Machine Learning

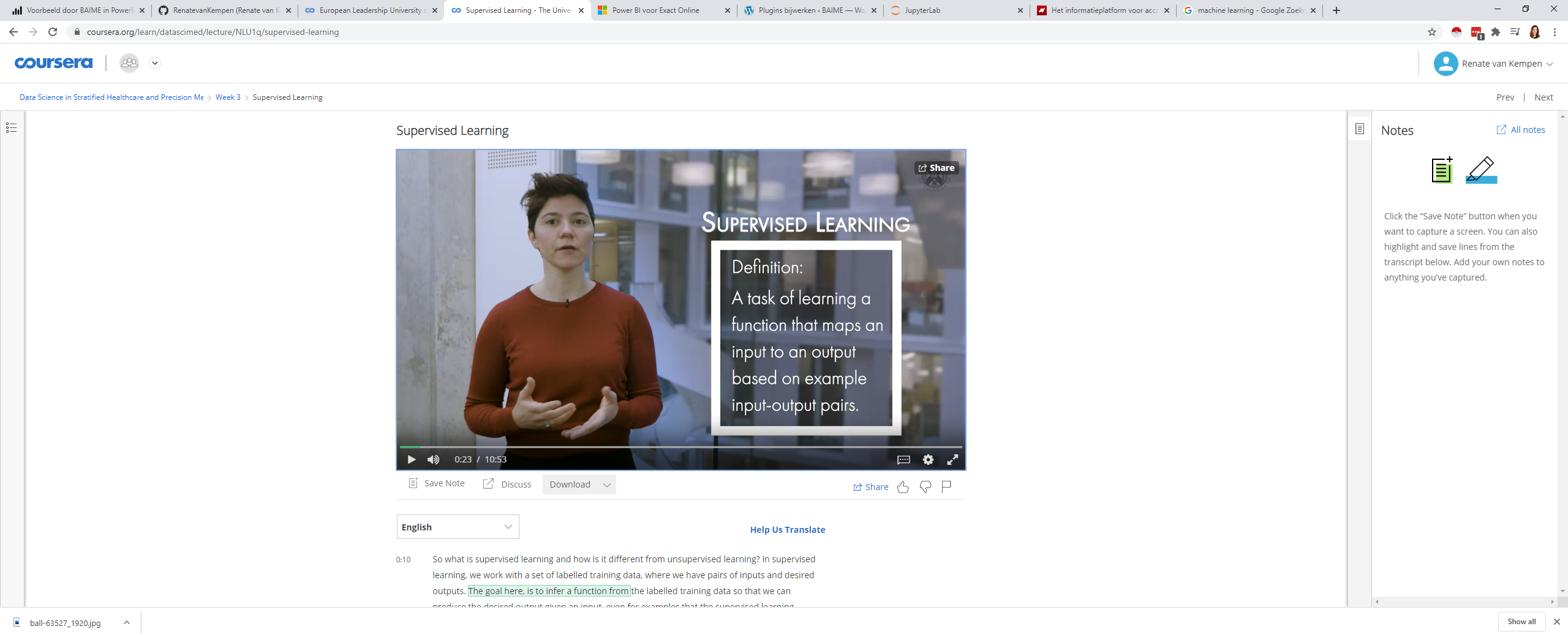
Machine learning is a subfield of Artificial Intelligence

Machine learning algorithms allow computers to identify patterns in data,

build models that explain the world, and make predictions without having explicit pre-programmed rules and models

Machine Learning also enables us to raise new questions of great significance.

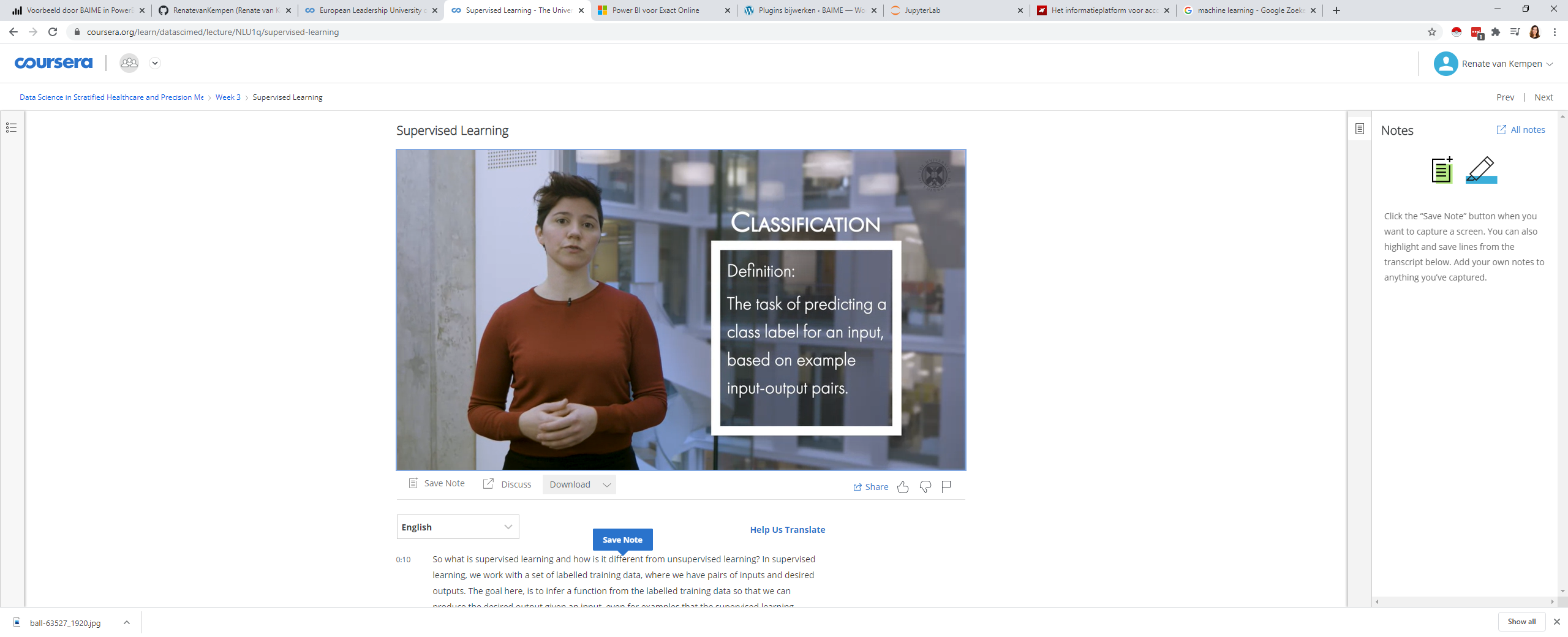
# Supervised learning



In supervised learning, we work with a set of labelled training data, where we have pairs of inputs and desired outputs. The goal here, is to infer a function from the labelled training data so that we can produce the desired output given an input, even for examples that the supervised learning algorithm has never seen before.

