Formal pre-requisites:

- Calculus
- Linear Algebra
- Probability

Recommended preparation:

- Multivariable calculus
- Programming experience

Course Syllabus Scavenger Hunt

1. What textbook (if any) are we using for this course?
2. How much are the math homeworks worth and what time are they due?
3. How much are the coding assignments worth and what time are they due?
4. Can you do the homework / assignment in a group?
5. What is "Homework 0"?
6. What software will we use for this course?
7. What is the late policy for homeworks and assignments? How do grace tokens work?
8. When is the midterm going to be and how much is it worth?
9. What happens if there is a snow storm and class gets cancelled in weeks 1-6? In week 7?
10. What is plagarism and how can you avoid it?

Note Taker

Accessibility Services is looking for reliable volunteers to serve as note-takers this semester.

See http://www.utm.utoronto.ca/accessibility/volunteer-resources/volunteer-note-taker

Pre-requisite Quiz

Supervised Learning

Supervised Learning

Supervised Learning: learning a function that maps an input to an output based on example input-output pairs

Given a set of labelled examples (the *training set*), determine/predict the labels of a set of unlabelled examples (the *test set*)

Supervised Learning Examples

- Age prediction given a headshot:
  - Input: headshot image
  - Output: person's age
- Sentiment classification given a tweet:
  - Input: tweet text
  - Output: whether the tweet is happy or sad
- Exam grade prediction:
  - Input: assignment grades
  - Output: exam grade

Q: Are these regression problems or classification problems?

We'll use the last example today.

Supervised Learning Setup

Input: represented using the vector $\mathbf{x}$

- Example: $\mathbf{x}$ represents assignment grades (0-100)
- To start, let's assume that x is a scalar, and that we only have the cumulative assignment grade

Output: represented using the scalar t

- Example: t represents the grade on an exam (0-100)
- We'll use the scalar y to denote a *prediction* of the value of t

Supervised Learning Idea

- We have some data $(\mathbf{x}^{(1)}, t^{(1)})$, $(\mathbf{x}^{(2)}, t^{(2)})$, ... $(\mathbf{x}^{(N)}, t^{(N)})$
- We want to be able to make prediction y (of an unseen t) for a new value of $\mathbf{x}$
  - For example, predict the exam grade of a person who missed their exam
- How can we build a *model* to solve the prediction problem?

Supervised Learning Models

In the first three weeks of class, we'll talk about two types of models:

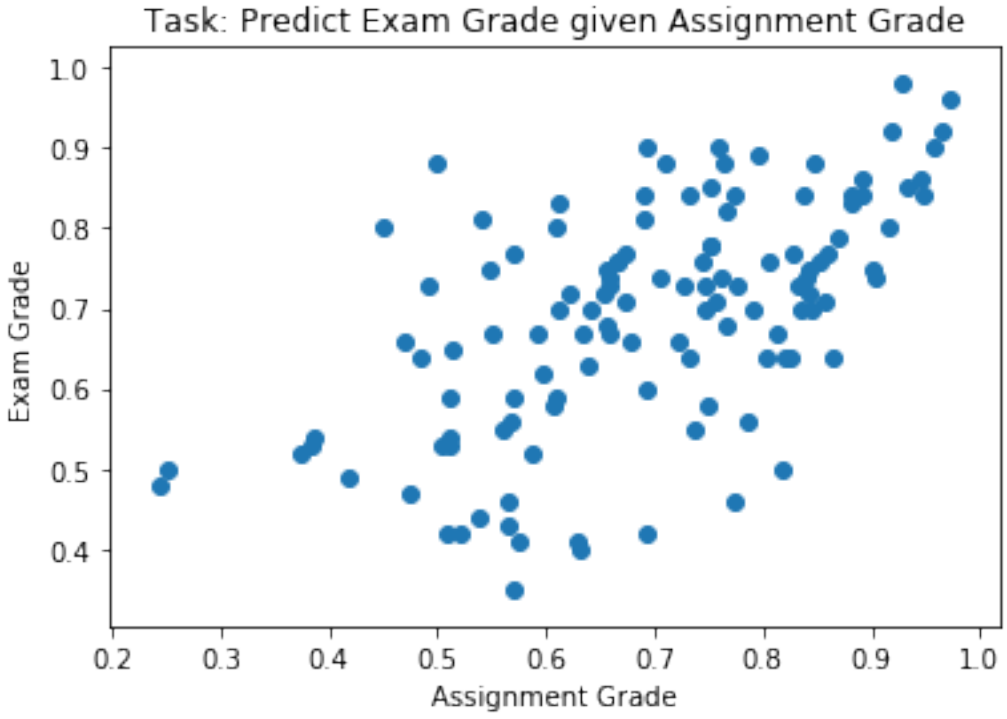
1. Linear Models
   - ...for regression: predict a scalar-valued target (Week 1-2)
   - ...for binary classification: predict a binary label (Week 2-3)
   - ...for multiway classification: predict a discrete label (Week 3)
2. k-Nearest Neighbours
   - ...for both regression and classification (Week 3)

