Introduction to Data Science

Agenda

Key Takeaways-

- What is Data Science?
- Need for Data Science
- Different components of Data Science
- Data Science vs Artificial Intelligence vs Machine Learning
- Data Science Project Life Cycle
- Applications of Data Science in various domains
- Scenario specific Case Study
- Different Data Science Platforms

Data Science Around us

Have you ever wondered about -

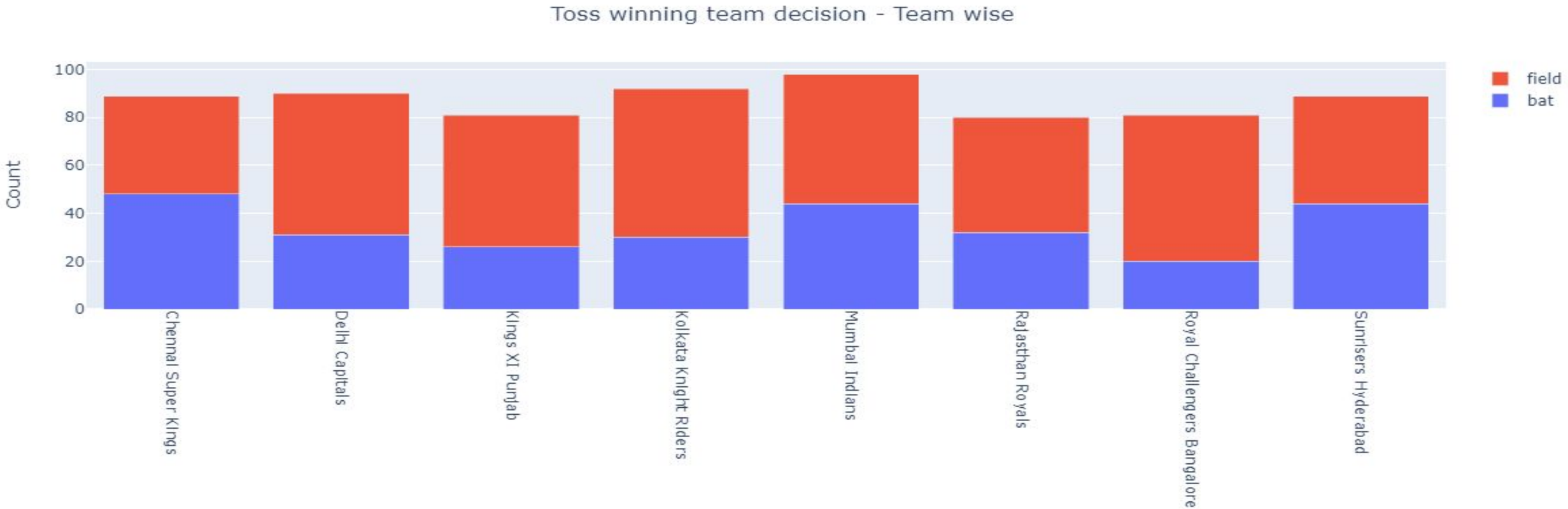
- How Uber, Ola estimates the fair amount after the ride?
- How Zomato, Swiggy predicts the delivery time after placing an order?
- How Amazon, Flipkart suggests items for you to buy?
- How Gmail auto-filters the mails to spam and non-spam categories?
- How Netflix, Hotstar recommends the shows of your choice?
- How cars24, carWale, carDekho computes resale price of used cars?

So, What is Data Science?

Data - *Consider IPL data*

How many times MI have won the toss?

season	city	date	team1	team2	toss_winner	winner
2008	Bangalore	4/18/2008	Kolkata Knight Riders	Royal Challengers Bangalore	Royal Challengers Bangalore	Kolkata Knight Riders
2008	Chandigarh	4/19/2008	Chennai Super Kings	Kings XI Punjab	Chennai Super Kings	Chennai Super Kings
2008	Delhi	4/19/2008	Rajasthan Royals	Delhi Daredevils	Rajasthan Royals	Delhi Daredevils
2008	Mumbai	4/20/2008	Mumbai Indians	Royal Challengers Bangalore	Mumbai Indians	Royal Challengers Bangalore
2008	Kolkata	4/20/2008	Deccan Chargers	Kolkata Knight Riders	Deccan Chargers	Kolkata Knight Riders
2008	Jaipur	4/21/2008	Kings XI Punjab	Rajasthan Royals	Kings XI Punjab	Rajasthan Royals
2008	Hyderabad	4/22/2008	Deccan Chargers	Delhi Daredevils	Deccan Chargers	Delhi Daredevils
2008	Chennai	4/23/2008	Chennai Super Kings	Mumbai Indians	Mumbai Indians	Chennai Super Kings
2008	Hyderabad	4/24/2008	Deccan Chargers	Rajasthan Royals	Rajasthan Royals	Rajasthan Royals
2008	Chandigarh	4/25/2008	Kings XI Punjab	Mumbai Indians	Mumbai Indians	Kings XI Punjab
2008	Bangalore	4/26/2008	Royal Challengers Bangalore	Rajasthan Royals	Rajasthan Royals	Rajasthan Royals
2008	Chennai	4/26/2008	Kolkata Knight Riders	Chennai Super Kings	Kolkata Knight Riders	Chennai Super Kings
2008	Mumbai	4/27/2008	Mumbai Indians	Deccan Chargers	Deccan Chargers	Deccan Chargers
2008	Chandigarh	4/27/2008	Delhi Daredevils	Kings XI Punjab	Delhi Daredevils	Kings XI Punjab
2008	Bangalore	4/28/2008	Chennai Super Kings	Royal Challengers Bangalore	Chennai Super Kings	Chennai Super Kings
2008	Kolkata	4/29/2008	Kolkata Knight Riders	Mumbai Indians	Kolkata Knight Riders	Mumbai Indians
2008	Delhi	4/30/2008	Delhi Daredevils	Royal Challengers Bangalore	Royal Challengers Bangalore	Delhi Daredevils
2008	Hyderabad	5/1/2008	Deccan Chargers	Kings XI Punjab	Kings XI Punjab	Kings XI Punjab
2008	Jaipur	5/1/2008	Rajasthan Royals	Kolkata Knight Riders	Rajasthan Royals	Rajasthan Royals



