Data Science : Brief understanding of Typical Project Life-cycle, Tools, Techniques and skills

Author, Dibyendu Banerjee

A Data Science Foundation White Paper

August 2019

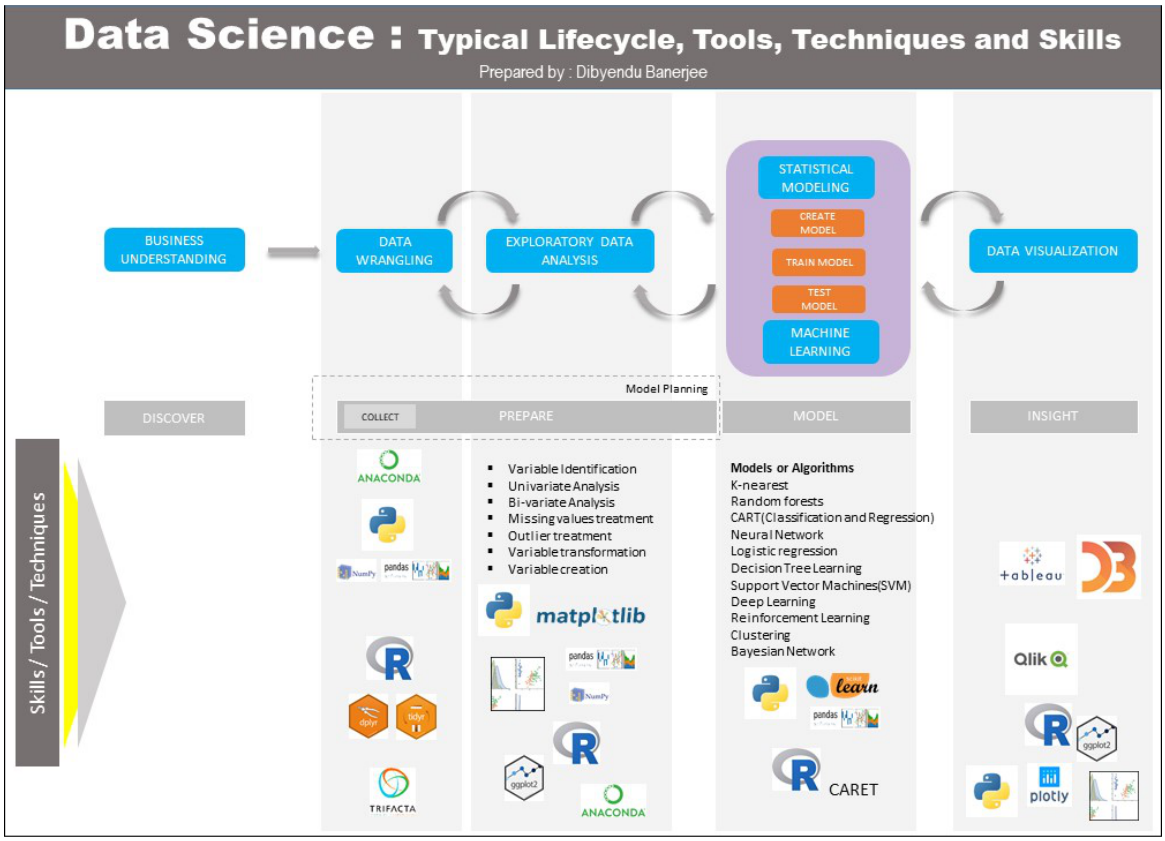
Data Science Foundation
Data Science Foundation, Atlantic Business Centre, Atlantic Street, Altrincham, WA14 5NQ
Tel: 0161 926 3641 Email: admin@datascience.foundation Web: www.datascience.foundation
Registered in England and Wales 4th June 2015, Registered Number 9624670