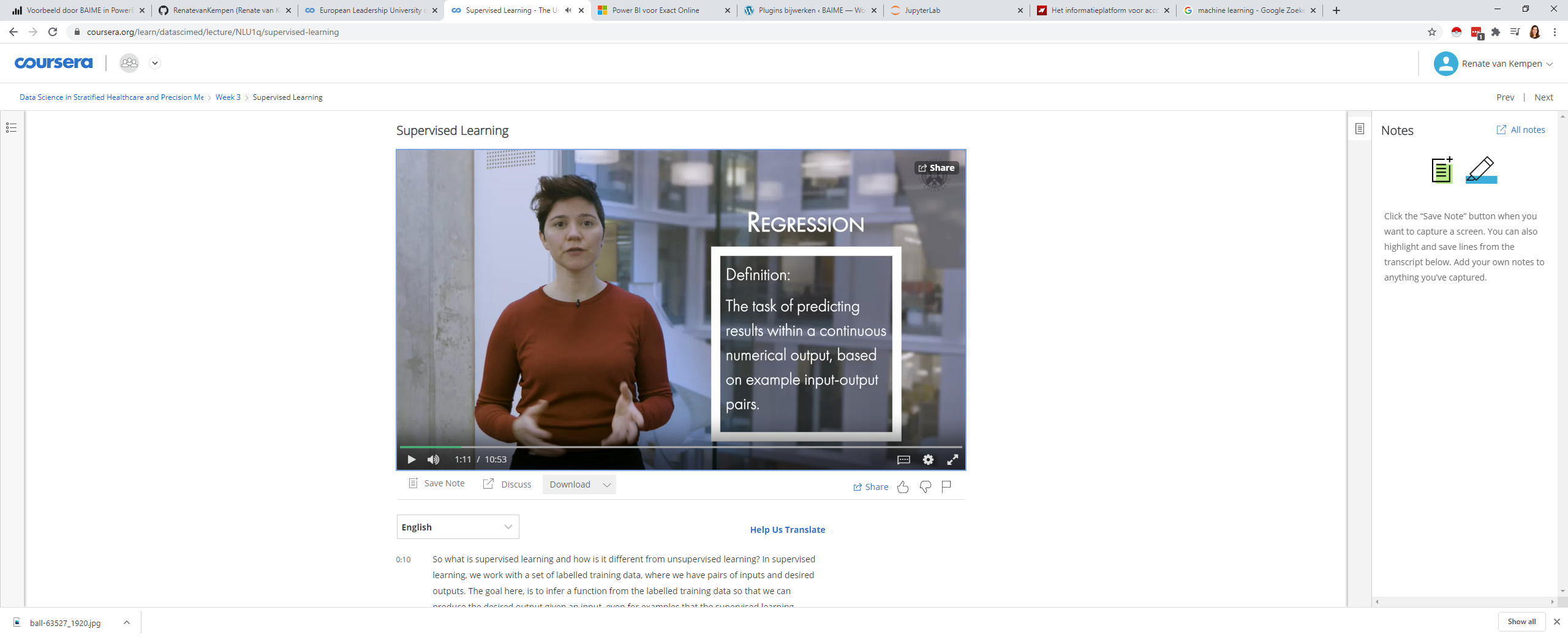
## Types

### Classification



In other words, we use our training data to build a model that assigns unseen observations to one or more classes.

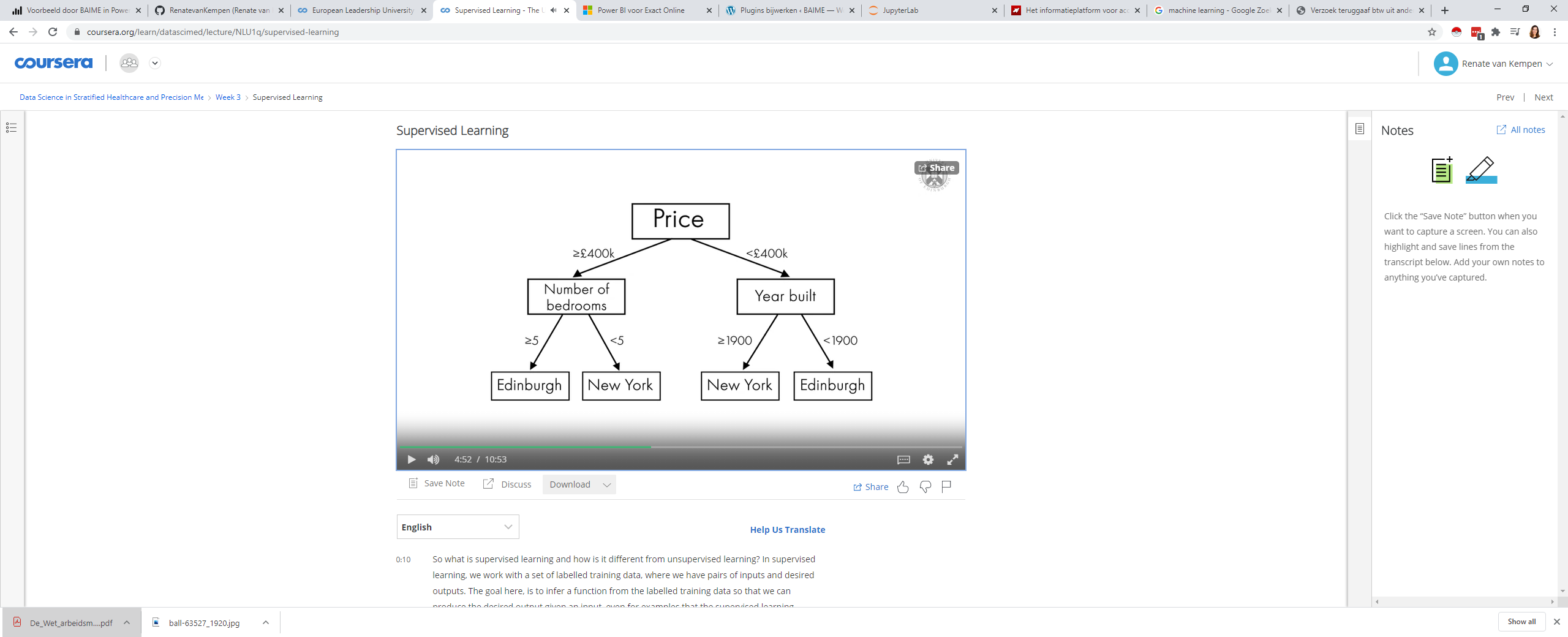
### Regression



we predict result within a continuous numerical output. Continuous means that there aren't any gaps in the value that the output can take

