

INTRODUCTORY CONCEPTS

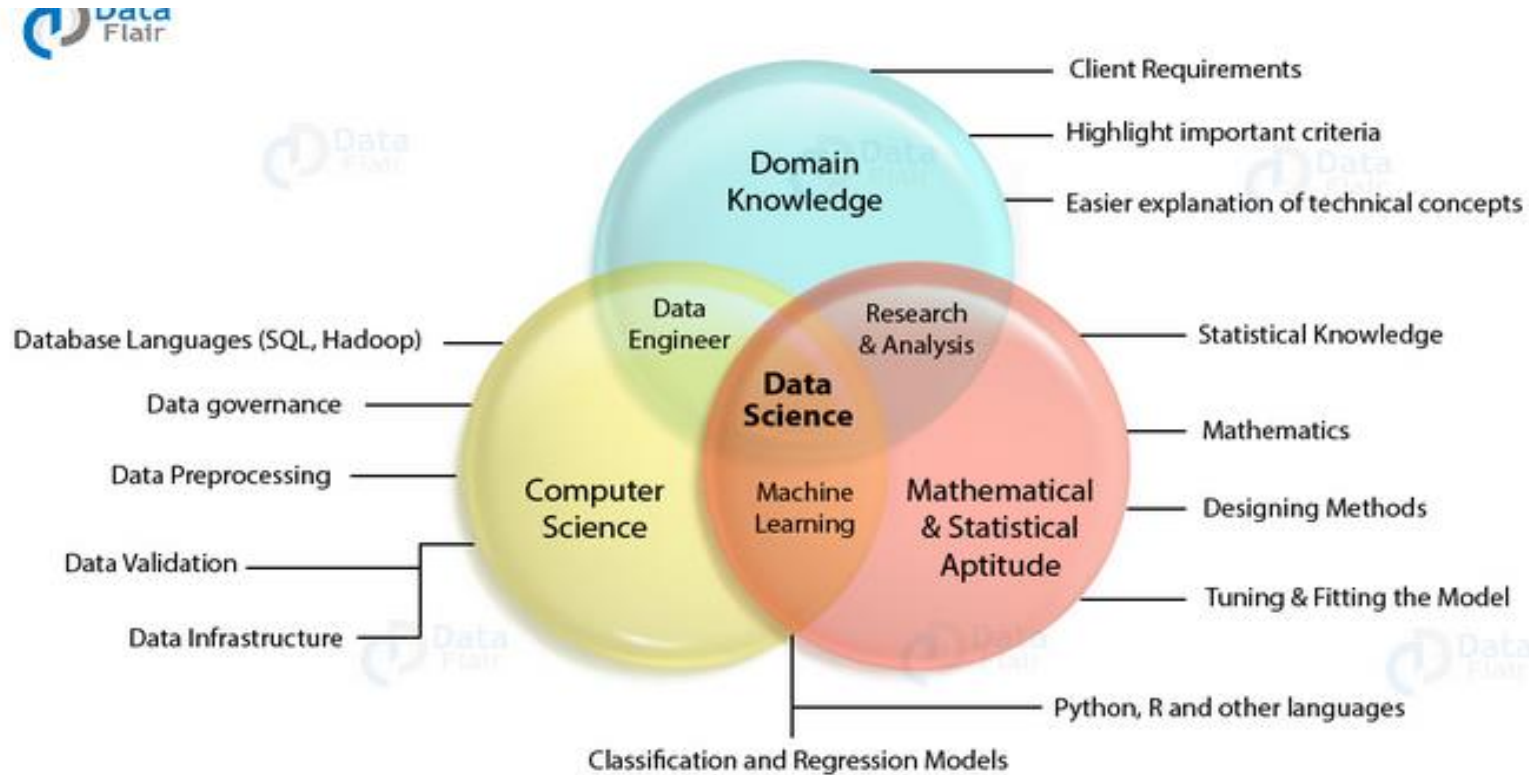
Introduction to Data Science

- Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions
- Data Science is a comprehensive process that involves preprocessing, analysis, visualization and prediction.
- Data science is about finding hidden patterns in the data.



Introduction to Data Science

- Data science involves various underlying fields like statistics, mathematics, computer science, predictive analytics, machine learning algorithm development, and new technologies to gain insights from big data.



Why Data Science ?

- Traditionally, the data that we had was mostly **structured** and **small** in size, which could be analyzed by using simple **BI tools**. Unlike data in the traditional systems which was mostly structured, today most of the data is **unstructured or semi-structured**.
- Year 2020 survey shows that, more than 80 % of the data will be unstructured. This data is generated from different sources like financial logs, text files, multimedia forms, sensors, and instruments. Simple BI tools are not capable of processing this huge volume and variety of data.
- This is why we need more complex and advanced analytical tools and algorithms for processing, analyzing and drawing meaningful insights out of it.

