
HALLIBURTON

Landmark

# Data Science Course

# 12th Oct 2022

# Contents

## Course Introduction

This course will provide the participants with a basic understanding of big data analytics and Machine Learning so that these emerging technologies can be applied to solve some of the problems facing the Oil & Gas industry.

Attendees will gain a strong foundation on the applicability, advantages and limitations of machine learning in the oil and gas industry in an efficient manner, because the program has an accelerated and intense pace. The 5 days of the boot camp will be used to the maximum to develop the skills and knowledge related to data science. The algorithms of artificial intelligence and the different machine learning workflows that they will learn are adapted to the problems of the industry. Finally, the participants will learn the Halliburton methodology to develop predictive models that has been successful in multiple use-cases all over the world.

## Navigating the Technology and Big Data Maze

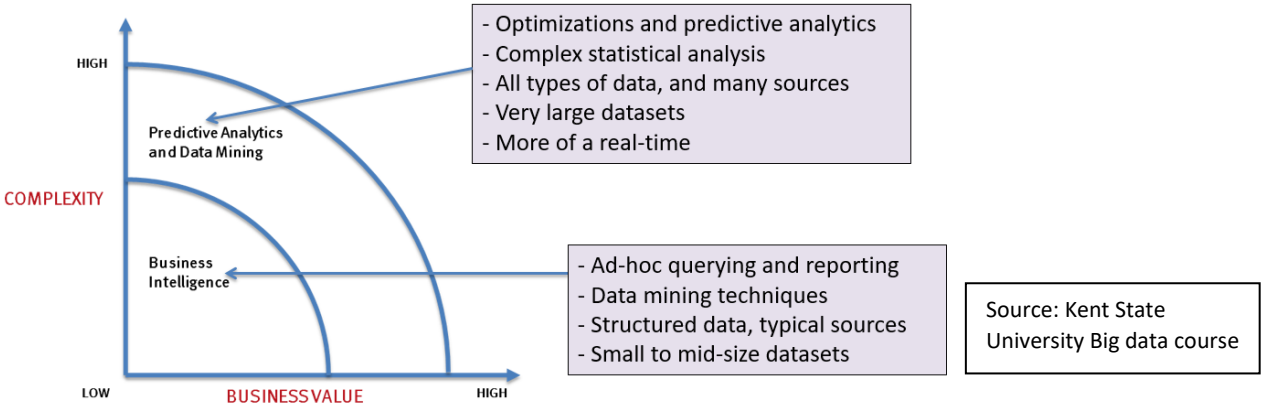
According to Gartner, Big data is data that contains greater variety, arriving in increasing volumes and with ever-higher velocity. The trend to larger data sets is due to the additional value that can be derived from the analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data. This means that more correlations can be found.

### The Three Vs of Big Data

| | |
|---|---|
| **Volume** | The amount of data matters. With big data, you'll have to process high volumes of low-density, unstructured data. This can be data of unknown value, such as Twitter data feeds, clickstreams on a webpage or a mobile app, or sensor-enabled equipment. For some organizations, this might be tens of terabytes of data. For others, it may be hundreds of petabytes. |
| **Velocity** | Velocity is the fast rate at which data is received and (perhaps) acted on. Normally, the highest velocity of data streams directly into memory versus being written to disk. Some internet-enabled smart products operate in real time or near real time and will require real-time evaluation and action. |
| **Variety** | Variety refers to the many types of data that are available. Traditional data types were structured and fit neatly in a relational database. With the rise of big data, data comes in new unstructured data types. Unstructured and semistructured data types, such as text, audio, and video, require additional preprocessing to derive meaning and support metadata. |

Source: IBM Data Science Blog

These are known as the three Vs.  Put simply, big data is all about large and complex datasets that are created from a variety of sources. These massive volumes of data can be used to address business problems that were impossible to tackle before.




The above figure illustrates the evolution of data science analytics because of the advent of big data. Earlier, it was the case that the field of business intelligence was focused on running searches on smaller databases that were not real-time based on human-intervention. With more complex databases and algorithms, it is possible to automate a lot of these tasks and actually generate more value. This is the core promise of big-data analytics and machine learning.

## Contextual Information

The history of big-data analytics can be traced to the massive amounts of data being created by companies like Google, Facebook and YouTube in the mid-2000s. Engineers in these companies used newly developed open-source frameworks like Hadoop (According to Apache, “software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models”) and NoSQL (da database format that can accommodate a wide variety of different data types) to better store and run operations on these huge datasets.

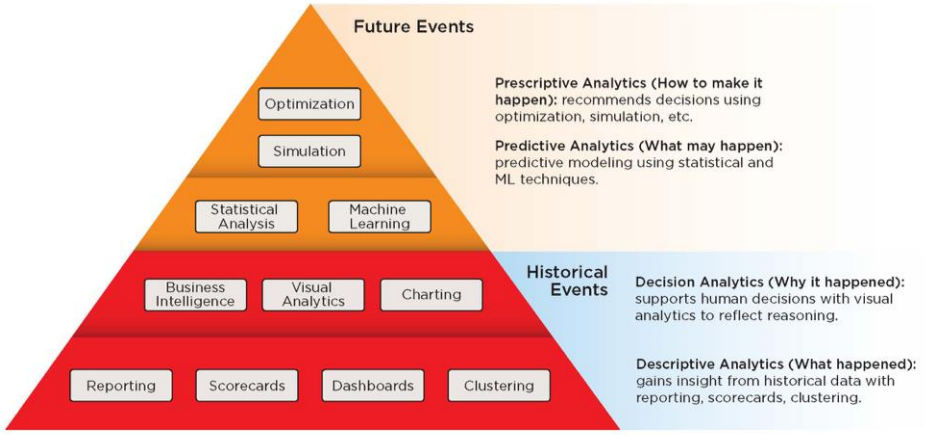
Nowadays, some common use-cases of big data analytics across industries are:

1. **Recommendation systems:** Various companies like YouTube, Rakuten and Amazon use usage data collected from all users to try to predict what content to show to a particular user based on what similar users liked the most. In this way, the recommender can try t to predict what they would respond to the best.
2. **Predictive Maintenance:** With the growth of Internet of Things and the number of sensors in the world, it is possible to collect a vast amount of data from real world systems such as assembly lines, retail stores and infrastructure projects. It is possible to use this data to predict mechanical and system failure by trying to find patterns that are visible in the data before a system fails. Systems like this are being developed and used to companies such as GE, Toyota and FANUC.

3. **Fraud Detection and Prevention:** In the financial sector, security and risk management is one of the most critical functions. Data collected and generated in this sector is based around transaction logs and client data regarding demographics, income, etc. With this information it is possible to detect anomalous transactions to limit losses from lost credit cards and hacking attempts through the use of pattern detection algorithms.

## Links to the Oil and Gas Industry

Oil and Gas companies have been adopting technologies for years, helping to increase the recovery of fossil resources, improve production processes, reduce costs and improve safety. As an industry that has been at the forefront of technology development and adoption for almost a 100 years, this data revolution promises to deliver enormous benefits.




Source: Equinix Blog

The above image highlights where the emerging technologies fit in with respect to the current data collection and storage systems. The higher-end technologies will allow better prediction and simulation. According to McKinsey, these applications could save the oil and gas industry as much as $50 billion in the coming decade. Since the collapse of the global oil price in late 2014, companies have increasingly been looking at technology to reduce costs, improve efficiency and minimize downtime. Finally, as the future remains challenging for our industry, these technologies will only get more important as time goes on.

