OHIO COMPUTER ACADEMY
Inspire, Educate & Evolve

Data Science Specialization

Trainer: Arbind Jain

What is Data?

- Data is a collection of information.
  - One purpose of Data Science is to structure data, making it interpretable and easy to work with.
  - Data can be categorized into two groups:
    - ✓ Structured data
    - ✓ Unstructured data

OHIO COMPUTER ACADEMY
Inspire, Educate & Evolve

Unstructured Data

- Unstructured data is not organized. We must organize the data for analysis purposes.

Unstructured data

Example of Unstructured data

Health Status 09082020 Inbox ×

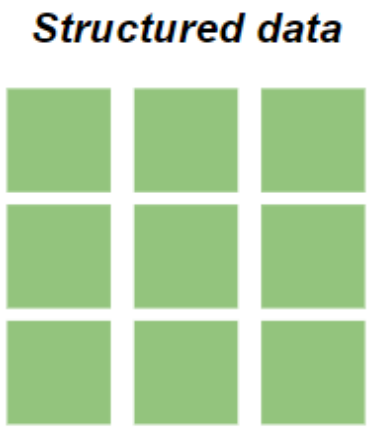
Date 09082020
Average pulse 70,
Max pulse 80,
Steps 10500

Structured Data

- Structured data is organized and easier to work with.
  - ✓ How to Structure Data?
    - We can use an array or a database table to structure or present data.
    - Example of an array: [80, 85, 90, 95, 100, 105, 110, 115, 120, 125]

- The following example shows how to create an array in Python:

```
Array = [80, 85, 90, 95, 100, 105, 110, 115, 120, 125]
print(Array)
```



What is Data Science?

- Data science is a deep study of the massive amount of data, which involves
  - ✓ Extracting meaningful insights from raw, structured, and unstructured data
  - ✓ That is processed using the scientific method, different technologies, and algorithms.
  - ✓ It is a multidisciplinary field that uses tools and techniques to manipulate the data so that you can find something new and meaningful.
  - ✓ Data science uses the most powerful hardware, programming systems, and most efficient algorithms to solve the data related problems.
  - ✓ It is the future of artificial intelligence.

What is Data Science?

Asking	Asking the correct questions and analyzing the raw data.
Modeling	Modeling the data using various complex and efficient algorithms.
Visualizing	Visualizing the data to get a better perspective.
Understanding	Understanding the data to make better decisions and finding the final result.

Prerequisite for Data Science

- **Non-Technical Prerequisite:**
  - ✓ **Curiosity**: To learn data science, one must have curiosities. When you have curiosity and ask various questions, then you can understand the business problem easily.
  - ✓ **Critical Thinking**: It is also required for a data scientist so that you can find multiple new ways to solve the problem with efficiency.
  - ✓ **Communication skills**: Communication skills are most important for a data scientist because after solving a business problem, you need to communicate it with the team.
- **Technical Prerequisite:**
  - ✓ **Machine learning**: To understand data science, one needs to understand the concept of machine learning. Data science uses machine learning algorithms to solve various problems.
  - ✓ **Mathematical modeling**: Mathematical modeling is required to make fast mathematical calculations and predictions from the available data.
  - ✓ **Statistics**: Basic understanding of statistics is required, such as mean, median, or standard deviation. It is needed to extract knowledge and obtain better results from the data.
  - ✓ **Computer programming**: For data science, knowledge of at least one programming language is required. R, Python, Spark are some required computer programming languages for data science.
  - ✓ **Databases**: The depth understanding of Databases such as SQL, is essential for data science to get the data and to work with data.

Data Science Jobs

- Data Scientist
- Data Analyst
- Machine learning expert
- Data engineer
- Data Architect
- Data Administrator
- Business Analyst
- Business Intelligence Manager