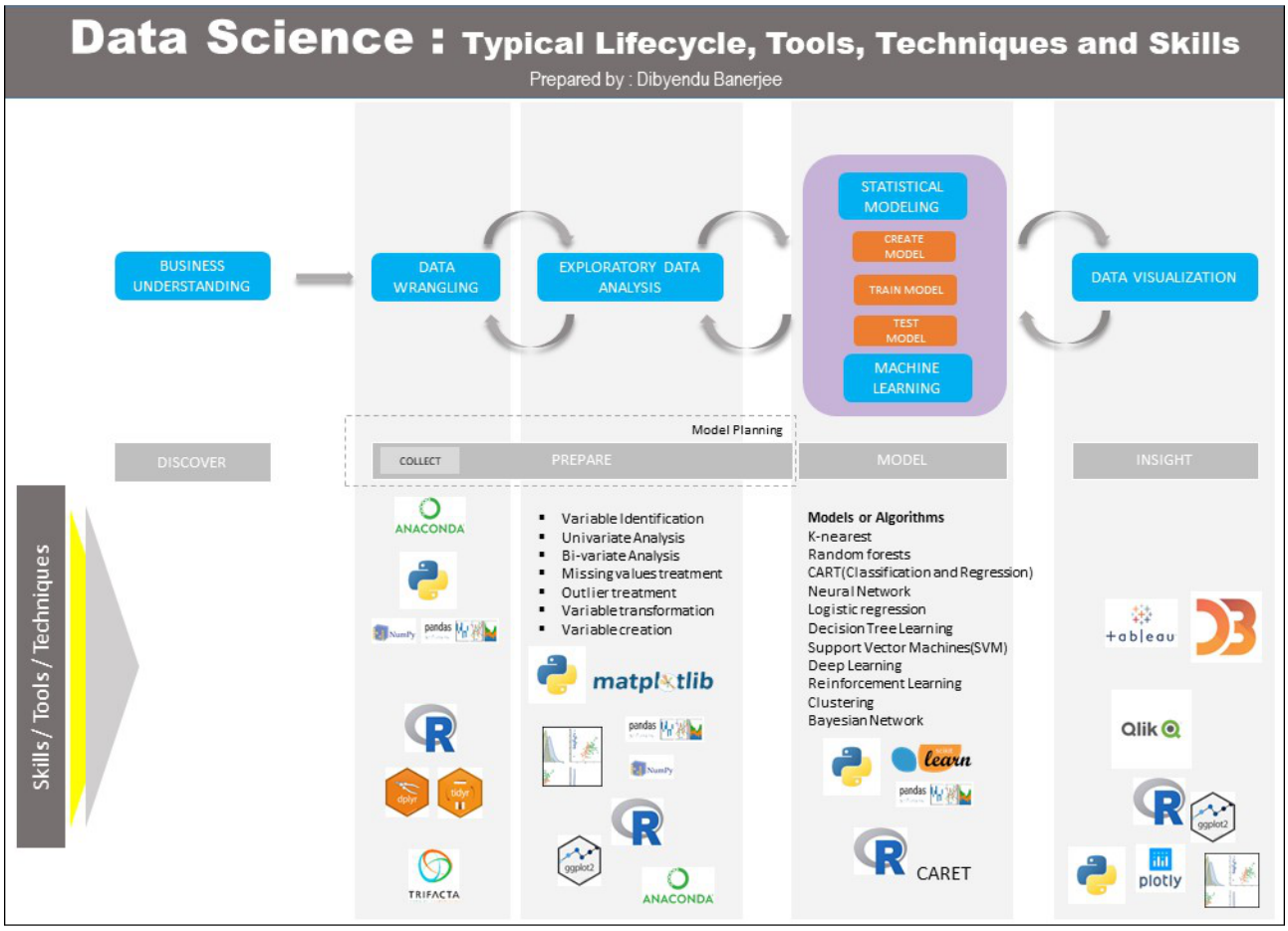
Data science projects do not have a nice clean lifecycle with well-defined steps like software development lifecycle (SDLC). We see in many forums discussion on Data Science, Business Analytics, Big Data, and Machine learning. However, what is the life cycle or steps for such projects. Nobody seems to be able to come up with a compact explanation of how the whole process goes.

People often confuse the lifecycle of a data science project with that of a software engineering project. That should not be the case, as data science is more of science and less of engineering. There is no one-size-fits-all workflow process for all data science projects and data scientists have to determine which workflow best fits the business requirements.

Every step in the lifecycle of a data science project depends on various data scientist skills and data science tools. The typical lifecycle of a data science project involves jumping back and forth among various interdependent science tasks using variety of tools, techniques (mostly statistical methods and formulae), programming etc.

Let us try to see what could be a typical life cycle.


Data Science : Typical Lifecycle, Tools, Techniques and Skills
Prepared by : Dibyendu Banerjee
BUSINESS UNDERSTANDING
DATA WRANGLING
EXPLORATORY DATA ANALYSIS
STATISTICAL MODELING
CREATE MODEL
TRAIN MODEL
TEST MODEL
MACHINE LEARNING
DATA VISUALIZATION
Model Planning
DISCOVER
COLLECT
PREPARE
MODEL
INSIGHT
Skills / Tools / Techniques
ANACONDA
TRIFACTA
Variable Identification
Univariate Analysis
Bi-variate Analysis
Missing values treatment
Outlier treatment
Variable transformation
Variable creation
matplotlib
ANACONDA
Models or Algorithms
K-nearest
Random forests
CART(Classification and Regression)
Neural Network
Logistic regression
Decision Tree Learning
Support Vector Machines(SVM)
Deep Learning
Reinforcement Learning
Clustering
Bayesian Network
CARET
+ableau
Qlik
plotly



Before you can even start on a data science project, it is critical that you understand the problem you are trying to solve.

Before you can even start on a data science project, it is critical that you understand the problem you are trying to solve.

It is important to understand the various specifications, requirements, priorities and required budget. You must possess the ability to ask the right questions. Here, you assess if you have the required resources present in terms of people, technology, time and data to support the project. In this phase, you also need to frame the business problem and formulate initial hypotheses (IH) to test.

According to Microsoft Azure's blog. we typically use data science to answer five type of questions:

1. How much or how many?(regression)
2. Which category? (classification)
3. Which group? (clustering)
4. is this weird? (anomaly detection)
5. Which option should be taken? (recommendation)

In this satge.you should also be identifying the central objective of your project by identifying the variables that need to be predict.

