Introduction to Machine Learning

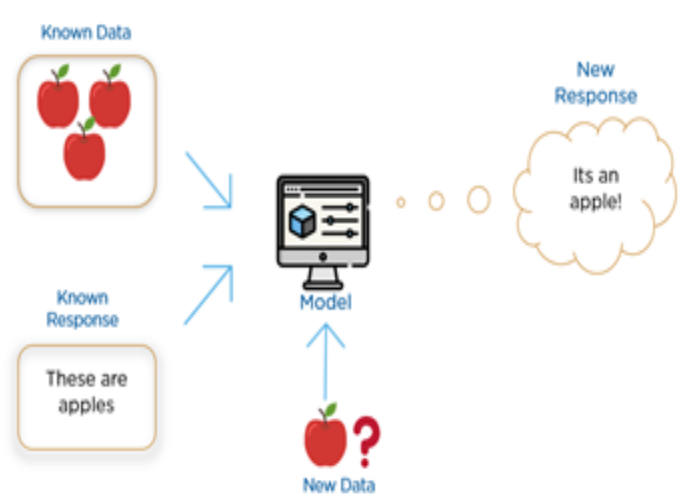
Topics

- Introduction to Machine Learning
- Applications of Machine Learning
- Simple Business Scenario
- Supervised Vs Unsupervised Learning
- Machine Learning Lifecycle
- Knowledge, Information and Data

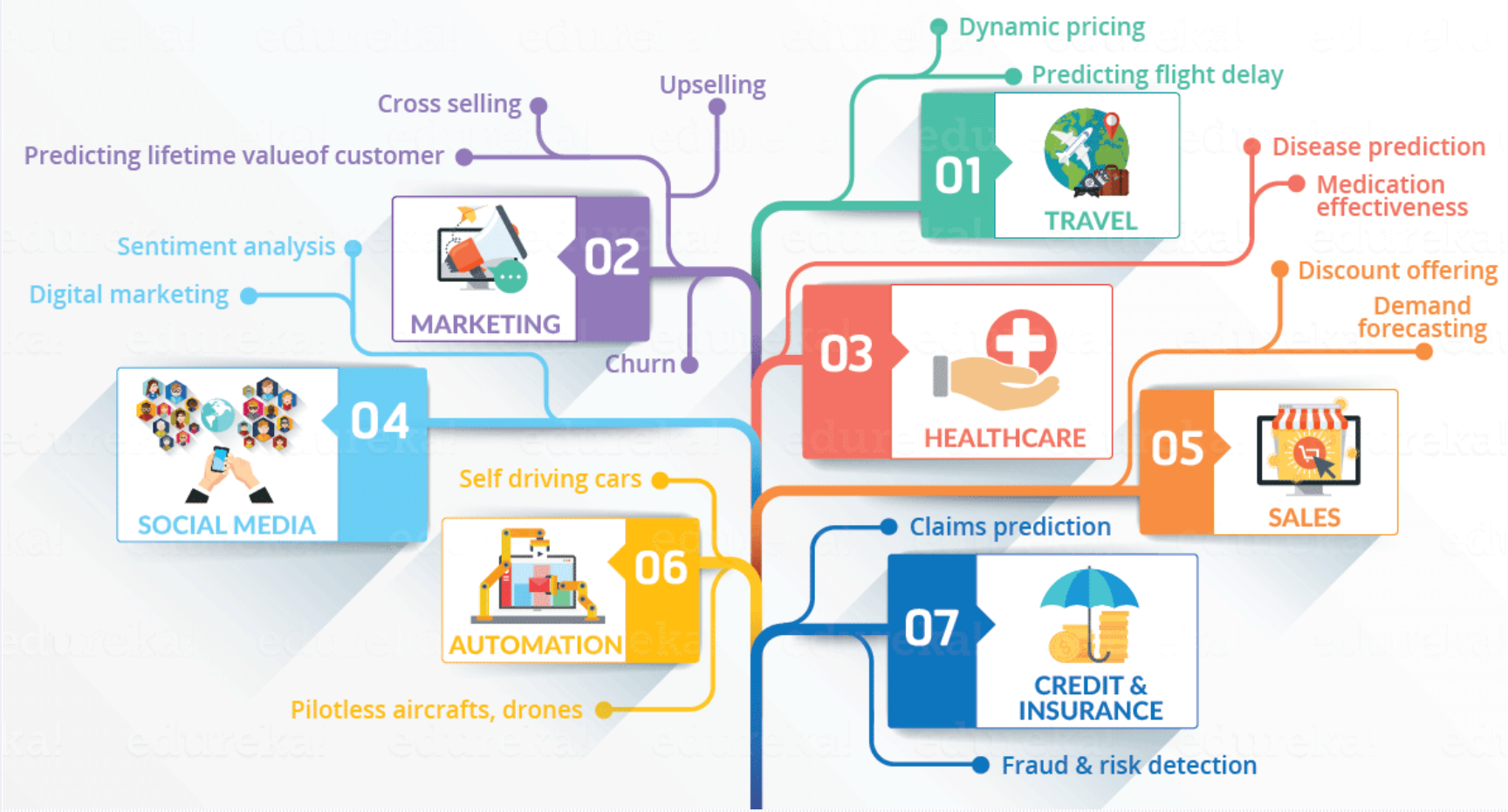
Introduction To Machine Learning

- **Why Machine Learning?**
  - Everyone wants to know (how to react to) the future.
  - Adapts and learns fast with changing scenario.
  - Acts fast with changing data.

