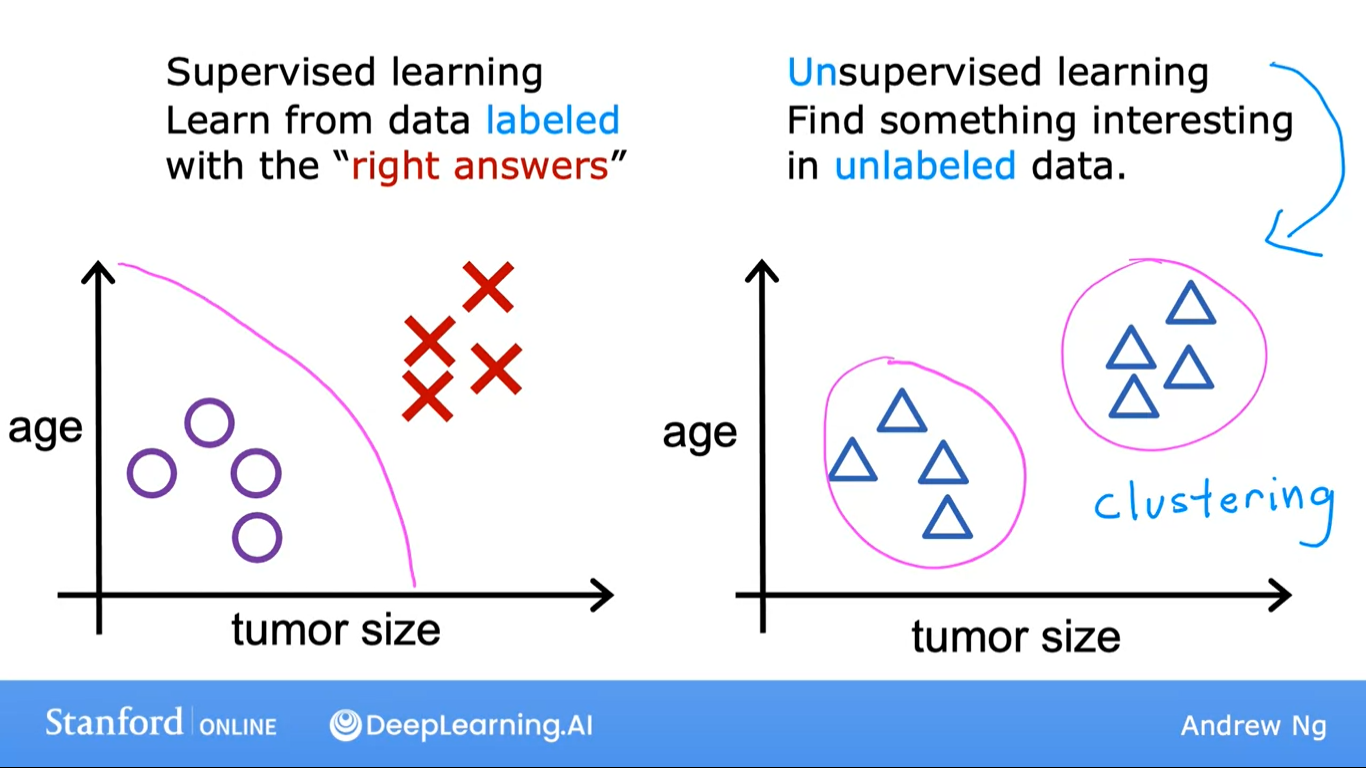
After supervised learning, the most widely used form of machine learning is unsupervised learning. Below is the comparison of supervised vs unsupervised learning.



In Supervised learning, each example was associated with an output label y such as benign or malignant, designated by the poles and crosses. In Unsupervised learning, we are given data that isn’t associated with any output label y, say you’re given data on patients and their tumor size and the patient’s age, but not whether the tumor was benign or malignant.

We’re not asked to diagnose whether the tumor is benign or malignant, because we’re not given any labels y in the dataset, instead our job is to find some structure or some pattern or just find something interesting in the data. This is unsupervised learning, and we call it unsupervised because we’re not trying to supervise the algorithm to give the right answer for every input.

Instead, we asked the algorithm to figure out all by yourself what’s interesting or what patterns or structures that might be in this data. With this particular dataset, an unsupervised learning algorithm might decide that the data can be assigned to two different groups or two different clusters. So, it might decide that there’s one cluster or group against each other.

This is a particular type of unsupervised learning, called a clustering algorithm because it places the unlabeled data into different clusters, and this turns out to be used in many applications. For example, clustering is used in Google news. Every day, what Google news does is everyday it goes at hundreds of thousands of news articles on the internet and groups related stories together.

For example, in the sample below from Google news, where the headline of the top article is, “Giant panda gives birth to rare twin clubs at Japan’s oldest zoo”, and looking at this you might notice that below it is other related articles. From the headline alone, you can start to guess what clustering might be doing.



