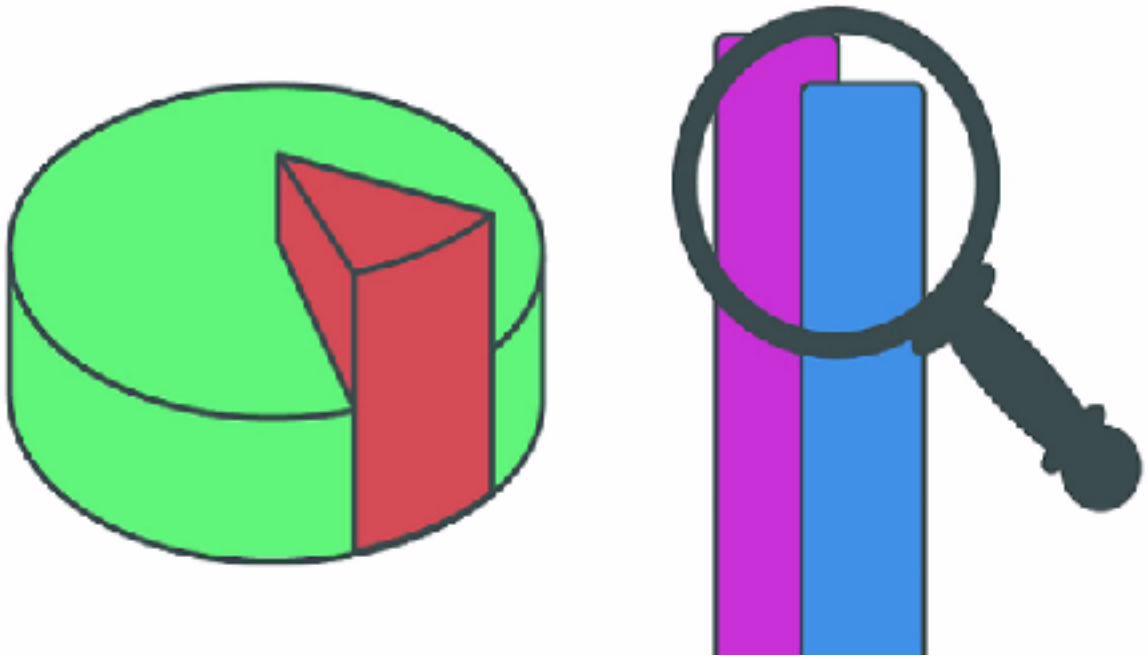
AN INTRODUCTION TO

MACHINE LEARNING

STATISTICS



- Aggregation
- Variation and SD
- Percentiles and movements
- Co-variance and Co-relation
- Linear Algebra
- Vectors and Matrices
- Probability Theory
- Data Distribution
- Calculus

ADD 'EM UP!

$$\sigma^2 = \frac{1}{N}\Sigma(X - \mu)^2$$

```
observations = [23, 40, 6, 74, 38, 1, 70]
mean = 36
difference_from_the_mean = [13, 4, 30, 38, 2, 35, 34]
square_of_the_differences = [169, 16, 900, 1444, 4, 1225, 1156]
variance = (169+16+900+1444+4+1225+1156)/7 = 4914/7 = 702
```

Variance and Standard Deviation are essentially a measure of the spread of the data in the data set.

$$\sigma^2 = 702$$

$$hence, \sigma = \sqrt{variance}$$

$$\sigma = 26.49$$

“how many sigmas away from the mean is this?”

A score percentile represents the percentage of scores
that are equal or below a certain score within a given sample.

Example: The 75th percentile SAT score for incoming freshmen is 1400.

75% of students
Scored 1400 or below

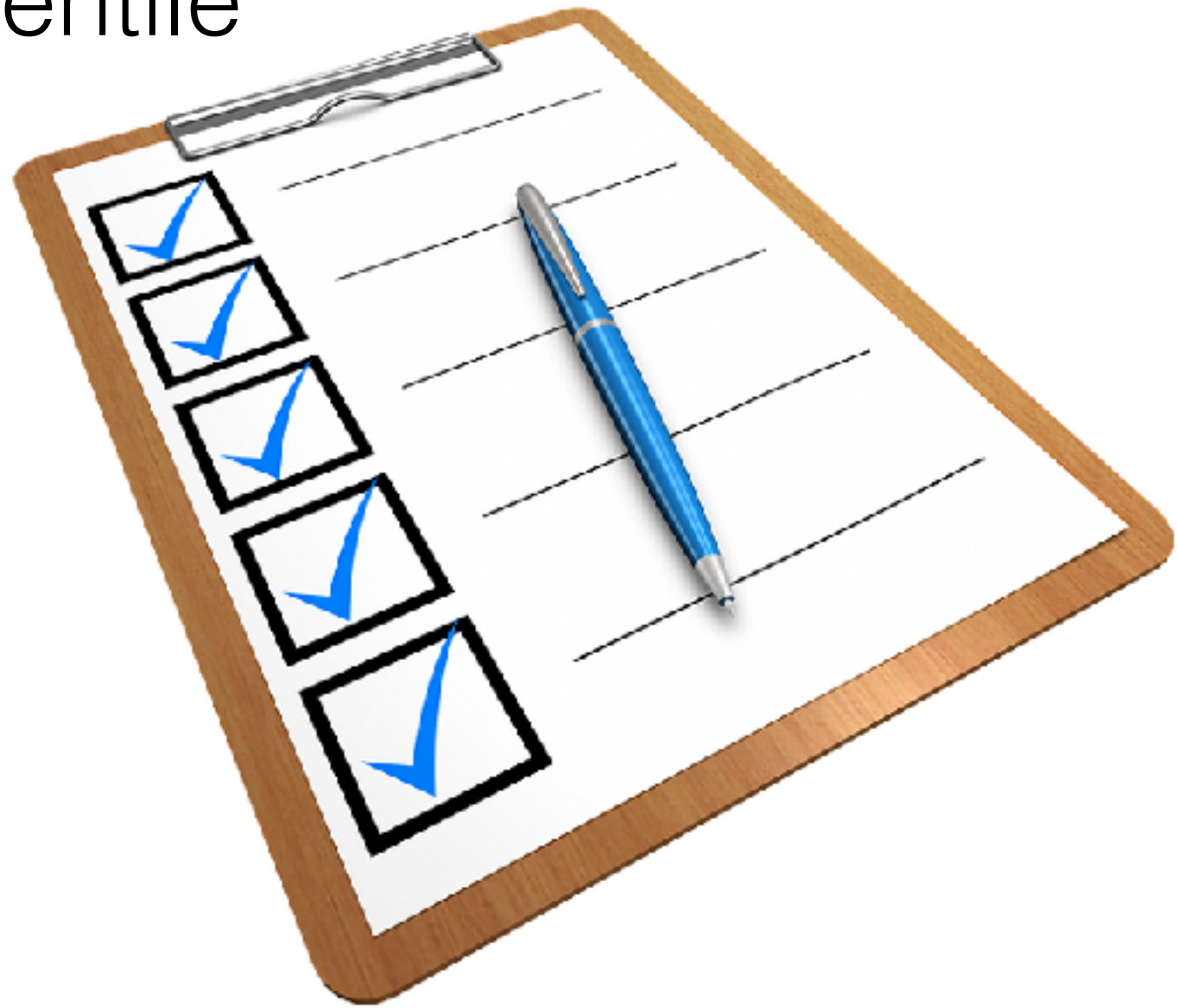
75th percentile
(25% of students)
Scored above 1400

How to Calculate

- Step 1: Order all the values in the data set from smallest to largest.
- Step 2: Multiply *k* percent by the total number of values, *n*. This number is called the index.
- Step 3: If the index obtained in Step 2 is not a whole number, round it up to the nearest whole number and go to Step 4a. If the index obtained in Step 2 is a whole number, go to Step 4b.
- Step 4a.Count the values in your data set from left to right (from the smallest to the largest value) until you reach the number indicated by Step 3. The corresponding value in your data set is the k^{th} percentile.
- Step 4b.Count the values in your data set from left to right until you reach the number indicated by Step 2. The k^{th} percentile is the average of that corresponding value in your data set and the value that directly follows it.