- **What is Machine Learning?**
  - An algorithm that learns from data, identifies patterns in data and stores the learnings in the form of a Model.
  - Applies the Model to make prediction on new data.
  - It has the ability to quickly change, refresh, and enhance the Model with changing data and newer datasets.



Applications



Simple Business Scenario

Scenario:

Let us assume you are working in a Bank and the Chief Marketing Officer suggests that he wants to run a campaign to promote an Investment Product.

Based on business filters, you have an eligible contactable base of **1,000,000 customers**. The Cost of Targeting each customer being **$1**.

It is expected that **0.5%** incremental customers will purchase the Investment Product because of the campaign.

The Expected Revenue per customer who purchases the product is **$250**.

Campaign Return on Marketing Investment without Analytics Approach

- Target Customer Base : 1,000,000
- Cost of Targeting per customer : $1
- Cost of Campaign = 1,000,000 * 1 = $1,000,000
- Expected Incremental Conversion Rate : 0.5%
- Expected Incremental Conversions = 1,000,000 * 0.5% = 5,000 customers
- Expected Revenue per Convert : $250.
- Expected Incremental Revenue = 5,000 * 250 = $1,250,000
- Expected Profit = $1,250,000 - $1,000,000 = $250,000

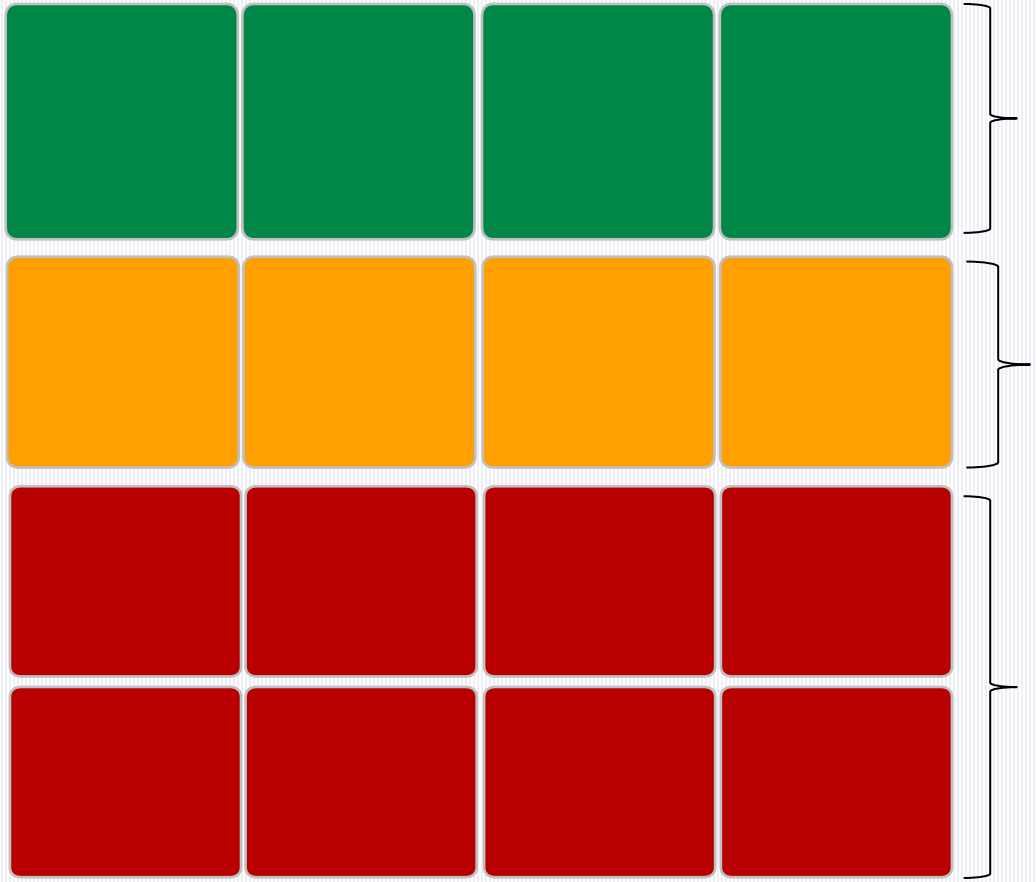
Campaign ROMI

$$\textbf{Return on Marketing Investment (ROMI)} = \frac{\text{Revenue} - \text{Cost}}{\text{Cost}}$$