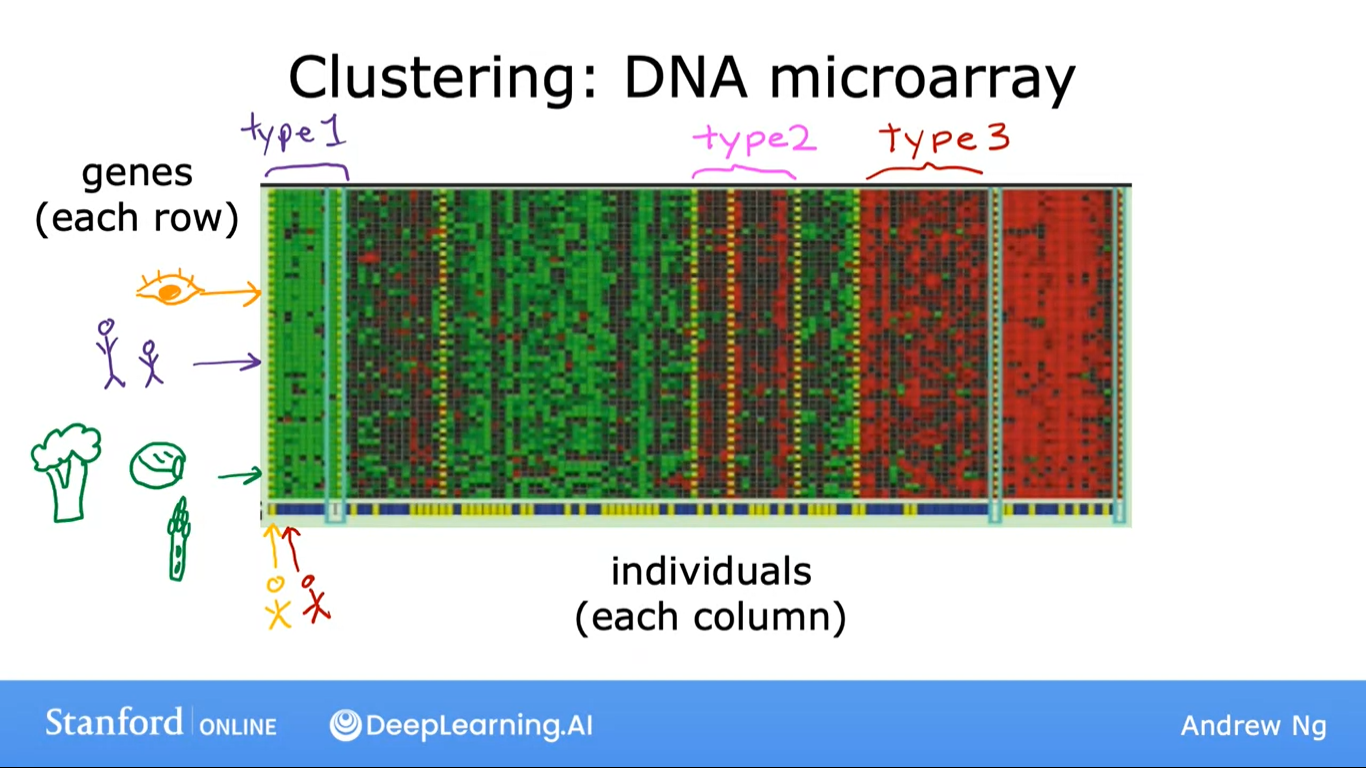
You can see the word Panda, Twin and Zoo appear in all of the articles. So, the clustering algorithm is finding articles, all of hundreds and thousands of new articles on the internet that day, finding the articles that mention similar words and grouping them into clusters. This clustering algorithm figures out on his own, which words suggest that certain articles are in the same group.

There isn’t an employee at Google news who’s telling the algorithm to find articles that have the word Panda, Twins or Zoo to put them into the same cluster. The news topics change every day and there are so many news stories, it just isn’t feasible for people to do this every single day for all the topics that use covers. Instead, the algorithm has to figure out on its own without supervision what are the clusters of news articles today. That’s why clustering algorithm is a type of unsupervised algorithm.

The second example of unsupervised learning is of clustering genetic or DNA data. The image below shows a picture of DNA micro array data, these look like tiny grids of a spreadsheet and each tiny column represents the genetic or DNA activity of one person. For example, the entire column is of one person’s DNA and the other column is of another person. Each row represents a particular gene. Therefore, one row may affect eye color and the other row may affect the height of a person.

Researchers have even found a genetic link to whether someone dislikes certain vegetables, such as broccoli or brussels sprouts. For DNA micro race, the idea is to measure how many certain genes are expressed for each individual person. So, these colors red, green, gray, and so on, show the degree to which different individuals do or do not have a specific gene active.

What you can do is then run a clustering algorithm to group individuals into different categories or different types of people like type 1, type 2 and so on. This is unsupervised learning because we’re not telling the algorithm in advance, that there is type 1 person with certain characteristics. Instead, what we are saying is here’s a bunch of data, I don’t know what the different types of people are, but can you automatically find structure into data and automatically figure out the major types of individuals.



Here's the third example, many companies have huge databases of customer information. Given this data can you automatically group your customers into different market segments so that you can more efficiently serve your customers?

Running clustering that is market segmentation found a few distinct groups of individuals, one group’s primary motivation is seeking knowledge to grow their skills, a second group’s primary motivation is looking for a way to develop their career, another group wants to stay updated on how AI impacts their field of work. This is the clustering that team used to try to better serve the community as we’re trying to figure out what the major categories of learners in the deeper AI community.