# Introduction to Machine Learning

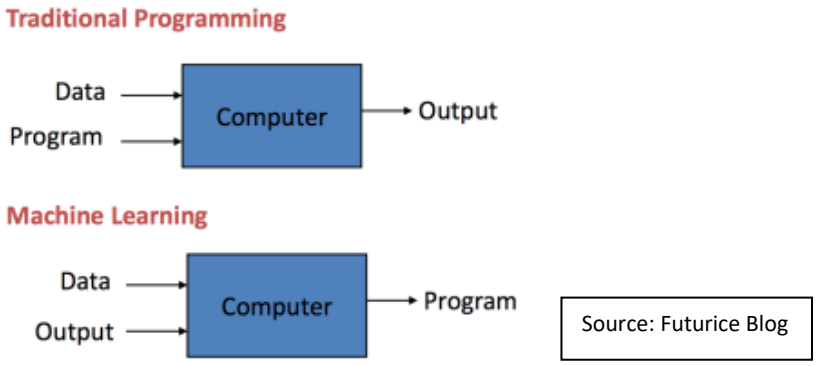
Machine Learning is one of the most important parts of big data analytics as it leads to the valuable insights that are hidden in the vast amounts of data. It refers to a process by which a computer can learn a general model from particular data without being explicitly programmed. This means that the computer is able to pick up on patterns from previously collected data. This allows for models to be created that can ML can forecast/predict events or classify/detect an item or a condition.

As mentioned earlier, Machine Learning is used anywhere from automating mundane tasks to offering intelligent insights and industries in every sector try to benefit from it. Some application areas from our industry include:

- **Drilling Performance Prediction:** In this use case, the rate of penetration can be predicted based on the data collected from wells that have been already drilled. This will allow for better performance monitoring and prognostics.
- **Production Forecasting:** The lifetime production of a well can be forecasted by comparing various parameters like drainage area, production and pressure to previously collected data.
- **Equipment failure/downtime prediction:** Anomalous behavior in equipment can be predicted using real-time data collected from temperature, torque and other kinds of sensors.

## What is Machine Learning?

Once the large datasets have been created, methodologies are needed that can extract valuable knowledge from these hordes of data. According to Arthur Samuel (coined the term ML), Machine Learning algorithms enable the computers to learn from data, and even improve themselves, without being explicitly programmed. ML is a category of an algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed.




Source: Futurice Blog

The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available. The procedure used to build these models is called training. In short, the data and output is used to create a model program that can then be reused to add value. This is visible in the above image. A traditional program is only capable of performing actions it is explicitly programmed to do, while a ML algorithm outputs another program that can take any input to generate the output based on the data that was used to train it.

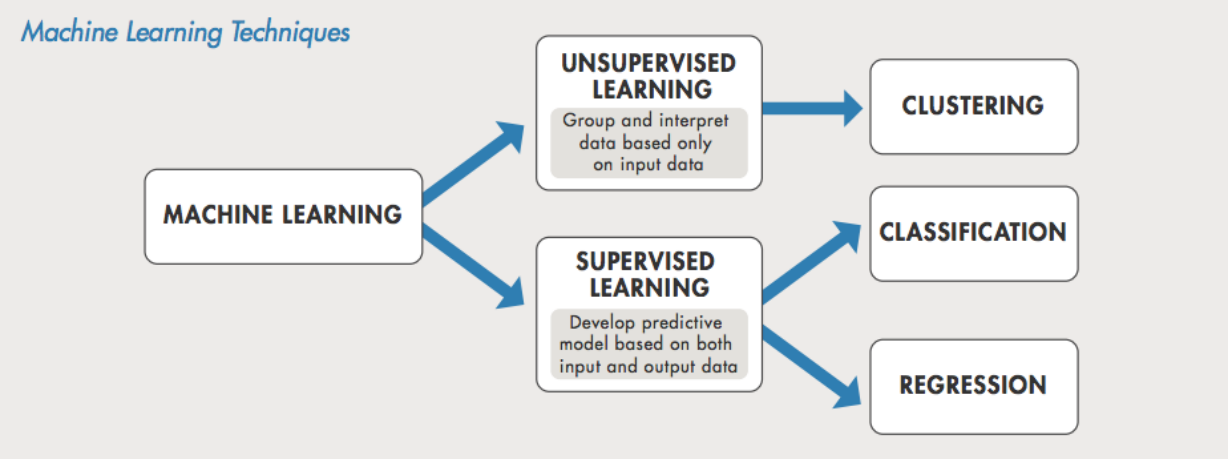

| area_type | availability | location | size | society | total_sqft | bath | balcony | price |
|---|---|---|---|---|---|---|---|---|
| Super built-up Area | 19-Dec | Electronic City Phase II | 2 BHK | Coomee | 1056 | 2 | 1 | 39.07 |
| Plot Area | Ready To Move | Chikka Tirupathi | 4 Bedroom | Theanmp | 2600 | 5 | 3 | 120 |
| Built-up Area | Ready To Move | Uttarahalli | 3 BHK | | 1440 | 2 | 3 | 62 |
| Super built-up Area | Ready To Move | Lingadheeranahalli | 3 BHK | Soiewre | 1521 | 3 | 1 | 95 |
| Super built-up Area | Ready To Move | Kothanur | 2 BHK | | 1200 | 2 | 1 | 51 |
| Super built-up Area | Ready To Move | Whitefield | 2 BHK | DuenaTa | 1170 | 2 | 1 | 38 |
| Super built-up Area | 18-May | Old Airport Road | 4 BHK | Jaades | 2732 | 4 | | 204 |
| Super built-up Area | Ready To Move | Rajaji Nagar | 4 BHK | Brway G | 3300 | 4 | | 600 |
| Super built-up Area | Ready To Move | Marathahalli | 3 BHK | | 1310 | 3 | 1 | 63.25 |
| Plot Area | Ready To Move | Gandhi Bazar | 6 Bedroom | | 1020 | 6 | | 370 |
| Super built-up Area | 18-Feb | Whitefield | 3 BHK | | 1800 | 2 | 2 | 70 |
| Plot Area | Ready To Move | Whitefield | 4 Bedroom | Prrry M | 2785 | 5 | 3 | 295 |
| Super built-up Area | Ready To Move | 7th Phase JP Nagar | 2 BHK | Shncyes | 1000 | 2 | 1 | 38 |
| Built-up Area | Ready To Move | Gottigere | 2 BHK | | 1100 | 2 | 2 | 40 |
| Plot Area | Ready To Move | Sarjapur | 3 Bedroom | Skityer | 2250 | 3 | 2 | 148 |

As an example of a simple machine learning problem, the image above represents a dataset which can be used to train a ML model that can predict the price of a house based on certain aspects of the property (such as location, house area, number of rooms, etc.). This dataset would need to be created through a process of data processing or engineering in which the raw data is structured, cleaned and collated into machine readable columns and rows.

The next step is termed feature selection. Some of the house parameters are selected from the data to be features based on the degree of correlation with the price of the house (higher correlations would make a better feature as it is more predictive). This can be done in a variety of manual or automated ways, depending on the tools or platform used.

The price is termed the label and the model will use the features for training so that the label can be predicted. After training, new data not included in the training dataset can be fed to the model, and the model will be able to make predictions about the price without any additional programming or tweaking. This describes a simple workflow that illustrates the power of a machine learning program.

