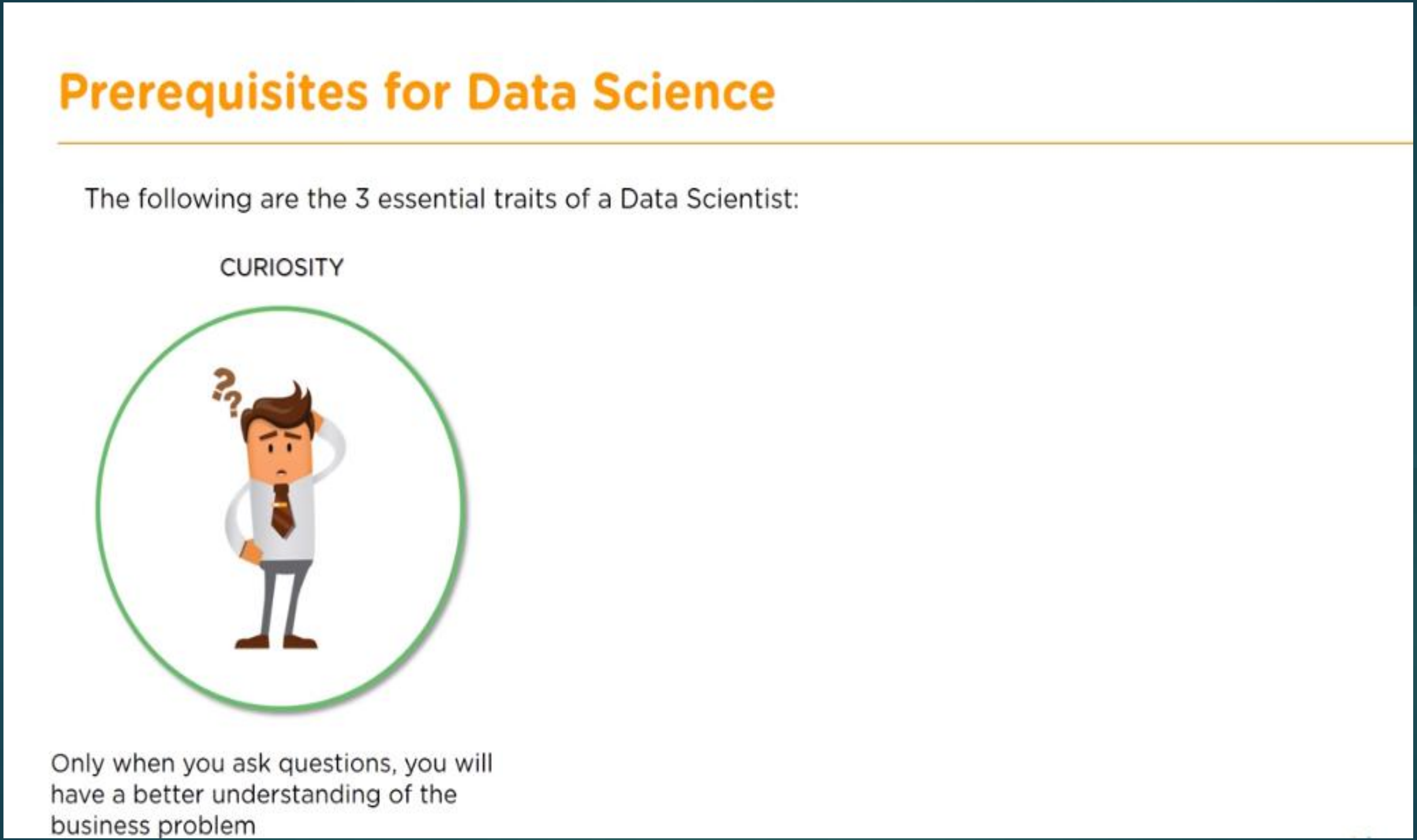
DATA SCIENCE -2

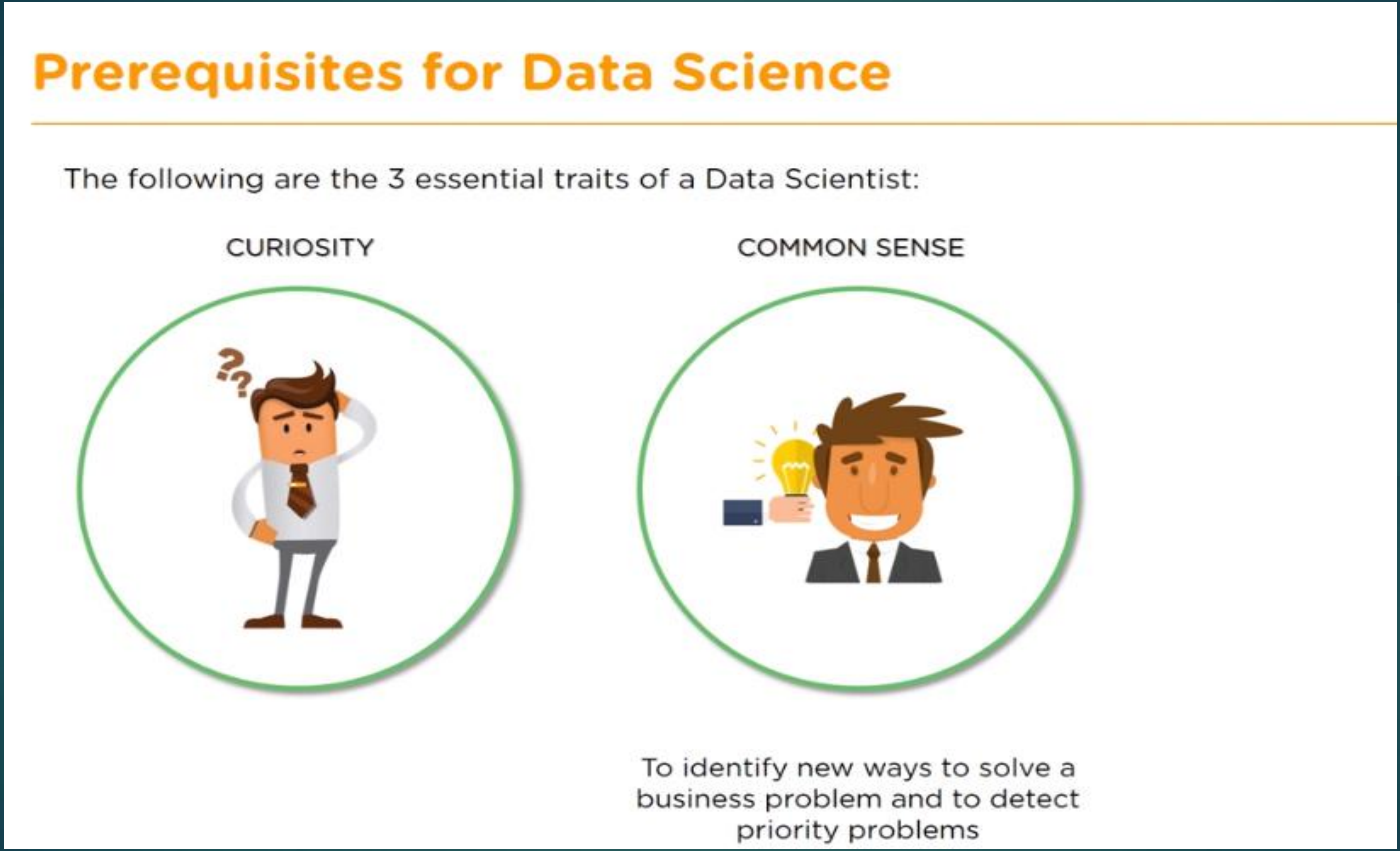


Prerequisites for Data Science

The following are the 3 essential traits of a Data Scientist:

CURIOSITY

Only when you ask questions, you will have a better understanding of the business problem



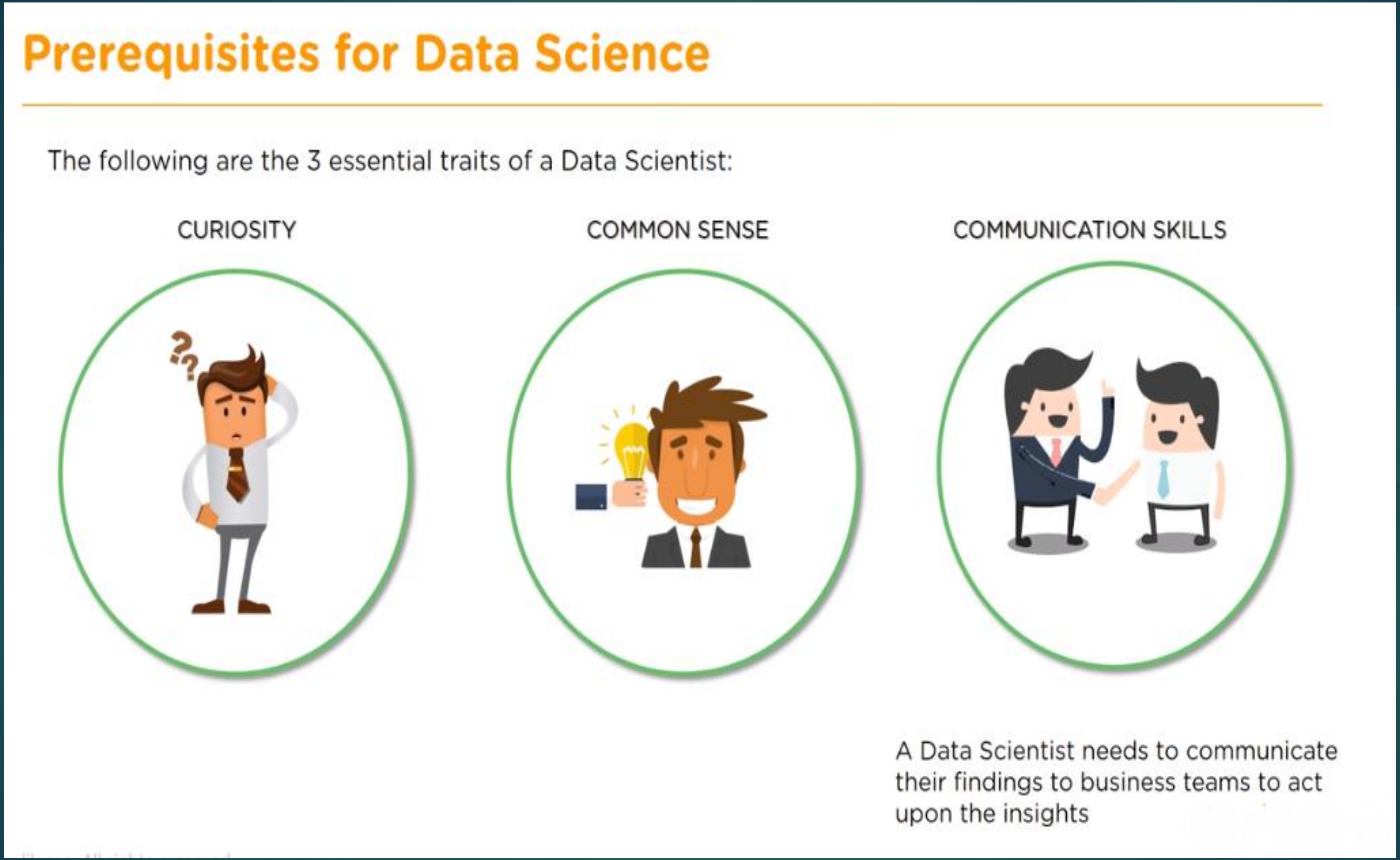
Prerequisites for Data Science

The following are the 3 essential traits of a Data Scientist:

CURIOSITY

COMMON SENSE

To identify new ways to solve a business problem and to detect priority problems



Prerequisites for Data Science

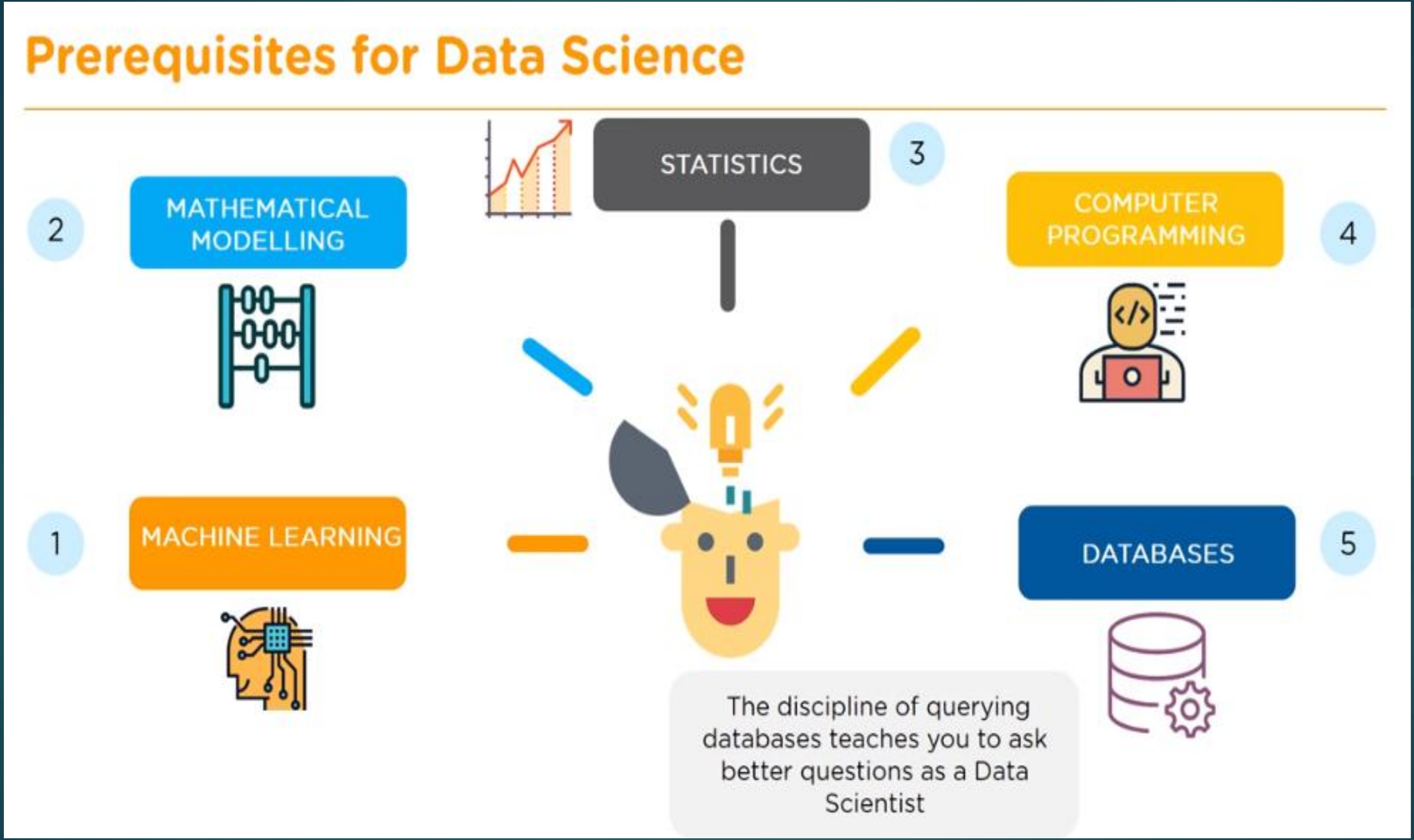
The following are the 3 essential traits of a Data Scientist:

CURIOSITY

COMMON SENSE

COMMUNICATION SKILLS

A Data Scientist needs to communicate their findings to business teams to act upon the insights



Prerequisites for Data Science

1 MACHINE LEARNING

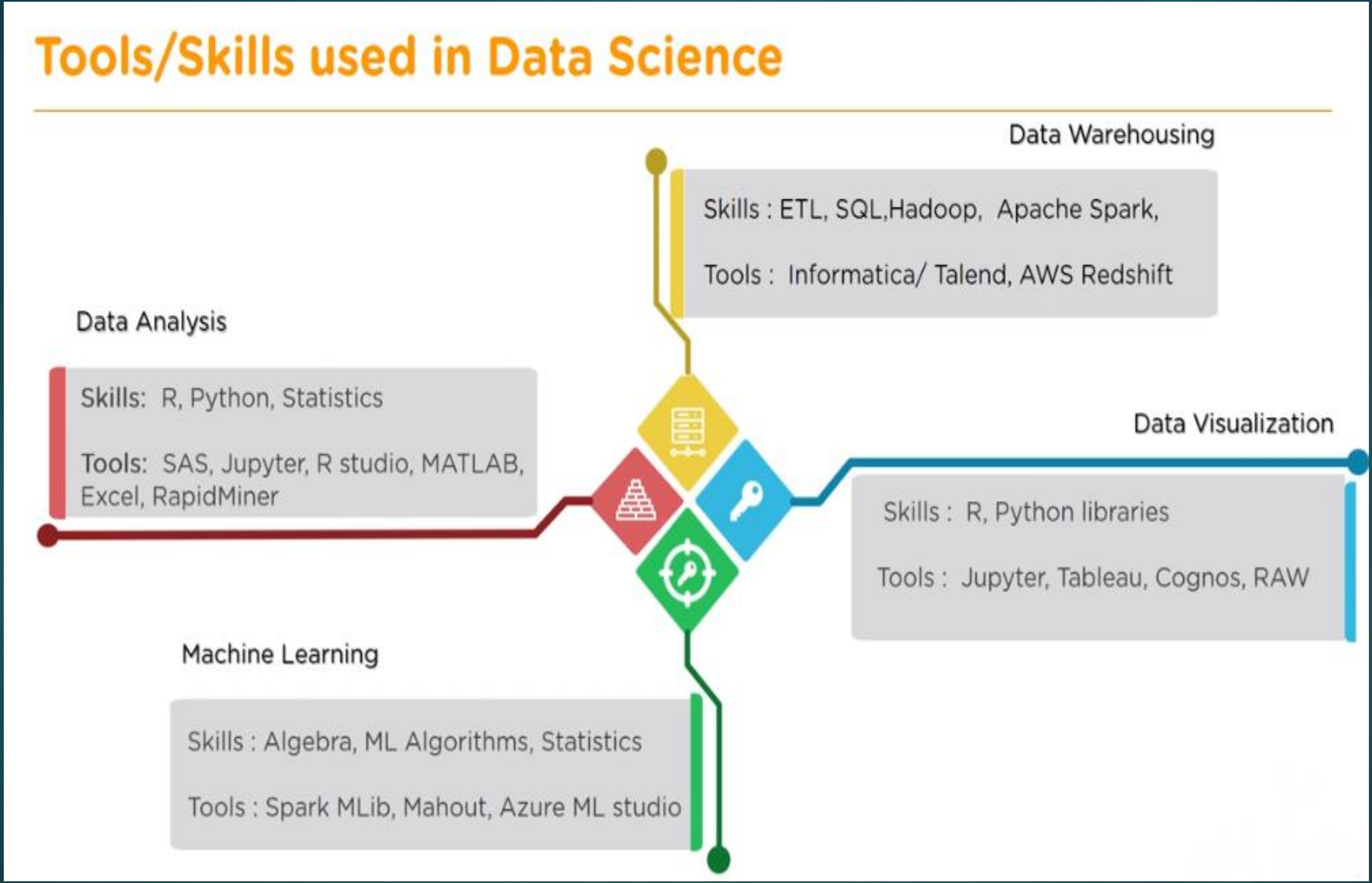
2 MATHEMATICAL MODELLING

3 STATISTICS

4 COMPUTER PROGRAMMING

5 DATABASES

The discipline of querying databases teaches you to ask better questions as a Data Scientist



Tools/Skills used in Data Science

Data Warehousing

Skills : ETL, SQL,Hadoop, Apache Spark,

Tools : Informatica/ Talend, AWS Redshift

Data Analysis

Skills: R, Python, Statistics

Tools: SAS, Jupyter, R studio, MATLAB, Excel, RapidMiner

Data Visualization

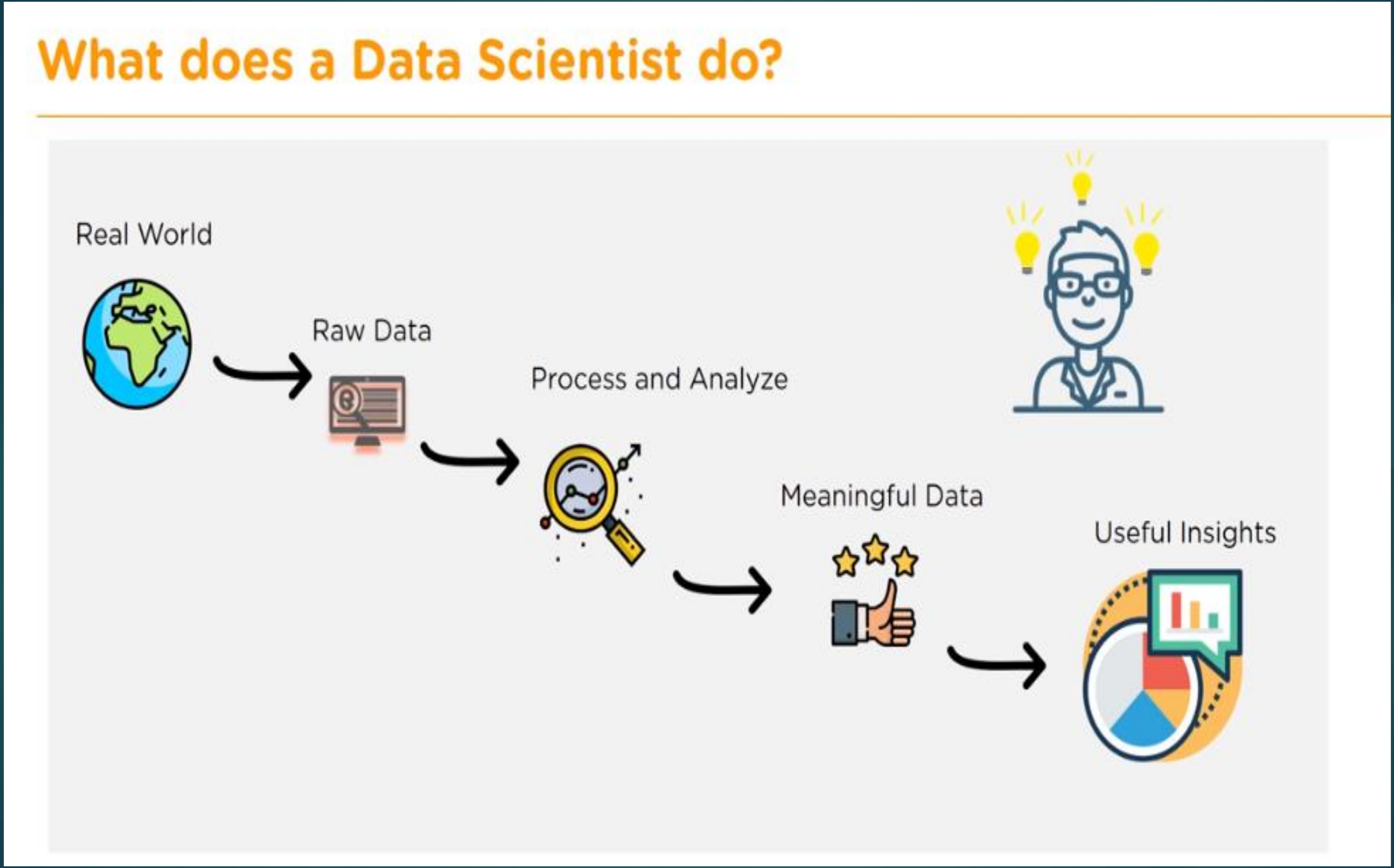
Skills : R, Python libraries

Tools : Jupyter, Tableau, Cognos, RAW

Machine Learning

Skills : Algebra, ML Algorithms, Statistics

Tools : Spark MLib, Mahout, Azure ML studio



What does a Data Scientist do?

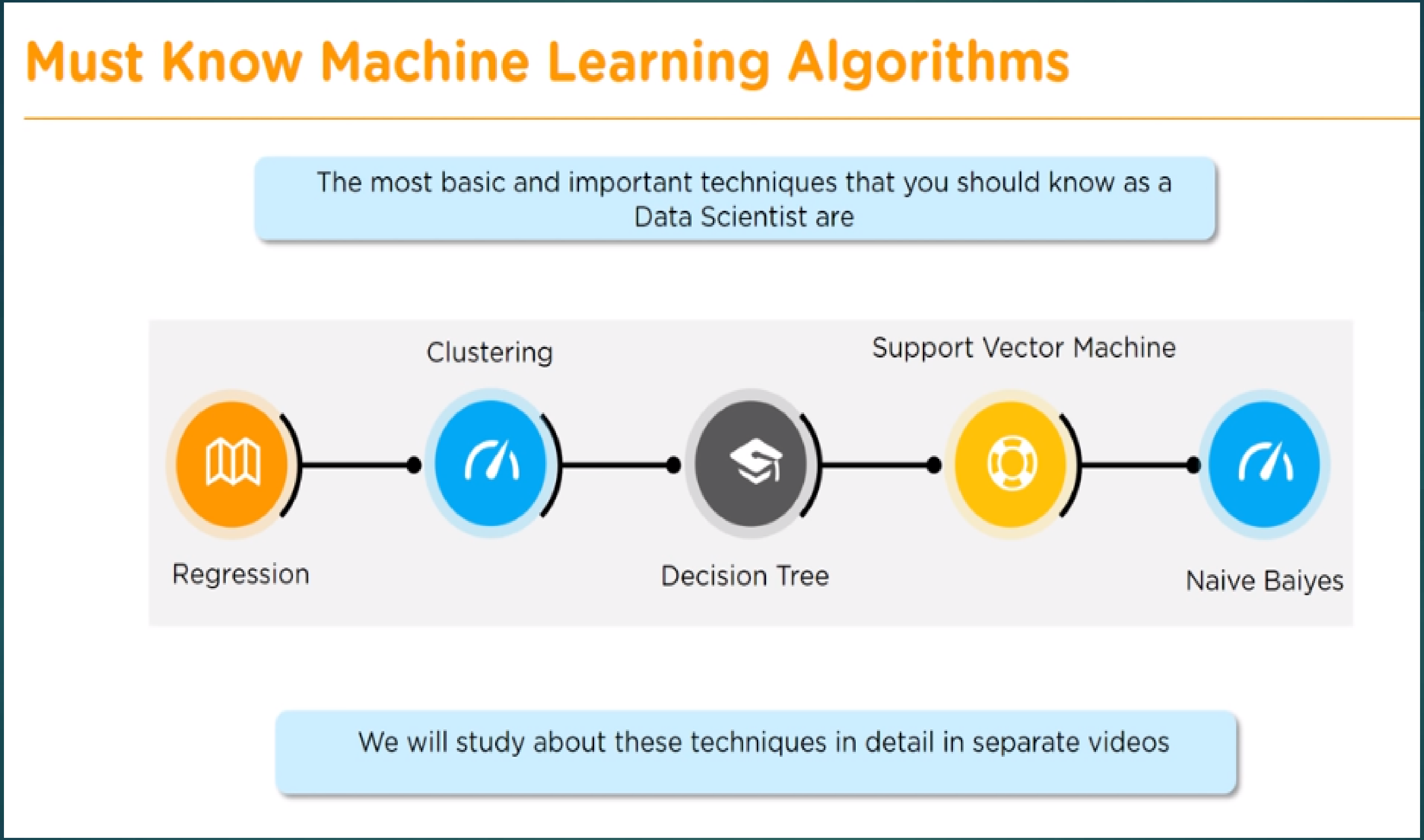
Real World

Raw Data

Process and Analyze

Meaningful Data

Useful Insights



Must Know Machine Learning Algorithms

The most basic and important techniques that you should know as a Data Scientist are