Find the 90th percentile

43, 54, 56, 61, 62,
66, 68, 69, 69, 70,
71, 72, 77, 78, 79,
85, 87, 88, 89, 93,
95, 96, 98, 99, 99



Moments

- **Moments** try to measure the shape of the probability distribution function.
- The zeroth moment is the total probability of the distribution which is 1.
- The first moment is the mean.
- The second moment is the variance.
- The third moment is the skew which measures how lopsided the distribution is.
- The fourth moment is kurtosis which is the measure of how sharp is the peak of the graph.

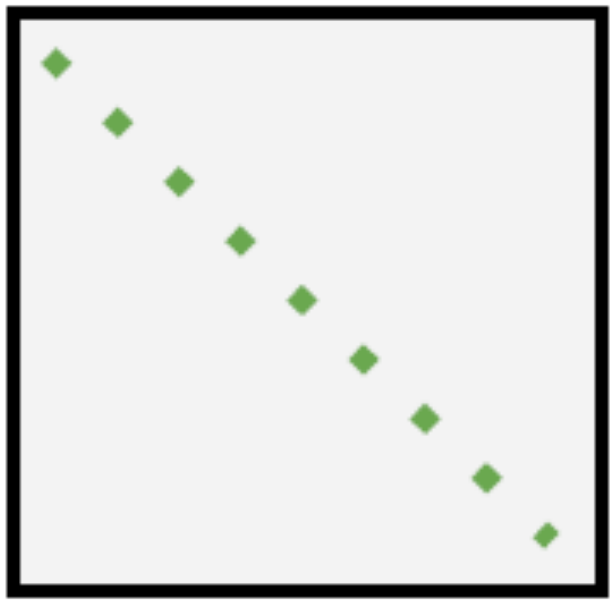
Covariance

Covariance measures how two variables vary in tandem to their means. The formula to calculate covariance is shown below.

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - E(X))(y_i - E(Y))$$

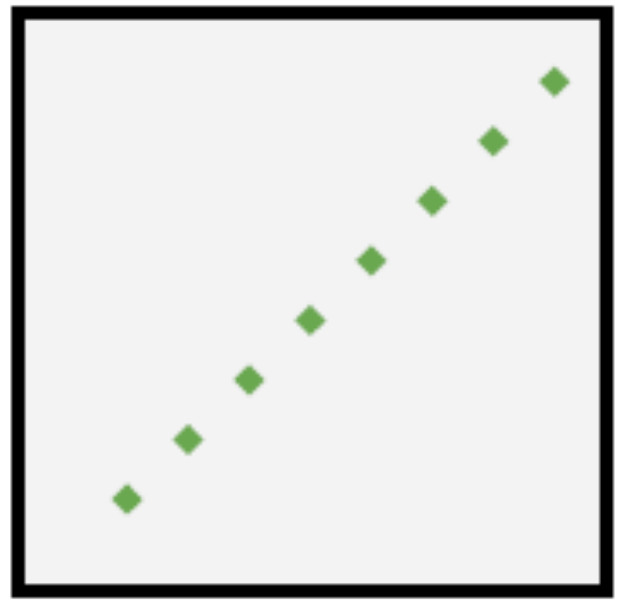
where x and y are the individual values of X and Y ranging from i = 1,2, .., n where the probability that each value may occur is equal and is equal to (1/n). E(x) and E(y) are the means of X and Y.

COVARIANCE



Large Negative Covariance

Nearly Zero Covariance

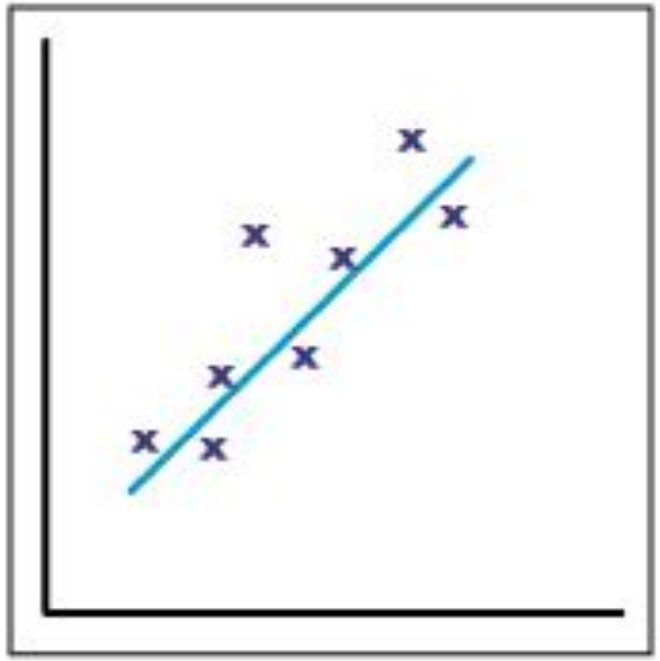
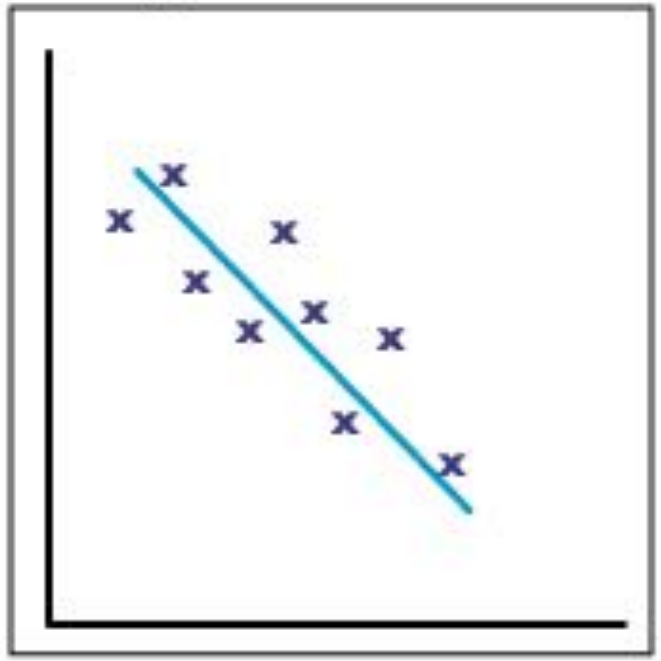
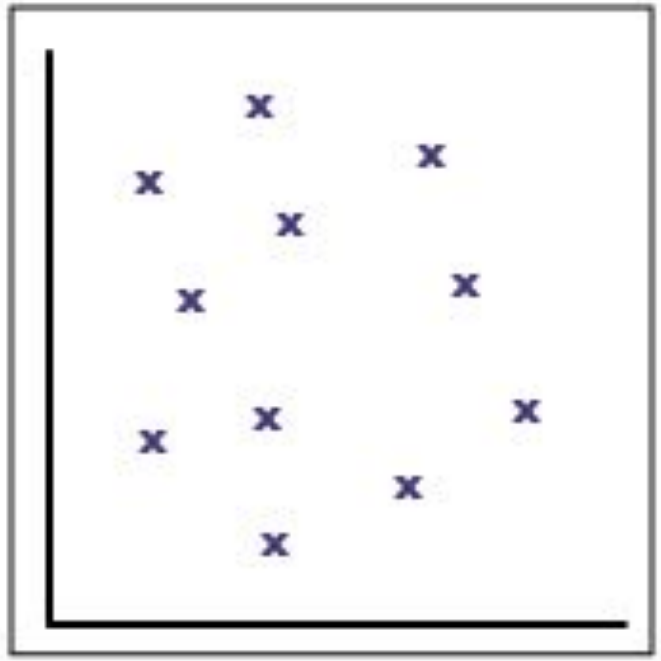


