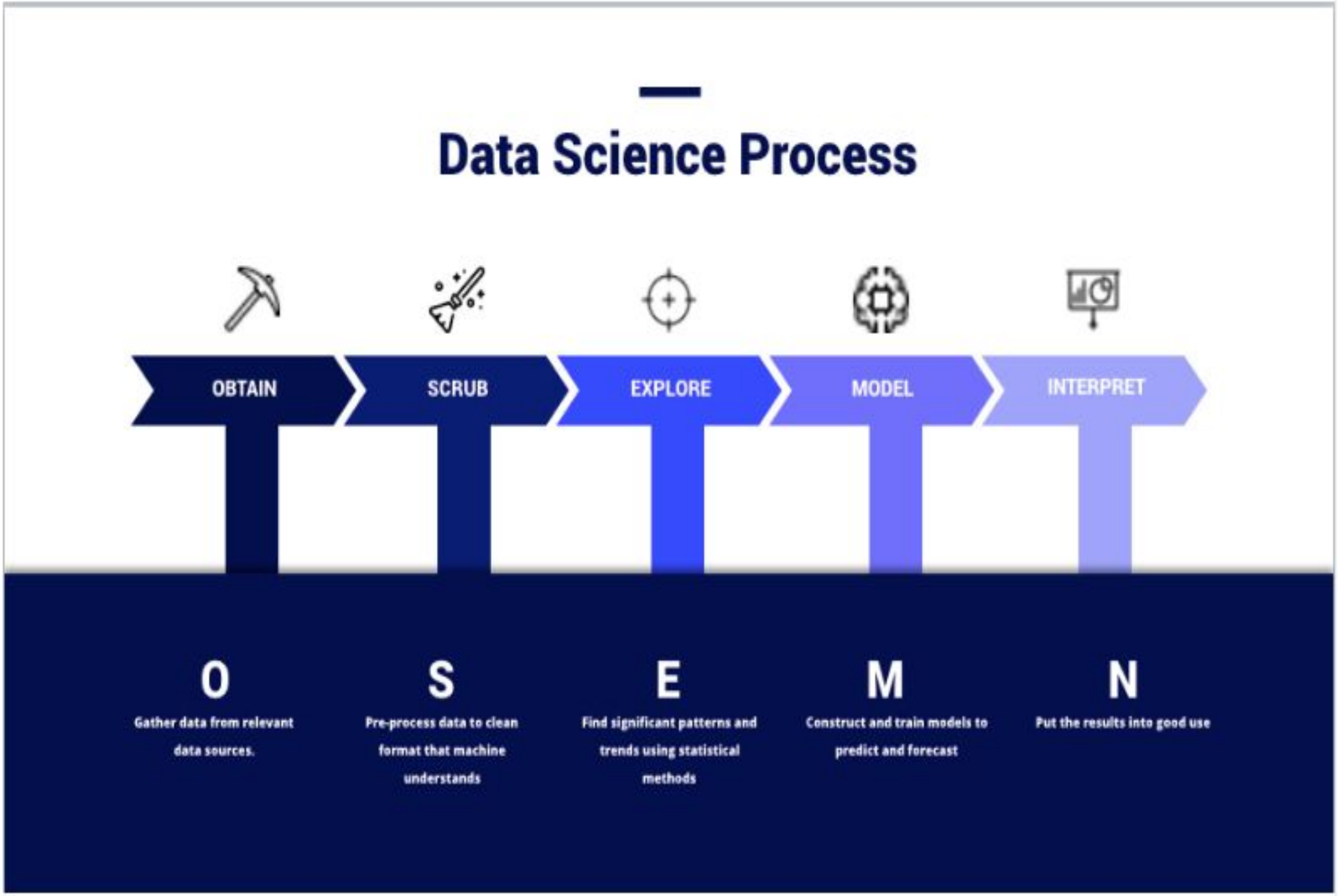
Unit-2

The Five Steps of Data Science

Overview of the Five Steps

- The five essential steps to perform data science are as follows:
  1. Obtaining the data
  2. Exploring the data
  3. Modeling the data
  4. Communicating and
  5. visualizing the results



Data Science Process

O

Gather data from relevant data sources.

S

Pre-process data to clean format that machine understands

E

Find significant patterns and trends using statistical methods

M

Construct and train models to predict and forecast

N

Put the results into good use

1) **Obtain the data**

- The very first step of any data science project is pretty much straightforward, that is to collect and obtain the data you need.

- If you do not have any data at all, you will not be able to have anything to process.

- In this step, you will need to query databases, and this will include a technical skillset like **MySQL** to process the data.

- You may even start out with simple formats like Microsoft Excel to obtain the data and then, later on, convert it into usable data.

- If you are using **Python** or **R**, they have specific packages that can directly read data from these platforms into the programmers.

2) **Exploring Data**

- Once your data is ready to be used, and right before you jump into AI and Machine Learning, you will have to explore the data.
- First of all, you will need to inspect the data and all its properties.
- There are different types of data like numerical data, categorical data, ordinal and nominal data etc. With that, there are different types of data characteristics which will require you to handle them differently.
- Following that, the next step would be to compute descriptive statistics to to extract features and test significant variables.
- Testing significant variables often times is done with correlation.

2) **Exploring Data**

- For example, exploring the correlation of the risk of someone getting high blood pressure in relations to their height and weight.
- Do note that some variables are correlated, but to significant in terms of the model.
- The term “Feature” used in Machine Learning or Modelling, is the data features to help you identify what are the characteristics that represent this database.
- For example, “Name”, “Age”, “Gender” are features of your dataset.

3) **Model Data**

- This is the most interesting stage of the data science project lifecycle. As many people would call it “where the magic happens”.
- One of the first things you need to do in modelling data, is to reduce the dimensionality of your data set.
- Not all your features or values are essential to predicting your model.
- what you need to do is to select the relevant ones that will contribute to the prediction of results you are looking for.
- Other than classification or prediction of the results, our purpose of this stage can also include the grouping of data to understand the logic behind those clusters.

3) **Model Data**

- For example, you would like to group your e-commerce customers to understand their behaviour on your website.

- So this would require you identify groups of data points with clustering algorithms, using methods like k-means; or make predictions using regressions like linear or logistic regressions.

- Lastly, in this step, you can also train models to perform classification. For example, like differentiating the mails you received as “Inbox” and “Spam mail”.

4) **Communicate**

- You usually have a client (internal or external) and you’re going to have to present the results in a way that makes sense to them.
- The more precise and clear the message is the better your chances are of seeing your model at work generating valuable insights.
- Communication of the results is not always the last step in a data science project.

5) Visualize **the results**

- This is arguably the most important step.
- While it might seem obvious and simple, the ability to conclude your results in a digestible format is much more difficult than it seems.
- Visualize the result is part of data science.
- Where data can be represented in the form of graphs like, line, pie, bar, scatter etc.
- Which helps end users to understand your data.