Regression

Clustering

Decision Tree

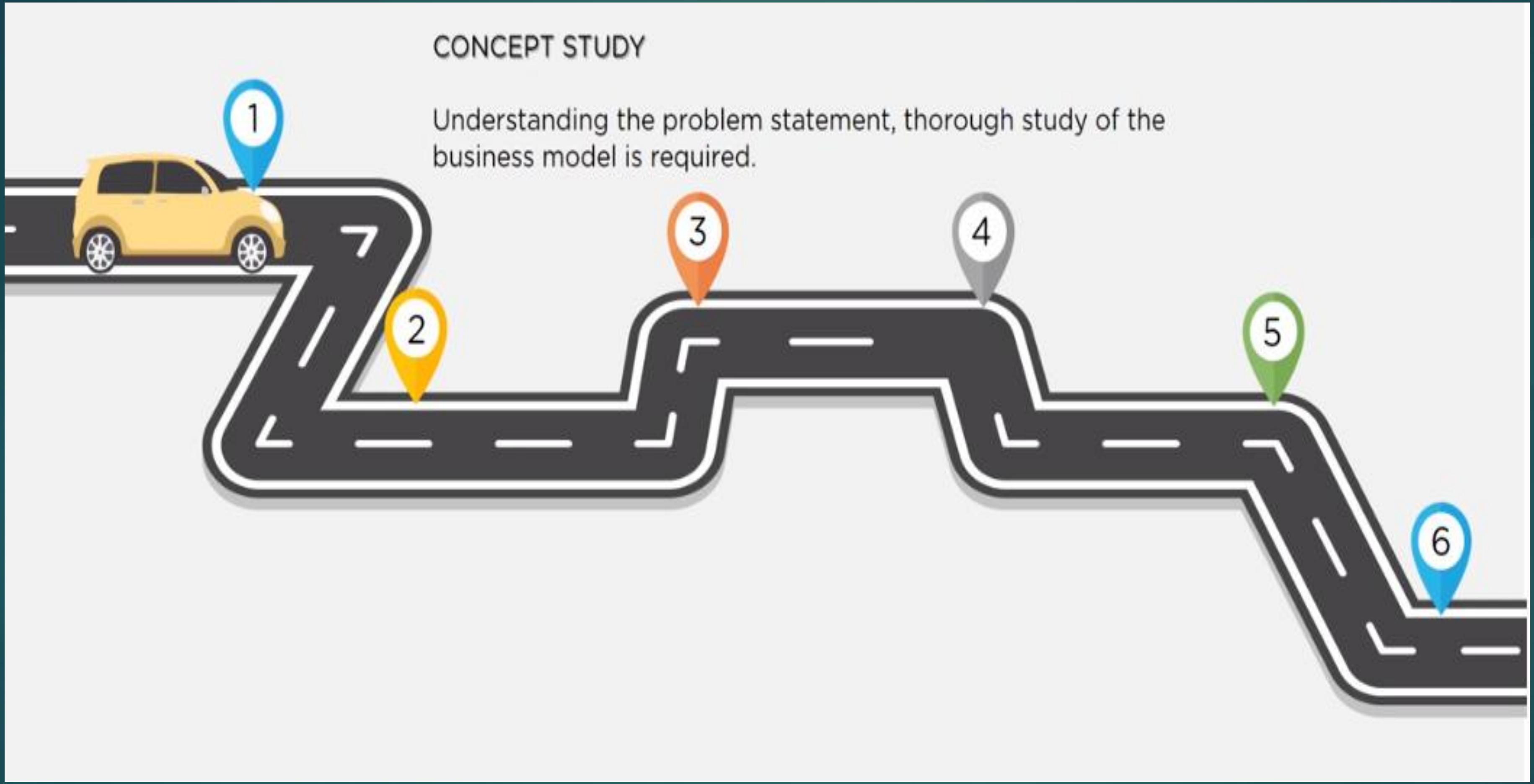
Support Vector Machine

Naive Baiyes

We will study about these techniques in detail in separate videos

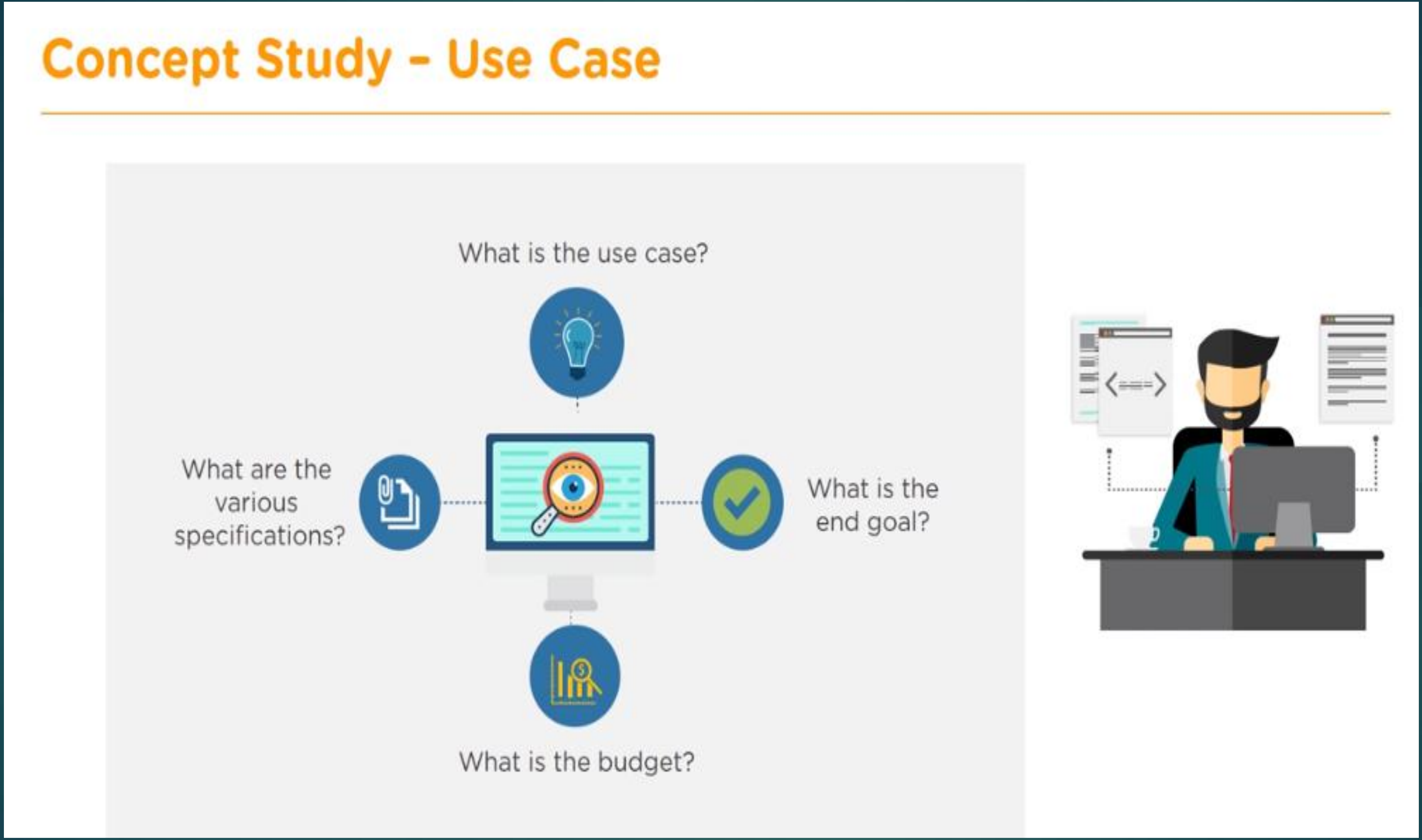
DATA SCIENCE LIFECYCLE WITH EXAMPLE

1) Concept Study



CONCEPT STUDY

Understanding the problem statement, thorough study of the business model is required.



Concept Study - Use Case

What is the use case?

What are the various specifications?

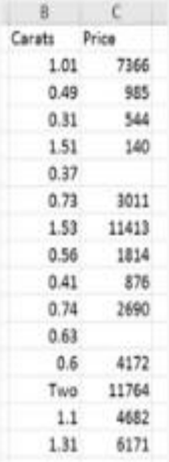
What is the end goal?

What is the budget?

Concept Study - Use Case

Concept of the task : Predict the price of 1.35 carat diamond

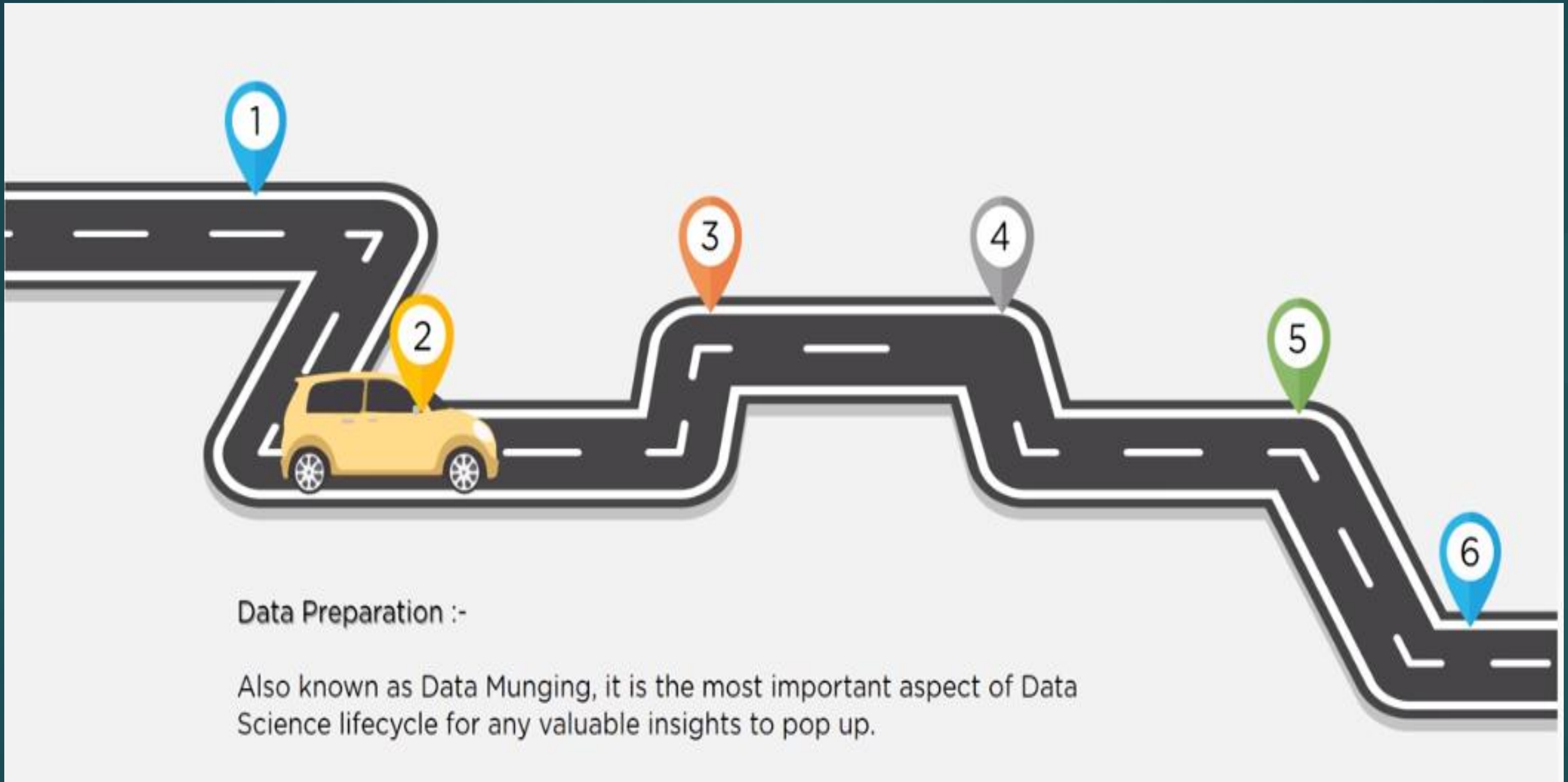
Get to know about the diamond industry, various terminologies used. Understand the business problem and collect RELEVANT and enough data



B	C
Carats	Price
1.01	7366
0.49	985
0.31	544
1.51	140
0.37	
0.73	3011
1.53	11413
0.56	1814
0.41	876
0.74	2690
0.63	
0.6	4172
Two	11764
1.1	4682
1.31	6171

