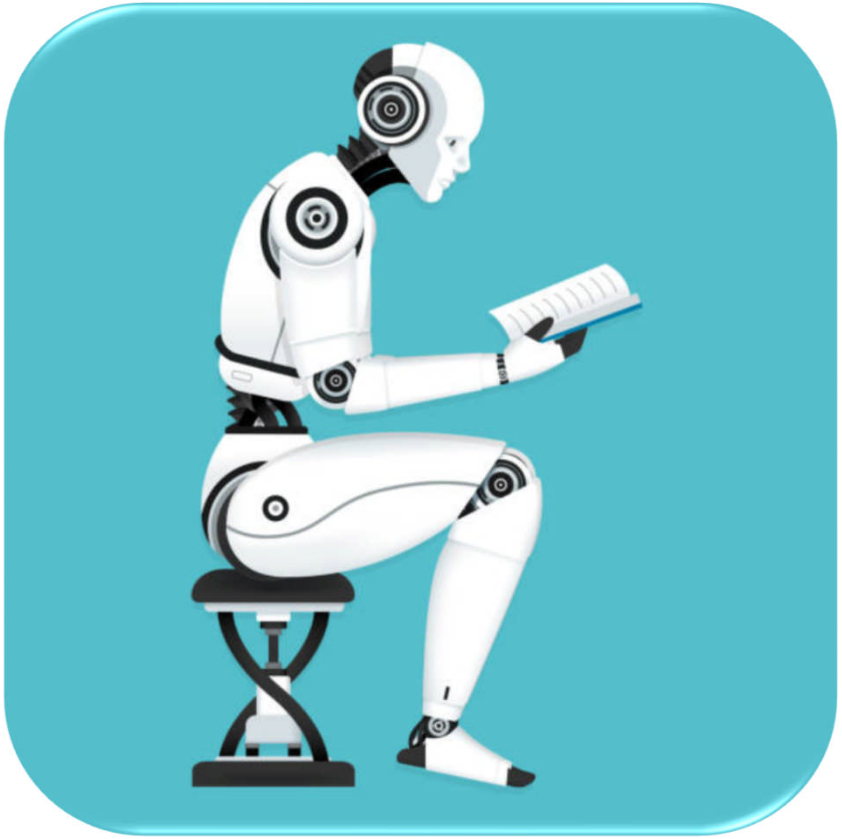
Machine Learning Algorithms



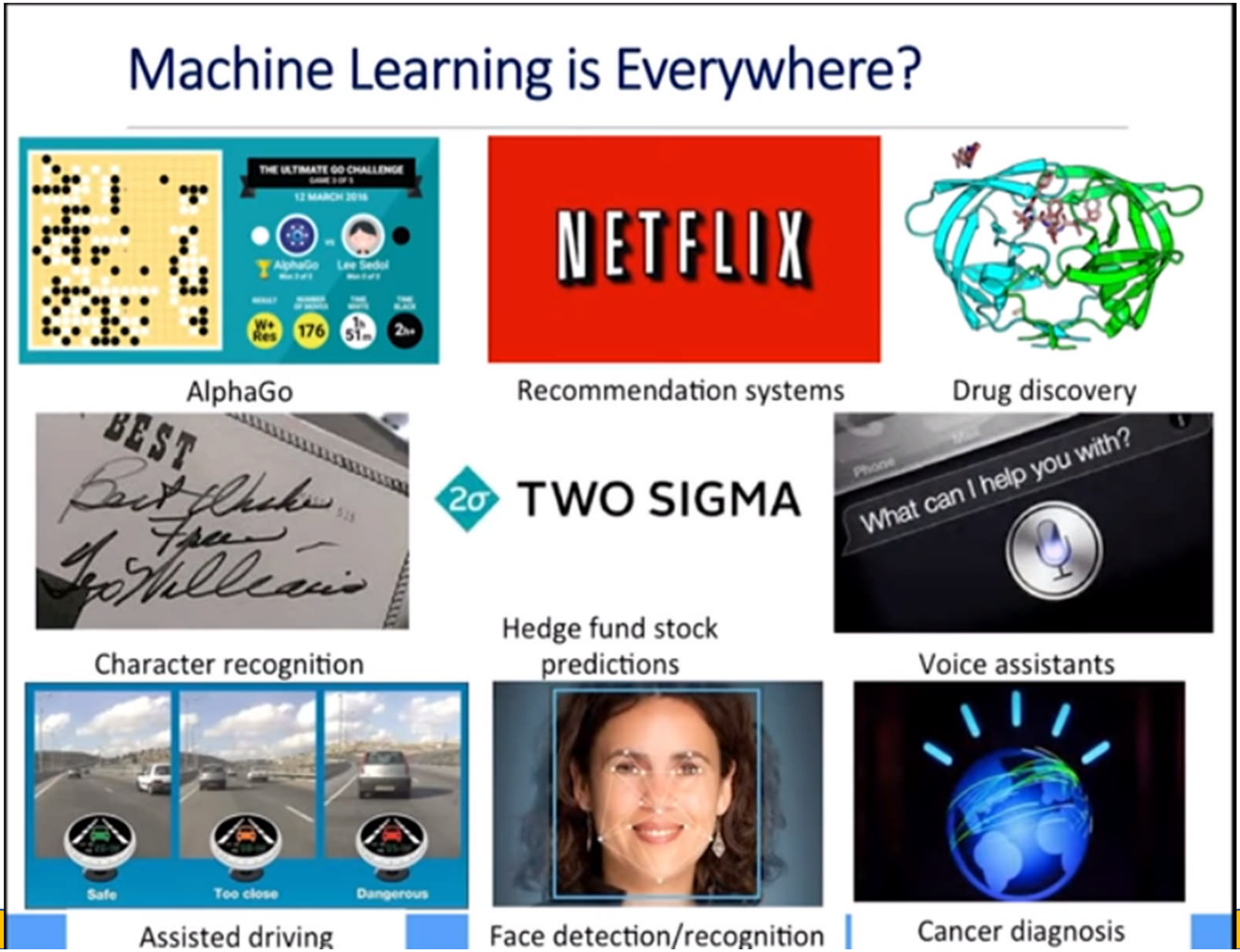
Machine Learning Clip Art - Bing

Machine Learning: Definition

Early definition of Machine Learning

"Field of study that gives computers the ability to learn without being explicitly programmed". Arthur Samuel (1959)

- **What do you mean by Explicitly Programmed ?**
- So, machine learning algorithms, inspired by the human learning process, iteratively learn from data, and allow computers to find hidden insights.