Science

"**Science** is the **intellectual** and practical activity encompassing the systematic study of the structure and behavior of the physical and natural world through observation and experiment." - Google dictionary

Machine Learning

Getting insights, Decision Making, Predictions

What are the chances for KXIP to win the match?

How many sixes will be hit in the next season?

How many times CSK won the matches without Dhoni's batting?

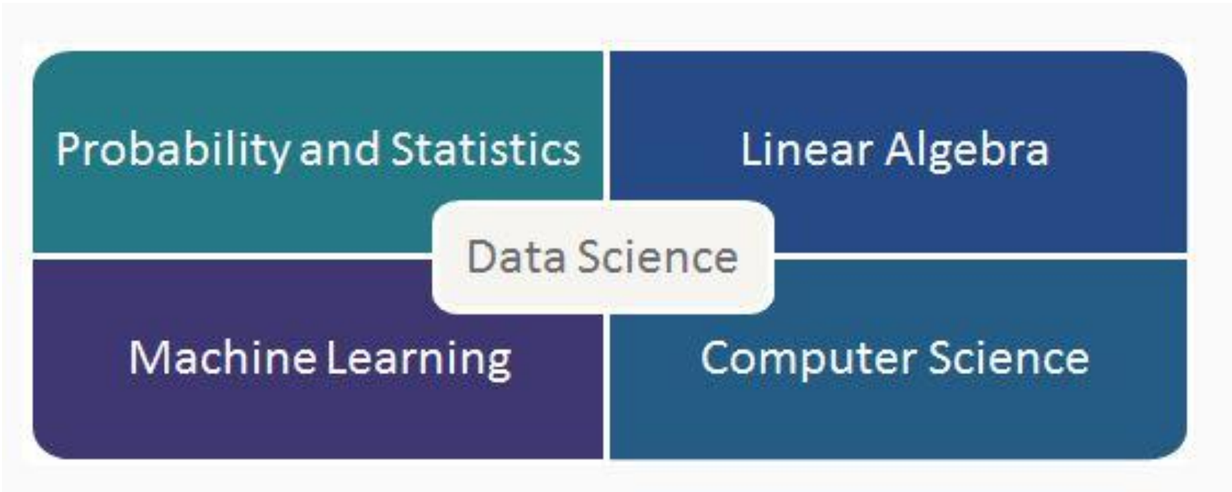
What are the chances for MI to win if Bumrah takes 3 wickets?

Which team has good homeground record?

Who will win the next IPL season?

What are the chances for RCB to lose if Kohli goes for a duck?

Definition



Data science, also known as **data**-driven **science**, is an interdisciplinary field about **scientific** methods, processes, and systems to extract knowledge or insights from **data** in various forms, either structured or unstructured.

It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, domain knowledge and information science.

Quiz 1

Which of the followings is(are) applications of Data Science in Banking?

1. Digital wallets
2. Fraud detection
3. Identifying the most suitable type of credit card for a customer
4. Regular transactions

Jargon Busting

Artificial Intelligence

Data Science

Computer Vision

Natural Language Processing

Machine Learning

Automation

Reinforcement Learning

Deep Learning

Is Data Science a subset of AI, a subset or neither?

What is AI?