## TEchniques

* Decision trees



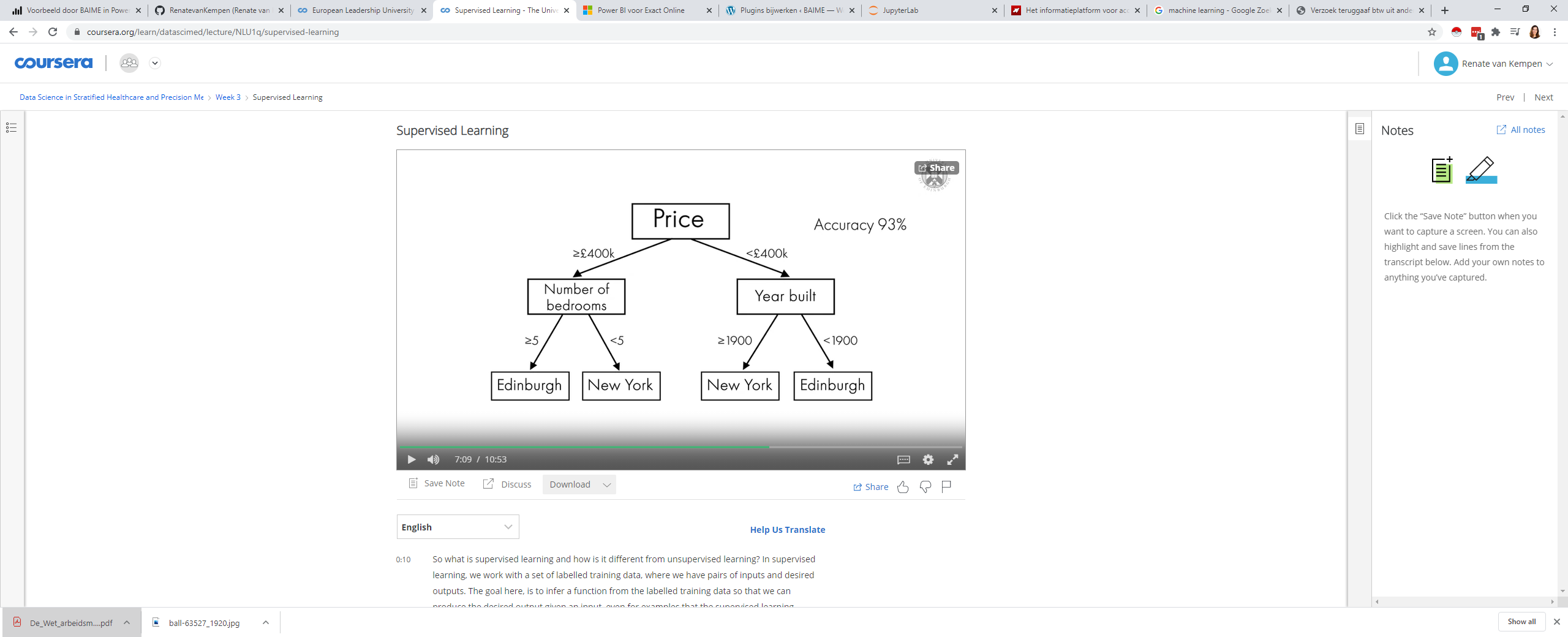
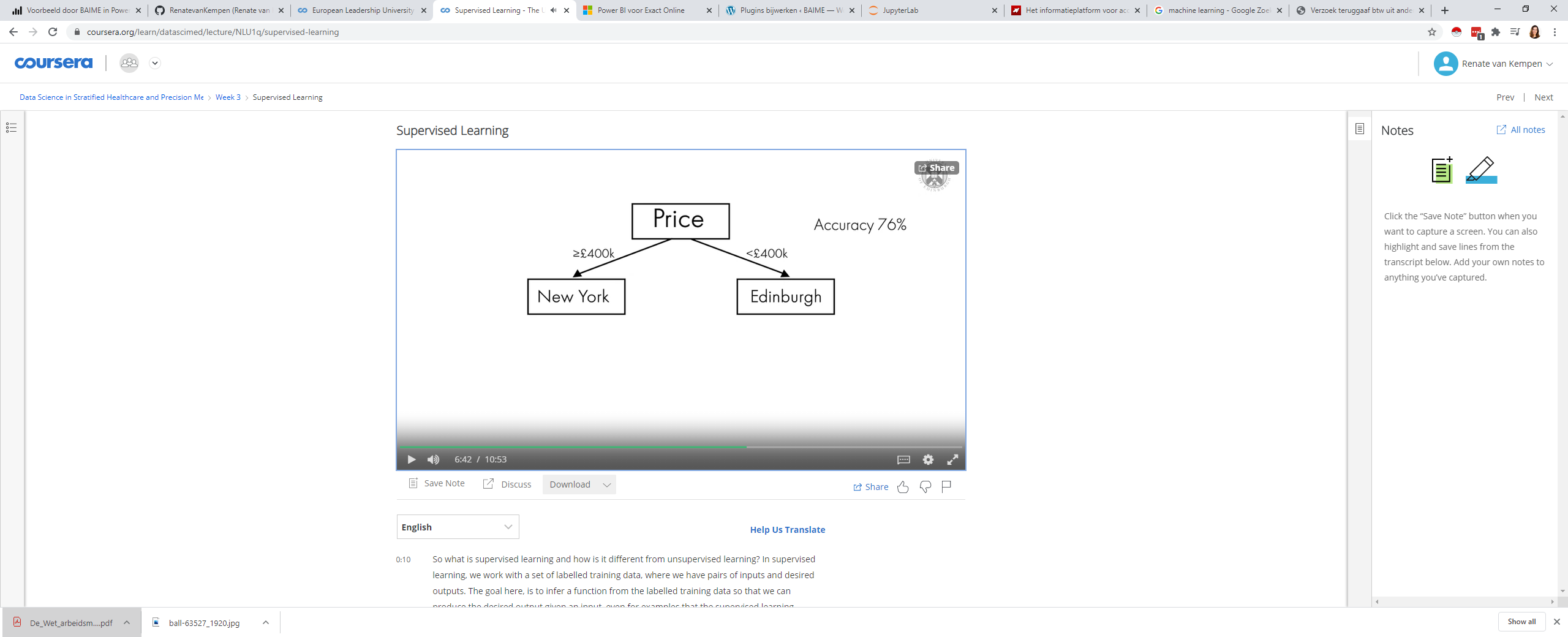
Learning a decision tree is about learning the sequence of questions.