- These models help us in a variety of tasks, such as object recognition, summarization, recommendation, and so on.



Machine Learning is Everywhere?

AlphaGo

Recommendation systems

Drug discovery

Character recognition

Hedge fund stock predictions

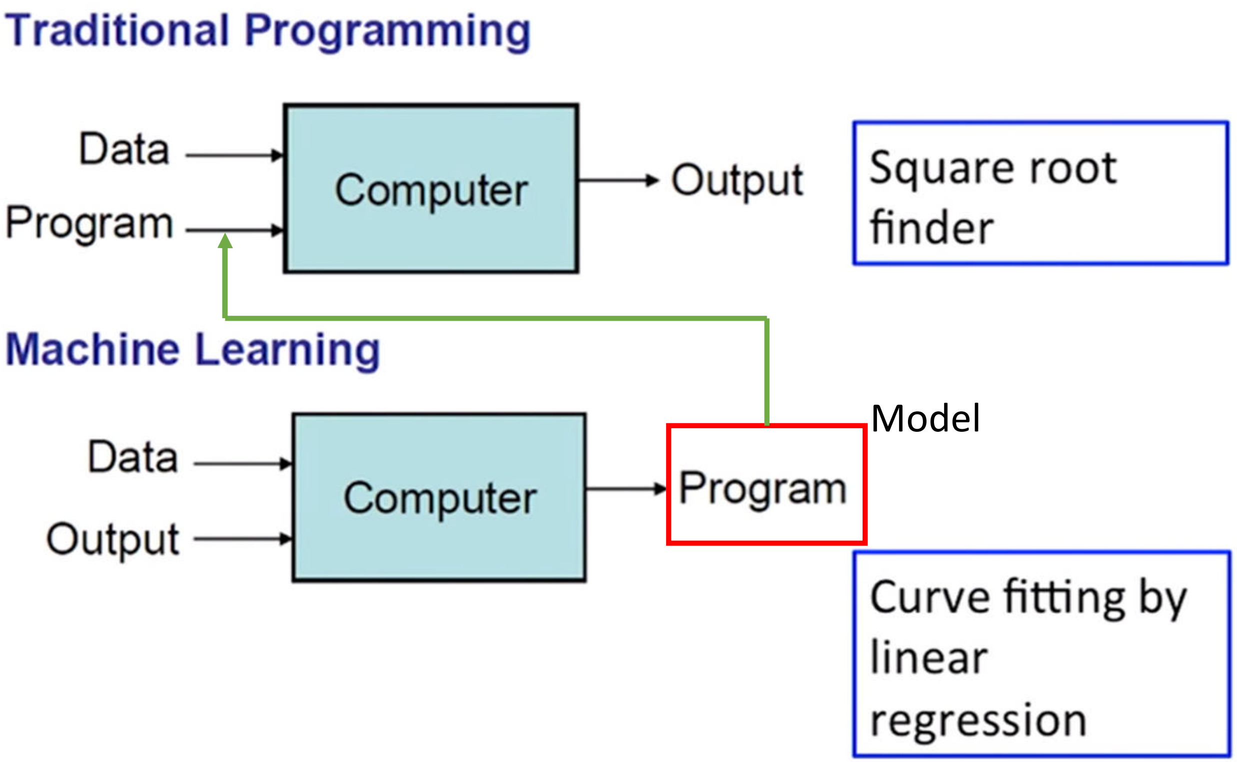
Voice assistants

Assisted driving

Face detection/recognition

Cancer diagnosis

What is Machine Learning ?