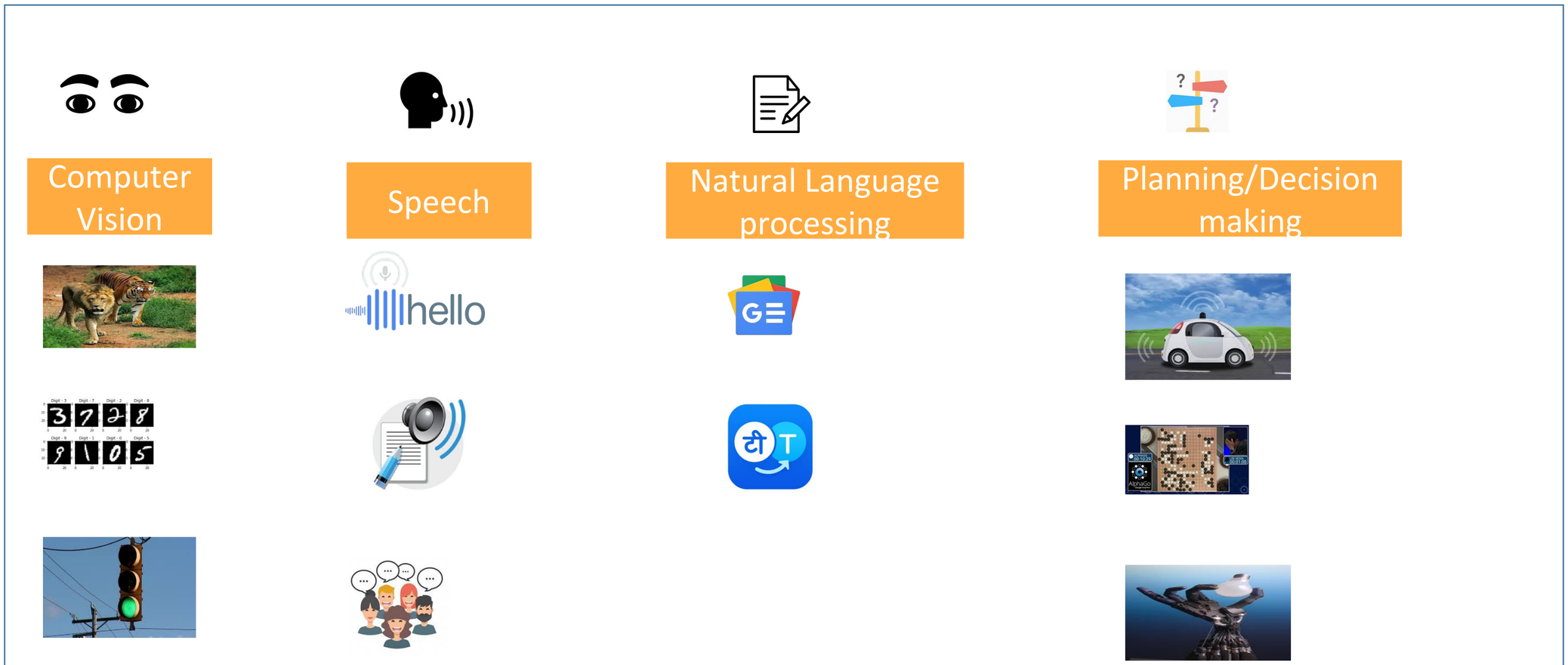
Let’s first talk about Natural Intelligence

Five Senses

- **Sight - Watch/Monitor**
- **Hearing - Listen/Talk/Interact**
- **Touch**
- **Smell**
- **Taste**

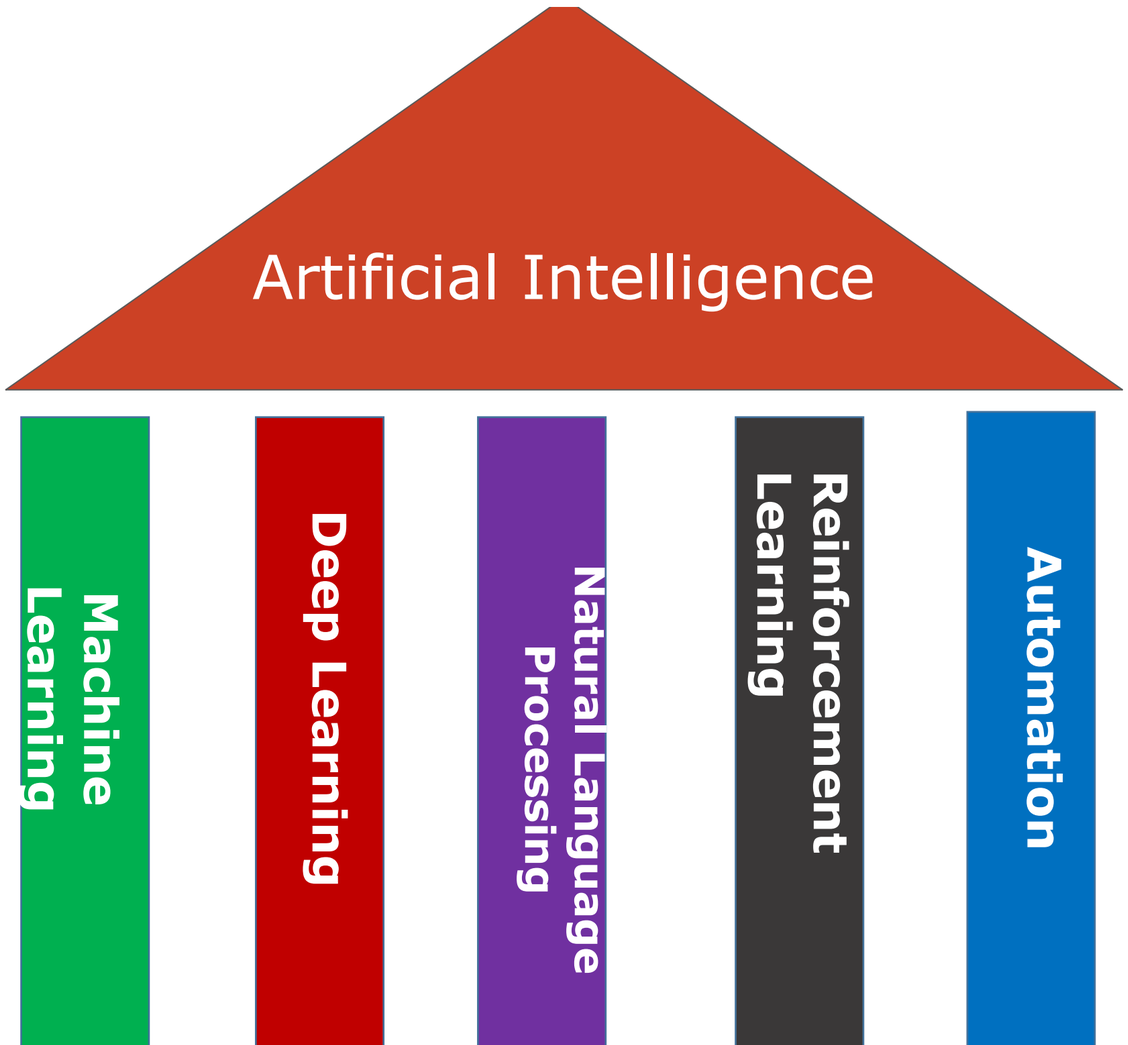
Is Data Science a subset of AI, a subset or neither?

