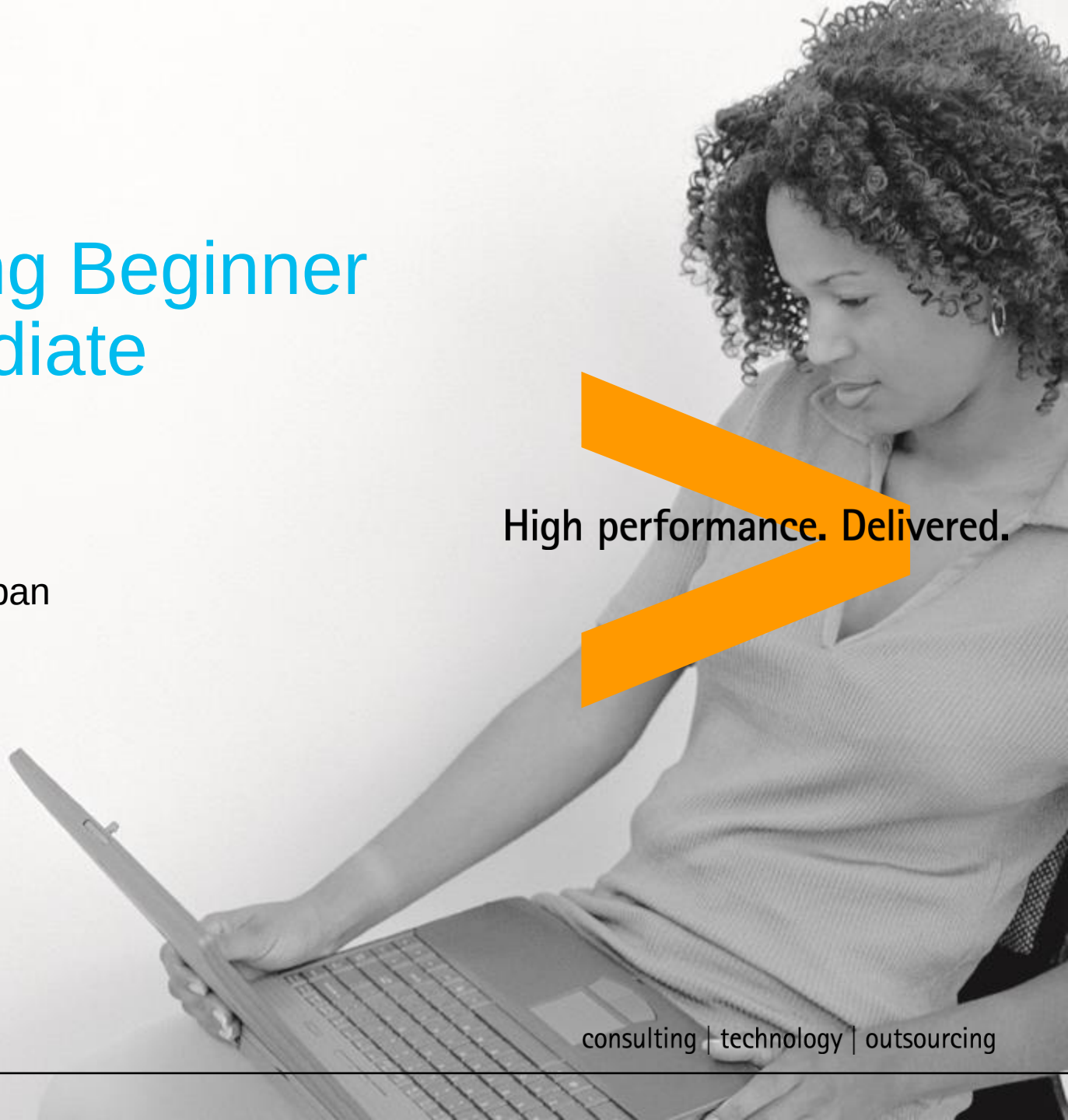
Data Mining Beginner to Intermediate



High performance. Delivered.

Parameswari Ettiappan

consulting | technology | outsourcing

Goals

DATA SCIENCE

- What is Machine Learning
- Use Cases
- Commonly used Terms
- Lifecycle of a ML Project
- Supervised Learning
- Unsupervised Learning

Goals

DATA SCIENCE

- Data Acquisition
- Types of Data
- Data Types
- Exploratory Data Analysis
- Data Pre-processing
- Data Quality assessment
- Feature Scaling

Goals

DATA SCIENCE

- Descriptive Statistics
- Methods to impute missing values
- Outlier/Anomaly Detection
- Data Visualization
- Histogram
- Bar Graph
- Scatter Plot
- Pie Chart

Goals

DATA SCIENCE

- Box Plot
- Feature Selection
- Univariate Selection
- Feature Importance
- Correlation matrix and Heat map
- Underfitting vs Overfitting
- Bias-Variance Trade-off

Goals

DATA SCIENCE

- Hypothesis Testing
- Statistical Assumptions
- Null Hypothesis
- Alternate Hypothesis
- One sample Z-test
- Z-test in Python
- T-test
- T-test in Python

Goals

DATA SCIENCE

- Pearson’s Chi Squared Test
- Confusion Matrix
- Absolute Error
- Relative Error
- RMSE
- Precision, Accuracy
- Recall
- Specificity
- F-Score
- ROC/AUC

Goals

DATA SCIENCE

- Cost Function
- Gradient Descent
- What is Regression
- Basic Idea
- Linear Regression Applications
- Linear Regression

Goals

DATA SCIENCE

- Types of Errors
- Better Regression Models
- Correlation is not Causation
- Polynomial Linear Regression
- Regularization
- Ridge Regression
- LASSO Regression

Goals

DATA SCIENCE

- Types of Classification Algorithms
- Applications of Classification Algorithms
- Logistic Function
- Logistic Regression
- Application of Logistic Regression
- Types of Logistic Regression
- Decision Trees
- Working of Decision Tree

Goals

DATA SCIENCE

- Attribute Selection measure
- Gini Index
- Information Gain
- Random Forests
- Working of RF
- Advantages and Disadvantages of RF algorithm
- Application of RF
- XGBoost
-

Goals

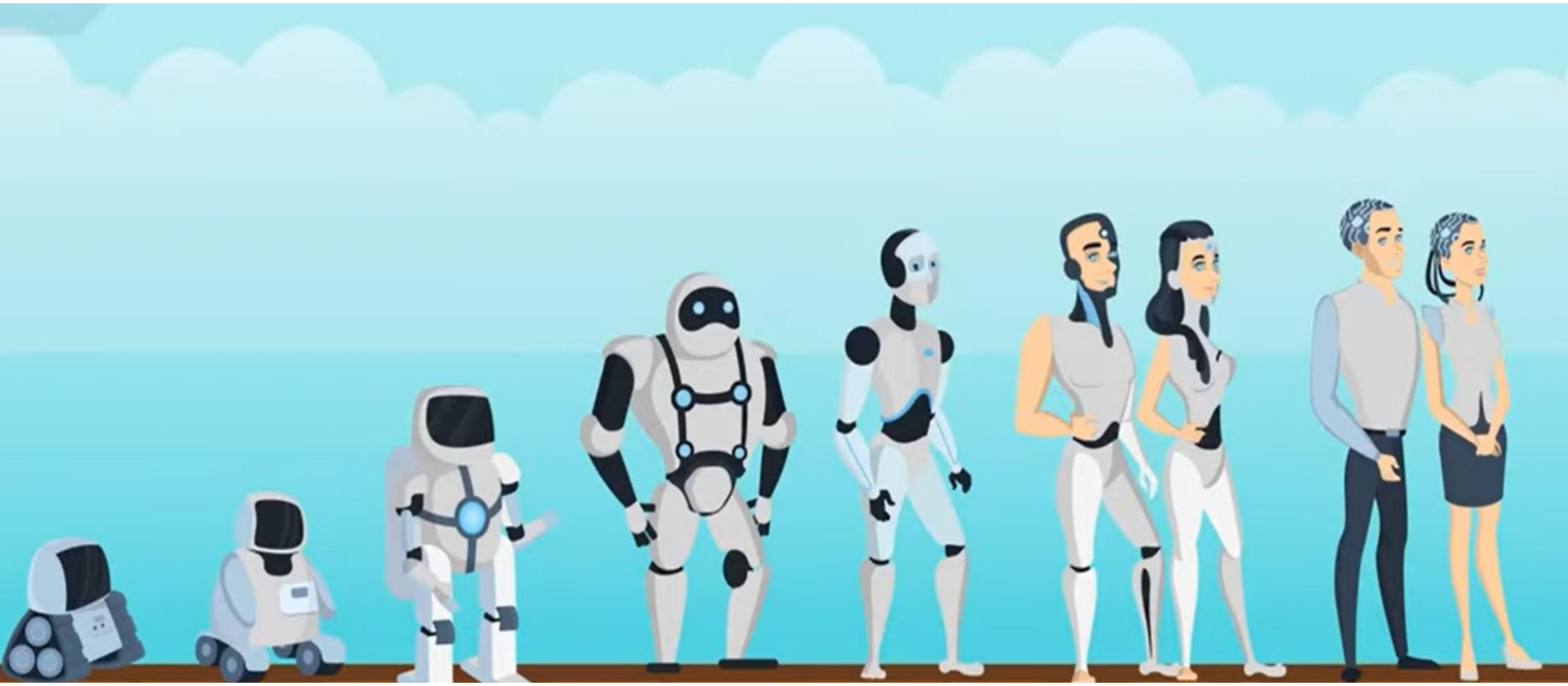
DATA SCIENCE

- Why Unsupervised Learning
- Applications
- Clustering
- Types of Clustering
- Singular value Decomposition
- Independent Component Analysis
- Association Rules

Machine Learning

DATA SCIENCE

Evolution of Machines



Machine Learning

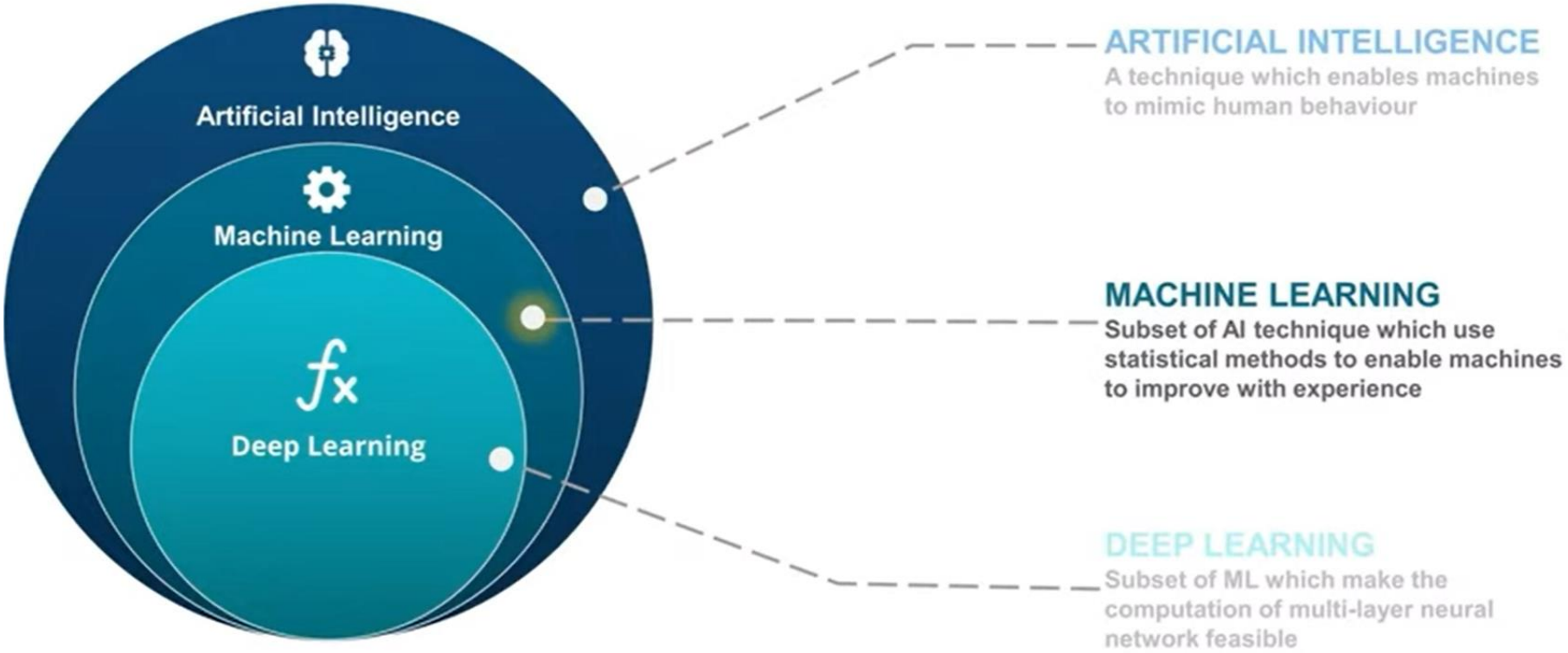
DATA SCIENCE

What is Machine Learning?

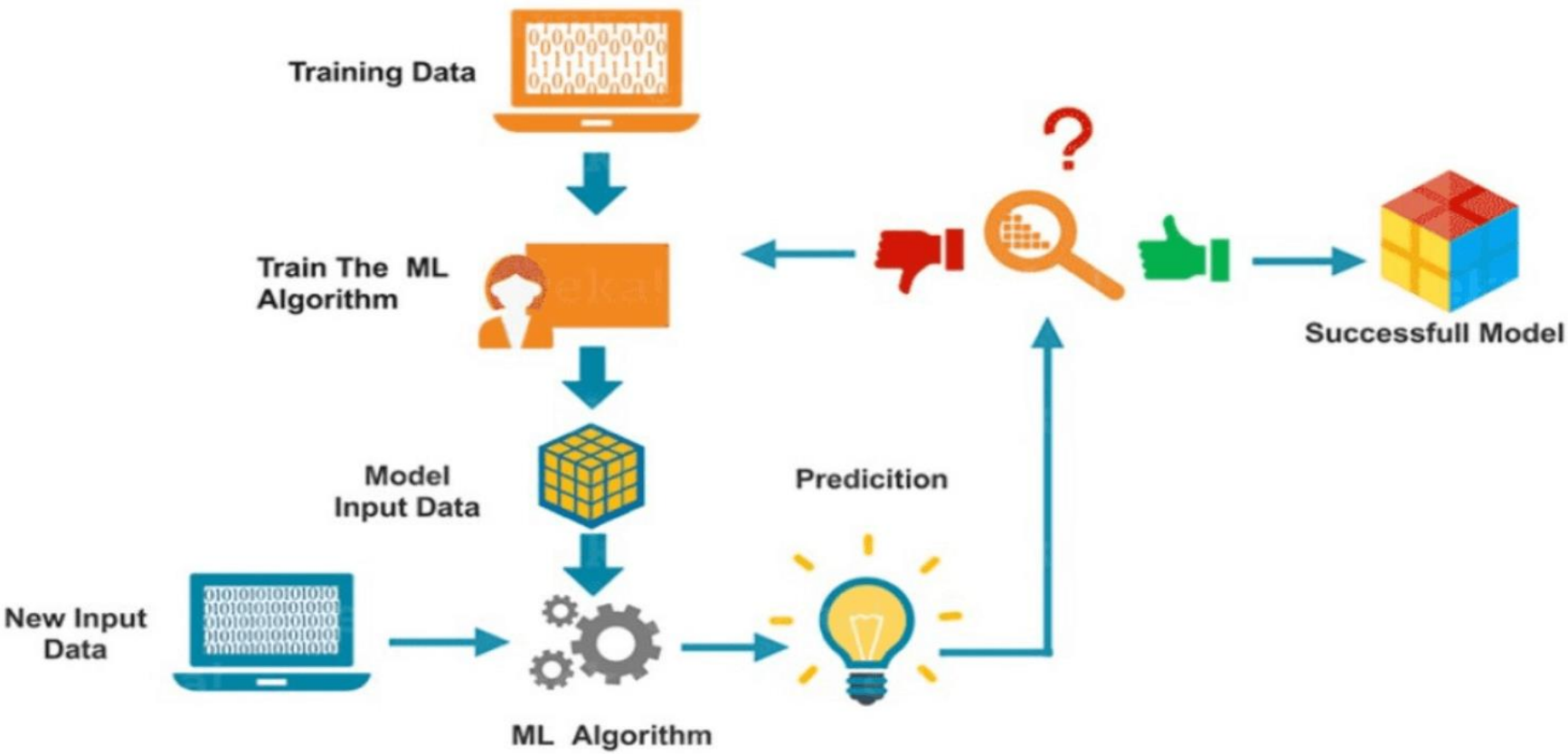
Machine Learning is a subset of artificial intelligence. It focuses mainly on the designing of systems, thereby allowing them to learn and make predictions based on some experience which is data in case of machines.

Big Confusion AI, ML and Deep Learning

DATA SCIENCE



Machine Learning

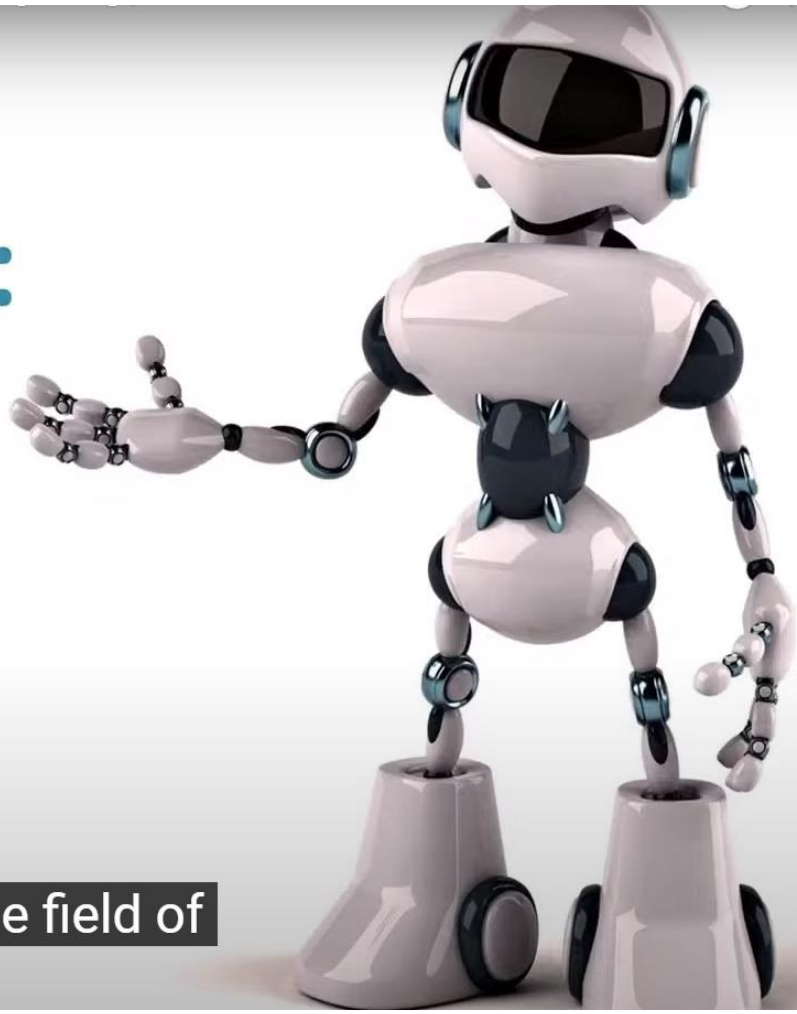


Types of Machine Learning

DATA SCIENCE

Machine Learning Types:

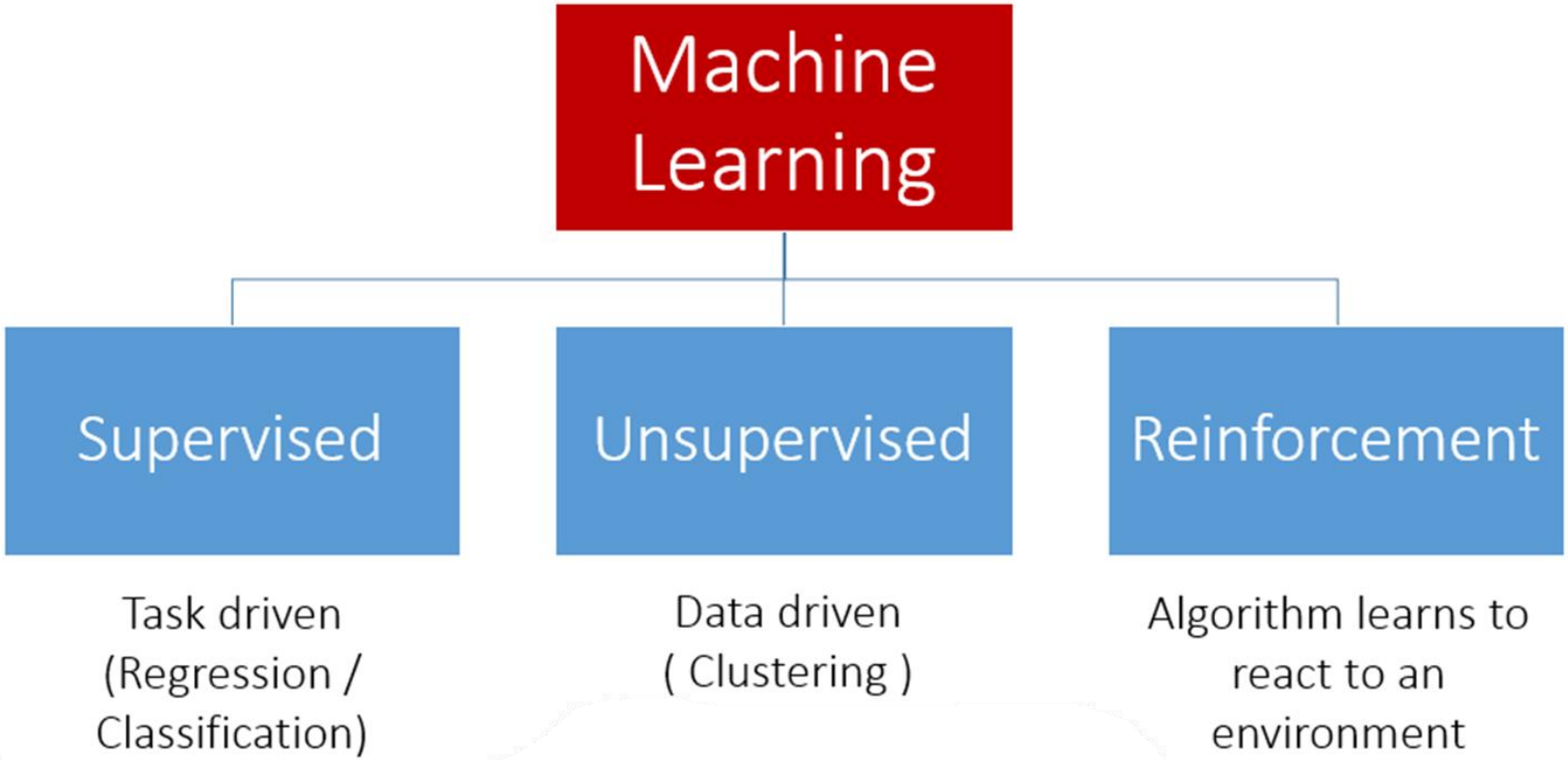
- Supervised
- Unsupervised
- Reinforcement



how each of them is used in the field of banking

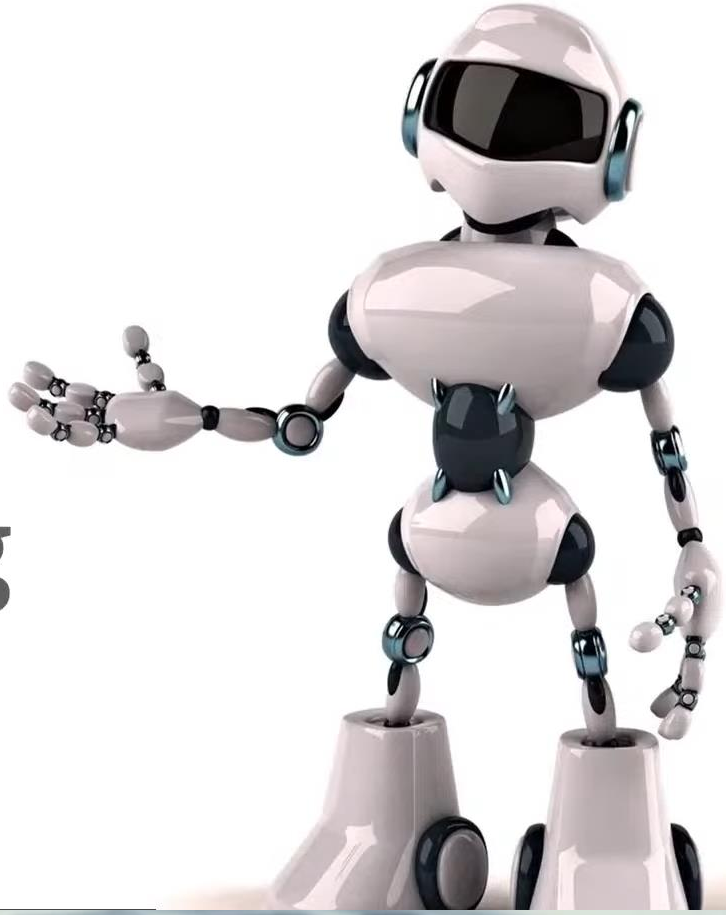
Types of Machine Learning

DATA SCIENCE



DATA SCIENCE

Supervised
Machine Learning



Supervised Learning

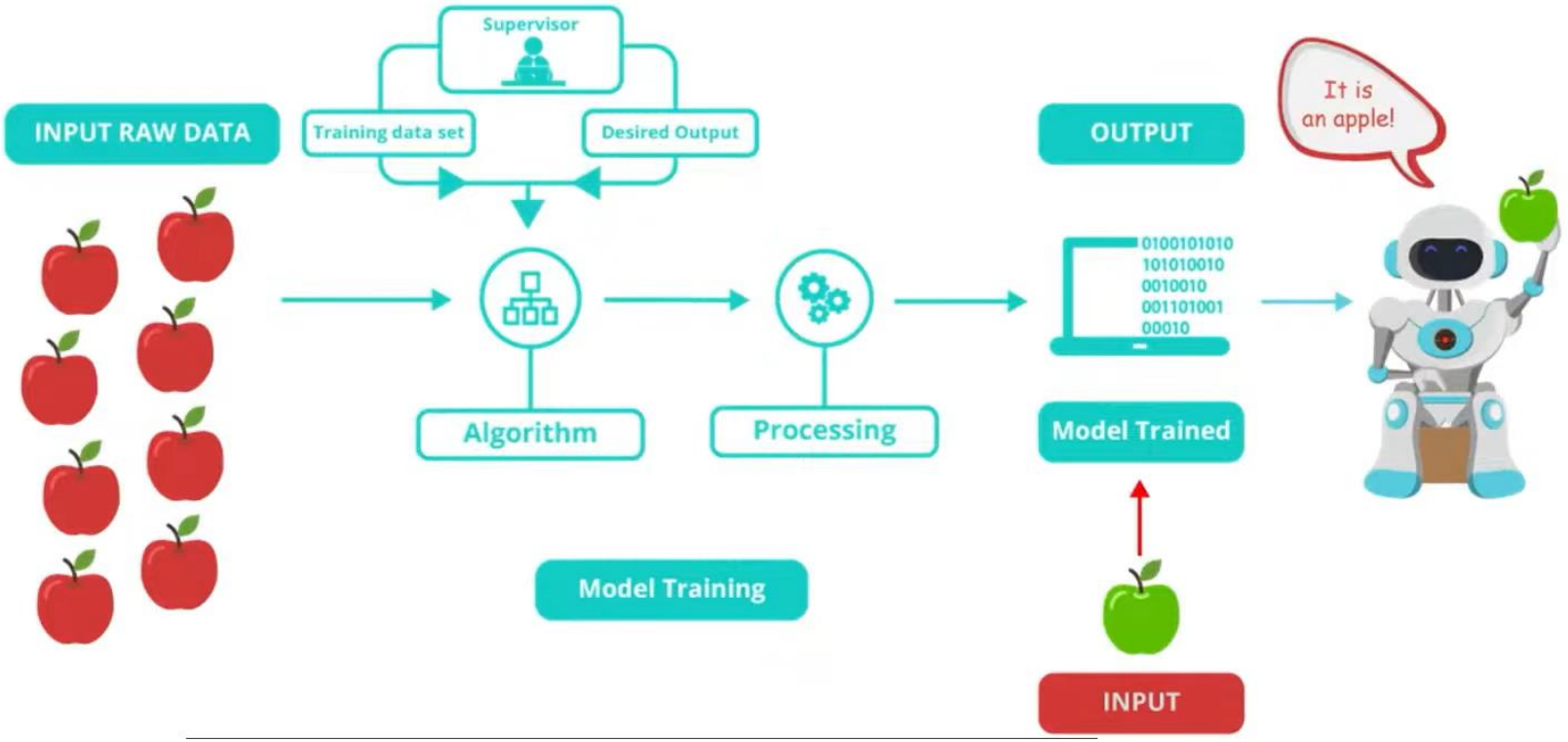
DATA SCIENCE

- Supervised learning categories and techniques
  - Linear classifier (numerical functions)
  - Parametric (Probabilistic functions)
  - Naïve Bayes, Gaussian discriminant analysis (GDA), Hidden Markov models (HMM), Probabilistic graphical models
  - Non-parametric (Instance-based functions)
  - K-nearest neighbors, Kernel regression, Kernel density estimation, Local regression
  - Non-metric (Symbolic functions)
  - Classification and regression tree (CART), decision tree

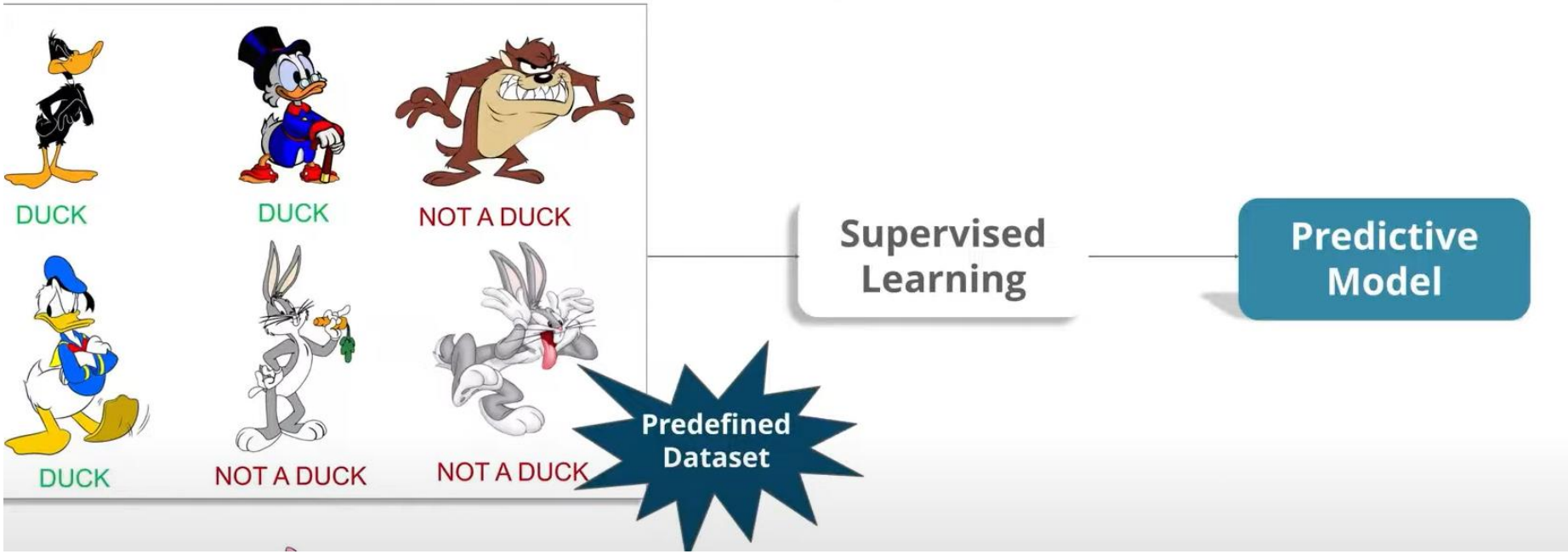
  - Aggregation
  - Bagging (bootstrap + aggregation), Adaboost, Random forest

Supervised Learning

DATA SCIENCE



Supervised Learning



Supervised Learning

DATA SCIENCE

Linear Regression

Random Forest

Support Vector

Supervised Learning Use Cases

Supervised Learning Use Cases

DATA SCIENCE

Supervised Learning Use Cases

Supervised Learning Use Cases

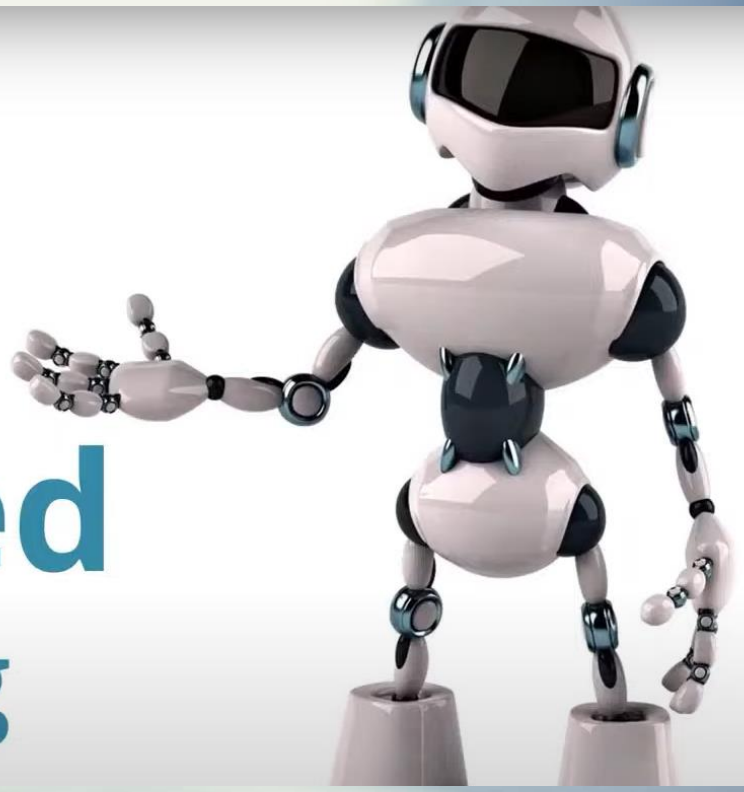
DATA SCIENCE

Supervised Learning Use Cases

Supervised Learning Use Cases

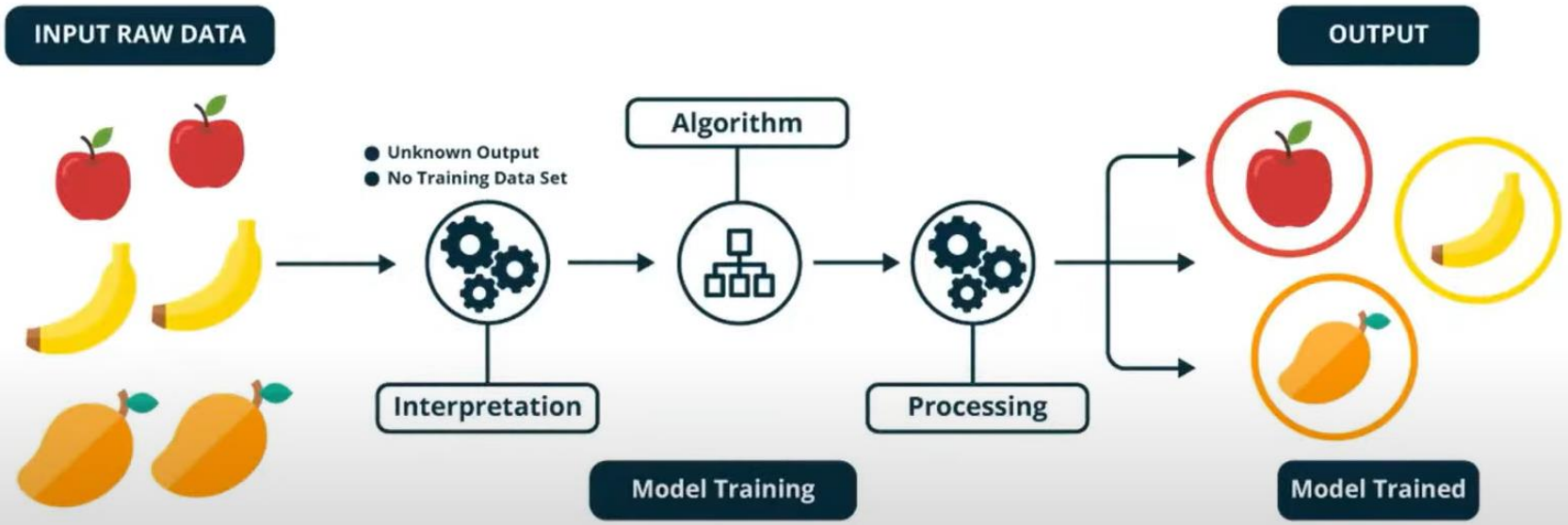
DATA SCIENCE

Unsupervised
Machine Learning

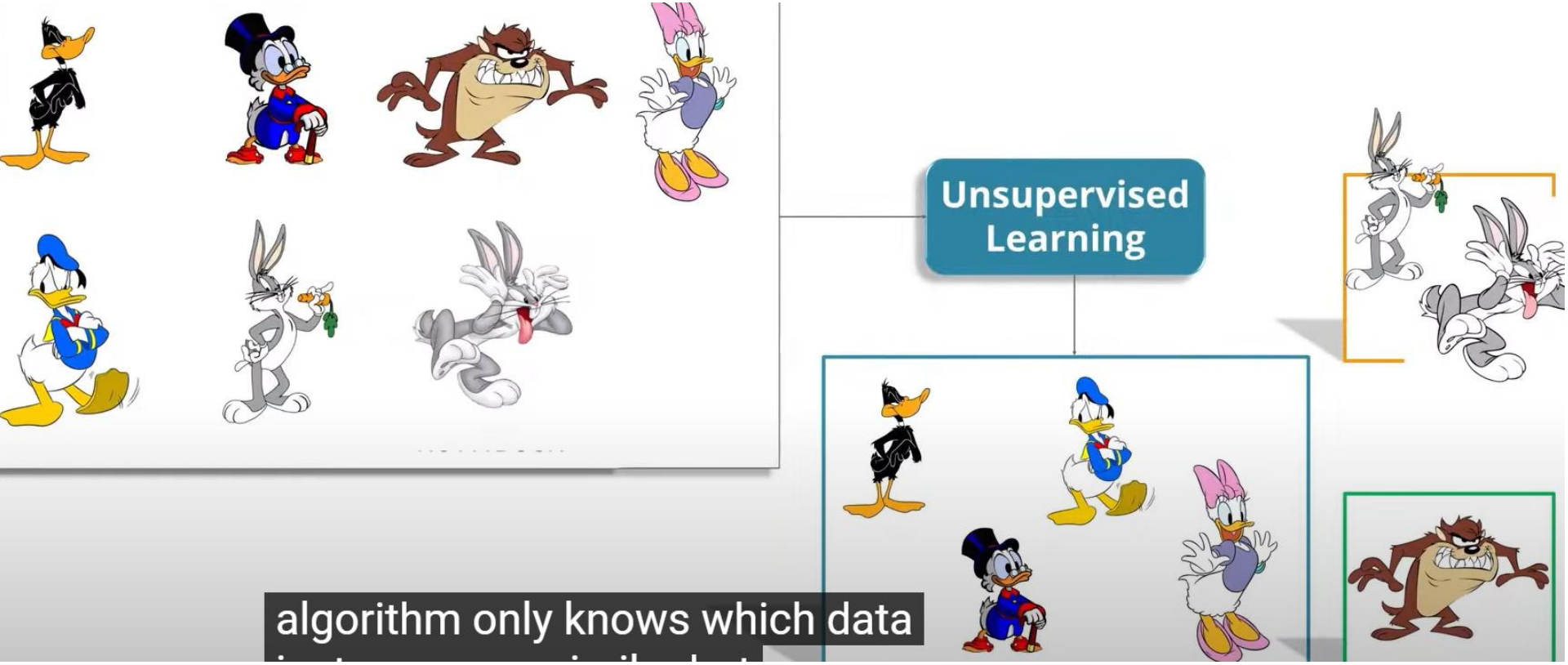


Unsupervised Learning

DATA SCIENCE



Unsupervised Learning



Types of Unsupervised Learning

DATA SCIENCE

Apriori Algorithm

K- Means Algorithm

Hierarchical Clustering

Types of Unsupervised Learning

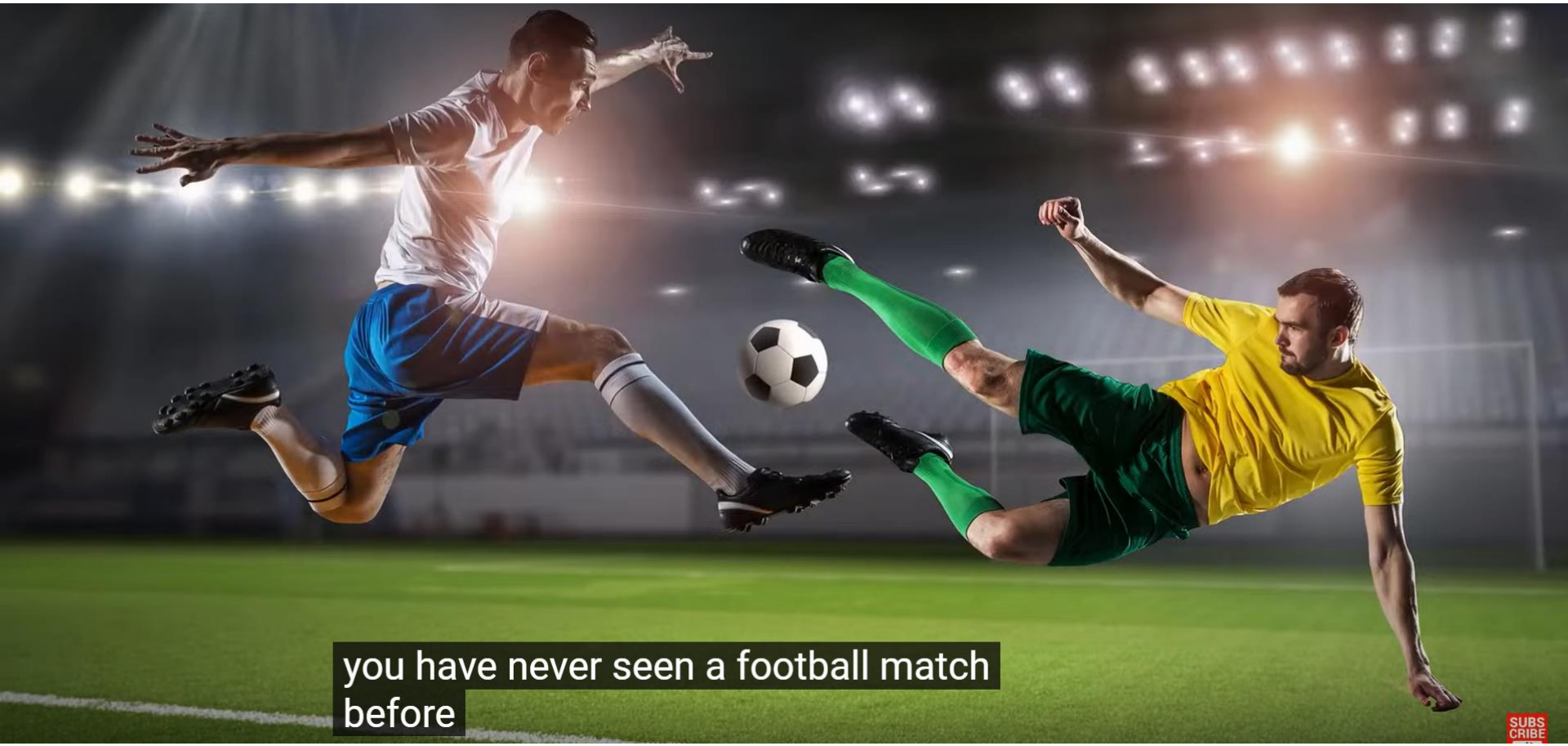
DATA SCIENCE

- Unsupervised learning categories and techniques
  - Clustering
  - K-means clustering
  - Spectral clustering
  - Density Estimation
  - Gaussian mixture model (GMM)
  - Graphical models
  - Dimensionality reduction
  - Principal component analysis (PCA)
  - Factor analysis