## Machine Learning Types



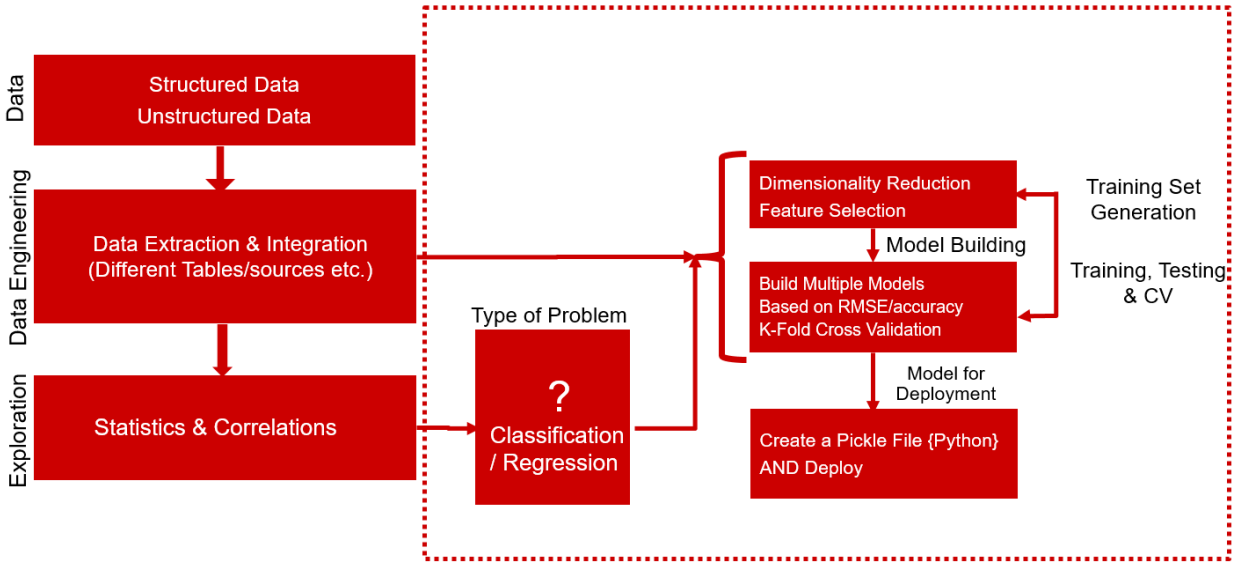

Source: Mathworks MATLAB Documentation

The main kinds of machine learning are supervised and unsupervised learning. They are differentiated by the kind of data available when the models are being trained. Both of these types of learning are used in a variety of different use cases.

Supervised learning is used to develop a predictive model when both the input and output data is available. This kind of data is termed labelled as it is possible to know the associated output with the features. The prediction of house prices based on house features would be an example of this technique. Supervised learning requires human intervention in the labelling stage as it is not possible for the computers to come up with them. It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. Knowing the correct answers, the algorithm repeat some steps to makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance.

Unsupervised learning is used to develop predictive models based only on the input data. This means that the dataset lacks labels. The task of the machine is to group unsorted information according to similarities, patterns or differences without any prior training of data. Systems like these can be used to segment customers and build recommendation systems. The two most important type of unsupervised learning algorithms perform principal component (PCA) and clustering analysis. PCA is used to simplify complex datasets while clustering allows for the data to be grouped based on underlying characteristics.

## Machine Learning Workflow




The machine learning workflow begins with the data that has been collected. If the ML project requires real-time data, it is possible to build an IoT (Internet of Things) system using different sensors. It is also possible to create a data set from various sources such as a file, archives and databases. These collections can rarely be used directly for performing the analysis process as there might be a lot of missing data, outliers, unorganized text data or noisy data. The process of data engineering or pre-processing is used to take care of these issues. In machine learning, there is an 80/20 rule. Every data scientist should spend 80% time for data pre-processing and 20% time to actually perform the analysis

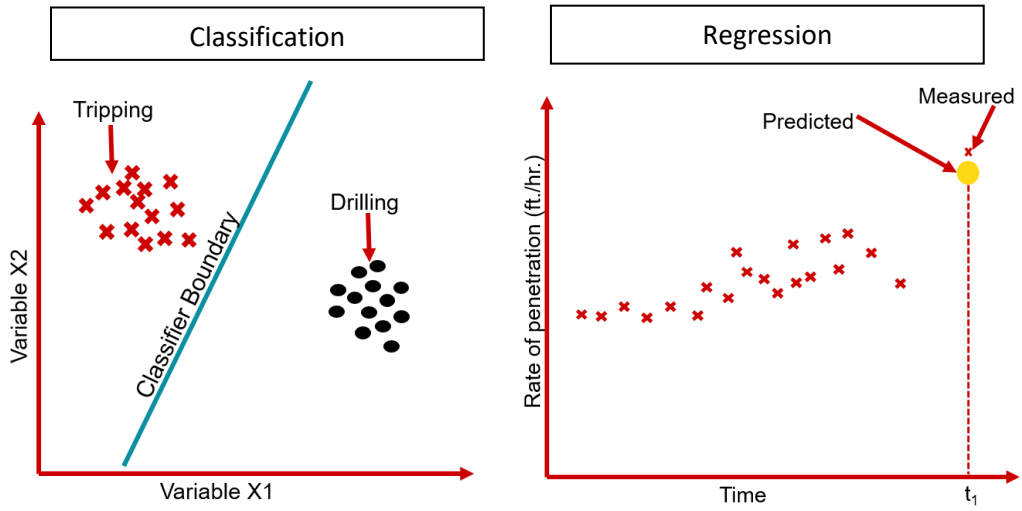
It is important to engineer this data so that the most important 'features' can be extracted for use by the ML algorithms. The selection of features can be done using techniques such as correlation matrices so that the data parameters that are the most correlated with the label data are selected for training purposes.

A ML algorithm is chosen based on the problem-case. For the next procedure, the processed data is split into a training and testing set. The training set is usually a larger section of the total dataset as more data leads to a better quality model.

Once the training has been completed, the model is considered built. After building, it is important to evaluate the model. There are various metrics and strategies that exist for this task. The choices for them is based on the kind of problem being solved and the type of ML algorithm used to construct the model. At this stage, it is possible to deploy and use it to make predictions for actual data.

## Types of Problems

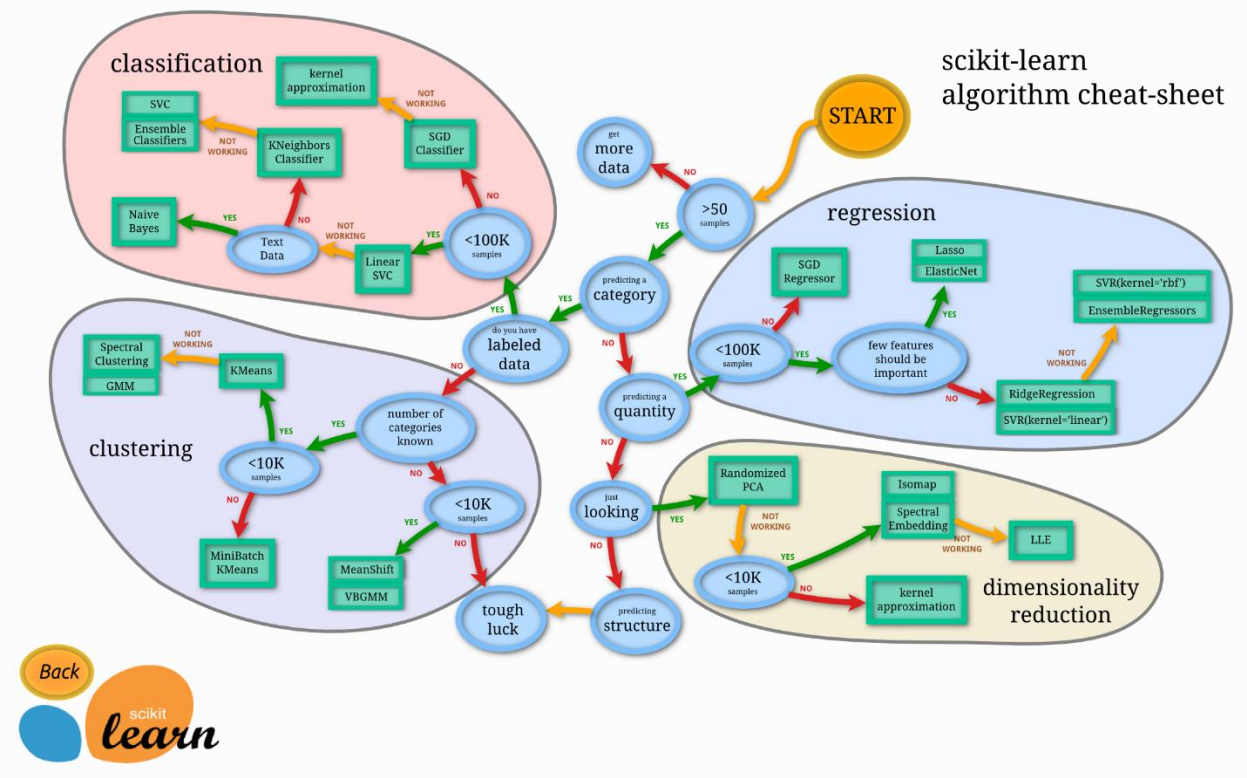
As mentioned in the workflow, the type of problem dictates the kind of algorithm used. In supervised learning, the two primary problems require either a classification or a regression model.




