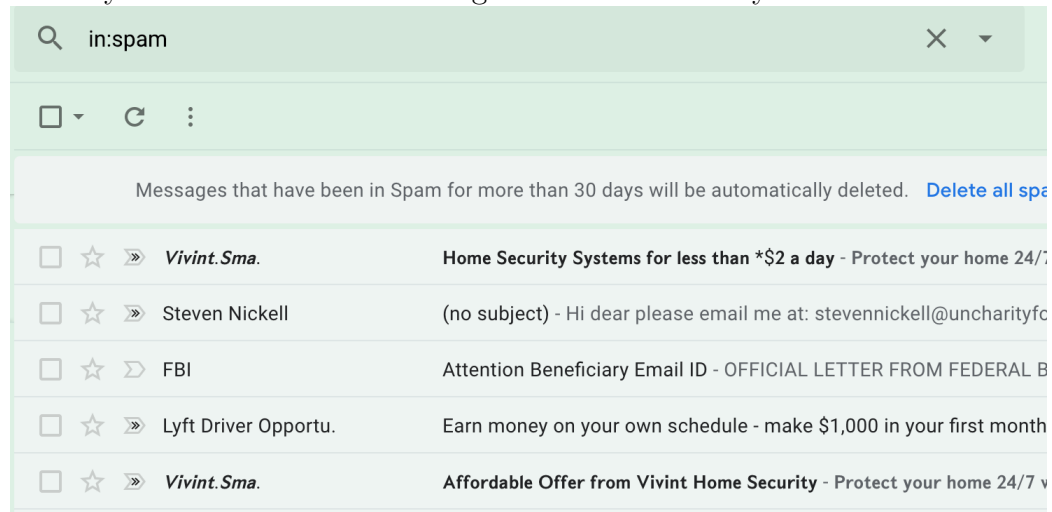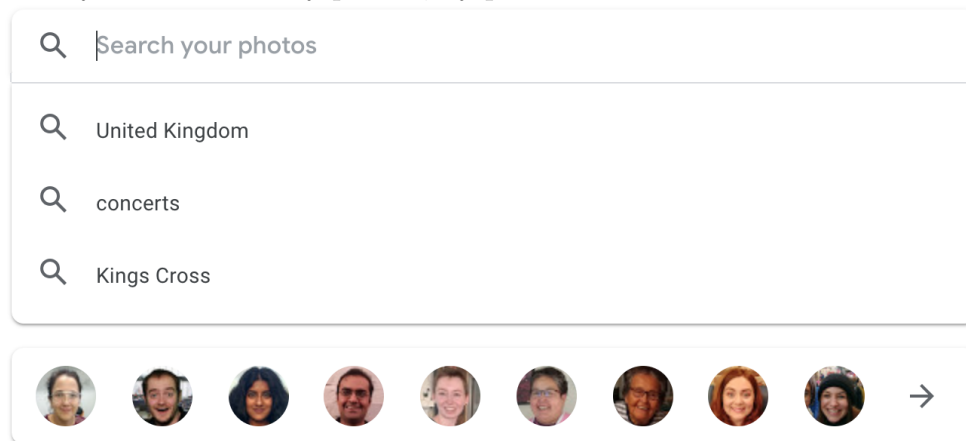# 1 Week 1 - Linear Regression

2 Introduction

//TODO add references to Coursera ML and Machine Learning for Humans book We come across machine learning all the time

//TODO: add overall outline and what we are intending to cover

It's why I don't have to sift through these emails in my inbox

It's why I can search my photos, by photos

2.1 What is Machine Learning?

According to wikipedia: Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead.

Machine learning algorithms can learn to do a particular task without being explicitly programmed by building a mathematical model based on sample data, known as "training data". Then, that model can be applied to new data not previously used to build the model.

2.2 What is an algorithm?

An algorithm is often described as a set of steps to accomplish a particular task. You could describe an algorithm for brushing your teeth, or making a grilled cheese sandwich

But it's a bit more general than just a set of steps

Valid input ——————▶ algorithm ——————▶ Valid output
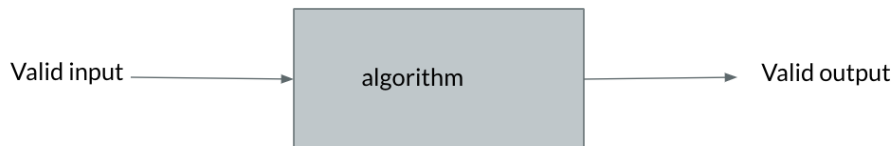
An algorithm is a way to solve a computational problem, and a computational problem just specifies valid input and desired output.