Data Wrangling, sometimes referred to as data **Munging**

Data Wrangling, sometimes referred to as data **Munging**

Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis. Transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.

In other words, it is a **Data Cleaning** activity so why not we call it as **Scrubbing**

MUNGING = SCRUBBING = DATA CLEANING

In this phase, you require analytical sandbox in which you can perform analytics for the entire

duration of the project. However, before that further, you will perform ETL (extract, transform and transform) to get data into the sandbox.

Data might need to be collected from multiple type of data sources.

Few Example of Data Source.

- File format Data(Spreadsheet, CSV, Text files, XML, jSON)
- Relational Database
- Non-relational Database(NoSQL)
- Scrapping Website Data using tools

SKILLS/Tools/Techniques

- Database Management: MySQL, PostgresSQL, MongoDB
- Querying Relational Databases
- Retrieving Unstructured Data: text, videos, audio files, documents
- Distributed Storage: Hadoops, Apache Spark/Flink
- R packages to read from file format
- Python libraries to read from files

EXPLORE… EXPLORE… EXPLORE

EXPLORATORY DATA ANALYSIS

EXPLORE… EXPLORE… EXPLORE

Object of this step is to apply scientific (Statistical) methods to make data more feasibly to feed into MODELS, in other words choosing baseline model is the outcome of this phase.

Exploratory analysis is often described as a philosophy, and there are no fixed rules for how you approach it. There are no shortcuts for data exploration.

Remember the quality of your inputs decide the quality of your output. Therefore, once you have got your business hypothesis ready, it makes sense to spend lot of time and efforts here.

Below are the some of the standard practices involved to understand, clean and prepare your data for building your predictive model:

1. Variable Identification
2. Univariate Analysis
3. Bi-variate Analysis
4. Missing values treatment
5. Outlier treatment
6. Variable transformation
7. Variable creation

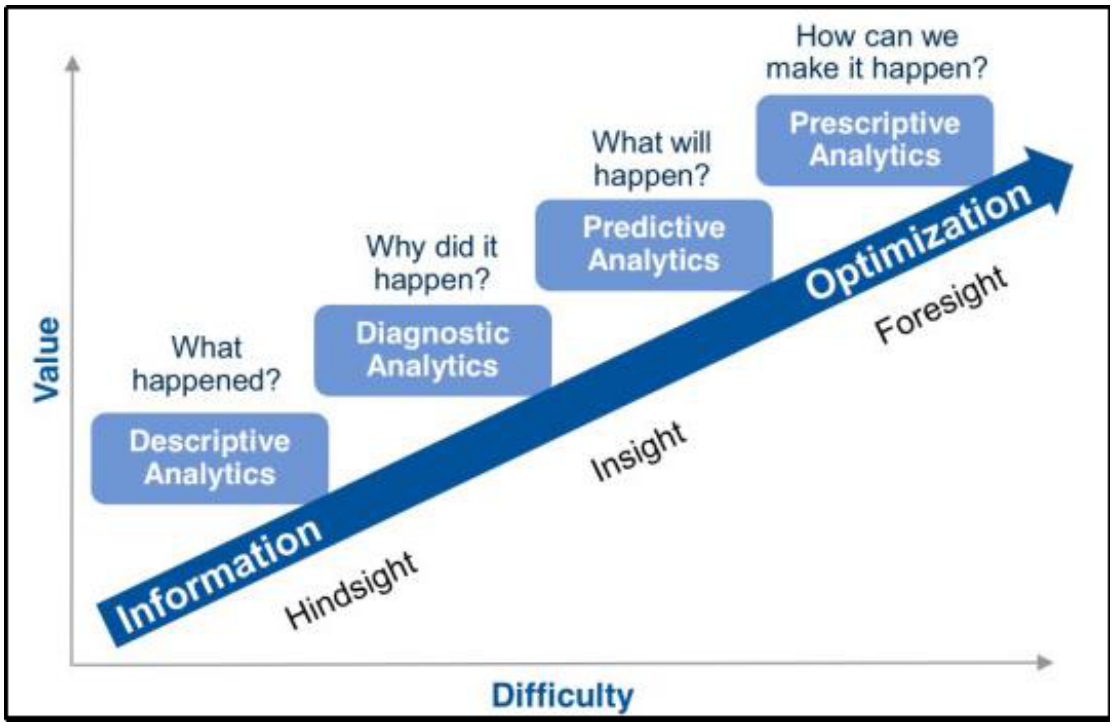
Finally, we will need to iterate over steps 4 – 7 multiple times before we come up with our refined model.