In other words, the sequence of variables used and their split of

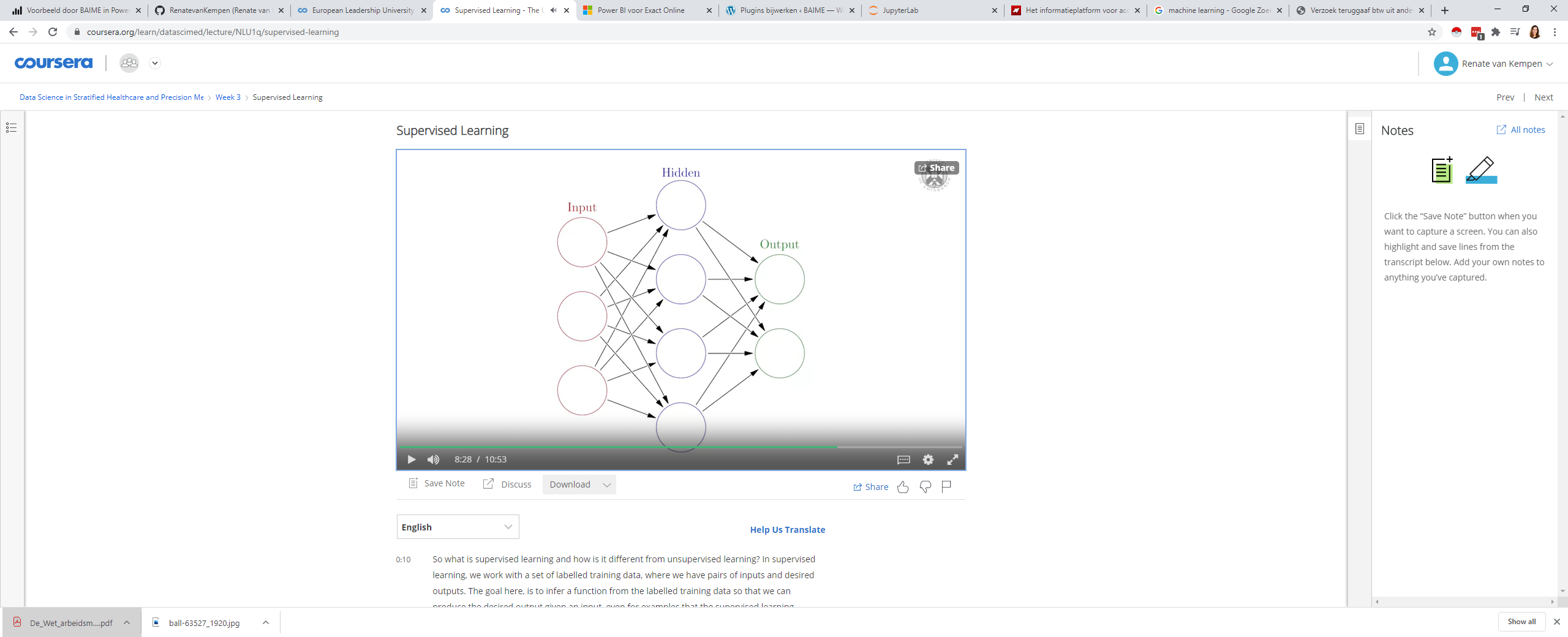
possible values that gets us to the correct answer more quickly

To learn a tree, the algorithm searches over all possible attributes or

questions, and picks the one that is most informative about the target variable



* Random forests
* Neural networks



A neural network is an interconnected group of nodes, called neurons,

which are captured here as circles.

The arrows here, represent connections between the neurons,

which can transmit a signal from one neuron to another.

The neuron that receives the signal can

process it and then signal neurons that are connected to it.

Neural networks are typically organized in layers.

There is an input and an output layer and then maybe one or more hidden layers

A neuron calculates a weighted sum of its inputs and then applies

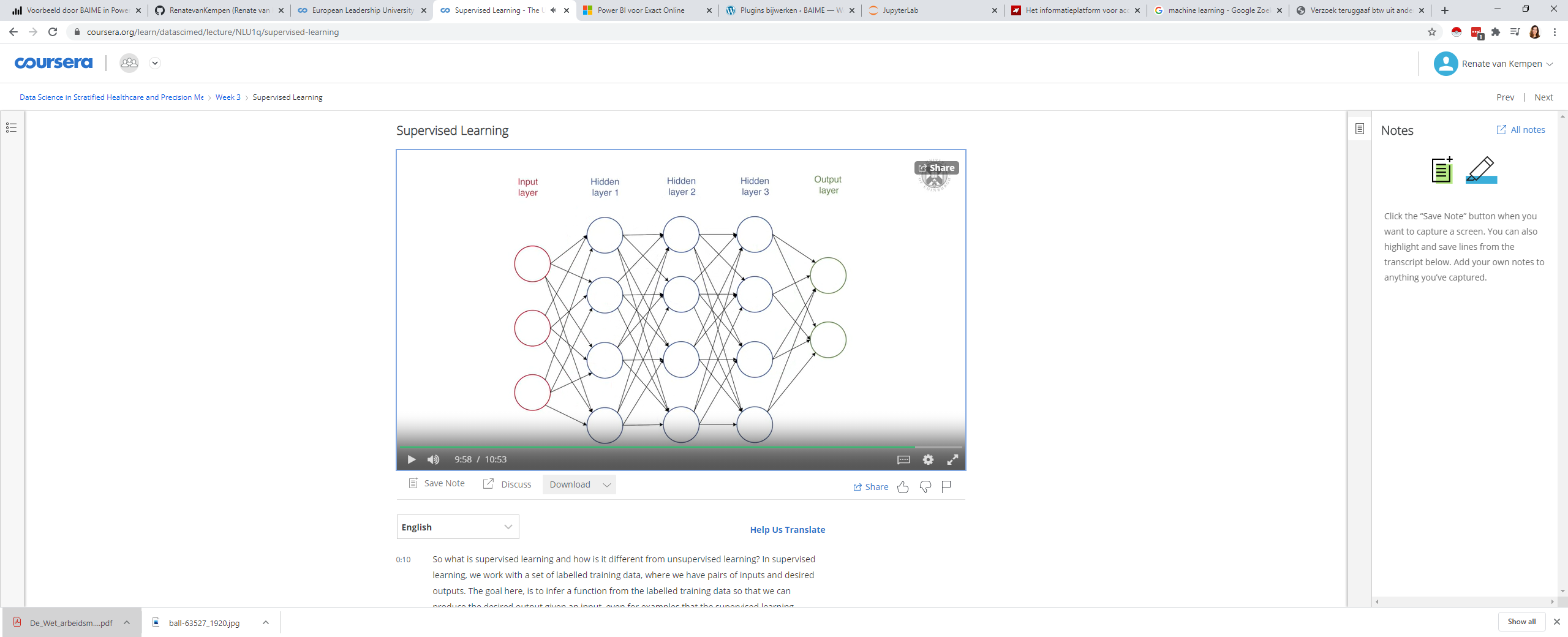
a threshold or activation function to determine whether the signal will be sent or not.