For example, for the problem of spam filtering the input would be an email message, and the output would be a classification (spam/not spam). The algorithm could be a set of steps. We could compose a series of regular expressions to the message that are based on previous messages that we know have turned out to be spam.

We could also train a model based on a dataset of emails and classifications, to learn patterns of what spam messages look like without explicitly writing spam identification rules. Then, we could use the model for new data without a classification. That's the approach we're more interested in here, but both approaches are algorithms.

2.3 What is a machine learning algorithm?

The ML Coursera course defines a *Well Posed Machine Learning Problem*

A computer program is said to learn from experience E with respect to some task T and some performance measure P if its performance on T, as measured by P improves with experience E

So, we can see that the rule based spam filtering approach wouldn't be a machine learning based approach because having more labelled data would not help us to classify spam any more accurately.

2.4 Supervised Learning

Here is an example of a supervised learning problem:

Suppose we have a dataset of houses that have sold containing how much they have sold for, and the area of the property in square feet.
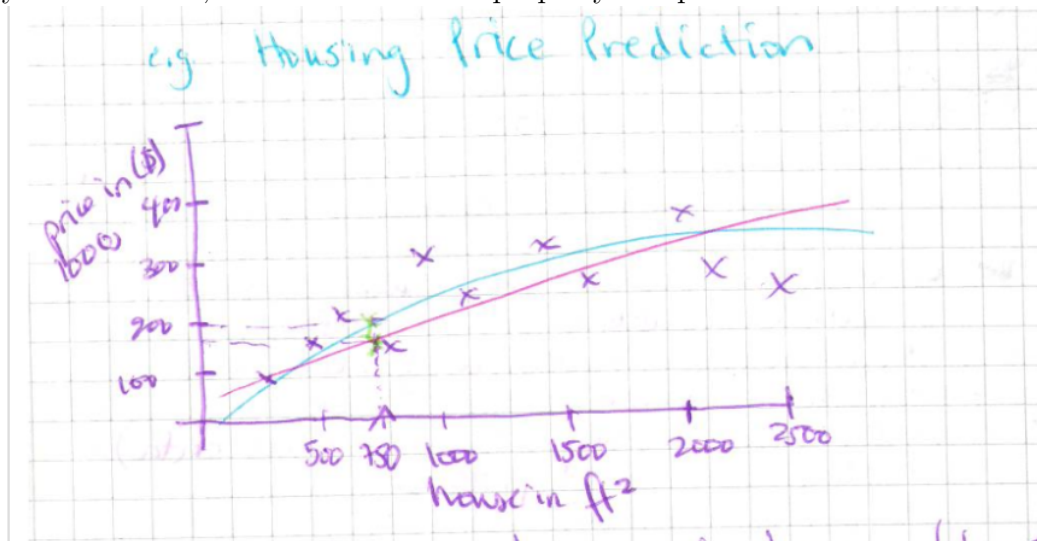
Then, we can use this data to build a model so that we can make a prediction like:

For this new 750 ft^2 property, how much will it sell for?

This is a *Supervised Learning* problem because existing labelled data (with the right answers) to work with and to test how accurate our model is.

This is also a *Linear Regression* problem because the prediction is a continuous valued output.

Another example of a supervised learning problem would be:

Suppose we want to predict whether a tumour is benign or malignant, and we have a dataset that contains tumours, their size and whether they are benign or malignant.
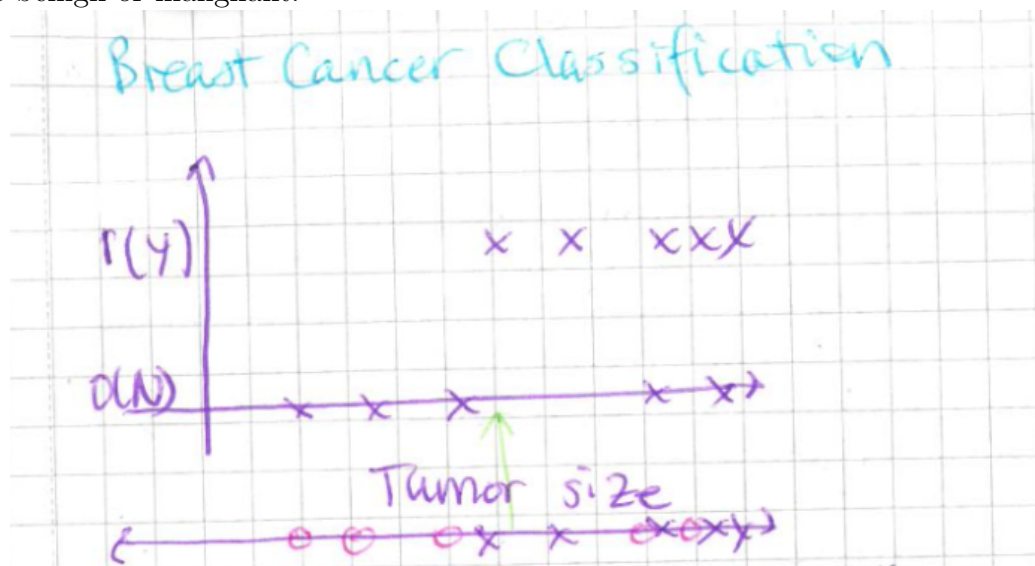
This would be a *classification* problem, because the prediction is a discrete valued output (in this case, either benign or malignant).

These might seem like toy examples so far. Cancer researchers will collect all sorts of data about tumours and patients that could all be useful in our model. But, we will aim to introduce concepts with a simplified view of the data to help our understanding, then move on to being able to use more features/data and more complicated techniques.

2.5   Unsupervised Learning

There is another area of Machine Learning, unsupervised learning that doesn't use labelled data to build a model. We won't cover unsupervised learning here, but let's introduce a couple unsupervised learning problems so that we'll get a feel for what these problems are like, and why they are different from supervised learning.

Here is an example of an unsupervised learning problem:

Headings

European media greet May's defeat with Groundhog Day fatigue

The Guardian · 1 hour ago

- **Brexit: Theresa May says UK can still leave EU with 'good deal'**
  BBC News · 3 hours ago

- **Brexit: EU points finger at UK for Theresa May's deal defeat**
  BBC News · 8 hours ago

- **A no-deal Brexit could still happen, even if MPs vote against it – and this is why**
  The Independent · 2 hours ago · **Opinion**

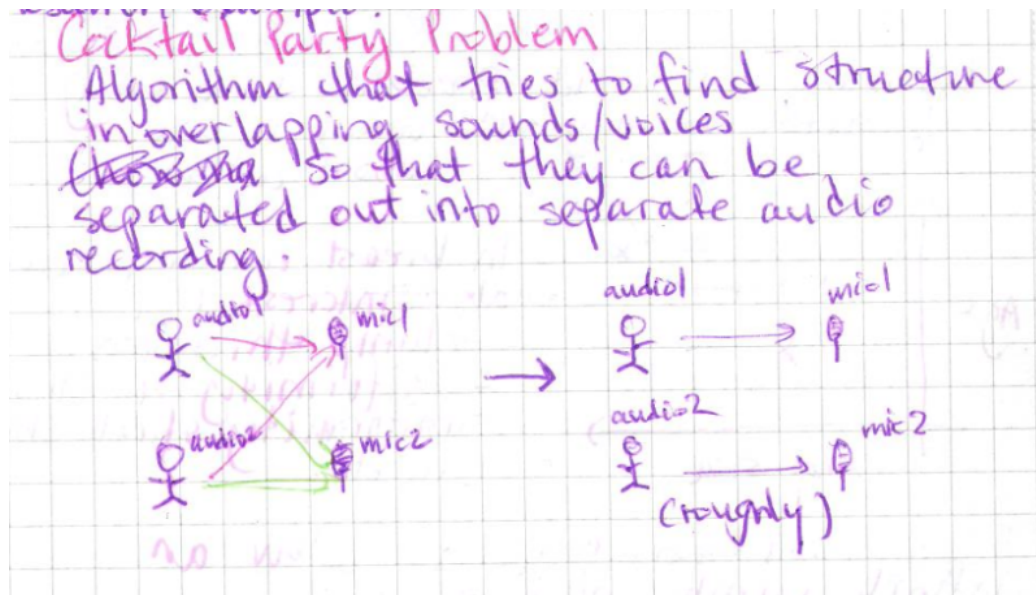- **An extension would be a national humiliation. The UK must leave the EU on time**
  Telegraph.co.uk · 5 hours ago · **Opinion**

View full coverage

News aggregators don't have a predefined list of topics, but will find an underlying structure in news articles to be able to group them together, as seen above.