Traditional Programming

Data

Program

Computer

Output

Square root finder

Machine Learning

Data

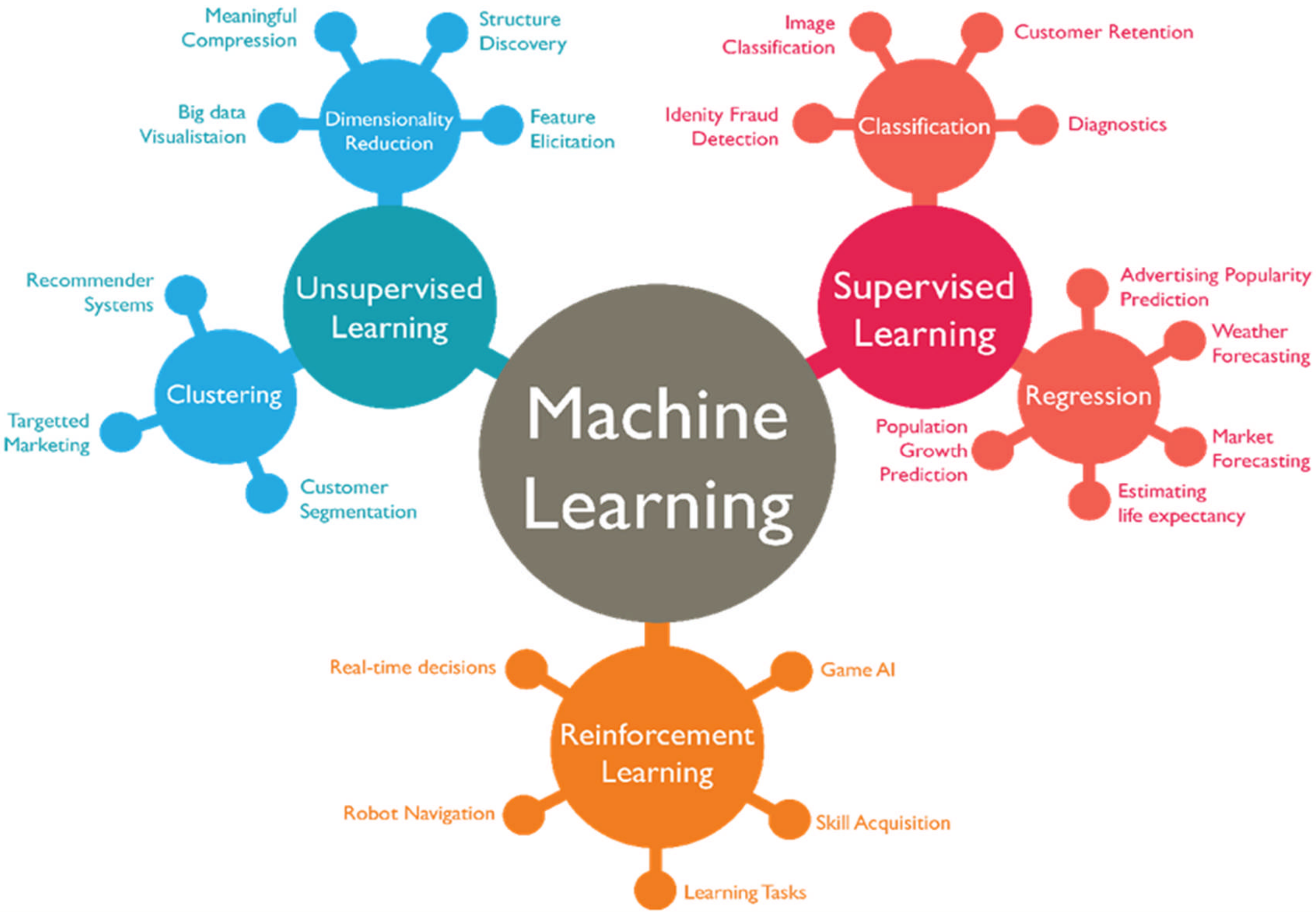
Output

Computer

Program

Model

Curve fitting by linear regression


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

ML Modeling

- **Generic Formulae:**

$$f(x) = y$$

where ‘f’ is the Machine learning algorithm which is applied on data cases ’x’, to generate ‘y’ as prediction.

- If we know what those ‘y’ values are when we're training the model, we call those the ‘labels’.