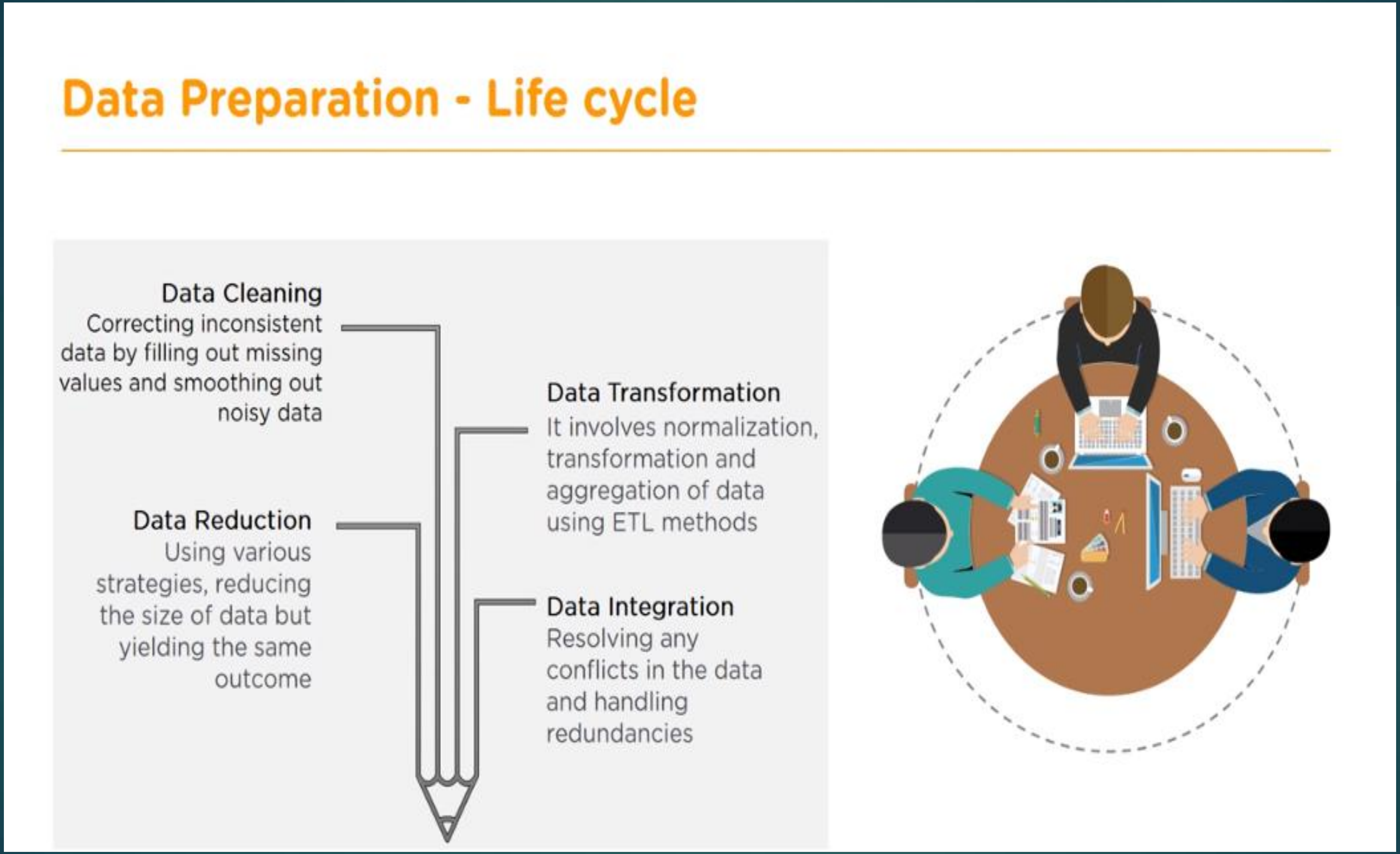
Suppose, we get the price of diamonds from different diamond retailers. But we want to find out the price of 1.35 carat diamond.

2)Data preparation:



Data Preparation :-

Also known as Data Munging, it is the most important aspect of Data Science lifecycle for any valuable insights to pop up.



Data Preparation - Life cycle

Data Cleaning
Correcting inconsistent data by filling out missing values and smoothing out noisy data

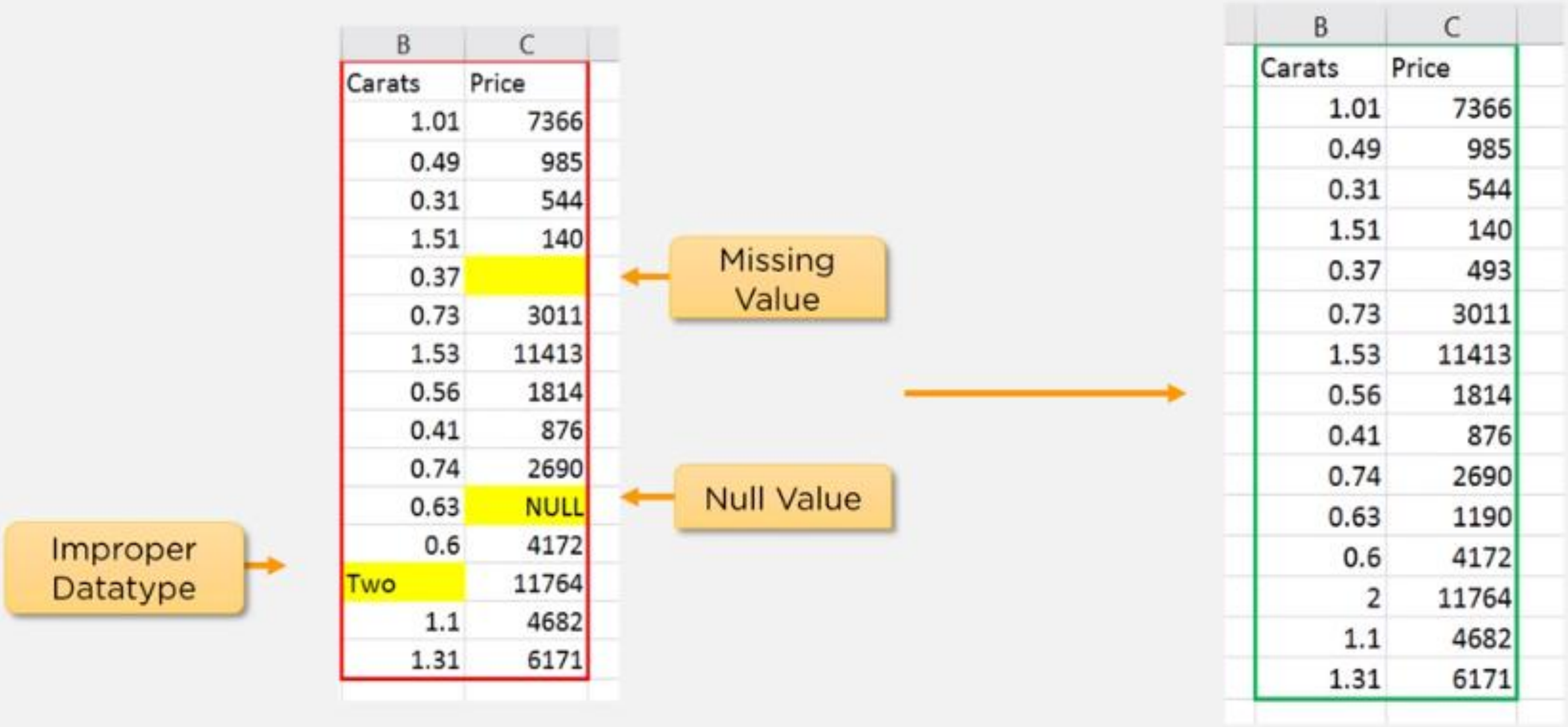
Data Reduction
Using various strategies, reducing the size of data but yielding the same outcome

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 Preparation - Use Case

Data preparation : Make the data clean and valuable.



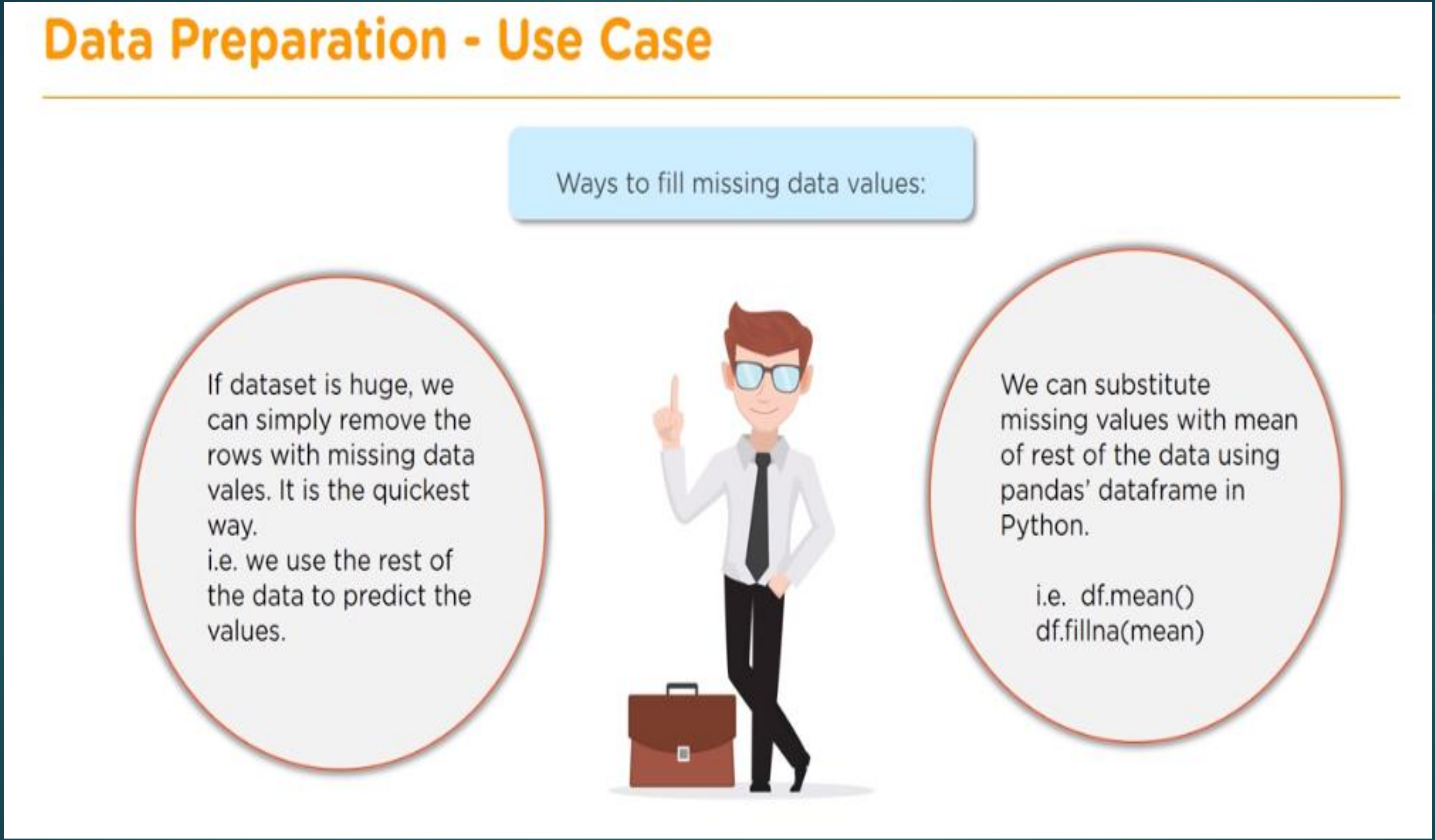
B	C
Carats	Price
1.01	7366
0.49	985
0.31	544
1.51	140
0.37	
0.73	3011
1.53	11413
0.56	1814
0.41	876
0.74	2690
0.63	NULL
0.6	4172
Two	11764
1.1	4682
1.31	6171

Missing Value

Null Value

Improper Datatype

B	C
Carats	Price
1.01	7366
0.49	985
0.31	544
1.51	140
0.37	493
0.73	3011
1.53	11413
0.56	1814
0.41	876
0.74	2690
0.63	1190
0.6	4172
2	11764
1.1	4682
1.31	6171



Data Preparation - Use Case

Ways to fill missing data values:

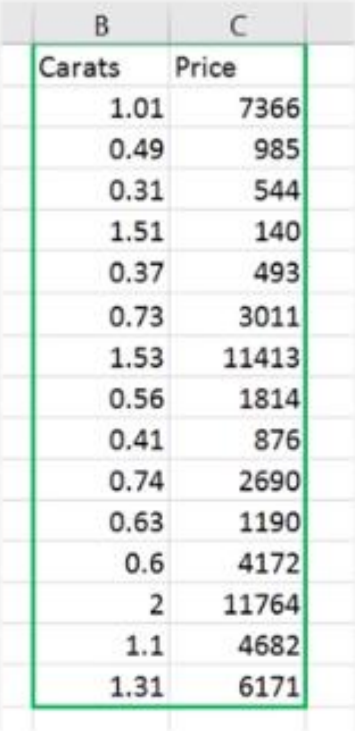
If dataset is huge, we can simply remove the rows with missing data vales. It is the quickest way.
i.e. we use the rest of the data to predict the values.

