What is Data Science

- Data science is the field of study that combines domain expertise, programming skills, and knowledge of math and statistics to extract meaningful insights from data.
- In turn, these systems generate insights that analysts and business users translate into business value.
- With data science, companies can predict the success rate of their strategies.

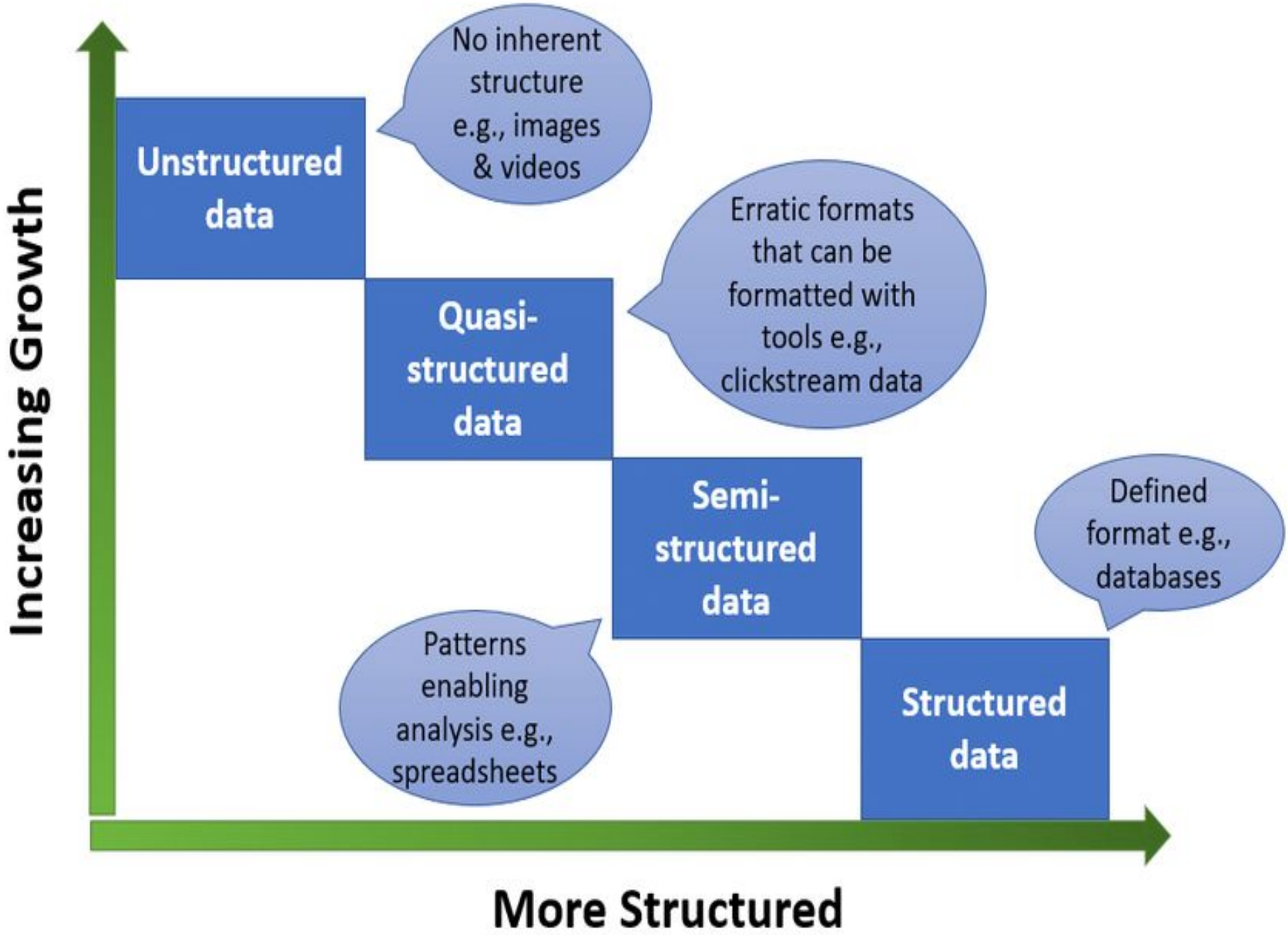
TYPES OF BIG DATA

Structured

Semi-structured

Quasi-structured

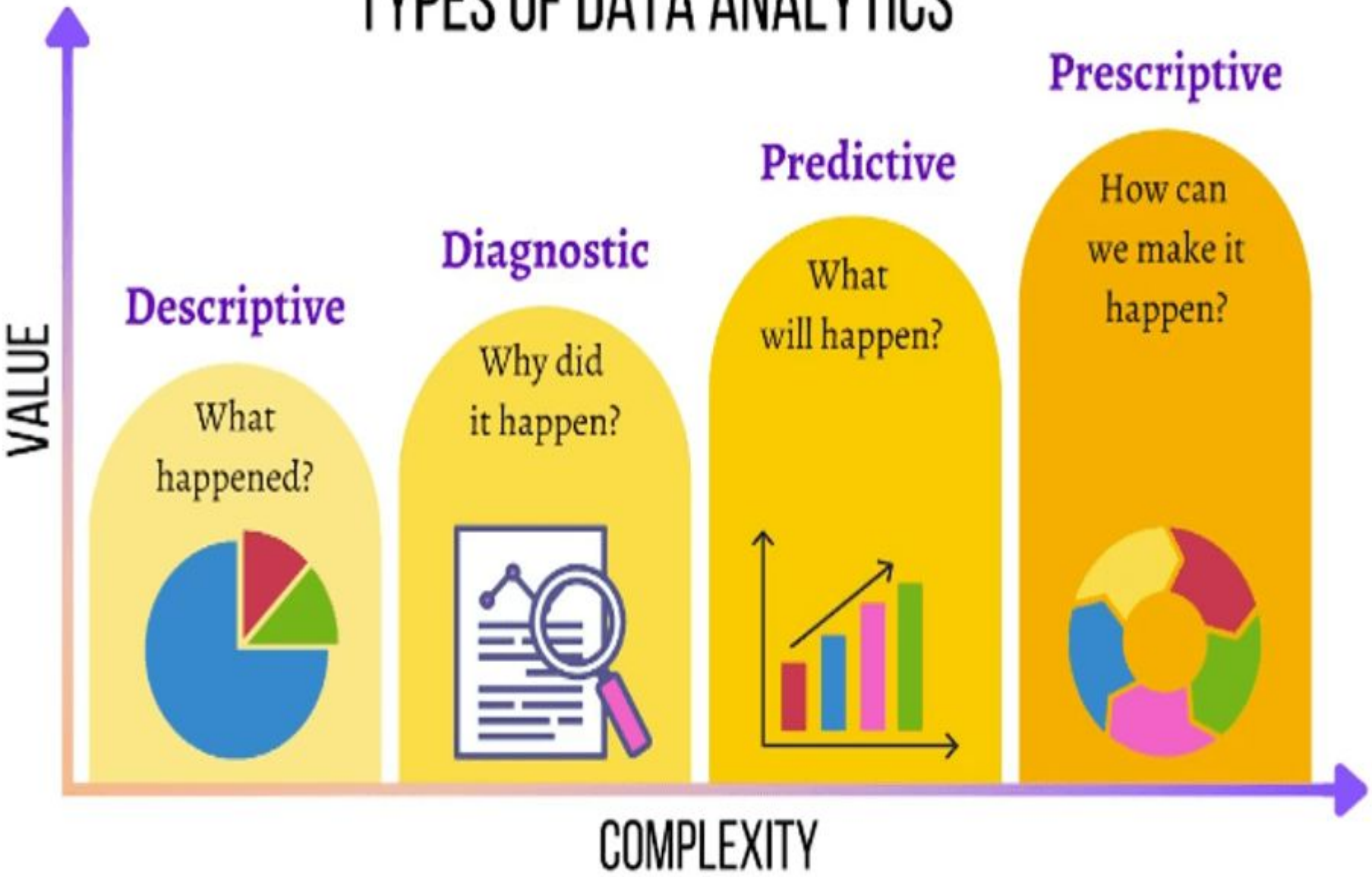
Unstructured


Increasing Growth
Unstructured data
No inherent structure e.g., images & videos
Quasi-structured data
Erratic formats that can be formatted with tools e.g., clickstream data
Semi-structured data
Defined format e.g., databases
Structured data
Patterns enabling analysis e.g., spreadsheets
More Structured

Data For Business Decisions

- In the past, many businesses would take poor decisions due to the lack of surveys , proper data
- It would result in some disastrous decisions leading to losses in millions.
- However, with the presence of a plethora of data and necessary data tools, it is now possible for the data industries to make calculated data-driven decisions.

