Performing Exploratory Data analysis

*PreviousNext
*Tutorial Playlist
*Table of Contents

What Is Exploratory Data Analysis?Steps Involved in Exploratory Data AnalysisMarket Analysis With Exploratory Data AnalysisConclusion
Before you start data analysis or run your data through a machine learning algorithm, you must clean your data and make sure it is in a suitable form. Further, it is essential to know any recurring patterns and significant correlations that might be present in your data. The process of getting to know your data in depth is called Exploratory Data Analysis.

Exploratory Data Analysis is an integral part of working with data. In this tutorial titled ‘ All the ins and outs of exploratory data analysis,’ you will explore how to perform exploratory data analysis on different data types.

Become a Data Scientist With Real-World Experience
Data Scientist Master’

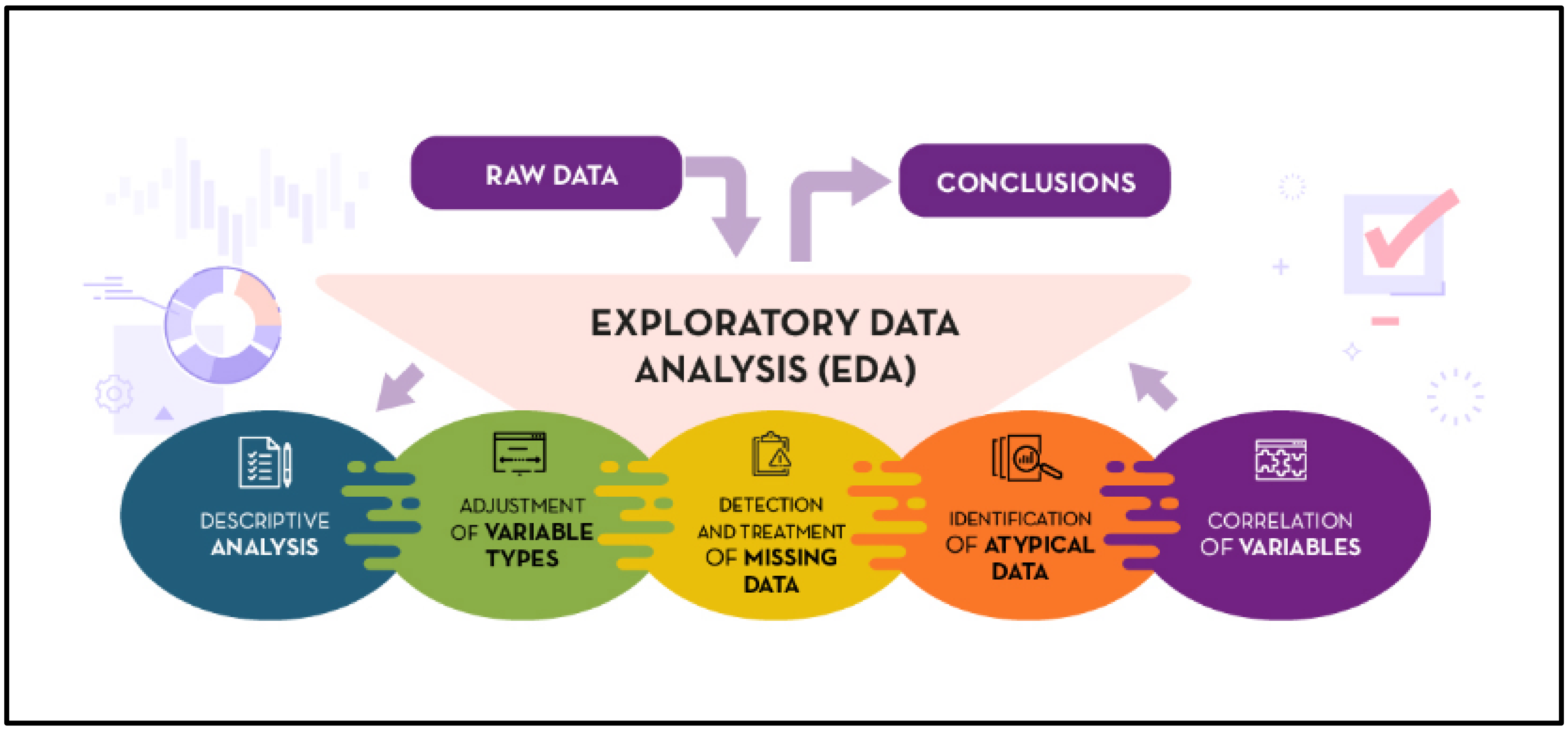
What Is Exploratory Data Analysis?
Exploratory Data Analysis is a data analytics process to understand the data in depth and learn the different data characteristics, often with visual means. This allows you to get a better feel of your data and find useful patterns in it.