Why Data Science ?

- Data science or data-driven science enables better decision making, predictive analysis, and pattern discovery.
- Allows companies to increase efficiencies, manage costs, identify new market opportunities, and boost their market advantage.
- Data Science has helped to create smarter systems that can take autonomous decisions based on historical datasets.

BI vs Data Science

- Difference between BI and Data science

Features	Business Intelligence (BI)	Data Science
Data Sources	Structured (Usually SQL, often Data Warehouse)	Both Structured and Unstructured (logs, cloud data, SQL, NoSQL, text)
Approach	Statistics and Visualization	Statistics, Machine Learning, Graph Analysis, Neuro- linguistic Programming (NLP)
Focus	Past and Present	Present and Future
Tools	Pentaho, Microsoft BI, QlikView, R	RapidMiner, BigML, Weka, R

Types of data scientists

- Type A Data Scientists
 - The A here stands for **Analysis**. This is a more static approach towards analysis of data or gaining insights from it. The work of a Type A data scientist is more closely related to that of a statistician.
 - A Type A Data Scientist is well versed with data cleaning, working with large data-sets, data visualization, domain knowledge, etc.

Types of data scientists

- Type B Data Scientists

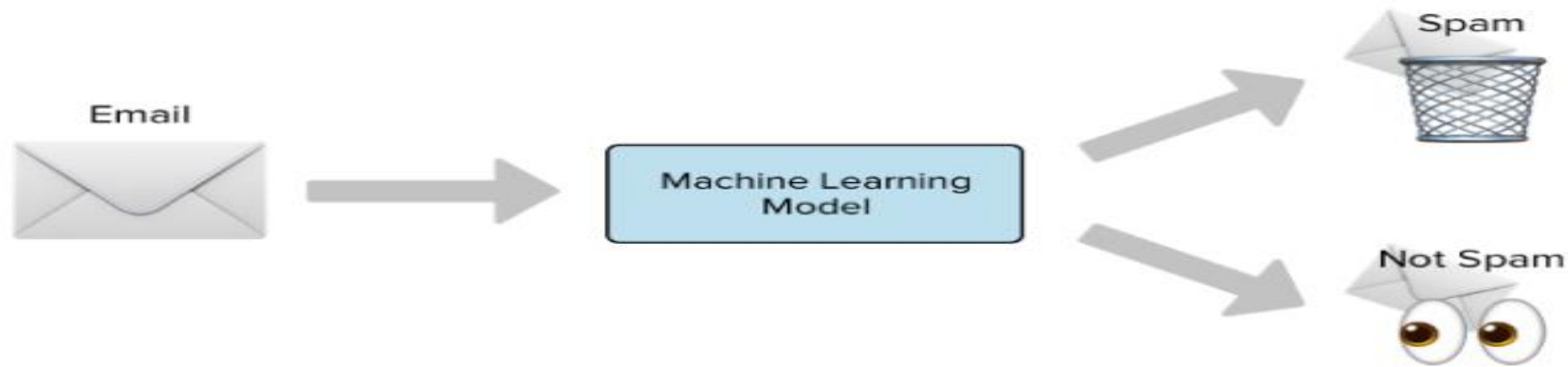
- The B here stands **Building**. While Type B Data Scientists share their background in statistics with Type A Data Scientists, they are well versed in coding and fundamentals of software engineering. They are responsible for building data products that directly interact with the user.
- This helps them to craft products that provide recommendations and other forms of interactive results to the user.

Application

- Examples
- Gmail filters your emails in the spam and non-spam categories
 - Before Google/Gmail decides to segregate the emails into spam or not spam category, before it arrives to your mailbox, hundreds of rules apply to those email in the data centres. These rules describe the properties of a spam email.
 - There are common types of spam filters which are used by Gmail/Google —Blatant Blocking, Bulk Email Filter, Category Filters, Null Sender Disposition, Null Sender Header Tag Validation.
 - There are ways to avoid spam filtering and send your emails straight to the inbox.

Application

- Examples
 - Gmail filters your emails in the spam and non-spam categories



- Spam detection is a supervised machine learning problem. This means you must provide your machine learning model with a set of examples of spam and ham messages and let it find the relevant patterns that separate the two different categories. Most email providers have their own vast data sets of labelled emails.

Case Study

- How Netflix Used Data Science to Improve its Recommendation System?



Case Study

- How Netflix Used Data Science to Improve its Recommendation System?
 - In order to make this happen, Netflix invested in a lot of algorithms to provide a flawless movie experience to its users. One of such algorithms is the **recommendation system** that is used by Netflix to provide suggestions to the users.
 - A recommendation system **understands the needs of the users** and provides suggestions of the various cinematographic products. System takes the information about the user as an input.
 - This information can be in the form of the **past usage of product or the ratings** that were provided to the product. It then processes this information to predict how much the user would rate or prefer the product. A recommendation system makes use of a variety of [machine learning algorithms](#).

