Data Science

(Past, Present and Future of Data)

Dr. Shahid Mahmood Awan

Outline

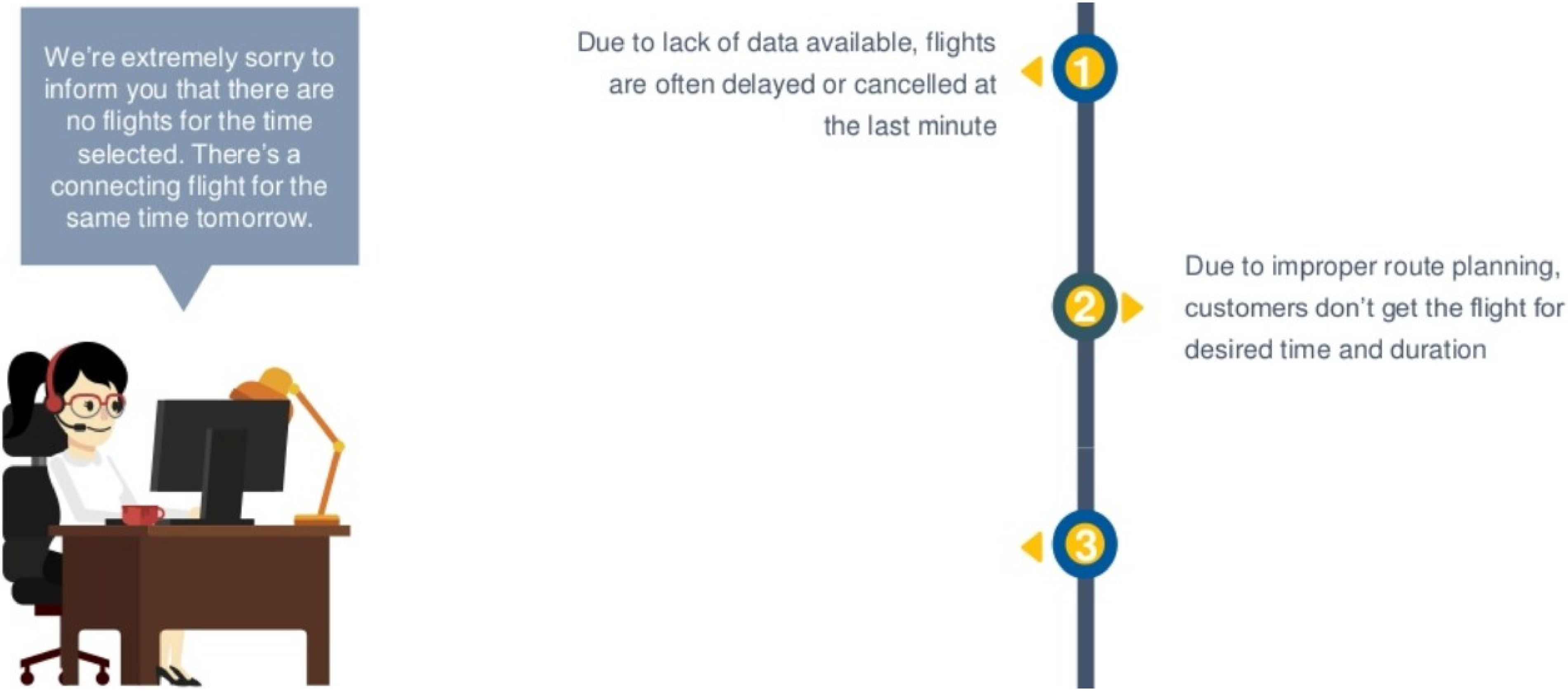
- Why Data Science
- Data Science Process
- Essential Technologies
- Use Cases
- The Way Forward..

Need of Data Science

We're extremely sorry to inform that your flight has been delayed by 4 hours due to bad weather conditions. Regret the inconvenience caused
1
2
3

Due to lack of data available, flights are often delayed or cancelled at the last minute

Need of Data Science

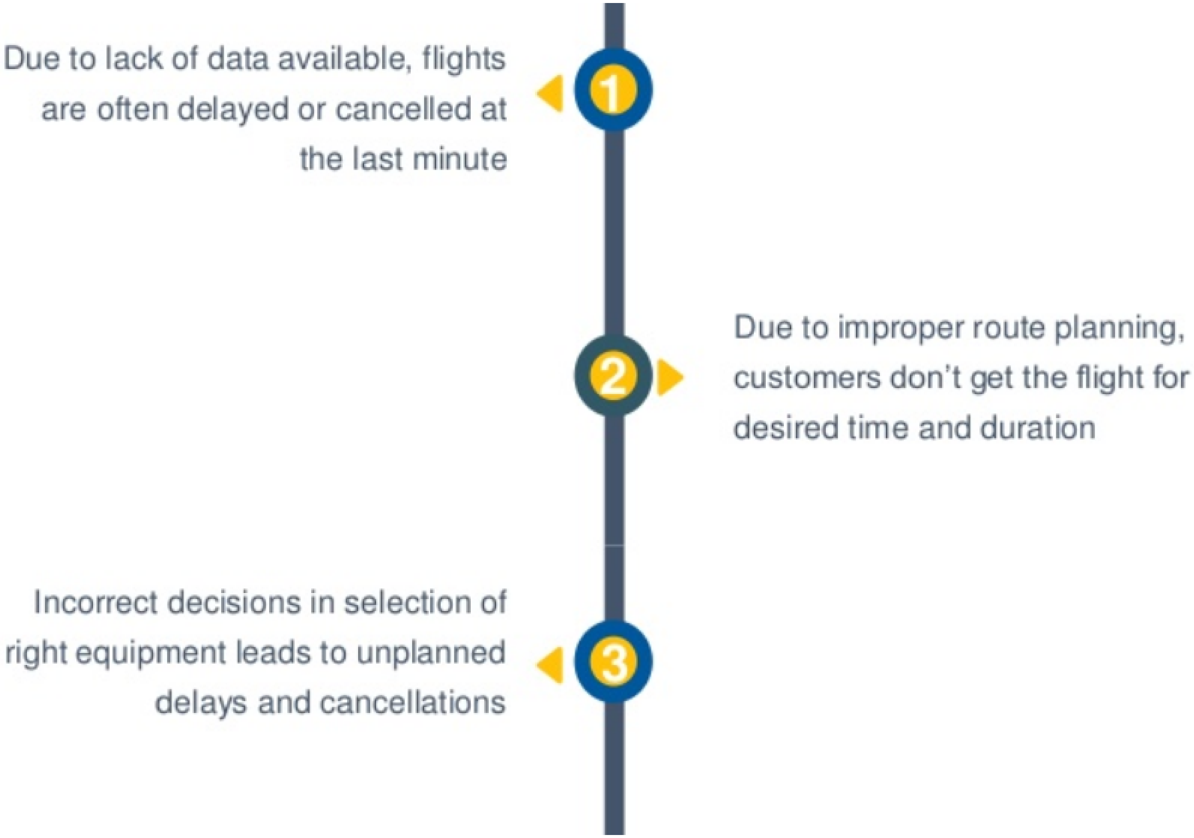

We're extremely sorry to inform you that there are no flights for the time selected. There's a connecting flight for the same time tomorrow.
1
2
3

Due to lack of data available, flights are often delayed or cancelled at the last minute

Due to improper route planning, customers don't get the flight for desired time and duration

Need of Data Science

Dear Flyer, We regret to inform you that your flight has been cancelled due to delay from Airbus on account of engine delivery

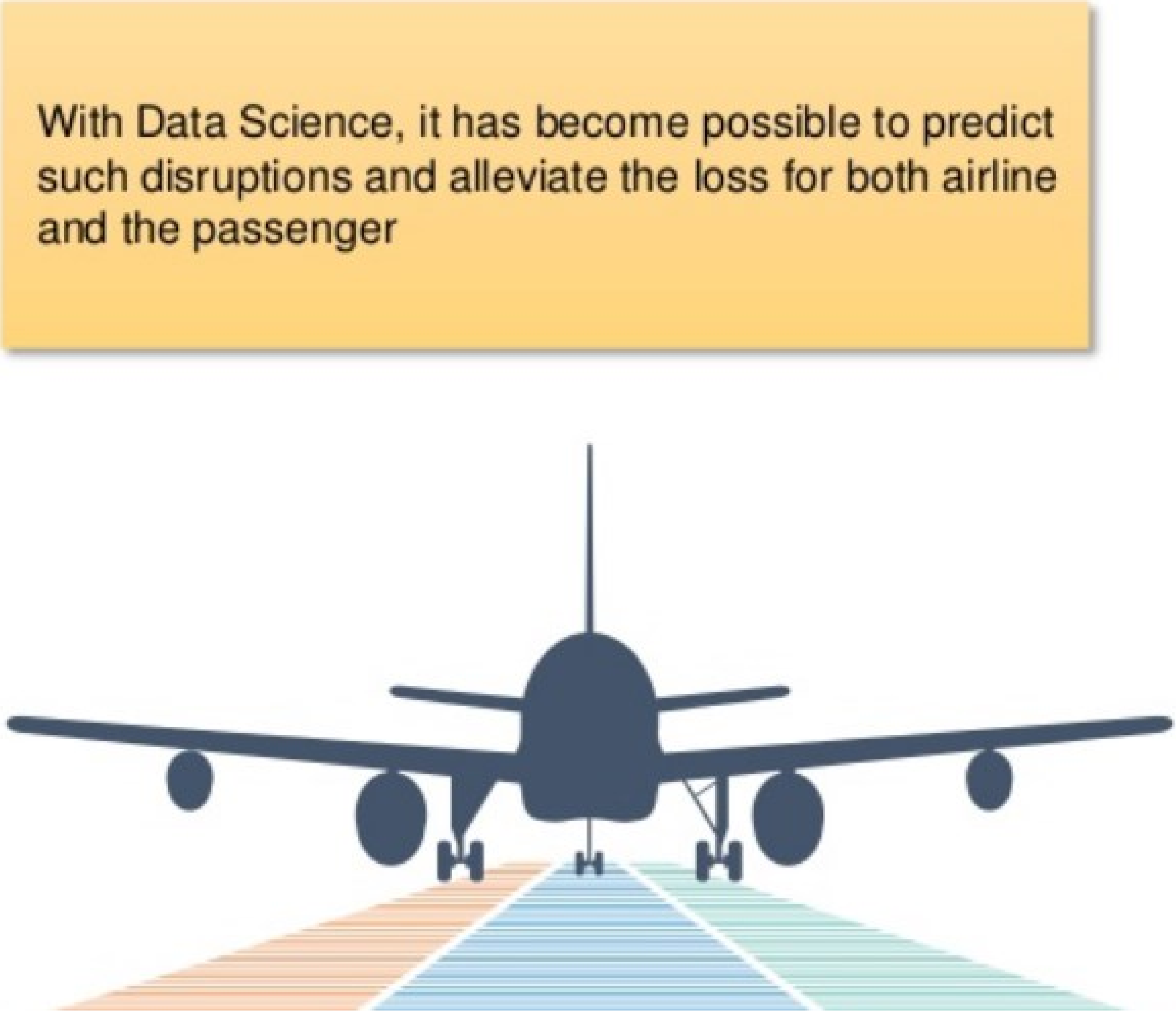


Due to lack of data available, flights are often delayed or cancelled at the last minute

Due to improper route planning, customers don't get the flight for desired time and duration

