DATA SCIENCE

Dr. Muhammad Nadeem Majeed
nadeem.majeed@pucit.edu.pk

Today's Agenda

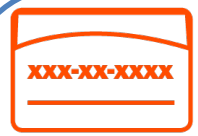
- Review of previous lecture
- How to Do Data Science?
- Languages, Tools and Techniques
- Life Cycle of a Data Science Project
- Industry Job Roles in Data Science

Recap of Previous Lecture

Structured Data

Structured Data

Examples



Social Security Number

Date

Phone Numbers

Customer Name

Transaction information

Credit Card number

Characteristics

Pre-defined Data Model

Easy to Search

Text-based

Applications

Airline reservation systems

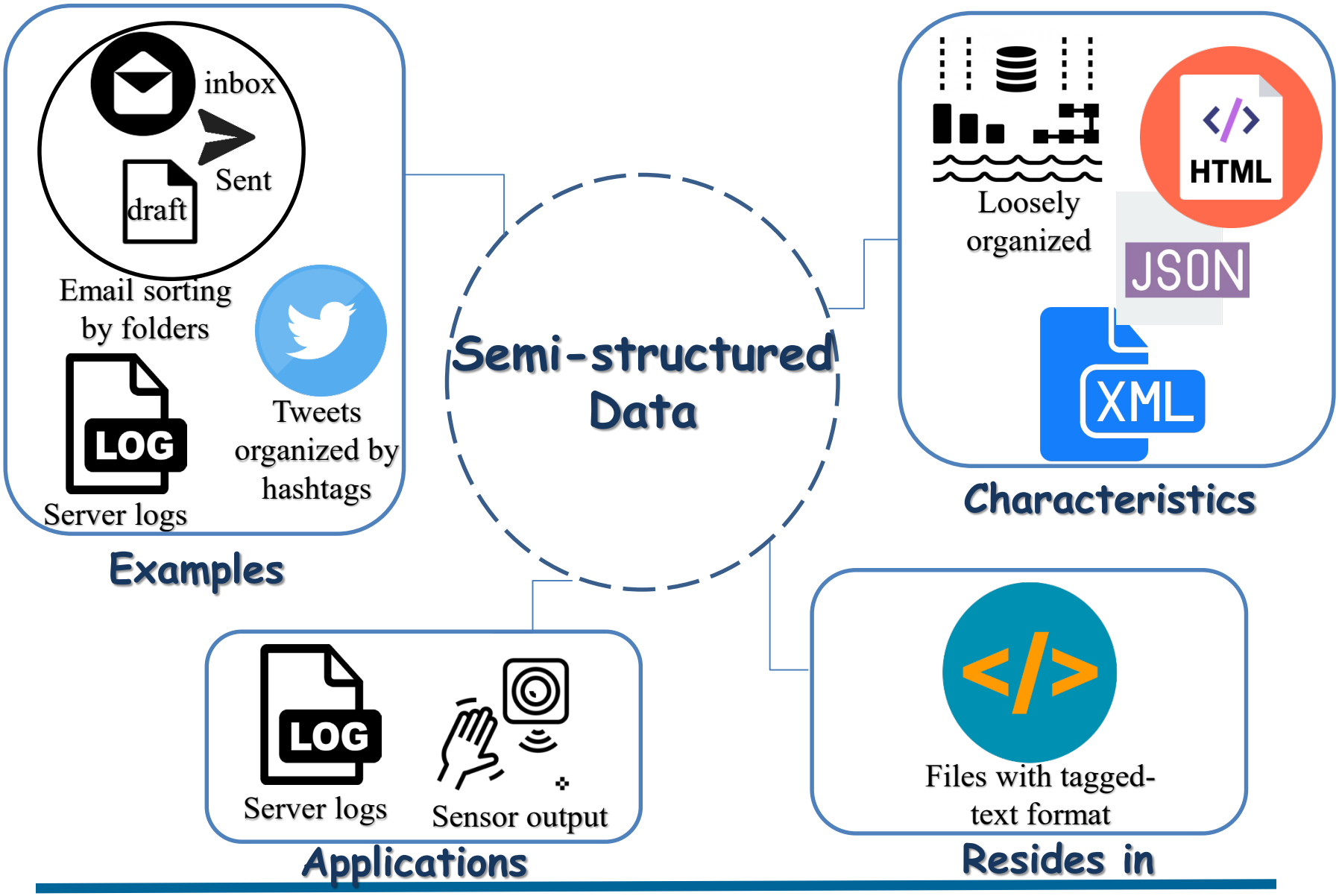
Inventory control

Resides in

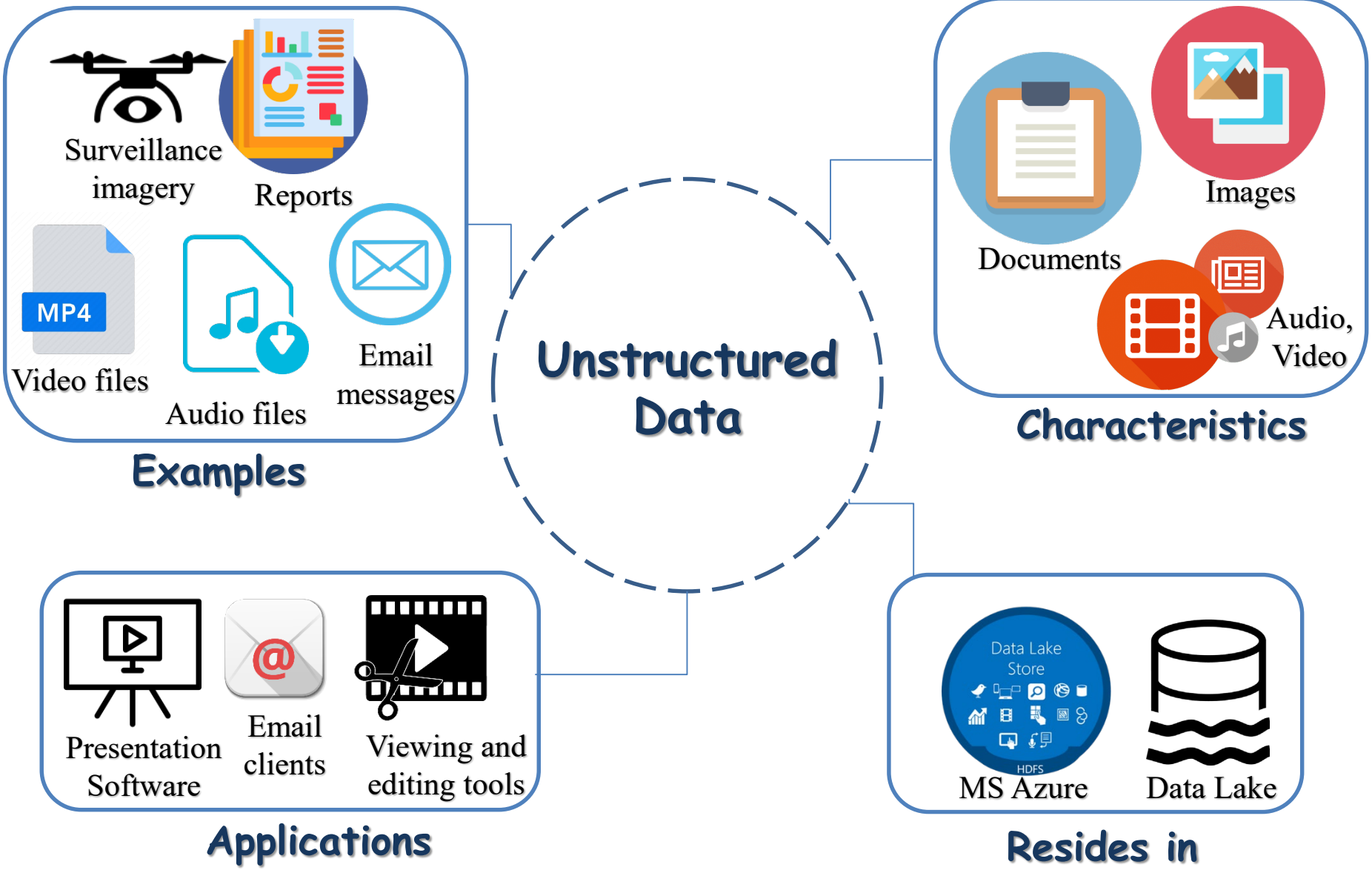
Data Mart

Database

Semi-structured Data

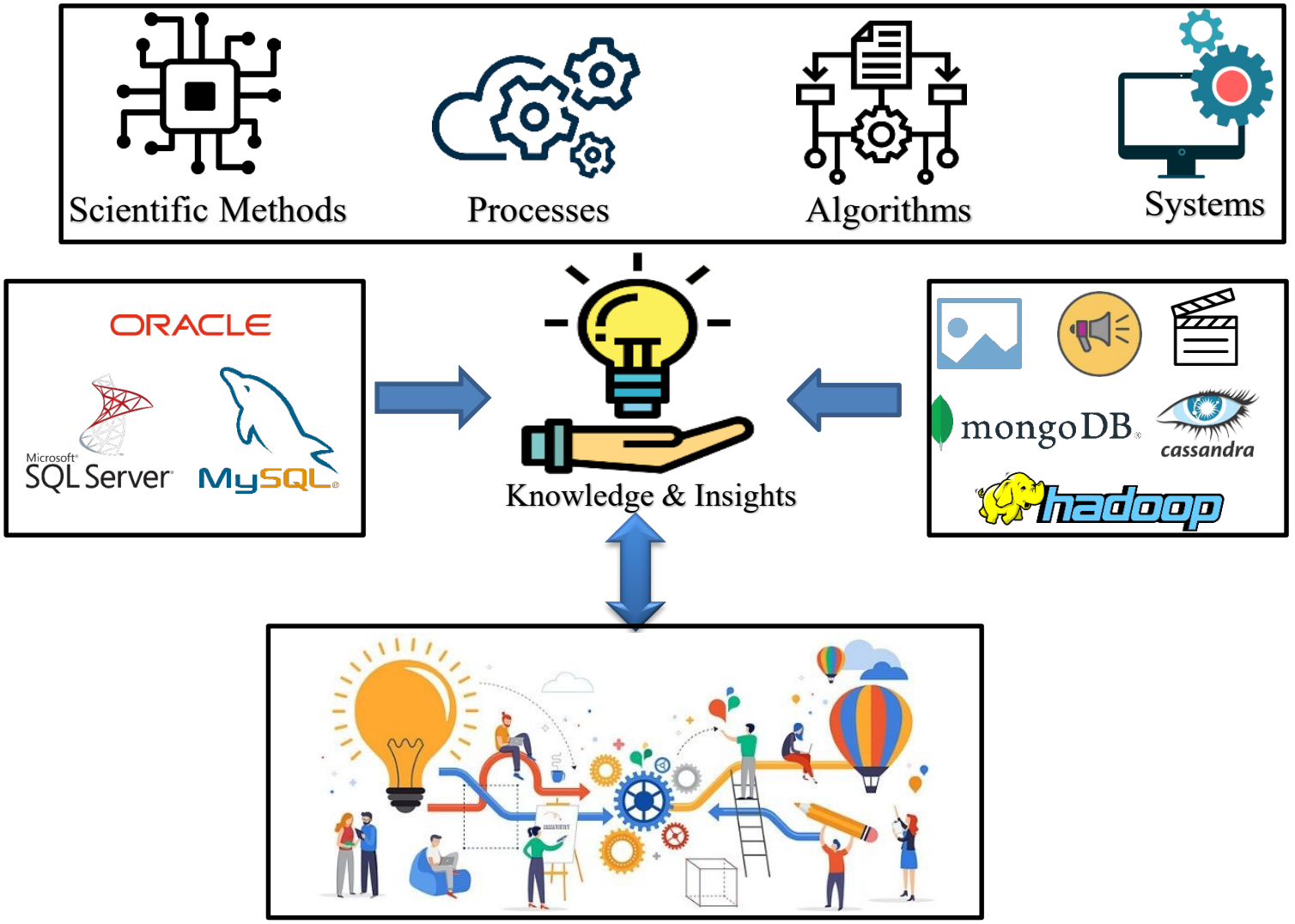


Unstructured Data

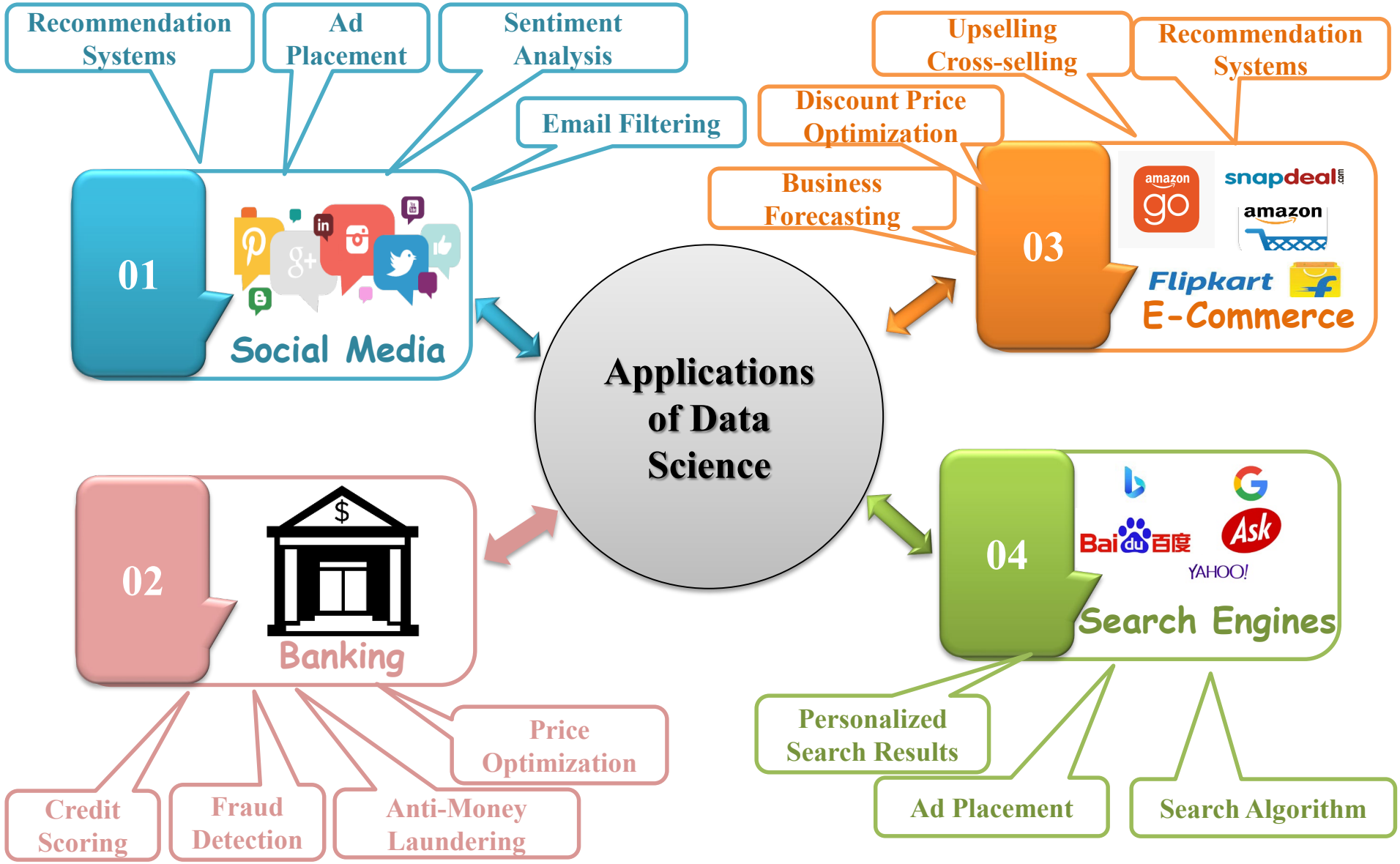


What is Data Science?