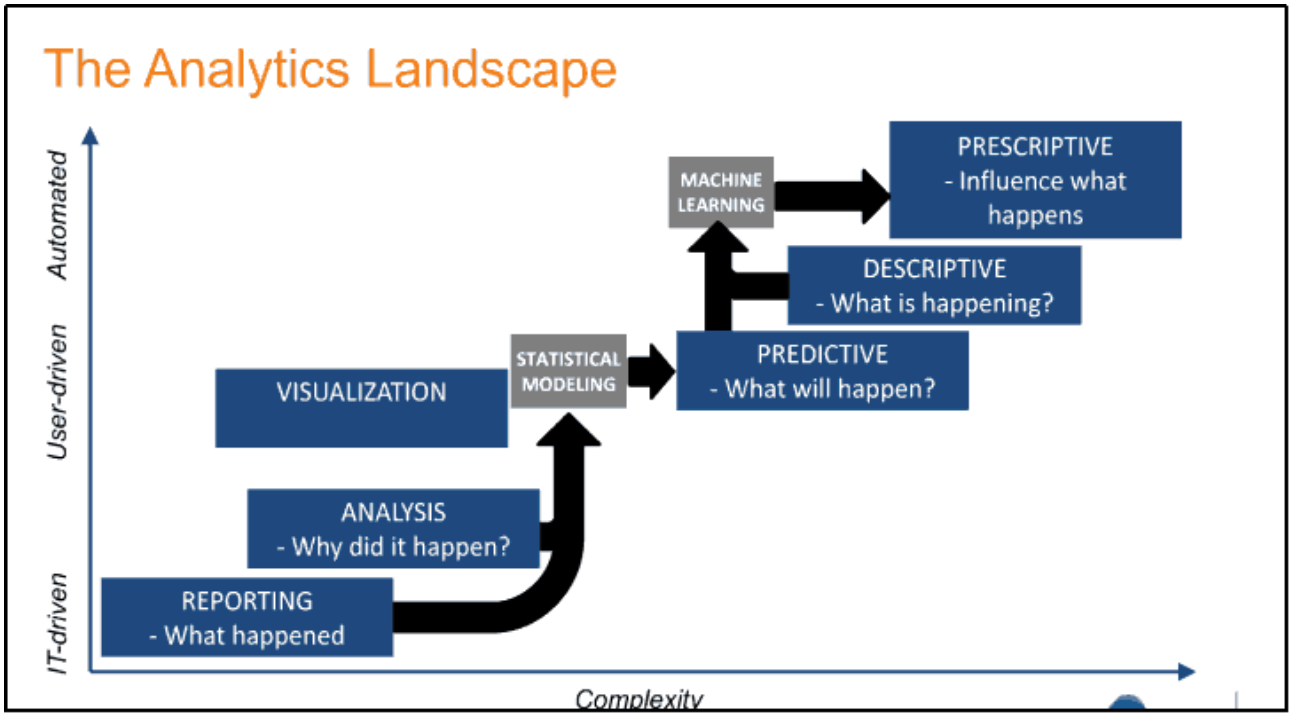
Model Building is the core activity of a data science project. It is carried out either Statistical Driven - Statistical Analytics or using Machine Learning Techniques.

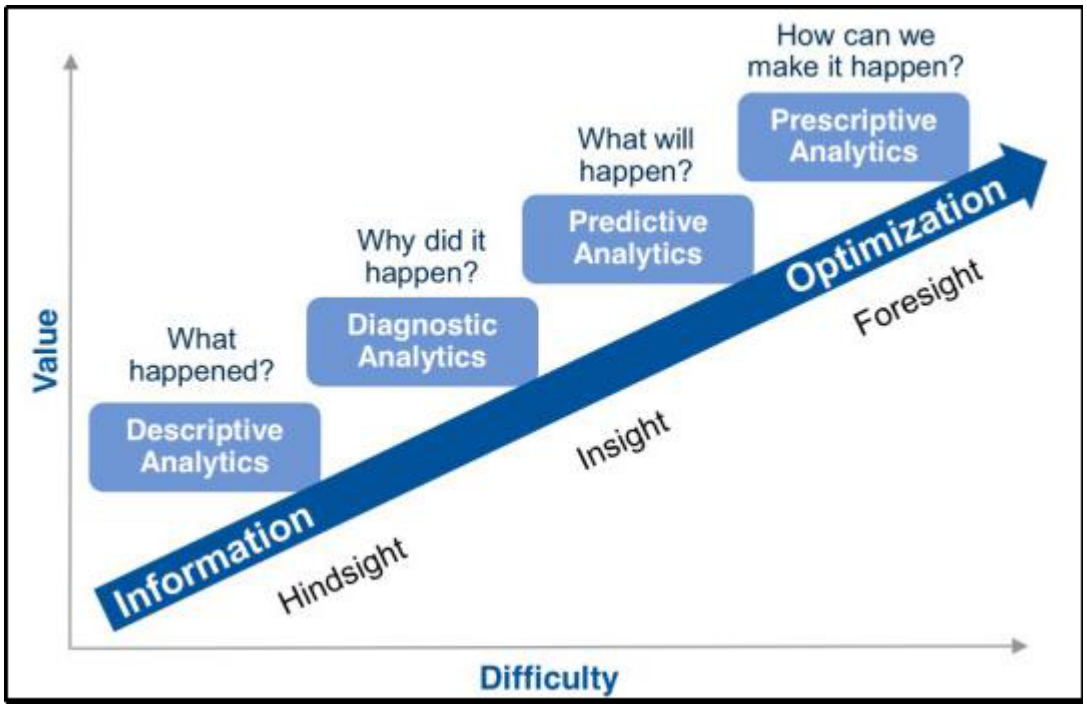
Model Building is the core activity of a data science project. It is carried out either Statistical Driven - Statistical Analytics or using Machine Learning Techniques.