In a classification model, the requirement is for the data to be categorized into various classes in a discrete manner. For example, a common problem would be identifying if an email is spam or not based on features such as length of the mail, sender address and email contents. An industry example from the figure shows how a trained classifier boundary can differentiate between tripping and drilling data.

A regression model, on the other hand, the required output is numerical. This means that the trained model should be able to approximate a mapping function from input variables to a continuous output variable. The house price prediction was an example of this as the function mapped the selected features to a predicted house price. An example from the oil and gas industry would try to predict rate of penetration (ROP) at time $t_1$ based on the historical data. If ROP is a $f(Drilling\ Parameters, Formation, BHA, Pressure\ regimes, Bit)$, then a regression model can be trained with those features and the labelled data to make a prediction.

## Algorithm Choice Flowchart




Scikit-learn is a library that provides machine learning tools in Python. The figure above shows a cheat-sheet flowchart for deciding on the ML algorithm based on the amount of data available, type of use-case, presence of absence of labelled data and the type of prediction needed.

The amount of data is an important factor in ML as the training process depends on the presence of large amount of data to actually be accurate. If the available datasets has a small number of examples of the selected features, it does not make sense to continue the ML project as valuable insights are unlikely. The next major choice is between predicting a category (classification) and a quantity (regression). An example related to weather would mean trying to predict if the next day would be hot or cold (category) versus trying to predict the actual temperature (quantity).

After that, the choice for supervised or unsupervised algorithms is made based on the availability of the labelled data. For category prediction problems, classification algorithms are used when data is labelled while clustering algorithms are used when it is not. In the same way for quantity prediction, regression algorithms are used when data is labelled and dimensionality reduction algorithms like principal component analysis are used otherwise.

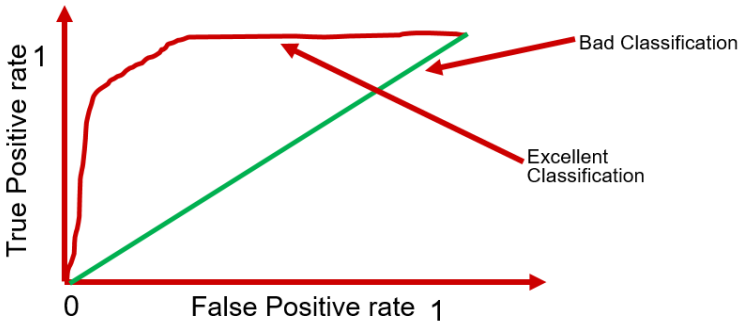
## Evaluating Machine Learning Algorithms

It is important to evaluate the performance of ML algorithms once they've been trained so that the outputs can be contextualized and the uncertainty around them can be reduced for the end-users. This process will allow the users to define what success looks like so that the models can be compared to other approaches and strategies for improvement can be considered.

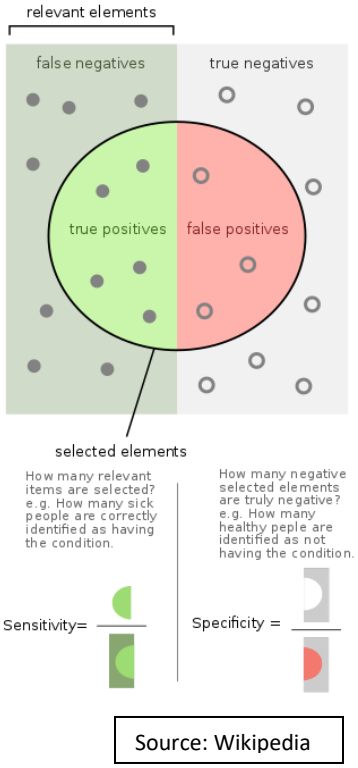
### Performance Metrics for Classification

When predictions are made based on categories, some of the commonly used performance metrics for classification are:

1. **Classification Accuracy:** This is the number of correct predictions made as a ratio of all predictions made, expressed as a percentage.
2. **Logarithmic Loss:** This measures the uncertainty of the prediction with regard to the true class and a lower value implies that a model is better at classification.
3. **Area under ROC Curve:** The area under the Receiver Operating Characteristic (ROC) curve is a measure of how well a parameter can distinguish between two groups.




a. The curve is generated by plotting the true positive rate (TPR), also known as the sensitivity, against the false positive rate (FPR) that is also called specificity. The sensitivity (TPR) measures the proportion of actual positives that are correctly identified as such, while the specificity (FPR) is calculated as the ratio between the numbers of negative events wrongly categorized as positive and the total number of actual negative events.
The adjacent image shows a visual representation of these two concepts.




Source: Wikipedia

b. The Area under the curve (AUC) can be thought of as a measure of the probability that the model ranks a random positive example more highly than a random negative example. AUC ranges in value from 0 to 1. A model whose predictions

are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.

4. **Confusion Matrix:** The confusion matrix is used to provide an easy to read visualization of the True Negative (a), False positive (b), False Negative (c) and True positive (d).

| Confusion Matrix | | Predicted | |
|---|---|---|---|
| | | Negative | Positive |
| Actual | Negative | a | b |
| | Positive | c | d |

   a.
   b. These terms can be defined as:
   c. True negative is the number of correct predictions that an instance is negative
   d. False positive is the number of incorrect predictions that an instance is positive
   e. False negative is the number of incorrect of predictions that an instance is negative
   f. True positive is the number of correct predictions that an instance is positive

## Performance Metrics for Regression

When a quantity is predicted or forecasted, some of the commonly used performance metrics for regression are:

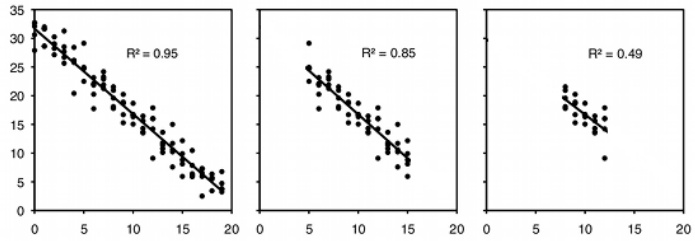
1. **Mean Absolute Error (MAE):** The MAE is the average over the test sample of the absolute differences between prediction and actual observation. It is calculated as:

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n},$$

2. **Root Mean Squared Error (RMSE):** The RMSE is the square root of the average of squared differences between prediction and actual observation. It is calculated as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(|y_i - x_i|)^2}{n}}.$$

3. $\mathbf{R^2}$ **value**: The R-squared is the percentage of the response variable variation that is explained by a linear model. It ranges from 0 where the data is completely random and completely unexplained by the linear model to 1 where the linear model is identical to the data. It is also commonly stated as a percentage, obtained by simply multiplying the r-squared value by 100.