In a neural network, this process is repeated several times,

first computing hidden units as an intermediate processing step,

which are again combined to yield the final result

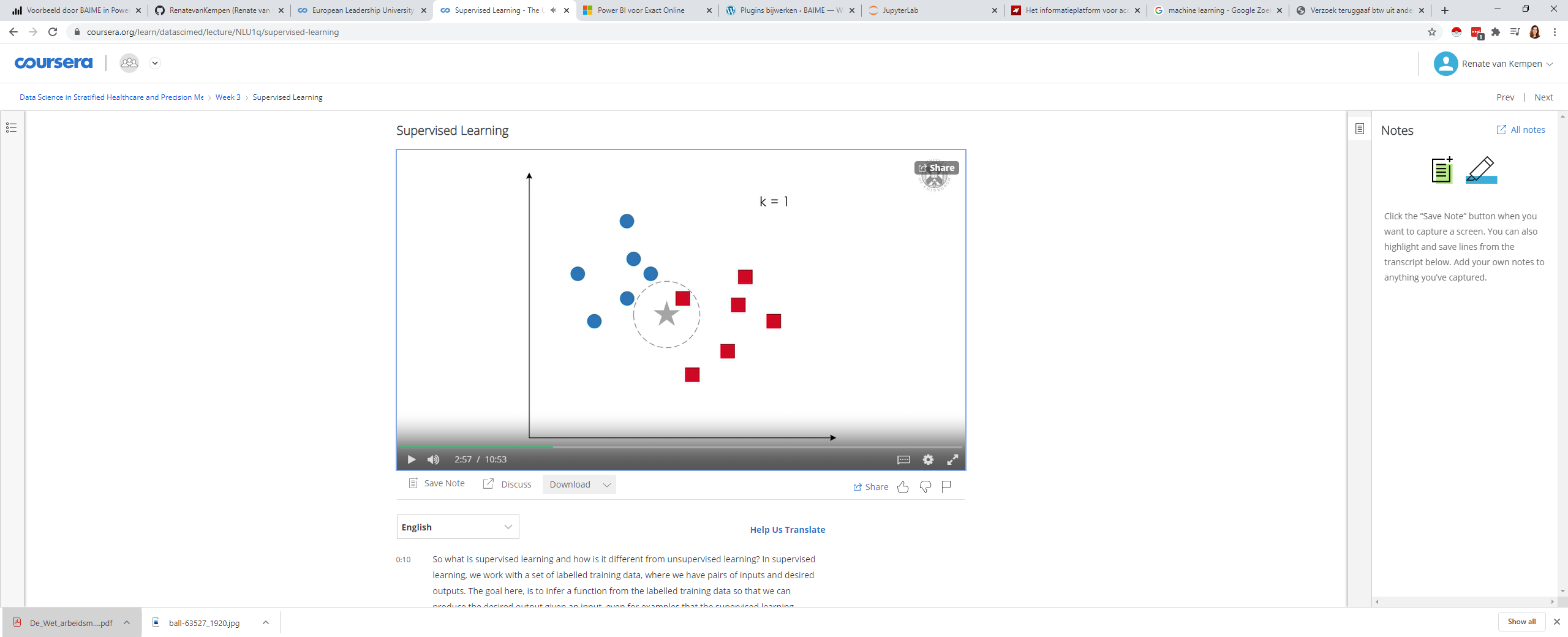
We should note that an important parameter that needs to be set by the user, is the number of nodes in the hidden layer. It is also possible to add more layers, as shown here. Having large neural networks that consist of many of these layers is referred to as deep learning

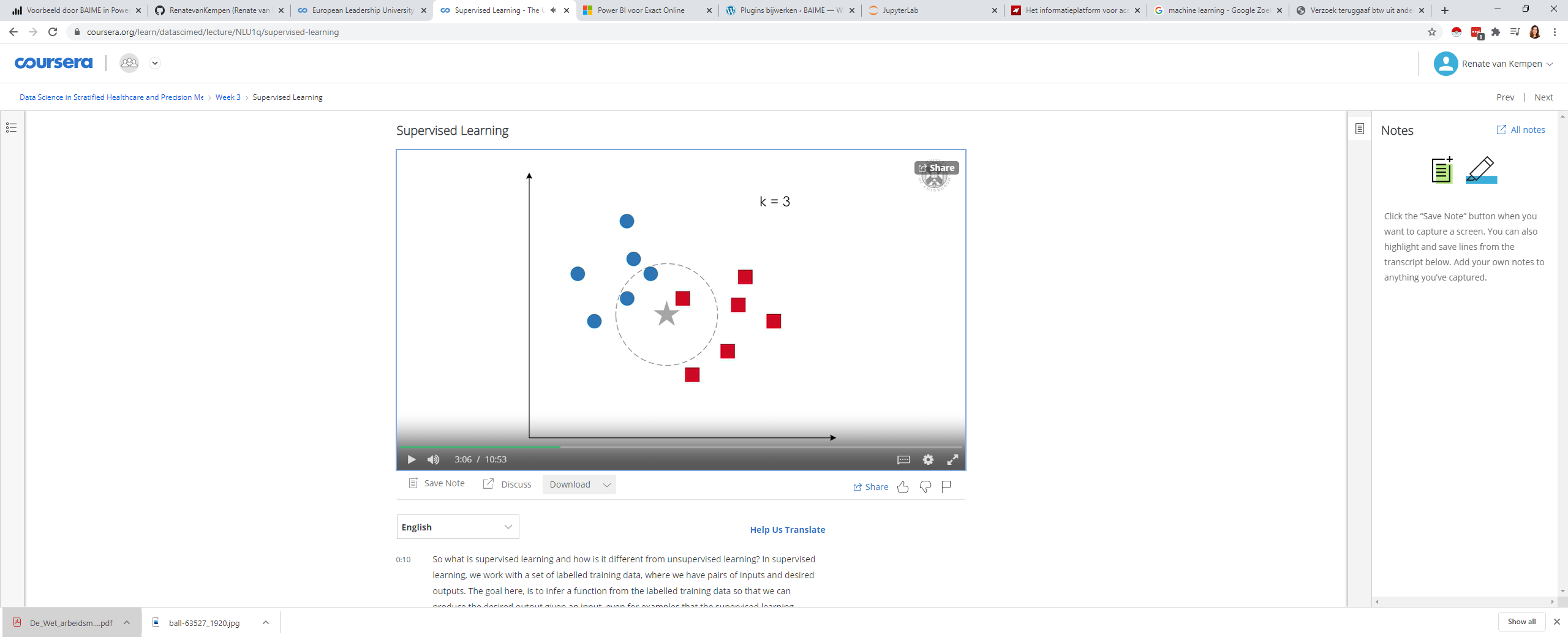


Neural networks and deep learning are extremely powerful. They can deal with large amounts of data and build incredibly complex models. This however comes at the cost of interpretability.

As the more hidden layers are added, the harder it is to understand what is happening inside the neural network. Training large and powerful neural networks can also be a challenging task.

* Support vector machines
* K-Nearest neighbors (simplest) 🡪 mode for classification and mean for regression,

 🡪 outcome a red square

 🡪 outcome a blue circle

* linear regression
* logistic regression

### Distance metric

A commonly used distance metric for continuous variables is euclidean distance,

while for discrete variables,

the Hamming distance is used

### Overfitting

Go as far to introduce further attributes until the accuracy is 100 percent.

However, this is not recommended, as our extended decision tree would then not perform well on previously unseen test data. This would be due to overfitting to the training data, a phenomenon that may occur in supervised learning, and which reduces the ability of the model to generalize from the training to the test data.

# Unsupervised learning

The data set is

unlabelled and hence there is no right or wrong answer.

