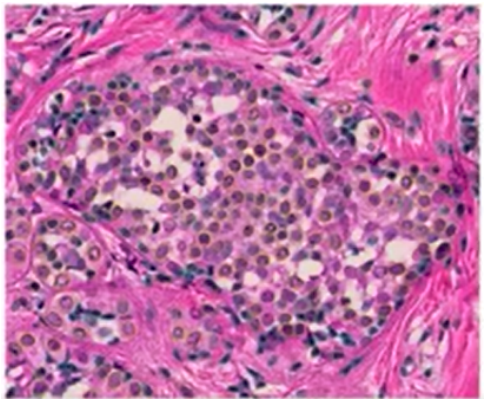
Introduction to Principal Component Analysis

In this lecture, we will start talking about the process after you have collected large-scale datasets - what can you do with them, and how do you even start thinking about data analysis?
Here, we will be presenting only one particular application, but you will also get an idea about how this relates to different kinds of industries as well, so that you can relate it to the applications that you care about in your work.

Let's take this example from the healthcare domain.
Assume we have a very large scale and a lot of microscopy images. The image here shows a microscopy image where you have many different cells, and we may want to identify in this image which cells might be abnormal cells.



So the question here is,
How can we visualize this dataset to find clusters or abnormal cells?
Here we have,
Input : x_1 ,..., $x_n \in R^D$
Output: y_1 ,..., $y_n \in R^d$, where d << D

So these are the microscopy images of human tissue slices that we are dealing with.
And the first thing we can do is crop out these cells (n cells) and summarize each cell by 100 different texture features (i.e, D = 100).
So every cell is represented as a high dimensional vector based on how many pixels you have and how many color channels you have.

Say every cell has like 10,000 pixels, then that means every cell here is represented as a point in 10,000-dimensional space. Now we humans are very bad at looking at very high dimensional spaces, we can look at maybe two or three-dimensional spaces. So if we need to get a first impression of what the data set actually includes, then we would like to represent these many many cells (where say we have 100,000 cells and every cell lives in a 10,000-dimensional space) into a **lower dimension**.

For example, we would like to represent it in two or three dimensions and by doing so we may see some clusters. There maybe we could find some groups of cells that are actually clustered together and maybe some of these clusters correspond to abnormal cells, maybe we can get some cancerous cells or maybe some of them might be abnormal in terms of going more towards the cancer state, etc.

Similarly, if you're thinking of other industries, for example, say you have a whole lot of data on your customers that you even have time-series data on who bought what and in terms of their buying patterns. So those could be long time series where you have like 10,000 different variables for every one of the customers. For example let's say you have 100,000 customers so that you have a data set of 100,000 samples and every sample (every customer) lives in 10,000-dimensional space, and we cannot look at 10,000-dimensional spaces. We would like to be able to represent this data in two or three dimensions so that we can actually look at it and see whether we can identify different groups of customers so that we can then maybe look and analyze a bit more carefully about what are these different clusters? What is it that makes these different people clustered together? What is it in their buying behavior that makes them clustered together? And maybe then we can target our treatment, for example, advertising, different kinds of processes that we want to use with different customers, etc. But we're really caring about actually identifying these different clusters. So first step, we just want to be able to visualize this data in two or three dimensions.

So now we will discuss two methods about how we can do this.

Principal Component Analysis (PCA): Projection that spreads data as much as possible.
Stochastic neighbor embedding(t-SNE): Non-linear embedding that tries to keep close-by points close.

Now, PCA is a very simple method, it's very, very fast, and that's the advantage. So it can be applied very quickly, it should always be one of the first things that you should do with your data.

And t-SNE is much slower but you will see the advantages of this method in order to be able to identify clusters.

t-SNE is one of the newer methods to have come out and is used a lot for finding clusters, and PCA is one of the standard methods that has a long history but is very, very fast.

PCA is a linear method, so it's very fast. And t-SNE is a non-linear function. But you will see how much better t-SNE is in general, in finding clusters or identifying them.

The previous version was only called SNE, and the newer version has a t in front and it's called t-SNE. We will also see the advantages and disadvantages of these two methods.

Let's talk about PCA first, because it's a very, very standard method and it's really something that everyone should do as a first thing when they're getting a dataset to just represent and see whether there are outliers in the space and to just find some general patterns in your dataset.