Source: Handbook of Biological Statistics

## Validation Techniques

Validation techniques in machine learning are used to get the error rate of the ML model, which can be considered as close to the true error rate of the population. In real-world scenarios, we work with samples of data that may not be a true representative of the population. This is where validation techniques come into the picture. Some common techniques are:

- **Resubstitution validation:** The model is trained with all available data and then tested on the same data.
- **Hold-out validation:** The data is split into two different datasets labeled as a training and a testing dataset. This can be a 60/40 or 70/30 or 80/20 split. After that, the model is trained on the training set and then tested on the testing dataset.
- **K-fold cross-validation:** In K-Fold CV, the data is partitioned into $k$ segments. Subsequently $k$ iterations of training and validation are performed such that within each iteration a different fold of the data is held-out for validation while the remaining $k - 1$ folds are used for learning. Upon completion, $k$ samples of performance metric will be available.
- **Leave-One-Out Cross Validation (LOOCV):** This method is a special case k-fold cv, where all the data except for a single observation are used for training and the model is tested on that single observation

## Exercise 1: Facies/Lithology classification without interpretations- using log data

Before we start doing this exercise, it is important to understand the importance of facies and lithology. We will also discuss Principal Component Analysis, commonly known as PCA and unsupervised machine learning with regards to lithology.

Facies and lithology can be used interchangeably for this exercise, but facies is a geologic term whereas lithology is a more general term for the properties of rocks. Let us define those two terms first:

**Facies:** Facies, as a geologic term, are a way to distinguish bodies of rock into mappable units in terms of physical characteristics, composition, formation, or various other attributes. Facies are used to establish different units of rock from adjacent units within a contiguous body of rock by physical, chemical, or biological means. Facies marked an important development in the concept of stratigraphy because compiled facies can generate a succession that can give insight into an assortment of different process and systems that acted on or within the region and rock record. When modeled, facies can give insight into regional biologic and ecologic activity, water chemistry and properties, igneous events, sedimentary processes, climate records, and tectonics movement. (source: wiki.seg.org)

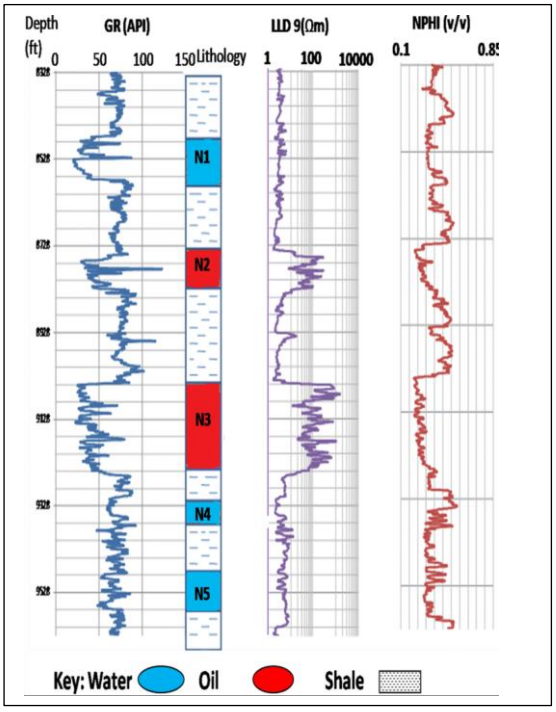
**Lithology:** Lithology is the general characteristics of sediments, rocks, and rock types present in a stratigraphic division of earth. In other words the study of rocks and their formation is called lithology. It helps in understanding and describing the physical characteristics of rock units such as their color, grain size, texture or composition. Lithology also helps in understanding porosity, permeability, water saturation, etc., and other petrophysical properties of rocks. (source: petropedia.com)

From the above definitions, we can see that both words can be used in synonymously for our exercise purpose.

Now, we should know why lithology/ facies identification is important in hydrocarbon exploration. For hydrocarbons to be accumulated they require reservoir, cap rock, source rock and migration path. Without any of these elements, oil and gas cannot be found inside the mother earth. If we need to find oil (or gas), we should look for a reservoir rock, that means a rock with certain reservoir properties like, porosity and permeability values within certain range that determine the rock type as a particular lithology (e.g. sand or shale).

The identification of a bed's lithology is fundamental to all reservoir characterization because the physical and chemical properties of the rock that holds hydrocarbons and/or water affect the response of every tool used to measure formation properties. Understanding reservoir lithology is the foundation from which all other petrophysical calculations are made. (source: petrowiki)

As mentioned it is the fundamental responsibility of any geoscientists, especially petrophysicists, to identify and delineate those rock units as different facies from the recorded well-log data. The process is called petrophysical interpretation. The below figure shows a typically interpreted petrophysical data or well-log data for lithology identification.




In this exercise, we will use the well-log data where no petrophysical interpretation is available and objective is to identify the different facies / lithology classes.

As there will be no previously interpreted log data for this exercise, this methodology will be known as unsupervised learning problem which is defined as:

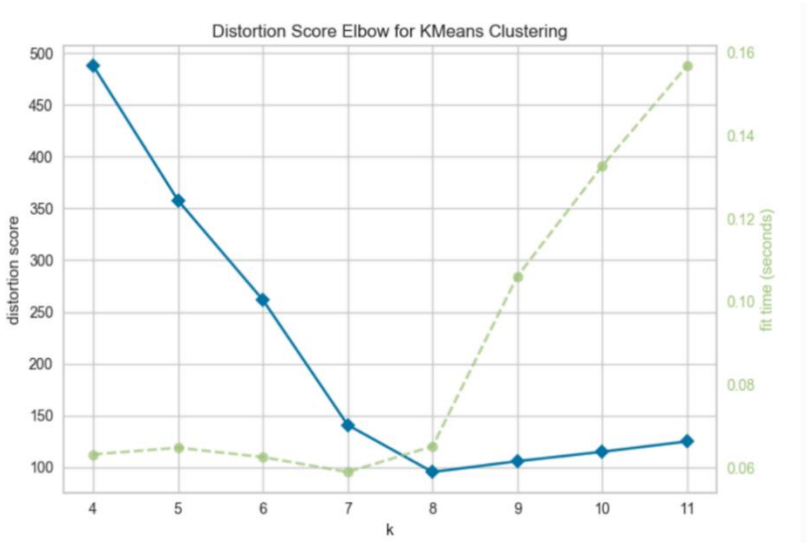
**Unsupervised learning** is the training of machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data. (Source: geeksforgeeks.org)

These types of problem can be solved using different algorithms and techniques. We will use principle component analysis (PCA), k-means clustering and elbow-plot technique. Let us discuss about the basics of those techniques and algorithms. During the exercise, we will learn more about them.

**K-means Clustering**: *K*-means clustering is a type of unsupervised learning that is used with unlabeled data. The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*. The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided. Data points are clustered based on feature

similarity. The *K*-means clustering algorithm is used to find groups which have not been explicitly labeled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets. Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the correct group.

**Elbow-plot Technique**: The *K*-Elbow Visualizer implements the "elbow" method to help data scientists select the optimal number of clusters by fitting the model with a range of values for *K*. If the line chart resembles an arm, then the "elbow" (the point of inflection on the curve) is a good indication that the underlying model fits best at that point. The below figure shows a representative elbow plot.




(source: scikit-yb.org)

## Exercise 2: PCA plus Facies/Lithology classification using log data

This exercise will be the continuation of Ex. 1. During last exercise, we have seen how log curves can be used for lithology or facies classification using machine learning algorithms. In the past exercise, we saw that there we had to use unsupervised classification techniques because no manual interpretation existed. Now, if an experienced petrophysicist or a geologist has already interpreted a few of the available log data to identify the lithology classes probable for that particular geological setup. The problem is if we can replicate similar interpretation using machine-learning algorithms?

In this case, the problem is classified as supervised learning. Now, let us discuss what is supervised learning and how it is different from the algorithms used in the previous exercise.