TYPES OF DATA ANALYTICS

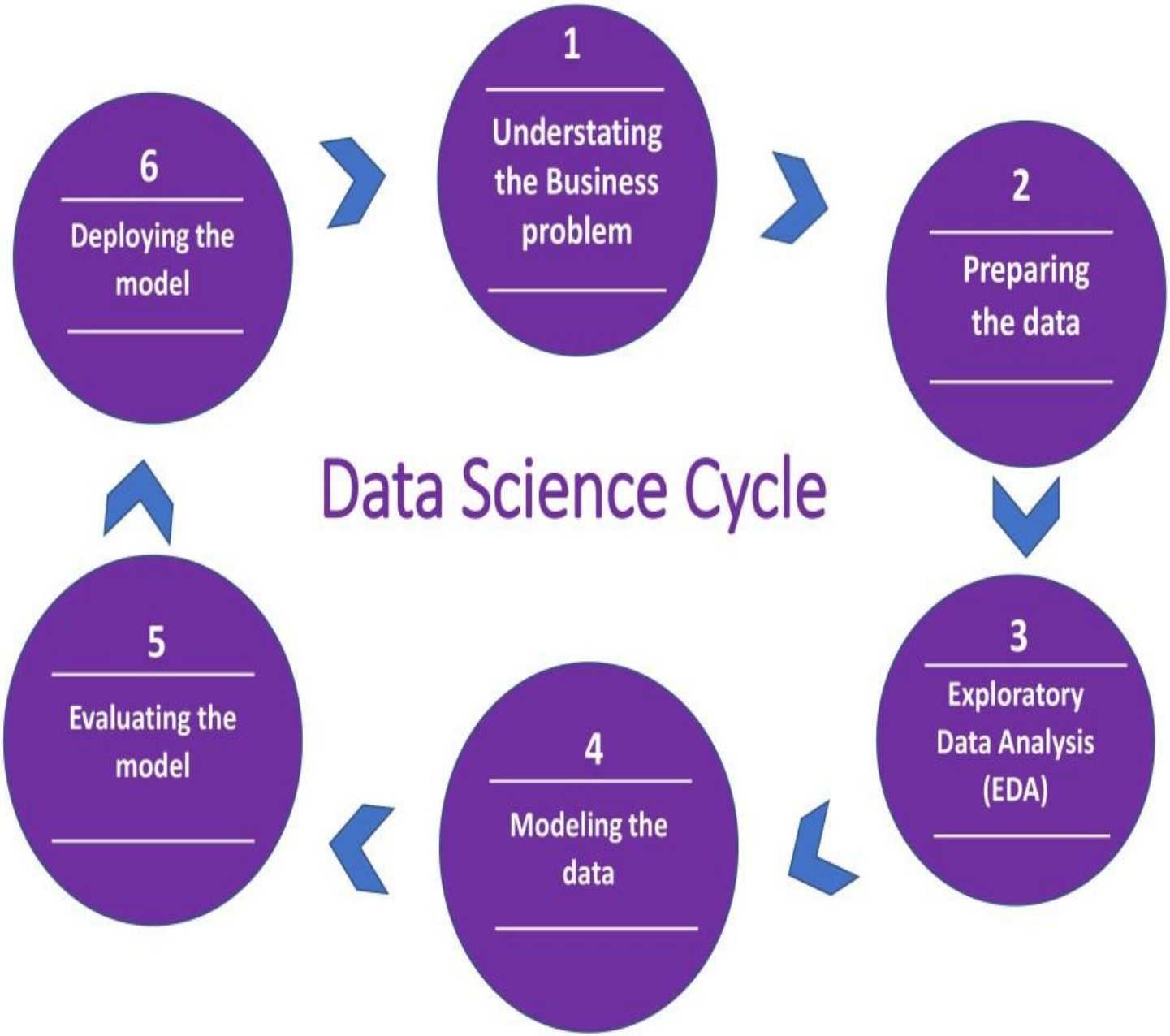


PREDICTIVE ANALYSIS

Predictive Analytics to Predict Outcomes

With the advent of advanced predictive tools and technologies, companies have expanded their capability to deal with diverse forms of data.

In formal terms, predictive analytics is the statistical analysis of data that involves several machine learning algorithms for predicting the future outcome using the historical data.


1
Understating the Business problem
2
Preparing the data
3
Exploratory Data Analysis (EDA)
4
Modeling the data
5
Evaluating the model
6
Deploying the model
Data Science Cycle

Data Science Lifecycle

1. Understand Problem

This phase focuses on clearly defining the problem and aligning it with the overall business objectives.

2. Data Collection

This phase involves acquiring relevant data from various sources, ensuring its quality and consistency.

3. Data Preparation

Raw data rarely comes ready-made. This phase involves cleaning and preparing the data for analysis.

4. Modeling the Data

This phase involves choosing the right algorithms and building models to extract insights from the data.

5. Model Evaluation

It involves testing the model's accuracy and performance, ensuring it provides reliable and actionable insights.

6. Model Deployment

It involves deploying the model into production, making it accessible to stakeholders & integrating it into applications.

BUSINESS UNDERSTANDING

The primary goal of the Business Understanding stage is to understand the business problem and determine the data needed to answer the core business question

The entire cycle revolves around the business goal.

It is essential to understand the business objective clearly because that will be your final goal of the analysis

You need to know if the client wants to reduce credit loss, or if they want to predict the price of a commodity, etc

DIFFERENT DATA COLLECTING SOURCES

- Collecting new data from internet and other sources
- Using the previously collected and stored data
- Reusing someone else's data
- Purchasing data

DATA COLLECTION

Data Collection refers to the systematic process of gathering, measuring, and analyzing information from various sources to get a complete and accurate picture of an area of interest.

This involves the collection of all the available data. It would help if you worked closely with the business team to know what data is present and what data could be used for this business problem, and other information.

This step involves describing the data, their structure, their relevance, their data type.

Explore the data using graphical plots. Basically, extracting any information that you can get about the data by just exploring the data.