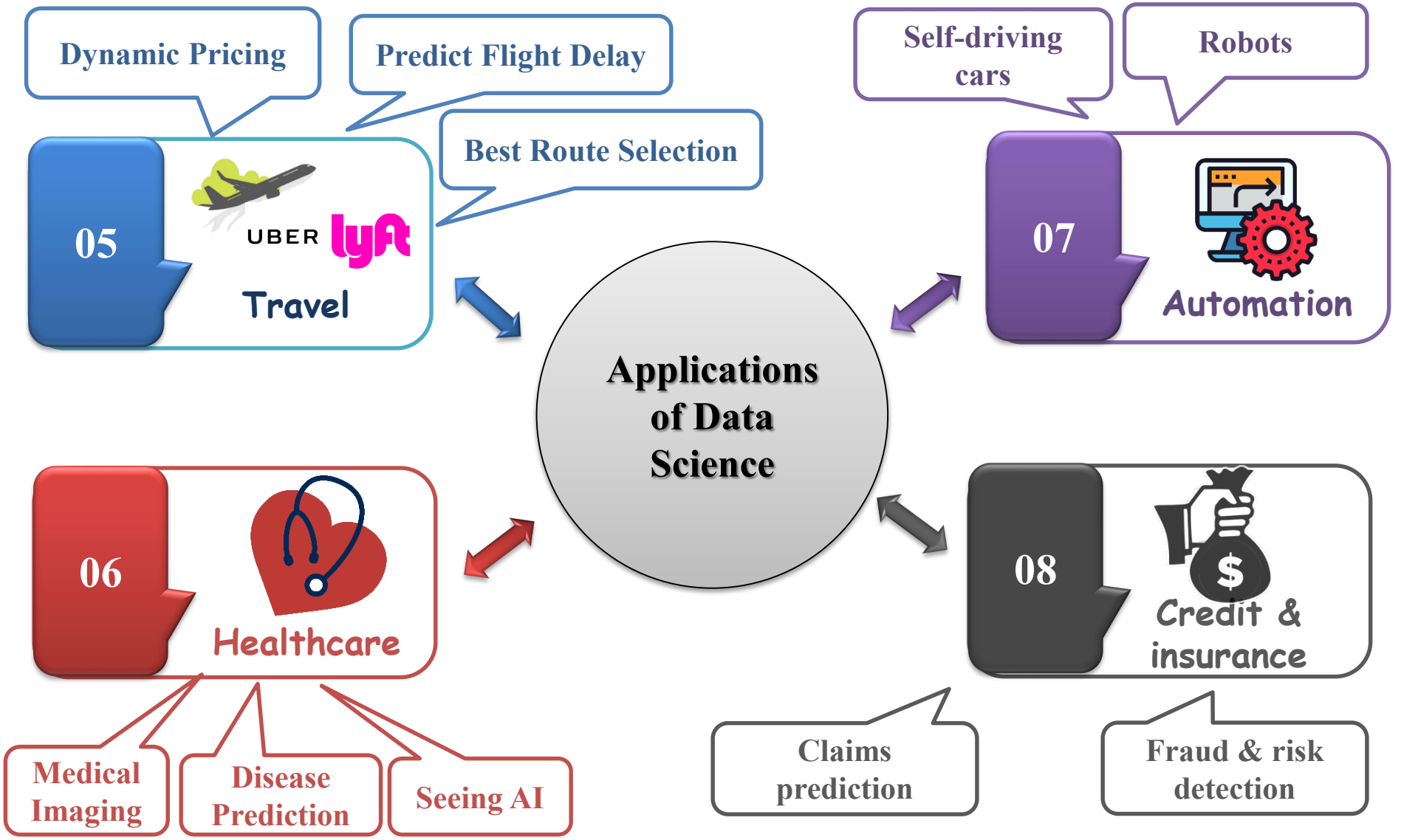
Data Science is an Inter-Disciplinary Field that uses



Applications of Data Science



Applications of Data Science (cont…)

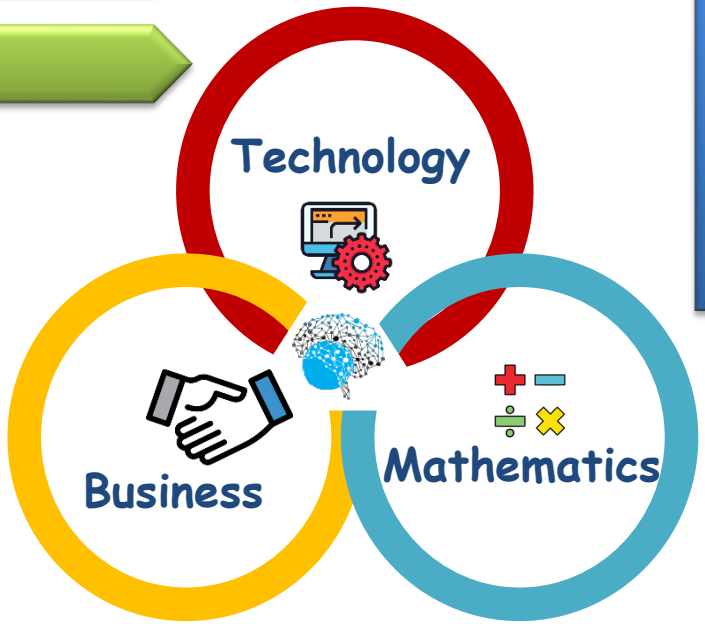


How to Do Data Science? Languages, Tools and Technologies

Who is a Data Scientist?

1. Data Scientist
2. Skill Set
3. Programming languages
4. Tools
5. Techniques

A data scientist is a professional responsible for **collecting**, **analyzing** and **interpreting** extremely large amounts of structured and unstructured data in order to gain useful insights to grow the business



Skill Sets of a Data Scientist

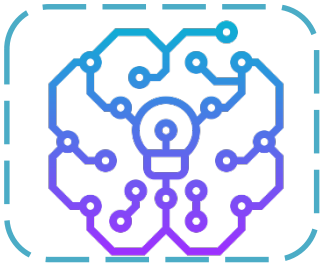
1. Data Scientist
2. **Skill Set**
3. Programming languages
4. Tools
5. Techniques

Statistics

Programming Languages

Data extraction & processing

Big Data processing framework



Machine Learning

Data Visualization

Data wrangling & exploration

Programming Languages for Data Science

1. Data Scientist
2. Skill Set
3. **Programming language**
4. Tools
5. Techniques

Python

R

Julia

Tools for Handling this Big Data (3Vs)

Tools are softwares that are used to apply DS techniques to perform a task.