**Supervised-learning:** Supervised learning indicates the presence of a supervisor as a teacher. Basically, supervised learning is a type of learning in which we teach or train the machine using data which is well labeled that means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples (data) so that supervised learning algorithm analyses the training data (set of training examples) and produces a correct outcome from labeled data.

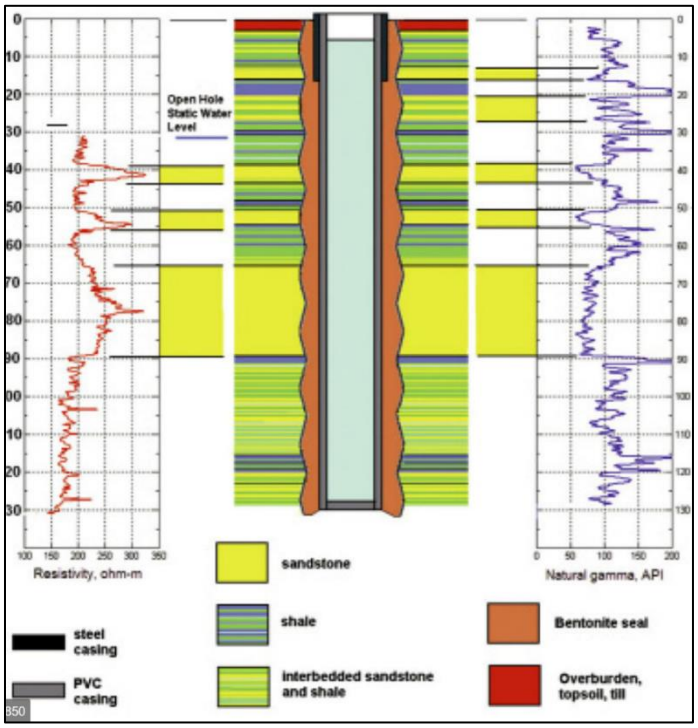
For instance, suppose you are given a basket filled with different kinds of fruits. Now the first step is to train the machine with all different fruits one by one like this:

- If shape of object is rounded and depression at top having color Red then it will be labelled as –Apple.
- If shape of object is long curving cylinder having color Green-Yellow then it will be labelled as –Banana.

(source: geeksforgeeks.org/supervised-unsupervised-learning)

Here the name of the fruits with their distinguishing characteristics are already provided during the training of the machines. This is the reason why this technique is called supervised learning.

Similarly, in our exercise, the well logs will be associated with a few interpretation of lithology, say, sand or shale. On the basis of those interpreted classes the algorithms train the machine to predict different lithology. The interpretation looks like similar to the below image:



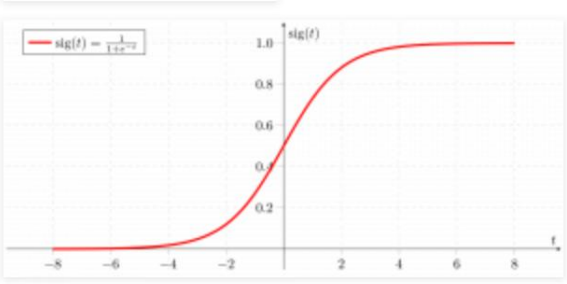

The lithology symbols can be digitized with a numerical value, say, sandstone as 0, shale as 1, and so on so forth and can be used in the supervised learning.

In this exercise, we will be using logistic regression, random forest algorithms, confusion matrix and F-score matrix techniques. So, let us discuss about these algorithms.

**Logistic regression:** Logistic regression is a supervised classification algorithm. In a classification problem, the target variable (or output), y, can take only discrete values for given set of features (or inputs), X.

Contrary to popular belief, logistic regression IS a regression model. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as "1". Just like linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function.

$$g(z) = \frac{1}{1+e^{-z}}$$



Logistic regression becomes a classification technique only when a decision threshold is brought into the picture. The setting of the threshold value is a very important aspect of Logistic regression and is dependent on the classification problem itself.

The decision for the value of the threshold value is majorly affected by the values of precision and recall. Ideally, we want both precision and recall to be 1, but this seldom is the case.

| Confusion Matrix | | Predicted | |
|---|---|---|---|
| | | Negative | Positive |
| Actual | Negative | a | b |
| | Positive | c | d |

**Confusion Matrix:** A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix.

Below is the process for calculating a confusion Matrix.

i. You need a test dataset or a validation dataset with expected outcome values.
ii. Make a prediction for each row in your test dataset.
iii. From the expected outcomes and predictions count:
   a. The number of correct predictions for each class.
   b. The number of incorrect predictions for each class, organized by the class that was predicted.

These numbers are then organized into a table, or a matrix as follows:

**Expected down the side:** Each row of the matrix corresponds to a predicted class.

**Predicted across the top:** Each column of the matrix corresponds to an actual class.

The counts of correct and incorrect classification are then filled into the table.

The total number of correct predictions for a class go into the expected row for that class value and the predicted column for that class value. In the same way, the total number of incorrect predictions for a class go into the expected row for that class value and the predicted column for that class value.

**F-score techniques:** The F score, also called the F1 score or F measure, is a measure of a test's accuracy. The F score is defined as the weighted harmonic mean of the test's precision and recall. This score is calculated according to:

$$F_1 = \left( \frac{\mathrm{recall}^{-1} + \mathrm{precisior}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\mathrm{precisior} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

With the precision and recall of a test taken into account. Precision, also called the positive predictive value, is the proportion of positive results that truly are positive. Recall, also called sensitivity, is the ability of a test to correctly identify positive results to get the true positive rate. The F score reaches the best value, meaning perfect precision and recall, at a value of 1. The worst F score, which means lowest precision and lowest recall, would be a value of 0.

The F score is used to measure a test's accuracy, and it balances the use of precision and recall to do it. The F score can provide a more realistic measure of a test's performance by using both precision and recall. The F score is often used in information retrieval for measuring search, document classification, and query classification performance.

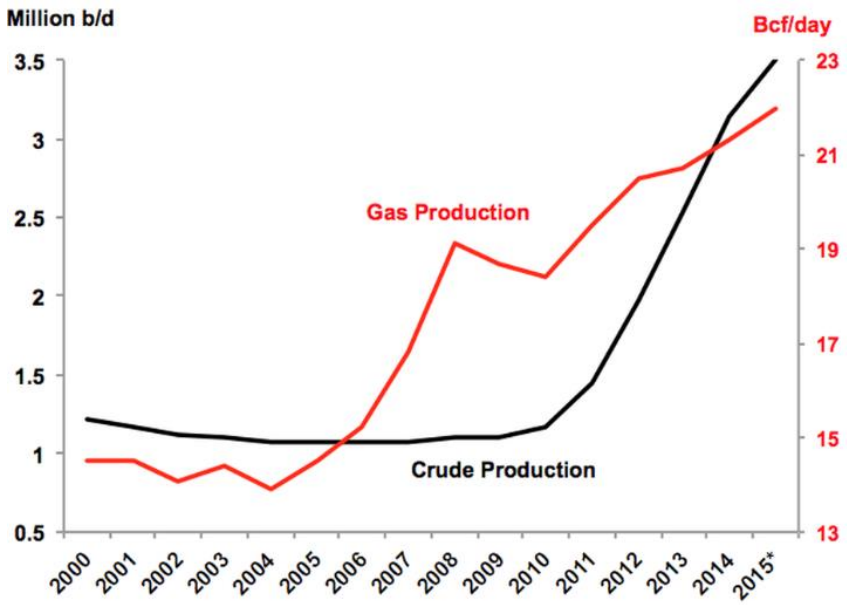
## Exercise 3: Short term production forecasting using ML