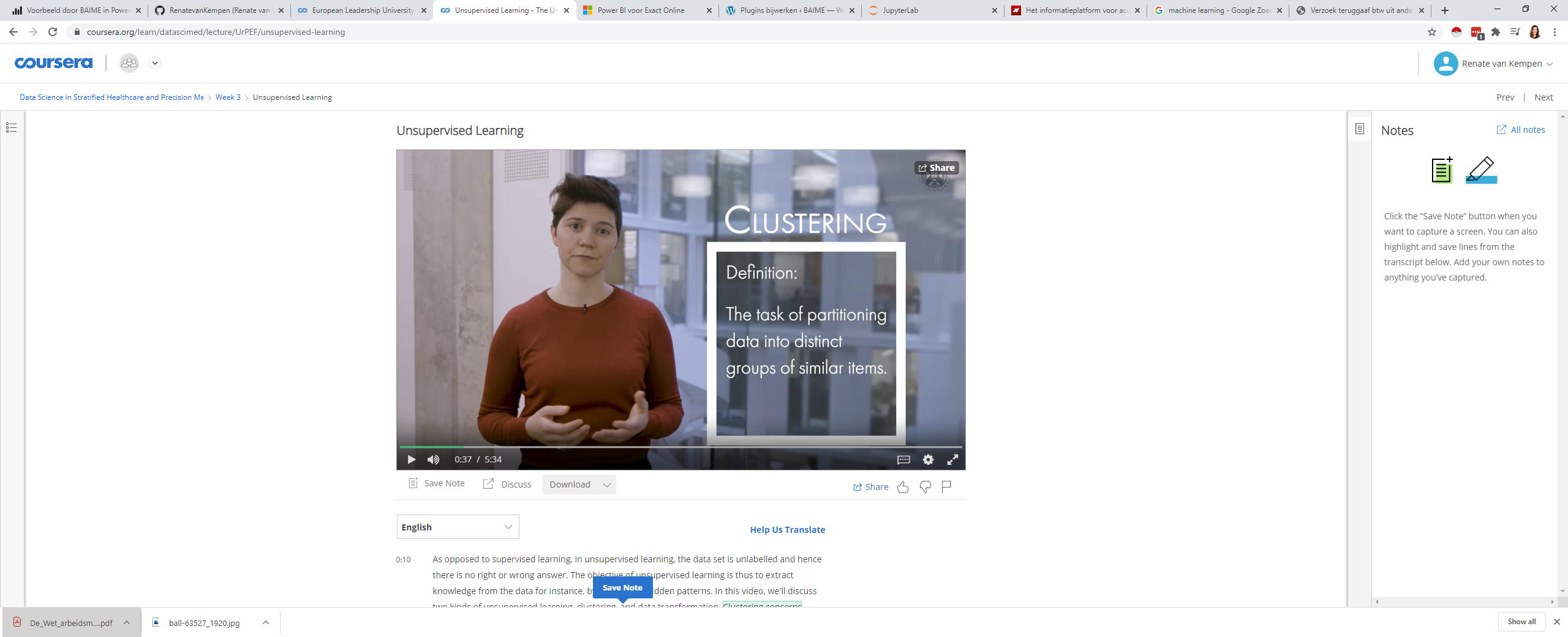
The objective of unsupervised learning is thus

to extract knowledge from the data for instance,

by discovering hidden patterns

## Types

### Clustering



Clustering concerns partitioning data into groups called clusters.

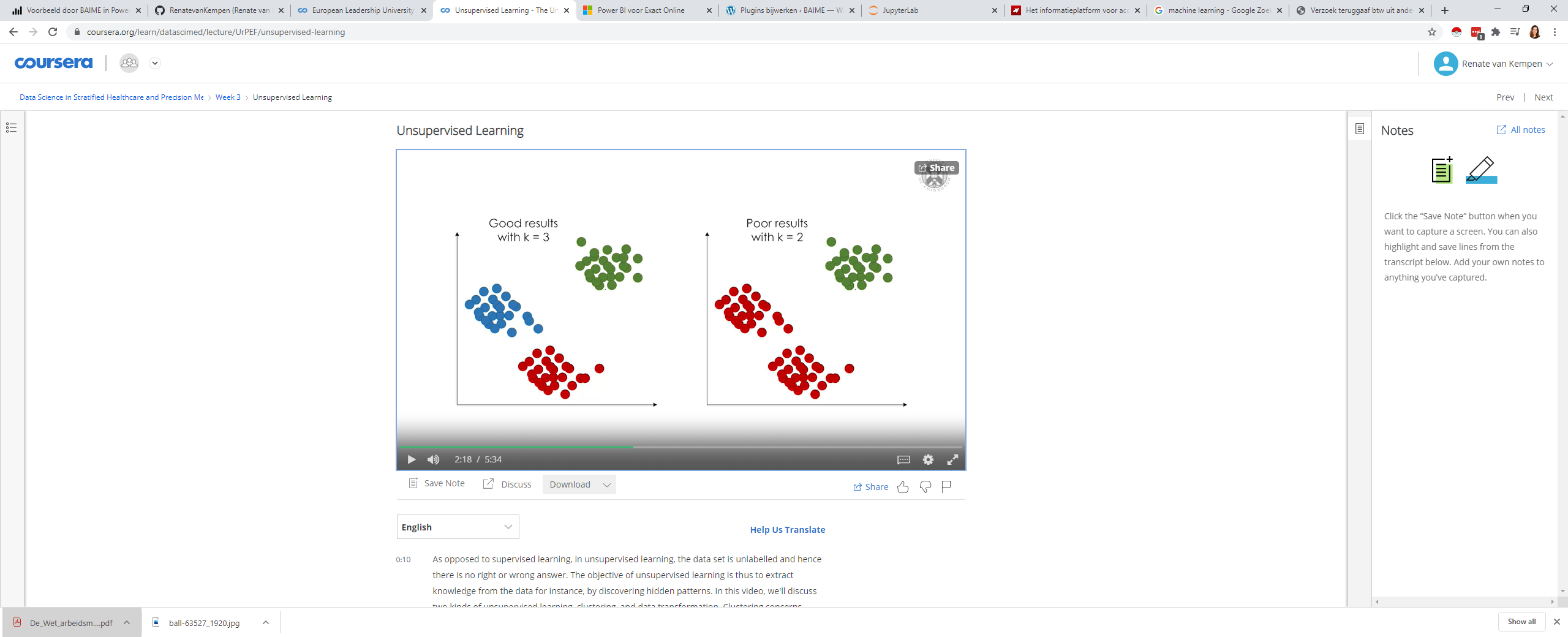
### Data transformation

We may carry out unsupervised learning so as to transform a data set. For instance, to compress the data and find a representation that is more informative for further data processing. This is commonly referred to as dimensionality reduction and feature extraction.