Exploratory Data Analysis:

It is crucial to understand it in depth before you perform data analysis and run your data through an algorithm. You need to know the patterns in your data and determine which variables are important and which do not play a significant role in the output. Further, some variables may have correlations with other variables. You also need to recognize errors in your data.

All of this can be done with Exploratory Data Analysis. It helps you gather insights and make better sense of the data, and removes irregularities and unnecessary values from data.

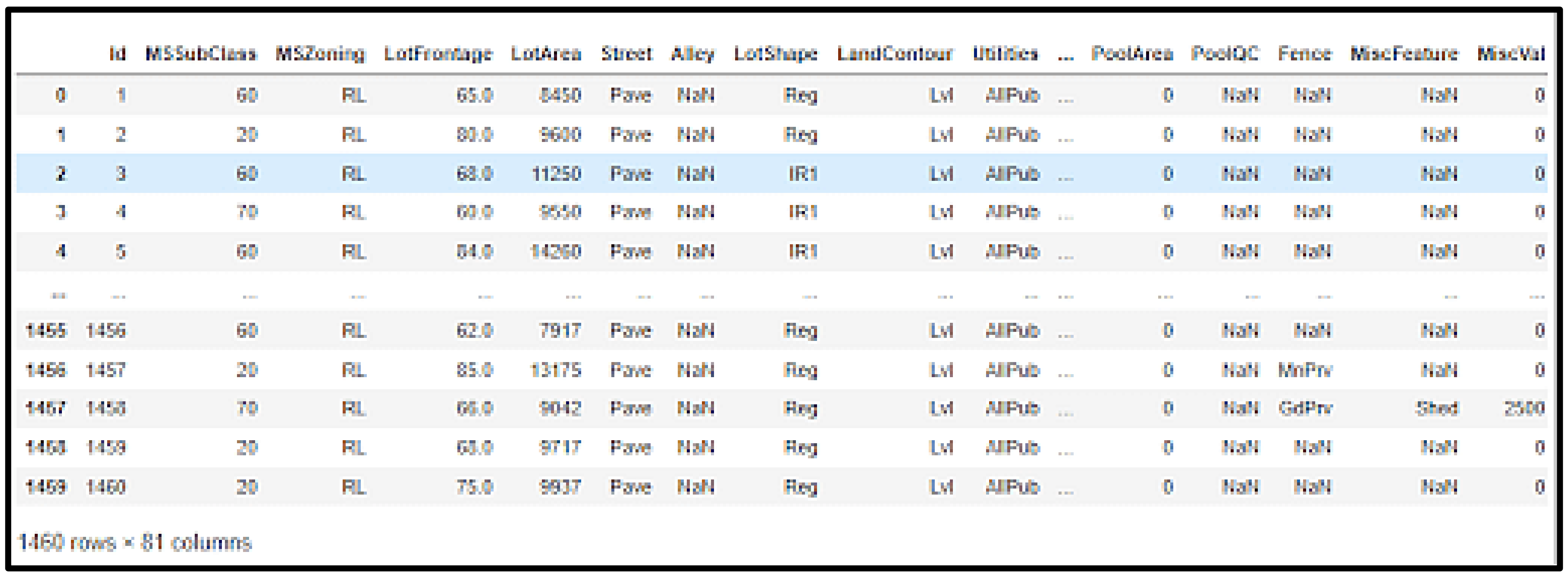
Helps you prepare your dataset for analysis.
Allows a machine learning model to predict our dataset better.
Gives you more accurate results.
It also helps us to choose a better machine learning model.