We can substitute missing values with mean of rest of the data using pandas' dataframe in Python.

i.e. df.mean()
df.fillna(mean)

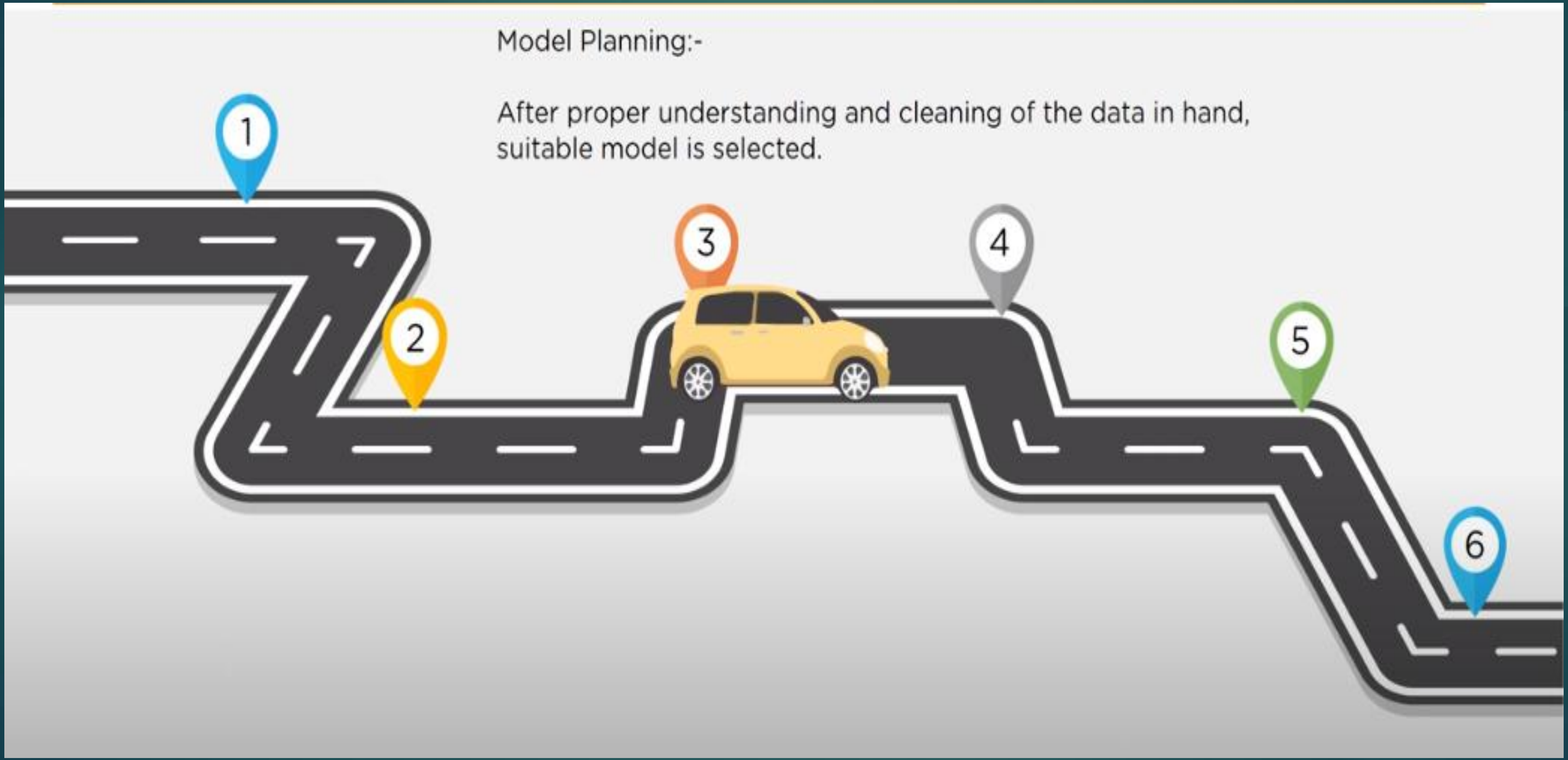
Data Preparation - Example

- Split the data into train data and test data in the ratio of 80:20
- It is generally advised to divide the dataset into two random partition



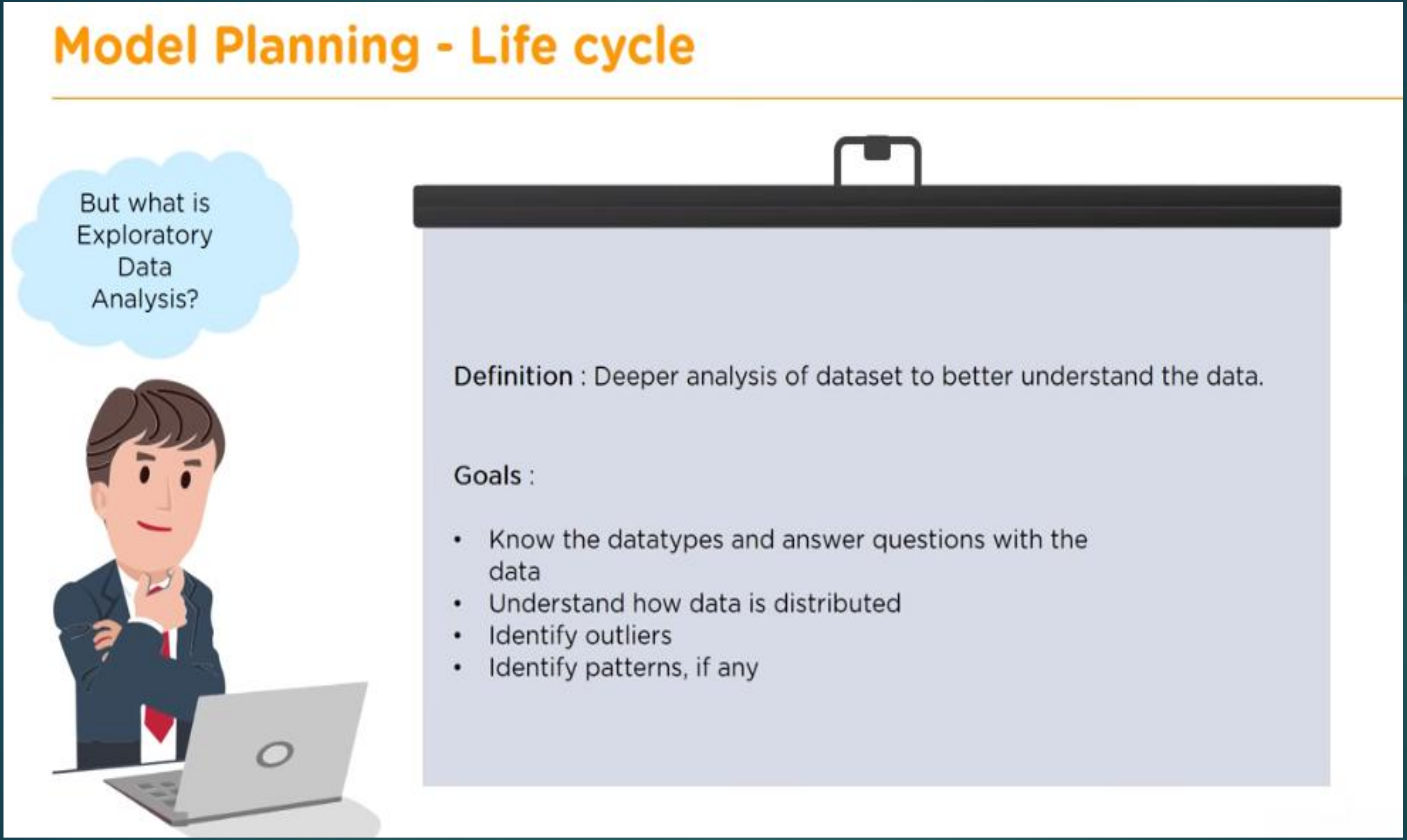
Carats	Price
1.01	7366
0.49	985
0.31	544
1.51	140
0.37	493
0.73	3011
1.53	11413
0.56	1814
0.41	876
0.74	2690
0.63	1190
0.6	4172
2	11764
1.1	4682
1.31	6171

3) Model planning



Model Planning:-

After proper understanding and cleaning of the data in hand, suitable model is selected.



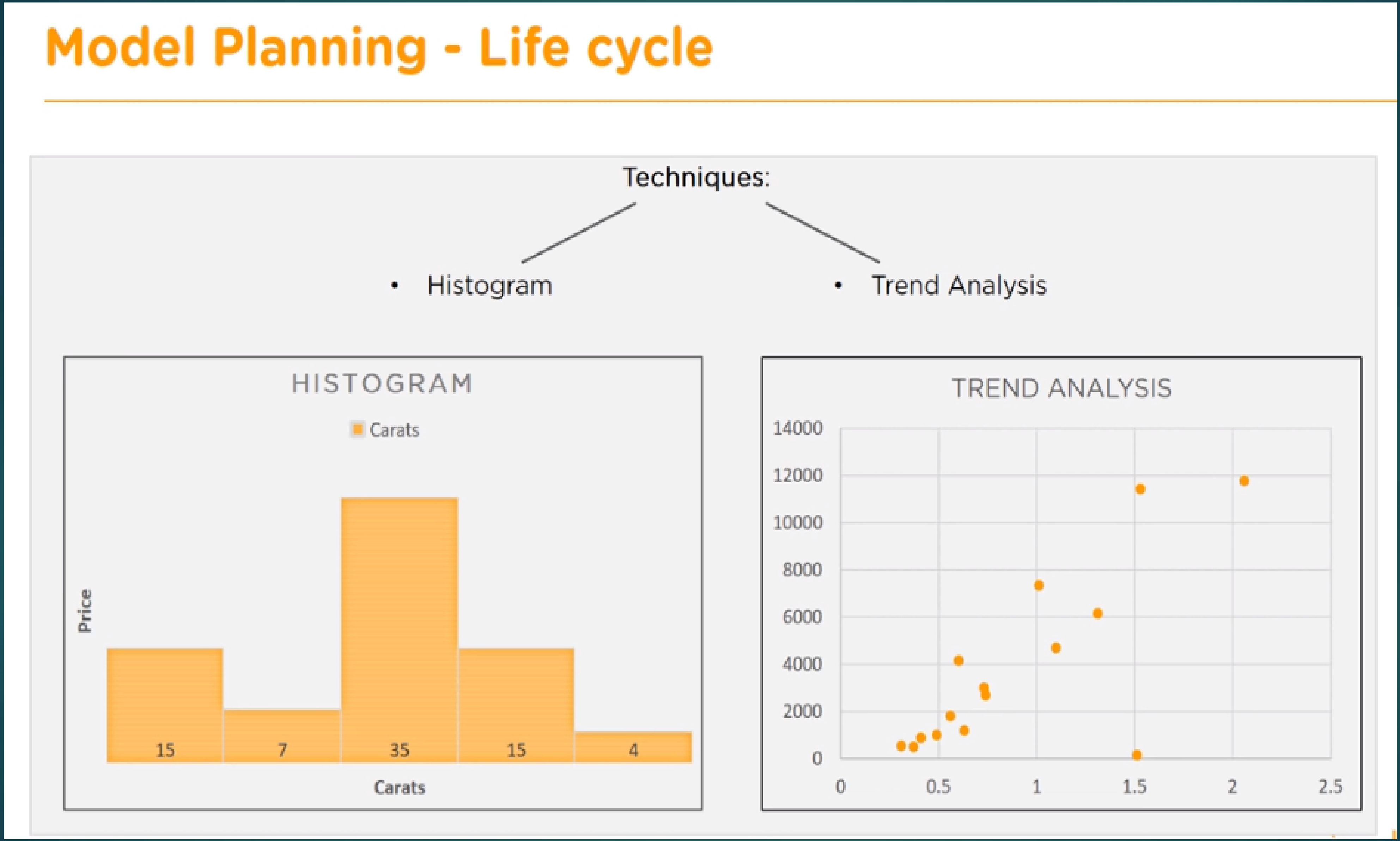
Model Planning - Life cycle

But what is Exploratory Data Analysis?

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

Goals :

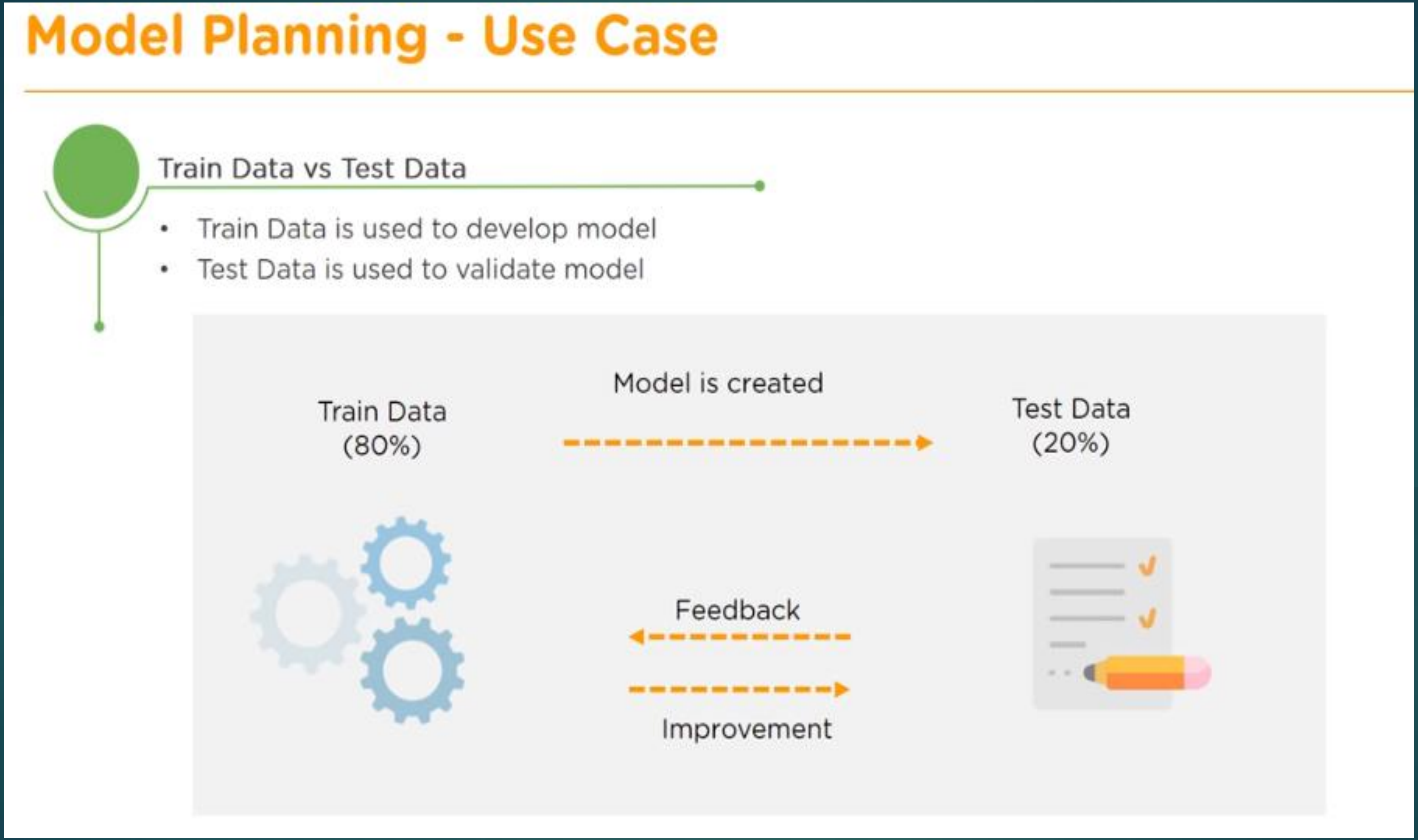
- Know the datatypes and answer questions with the data
- Understand how data is distributed
- Identify outliers
- Identify patterns, if any