## Techniques

* K-means clustering (clustering)

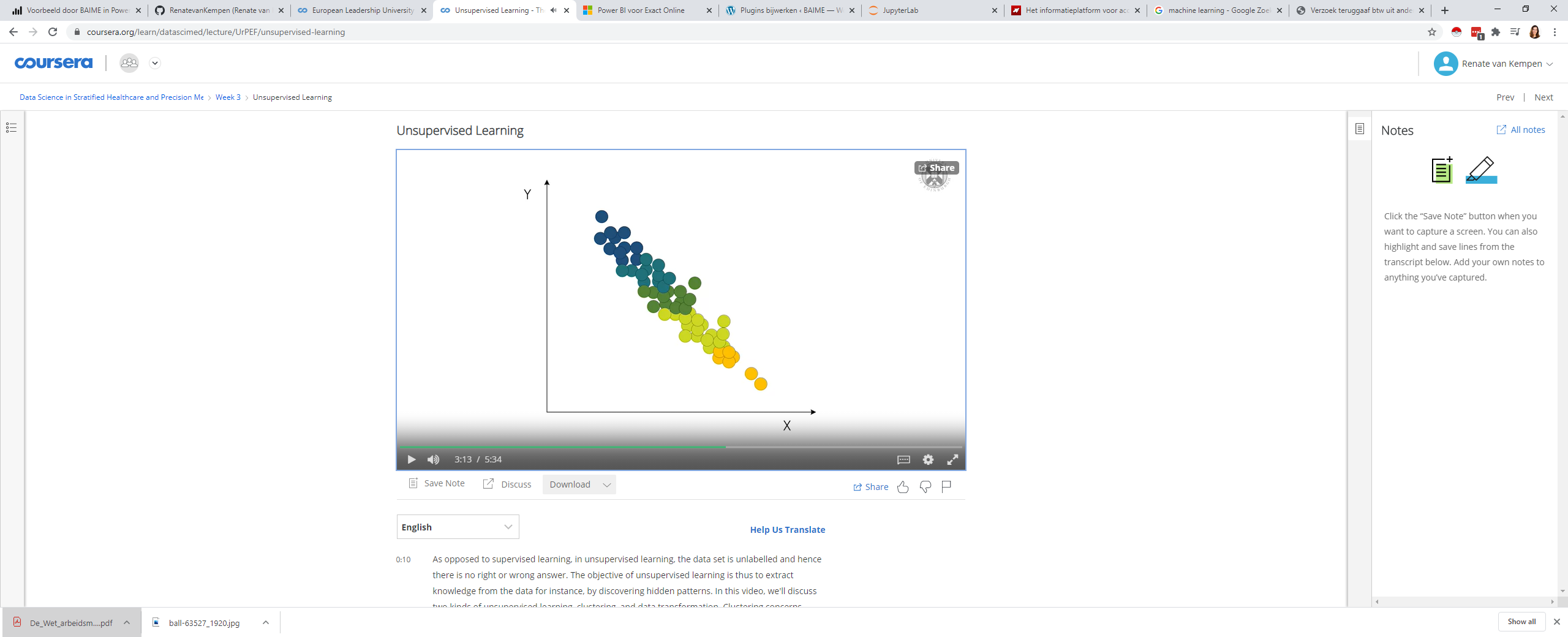
The main idea is to find cluster centers that are representative of certain regions of the data. This is done by alternating between two steps, assigning each data point to the closest cluster centre, and then setting each cluster centre as the mean of the data points that are assigned to it



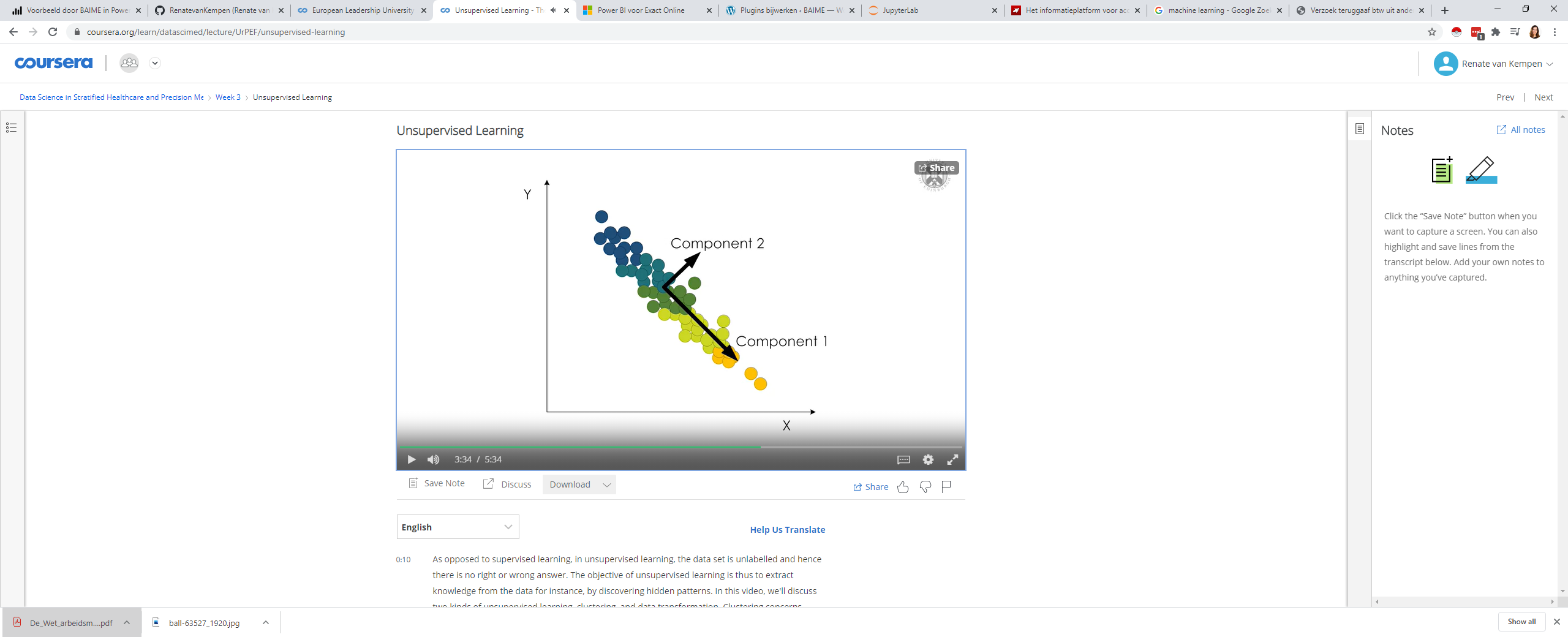
* Principal component analysis (PCA) for data transformation

A simple method that is commonly used for dimensionality reduction is Principal Component Analysis or PCA, as it's typically called. The goal here is to reduce the number of features, while retaining as much useful information as possible.

Original dataset:



The first thing that the algorithm does is find the direction of maximum variance. This is the direction of the data that contains most of the information, or in other words, the direction along which the features are most correlated with each other. We call this component one.



Next, the algorithm finds a direction orthogonal to

component one that has a maximum variance. This is component two.

