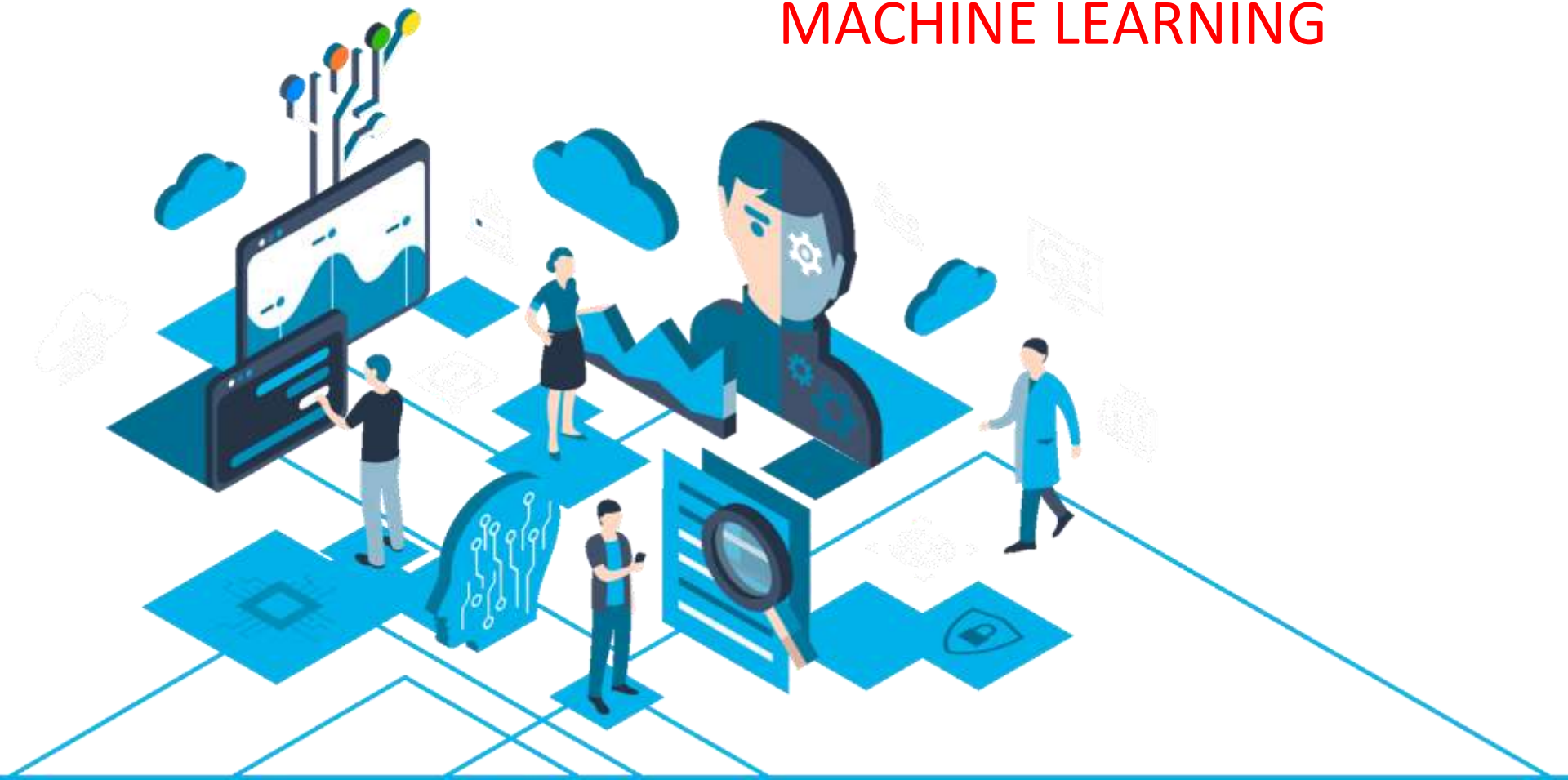
MACHINE LEARNING

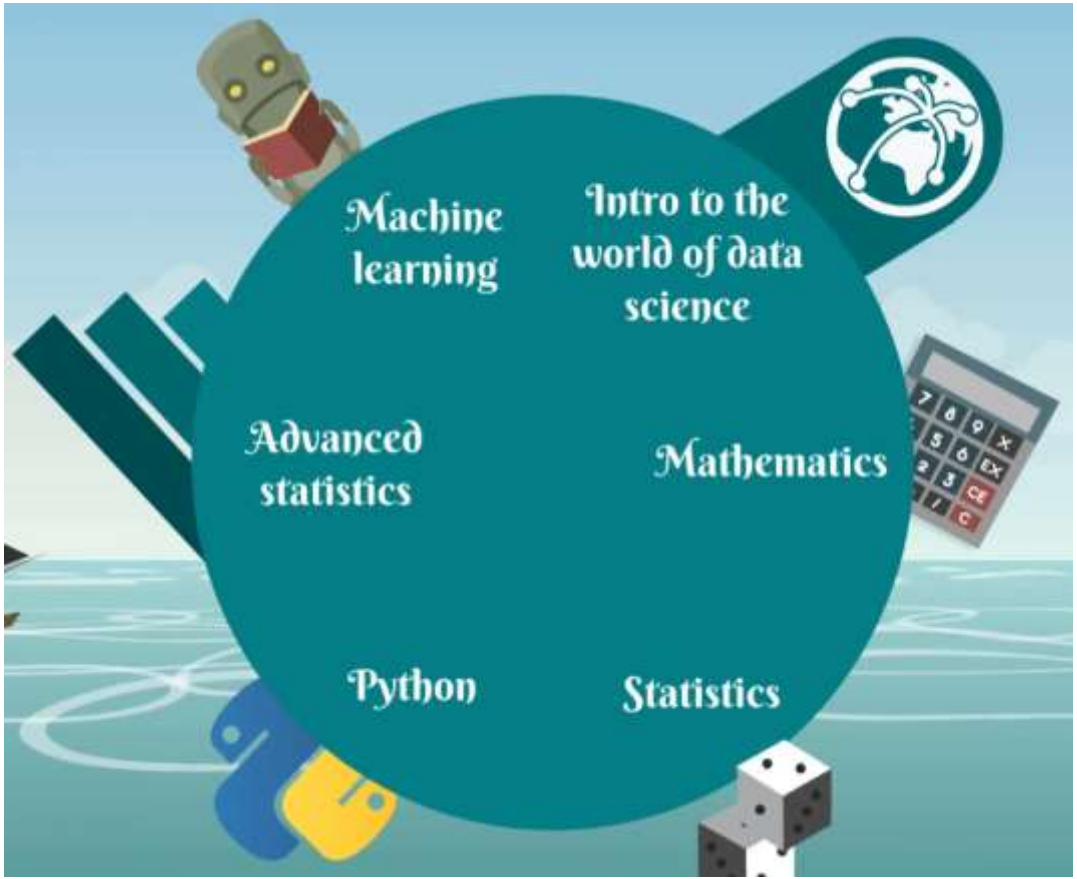


ML Is The Future

What is Data Science?

- **Data Science is the science of analysing raw data** using statistics and machine learning techniques with the purpose of drawing insights from the data.
- **Data Science** is used in many industries to allow them to make better business decisions, and in the sciences to test models or theories.
- This requires a process of inspecting, cleaning, transforming, modelling, analyzing, and interpreting raw data.

Important Disciplines Under Data Science



The Constant Evolution Of Data Science Industry

Statistician

Data Mining Specialist

Predictive Analytics Specialist

Data Scientist

- **Data science mainly needed for:**

  Better decision making (Whether A or B?)
  Predictive Analysis (What will happen next?)
  Pattern Discovery (Is there any hidden information in the data?)

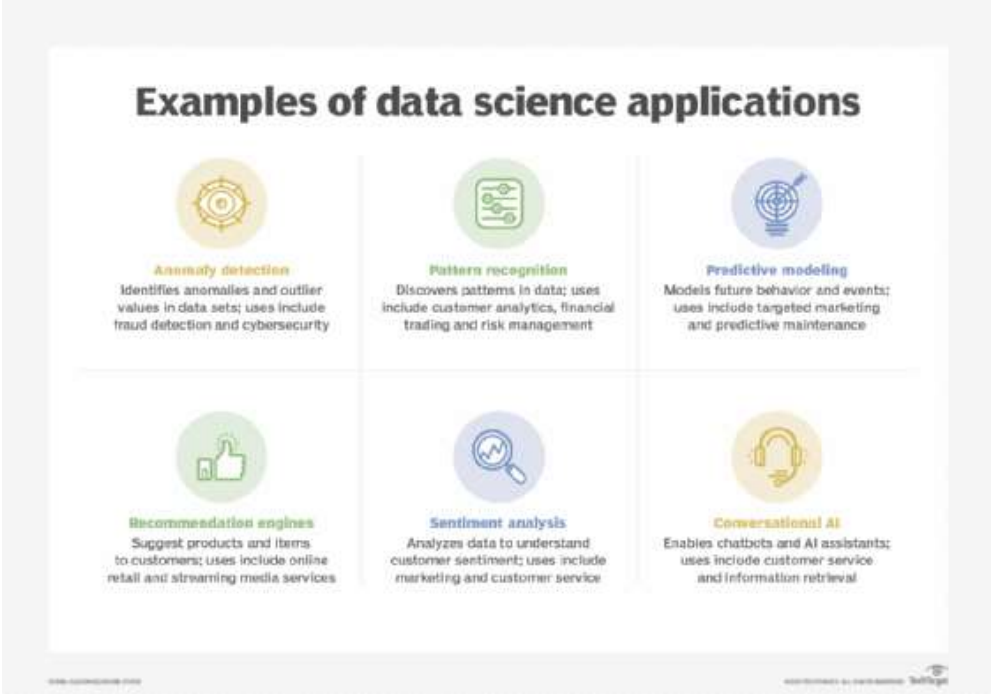
- **So data science is about**

  Asking right questions and exploring the data

  Modelling the data using various algorithms

  Finally, communicating and visualising results

- **Examples:**
  1) **Self-driving cars**
  2) **Airlines**
  3) **Logistic companies** like FedEx



The Various Data Science Disciplines

Not everyone in the field of Data Science is a Data Scientist!

Data Engineer

Data Engineers are software engineers who handle the design, building, integration of data from various data sources and also manage them.

Big Data Engineers

The set of engineers handle Data Warehousing process, by running the Extract-Transform-Load (ETL) procedure on data. They are also known as Big Data Engineers.

Big Data is data that contains greater variety arriving in increasing volumes and with ever-higher velocity ~ *Gartner*

Data Analyst

A Data Analyst is someone who processes and does statistical analysis on data to discover possible patterns, trends and also appropriately communicate the insights gotten for proper understanding.

Data Analysts are sometimes called "Junior Data Scientists" or "Data Scientists in Training"

Machine Learning Engineer

A Machine Learning (ML) Engineer is a software engineer that specializes in making data products work in production.

They are **involved in software architecture and design**; they understand and carryout practices like A/B testing (A/B testing is a user experience research methodology

Data Visualization Engineer

This is someone that tells visually stunning stories with data, create dynamic data visualizations to help businesses/customers make meaningful decisions in an interactive format.
They basically collaborate with Data Analysts and Data Scientists to make visualizations which effectively communicates the insights gotten from data to the business.

Data Scientist

A Data Scientist is an analytical data expert who has the technical skills to solve complex problems and the curiosity to explore what problems needs to be solved.

Data Scientists apply Statistics, Machine Learning, and analytical approaches to solve critical business problems.

A Data Scientist is also known as a mathematician, a statistician, a computer programmer and an analyst equipped with a diverse and wide-ranging skill set, balancing knowledge in different computer programming languages with advanced experience

Difference between Analysis and Analytics

Analysis

Consider you have a huge data set containing data of various types.

Instead of tackling the entire dataset and running the risk of becoming overwhelmed, you separated into easier to digest chunks and study them individually and examine how they relate to other parts and that's **analysis.**

One important thing to remember however is that **you perform analyses on things that have already happened in the past** such as using an analysis to explain how a story ended the way it did or how there was a decrease in the cells last summer.