Case Study

- How Netflix Used Data Science to Improve its Recommendation System?
- Another important role that a recommendation system plays today is to **search for similarity** between different products. In the case of Netflix, the recommendation system searches for movies that are similar to the ones you have watched or have liked previously.
- Therefore, based on the movies that are watched, Netflix provides recommendations of the films that share a degree of similarity.
- Now a days Netflix uses Hybrid Recommendation System (Content + Collaborative filtering) for suggesting content to its users.

Recommendation System

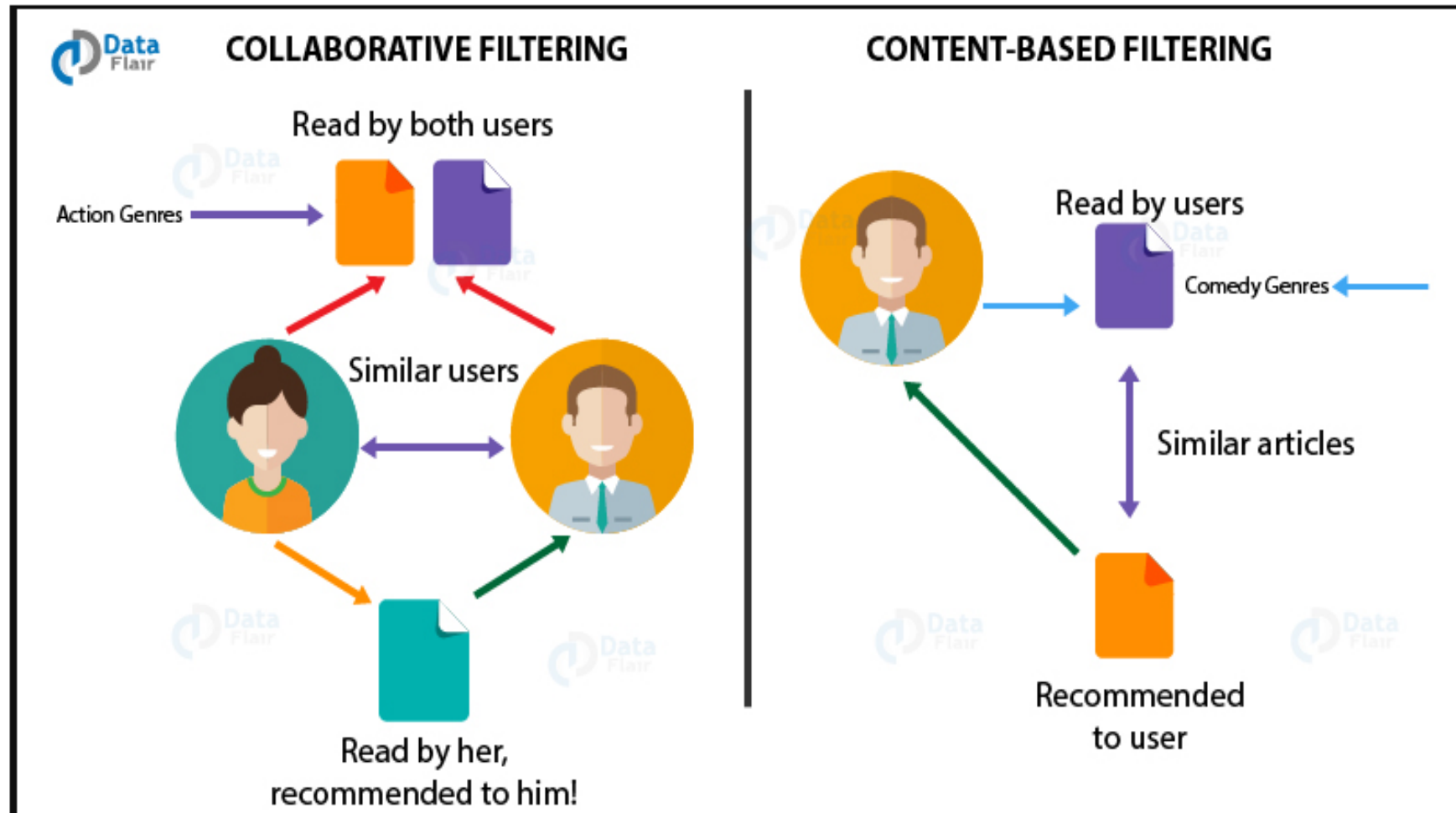
- There are two main types of Recommendation Systems –
- **Content-based recommendation systems**
 - In a content-based recommendation system, the **background knowledge of the products and customer information** are taken into consideration. Based on the content that you have viewed on Netflix, it provides you with similar suggestions.
- **For example**, if you have watched a film that has a sci-fi genre, the content-based recommendation system will provide you with suggestions for similar films that have the same genre.

