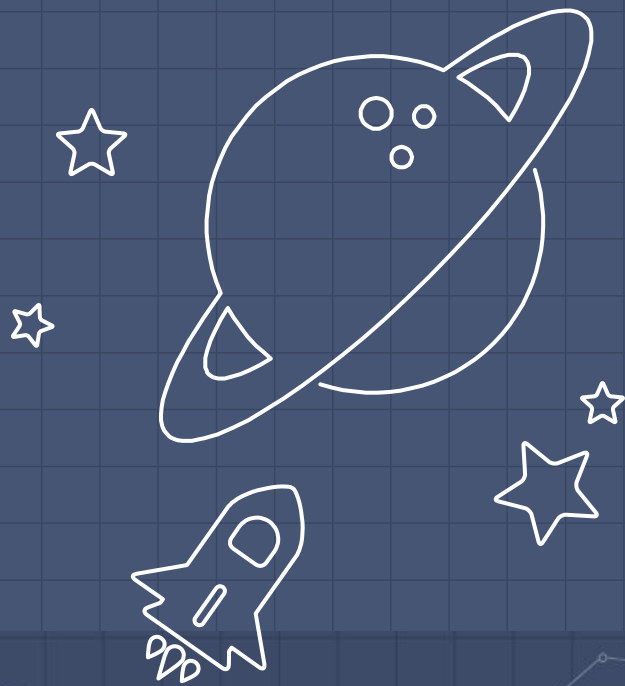
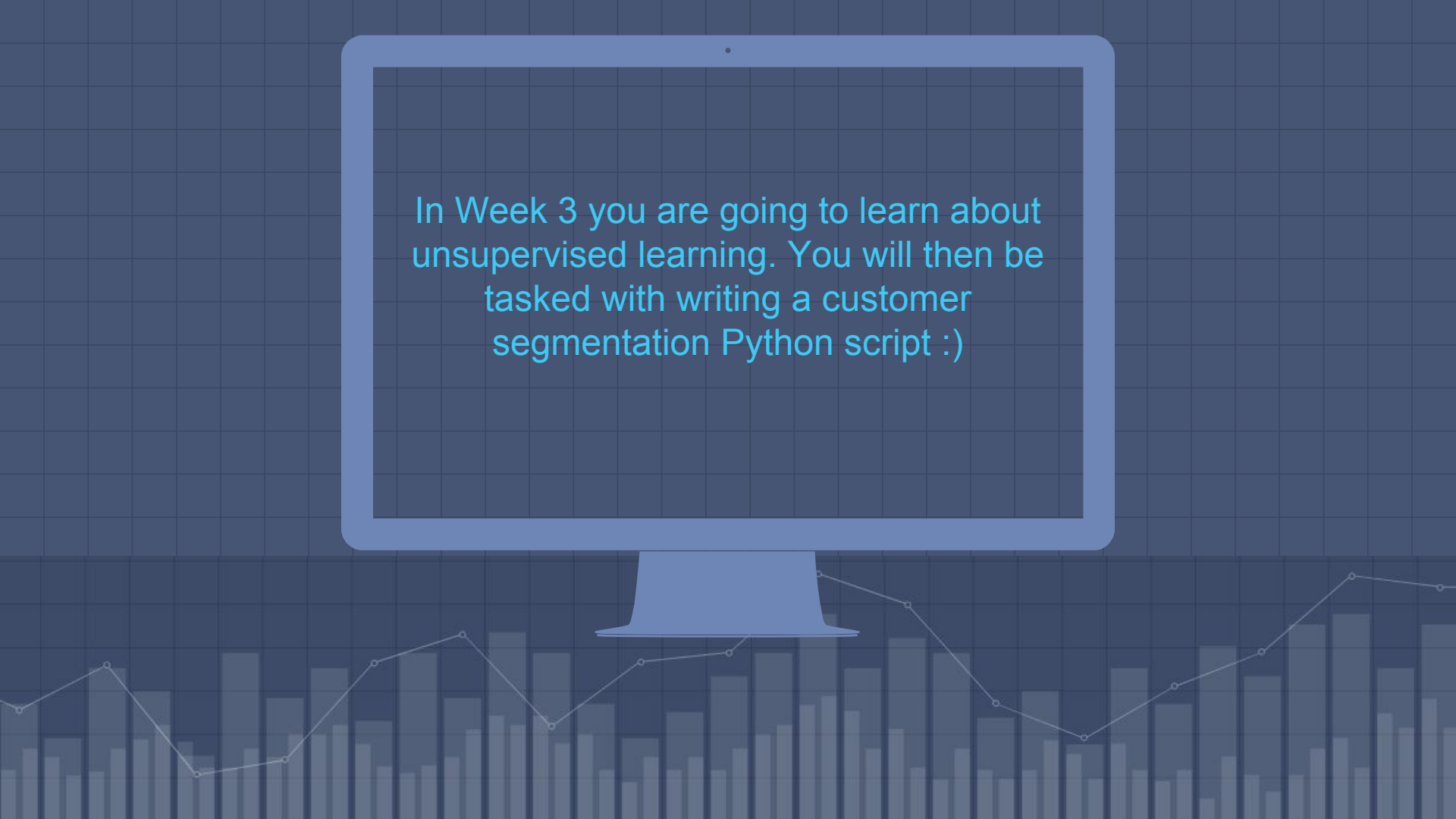


