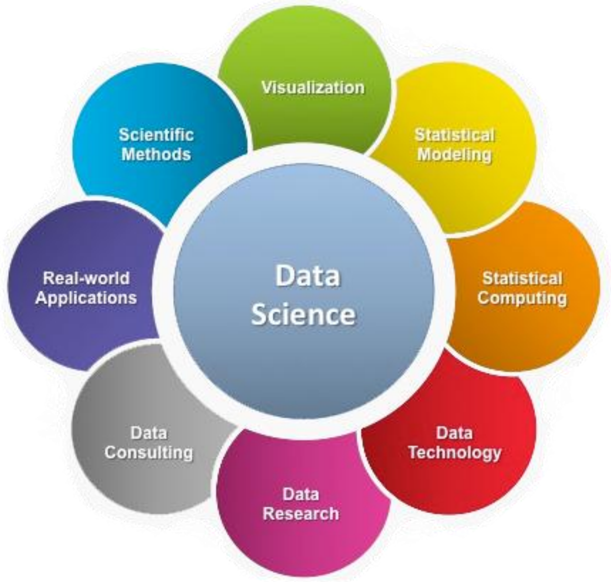
Data Science and its own importance

Introduction: Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to data mining, machine learning and big data.

What is Data?

- Data is often viewed as the lowest level of abstraction from which information and knowledge are derived.
- Data can be numbers, words, measurements, observations or even just descriptions of things. Also, data is a representation of a fact, figure and idea.
- Data on its own carries no meaning. If data to be an information, it must be interpreted and take on a meaning.



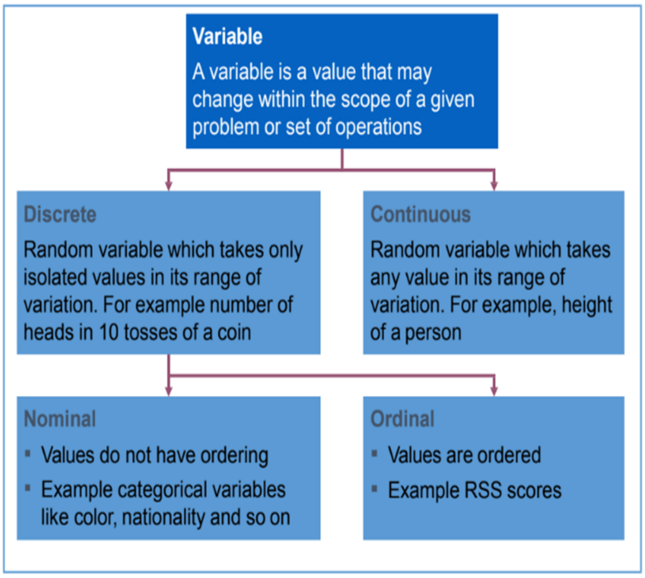
Types of Data/Variables:

Catagorical data is the data that is non numeric.

e.g.. Favourite color, Place of Birth, Types of Car

Quantitative Data is numeric. There are 2 types of quantitative data.

1. Discreate data can only take specific values;
   e.g. shoe size,number of brothers, number ofcars in a car park.
2. Continuous data can take any numerical value;
   e.g. height, mass,length.



Variable
A variable is a value that may change within the scope of a given problem or set of operations

Discrete
Random variable which takes only isolated values in its range of variation. For example number of heads in 10 tosses of a coin

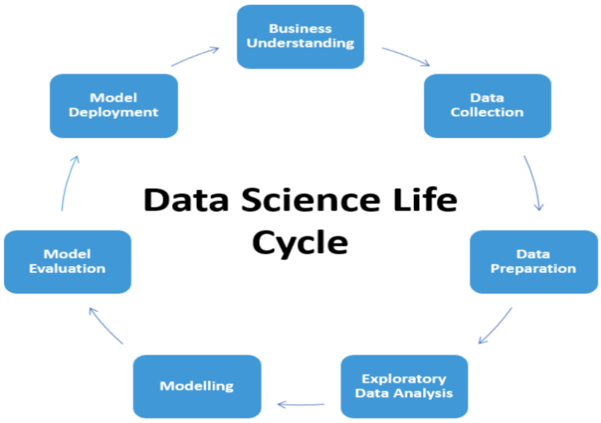
Continuous
Random variable which takes any value in its range of variation. For example, height of a person

Nominal
- Values do not have ordering
- Example categorical variables like color, nationality and so on

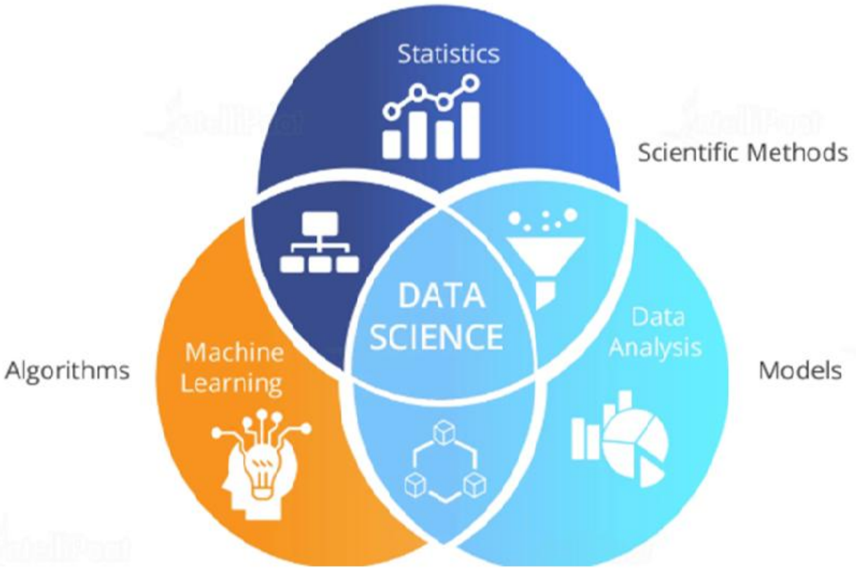
Ordinal
- Values are ordered
- Example RSS scores