Recommendation System

- There are two main types of Recommendation Systems –
- **Collaborative filtering recommendation systems**
 - Unlike the content based filtering that provided recommendations of similar products, Collaborative Filtering provides recommendations based on the **similar profiles of its users**. One key advantage of collaborative filtering is that it is independent of the product knowledge.
 - Rather, it relies on the users with a basic assumption that what the users liked in the past will also like in the future. **For example**, if a person A watches crime, sci-fi and thriller genres and B watches sci-fi, thriller and action genres then A will also like action and B will like crime genre.

Recommendation System

- There are two main types of Recommendation Systems –



Applications

- Examples
 - Companies like Google and Amazon are using Data Science to develop powerful recommendation systems for their users.
 - Various financial companies are using predictive analytics and forecasting methods to predict stock prices.
 - Asking a personal assistant like Alexa or Siri for a recommendation

Applications

- Data science is all about using data to solve problems.
- **Decision making:** The problem could be decision making such as identifying which email is spam and which is not. OR understand the precise requirements of your customers from the existing data like the customer's past browsing history, purchase history, age and income.
- **Product recommendation:** product recommendation such as which movie to watch?
- **predictive analytics:** Predicting the outcome such as who will be the next President of the USA?
- So, the core job of a data scientist is to **understand** the data, **extract** useful information out of it and **apply** this in solving the problems.

Applications

- Self-driving car: self-driving cars collect live data from sensors, including radars, cameras, and lasers to create a map of its surroundings. Based on this data, it takes decisions like when to speed up, when to speed down, when to overtake, where to take a turn – making use of advanced machine learning algorithms
- Weather forecasting: Data from ships, aircraft, radars, satellites can be collected and analyzed to build models. These models will not only forecast the weather but also help in predicting the occurrence of any natural calamities. It will help you to take appropriate measures beforehand and save many precious lives.
- Interest based search engine
- Talking to a Chatbot for customer service

Applications

- With the help of data science, Airlines can optimize operations in many ways, including:
 - Plan routes and decide whether to schedule direct or connecting flights
 - Build predictive analytics models to forecast flight delays
 - Offer personalized promotional offers based on customers booking patterns
 - Decide which class of planes to purchase for better overall performance
- To choose insurance plan
- To find restaurant
- To select internet plan
- To create marketing campaign
- To choose next destination for vacation

Applications

- All the domains where Data Science is creating its impression.



Applications

- **Data Science in Healthcare**

- With the help of classification algorithms, doctors are able to detect cancer and tumors at an early stage using Image Recognition software.
- Genetic Industries use Data Science for analyzing and classifying patterns of genomic sequences. Various virtual assistants are also helping patients to resolve their physical and mental ailments.
- Drug Discovery with Data Science
- Predictive Analytics in Healthcare: Finds various correlations and association of symptoms, finds habits, diseases and then makes meaningful predictions.
- Monitoring Patient Health

Applications

- **Data Science in E-commerce**

- Amazon uses a recommendation system that recommends users various products based on their historical purchase. Data Scientists have developed recommendation systems predict user preferences using Machine Learning.

- **Data Science in Manufacturing**

- Industrial robots have made taken over mundane and repetitive roles required in the manufacturing unit. These industrial robots are autonomous in nature and use Data Science technologies such as Reinforcement Learning and Image Recognition.

Applications

- **Data Science as Conversational Agents**

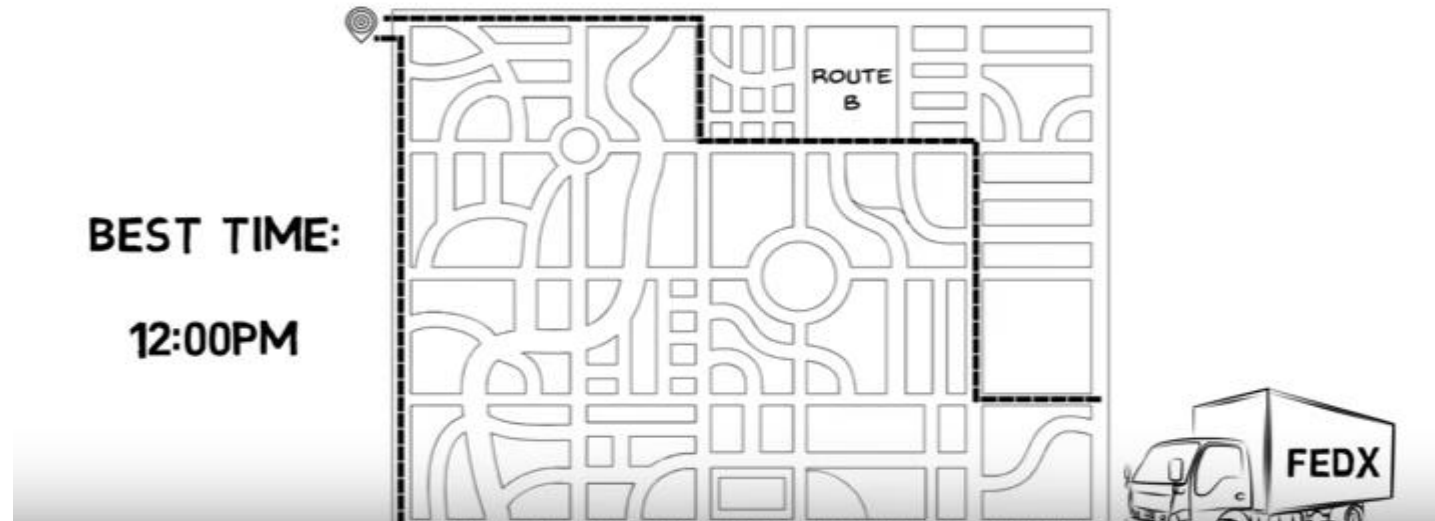
- Amazon's Alexa and Siri by Apple use Speech Recognition to understand users. Data Scientists develop this speech recognition system, that converts human speech into textual data. Also, it uses various Machine Learning algorithms to classify user queries and provide an appropriate response.

