MACHINE LEARNING

- Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

- Machine learning is a growing technology which enables computers to learn automatically from past data.

- Machine learning uses various algorithms for building mathematical models and making predictions using historical data or information.

- Used for various tasks such as image recognition, speech recognition, email filtering, recommender system, and many more.

Artificial Intelligence

Machine Learning

Data Science

Deep Learning
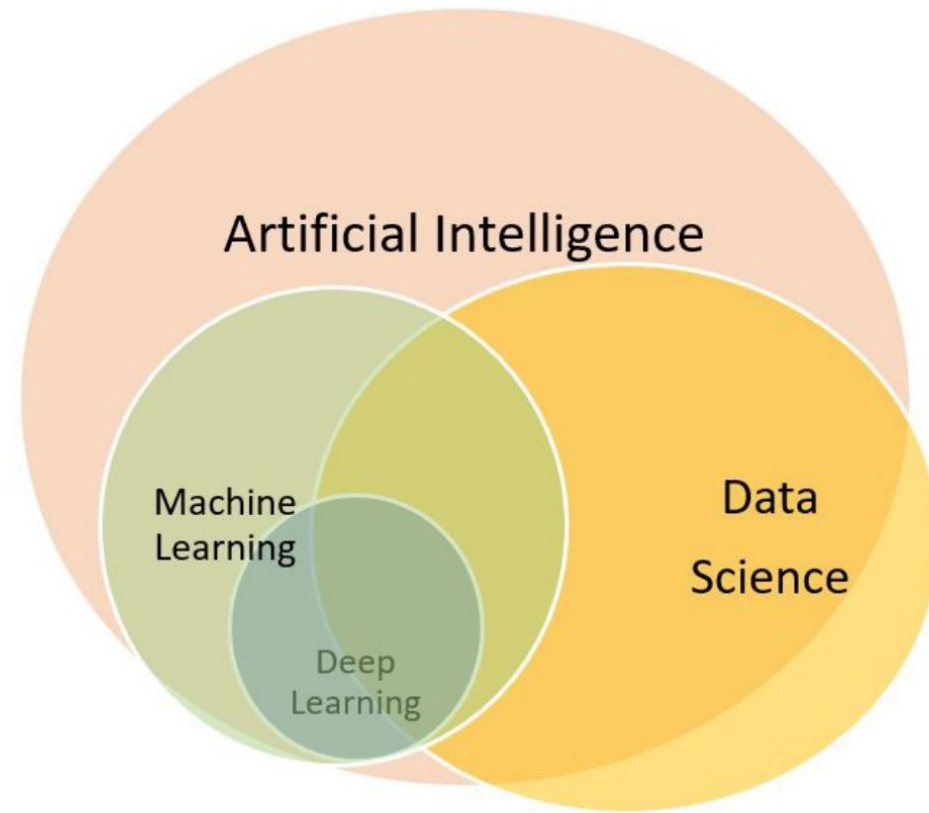
How a human being learn

This is a shirt we used to wear.

Color: green
Size : Large
Type : Formal

Input (X) → **Program** → Output (y)
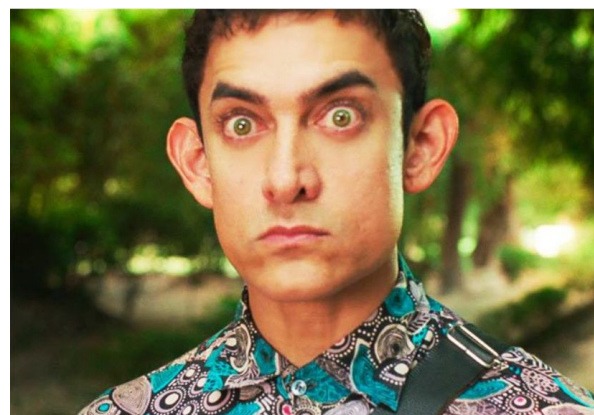
Input (X)
Output (y) → **Machine learning** → Representative Model of the program

Traditional program and machine learning.

Table 2.1 Dataset

Borrower ID	Credit Score	Interest Rate (%)
01	500	7.31
02	600	6.70
03	700	5.95
04	700	6.40
05	800	5.40
06	800	5.70
07	750	5.90
08	550	7.00
09	650	6.50
10	825	5.70

Table 2.2 New Data With Unknown Interest Rate

Borrower ID	Credit Score	Interest Rate
11	625	?

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0
8	99	84	0	0	35.4	0.388	50	0
7	196	90	0	0	39.8	0.451	41	1
9	119	80	35	0	29	0.263	29	1
11	143	94	33	146	36.6	0.254	51	1
10	125	70	26	115	31.1	0.205	41	1
7	147	76	0	0	39.4	0.257	43	1

- A dataset is a collection of data with a defined structure.

This structure is also sometimes referred to as a "data frame".

● A data point (record, object) is a single instance in the dataset. Each instance contains the same structure as the dataset.

● An attribute (feature, input, dimension, variable, or predictor) is a single property of the dataset. Attributes can be numeric, categorical, date-time, text, or Boolean data types.

- A label (class label, output, prediction, target, or response) is the special attribute to be predicted based on all the input attributes. In Table interest rate is the output variable.
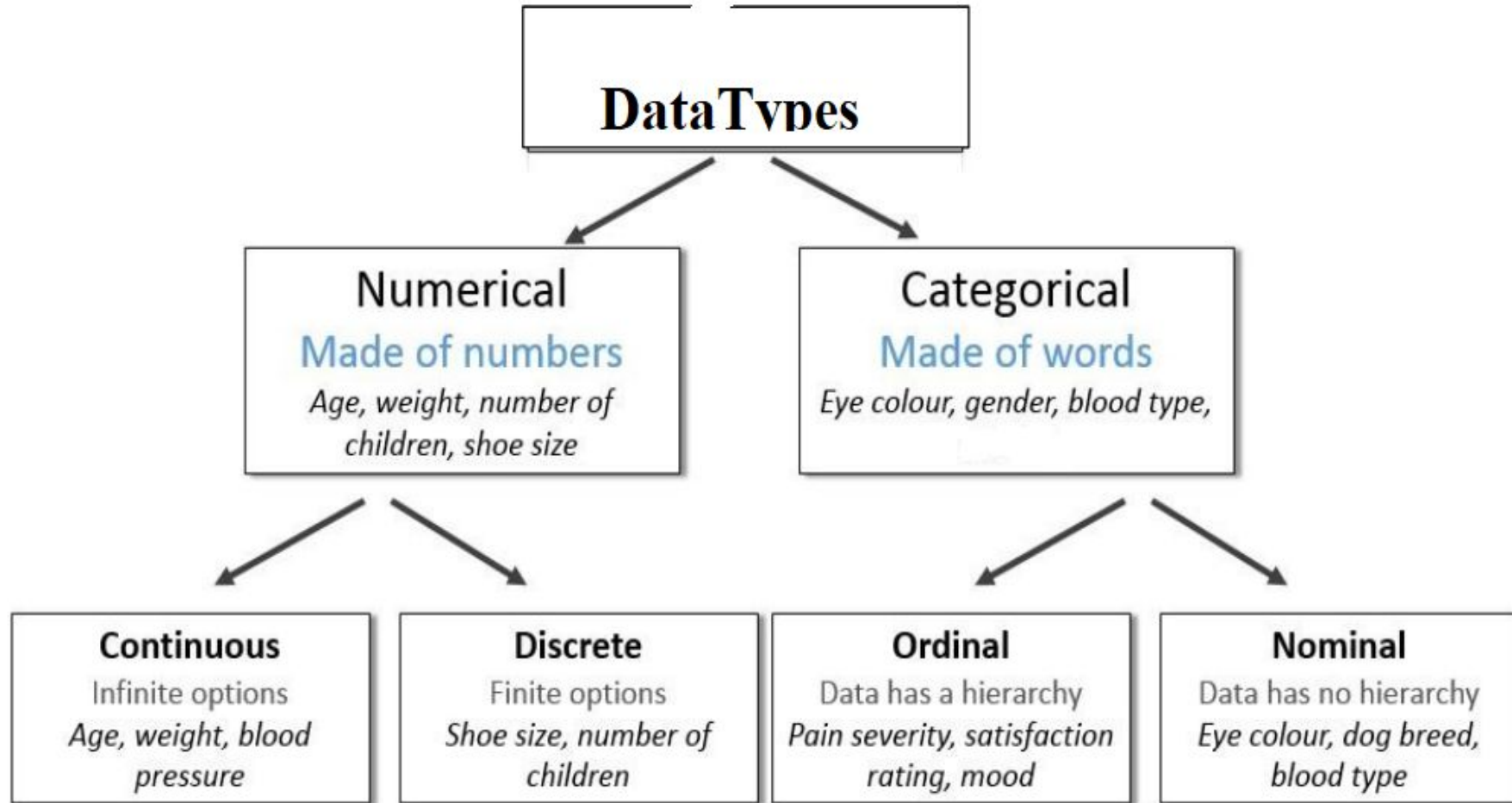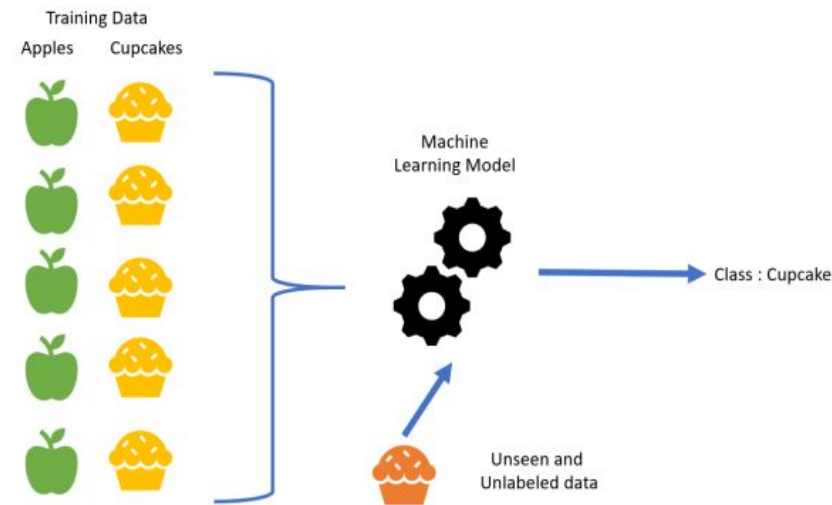
2 types of machine learning

Supervised Models

- Supervised techniques predict the value of the output variables based on a set of input variables.
- To do this, a model is developed from a training dataset where the values of input and output are previously known.
- The model generalizes the relationship between the input and output variables and uses it to predict for a dataset where only input variables are known.
- The output variable that is being predicted is also called a class label or target variable
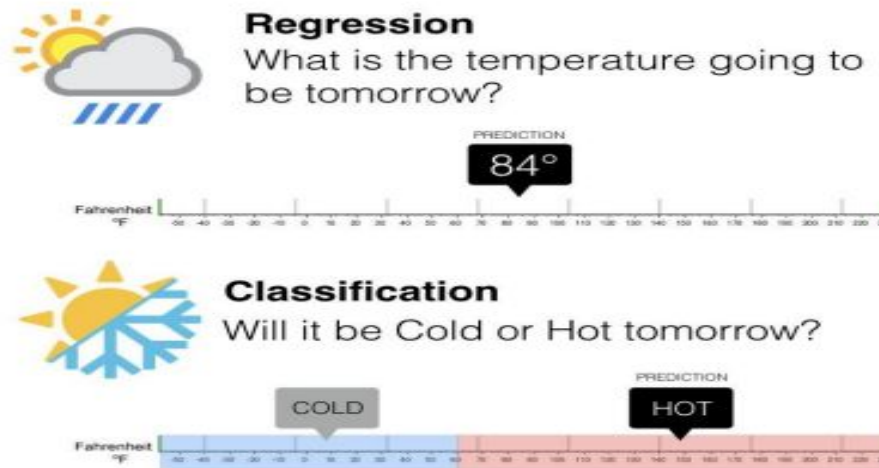
Input : Labeled Data

X (features)	Y (labels)
$x_{11}, x_{12}, x_{13}, \ldots\ldots\ldots\ldots\ldots x_{1n}$	y_1
.	.
.	.
.	.
$x_{k1}, xk_2, x_{k3}, \ldots\ldots\ldots\ldots\ldots x_{kn}$	y_k

Training Data

Apples   Cupcakes

Machine
Learning Model

Class : Cupcake

Unseen and
Unlabeled data

Classification and *regression* techniques predict a target variable based on input variables.

The prediction is based on a generalized model built from a previously known dataset. In regression tasks, the output variable is <u>numeric</u>. Classification tasks predict output variables, which are <u>categorical</u>.
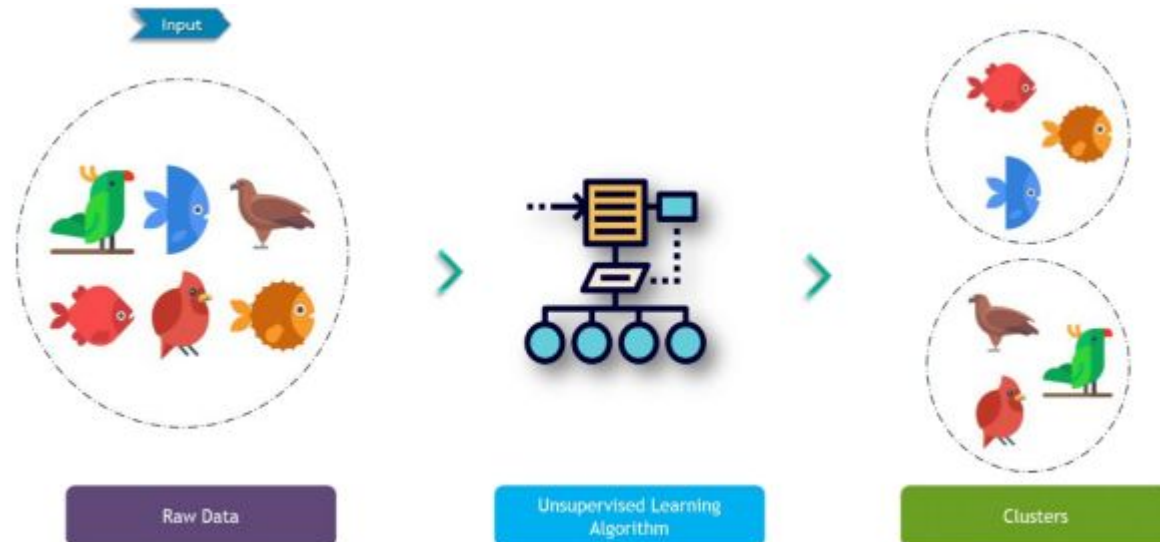
Unsupervised Models

- Unsupervised or undirected data science uncovers hidden patterns in unlabeled data.

- In unsupervised data science, there are no output variables to predict.

- The objective of this class of data science techniques, is to find patterns in data based on the relationship between data points themselves

Input : Unlabeled Data

X (features)
$x_{11}, x_{12}, x_{13}, \ldots\ldots\ldots\ldots\ldots\ldots x_{1n}$
.
.
.
$x_{k1}, xk_2, x_{k3}, \ldots\ldots\ldots\ldots\ldots\ldots x_{kn}$

Input

Raw Data

Unsupervised Learning Algorithm

Clusters

Clustering is the process of identifying the natural groupings in a dataset.

Raw Data   Algorithm   Output