Data Science Steps:

- Data Gathering: Unless you're at a company with great data governance you're likely going to have some trouble accessing the data you want. Whether that's because your company has neglected to put the necessary systems in place to gather data, or the data that they are collecting is fragmented and scattered across the organization, you'll have to first spend some time gathering whatever data you'll need to do your job.
- Data Preparation: Once you have access to data, you'll need to spend some time cleaning and formatting it. This is where Data Science can often become more of an art, then a science. Unlike datasets you'll find in competitions, the real world has very messy data sets. Missing values, error in data collection, data formatting, normalization, outliers - these are all issues that you'll have to learn to deal with.
- Exploration: Before diving into building any models, you'll want to explore the data to try to glean some insights. Clustering algorithms, scatterplots, bar graphs, Chernoff faces are all interesting ways of visualizing data that will lead to a better understanding of the structure of your data and aid you in your model building step.
- Model Building: With your data cleaned and formatted, you'll have an opportunity to explore a variety of models to see which one works best. Random Forests, SVM's, Bayesian Predictors Neural Networks, Deep Learning, K-Nearest Neighbours - all models you should familiarize yourself with. There is no one model fits all, and so you again will need to develop intuition on which model suits your particular problem.
- Model Validation: Prediction accuracy is a common benchmark for whether your model is performing well, however often times there are other evaluation metrics to consider. False positives and false negatives are important to think about from the perspective of the problem you're working on. If you're predicting disease, you'll care more about minimizing false negative, since it may result in a persons death -whereas a false positive will only lead to additional testing.
- Model Deployment: Finally you'll deploy your model into the wild, as you gather more data and feedback on how its doing you'll be able to tweak and improve it as time goes on.

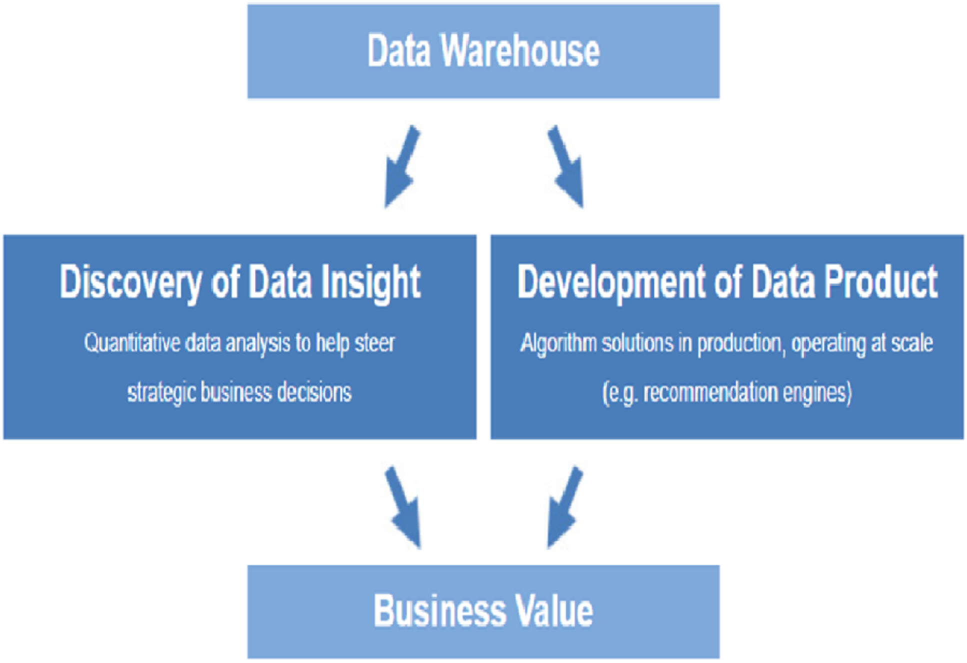


Key Elements of Data Science:



Data Science & its importance:

- Data science helps brands to understand their customers in a much enhanced and empowered manner.
- It allows brands to communicate their story in such a engaging and powerful manner.
- Big Data is a new field that is constantly growing and evolving.
- Its findings and results can be applied to almost any sector like travel, healthcare and education among others.
- Data science is accessible to almost all sectors.



Data Warehouse

Discovery of Data Insight

Quantitative data analysis to help steer strategic business decisions

Development of Data Product

Algorithm solutions in production, operating at scale (e.g. recommendation engines)

Business Value

Python Libraries:

Numpy, Pandas, Seaborn, Bokeh, SciPy, Scrapy, Keras, SciKit-Learn, PyTorch, TensorFlow, XGBoost, Matplotlib, Plotly, Pydot, StatsModel.

Open Source IDE for Programming:

- Google CoLab
- Anaconda (Jupyter Notebook)
- Pycharm

Data Repositories:

- Kaggle
- GitHub
- UCI Machine Learning