All this means that we do analyses to explain how and or why something happened

Analytics

Analytics generally refers to the future instead of explaining past events.

It explores potential future ones.

Analytics is essentially the application of logical and computational reasoning to the component parts obtained in an analysis and in doing this you are looking for patterns in exploring what you can do with them in the future.

Here analytics branches off into two areas.

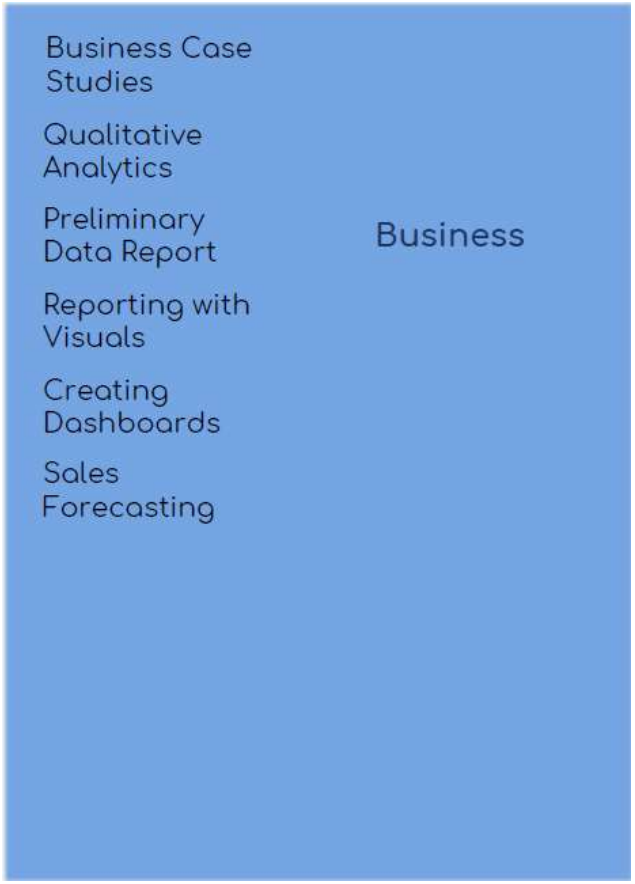
Qualitative analytics

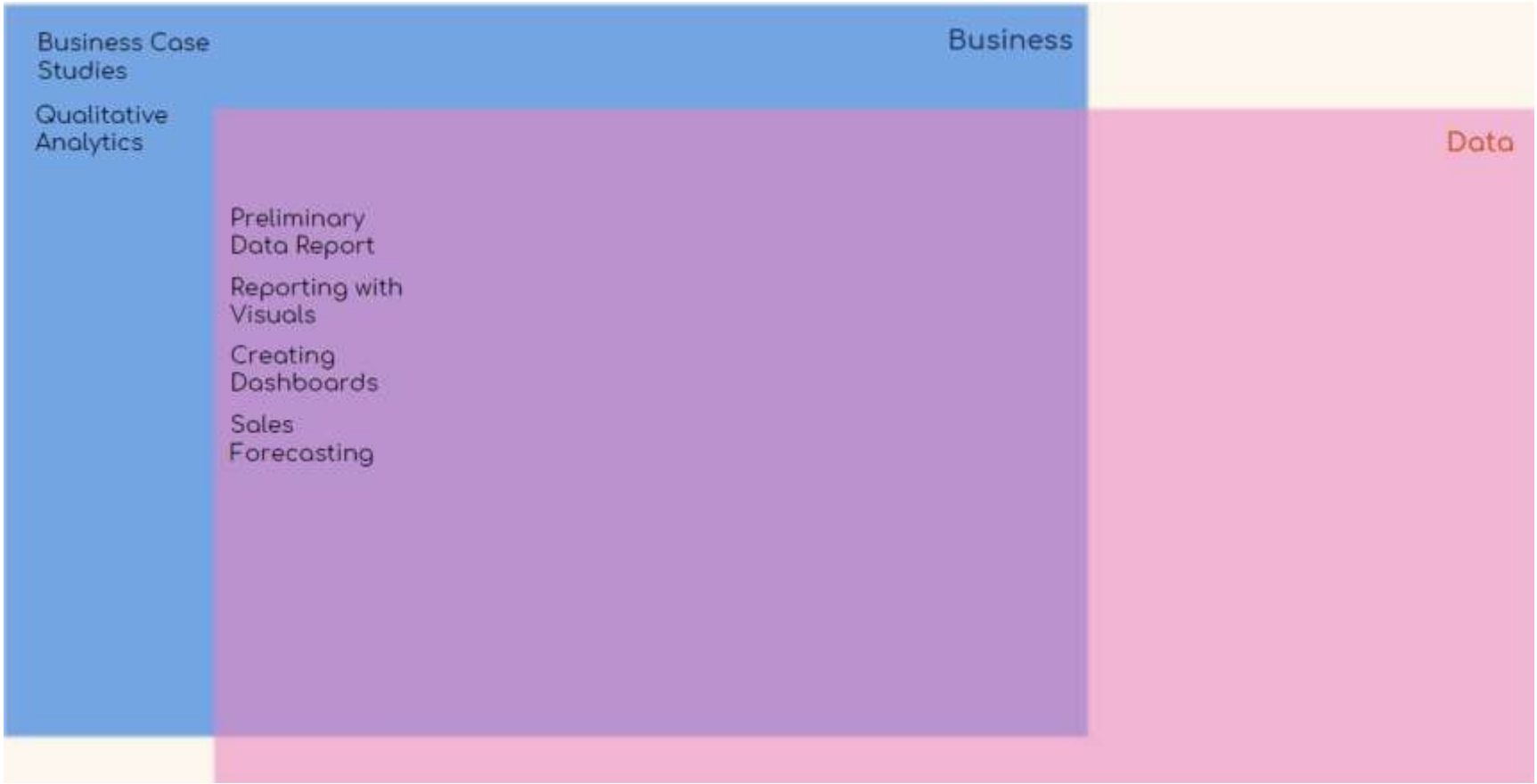
This is using your intuition and experience in conjunction with the analysis to plan your next business move

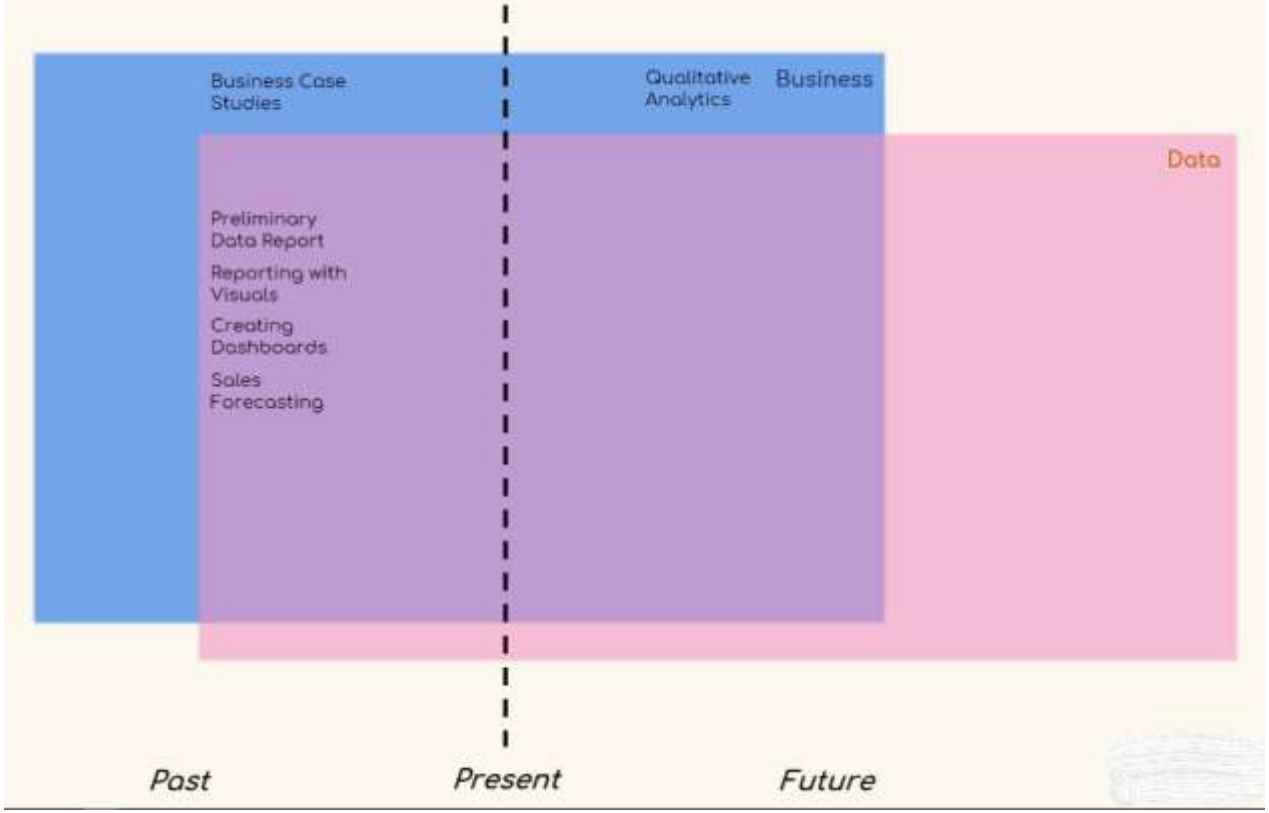
Quantitative analytics

This is applying formulas and algorithms to numbers you have gathered from your analysis.

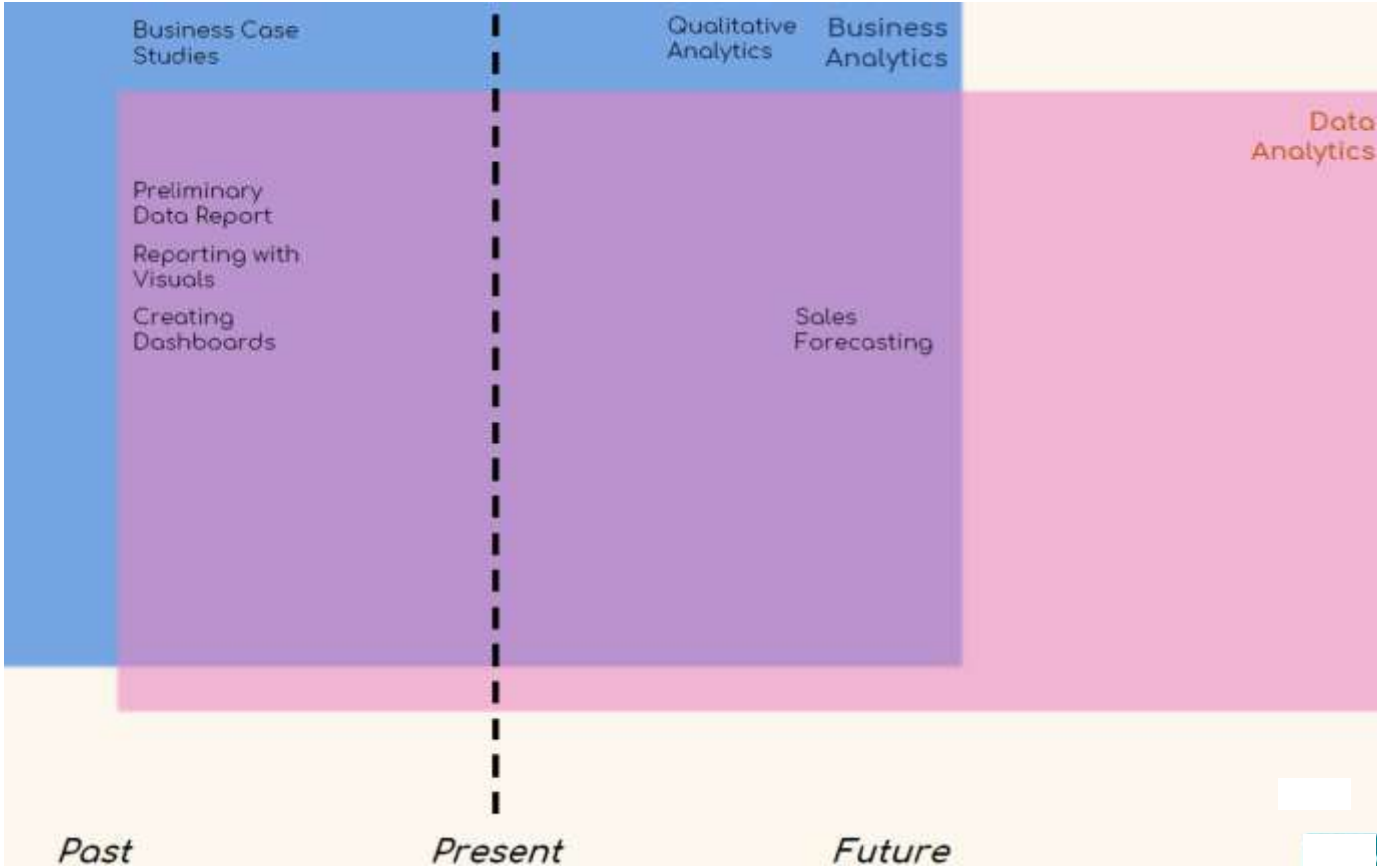
Business Analytics, Data Analytics, Data Science: An Introduction