1. Data Scientist
2. Skill Set
3. Programming language
4. **Tools**
5. Techniques

VOLUME

VARIETY

VELOCITY

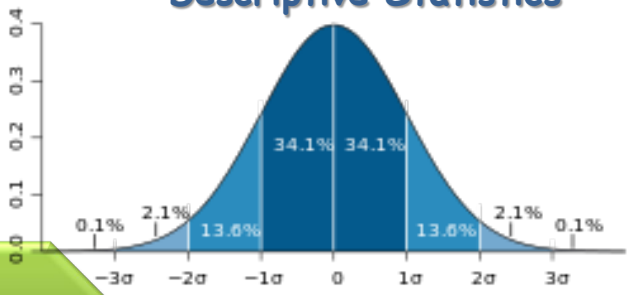
Python Libraries for Data Science Tasks

Techniques for Data Science

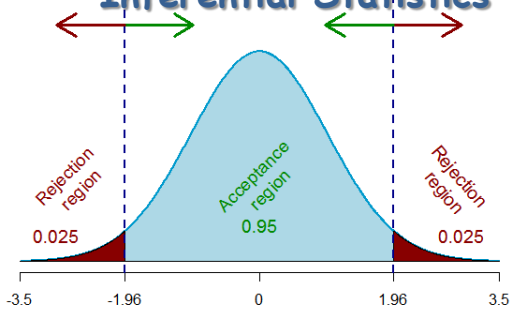
1. Data Scientist
2. Skill Set
3. Programming language
4. Tools
5. **Techniques**

Techniques are set of procedures that are followed to perform a task. Tools and techniques together helps in data collection, data storage, data preparation, data analysis, data modeling and data visualization

Descriptive Statistics

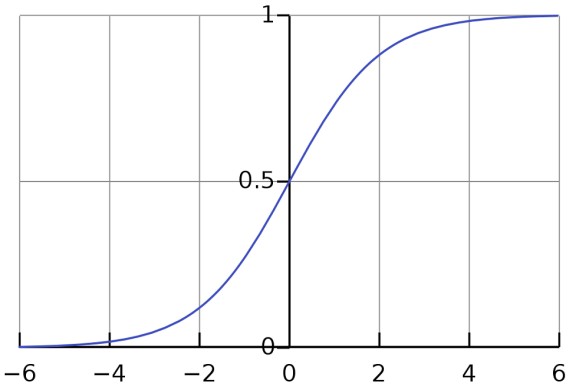


Inferential Statistics

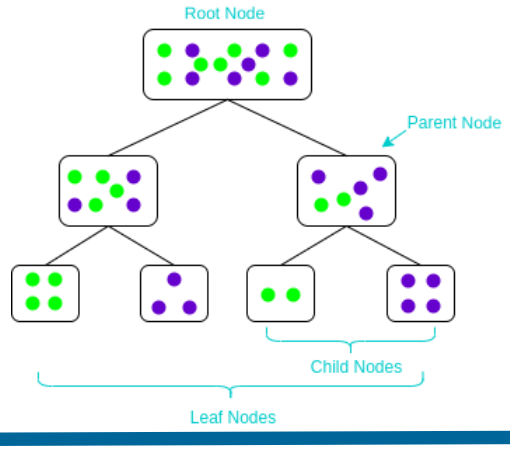


Classification Techniques

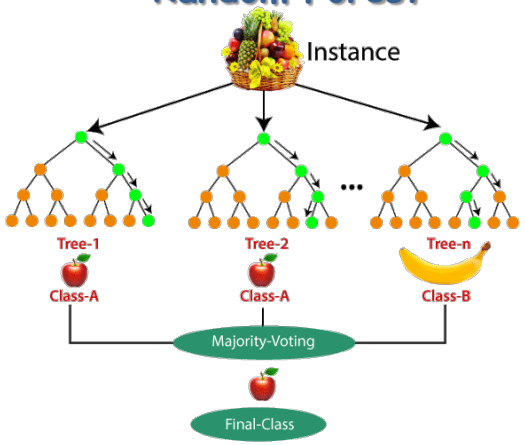
Logistic Regression



Decision Tree



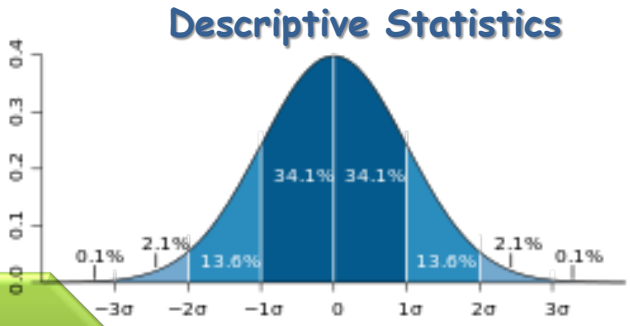
Random Forest

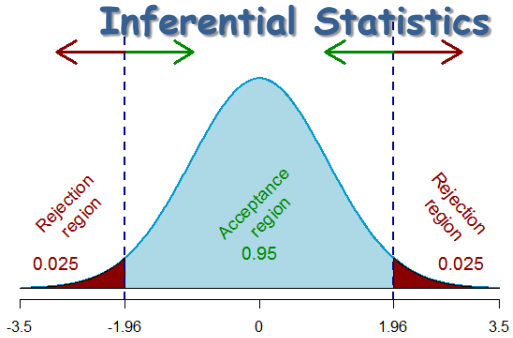


Techniques for Data Science

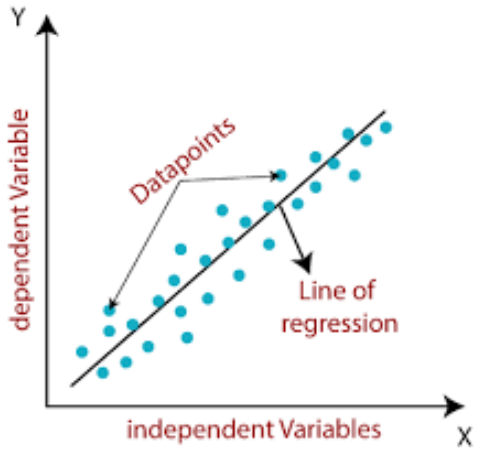
1. Data Scientist
2. Skill Set
3. Programming language
4. Tools
5. **Techniques**

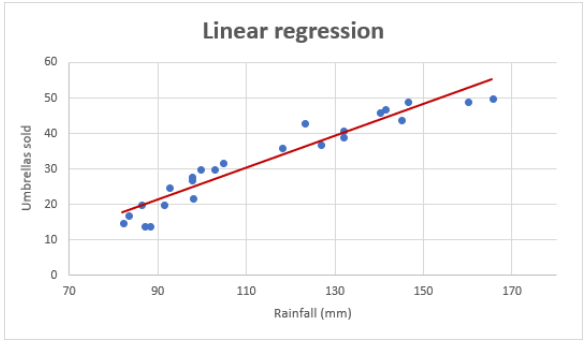
Techniques are set of procedures that are followed to perform a task. Tools and techniques together helps in data collection, data storage, data preparation, data analysis, data modeling and data visualization

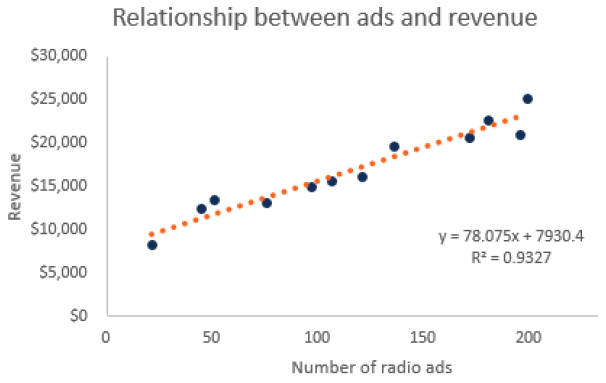




Regression Techniques





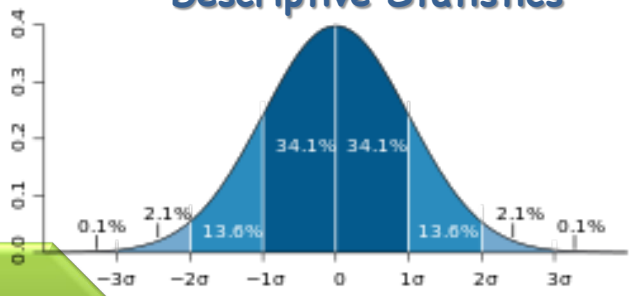


Techniques for Data Science

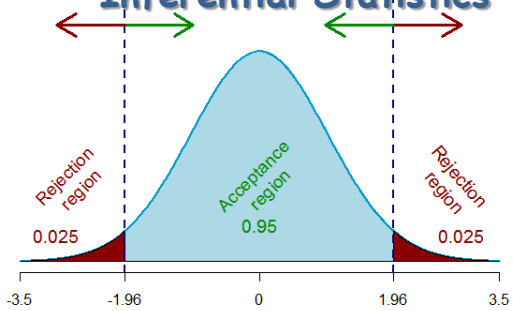
1. Data Scientist
2. Skill Set
3. Programming language
4. Tools
5. **Techniques**