Here is another example of an unsupervised learning problem:

Given fixed mics in a room, and conversations happening around them, the mics will collect audio data for all of the conversations mixed together. If you want to be able to follow one of the conversations, an algorithm that identifies the underlying structure to be able to extract just the parts of the audio that belong to one conversation is an unsupervised learning problem.

2.6   Linear Regression

Let's move on to working through our first machine learning problem.

Suppose we want to predict a person's weight, given their height. This isn't something I actually want to do, but it's two variables that are clearly dependent on each other.

Here is a sample from a dataset of heights and weights:

Let's plot height vs weight for the whole dataset.

Table 1: Sample from a dataset of heights and weights

Gender	Age	Height (inches)	Weight (lbs)	How would you describe your weight?
Female	18	67	150	Slightly overweight
Male	17	67	140	About the right weight
Female	16	67	100	Slightly underweight
Female	18	62	185	Slightly overweight
Male	17	62	140	Slightly overweight
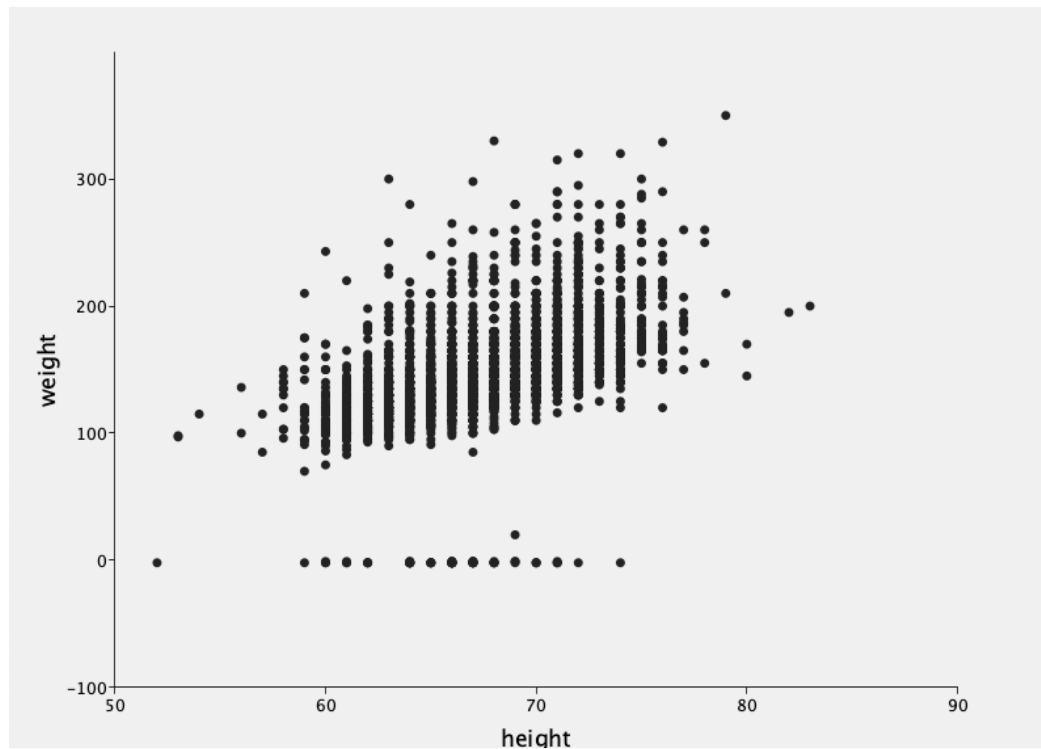
It does seem reasonable to presume that height and weight have a linear relationship.

Let's begin setting up our model a little more formally.