Model Planning - Life cycle

Techniques:

- Histogram
- Trend Analysis



Model Planning - Use Case

Train Data vs Test Data

- Train Data is used to develop model
- Test Data is used to validate model

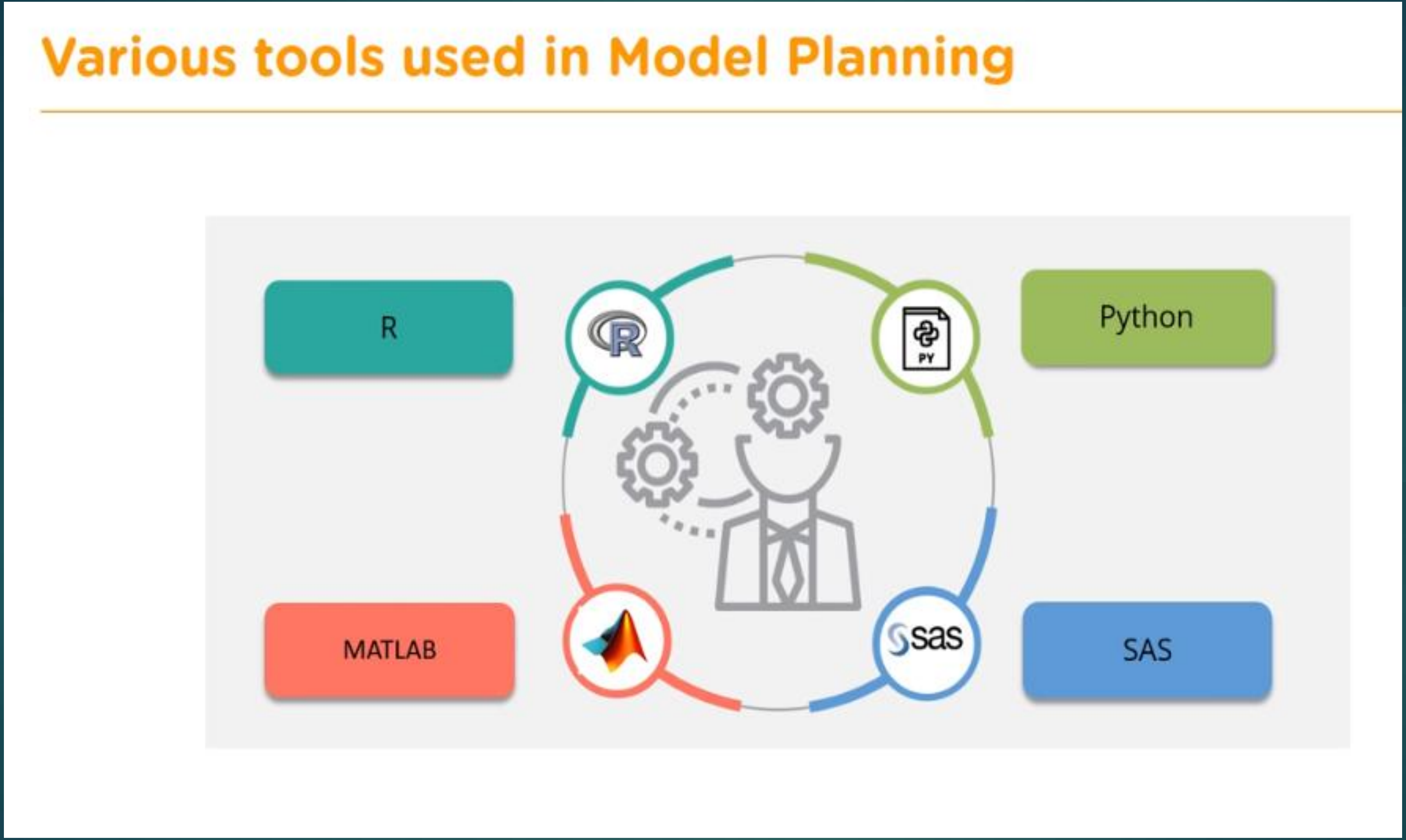
Train Data
(80%)

Model is created

Test Data
(20%)