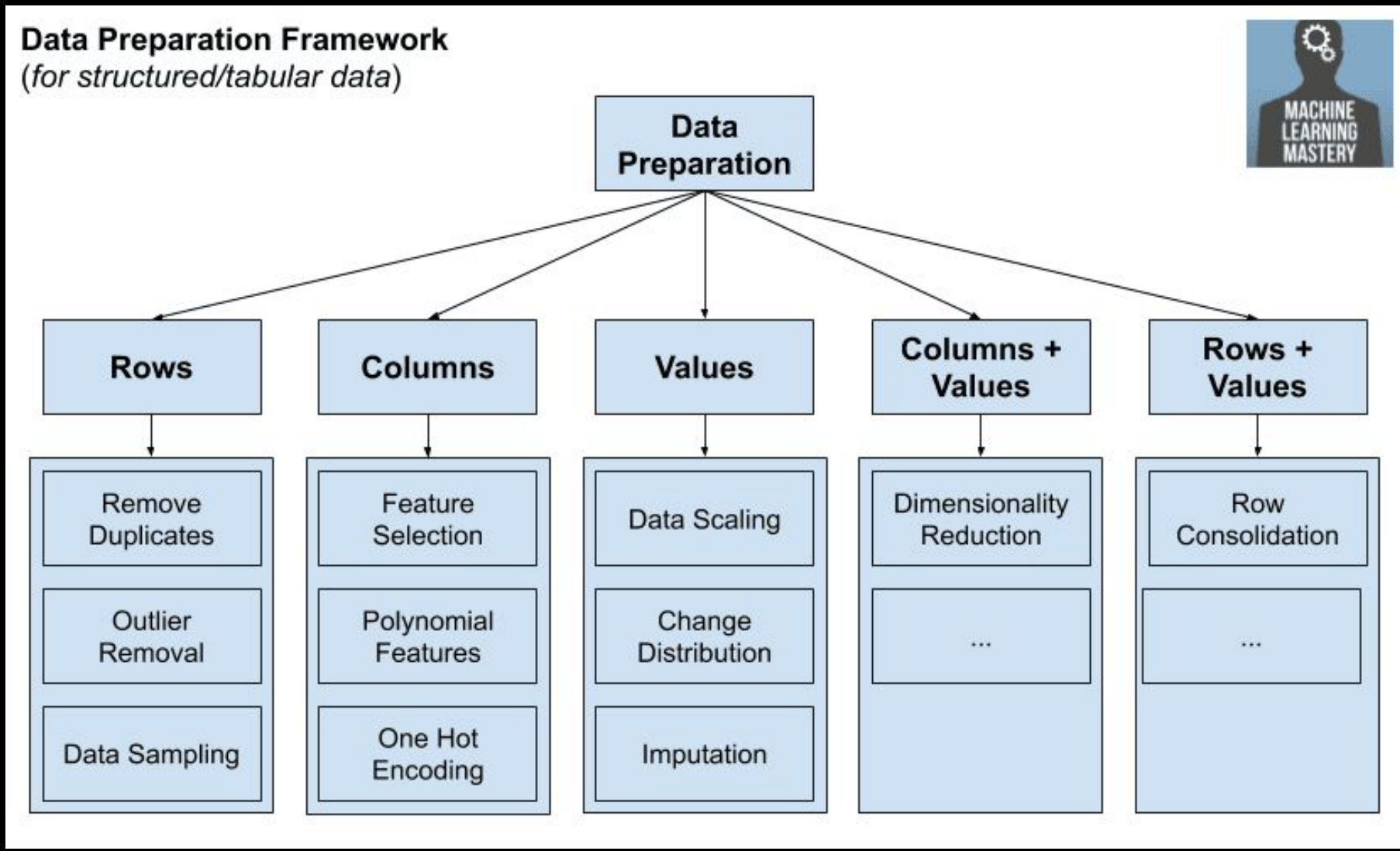
RELATIONSHIP BETWEEN DATA

During the Data Understanding stage, scientists might use descriptive statistics, predictive statistics, or both.