Applications

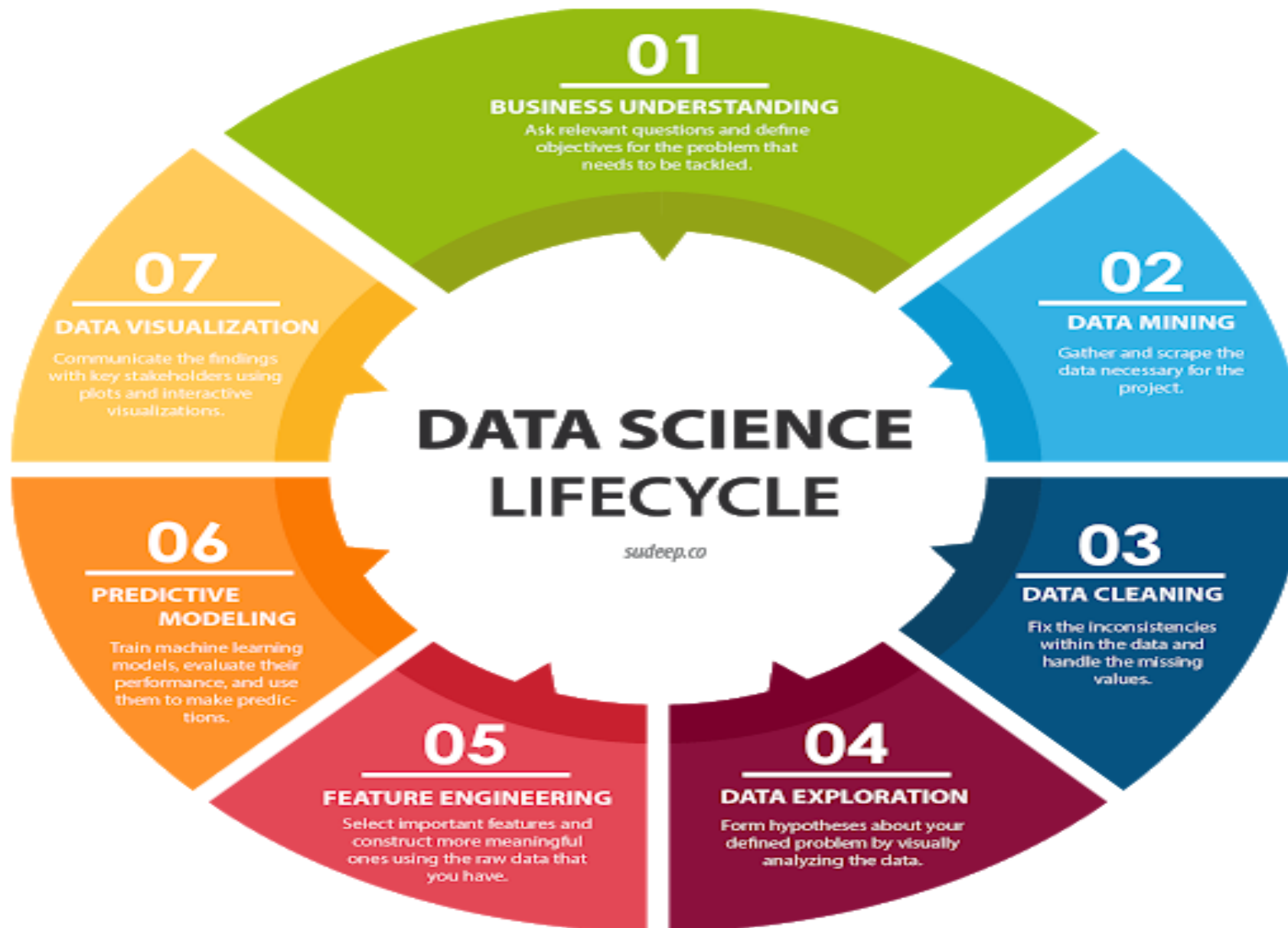
- **Data Science in Transport**

- Self Driving Cars use autonomous agents that utilize Reinforcement Learning and Detection algorithms. Self-Driving Cars are no longer fiction due to advancements in Data Science.