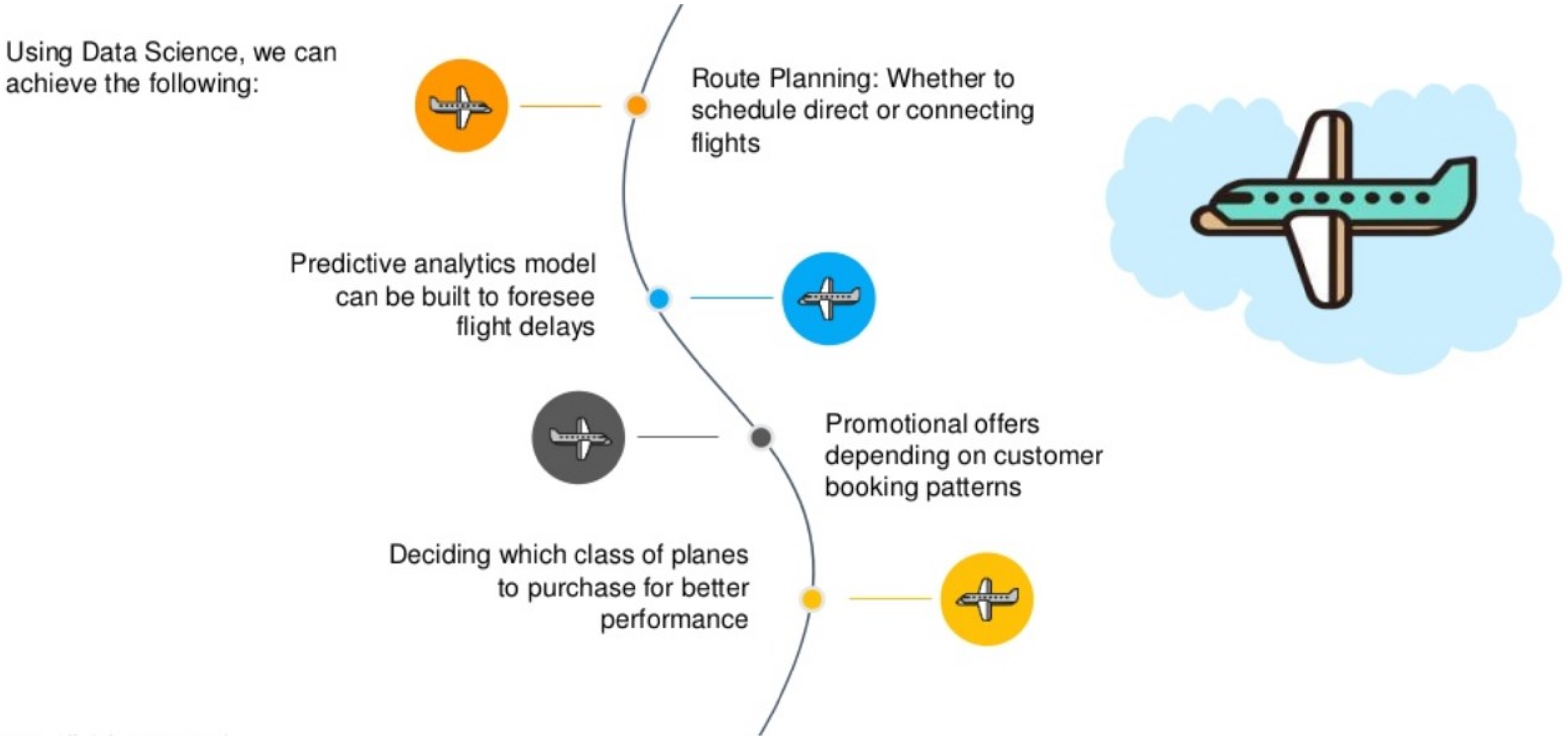
Incorrect decisions in selection of right equipment leads to unplanned delays and cancellations

Need of Data Science



With Data Science, it has become possible to predict such disruptions and alleviate the loss for both airline and the passenger

Need of Data Science



Using Data Science, we can achieve the following:

Route Planning: Whether to schedule direct or connecting flights

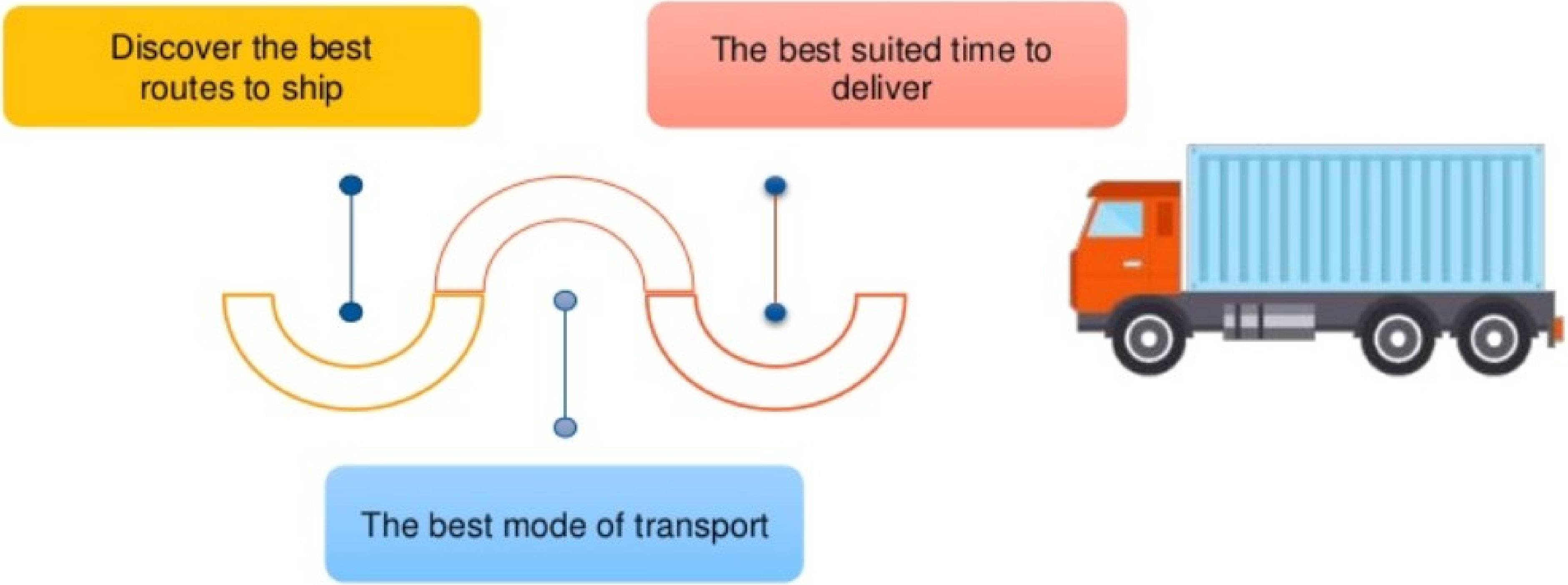
Predictive analytics model can be built to foresee flight delays

Promotional offers depending on customer booking patterns

Deciding which class of planes to purchase for better performance

Need of Data Science

Logistics companies like FedEx are using Data Science models for operational efficiency

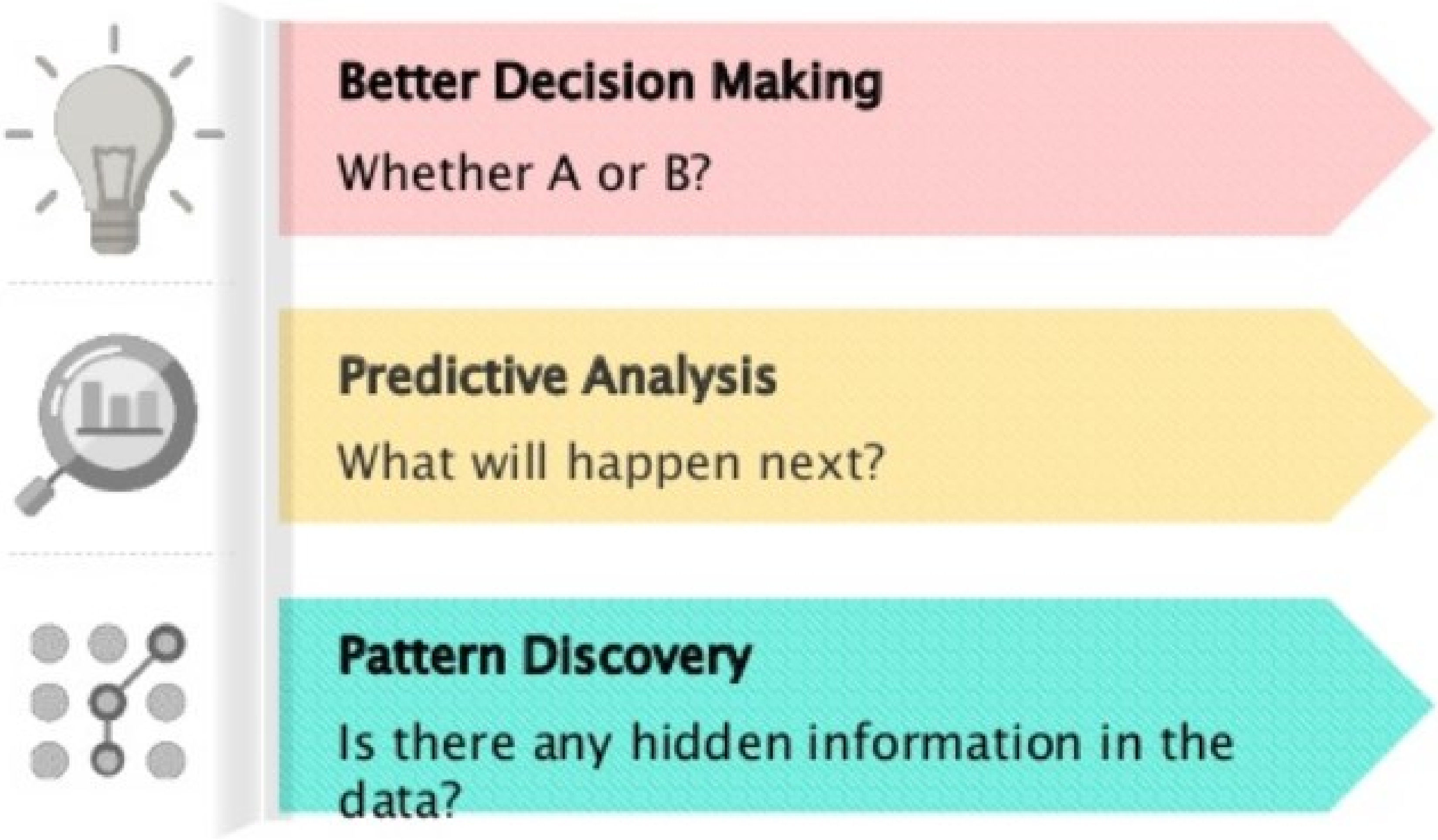


Discover the best routes to ship

The best suited time to deliver

The best mode of transport

So Data Science is mainly needed for:



Better Decision Making

Whether A or B?

Predictive Analysis

What will happen next?