**Production** is the process of extracting the hydrocarbons and separating the mixture of liquid hydrocarbons, gas, water, and solids, removing the constituents that are non-saleable, and selling the liquid hydrocarbons and gas. Production sites often handle crude oil from more than one well. Oil is nearly always processed at a refinery; natural gas may be processed to remove impurities either in the field or at a natural gas processing plant.
(Source: http://www.oilandgasbmps.org/resources/development.php)

Crude oil production is defined as the quantities of oil extracted from the ground after the removal of inert matter or impurities. It includes crude oil, natural gas liquids (NGLs) and additives. This indicator is measured in thousand tonne of oil equivalent (toe).Crude oil is a mineral oil consisting of a mixture of hydrocarbons of natural origin, yellow to black in colour, and of variable density and viscosity. NGLs are the liquid or liquefied hydrocarbons produced in the manufacture, purification and stabilisation of natural gas. Additives are non-hydrocarbon substances added to or blended with a product to modify its properties, for example, to improve its combustion characteristics (e.g. MTBE and tetraethyl lead).Refinery production refers to the output of secondary oil products from an oil refinery.
(Source: https://data.oecd.org/energy/crude-oil-production.htm)

In Oil and Gas industry, we have collected lot of production data. The industry has created many tools and software, which help in calculating production forecasting. In this exercise we will see how a Machine learning and Artificial Intelligence method is applied in predicting the production.

A typical production graph looks like:




Source: https://www.forbes.com/sites/judeclemente/2015/05/17/the-importance-of-texas-oil-and-natural-gas-surge/#eb3ac915c9b6

The various sensors captures many parameters while production. Some of the parameters used in this exercise are:

1. WaterRate
2. CasingHeadPressure
3. TubingHeadPressure
4. PumpSpeed
5. Torque

Using the historical data of these parameters, we will predict the Gas Production (GasRate).

**Regression** is a technique used to model and analyze the relationships between variables and often times how they contribute and are related to producing a particular outcome together. A linear regression refers to a regression model that is completely made up of linear variables.

You might have observed that we have mentioned linear regression just now. That means there are different types of regression, namely:

- Linear Regression
- Polynomial Regression
- Ridge Regression
- Lasso Regression
- Elastic-net Regression

In this exercise, we will focus on linear and polynomial regression as the model validation demands. An advanced technique called neural network is also very common in machine learning problem.

## Exercise 4: Text Analytics

Any scientific studies demand consultation of scientific literatures and journals for further understanding of the problems and innovations in that particular domain. Hydro-carbon exploration is not an exception. Before we start discussing about the problem, let us spend a few words on different types of data available in any organization.

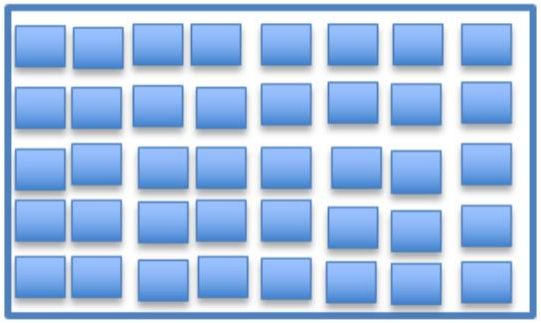
In a boarder sense, the data can be divided into two groups:

1. Structured data
2. Un-structured data

Let us define them, in other words, we need to understand what unstructured or structured data is. We will also try to figure out whether geophysical literature data comes under structured category or unstructured data category? What are the usages of those different data types.

**Structured Data:** Structured data is data that has been organized into a formatted repository, typically a database, so that its elements can be made addressable for more effective processing and analysis. A data structure is a kind of repository that organizes information for that purpose. (source: techtarget.com)

For simple understanding, we can say that the data with high degree of organization is called structured data.



In our day-to-day activities, we encounter a lot of structured data. Earlier we have used log files (LAS format), these are structured data. Similarly, spreadsheets, tables, and relational database data, like OpenWorks™ can be good examples of structured data. These data are easy to used as they follow some rules for storing information.

**Unstructured Data:** Unstructured data (or unstructured information) is information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts as well. This results in irregularities and ambiguities that make it difficult to understand using traditional programs as compared to data stored in fielded form in databases or annotated (semantically tagged) in documents. (source: Wikipedia)

In other words, the data which is difficult to organize using traditional mechanism is called unstructured data.

Daily drilling reports (DDR), webpages, emails, media files like, pictures, video files, text files, pdf files, presentation slides, etc. From the above examples, it is very clear that we use a lot of unstructured data everyday. Unfortunately, the machine does not understand the unstructured data easily.

Almost all the scientific literatures come into these unstructured formats, like pdf or word (text) files. A significant amount of time is spent to understand or extract the useful information from those literatures by reading. Here the reading is purely performed by a human. So, why not use the machine for this purpose? This exercise will show how similar information extraction, known popularly as **text mining** can be performed using a computer language, like, Python.

In this purpose, we will use NLTK, natural language toolkit, package in Python.

NLTK is a tool specially designed for natural language processing (NLP). NLP is basically an intersection of the following:

i. Computer science
ii. Linguistic and
iii. Machine Learning

In other words, Natural Language Processing is manipulation or understanding text or speech by any software or machine. An analogy is that humans interact, understand each other views, and respond with the appropriate answer. In NLP, this interaction, understanding, the response is made by a computer instead of a human. On the other hand, NLTK toolkit is one of the most powerful NLP libraries which contains packages to make machines understand human language and reply to it with an appropriate response. Tokenization, Stemming, Lemmatization, Punctuation, Character count, word count are some of these packages which will be discussed in this tutorial.

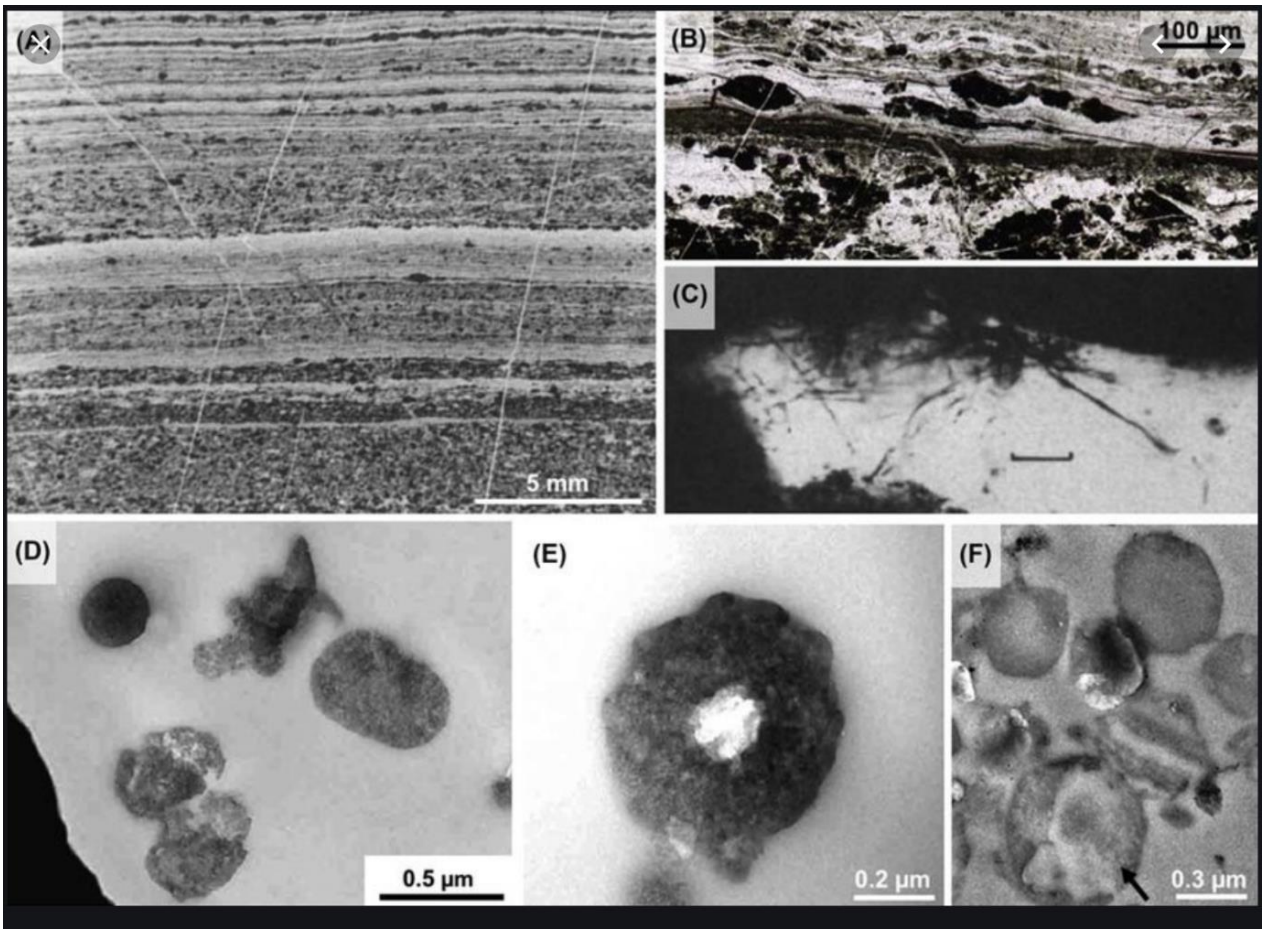
## Exercise 5: Fossil Identification through Computer Vision