Large Positive Covariance

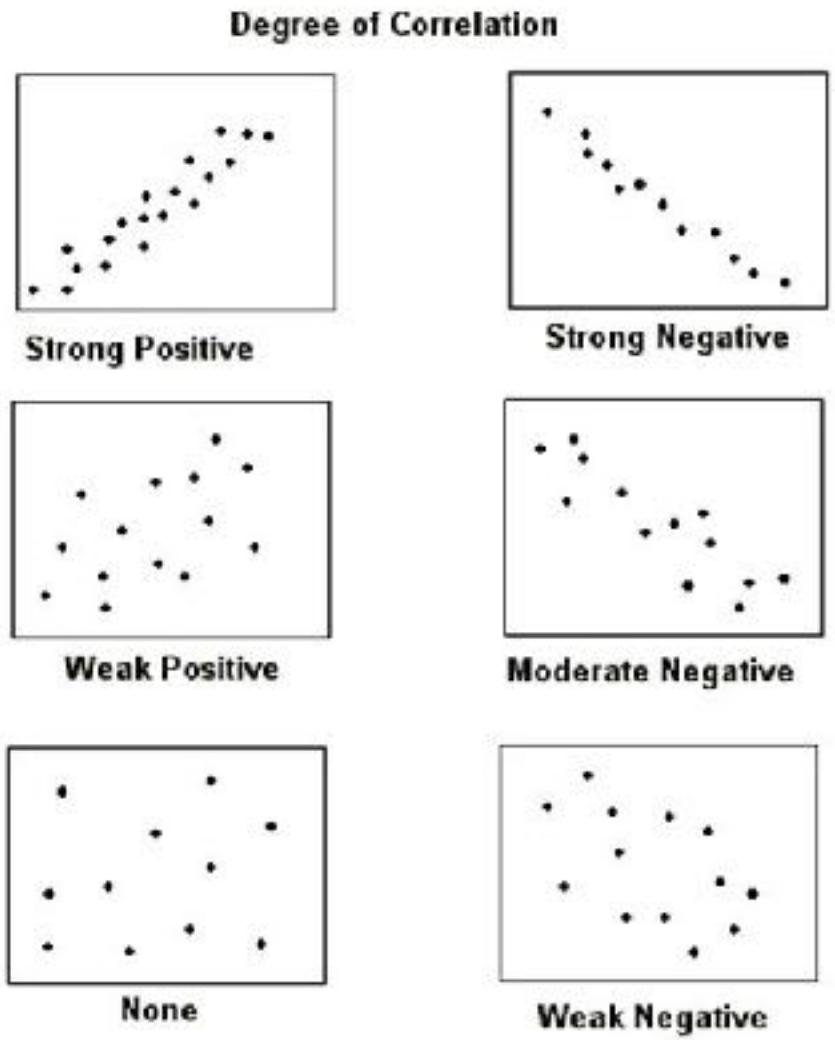
Correlation

- **Correlation** also measures how two variables move with respect to each other.
- A perfect positive correlation means that the correlation coefficient is 1.
- A perfect negative correlation means that the correlation coefficient is -1.
- A correlation coefficient of 0 means that the two variables are independent of each other.
  The formula for finding the correlation coefficient can be found using the following formula.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})\sum_{i=1}^{n}(y_i - \bar{y})}}$$

Positive correlation	Negative correlation	No correlation
		
The points lie close to a straight line, which has a positive gradient.	The points lie close to a straight line, which has a negative gradient.	There is no pattern to the points.
This shows that as one variable **increases** the other **increases**.	This shows that as one variable **increases**, the other **decreases**.	This shows that there is **no connection** between the two variables.