Data is Important

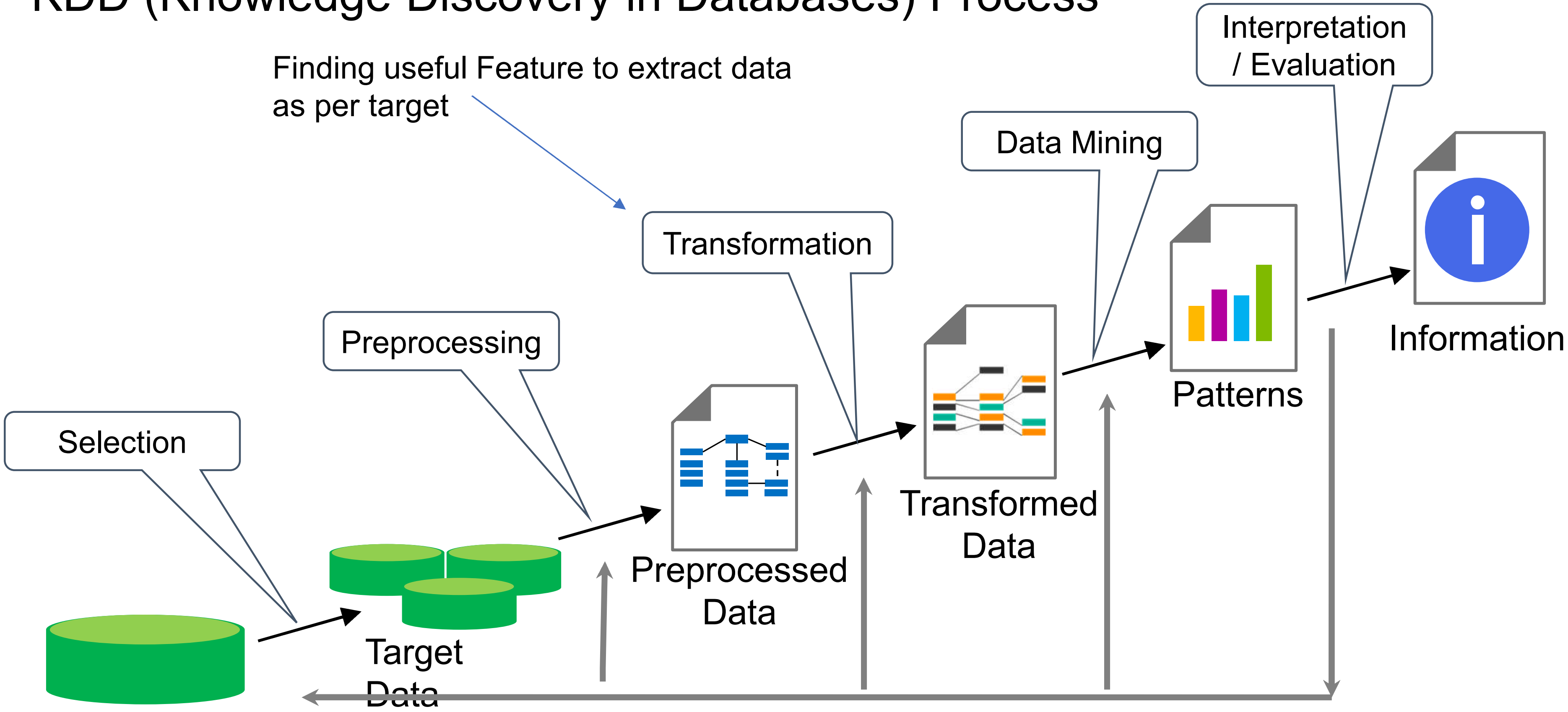
- Data must be:
  - Relevant: relationship between data
  - Connected : No Missing values in data
  - Accurate
  - Enough to work with: To make detailed decisions

- Structured Vs Unstructured Data

Data Science Stages

- The stages are basically the same no matter who invents or reinvents the (knowledge discovery / data mining / big data / data science) process.
- You may not always need all the stages.
- Data science is an iterative process.
- Backwards arrows on most process diagrams.

KDD (Knowledge Discovery in Databases) Process



Based on content in “From Data Mining to Knowledge Discovery”, AI Magazine, Vol 17, No. 3 (1996)

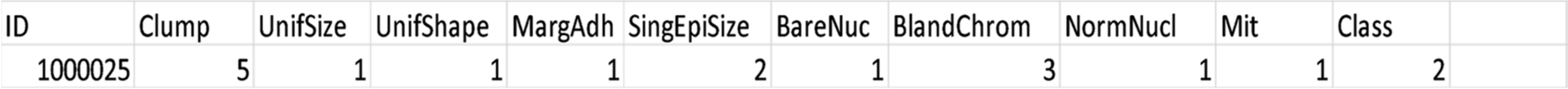
http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230

Model Building

- **Often predictive modeling means machine learning or statistical modeling**
- If you want to answer a yes/no question, this is classification.
  - For manholes, will the manhole explode next year? Y/N
- If you want to predict a numerical value, this is regression.
  - Ex: Stock prediction at BSE Sensex tomorrow ?
- If you want to group observations into similar-looking groups, this is clustering.
- If you want to recommend someone an item (e.g., book/movie/product) based on ratings data from customers, this is a recommender system.
- Note: There are many other machine learning problems.

Real World Example

- This is a human cell sample extracted from a patient.



ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	2

- One of the interesting questions we can ask, at this point is: "Is this a Benign or Malignant cell?”
- Malignant Tumour: invades surrounding tissue or spread around body; MUST be diagnosed early
- Benign Tumour: Non-cancerous

Real World Example

- Only a doctor with years of experience could diagnose this. Right?
- Imagine that you have obtained a dataset containing characteristics of thousands of human cell samples extracted from patients who were believed to be at risk of developing cancer.
- **You can use the values of these cell characteristics in samples from other patients to give an early indication of whether a new sample might be benign or malignant.**
- This might be a potent situation where machine learning can help !

Any Mandate?