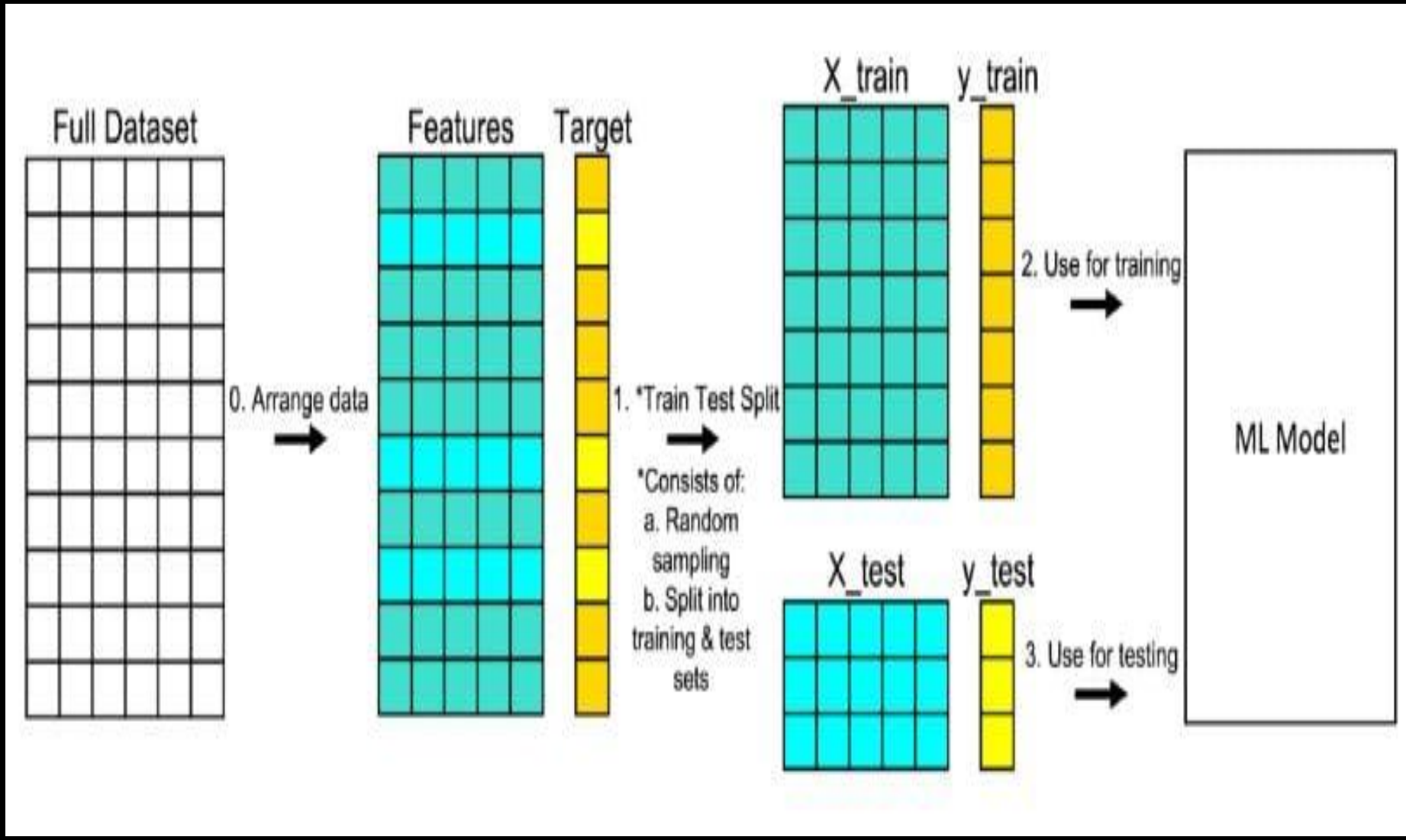
Data scientists commonly apply univariates, and other statistics on each variable, such as mean, median, minimum, maximum, pairwise correlation, and histograms.

Data scientists also use statistics, and histograms to assess data quality.

DATA PREPROCESSING



TRAIN TEST SPLIT



MODEL BUILDING

A model takes the prepared data as input and provides the desired output This step includes choosing the appropriate type of model, whether the problem is a classification problem, or a regression problem or a clustering problem.

After choosing the model family, amongst the various algorithm amongst that family, we need to choose the algorithms to implement and implement them carefully.

We need to tune the hyperparameters of each model to achieve the desired performance

EVALUATING YOUR DATA MODEL'S PERFORMANCE

To evaluate the performance of a data model, data scientists employ various evaluation metrics, such as accuracy, precision, recall, and F1 score,rmse etc.

These metrics quantify the model's predictive accuracy and allow for the comparison of different models or approaches.

If we do not obtain a satisfactory result in the evaluation, we must re-iterate the entire modelling process until the desired level of metrics is achieved.

Model Deployment

The model, after a rigorous evaluation, is finally deployed in the desired format and channel.

- The model can be deployed in a variety of ways, such as a web service, a mobile app, or a desktop application.
- Making predictions: Once the model has been deployed, it can be used to make predictions on new data.

Monitoring the model:

It is important to monitor the model's performance in production to ensure that it is still performing as expected.

Sooner or later, even the most accurate and carefully tested solution starts to degrade

- Instacart's model's accuracy predicting item availability at stores dropped from 93% to 61% due to a drastic shift in shopping habits.
- Bankers question whether credit models trained on good times can adapt to the stress scenarios.
- Trading algorithms misfired in response to market volatility. Some funds had a 21% fall.
- Image classification models had to learn the new normal: a family at home in front of laptops can now mean "work," not "leisure."

