Spark MLlib

Tushar B. Kute,

http://tusharkute.com

Machine Learning

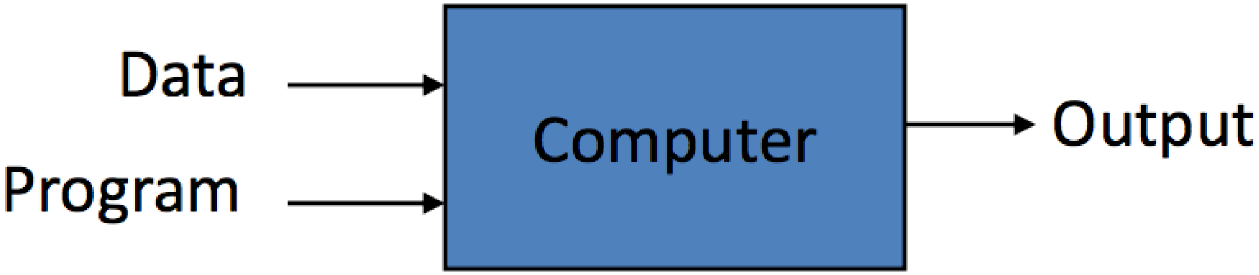
- Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.
- The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide.
- The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Origins of Machine Learning

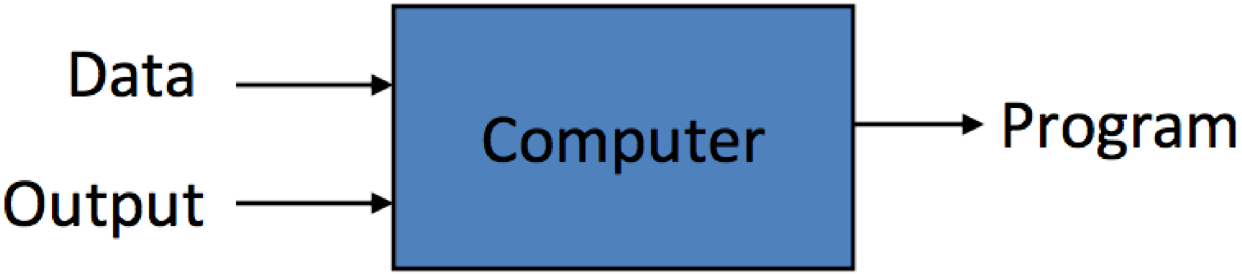
- The earliest databases recorded information from the observable environment.
- Astronomers recorded patterns of planets and stars; biologists noted results from experiments crossbreeding plants and animals; and cities recorded tax payments, disease outbreaks, and populations. Each of these required a human being to first observe and second, record the observation.
- Today, such observations are increasingly automated and recorded systematically in ever-growing computerized databases.

Machine Learning

Traditional Programming

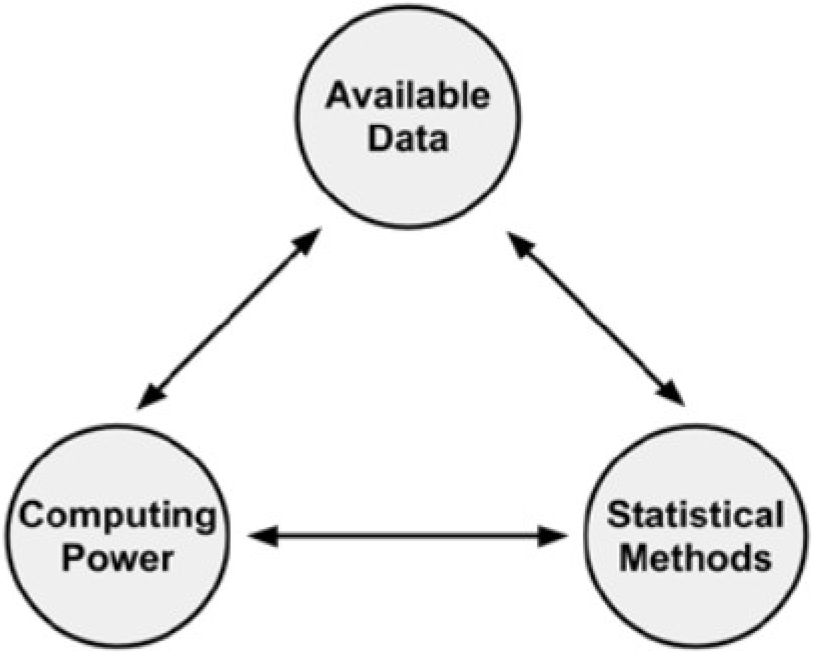


Machine Learning



Machine Learning

- The field of study interested in the development of computer algorithms for transforming data into intelligent action is known as machine learning.