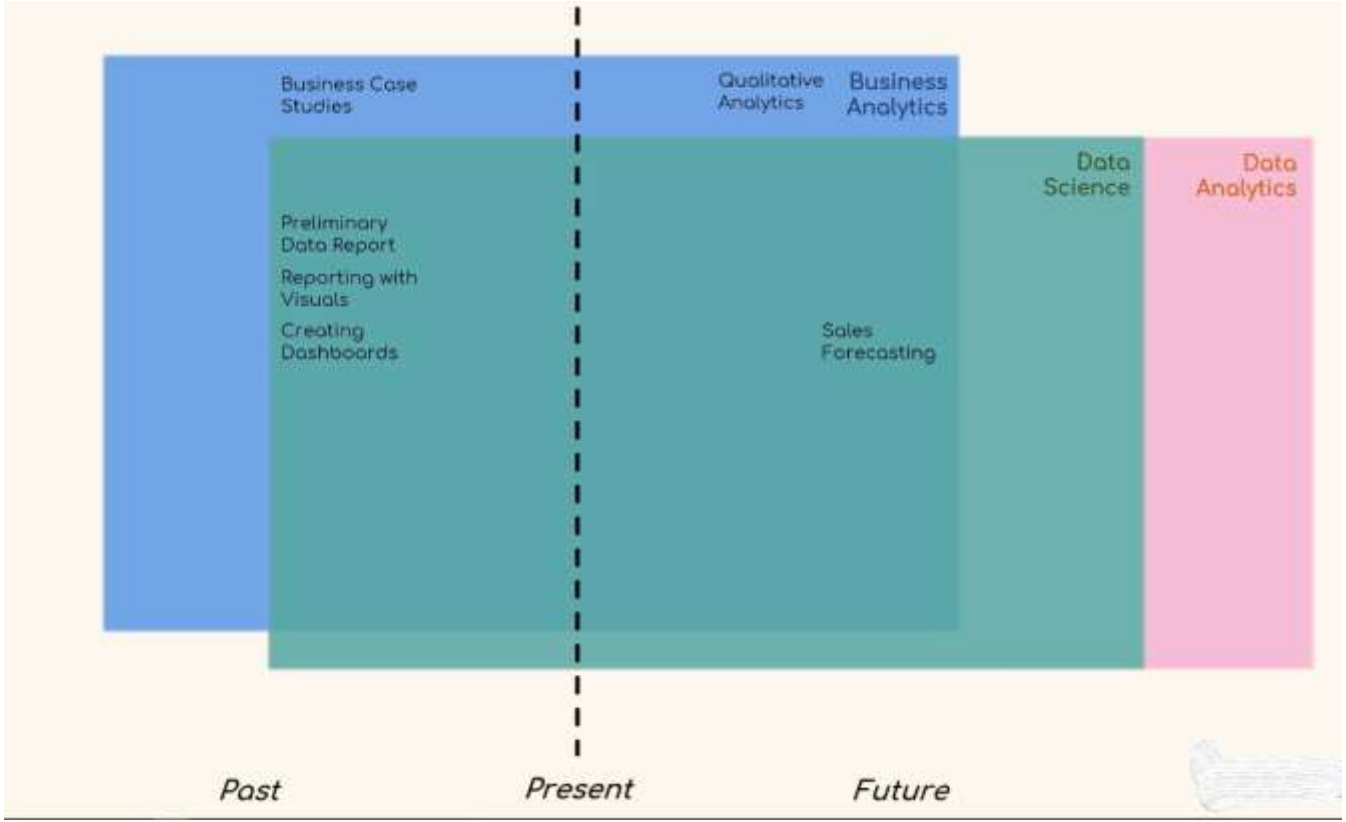

Business Case Studies
Business
Qualitative Analytics
Data
Preliminary Data Report
Reporting with Visuals
Creating Dashboards
Sales Forecasting



- **Qualitative analytics.**
- This includes working with tools that help predict future behaviour.
- Therefore must be placed on the right.
- In essence what we have now is qualitative Analytics which belongs to the area of business analytics.

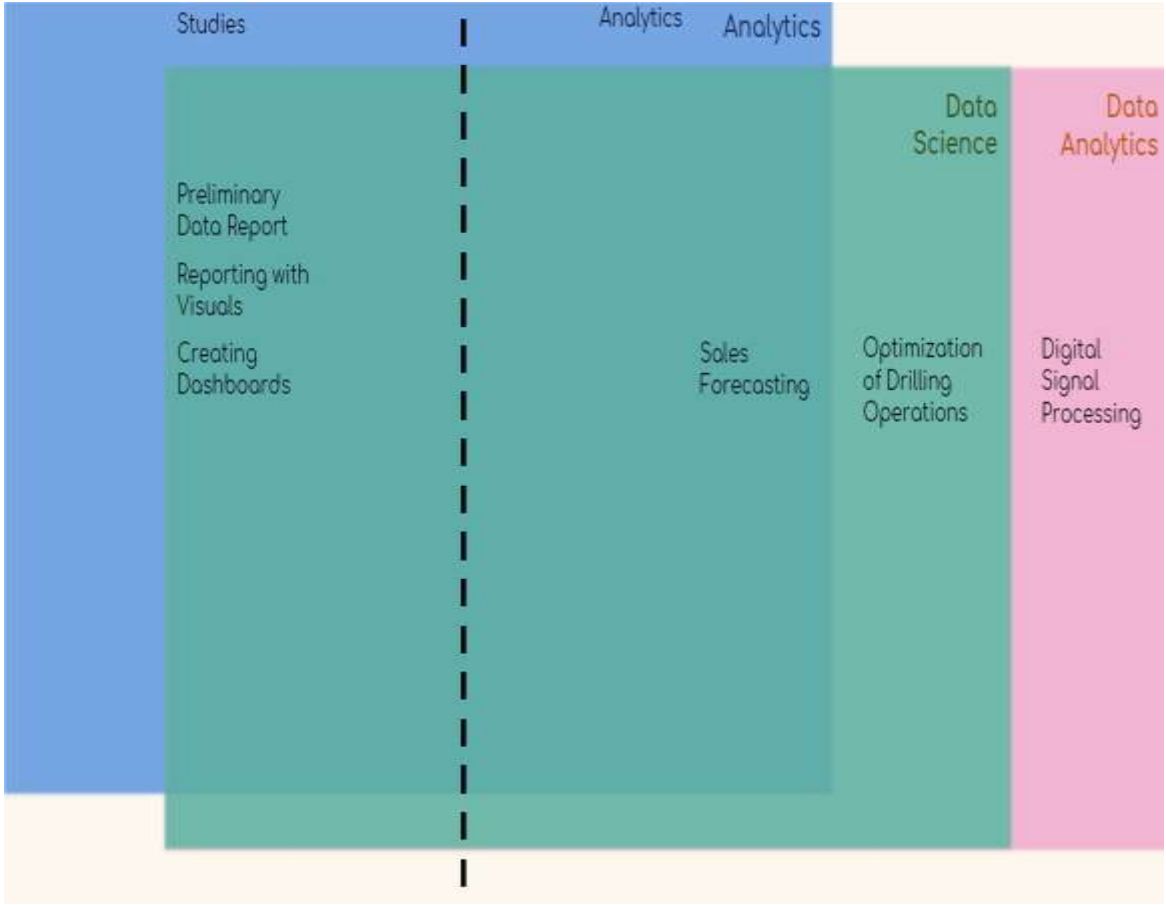


- **Sales Forecasting** though is a future oriented activity so we can move it to the right of the black line but not too much.
- It must still be long on the sphere of business.
- So it must be in the area where business analytics and data intersect.