Techniques are set of procedures that are followed to perform a task. Tools and techniques together helps in data collection, data storage, data preparation, data analysis, data modeling and data visualization

Descriptive Statistics

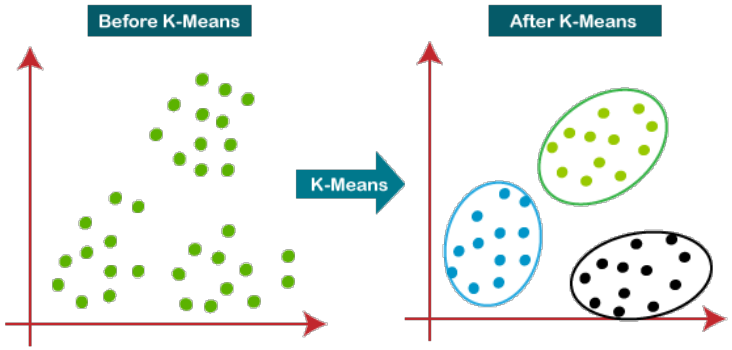


Inferential Statistics

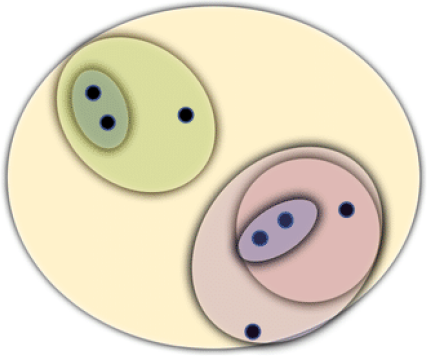


Clustering Techniques

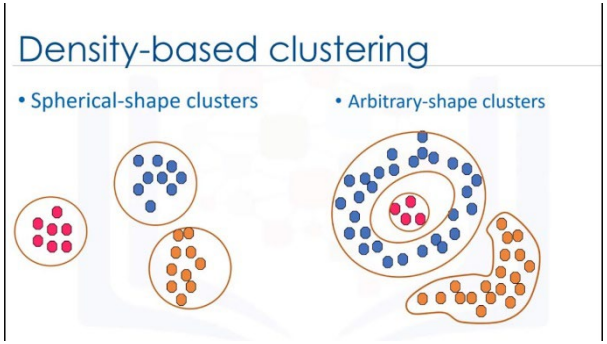
K-Means Clustering



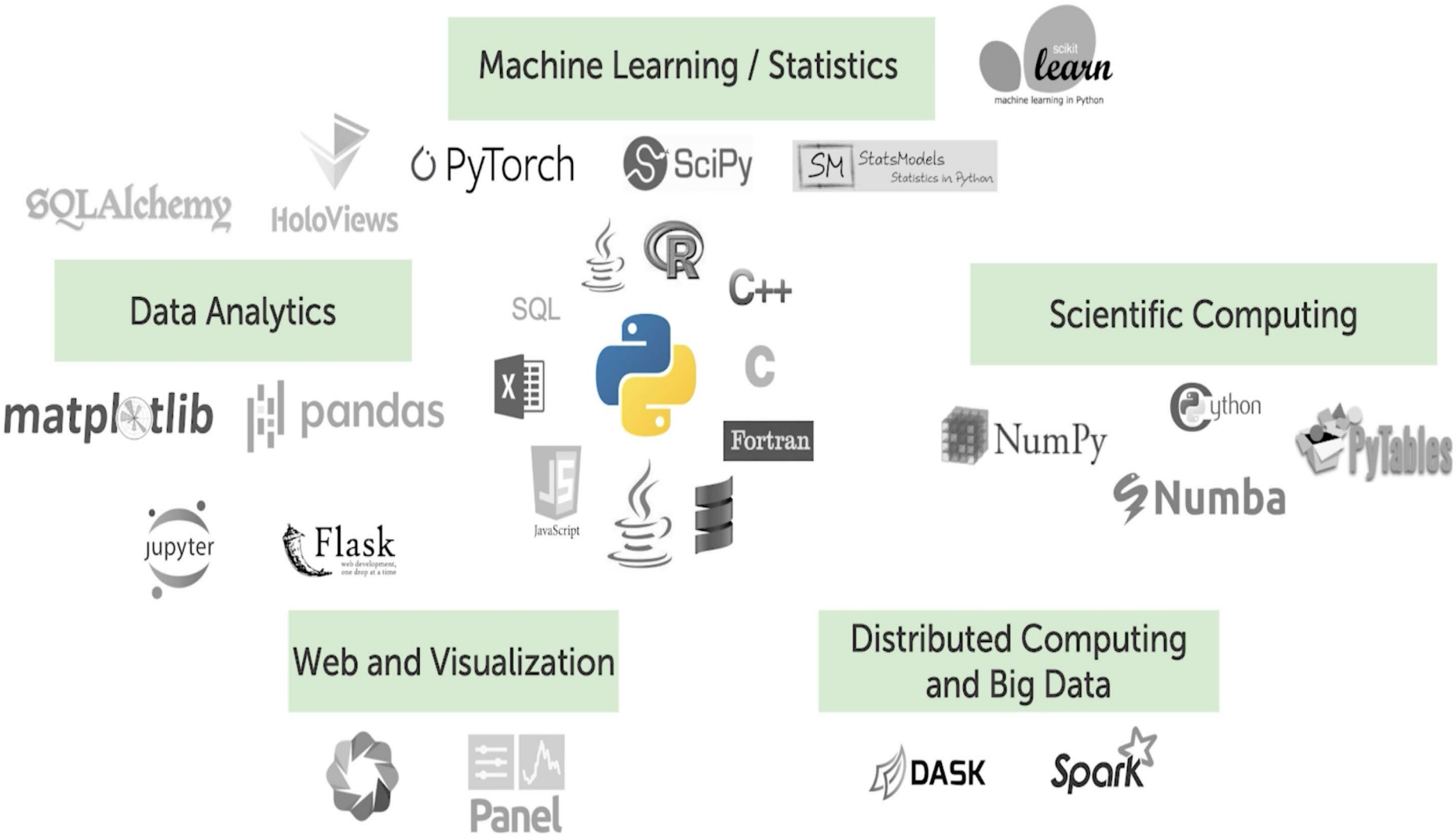
Hierarchical Clustering



DB SCAN

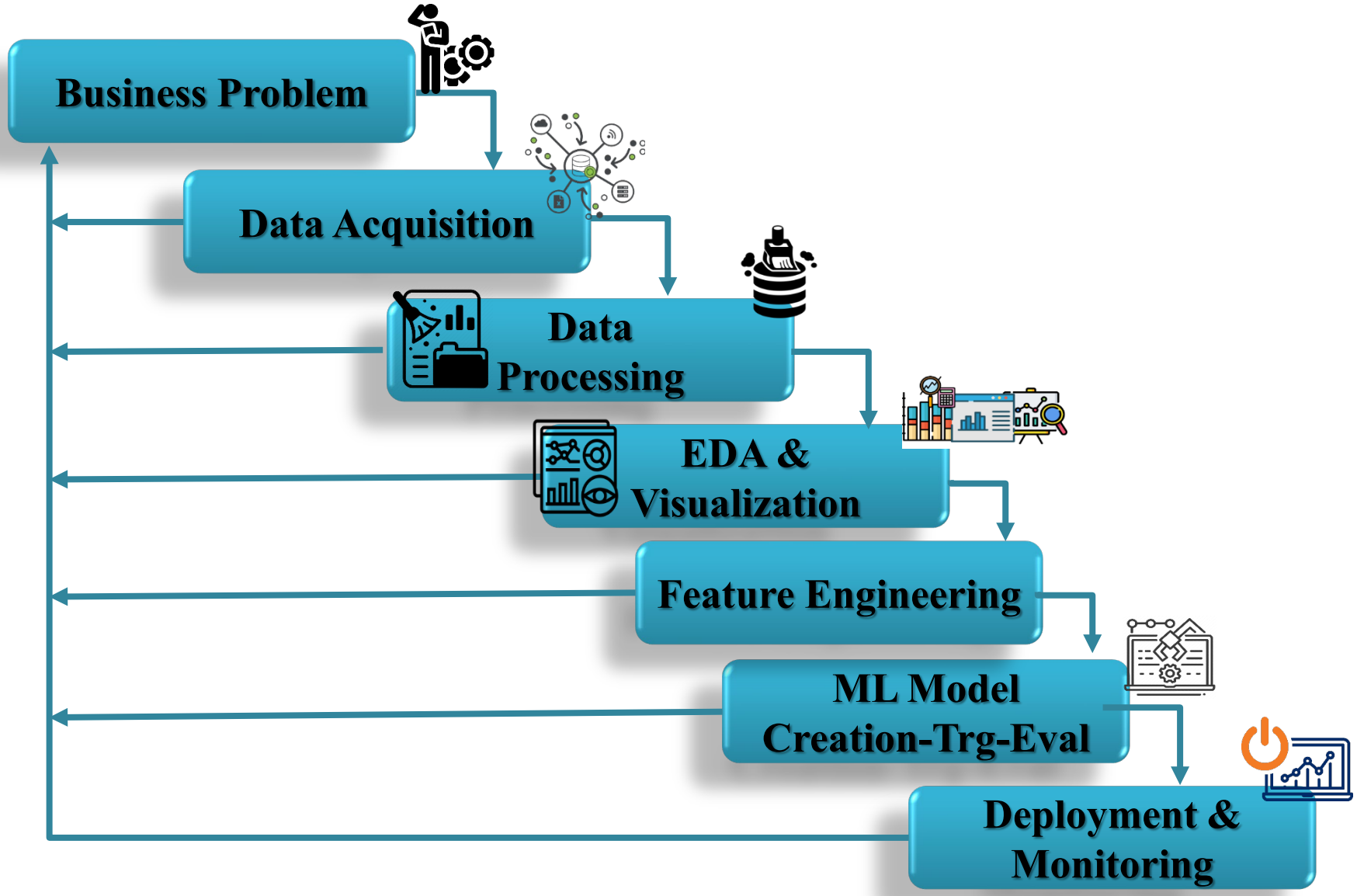


Why is Data Science so Complicated?

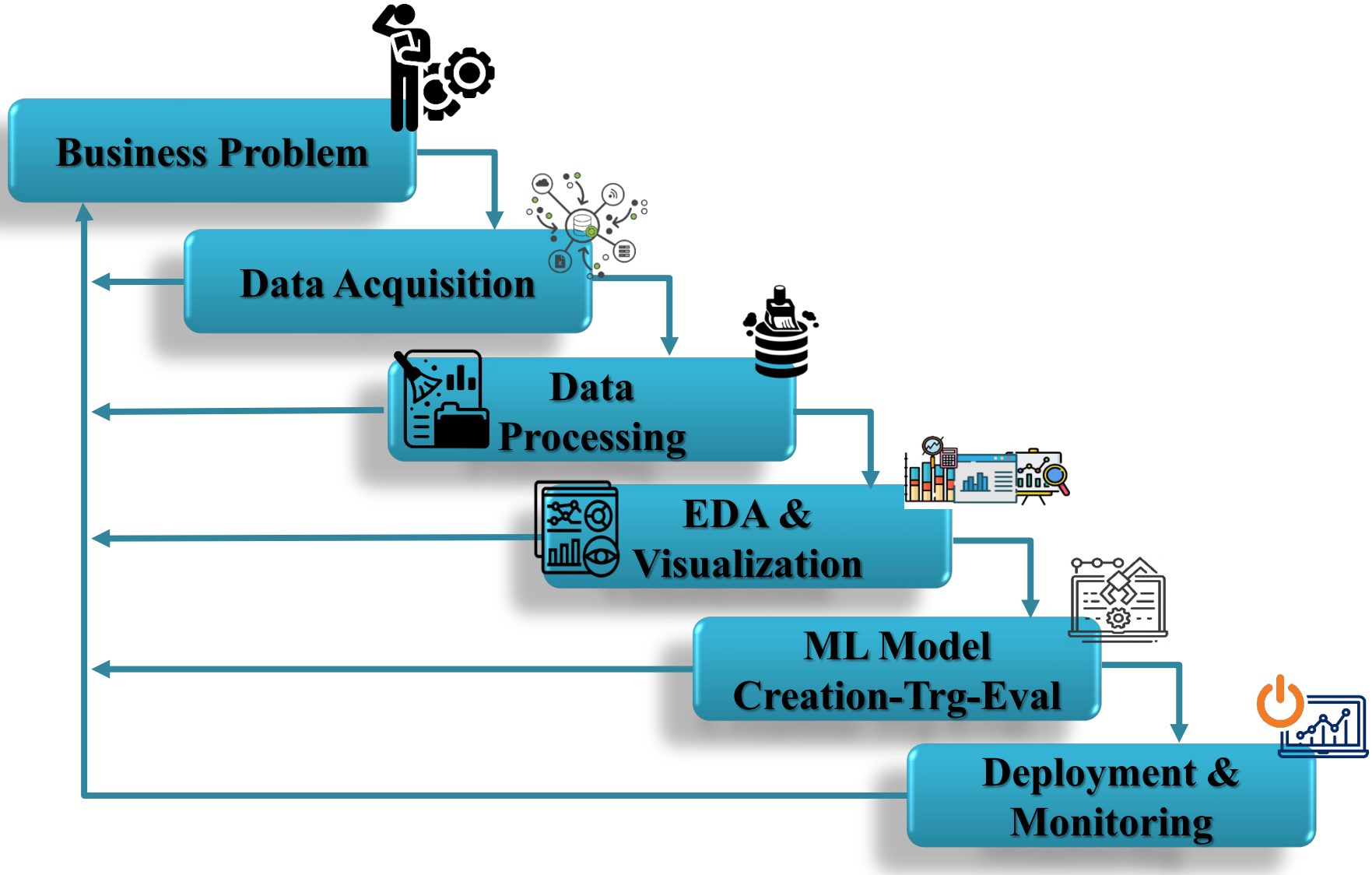


Data Science Life Cycle

Overview of Data Science Life Cycle

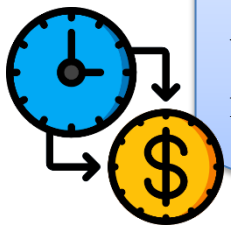


Overview of Data Science Life Cycle



Understanding Business Problem

1. **Business Problem**
2. Data Acquisition
3. Data Processing
4. EDA & Visualization
5. Model Creation-Trg-Eval
6. Deployment & Monitoring



Most critical phase of a Data Science Life Cycle, if conducted will saves lot of time, money and resources.

Understand the problem by talking to the stake holders & domain experts to get the clear understanding of the problem and document all the requirements.

Identify the key business variables that need to be predicted

Define the success criteria and success measuring metrics (KPIs & SLAs)

Data Acquisition

1. Business Problem
2. **Data Acquisition**
3. Data Processing
4. EDA & Visualization
5. Model Creation-Trg-Eval
6. Deployment & Monitoring

What data do we need for our project?

What are the data sources and data format?
Where is the data located?

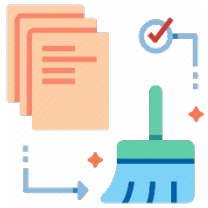
How can we obtain the data?

What is the most efficient way to store and access all of it for later processing?