What is AI?

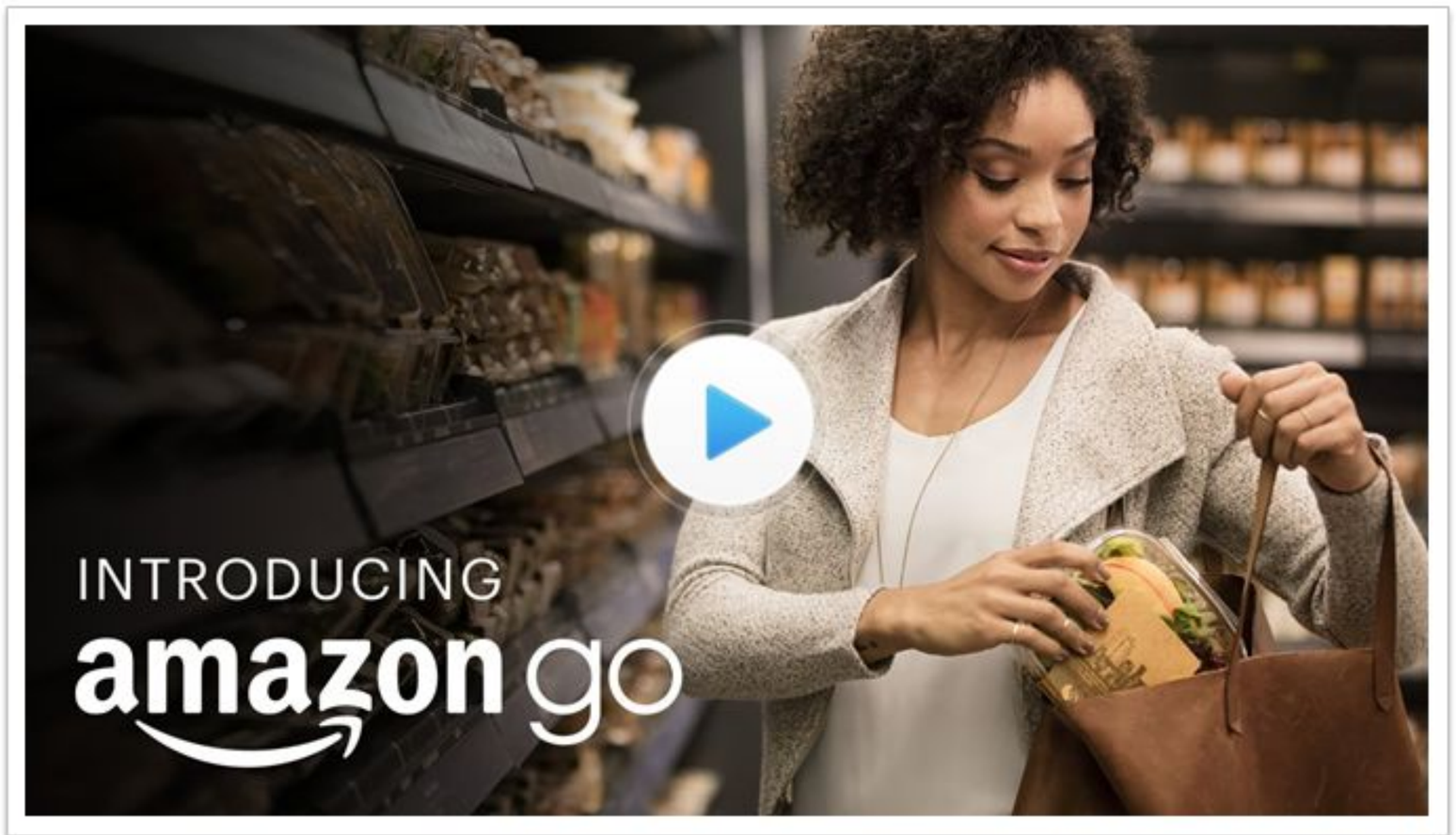


- **Abilities**
- **Tasks**
- **Methods**

So AI is



Amazon Go – Simplifying everyday life



Amazon Go Benefits

- Just walk in and go
- No lines, no checkout, no register
- Use amazon go application to enter
- Anything you pick up is added to your virtual cart and keep back is removed
- Walk out technology use computer vision, deep learning and many more technology
- To pay just move out of store and app deduct the charges from your amazon app.

Quiz 2

Chatbot(Virtual Assistant) is an example of -

1. Machine Learning
2. Deep Learning
3. Artificial Intelligence
4. Natural Language Processing

Quiz 3

Recommender System is an example of -

1. Machine Learning
2. Deep Learning
3. Artificial Intelligence
4. Natural Language Processing

What top business leaders think about DS, ML & AI

"Artificial Intelligence & Machine learning will lead to Ultimate breakthroughs."

- Satya Nadella(CEO- Microsoft)

"Machine learning is not important but also essential for Facebook's existence."

- Mark Zuckerberg(CEO- Facebook)

"We're moving from a mobile first world to AI first worlds."

- Sundar Pichai(CEO- Google)

"I have exposure to the most cutting edge AI, and I think people should be concerned by it."

- Elon Musk

Not only the Tech giants but the countries, governments also

- US has Chief Data Scientist of the United States Office of Science and Technology Policy.

- UAE has now AI minister.

- **"**Artificial intelligence is the future, not only for Russia, but for all humankind, Whoever becomes the leader in this sphere will become the ruler of the world.**"**

  - Vladimir Putin(Russian president)