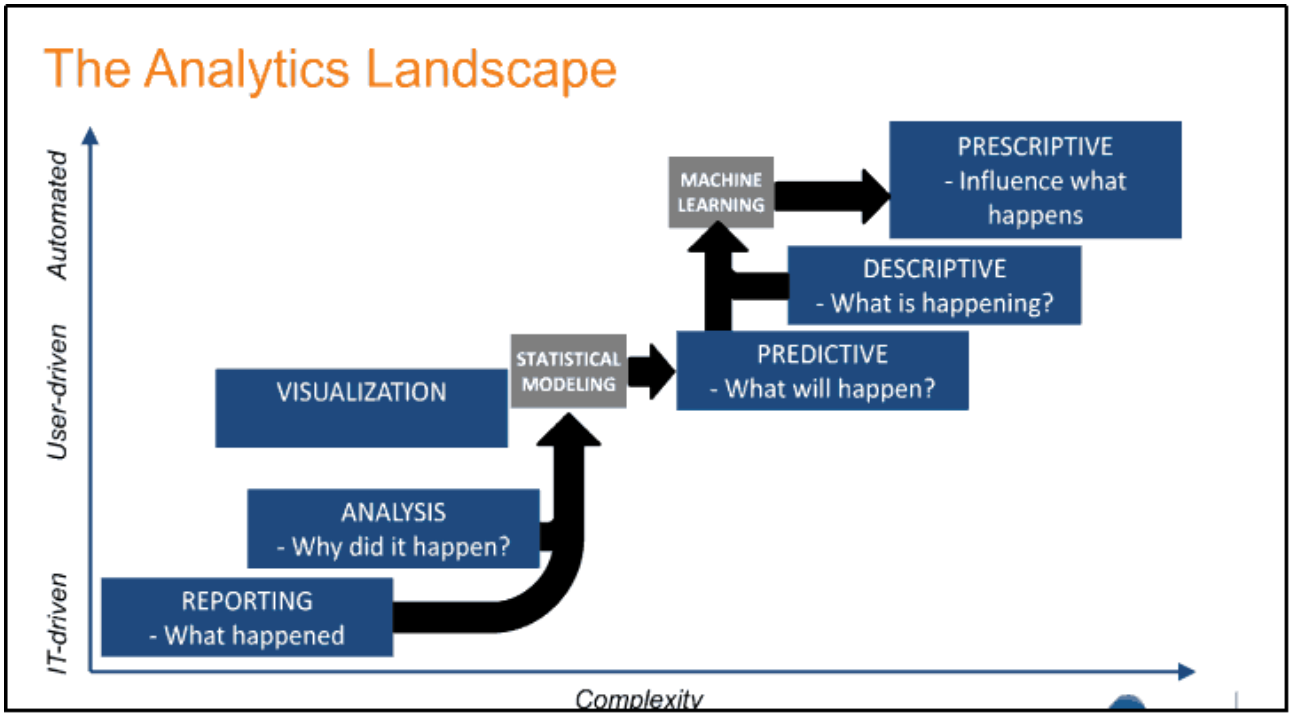
The below first picture explains different stages of analytics. Second picture is typical flow of Data Science activities, which shows statistical modeling, are followed by ML.

The below first picture explains different stages of analytics. Second picture is typical flow of Data Science activities, which shows statistical modeling, are followed by ML.




The Analytics Landscape
Automated
User-driven
IT-driven
Complexity
MACHINE LEARNING
PRESCRIPTIVE
- Influence what happens
DESCRIPTIVE
- What is happening?
STATISTICAL MODELING
PREDICTIVE
- What will happen?
VISUALIZATION
ANALYSIS
- Why did it happen?
REPORTING
- What happened


How can we make it happen?
Prescriptive Analytics
What will happen?
Predictive Analytics
Why did it happen?
Diagnostic Analytics
What happened?
Descriptive Analytics
Value
Optimization
Foresight
Insight
Information
Hindsight
Difficulty



Difference between statistical modeling and ML

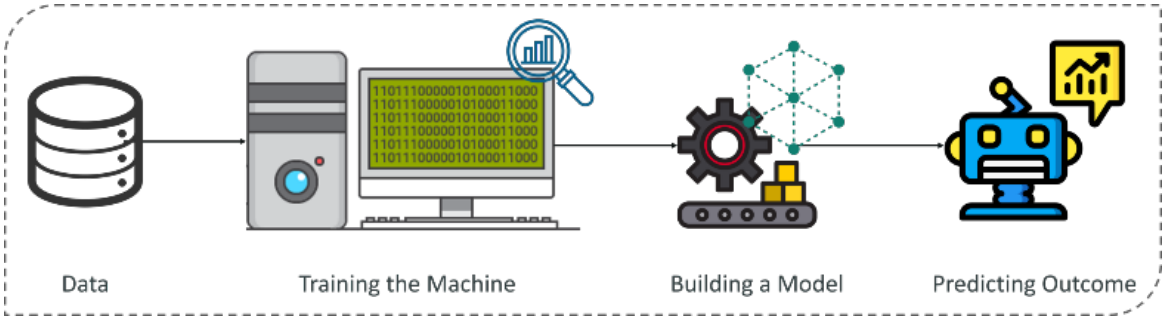
Machine Learning is An algorithm that can learn from data without relying on rules-based programming.Statistical Modelling is Formalization of relationships between variables in the form of mathematical equations.