**LOGISTICS COMPANIES LIKE DHL, FEDEX HAVE
DISCOVERED THE BEST TIME AND ROUTES TO SHIP**



Data Science Life Cycle



Data Science Life Cycle

- **Step 1: Define Problem Statement**

- Creating a well-defined problem statement is a first and critical step in data science. It is a brief description of the problem that you are going to solve.

why do we need a well-defined problem statement?

- Most of the times, these initial set of a problem shared with you is vague and ambiguous.
- For example, the problem statement: “I want to increase the revenue”, doesn’t tell you how much to increase the revenue such as 20% or 30%, for which products to increase revenue and what is the time frame to increase the revenue.
- You have to make the problem statement clear, goal-oriented and measurable. This can be achieved by asking the right set of questions.

Data Science Life Cycle

- **Step 2: Data Collection**

- Data collection is a systematic approach to gather relevant information from a variety of sources.
- Depending on the problem statement, the data collection method is broadly classified into two categories.
 - Primary data collection
 - Secondary data collection
- **Primary data collection:** When we have some unique problem and no related research is done on the subject. Then, we need to collect new data. This method is called as **primary data collection**.
- For **example**, we want information on the average time that employees spend in a cafeteria across companies. There is no public data available of these. But we can collect the data through various methods such as surveys, interviews of employees and by monitoring the time spent by employees in cafeteria. This method is time-consuming.

Data Science Life Cycle

- **Step 2: Data Collection**

- **Secondary data collection:** To use the data which is readily available or collected by someone else. These data can be found on the internet, news articles, government census, magazines and so on. This method is called as **secondary data collection**.
- This method is less time-consuming than the primary method.

Data Science Life Cycle

- **Step 3: Data Preparation**

- Since raw data may not be usable, data preparation is the most crucial aspect of the data science lifecycle.
- A data scientist must first examine the data to identify any gaps or data that do not add any value. During this process, you must go through several steps, including:
 - **Data Integration:** Resolve any conflicts in the dataset and eliminate redundancies
 - **Data Transformation:** Normalize, transform and aggregate data using ETL (extract, transform, load) methods
 - **Data Reduction:** Using various strategies, reduce the size of data without impacting the quality or outcome
 - **Data Cleaning:** Correct inconsistent data by filling out missing values and smoothing out noisy data

Data Science Life Cycle

- **Step 4: Exploratory Data Analysis**

- it's important to analyze the data. It is the most exciting step as it helps you to build familiarity with the data and extract useful insights.
- If you skip this step then you might end up generating inaccurate models and choosing the insignificant variables in your model.



DEFINES AND REFINES
THE SELECTION OF FEATURE
VARIABLES THAT WILL BE USED
IN THE MODEL DEVELOPMENT

Data Science Life Cycle

- **Step 4: Exploratory Data Analysis**

- A Data Scientist analyzes the data through various statistical procedures. In particular, two types of procedures used are:
 - Descriptive Statistics
 - Inferential Statistics
- Assume that you are a Data Scientist working for a company that manufactures cell phones. You have to analyze customers using the mobile phones of your company. In order to do so, you will first take a thorough look at the data and understand various trends and patterns involved.
- In the end, you will summarize the data and present it in the form of a graph or a chart. You therefore, apply **Descriptive Statistics** to solve the problem.