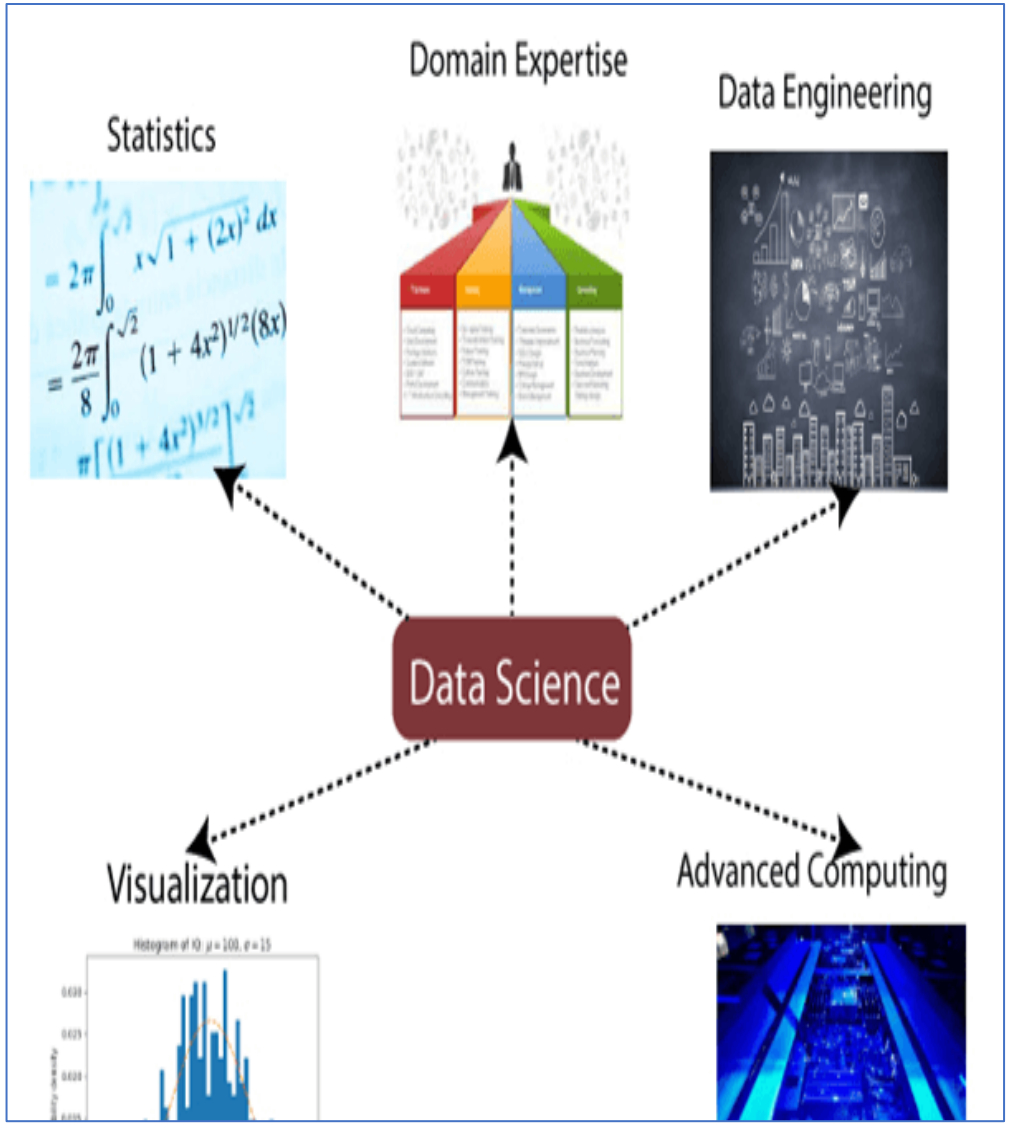
Difference between BI and Data Science

BI stands for business intelligence, which is also used for data analysis of business information

Criterion	Business intelligence	Data science
Data Source	Business intelligence deals with structured data, e.g., data warehouse.	Data science deals with structured and unstructured data, e.g., weblogs, feedback, etc.
Method	Analytical(historical data)	Scientific(goes deeper to know the reason for the data report)
Skills	Statistics and Visualization are the two skills required for business intelligence.	Statistics, Visualization, and Machine learning are the required skills for data science.
Focus	Business intelligence focuses on both Past and present data	Data science focuses on past data, present data, and also future predictions.

