CS 4850/6850: Introduction to Machine Learning **Fall 2018**

Topic 1: Introduction

Instructor: Daniel L. Pimentel-Alarcón

1.1 Introduction

In a nutshell, Machine Learning is about programming computers to replicate human tasks, for example:

- Determine whether a person is sick, similar to what a Doctor would do.
- Predict a stock price, similar to what a broker would do.
- Drive vehicles, like cars or planes.
- Classify/recognize things, like people faces.
- Identify suspicious activities, like credit fraud.
- Detect a crime in a surveillance video.

Modern Machine Learning mostly requires a combination of

- **Mathematical and statistical knowledge.** To derive and infer the models that will emulate human reasoning.
- **Computer science skills.** To code these models in efficient ways.

However, Machine Learning is a highly interdisciplinary field with many other applications in engineering, biology, chemistry, medicine, and virtually every field of knowledge.

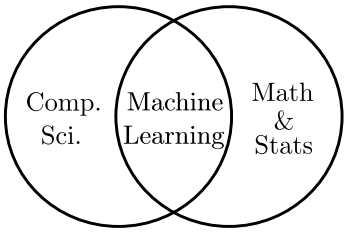


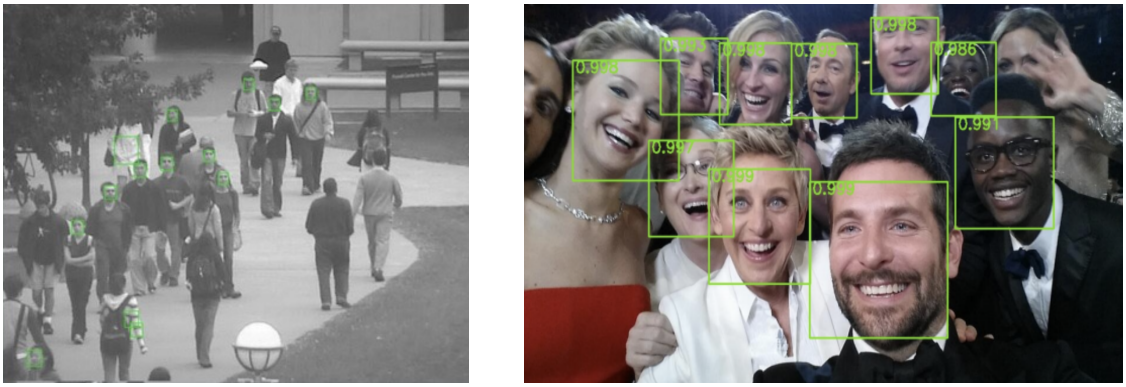
Figure 1.1: Modern Machine Learning mostly requires computer science, mathematics and statistics.

1.2 An Intuitive Example

Consider the following images:

Most humans will immediately recognize that the first two images correspond to the same person, and that the third image corresponds to a different person. However, a computer can only *see* these images as matrices containing numbers. Without any further programming, a computer cannot even recognize that these images contain faces! How would you code a computer so that it were able to classify face images like these? In other words, what would be your algorithm? How would you code a computer to determine which of the following images contain a human face?

In images like the following, how would you code a computer to locate all human faces?



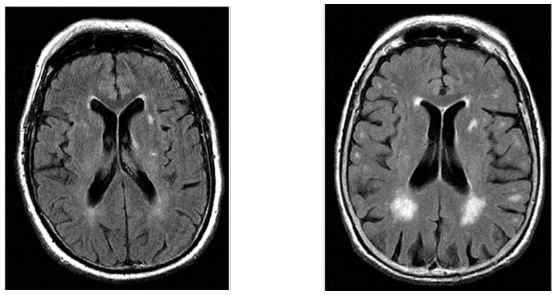
These are the type of the questions that Machine Learning aims to address. Most of these tasks are simple enough for most humans. However, they can be tremendously challenging to replicate with computers.

As you can imagine, these task are crucial towards self-driving vehicles and robotics. For example, an autonomous car must be able to recognize and locate obstacles (e.g., people) to determine its actions, just as a human driver does.

Figure 1.2: View of a self-driving vehicle during testing.

1.3 Other Motivating Applications

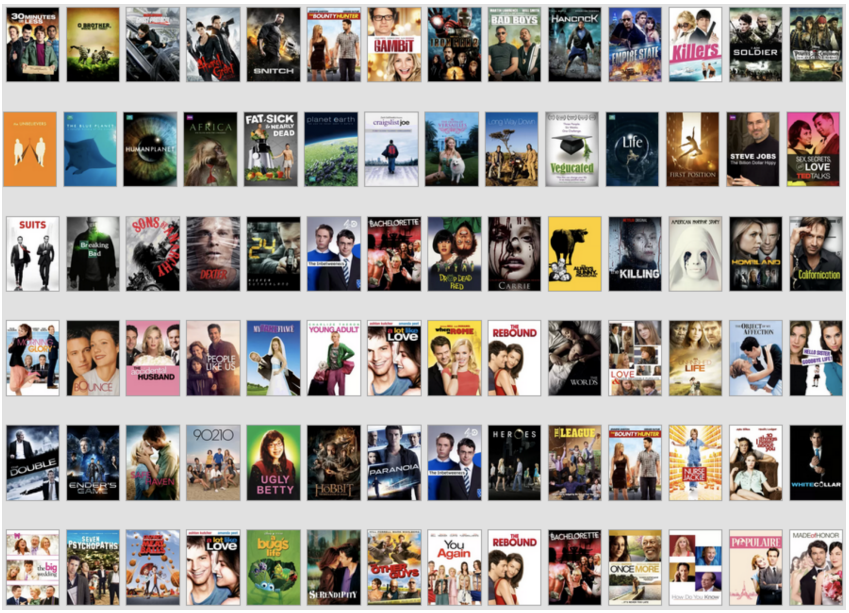
Example 1.1 (Precision Medicine)**.** Machine Learning also deals with problems that are not necessarily easy for *all* humans. For example, can you tell which of these magnetic resonance images (MRIs) corresponds to an individual with Alzheimer?