$$= \frac{1.25 - 1}{1}$$

$$= 25\%$$

Analytics Based Approach



High Response Segment
25% of Base
With expected conversion rate of 1.3%

Medium Response Segment
25% of Base
With expected conversion rate of 0.4%

Low Response Segment
50% of Base
With expected conversion rate of 0.15%

Analytics Based ROMI

Segment	# Customer (A)	Exp. Conv. Rate (B)	# Conv's (C = A * B)	Cost of Targeting (D = A * 1)	Exp. Revenue (E = C * 250)	Profit (F = E – D)	ROMI G = F / D
High Response Segment	250,000	1.3%	3250	250,000	812,500	562,500	225%
Medium Response Segment	250,000	0.4%	1000	250,000	250,000	0	0%
Low Response Segment	500,000	0.15%	750	500,000	187,500	-312,500	-ve
Total	1,000,000	0.5%	5000	1,000,000	1,250,000	250,000	25%

Campaign ROMI

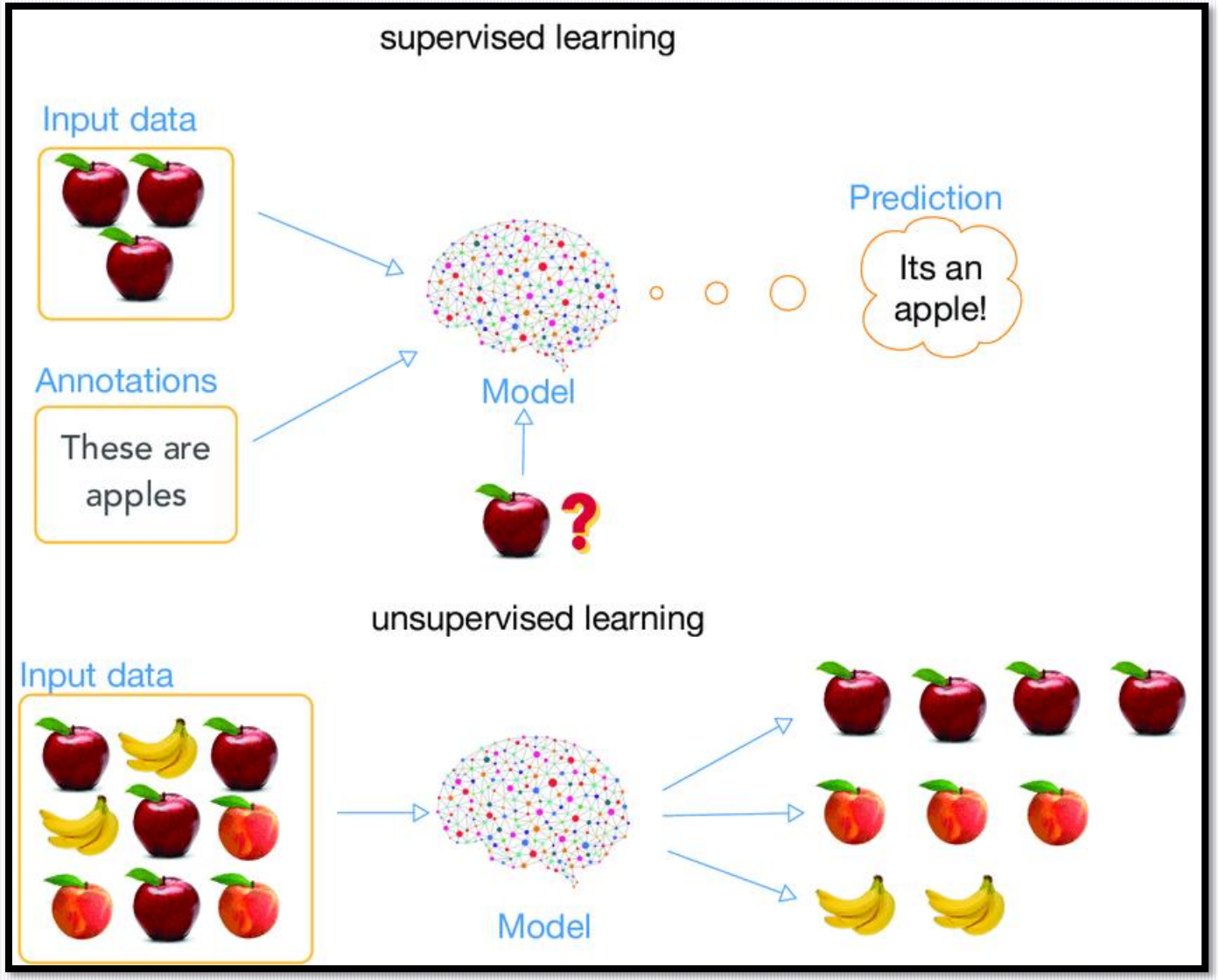
Your recommendation to the CMO:

- Target only the High Response Segment.
- Benefits of the strategy
  - A) It will reduce Marketing Cost by 75%
  - B) It will increase Profits by 200%
  - A) 9X increase in ROMI

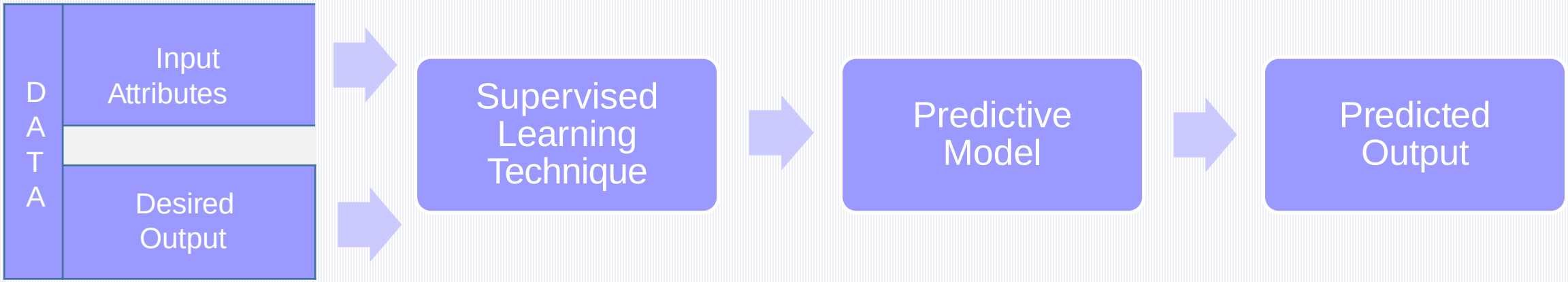
Supervised vs Unsupervised Learning

- Supervised Learning is the Machine Learning task of finding a function from a Labeled Data
  - Labeled Data is a dataset which has Independent Variable/s and a Dependent Variable.
- Unsupervised Learning is the Machine Learning task of exploring the data to derive some inferences / insights from the dataset
  - The “Target Variable” or the “Labeled Class” is not present in the Unsupervised Learning dataset.

Example:


supervised learning
Input data
Prediction
Its an apple!
Annotations
These are apples
Model
unsupervised learning
Input data
Model

Supervised Learning



- Supervised Learning Techniques
  - Classification
  - Regression

Application of Supervised Learning

- Assume you are working in a bank (say MyBank)
- The Chief Marketing Officer has assigned you the task of growing the Personal Loans Portfolio by **cross-selling** the loans to existing Customers.
- Data of past promotional campaigns and offers sent to the Customers, their behavioral data and those who took the loan are all made available with you.
- This is an example where **Supervised Learning** can be applied.

Marketing Modelling Dataset

- Sample Predictive Modelling Dataset

	Cust_ID	Target	Age	Gender	Balance	Occupation	No_OF_CR_TXNS	AGE_BKT	SCR	Holding_Period
1	C1	0	30	M	160378.60	SAL	2	26-30	826	9
2	C10	1	41	M	84370.59	SELF-EMP	14	41-45	843	9
3	C100	0	49	F	60849.26	PROF	49	46-50	328	26
4	C1000	0	49	M	10558.81	SAL	23	46-50	619	19
5	C10000	0	43	M	97100.48	SENP	3	41-45	397	8
6	C10001	0	30	M	160378.60	SAL	2	26-30	781	11
7	C10002	0	43	M	26275.55	PROF	23	41-45	354	12
8	C10003	0	53	M	33616.47	SAL	45	>50	239	5
9	C10004	0	45	M	1881.37	PROF	3	41-45	339	13

Some examples of Supervised Learning Applications

Industry/ Vertical	Supervised Learning Technique Applications	Labeled Class
HR	To predict whether a good employee is likely to resign or not	Resign / Not-Resign
Telecom	To classify customers who are likely to be Churners	Churn / Not-Churn
Retail / Ecommerce	To find potential customers from churned base who can be won back again	Win-back Yes / No
Banking	To build a model that will help assign the probability to a customer to take a product / service	Respond / Not-Respond
Insurance	To build a model to assess the likelihood of customer not renewing his / her policy	Lapse / Not-Lapse

Unsupervised Learning

- Unsupervised Learning Techniques
  - Dimension Reduction Techniques like Principal Component Analysis (PCA), Factor Analysis.
  - Clustering.
  - Association (Rules) Analysis.