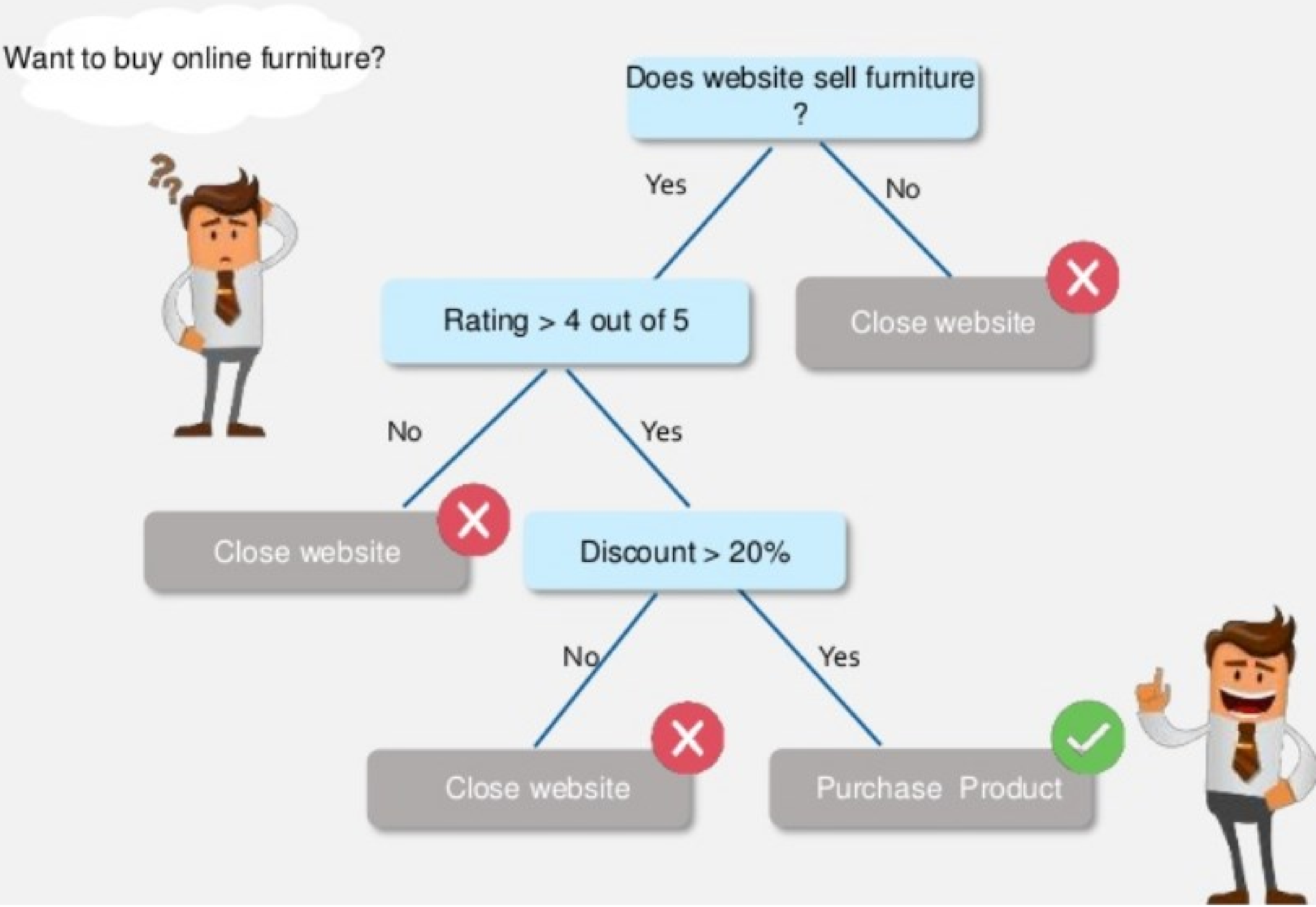
Pattern Discovery

Is there any hidden information in the data?

What is Data Science?

Suppose, you have decided to buy furniture online for your new office

How do you choose the right website?



Want to buy online furniture?

Does website sell furniture ?

Yes

No

Rating > 4 out of 5

Close website

No

Yes

Close website

Discount > 20%

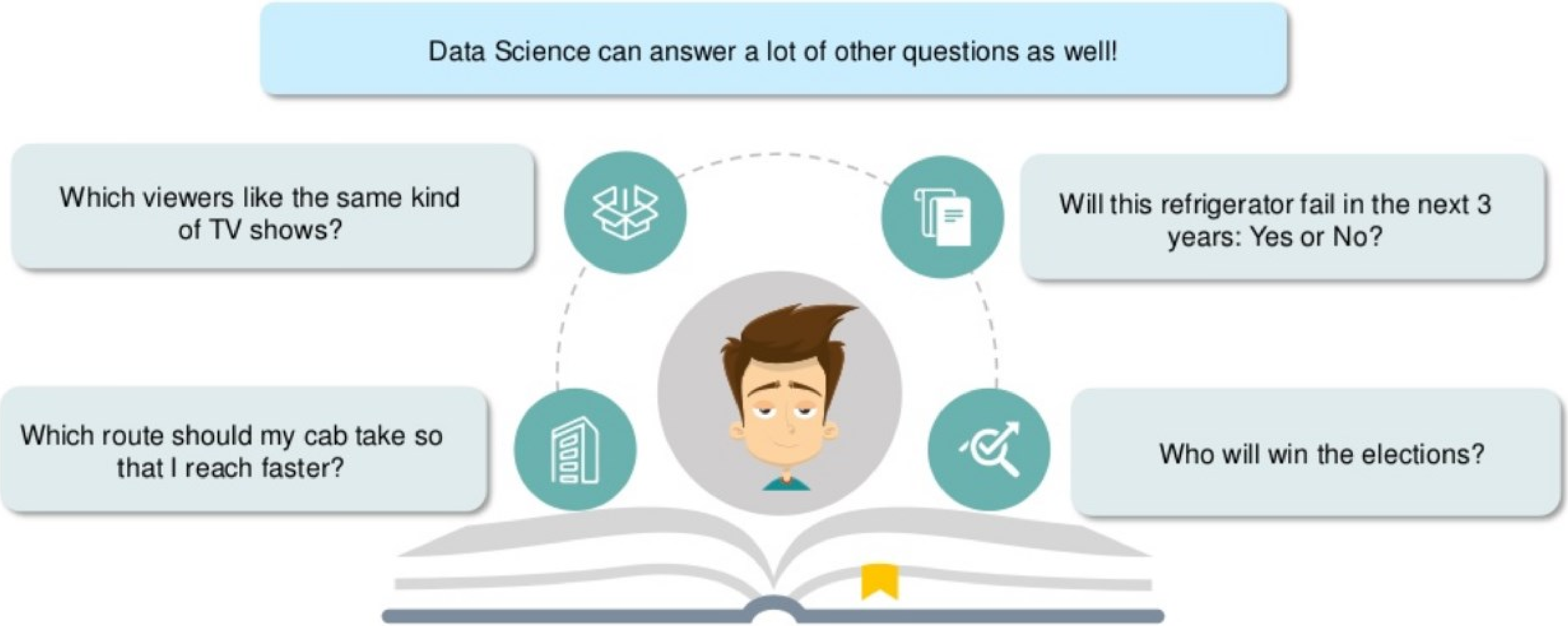
No

Yes

Close website

Purchase Product

What is Data Science?



Data Science can answer a lot of other questions as well!

Which viewers like the same kind of TV shows?

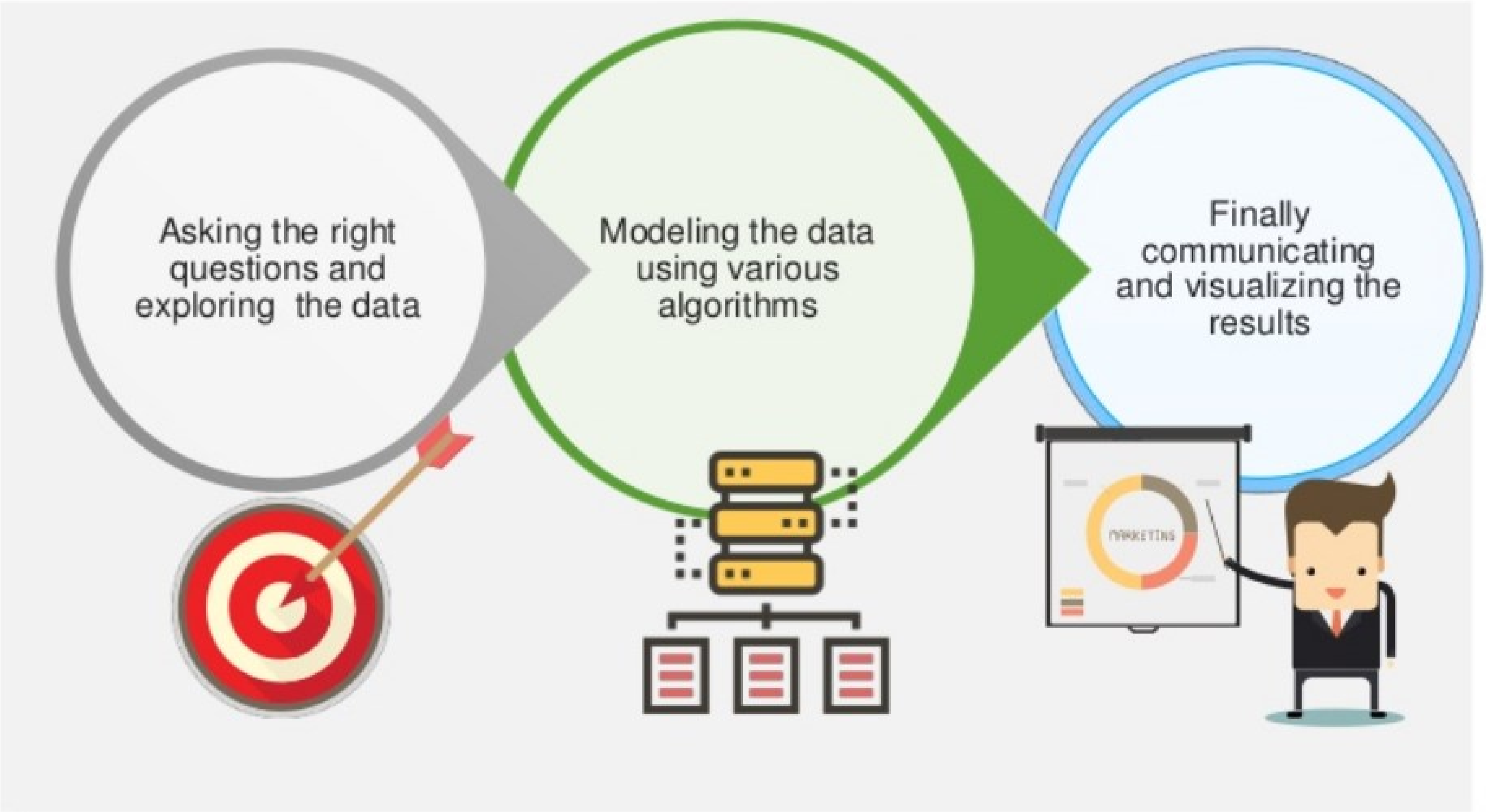
Will this refrigerator fail in the next 3 years: Yes or No?

Which route should my cab take so that I reach faster?

Who will win the elections?

What is Data Science?

So, Data Science or Data-driven Science is about:



Asking the right questions and exploring the data

Modeling the data using various algorithms

Finally communicating and visualizing the results

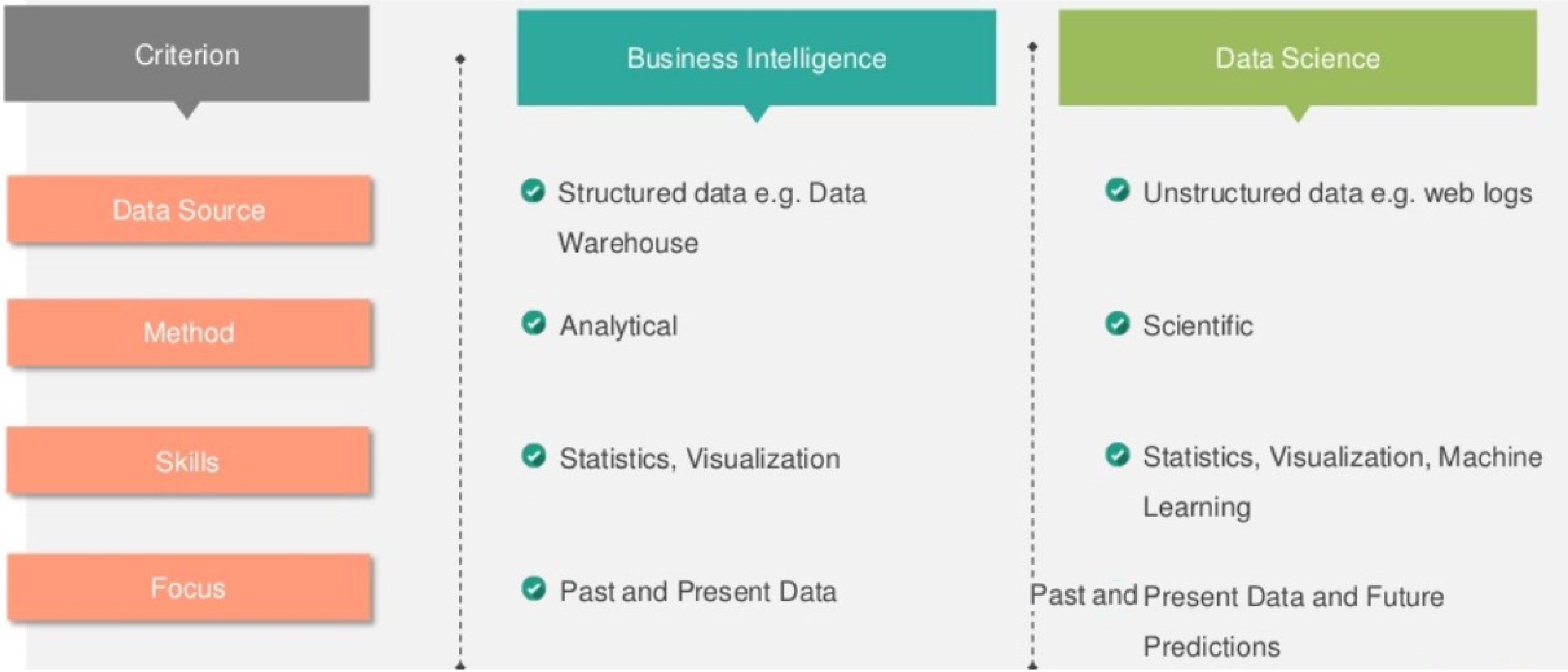
Data Analytics

- **Data analytics is the process of**
  - **collecting,**
  - **organizing and**
  - **analyzing**

data

- **To uncover**
  - **hidden patterns,**
  - **correlations,**
  - **market trends,**
  - **customer preferences and**
  - **other useful business insights.**
- **The analytical findings can lead to**
  - **more effective marketing,**
  - **new revenue opportunities,**
  - **better customer service,**
  - **improved operational efficiency,**
  - **competitive advantages over rival organizations**
  - **and other business benefits.**

Data Science Vs Business Intelligence



Criterion	Business Intelligence	Data Science
Data Source	✓ Structured data e.g. Data Warehouse	✓ Unstructured data e.g. web logs
Method	✓ Analytical	✓ Scientific
Skills	✓ Statistics, Visualization	✓ Statistics, Visualization, Machine Learning
Focus	✓ Past and Present Data	Past and Present Data and Future Predictions

Pre-Requisites 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

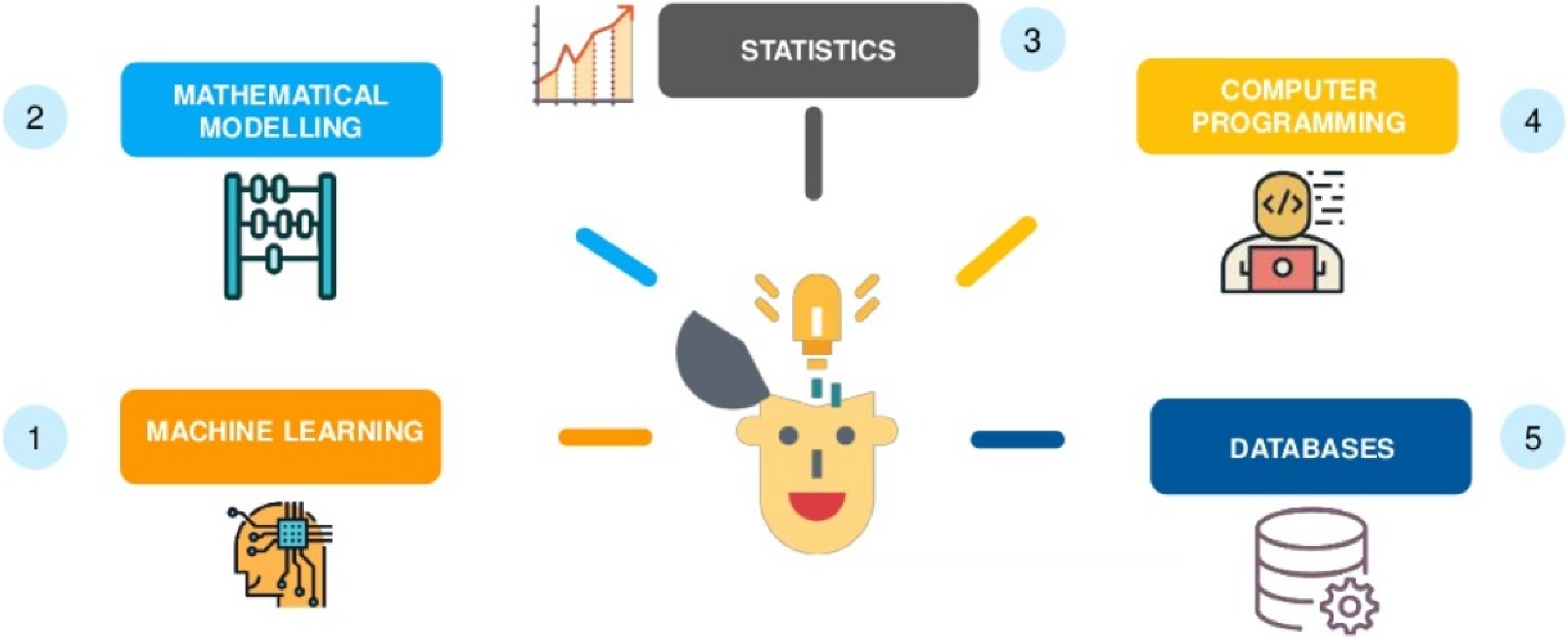
COMMON SENSE

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

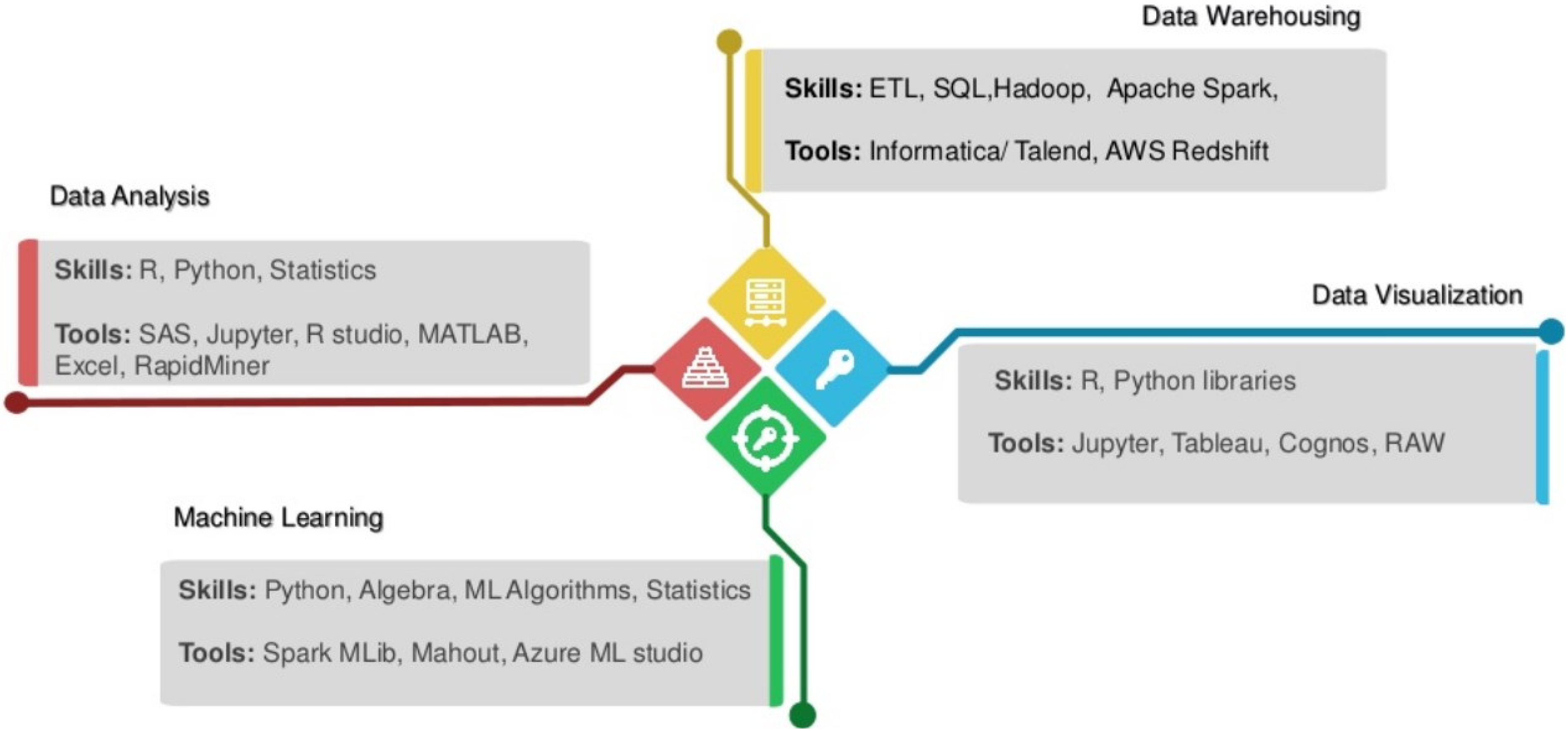
COMMUNICATION SKILLS

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

Pre-Requisites for Data Science



Tools 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

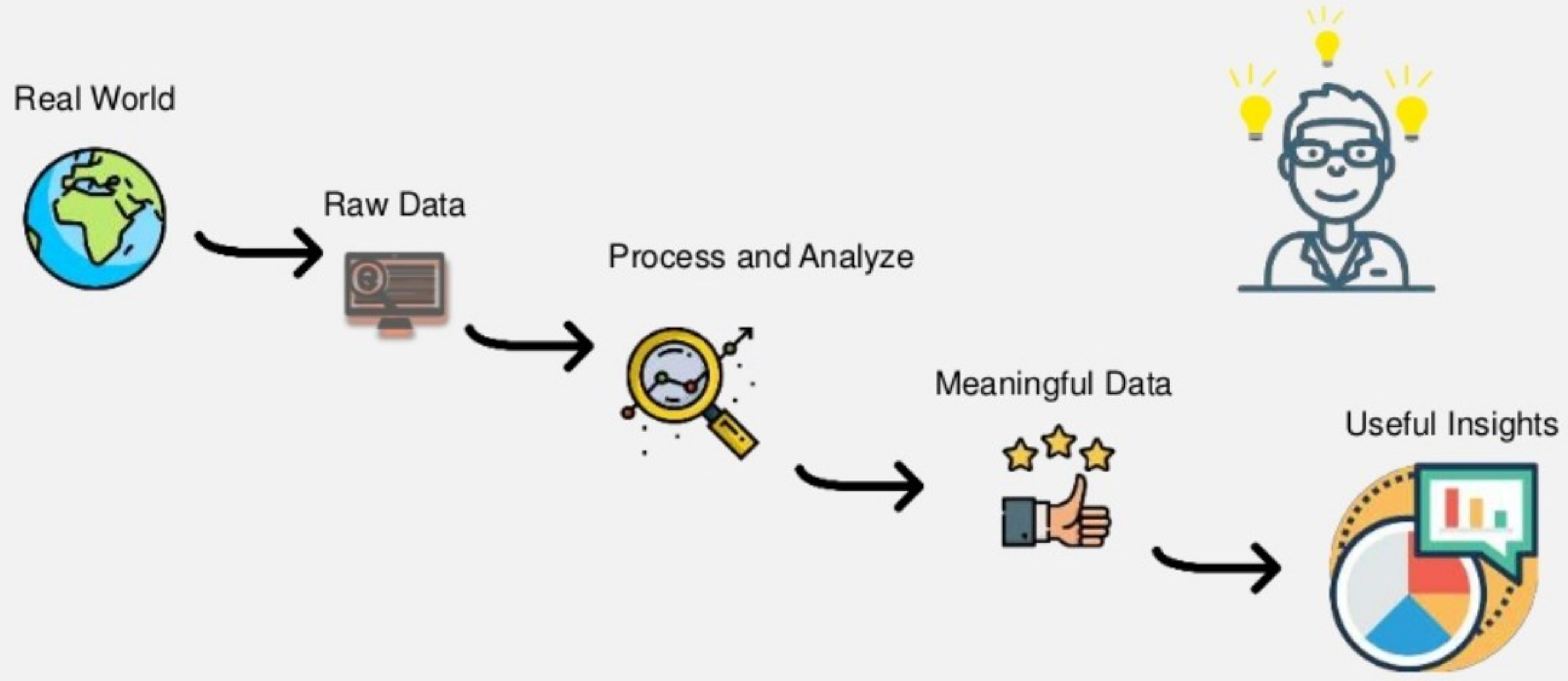
Tools: Jupyter, Tableau, Cognos, RAW

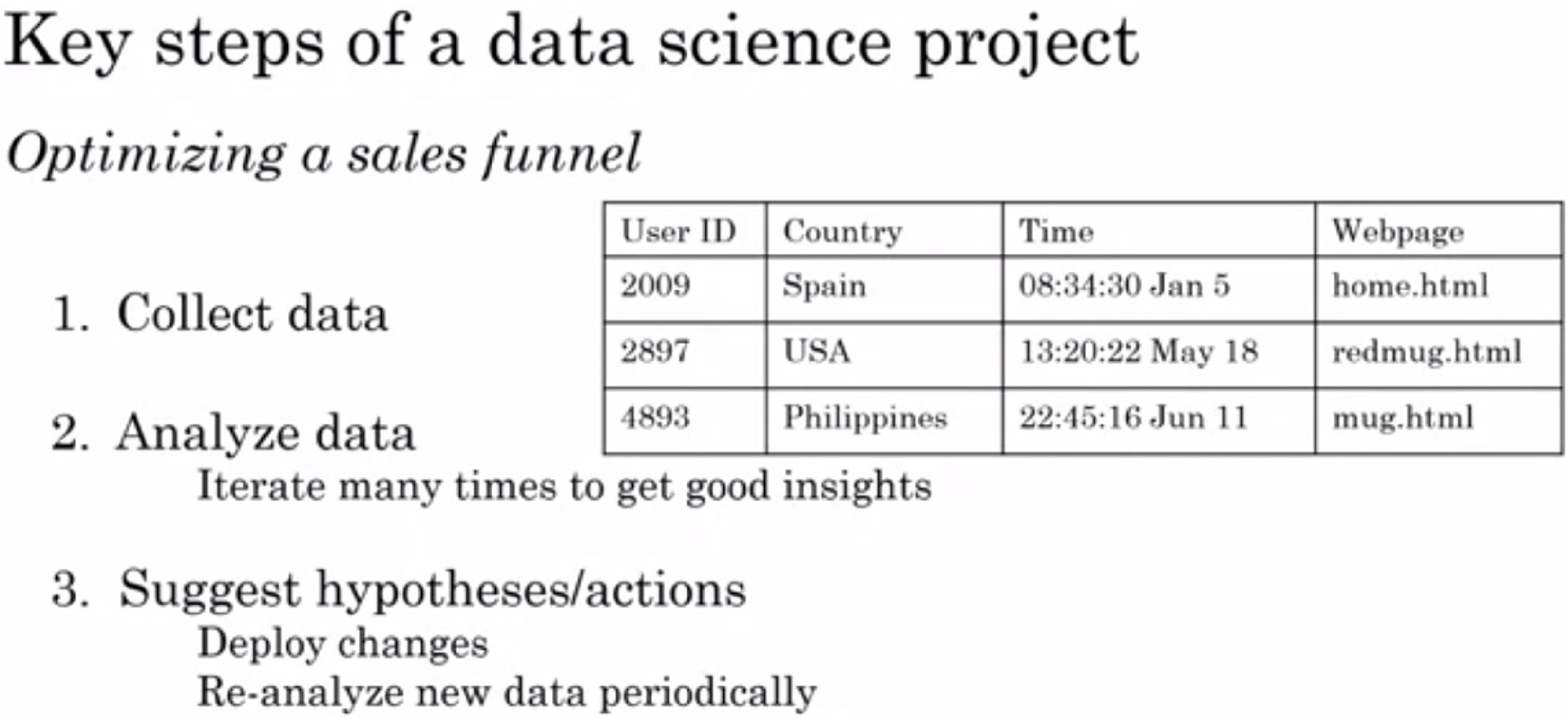
Machine Learning

Skills: Python, Algebra, ML Algorithms, Statistics

Tools: Spark MLib, Mahout, Azure ML studio

Data Science Process





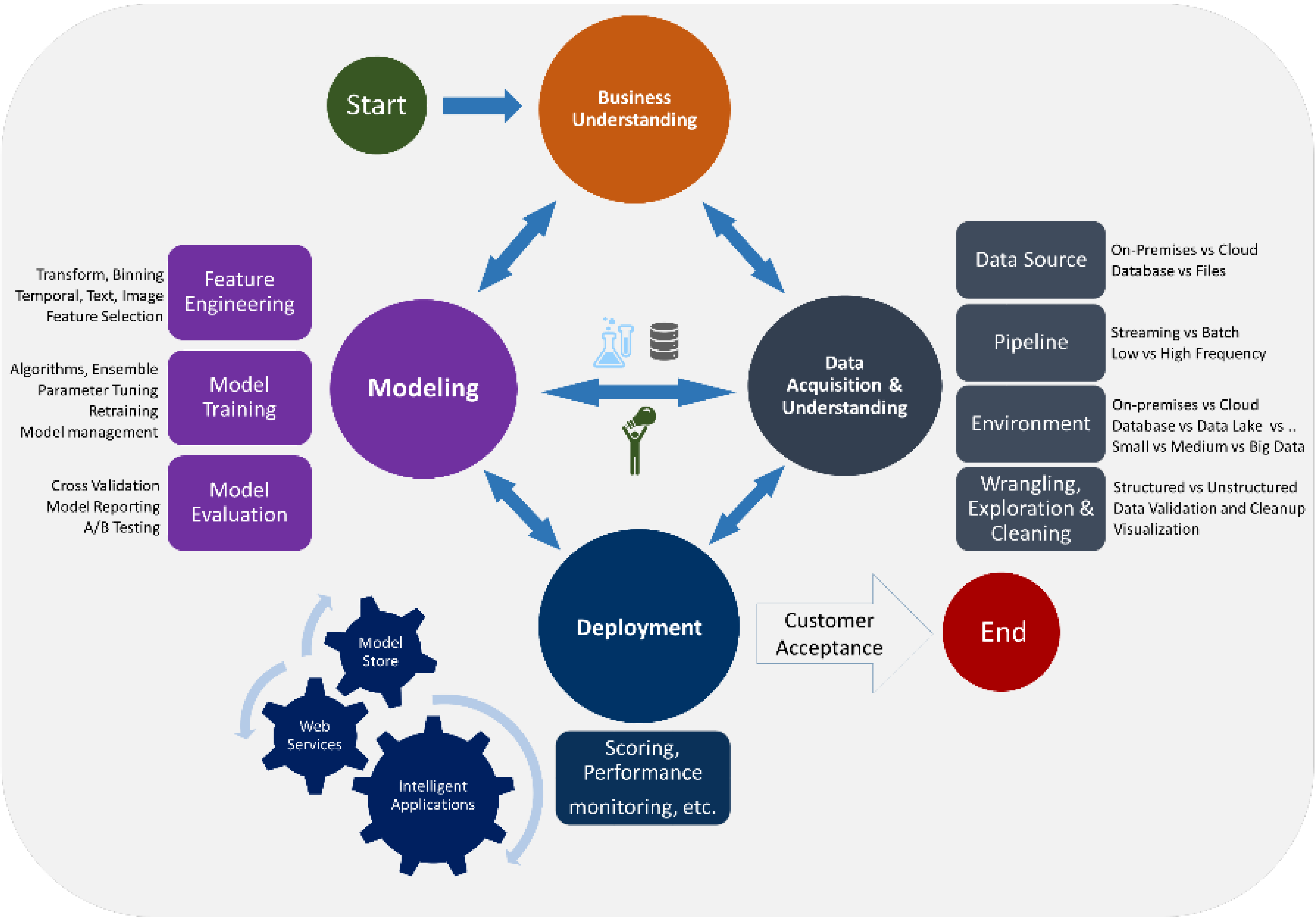