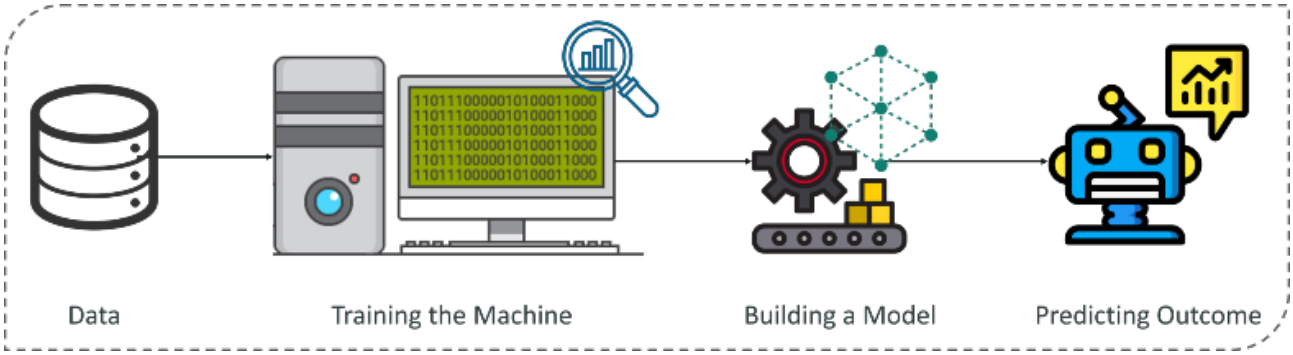
MACHINE LEARNING?

Undoubtedly, Machine Learning is the most in-demand technology in today's market. Its applications range from self-driving cars to predicting deadly diseases.

Machine Learning Terminologies

Algorithm: A Machine Learning algorithm is a set of rules and statistical techniques used to learn patterns from data and draw significant information from it. It is the logic behind a Machine Learning model. An example of a Machine Learning algorithm is the Linear Regression algorithm.





Model: A model is the main component of Machine Learning. A model is trained by using a Machine Learning Algorithm. An algorithm maps all the decisions that a model is supposed to take based on the given input, in order to get the correct output.

Predictor Variable: It is a feature(s) of the data that can be used to predict the output.

Response Variable: It is the feature or the output variable that needs to be predicted by using the predictor variable(s).

Training Data: The Machine Learning model is built using the training data. The training data helps the model to identify key trends and patterns essential to predict the output.

Testing Data: After the model is trained, it must be tested to evaluate how accurately it can predict an outcome. This is done by the testing data set.

Machine Learning Types

A machine can learn to solve a problem by following any one of the following three approaches.

These are the ways in which a machine can learn:

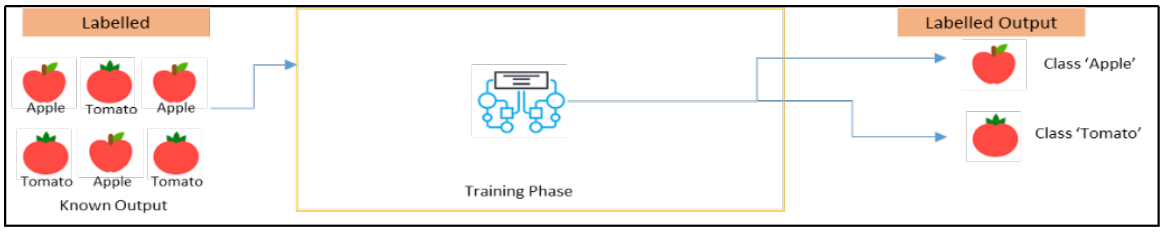
- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning(Out of Scope of this document)

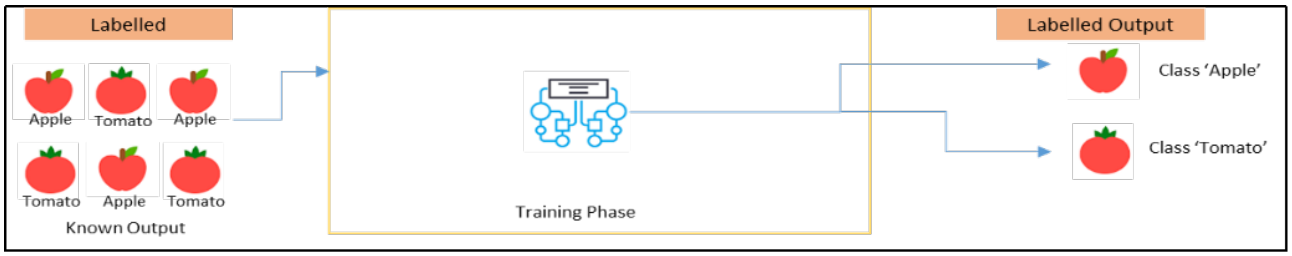
Supervised Learning

Supervised learning is a technique in which we teach or train the machine using data, which is well **labeled**.