Uses and Abuses

- Predict the outcomes of elections
- Identify and filter spam messages from e-mail
- Foresee criminal activity
- Automate traffic signals according to road conditions
- Produce financial estimates of storms and natural disasters
- Examine customer churn
- Create auto-piloting planes and auto-driving cars
- Stock market predition
- Target advertising to specific types of consumers

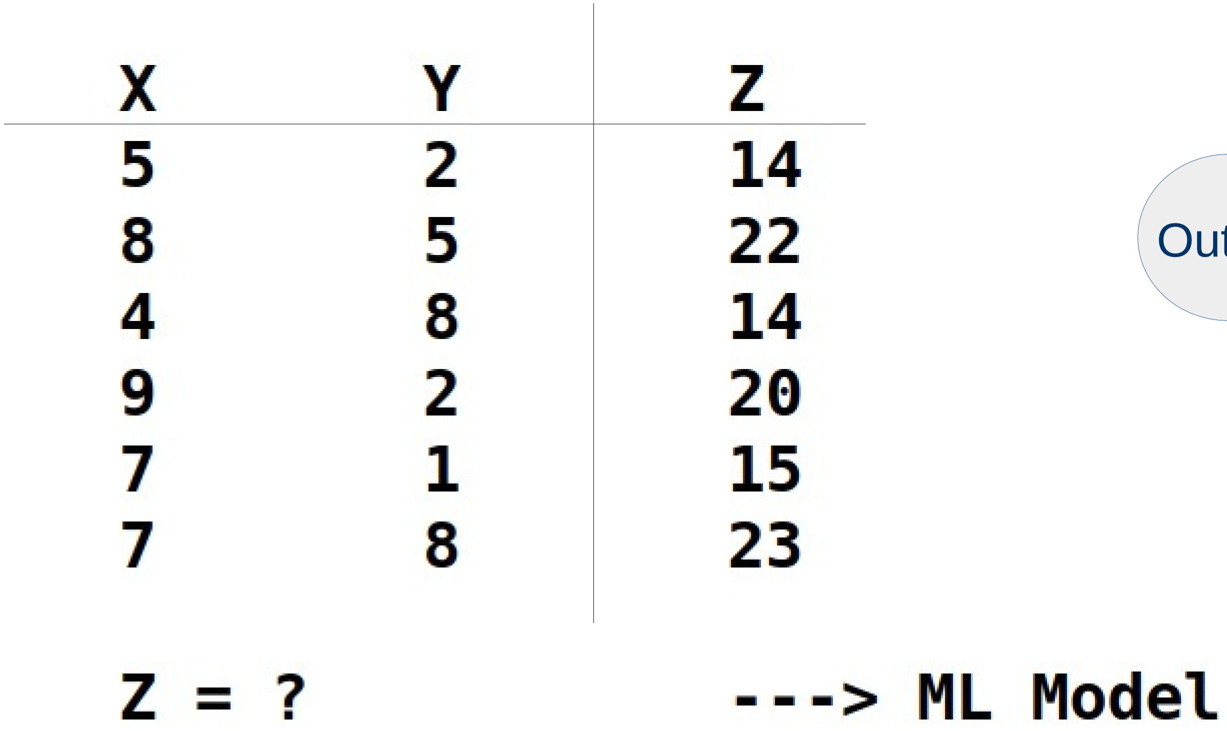
How do machine learn ?

- A commonly cited formal definition of machine learning, proposed by computer scientist Tom M. Mitchell, says that a machine is said to learn if it is able to take experience and utilize it such that its performance improves up on similar experiences in the future.
- This definition is fairly exact, yet says little about how machine learning techniques actually learn to transform data into actionable knowledge.

Training a dataset

- The process of fitting a particular model to a dataset is known as training.
- Why is this not called learning? First, note that the learning process does not end with the step of data abstraction.
- Learning requires an additional step to generalize the knowledge to future data.
- Second, the term training more accurately describes the actual process undertaken when the model is fitted to the data.