Key steps of a data science project

Optimizing a sales funnel

User ID	Country	Time	Webpage
2009	Spain	08:34:30 Jan 5	home.html
2897	USA	13:20:22 May 18	redmug.html
4893	Philippines	22:45:16 Jun 11	mug.html

1. Collect data
2. Analyze data
   Iterate many times to get good insights
3. Suggest hypotheses/actions
   Deploy changes
   Re-analyze new data periodically

Data Science Lifecycle



Start

Business Understanding

Data Acquisition & Understanding

Data Source
On-Premises vs Cloud
Database vs Files

Pipeline
Streaming vs Batch
Low vs High Frequency

Environment
On-premises vs Cloud
Database vs Data Lake vs ..
Small vs Medium vs Big Data

Wrangling, Exploration & Cleaning
Structured vs Unstructured
Data Validation and Cleanup
Visualization

Modeling

Feature Engineering
Transform, Binning
Temporal, Text, Image
Feature Selection

Model Training
Algorithms, Ensemble
Parameter Tuning
Retraining
Model management

Model Evaluation
Cross Validation
Model Reporting
A/B Testing

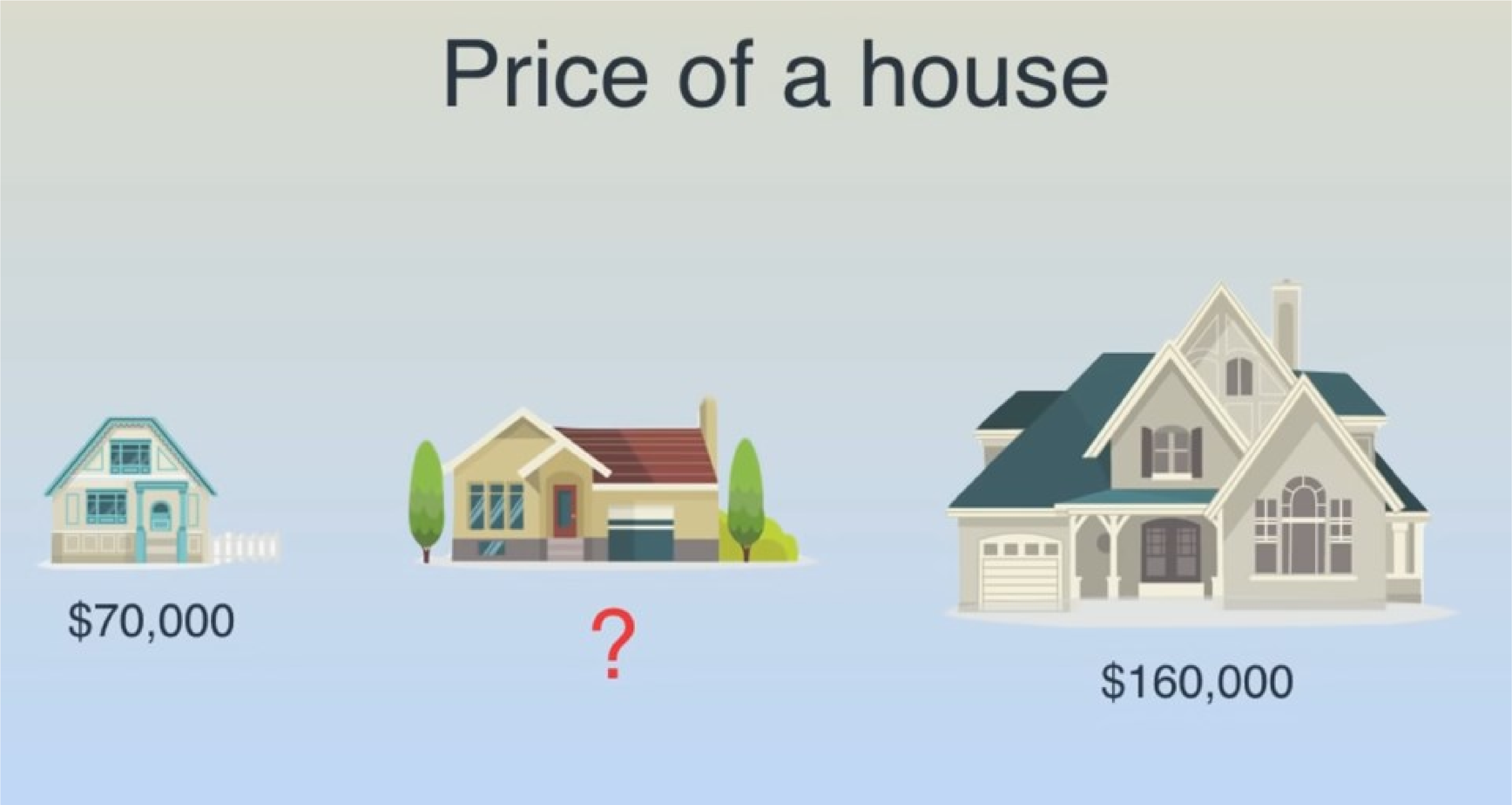
Deployment

Scoring, Performance monitoring, etc.

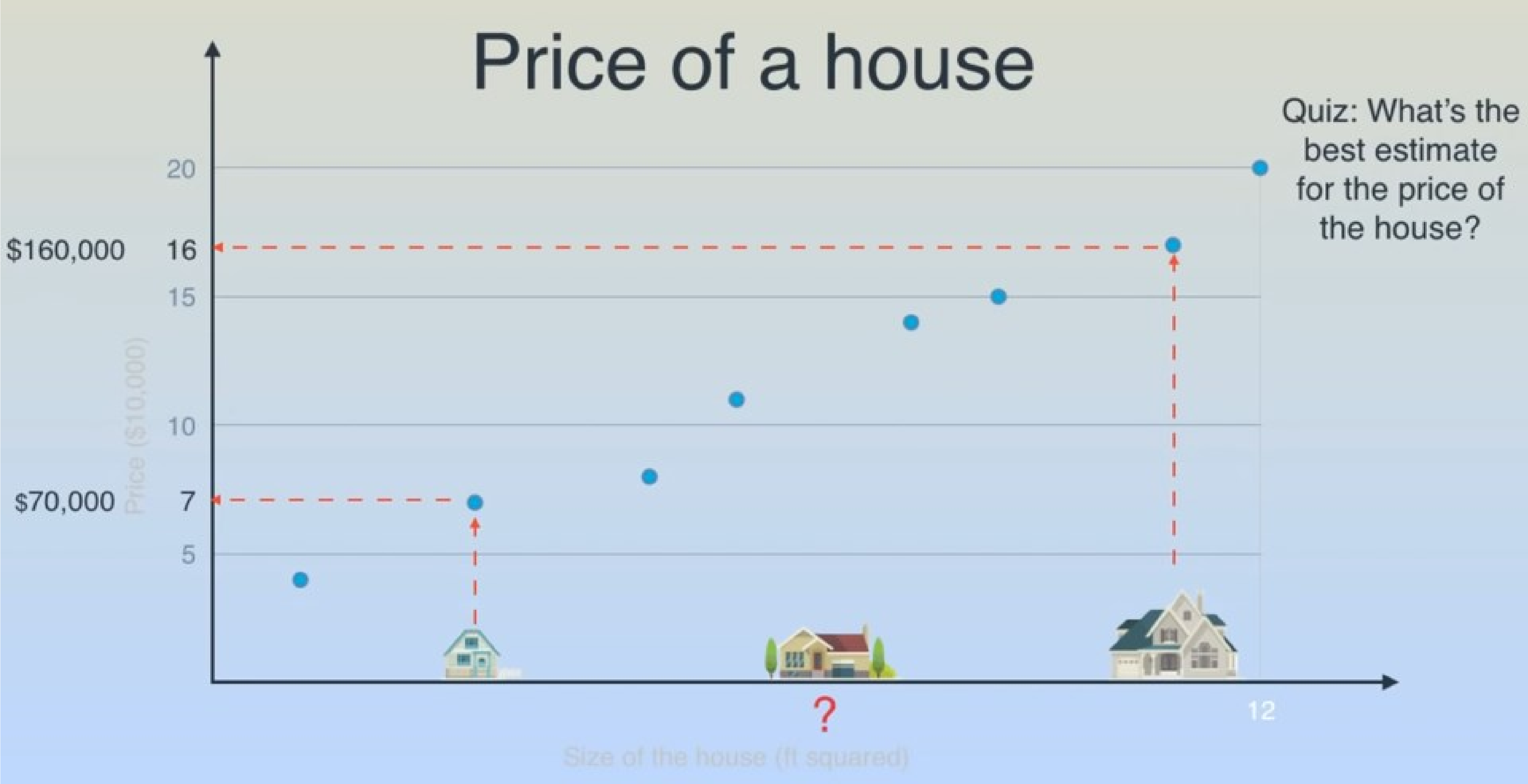
Customer Acceptance

End

Concept Task



Concept Task

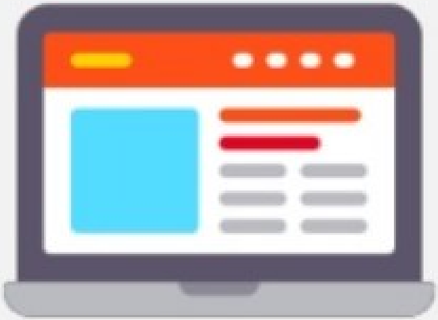


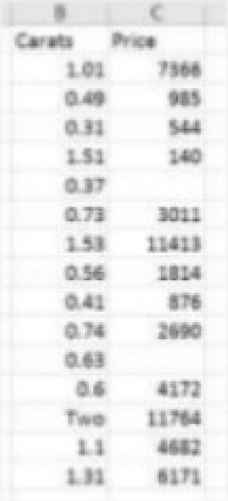
Quiz: What's the best estimate for the price of the house?

Concept Task

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

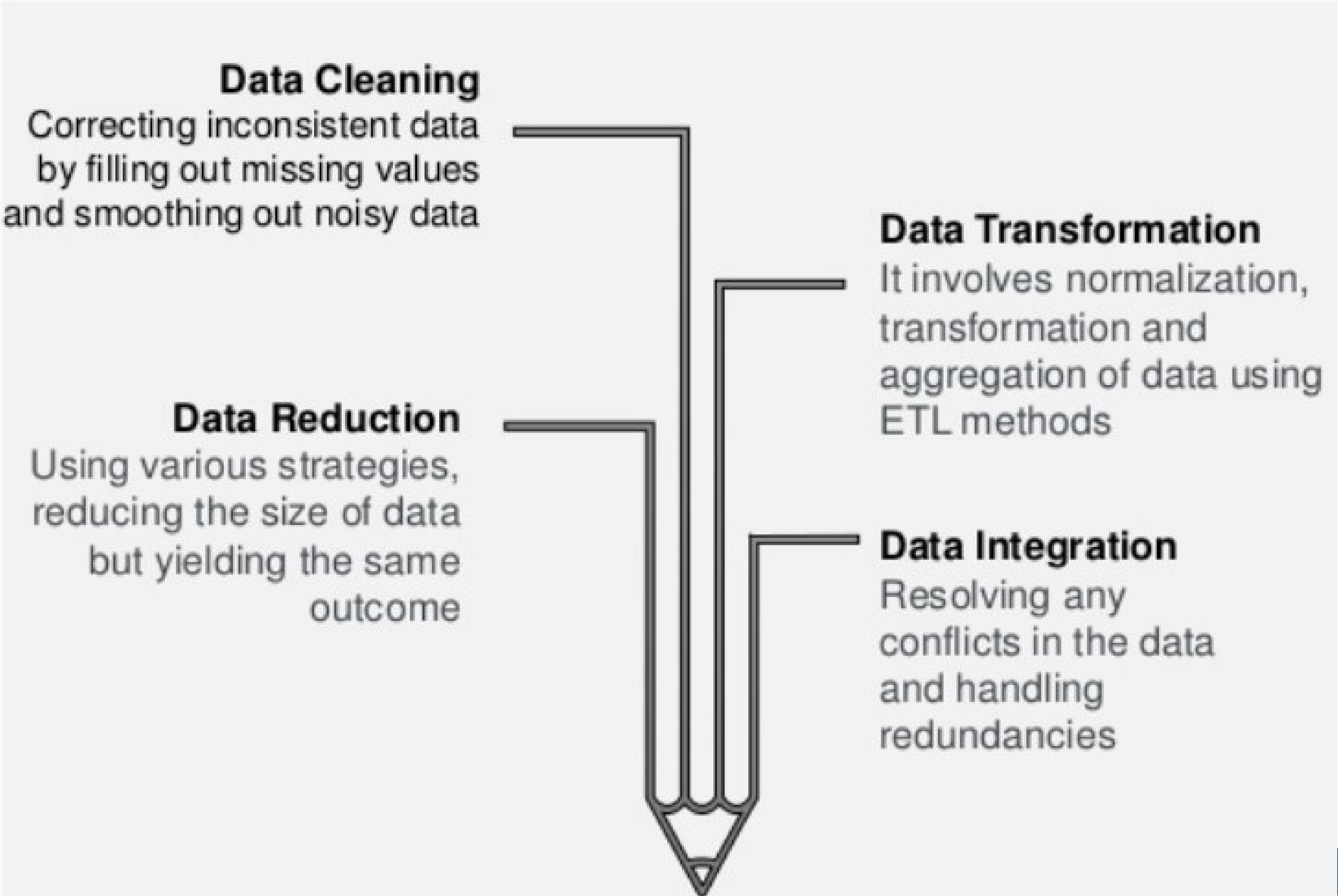




B	C
Carats	Price
1.01	7366
0.49	985
0.31	544
1.51	140
0.37	
0.73	3011
1.53	11413
0.56	1814
0.41	876
0.74	2690
0.63	
0.6	4172
Two	11784
1.1	4682
1.31	6171

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

Data Preparation



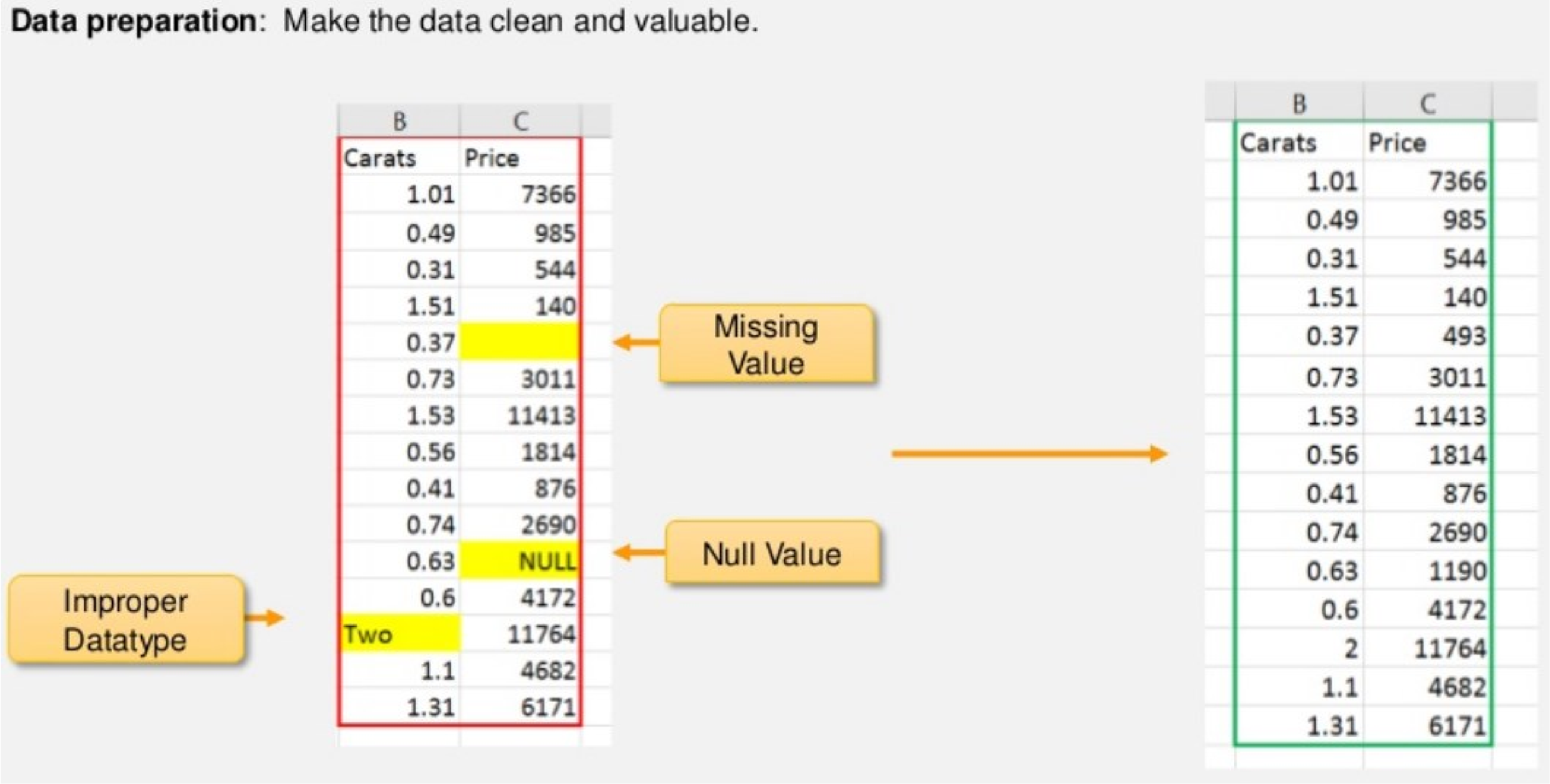
Data Cleaning
Correcting inconsistent data by filling out missing values and smoothing out noisy data

Data Transformation
It involves normalization, transformation and aggregation of data using ETL methods

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

Data Integration
Resolving any conflicts in the data and handling redundancies

Data Preparation: Example



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

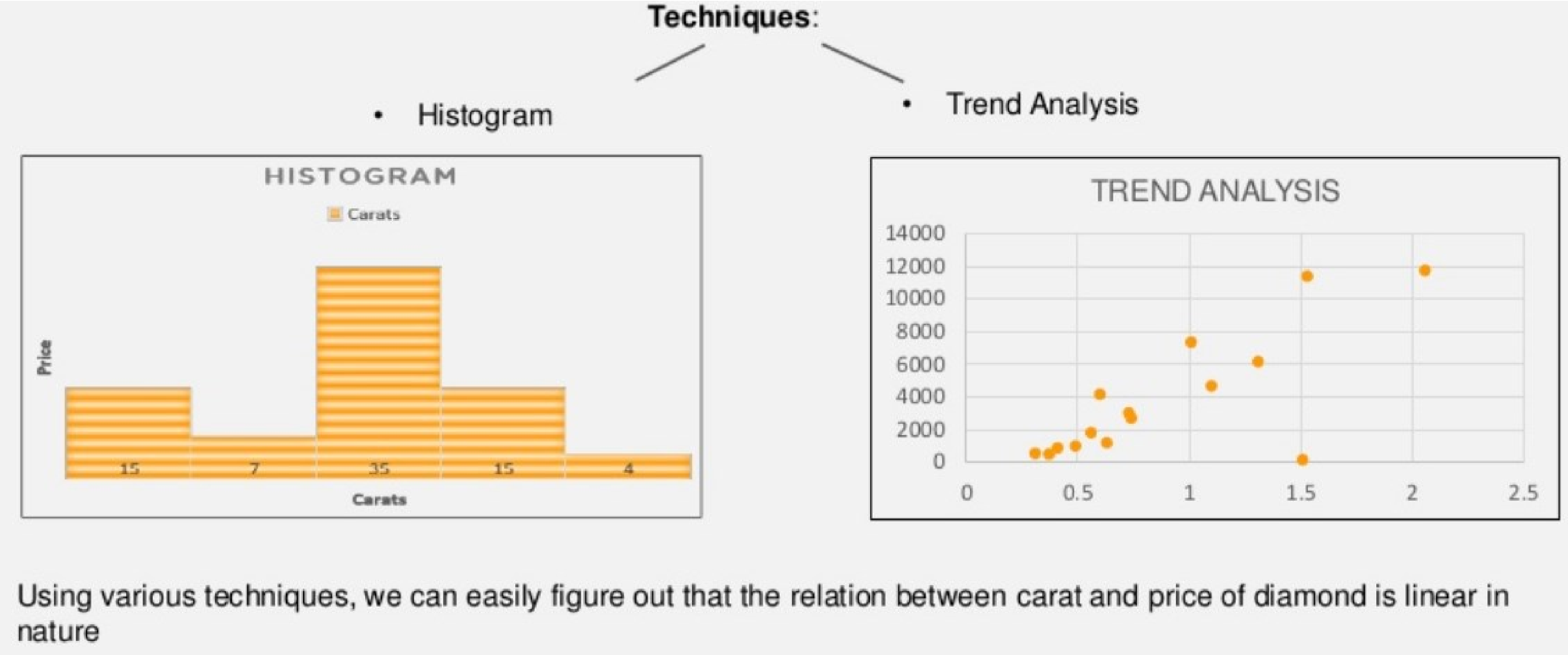
Missing Value

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

Model Planning: Exploratory Data Analysis

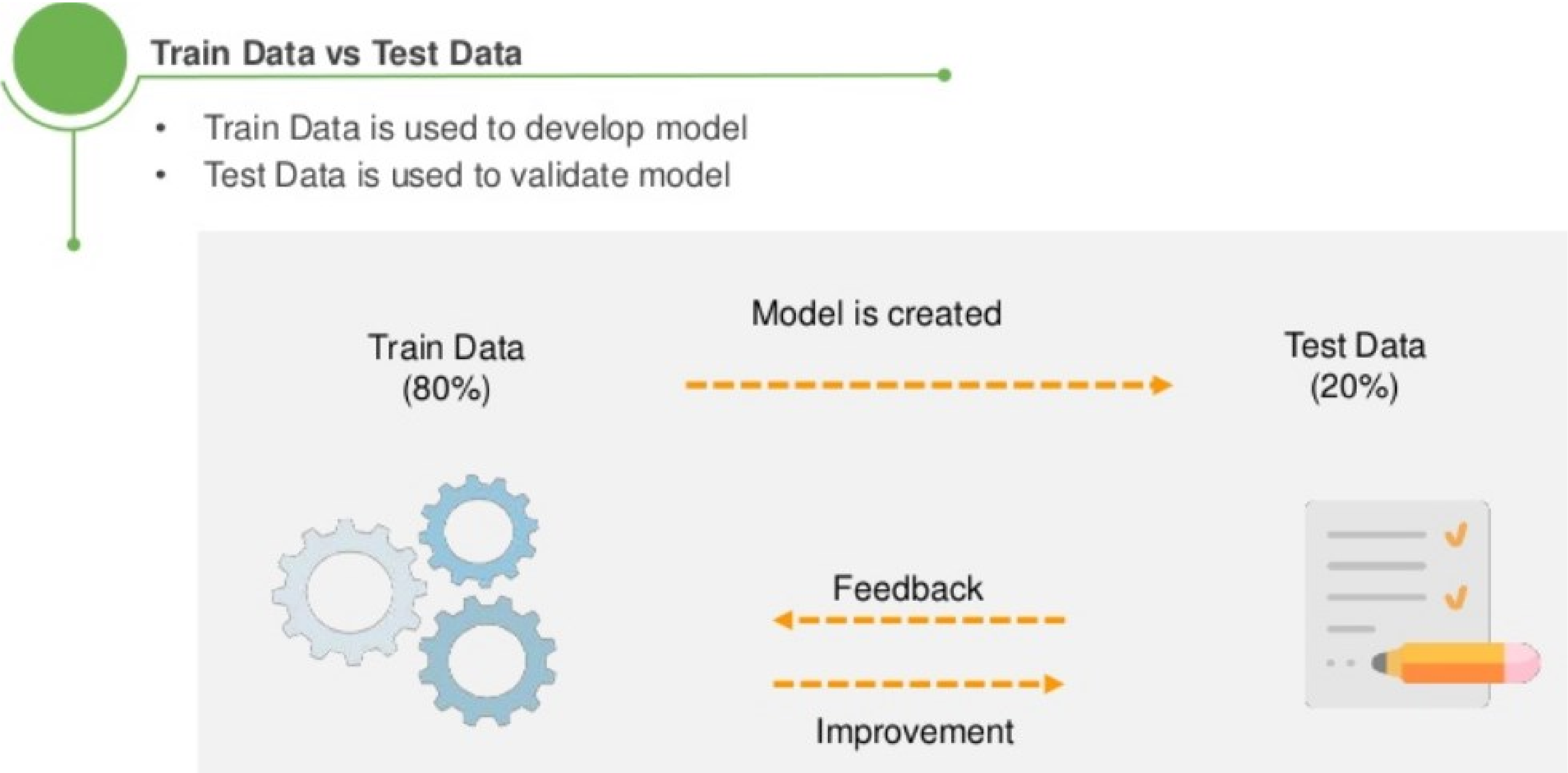


Techniques:

- Histogram
- Trend Analysis

Using various techniques, we can easily figure out that the relation between carat and price of diamond is linear in nature

Model Planning



Train Data vs Test Data

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

Train Data (80%)

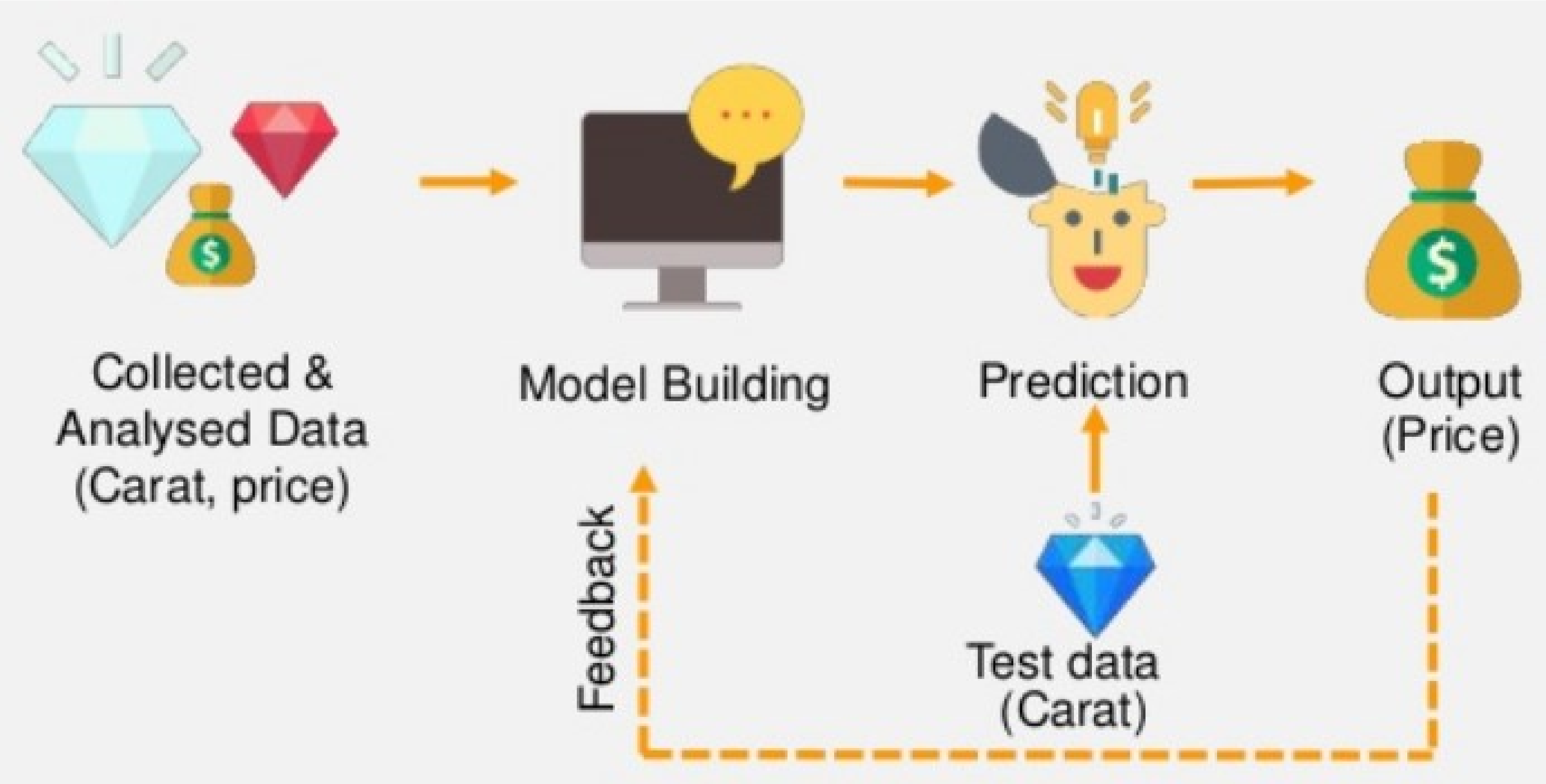
Model is created

Test Data (20%)

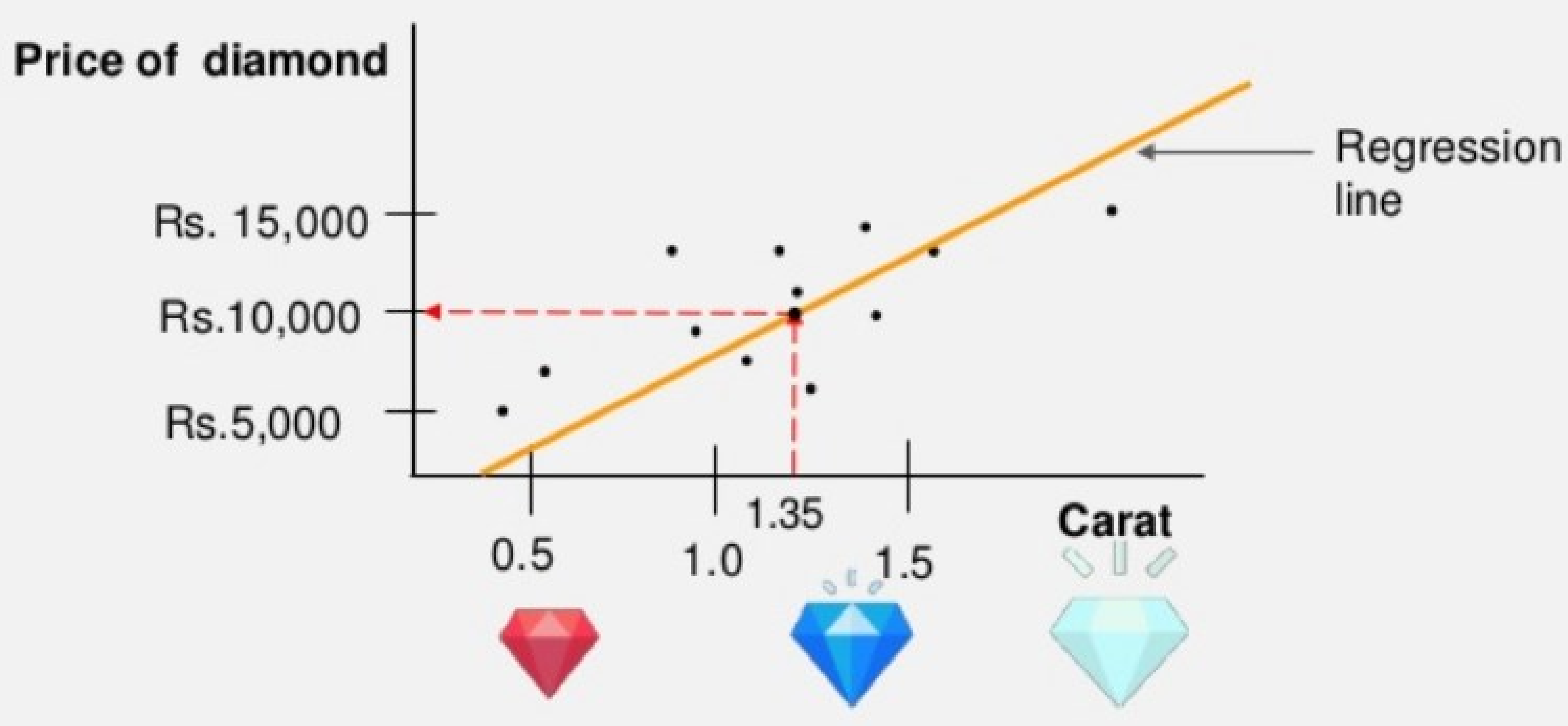
Feedback

Improvement

Machine Learning Model



Model Output



Types of Analytics

- Analytics is generally broken down into one of four types:
- ***Descriptive*** –Helping to understand what is currently happening based on incoming data.
- ***Diagnostic*** – Helping to understand what outcomes were achieved and why, given a particular data set.
- ***Predictive*** – Helping to infer what scenarios are likely to happen given a particular data set.
- ***Prescriptive*** – Helping to infer the kinds of actions that *should* be taken.

Descriptive Analytics

- **1. Heterogeneous Data**

- **2. Data dispersion characteristics**
  - ➢ **Median, Mode, Max, Min, Quantiles, Range, MidRange, Variance, Standard Deviation**

- **3. Data Visualization**
  - ➢ **Line Chart, Box Plot, Q Plot, Heat Maps, Histograms**

Diagnostic Analytics

- **Q-Q Plot**
- **Covariance, correlations**
- **Frequent Patterns**
- **Association Mining**

Association and Correlation Analysis

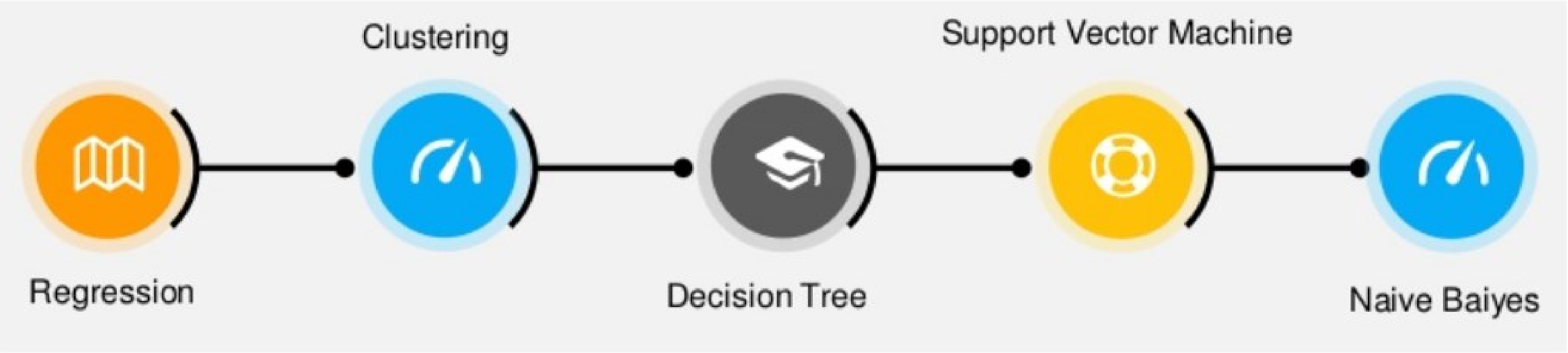
- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association, correlation vs. causality
  - A typical association rule
    - Diaper → Beer [0.5%, 75%] (support, confidence)
  - Are strongly associated items also strongly correlated?

Cluster Analysis

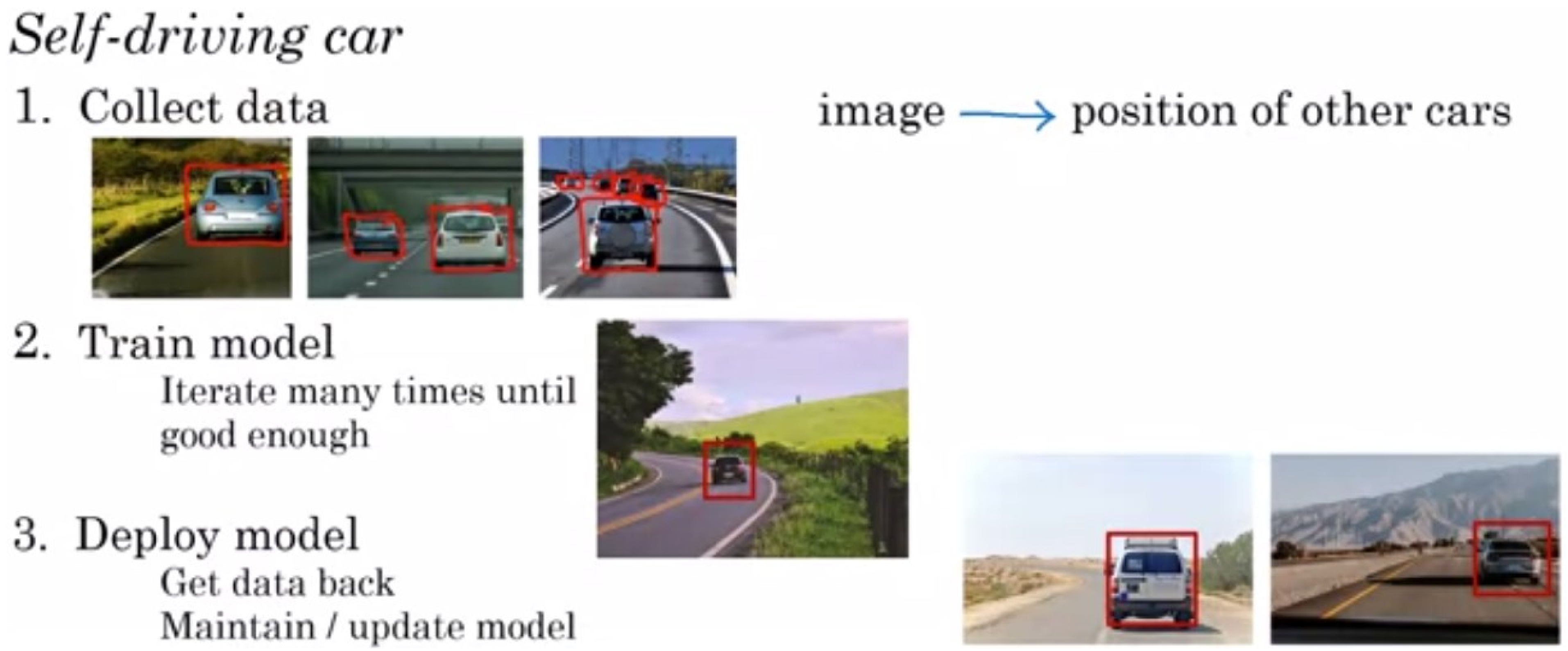
- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications

Predictive Analytics

- **Classification**
- **Regression**



Machine Learning: WorkFlow



Self-driving car

1. Collect data

image ⟶ position of other cars

2. Train model

   Iterate many times until good enough

3. Deploy model

   Get data back

   Maintain / update model

Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

- **Sequence, trend and evolution analysis**
  - Trend, time-series, and deviation analysis: e.g., regression and value prediction
  - **Sequential pattern mining**
    - e.g., first buy digital camera, then buy large SD memory cards
  - **Periodicity analysis**
  - Motifs and biological sequence analysis
    - Approximate and consecutive motifs
  - Similarity-based analysis
- **outlier analysis,**

Prescriptive Analytics

- Rules
- Recommendations

Thank You

shahidmawan@gmail.com

https://sites.google.com/site/shahidmawan/machine-learning

https://github.com/shahidmawan

Course Resources

- https://github.com/shahidmawan/LearnPython/blob/master/Python_Language.pdf
- Slides: https://github.com/shahidmawan/Machine-Learning

- https://github.com/shahidmawan/practicalAI/blob/master/notebooks/01_Python.ipynb
- https://github.com/shahidmawan/LearnPython/blob/master/Introduction%20session%20of%20Python%20.ipynb
- https://github.com/shahidmawan/LearnPython/blob/master/Advance%20Python%20session%20.ipynb

- https://github.com/shahidmawan/practicalAI/blob/master/notebooks/03_Pandas.ipynb Exploratory Analysis

- https://github.com/shahidmawan/LearnPython/blob/master/Charts.ipynb
- https://github.com/shahidmawan/DataVisualization-Python/blob/master/DataVisualization-Python.ipynb DataVisualization-Python
- https://github.com/shahidmawan/PythonDataScienceHandbook/tree/master/notebooks

- https://github.com/shahidmawan/LearnPython/blob/master/Numpy%20Exercise%20-%20Solutions.ipynb

Today's Task

- https://github.com/shahidmawan/numpy-100/blob/master/100_Numpy_exercises.ipynb