Application of Unsupervised Learning

- Assume you are working in a Retail Company.
- You have 1,000,000 Loyalty Members.
- You have been asked to segment them based on their Buying Behavior Pattern.
- This is an example of **Unsupervised Learning** Application.

Clustering Modeling Dataset

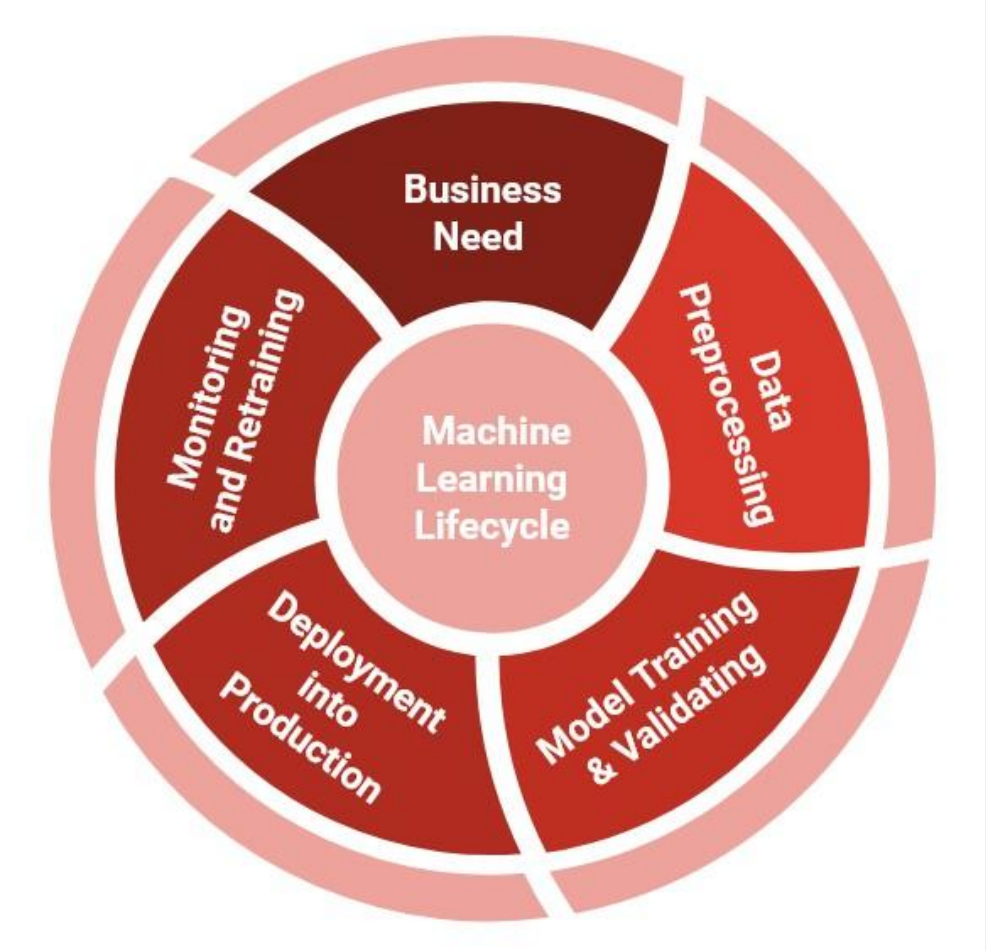
- Sample Clustering Modeling Dataset (**Note**: No target class).

	Cust_ID	Name	Avg_Mthly_Spend	No_Of_Visits	Apparel_Items	FnV_Items	Staples_Items
1	1	A	10000	2	1	1	0
2	2	B	7000	3	0	10	9
3	3	C	7000	7	1	3	4
4	4	D	6500	5	1	1	4
5	5	E	6000	6	0	12	3
6	6	F	4000	3	0	1	8
7	7	G	2500	5	0	11	2
8	8	H	2500	3	0	1	1
9	9	I	2000	2	0	2	2
10	10	J	1000	4	0	1	7