Data Processing

1. Business Problem
2. Data Acquisition
3. **Data Processing**
4. EDA & Visualization
5. Model Creation-Trg-Eval
6. Deployment & Monitoring

Extract: Acquire data from single or multiple sources



Transform

Data Wrangling/Munging: Transform collected data into desired format for later analysis

Data Cleansing: Handling missing data, duplicate values, null values, mis-spelled attributes, inconsistent data types and outliers

rapidminer

Informatica

NumPy

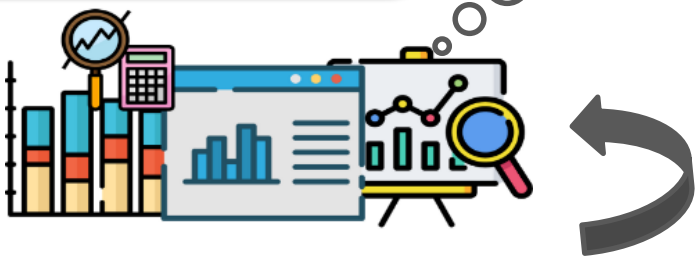
pandas

Load: The transformed data is loaded into the target data source or data warehouse

Exploratory Data Analysis & Visualization

1. Business Problem
2. Data Acquisition
3. Data Processing
4. **EDA & Visualization**
5. Model Creation-Trg-Eval
6. Deployment & Monitoring

EDA involves understanding your data and identifying patterns. It involves identifying relationships and correlations between variables using visual as well as statistical techniques



These patterns are not evident when you are looking at data in tables. A correct visualization tool can help you quickly gain a deeper understanding of your data

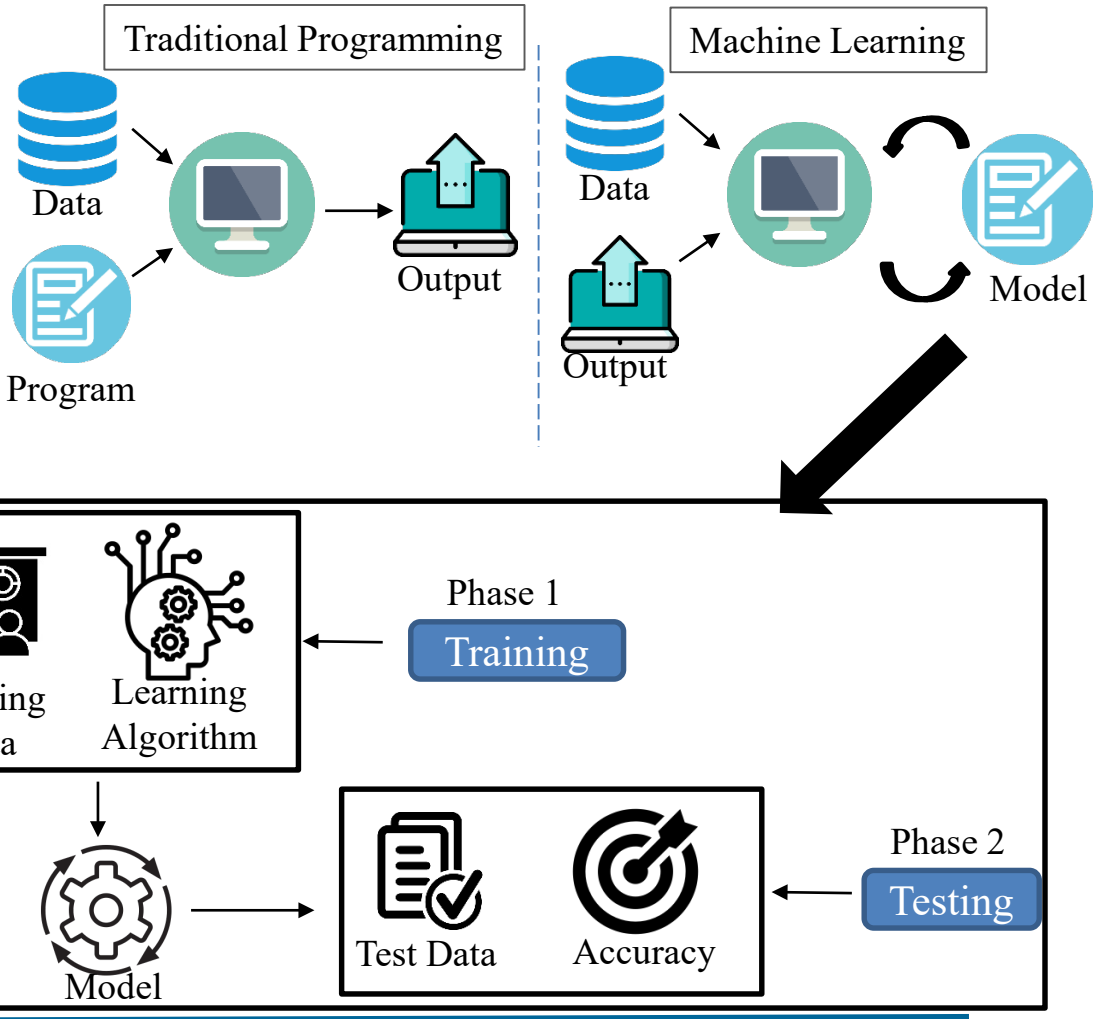
Finally EDA involves Feature Engineering, which performs feature creation, transformation, extraction and selection before creation of ML model

Data Analyst's Job Ends Here

ML Model: Creation-Training-Evaluation

1. Business Problem
2. Data Acquisition
3. Data Processing
4. EDA & Visualization
5. **ML Model Creation-Trg-Eval**
6. Deployment & Monitoring

ML is an application of AI that gives computers the ability to learn without being explicitly programmed. [Arthur Samuel]



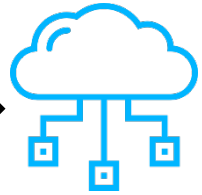
Use different but appropriate machine learning algorithms like Decision Tree, Linear Regression, K-Nearest Neighbour to the data to identify the model that best fits the business requirements

Model Deployment and Monitoring

1. Business Problem
2. Data Acquisition
3. Data Processing
4. EDA & Visualization
5. Model Creation-Trg-Eval
6. **Deployment & Monitoring**

After a model is trained, tuned and tested, you can deploy the model into production and make inferences (predictions)

Check the deployment environment for dependency issues
Deploy the model first in the test and then in the production environment



Cloud Deployment

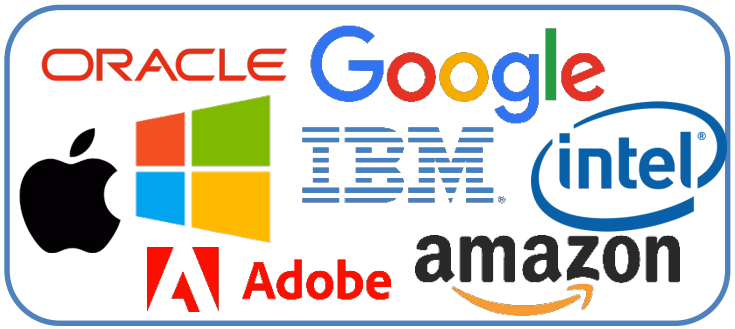
Most of the times the live real world data differ from the data that was used to train the model, thus making the model less accurate. To handle this, build a model monitor that detects deviations such as data drift and alerts you to take remedial actions

Industry Job Roles in Data Science

Industry Job Roles: Data Scientist

1. **Data Scientist**
2. Data Engineer
3. Data Analyst
4. Database Administrator
5. ML Engineer

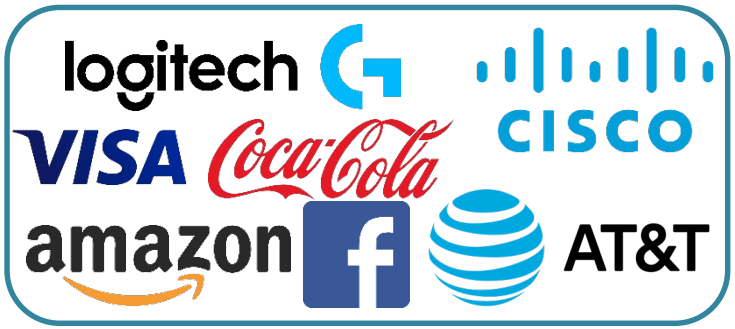
- Senior most in the team and take inputs from the rest to formulate actionable insight for the business
- Makes use of the latest tools and technologies in finding solutions and reaching conclusions that are crucial for an organization's growth and development