# arise:

## Unsupervised Learning

### Week 3





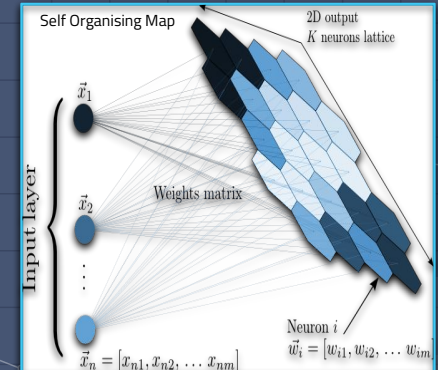
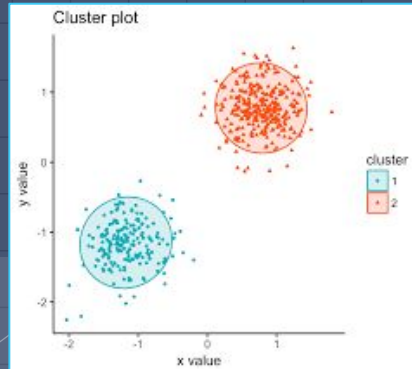
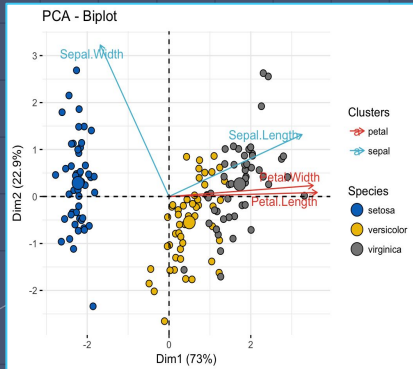
In Week 3 you are going to learn about unsupervised learning. You will then be tasked with writing a customer segmentation Python script :)

# Unsupervised Learning.



Unsupervised machine learning algorithms infer patterns within a dataset. Unlike supervised learning, unsupervised algorithms cannot be directly applied to a regression or a classification problem as we have no idea what the values for the output data might be, making it impossible for you to train the algorithm the way we “normally would”. Instead, we use unsupervised algorithms to detect the underlying structure of the data.

The most common unsupervised learning techniques are dimensionality reduction and clustering techniques. In more recent times unsupervised deep learning techniques such as self-organising maps and auto-encoders have been used to perform unsupervised analysis.



# Dimensionality Reduction



Since Big Data became prominent, we have had an ever growing access to features from which we can create models. This is both a good and a bad thing. Using a large number of features within a model can lead to:

- a model having a high variance, which in turn makes a model more prone to overfitting (this phenomenon is often referred to as the curse of dimensionality)
- a model requiring a long time to train and a large amount of processing power to compute.