- You should clean your data, select a proper algorithm for building a prediction model, and train your model to understand patterns of benign or malignant cells within the data.

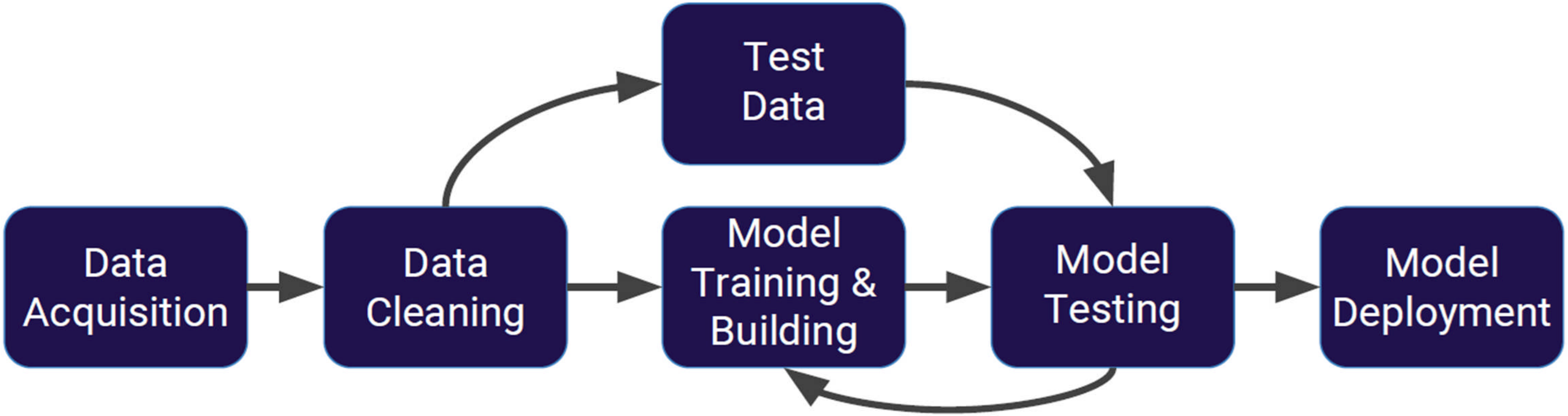
- Once the model has been trained by going through data iteratively, it can be used to predict your new or unknown cell with a rather high accuracy.

- This is how machine learning helps!

More Real World Examples

- **Targeted Marketing--**NETFLIX, AMAZON, YouTube etc.
- **Fraud Detection--**Banks making decision while approving a loan application.
- **Customer Churn--C**ompanies use their customer's demographic data to segment them, or predict if they will unsubscribe from their company the next month.
- SPAM Detection
- And many more!

Machine Learning Process



Python for Machine Learning: Libraries

- There are a lot of modules and libraries already implemented in python that can make out life much easier.
- Few necessary python packages and libraries which are required are:
  - **Numpy,** which is a math library to work with n-dimensional arrays in Python.
  - It enables you to do computation efficiently and effectively.
  - For example, for working with arrays, dictionaries, functions, datatypes, and working with images, you need to know Numpy.

Python for Machine Learning: Libraries

- **SciPy** is a collection of numerical algorithms and domain-specific toolboxes, including signal processing, optimization, statistics and much more.

- **Matplotlib** is a very popular plotting package that provides 2D plotting as well as 3D plotting.

- Basic Knowledge about these 3 packages, which are built on top of python, is a good asset for data scientists who want to work with real world problems.

Python for Machine Learning: Libraries

- **Pandas** library, is a very high-level python library that provides high-performance, easy to use data structures.

- It has many functions for data importing, manipulation and analysis.

- In particular, it offers data structures and operations for manipulating numerical tables and time series.

- **Scikit-learn** is a collection of algorithms and tools for machine learning.

Scikit-Learn Library

- Scikit-learn is a free machine learning library for the Python programming language.

- **It has most of the classification, regression and clustering algorithms, and it is designed to work with the Python numerical and scientific libraries, NumPy and SciPy.**

- Most of the tasks that need to be done in a machine learning pipeline are implemented already in scikit learn, including, pre-processing of data, feature selection, feature extraction, train/test splitting, defining the algorithms, fitting models, tuning parameters, prediction, evaluation, and exporting the model.

Implementing Scikit-learn Library : Example

- Let’s look how to use Scikit-Learn library.
- Basically, **Machine learning algorithms benefit from standardization of the data set.**
- **If there are some outliers, or different scales fields in your data set, you have to fix them.**
- The preprocessing package of scikit learn provides several common utility functions and transformer classes to *change raw feature vectors into a suitable form of vector for modeling.*

```
from sklearn import preprocessing
X = preprocessing.StandardScaler().fit(X).transform(X)
```

Implementing Scikit-learn Library : Example

- Now, you have to split your dataset into train and test sets to train your model, and then test the model's accuracy separately.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test=train_test_split(X, y, test_size=0.33)
```

- Then, you can setup your algorithm.

Implementing Scikit-learn Library : Example

- For example, you can build a classifier using a support vector classification algorithm.