The k-Nearest Neighbour model is arguably a lot simpler, but to get you started on homework 1 we'll talk about linear regression first.

Linear Regression

Supervised Learning Task: Exam Grade Prediction

(Definitely not real data from last term)



- Data: $(x^{(1)}, t^{(1)}), (x^{(2)}, t^{(2)}), \ldots \ (x^{(N)}, t^{(N)})$
- The $x^{(i)}$ are called *inputs*
- The $t^{(i)}$ are called *targets*

Linear Regression Model

A **model** is a set of assumptions about the underlying nature of the data we wish to learn about. The **model**, or **architecture** defines the set of allowable **hypotheses**.

In linear regression, our **model** will look like this

$$y = \sum_j w_j x_j + b$$

Where y is a prediction for t, and the w_j and b are **parameters** of the model, to be determined based on the data.

Linear Regression for Exam Grade Prediction

For the exam prediction problem, we only have a single feature, so we can simplify our model to:

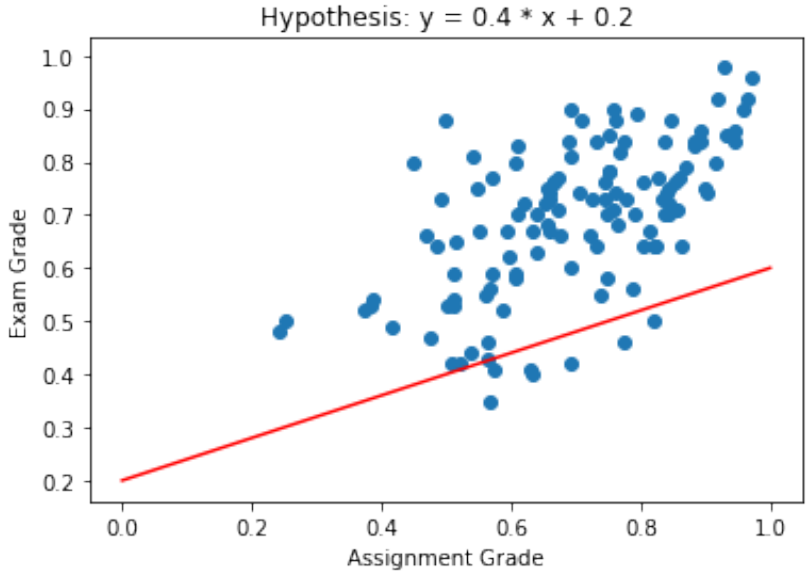
$$y = wx + b$$

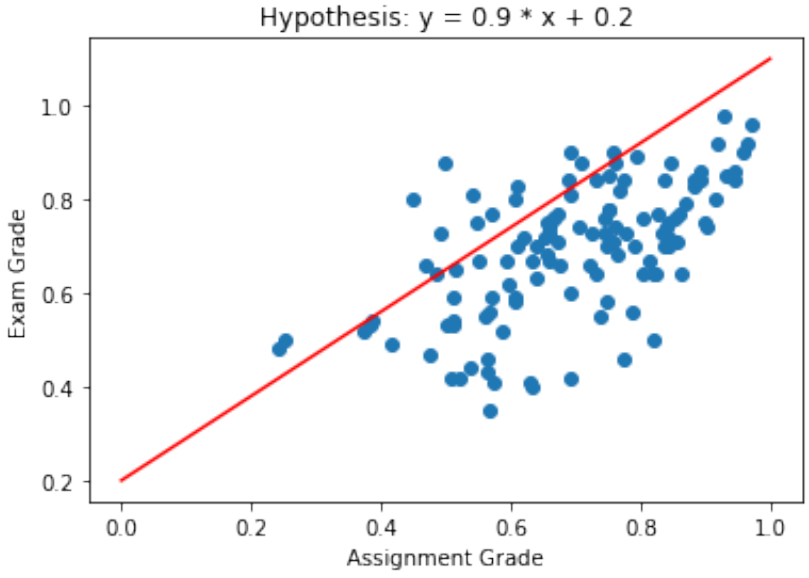
Our **hypothesis space** includes all functions of the form $y = wx + b$. Here are some examples:

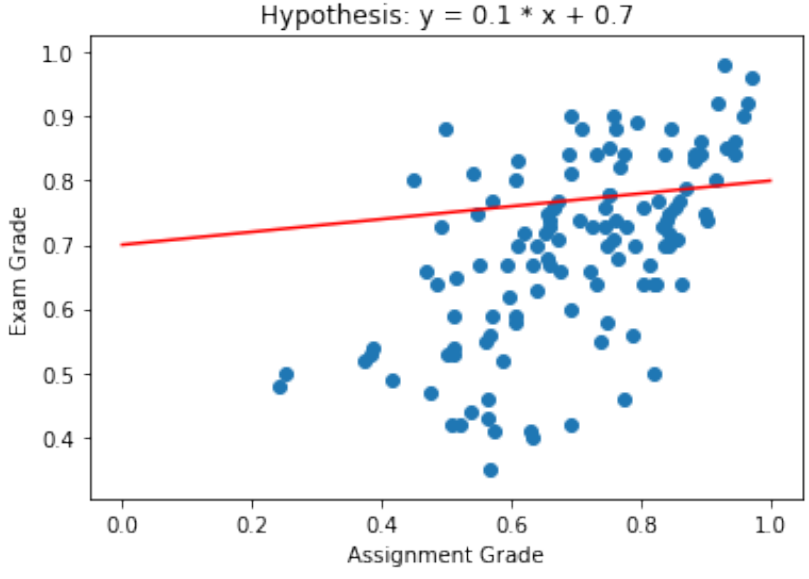
- $y = 0.4x + 0.2$
- $y = 0.9x + 0.2$
- $y = 0.1x + 0.7$
- $y = -x - 1$
- ...

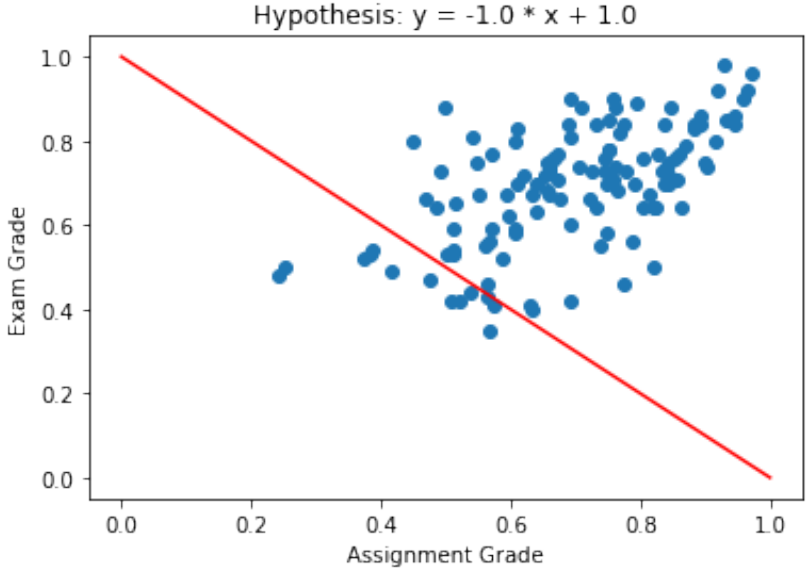
The variables w and b are called **weights** or **parameters** of our model. (Sometimes w and b are referred to as coefficients and intercept, respectively.)