Use cases - Unsupervised Learning

Friend Invites You to a Party Full of Strangers and You are going to group them by age, qualifications and behaviour

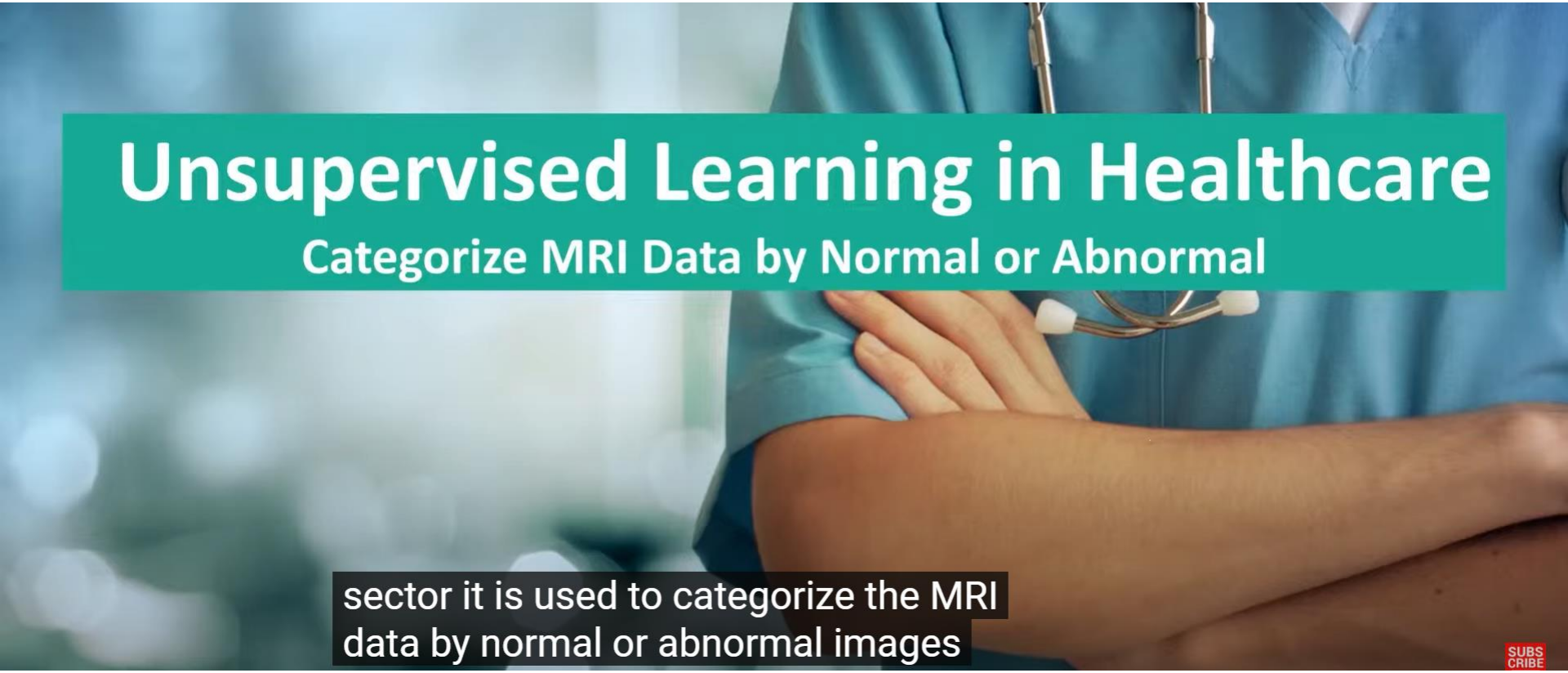
Use cases - Unsupervised Learning



You start classifying them by the way you observe either by jersey, playing style or by defender etc.,

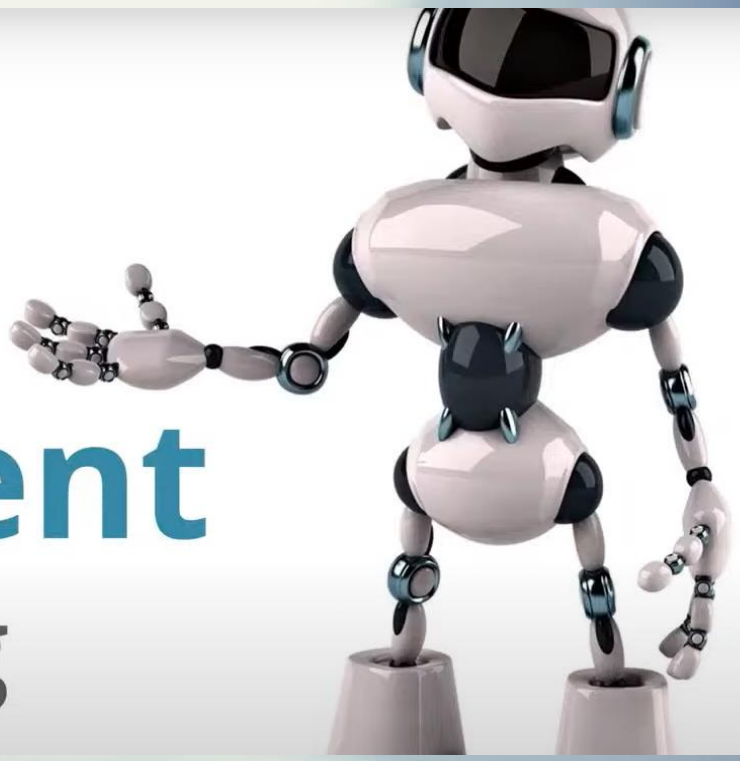
Use cases - Unsupervised Learning

Use cases - Unsupervised Learning



Use cases - Unsupervised Learning

DATA SCIENCE

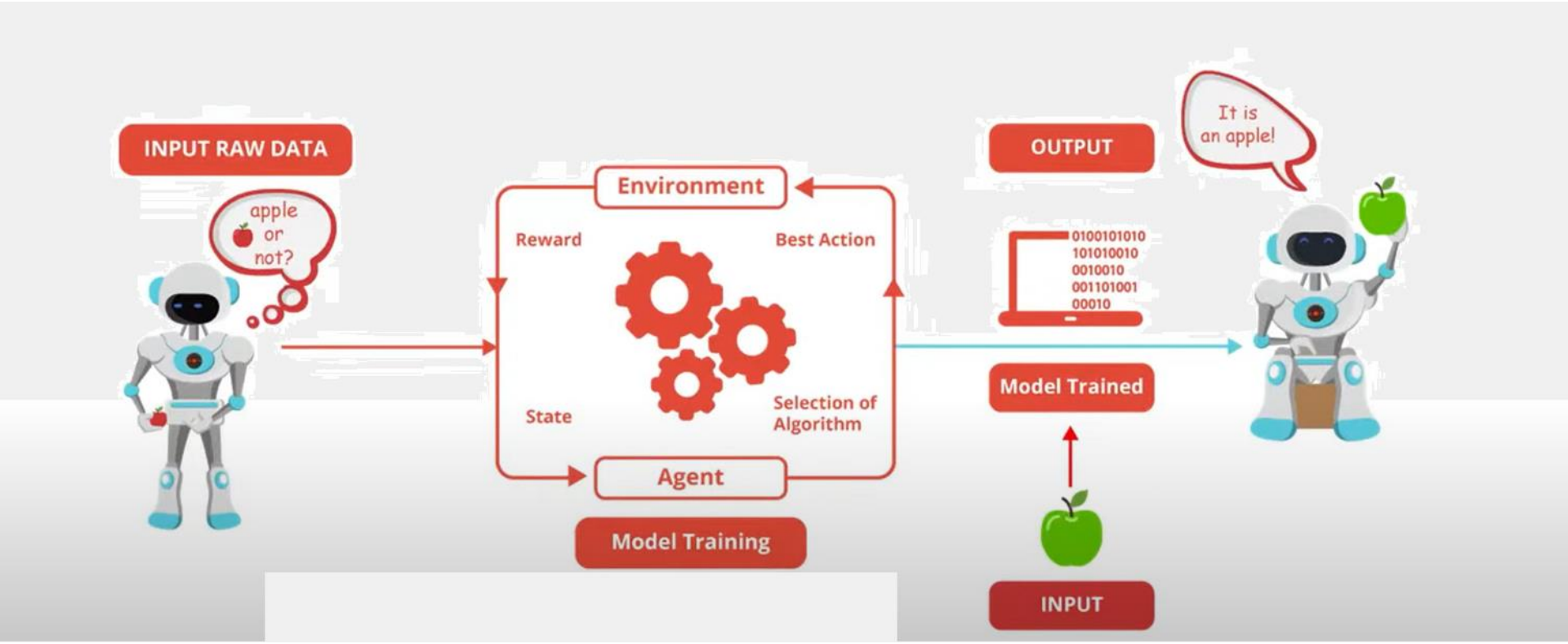


Reinforcement

Machine Learning

DATA SCIENCE

Exploration and Exploitation



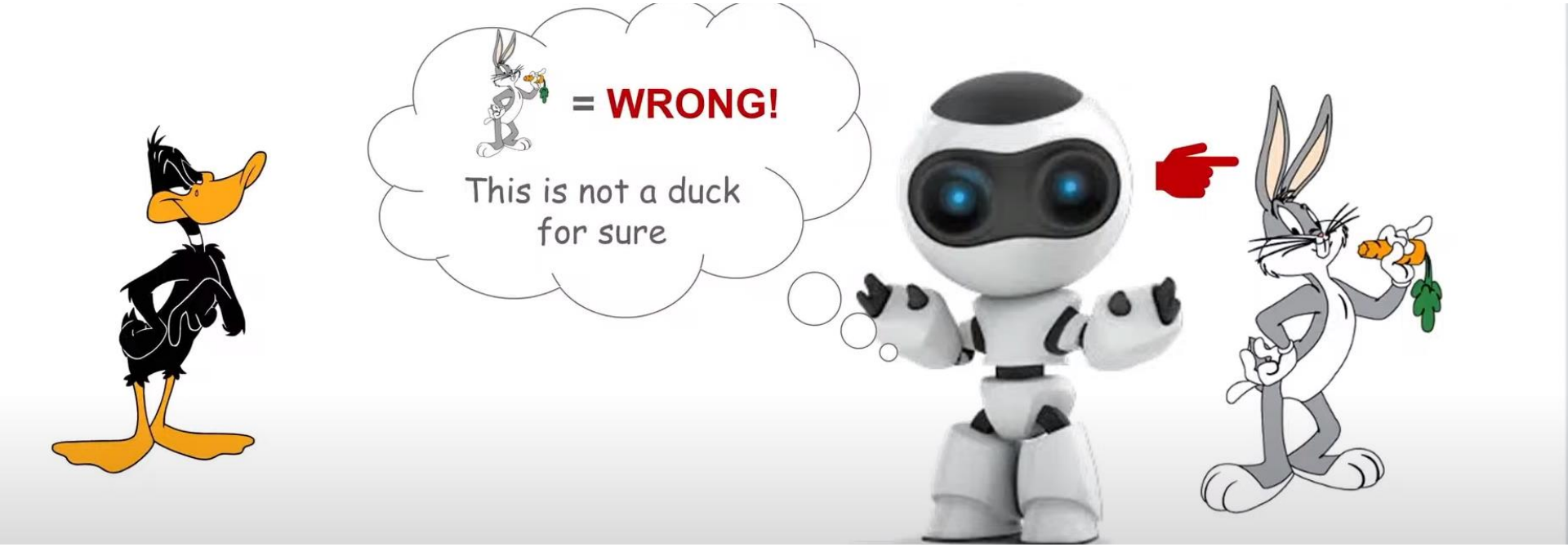
Reinforcement Learning

Reinforcement Learning

DATA SCIENCE

Machine made wrong decision by choosing bunny as duck so penalized

Reinforcement Learning

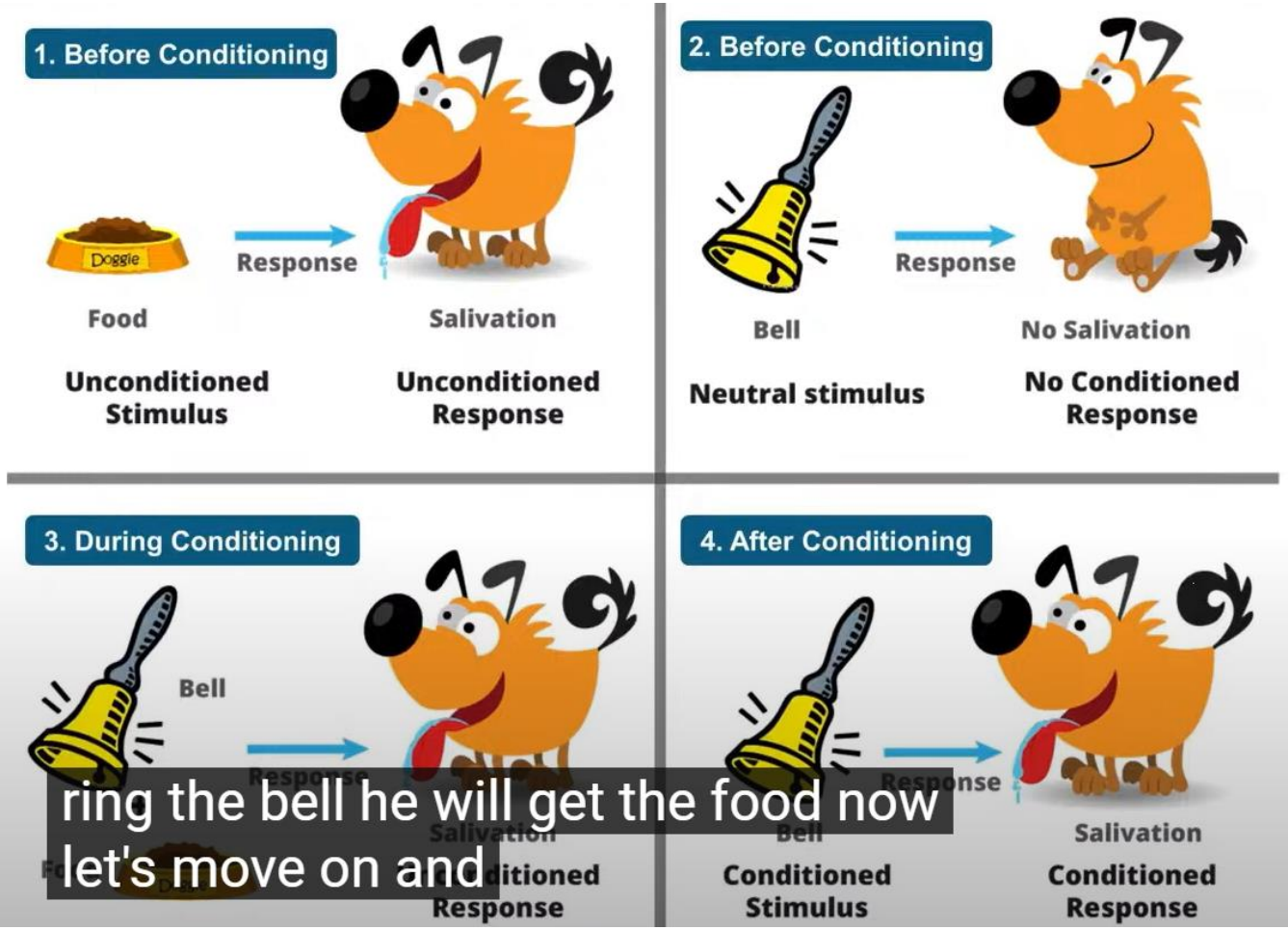


Reinforcement Learning

DATA SCIENCE

Reinforcement Examples & Use Cases

Reinforcement Learning



ring the bell he will get the food now
let's move on and

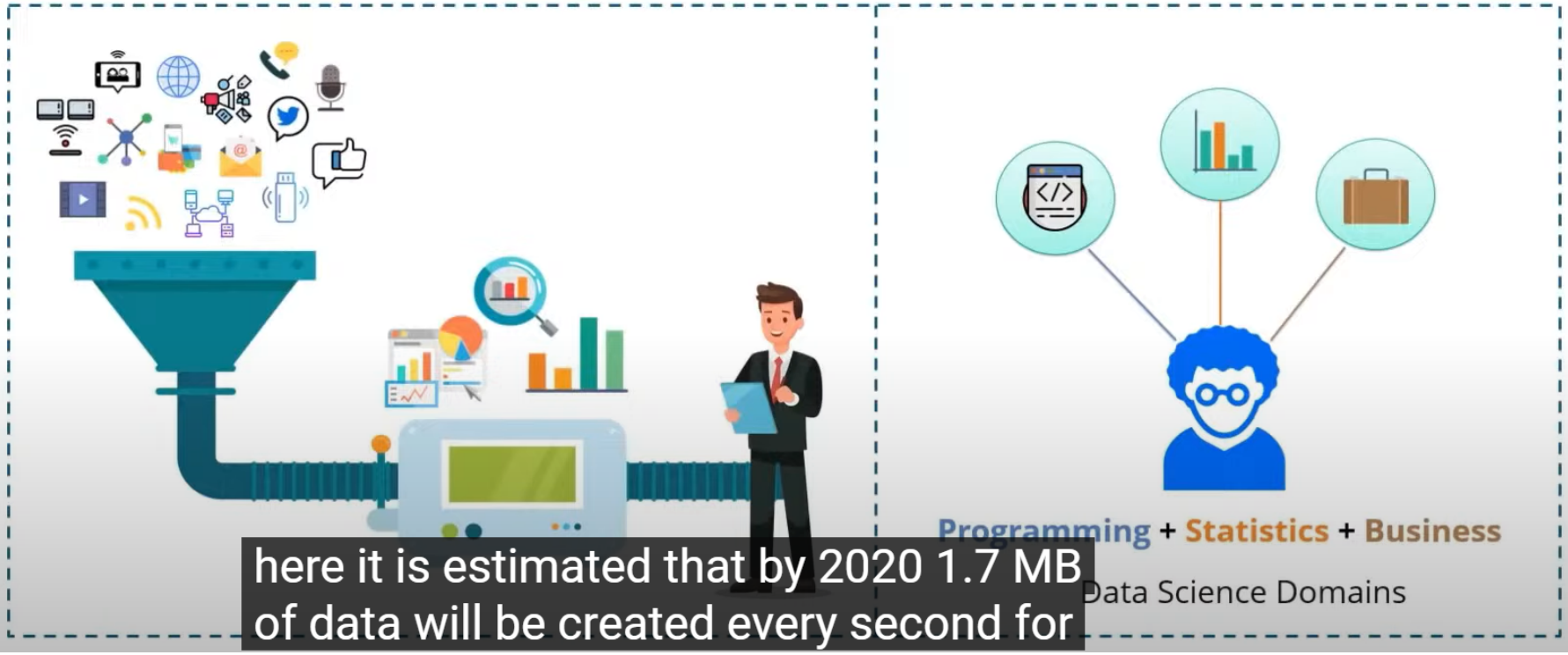
Reinforcement Learning Use cases

Reinforcement Learning Use cases

Reinforcement Learning Use cases

Data Science

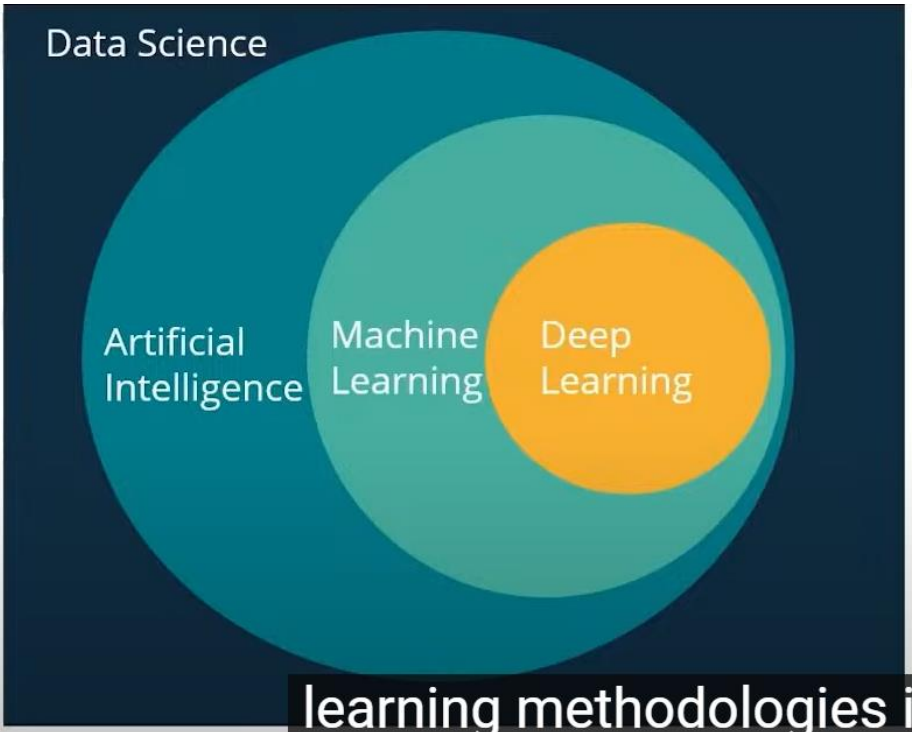
Data Science is the process of extracting useful insights from data by using a variety of tools, algorithms and Machine Learning fundamentals.



Netflix analyses the viewers interest based on that movies, web series are produced


DATA SCIENCE
Fields Of Data Science

Data Science



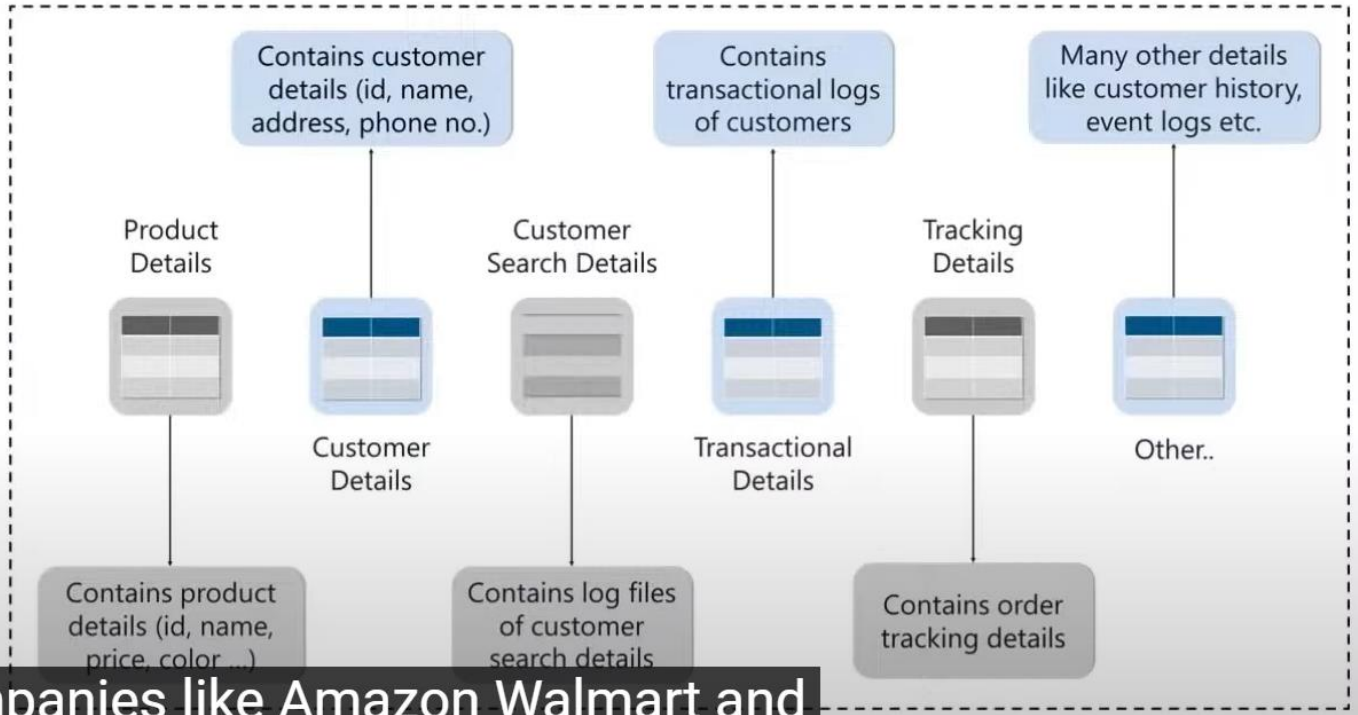
- *Data science* is the extraction of knowledge from data by using different techniques and algorithms
- *Artificial Intelligence* is a technique which enables machines to mimic human behaviour
- *Machine Learning* is a subset of AI technique which uses statistical methods to enable machines to improve with experience
- *Deep learning* is a subset of ML which make the computation of multi-layer neural network feasible

learning methodologies in order to analyze data and then extract useful

Data Science

DATA SCIENCE

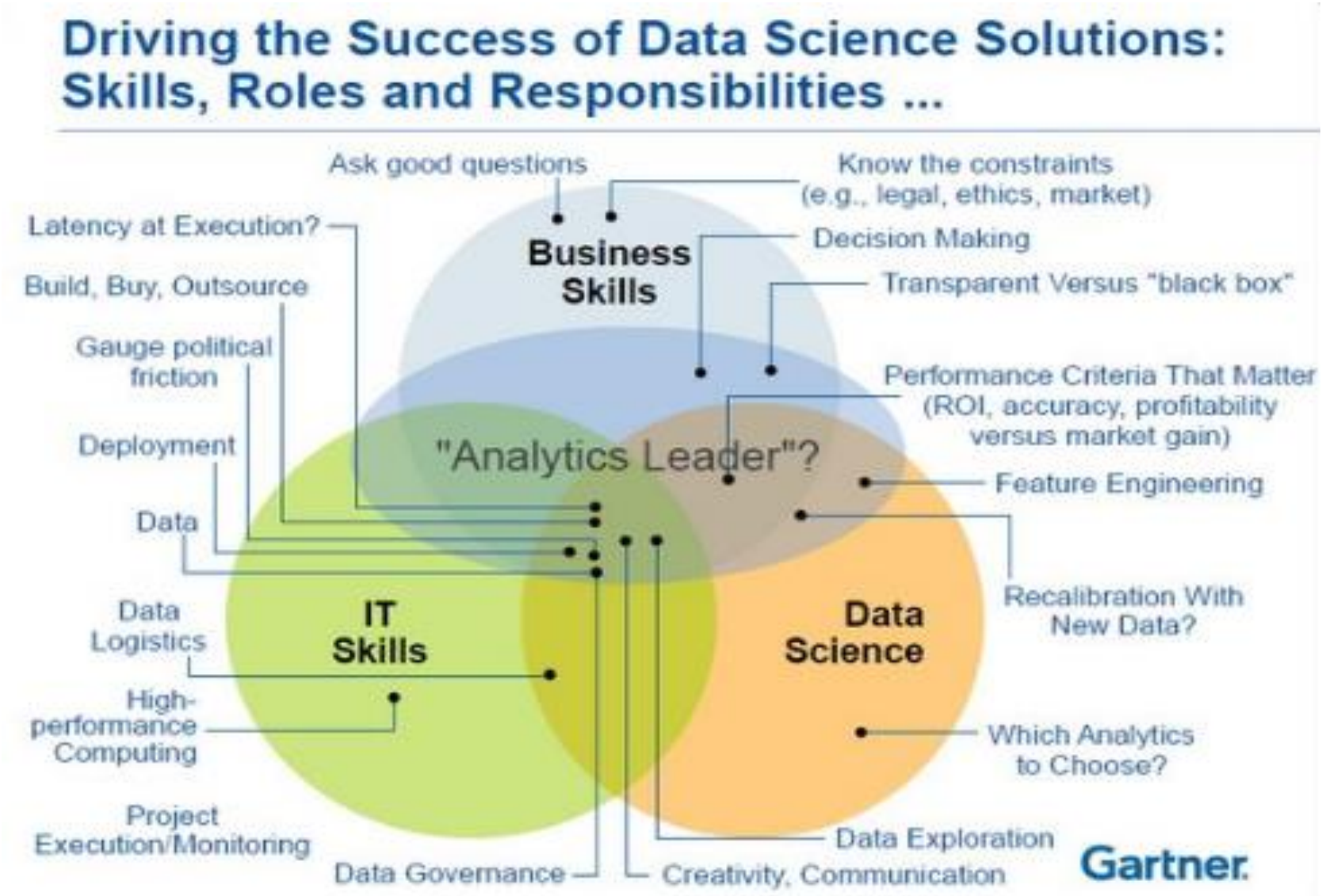
Recommendation System which suggest laptop bag when you buy laptop



why companies like Amazon Walmart and Netflix are doing so well is because of

Data Science

DATA SCIENCE

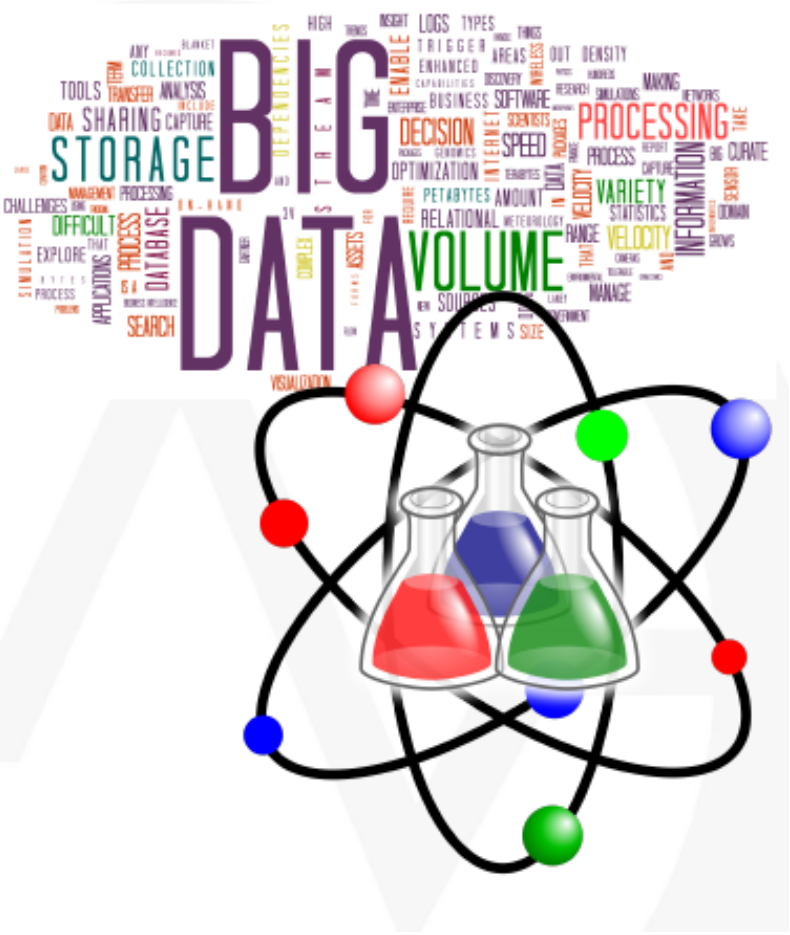


Data Science

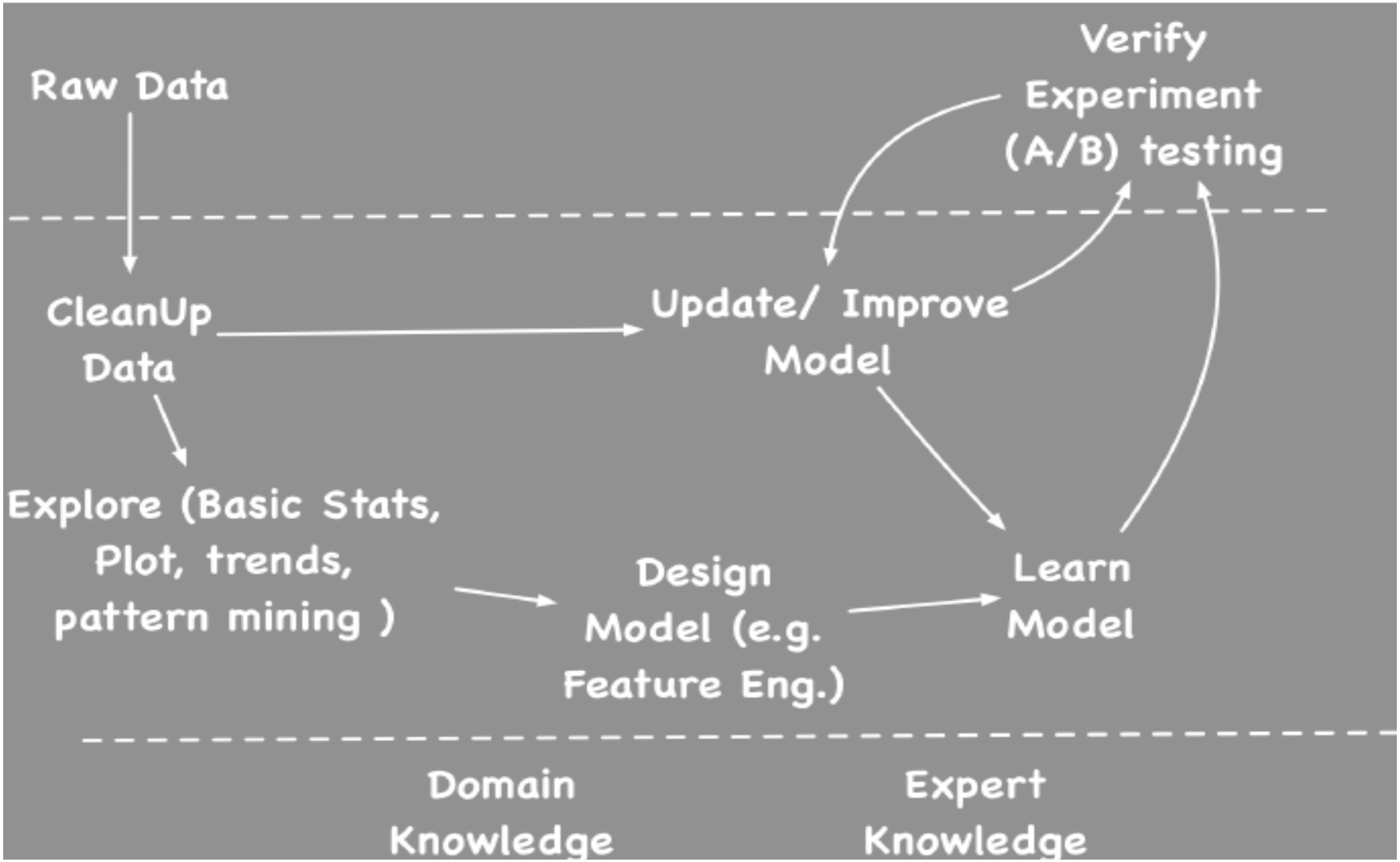
DATA SCIENCE

Extraction of knowledge from large **volumes of data** that are structured or unstructured.