Dimensionality Reduction is used to **reduce the number of features used within a model without losing information**. This is done to reduce the variance of a model, and to reduce the training and computing time of a model.

Humans can only see in 3 dimensions, therefore we would only ever be able to display the relationship between 3 variables at any one time. Another use of dimensionality reduction techniques is to visualise the entire variance contained within a dataset in only two or three dimensions.

The most common dimensionality reduction techniques are:

- Principal Components Analysis (read [here](#) for an explanation of the technique)
- Multidimensional Scaling (read [here](#) for an explanation of the technique)
- t-distributed Stochastic Neighbor Embedding (read [here](#) for an explanation of the technique)

# Dimensionality Reduction



We are going to run through a principal components analysis example. To do this please open the Jupyter notebook titled `pca.ipynb` in the week 3 folder and work through it. We are going to use the iris dataset for the PCA and clustering tutorials but we are going to load the dataset in two different ways.

# Clustering Techniques



The aim of clustering techniques is to group data observations into divided clusters that contain similar features. Clustering is often used for customer segmentation and to support product design.

We are going to run through three common clustering techniques; K-means, Hierarchical and DBSCAN. Please open the Jupyter notebook titled **`clustering.ipynb`** in the week 3 folder and work through it. The notebook contains links to articles and videos that will help with your understanding of the three algorithms.

# Deep Learning Techniques



Deep learning has many applications within unsupervised learning, but it is mainly being used to explore the features used in models, remove noise from models and to group observations. The most common deep learning tools used in unsupervised learning are autoencoders and self-organising maps (SOM). Autoencoders are used mainly to remove noise from features, while self-organising maps are used to cluster observations.

You can read more about autoencoders [here](#) and how to code them in Python [here](#). You can read more about self-organizing maps [here](#). Both types of models are types of neural networks. You will not be required to create an autoencoder or SOM in the program but the above articles are very interesting.

# Challenge



Your challenge this week is a tough one. You are being tasked with investigating a collection of clients that applied for a loan but were declined. You are to use a K-means clustering to identify clusters of clients within the collection. The data set is titled '**week3\_challenge.csv**' and can found in the week 3 folder.

You are to use a Jupyter notebook to develop your solution. Please title your notebook as 'yourEmailAddress.ipynb' and submit it to the Week 3 Answer Form [here](#).

The steps you will have to follow:

- Install the following packages on your machine and then import them into your notebook:
  - pandas
  - numpy
  - sklearn.preprocessing
  - PCA from sklearn.decomposition
  - KMeans from sklearn.cluster
  - matplotlib.pyplot
  - Seaborn



# Challenge



- Use **pandas** to **read** in the dataset and store it as a **pandas dataframe**. (hint: `read_csv()`).
- **Remove** columns from the dataframe that have **more than 30%** of their values **missing** (**answer** for number of columns removed is **required** in form).
- Create a **new dataframe** containing only the **numeric columns** (integers and floats) in the dataset (hint: `select_dtypes(include=[])`) (**answer** for number of numeric columns, remaining after removing the columns with too many missing values, is **required** in form).
- **Replace null** values in the numeric dataframe with 0.
- **Scale** the numeric dataframe.
- Perform a **PCA** reduction on the numeric dataframe using `sklearn.decomposition`'s PCA function. Only include the first **5 principal components** and **store** the results in a pandas **dataframe**. Print the total variance explained by the first **5 principal components** (**answer required** in form).
- Use **sklearn's KMeans** function to cluster the numeric PCA dataframe into **3 clusters** and **plot the clusters**.
- **Reconnect** the **final clusters** of each **observation** to the original dataframe and perform some exploratory analysis. An example would be to calculate the mean income of each cluster (this can be done using a pandas group by).

# Challenge



Please make sure you have submitted the necessary answers and your notebook in the Week 3 answer form.

Please submit your answers and your notebook **by midnight on Sunday the 10th of March.**



# Next Week



Next week you are going to learn about classification techniques!!!