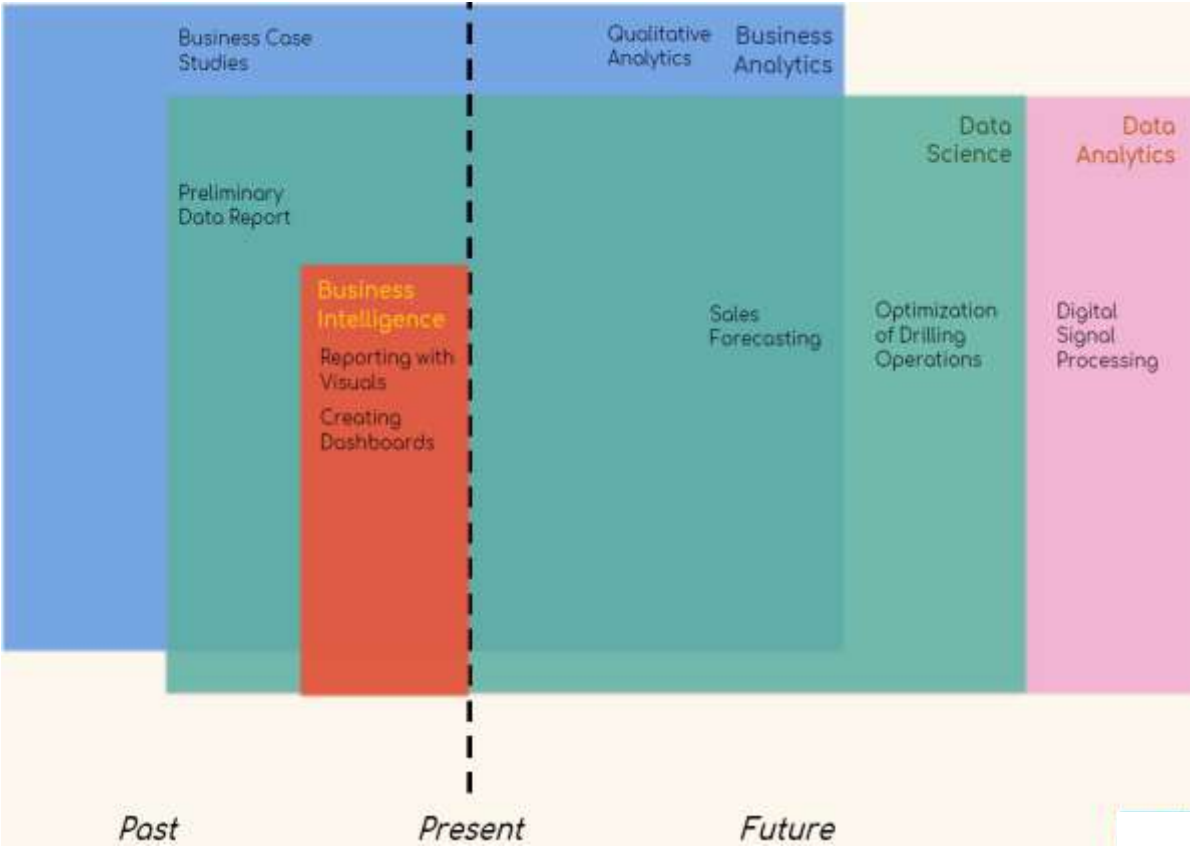


❑ **Data Science**

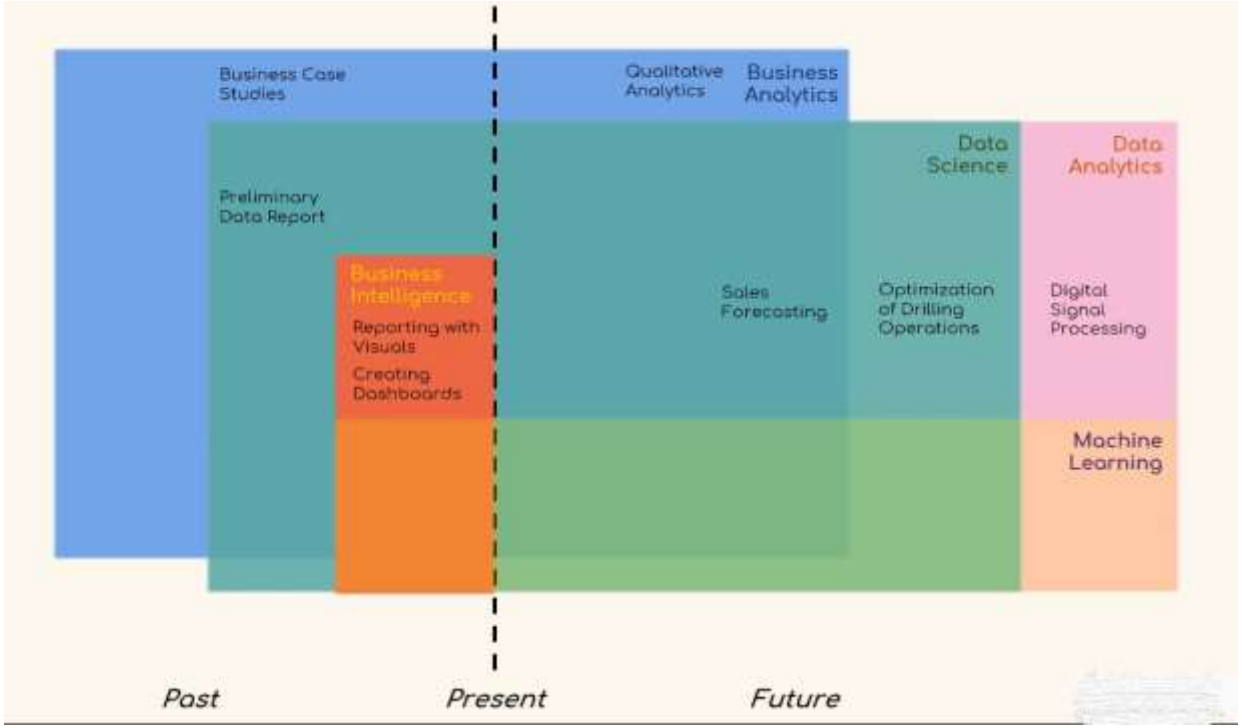
- The most sparkly of them all is data science.
- Data science is a discipline reliant on data availability while business analytics does not completely rely on data.
- However, data science incorporates part of data analytics mostly the part that uses complex mathematical statistical and programming tools.
- Consequently, this green rectangle representing data science on our diagram will not overlap with data analytics completely but it will reach a point beyond the area of business analytics.



- An example of a discipline that belongs to the field of data science and is considered data analytics but not business analytics is the **oil and gas industry and the optimization of drilling operations** (It aims to optimize weight on bit, bit rotation for obtaining maximum drilling rate as well as minimizing drilling cost).
- This is a perfect fit for this sub area data science can be used to improve the accuracy of predictions based on data extracted from various activities typical for drilling efficiency.
- Something that involves data analytics but neither data science nor business analytics can be **digital signal processing**.
- Digital signal is used to represent data in the form of discrete values which is an example of numeric data.
- Therefore data analytics can be applied

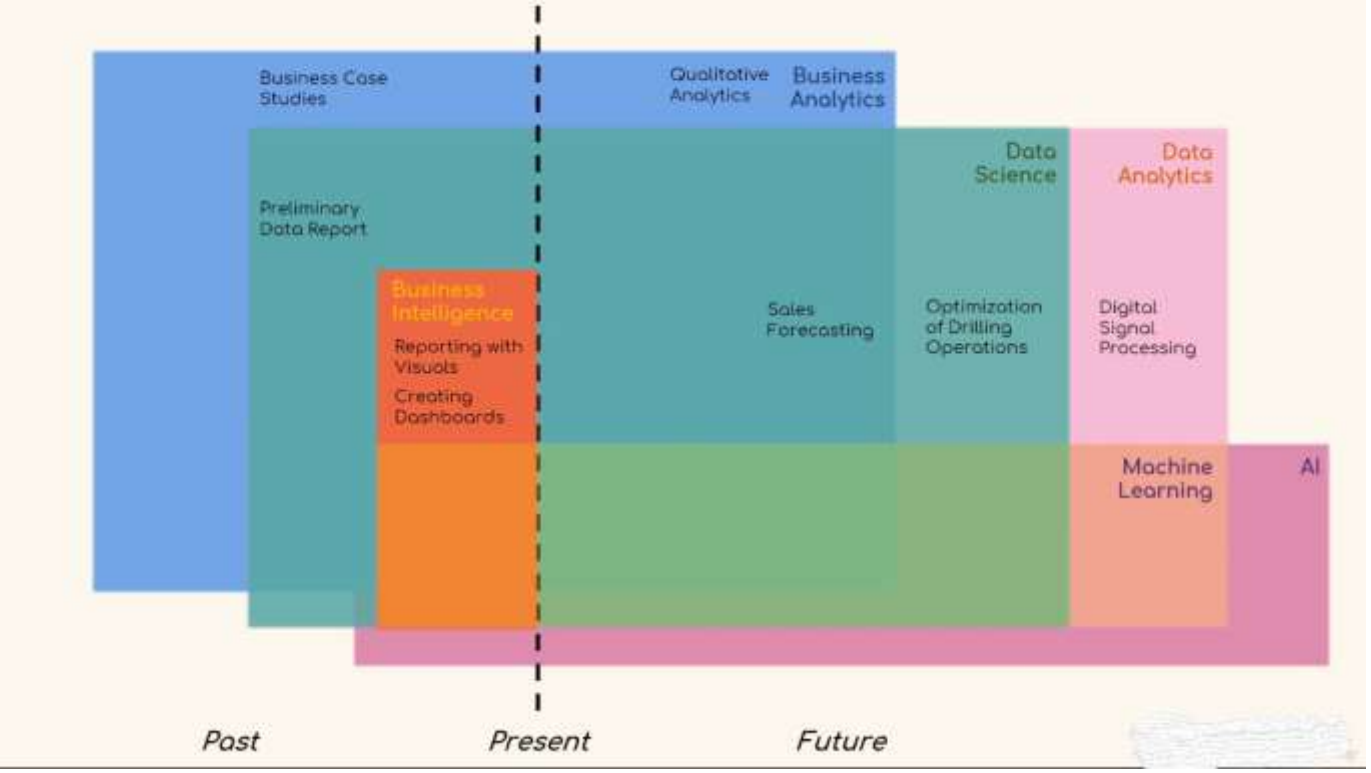


- The **business intelligence** or BI is the process of analysing and reporting historical business data after reports and dashboards have been prepared.
- They can be used to make an informed strategic and tactical business decisions by end users such as the general manager.
- **Business intelligence aims to explain past events using business data**.
- It must go on the left of the timeline as it deals only with past events and it must sit within the data science rectangle as a subfield business intelligence fits comfortably within data science because it is the preliminary step of predictive analytics.
- First you must analyse past data and extract useful insights using these inferences will allow you to create appropriate models that could predict the future of your business accurately.
- As with reporting and creating dashboards these are precisely what business intelligence is all about. So we will neatly place these two into the orange rectangle.