In fact, there is only a handful of neuroscientist experts that can make these determinations, and they cannot possibly analyze the MRIs of *all* people. Hence there is an increasing need to have computers aid making these diagnoses.

Example 1.2 (Recommender Systems)**.** Machine Learning also aims to solve problems that are difficult for all humans! For example, consider companies like Amazon, Netflix, Pandora, Spotify, Pinterest, Yelp, Apple, etc. These companies keep information of their users, such as age, gender, income level, and very importantly, ratings of their products. Their goal is to predict which users will like which items, in order to make good recommendations. If Amazon recommends you an item you will like, you are more likely to buy it. You can see why all these companies have a great interest in this problem, and they are paying *a lot* of money to Machine Learning Scientists who work on this. Humans cannot process these overwhelming amount of data (imagine how many movies are on Netflix, how many songs

on Spotify, how many products in Amazon), and so we rely on computers to process these data, and make predictions. How would you code a computer to make good predictions/recommendations? In particular, what would be your algorithm to recommend good movies to a Netflix user?

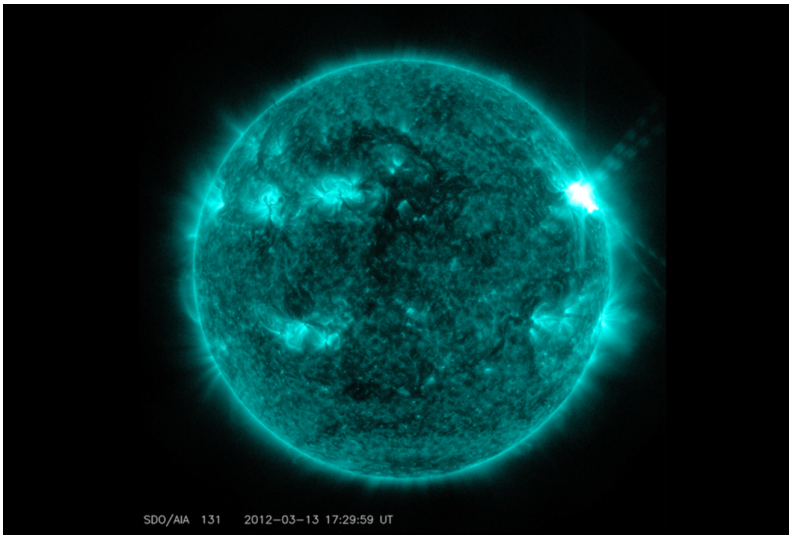


Example 1.3 (Genomics)**.** Similar problems, where data is overwhelmingly big for humans to analyze, also arise in genomics. For example, scientists often want to determine which genes are associated to which diseases (or features, like height or weight).

Example 1.4 (Solar Flares)**.** The Sun, like all active stars, is constantly producing huge electromagnetic *flares*. Every now and then, these flares hit the Earth. Last time this happened was in 1859, and all that happened was that you could see the northern lights all the way down to Mexico — not a bad secondary effect! However, back in 1859 we didn't have a massive power grid, satellites, wireless communications, GPS, airplanes, space stations, etc. If a flare hits the Earth now, all these systems would be crippled,

and repairing them could take *years* and would cost *trillions* of dollars to the U.S. alone! To make things worse, it turns out that these flares are not rare at all! It is estimated that the chance that a flare hits the earth in the next decade is about 12%.

Of course, we cannot stop these flares any more than we can stop an earthquake. If it hits us, it hits us. However, like with an earthquake, we can act ahead. If we know that one flare is coming, we can turn everything off, let it pass, and then turn everything back on, like nothing happened. Hence the NASA and other institutions are investing a great deal of time, effort and money to develop techniques that enable machines to *predict* that a flare is coming. So essentially, we want to device a sort of flares *radar* or *detector*.



Example 1.5 (Fraud Detection)**.** Credit cards are a classical example of fraud detection. The main idea is to look at usage patterns, and identify *outliers* (unusual activity).

For a more interesting example, consider *click fraud*. Companies pay popular websites to advertise their products. How much they pay depends on the popularity of each website, measured in number of clicks. Hence, companies often cheat, using *bots* that click their websites (to have a higher click count, and charge more for advertising). How would you detect whether a click is genuine or fraudulent?

1.4 Notation

Throughout this course we will use standard mathematical notation. You should get familiar with it:

	Examples	Regular	Bold	Lower	Capital	Roman	Script
Scalar	x	✓		✓		✓	
[Column] vector	$\mathbf{x}$		✓	✓		✓	
Matrix	$\mathbf{X}$		✓		✓	✓	
Random variable	x	✓		✓			✓
Random vector	$\boldsymbol{x}$		✓	✓			✓
Random matrix	$\boldsymbol{X}$		✓		✓		✓

1.5 Summary

This lecture shows some motivating applications of Machine Learning. In all these applications, the task can be summarized as replicating human tasks. This will often involve pre-processing (for example transforming real-valued glucose levels to a binary *label* indicating healthy/diabetic), visualization, analysis, classification, modeling, and several other tasks that we will study in upcoming lectures. Notice that in most modern applications, data tends to be big — high-dimensional, and with a large number of samples. Without efficient computer processing techniques (e.g., distributed systems), many Machine Learning tasks would be impossible.