To understand Supervised Learning let us consider an analogy. As kids we all needed guidance to solve math problems. Our teachers helped us understand what addition is and how it is done. Similarly, you can think of supervised learning as a type of Machine Learning that involves a guide. The labeled data set is the teacher that will train you to understand patterns in the data. The labeled data set is nothing but the training data set.

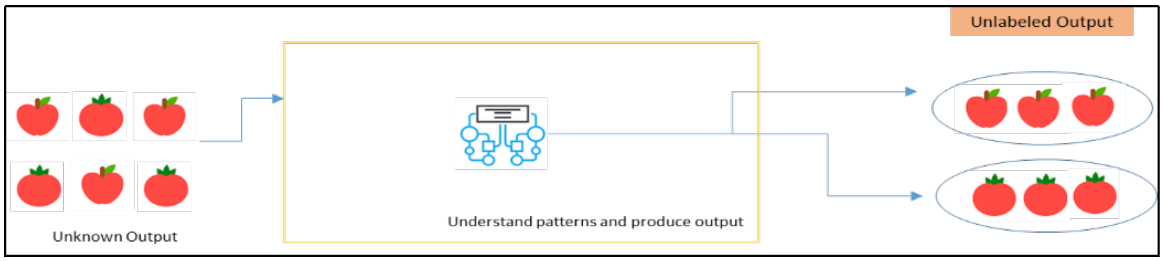
The pic below shows Supervised Learning. By doing so, you are training the machine by using labeled data. In Supervised Learning, there is a well-defined training phase done with the help of labeled data.