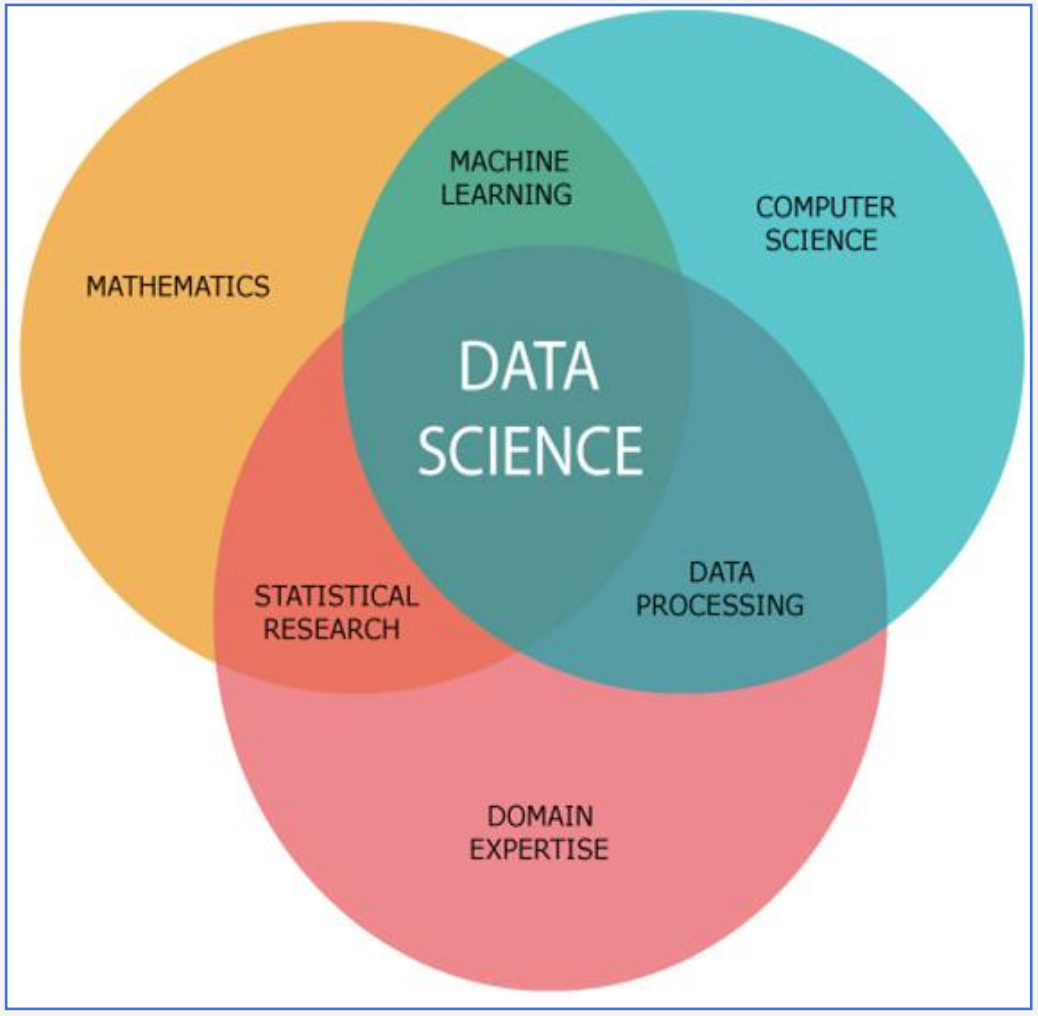
Data Science Components

- ✓**Statistics** is a way to collect and analyze the numerical data in a large amount and finding meaningful insights from it.
- ✓**Domain expertise** means specialized knowledge or skills of a particular area.
- ✓**Data engineering** involves acquiring, storing, retrieving, and transforming the data, it includes metadata (data about data) to the data.
- ✓**Data visualization** meant by representing data in a visual context so that people can easily understand the significance of data. Data visualization makes it easy to access the huge amount of data in visuals.
- ✓**Advanced computing** is heavy lifting of data science is advanced computing, it involves designing, writing, debugging, and maintaining the source code of computer programs.



Data Science Components

- **Mathematics**:
  - ✓ Mathematics involves the study of quantity, structure, space, and changes.
  - ✓ For a data scientist, knowledge of good mathematics is essential.
- **Machine learning**:
  - ✓ Machine learning is backbone of data science. Machine learning is all about to provide training to a machine so that it can act as a human brain.
  - ✓ In data science, we use various machine learning algorithms to solve the problems.



Tools for Data Science

- **Data Analysis tools:** R, Python, Statistics, SAS, Jupyter, R Studio, MATLAB, Excel, RapidMiner.
- **Data Warehousing / Data Lake:** ETL, SQL, Hadoop, Informatica/Talend, AWS Redshift, Amazon Glue, AWS Data lake etc.
- **Data Visualization tools:** R, Jupyter, Tableau, Cognos.
- **Machine learning tools:** Spark, Mahout, Azure ML studio.
- **Big Data Storage**: Hadoop HDFS, Amazon S3 etc.