Correlation

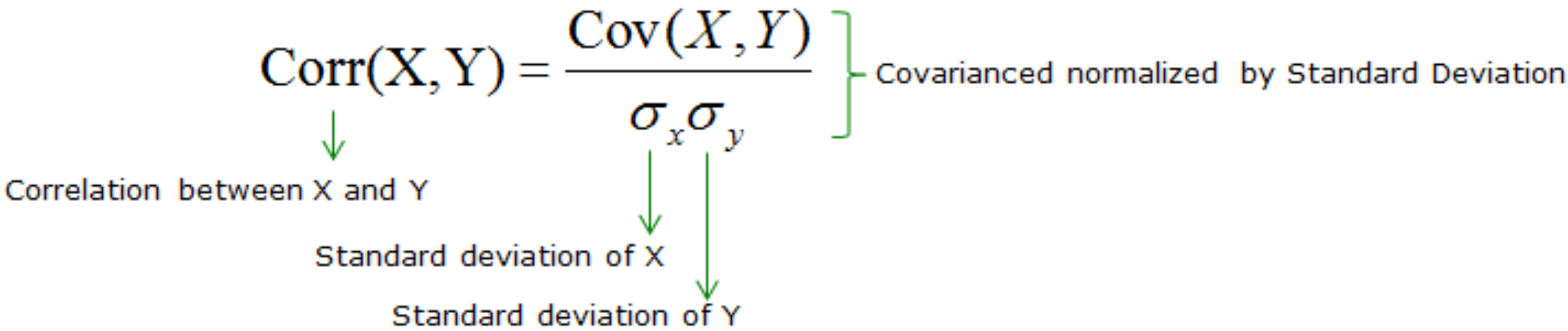


$$r = \frac{\mathrm{COV}_{xy}}{s_x s_y}$$

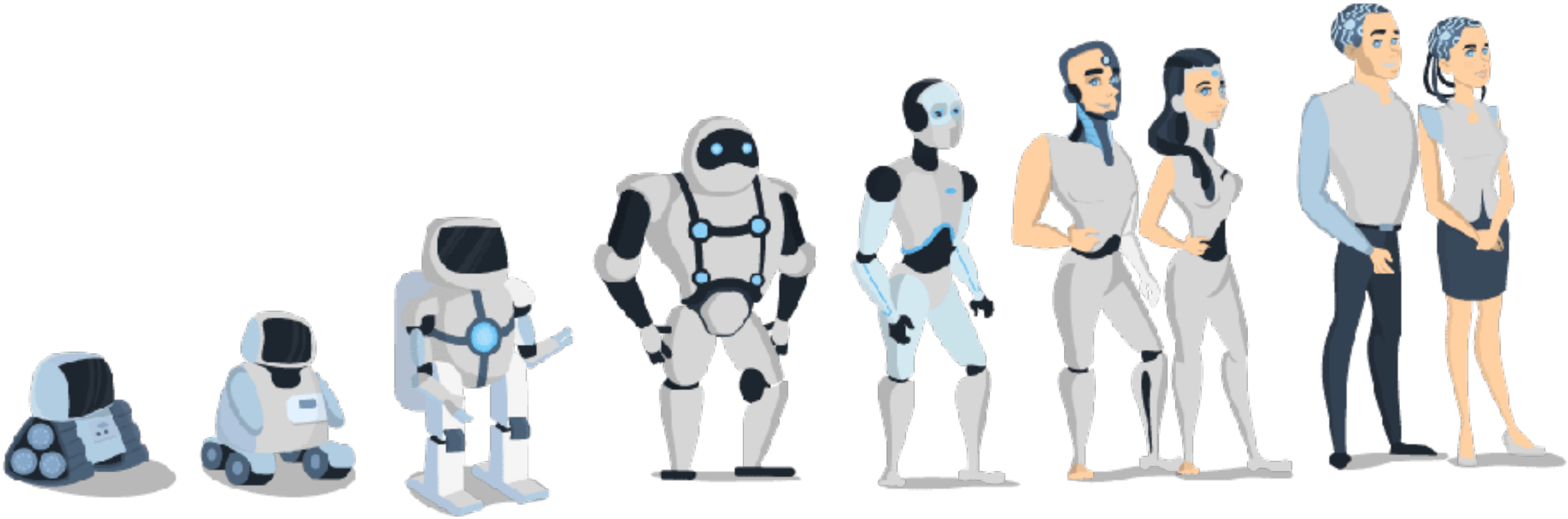
Covariance is High: r ~1

Covariance is Low: r ~0

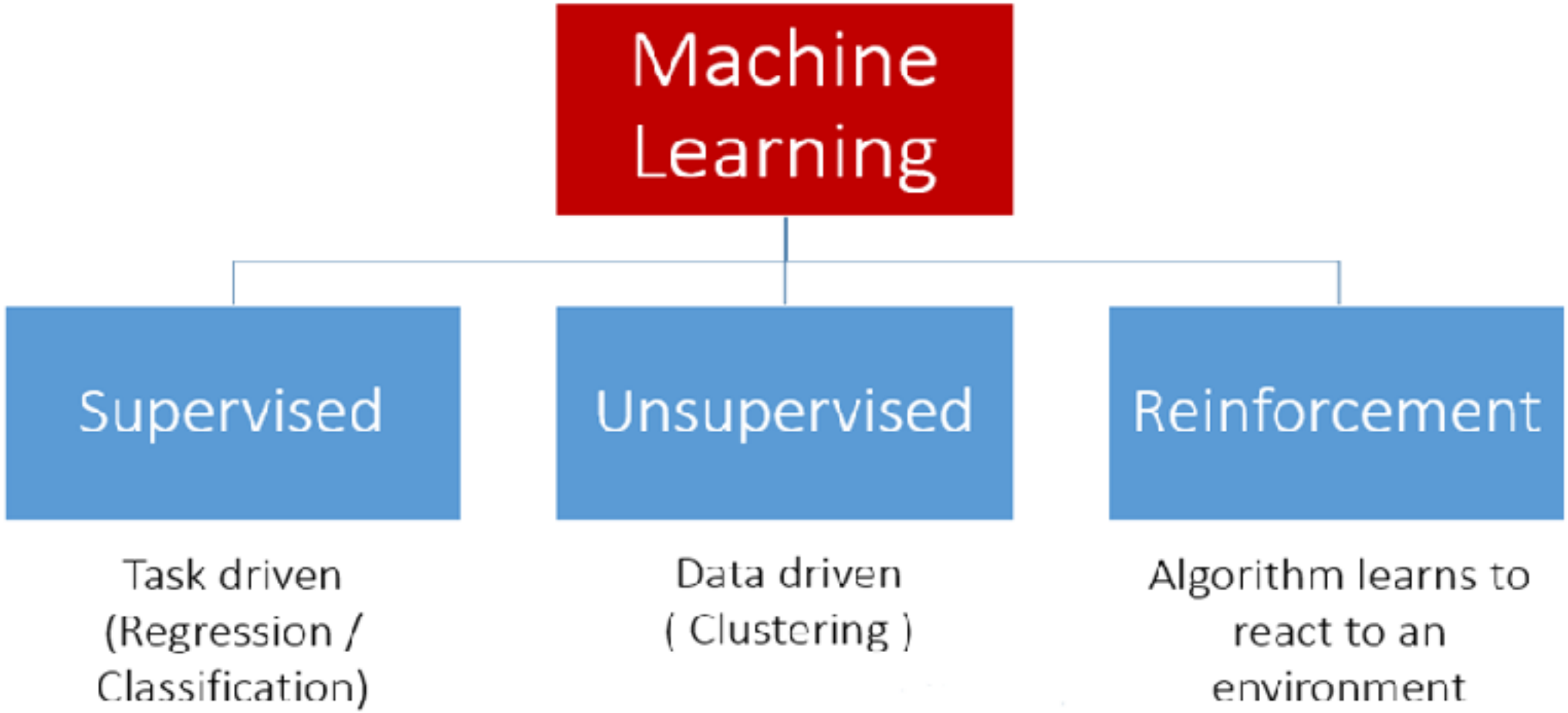
- Both correlation and covariance only measure the linear relationship between data.
- They will fail to discover any nth order relationship between the two. Correlation is a special case of covariance when the data is standardized.
- If we are interested in knowing if there is a relationship then correlation is a better measure as they also measure the extent of the relationship.


$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_x \sigma_y}$
Covarianced normalized by Standard Deviation
Correlation between X and Y
Standard deviation of X
Standard deviation of Y