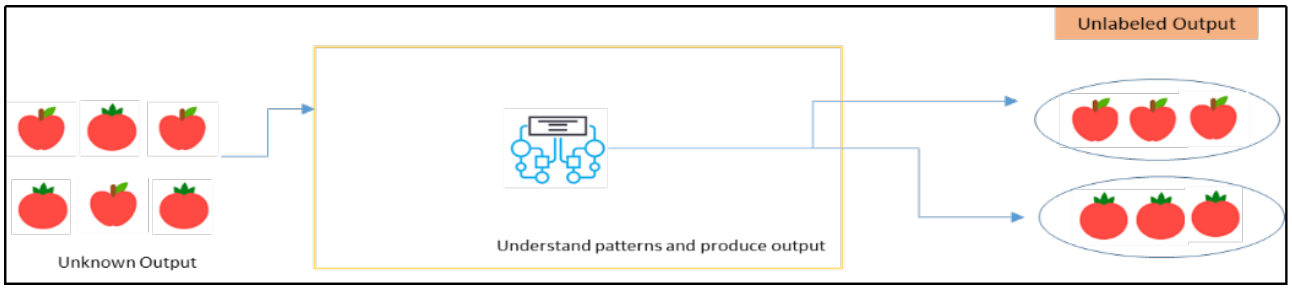


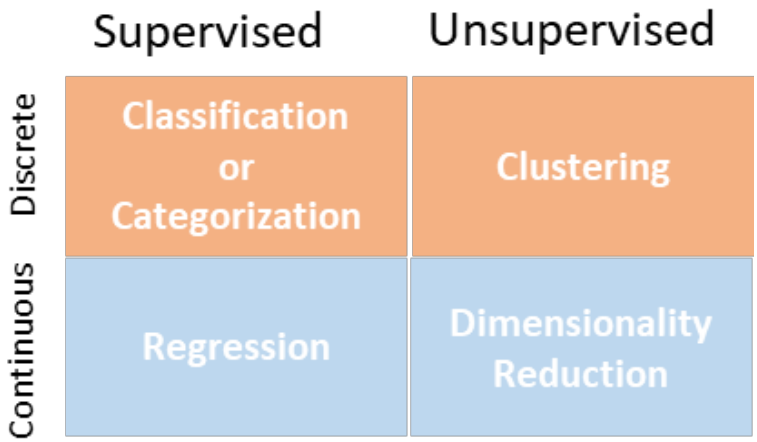


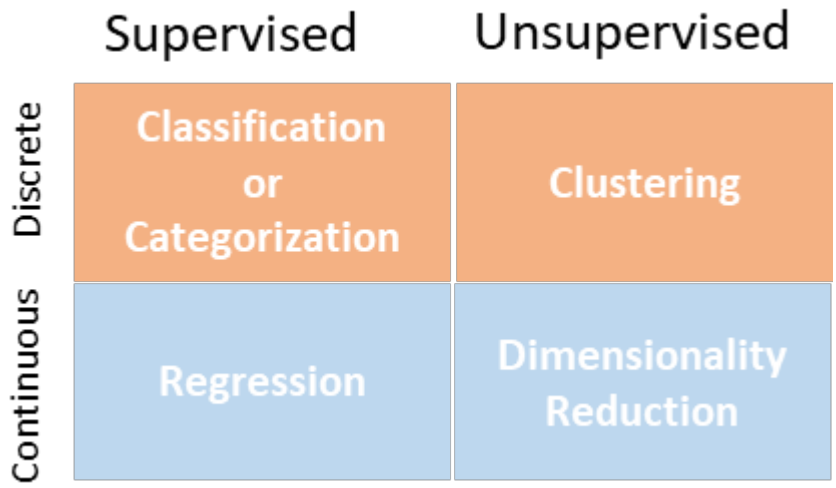
Unsupervised Learning

Unsupervised learning involves training by using unlabeled data and allowing the model to act on that information without guidance. Think of unsupervised learning as a smart kid that learns without any guidance.









List of Common Machine Learning Algorithms

Here is the list of commonly used machine learning algorithms. These algorithms can be applied to almost any data problem:

Regression Algorithms

Regression is concerned with modeling the relationship between variables that is iteratively refined using a measure of error in the predictions made by the model. Regression methods are a workhorse of statistics and have been co-opted into statistical machine learning.

The most popular regression algorithms are:

- Ordinary Least Squares Regression (OLSR)
- Linear Regression
- Logistic Regression
- Stepwise Regression
- Multivariate Adaptive Regression Splines (MARS)
- Locally Estimated Scatterplot Smoothing (LOESS)

Clustering Algorithms

Clustering, like regression, describes the class of problem and the class of methods.Clustering methods are typically organized by the modeling approaches such as centroid-based and hierarchal. All methods are concerned with using the inherent structures in the data to best organize the data into groups of maximum commonality.

The most popular clustering algorithms are:

- k-Means
- k-Medians

Dimensionality Reduction Algorithms

Like clustering methods, dimensionality reduction seek and exploit the inherent structure in the data, but in this case in an unsupervised manner or order to summarize or describe data using less information.

This can be useful to visualize dimensional data or to simplify data, which can then be used in a supervised learning method. Many of these methods can be adapted for use in classification and regression.

- Principal Component Analysis (PCA)
- Principal Component Regression (PCR)
- Partial Least Squares Regression (PLSR)
- Linear Discriminant Analysis (LDA)
- Mixture Discriminant Analysis (MDA)

Instance-based Algorithms

Instance-based learning model is a decision problem with instances or examples of training data that are deemed important or required to the model. Such methods typically build up a database of example data and compare new data to the database using a similarity measure in order to find the best match and make a prediction. For this reason, instance-based methods are also called winner-take-all methods and memory-based learning. Focus is put on the representation of the stored instances and similarity measures used between instances.

The most popular instance-based algorithms are:

- k-Nearest Neighbor (kNN)
- Learning Vector Quantization (LVQ)
- Self-Organizing Map (SOM)
- Locally Weighted Learning (LWL)

Decision Tree Algorithms

Decision tree methods construct a model of decisions made based on actual values of attributes in the data.

Decisions fork in tree structures until a prediction decision is made for a given record. Decision trees are trained on data for classification and regression problems. Decision trees are often fast and accurate and a big favorite in machine learning.

The most popular decision tree algorithms are:

- Classification and Regression Tree (CART)

Other important algorithms are:

1. Naive Bayes
2. Random Forest
3. Dimensionality Reduction Algorithms
4. Neural Network Algorithms
5. Natural Language Processing (NLP)

Tools

There are various R packages available

Python - Scikit-learn – It is a free library, which contains simple and efficient tools for data analysis and mining purposes. You can implement various algorithm, such as logistic regression, time series algorithm using scikit-learn.

Interpreting data refers to the presentation of your data to a non-technical layman. We deliver the results in to answer the business questions we asked when we first started the project, together with the actionable insights that we found through the data science process.

Interpreting data refers to the presentation of your data to a non-technical layman. We deliver the results in to answer the business questions we asked when we first started the project, together with the actionable insights that we found through the data science process.

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

On top of that, you will need to visualize your findings accordingly, keeping it driven by your business questions. It is essential to present your findings in such a way that is useful to the organization, or else it would be pointless to your stakeholders.

Note : Data Visualization is a technique where data is visualized using certain tools. Visualization is used by data scientist as and when required say it is EDA or Data Wrangling etc. Hence, form general life cycle perspective DATA VISUALIZATION can be more generically called and as getting INSIGHTS.

SKILLS

In this process, technical skills only are not sufficient. One essential skill you need is to be able to tell a clear and actionable story. If your presentation does not trigger actions in your audience, it means that your communication was not efficient. Remember that you will be presenting to an audience with no technical background, so the way you communicate the message is key.

- Tools
- Tableau
- Power BI
- R – ggplot2, lattice,
- Python – Matpoltlib, Seaborn, Plotly.

Range of Technologies Brands in Data Science

About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

Contact Data Science Foundation

Email:admin@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation