Data Analysis For Energy Management

Sahaana Venkat | 04-09-2021

Data All Around

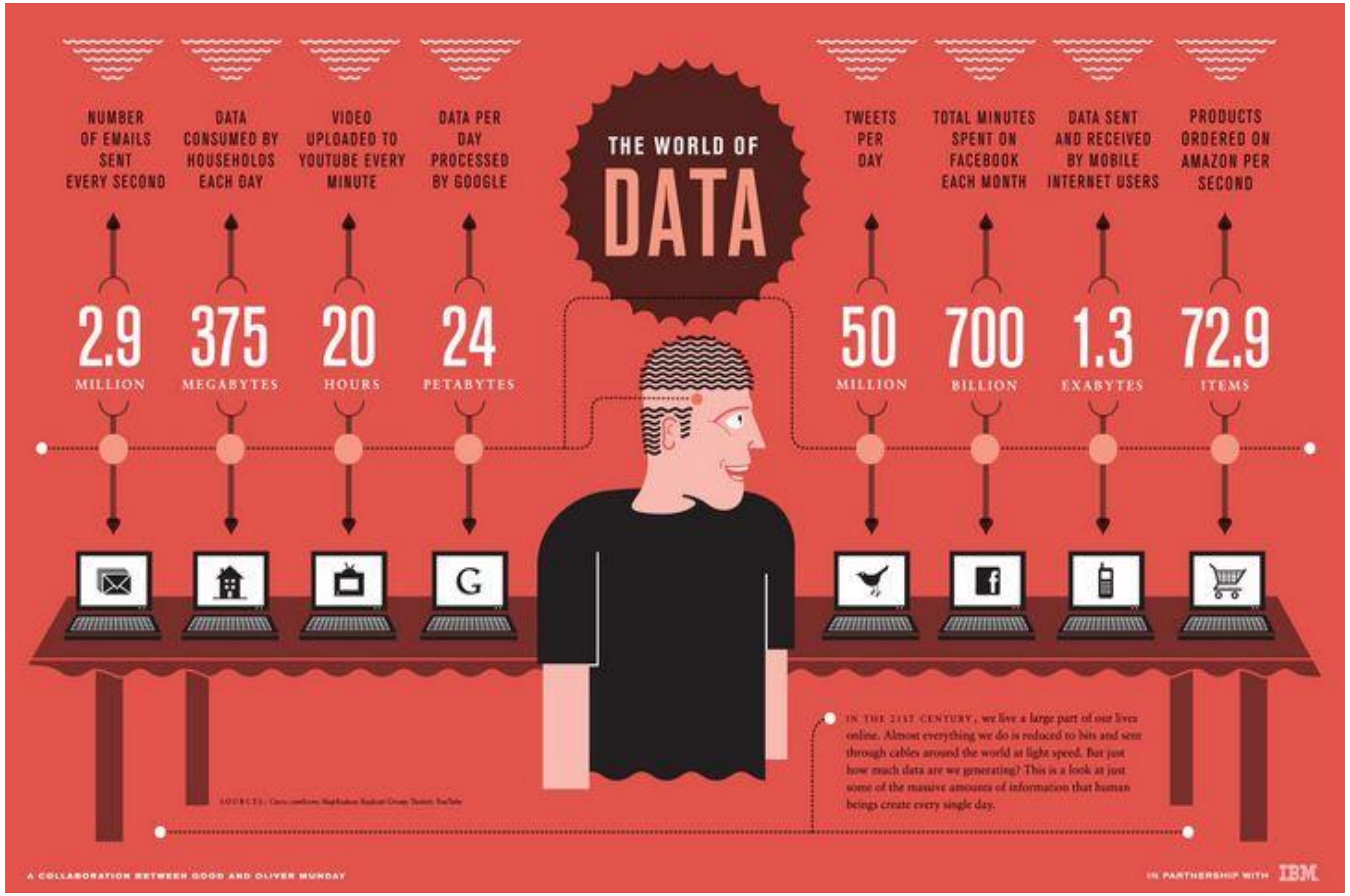
Lots of data is being collected and warehoused

- Web data, e-commerce
- Financial transactions, bank/credit transactions
- Online trading and purchasing
- Social Network

How Much Data Do We have?

- Google processes 20 PB a day (2008)
- Facebook has 60 TB of daily logs
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- 1000 genomes project: 200 TB

- Cost of 1 TB of disk: $35
- Time to read 1 TB disk: 3 hrs (100 MB/s)


THE WORLD OF DATA
NUMBER OF EMAILS SENT EVERY SECOND
2.9 MILLION
DATA CONSUMED BY HOUSEHOLDS EACH DAY
375 MEGABYTES
VIDEO UPLOADED TO YOUTUBE EVERY MINUTE
20 HOURS
DATA PER DAY PROCESSED BY GOOGLE
24 PETABYTES
TWEETS PER DAY
50 MILLION
TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH
700 BILLION
DATA SENT AND RECEIVED BY MOBILE INTERNET USERS
1.3 EXABYTES
PRODUCTS ORDERED ON AMAZON PER SECOND
72.9 ITEMS
G
SOURCES: Cisco, comScore, [illegible] Radicati Group, Twitter, YouTube
A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY
IN PARTNERSHIP WITH IBM

IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.

Big Data

- Big Data is any data that is expensive to manage and hard to extract value from

- **Volume**

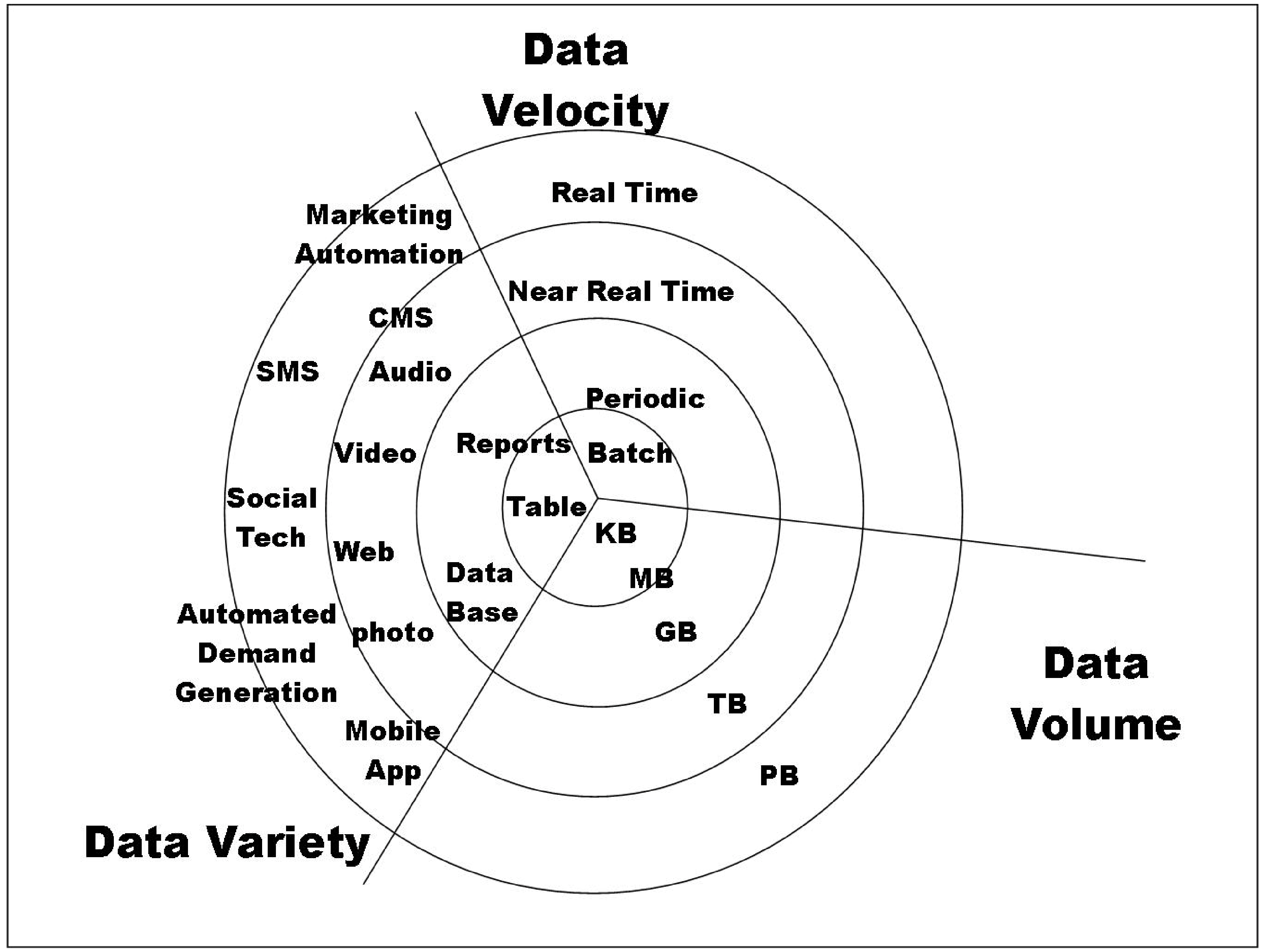
  The size of the data

- **Velocity**

  The latency of data processing relative to the growing demand for interactivity

- **Variety and Complexity**

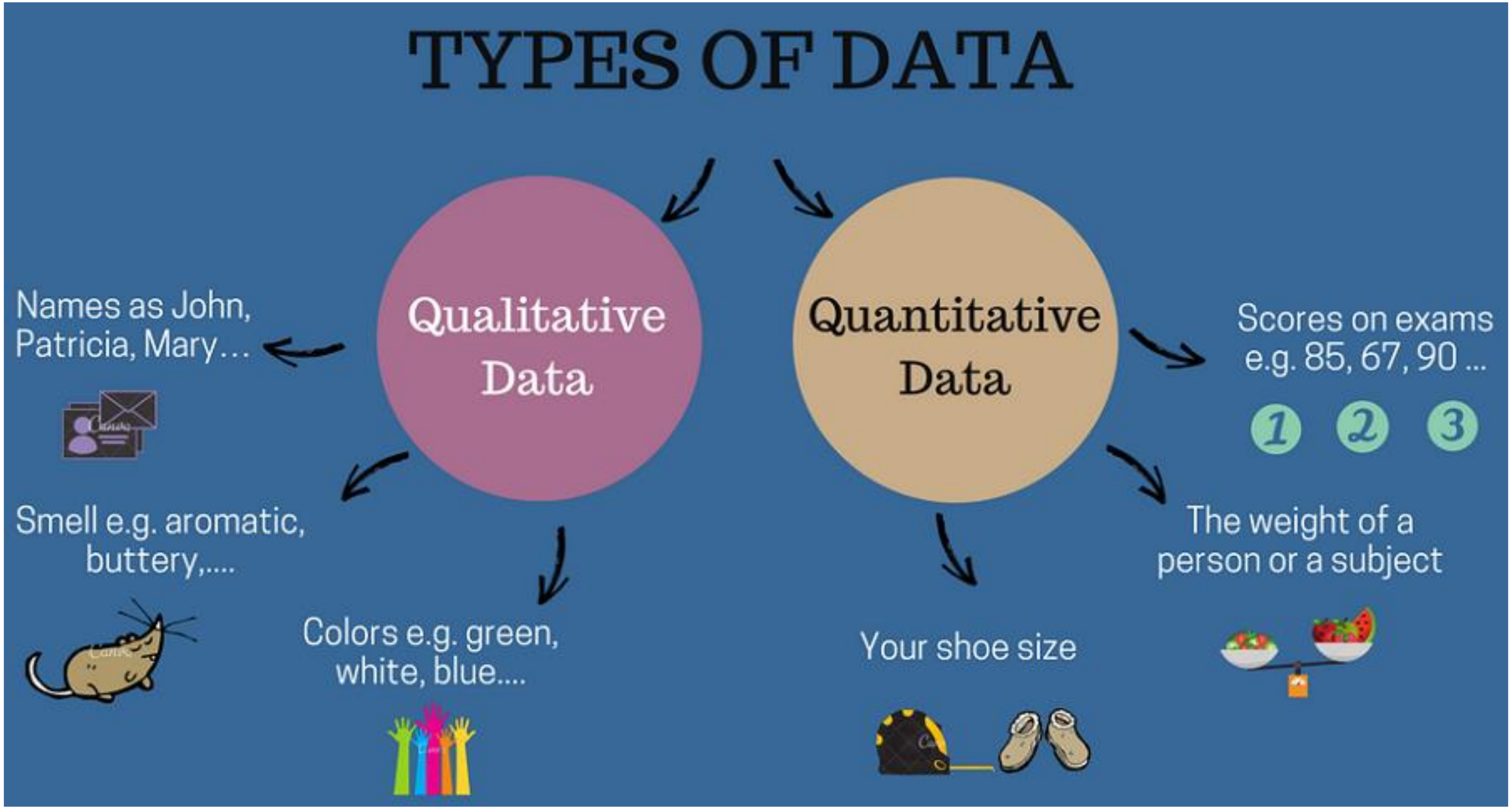
  The diversity of sources, formats, quality, structures.



Types of Data We Have

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
- Social Network, Semantic Web (RDF), …
- Streaming Data
- You can afford to scan the data once

Types Of Data


TYPES OF DATA
Qualitative Data
Quantitative Data
Names as John, Patricia, Mary…
Smell e.g. aromatic, buttery,….
Colors e.g. green, white, blue….
Scores on exams e.g. 85, 67, 90 ...
1
2
3
The weight of a person or a subject
Your shoe size

Data Analysis

- Data Analysis

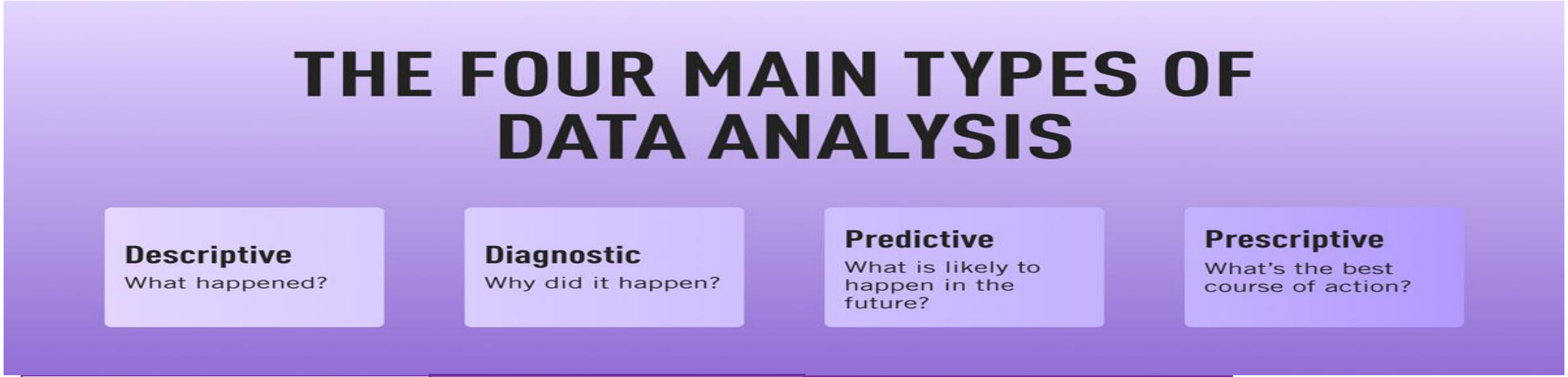
  Data analysis is defined as a process of cleaning, transforming, and modeling data to discover useful information for business decision-making. The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis.

- Tools Used for Data Analysis

R
sas
Python
Java
SQL
MATLAB

Data Analysis Tools

Type Of Data Analysis



THE FOUR MAIN TYPES OF DATA ANALYSIS

Descriptive
What happened?

Diagnostic
Why did it happen?

Predictive
What is likely to happen in the future?

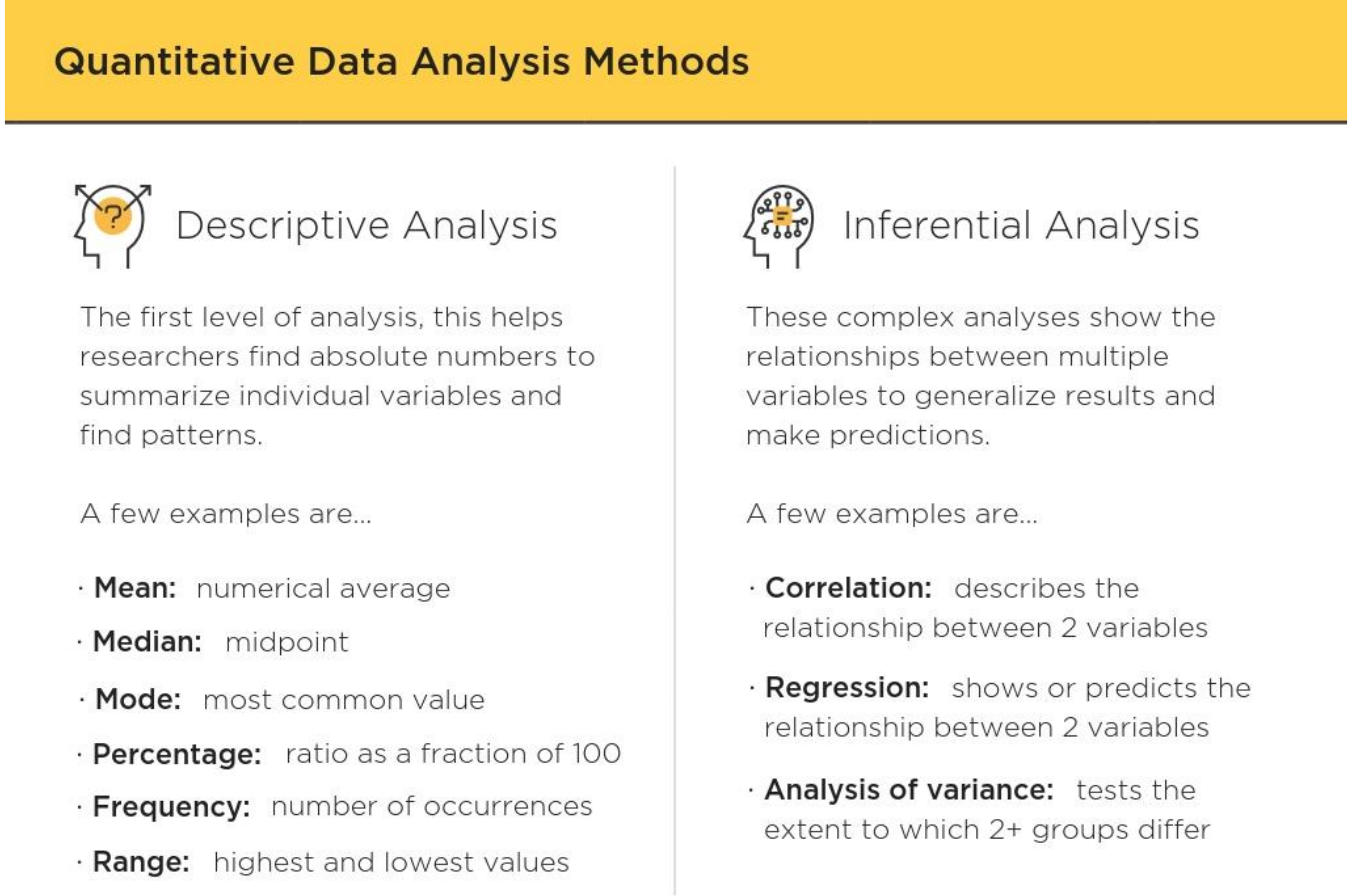
Prescriptive
What's the best course of action?

- KPI dashboards
- Monthly revenue reports
- Sales leads overview

- A freight company investigating the cause of slow shipments in a certain region
- A SaaS company drilling down to determine which marketing activities increased trials

- Risk Assessment
- Sales Forecasting
- Using customer segmentation to determine which leads have the best chance of converting
- Predictive analytics in customer success teams

Quantitative Data Analysis Methods



Quantitative Data Analysis Methods

Descriptive Analysis

The first level of analysis, this helps researchers find absolute numbers to summarize individual variables and find patterns.

A few examples are...

- **Mean:** numerical average
- **Median:** midpoint
- **Mode:** most common value
- **Percentage:** ratio as a fraction of 100
- **Frequency:** number of occurrences
- **Range:** highest and lowest values

Inferential Analysis

These complex analyses show the relationships between multiple variables to generalize results and make predictions.

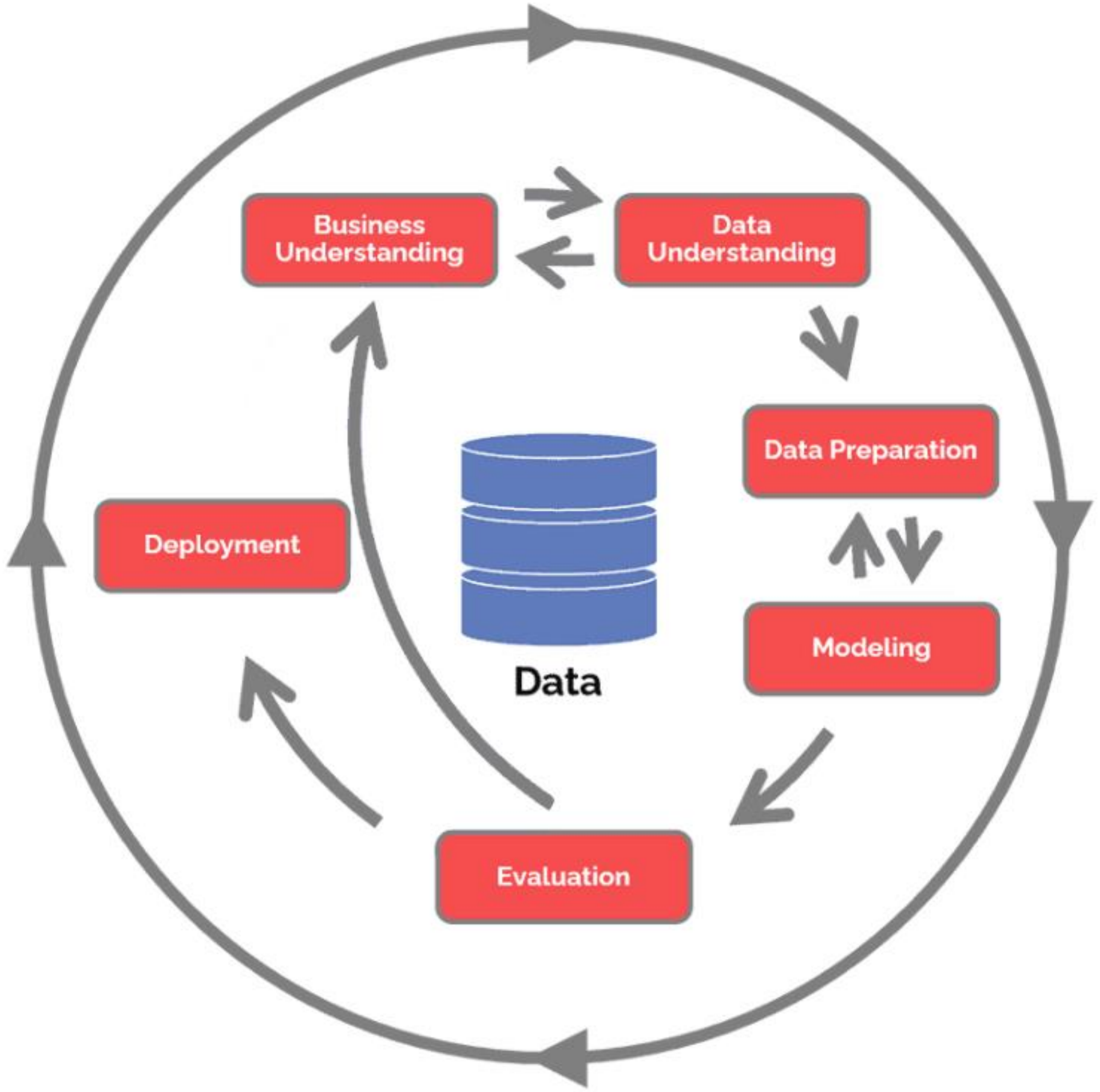
A few examples are...

- **Correlation:** describes the relationship between 2 variables
- **Regression:** shows or predicts the relationship between 2 variables
- **Analysis of variance:** tests the extent to which 2+ groups differ

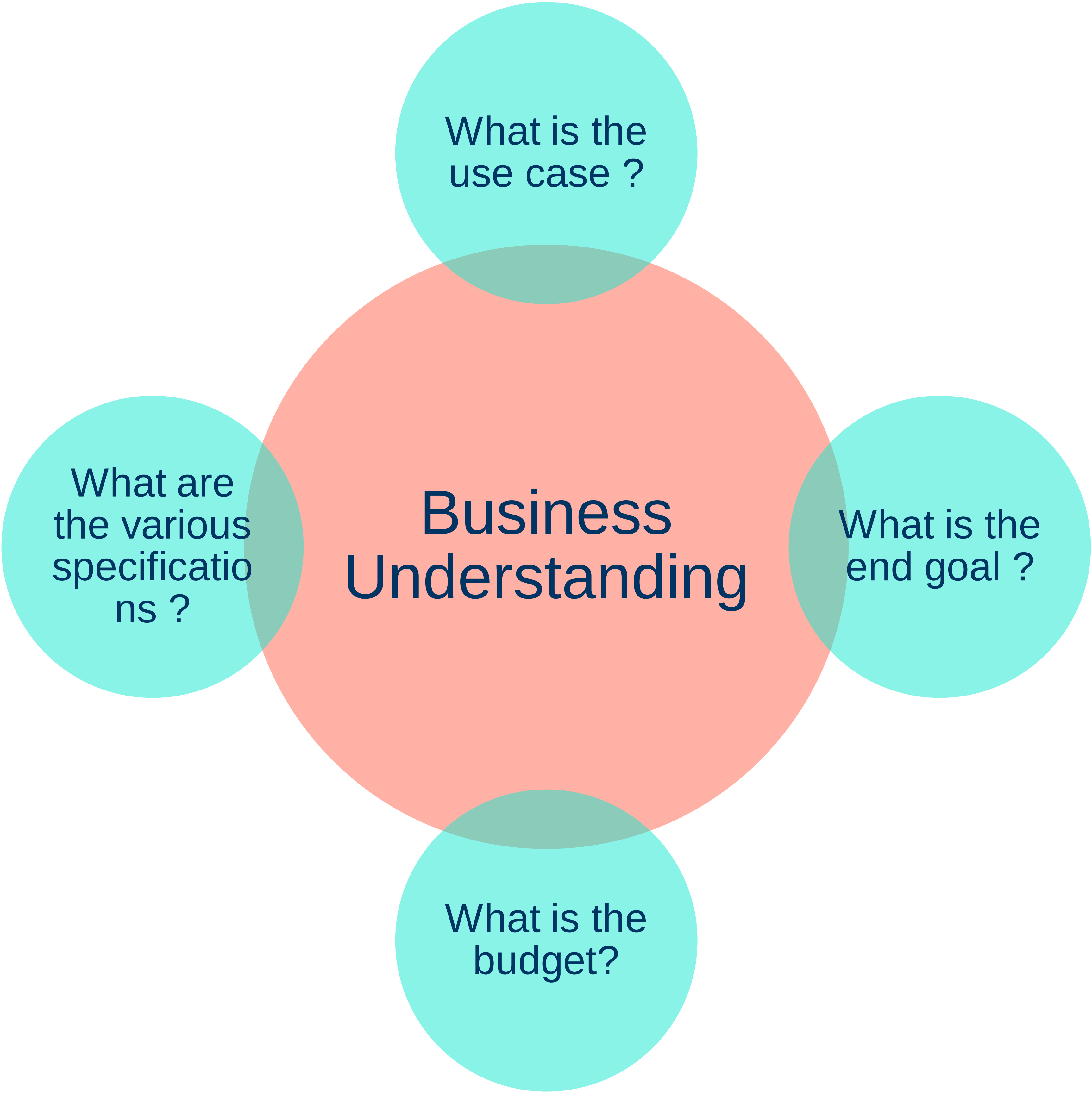
CRISP – DM Framework

The **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (**CRISP-DM**) is a process model with six phases that naturally describes the data science life cycle. It's like a set of guardrails to help you plan, organize, and implement your data science (or machine learning) project.

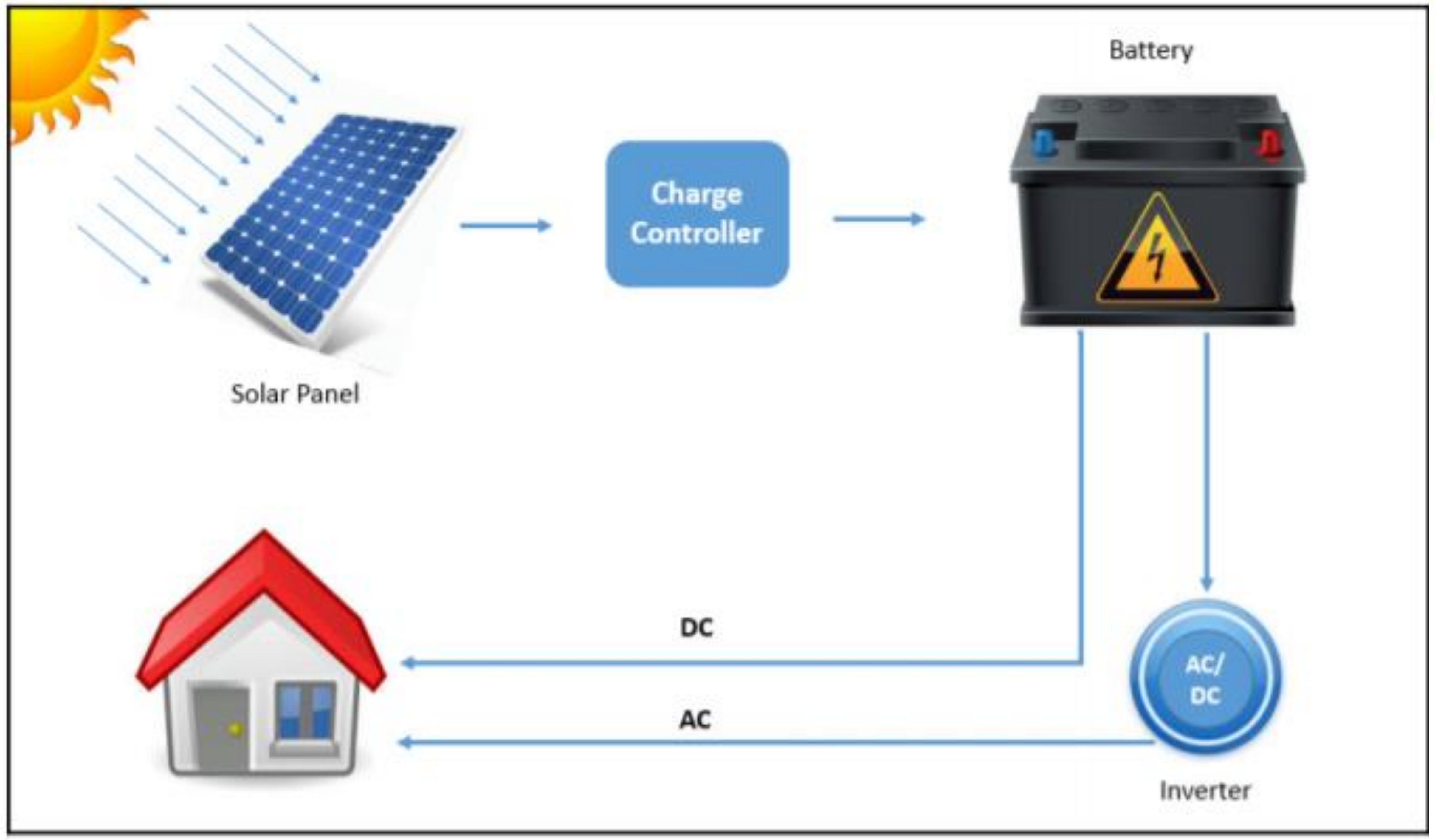
1. Business understanding – What does the business need?
2. Data understanding – What data do we have / need? Is it clean?
3. Data preparation – How do we organize the data for modeling?
4. Modeling – What modeling techniques should we apply?
5. Evaluation – Which model best meets the business objectives?
6. Deployment – How do stakeholders access the results?

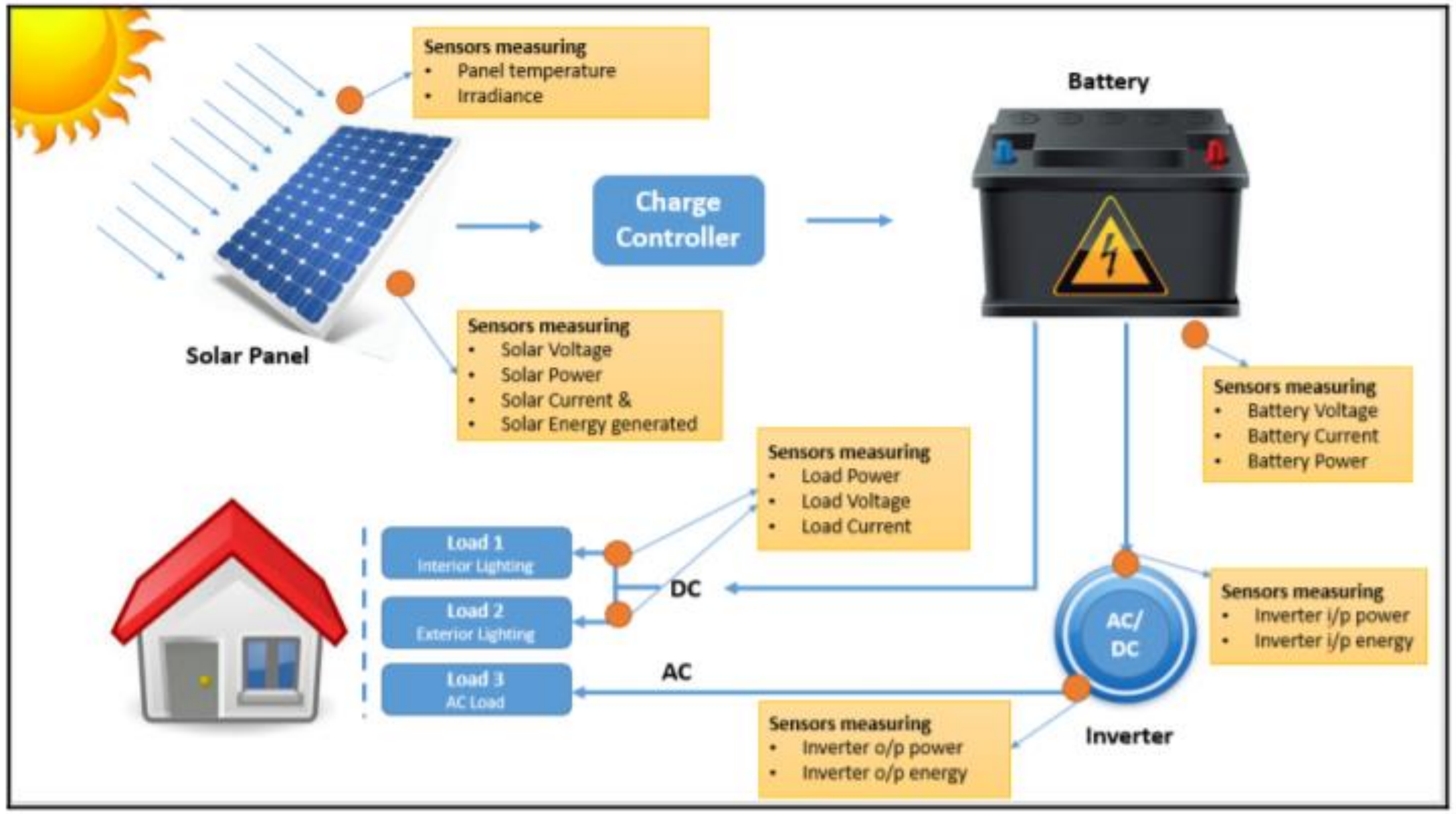


Business Understanding



Solar Energy Setup In Africa

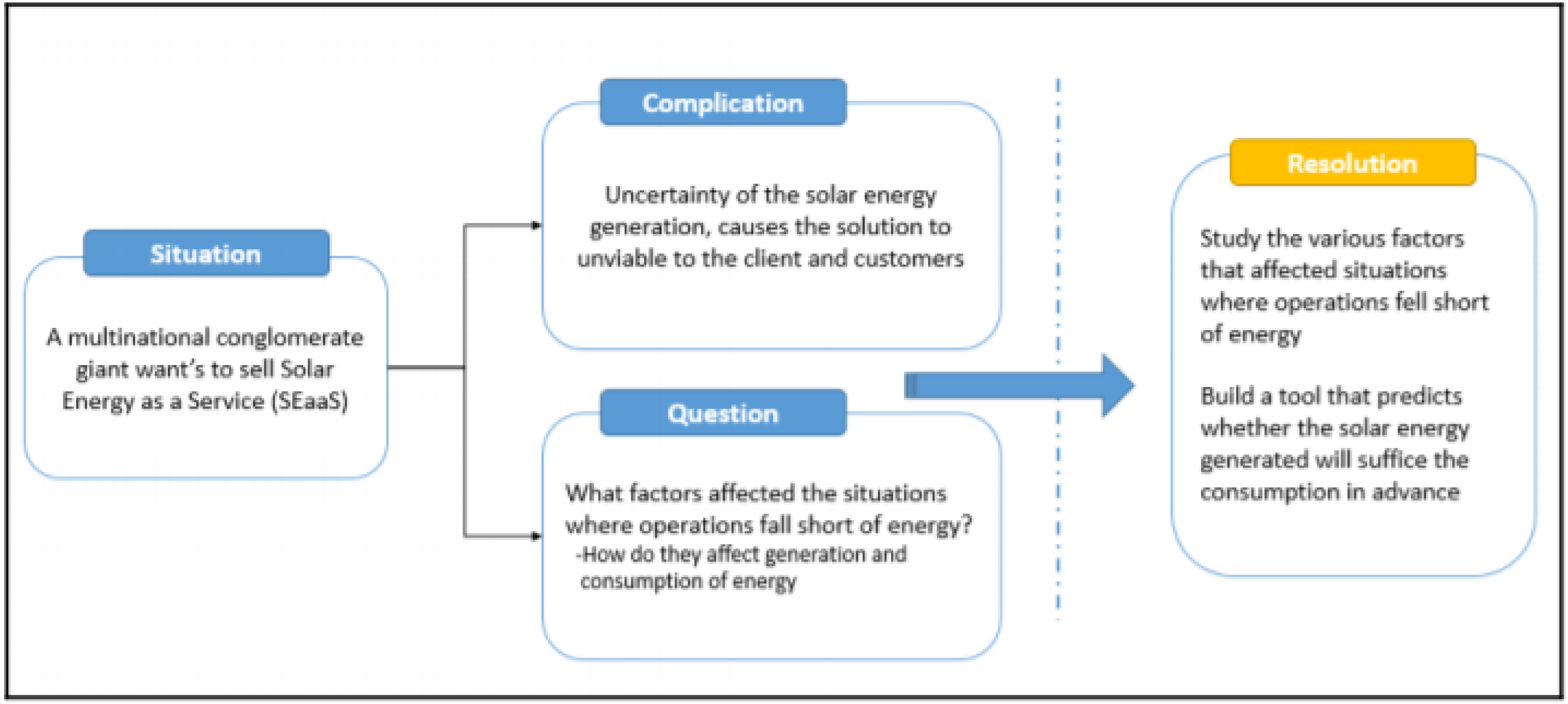




Problem To Identify

- Whether solar panel is enough to give current to whole city ?
- What happens when the climate is bad ?
- Whether we have to make the solar panel backup more or we have to keep the diesel generator as a backup ?
- If we keep the diesel generator as a backup do we have to hire the person to maintain the generator ?

Building the SCQ: Situation – Complication – Question



Situation

A multinational conglomerate giant want's to sell Solar Energy as a Service (SEaaS)

Complication

Uncertainty of the solar energy generation, causes the solution to unviable to the client and customers

Question

What factors affected the situations where operations fall short of energy?

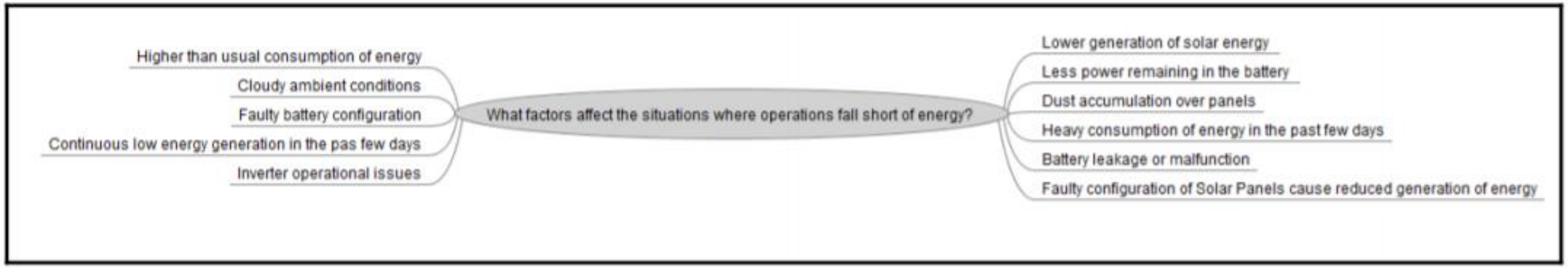
-How do they affect generation and consumption of energy

Resolution

Study the various factors that affected situations where operations fell short of energy

Build a tool that predicts whether the solar energy generated will suffice the consumption in advance

Designing the approach



What factors affect the situations where operations fall short of energy?

Higher than usual consumption of energy

Cloudy ambient conditions

Faulty battery configuration

Continuous low energy generation in the pas few days

Inverter operational issues

Lower generation of solar energy

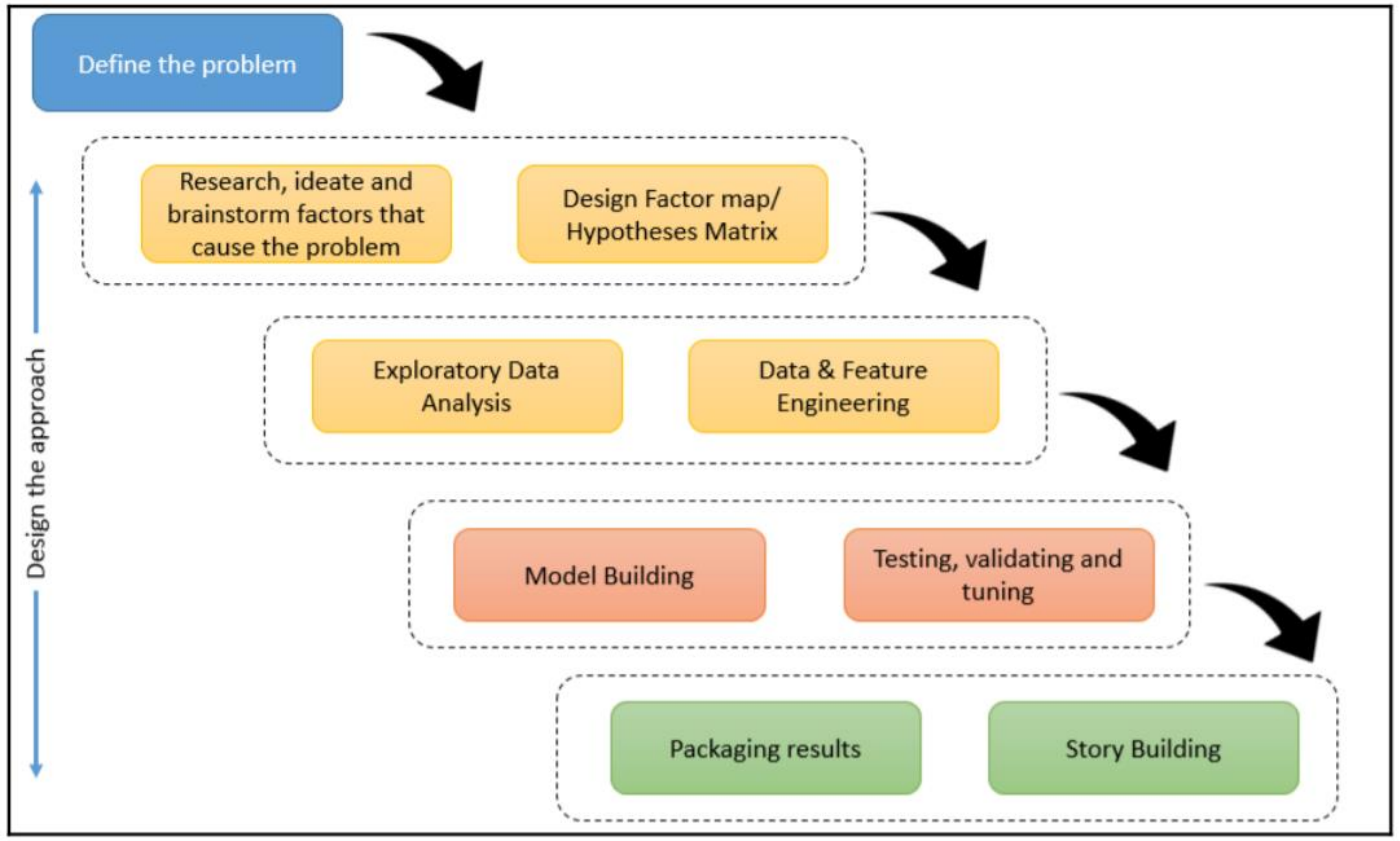
Less power remaining in the battery

Dust accumulation over panels

Heavy consumption of energy in the past few days

Battery leakage or malfunction

Faulty configuration of Solar Panels cause reduced generation of energy



Define the problem

Research, ideate and brainstorm factors that cause the problem

Design Factor map/ Hypotheses Matrix

Exploratory Data Analysis

Data & Feature Engineering

Model Building

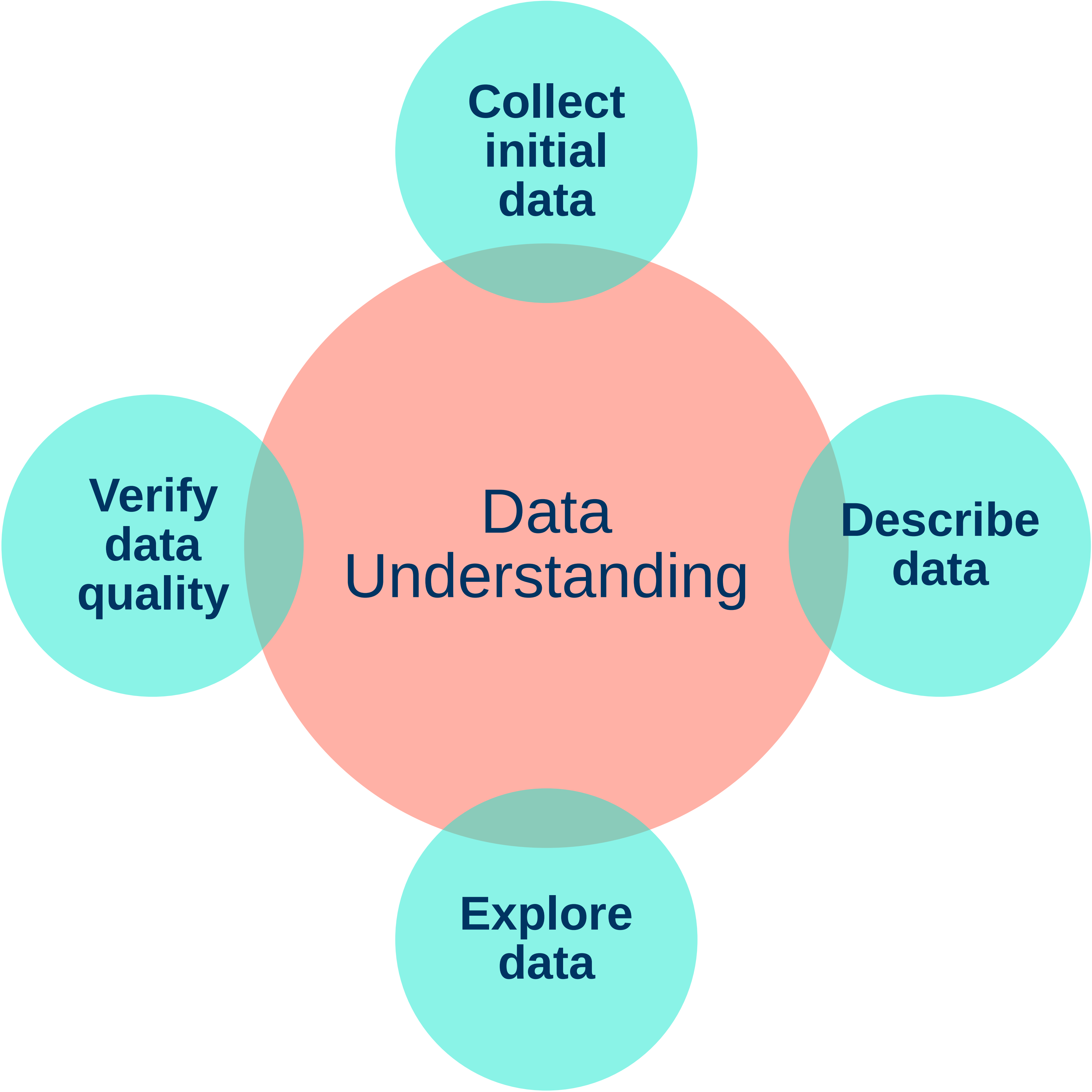
Testing, validating and tuning

Packaging results

Story Building

Design the approach

Data Understanding



Data Preparation

Data Cleaning

- Correcting inconsistent data by filling out missing values and smoothing out noisy data.

Data Transformation

- It involves normalization, transformation and aggregation of data using ETL methods.

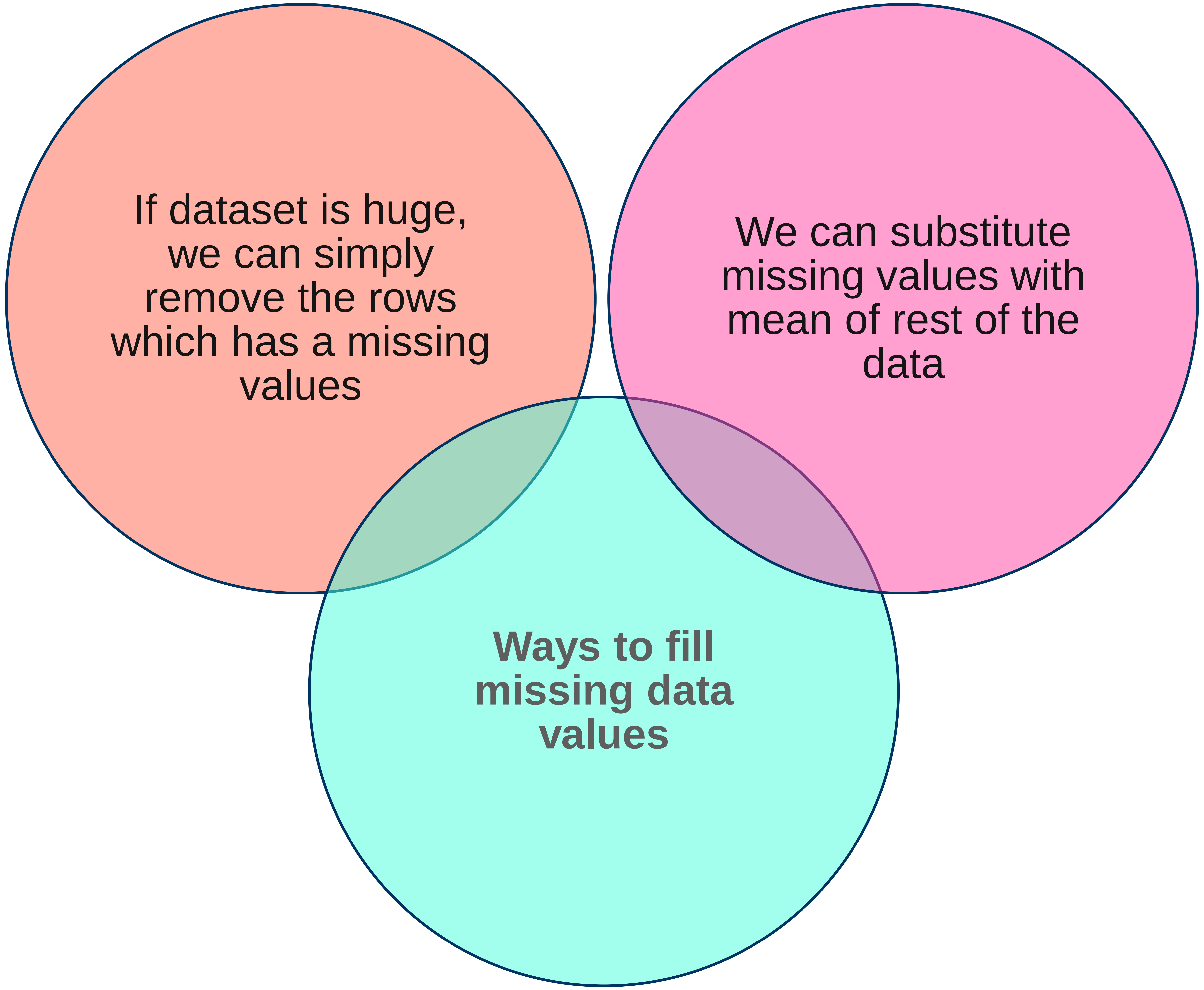
Data Integration

- Resolving any conflicts in the data and handling redundancies.

Data Reduction

- Using various strategies, reducing the size of data but yielding the same outcome.

Data Preparation



If dataset is huge, we can simply remove the rows which has a missing values

We can substitute missing values with mean of rest of the data

Ways to fill missing data values

Model Planning

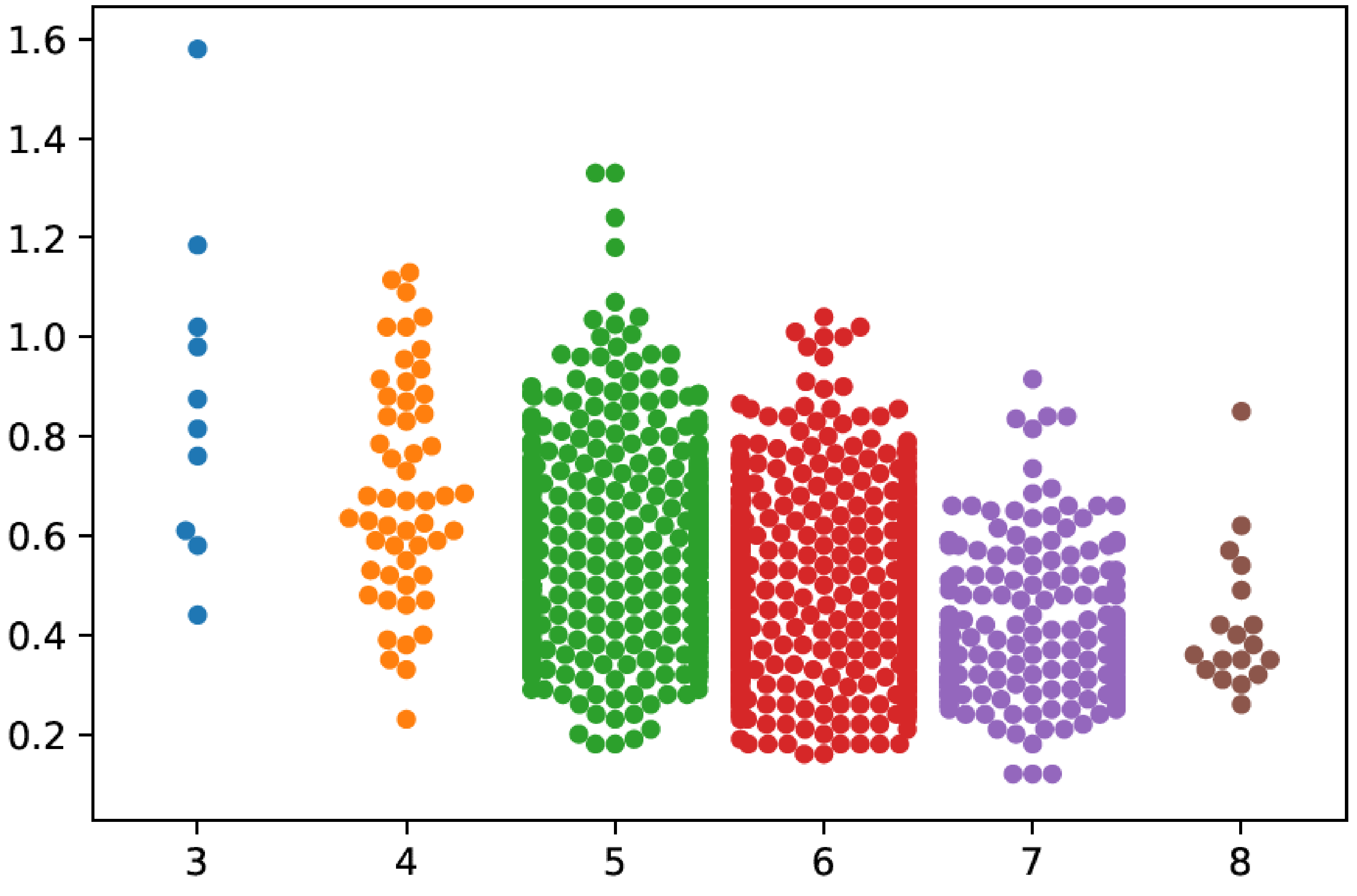
Exploratory Data Analysis (EDA) :

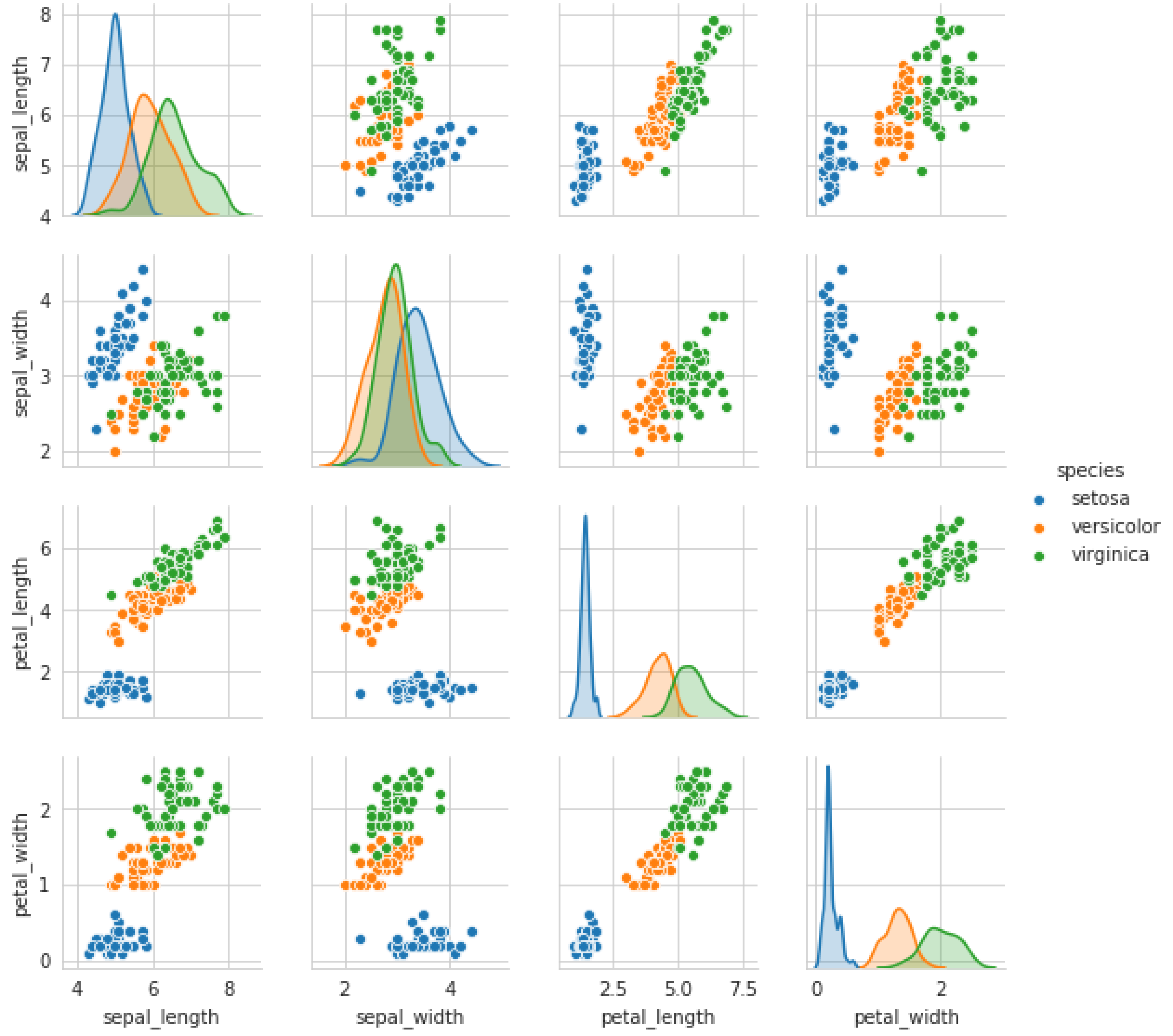
In model Planning the main step is Exploratory Data Analysis (EDA) to understand the relation between variables and to see what the data can tell us.

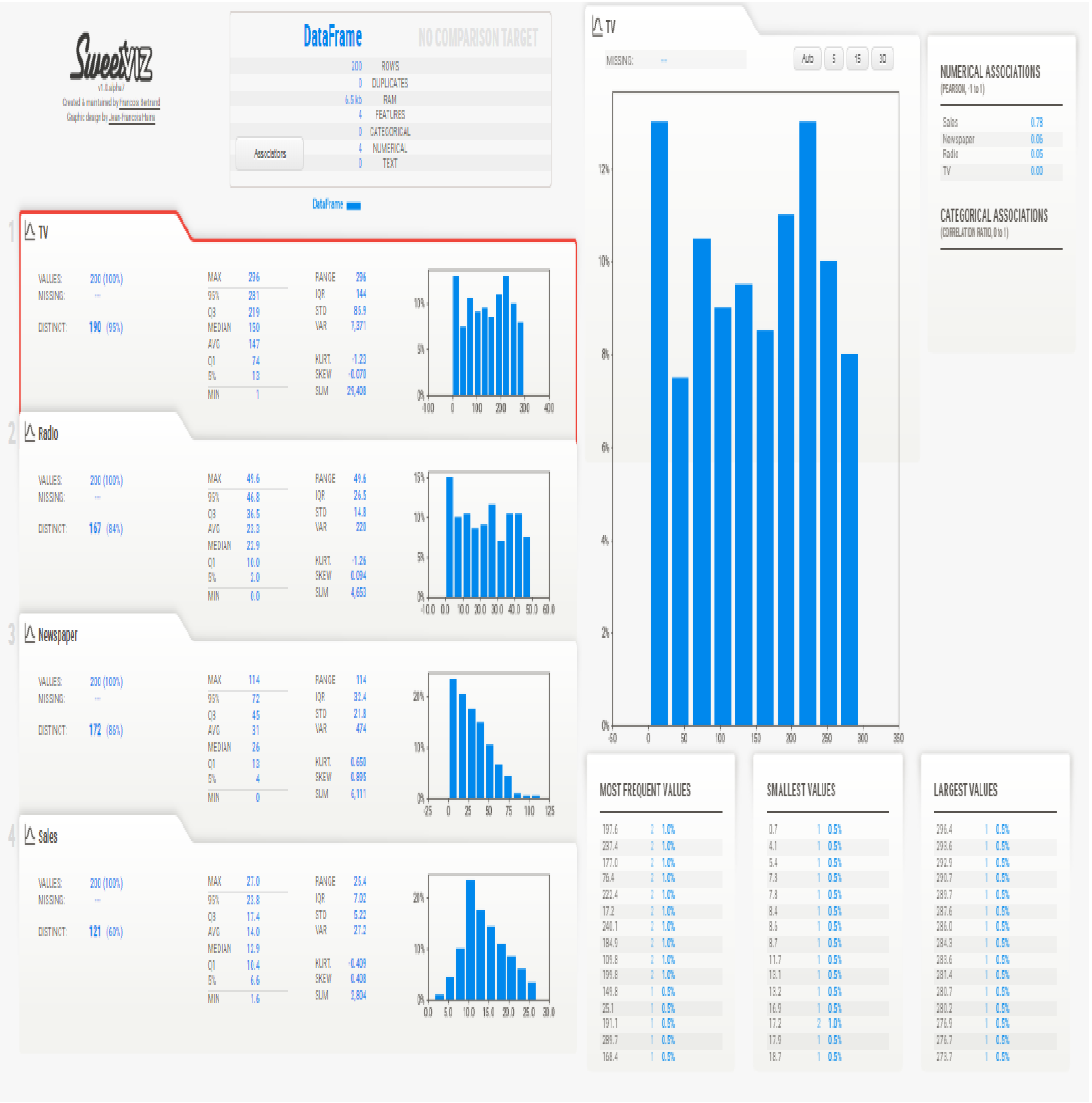
Definition : Deeper analysis of dataset to better understand the data.

Goals :

- Know the datatypes and answer questions with the data
- Understand how data is distributed
- Identify outliers
- Identify patterns , if any






SweetVIZ
v1.0 alpha7
Created & maintained by Francois Bertrand
Graphic design by Jean-Francois Hains
Associations
DataFrame
Auto
5
15
30

DataFrame NO COMPARISON TARGET

200	ROWS
0	DUPLICATES
6.5 kb	RAM
4	FEATURES
0	CATEGORICAL
4	NUMERICAL
0	TEXT

1 TV

VALUES: 200 (100%)
MISSING: ---
DISTINCT: 190 (95%)

MAX	296	RANGE	296
95%	281	IQR	144
Q3	219	STD	85.9
MEDIAN	150	VAR	7,371
AVG	147		
Q1	74	KURT.	-1.23
5%	13	SKEW	-0.070
MIN	1	SUM	29,408

2 Radio

VALUES: 200 (100%)
MISSING: ---
DISTINCT: 167 (84%)

MAX	49.6	RANGE	49.6
95%	46.8	IQR	26.5
Q3	36.5	STD	14.8
AVG	23.3	VAR	220
MEDIAN	22.9		
Q1	10.0	KURT.	-1.26
5%	2.0	SKEW	0.094
MIN	0.0	SUM	4,653

3 Newspaper

VALUES: 200 (100%)
MISSING: ---
DISTINCT: 172 (86%)

MAX	114	RANGE	114
95%	72	IQR	32.4
Q3	45	STD	21.8
AVG	31	VAR	474
MEDIAN	26		
Q1	13	KURT.	0.650
5%	4	SKEW	0.895
MIN	0	SUM	6,111

4 Sales

VALUES: 200 (100%)
MISSING: ---
DISTINCT: 121 (60%)

MAX	27.0	RANGE	25.4
95%	23.8	IQR	7.02
Q3	17.4	STD	5.22
AVG	14.0	VAR	27.2
MEDIAN	12.9		
Q1	10.4	KURT.	-0.409
5%	6.6	SKEW	0.408
MIN	1.6	SUM	2,804

TV

MISSING: ---

NUMERICAL ASSOCIATIONS
(PEARSON, -1 to 1)

Sales	0.78
Newspaper	0.06
Radio	0.05
TV	0.00

CATEGORICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)

MOST FREQUENT VALUES

197.6	2	1.0%
237.4	2	1.0%
177.0	2	1.0%
76.4	2	1.0%
222.4	2	1.0%
17.2	2	1.0%
240.1	2	1.0%
184.9	2	1.0%
109.8	2	1.0%
199.8	2	1.0%
149.8	1	0.5%
25.1	1	0.5%
191.1	1	0.5%
289.7	1	0.5%
168.4	1	0.5%

SMALLEST VALUES

0.7	1	0.5%
4.1	1	0.5%
5.4	1	0.5%
7.3	1	0.5%
7.8	1	0.5%
8.4	1	0.5%
8.6	1	0.5%
8.7	1	0.5%
11.7	1	0.5%
13.1	1	0.5%
13.2	1	0.5%
16.9	1	0.5%
17.2	2	1.0%
17.9	1	0.5%
18.7	1	0.5%

LARGEST VALUES

296.4	1	0.5%
293.6	1	0.5%
292.9	1	0.5%
290.7	1	0.5%
289.7	1	0.5%
287.6	1	0.5%
286.0	1	0.5%
284.3	1	0.5%
283.6	1	0.5%
281.4	1	0.5%
280.7	1	0.5%
280.2	1	0.5%
276.9	1	0.5%
276.7	1	0.5%
273.7	1	0.5%

Packages Used For Data Analysis In Python

NumPy

pandas

SciPy

matplotlib

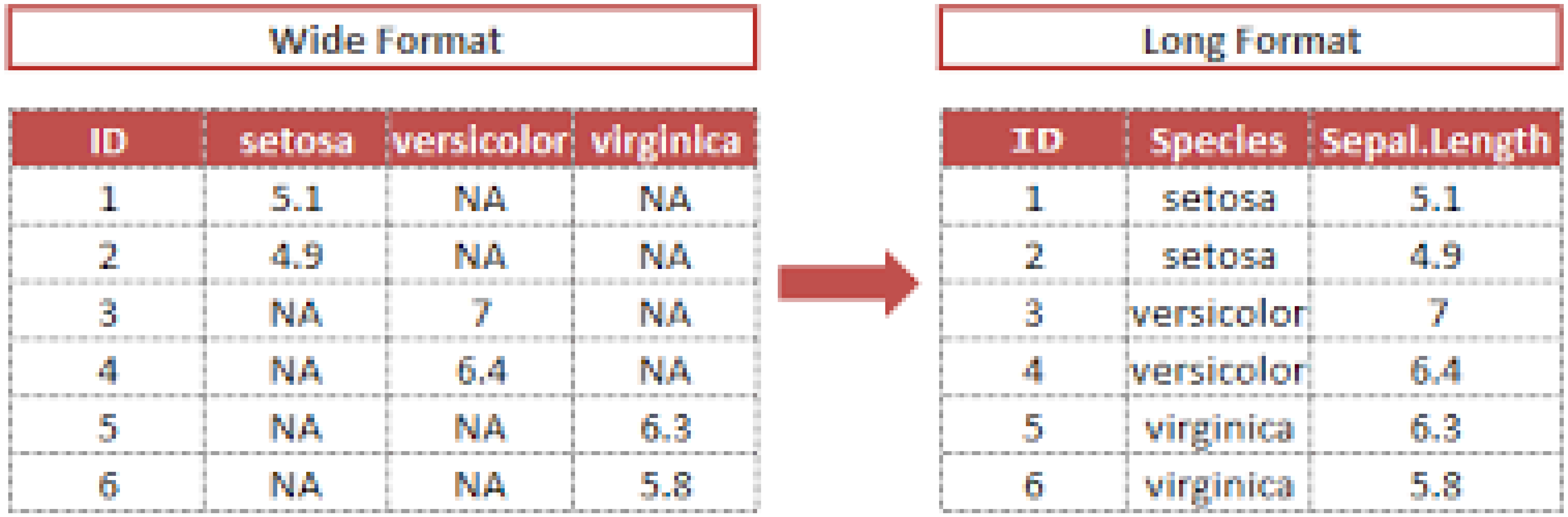
seaborn

plotly

IDE Used For Python

Python Code For Data Analysis

Wide And Long Format



Wide Format

ID	setosa	versicolor	virginica
1	5.1	NA	NA
2	4.9	NA	NA
3	NA	7	NA
4	NA	6.4	NA
5	NA	NA	6.3
6	NA	NA	5.8

Long Format

ID	Species	Sepal.Length
1	setosa	5.1
2	setosa	4.9
3	versicolor	7
4	versicolor	6.4
5	virginica	6.3
6	virginica	5.8

Data Exploration of Solar Panel

- Understand the data distribution of power, voltage, current and energy generated based on sun.

- Behavior of all the above parameters were in sync with sunrise and sunset.

- Cumulative energy generation trend for a sample day across time and also across the time period at a day level and found that energy generation increases almost linearly during daytime (6 A.M. to 6 P.M.)

- Around 8-11 units of energy are generated on a daily basis.

- Maximum consumption was mostly seen from Load 2 and the least from Load 1.

- AC loads were more or less in the middle throughout.