What is Machine Learning - Evolution of Machines



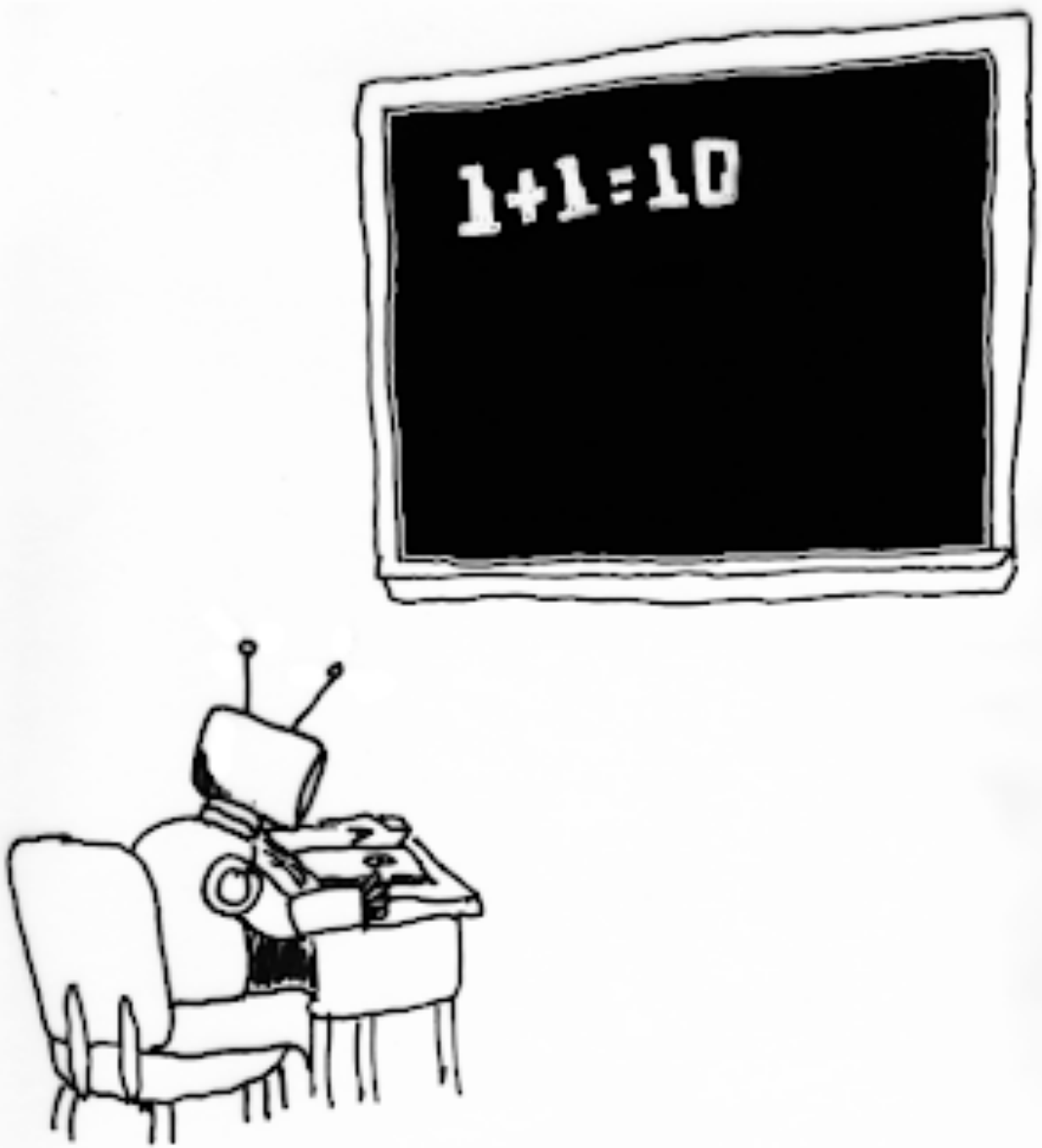
Types of Machine Learning



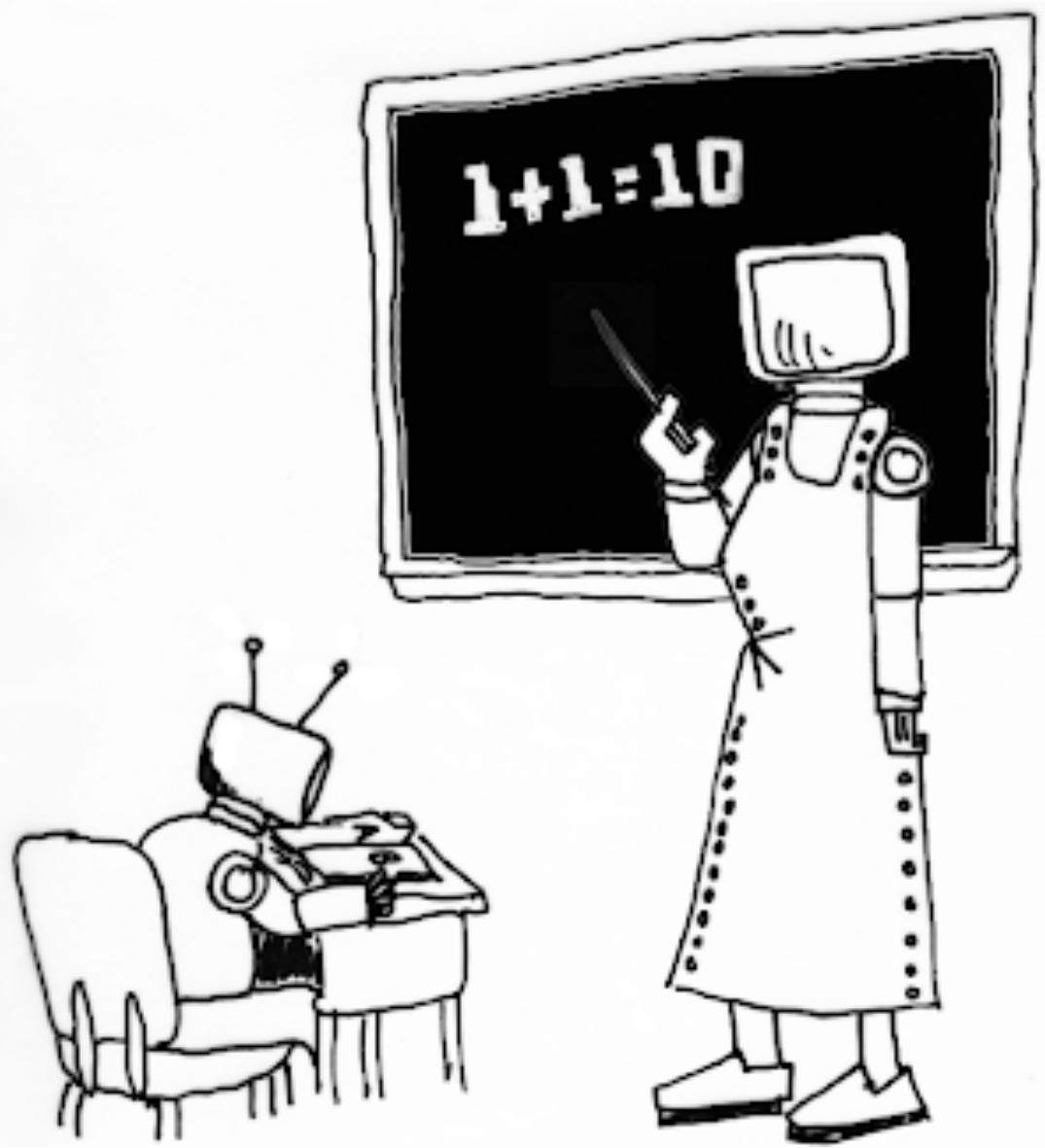
HOW DO I WORK?

- Supervised Learning – Train Me!
- Unsupervised Learning – I am self sufficient in learning
- Reinforcement Learning – My life My rules! (Hit & Trial)

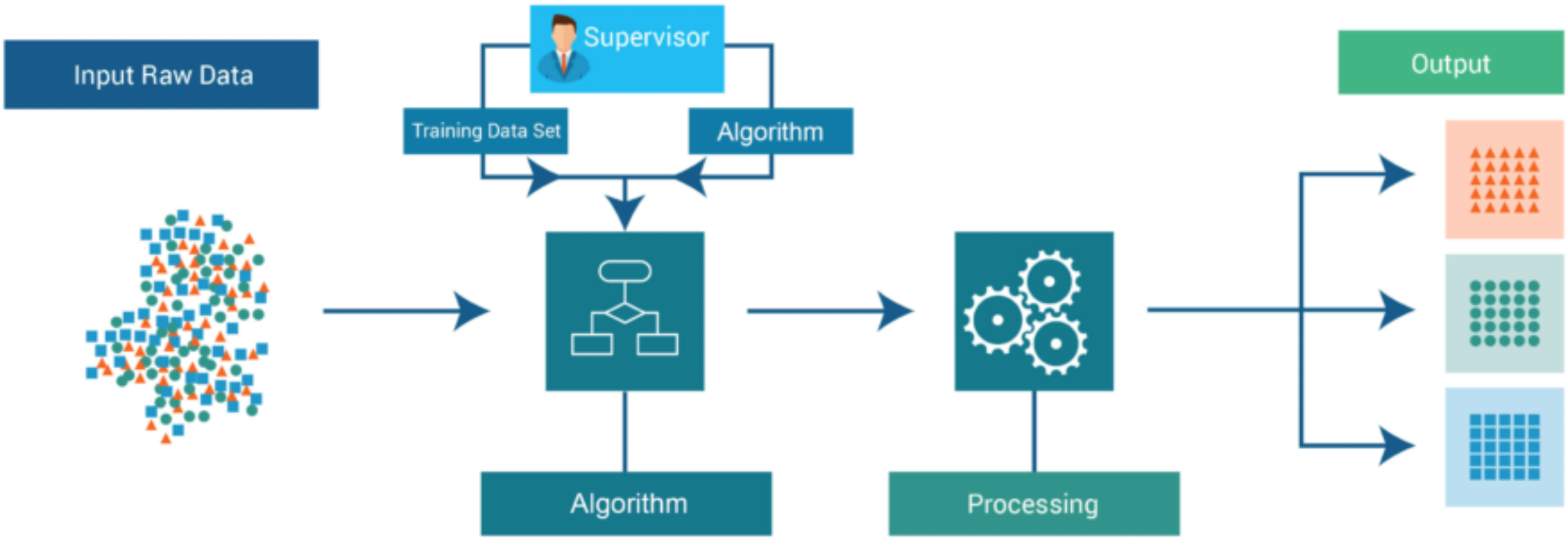
UNSUPERVISED MACHINE LEARNING

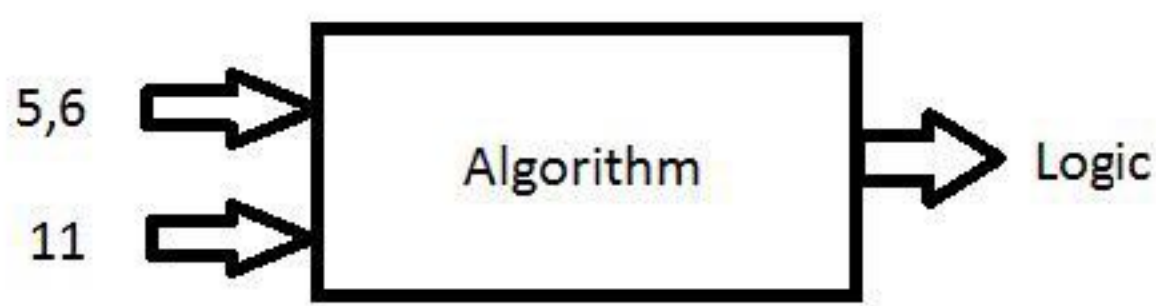


SUPERVISED MACHINE LEARNING

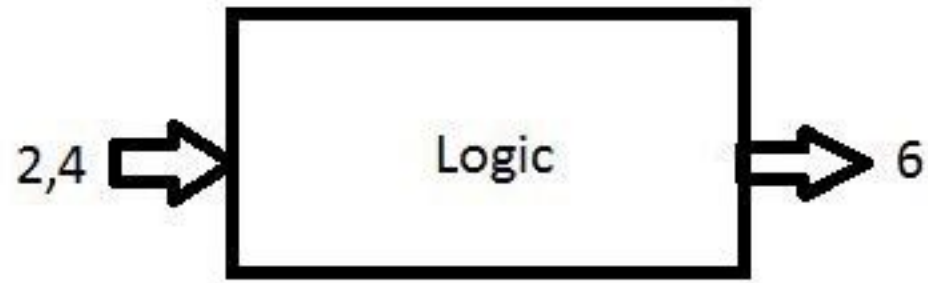


Supervised Learning





Training with Training data



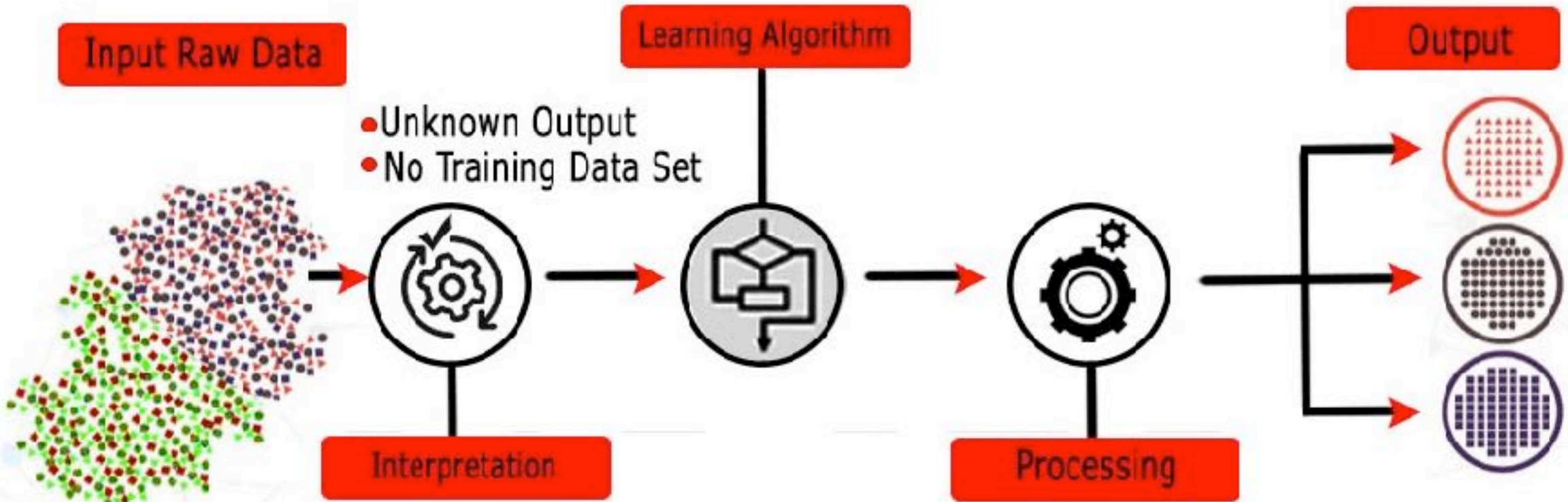
Predicting with new data

Examples

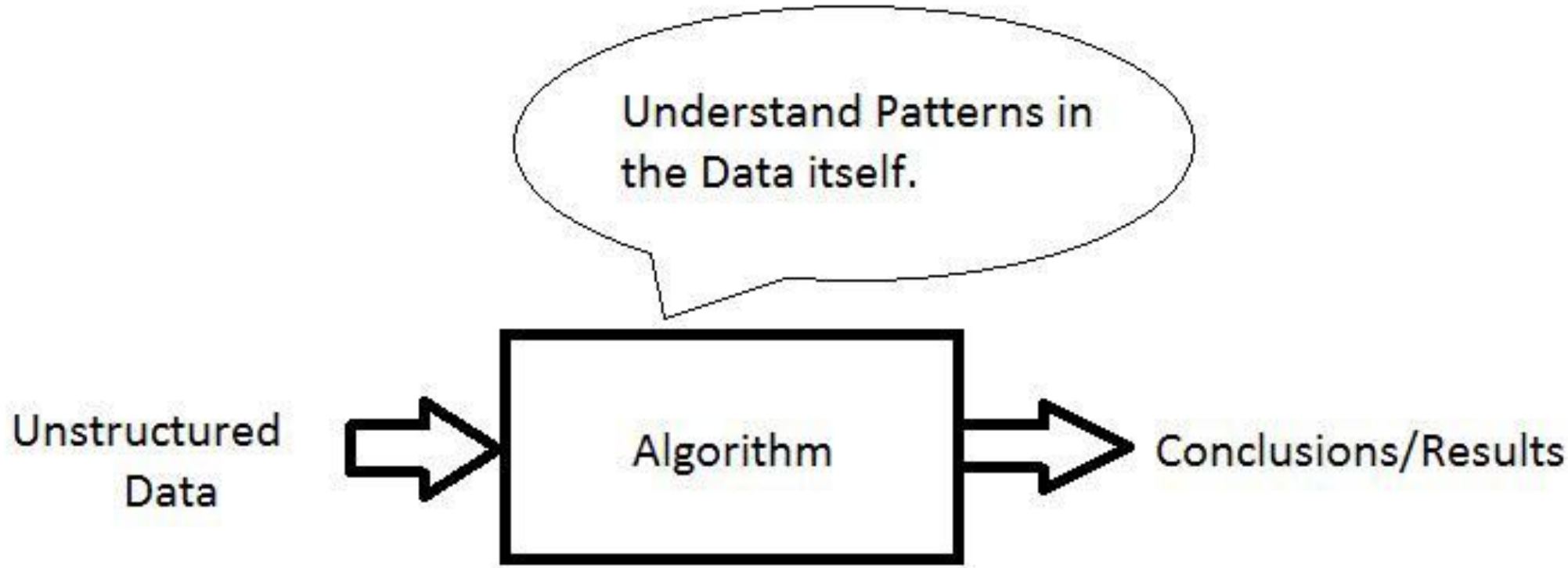
- what is the price of house in a specific city?
- what is the value of the stock?
- how many total runs can be on board in a cricket game?

- this mail is spam or not?
- will it rain today or not?
- is this picture a cat or not?
- this mail is spam or important or promotion?
- is this picture a cat or a dog or a tiger?

Unsupervised Learning

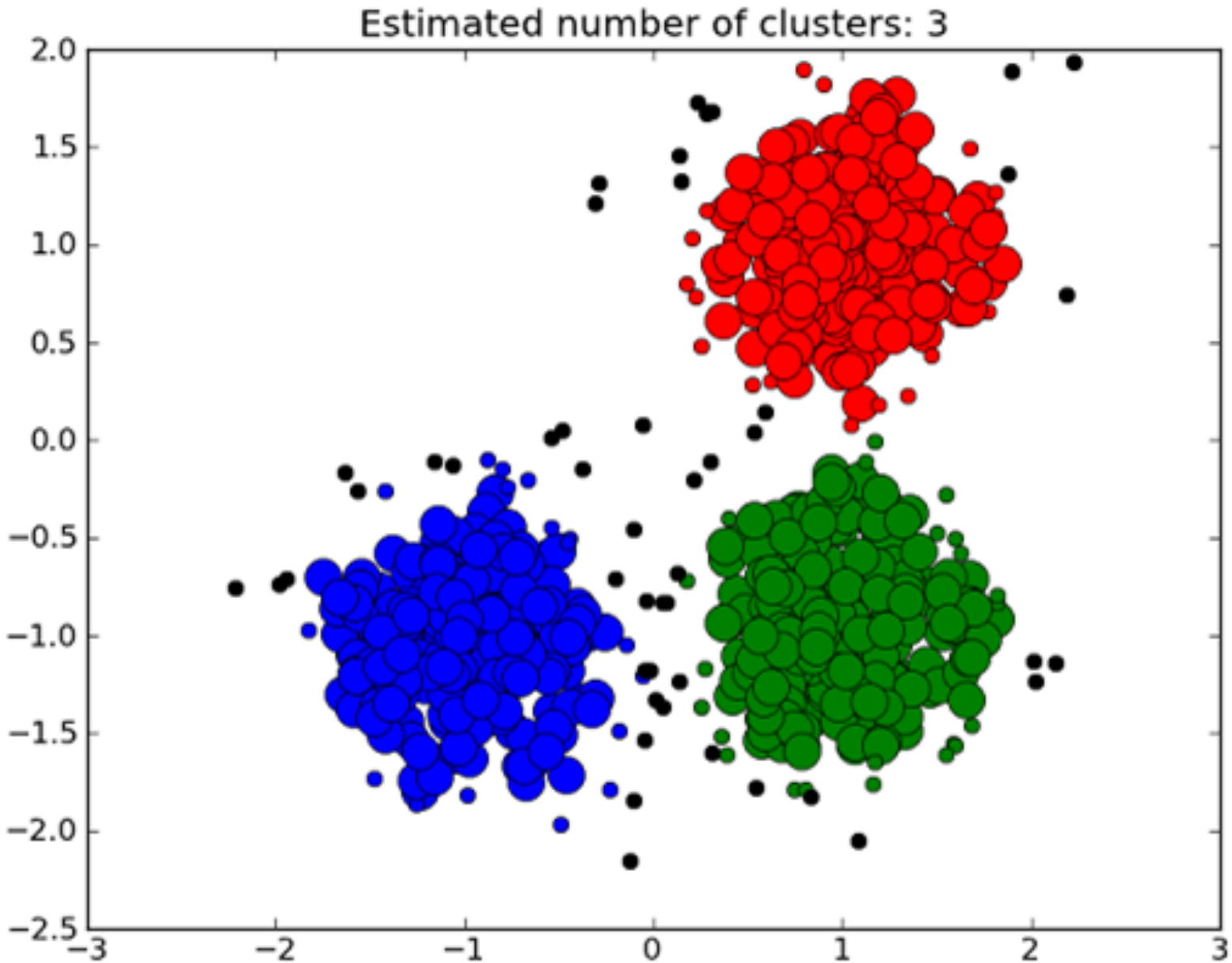


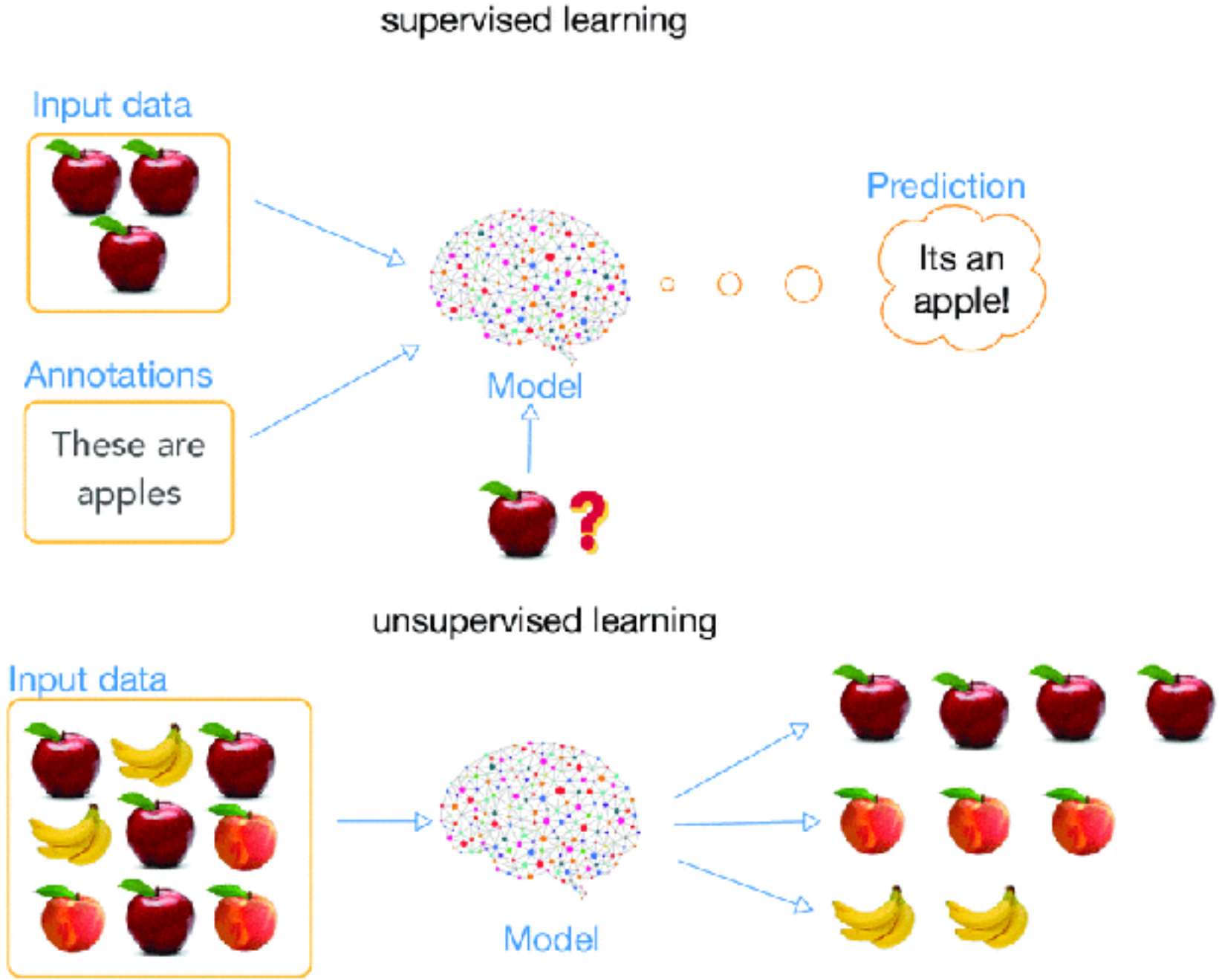

vnessa-hudgens
I gotta go my own way.


Understand Patterns in the Data itself.
Unstructured Data
Algorithm
Conclusions/Results

Examples

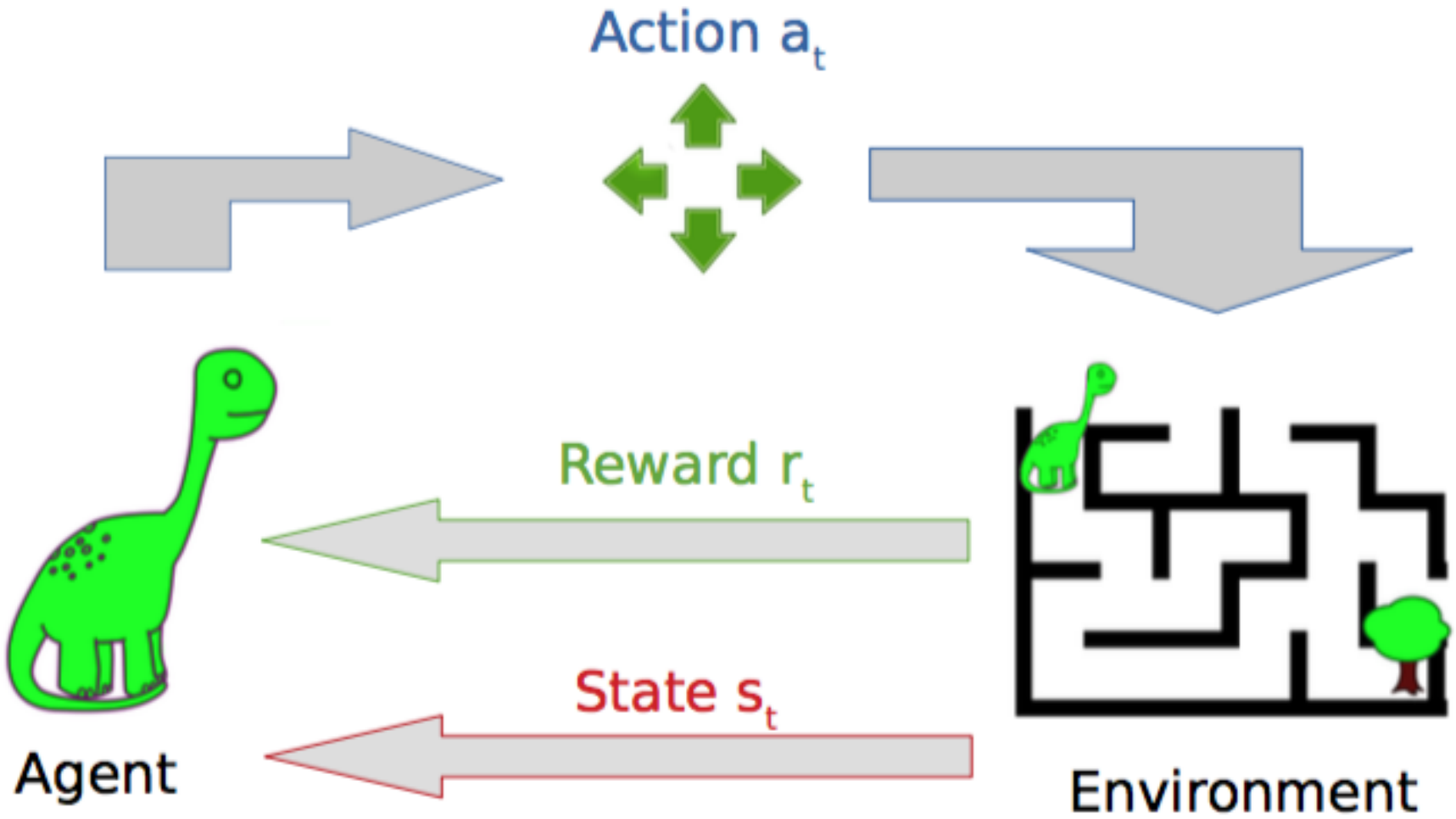
- given news articles,cluster into different types of news
- given a set of tweets ,cluster based on content of tweet
- given a set of images, cluster them into different objects


Estimated number of clusters: 3
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3
−2
−1
0
1
2
3