Machine learning in Data Science

- To become a data scientist, one should also be aware of machine learning and its algorithms, as in data science, there are various machine learning algorithms which are broadly being used.
- Following are the name of some machine learning algorithms used in data science:
  - ✓ Regression
  - ✓ Decision tree
  - ✓ Clustering
  - ✓ Principal component analysis
  - ✓ Support vector machines
  - ✓ Naive Bayes
  - ✓ Artificial neural network
  - ✓ Apriori

Applications of Data Science

- **Image recognition and speech recognition**
  - ✓ When you upload an image on Facebook and start getting the suggestion to tag to your friends.
  - ✓ This automatic tagging suggestion uses image recognition algorithm, which is part of data science.
  - ✓ When you say something using, "Ok Google, Siri, Cortana", etc., and these devices respond as per voice control, so this is possible with speech recognition algorithm.

- **Gaming world**
  - ✓ In the gaming world, the use of Machine learning algorithms is increasing day by day.
  - ✓ EA Sports, Sony, Nintendo, are widely using data science for enhancing user experience.

- **Internet search**
  - ✓ When we want to search for something on the internet, then we use different types of search engines such as Google, Yahoo, Bing, Ask, etc.
  - ✓ All these search engines use the data science technology to make the search experience better, and you can get a search result with a fraction of seconds.