Which hypothesis is better suited to the data?



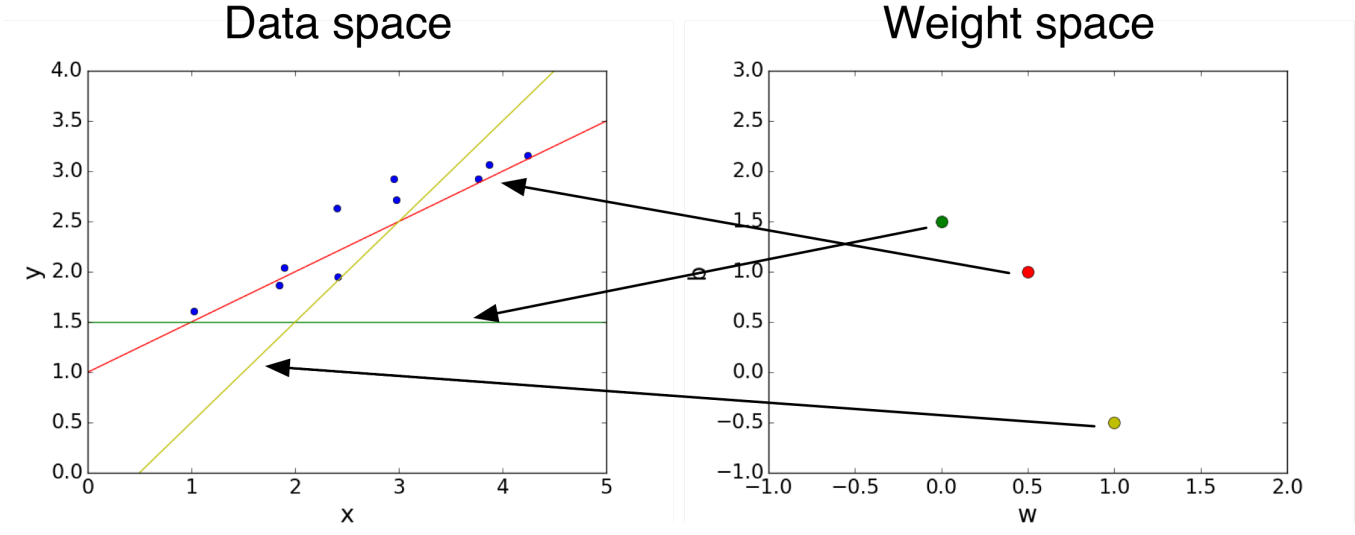






Hypothesis Space

We can visualize the hypothesis space or **weight space**:



Each *point* in the weight space represents a hypothesis.

Quantifying the "badness" of a hypothesis

Idea:

- A good hypothesis should make good predictions about our labeled data $(x^{(1)}, t^{(1)})$, $(x^{(2)}, t^{(2)})$, ... $(x^{(N)}, t^{(N)})$

Quantifying the "badness" of a hypothesis

Idea:

- A good hypothesis should make good predictions about our labeled data $(x^{(1)}, t^{(1)})$, $(x^{(2)}, t^{(2)})$, ... $(x^{(N)}, t^{(N)})$
- That is, $y^{(i)} = wx^{(i)} + b$ should be "close to" $t^{(i)}$

Quantifying the "badness" of a hypothesis

Idea:

- A good hypothesis should make good predictions about our labeled data $(x^{(1)}, t^{(1)})$, $(x^{(2)}, t^{(2)})$, ... $(x^{(N)}, t^{(N)})$
- That is, $y^{(i)} = wx^{(i)} + b$ should be "close to" $t^{(i)}$
- But how do we define the notion of "close to"?