Data Science Life Cycle

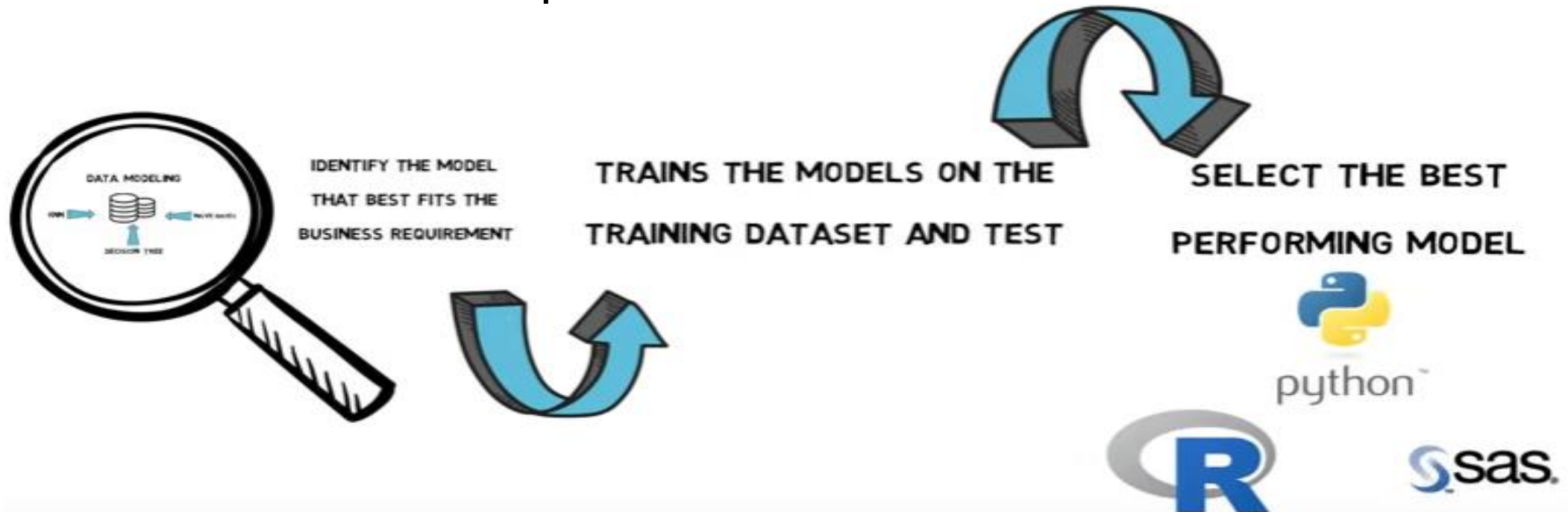
- **Step 4: Exploratory Data Analysis**

- You will then draw 'inferences' or conclusions from the data. To understand inferential statistics through the following example – Assume that you wish to find out a number of defects that occurred during manufacturing.
- However, individual testing of mobile phones can take time. Therefore, you will consider a sample of the given phones and make a generalization about the number of defective phones in the total sample.

Data Science Life Cycle

- **Step 5: Data Modeling**

- Modeling means formulating every step and gather the techniques required to achieve the solution.
- The important factor is how to perform the calculations. There are various techniques under Statistics and Machine Learning that you can choose based on the requirement.



Data Science Life Cycle

- **Step 6: Communicate Results:** Data reporting, Data visualization, Business Intelligence, Decision making
 - This is the final step where you present the results from your analysis to the stakeholders. You explain to them how you came to a specific conclusion and your critical findings.
 - Uses tools like tableau, Power BI, QlikView to create powerful reports and dashboards

Data Science Life Cycle

- **Step 7: Deploys and maintains**

- This is the final step where you present the results from your analysis to the stakeholders. You explain to them how you came to a specific conclusion and your critical findings.
- Test the selective model in a pre production environment before deploying in production environment.
- After successfully deployment, uses reports and dashboards to get real time analytics.
- Further monitor and maintain project performance

Data Science Use Cases

- Big companies are using data science for different purposes.



Data Science Use Cases

- **Facebook – Using Data to Revolutionize Social Networking & Advertising**
 - Facebook using advanced techniques like deep learning in data science to study user behaviour and gain insights to improve their product.
 - Using deep learning, Facebook makes use of facial recognition and text analysis.
 - In facial recognition, Facebook uses powerful neural networks to classify faces in the photographs. It uses its own text understanding engine called “DeepText” to understand user sentences.
 - It also uses Deep Text to understand people’s interest and aligning photographs with texts.

Data Science Use Cases

- **Facebook – Using Data to Revolutionize Social Networking & Advertising**
 - Facebook uses deep learning for targeted advertising. Using this, it decides what kind of advertisements the users should view.
 - It uses the insights gained from the data to cluster users based on their preferences and provides them with the advertisements that appeal to them.

Data Science Use Cases

- **Uber – Using Data to Make Rides Better**

- Uber is a popular smartphone application that allows you to book a cab. Uber makes extensive use of Big Data because Uber has to maintain a large database of drivers, customers, and several other records.
- Uber contains a database of drivers. Therefore, whenever you hail for a cab, Uber matches your profile with the most suitable driver. What differentiates Uber from other cab companies is that Uber charges you based on the time it takes to cover the distance and not the distance itself.
- It calculates the time taken through various algorithms that also make use of data related to traffic density and weather conditions.

Data Science Use Cases

- **Uber – Using Data to Make Rides Better**

- Uber makes the best use of data science to calculate its surge pricing. When there are less drivers available to more riders, the price of the ride goes up. This happens only during the scarcity of drivers in any given area.
- However, if the demand for Uber rides is less, then Uber charges a lower rate. This dynamic pricing is rooted in Big Data and makes excellent usage of data science to calculate the fares based on the parameters.