Components one and two are the principal components of our data set,

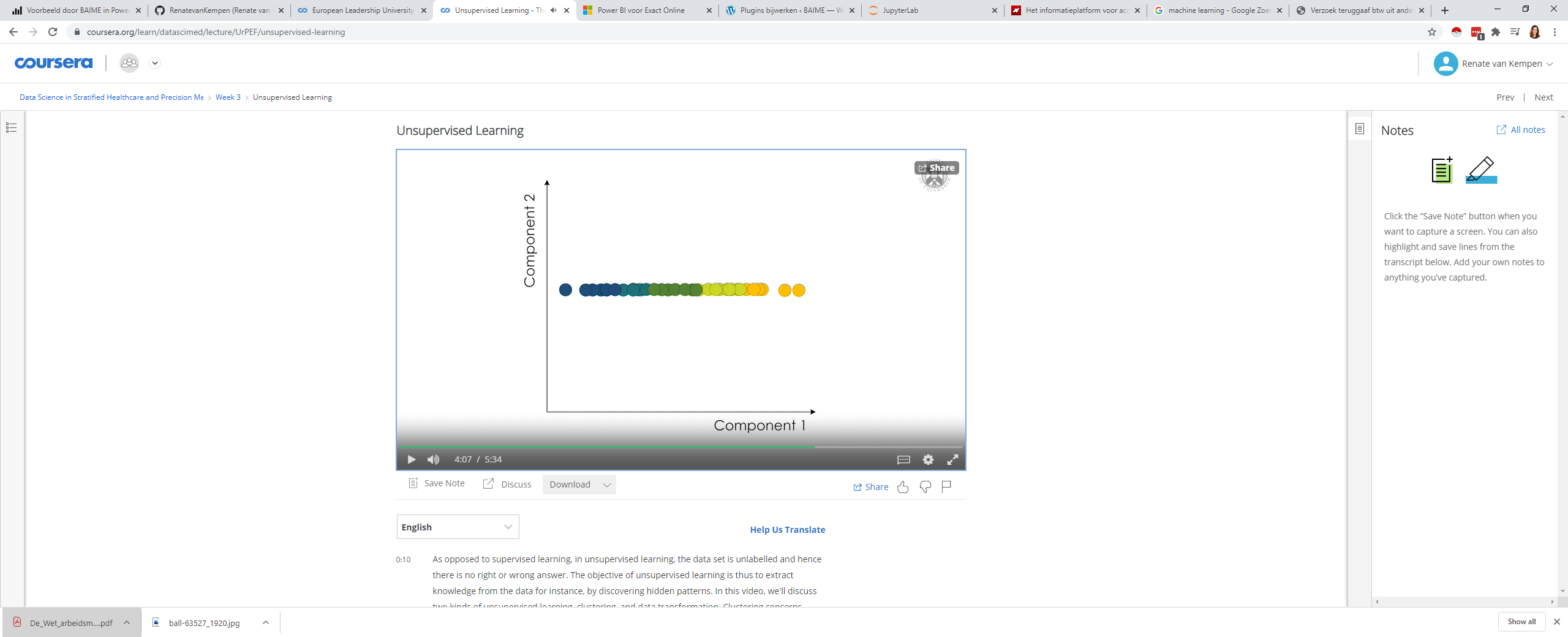
which are the main directions of variance in the data set.

We can see that if we return these dimensions,

were not going to lose any information, simply a rotation = transform



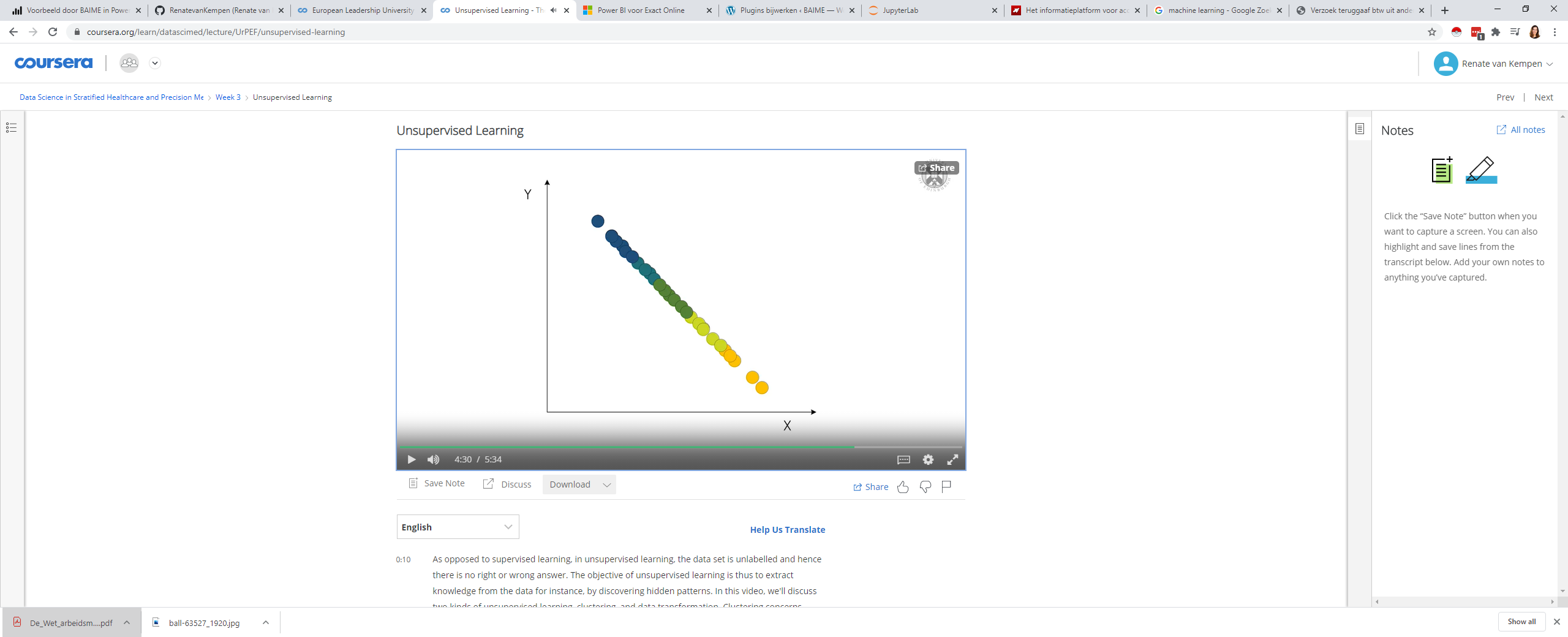
Dimensionality reduction



we can use PCA to carry out dimensionality reduction by dropping one, or more components that have the smallest maximum variance

In our two dimensional dataset this means retaining only component one. This is the reduction that gives the minimum error in the case that we decide to reconstruct the original dataset

Transformed dataset:



PCA allowed us to reduce the number of dimensions in our original dataset from

two, to one and get a transformed dataset that retains as much useful information as possible

PCA is commonly used to speed up supervised learning

## control

A big challenge here, is evaluating whether the algorithm learned something useful, or not.

So, performance here is often subjective and domain specific, therefore, unsupervised learning is often carried out in an explorative setting or as a preprocessing step for supervised learning

# conclusion

Effective machine learning is hard and finding good quality data is a big challenge. This is crucial. If you feed a machine learning algorithm with junk, all you're going to get back is junk.

So, it is important that we, as a community, work towards better data. Only then, can we make the most of what machine learning technologies can bring.