RAW DATA
CONCLUSIONS
EXPLORATORY DATA ANALYSIS (EDA)
DESCRIPTIVE ANALYSIS
ADJUSTMENT OF VARIABLE TYPES
DETECTION AND TREATMENT OF MISSING DATA
IDENTIFICATION OF ATYPICAL DATA
CORRELATION OF VARIABLES

DataCollection:

Data collection is an essential part of exploratory data analysis. It refers to the process of finding and loading data into our system. Good, reliable data can be found on various public sites or bought from private organizions. Some reliable sites for data collection are Kaggle, Github, Machine Learning Repository

The data depicted below represents the housing dataset that is available on Kaggle. It contains informationonhousesandthe pricethattheyweresold for.



	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1455	1456	60	RL	62.0	7917	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0
1456	1457	20	RL	85.0	13175	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	MnPrv	NaN	0
1457	1458	70	RL	66.0	9042	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	GdPrv	Shed	2500
1458	1459	20	RL	68.0	9717	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0
1459	1460	20	RL	75.0	9937	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0

1460 rows × 81 columns

Feature engineering

What is Feature Engineering?

Feature engineering is the process of transforming raw data into features that are suitable for machine learning models. In other words, it is the process of selecting, extracting, and transforming the most relevant features from the available data to build more accurate and efficient machine learning models.

The success of machine learning models heavily depends on the quality of the features used to train them. Feature engineering involves a set of techniques that enable us to create new features by combining or transforming the existing ones. These techniques help to highlight the most important patterns and relationships in the data, which in turn helps the machine learning model to learn from the data more effectively.

Types of Feature Creation:

Domain-Specific:
Creating new features based on domain knowledge, such as creating features based on business rules or industry standards.
Data-Driven:
Creating new features by observing patterns in the data, such as calculating aggregations or creating interaction features.
Synthetic:
Generating new features by combining existing features or synthesizing new data points.
Previous
Tutorial Playlist
Table of Contents

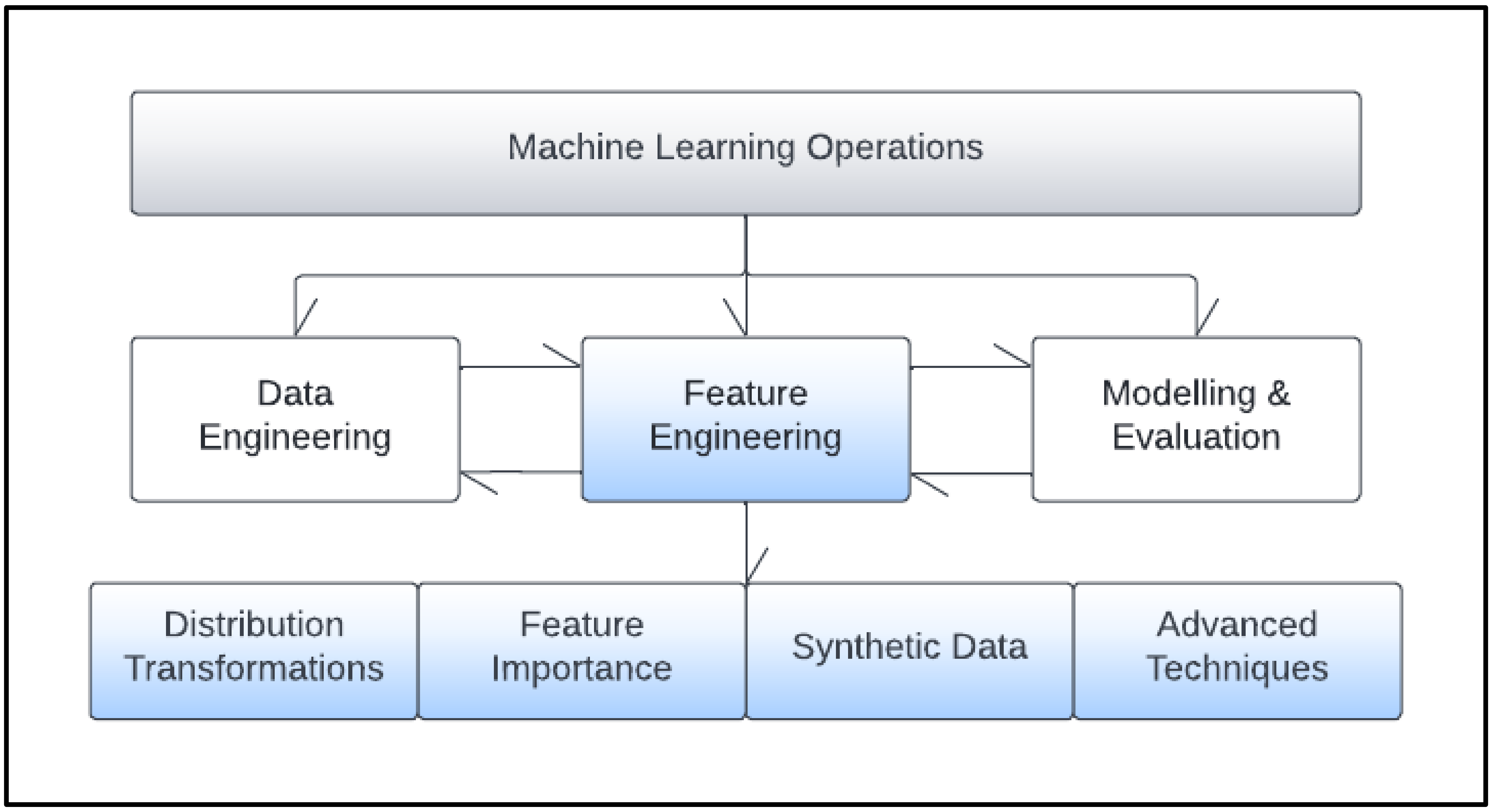
In machine learning, the prowess algorithms are effective not only in the quantity of data but also in the quality of the functions within that data. Feature engineering is pivotal in shaping raw information into meaningful attributes that empower machine learning models to extract precious insights and make accurate predictions. The essence of feature engineering, its importance, methodologies, techniques, and equipment contribute to its achievement.