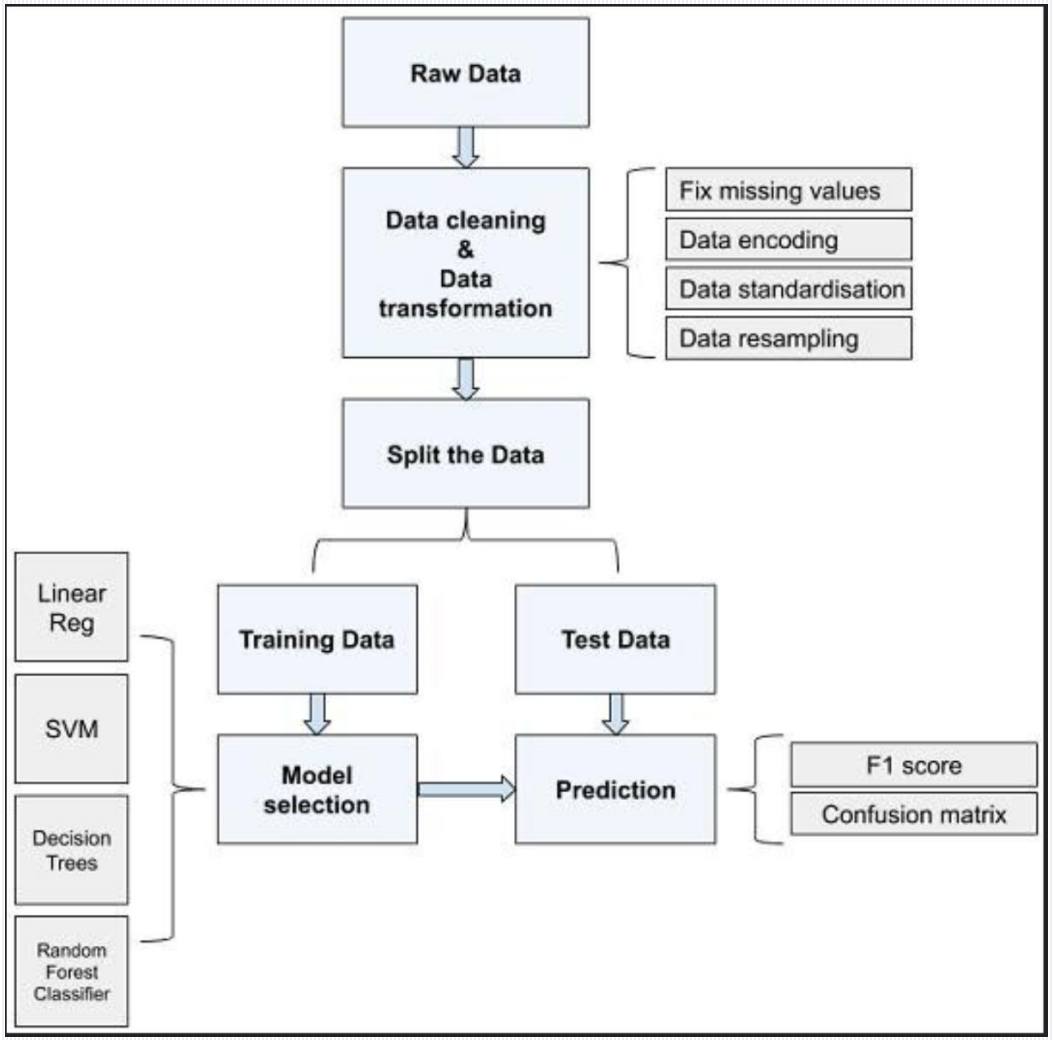
Some examples of Unsupervised Learning Applications

Industry/ Vertical	Unsupervised Learning Technique Applications
Healthcare	• Clustering patients based on symptoms or treatment outcomes. • Identifying genetic similarities for drug discovery. • Detecting anomalies in medical imaging (e.g., X-rays, MRIs).
Retail / Ecommerce	• Customer segmentation for personalized marketing. • Product recommendation based on purchase history (collaborative filtering). • Market basket analysis to identify frequent item sets (association rule mining).
Finance	• Fraud detection by identifying anomalies in transaction data. • Segmentation of customers based on spending behavior. • Risk profiling and grouping investors by portfolio preferences.
Education	• Grouping students based on learning behaviors or performance. • Identifying patterns in course engagement for curriculum improvement. • Clustering schools or institutions by demographic or performance indicators.

Machine Learning Lifecycle



Supervised Learning Flow Diagram



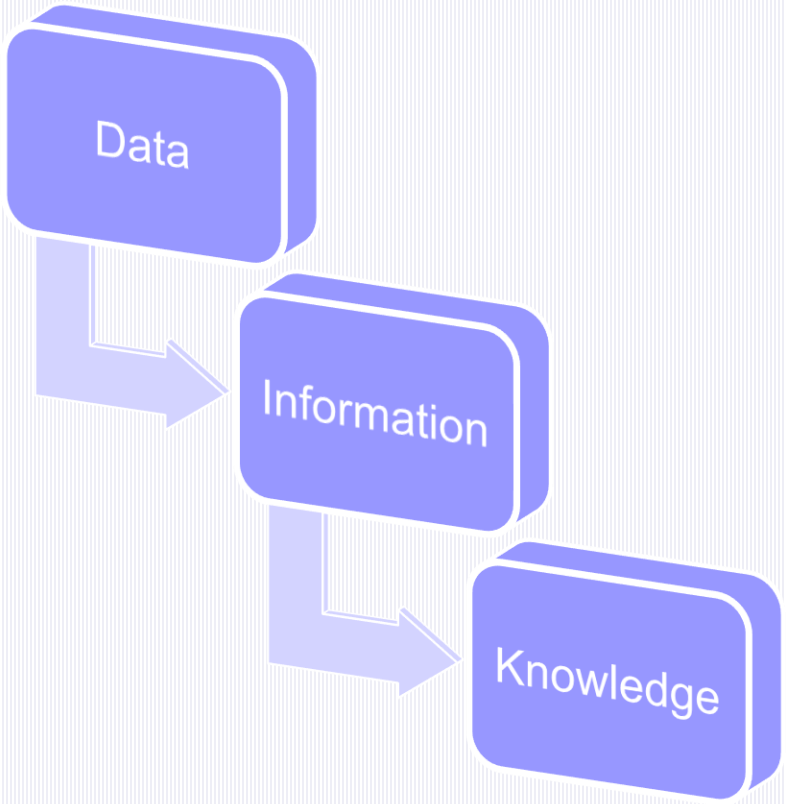
Data, Information and Knowledge

What is Data?

The concept of data is ‘raw’ data – a collection of text, numbers and symbols with no meaning. Data therefore must be processed, or provided with a context, before it can have meaning.

Example:

- 3, 6, 9, 12
- cat, dog, gerbil, rabbit, cockatoo
- 161.2, 175.3, 166.4, 164.7, 169.3



Cont.

What is Information?

Information is the result of processing data, usually by computer. This results in facts, which enables the processed data to be used in context and have meaning. Information is data that has meaning.

Data + Meaning = Information

Example:

Looking at the examples given for data:

- 3, 6, 9, 12
- cat, dog, gerbil, rabbit, cockatoo
- 161.2, 175.3, 166.4, 164.7, 169.3

Only when we assign a context or meaning does the data become information. It all becomes meaningful :

- 3, 6, 9 and 12 are the first four answers in the 3 x table.
- cat, dog, gerbil, rabbit, cockatoo is a list of household pets.
- 161.2, 175.3, 166.4, 164.7, 169.3 are the heights of 15-year-old students.

Cont.

What is Knowledge?

When someone memorizes information, this is often referred to as ‘rote-learning’ or ‘learning by heart’. We can say that they have acquired some knowledge. Another form of knowledge is produced as a result of understanding information that has been given to us and using that information to gain knowledge of how to solve problems.

Knowledge can therefore be:

- acquiring and remembering a set of facts, or
- the usage of information to solve problems.

Information + application / usage = Knowledge

Example:

Looking at the examples given for data:

- 3, 6, 9, 12
- cat, dog, gerbil, rabbit, cockatoo
- 161.2, 175.3, 166.4, 164.7, 169.3

Only when we assign a context or meaning does the data become information.

- 3, 6, 9 and 12 are the first four answers in the 3 x table.
- cat, dog, gerbil, rabbit, cockatoo is a list of household pets.
- 161.2, 175.3, 166.4, 164.7, 169.3 are the heights of the five tallest 15-year-old students in a class.

If we now apply this information to gain further knowledge we could say that:

- 4, 8, 12 and 16 are the first four answers in the 4 x table (because the 3 x table starts at three and goes up in threes the 4 x table must start at four and go up in fours)
- The tallest student is 175.3cm.
- A lion is not a household pet as it is not in the list and it lives in the wild.

Different Types of Data in Statistics

Categorical Data

Categorical data represents characteristics. Therefore, it can represent things like a person's gender, language etc. Categorical data can also take on numerical values (Example: 1 for female and 0 for male). Note that the numbers in this context don't have mathematical meaning.

- **Nominal Data**

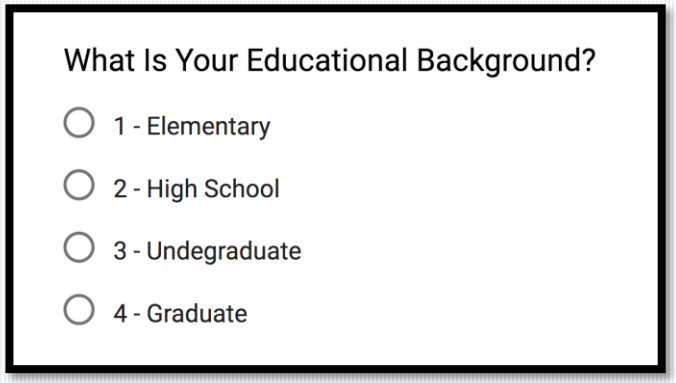
Nominal values represent discrete units and are used to label variables, that have no quantitative value. Just think of them as "labels". Note that nominal data that has no order. Therefore, if you change the order of its values, the meaning will not change.

Example:

- **Ordinal Data**

Ordinal values represent discrete and ordered units. It is therefore nearly the same as nominal data, except that it's ordering matters.

Example:



Cont.