We'll choose **square vertical distance**:

$$\mathcal{L}(y, t) = \frac{1}{2}(y - t)^2$$

This choice has some nice mathematical and statistical properties.

Cost Function (Loss Function)

The "badness" of an entire hypothesis is the average badness across our labeled data.

$$\begin{aligned}\mathcal{E}(w, b) &= \frac{1}{N} \sum_i \mathcal{L}(y^{(i)}, t^{(i)}) \\ &= \frac{1}{2N} \sum_i (y^{(i)} - t^{(i)})^2 \\ &= \frac{1}{2N} \sum_i ((wx^{(i)} + b) - t^{(i)})^2\end{aligned}$$

This is called the **loss** of a particular hypothesis.

Since the loss depends on the choice of w and b, we call $\mathcal{E}(w, b)$ the **loss function**.

Summary so far

Hypothesis	$y = wx + b$
Parameters	$w,\ b$
Loss Function	$\mathcal{E}(w, b) = \frac{1}{2N} \sum_i ((wx^{(i)} + b) - t^{(i)})^2$
Goal	Find w, b that minimize $L(w, b)$

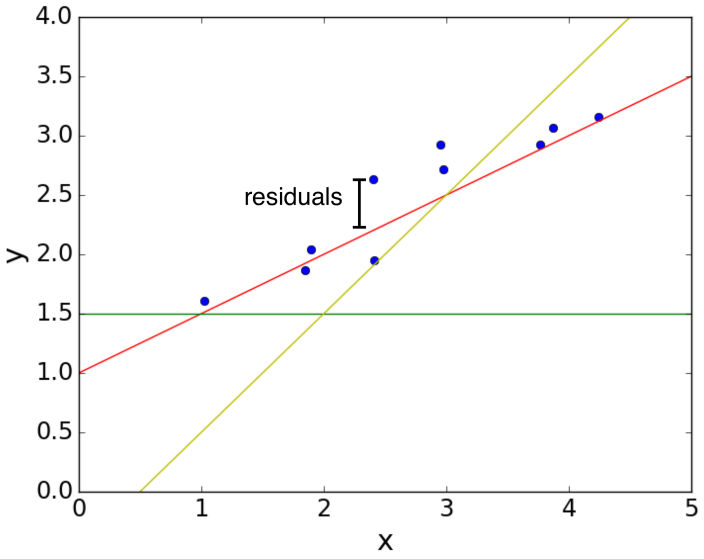
Minimizing the Loss Function

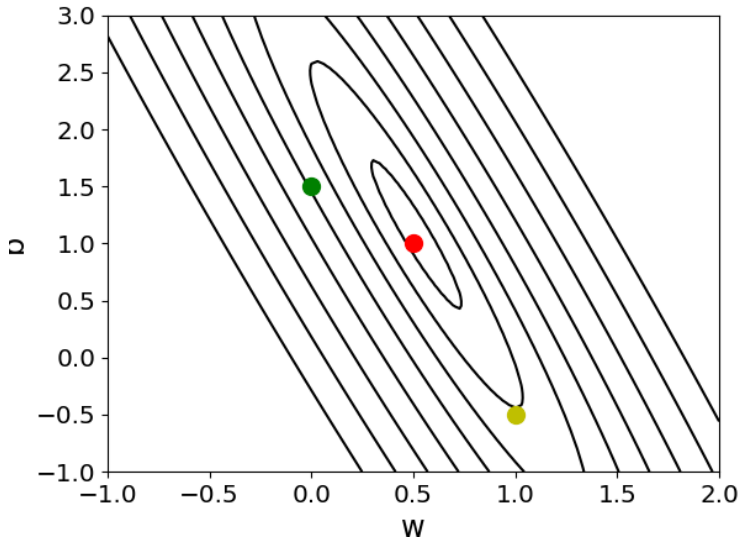
Task: Find w and b that minimize the loss function:

$$\mathcal{E}(w, b) = \frac{1}{2N} \sum_i ((wx^{(i)} + b) - t^{(i)})^2$$

Potential Strategy: Grid search

Search through combinations of (w, b).