This is an interesting exercise where we need to classify fossils using computer vision, meaning that the machine will recognize the different types of fossil images. Before we go into the details of the exercise, we need to know what the definition of a fossil is and why fossil recognition is required or important in oil and gas industry.

So, what is fossil? In geology, fossil can be defined as below:

**Fossils** are physical evidence of preexisting organisms, either plant or animal. The most common and obvious fossils are the preserved skeletal remains of animals. Other fossils, which are also evidence of past organisms, include leaf impressions, tracks and trails, burrows, droppings, and root casts.

For the exploration of hydro-carbon, fossils play a very important role, and we are more interested in identifying the micro-fossils which are basically very small in size and also happen to be in abundance, especially in marine rocks, which are the most common form of sedimentary rock in the crust of the Earth. The below figure shows a sediment section and a few micro-fossils.



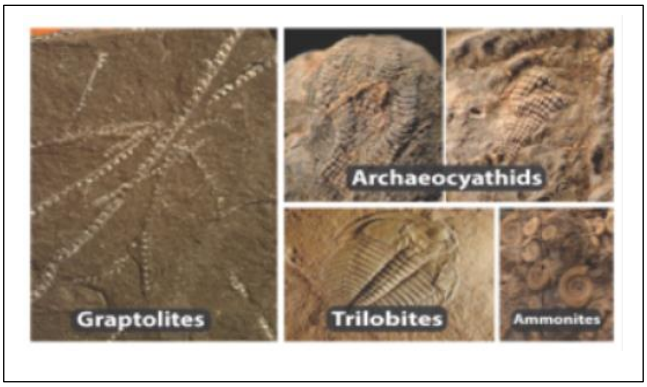

(credit: Keyron Hickman-Lewis)

**Importance of fossils:**

Fossils are used for bio-stratigraphic studies or analysis in oil and gas industry to understand the depositional environment and to denote the age of deposition for a particular rock unit. Biostratigraphy is the branch of stratigraphy which focuses on correlating and assigning relative ages of rock strata by using the fossil assemblages contained within them.

From the above statement, it is very clear that we need to know the different fossils preserved in a particular rock unit, so that the favourable conditions for hydro-carbon accumulation or generation can be identified as per the bio-marker fossils and the paleo-environmental conditions associated with those fossils.

In biostratigraphy, biozones are important. Biozones or biostratigraphic units are defined as bodies of strata that are characterized on the basis of their contained fossils. Ammonites, graptolites, archeocyathids and trilobites are index fossils that are widely used in biostratigraphy.



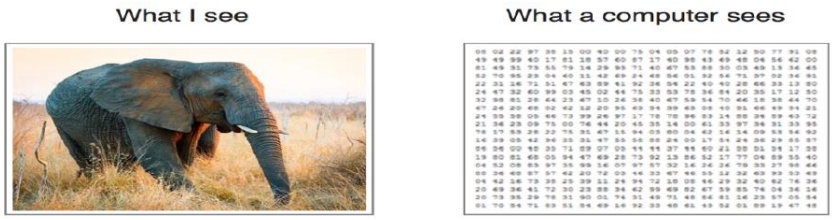

The figure shows a representative image of those index fossils, which we mentioned just now. In reality those fossils images are captured through microscopic sections of rock collected during drilling as core samples. Presence of index fossils confirm the geological period for deposition, as well as paleo environment.

The following points can be remembered in the context of Computer Vision:

- Computer vision enables computer to understand the content of images and videos.
- The goal of computer vision is to extract meaningful information from images or videos, such as, whether a certain object is present or not in a particular scene.
- Computer vision is not limited to pixel-wise operations; it can be complex. In actual, it is far more complex than image processing.
- Those complex operations can be summarized into feature detectors which can provide rich information about the contents of the image or video.
- The goal of machine learning is to optimize differentiable parameters so that a certain loss / cost function is minimized.
- Machine learning can be used in both image processing and computer vision but it has found more use in computer vision compared to other one.
  (source: quora.com/What-is-the-relation-between-machine-learning-image-processing-and-computer-vision)

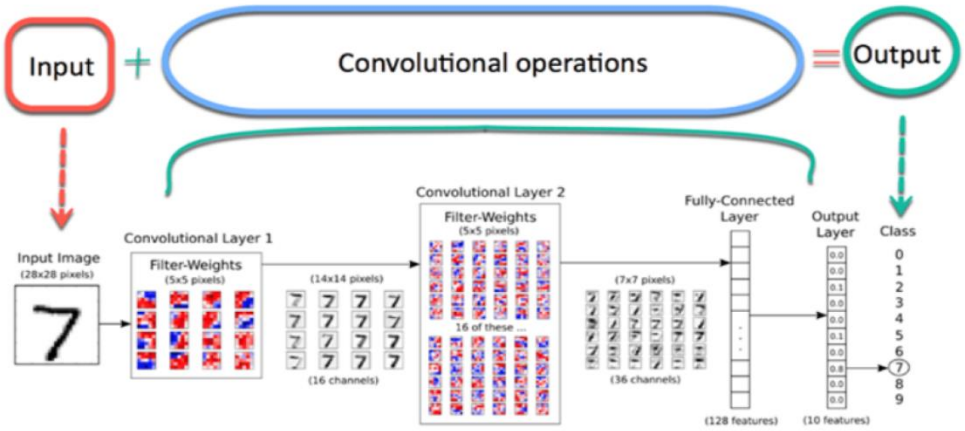
For this exercise, we will be using Orange software. Before that we need to know that computer vision problems will be solved through convolutional neural network, which is a common

machine learning algorithm. We have discussed earlier about artificial neural network (ANN). Here we will use Convolutional Neural Network (CNN) through Orange. CNN is a special architecture of ANN. The main task of image classification of the input image and the following definition of its class. This is a skill that people learn from their birth and are able to easily determine that the image in the picture is an elephant (see below figure). But the computer sees the picture quite differently.




Source : http://www.adobepress.com/articles/article.asp?p=2240988&seqNum=10

A representation of the CNN architecture is given below:




**Source:**https://www.guru99.com/convnet-tensorflow-image-classification.html#1 (https://www.guru99.com/convnet-tensorflow-image-classification.html#1)