Practical Machine Learning



X	Y	Z
5	2	14
8	5	22
4	8	14
9	2	20
7	1	15
7	8	23

```
Z = ?                  ---> ML Model
```

Inputs

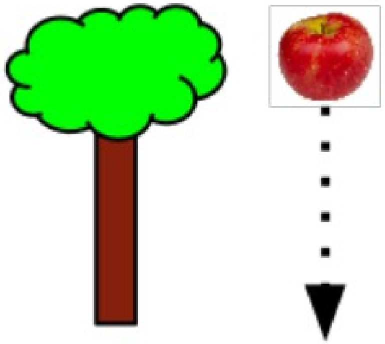
Output

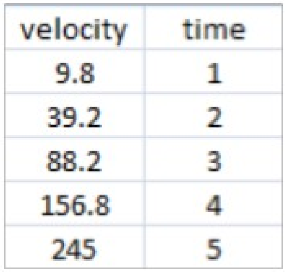
X	Y	Z	Pre	Error
5	2	14	12	-2
8	5	22	21	-1
4	8	14	16	+2
9	2	20	20	0
7	1	15	15	0
7	8	23	22	-1

```
Z = 2X + Y                ---> ML Model

Prediction --->   X = 6   Y = 8   Z = ?

if 20 == 19:              95%
```

Training a dataset

Observations → **Data** → **Model**

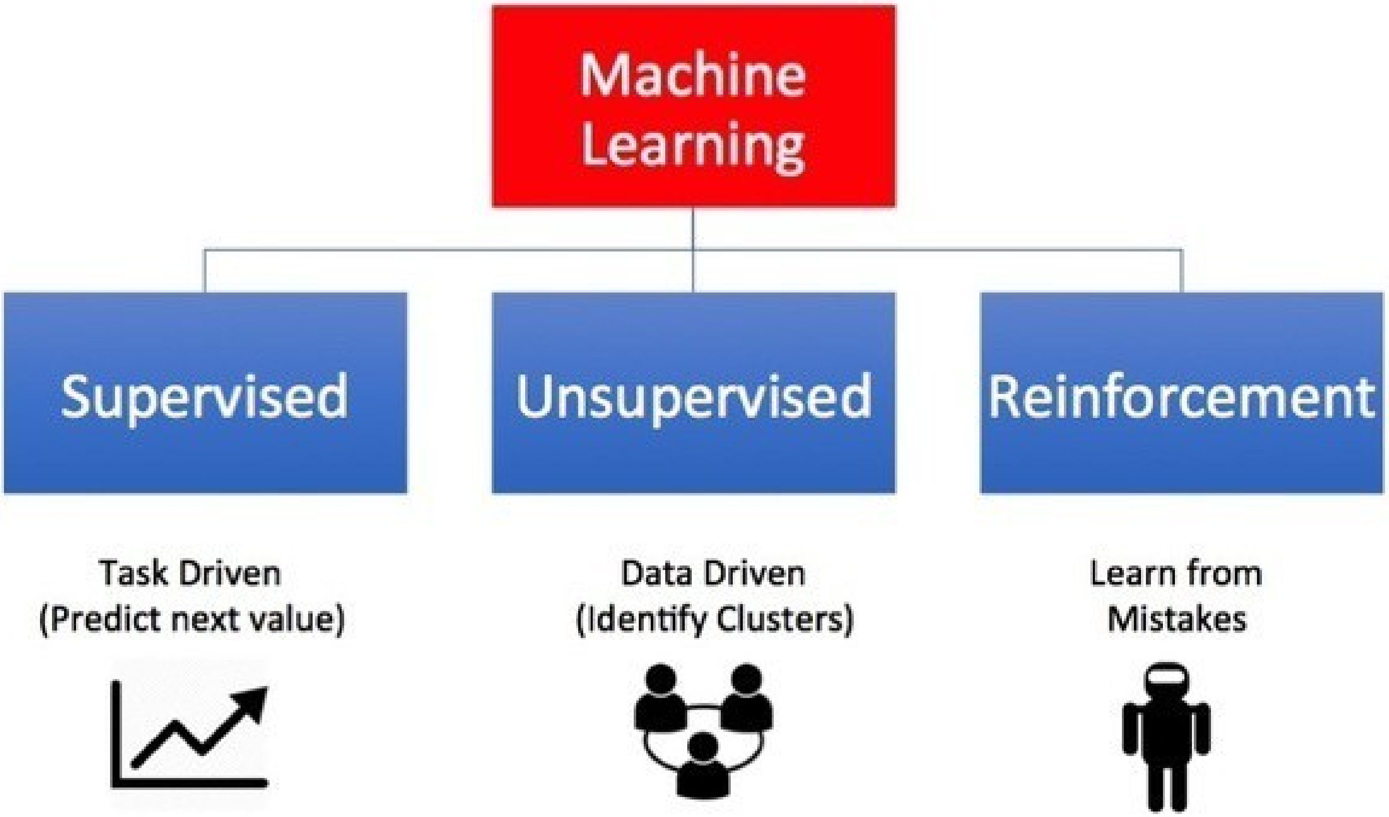




velocity	time
9.8	1
39.2	2
88.2	3
156.8	4
245	5

$$g = 9.8\ m/s^2$$

Types of Machine Learning



Machine Learning

Supervised

Unsupervised

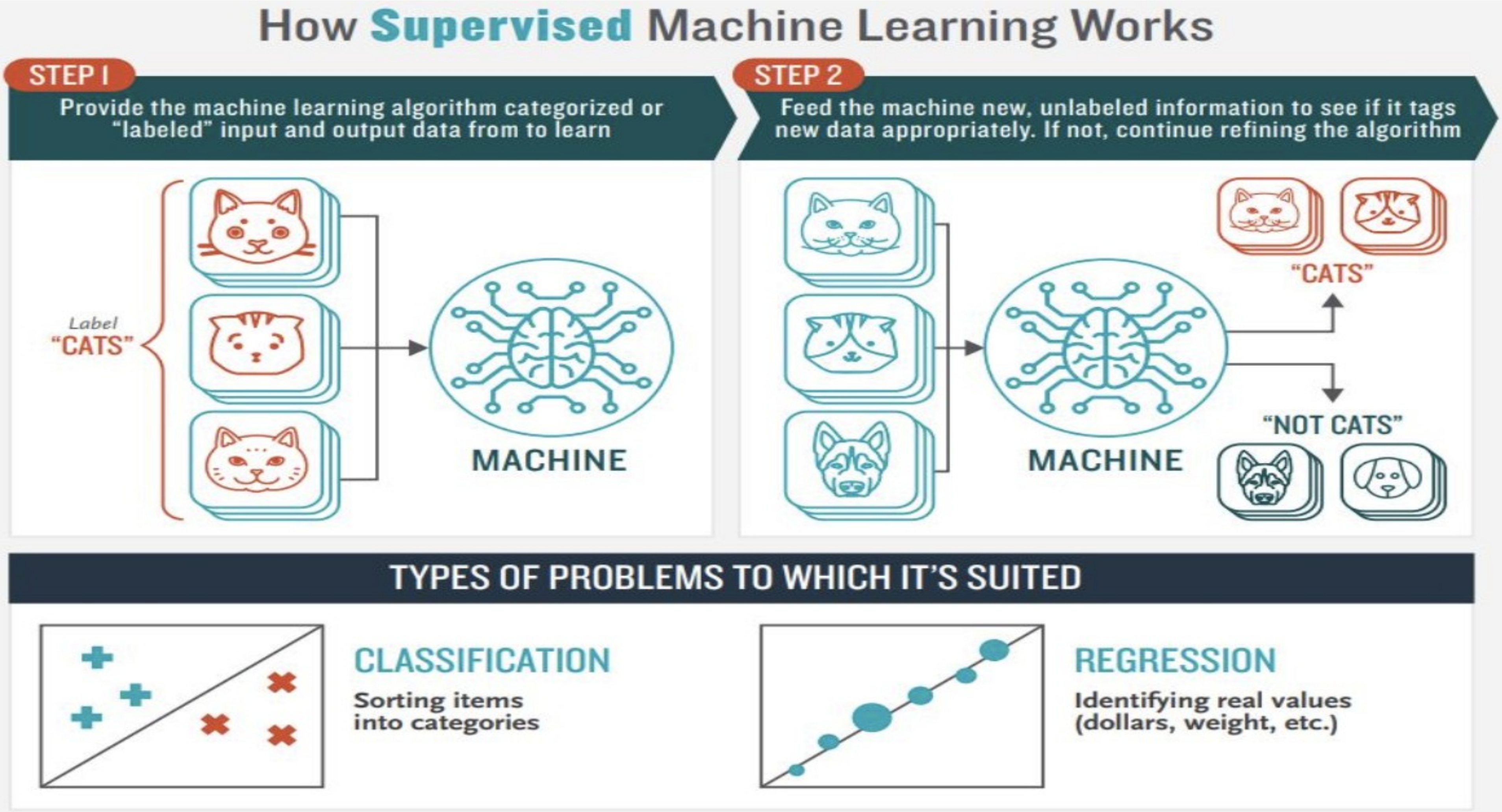
Reinforcement

Task Driven
(Predict next value)

Data Driven
(Identify Clusters)