How Does Feature Engineering Work?

Feature engineering is a multi-faceted method that entails creativity, domain understanding, and analytical skills. The steps included in feature engineering encompass the following

Feature Engineering
Previous
Tutorial Playlist
Table of Contents

In machine learning, the prowess algorithms are effective not only in the quantity of data but also in the quality of the functions within that data. Feature engineering is pivotal in shaping raw information into meaningful attributes that empower machine learning models to extract precious insights and make accurate predictions. The essence of feature engineering, its importance, methodologies, techniques, and equipment contribute to its achievement.



Predictive modeling

What Is Predictive Modeling?

In short, predictive modeling is a statistical technique using machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and existing data. It works by analyzing current and historical data and projecting what it learns on a model generated to forecast likely outcomes. Predictive modeling can be used to predict just about anything, from TV ratings and a customer’ s next purchase to credit risks and corporate earnings.

A predictive model is not fixed
it is validated or revised regularly to incorporate changes in the underlying data. In other words, it’ s not a one-and-done prediction. Predictive models make assumptions based on what has happened in the past and what is happening now. If incoming, new data shows changes in what is happening now, the impact on the likely future outcome must be recalculated, too. For example, a software company could model historical sales data against marketing expenditures across multiple regions to create a model for future revenue based on the impact of the marketing spend.

Most predictive models work fast and often complete their calculations in real time. That’ s why banks and retailers can, for example, calculate the risk of an online mortgage or credit card application and accept or decline the request almost instantly based on that prediction.

Types of Predictive Models:

Classification model:
Considered the simplest model, it categorizes data for simple and direct query response. An example use case would be to answer the question “ Is this a fraudulent transaction