Feedback

Improvement



Various tools used in Model Planning

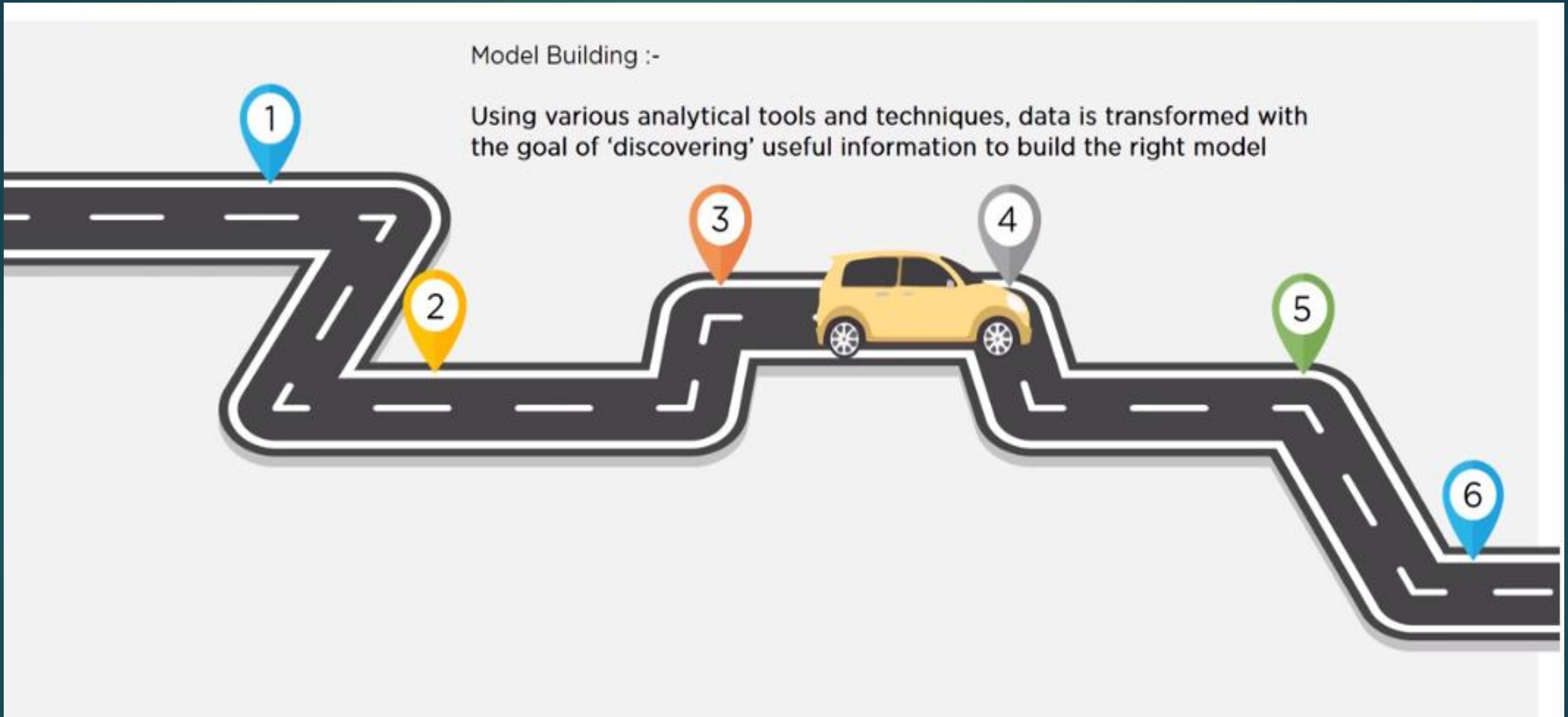
R

Python

MATLAB

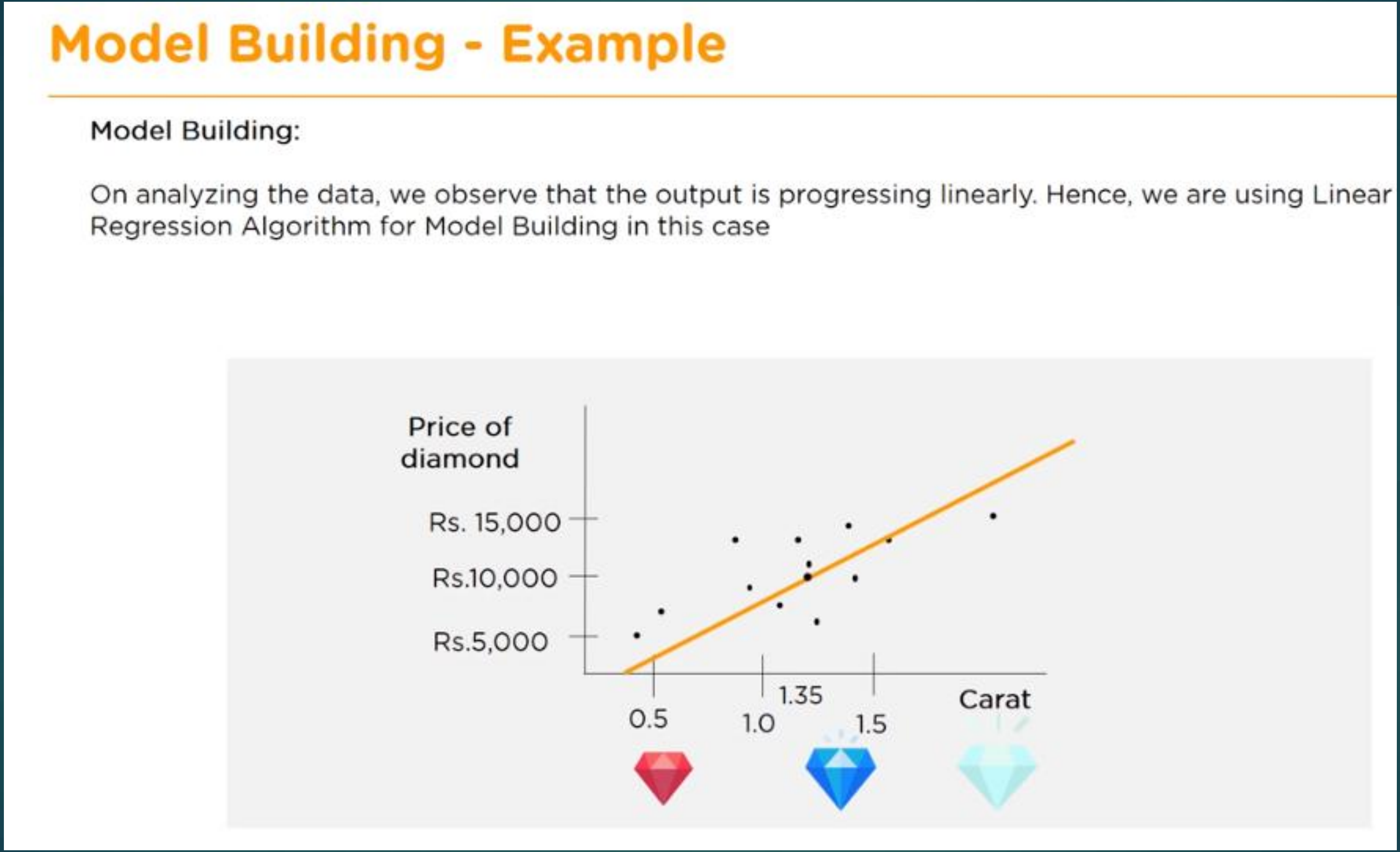
SAS

4) MODEL BUILDING



Model Building :-

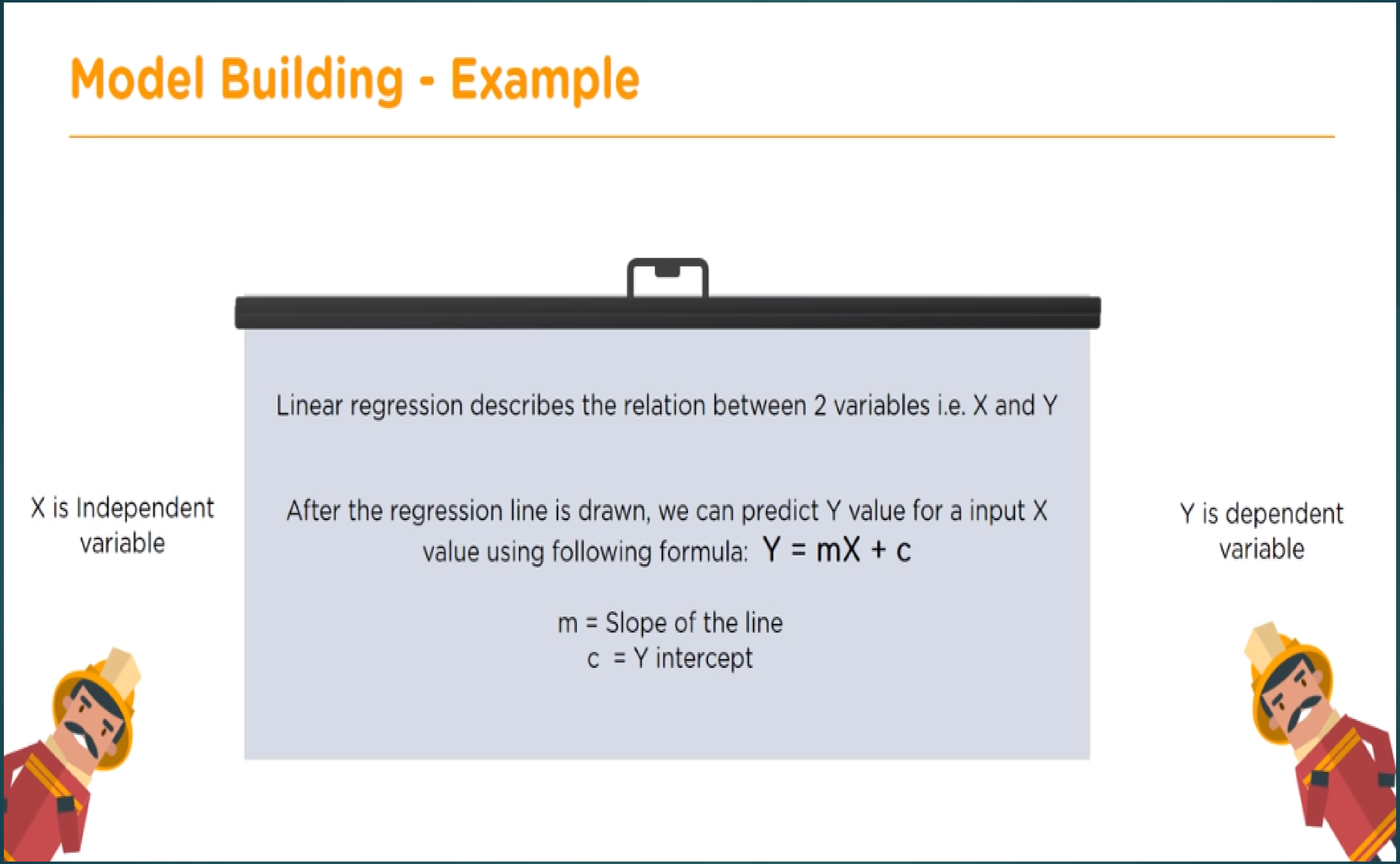
Using various analytical tools and techniques, data is transformed with the goal of ‘discovering’ useful information to build the right model



Model Building - Example

Model Building:

On analyzing the data, we observe that the output is progressing linearly. Hence, we are using Linear Regression Algorithm for Model Building in this case



Model Building - Example

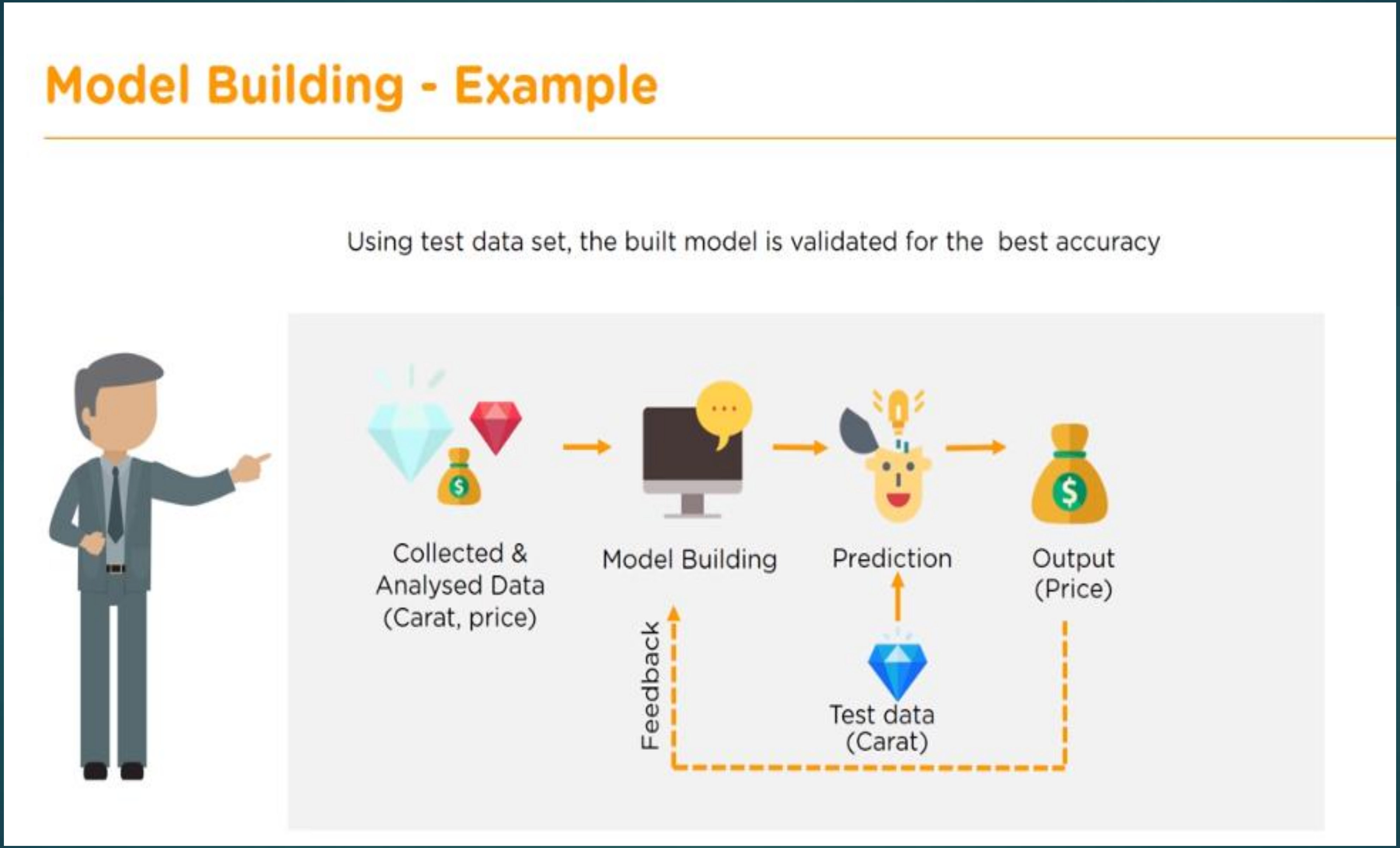
X is Independent variable

Linear regression describes the relation between 2 variables i.e. X and Y

After the regression line is drawn, we can predict Y value for a input X value using following formula: $Y = mX + c$

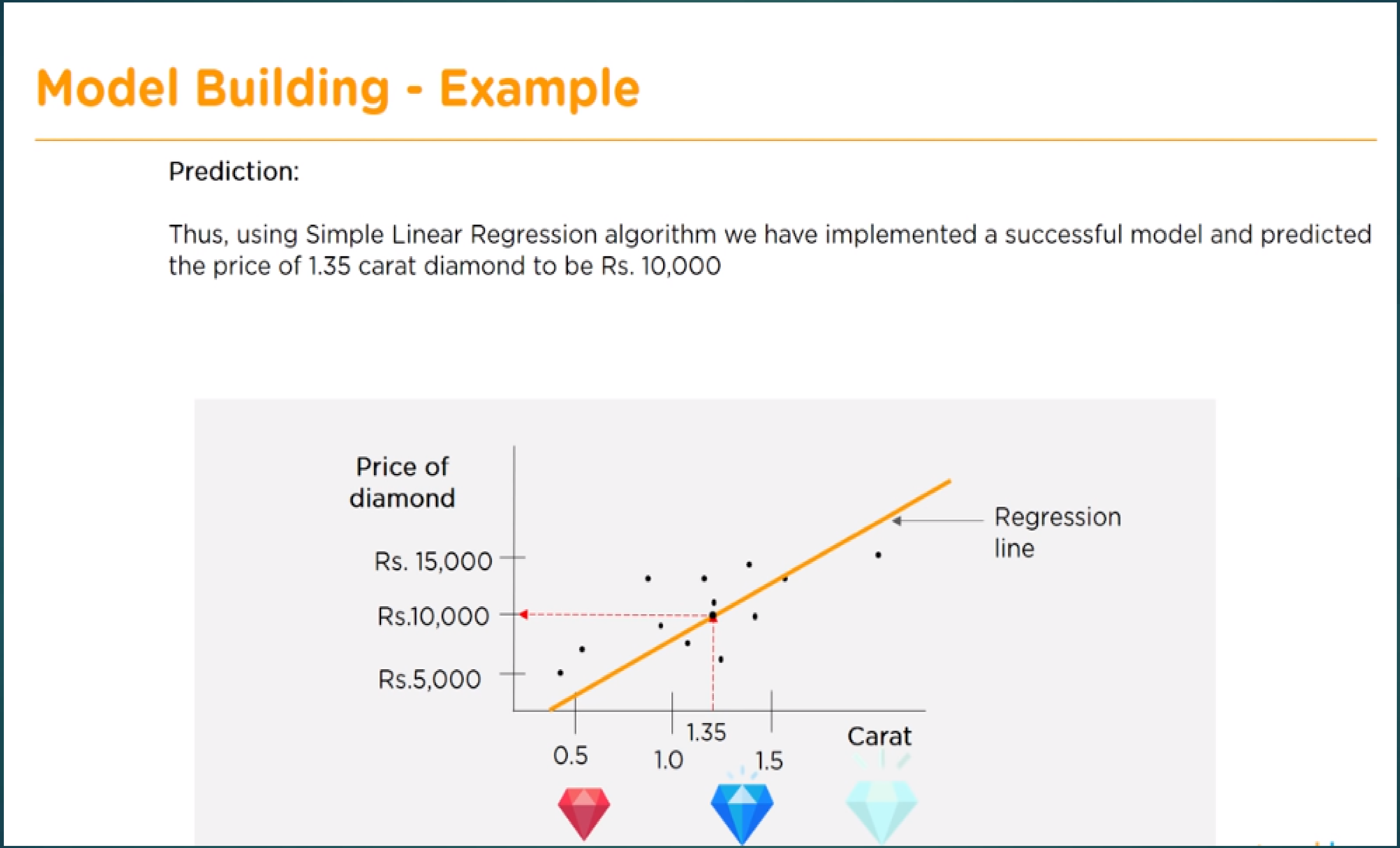
m = Slope of the line
c = Y intercept

Y is dependent variable



Model Building - Example

Using test data set, the built model is validated for the best accuracy



Model Building - Example

Prediction:

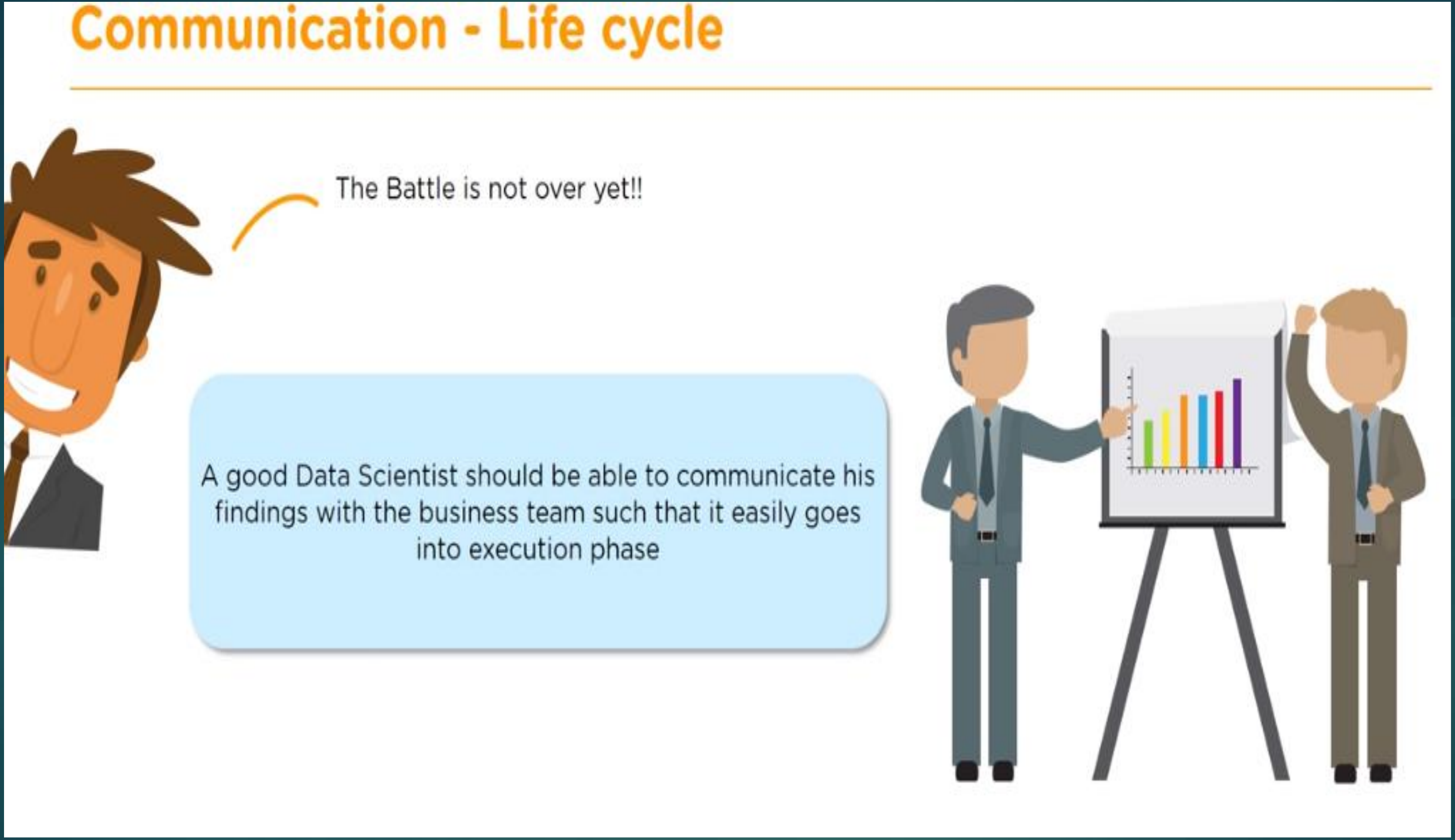
Thus, using Simple Linear Regression algorithm we have implemented a successful model and predicted the price of 1.35 carat diamond to be Rs. 10,000