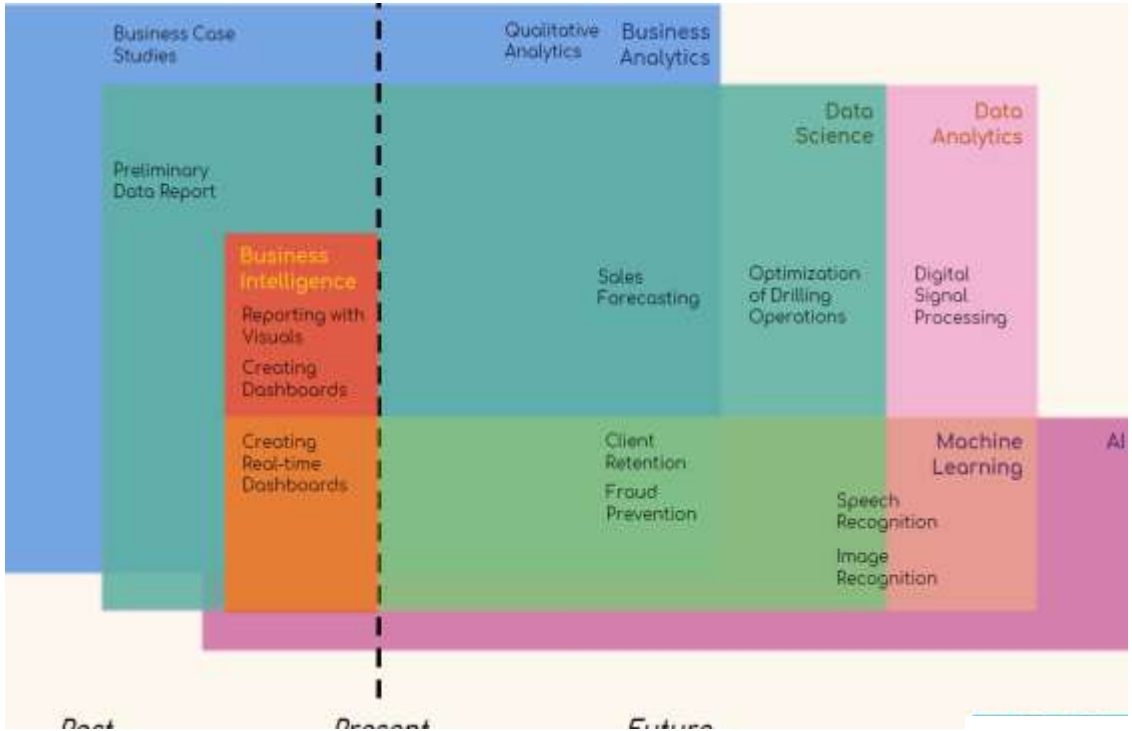


❑ Machine Learning

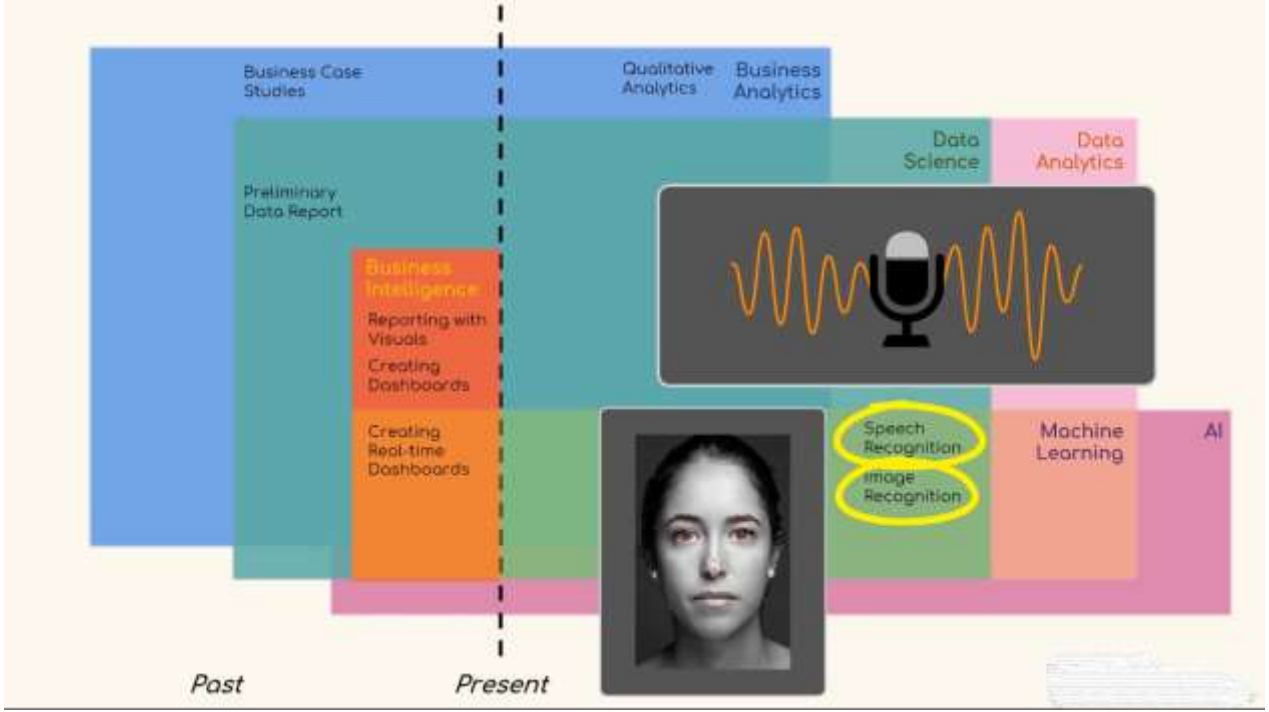
- The ability of machines to predict outcomes without being explicitly programmed to do so is regarded as machine learning.
- Expanding on this is about creating and implementing algorithms that let machines receive data and use this data to make predictions analyse patterns and give recommendations on their own.
- Machine learning cannot be implemented without data. Hence it should stay within Data analytics completely.
- By definition it is about simulating human knowledge and decision making with computers.



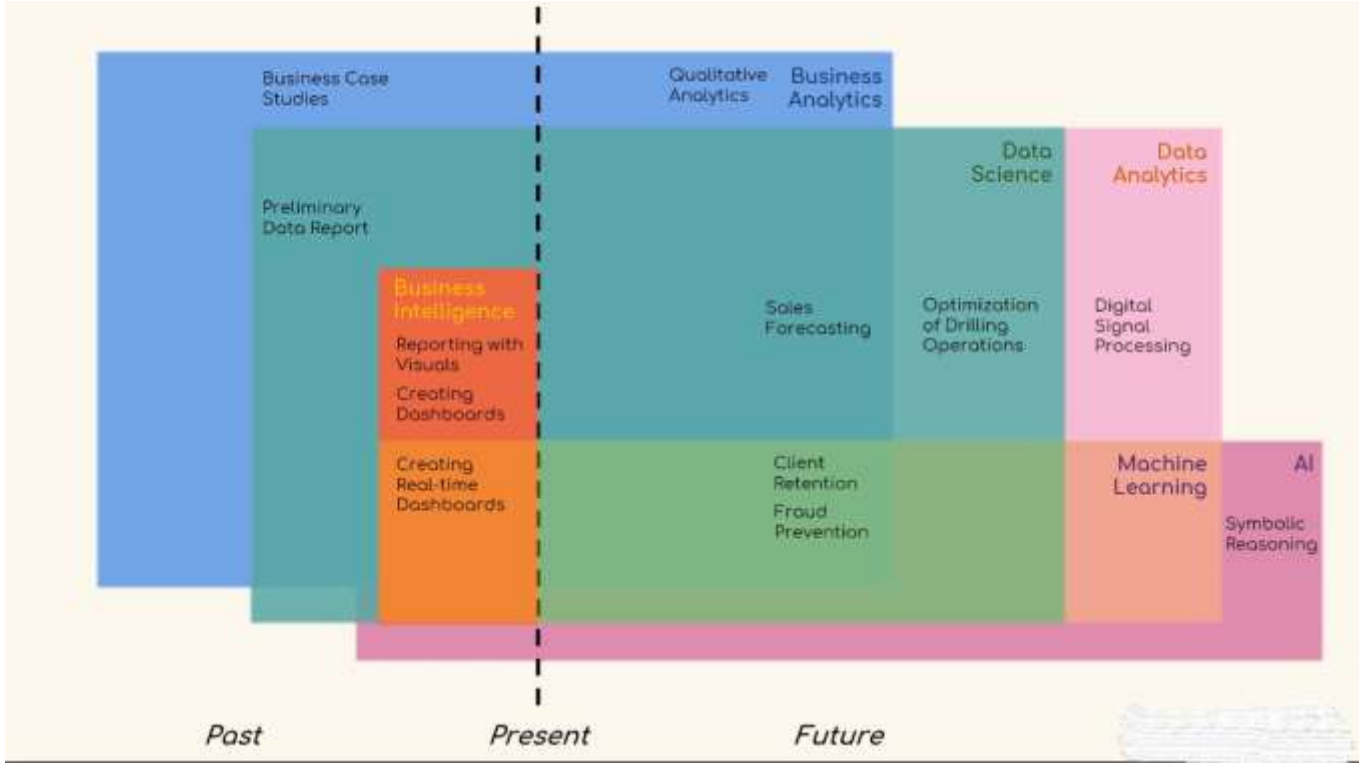
- We as humans have only managed to reach AI through machine learning the discipline we just talked about and as the data scientists we are interested in how tools from machine learning can help us improve the accuracy of our estimations.
- AI is beyond our expertise
- **Artificial intelligence is intelligence demonstrated by machines, as opposed to the natural intelligence displayed by humans or animals.**



- The **client retention**(process of engaging existing customers to continue buying products or services from your business) and **acquisition**(process of gaining new customers) are two typical business activities where machine learning is involved. It helps develop models that predict what a client's next purchase would be.
- For example since we could say data analytics and data science are applied in client retention and acquisition as well we can leave this term right over here.
- ML can be applied to **fraud prevention** as another example we can feed a machine learning algorithm with prior fraudulent activity data. It will find patterns which the human brain is incapable of seeing.
- Having a model which can detect such transactions or operations in real time it has helped the financial system prevent a



- When talking AI and ML usually **speech and image recognition** are among the most popular examples as they are already being implemented in products like Siri, Cortana, Google's assistant and more impressively self-driving cars.



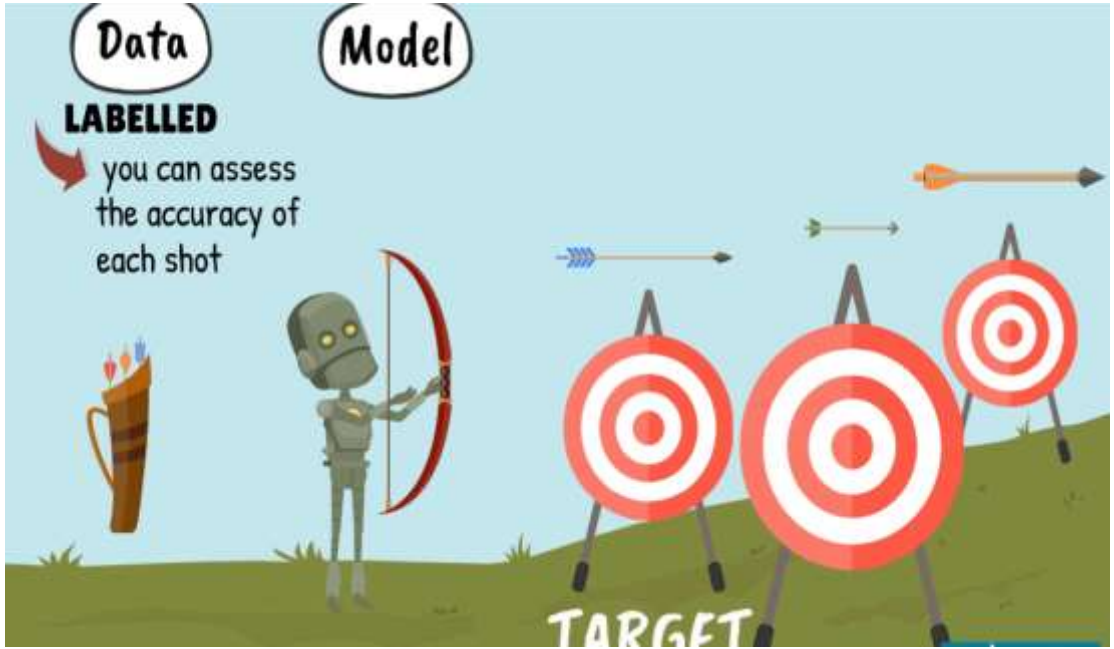
- Finally an example that is considered artificial intelligence but not machine learning is **symbolic reasoning.**
- It is based on the high level human readable representations of problems in logic.

Machine Learning (ML) Techniques

- The core of machine learning is creating an algorithm which a computer then uses to find a model that fits the data as best as possible and makes very accurate predictions based on that and how is that different from conventional methods. We provided with algorithms which give the machine directions on how to learn on its own.
- A machine learning algorithm is like a trial-and-error process. Each consecutive trial is at least as good as the previous one.
- Technically speaking there are four ingredients data, model, objective function and optimization algorithm.
- Example. Imagine a robot holding a bow. We want to find the best way to use that bow to fire accurately. In other words the usage of the bow is our ***model***, the best way to learn archery is to train right. We train by taking different arrows and trying to hit the target. So, the quiver of arrows will be or ***data*** or more precisely the data that the robot will use for training.
- They are all arrows but they have their subtleties. There are straight ones, crooked ones, light ones, heavy ones. So we can safely say the arrows represent different data values.
- We said the robot will be firing at a target. In machine learning or at least in the most common type supervised learning, we know what we are aiming for and we call it a target.
- The ***objective function*** will calculate how far from the target the robot shots were on average.
- Here comes the fourth ingredient the optimization algorithm.
- It steps on the findings of the objective function and consists of the mechanics that will improve the robot's archery skills somehow. It's posture the way it holds the bow how strong it pulls the bowstring etc. Then the robot will take the exact same data or arrows and fire them once again with its adjusted posture.
- This time the shots will be on average closer to the centre of the target. Normally the improvement will be almost unnoticeable. This entire process could have been hundreds or thousands of times until the robot finds the optimal way to fire this set of arrows and hit the centre every single time.

- Nevertheless, **it is important to remember that while training you won't provide the robot with a set of rules that is you won't have programmed a set of instructions like place the arrow in the middle of the bow pull the bow string and so on**.
- Instead, you will have given the machine a final goal to place the arrow in the centre of the target.
- So you don't care if it places the arrow in the middle or in the bottom of the bow as long as it hits the target.
- **Another important thing is that it won't learn to shoot well right away but after a hundred thousand tries it may have learned how to be the best archer out there.**
- Now there can be infinite possibilities to trial, when will the robots stop training first.
- The robot will learn certain things on the way and will take them into consideration for the next shots at fires for instance if it learns that it must look towards the target it will stop firing in the opposite direction.
- That is the purpose of the **optimization algorithm**. Second it cannot fire arrows forever.
- However, hitting the centre nine out of 10 times may be good enough. So, we can choose to stop it after it reaches a certain level of accuracy or fires a certain number of arrows.
- So, let us follow the four ingredients at the end of the training. Our robot or model is already trained on this data. With this set of arrows most shots hit the centre so the air or the objective function is quite low or minimized as we like to say the posture the technique and all other factors cannot be improved.
- So, the optimization algorithm has done its best to improve the shooting ability of the machine.
- We own a robot that is an amazing Archer. So, what can you do? Give it a different bag of arrows. If they had seen most types of arrows while training it will do great with the new ones.
- However, if we give it half an arrow or a longer arrow than it has seen it will not know what to do with it.
- In all ordinary cases though we would expect the robot to hit the centre or at least get close.
- The benefit of using machine learning is that the robot can learn to fire more effectively than a human.