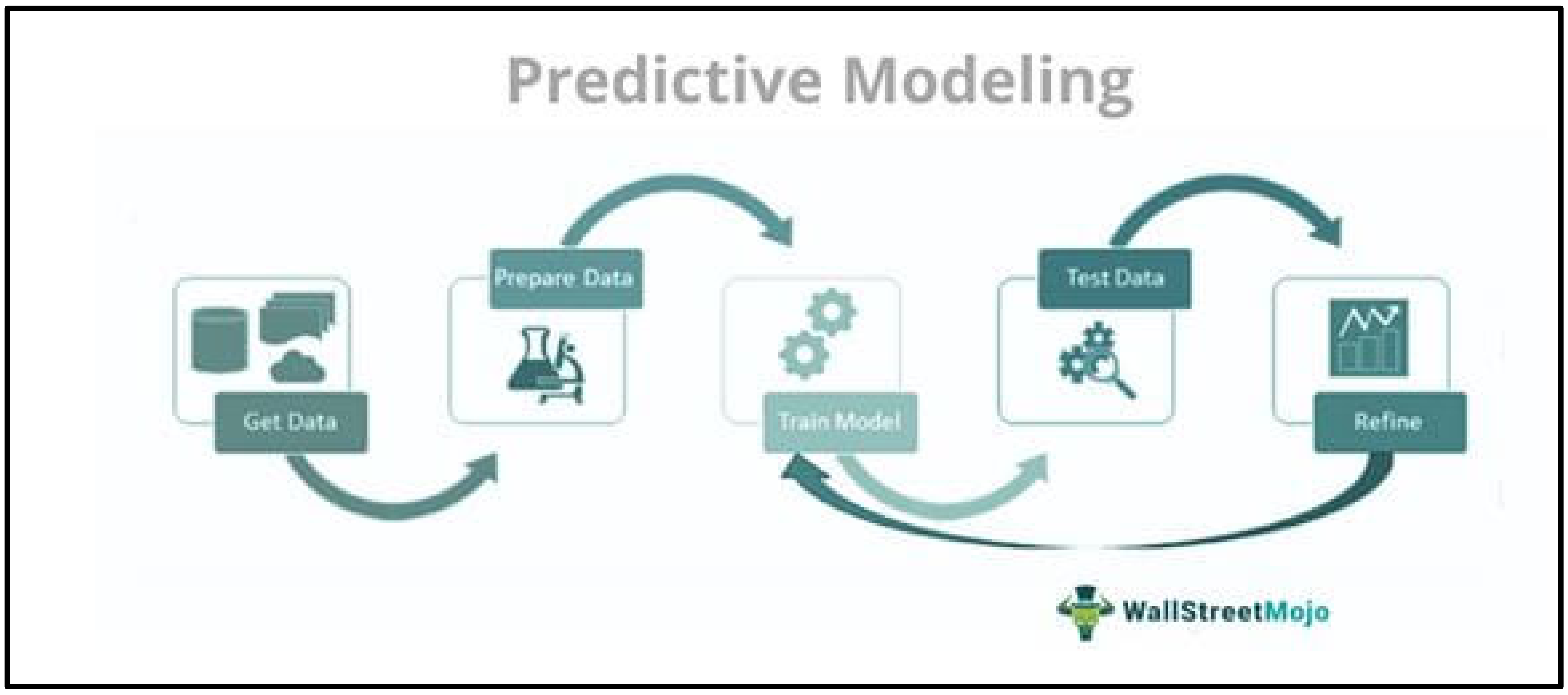
Clustering model:
This model nests data together by common attributes. It works by grouping things or people with shared characteristics or behaviors and plans strategies for each group at a larger scale. An example is in determining credit risk for a loan applicant based on what other people in the same or a similar situation did in the past.

Forecast model:
This is a very popular model, and it works on anything with a numerical value based on learning from historical data. For example, in answering how much lettuce a restaurant should order next week or how many calls a customer support agent should be able to handle per day or week, the system looks back to historical data.

Outliers model:
This model works by analyzing abnormal or outlying data points. For example, a bank might use an outlier model to identify fraud by asking whether a transaction is outside of the customer’ s normal buying habits or whether an expense in a given category is normal or not. For example, a $1,000 credit card charge for a washer and dryer in the cardholder’ s preferred big box store would not be alarming, but $1,000 spent on designer clothing in a location where the customer has never charged other items might be indicative of a breached account.

Time series model:

This model evaluates a sequence of data points based on time. For example, the number of stroke patients admitted to the hospital in the last four months is used to predict how many patients the hospital might expect to admit next week, next month or the rest of the year. A single metric measured and compared over time is thus more meaningful than a simple average.

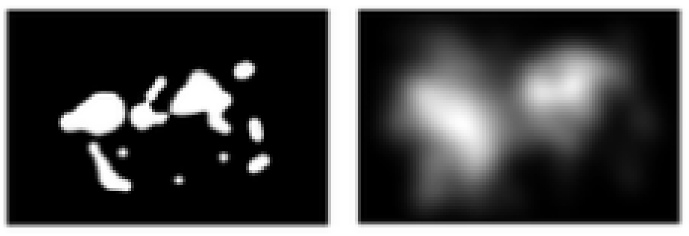


The Future of Predictive Modeling:

Predictive modeling, also known as predictive analytics, and machine learning are still young and developing technologies, meaning there is much more to come. As techniques, methods, tools and technologies improve, so will the benefits to businesses and societies.