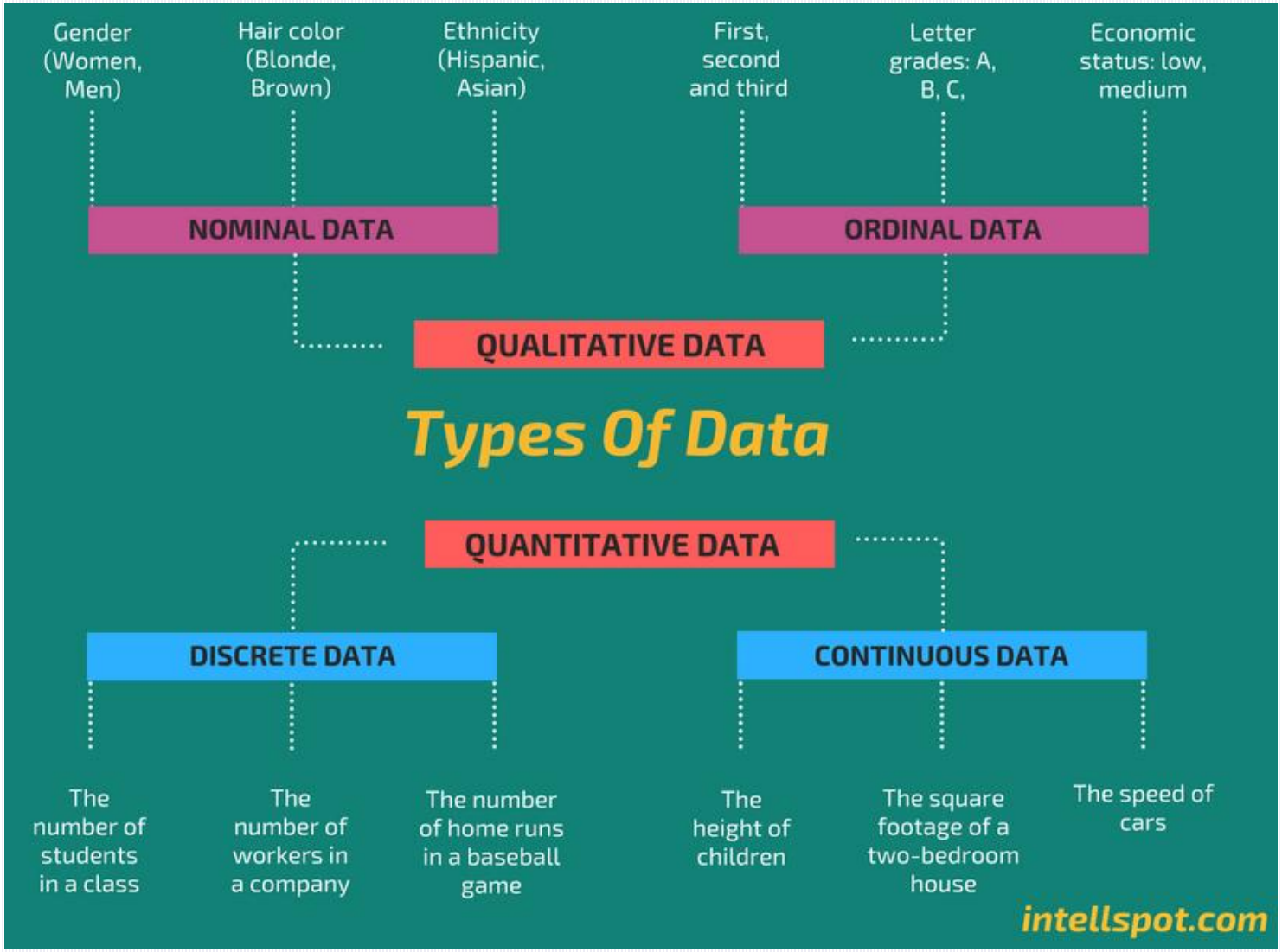
Numerical Data

1. Discrete Data

Discrete data if its values are distinct and separate. In other words: We speak of discrete data if the data can only take on certain values. This type of data **can't be measured but it can be counted**. It basically represents information that can be categorized into a classification. An example is the number of heads in 100 coin flips.

2. Continuous Data

Continuous Data represents measurements and therefore their values can't be counted but they can be measured. An example would be the height of a person, which you can describe by using intervals on the real number line.


Gender (Women, Men)
Hair color (Blonde, Brown)
Ethnicity (Hispanic, Asian)
First, second and third
Letter grades: A, B, C,
Economic status: low, medium
NOMINAL DATA
ORDINAL DATA
QUALITATIVE DATA
Types Of Data
QUANTITATIVE DATA
DISCRETE DATA
CONTINUOUS DATA
The number of students in a class
The number of workers in a company
The number of home runs in a baseball game
The height of children
The square footage of a two-bedroom house
The speed of cars
intellspot.com

THANK YOU