- The study of total energy generated and total energy consumed on a day-to-day basis revealed that there have been enough cases where energy generated for a day was lower than the energy consumed and vice versa

Data Exploration of Solar Panel

- Except for battery voltage, the other parameters behave in accordance with the solar panel behavior.
- The voltage of the battery decreases while discharging and increases while charging that was seen consistent for a normal day during the sun hours.
- Also, the power, current, and voltage parameters for the DC loads are intermittent and completely depend on the kind of devices consuming energy
- Power is in a linear relationship with current, we can see a similar trend for both the parameters.
- Lastly, exploring the inverter parameters, we studied that the inverter power trends are very intermittent again due to the sporadic use of the AC load during the day.

Solving the Problem

- The major problem or the pain point faced by the clinic with the solar panel installations is the uncertainty about the sustenance of energy for the next day. So basically, we need to predict whether there will be enough energy for the next day or not. Finding which day there was a power outage is something we cannot directly calculate from the data. This is because, apart from the energy consumed and energy generated difference, there is also a finite amount of energy stored in the battery from the past generation.

We have a separate dataset that has recorded the power outage scenario for the same time period and location. The data is a power outage flag, say 1 or 0 (1 indicating that there was a power outage) for each day. We therefore need to build a model where we can have all the metrics or features at day level. With data at this level, we can engineer the data to predict the condition for the next day based on different features, metrics, and other data points for the current day.

Feature Engineering

First and foremost, the easiest and most important features we can create are as follows:

  - Total solar energy generated for a day
  - Total energy consumed for a day.

- The maximum value for most parameters on a day level will have relatively good variations. The minimum, however, will be 0 for most of the parameters, so let's chuck this for now.

- The duration for which these parameters were active, such as the solar current will be 0 in the absence of the sun, but will have values beyond the threshold when the sun rays are powerful enough to generate energy, is valuable.

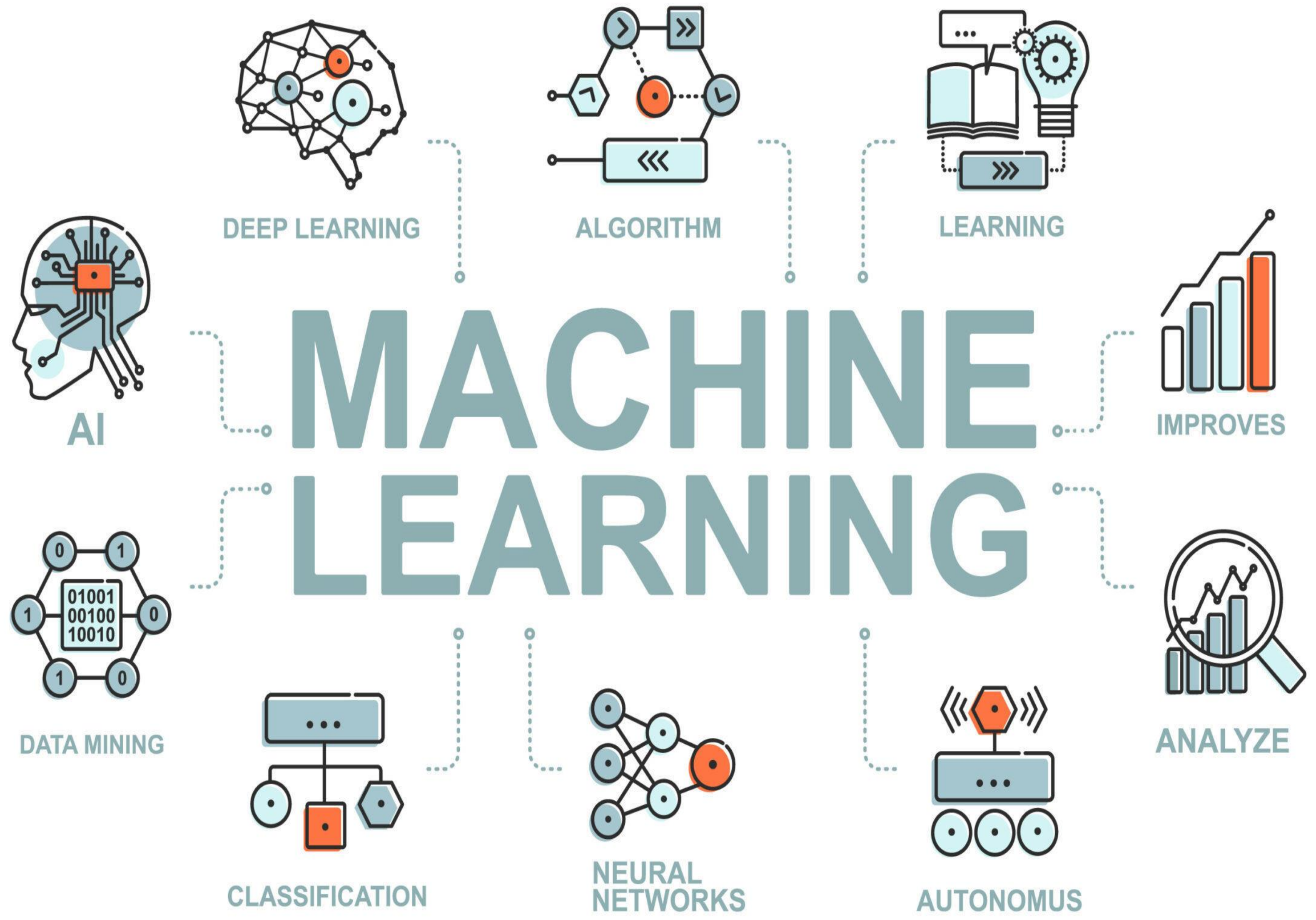
- the amount of energy present in the battery at the start of the day and end of the day will be helpful in deciding the chances for a power outage the next day. We have battery voltage values for every minute. This can be used to find out the percentage of energy left in the battery at a particular instant.

Feature Engineering

- The battery's maximum voltage is 112 and minimum is 88V.
- The battery is never allowed to drop below 30% of its capacity for performance reasons.
- Here, 112V indicates 100% energy and 88V indicates 30% energy
- We can therefore calculate the percentage of energy left in the battery at any given instant with the voltage alone.
- Solar panel is designed to work best when it receives an irradiance of 1,000 w/m2 at 25° C.
- An increase or decrease in temperature causes a small reduction in the energy generated; similarly, an irradiance value below 1,000 w/m2 will also reduce the generation of energy.
- We have chosen 5 Amperes for Solar Current, 120V for Solar Voltage, 1,000 Watts for Solar Power, 10 Amperes for battery current, and 800 Watts for battery power.

Python Code For Data Analysis

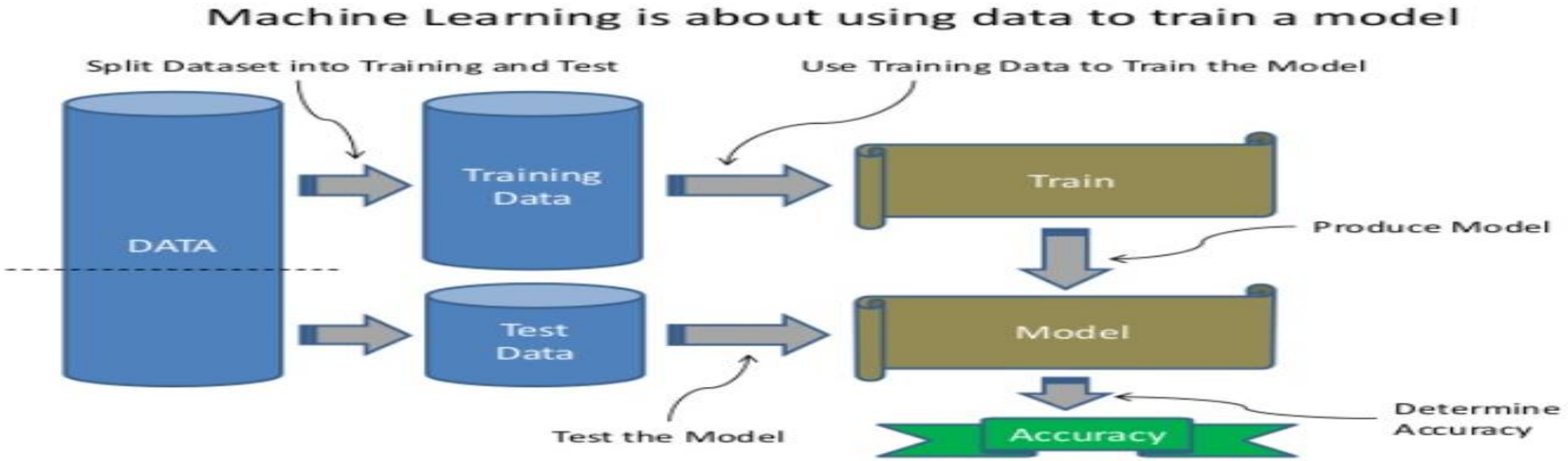
Machine Learning



Machine Learning is the 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 machine.

Machine Learning Types:

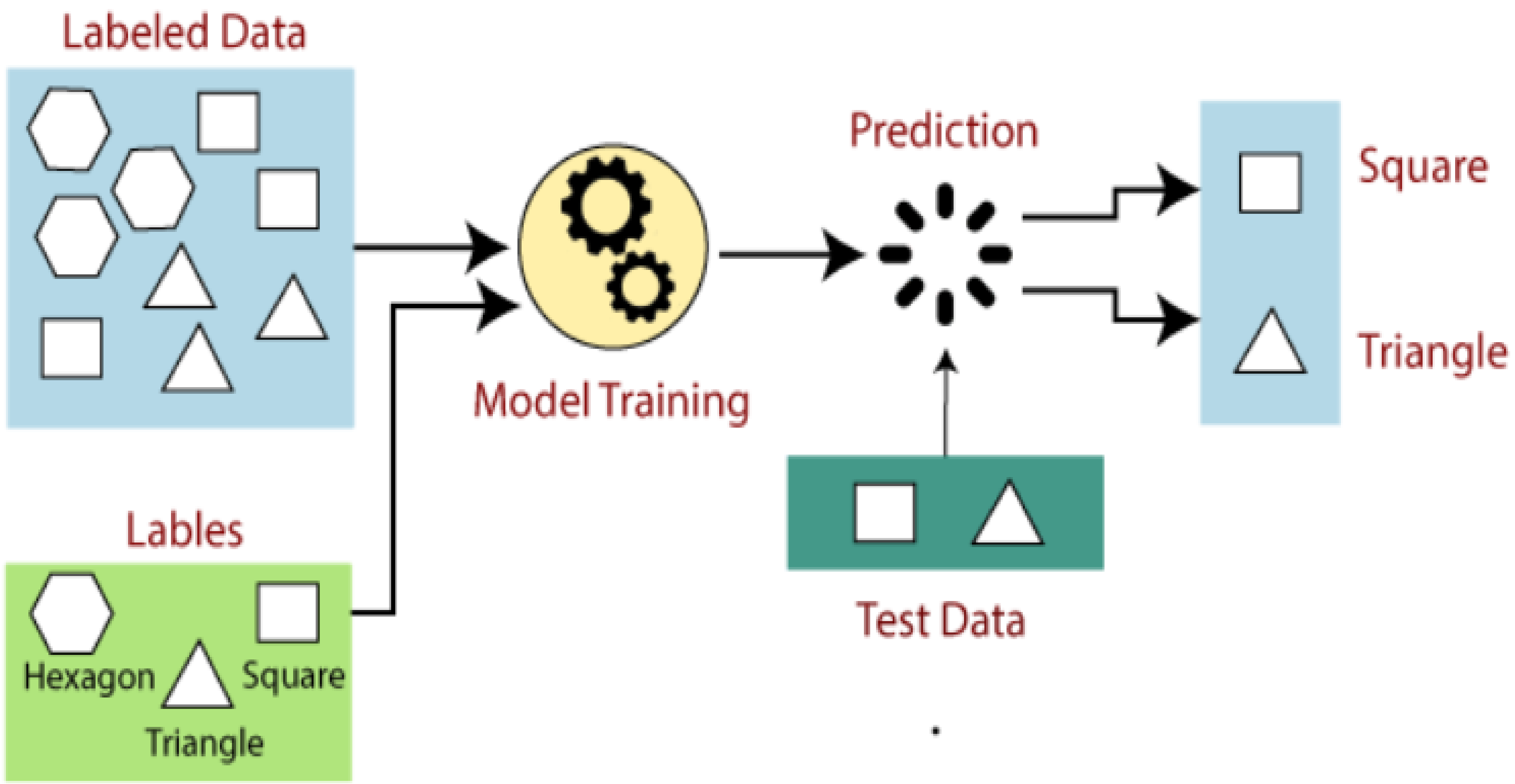
- ❑ Supervised
- ❑Unsupervised
- ❑Reinforcement