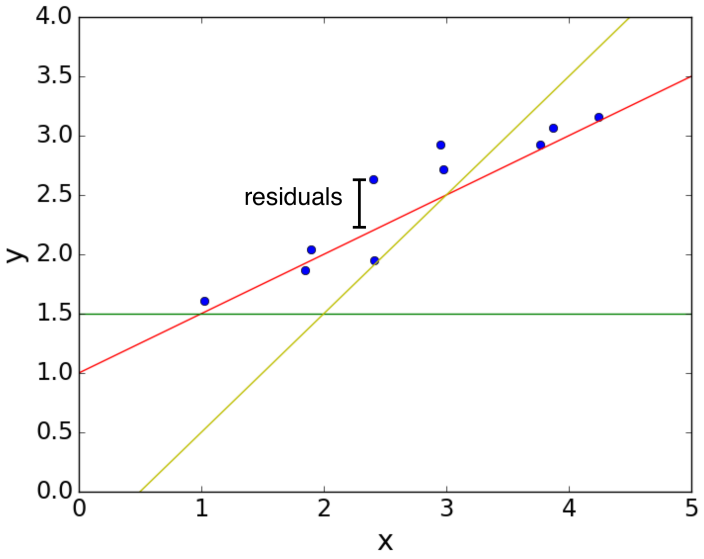


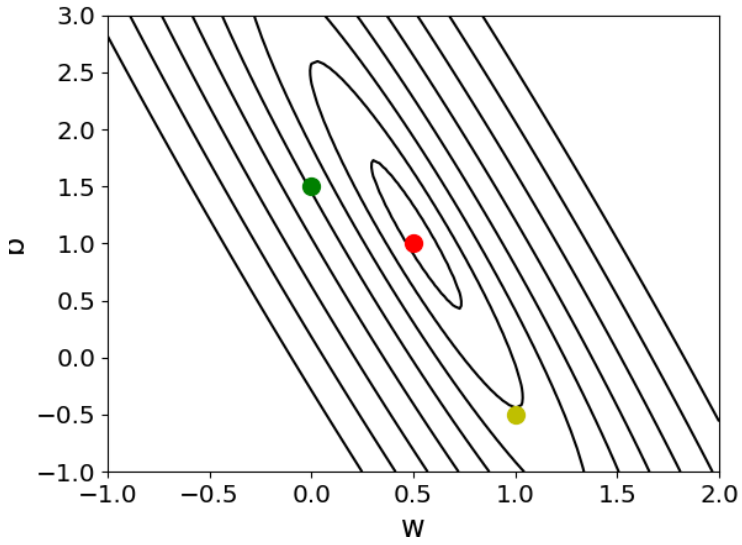


Why is this strategy poor?

Potential Strategy: Grid search

Search through combinations of (w, b).





Why is this strategy poor?

Slow! Especially if **x** is high dimensional.

Potential Strategy: Direct Solution

Find a *critical point* by setting

$$\frac{\partial \mathcal{E}}{\partial w} = 0$$
$$\frac{\partial \mathcal{E}}{\partial b} = 0$$

Possible for our hypothesis space, and are covered in the notes . . . and the pre-requisite quiz! See what we did there?

However, let's use a technique that can also be applied to more general models.

Strategy: Gradient Descent

. . . next class

Summary

- We started with a **prediction problem**: predict y for a given x.
- We restricted ourselves to one type of **model** or **architecture**.
- We defined a continuous **loss function** to frame the problem as an **optimization problem**.
- We will solve the optimization problem using **gradient descent**.

This strategy of turning a prediction problem into an optimization problem is key in machine learning.

What to do

Homework 0

- Math pre-requisite problems

Homework 1

- Due next Thursday 9pm
- You have everything necessary to finish this homework!

Project 1

- Start reading the handout
- Find a partner