```
from sklearn import svm
clf=svm.SVC(gamma=0.001, C=100)
```

- We call our estimator instance clf, and initialize its parameters.
- Now, you can train your model with the train set.
- By passing our training set to the fit method, the clf model learns to classify unknown cases.

```
clf.fit(X_train, y_train)
```

Making Predictions

- Then, we can use our test set to run predictions.

```
clf.predict(X_test)
```

- And, the result tells us what the class of each unknown value is.
- Also, you can use different metrics to evaluate your model accuracy, for example, using a confusion matrix to show the results.

```
from sklearn.metrics import confusion_matrix
print(confusion_matrix(y_test, yhat, labels=[1,0]))
```

- And finally, you save your model.

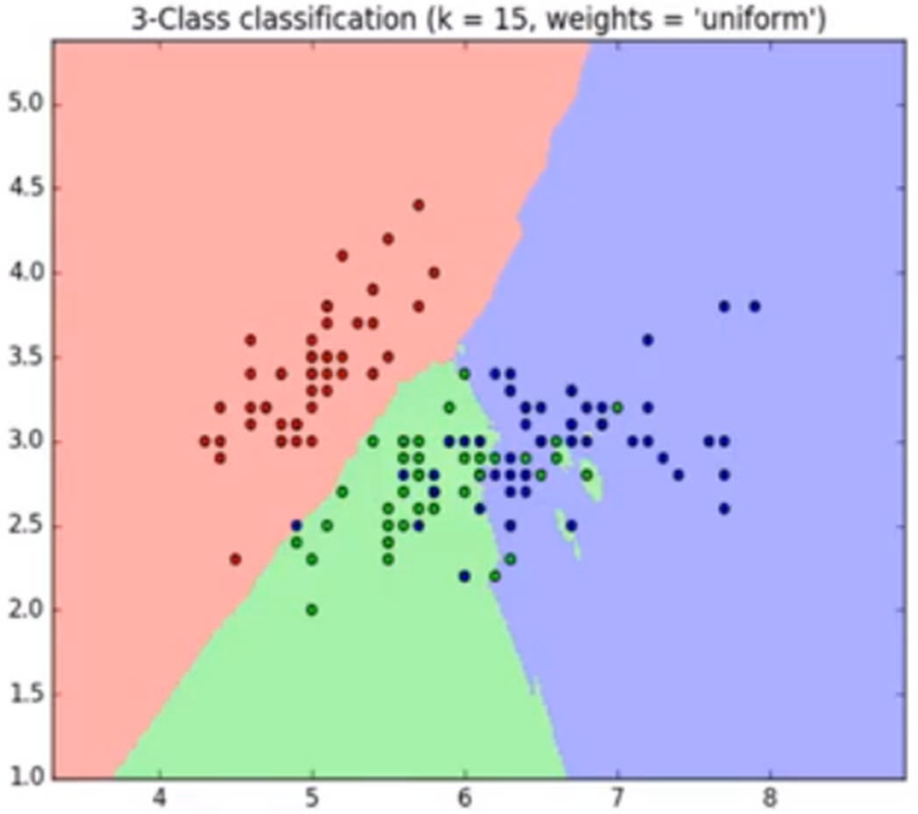
```
import pickle
s=pickle.dumps(clf)
```

Supervised algorithms Vs Unsupervised algorithms

Supervised & Unsupervised algorithms

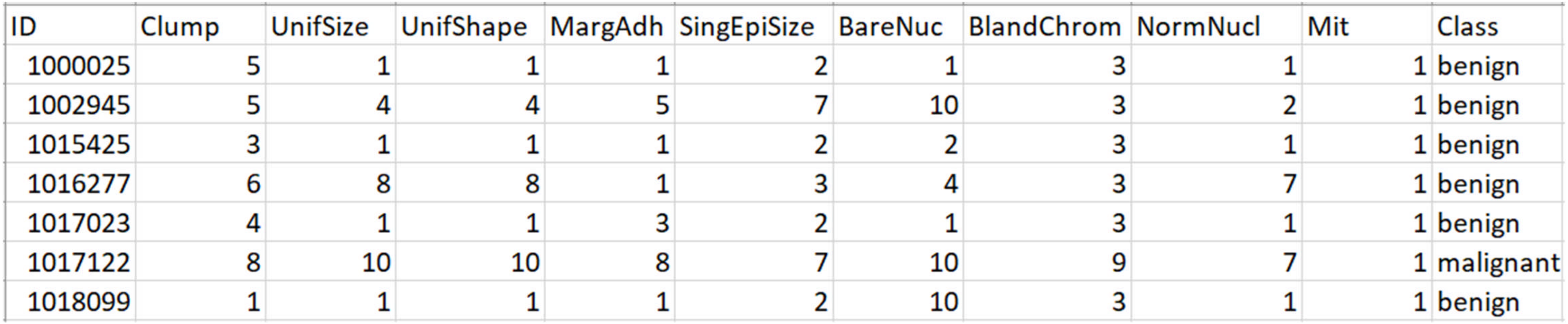
- We supervise a machine learning model by teaching (or training) the model with some data, from a **labeled dataset**.

- That is, **we teach the model (load the model) with knowledge so that we can use knowledge to predict unknown or future instances.**



Features and Observation

- The columns are called **Features**, which include the data.
- If you plot this data, and look at a single data point on a plot, it have all of these attributes.
- That would make a row on this chart, also referred to as an **observation**.



ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	benign
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10	9	7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign

Attribute Types

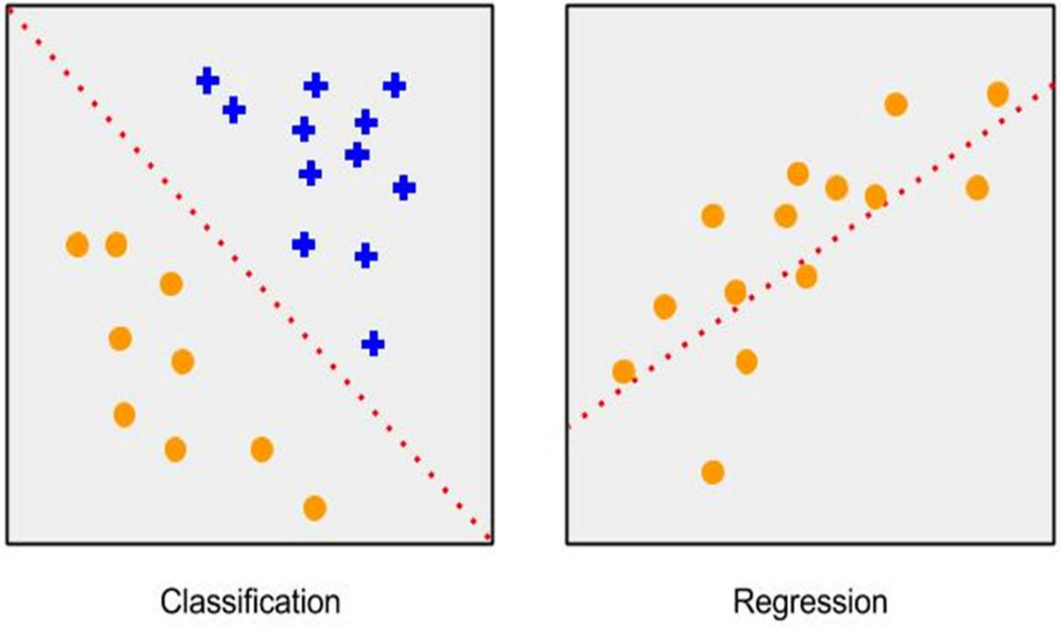