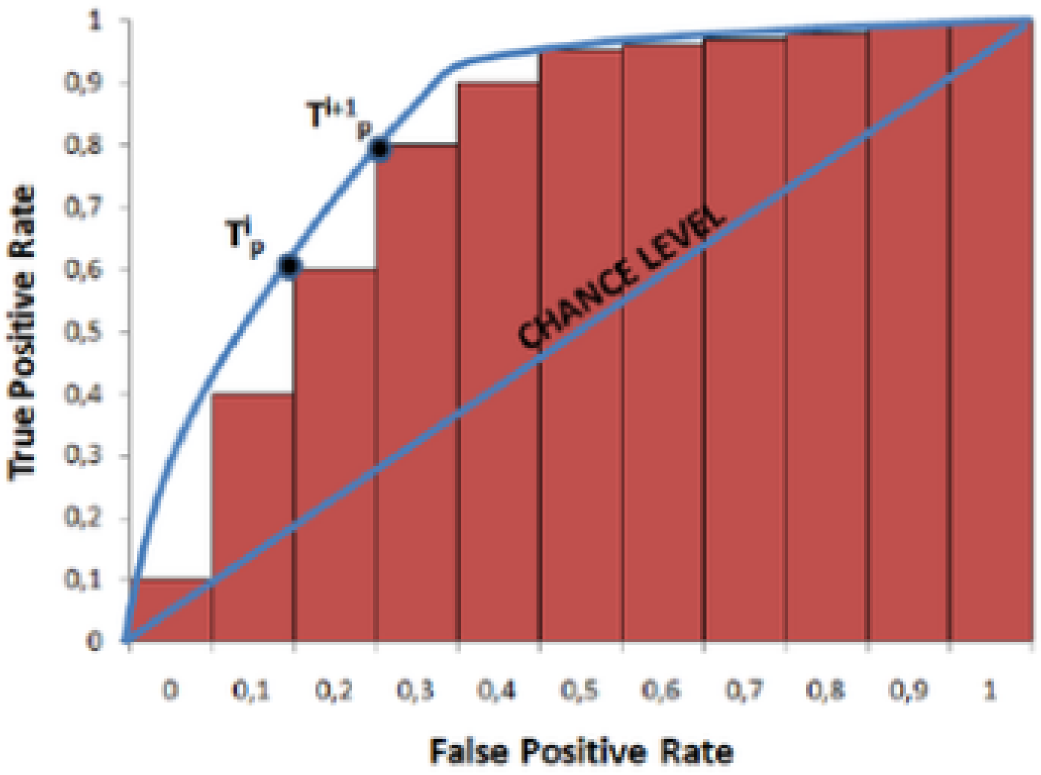
However, these are not technologies that businesses can afford to adopt later, after the tech reaches maturity and all the kinks are worked out.

Given a thresholded map SM_G^{BIN} and a predicted map SM_P:



```
                    PSEUDO CODE
Define a set of threshold $T = \{T_p^i\}_{i \in N}$
FOR ALL THRESHOLDS in T
 Binarization of the predicted map with the threshold $T_p^i$
  FOR ALL PIXELS in binarized maps
   IF $SM_G^{BIN} = 255$ //FIXATED
    //BIN(i) indicates the thresholdeding is done with $T_p^i$
    IF $SM_p^{BIN(i)} = 255$
        TP++
    ELSE
        FN++
    ENDIF
   ELSE // NON FIXATED
    IF $SM_p^{BIN(i)} = 255$
        FP++
    ELSE
        TN++
    ENDIF
   ENDIF
  END FOR
 TruePositveRate($T_p^i$)=TP/(TP+FN)
 FalsePositveRate($T_p^i$)=FP/(TP+FN)
END FOR
PLOT(TruePositveRate, FalsePositveRate) //for each $T_p^i$
```

(a) Pseudo code



(b) ROC curve

		actual value		
		p	n	total
prediction outcome	p'	True Positive	False Positive	P'
	n'	False Negative	True Negative	N'
	total	P	N	

(c) Confusion Matrix