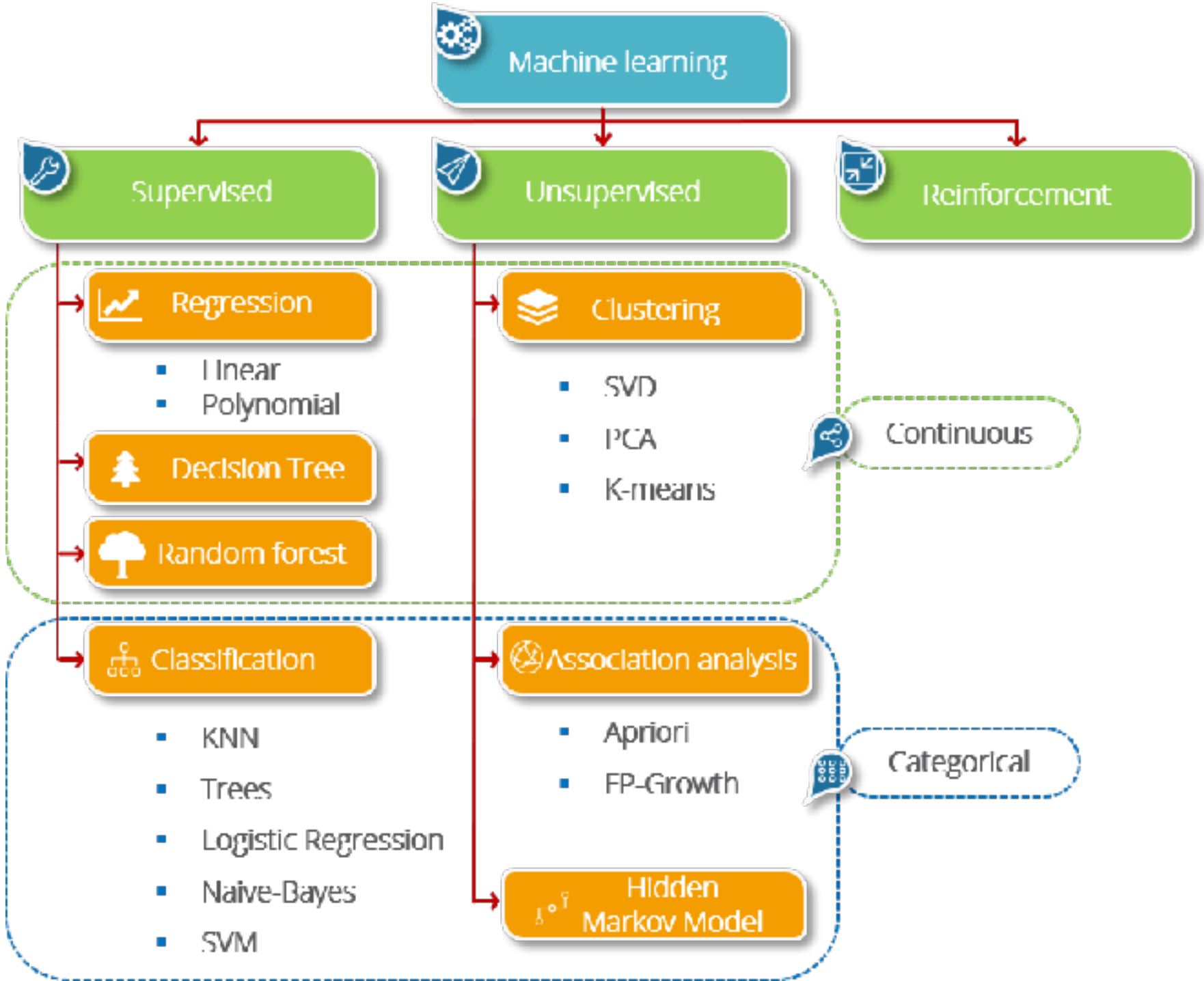

supervised learning
Input data
Annotations
These are apples
Model
Prediction
Its an apple!
unsupervised learning
Input data
Model

Reinforcement Learning

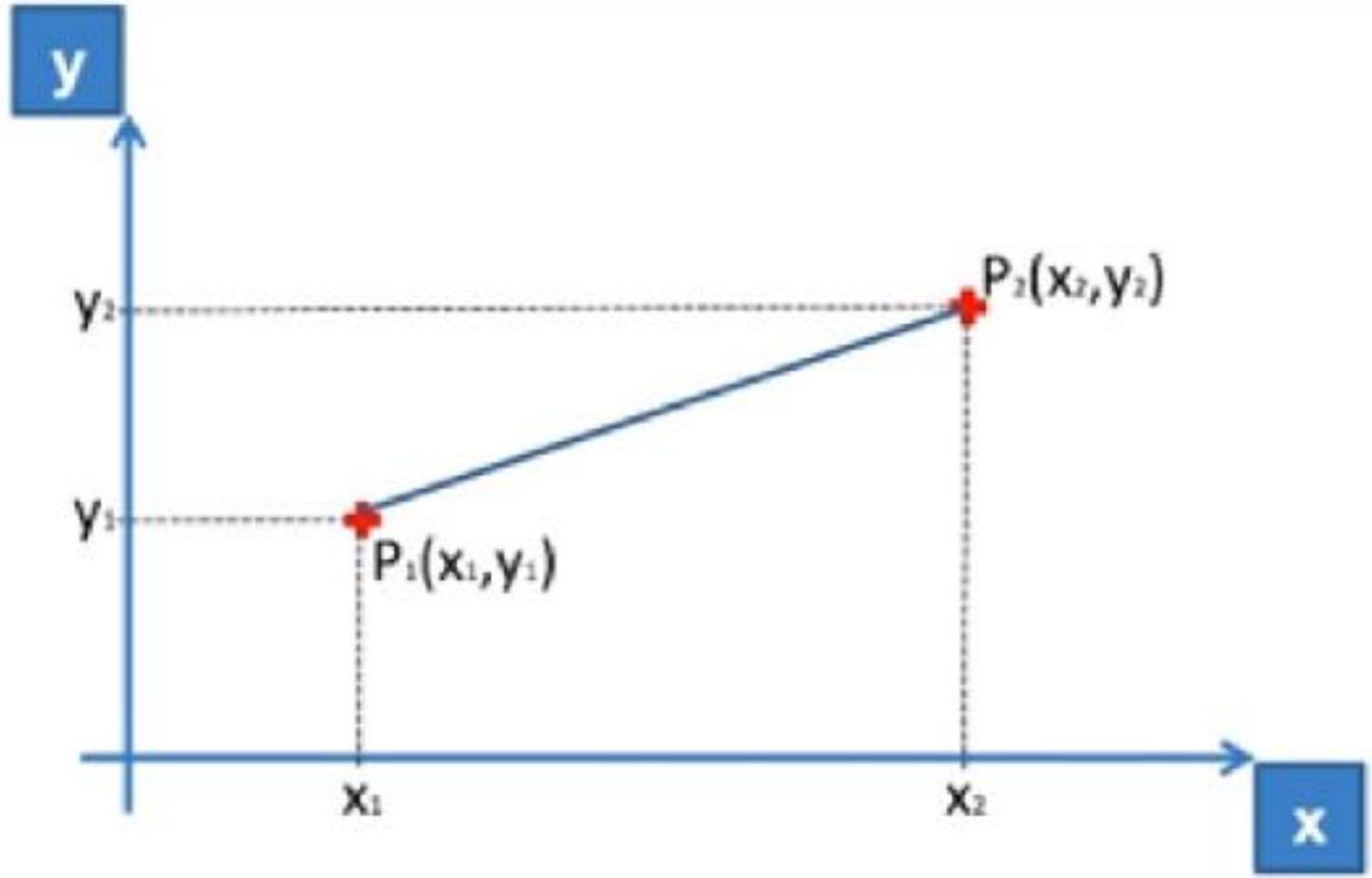
- It is the ability of an agent to interact with the environment and find out what is the best outcome. It follows the concept of hit and trial method. The agent is rewarded or penalized with a point for a correct or a wrong answer, and on the basis of the positive reward points gained the model trains itself. And again once trained it gets ready to predict the new data presented to it.


Action a_t
Reward r_t
State s_t
Agent
Environment


State & Reward
Agent
Environment
Actions


Machine learning
Supervised
Unsupervised
Reinforcement
Regression
Linear
Polynomial
Decision Tree
Random forest
Clustering
SVD
PCA
K-means
Continuous
Classification
KNN
Trees
Logistic Regression
Naive-Bayes
SVM
Association analysis
Apriori
FP-Growth
Hidden Markov Model
Categorical

Feature Scaling



Euclidean Distance between P_1 and $P_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

Standardisation	Normalisation
$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$	$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$