You may have two kinds of value of the data.

1. Numerical.
2. Categorical.

Supervised Algorithms

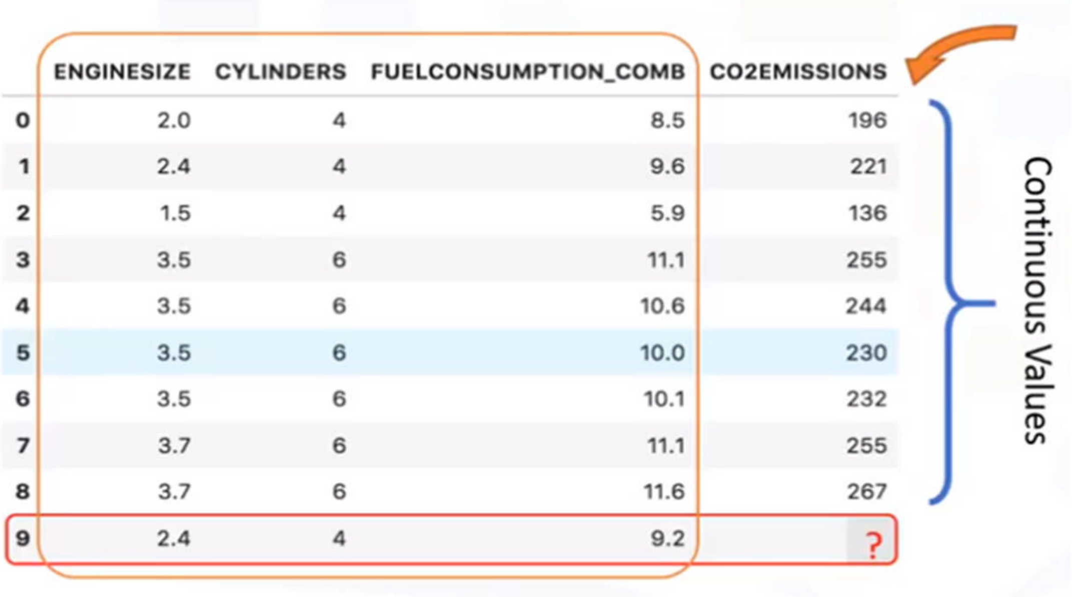
- There are two types of Supervised Learning techniques.
    - **Classification and regression.**

- **Classification** is the process of predicting a **discrete class label** or category.
- **Regression** is the process of predicting a **continuous value** as opposed to predicting a categorical value in Classification.



Problem Type?

- This dataset is related to Co2 emissions of different cars.
- Can use regression to predict the Co2 emission of a new car by using other fields, such as Engine size or number of Cylinders.



	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Continuous Values

Unsupervised Algorithms

- In Unsupervised learning, we do not supervise the model, but we train the model on the dataset to discover information that may not be visible to the human eye, on Unlabeled Data.

- The system is not told the "right answer" ; algorithm must figure out what is being shown.

- The goal is to explore the data and find some structure within or it can find the main attributes that separate segments from each other.

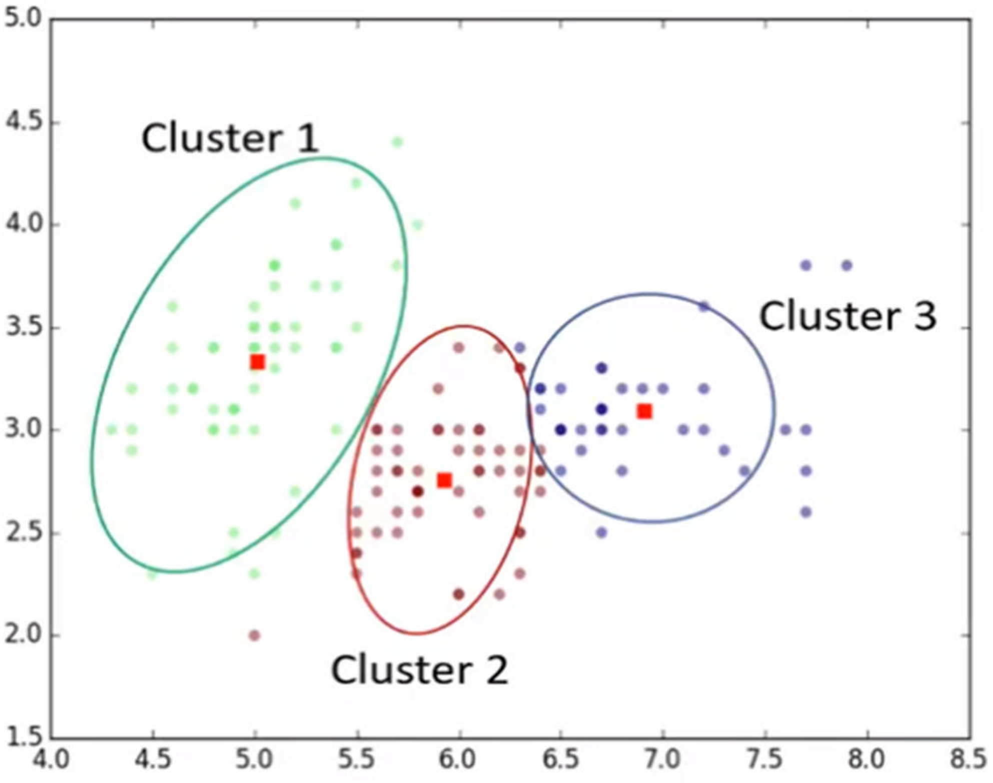
Popular Machine Learning Techniques

- **Dimensionality Reduction and/or feature selection** play a large role by reducing redundant features to make the classification easier.
- **Market basket analysis** is a modelling technique based upon the theory that if you buy a certain group of items, you are more likely to buy another group of items.
- **Density estimation** is a very simple concept that is mostly used to explore the data to find some structure within it.