On top of this, all sorts of issues occur with live data.

There are input errors and database outages.

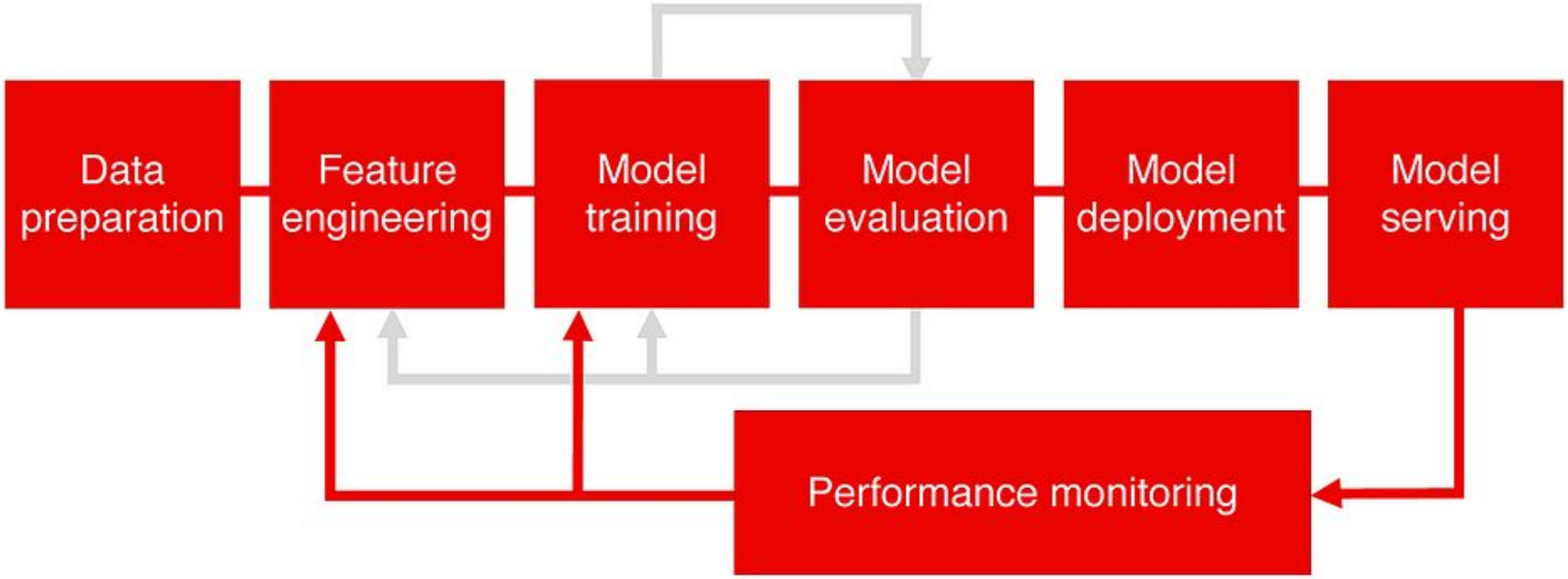
Data pipelines break.

User demographic changes. If a model receives wrong or unusual input, it will make an unreliable prediction.

We already know that they will break and degrade. To operate them successfully, we need a real-time view of their performance.

If we detect a quality drop, we can trigger retraining or step back into the research phase to issue a model remake.