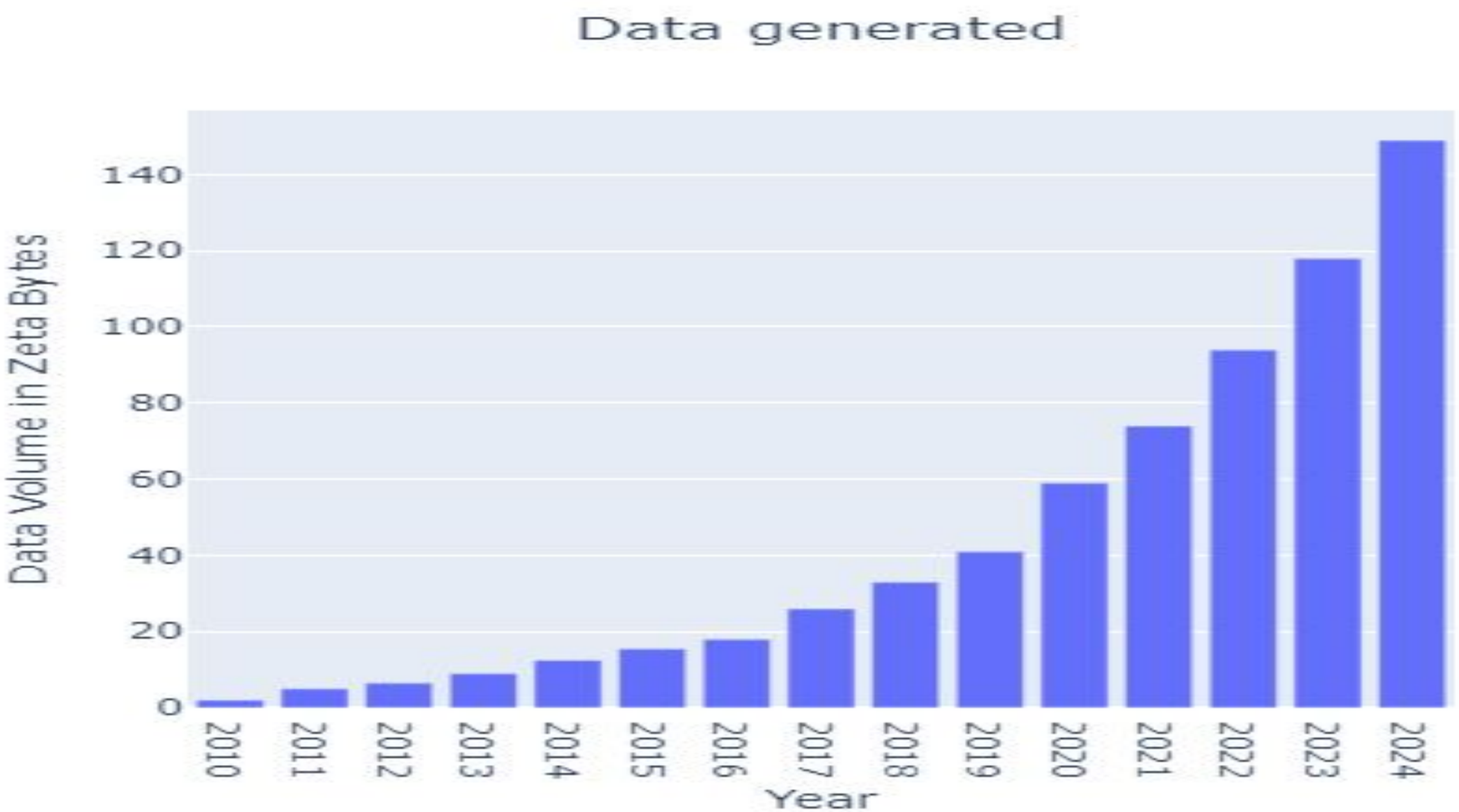
Evolution of Data Science

Factors that contributed to the evolution of Data Science -

- Abundance of data available
  - 2.5 quintillion bytes of data are produced by humans every day
- Better computation power
  - multicores, GPUs, TPUs, Cloud
- Better data management capabilities
  - NoSQL, Big Data(Hadoop and Spark), Cloud
- Drastic changes in hardware costs
  - Processors, Memory, I/O, Network
- Availability of high speed communication/cheaper internet access

Data Availability

- 90% of the data in the world today has been created in the last two years.
- Every minute on Facebook: 510,000 comments are posted, 293,000 statuses are updated, and 136,000 photos are uploaded.



Types of Data

	Structured	Semi-Structured	Unstructured
Rationale	Data is formatted, organized and transformed in a well-defined data model/manner.	It lies between Structured and Unstructured, i.e. partially structured. It does not confine to a rigid structure as of Structured Data.	Data is present in absolute raw form. It does not have a predefined format.
Examples	Relational data, SQL tables, Spreadsheets	XML,HTML data	Text, PDF, Word, Images, Audio, Video
Flexibility & Scalability	Less flexible and scalable	More flexible than Structured Data but less flexible and scalable as compare to Unstructured Data.	More flexible and scalable as compare to Structured and Semi Structured Data.