Types of Machine Learning



1) Supervised Machine learning

- This name derives from the fact that training an algorithm resembles a teacher supervising her students.
- In Supervised machine learning, it is important to mention you have been dealing with label data. In other words, you can assess the accuracy of each shot.
- Consider previous example, where there isn't a single target different arrows have their own targets.
- Let's check what the robot sees when shooting the ground, a target at a short distance a target at a further distance a target hanging on a tree far behind it a house to the side and the sky.
- ***So, having labelled data means the associating or labelling a target to a type of Arrow.***
- You know that with a small arrow the robot is supposed to hit the closest target with a medium arrow it can reach the target located further away while with a larger arrow the target that's hanging on the tree. Finally, a crooked arrow is expected to hit the ground not reaching any target during the training process.
- The robot will be shooting arrows at the respective targets as well as it can. After training is finished.
- Ideally the robot will be able to fire the small arrow at the centre of the closest target the middle arrow at the centre of the one further away and so on.
- To summarize label data means we know the target prior to the shot, and we can associate that shot with the target this way. We're sure where the arrow should hit.
- This allows us to measure the inaccuracy of the shot through the objective function and improve the way the robot shoots through the optimization algorithm. So, what we supervise is the training itself. If a shot is far off from its target, we correct the posture. Otherwise, we don't get.

Past data is used to make predictions in supervised machine learning.

Example of supervised machine learning is the spam filtering of emails. We all use Gmail, Yahoo, or Outlook. Machine learning algorithms are used for deciding which email is spam and which is not.

Based on the previous data like received emails, data that we use etc., the system makes predictions about an email as for whether it is a spam or not. These predictions may not be perfect, but they are accurate most of the times.

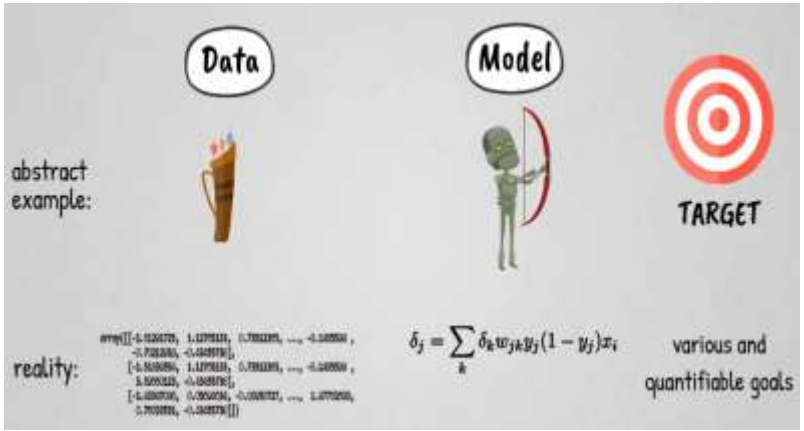
Classification and Regression are the ML algorithms that come under Supervised ML.

2) Unsupervised Machine learning

- In practice though it might happen that you won't have the time or the resources to associate the arrows with targets before giving them to the robot.
- In that case you could apply the other major type of M-L unsupervised learning here you will just give your robot a bag of arrows with unknown physical properties unlabelled data. This means neither you nor the robot will have separated the arrows into groups.
- Then you'd ask the machine to simply fire in a direction without providing it with targets. Therefore, in this case you won't be looking for a model that helps you shoot better rather you'll be looking for one which divides the arrows in a certain way.
- The robot will see just the ground the tree the House and the sky. Remember there are no targets. So, after firing thousands of shots during the training process we will end up having different types of arrows stuck in different areas.
- For instance, you may identify all the broken arrows by noticing they have fallen on the ground nearby the others you may realise are divided into small medium and large arrows.
- There may be anomalies like crossbow bolts in your bag that after being shot may have accumulated in a pile over here.
- You wouldn't want to use them with a simple bow would you. At the end of the training the robot will have fired so many times that it could discover answers that may surprise you.
- The machine may have managed to split the arrows not into four but into five sized categories due to discovering the crossbow bolt. Or it may have identified that some arrows are going to break soon by placing them in the Broken Arrow pile.
- It is worth mentioning that supervised learning can deal with such problems too and it does very often. However, if you have one million arrows you don't really have the time to assign targets to all of them do you.
- To save time and resources you should apply unsupervised learning.

3) Rainforcement learning

- The third major type of machine learning is called reinforcement learning. This time we introduce a reward system.
- Every time the robot fires an arrow better than before it will receive an award say a chocolate it will receive nothing if it fires worse.
- So instead of minimizing an error we are maximizing a reward or in other words maximizing the objective function.
- If you put yourselves in the shoes of the machine, you'll be reasoning in the following way. I fire an arrow and receive a reward. I'll try to figure out what I did correctly.
- So, I get more chocolate with the next shot or I fire an arrow and don't receive a reward. There must be something I need to improve.
- For me to get some chocolate on my next shot positive reinforcement..



- In addition, don't forget the robot Archer was an abstract depiction of what a machine learning model can do.
- In reality there are robots, but the model will be a highly complex mathematical formula the arrows will be a data set and the goals will be various and quantifiable
- Here are the most notable approaches you will encounter when talking about machine learning support vector machines neural networks deep learning random forced models and Bazy and networks are all types of supervised learning.
- There are neural networks that can be applied to an unsupervised type of machine learning, but K means is the most common unsupervised approach.
- By the way you may have noticed we have placed deep learning in both categories.
- This is a relatively new revolutionary computational approach which is acclaimed as the State-of-the-art email today.
- Describing it briefly we can say it is fundamentally different from the other approaches.
- However, it has a broad practical scope of application in all M-L areas because of the extremely high accuracy of its models.
- Note that deep learning is still divided and supervised, unsupervised and reinforcement, so it solves the same problems but in a conceptually different way.

Real Life Examples of Machine Learning (ML)

The ***financial sector and banks*** have ginormous data sets of credit card transactions.
Unfortunately, banks are facing issues with fraud daily. They are tasked with preventing fraudsters from acquiring customer data and in order to keep customers funds safe they use machine learning algorithms.
They take past data and because they can tell the computer which transactions in their history were legitimate and which were found to be fraudulent, they can label the data as such.
So through supervised learning they train models that detect fraudulent activity when these models detect even the slightest probability of theft.
They flagged the transactions and prevent the fraud in real time. Although no one in the sector has reached a perfect solution.

Another example of using supervise machine learning with label data can be found in ***client retention***.

A focus of any business be it a global supermarket chain or an online clothing shop is to retain its customers.

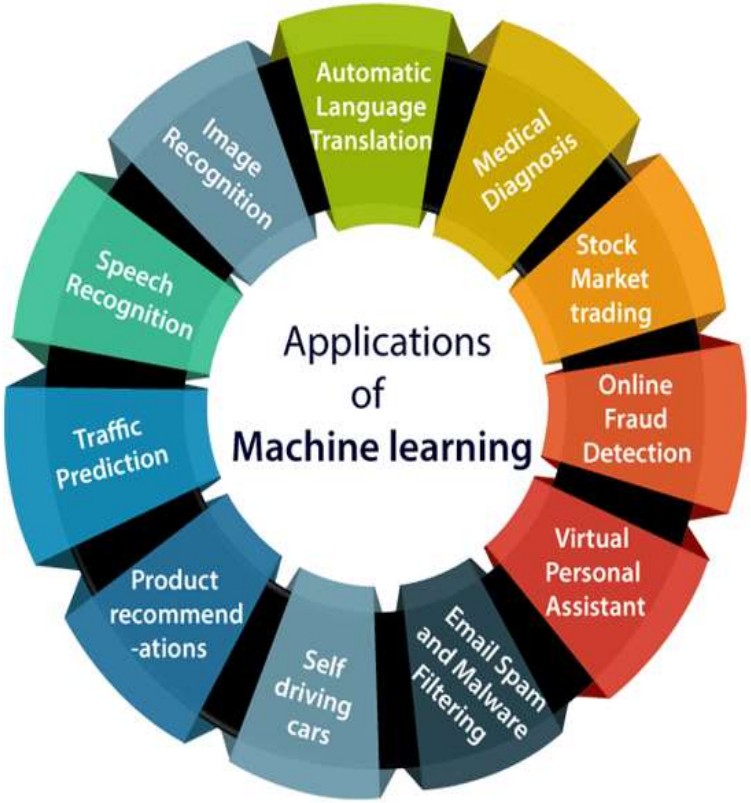
But the larger a business grows the harder it is to keep track of customer trends. A local corner shop owner will recognize and get to know their most loyal customers. They will offer them exclusive discounts to thank them for their custom.

And by doing so keep them returning on a larger scale. Companies can use machine learning and past label data to automate the practice.

And with this they can know which customers may purchase goods from them. This means the store can offer discounts and a personal touch in an efficient way minimizing marketing costs and maximizing profits.

Applications of Machine learning

Machine learning is a buzzword for today's technology, and it is growing very rapidly day by day. We are using machine learning in our daily life even without knowing it such as Google Maps, Google assistant, Alexa, etc. Below are some most trending real-world applications of Machine Learning:



1. Image Recognition:

Image recognition is one of the most common applications of machine learning. It is used to identify objects, persons, places, digital images, etc. The popular use case of image recognition and face detection is, **Automatic friend tagging suggestion**:

Facebook provides us a feature of auto friend tagging suggestion. Whenever we upload a photo with our Facebook friends, then we automatically get a tagging suggestion with name, and the technology behind this is machine learning **face detection** and **recognition algorithm**.

It is based on the Facebook project named "**Deep Face**," which is responsible for face recognition and person identification in the picture.

2. Speech Recognition

While using Google, we get an option of "**Search by voice**," it comes under speech recognition, and it's a popular application of machine learning.

Speech recognition is a process of converting voice instructions into text, and it is also known as "**Speech to text**", or "**Computer speech recognition**." At present, machine learning algorithms are widely used by various applications of speech recognition. **Google assistant**, **Siri**, **Cortana**, and **Alexa** are using speech recognition technology to follow the voice instructions.

3. Traffic prediction:

If we want to visit a new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions.

It predicts the traffic conditions such as whether traffic is cleared, slow-moving, or heavily congested with the help of two ways:

- **Real Time location** of the vehicle form Google Map app and sensors
- **Average time has taken** on past days at the same time.

Everyone who is using Google Map is helping this app to make it better. It takes information from the user and sends back to its database to improve the performance.

4. Product recommendations:

Machine learning is widely used by various e-commerce and entertainment companies such as **Amazon**, **Netflix**, etc., for product recommendation to the user. Whenever we search for some product on Amazon, then we started getting an advertisement for the same product while internet surfing on the same browser and this is because of machine learning.

Google understands the user interest using various machine learning algorithms and suggests the product as per customer interest.

As similar, when we use Netflix, we find some recommendations for entertainment series, movies, etc., and this is also done with the help of machine learning.

5. Self-driving cars:

One of the most exciting applications of machine learning is self-driving cars. Machine learning plays a significant role in self-driving cars. Tesla, the most popular car manufacturing company is working on self-driving car. It is using unsupervised learning method to train the car models to detect people and objects while driving.

6. Email Spam and Malware Filtering:

Whenever we receive a new email, it is filtered automatically as important, normal, and spam. We always receive an important mail in our inbox with the important symbol and spam emails in our spam box, and the technology behind this is Machine learning. Below are some spam filters used by Gmail:

- Content Filter
- Header filter
- General blacklists filter
- Rules-based filters
- Permission filters

Some machine learning algorithms such as **Multi-Layer Perceptron**, **Decision tree**, and **Naïve Bayes classifier** are used for email spam filtering and malware detection.