5) COMMUNICATION

1
2
3
4
5
6
Communicate results

Communicate results:

Keys findings are identified and conveyed to the stakeholders

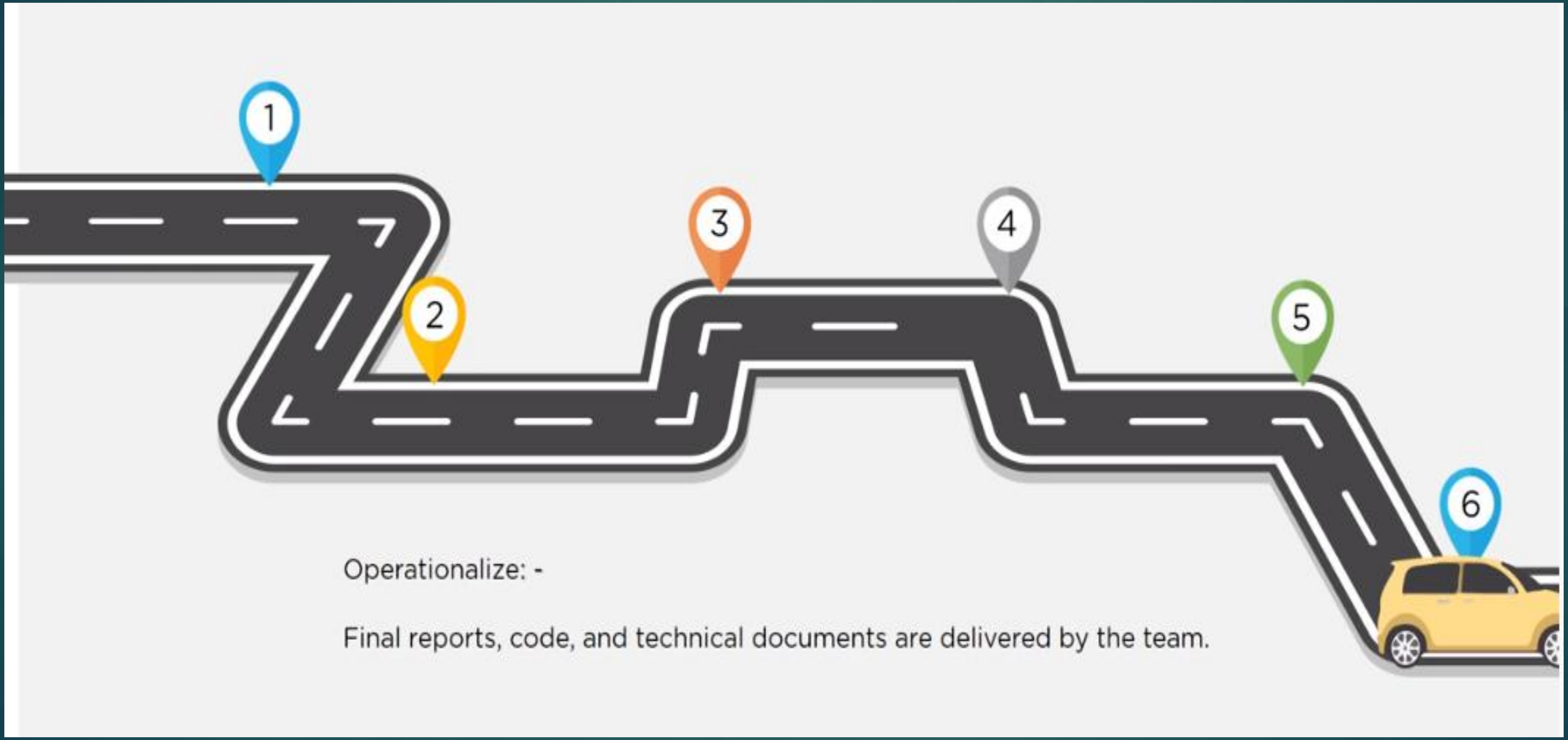


Communication - Life cycle

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

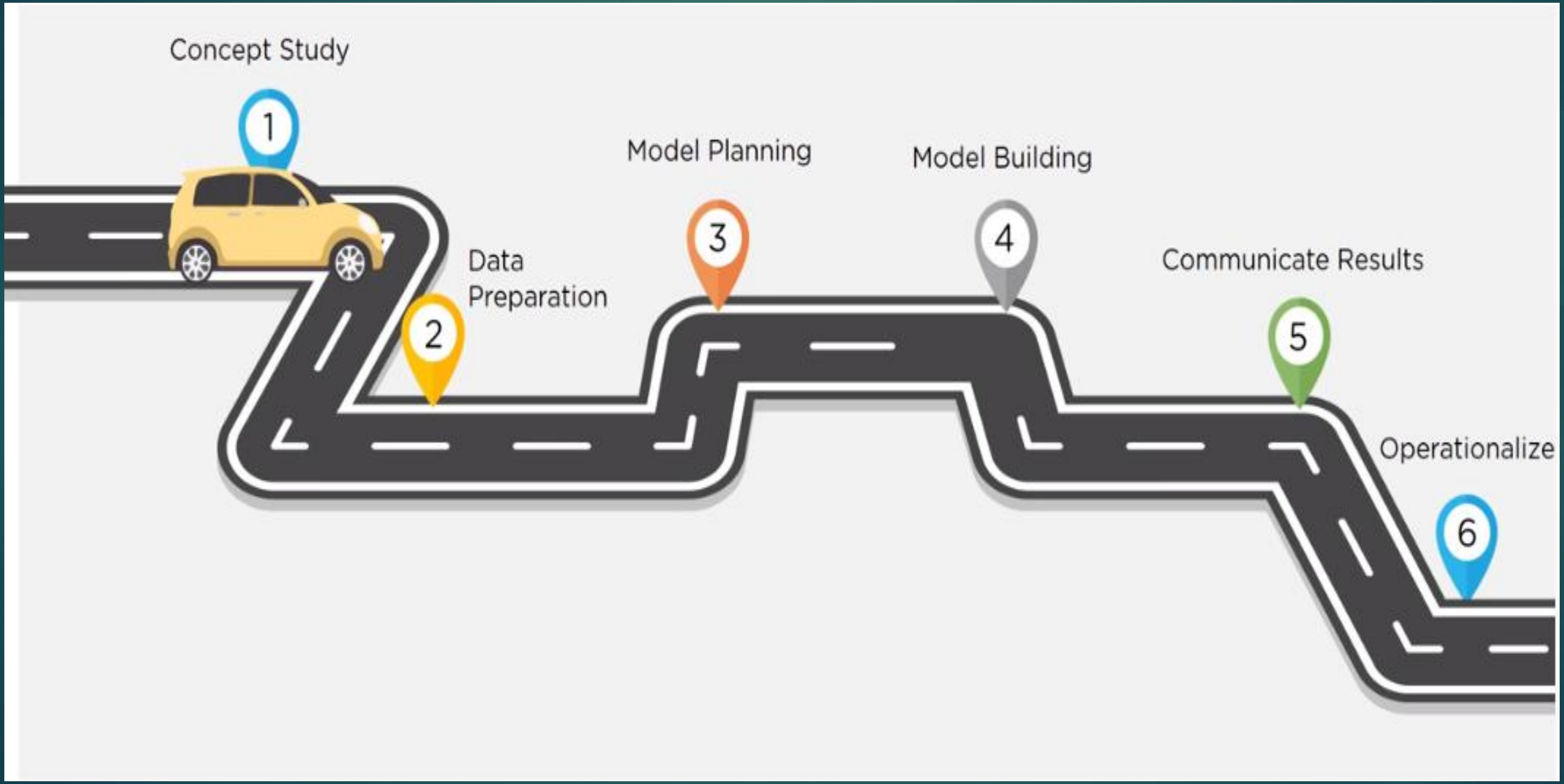
6) OPERATIONALIZE



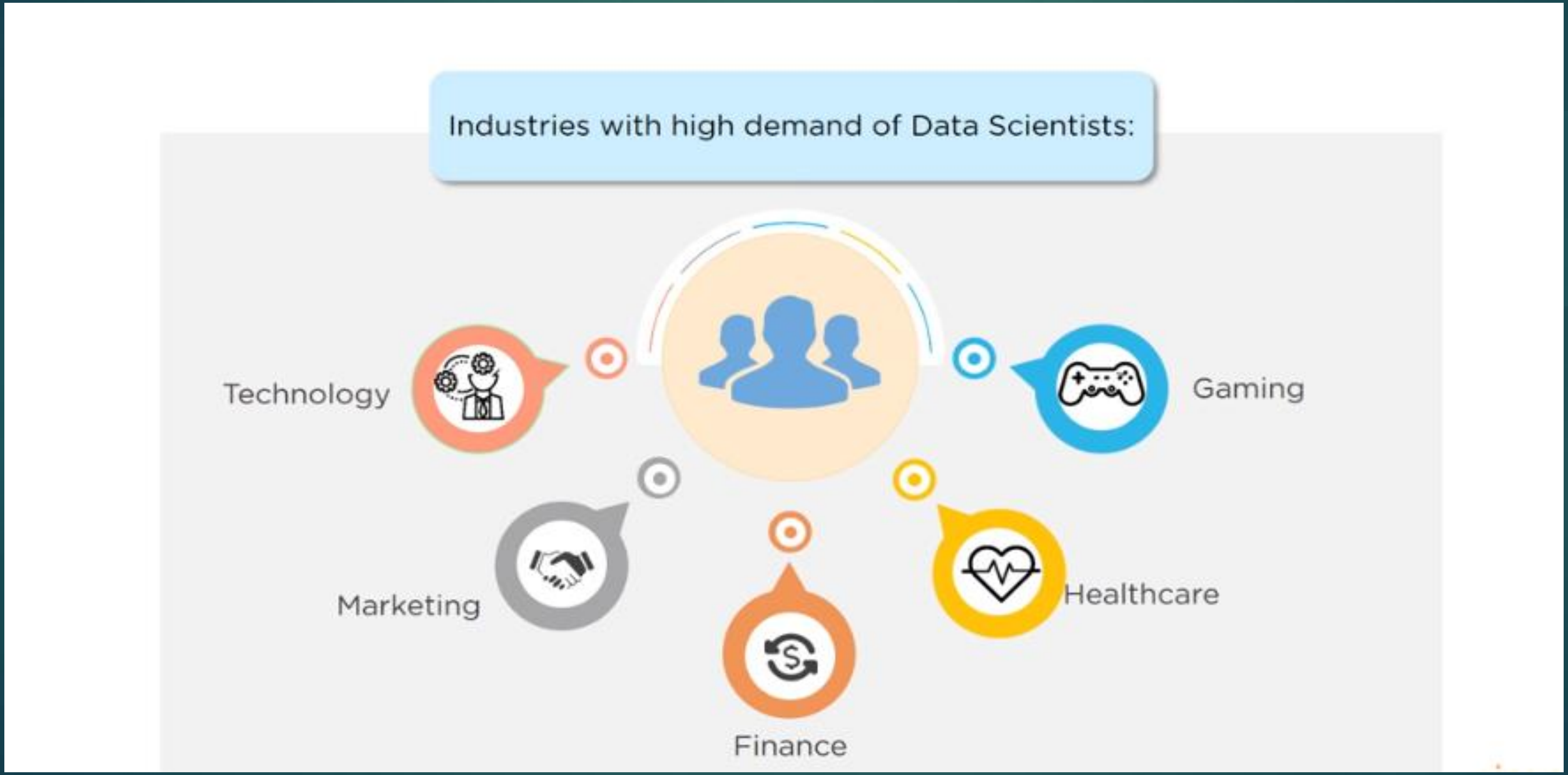
Operationalize: -

Final reports, code, and technical documents are delivered by the team.

SUMMARY - LIFECYCLE



FUTURE OF DATA SCIENTISTS


Industries with high demand of Data Scientists:
Technology
Gaming
Marketing
Healthcare
Finance