It is a continuation of the fields **data mining** and **predictive analytics**



Data Science Pipeline



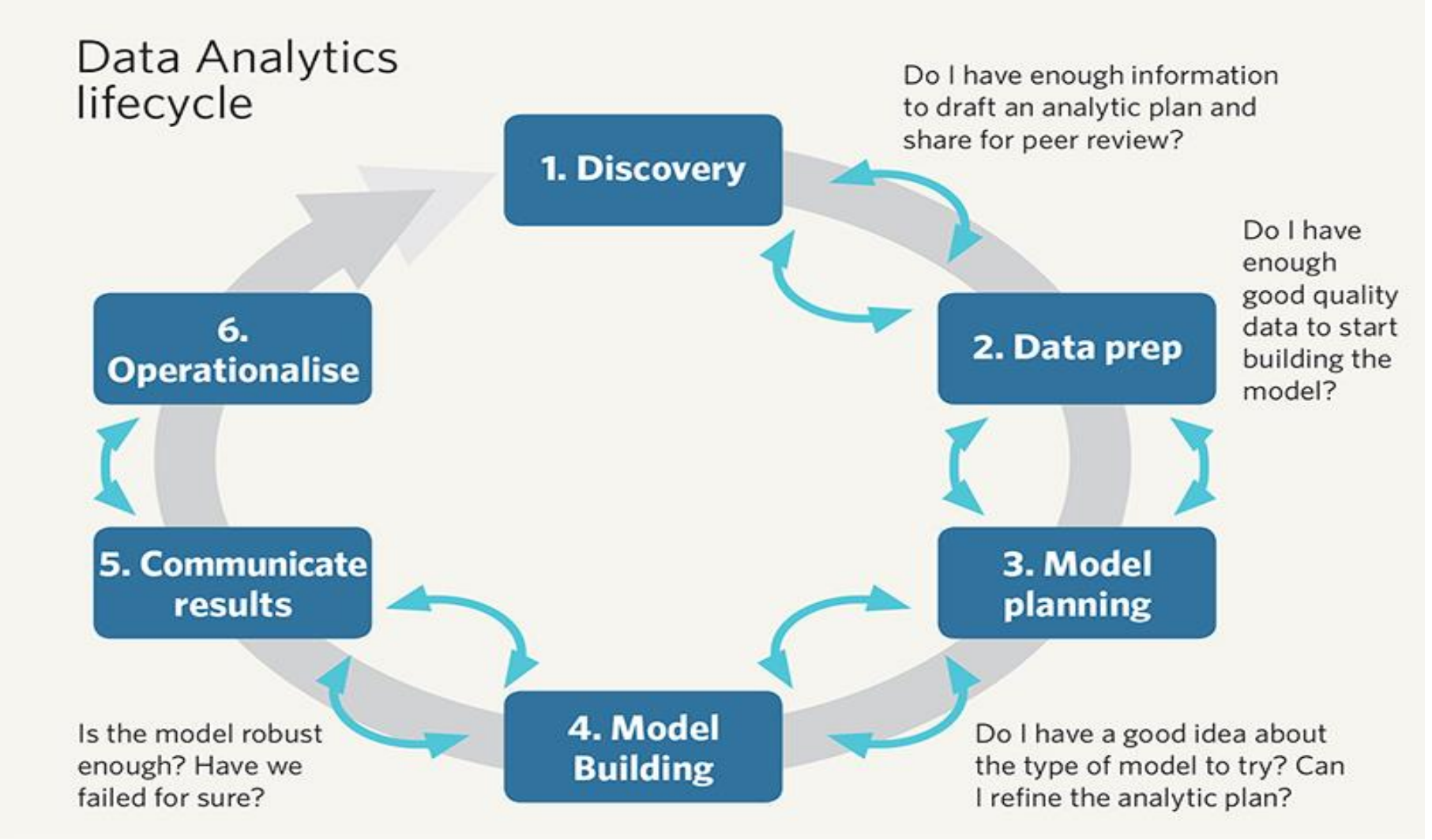
Examples of Big Data Analytics

- Using Big Data to win elections
- Big Data for finding a perfect match(Matrimony.com)
- Big Data for detecting water leakages(Bangalore Water supply and sewage system)
- Big Data for gaining insights into shopping behavior

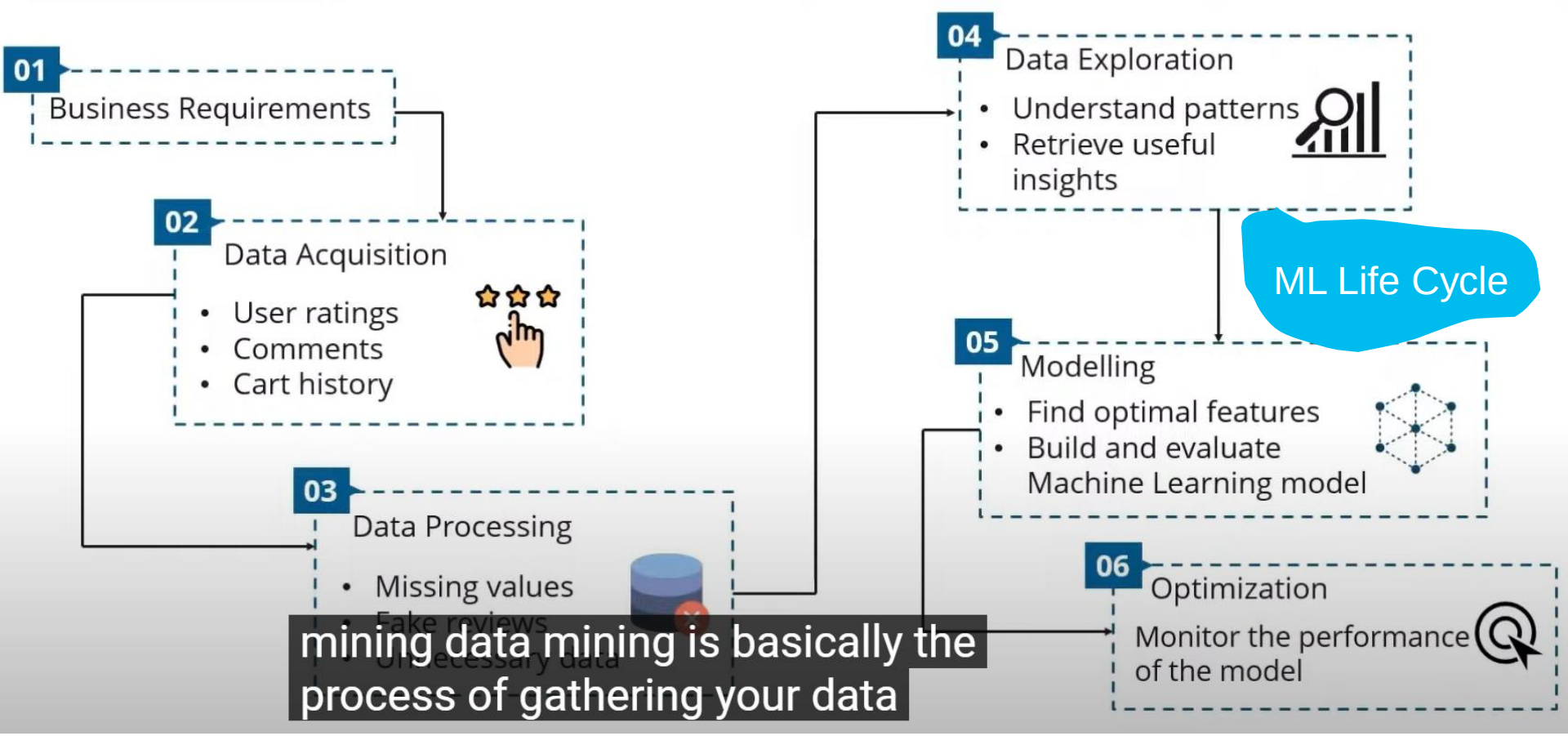
Examples of Big Data Analytics

- Big Data for ensuring proper water supply(Kerala Water Supply System)
- Using Big Data to improve India’s financial inclusion ratio**(**Micro-finance firm, Janalakshmi Financial Services)
- Using Big Data to improve product development(Reliance Games)
- Using Big Data to predict ticket confirmations for trains**(**PNR prediction)

Data Science Lifecycle

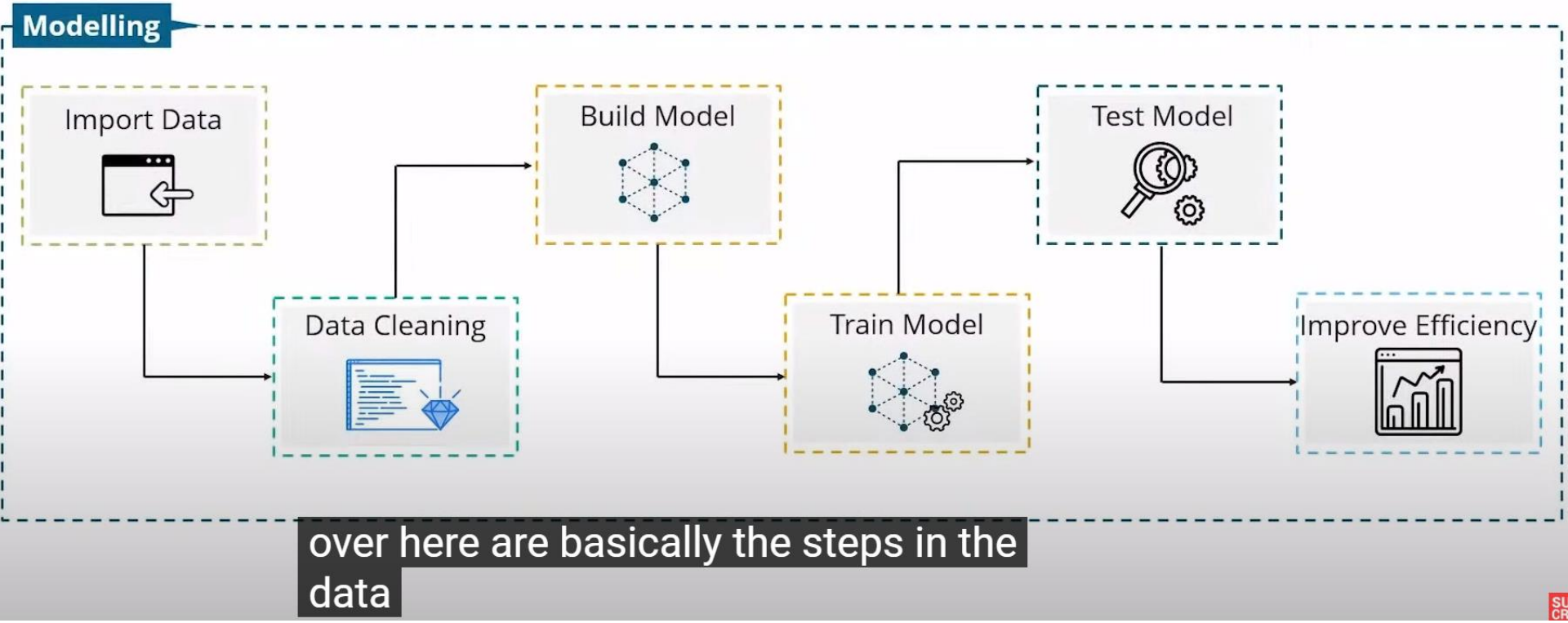


Data Science Life Cycle

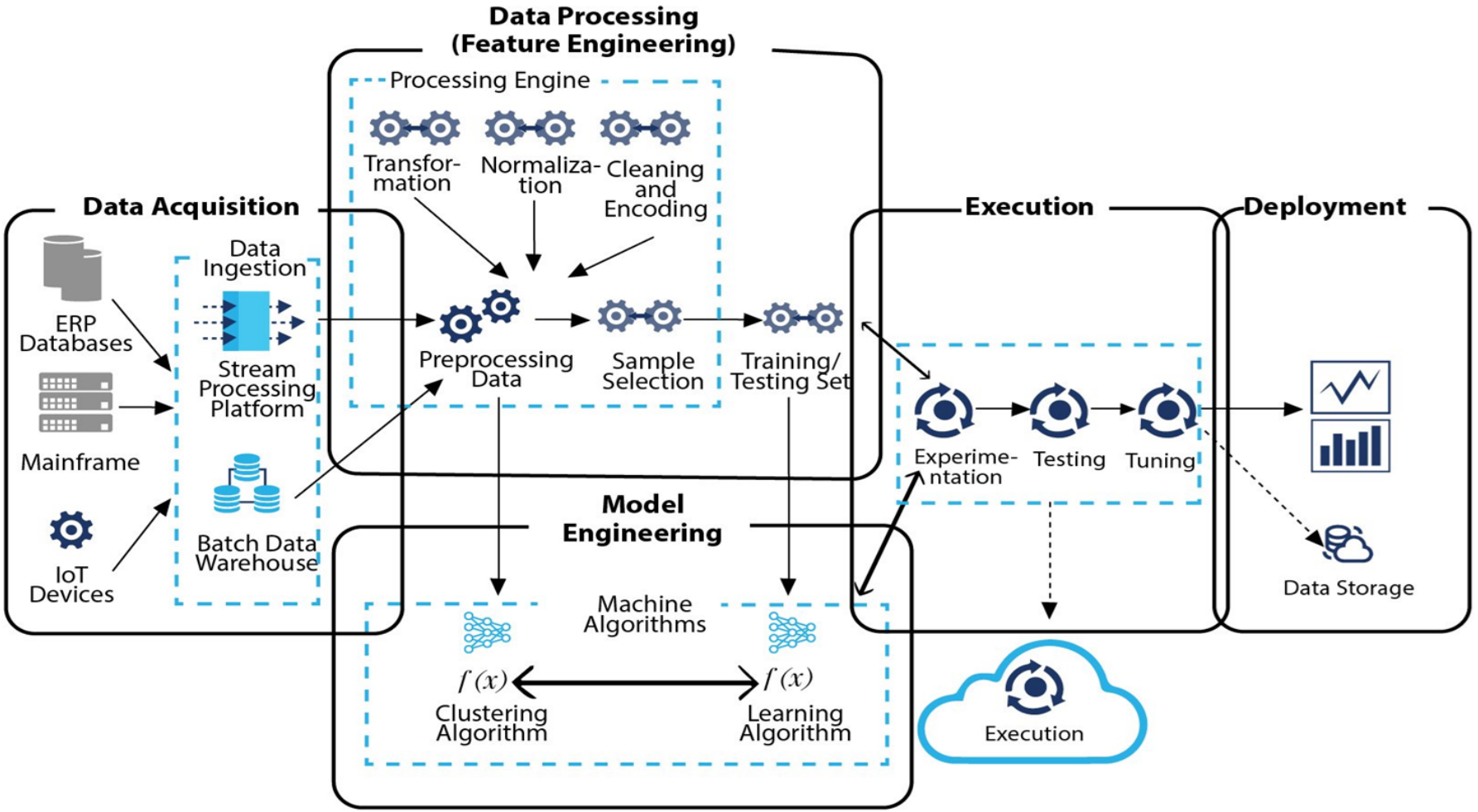


Machine Learning Lifecycle

DATA SCIENCE



Machine Learning Lifecycle



AI and ML Use cases

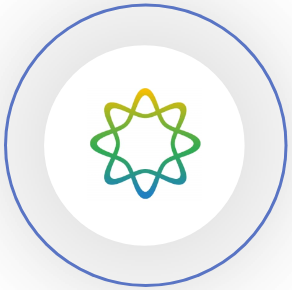
DATA SCIENCE

The Most Popular AI Solutions

Siri

Cortana

Google Assistant



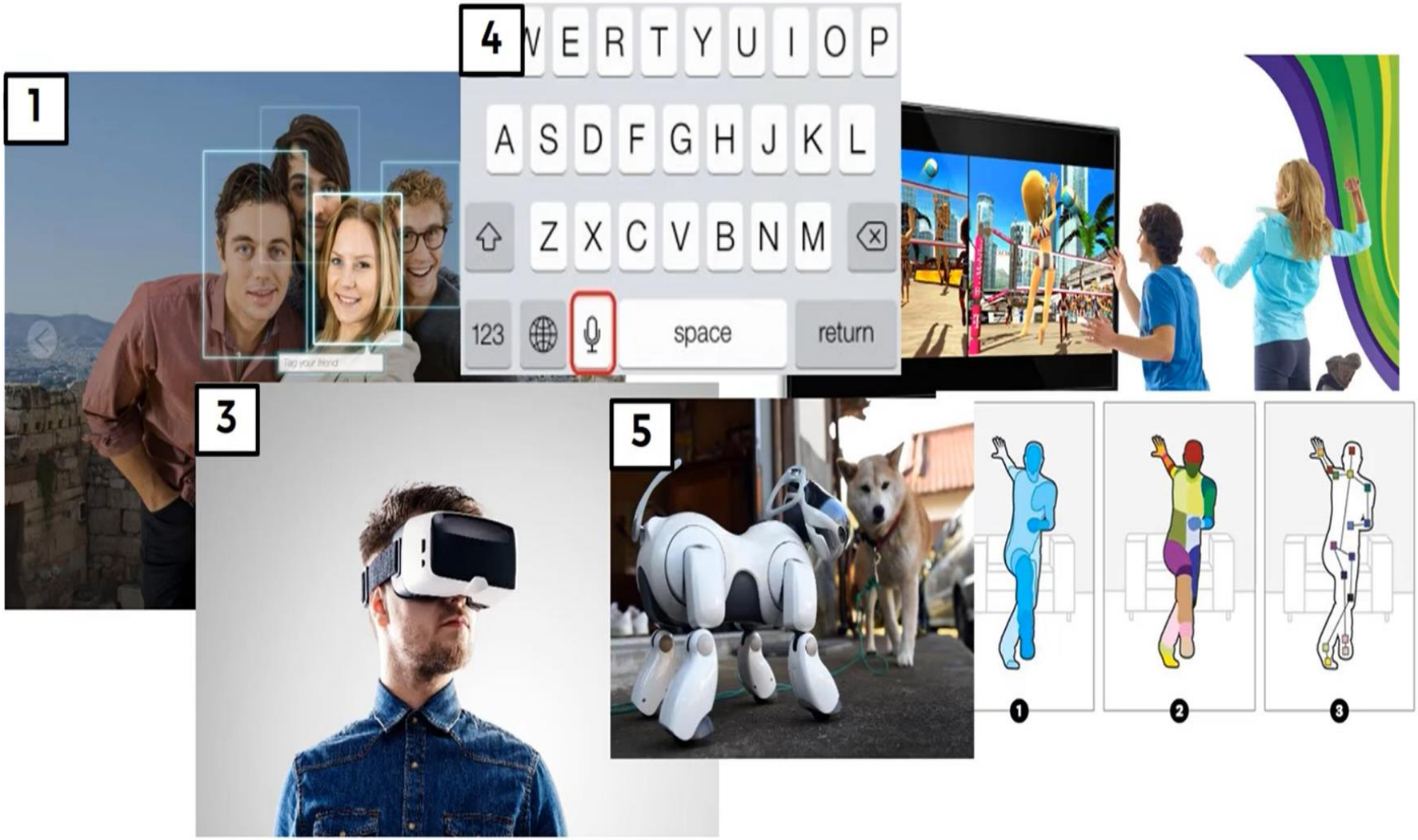
Elsa

FaceApp

Alexa

AI and ML Use cases

DATA SCIENCE

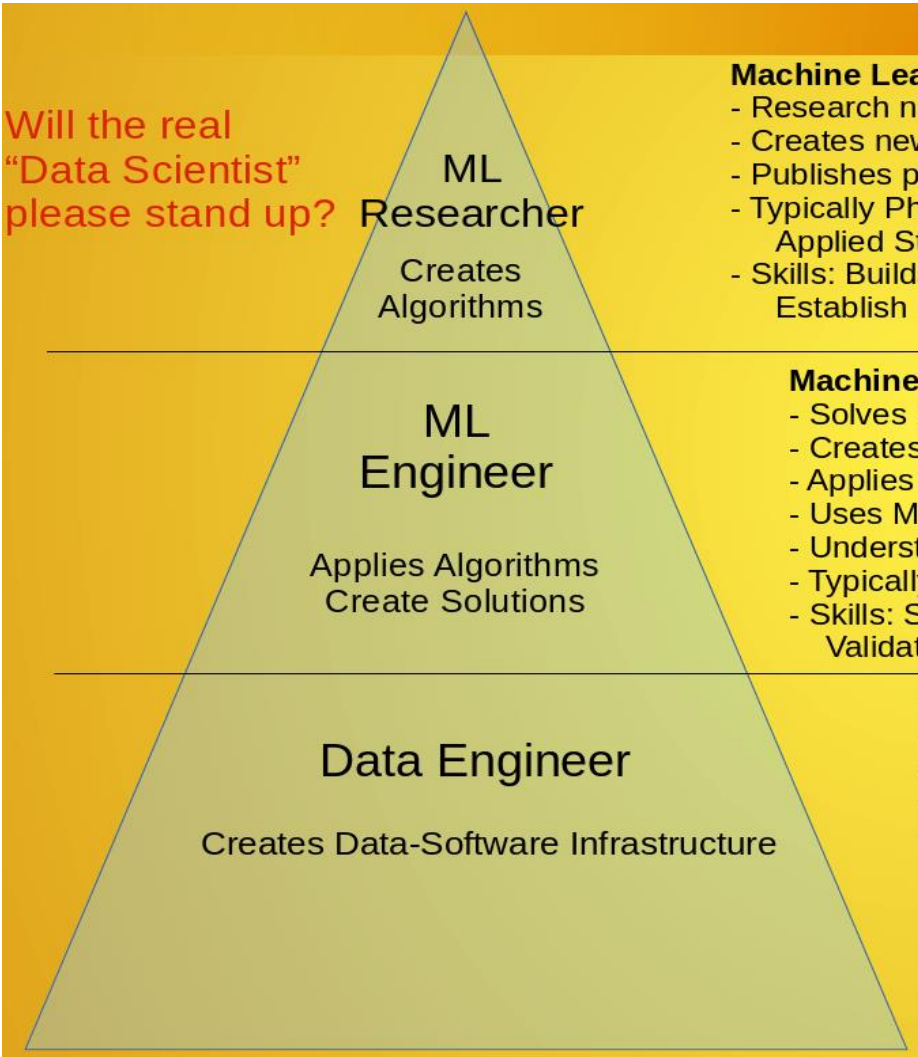


AI and ML Use cases

DATA SCIENCE

ML Skill Pyramid

DATA SCIENCE



Machine Learning Researcher/Scientist:
- Research novel machine learning problems
- Creates new mathematical models and algorithms
- Publishes papers on research results
- Typically PhD/MA Level: Robotics, Machine Learning, Cognitive Science, Applied Statistics, Engineering, Operations Research, Math, etc.
- Skills: Builds mathematical models, Breaks ground in research, Establish new paradigms, Scientific Formalism, Experiment design

Machine Learning Engineer:
- Solves business/data learning problems
- Creates ML solutions to achieve an organization's objective
- Applies established algorithms
- Uses ML algorithm libraries
- Understands strengths and weaknesses of different algorithms
- Typically BS/MA Level: Computer Science, Math, Other Technical
- Skills: Software Eng. PLUS Data Analysis, ML Algorithm Selection, Cross Validation, Metrics/Scoring, Feature Engineering

Data Engineer:
- Develops code in support of Machine Learning Solutions
- Data extraction, transformation, scraping, joining, cleaning
- Summary Statistics, counting, sampling on request
- Skills: Platform/DB/Language specific expertise, Performance, Parallel and Distributed Computing, Quality, Reliability, Map/Reduce-Hadoop, VMs/Cloud, SQL/noSQL, Production Scaling etc.

How to become ML Engineer

Machine Learning Engineer

DATA SCIENCE

Who is a Machine Learning (ML) Engineer?

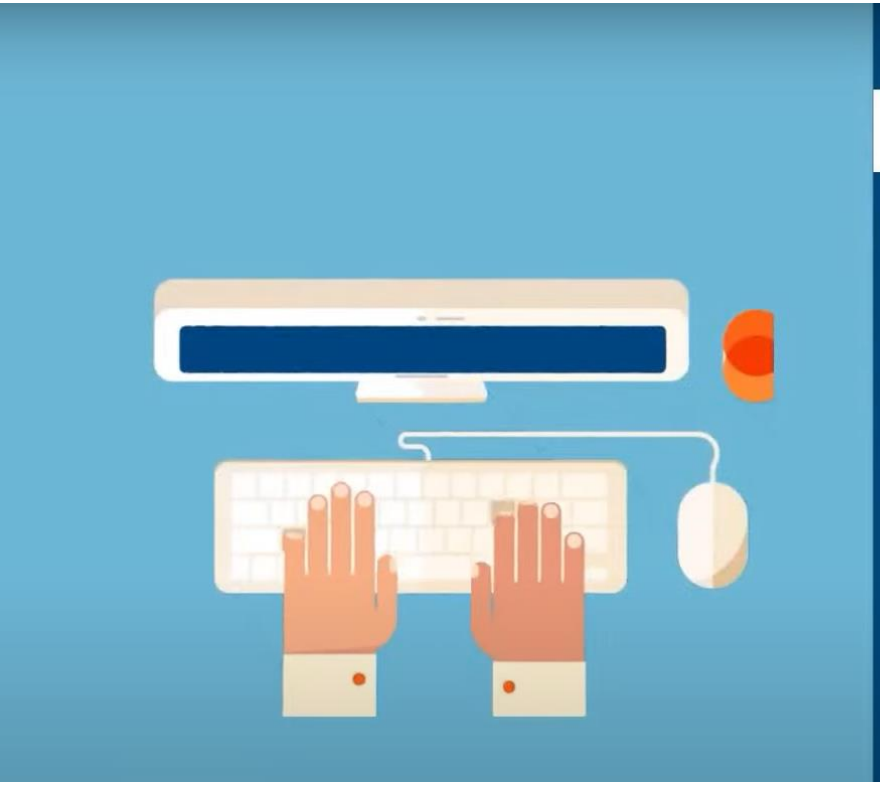
Machine learning engineers are sophisticated programmers who **develop** machines and systems that can learn and apply knowledge without specific direction

Machine Learning Engineer

What does an ML Engineer do?

Machine learning engineers are creators of the algorithms that allow a machine to find patterns in its own programming data, teaching it to understand commands and even think for itself

Machine Learning Engineer



Skills Needed to Become an ML Engineer

Programming Skills

- **R** : Used for developing Statistical Software and Data Analysis
- **Python**: Lets you Create, Analyse and Organize large chunks of Data with ease
- **Java Programming**: Data Description

Machine Learning Engineer

DATA SCIENCE

Roles & Responsibilities of an ML Engineer

- ✓ Study some Prototypes & Transform them into Applications
- ✓ Design and Build Machine Learning Systems
- ✓ Find Appropriate Algorithms & Tools
- ✓ Develop Machine Learning Applications
- ✓ Select the right dataset & find the correct Data Representation Methods
- ✓ Run Machine Learning Tests and Experiments
- ✓ Train the Systems for Top-notch Accuracy

lastly we need to train the systems for top Notch

Machine Learning Engineer

ML is the Future

DATA SCIENCE

2005 – 130 EXABYTES

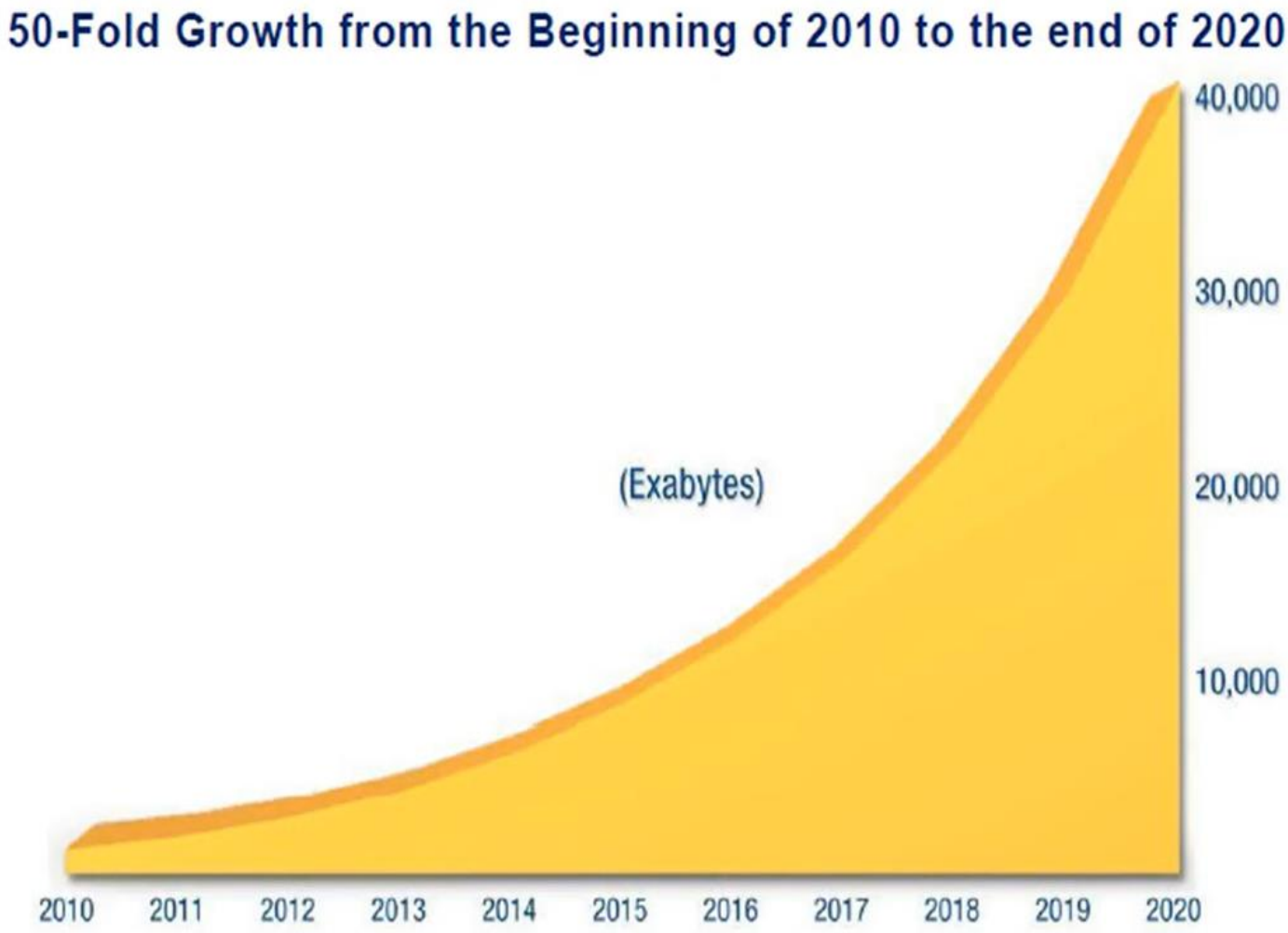
2010 – 1,200 EXABYTES

2015 – 7,900 EXABYTES

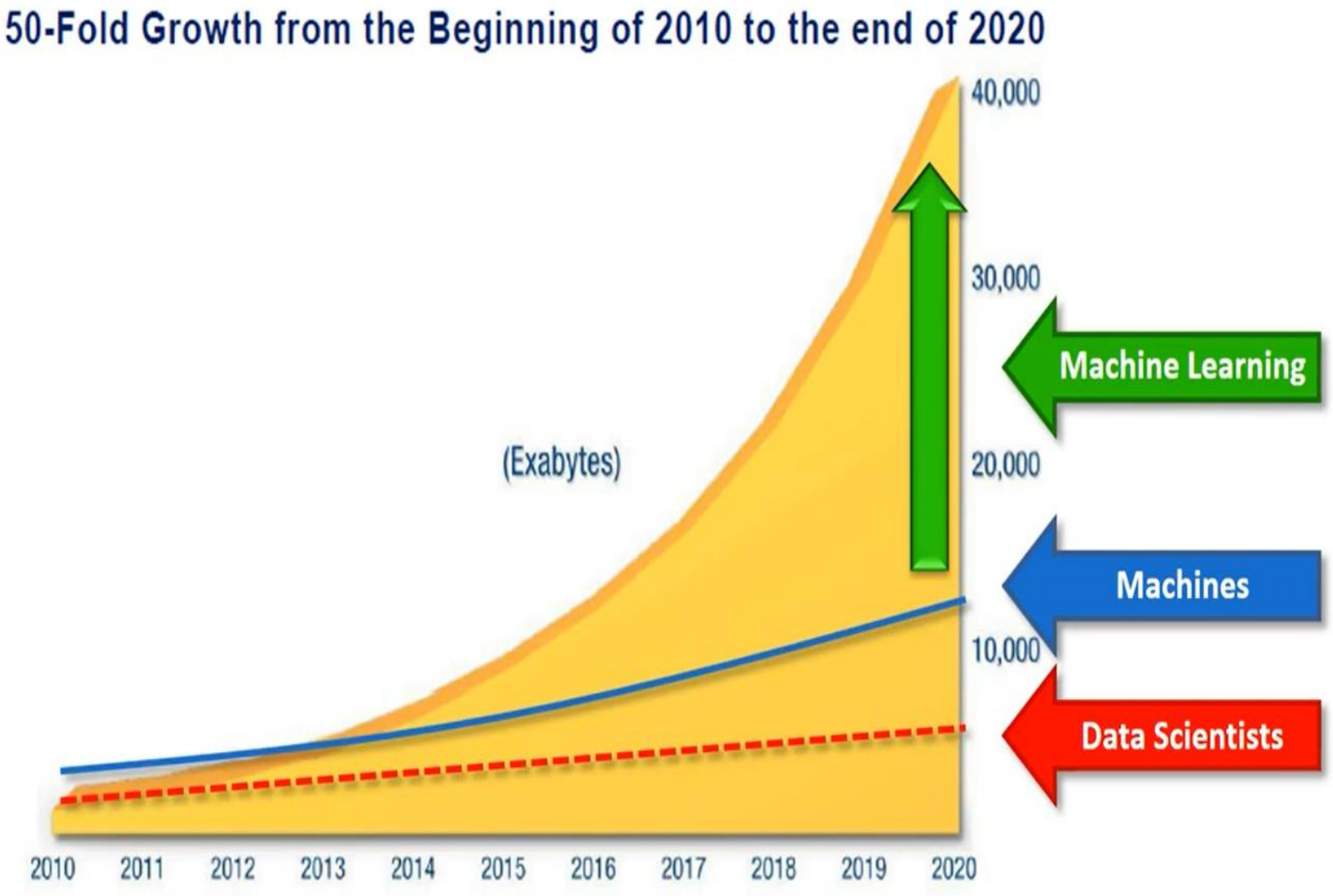
2020 – 40,900 EXABYTES

ML is the Future

DATA SCIENCE



ML is the Future



Machine Learning Algorithms How to Choose

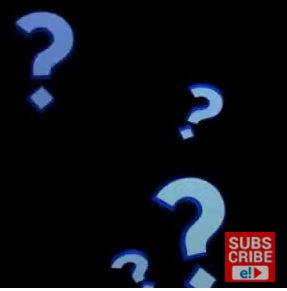
DATA SCIENCE

Classification Algorithm

Category is predicted using the data

- Is the person a male or a female?
- Is the mail spam or non-spam?

these category of question would fall under the classification algorithm



Machine Learning Algorithms How to Choose

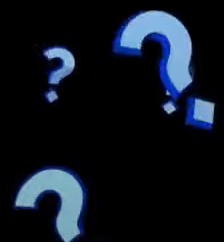
DATA SCIENCE

Anomaly Detection Algorithm

Identify unusual data points

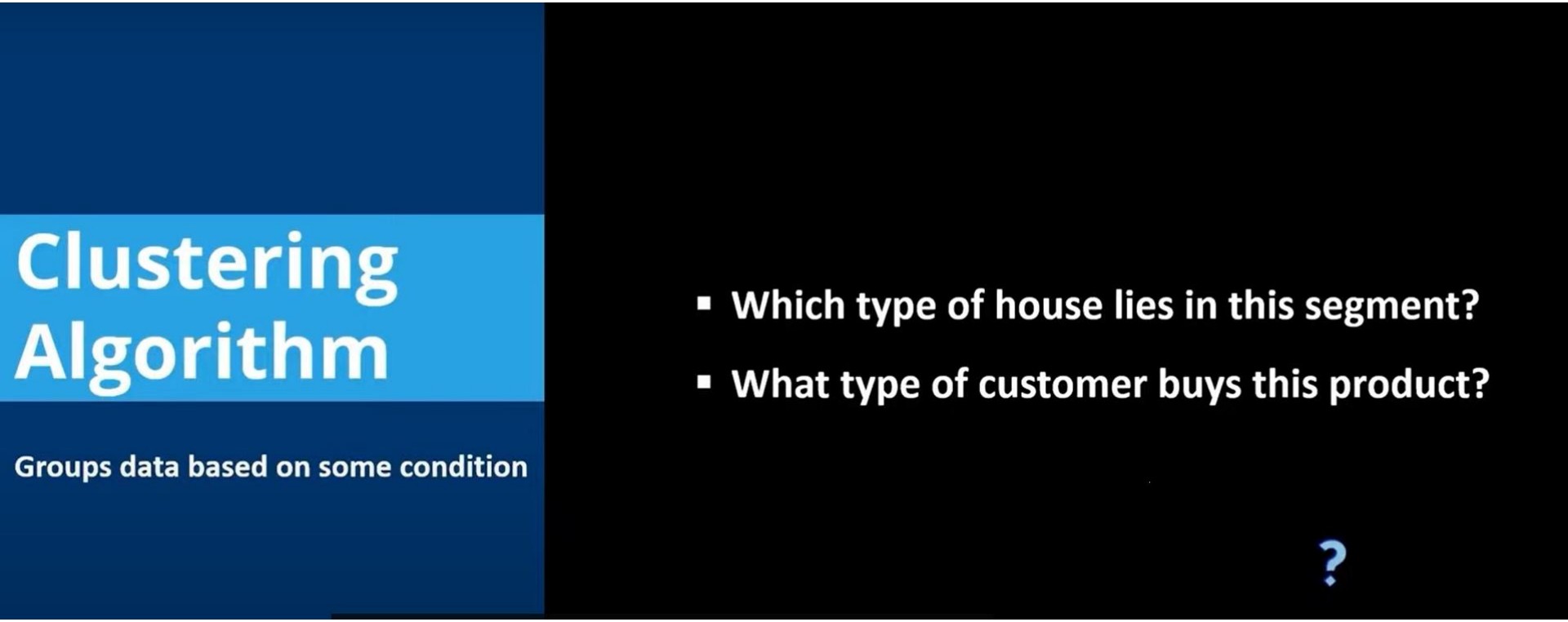
- Is there any fraud in this transaction?
- Is someone trying to hack our network?

in business like intrusion detection
like identifying strange patterns in the



Machine Learning Algorithms How to Choose

DATA SCIENCE



Machine Learning Algorithms How to Choose

DATA SCIENCE

Regression Algorithm

Data itself is predicted

- What is the market value of this house?
- Is it going to rain tomorrow?

important and broadly used machine learning and statistics tool it

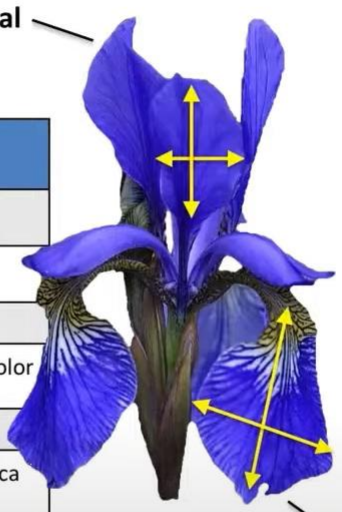
Machine Learning Algorithms How to Choose

DATA SCIENCE

Iris Dataset

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

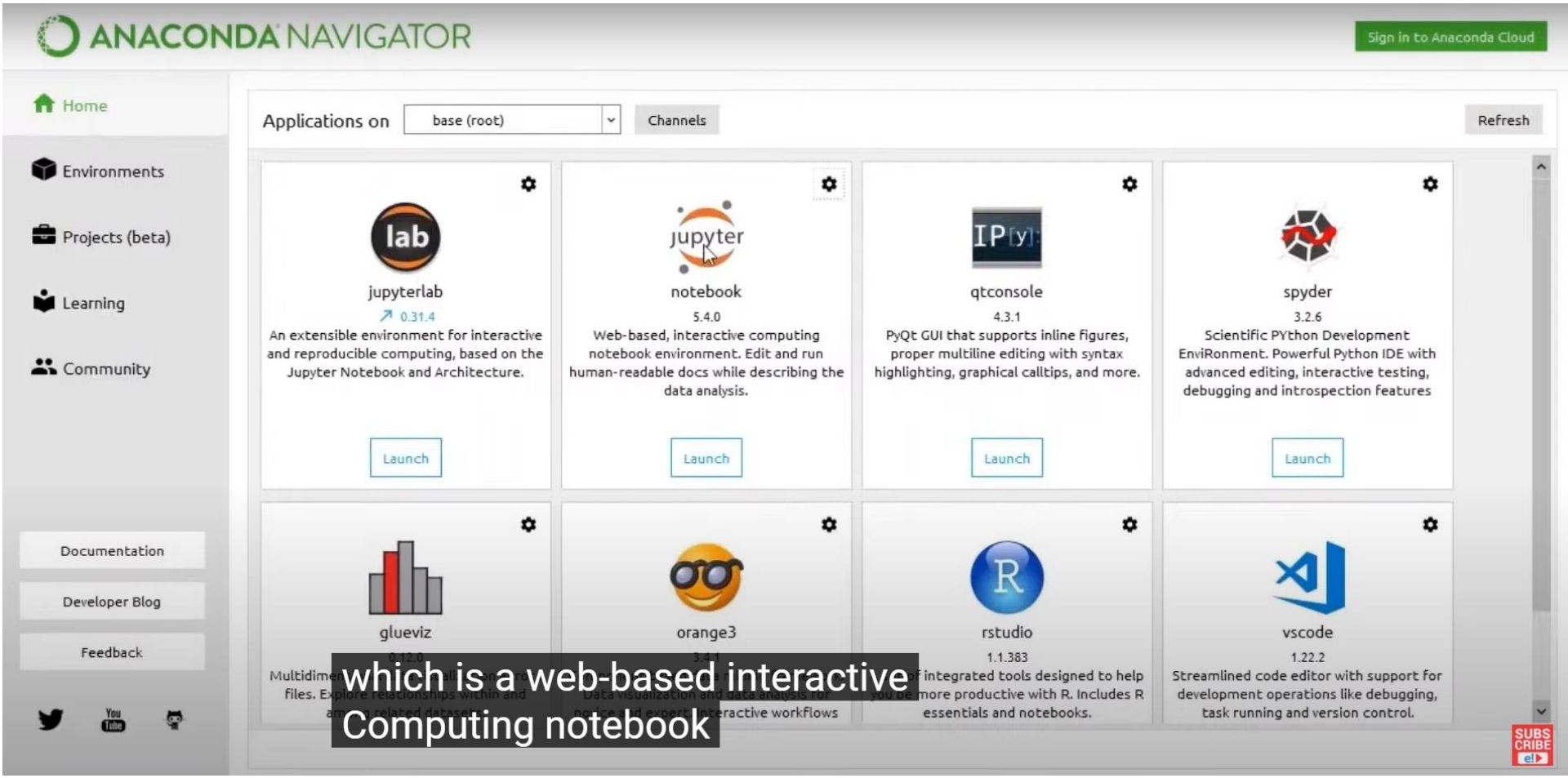


Class labels
(targets)

Features
(attributes, measurements, dimensions)

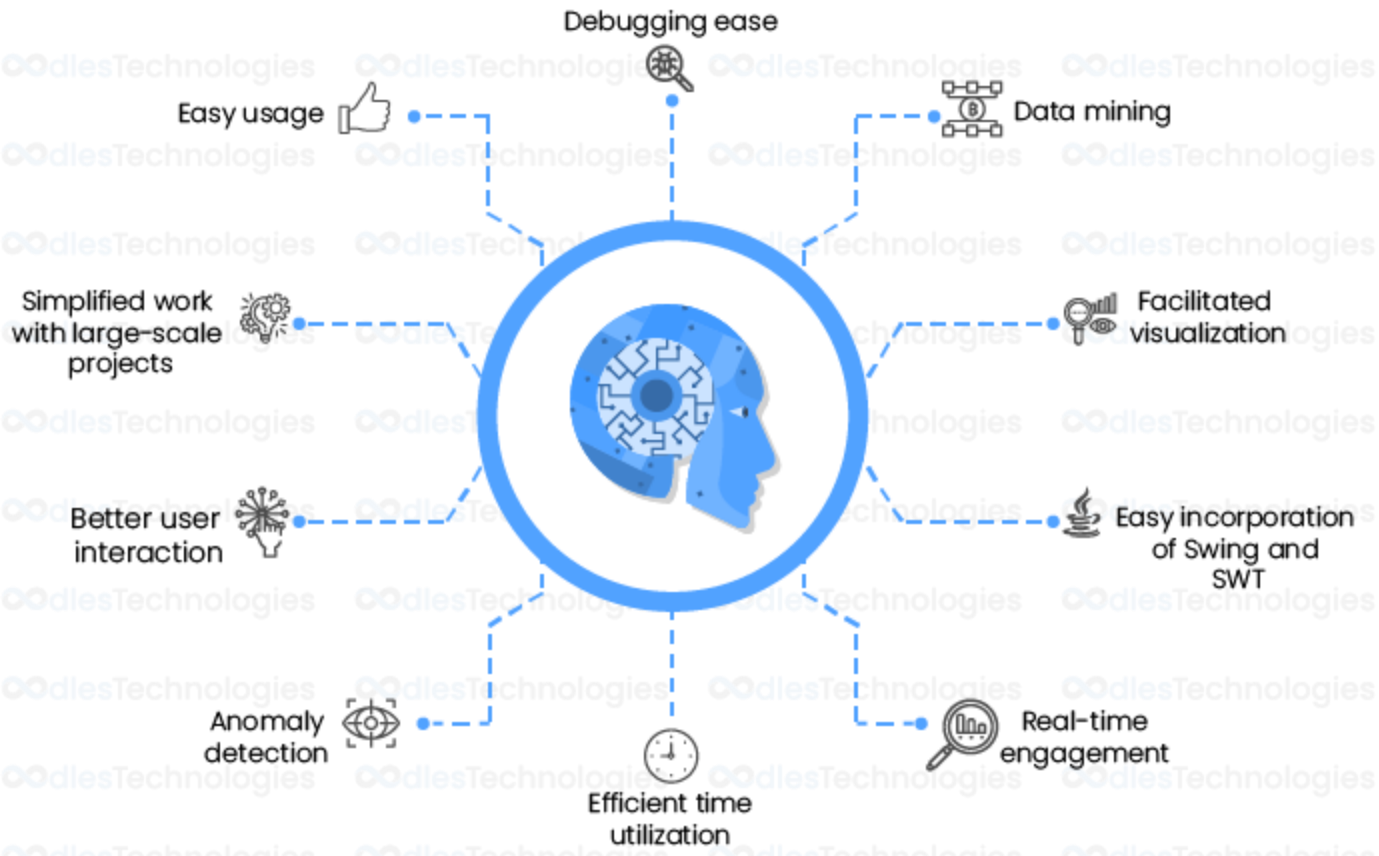
the flower observed all the observed flowers belong to one of

Machine Learning Algorithms How to Choose



Benefits of Programming AI in Java

DATA SCIENCE



For Expert Systems

- Apache Jena — a framework to create web and linked data applications.
- PowerLoom — a platform to build knowledge-based applications and reasoning systems.
- d3web — a reasoning engine with many algorithms to solve given problems.
- Eye — a reasoning engine to perform semi-backward reasoning.
- Tweety — a collection of frameworks for logical aspects of AI and knowledge representation.

For Neural Networks

- Neuroph — an open-source framework for neural network creation.
- Deeplearning4j — a deep learning library for JVM that also provides API for neural network creation.
- For Natural Language Processing
  - Apache OpenNLP — a toolkit to process the natural language text.
  - Stanford CoreNLP — a framework to perform NLP tasks.

For ML

- Java-ML — a collection of machine learning algorithms.
- RapidMiner — a data science platform that provides machine learning algorithms through GUI and Java API.
- Weka — a collection of machine learning algorithms.
- Encog — a collection of advanced algorithms.
- For Genetic Algorithms
  - Jenetics — an advanced genetic algorithm.

For ML

- Watchmaker — a framework for implementing genetic algorithms.
- ECJ 23 — a research framework with support for genetic algorithms.
- JGAP (Java Genetic Algorithms Package) — a genetic programming component.
- Eva — a simple OOP evolutionary algorithm framework.
- For automatic programming
  - Spring Roo — a lightweight developer tool.
  - Acceleo — a code generator for Eclipse which creates code from EMF models.

Data Acquisition

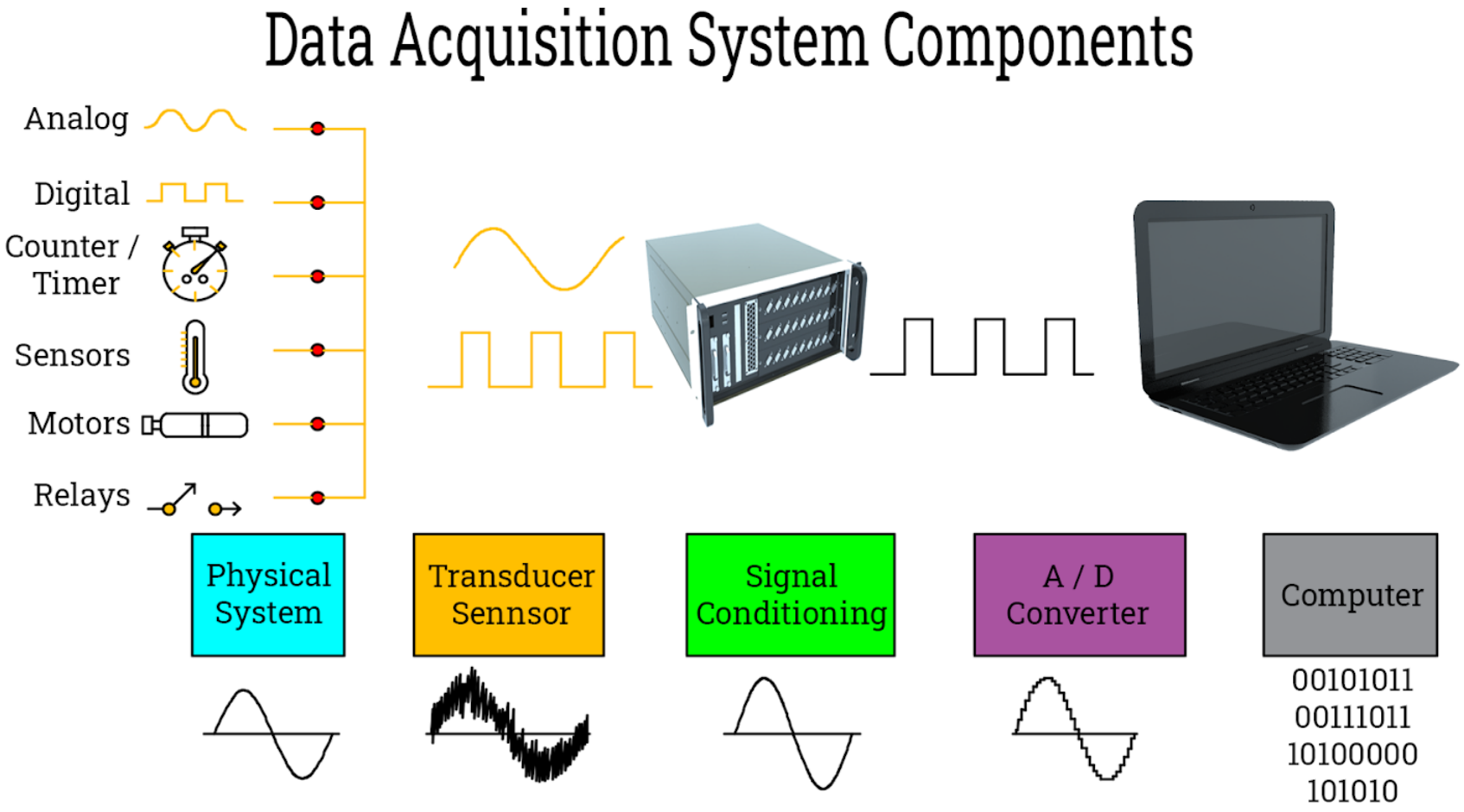
DATA SCIENCE

- A data acquisition system is a system that comprises sensors, measurement devices, and a computer.
- A data acquisition system is used for processing acquired data, which involves collecting the information required to understand electrical or physical phenomena.

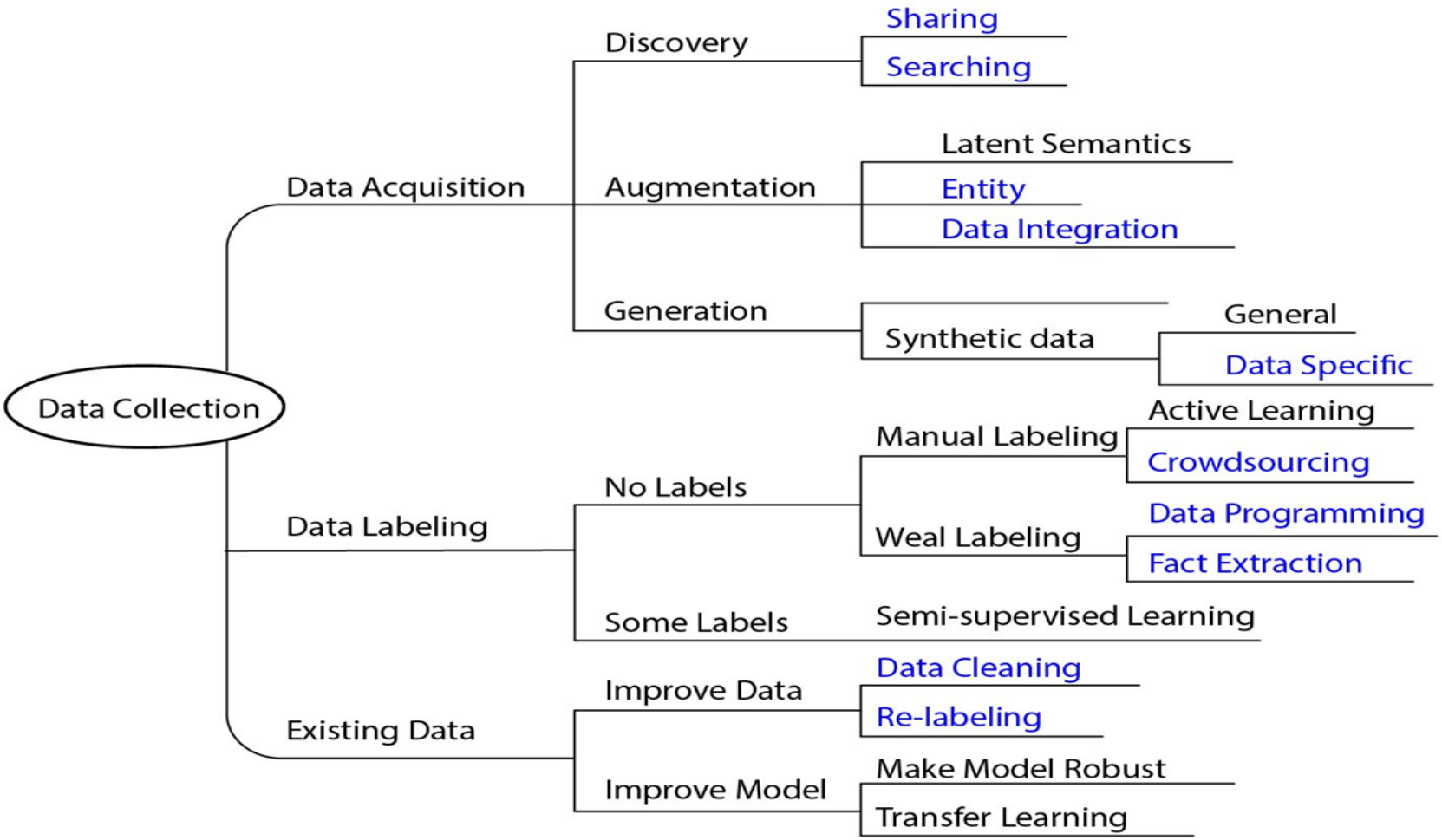
Data Acquisition

DATA SCIENCE

Basic Components of Data Acquisition System



Basic Components of Data Acquisition System



Data Ingestion

DATA SCIENCE

- Data ingestion is the transportation of data from assorted sources to a storage medium where it can be accessed, used, and analyzed by an organization.
- The destination is typically a data warehouse, data mart, database, or a document store.
- Sources may be almost anything — including SaaS data, in-house apps, databases, spreadsheets, or even information scraped from the internet.

Data Ingestion

- In a broader sense, data ingestion can be understood as a directed dataflow between two or more systems that result in a smooth, and independent, operation (a definition which already implies some independence or automation).
- Ingestion can occur in real-time, as soon as the source produces it, or in batches, when data is input in specific chunks at set periods.
- Generally, three steps occur within data ingestion:
- Data extraction – retrieving data from sources
- Data transformation – validating, cleaning, and normalizing data to ensure accuracy and reliability (sometimes known as trustworthiness)
- Data loading – routing or placing the data in its correct silo or database for analysis

Reasons to automate data ingestion

- Automating data ingestion improves time-to-market goals.
- Automating data ingestion increases scalability.
- Automating data ingestion refocuses on necessary work.
- Automating data ingestion mitigates risk.

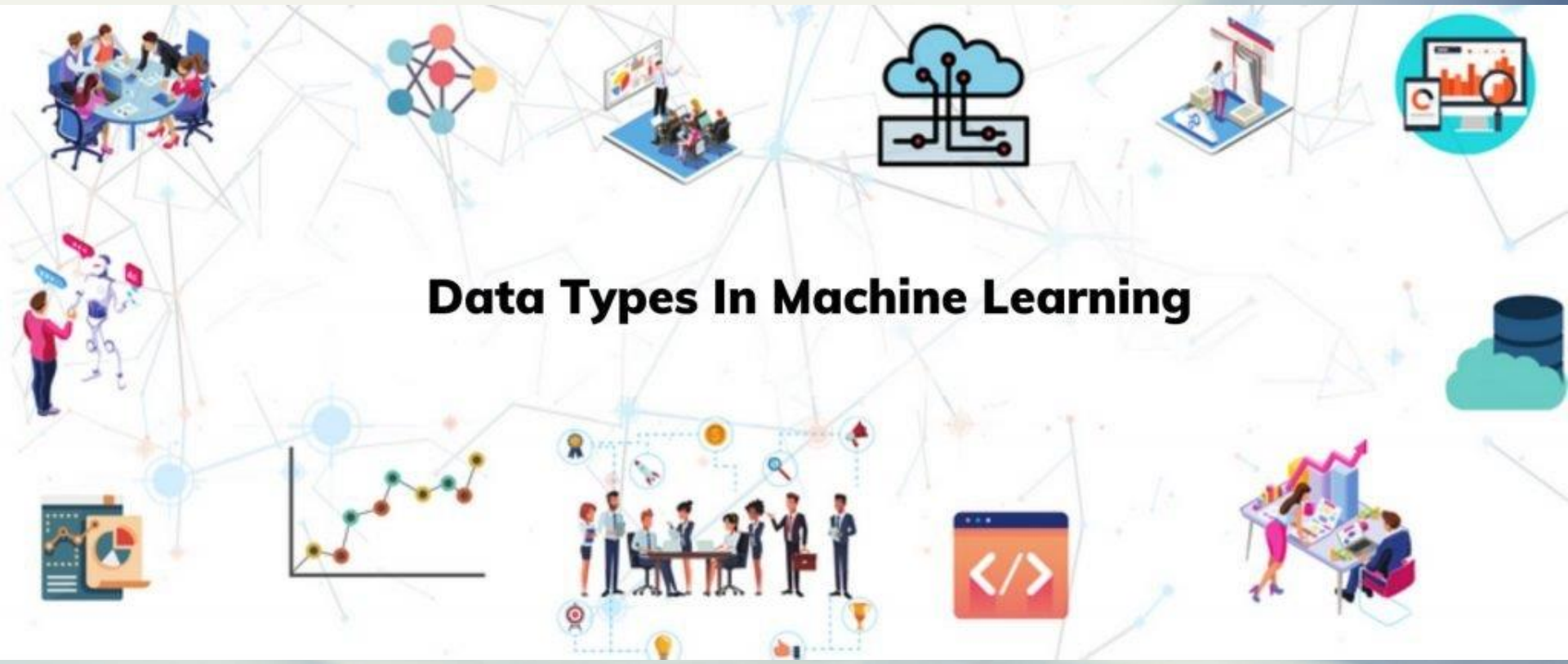
Data Ingestion Tools

- Apache NiFi (a.k.a. Hortonworks DataFlow)
- Stream Sets Data Collector (SDC)
- Gobblin
- Sqoop
- Flume
- Kafka
- Airflow
- Spark
- petl

Data Ingestion Tools

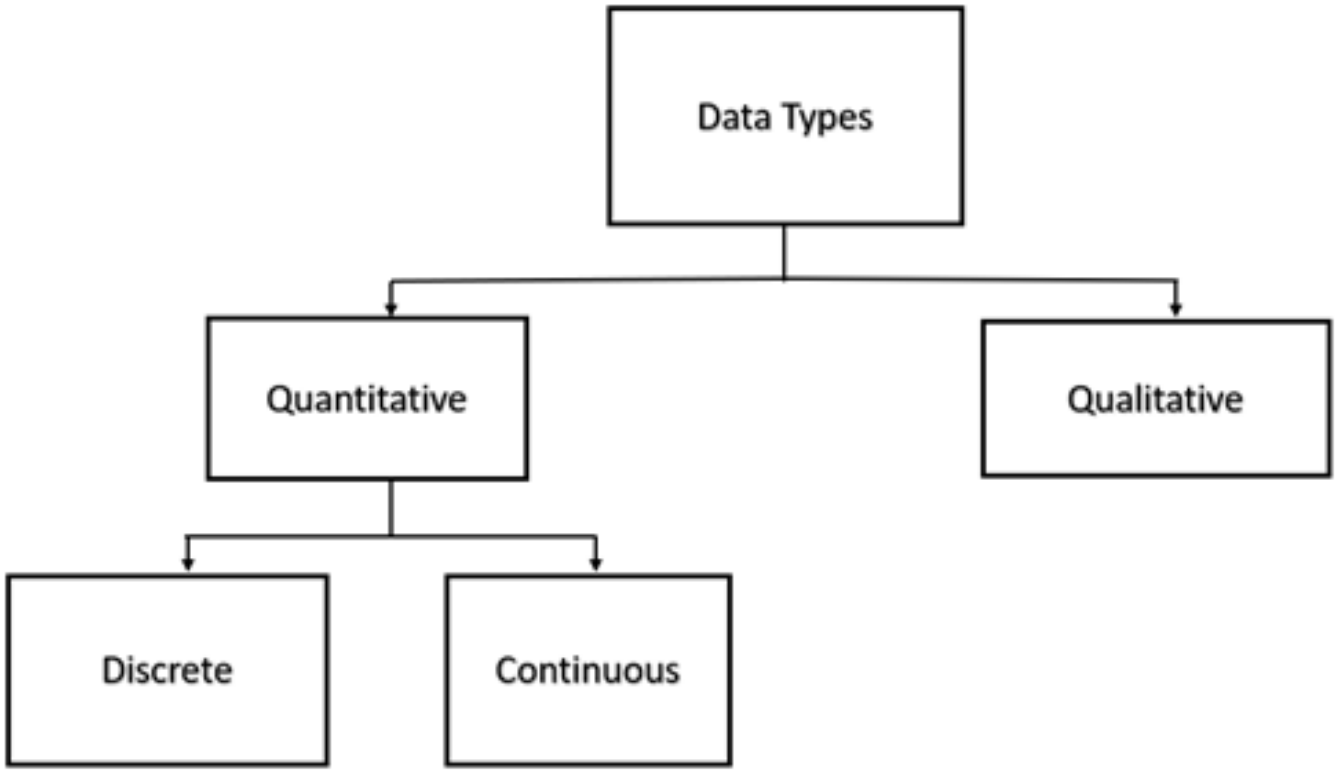
- Panoply
- Pandas
- Bubbles
- Bonobo
- Luigi
- etlalchemy

DATA SCIENCE



Data Types in Machine Learning

DATA SCIENCE



Data Types in Machine Learning

DATA SCIENCE

- Quantitative Data Type: –
  - This type of data type consists of numerical values. Anything which is measured by numbers.
  - ex., Profit, Quantity Sold, Height, Weight, Temperature, etc.

Data Types in Machine Learning

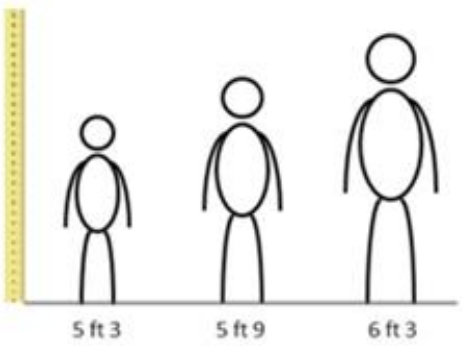
- A.) Discrete Data Type: –
- The numeric data which have discrete values or whole numbers.
- This type of variable value if expressed in decimal format will have no proper meaning.
- Their values can be counted.
- ex: – No. Of Cars You Have, No. Of Marbles In Containers, Students In A Class, etc.

No. of Laptops

No. of Cars

Data Types in Machine Learning

- B.) Continuous Data Type: –
- The numerical measures which can take the value within A certain range.
- This type of variable value if expressed in decimal format has true meaning.
- Their values can not be counted but measured. The value can be infinite
- ex: – Height, Weight, Time, Area, Distance, Measurement Of Rainfall, etc.



Height

Time

Data Types in Machine Learning

- Qualitative Data Type: –
- These are the data types that cannot be expressed in numbers.
- This describes categories or groups and is hence known as the categorical data type.

Data Types in Machine Learning

DATA SCIENCE

- Structured Data:
- This type of data is either number or words.
- This can take numerical values, but mathematical operations cannot be performed on it.
- This type of data is expressed in tabular format.
- ex) Sunny=1, Cloudy=2, Windy=3 Or Binary Form Data Like 0 Or1, Good Or Bad, Etc.

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

Data Types in Machine Learning

DATA SCIENCE

- Unstructured Data:
- This type of data does not have the proper format and therefore known as unstructured data.
- This comprises textual data, sounds, images, videos, etc.

Data Types in Machine Learning

- This, there are also other types refer as data types preliminaries or data measures:-
- Nominal
- Ordinal
- Interval
- Ratio

Data Types in Machine Learning

DATA SCIENCE

- Nominal Data Type:
- This is in use to express names or labels which are not order or measurable.
- ex., Male or female (gender), race, country, etc.

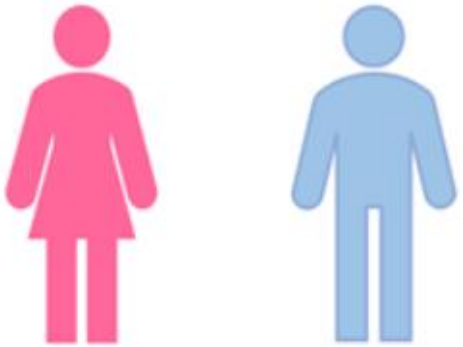


Fig: Gender (Female, Male), An Example Of Nominal Data Type

Data Types in Machine Learning

DATA SCIENCE

- Ordinal Data Type:
- This is also A categorical data type like nominal data but has some natural ordering associated with it.
- Ex., Likert rating scale, shirt sizes, ranks, grades, etc.,

Fig: Rating (Good, Average, Poor), An Example Of Ordinal Data Type

Data Types in Machine Learning

DATA SCIENCE

- Interval Data Type:
- This is numeric data which has proper order and the exact zero means the true absence of A value attached.
- Here zero means not A complete absence but has some value. This is the local scale.
- ex., Temperature measured in degree Celsius, time, sat score, credit score, PH, etc. Difference between values is familiar. In this case, there is no absolute zero. Absolute

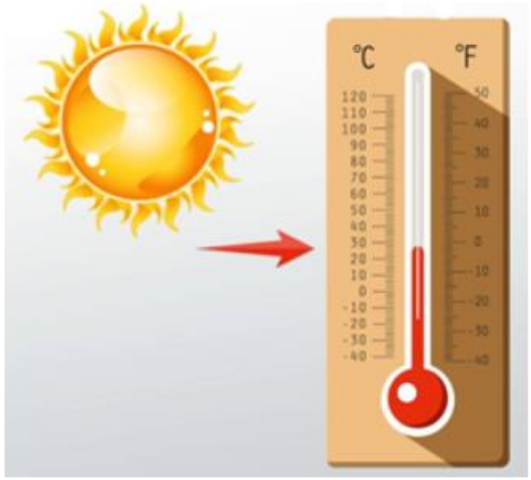


Fig: Temperature, An Example Of Interval Data Type

Data Types in Machine Learning

DATA SCIENCE

- Ratio Data Type:
- This quantitative data type is the same as the interval data type but has the absolute zero. Here zero means complete absence and the scale starts from zero. This is the global scale.
- ex., Temperature in kelvin, height, weight, etc.

Fig: Weight, An Example Of Ratio Data Type

Difference between Structured, Semi-structured and Unstructured data

PROPERTIES	STRUCTURED DATA	SEMI-STRUCTURED DATA	UNSTRUCTURED DATA
Technology	It is based on Relational database table	It is based on XML/RDF	It is based on character and binary data
Transaction management	Matured transaction and various concurrency technique	Transaction is adapted from DBMS not matured	No transaction management and no concurrency
Version management	Versioning over tuples,row,tables	Versioning over tuples or graph is possible	Versioned as whole
Flexibility	It is sehema dependent and less flexible	It is more flexible than structuded data but less than flexible than unstructured data	it very flexible and there is abbsence of schema
Scalability	It is very difficult to scale DB schema	It’s scaling is simpler than sstructured data	It is very scalable
Robustness	Very robust	New technology, not very spread	—
Query performance	Structured query allow complex joining	Queries over anonymous nodes are possible	Only textual query are possible

Data from Various Sources

- CSV file
- Flat File (tab, space, or any other separator)
- Text File (In a single file — reading data all at once)
- ZIP file
- Multiple Text Files (Data is split over multiple text files)
- Download File from Internet (File hosted on a server)
- Webpage (scraping)
- APIs (JSON)
- Text File (Reading data line by line)
- RDBMS (SQL Tables)

Common Data Platforms

DATA SCIENCE

- Inconsistent column names
- Missing data
- Outliers
- Duplicate rows
- Untidy
- Need to process columns
- Column types can signal unexpected data values

Exploratory Data Analysis (EDA)

DATA SCIENCE

- Exploratory Data Analysis (EDA) refers to the method of studying and exploring record sets.
- To apprehend their predominant traits, discover patterns, locate outliers, and identify relationships between variables.
- EDA is normally carried out as a preliminary step before undertaking extra formal statistical analyses or modeling.

The Foremost Goals of EDA

- Data Cleaning: EDA involves examining the information for errors, lacking values, and inconsistencies.
- It includes techniques including records imputation, managing missing statistics, and figuring out and getting rid of outliers.

The Foremost Goals of EDA

- Descriptive Statistics: EDA utilizes precise records to recognize the important tendency, variability, and distribution of variables.
- Measures like suggest, median, mode, preferred deviation, range, and percentiles are usually used.

The Foremost Goals of EDA

- Data Visualization: EDA employs visual techniques to represent the statistics graphically.
- Visualizations consisting of histograms, box plots, scatter plots, line plots, heatmaps, and bar charts assist in identifying styles, trends, and relationships within the facts.

The Foremost Goals of EDA

DATA SCIENCE

- Feature Engineering: EDA allows for the exploration of various variables and their adjustments to create new functions or derive meaningful insights.
- Feature engineering can contain scaling, normalization, binning, encoding express variables, and creating interplay or derived variables.

The Foremost Goals of EDA

DATA SCIENCE

- Correlation and Relationships: EDA allows discover relationships and dependencies between variables.
- Techniques such as correlation analysis, scatter plots, and pass-tabulations offer insights into the power and direction of relationships between variables.

The Foremost Goals of EDA

- Data Segmentation: EDA can contain dividing the information into significant segments based totally on sure standards or traits.
- This segmentation allows advantage insights into unique subgroups inside the information and might cause extra focused analysis.

The Foremost Goals of EDA

DATA SCIENCE

- Hypothesis Generation: EDA aids in generating hypotheses or studies questions based totally on the preliminary exploration of the data.
- It facilitates form the inspiration for in addition evaluation and model building.

The Foremost Goals of EDA

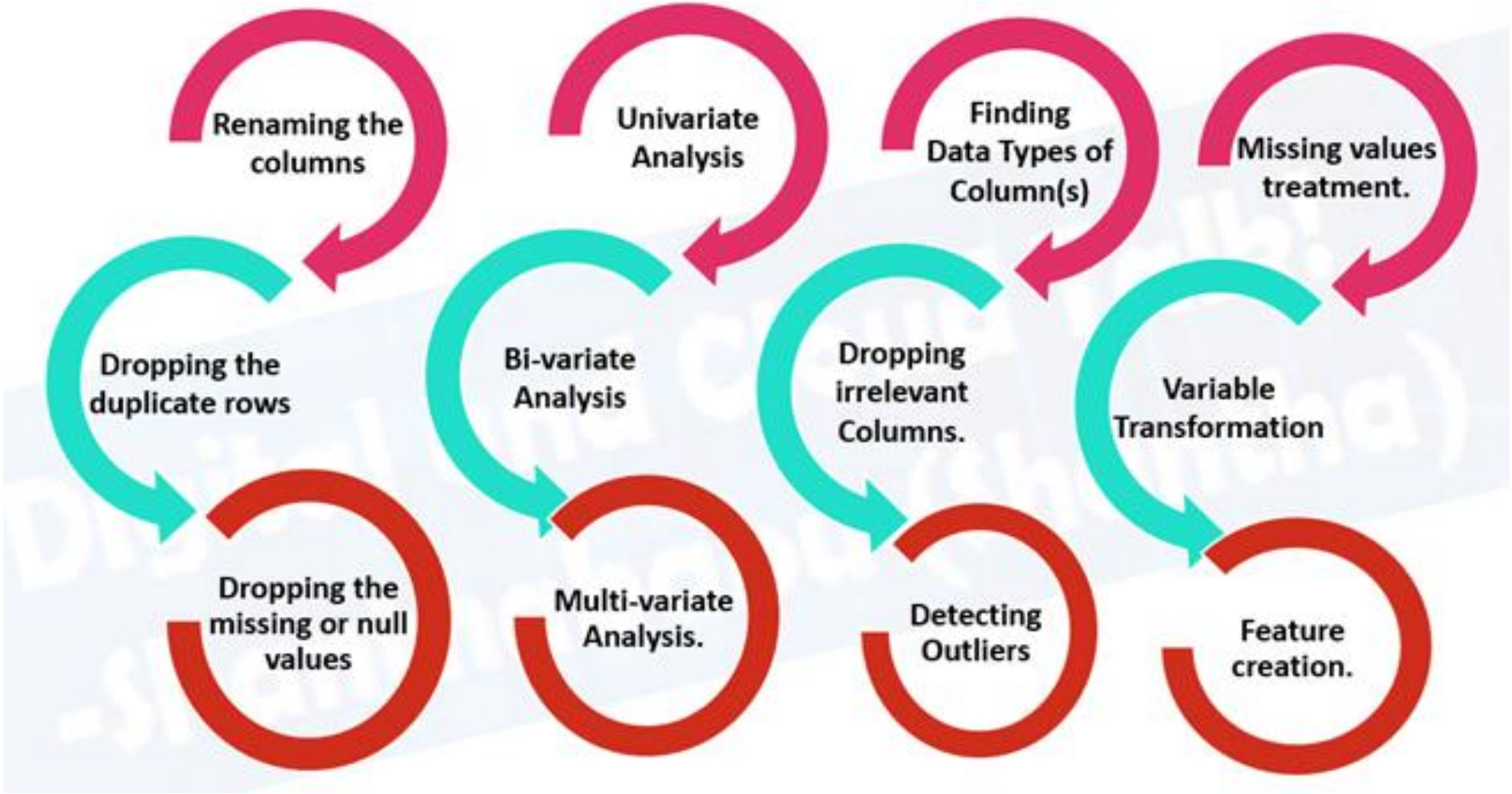
- Data Quality Assessment: EDA permits for assessing the nice and reliability of the information.
- It involves checking for records integrity, consistency, and accuracy to make certain the information is suitable for analysis.

Outcome of EDA

- Understanding the given dataset and helps clean up the given dataset.
- It gives you a clear picture of the features and the relationships between them.
- Providing guidelines for essential variables and leaving behind/removing non-essential variables.
- Handling Missing values or human error.
- Identifying outliers.
- EDA process would be maximizing insights of a dataset.

Types of EDA

DATA SCIENCE



Types of EDA

DATA SCIENCE

- Univariate Analysis: This sort of evaluation makes a speciality of analyzing character variables inside the records set.
- It involves summarizing and visualizing unmarried variable at a time to understand its distribution, relevant tendency, unfold, and different applicable records.
- Techniques like histograms, field plots, bar charts, and precis information are generally used in univariate analysis.

Types of EDA

DATA SCIENCE

- Bivariate Analysis: Bivariate evaluation involves exploring the connection between variables.
- It enables find associations, correlations, and dependencies between pairs of variables.
- Scatter plots, line plots, correlation matrices, and move-tabulation are generally used strategies in bivariate analysis.

Types of EDA

DATA SCIENCE

- Multivariate Analysis: Multivariate analysis extends bivariate evaluation to encompass greater than variables.
- It ambitions to apprehend the complex interactions and dependencies among more than one variables in a records set.
- Techniques inclusive of heatmaps, parallel coordinates, aspect analysis, and primary component analysis (PCA) are used for multivariate analysis.

Types of EDA

DATA SCIENCE

- Time Series Analysis: This type of analysis is mainly applied to statistics sets that have a temporal component.
- Time collection evaluation entails inspecting and modeling styles, traits, and seasonality inside the statistics through the years.
- Techniques like line plots, autocorrelation analysis, transferring averages, and ARIMA (AutoRegressive Integrated Moving Average) fashions are generally utilized in time series analysis.

Types of EDA

- Missing Data Analysis: Missing information is a not unusual issue in datasets, and it may impact the reliability and validity of the evaluation.
- Missing statistics analysis includes figuring out missing values, know-how the patterns of missingness, and using suitable techniques to deal with missing data.
- Techniques along with lacking facts styles, imputation strategies, and sensitivity evaluation are employed in lacking facts evaluation.

Types of EDA

DATA SCIENCE

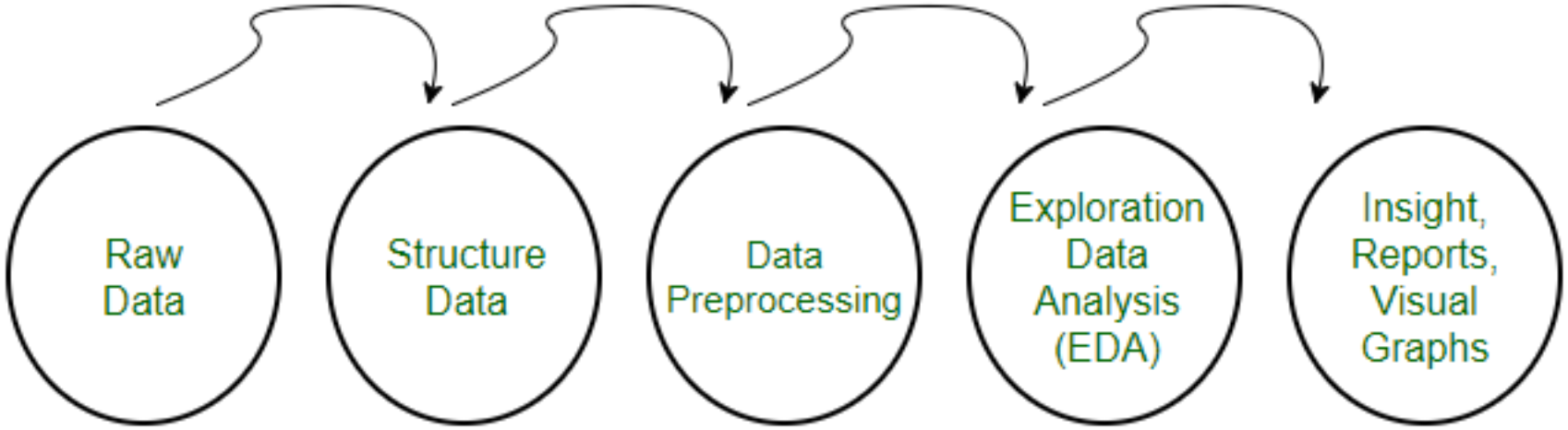
- Outlier Analysis: Outliers are statistics factors that drastically deviate from the general sample of the facts.
- Outlier analysis includes identifying and knowledge the presence of outliers, their capability reasons, and their impact at the analysis.
- Techniques along with box plots, scatter plots, z-rankings, and clustering algorithms are used for outlier evaluation.

Types of EDA

DATA SCIENCE

- Data Visualization: Data visualization is a critical factor of EDA that entails creating visible representations of the statistics to facilitate understanding and exploration.
- Various visualization techniques, inclusive of bar charts, histograms, scatter plots, line plots, heatmaps, and interactive dashboards, are used to represent exclusive kinds of statistics.

Data preprocessing



Steps in Data Preprocessing

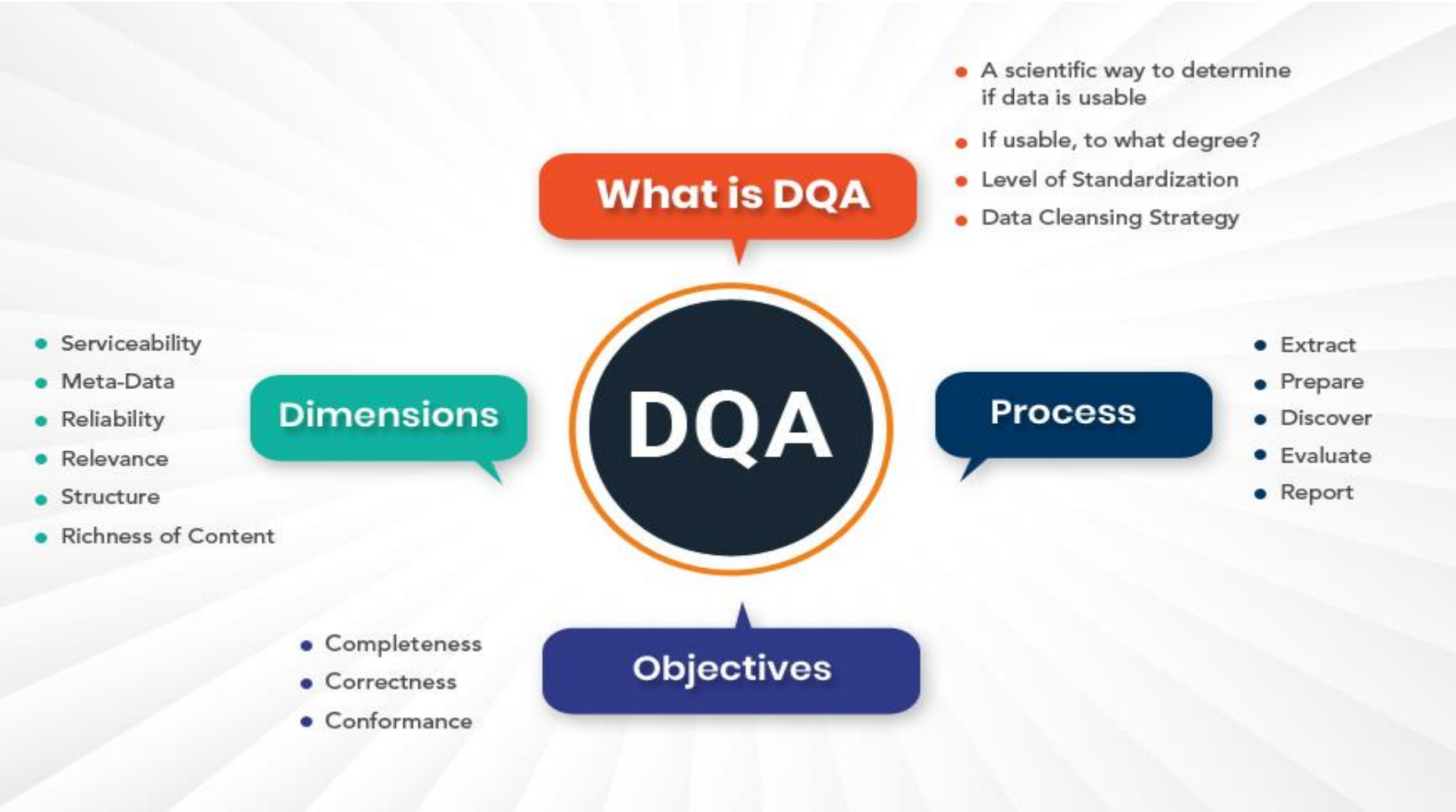
- Step 1: Import the necessary libraries
- Step 2: Load the dataset
- Step 3: Statistical Analysis
- Step 4: Check the outliers
- Step 5: Correlation
- Step 6: Separate independent features and Target Variables
- Step 7: Normalization or Standardization

Data Quality

Measuring
Knowing
Improving

Data Quality assessment

DATA SCIENCE



Our 5-step DQA Process

DATA SCIENCE

Extract

- Identify all data sources
- Identify data size – volume, duration, no. of samples
- Convert / Format data

Discover

- Perform EDA (Exploratory Data Analysis)
- Data visualization
- Identify external data that could be useful

Report

- Comprehensiveness Quotient
- Accuracy
- Richness of content

Prepare

- Clean the data – Outliers and Missing values
- Feature Engineering
- Feature scaling – Standardization & Normalization

Evaluate

- Feature importance using ML models
- Feature correlations and optimization
- Measure performance of predictive models

Benefits of DQA

DATA SCIENCE

Feature Scaling

DATA SCIENCE

- Feature scaling is a method used to normalize the range of independent variables or features of data.
- In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.
- Just to give you an example — if you have multiple independent variables like age, salary, and height; With their range as (18–100 Years), (25,000–75,000 Euros), and (1–2 Meters) respectively.
- Feature scaling would help them all to be in the same range, for example- centered around 0 or in the range (0,1) depending on the scaling technique.

Why Feature Scaling

DATA SCIENCE

#	Emp	Age	Salary
1	Emp1	44	73000
2	Emp2	27	47000
3	Emp3	30	53000
4	Emp4	38	62000
5	Emp5	40	57000
6	Emp6	35	53000
7	Emp7	48	78000

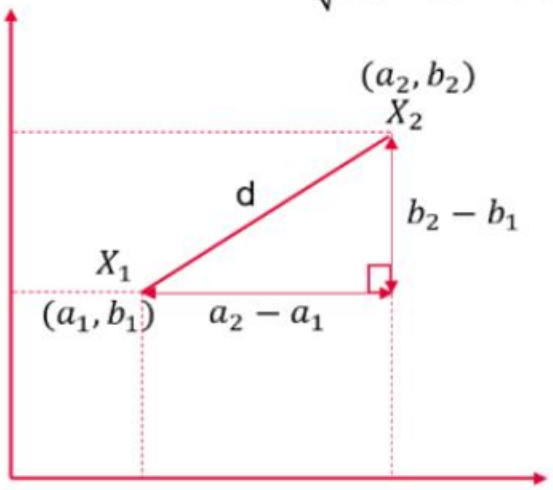
Age Range: 27-48

Salary Range: 47000-78000

Why Feature Scaling

DATA SCIENCE

$$\textbf{Euclidean Distance } d = \sqrt{(a_2-a_1)^2 + (b_2-b_1)^2}$$



- X_1 and X_2 are two data points means two different rows in data set. And Data set is two dimensional feature space
- For three dimensional feature space suppose two data points are $X_1(a_1, b_1, c_1)$ *and* $X_2(a_2, b_2, c_2)$ then distance can be calculated as below:

$$\boldsymbol{d} = \sqrt{(a_2-a_1)^2 + (b_2-b_1)^2 + (c_2-c_1)^2}$$

- In case of multi-dimensional vector space, More generic formula will be

$$\textbf{Euclidean Distance } d = \sum_{i=1}^{k} \sqrt{(X1_i - X2_i)^2}$$

Why Feature Scaling

DATA SCIENCE

#	Emp	Age	Salary
1	Emp1	44	73000
2	Emp2	27	47000
3	Emp3	30	53000
4	Emp4	38	62000
5	Emp5	40	57000
6	Emp6	35	53000
7	Emp7	48	78000

Distance between Emp2 and Emp1 = $\sqrt{(27-44)^2+(47000-73000)^2} = 31.06$

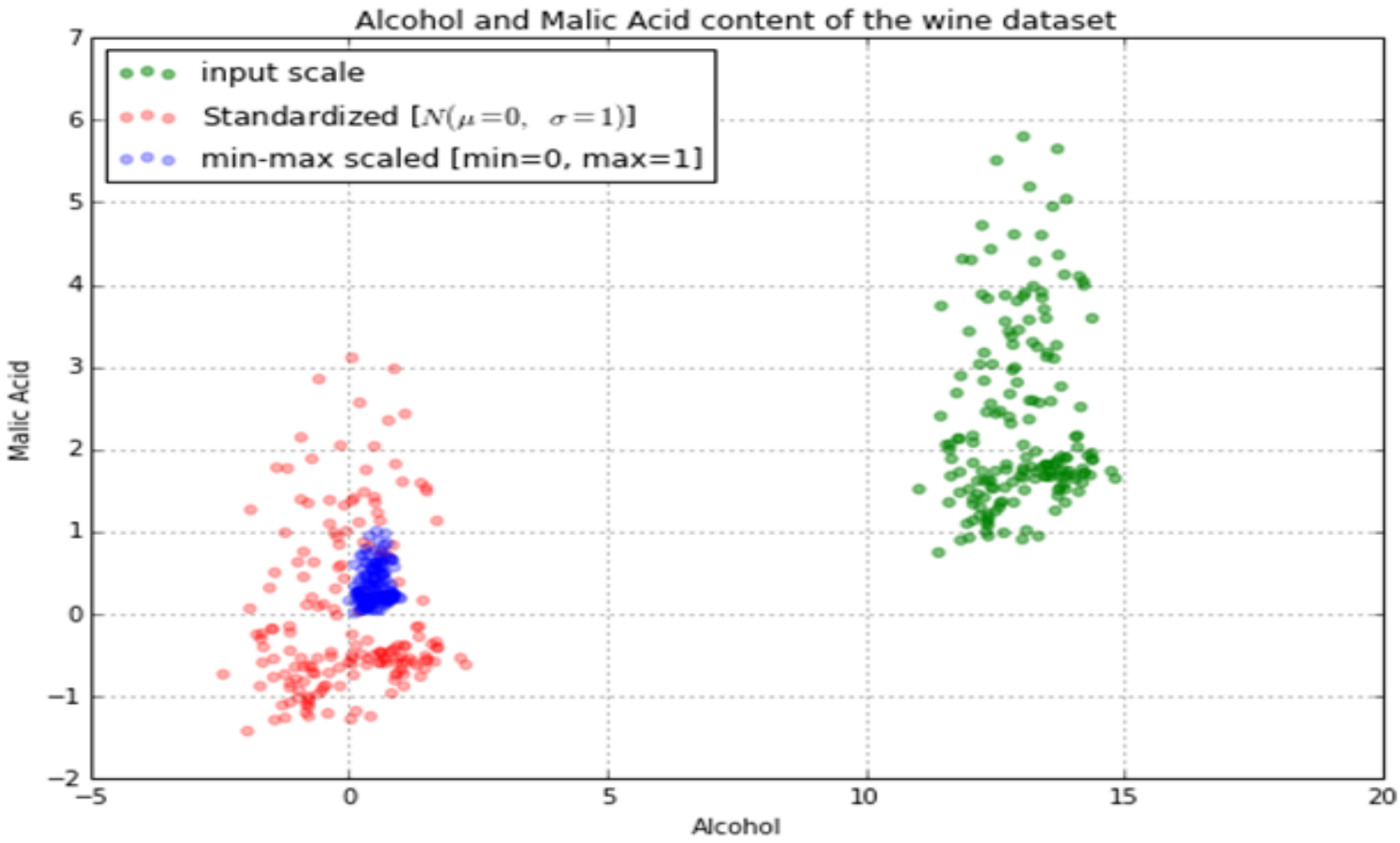
Distance between Emp2 and Emp3 = $\sqrt{(30-27)^2+(53000-47000)^2} = 6.70$

Feature Scaling

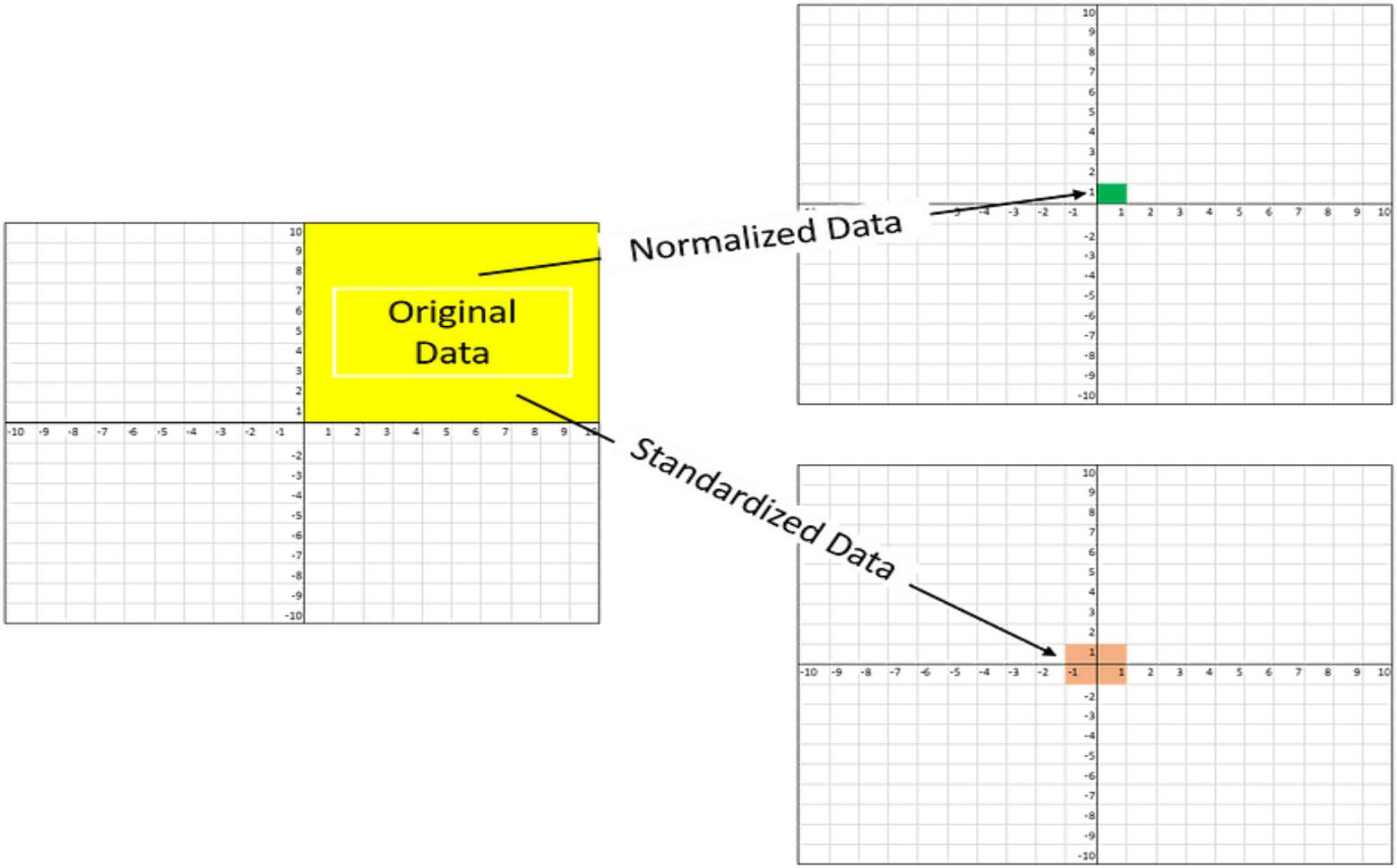
DATA SCIENCE

- To visualize the above, let us take an example of the independent variables of alcohol and Malic Acid content in the wine dataset from the “Wine Dataset” that is deposited on the UCI machine learning repository.
- Below you can see the impact of the two most common scaling techniques (Normalization and Standardization) on the dataset.

Feature Scaling



Methods for Scaling



Methods for Scaling

DATA SCIENCE

Normalization

Also known as min-max scaling or min-max normalization, it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Here, max(x) and min(x) are the maximum and the minimum values of the feature respectively.

We can also do a normalization over different intervals, e.g. choosing to have the variable laying in any [a, b] interval, a and b being real numbers. To rescale a range between an arbitrary set of values [a, b], the formula becomes:

Methods for Scaling

DATA SCIENCE

We can also do a normalization over different intervals, e.g. choosing to have the variable laying in any [a, b] interval, a and b being real numbers. To rescale a range between an arbitrary set of values [a, b], the formula becomes:

$$x' = a + \frac{(x - \min(x))(b - a)}{\max(x) - \min(x)}$$

Methods for Scaling

DATA SCIENCE

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

#	Emp	Age	Salary
1	Emp1	44	73000
2	Emp2	27	47000
3	Emp3	30	53000
4	Emp4	38	62000
5	Emp5	40	57000
6	Emp6	35	53000
7	Emp7	48	78000

Normalization

Age	Normalized Age	Salary	Normalized Salary
44	0.80952381	73000	0.838709677
27	0	47000	0
30	0.142857143	53000	0.193548387
38	0.523809524	62000	0.483870968
40	0.619047619	57000	0.322580645
35	0.380952381	53000	0.193548387
48	1	78000	1
	Range 0-1		**Range 0-1**

How to calculate Normalized value?
X = 35, min = 27, max = 48 for column Age.
Xnorm(for 35) = $\frac{35-27}{48-27} = 0.3809$

After Normalization

Distance between Emp2 and Emp1 = $\sqrt{(0-.80)^2+(0-.83)^2} = 1.15$

Distance between Emp2 and Emp3 = $\sqrt{(.14-0)^2+(.19-0)^2} = 0.23$

Comparison will be more significant

Methods for Scaling

DATA SCIENCE

Standardization

Feature standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point by the following formula:

$$x' = \frac{x - \bar{x}}{\sigma}$$

Here, σ is the standard deviation of the feature vector, and $\bar{x}$ is the average of the feature vector.

Methods for Scaling

DATA SCIENCE

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

#	Emp	Age	Salary
1	Emp1	44	73000
2	Emp2	27	47000
3	Emp3	30	53000
4	Emp4	38	62000
5	Emp5	40	57000
6	Emp6	35	53000
7	Emp7	48	78000
		Mean = 37.42857 Std. Dev. = 6.883876	Mean = 60428.5714 Std. Dev = 10499.7570

Standardization

How to calculate Standardized value?
X = 35, mean = 37.42, Std. Dev. = 6.88
for column Age.
Xstd(for 35) = $\frac{35-37.42}{6.88} = -0.3527$

Age	Standardized Age	Salary	Standardized Salary
44	0.954611636	73000	1.197306616
27	-1.514927162	47000	-1.278941158
30	-1.079126198	53000	-0.707499364
38	0.083009708	62000	0.149663327
40	0.373543684	57000	-0.326538168
35	-0.352791257	53000	-0.707499364
48	1.535679589	78000	1.673508111
	Mean = 0 Std. dev. = 1		**Mean = 0 Std. dev. = 1**

Methods for Scaling

DATA SCIENCE

After Standardization

Distance between Emp2 and Emp1 = $\sqrt{(-1.51 - 0.95)^2 + (-1.27 - 1.19)^2} = 3.47$

Distance between Emp2 and Emp3 = $\sqrt{(-1.07 + 1.51)^2 + (-0.70 + 1.27)^2} = 0.71$

Comparison will be more significant

Normalization vs Standardization

- If you have outliers in your feature (column), normalizing your data will scale most of the data to a small interval, which means all features will have the same scale and hence it will not handle outliers well.
- Standardization is more robust to outliers, and in many cases, it is preferable over Max-Min Normalization.
- Normalization is good to use when your data does not follow a Normal distribution.
- This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
- Standardization, on the other hand, can be helpful in cases where the data follows a Normal distribution.
- However, this does not have to be necessarily true.
- Also, unlike normalization, standardization does not have a bounding range.
- So, even if you have outliers in your data, they will not be affected by standardization.

Types of Feature Scaling

DATA SCIENCE

- Standardization:
- Standardization is also known as z-score Normalization.
- In standardization, features are scaled to have zero-mean and one-standard-deviation.
- It means after standardization features will have mean = 0 and standard deviation = 1.
  - Standard Scaler
- Normalization:
- Normalization is also known as min-max normalization or min-max scaling.
- Normalization re-scales values in the range of 0-1
  - Min Max Scaling
  - Mean Normalization
  - Max Absolute Scaling
  - Robust Scaling etc.

Which Algorithms Need Feature Scaling

DATA SCIENCE

Algorithms	Reasons for applying feature scaling
K-means	Use Euclidean Distance measure
K-nearest neighbours	Measure the distance between pairs of samples and these distances are influenced by the measurement units
Principal Component Analysis (PCA)	Get the features with maximum variance
Artificial Neural Network	Apply gradient descent
Gradient Descent	Theta calculation becomes faster after feature scaling and the learning rate in the update equation of Stochastic gradient descent is the same for every parameter.

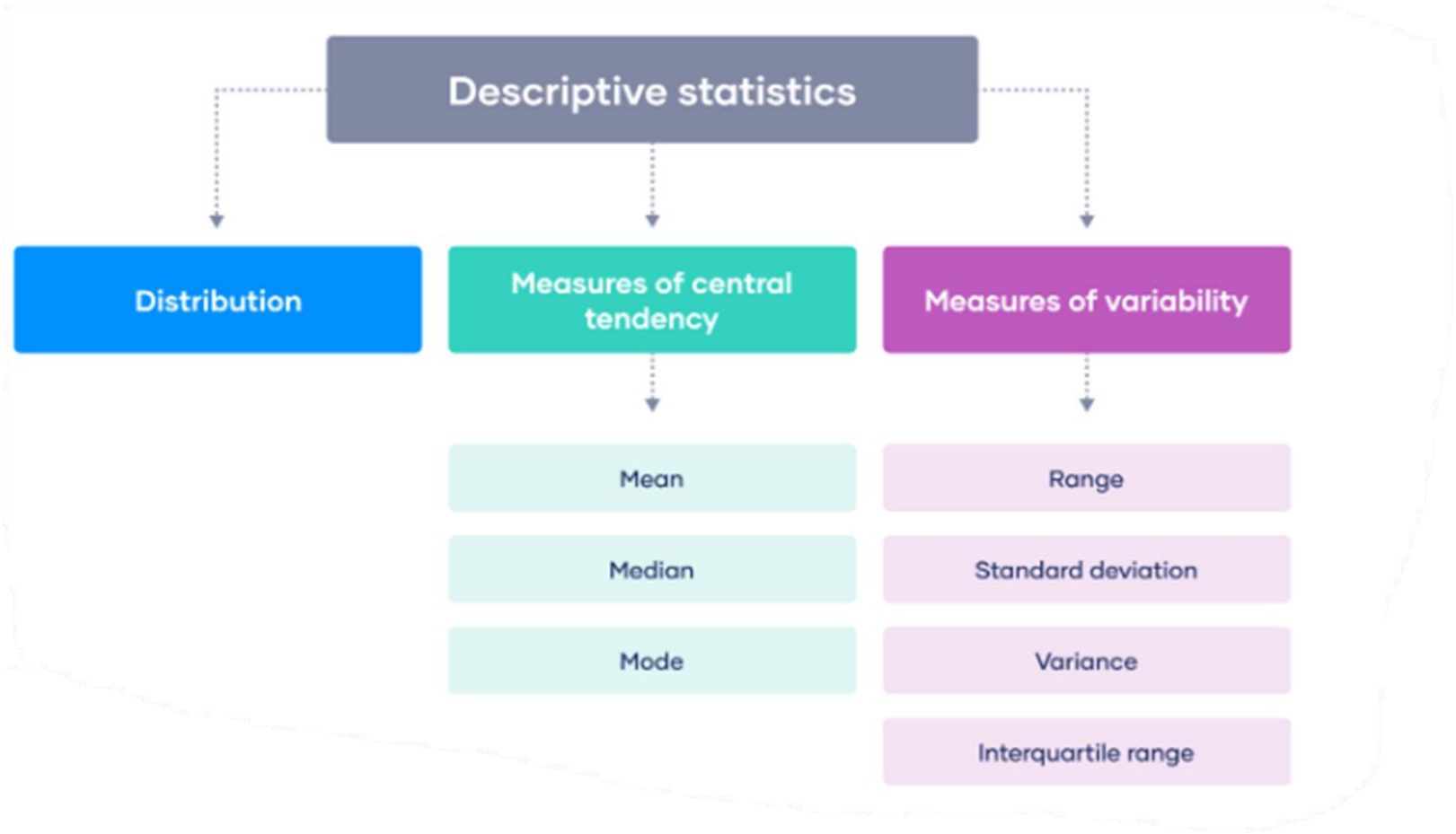
Descriptive Statistics

DATA SCIENCE

- Descriptive statistics summarize and organize characteristics of a data set.
- A data set is a collection of responses or observations from a sample or entire population.
- In quantitative research, after collecting data, the first step of statistical analysis is to describe characteristics of the responses, such as the average of one variable (e.g., age), or the relation between two variables (e.g., age and creativity).
- The next step is inferential statistics, which help you decide whether your data confirms or refutes your hypothesis and whether it is generalizable to a larger population.

Types of descriptive statistics

DATA SCIENCE



Frequency distribution

DATA SCIENCE

A data set is made up of a distribution of values, or scores. In tables or graphs, you can summarize the frequency of every possible value of a variable in numbers or percentages. This is called a **frequency distribution**.

Simple frequency distribution table | **Grouped frequency distribution table**

For the variable of gender, you list all possible answers on the left hand column. You count the number or percentage of responses for each answer and display it on the right hand column.

Gender	Number
Male	182
Female	235
Other	27

From this table, you can see that more women than men or people with another gender identity took part in the study.

Measures of central tendency

DATA SCIENCE

Measures of central tendency estimate the center, or average, of a data set. The mean, median and mode are 3 ways of finding the average.

Here we will demonstrate how to calculate the mean, median, and mode using the first 6 responses of our survey.

Mean **Median** **Mode**

The **mean**, or *M*, is the most commonly used method for finding the average.

To find the mean, simply add up all response values and divide the sum by the total number of responses. The total number of responses or observations is called *N*.

Mean number of library visits

Data set	15, 3, 12, 0, 24, 3
Sum of all values	15 + 3 + 12 + 0 + 24 + 3 = 57
Total number of responses	*N* = 6
Mean	Divide the sum of values by *N* to find *M*: 57/6 = **9.5**

Measures of central tendency

Measures of central tendency estimate the center, or average, of a data set. The mean, median and mode are 3 ways of finding the average.

Here we will demonstrate how to calculate the mean, median, and mode using the first 6 responses of our survey.

Mean **Median** **Mode**

The **mean**, or *M*, is the most commonly used method for finding the average.

To find the mean, simply add up all response values and divide the sum by the total number of responses. The total number of responses or observations is called *N*.

Mean number of library visits

Data set	15, 3, 12, 0, 24, 3
Sum of all values	15 + 3 + 12 + 0 + 24 + 3 = 57
Total number of responses	*N* = 6
Mean	Divide the sum of values by *N* to find *M*: 57/6 = **9.5**

Measures of variability

DATA SCIENCE

Measures of variability give you a sense of how spread out the response values are. The range, standard deviation and variance each reflect different aspects of spread.

Range

The range gives you an idea of how far apart the most extreme response scores are. To find the range, simply subtract the lowest value from the highest value.

Range of visits to the library in the past year

Ordered data set: 0, 3, 3, 12, 15, 24

Range: 24 − 0 = **24**

Standard deviation

DATA SCIENCE

- The standard deviation (s or SD) is the average amount of variability in our dataset.
- It tells us, on average, how far each score lies from the mean.
- The larger the standard deviation, the more variable the data set is.
- There are six steps for finding the standard deviation:
  - List each score and find their mean.
  - Subtract the mean from each score to get the deviation from the mean.
  - Square each of these deviations.
  - Add up all of the squared deviations.
  - Divide the sum of the squared deviations by N – 1.
  - Find the square root of the number you found.

Standard deviation

DATA SCIENCE

Standard deviations of visits to the library in the past year

In the table below, you complete **Steps 1 through 4**.

Raw data	Deviation from mean	Squared deviation
15	15 – 9.5 = 5.5	30.25
3	3 – 9.5 = -6.5	42.25
12	12 – 9.5 = 2.5	6.25
0	0 – 9.5 = -9.5	90.25
24	24 – 9.5 = 14.5	210.25
3	3 – 9.5 = -6.5	42.25
M = 9.5	Sum = 0	Sum of squares = 421.5

Step 5: 421.5/5 = 84.3

Step 6: √84.3 = 9.18

From learning that s = **9.18**, you can say that on average, each score deviates from the mean by 9.18 points.

Variance

DATA SCIENCE

Variance

The variance is the average of squared deviations from the mean. Variance reflects the degree of spread in the data set. The more spread the data, the larger the variance is in relation to the mean.

To find the variance, simply square the standard deviation. The symbol for variance is s^2.

Variance of visits to the library in the past year

Data set: 15, 3, 12, 0, 24, 3

$s = 9.18$

$s^2 = \mathbf{84.3}$

Univariate descriptive statistics

Univariate descriptive statistics focus on only one variable at a time. It's important to examine data from each variable separately using multiple measures of distribution, central tendency and spread. Programs like SPSS and Excel can be used to easily calculate these.

	Visits to the library
N	6
Mean	9.5
Median	7.5
Mode	3
Standard deviation	9.18
Variance	84.3
Range	24

If you were to only consider the mean as a measure of central tendency, your impression of the "middle" of the data set can be skewed by outliers, unlike the median or mode.

Likewise, while the range is sensitive to outliers, you should also consider the standard deviation and variance to get easily comparable measures of spread.

Bivariate descriptive statistics

- If we've collected data on more than one variable, we can use bivariate or multivariate descriptive statistics to explore whether there are relationships between them.
- In bivariate analysis, we simultaneously study the frequency and variability of two variables to see if they vary together.
- We can also compare the central tendency of the two variables before performing further statistical tests.
- Multivariate analysis is the same as bivariate analysis but with more than two variables.

Contingency table

DATA SCIENCE

In a contingency table, each cell represents the intersection of two variables. Usually, an independent variable (e.g., gender) appears along the vertical axis and a dependent one appears along the horizontal axis (e.g., activities). You read "across" the table to see how the independent and dependent variables relate to each other.

	Number of visits to the library in the past year				
Group	0–4	5–8	9–12	13–16	17+
Children	32	68	37	23	22
Adults	36	48	43	83	25

Interpreting a contingency table is easier when the raw data is converted to percentages. Percentages make each row comparable to the other by making it seem as if each group had only 100 observations or participants. When creating a percentage-based contingency table, you add the *N* for each independent variable on the end.

Contingency table

DATA SCIENCE

Visits to the library in the past year (Percentages)						
Group	0–4	5–8	9–12	13–16	17+	*N*
Children	18%	37%	20%	13%	12%	182
Adults	15%	20%	18%	35%	11%	235

From this table, it is more clear that similar proportions of children and adults go to the library over 17 times a year. Additionally, children most commonly went to the library between 5 and 8 times, while for adults, this number was between 13 and 16.

What is Data Imputation

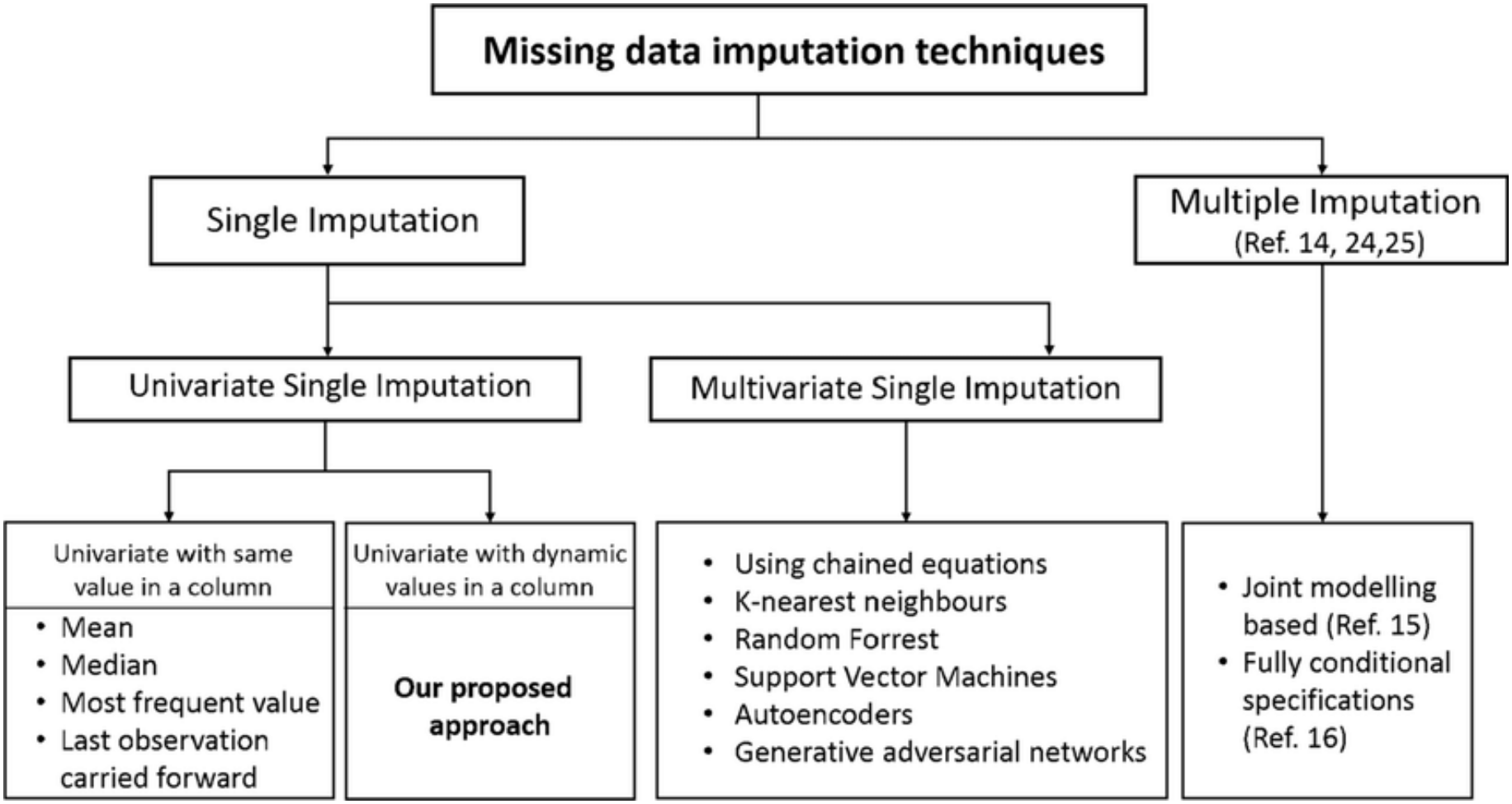
DATA SCIENCE

- Data imputation is a method for retaining the dataset's data and information by substituting missing data with a different value.
- These methods are employed because it would be impractical to remove data from a dataset each time.
- Additionally, doing so would substantially reduce the dataset's size, raising questions about bias and impairing analysis.

Why Data Imputation

- Distorts Dataset: Large amounts of missing data can lead to anomalies in the variable distribution, which can change the relative importance of different categories in the dataset.
- Unable to work with machine learning-related Python libraries: When utilizing ML libraries (SkLearn is the most popular), mistakes may occur because there is no automatic handling of these missing data.
- Impacts on the Final Model: Missing data may lead to bias in the dataset, which could affect the final model's analysis.
- Desire to restore the entire dataset: This typically occurs when we don't want to lose any of the data in our dataset because all of it is crucial. Additionally, while the dataset is not very large, eliminating a portion of it could have a substantial effect on the final model.

Missing Data Imputation Techniques



Mark Missing Values

DATA SCIENCE

```
# load and summarize the dataset
from pandas import read_csv
# load the dataset
dataset = read_csv('pima-indians-diabetes.csv', header=None)
# summarize the dataset
print(dataset.describe())
```

Running this example produces the following output:

```
                0           1           2  ...           6           7           8
count  768.000000  768.000000  768.000000  ...  768.000000  768.000000  768.000000
mean     3.845052  120.894531   69.105469  ...    0.471876   33.240885    0.348958
std      3.369578   31.972618   19.355807  ...    0.331329   11.760232    0.476951
min      0.000000    0.000000    0.000000  ...    0.078000   21.000000    0.000000
25%      1.000000   99.000000   62.000000  ...    0.243750   24.000000    0.000000
50%      3.000000  117.000000   72.000000  ...    0.372500   29.000000    0.000000
75%      6.000000  140.250000   80.000000  ...    0.626250   41.000000    1.000000
max     17.000000  199.000000  122.000000  ...    2.420000   81.000000    1.000000

[8 rows x 9 columns]
```

Mark Missing Values

DATA SCIENCE

```
# load the dataset and review rows
from pandas import read_csv
# load the dataset
dataset = read_csv('pima-indians-diabetes.csv', header=None)
# print the first 20 rows of data
print(dataset.head(20))
```

Mark Missing Values

Running the example, we can clearly see 0 values in the columns 2, 3, 4, and 5.

		0	1	2	3	4	5	6	7	8
2	0	6	148	72	35	0	33.6	0.627	50	1
3	1	1	85	66	29	0	26.6	0.351	31	0
4	2	8	183	64	0	0	23.3	0.672	32	1
5	3	1	89	66	23	94	28.1	0.167	21	0
6	4	0	137	40	35	168	43.1	2.288	33	1
7	5	5	116	74	0	0	25.6	0.201	30	0
8	6	3	78	50	32	88	31.0	0.248	26	1
9	7	10	115	0	0	0	35.3	0.134	29	0
10	8	2	197	70	45	543	30.5	0.158	53	1
11	9	8	125	96	0	0	0.0	0.232	54	1
12	10	4	110	92	0	0	37.6	0.191	30	0
13	11	10	168	74	0	0	38.0	0.537	34	1
14	12	10	139	80	0	0	27.1	1.441	57	0
15	13	1	189	60	23	846	30.1	0.398	59	1
16	14	5	166	72	19	175	25.8	0.587	51	1
17	15	7	100	0	0	0	30.0	0.484	32	1
18	16	0	118	84	47	230	45.8	0.551	31	1
19	17	7	107	74	0	0	29.6	0.254	31	1
20	18	1	103	30	38	83	43.3	0.183	33	0
21	19	1	115	70	30	96	34.6	0.529	32	1

Count Column wise Missing Values

DATA SCIENCE

```
# example of summarizing the number of missing values for each variable
from pandas import read_csv
# load the dataset
dataset = read_csv('pima-indians-diabetes.csv', header=None)
# count the number of missing values for each column
num_missing = (dataset[[1,2,3,4,5]] == 0).sum()
# report the results
print(num_missing)
```

Running the example prints the following output:

```
1      5
2     35
3    227
4    374
5     11
```

Replace Missing Values

Pandas provides the fillna() function for replacing missing values with a specific value.

For example, we can use fillna() to replace missing values with the mean value for each column, as follows:

```
# manually impute missing values with numpy
from pandas import read_csv
from numpy import nan
# load the dataset
dataset = read_csv('pima-indians-diabetes.csv', header=None)
# mark zero values as missing or NaN
dataset[[1,2,3,4,5]] = dataset[[1,2,3,4,5]].replace(0, nan)
# fill missing values with mean column values
dataset.fillna(dataset.mean(), inplace=True)
# count the number of NaN values in each column
print(dataset.isnull().sum())
```

Running the example provides a count of the number of missing values in each column, showing zero missing values.

```
0    0
1    0
2    0
3    0
4    0
5    0
6    0
7    0
8    0
dtype: int64
```

Replace Missing Values

Pandas provides the fillna() function for replacing missing values with a specific value.

For example, we can use fillna() to replace missing values with the mean value for each column, as follows:

```
# manually impute missing values with numpy
from pandas import read_csv
from numpy import nan
# load the dataset
dataset = read_csv('pima-indians-diabetes.csv', header=None)
# mark zero values as missing or NaN
dataset[[1,2,3,4,5]] = dataset[[1,2,3,4,5]].replace(0, nan)
# fill missing values with mean column values
dataset.fillna(dataset.mean(), inplace=True)
# count the number of NaN values in each column
print(dataset.isnull().sum())
```

Running the example provides a count of the number of missing values in each column, showing zero missing values.

```
0    0
1    0
2    0
3    0
4    0
5    0
6    0
7    0
8    0
dtype: int64
```

Replace Missing Values

The example below uses the SimpleImputer class to replace missing values with the mean of each column then prints the number of NaN values in the transformed matrix.

```
# example of imputing missing values using scikit-learn
from numpy import nan
from numpy import isnan
from pandas import read_csv
from sklearn.impute import SimpleImputer
# load the dataset
dataset = read_csv('pima-indians-diabetes.csv', header=None)
# mark zero values as missing or NaN
dataset[[1,2,3,4,5]] = dataset[[1,2,3,4,5]].replace(0, nan)
# retrieve the numpy array
values = dataset.values
# define the imputer
imputer = SimpleImputer(missing_values=nan, strategy='mean')
# transform the dataset
transformed_values = imputer.fit_transform(values)
# count the number of NaN values in each column
print(f'Missing: {isnan(transformed_values).sum()}')
```

Running the example shows that all NaN values were imputed successfully.

```
Missing: 0
```

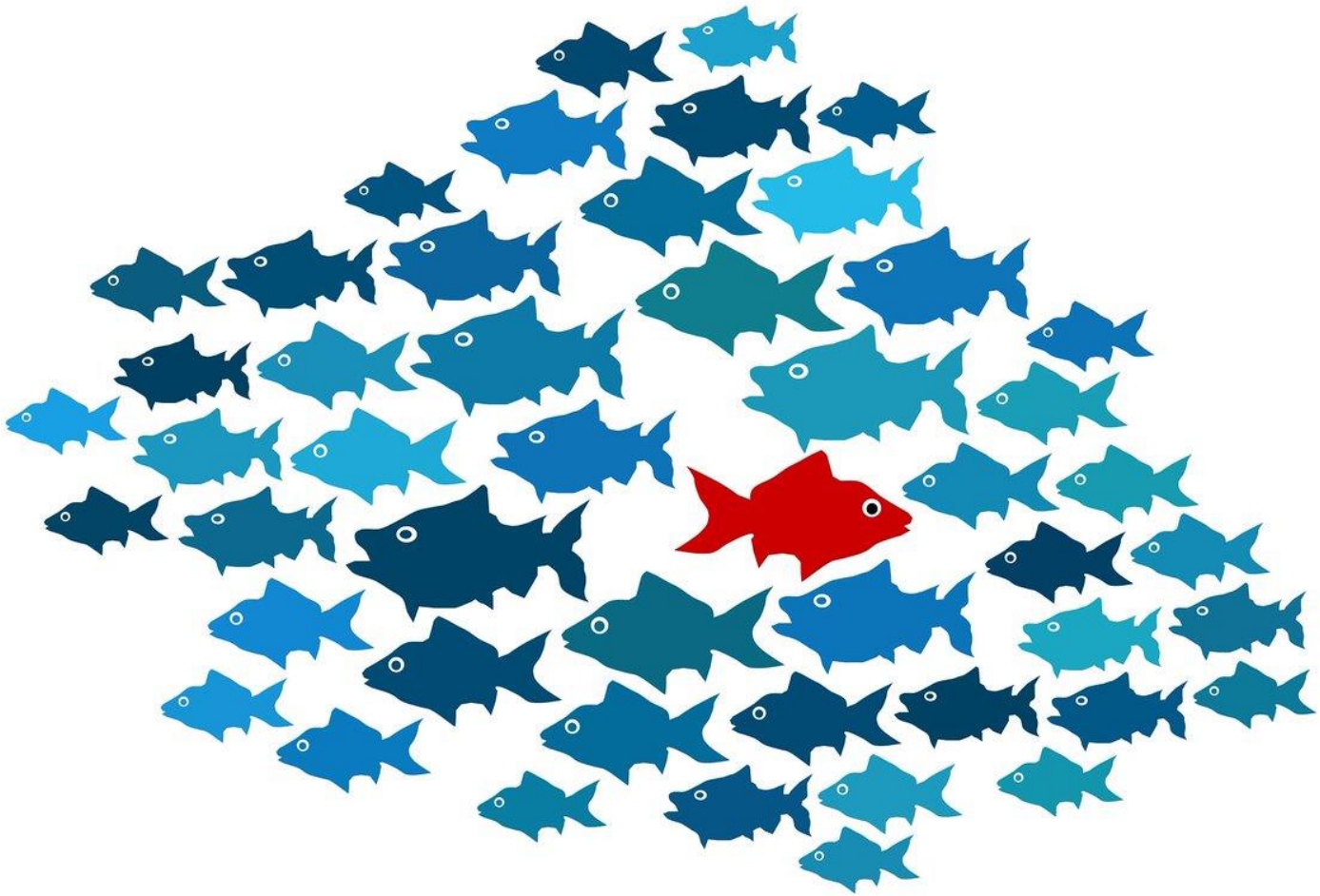
Outlier/Anomaly Detection

DATA SCIENCE

- Anomaly detection, sometimes called outlier detection, is a process of finding patterns or instances in a dataset that deviate significantly from the expected or “normal behavior.”
- The definition of both “normal” and anomalous data significantly varies depending on the context.

Outlier/Anomaly Detection

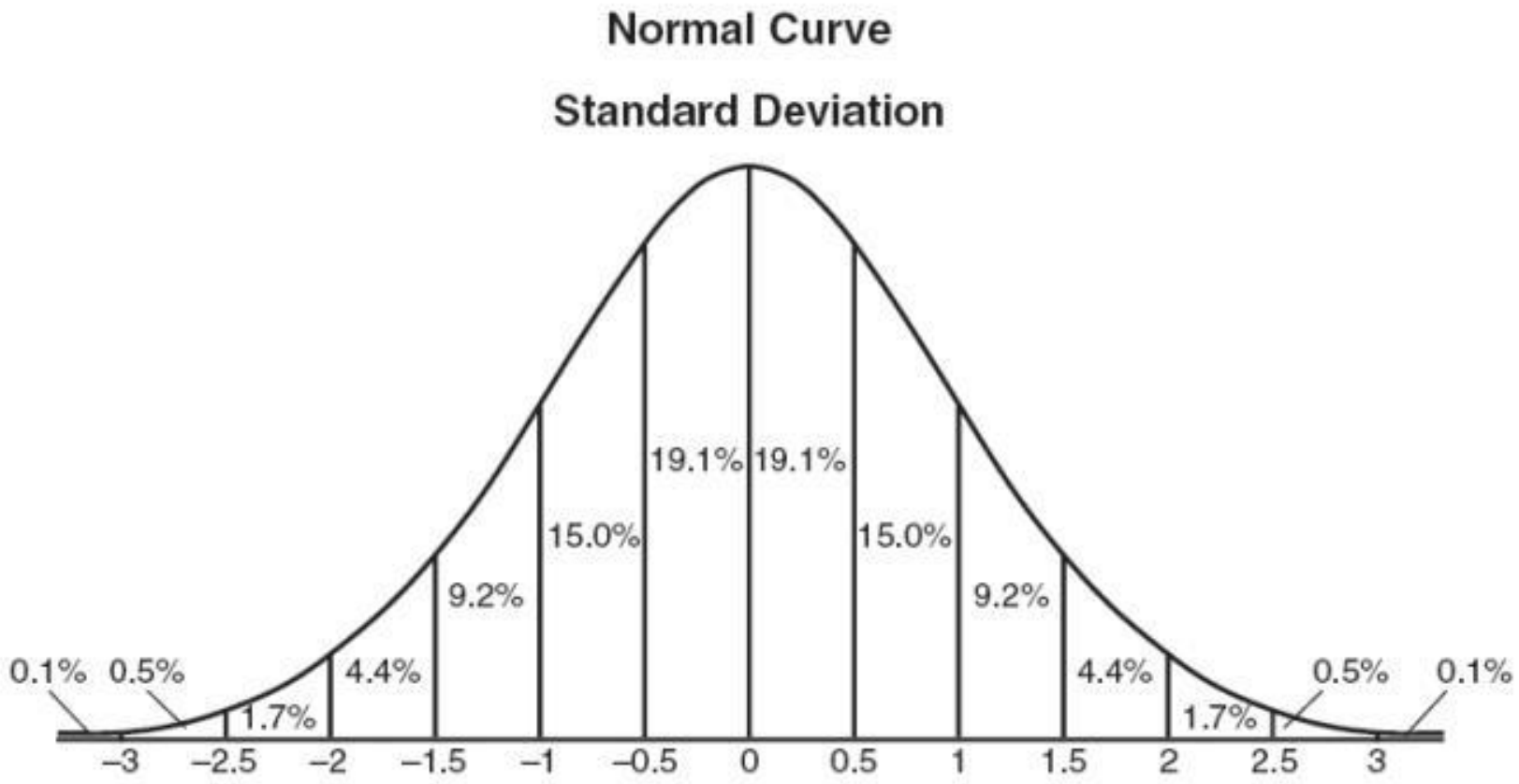
DATA SCIENCE



Outlier/Anomaly Detection

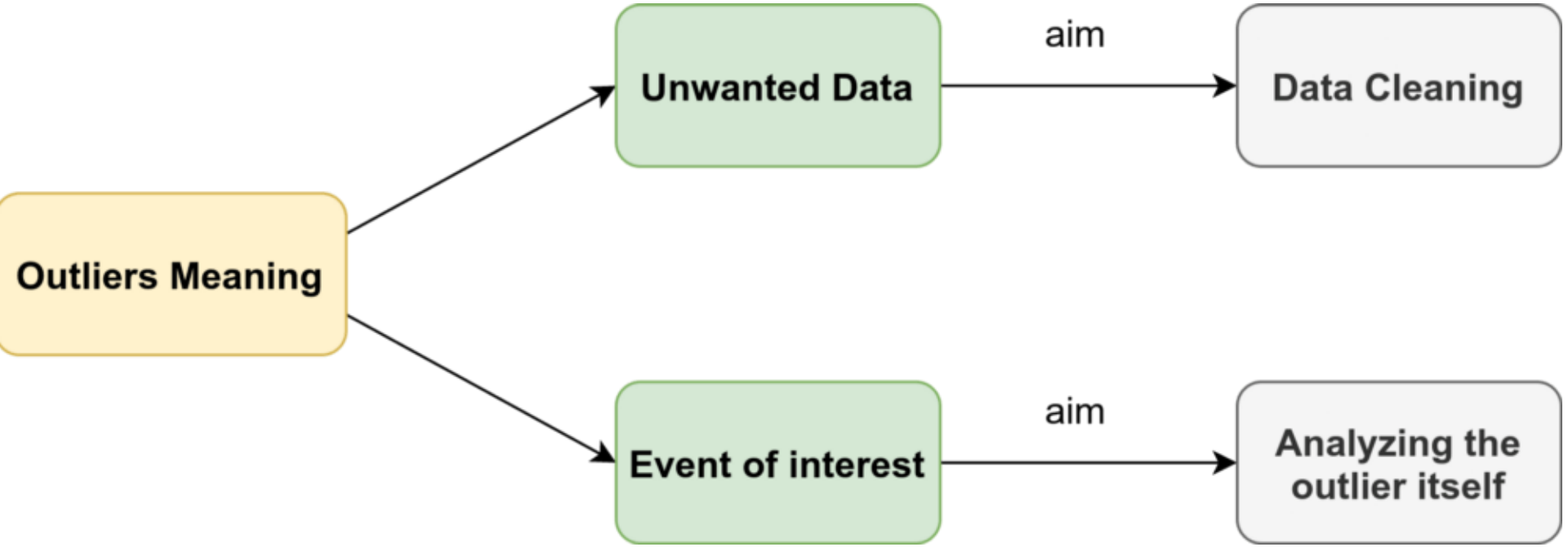
DATA SCIENCE

The data points which fall below mean-3(sigma) or above mean+3(sigma) are outliers.



Outlier/Anomaly Detection

- An observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism.



Outlier/Anomaly Detection

- Financial transactions
  - Normal: Routine purchases and consistent spending by an individual in London.
  - Outlier: A massive withdrawal from Ireland from the same account, hinting at potential fraud.

Outlier/Anomaly Detection

DATA SCIENCE

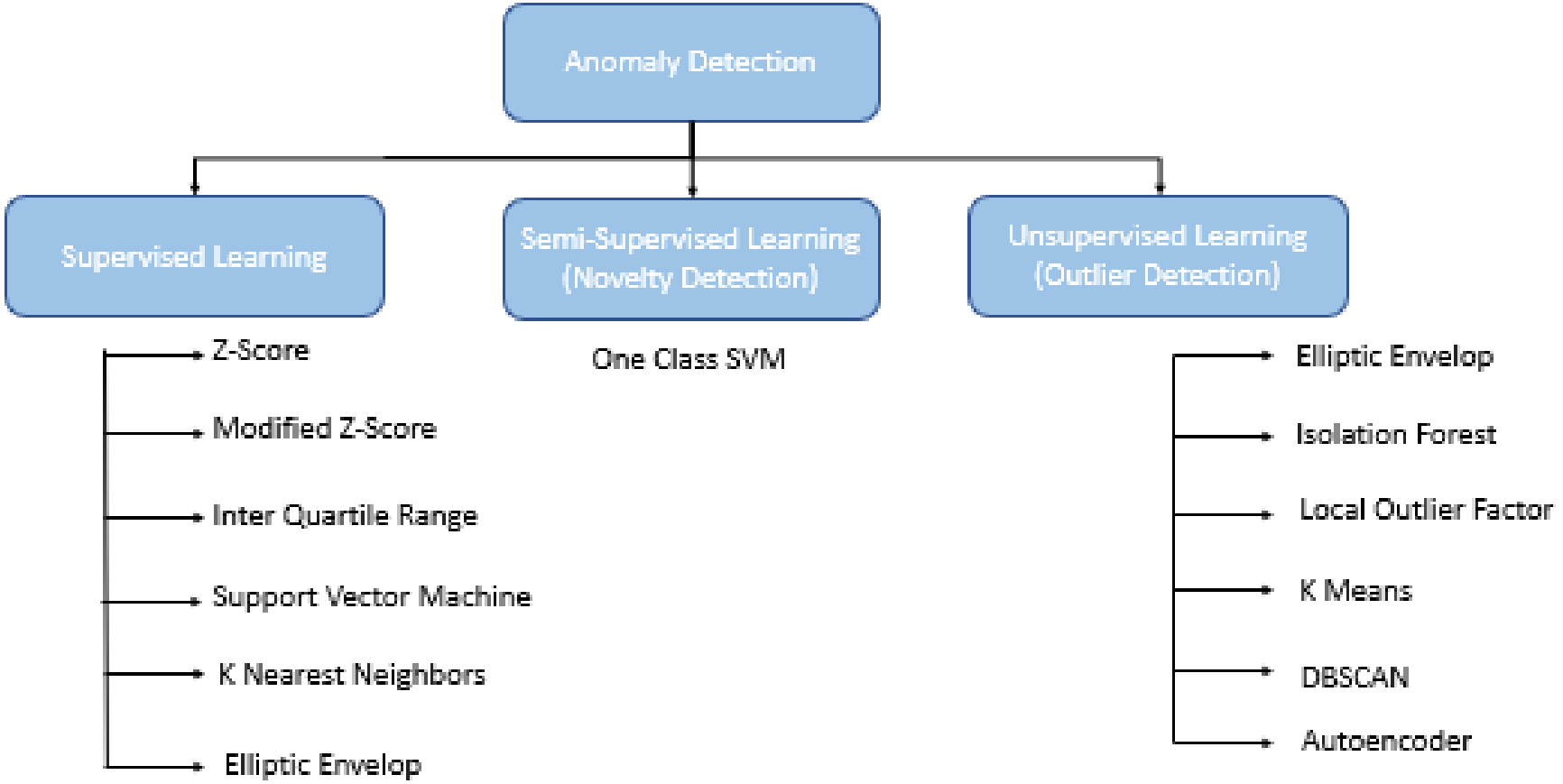
- Network traffic in cybersecurity
  - Normal: Regular communication, steady data transfer, and adherence to protocol.
  - Outlier: Abrupt increase in data transfer or use of unknown protocols signaling a potential breach or malware.

Outlier/Anomaly Detection

DATA SCIENCE

- Patient vital signs monitoring
  - Normal: Stable heart rate and consistent blood pressure
  - Outlier: Sudden increase in heart rate and decrease in blood pressure, indicating a potential emergency or equipment failure.

Types of Outliers



Types of Outliers

DATA SCIENCE

- Univariate outliers exist in a single variable or feature in isolation.
- Univariate outliers are extreme or abnormal values that deviate from the typical range of values for that specific feature.
- Multivariate outliers are found by combining the values of multiple variables at the same time.

Types of Outliers

DATA SCIENCE

- For example, consider a dataset of housing prices in a neighborhood.
- Most houses cost between $200,000 and $400,000, but there is House A with an exceptionally high price of $1,000,000.
- When we analyze only the price, House A is a clear outlier.

Types of Outliers

DATA SCIENCE

- Now, let's add two more variables to our dataset: the square footage and the number of bedrooms.
- When we consider the square footage, the number of bedrooms, and the price, it's House B that looks odd:
  - It has half the square footage of the mean house price.
  - It has only one bedroom.
  - It costs the top of the range $380.000.
- When we look at these variables individually, they seem ordinary.
- Only when we put them together do we find out that House B is a clear multivariate outlier.

Anomaly Detection Methods And When to Use Each One

- For univariate outlier detection, the most popular methods are:
  - Z-score (standard score): the z-score measures how many standard deviations a data point is away from the mean.
  - Generally, instances with a z-score over 3 are chosen as outliers.
  - Interquartile range (IQR): The IQR is the range between the first quartile (Q1) and the third quartile (Q3) of a distribution.
  - When an instance is beyond Q1 or Q3 for some multiplier of IQR, they are considered outliers. The most common multiplier is 1.5, making the outlier range [Q1–1.5 * IQR, Q3 + 1.5 * IQR].

Anomaly Detection Methods And When to Use Each One

- Modified z-scores: like z-scores, but modified z-scores use the median and a measure called Median Absolute Deviation (MAD) to find outliers.
- Since mean and standard deviation are easily skewed by outliers, modified z-scores are generally considered more robust.

Z-Score

$$Z = (X - \mu)\sigma$$

where,

- X=a single raw data value
- μ = population mean
- σ = population standard deviation

Refer AnamolyOutliers.py

Real-World Applications of Anomaly Detection

- Cybersecurity
- Healthcare
- Industrial equipment monitoring
- Network intrusion detection
- Energy grid monitoring
- E-commerce and user behavior analysis
- Quality control in manufacturing

Data Visualization

DATA SCIENCE

- Data visualization provides a good, organized pictorial representation of the data which makes it easier to understand, observe, analyze.

Data Visualization Libraries

- Matplotlib: low level, provides lots of freedom
- Pandas Visualization: easy to use interface, built on Matplotlib
- Seaborn: high-level interface, great default styles
- Plotnine: based on R's ggplot2, uses Grammar of Graphics
- Plotly: can create interactive plots

Feature Selection

DATA SCIENCE

- Feature selection is the process of reducing the number of input variables when developing a predictive model.
- It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model.
- Statistical-based feature selection methods involve evaluating the relationship between each input variable and the target variable using statistics.
- Selecting those input variables that have the strongest relationship with the target variable

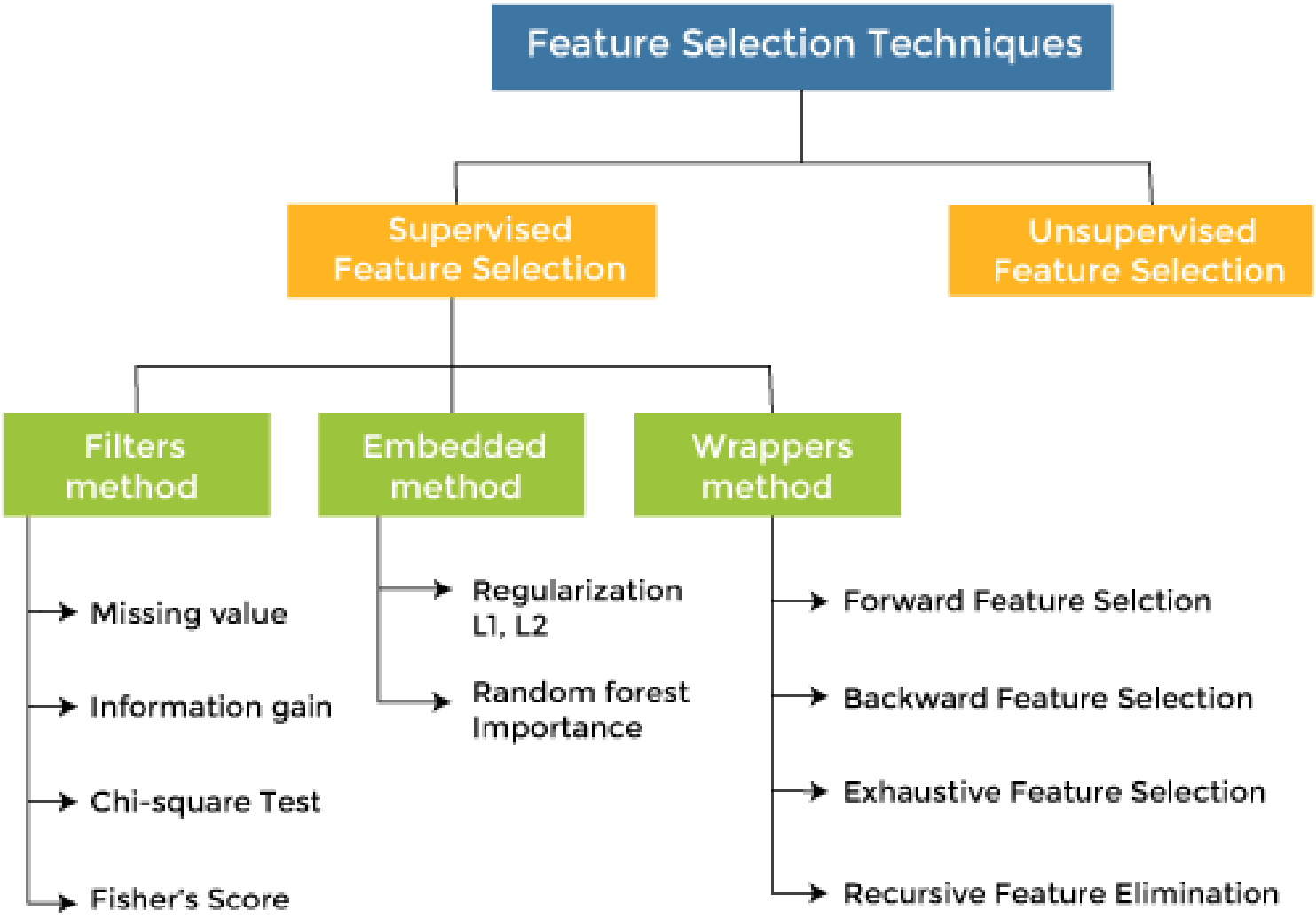
Feature Selection

DATA SCIENCE

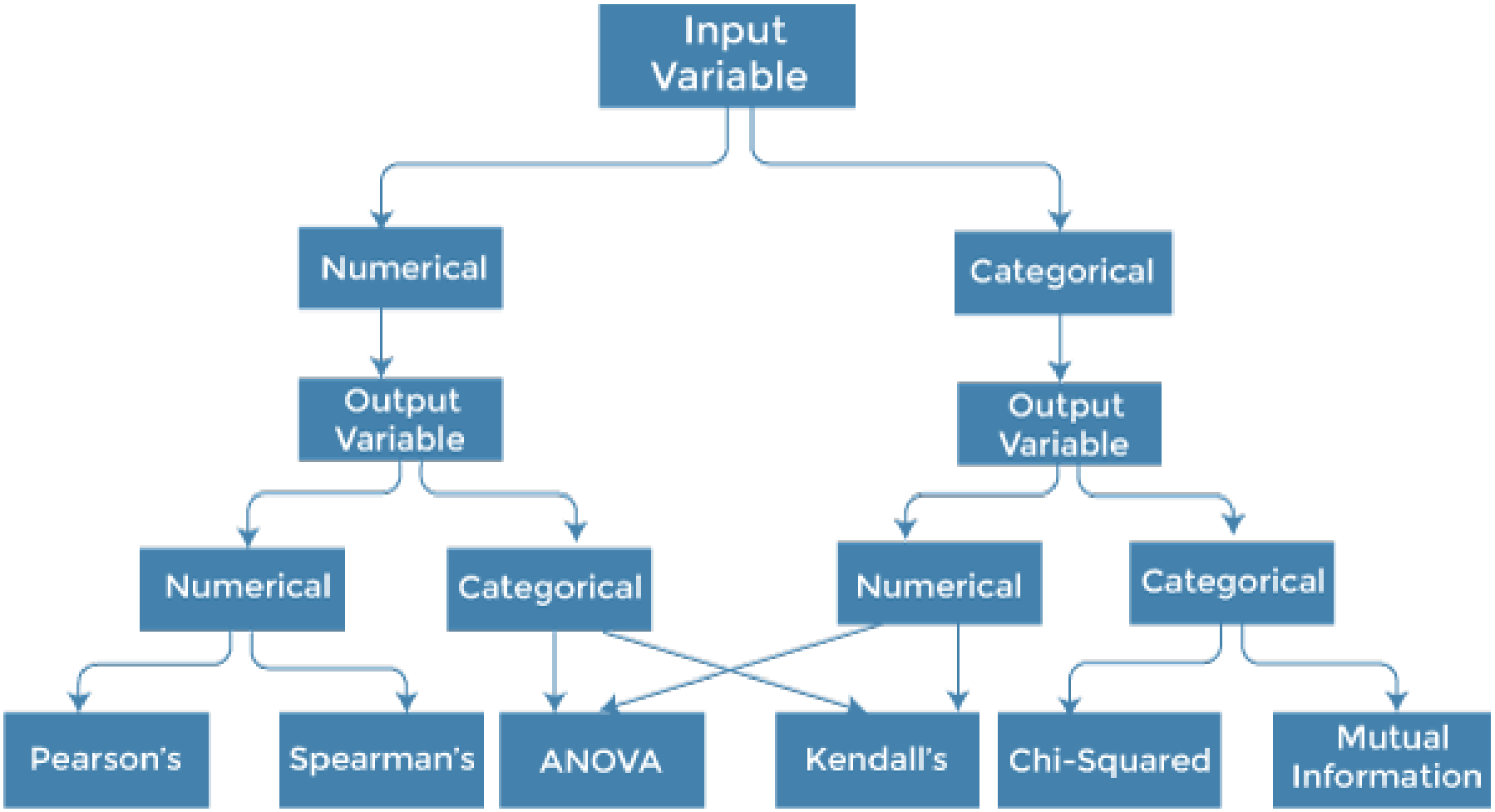
- These methods can be fast and effective, although the choice of statistical measures depends on the data type of both the input and output variables.

Feature Selection Techniques

DATA SCIENCE



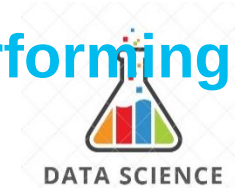
Statistics for Filter-Based Feature Selection Methods



Statistics for Filter-Based Feature Selection Methods

Input Variable	Output Variable	Feature Selection technique
Numerical	Numerical	○ Pearson's correlation coefficient (For linear Correlation). ○ Spearman's rank coefficient (for non-linear correlation).
Numerical	Categorical	○ ANOVA correlation coefficient (linear). ○ Kendall's rank coefficient (nonlinear).
Categorical	Numerical	○ Kendall's rank coefficient (linear). ○ ANOVA correlation coefficient (nonlinear).
Categorical	Categorical	○ Chi-Squared test (contingency tables). ○ Mutual Information.

How to select features and what are Benefits of performing feature selection before modeling your data?



- Reduces Overfitting: Less redundant data means less opportunity to make decisions based on noise.
- Improves Accuracy: Less misleading data means modeling accuracy improves.
- Reduces Training Time: fewer data points reduce algorithm complexity and algorithms train faster.

Univariate Selection

- Statistical tests can be used to select those features that have the strongest relationship with the output variable.
- The scikit-learn library provides the SelectKBest class that can be used with a suite of different statistical tests to select a specific number of features.
- Many different statistical test scan be used with this selection method.
- For example the ANOVA F-value method is appropriate for numerical inputs and categorical data, as we see in the Pima dataset.

Feature Importance

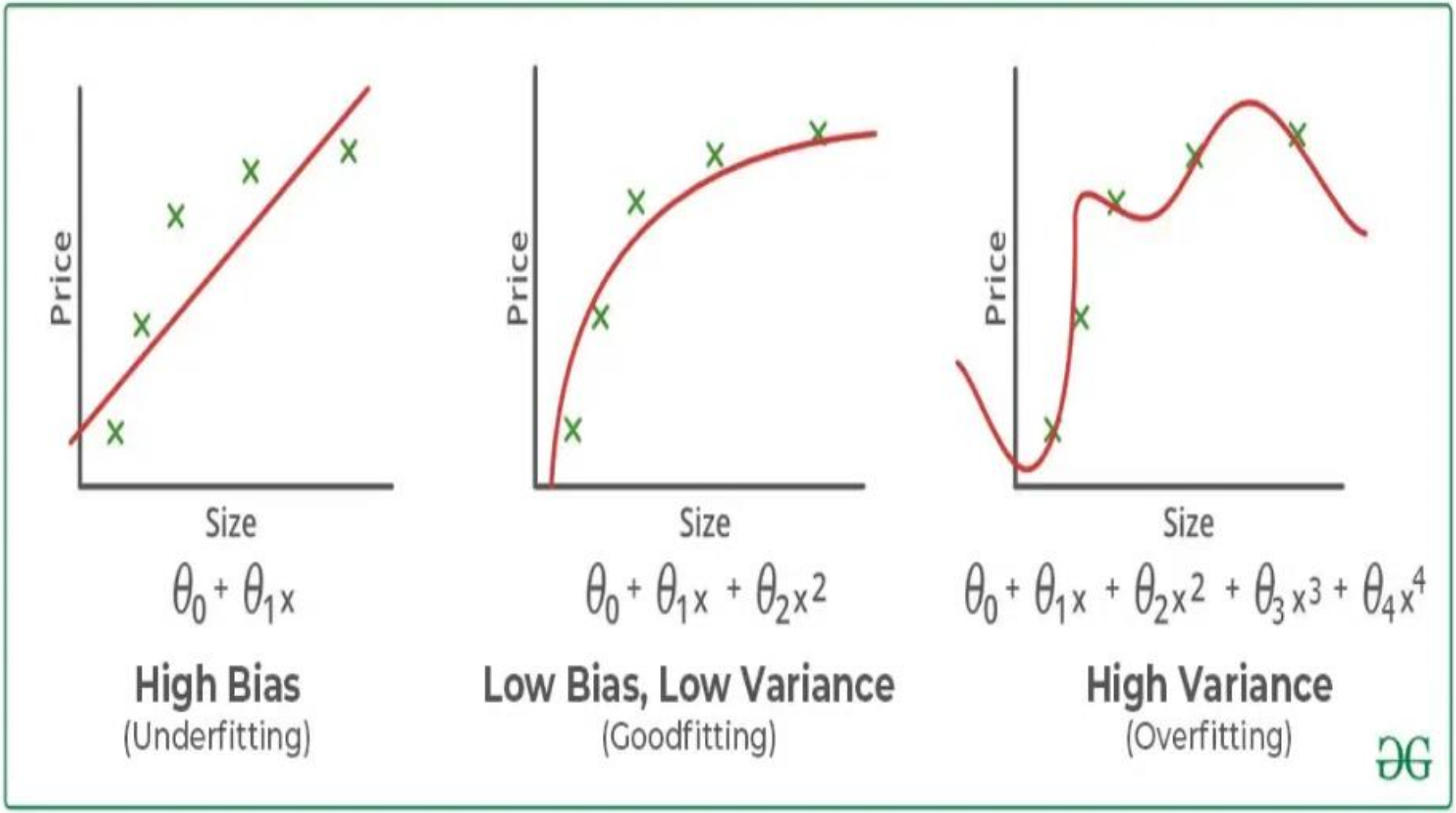
DATA SCIENCE

- We can get the feature importance of each feature of your dataset by using the feature importance property of the model.
- Feature importance gives us a score for each feature of our data, the higher the score more important or relevant is the feature towards your output variable.
- Feature importance is an inbuilt class that comes with Tree Based Classifiers, we will be using Extra Tree Classifier for extracting the top 10 features for the dataset.

Correlation Matrix with Heatmap

- Correlation states how the features are related to each other or the target variable.
- Correlation can be positive (increase in one value of feature increases the value of the target variable) or negative (increase in one value of feature decreases the value of the target variable)
- Heatmap makes it easy to identify which features are most related to the target variable, we will plot heatmap of correlated features using the seaborn library.

Under Fitting and Over Fitting



Under fitting

DATA SCIENCE

- A statistical model or a machine learning algorithm is said to have underfitting when a model is too simple to capture data complexities.
- It represents the inability of the model to learn the training data effectively result in poor performance both on the training and testing data.
- In simple terms, an underfit model's are inaccurate, especially when applied to new, unseen examples.
- It mainly happens when we uses very simple model with overly simplified assumptions.
- To address underfitting problem of the model, we need to use more complex models, with enhanced feature representation, and less regularization.

Reasons For Under fitting

DATA SCIENCE

- The model is too simple, So it may be not capable to represent the complexities in the data.
- The input features which is used to train the model is not the adequate representations of underlying factors influencing the target variable.
- The size of the training dataset used is not enough.
- Excessive regularization are used to prevent the overfitting, which constraint the model to capture the data well.
- Features are not scaled.

Techniques to Reduce Underfitting

DATA SCIENCE

- Increase model complexity.
- Increase the number of features, performing feature engineering.
- Remove noise from the data.
- Increase the number of epochs or increase the duration of training to get better results.

Overfitting in Machine Learning

- A statistical model is said to be overfitted when the model does not make accurate predictions on testing data.
- When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set.
- when testing with test data results in High variance.
- Then the model does not categorize the data correctly, because of too many details and noise.
- The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models.
- A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.

Reasons for Overfitting:

DATA SCIENCE

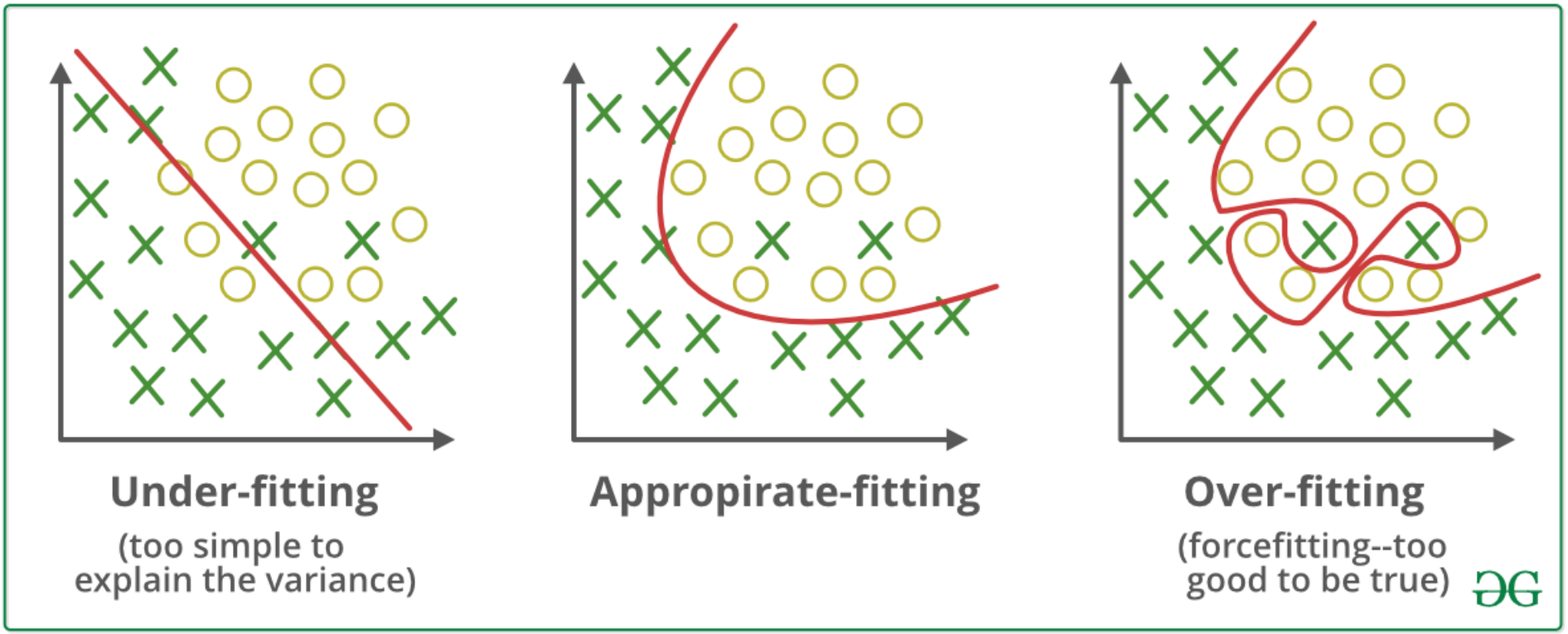
- High variance and low bias.
- The model is too complex.
- The size of the training data.

Techniques to Reduce Overfitting

- Increase training data.
- Reduce model complexity.
- Early stopping during the training phase (have an eye over the loss over the training period as soon as loss begins to increase stop training).
- Ridge Regularization and Lasso Regularization.
- Use dropout for neural networks to tackle overfitting.

Under fitting and Over fitting

DATA SCIENCE



Good Fit in a Statistical Model

- Ideally, the case when the model makes the predictions with 0 error, is said to have a good fit on the data.
- This situation is achievable at a spot between overfitting and underfitting.
- With the passage of time, our model will keep on learning, and thus the error for the model on the training and testing data will keep on decreasing.
- If it will learn for too long, the model will become more prone to overfitting due to the presence of noise and less useful details.
- Hence the performance of our model will decrease.
- To get a good fit, we will stop at a point just before where the error starts increasing.
- At this point, the model is said to have good skills in training datasets as well as our unseen testing dataset.

Bias-Variance Trade-off

DATA SCIENCE

- Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.
- Model with high bias pays very little attention to the training data and oversimplifies the model.
- It always leads to high error on training and test data.

Bias-Variance Trade-off

DATA SCIENCE

- Variance is the variability of model prediction for a given data point or a value which tells us spread of our data.
- Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn’t seen before.
- As a result, such models perform very well on training data but has high error rates on test data.

Mathematically

DATA SCIENCE

- Let the variable we are trying to predict as Y and other covariates as X. We assume there is a relationship between the two such that
- $Y=f(X) + e$
- Where e is the error term and it's normally distributed with a mean of 0.
- We will make a model f^(X) of f(X) using linear regression or any other modeling technique.

Mathematically

DATA SCIENCE

So the expected squared error at a point x is

$$Err(x) = E\left[(Y - \hat{f}(x))^2\right]$$

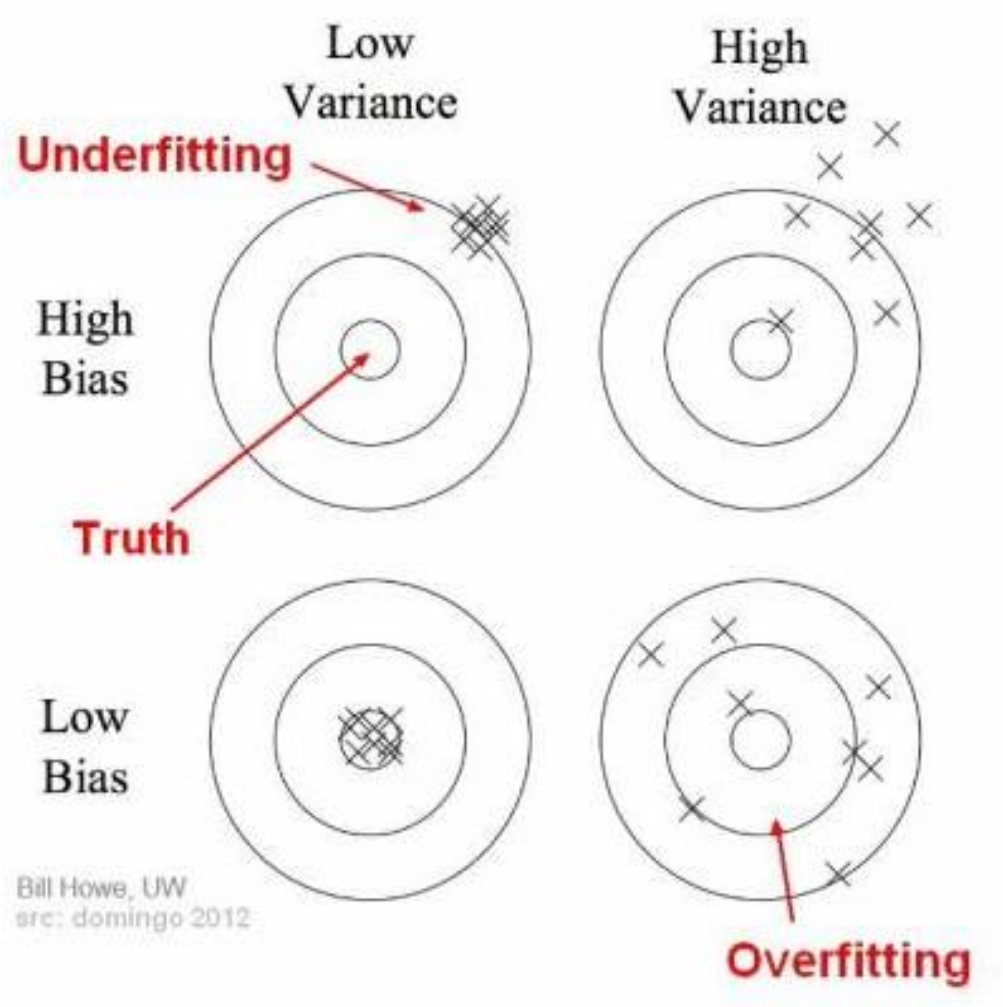
The Err(x) can be further decomposed as

$$Err(x) = \left(E[\hat{f}(x)] - f(x)\right)^2 + E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right] + \sigma_e^2$$

$$Err(x) = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

Err(x) is the sum of Bias^2, variance and the irreducible error.

Mathematically



Why is Bias Variance Tradeoff?

DATA SCIENCE

- If our model is too simple and has very few parameters, then it may have high bias and low variance.
- On the other hand, if our model has large number of parameters then it’s going to have high variance and low bias.
- So, we need to find the right/good balance without overfitting and underfitting the data.
- This tradeoff in complexity is why there is a tradeoff between bias and variance.
- An algorithm can’t be more complex and less complex at the same time.

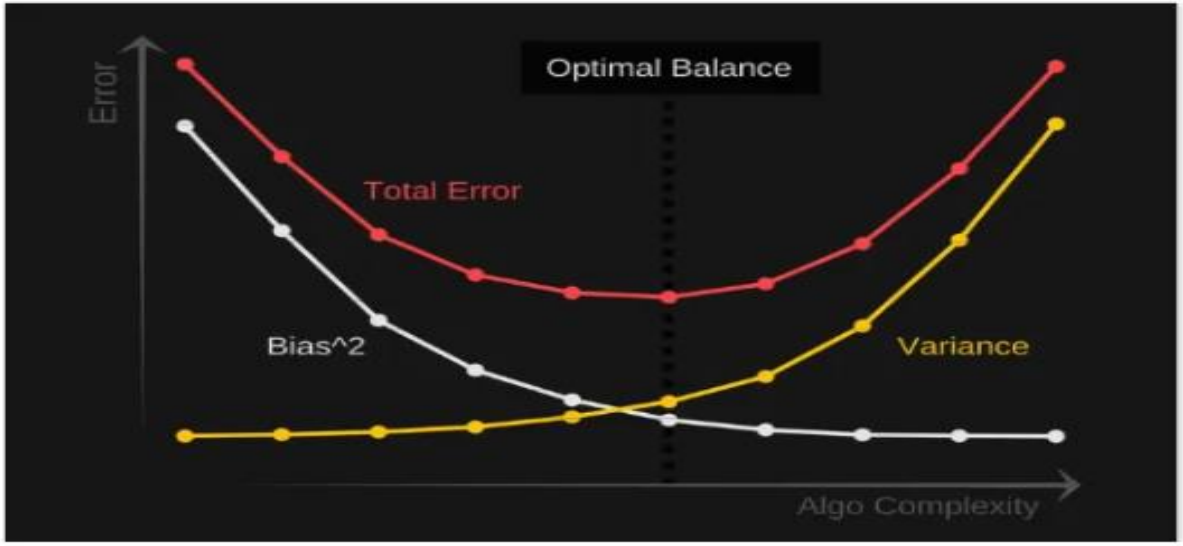
Total Error

DATA SCIENCE

Total Error

To build a good model, we need to find a good balance between bias and variance such that it minimizes the total error.

Total Error = Bias^2 + Variance + Irreducible Error



An optimal balance of bias and variance would never overfit or underfit the model.

Evaluation Metrics

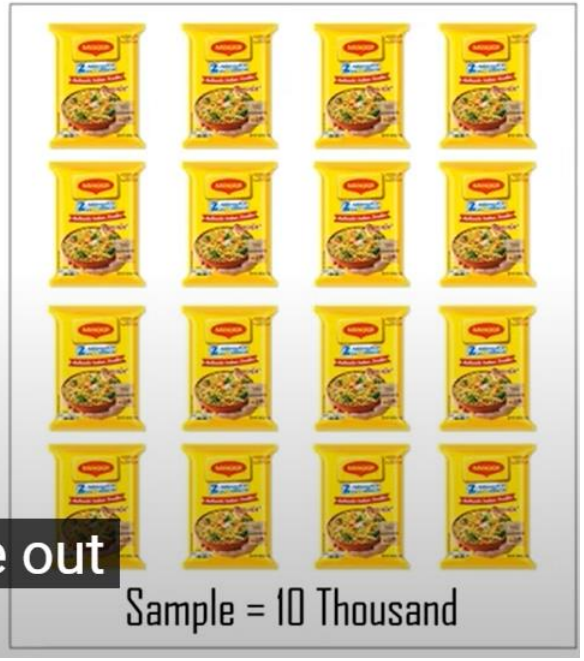
Hypothesis Testing

DATA SCIENCE

- Hypothesis testing refers to a process used by analysts to assess the plausibility of a hypothesis by using sample data.
- In hypothesis testing, statisticians formulate two hypotheses: the null hypothesis and the alternative hypothesis.
- A null hypothesis determines there is no difference between two groups or conditions, while the alternative hypothesis determines that there is a difference.
- Researchers evaluate the statistical significance of the test based on the probability that the null hypothesis is true.

Hypothesis Testing

Lead content <= 2.5 PPM



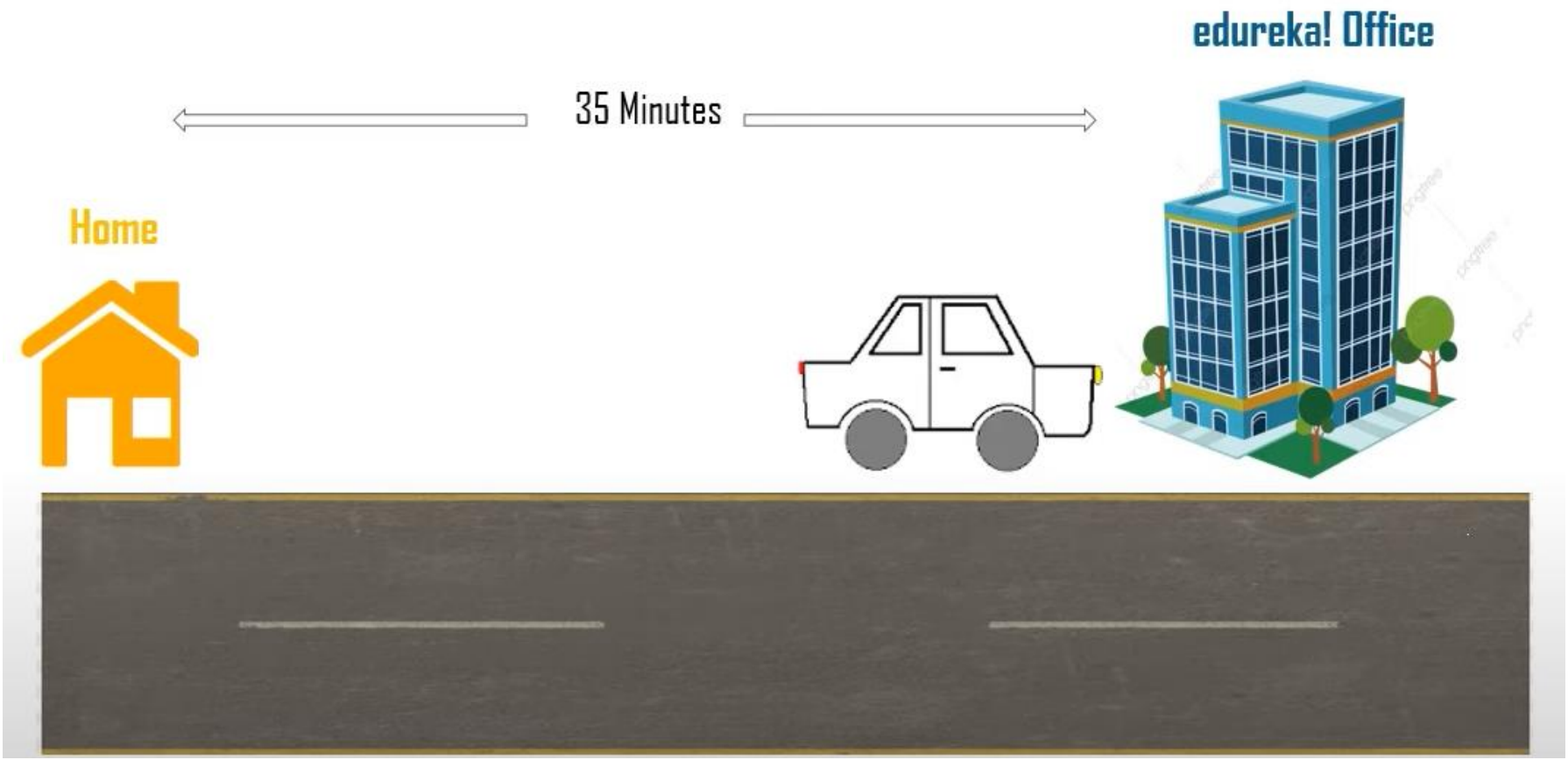
so they extracted out a small sample out of a very book population

Hypothesis Testing

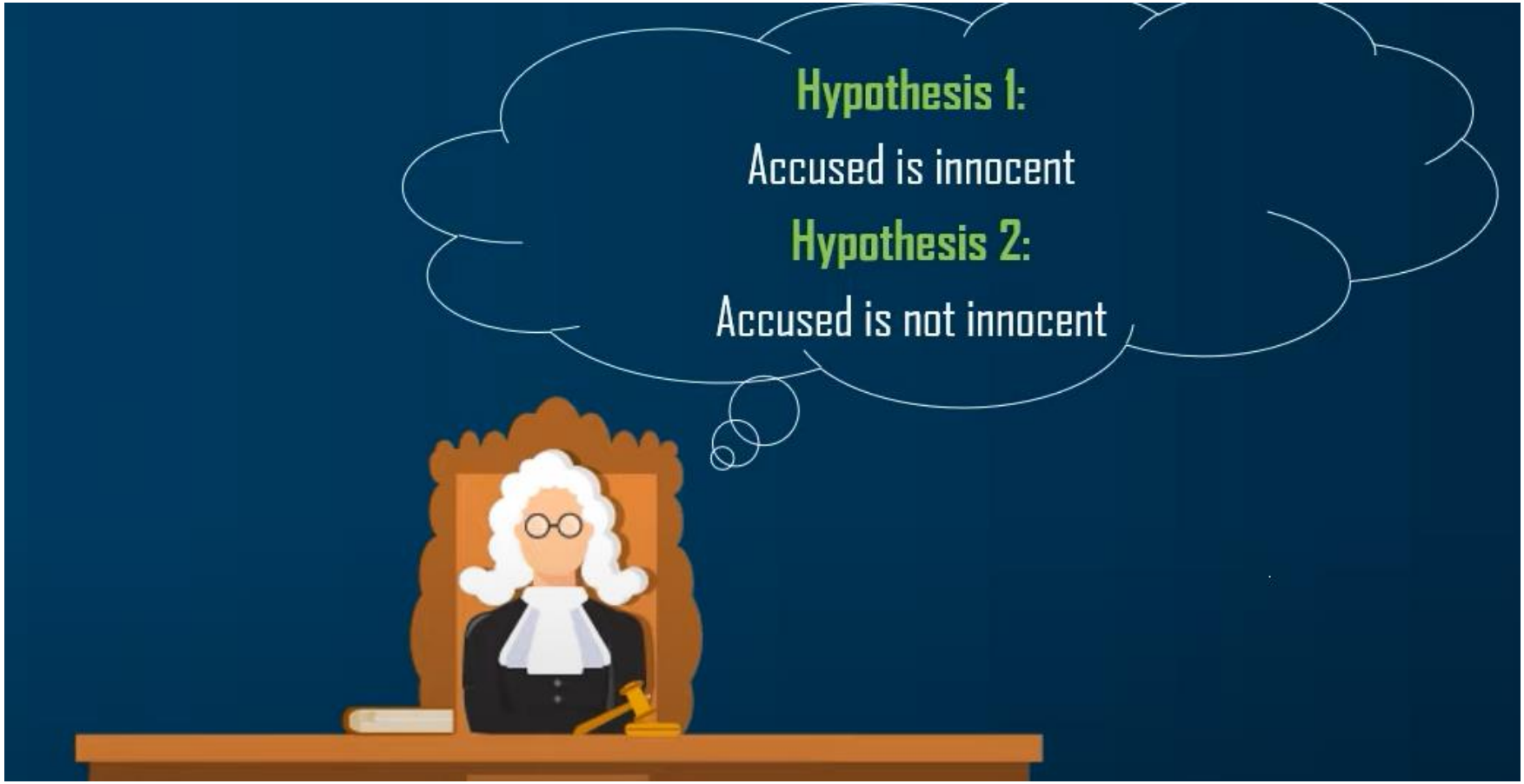
DATA SCIENCE

- Hypothesis testing is a statistical method used to determine if there is enough evidence in a sample data to draw conclusions about a population.
- Let's discuss few examples of statistical hypothesis from real-life -
- A teacher assumes that 60% of his college's students come from lower-middle-class families.
- A doctor believes that 3D (Diet, Dose, and Discipline) is 90% effective for diabetic patients.

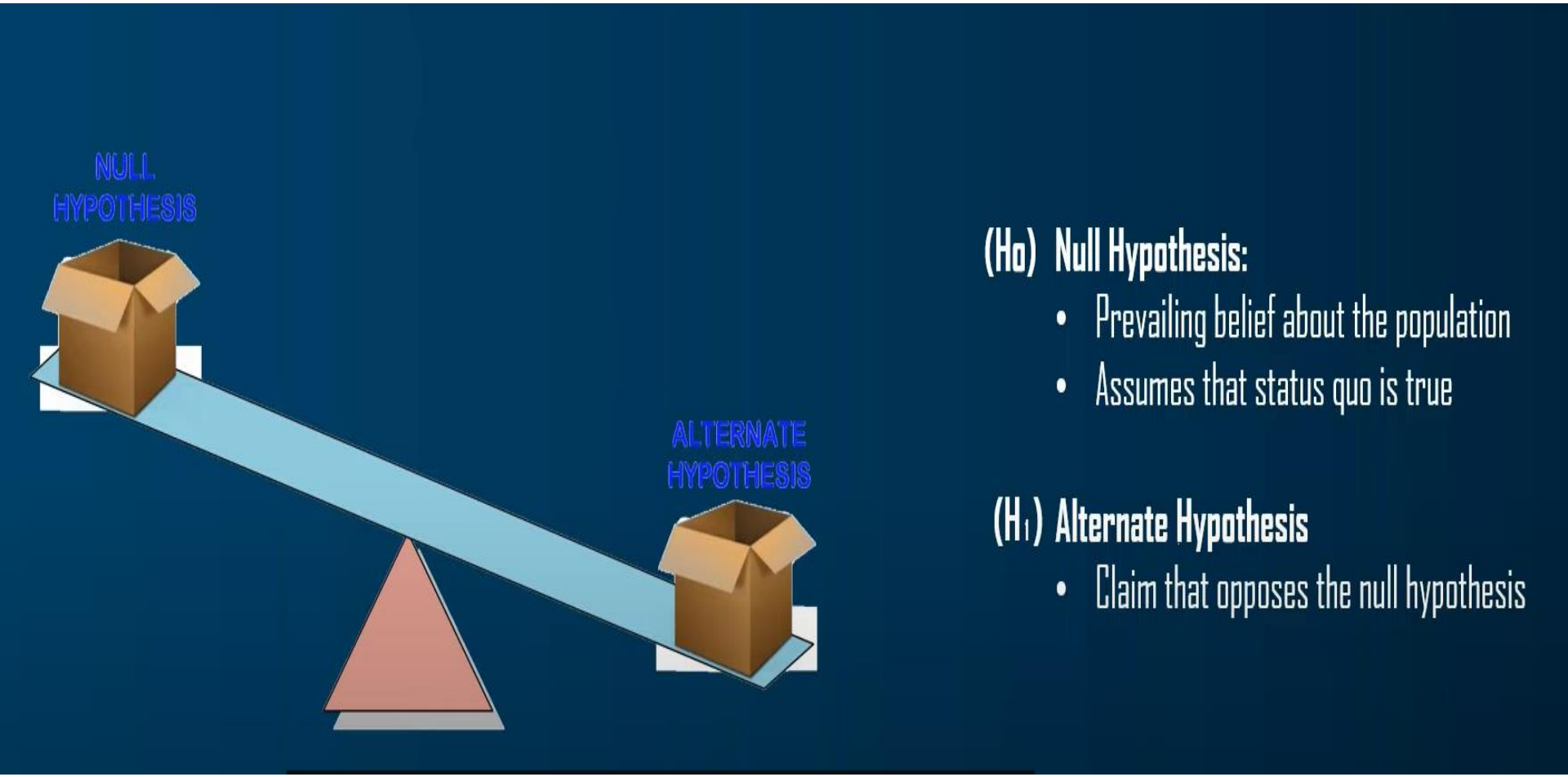
Hypothesis Testing



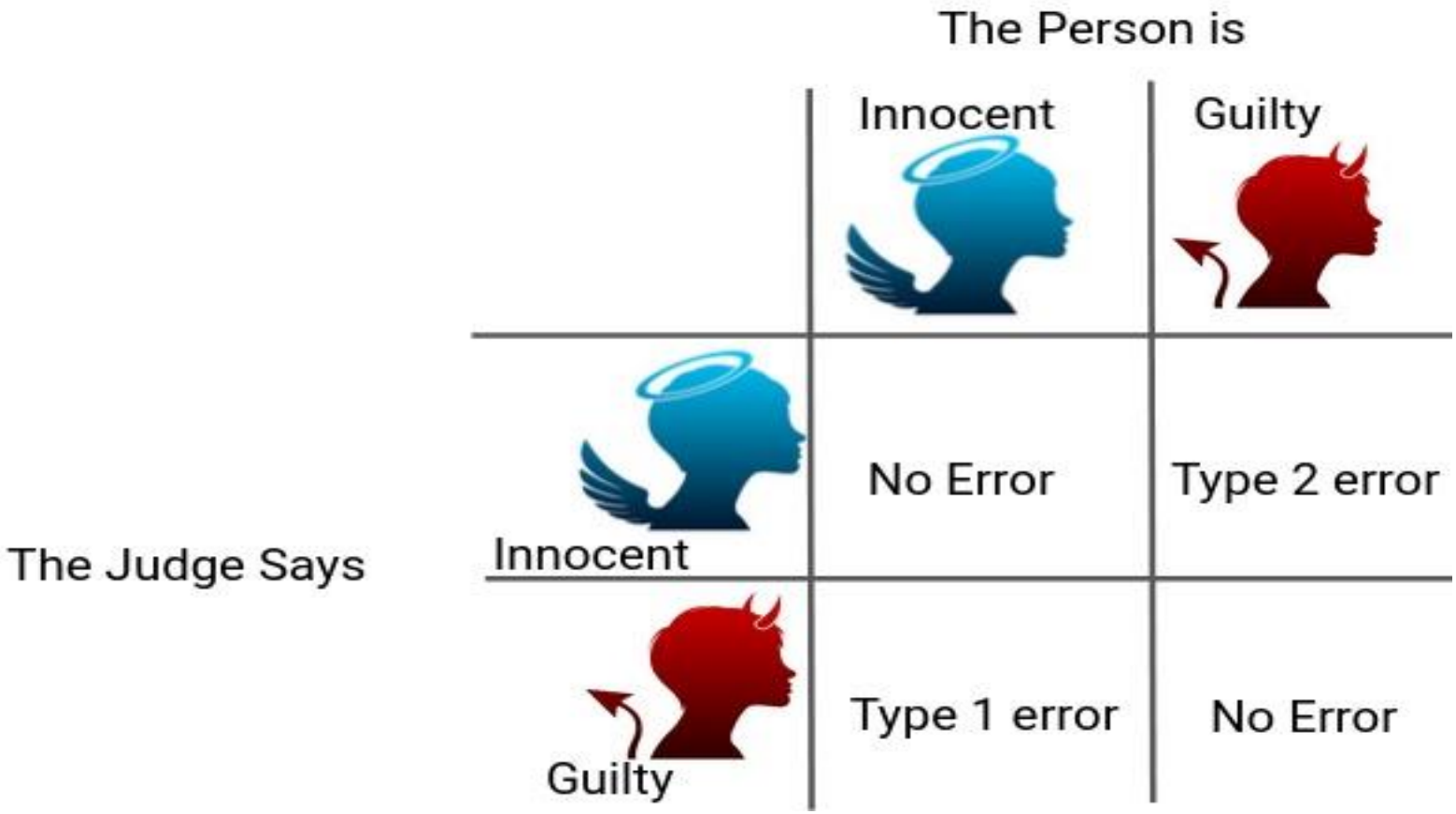
Null & Alternate Hypothesis



Null & Alternate Hypothesis



Null & Alternate Hypothesis



Null & Alternate Hypothesis

DATA SCIENCE

- First Case: The person is innocent, and the judge identifies the person as innocent
- Second Case: The person is innocent, and the judge identifies the person as guilty
- Third Case: The person is guilty, and the judge identifies the person as innocent
- Fourth Case: The person is guilty, and the judge identifies the person as guilty

Null & Alternate Hypothesis

Formulation of Null & Alternate Hypothesis:

If your claim statement has words like "at least", "at most", "less than", or "greater than", you cannot formulate the null hypothesis just from the claim statement

Rule to formulate the null and alternate hypotheses:

- Null hypothesis signs: = OR ≤ OR ≥
- Alternate hypothesis signs: ≠ OR > OR <

Situation 1:

Situation 2:

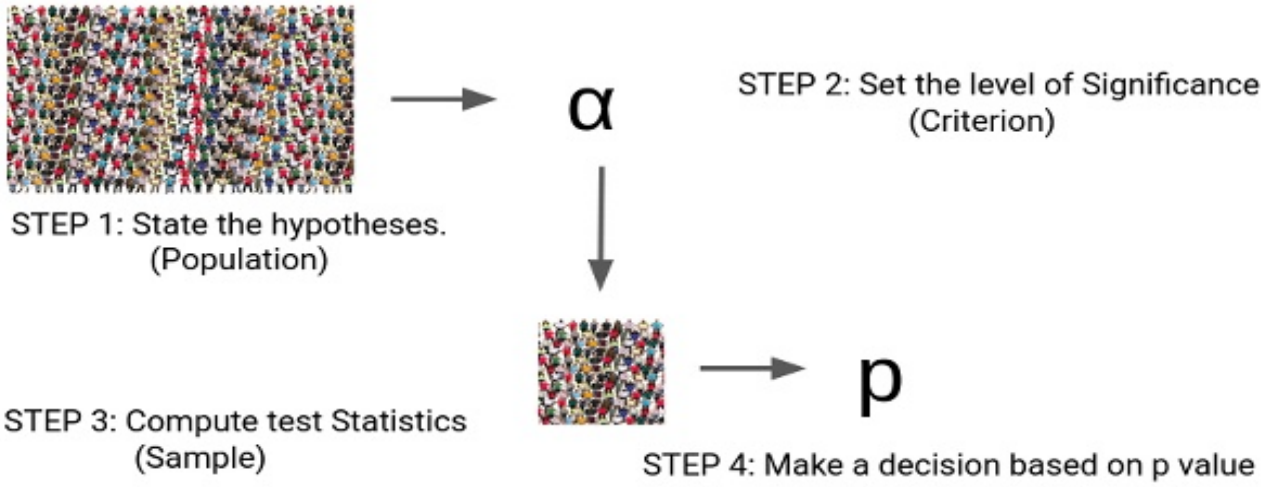
Steps to Perform Hypothesis Testing

DATA SCIENCE

Steps to Perform Hypothesis Testing

There are four steps to performing Hypothesis Testing:

1. Set the Null and Alternate Hypotheses
2. Set the Significance Level, Criteria for a decision
3. Compute the test statistic
4. Make a decision



Steps to Perform Hypothesis Testing

- It must be noted that z-Test & t-Tests are Parametric Tests, which means that the Null Hypothesis is about a population parameter, which is less than, greater than, or equal to some value.
- Steps 1 to 3 are quite self-explanatory but on what basis can we make decision in step 4? What does this p-value indicate?
- We can understand this p-value as the measurement of the Defense Attorney's argument.
- If the p-value is less than α , we reject the Null Hypothesis, and if the p-value is greater than α, we fail to reject the Null Hypothesis.

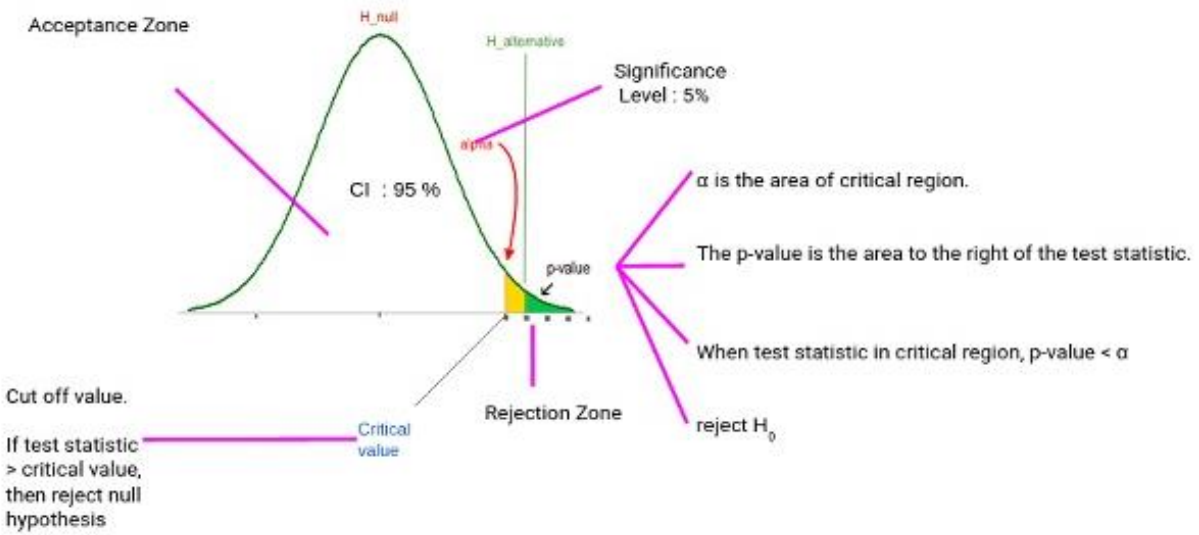
Critical P-Value

- Critical Value is the cut off value between Acceptance Zone and Rejection Zone.
- We compare our test score to the critical value and if the test score is greater than the critical value, that means our test score lies in the Rejection Zone and we reject the Null Hypothesis.
- On the other hand, if the test score is less than the Critical Value, that means the test score lies in the Acceptance Zone and we fail to reject the null Hypothesis.

Critical P-Value

DATA SCIENCE

Let's understand the logic of Hypothesis Testing with the graphical representation for Normal Distribution.



The above visualization helps to understand the z-value and its relation to the critical value. Typically, we set the Significance level at 10%, 5%, or 1%. If our test score lies in the Acceptance Zone, we fail to reject the Null Hypothesis. If our test score lies in the Critical Zone, we reject the Null Hypothesis and accept the Alternate Hypothesis.

Z-Test

DATA SCIENCE

- z tests are a statistical way of testing a Null Hypothesis when either:
  - We know the population variance, or
  - We do not know the population variance, but our sample size is large n ≥ 30
- If we have a sample size of less than 30 and do not know the population variance, we must use a t-test.
- This is how we judge when to use the z-test vs the t-test.
- Further, it is assumed that the z-statistic follows a standard normal distribution.
- In contrast, the t-statistics follows the t-distribution with a degree of freedom equal to n-1, where n is the sample size.
- It must be noted that the samples used for z-test or t-test must be independent sample and must have a distribution identical to the population distribution.
- This makes sure that the sample is not “biased” to/against the Null Hypothesis which we want to validate/invalidate.

Types of Hypothesis Testing

DATA SCIENCE

- Z Test
- T Test
- Chi Square Test

Finding P-Value

DATA SCIENCE

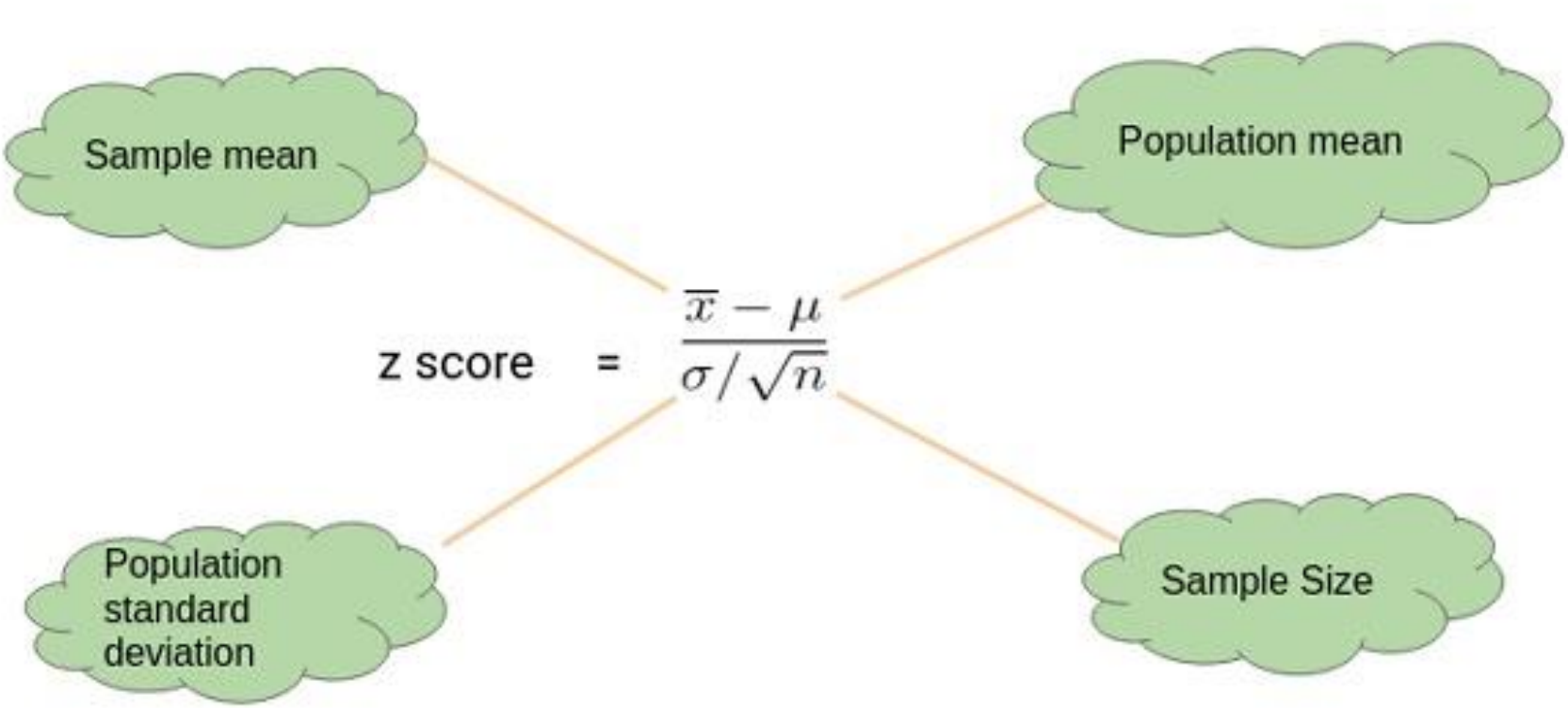
- P stands for probability here.
- To calculate the p-value, the chi-square test is used in statistics.
- The different values of p indicates the different hypothesis interpretation, are given below:
- P≤ 0.05; Hypothesis rejected
- P>.05; Hypothesis Accepted
- Probability is all about chance or risk or uncertainty.
- It is the possibility of the outcome of the sample or the occurrence of an event.

Finding P-Value

DATA SCIENCE

- But when we talk about statistics, it is more about how we handle various data using different techniques.
- It helps to represent complicated data or bulk data in a very easy and understandable way.
- It describes the collection, analysis, interpretation, presentation, and organization of data.
- The concept of both probability and statistics is related to the chi-squared test.

Z-Test


Sample mean
Population mean
z score = $\frac{\overline{x} - \mu}{\sigma / \sqrt{n}}$
Population standard deviation
Sample Size

Z-Test

DATA SCIENCE

Let's say we need to determine if girls on average score higher than 600 in the exam. We have the information that the standard deviation for girls' scores is 100. So, we collect the data of 20 girls by using random samples and record their marks. Finally, we also set our α value (significance level) to be 0.05.

Score
650
730
510
670
480
800
690
530
590
620
710
670
640
780
650
490
800
600
510
700

Z-Test

DATA SCIENCE

In this example:

- Mean Score for Girls is 641
- The number of data points in the sample is 20
- The population mean is 600
- Standard Deviation for Population is 100

$$\text{z score} = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}}$$

$$= \frac{641 - 600}{100 / \sqrt{20}}$$

$$= 1.8336$$

p value = .033357.

Critical Value = 1.645

Z score > Critical Value

P value < 0.05

 # P-Value : Probability of getting less than a Z-score
p_value = 1-stats.norm.cdf(z_score)

$H_0 : \mu \leq 600$

$H_1 : \mu > 600$

Z-Test

DATA SCIENCE

P Value from Z Score Calculator

This is very easy: just stick your Z score in the box marked Z score, select your significance level and whether you're testing a one or two-tailed hypothesis (if you're not sure, go with the defaults), then press the button!

If you need to derive a Z score from raw data, you can find a Z test calculator here.

Z score: 1.8335

Significance Level:

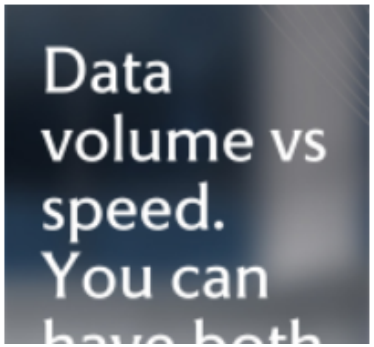
- ○ 0.01
- ◉ 0.05
- ○ 0.10

One-tailed or two-tailed hypothesis?:

- ◉ One-tailed
- ○ Two-tailed

The P-Value is .033364.

The result is significant at p < .05.



Types of Hypothesis Testing

DATA SCIENCE

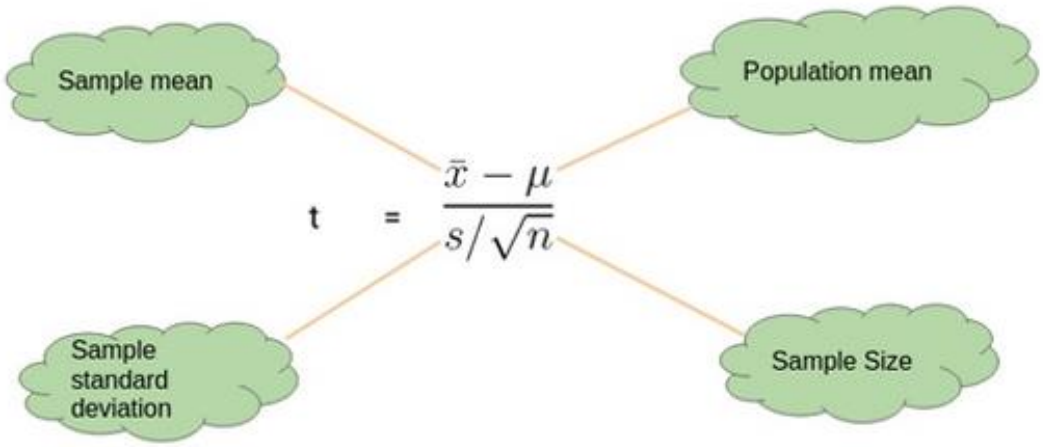
- T Test
  - T-tests are a statistical way of testing a hypothesis when:
  - We do not know the population variance
  - Our sample size is small, n < 30.

Types of Hypothesis Testing

DATA SCIENCE

One-Sample T-Test

We perform a One-Sample t-test when we want to **compare a sample mean with the population mean.** The difference from the z-Test is that we do **not have the information on Population Variance** here. We use the **sample standard deviation** instead of population standard deviation in this case.



Types of Hypothesis Testing

DATA SCIENCE

Here's an Example to Understand a One Sample T-Test

Let's say we want to determine if on average girls score more than 600 in the exam. We do not have the information related to variance (or standard deviation) for girls' scores. To a perform t-test, we randomly collect the data of 10 girls with their marks and choose our α value (significance level) to be 0.05 for Hypothesis Testing.

Girls_Score
587
602
627
610
619
622
605
608
596
592

Types of Hypothesis Testing

DATA SCIENCE

In this example:

- Mean Score for Girls is 606.8
- The size of the sample is 10
- The population mean is 600
- Standard Deviation for the sample is 13.14

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$= \frac{606.8 - 600}{13.14/\sqrt{10}}$$

$$= 1.64$$

$H_0 : \mu \leq 600$

$H_1 : \mu > 600$

Critical Value = 1.833

t score < Critical Value

P value = 0.0678

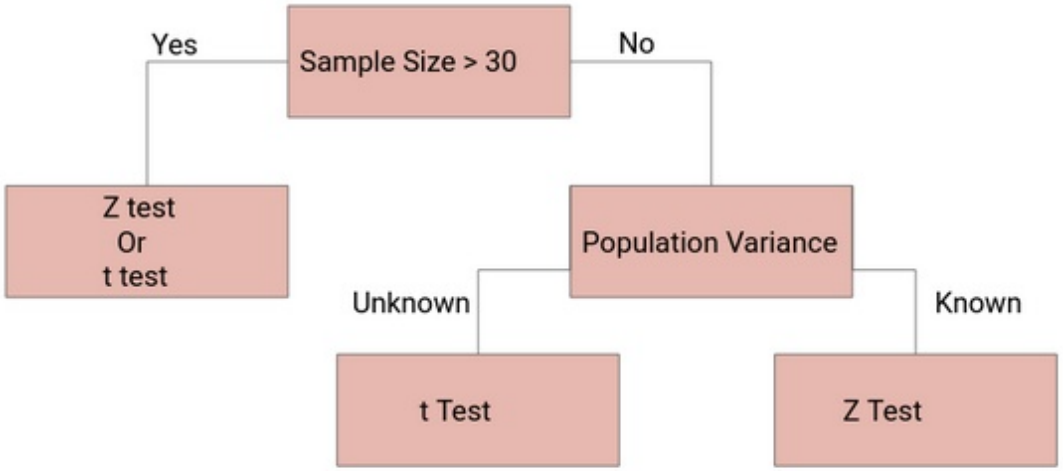
P value > 0.05

Our **p-value is greater than 0.05 thus we fail to reject the null hypothesis** and don't have enough evidence to support the hypothesis that on average, girls score more than 600 in the exam.

Deciding Between Z-Test and T-Test

DATA SCIENCE

So when should we perform the z-test, and when should we perform the t-Test? It's a critical question we need to answer if we want to master statistics.



If the sample size is large enough, then the z-Test and t-Test will conclude with the same results. For a **large sample size**, **Sample Variance will be a better estimate** of Population variance, so even if population variance is unknown we can **use the z-test using sample variance.**

Similarly, for a **Large Sample**, we have a high degree of freedom. And since t-**distribution approaches the normal distribution**, the difference between the z score and t score is negligible.

Deciding Between Z-Test and T-Test

DATA SCIENCE

	Z Test	T Test
Assumption	Population standard deviation is known	Population standard deviation is unknown
Sample Size	Large sample size (n > 30)	Small sample size (n < 30)
Distribution	Z-distribution	T-distribution
Test Statistic	(Sample mean – Population mean) / (Population SD / √n)	(Sample mean – Population mean) / (Sample SD / √n)
Hypothesis Testing	Test for a population mean or proportion	Test for a population mean
Degrees of Freedom	Not applicable	n – 1
Application	Used when the population standard deviation is known and the sample size is large	Used when the population standard deviation is unknown or the sample size is small
Example	Testing whether the average height of male adults is significantly different from a known value	Testing whether a new teaching method improves student test scores compared to the old method

Pearson's Chi Square Test

DATA SCIENCE

- A chi-squared test (symbolically represented as $\chi 2$) is basically a data analysis on the basis of observations of a random set of variables.
- Usually, it is a comparison of two statistical data sets.
- This test was introduced by Karl Pearson in 1900 for categorical data analysis and distribution.
- So, it was mentioned as Pearson's chi-squared test.
- The chi-square test is used to estimate how likely the observations that are made would be, by considering the assumption of the null hypothesis as true.

Chi-Square Test of Independence

DATA SCIENCE

- The chi-square test of independence also known as the chi-square test of association which is used to determine the association between the categorical variables.
- It is considered as a non-parametric test.
- It is mostly used to test statistical independence.

Chi-Square Test of Independence

DATA SCIENCE

- The chi-square test of independence is not appropriate when the categorical variables represent the pre-test and post-test observations.
- For this test, the data must meet the following requirements:
- Two categorical variables
- Relatively large sample size
- Categories of variables (two or more)
- Independence of observations.

Chi-Square Test of Independence

- Let us take an example of a categorical data where there is a society of 1000 residents with four neighbourhoods, P, Q, R and S.
- A random sample of 650 residents of the society is taken whose occupations are doctors, engineers and teachers.
- The null hypothesis is that each person's neighbourhood of residency is independent of the person's professional division.

Finding P-Value

DATA SCIENCE

Formula

The chi-squared test is done to check if there is any difference between the observed value and expected value. The formula for chi-square can be written as;

$$X^2 = \sum \frac{(\text{Observed value} - \text{Expected value})^2}{\text{Expected value}}$$

or

$$\chi^2 = \sum (O_i - E_i)^2 / E_i$$

where O_i is the observed value and E_i is the expected value.

Chi-Square Test of Independence

DATA SCIENCE

Categories	P	Q	R	S	Total
Doctors	90	60	104	95	349
Engineers	30	50	51	20	151
Teachers	30	40	45	35	150
Total	150	150	200	150	650

Chi-Square Test of Independence

DATA SCIENCE

- Assume the sample living in neighborhood P, 150, to estimate what proportion of the whole 1,000 people live in neighborhood P.
- In the same way, we take 349/650 to calculate what ratio of the 1,000 are doctors.
- By the supposition of independence under the hypothesis, we should “expect” the number of doctors in neighbourhood P is;
- 150 x 349/650 ≈ 80.54

Chi-Square Test of Independence

- So, by the chi-square test formula for that particular cell in the table, we get;
- (Observed – Expected)2/Expected Value = (90-80.54)2/80.54 ≈ 1.11
- Some of the exciting facts about the Chi-square test are given below:
- The Chi-square statistic can only be used on numbers.
- We cannot use them for data in terms of percentages, proportions, means or similar statistical contents.
- Suppose, if we have 20% of 400 people, we need to convert it to a number, i.e. 80, before running a test statistic.
- A chi-square test will give us a p-value. The p-value will tell us whether our test results are significant or not.

Chi-Square Test of Independence

- However, to perform a chi-square test and get the p-value, we require two pieces of information:
- (1) Degrees of freedom. That's just the number of categories minus 1.
- (2) The alpha level(α). You or the researcher chooses this. The usual alpha level is 0.05 (5%), but you could also have other levels like 0.01 or 0.10.
- In elementary statistics, we usually get questions along with the degrees of freedom(DF) and the alpha level.
- Thus, we don't usually have to figure out what they are. To get the degrees of freedom, count the categories and subtract 1.

Confusion Matrix

DATA SCIENCE

- A confusion matrix is a matrix that summarizes the performance of a machine learning model on a set of test data.
- It is often used to measure the performance of classification models, which aim to predict a categorical label for each input instance.
- The matrix displays the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) produced by the model on the test data.

Confusion Matrix

DATA SCIENCE

- A confusion matrix is a matrix that summarizes the performance of a machine learning model on a set of test data.
- It is often used to measure the performance of classification models, which aim to predict a categorical label for each input instance.
- The matrix displays the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) produced by the model on the test data.

Confusion Matrix

DATA SCIENCE

		Actual	
		Dog	**Not Dog**
Predicted	**Dog**	True Positive (TP)	False Positive (FP)
	Not Dog	False Negative (FN)	True Negative (TN)

- **True Positive (TP):** It is the total counts having both predicted and actual values are Dog.
- **True Negative (TN):** It is the total counts having both predicted and actual values are Not Dog.
- **False Positive (FP):** It is the total counts having prediction is Dog while actually Not Dog.
- **False Negative (FN):** It is the total counts having prediction is Not Dog while actually, it is Dog.

Confusion Matrix

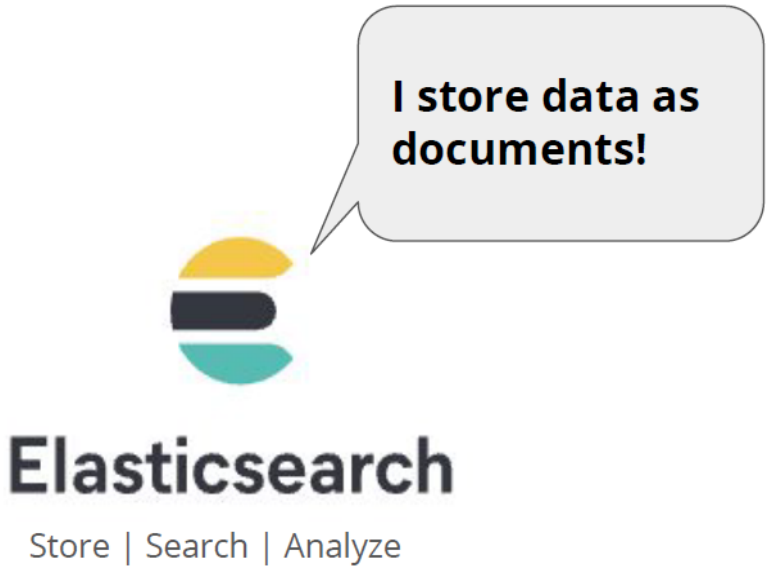
DATA SCIENCE

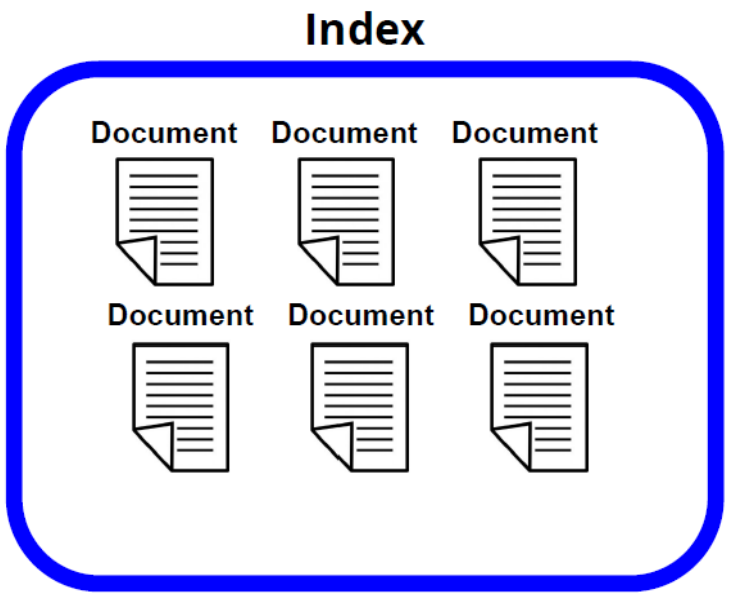
Example

Index	1	2	3	4	5	6	7	8	9	10
Actual	Dog	Dog	Dog	Not Dog	Dog	Not Dog	Dog	Dog	Not Dog	Not Dog
Predicted	Dog	Not Dog	Dog	Not Dog	Dog	Dog	Dog	Dog	Not Dog	Not Dog
Result	TP	FN	TP	TN	TP	FP	TP	TP	TN	TN

- Actual Dog Counts = 6
- Actual Not Dog Counts = 4
- True Positive Counts = 5
- False Positive Counts = 1
- True Negative Counts = 3
- False Negative Counts = 1

Relevance



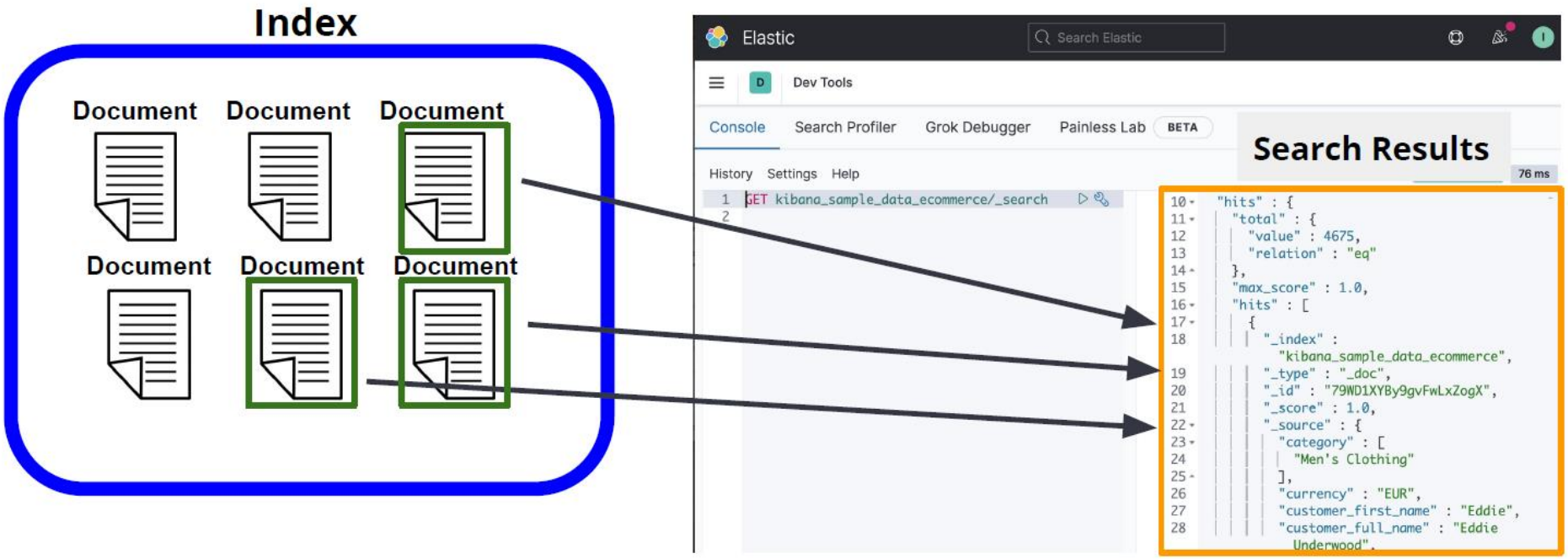


Documents with similar traits are grouped into an index!

Relevance

DATA SCIENCE

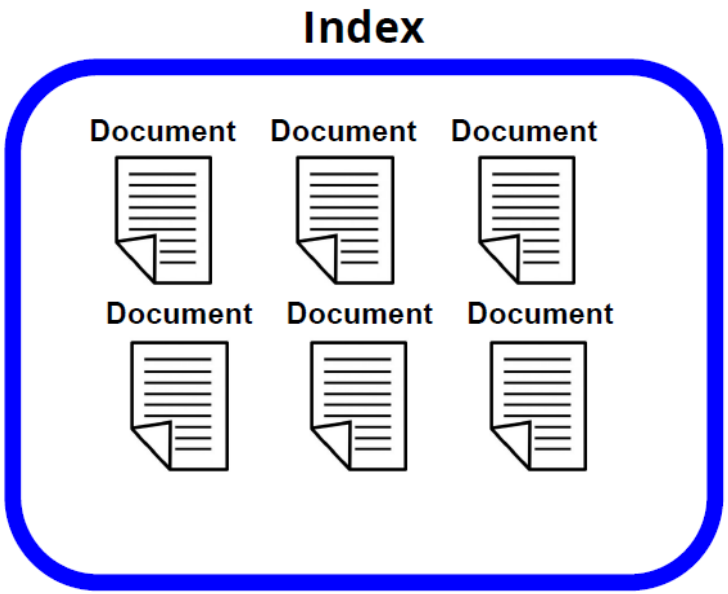
When search query is sent, Elasticsearch retrieves relevant documents and presents the documents as search results.

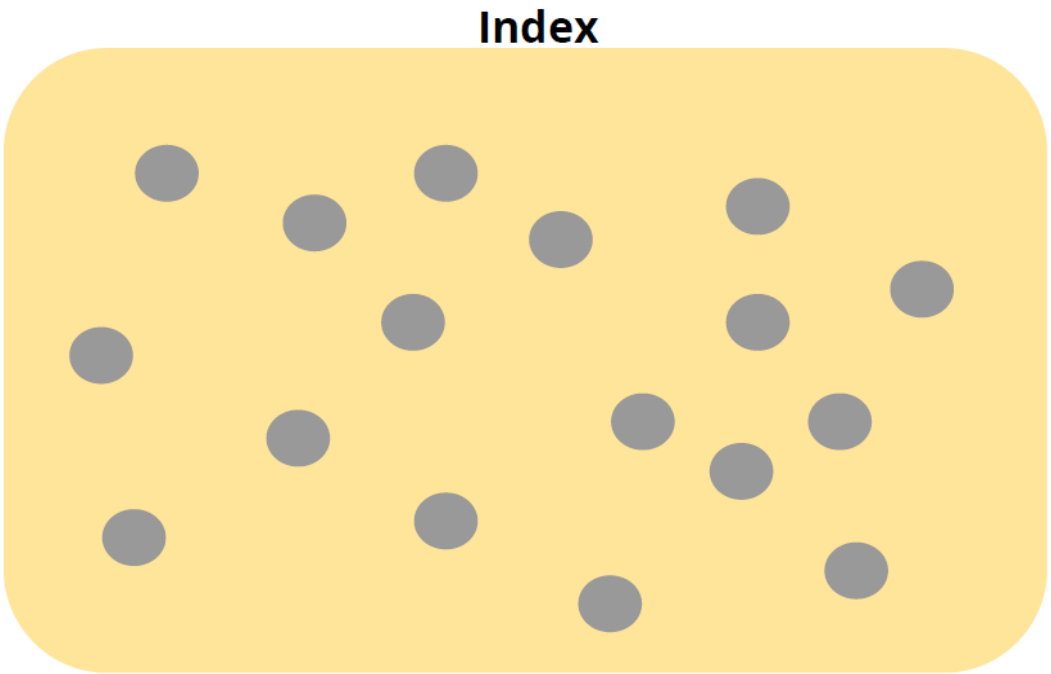


Relevance

DATA SCIENCE

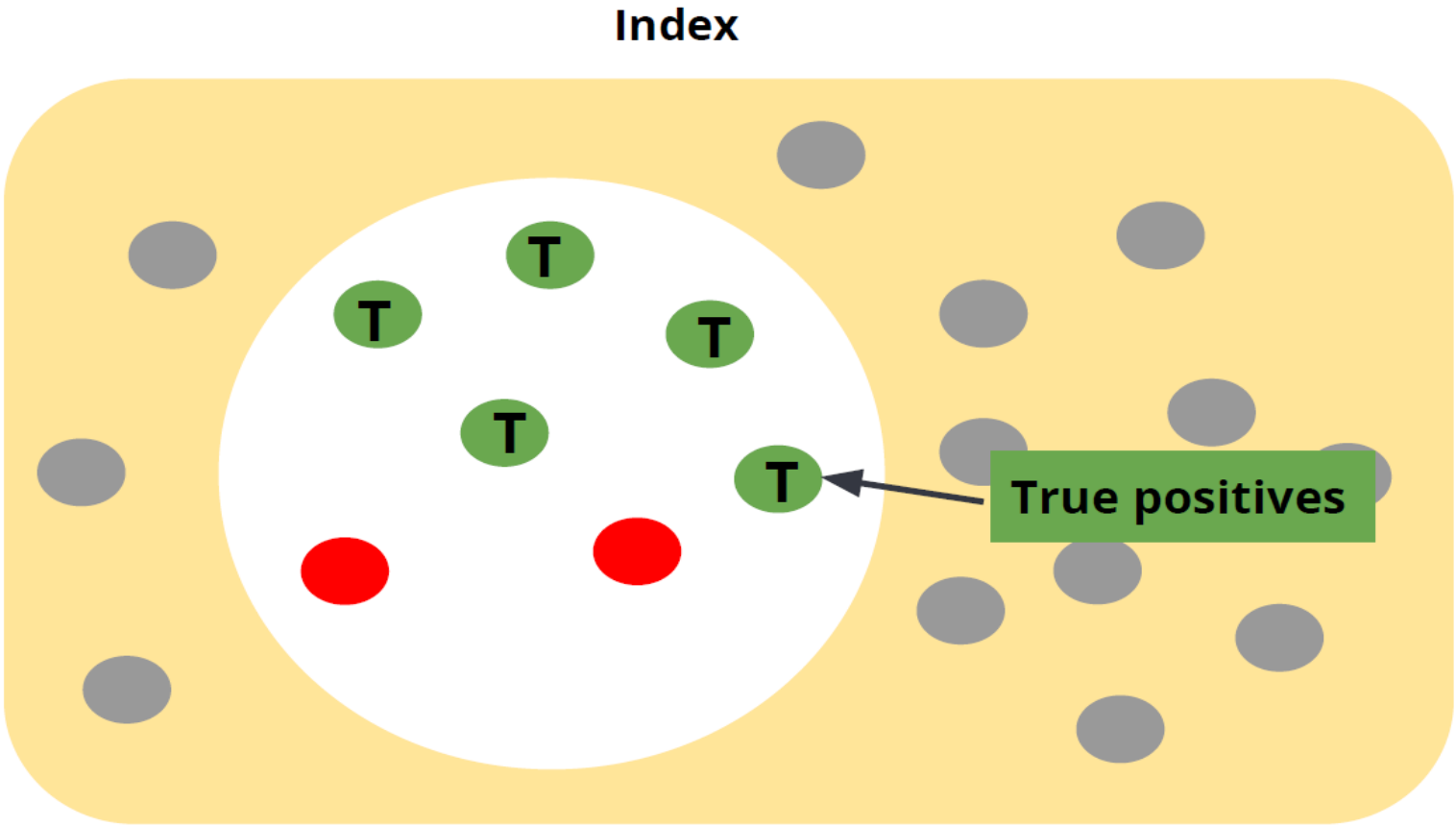
These two diagrams depict the same thing!





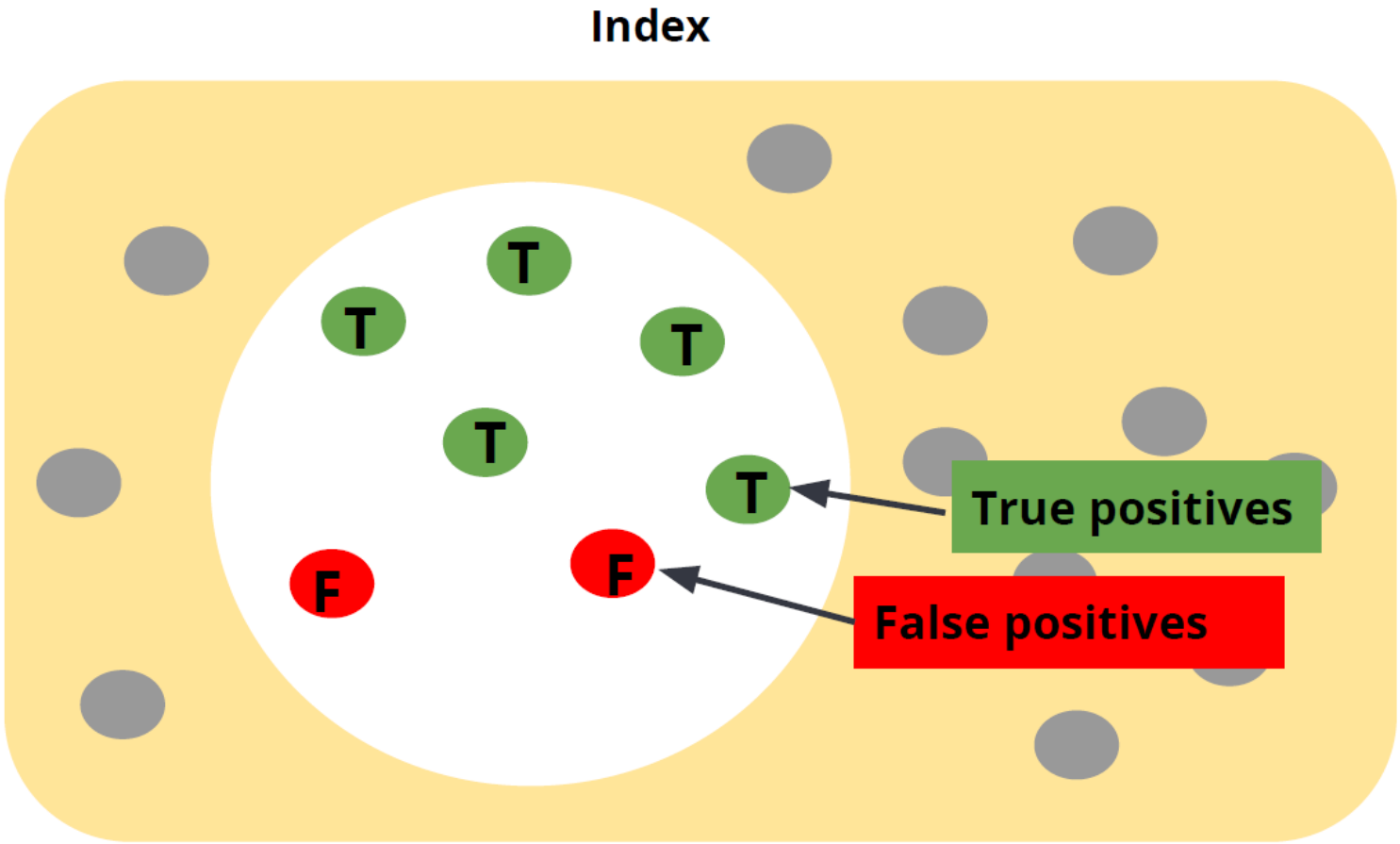
Relevance

True positives are relevant documents that are returned to the user.



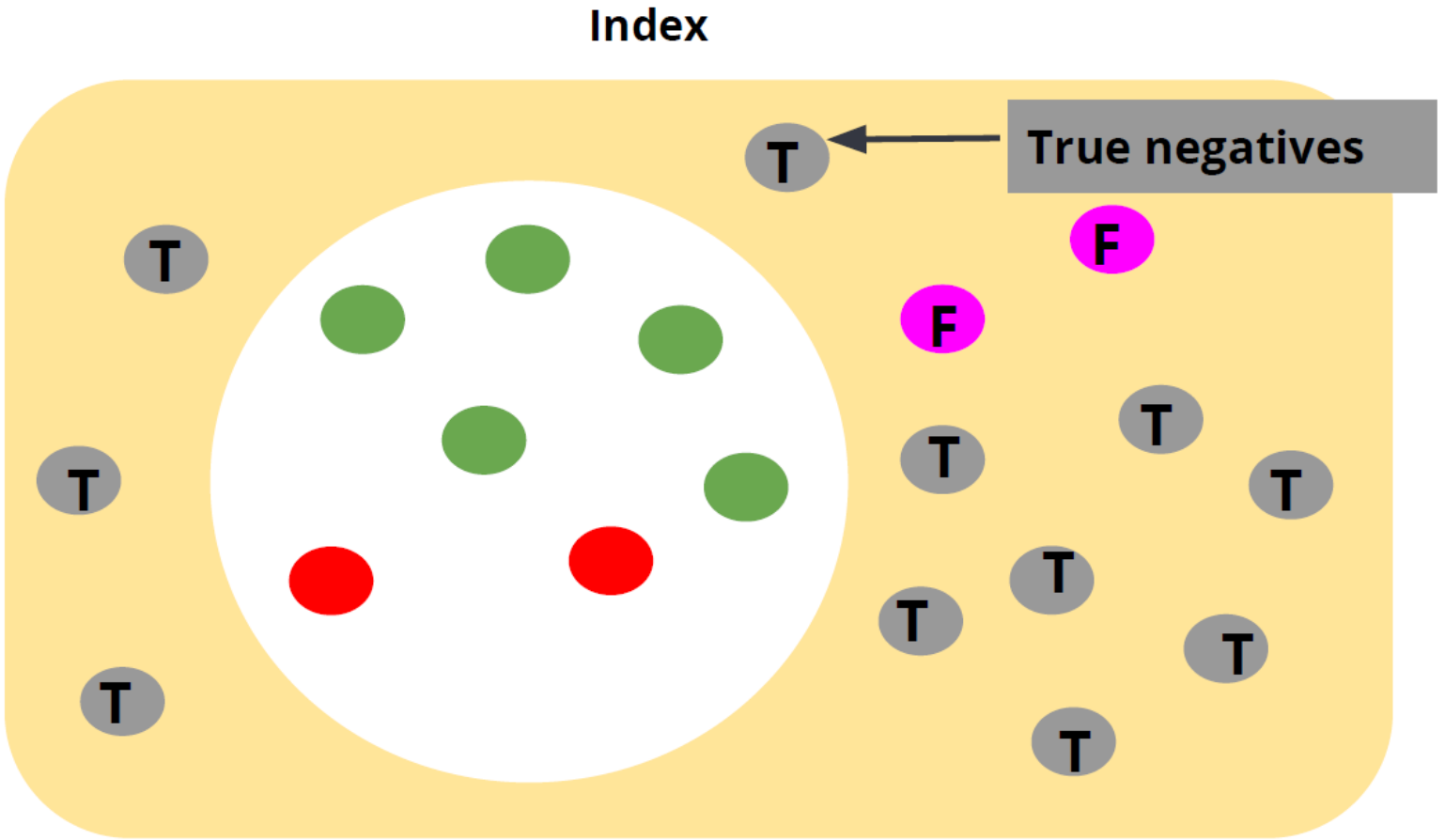
Relevance

False positives are irrelevant documents that are returned to the user.



Relevance

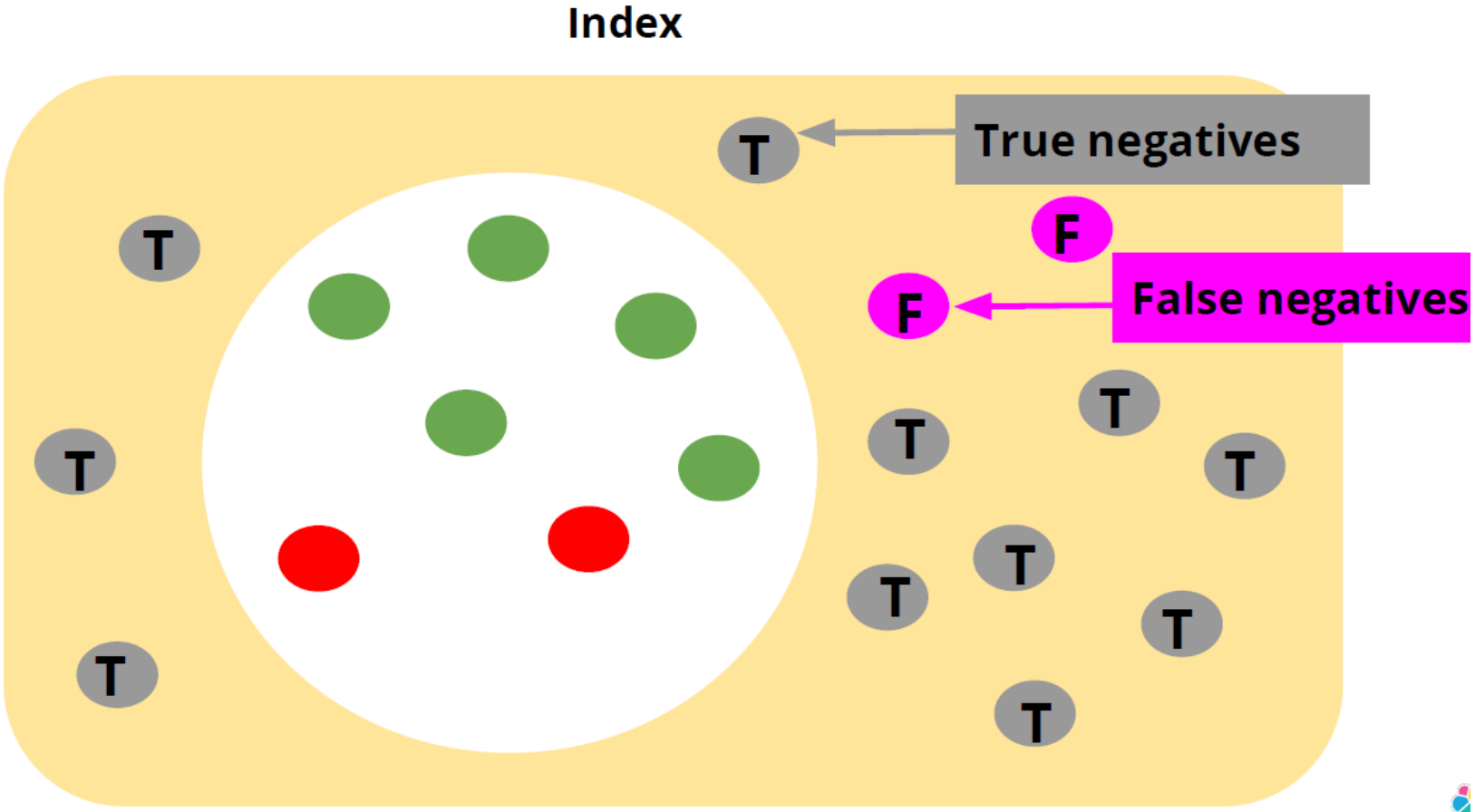
True negatives are irrelevant documents that are not returned to the user.



Relevance

DATA SCIENCE

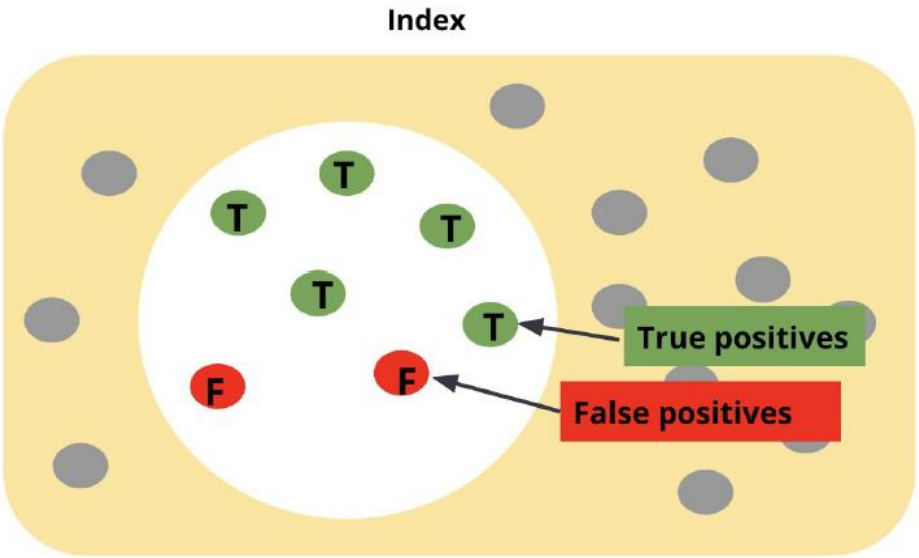
False negatives are relevant documents that were not returned to the user.



Relevance

DATA SCIENCE

What is precision?



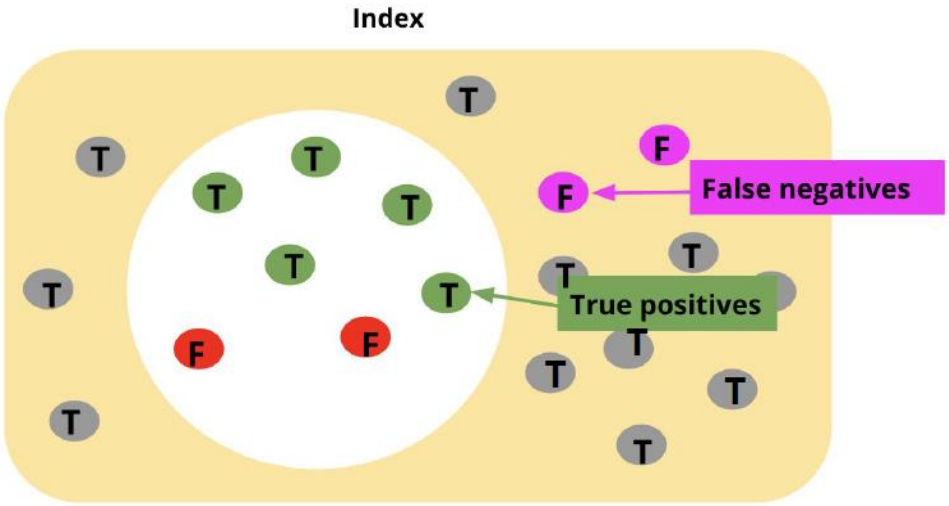
$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

What portion of the retrieved data is actually relevant to the search query?

Relevance

DATA SCIENCE

What is recall?



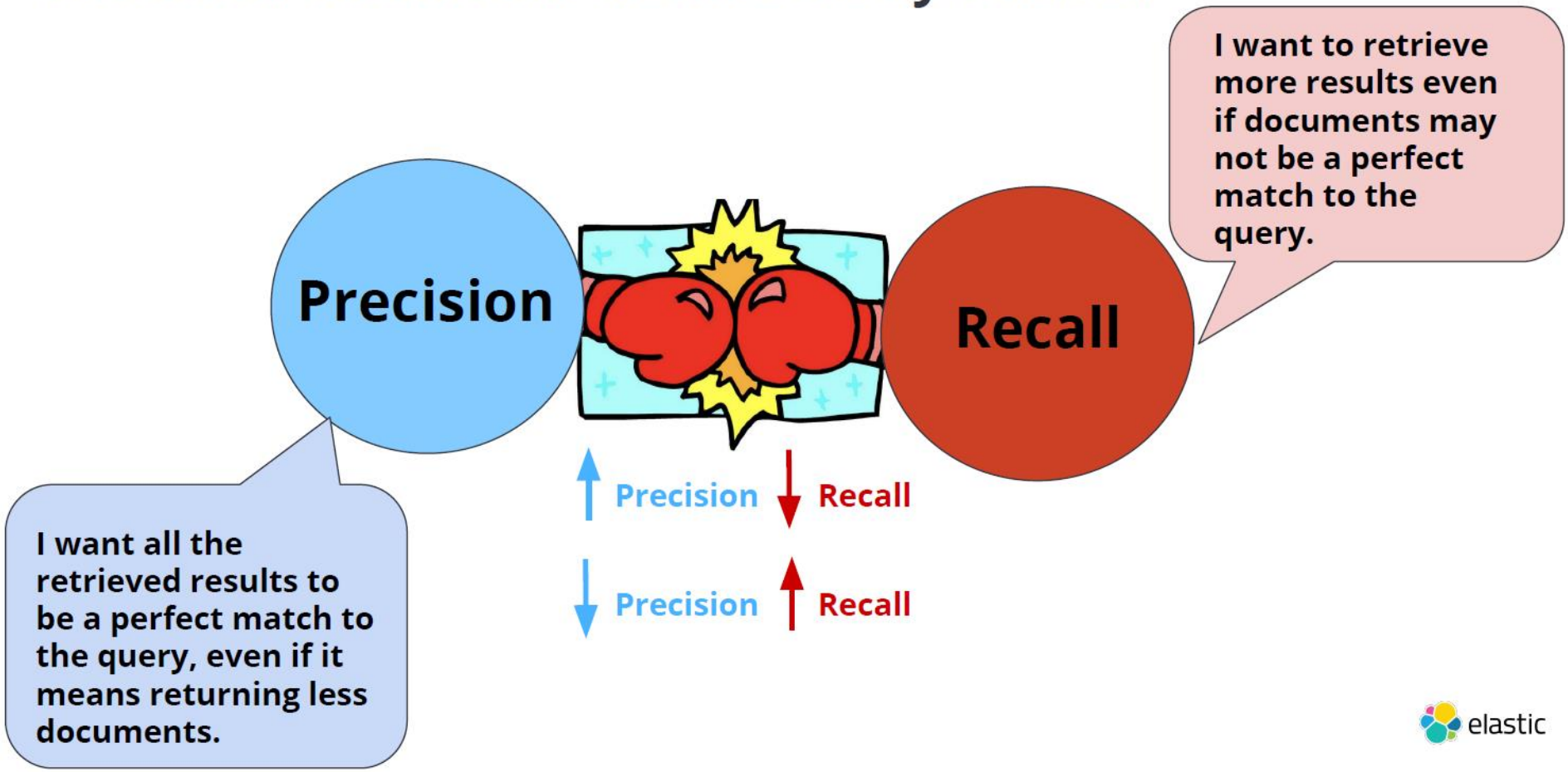
$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

What portion of relevant data is being returned as search results?

Relevance

DATA SCIENCE

Precision and Recall are inversely related

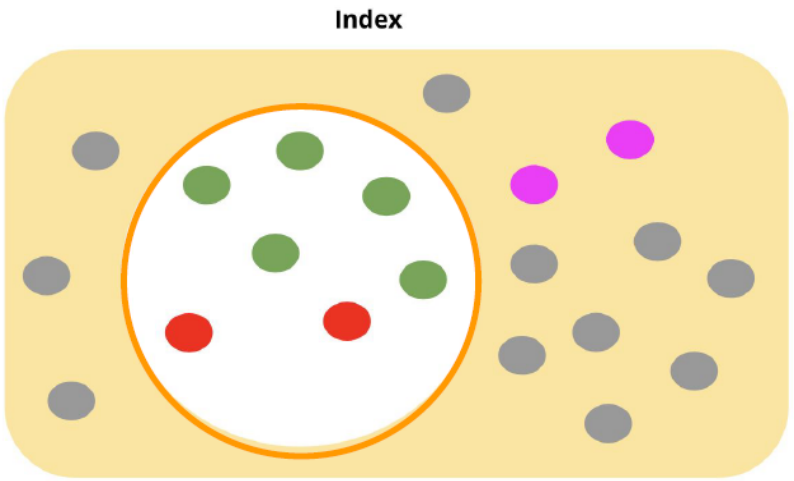


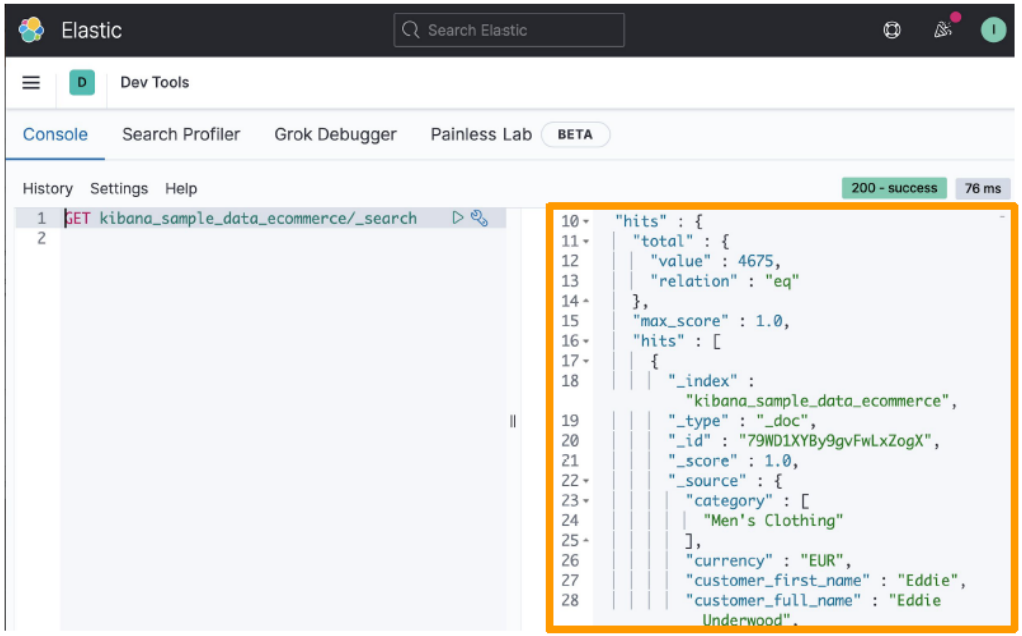
elastic

Relevance

DATA SCIENCE

Precision and recall determine which documents are included in the search results.





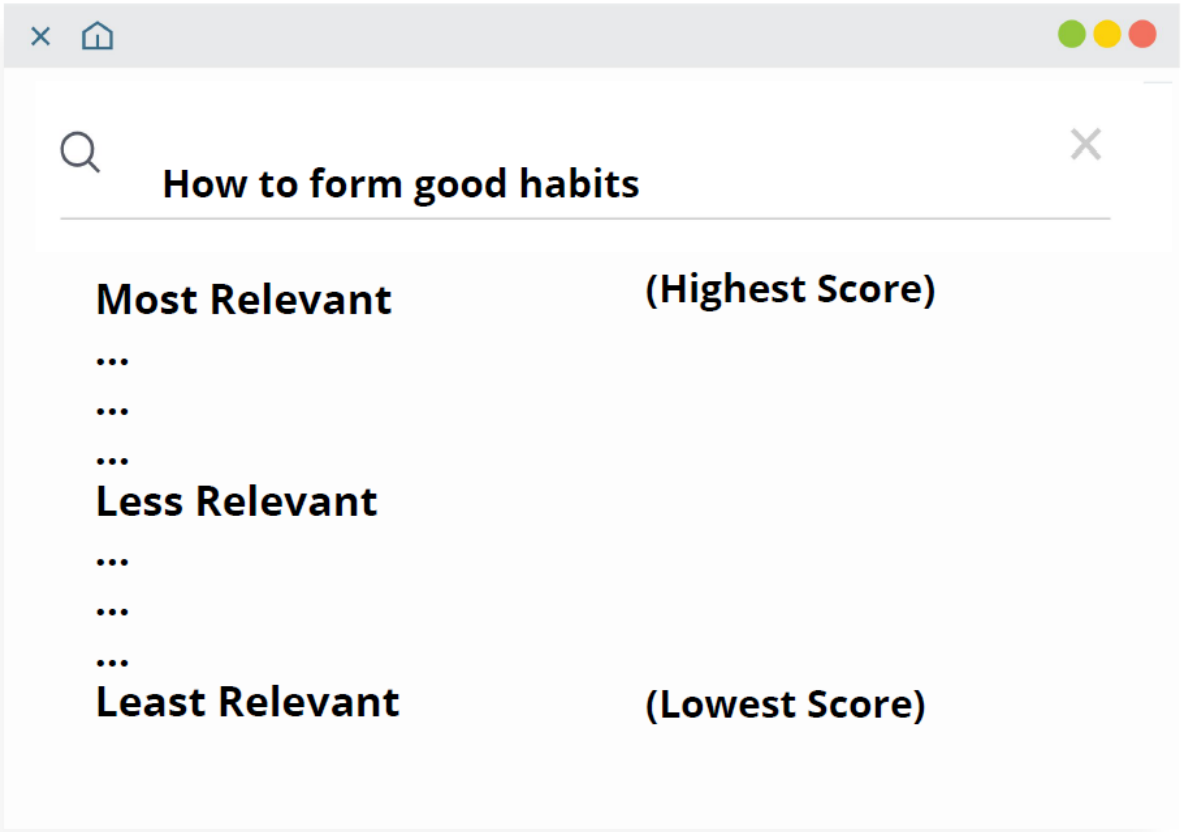
Precision and recall do not determine which of the returned documents are more relevant compared to the other!

elastic

Relevance

DATA SCIENCE

Ranking refers to ordering of the results (from most relevant results at the top, to least relevant at the bottom).



elastic

Score

DATA SCIENCE

The score is a value that represents how relevant a document is to that specific query

A score is computed for each document that is a hit

Specificity

DATA SCIENCE

- Specificity of a classifier is the ratio between how much were correctly classified as negative to how much was negative.
- Specificity = TN/FP+TN
- Where is it used?
- Places where classification of negatives are high priority.
- Eg: Diagnosing for a health condition before treatment.

Sensitivity

DATA SCIENCE

- Sensitivity of a classifier is the ratio between how much were correctly identified as positive to how much were positive.
- Sensitivity = TP / FN+TP
- Where is it used ?
- Places where classification of positives are high priority.
- Eg: Security checks in airports.

Specificity

DATA SCIENCE

- Specificity of a classifier is the ratio between how much were correctly classified as negative to how much was negative.
- Specificity = TN/FP+TN
- Where is it used?
- Places where classification of negatives are high priority.
- Eg: Diagnosing for a health condition before treatment.

Precision

DATA SCIENCE

- How much were correctly classified as positive out of all positives.
- Precision = TP/TP+FP

Recall

DATA SCIENCE

- **Recall and sensitivity are one and the same.**
  Recall = TP / FN+TP

Where does precision and recall are used ?

- **The harmonic mean of precision and recall gives a score call f1 score which is a measure of performance of the model's classification ability.**

- **F1 score = 2 * (precision * recall)/ (precision + recall)**

- **F1 score is considered a better indicator of the classifier's performance than the regular accuracy measure.**

AUC-ROC curve

DATA SCIENCE

- ROC stands for Receiver Operating Characteristics, and the ROC curve is the graphical representation of the effectiveness of the binary classification model.
- It plots the true positive rate (TPR) vs the false positive rate (FPR) at different classification thresholds.

AUC-ROC curve

- AUC stands for Area Under the Curve, and the AUC curve represents the area under the ROC curve.
- It measures the overall performance of the binary classification model.
- As both TPR and FPR range between 0 to 1, So, the area will always lie between 0 and 1, and A greater value of AUC denotes better model performance.
- Our main goal is to maximize this area to have the highest TPR and lowest FPR at the given threshold.
- The AUC measures the probability that the model will assign a randomly chosen positive instance a higher predicted probability compared to a randomly chosen negative instance.

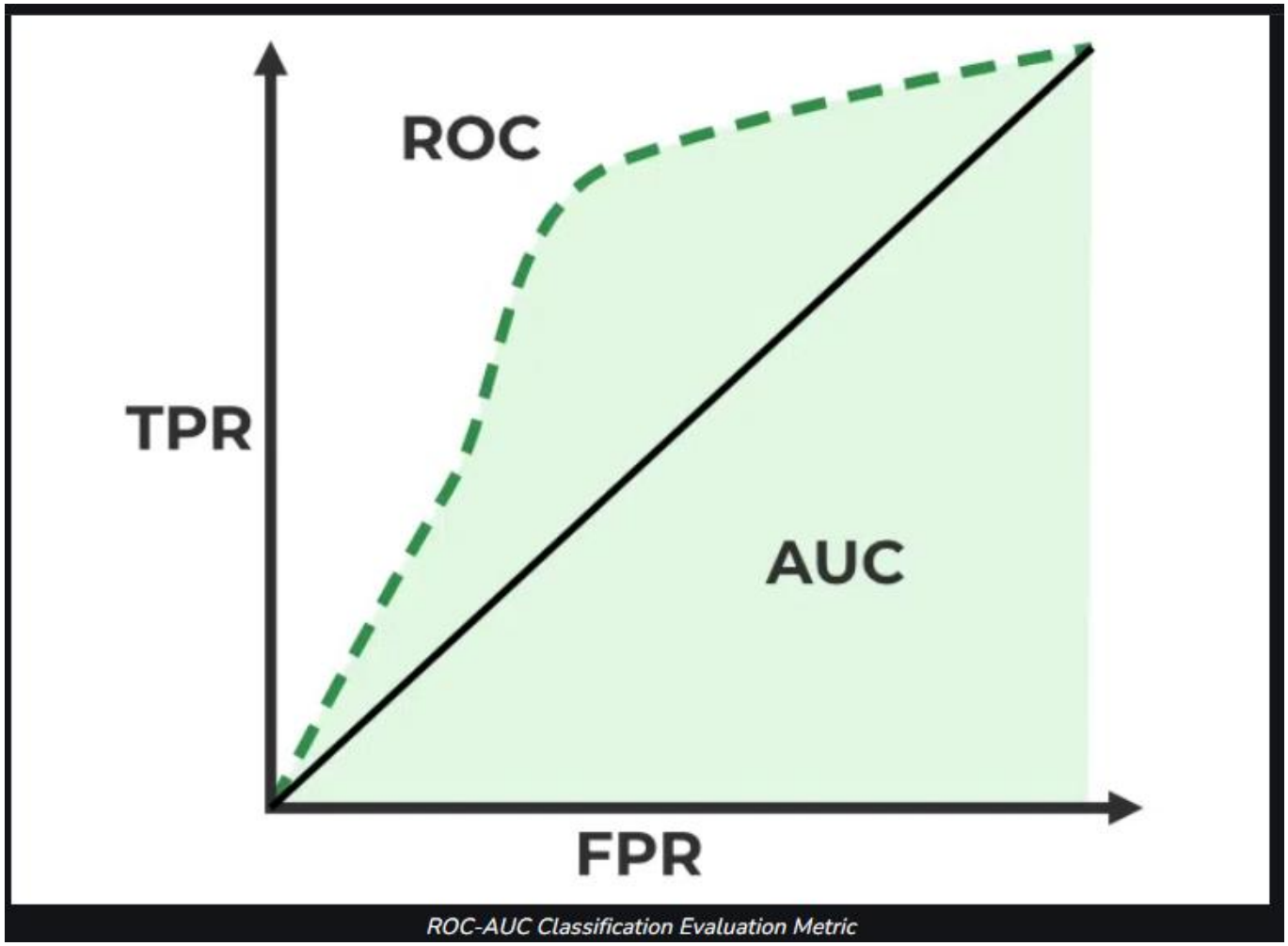
AUC-ROC curve

DATA SCIENCE

- TPR and FPR
- This is the most common definition that you would have encountered when you would Google AUC-ROC
- Basically, the ROC curve is a graph that shows the performance of a classification model at all possible thresholds(threshold is a particular value beyond which you say a point belongs to a particular class).
- The curve is plotted between two parameters

- TPR – True Positive Rate
- FPR – False Positive Rate

AUC-ROC curve



Absolute Error Definition

- Absolute error is the difference between measured or inferred value and the actual value of a quantity.
- The absolute error is inadequate since it does not give any details regarding the importance of the error.
- While measuring distances between cities kilometers apart, an error of a few centimeters is negligible and is irrelevant.
- Consider another case where an error of centimeters when measuring small machine parts is a very significant error.
- Both the errors are in the order of centimeters, but the second error is more severe than the first one.

Absolute Error Definition

DATA SCIENCE

- For example, 24.13 is the actual value of a quantity and 25.09 is the measure or inferred value, then the absolute error will be:
- Absolute Error = 25.09 – 24.13
- = 0.86
- Most of the time it is sufficient to record only two decimal digits of the absolute error.
- Thus, it is sufficient to state that the absolute error of the approximation 4.55 to the correct value 4.538395 is 0.012.

Relative Error

DATA SCIENCE

- The relative error is defined as the ratio of the absolute error of the measurement to the actual measurement.
- Using this method, we can determine the magnitude of the absolute error in terms of the actual size of the measurement.
- If the true measurement of the object is not known, then the relative error can be found using the measured value.
- The relative error gives an indication of how good measurement is relative to the size of the object being measured.

Relative Error

DATA SCIENCE

- If x is the actual value of a quantity, x0 is the measured value of the quantity and Δx is the absolute error, then the relative error can be measured using the below formula.

- Relative error = (x0-x)/x = (Δx)/x

- An important note that relative errors are dimensionless.
- When writing relative errors,it is usual to multiply the fractional error by 100 and express it as a percentage.

Relative Error

DATA SCIENCE

- Example 1:
- Find the absolute and relative errors of the approximation 125.67 to the value 119.66.
- Solution:
- Absolute error = |125.67-119.66|=6.01
- Relative error = |125.67-119.66|/119.66 = 0.05022

Mean Absolute Error

DATA SCIENCE

- The mean absolute error is the average of all absolute errors of the data collected.
- It is abbreviated as MAE (Mean Absolute Error).
- It is obtained by dividing the sum of all the absolute errors with the number of errors.
- The formula for MAE is:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_i - x|$$

Here,

$|x_i - x|$ = absolute errors

n = number of errors