Data Science Use Cases

- **Bank of America – Using Data to Leverage Customer Experience**

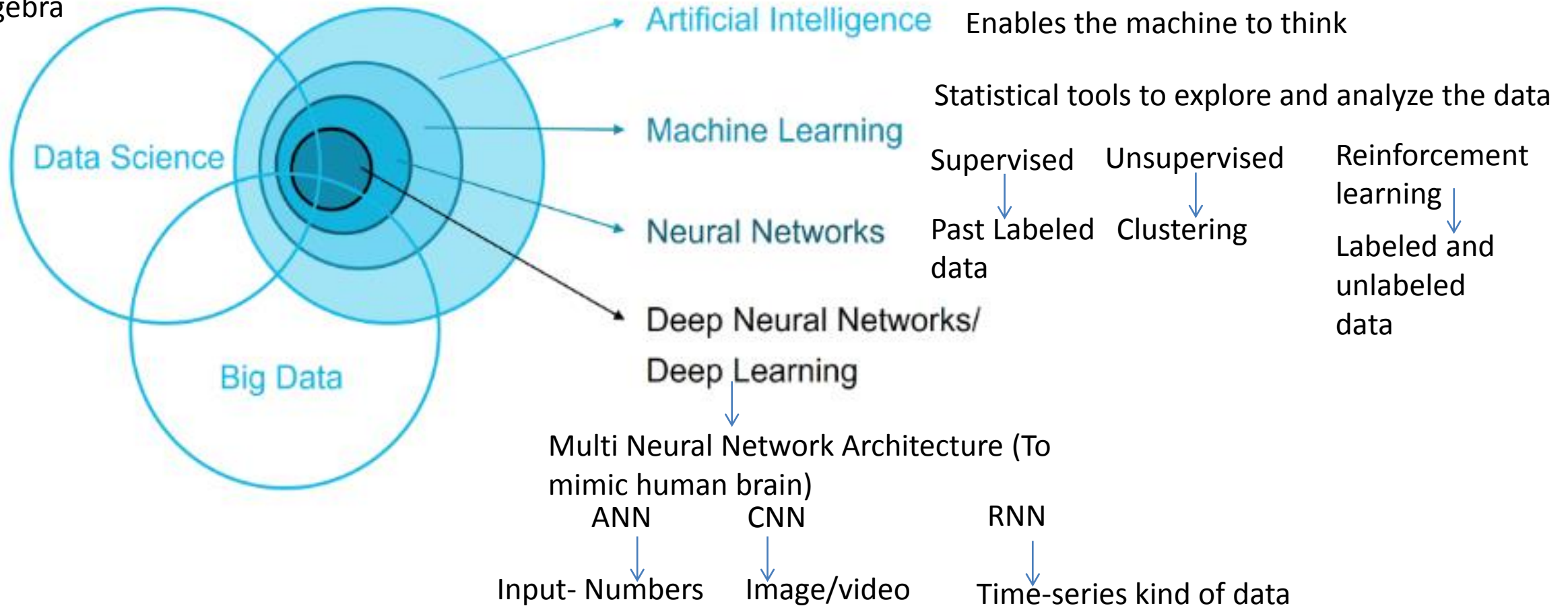
- Using data science, banking industries are able to detect frauds in payments and customer information. It also prevents frauds regarding insurances, credit cards, and accounting.
- In order to minimize the losses, a bank needs to detect fraud sooner. In order to carry this out, banks employ data scientists to use their quantitative knowledge where they apply algorithms like association, clustering, forecasting, and classification.

Data Science Use Cases

- **Spotify – Revolutionizing Music Streaming**
 - It is an online music streaming giant that uses Data Science for providing personalized music recommendations
 - Spotify is a data-driven company that leverages big data to provide personalized playlists to its users.
 - In the year 2017, Spotify used data science to gain insights about which universities had the highest percentage of party playlists and which ones spent the most time on it. It publishes its findings on their page “Spotify Insights” to provide information about the on-going trends in the music.

AI vs ML vs DL vs Data science

Uses tools like Statistics, probability,
linier algebra



Tools and Libraries

- Programming language
 - R
 - Python
 - Java
- Machine learning Algorithms
 - Classification and clustering
 - Regression
 - Reinforcement
 - Deep learning
 - Dimensionality reduction
- IDE (Integrated development environment)
 - Pycharm
 - Jupyter
 - Spyder
 - R Studio

Tools and Libraries

- Web Scraping
 - BeautifulSoup library
 - Scrapy tool
 - URLLib library
- Math
 - Statistics
 - Linear Algebra
 - Differential Calculus
- Data Visualization tools
 - Tableau
 - Power BI
 - Matplotlib library
- Data analysis
 - Feature engineering
 - Data Wrangling
 - Exploratory data analysis

Thank you.