7. Virtual Personal Assistant:

We have various virtual personal assistants such as **Google assistant**, **Alexa**, **Cortana**, **Siri**. As the name suggests, they help us in finding the information using our voice instruction. These assistants can help us in various ways just by our voice instructions such as Play music, call someone, Open an email, Scheduling an appointment, etc.

These virtual assistants use machine learning algorithms as an important part.

These assistant record our voice instructions, send it over the server on a cloud, and decode it using ML algorithms and act accordingly.

8. Online Fraud Detection:

Machine learning is making our online transaction safe and secure by detecting fraud transaction. Whenever we perform some online transaction, there may be various ways that a fraudulent transaction can take place such as **fake accounts**, **fake ids**, and **steal money** in the middle of a transaction. So to detect this, **Feed Forward Neural network** helps us by checking whether it is a genuine transaction or a fraud transaction.

For each genuine transaction, the output is converted into some hash values, and these values become the input for the next round. For each genuine transaction, there is a specific pattern which gets change for the fraud transaction hence, it detects it and makes our online transactions more secure.

9. Stock Market trading:

Machine learning is widely used in stock market trading. In the stock market, there is always a risk of up and downs in shares, so for this machine learning's **long short term memory neural network** is used for the prediction of stock market trends.

10. Medical Diagnosis:

In medical science, machine learning is used for diseases diagnoses. With this, medical technology is growing very fast and able to build 3D models that can predict the exact position of lesions in the brain.

It helps in finding brain tumors and other brain-related diseases easily.

11. Automatic Language Translation:

Nowadays, if we visit a new place and we are not aware of the language then it is not a problem at all, as for this also machine learning helps us by converting the text into our known languages. Google's GNMT (Google Neural Machine Translation) provide this feature, which is a Neural Machine Learning that translates the text into our familiar language, and it called as automatic translation.

The technology behind the automatic translation is a sequence to sequence learning algorithm, which is used with image recognition and translates the text from one language to another language.

TOOLS IN MACHINE LEARNING

1) Scikit-learn

Scikit-learn is for machine learning development in python. It provides a library for the Python programming language.

Features:

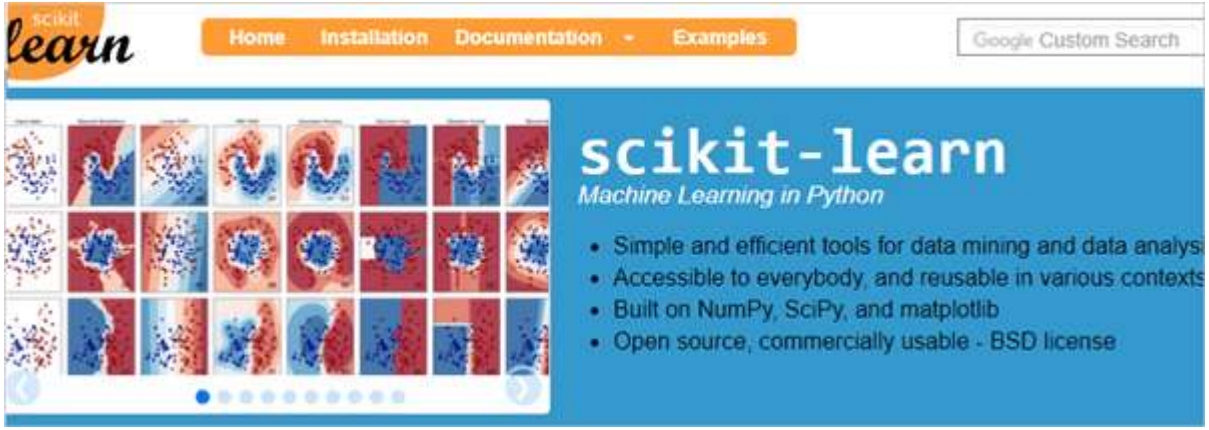
- It helps in data mining and data analysis.
- It provides models and algorithms for Classification, Regression, Clustering, Dimensional reduction, Model selection, and Pre-processing.

Pros:

- Easily understandable documentation is provided.
- Parameters for any specific algorithm can be changed while calling objects.

Tool Cost/Plan Details: Free.

Official Website: scikit-learn



2) PyTorch

PyTorch is a Torch based, Python machine learning library. The torch is a Lua based computing framework, scripting language, and machine learning library.

Features:

- It helps in building neural networks through Autograd Module.
- It provides a variety of optimization algorithms for building neural networks.
- PyTorch can be used on cloud platforms.
- It provides distributed training, various tools, and libraries.

Pros:

- It helps in creating computational graphs.
- Ease of use because of the hybrid front-end.

Tool Cost/Plan Details: Free

Official Website: **Pytorch**

3) TensorFlow

TensorFlow provides a JavaScript library that helps in machine learning. APIs will help you to build and train the models.

Features:

- Helps in training and building your models.
- You can run your existing models with the help of TensorFlow.js which is a model converter.
- It helps in the neural network.

Pros:

- You can use it in two ways, i.e. by script tags or by installing through NPM.
- It can even help for human pose estimation.

Cons:

- It is difficult to learn.

Tool Cost/Plan Details: Free

Official Website: **Tensorflow**

4) Weka

These machine learning algorithms help in data mining.

Features:

- Data preparation
- Classification
- Regression
- Clustering
- Visualization and
- Association rules mining.

Pros:

- Provides online courses for training.
- Easy to understand algorithms.
- It is good for students as well.

Cons:

- Not much documentation and online support are available.

Tool Cost/Plan Details: Free

Official Website: **Waikato-weka**

5) KNIME

KNIME is a tool for data analytics, reporting and integration platform. Using the data pipelining concept, it combines different components for machine learning and data mining.

Features:

- It can integrate the code of programming languages like C, C++, R, Python, Java, JavaScript etc.
- It can be used for business intelligence, financial data analysis, and CRM.

Pros:

- It can work as a SAS alternative.
- It is easy to deploy and install.
- Easy to learn.

Cons:

- Difficult to build complicated models.
- Limited visualization and exporting capabilities.

Tool Cost/Plan Details: Free

Official website: KNIME

6) Colab

Google Colab is a cloud service which supports Python. It will help you in building the machine learning applications using the libraries of PyTorch, Keras, TensorFlow, and OpenCV

Features:

- It helps in machine learning education.
- Assists in machine learning research.

Pros:

- You can use it from your google drive.

Tool Cost/Plan Details: Free

Official Website: Colab

7) Apache Mahout

Apache Mahout helps mathematicians, statisticians, and data scientists for executing their algorithms.

Features:

- It provides algorithms for Pre-processors, Regression, Clustering, Recommenders, and Distributed Linear Algebra.
- Java libraries are included for common math operations.
- It follows Distributed linear algebra framework.

Pros:

- It works for large data sets.
- Simple
- Extensible

Cons:

- Needs more helpful documentation.
- Some algorithms are missing.

Tool Cost/Plan Details: Free

Official Website: **Mahout – Apache**

8) Accord.Net

Accord.Net provides machine learning libraries for image and audio processing.

Features:

It provides algorithms for:

- Numerical linear algebra.
- Numerical optimization
- Statistics
- Artificial Neural networks.
- Image, audio, & signal processing.
- It also provides support for graph plotting & visualization libraries.

Pros:

- Libraries are made available from the source code and also through executable installer & NuGet package manager.

Cons:

- It supports only. Net supported languages.

Tool Cost/Plan Details: Free

Official Website: **Accord.net**

9) Shogun

Shogun provides various algorithms and data structures for machine learning. These machine learning libraries are used for research and education.

Features:

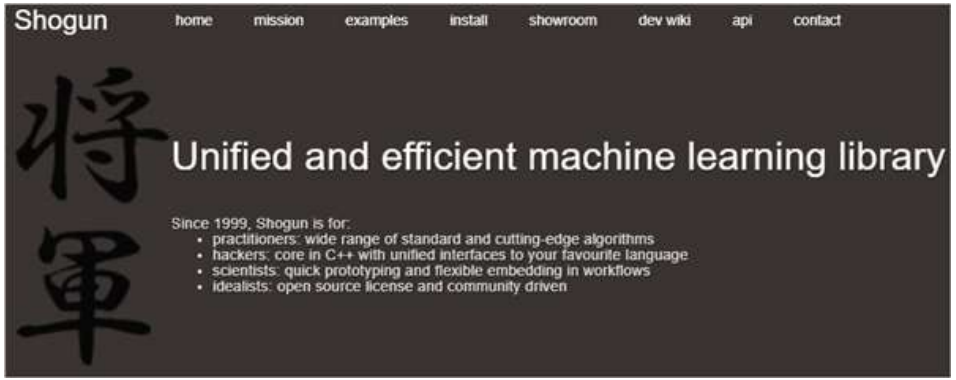
- It provides support vector machines for regression and classification.
- It helps in implementing Hidden Markov models.
- It offers support for many languages like – Python, Octave, R, Ruby, Java, Scala, and Lua.

Pros:

- It can process large data-sets.
- Easy to use.
- Provides good customer support.
- Offers good features and functionalities.

Tool Cost/Plan Details: Free

Official Website: **Shogun**



10) Keras.io

Keras is an API for neural networks. It helps in doing quick research and is written in Python.

Features:

- It can be used for easy and fast prototyping.
- It supports convolution networks.
- It assists recurrent networks.
- It supports a combination of two networks.
- It can be run on the CPU and GPU.

Pros:

- User-friendly
- Modular
- Extensible

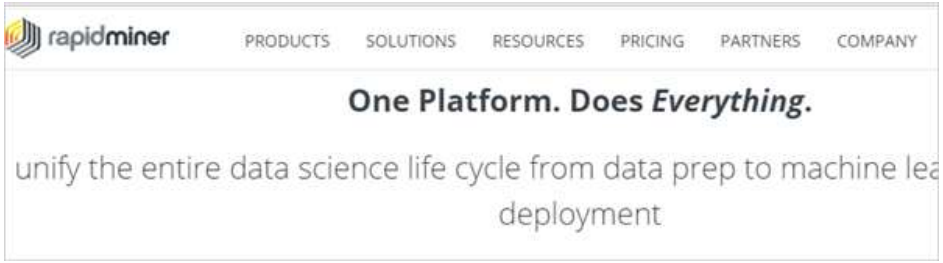
Cons:

- In order to use Keras, you must need TensorFlow, Theano, or CNTK.

Tool Cost/Plan Details: Free

Official Website: **Keras**

11) Rapid Miner



Rapid Miner provides a platform for machine learning, deep learning, data preparation, text mining, and predictive analytics. It can be used for research, education and application development.

Features:

- Through GUI, it helps in designing and implementing analytical workflows.
- It helps with data preparation.
- Result Visualization.
- Model validation and optimization.

Pros:

- Extensible through plugins.
- Easy to use.
- No programming skills are required.

Cons:

- The tool is costly.

Tool Cost/Plan Details:

It has four plans:

- Free plan
- **Small:** $2500 per year.
- **Medium:** $5000 per year.
- **Large:** $10000 per year.

Machine Learning Step

Preparing to model:-

1. Collecting Data:

- Machines initially learn from the data that you give them. It is of the utmost importance to collect reliable data so that your machine learning model can find the correct patterns.
- The quality of the data that you feed to the machine will determine how accurate your model is.
- If you have incorrect or outdated data, you will have wrong outcomes or predictions which are not relevant.
- Make sure you use data from a reliable source, as it will directly affect the outcome of your model.
- Good data is relevant, contains very few missing and repeated values, and has a good representation of the various subcategories/classes present.