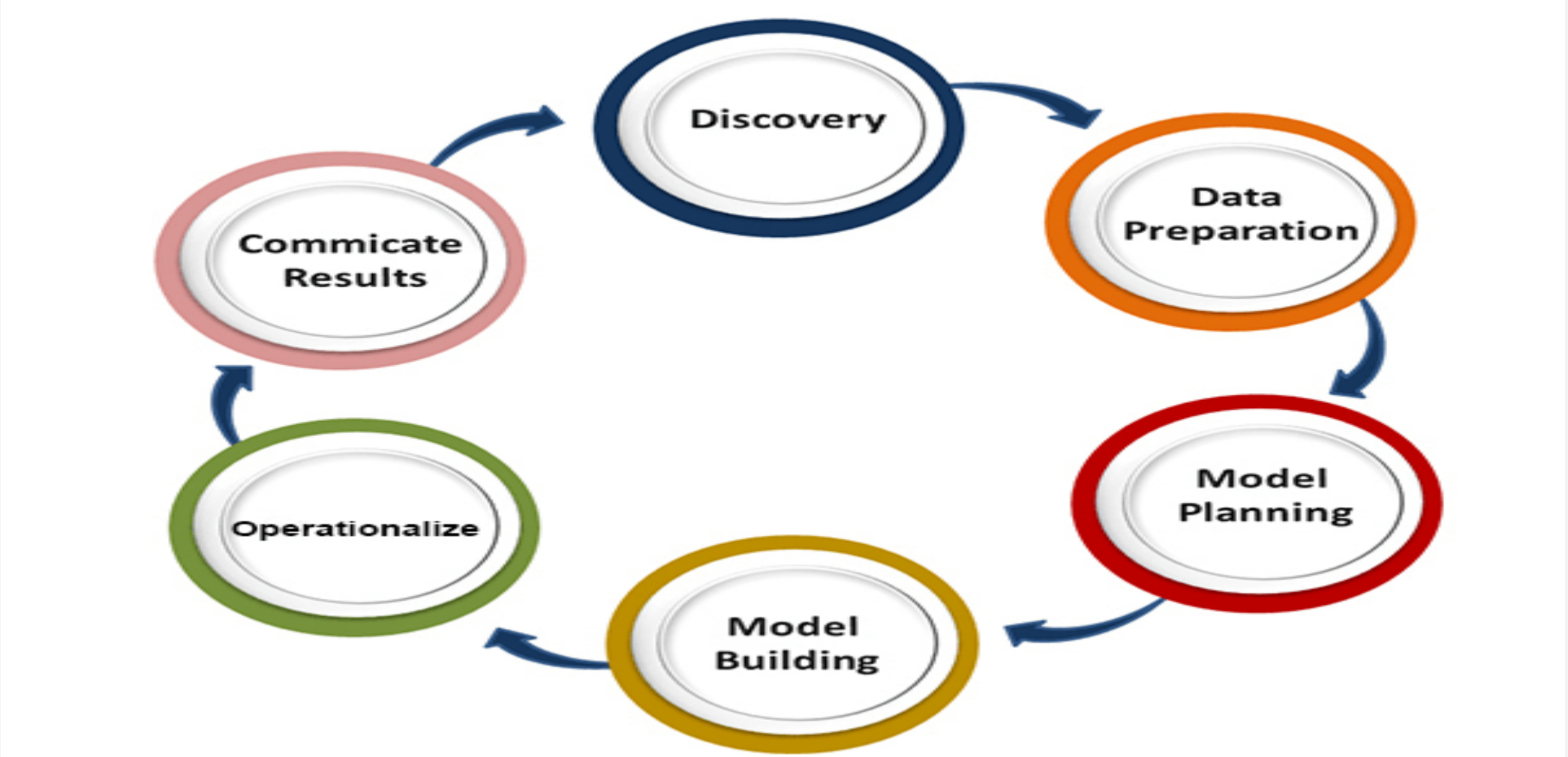
Applications of Data Science

- **Transport**
  - ✓ Transport industries also using data science technology to create self-driving cars.
  - ✓ With self-driving cars, it will be easy to reduce the number of road accidents.
- **Healthcare**
  - ✓ In the healthcare sector, data science is providing lots of benefits.
  - ✓ Data science is being used for tumor detection, drug discovery, medical image analysis, virtual medical bots, etc.
- **Recommendation systems**
  - ✓ Most of the companies, such as Amazon, Netflix, Google Play, etc., are using data science technology for making a better user experience with personalized recommendations.
  - ✓ Such as, when you search for something on Amazon, and you started getting suggestions for similar products, so this is because of data science technology.
- **Risk detection**
  - ✓ Finance industries always had an issue of fraud and risk of losses, but with the help of data science, this can be rescued.
  - ✓ Most of the finance companies are looking for the data scientist to avoid risk and any type of losses with an increase in customer satisfaction.

Challenges of Data Science Technology

- A high variety of information & data is required for accurate analysis
- Not adequate data science talent pool available
- Management does not provide financial support for a data science team
- Unavailability of/difficult access to data
- Business decision-makers do not effectively use data Science results
- Explaining data science to others is difficult.
- Privacy issues.
- Lack of significant domain expert.
- If an organization is very small, it can’t have a Data Science team.

Data Science Lifecycle



Data Science Lifecycle

- **Discovery**: The first phase is discovery, which involves asking the right questions. When you start any data science project, you need to determine what are the basic requirements, priorities, and project budget.