Performance	Highest	Lower than Structured Data but more than that of Unstructured Data	Lower than both Structured and Semi Structured Data.

Quiz 4

Email message field is an example of -

1. Structured Data
2. Unstructured Data
3. Semi-Structured Data
4. Both Semi-Structured and Unstructured Data

Hardware cost comparison

- 5MB hard disk in 1956
- **Size** - 50 magnetic disks each 24 inches in diameter
- **Cost** - $3200 per month

- 128GB micro sd card today
- **Size** - 15mm x 11mm x 1.0mm
- **Cost** - $15 dollar

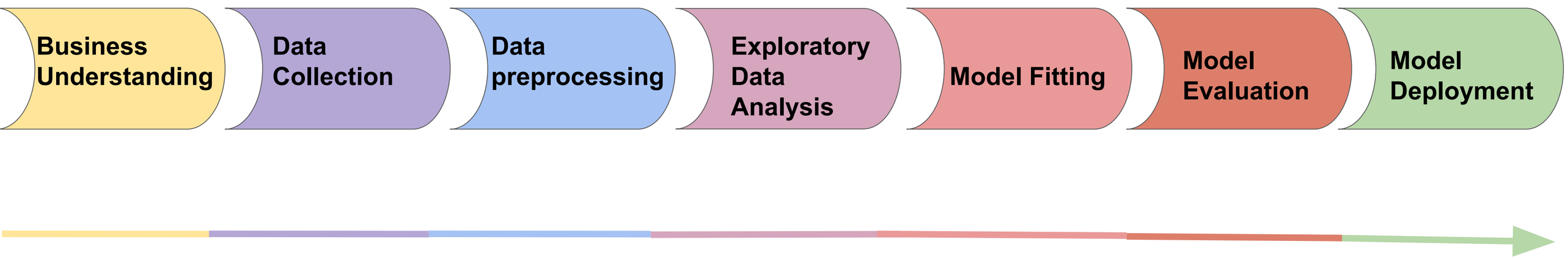
DS, ML & AI Adoption by Tech giants

Organization	Few Sample Applications
Google	• Speech recognition • Google translator, • Image search • Search suggestion
Facebook	• Facial recognition • Friend suggestion • Targeted advertisement
Microsoft	• Speech recognition • Text recognition • Text generation
Amazon	• Speech recognition • Targeted advertisement • Image recognition
IBM	• Image recognition • Speech recognition

All the Tech giants are making Multi-Billion Dollar investments to gain the first mover advantage.

- Facebook in 2013: AI Lab, DeepFace
- Yahoo in 2013: LookFlow
- Google in 2014: DeepMind $500 million
- IBM in 2014: $1 billion in Watson
- Toyota in 2014: $1 billion AI and Robotics Lab, Silicon Valley

Data Science Life Cycle



Business Understanding/Defining the Problem Statement(s)

Business Understanding plays a key role in the success of a Data Science project. It can make or break the entire project. So the problem needs to be well defined.