- **Clustering** is considered to be one of the most popular unsupervised machine learning techniques used for grouping data points or objects that are somehow similar.

Cluster Analysis

- **Cluster analysis** has many applications in different domains.

- Bank's desire to segment its customers based on certain characteristics, or helping an individual to organize and group his/her favorite types of music!



Reinforcement Learning

- **Reinforcement learning** is often used for robotics, gaming and navigation.
- With reinforcement learning, the algorithm discovers through trial and error which actions yield the greatest rewards.
- This type of learning has **three primary components**:
  1. the agent (the learner or decision maker)
  2. the environment (everything the agent interacts with) and
  3. actions (what the agent can do).

Start Writing “Hello World” Machine Learning Code in Only 6 Lines!

Write a code to differentiate between Apples & Oranges ?

Training Data

Features

Input of classifier		Output of classifier
Weight	Texture	Label
150g	Bumpy	Orange
170g	Bumpy	Orange
140g	Smooth	Apple
130g	Smooth	Apple
…	…	…

1. import sklearn
2. features = [[140,"smooth"],[130,"smooth"],[150,"bumpy"],[170,"bumpy"]]
3. labels = ["apples", "apples", "orange", "orange"]

Change strings to integers

First 3 lines of code !

1. import sklearn
2. features = [[140, 1], [130, 1], [150, 0], [170, 0]]
3. labels = [0, 0, 1, 1]

0: bumpy ; 1: smooth

0: apple ; 1: orange

Methodology

```
from sklearn import tree
features = [[140, 1], [130, 1], [150, 0], [170, 0]]
labels = [0, 0, 1, 1]
clf = tree.DecisionTreeClassifier()
clf = clf.fit(features, labels)
print(clf.predict([[150, 0]]))
```

0: bumpy ; 1: smooth

0: apple ; 1: orange

Classifier gets trained on input data

Hands On