Now, what is the goal of PCA or Principal Component Analysis?
Goal: Dimensionality reduction to a few dimensions.
Intuition: Find the low-dimensional projection with the largest spread.

For example, say you have a dataset of your customers and every customer is a vector in 10,000-dimensional space and you would like to represent each one of your customers in two or three dimensions so that you can look at it.
So there are a couple of different ways of thinking about what PCA does. So first of all, it's a linear method. So it just projects the data from 10,000-dimensional space into whatever dimension you choose. It may be one dimensional, two dimensional, or three dimensional. So, the dimensions should be few so that you can still look at them.

And now the question is, what does it try to achieve when it tries to find such a linear embedding of such a projection?
There are different ways of seeing what it tries to do, and they're all equivalent.
So one way of trying to see what it does is that it actually tries to find the projection so that the data is still as spread out as possible.
Now, why is this important i.e why would you want to capture the directions of the largest variation in your dataset?
So think again about your customer base and think about the fact that maybe you're measuring a few variables. Age group could be one of the variables, and maybe another variable is the country they're living in.

Let us say, you are a company that is based in the US and in your customer pool, basically everyone is in the US. Then for that variable of where these people are living, there is nothing changing in that variable, since everyone is in the US, so this variable really doesn't tell you anything. If you still see differences in your customers in terms of their buying behavior still that variable cannot explain it since it doesn't change at all.

So these variables have very little variation in your data set. They don't matter as they don't tell you a lot about the dataset. So that's the intuition for removing all these variables that have very little variation, and only keeping the ones that actually have differences in the samples that you're looking at.

Also in the microscopy image that we saw previously, we don't have to keep the pixels that are always the same in every one of the cells. We want to concentrate on the pixels that are actually different in different cells because these can tell us something about whether a cell is indicative of being abnormal versus normal.

So that's the intuition for trying to find a projection that preserves the variables or the directions that show the biggest amount of variation.

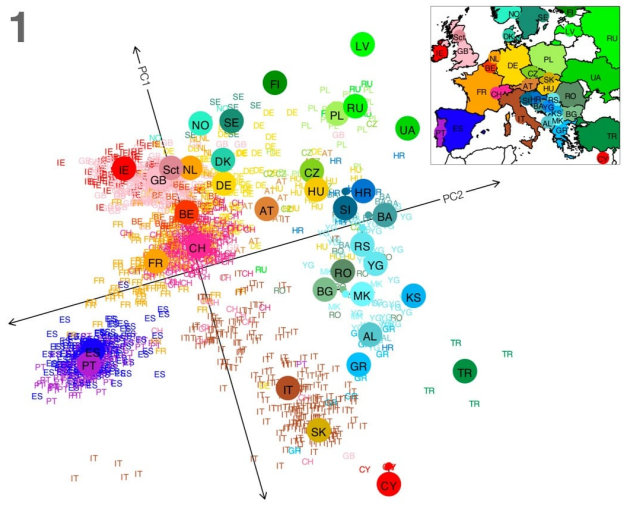
Let's look at an application of PCA :

There are different kinds of companies (for e.g. 23andMe is one such company) that offer to actually sequence your DNA. And so, you can get all the snippets of the values at all of the positions that differ between different people. So there are many sequence structures that are the same between you and me. They justify that we are humans, and they will all be the same for all of us. But, these positions don't matter much. They don't contain any information about our ancestry or about any of our phenotypic values. So what we really want to measure are the positions that actually change between different people. Those are the ones that carry information. So those are the ones that are usually sequenced.

And what they did in the experiment is they took people that got sequenced in different parts of Europe. So, they got a very very long 100,000-dimensional vector for every person of what is the value at that particular location in the DNA. And all they did is do a Principal Component Analysis.

You can see in the figure below the vertical axis **PC1 (Principal Component 1)** and the horizontal axis **PC2 (Principal Component 2).**
So this is just a projection along with the two directions that vary the most.

1



What we see is really quite amazing, where basically every person was mapped to the location they are coming from. The map that they get here represents quite well the map of Europe.

So without putting in any information about where this person comes from, all that was done here is to take DNA from every person, and then a Principal Component Analysis was done, and that outcome is basically the map of Europe. So just PCA, which is something so simple, was able to recreate and actually map every one of the dots back to where they're coming from just by choosing two directions in the data set which maximize the spread.