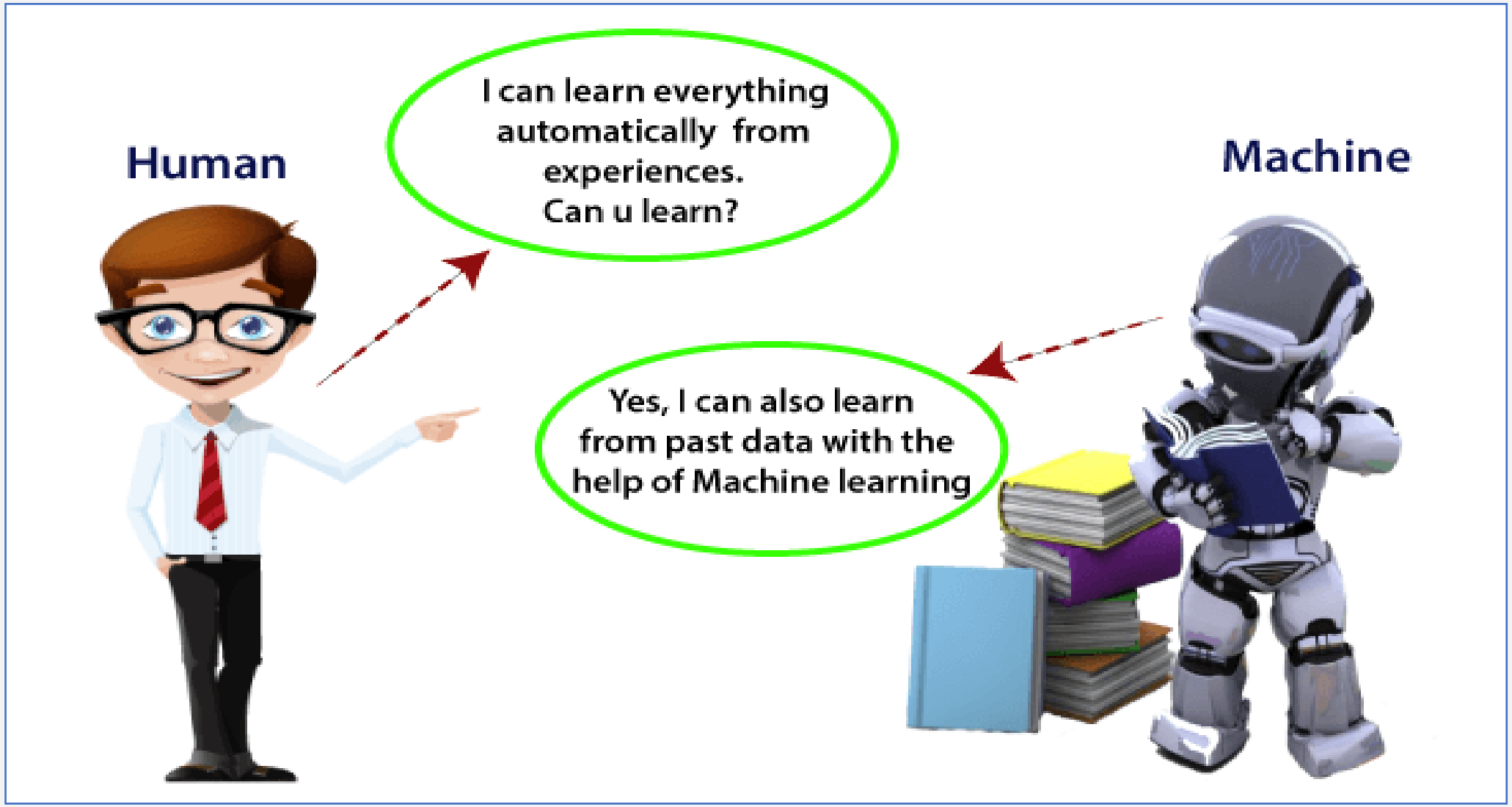
- **Data preparation**: Data preparation is also known as Data Munging. In this phase, we need to perform the following tasks:
  - ✓Data cleaning
  - ✓Data Reduction
  - ✓Data integration
  - ✓Data transformation

- **Model Planning**: Determines various methods and techniques to establish the relation between input variables.
  - ✓Apply Exploratory data analytics(EDA) by using various statistical formula and visualization tools to understand the relations between variable and to see what data can inform us. Common tools used for model planning are:
    - ✓SQL Analysis Services, R,SAS,Python etc,

Data Science Lifecycle

- **Model-building**: In this phase, the process of model building starts.
  - ✓Create datasets for training and testing purpose.
  - ✓Apply different techniques such as association, classification, and clustering, to build the model. Following are some common Model building tools:
    - SAS Enterprise Miner, WEKA, SPCS Modeler, MATLAB
- **Operationalize**: In this phase,
  - Deliver the final reports of the project, along with briefings, code, and technical documents.
  - This phase provides a clear overview of complete project performance and other components on a small scale before the full deployment.
- **Communicate results**: In this phase
  - ✓Check if we reach the goal, which we have set on the initial phase.
  - ✓Communicate the findings and final result with the business team.

What is Machine Learning

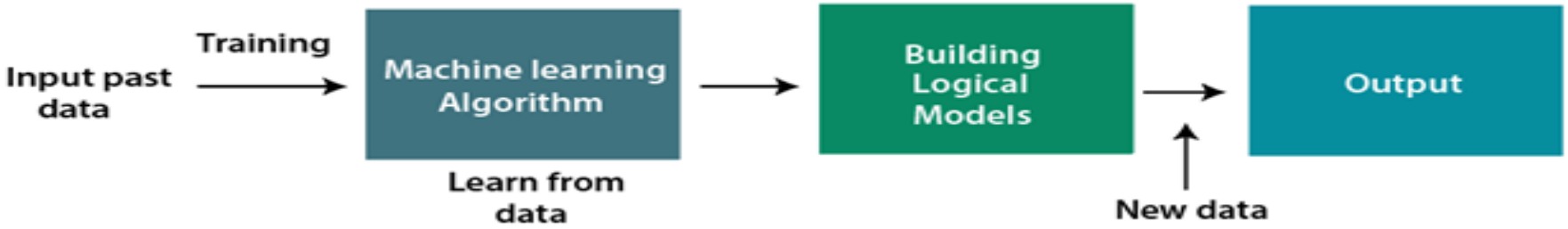


What is Machine Learning?