2. Preparing the Data:

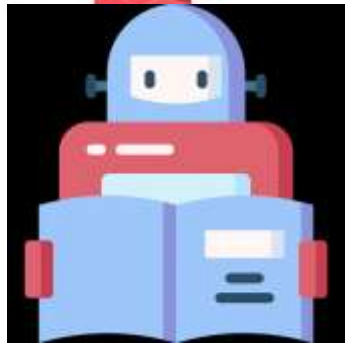
After you have your data, you have to prepare it. You can do this by :

- Putting together all the data you have and randomizing it. This helps make sure that data is evenly distributed, and the ordering does not affect the learning process.
- Cleaning the data to remove unwanted data, missing values, rows, and columns, duplicate values, data type conversion, etc. You might even have to restructure the dataset and change the rows and columns or index of rows and columns.
- Visualize the data to understand how it is structured and understand the relationship between various variables and classes present.
- Splitting the cleaned data into two sets - a training set and a testing set. The training set is the set your model learns from. A testing set is used to check the accuracy of your model after training.

Modelling and Evaluation:

3. Choosing a Model:

- A machine learning model determines the output you get after running a machine learning algorithm on the collected data.
- It is important to choose a model which is relevant to the task at hand.
- Over the years, scientists and engineers developed various models suited for different tasks like speech recognition, image recognition, prediction, etc.
- Apart from this, you also have to see if your model is suited for numerical or categorical data and choose accordingly.



4. Training the Model:

- Training is the most important step in machine learning.
- In training, you pass the prepared data to your machine learning model to find patterns and make predictions.
- It results in the model learning from the data so that it can accomplish the task set. Over time, with training, the model gets better at predicting.

5. Evaluating the Model:

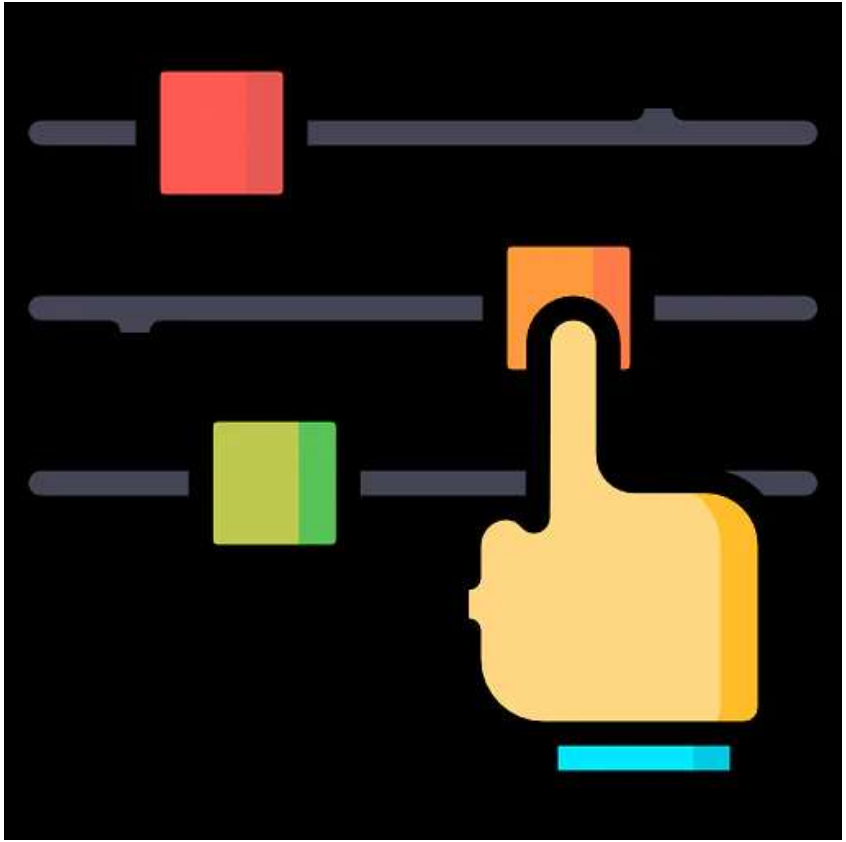
- After training your model, you have to check to see how it's performing. This is done by testing the performance of the model on previously unseen data.
- The unseen data used is the testing set that you split our data into earlier. If testing was done on the same data which is used for training, you will not get an accurate measure, as the model is already used to the data, and finds the same patterns in it, as it previously did.
- This will give you disproportionately high accuracy.
- When used on testing data, you get an accurate measure of how your model will perform and its speed.

6. Parameter Tuning:

- Once you have created and evaluated your model, see if its accuracy can be improved in any way.
- This is done by tuning the parameters present in your model.
- Parameters are the variables in the model that the programmer generally decides.
- At a particular value of your parameter, the accuracy will be the maximum. Parameter tuning refers to finding these values.

7. Making Predictions

- In the end, you can use your model on unseen data to make predictions accurately.



Feature Engineering for Machine Learning

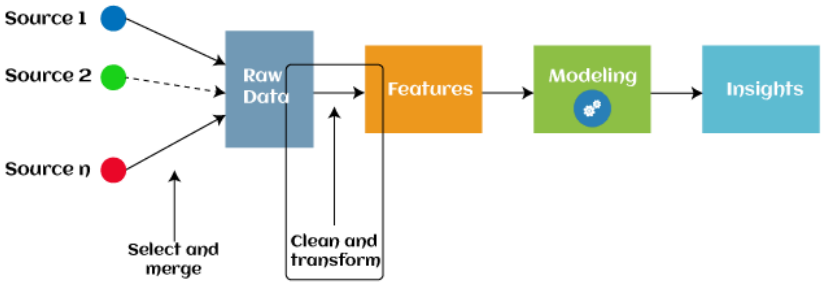
- **Feature engineering is the pre-processing step of machine learning, which is used to transform raw data into features that can be used for creating a predictive model using Machine learning or statistical Modelling**.
- Feature engineering in machine learning aims to improve the performance of models.

What is a feature?

- Generally, all machine learning algorithms take input data to generate the output. The input data remains in a tabular form consisting of rows (instances or observations) and columns (variable or attributes), and these attributes are often known as **features**.
- For example, an image is an instance in computer vision, but a line in the image could be the feature. Similarly, in NLP, a document can be an observation, and the word count could be the feature.
- So, we can say **a feature is an attribute that impacts a problem or is useful for the problem**.

What is Feature Engineering?

- **Feature engineering is the pre-processing step of machine learning, which extracts features from raw data**.
- It helps to represent an underlying problem to predictive models in a better way, which as a result, improve the accuracy of the model for unseen data.
- The predictive model contains predictor variables and an outcome variable, and while the feature engineering process selects the most useful predictor variables for the model.
- Feature engineering in ML contains mainly four processes: **Feature Creation, Transformations, Feature Extraction, and Feature Selection.**



These processes are described as below:

1. **Feature Creation**: Feature creation is finding the most useful variables to be used in a predictive model. The process is subjective, and it requires human creativity and intervention. The new features are created by mixing existing features using addition, subtraction, and ration, and these new features have great flexibility.
2. **Transformations**: The transformation step of feature engineering involves adjusting the predictor variable to improve the accuracy and performance of the model. For example, it ensures that the model is flexible to take input of the variety of data; it ensures that all the variables are on the same scale, making the model easier to understand. It improves the model's accuracy and ensures that all the features are within the acceptable range to avoid any computational error.
3. **Feature Extraction**: Feature extraction is an automated feature engineering process that generates new variables by extracting them from the raw data. The main aim of this step is to reduce the volume of data so that it can be easily used and managed for data modelling. Feature extraction methods include **cluster analysis, text analytics, edge detection algorithms, and principal components analysis (PCA**).
4. **Feature Selection:** While developing the machine learning model, only a few variables in the dataset are useful for building the model, and the rest features are either redundant or irrelevant. If we input the dataset with all these redundant and irrelevant features, it may negatively impact and reduce the overall performance and accuracy of the model. Hence it is very important to identify and select the most appropriate features from the data and remove the irrelevant or less important features, which is done with the help of feature selection in machine learning. ***"Feature selection is a way of selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant, or noisy features."***

Below are some benefits of using feature selection in machine learning:

- It helps in avoiding the curse of dimensionality.
- It helps in the simplification of the model so that the researchers can easily interpret it.
- It reduces the training time.
- It reduces overfitting hence enhancing the generalization.

Need for Feature Engineering in Machine Learning

In machine learning, the performance of the model depends on data pre-processing and data handling. But if we create a model without pre-processing or data handling, then it may not give good accuracy. Whereas, if we apply feature engineering on the same model, then the accuracy of the model is enhanced. Hence, feature engineering in machine learning improves the model's performance. Below are some points that explain the need for feature engineering:

- **Better features mean flexibility.**
  In machine learning, we always try to choose the optimal model to get good results. However, sometimes after choosing the wrong model, still, we can get better predictions, and this is because of better features. The flexibility in features will enable you to select the less complex models. Because less complex models are faster to run, easier to understand and maintain, which is always desirable.
- **Better features mean simpler models.**
  If we input the well-engineered features to our model, then even after selecting the wrong parameters (Not much optimal), we can have good outcomes. After feature engineering, it is not necessary to do hard for picking the right model with the most optimized parameters. If we have good features, we can better represent the complete data and use it to best characterize the given problem.
- **Better features mean better results.**
  As already discussed, in machine learning, as data we will provide will get the same output. So, to obtain better results, we must need to use better features.

Steps in Feature Engineering

The steps of feature engineering may vary as per different data scientists and ML engineers. However, there are some common steps that are involved in most machine learning algorithms, and these steps are as follows:

- **Data Preparation:** The first step is data preparation. In this step, raw data acquired from different resources are prepared to make it in a suitable format so that it can be used in the ML model. The data preparation may contain cleaning of data, delivery, data augmentation, fusion, ingestion, or loading.
- **Exploratory Analysis:** Exploratory analysis or Exploratory data analysis (EDA) is an important step of features engineering, which is mainly used bydata scientists. This step involves analysis, investing data set, and summarization of the main characteristics of data. Different data visualization techniques are used to better understand the manipulation of data sources, to find the most appropriate statistical technique for data analysis, and to select the best features for the data.
- **Benchmark**: Benchmarking is a process of setting a standard baseline for accuracy to compare all the variables from this baseline. The benchmarking process is used to improve the predictability of the model and reduce the error rate.

Feature Engineering Techniques

Some of the popular feature engineering techniques include:

1. Imputation

Feature engineering deals with inappropriate data, missing values, human interruption, general errors, insufficient data sources, etc. Missing values within the dataset highly affect the performance of the algorithm, and to deal with them "Imputation" technique is used. **Imputation is responsible for handling irregularities within the dataset.**

For example, removing the missing values from the complete row or complete column by a huge percentage of missing values. But at the same time, to maintain the data size, it is required to impute the missing data, which can be done as:

- For numerical data imputation, a default value can be imputed in a column, and missing values can be filled with means or medians of the columns.
- For categorical data imputation, missing values can be interchanged with the maximum occurred value in a column.

2. Handling Outliers

Outliers are the deviated values or data points that are observed too away from other data points in such a way that they badly affect the performance of the model. Outliers can be handled with this feature engineering technique. This technique first identifies the outliers and then remove them out.

Standard deviation can be used to identify the outliers. For example, each value within a space has a definite to an average distance, but if a value is greater distant than a certain value, it can be considered as an outlier. **Z-score** can also be used to detect outliers.

3. Log transform

Logarithm transformation or log transform is one of the commonly used mathematical techniques in machine learning. Log transform helps in handling the skewed data, and it makes the distribution more approximate to normal after transformation. It also reduces the effects of outliers on the data, as because of the normalization of magnitude differences, a model becomes much robust.

4. Binning

In machine learning, overfitting is one of the main issues that degrade the performance of the model and which occurs due to a greater number of parameters and noisy data. However, one of the popular techniques of feature engineering, "binning", can be used to normalize the noisy data. This process involves segmenting different features into bins.

5. Feature Split

As the name suggests, feature split is the process of splitting features intimately into two or more parts and performing to make new features. **This technique helps the algorithms to better understand and learn the patterns in the dataset.**

The feature splitting process enables the new features to be clustered and binned, which results in extracting useful information and improving the performance of the data models.

6. One hot encoding

One hot encoding is the popular encoding technique in machine learning. It is a technique that converts the categorical data in a form so that they can be easily understood by machine learning algorithms and hence can make a good prediction. It enables group the of categorical data without losing any information.