Machine Learning is about using data to train a model
Split Dataset into Training and Test
Use Training Data to Train the Model
DATA
Training Data
Train
Produce Model
Test Data
Model
Test the Model
Accuracy
Determine Accuracy

Supervised Learning

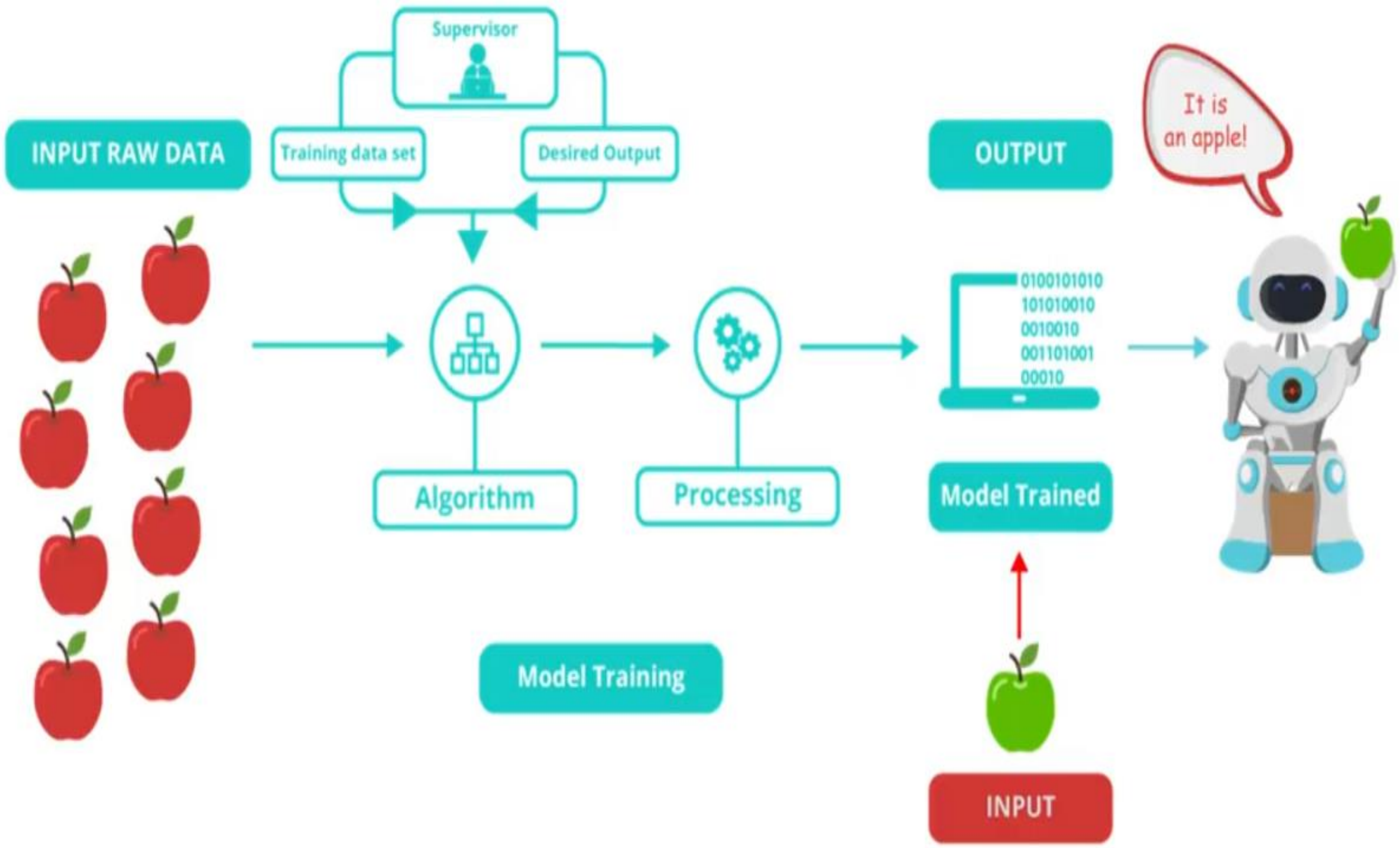
Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.



Algorithm Example : Linear Regression, Logistic Regression ,Decision Tree

Supervised Learning



General Fields:

❖ Speech recognition
❖ Weather forecasting
❖ Spam/Ham mail

Bank Fields:

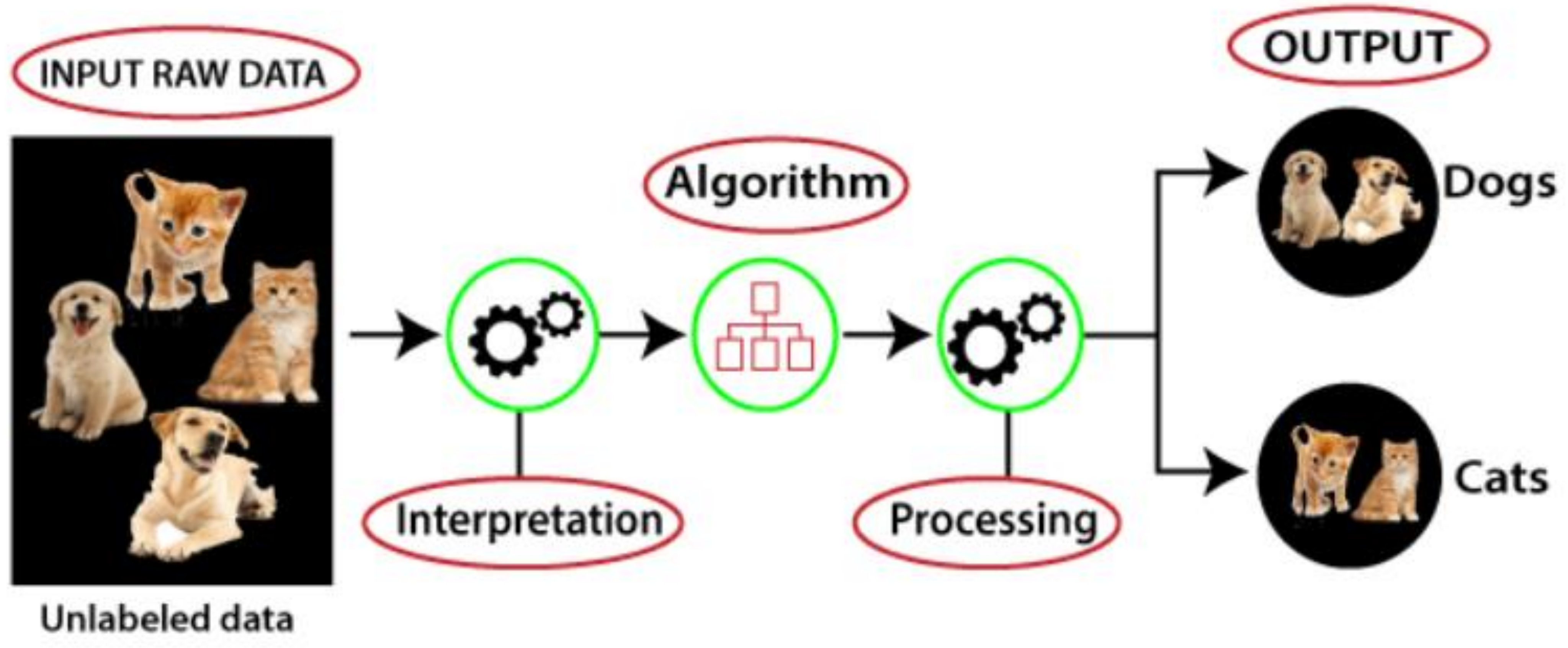
❖ Credit card

Health Care:

❖ Patient diabetes rate

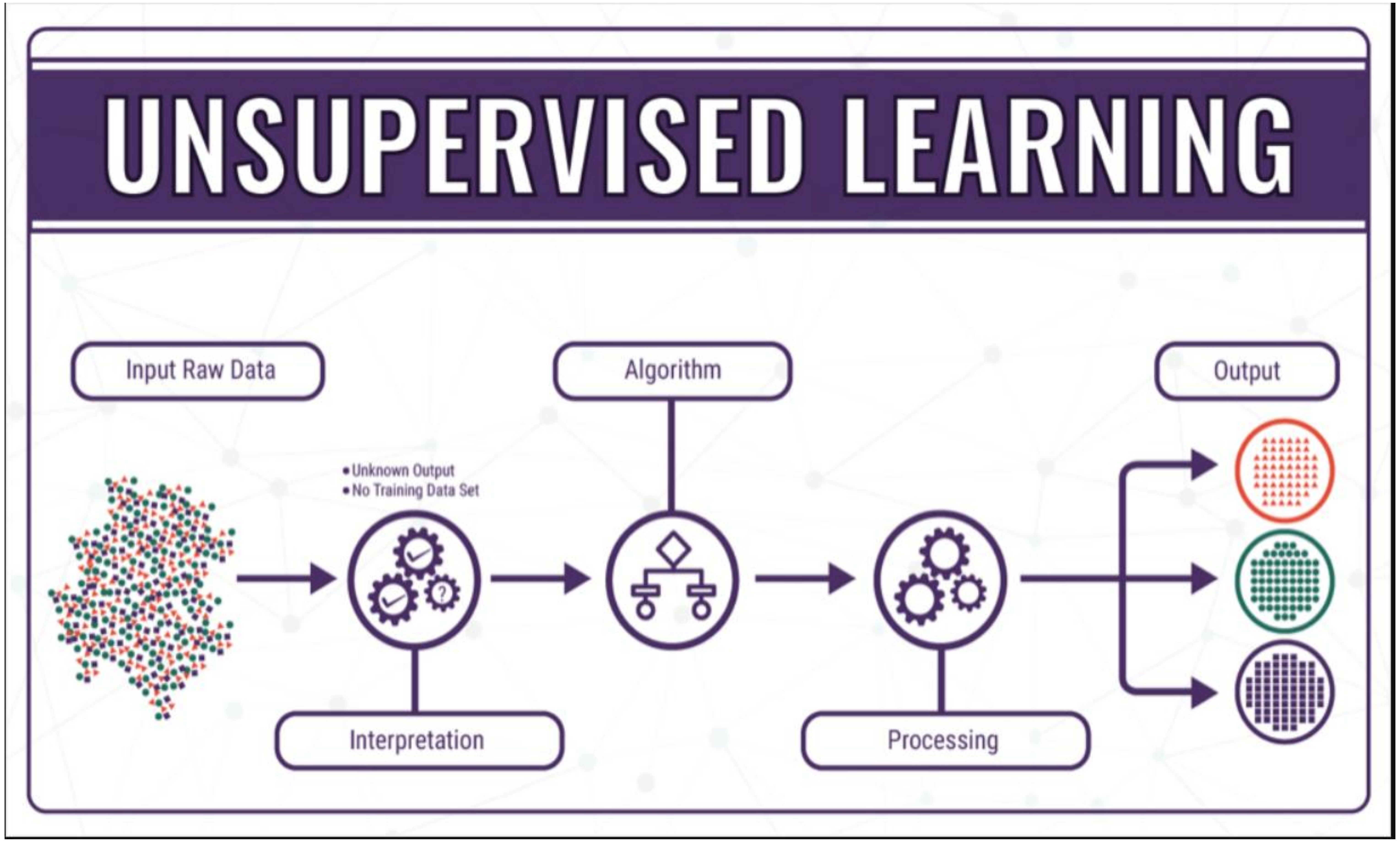
Unsupervised Learning

Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. Its ability to discover similarities and differences in information make it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation, and image recognition.



Algorithm Example : K means, One Class SVM, Isolation forest,etc

Unsupervised Learning



General fields:

- ❖E commerce
- ❖Anomaly Detection
- ❖You Tube
- ❖Media services

Evaluation

The Battle is not over yet!!

A good Data Scientist should be able to communicate his findings with the business team such that it easily goes into execution phase

Any Queries ?

Thank you

Sahaana Venkat
Mail ID : sagu1995@gmail.com
LinkedIn ID : https://www.linkedin.com/in/sahaana-venkat-a31997127/