- Machine Learning is defined as a technology that is used to train machines to perform various actions such as predictions, recommendations, estimations, etc., based on historical data or past experience.
- Machine Learning enables computers to behave like human beings by training them with the help of past experience and predicted data.
- Three key aspects of Machine Learning,
  - ✓Task: A task is defined as the main problem in which we are interested. This task/problem can be related to the predictions and recommendations and estimations, etc.
  - ✓Experience: It is defined as learning from historical or past data and used to estimate and resolve future tasks.
  - ✓Performance: It is defined as the capacity of any machine to resolve any machine learning task or problem and provide the best outcome for the same. However, performance is dependent on the type of machine learning problems.
- Currently ML is being used in image recognition, speech recognition, email filtering, Facebook auto-tagging, recommender system, and many more.

How does Machine Learning work?

- A Machine Learning system learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it.
- The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.
- Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output.
- Machine learning has changed our way of thinking about the problem. The below block diagram explains the working of Machine Learning algorithm:



Features of Machine Learning

- Machine learning uses data to detect various patterns in a given dataset.
- It can learn from past data and improve automatically.
- It is a data-driven technology.
- Machine learning is much similar to data mining as it also deals with the huge amount of the data.

Types of ML

- Supervised learning,
- Semi supervised learning,
- Unsupervised learning
- Reinforcement learning.

Supervised Learning

- Supervised learning is applicable when a machine has sample data, i.e., input as well as output data with correct labels.
- Correct labels are used to check the correctness of the model using some labels and tags. Supervised learning technique helps us to predict future events with the help of past experience and labeled examples.
- Initially, it analyses the known training dataset, and later it introduces an inferred function that makes predictions about output values. Further, it also predicts errors during this entire learning process and also corrects those errors through algorithms.
- Example: Let's assume we have a set of images tagged as ''dog''. A machine learning algorithm is trained with these dog images so it can easily distinguish whether an image is a dog or not.

Unsupervised Learning

- In unsupervised learning, a machine is trained with some input samples or labels only, while output is not known.
- The training information is neither classified nor labeled; hence, a machine may not always provide correct output compared to supervised learning.
- Although Unsupervised learning is less common in practical business settings, it helps in exploring the data and can draw inferences from datasets to describe hidden structures from unlabeled data.
- Example: Let's assume a machine is trained with some set of documents having different categories (Type A, B, and C), and we have to organize them into appropriate groups. Because the machine is provided only with input samples or without output, so, it can organize these datasets into type A, type B, and type C categories, but it is not necessary whether it is organized correctly or not.

Reinforcement Learning

- Reinforcement Learning is a feedback-based machine learning technique.
- In such type of learning, agents (computer programs) need to explore the environment, perform actions, and on the basis of their actions, they get rewards as feedback.
- For each good action, they get a positive reward, and for each bad action, they get a negative reward.
- The goal of a Reinforcement learning agent is to maximize the positive rewards.
- Since there is no labeled data, the agent is bound to learn by its experience only.

Semi-supervised Learning

- Semi-supervised Learning is an intermediate technique of both supervised and unsupervised learning.
- It performs actions on datasets having few labels as well as unlabeled data. However, it generally contains unlabeled data. Hence, it also reduces the cost of the machine learning model as labels are costly, but for corporate purposes, it may have few labels. Further, it also increases the accuracy and performance of the machine learning model.
- Sem-supervised learning helps data scientists to overcome the drawback of supervised and unsupervised learning.
- Speech analysis, web content classification, protein sequence classification, text documents classifiers., etc., are some important applications of Semi-supervised learning.

Difference between Artificial intelligence and Machine learning

- Artificial intelligence and machine learning are the part of computer science that are correlated with each other. These two technologies are the most trending technologies which are used for creating intelligent systems.
- Although these are two related technologies and sometimes people use them as a synonym for each other, but still both are the two different terms in various cases.
- On a broad level, we can differentiate both AI and ML as:
- *AI is a bigger concept to create intelligent machines that can simulate human thinking capability and behavior, whereas, machine learning is an application or subset of AI that allows machines to learn from data without being programmed explicitly.*

AI vs ML