So the problem statement must be simple and clear as per the business objectives and constraints.

A good data science problem should be relevant, specific, and unambiguous. It should align with the business strategy.

Understand better by asking questions like -

- What are the challenges in the existing solution?
- What are the business objectives and vision?
- What resources are available?
- What kind of data science problem is it?
- What are the potential benefits?
- What are the risk associated in continuing this project?
- What should be the duration to complete this project?

Data Collection

Since there is no Data Science without Data so once we understand the business objectives and constraints, we should start collecting the data.

Here, we need to determine -

- What are the potential data sources??
- How much data is sufficient to proceed further?
- Whether the collected data is good representation of the problem or not?

***Data Collection** process is often iterative and generally done through **ETL**(**Extraction**, **transformation** and **Loading**) pipeline.*

Data Preprocessing

Making the data ready for analysis. Majorly involved steps are -

- Checking for Data Consistency
- Checking for Missing values and imputing them
- Removing the duplicate records/data samples
- Outlier detection
- Mapping different variants of an entity to the root
- Data Encoding
- Data Normalization
- Shuffling/Sorting the data

Exploratory Data Analysis

Some of standard practices involved here are -

- Variable Identification and its properties
- Finding relationships among the variables
- Uni-Variate Analysis
- Bi-Variate Analysis
- Multivariate Analysis
- Exploring through Data Visualization
- Feature Engineering

Model Fitting

This is the most interesting and core step of a Data Science project. Many people call it "*a stage where magic happens*".

Modelling is done to find patterns or behaviour from the data and making predictions based on that.

Based on the type of the target variable and business problem, various algorithms are fitted in this step like -

- Regression
- Classification
- Clustering
- Recommendation

...etc.

Model Evaluation

Built models are evaluated to assess their usefulness/goodness using different evaluation measures.

Similarly, models can be characterized based on the business constraints such as

- Accuracy
- Relevance
- Latency
- Interpretability

Some of the commonly used evaluation measures are -

- **Regression** - Mean Squared Error(MSE), Root Mean Squared Error(RMSE), Mean Absolute Error(MAE)
- **Classification** - Accuracy, Precision, Recall, F1-score, Log loss etc.
- **Clustering** - Elbow method, Inter cluster distance, Intra cluster distance, Silhouette analysis

Model Deployment

Making the fine-tuned model available to end users/client/stakeholders.

Model deployment also completes the Data Science project Life Cycle.

> *"No machine learning model is valuable, unless it’s deployed to production." – Luigi Patruno*

Quiz 5

Choose the right order of stages under Data Science project Life Cycle.

1. Business Understanding -> Data Collection -> Exploratory Data Analysis -> Data Preprocessing -> Model Fitting -> Model Deployment -> Model Evaluation

2. Business Understanding -> Data Collection -> Data Preprocessing -> Exploratory Data Analysis -> Model Fitting -> Model Deployment -> Model Evaluation

3. Business Understanding -> Data Collection -> Data Preprocessing -> Exploratory Data Analysis -> Model Fitting -> Model Evaluation -> Model Deployment

4. Business Understanding -> Data Preprocessing -> Data Collection -> Exploratory Data Analysis -> Model Fitting -> Model Evaluation -> Model Deployment

Data Science in various domains

Finance

Retail & ecommerce

Airline

Energy

Medical/Healthcare

Telecom

Finance

- Fraud detection
  - ➔ Identifying frauds involving credit/debit cards, forging checks.
  - ➔ Identifying misleading accounting/transaction practices.
- Loan And Insurance Underwriting
  - ➔ Better loan amount prediction
  - ➔ Prediction of insurance types and coverage plans
  - ➔ Better premiums and policy updates
- Portfolio management
  - ➔ Matching investments to objectives
  - ➔ Balancing risk against performance and better decision making system
- Algorithmic trading
  - ➔ Defining set of instructions are based on timing, price, quantity or any mathematical model

Retail & ecommerce

- Recommending most similar items/products
- Product pricing, promotions & discounts
- Customer segmentation or targeting potential customers
- Sales forecasting
- Fraud detection in online purchases and returns
- Target campaigning
- A/B Testing

Airline

- Predicting flight delays
- Route optimization
- Fare estimates
- Forecasting the demand, no. of passengers
- Predicting the maintenance

Energy

- Predicting the influence of weather conditions on the power grid
- Predicting and managing the load accurately
- Improving the efficiency of building appliances and materials
- Monitoring the equipment conditions and performance level
- Real-time customer billing
- Optimizing asset performance

Medical/Healthcare

- Personalized health monitoring
- Detecting rare diseases
- Automating image/reports diagnosis
- Monitoring risk factors
- Better Clinical trial of new drugs

Telecom

- Churn analytics & prevention
- Customer acquisition strategies
- Network management and optimization
- Social media and sentiment analysis
- Location based initiatives

Quiz 6

Choose the correct match(es) of applications w.r.t their domains.