Industry Job Roles: Data Engineer/Architect

1. Data Scientist
2. **Data Engineer**
3. Data Analyst
4. Database Administrator
5. ML Engineer

- Scrape data and store it in warehouses using ETL
- Handle databases and create data warehouses
- Design, build, and manage the big data infrastructure
- Build data pipelines for easy access of data
- Big Data Tools (Apache Spark, Apache Hive, Hadoop)
- Cloud Platforms (AWS, Google Cloud Platform)



Industry Job Roles: Data Analyst

1. Data Scientist
2. Data Engineer
3. **Data Analyst**
4. Database Administrator
5. ML Engineer

- Data Analyst is an entry level member into the data analytics team
- Needs to have good technical skills and know the basics of statistics, data munging, data utilization, and exploratory data analysis
- Generate reports after analyzing the data
- Can move to the role of Data engineer and Data scientist with more experience

Industry Job Roles: Database Administrator

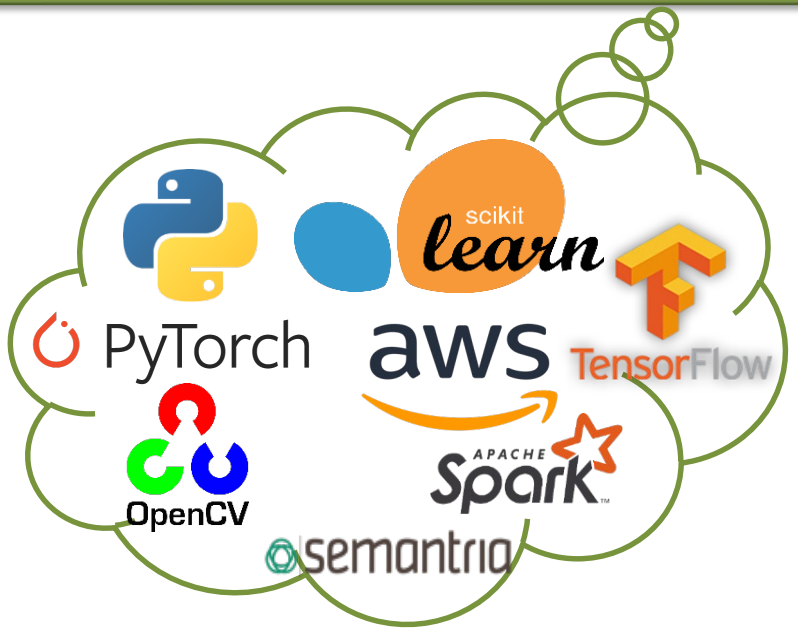
1. Data Scientist
2. Data Engineer
3. Data Analyst
4. **Database Administrator**
5. ML Engineer

- Responsible for administering the collected data by installing, configuring, monitoring, operating, and maintaining database
- Ensure that all databases are available to all relevant users, and is protected securely from any malicious activity

Industry Job Roles: Machine Learning Engineer

1. Data Scientist
2. Data Engineer
3. Data Analyst
4. Database Administrator
5. **ML Engineer**

- Machine learning engineer works as a part of large data science team
- Responsible to design and create all algorithms capable of learning and making predictions
- They are expected to perform A/B testing, build data pipelines, and implement algorithms for classification, clustering, regression, anomaly detection etc.



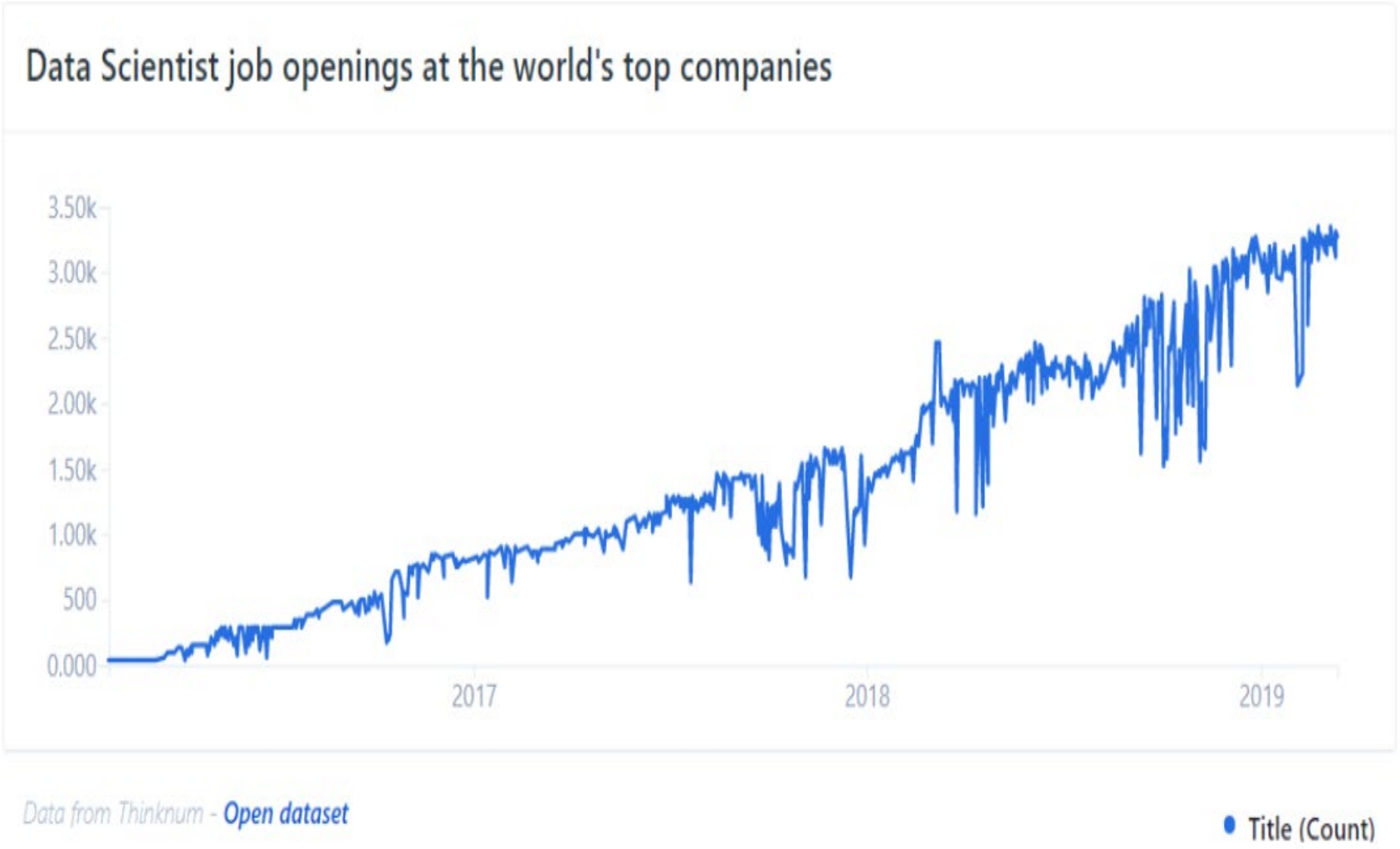
History: Data Science Salary Trends

National average salary for different job roles in data science

Source - Glassdoor

Source: https://towardsdatascience.com/why-learn-data-science-in-2020-d3f54123b2e4

History: Job trends



Source: https://www.tecla.io/blog/the-high-demand-for-data-scientists-and-how-to-hire-for-them/

Things To Do

- Visit all the hyperlinked tools and technologies in todays lecture slides. You should be able to give a single line description of each.
- Have a very clear understanding of Data Science Life Cycle, the tools & the technologies used in each phase.
- Think of few use cases where you can apply Data Science, Machine Learning and Deep Learning technologies and make a list of the skill set you need to develop/learn to implement and deploy such projects.

Coming to office hours does NOT mean you are academically weak!