MONITORING AND FEEDBACK



Most Common Applications of Data Science

Business Analytics

- Customer Analytics
- Market Segmentation
- Churn Prediction

Healthcare

- Disease Prediction
- Drug Discovery
- Healthcare Management

Social Media

- User Behavior Analysis
- Sentiment Analysis
- Content Personalization

Finance

- Risk Assessment
- Algorithmic Trading
- Financial Forecasting

E-commerce

- Recommendation Systems
- Supply Chain Optimization
- Price Optimization

Manufacturing

- Predictive Maintenance
- Quality Control
- Supply Chain Management

Entertainment

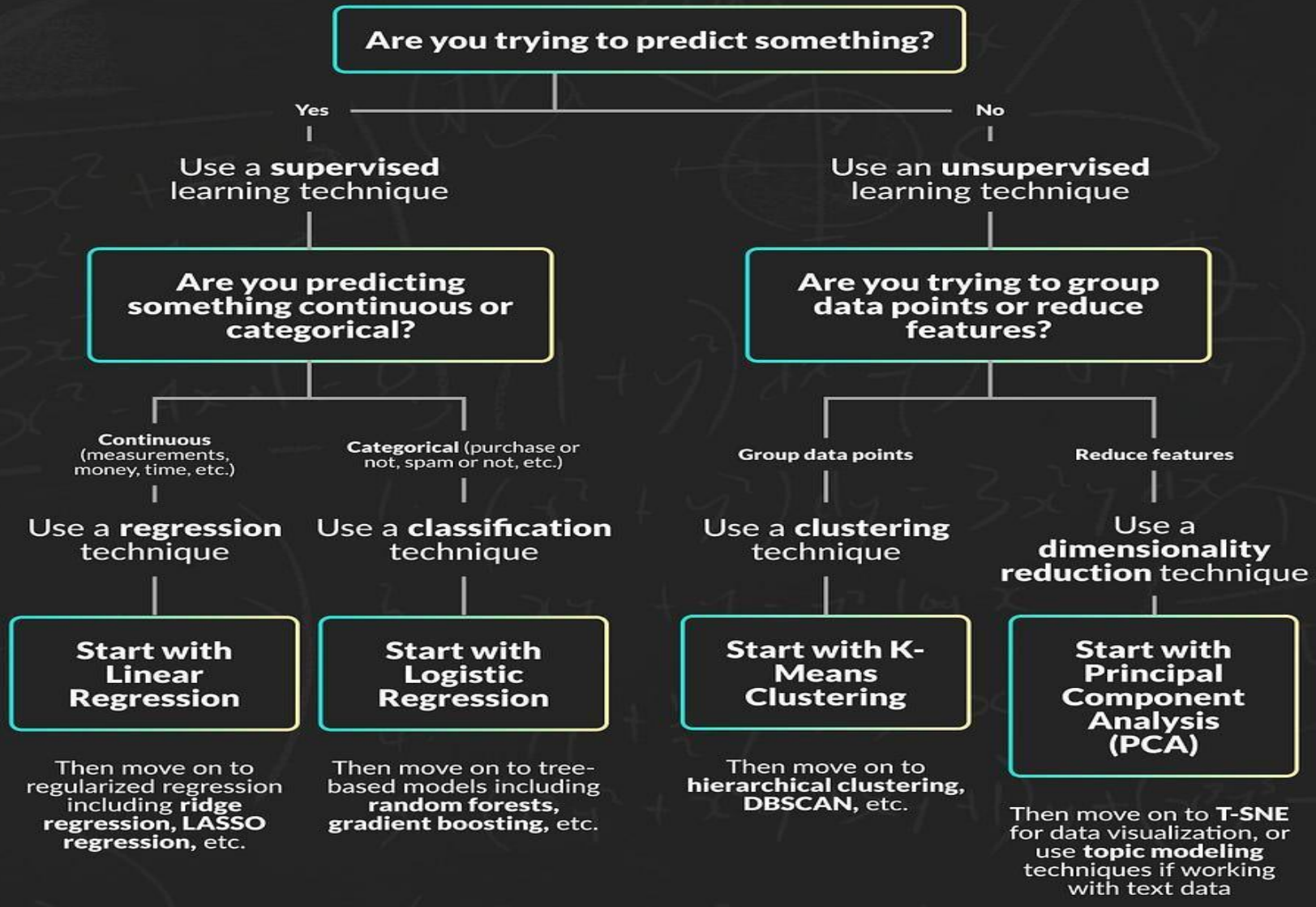
- Content Recommendation
- Box Office Prediction
- Gaming

Transportation

- Route Optimization
- Demand Forecasting
- Fleet Management

FLOWCHART

Which machine learning algorithm should I use for my data science project?



ASSIGNMENTS

DESIGN 10 USE CASES USING MACHINE LEARNING ALGORITHMS WHATEVER DISCUSSED IN OUR CURRICULUM