1. Loan amount prediction -> Finance and Banking
2. Recommending most similar products -> Retail & ecommerce
3. Forecasting the number of passengers -> Energy
4. Detecting rare diseases -> Medical/Healthcare

Case Study - Bike Portals

388K followers

Portal **A**

392K followers

Portal **B**

289K followers

Portal **C**

User enter the details
&
Immediately gets the price

Data Cleaning/Pre-processing

- Collecting data from multiple sources
- Data Consistency
- Data Mapping
- Checking for Missing values
- Outlier detection
- Data Encoding
- Data Normalization

Outliers example

Case 1

Very rare and unique registration number
e.g. AK-0047, 0000, 1111 etc.

Case 2

Premium bikes with less reselling price

Few Driver features

- Age/Registration Year
- Manufacturer
- Number of Kilometers driven
- Ownership type(First/Second/Third etc.)
- Insurance(active or not)
- Warranty period
- Engine Capacity
- Fuel Tank Capacity
- No. of Gears
- Self start(Yes/No)

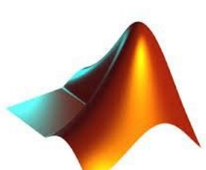
Most commonly used Programming Languages in Data Science

Some of the most popular languages in this domain are -

- Python
- R
- Julia
- MATLAB
- SQL
- SAS
- Scala







Python

Python is the default choice for a range of Data Science tasks such as data manipulation, data visualization, machine learning, deep learning, NLP etc.

It is open source, object-oriented, easy to use and developer friendly.

Python Features

- Supports various file imports, exports and sharing options
- Rich set of useful libraries
- Perfectly suited for tasks such as data analysis, visualization, modelling etc.
- Powerful libraries for data science workloads such as Pandas, matplotlib, scikit-learn, keras, tensorflow etc.
- Wide and strong community support

R

R is an open source language and software environment for statistical computing and graphics. Time series analysis, clustering, statistical tests, linear and non-linear modeling are some of the many statistical computing and analysis options provided by R.

It was developed by statisticians for statisticians!

R Features

- Excellent plots for data analysis, e.g. using ggplot.
- Being a vector language, R can do many things at once, functions can be added to a single vector without putting it in a loop.
- The public R package archive consists of contributed packages from almost 8,000 networks.
- An active community of contributors.

Data Science Platforms

The most popular Data Science platforms are:

- Google Cloud Platform
- H2O
- IBM Watson
- AWS
- Apache Spark MLlib

Google Cloud Platform

Google Cloud Platform, offered by Google, is a suite of cloud computing services that runs on the same infrastructure that Google uses internally for its end-user products, such as Google Search and YouTube.

It provides a series of modular cloud services including computing, data storage, data analytics and machine learning.

GCP Features:

- Server less, fully managed computing
- Secured platform
- Better data center
- Powerful for data analytics
- Infrastructure developed keeping future in mind

H2O

H2O is open-source software for big-data analysis. It is produced by the company H2O.ai. H2O allows users to fit thousands of potential models as part of discovering patterns in data.

It is used for exploring and analyzing datasets held in cloud computing systems and in the Apache Hadoop Distributed File System as well as in the conventional operating-systems.

H2O Features:

- Algorithms are developed from the ground up for distributed computing and for both supervised and unsupervised approaches.
- You can access it from Python, R, Flow and many more.
- AutoML can be used for automating the machine learning workflow.

IBM Watson

Watson is a question-answering computer system capable of answering questions posed in natural language.

It was created as a question answering computing system that IBM built to apply advanced natural language processing, information retrieval, knowledge representation, automated reasoning, and machine learning technologies to the field of open domain question answering.

Watson Features:

- Accelerate research and discovery
- Detect liabilities and migrate risk
- Scale expertise and learning
- Learn more with less data
- Reimagine your workflow

Amazon Web Services

Amazon Web Services is a subsidiary of Amazon.com that provides on-demand cloud computing platforms to individuals, companies and governments, on a paid subscription basis.

The technology allows subscribers to have at their disposal a virtual cluster of computers, available all the time, through the Internet.

AWS Features:

- Provide best practice recommendations to improve performance and efficiency.
- It has AWS Trusted Advisor that draws upon best practices learned from operational history.
- The AWS Support API allows you to programmatically interact with your Support cases.
- It provides third party software support.

Apache Spark MLlib

Spark MLlib is Apache Spark's Machine Learning component. One of the major attractions of **Spark** is the ability to scale computation massively, and that is exactly what you need for machine learning algorithms.

Spark Mllib Features:

- It is easy to use at it support Java, Scala, Python and R.
- It has better performance and is faster than MapReduce.
- It can run anywhere like on Hadoop Apache Mesos, Kubernetes, standalone, or in the cloud, against diverse data sources.

Summary

In this session, we have learnt about :

- What Data Science means and how it has evolved over the period of time
- Why we need to use Data Science
- Different components of Data Science
- How Data Science is different from Machine Learning and Artificial Intelligence
- Data Science Project Life Cycle
- Applications of Data Science in various domains
- Scenario specific Case Study
- Different Data Science platforms