Learn from
Mistakes

Supervised Machine Learning



How Supervised Machine Learning Works

STEP 1
Provide the machine learning algorithm categorized or "labeled" input and output data from to learn

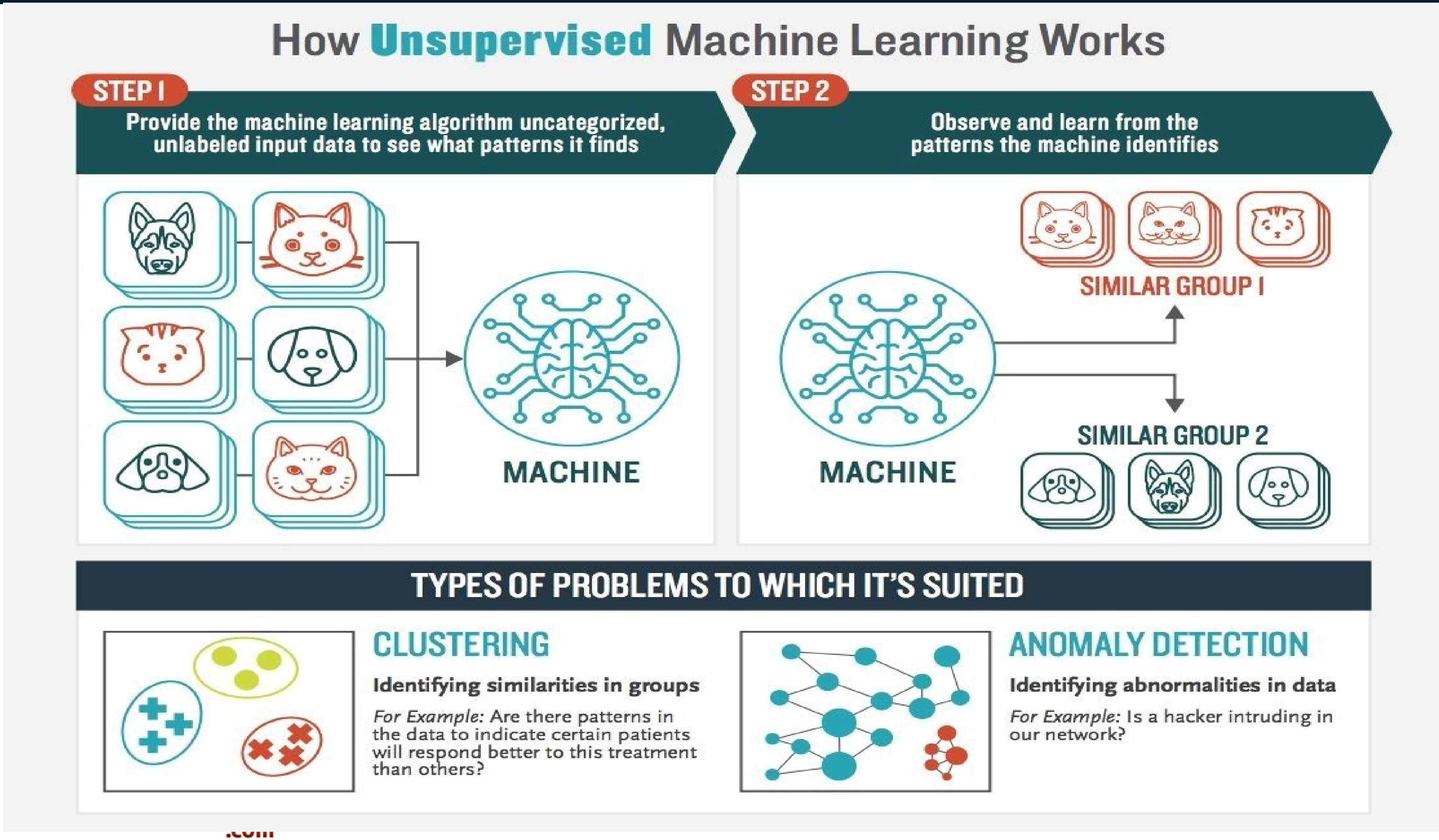
STEP 2
Feed the machine new, unlabeled information to see if it tags new data appropriately. If not, continue refining the algorithm

TYPES OF PROBLEMS TO WHICH IT'S SUITED

CLASSIFICATION
Sorting items into categories

REGRESSION
Identifying real values (dollars, weight, etc.)

Unsupervised Machine Learning



How Unsupervised Machine Learning Works

STEP 1
Provide the machine learning algorithm uncategorized, unlabeled input data to see what patterns it finds

STEP 2
Observe and learn from the patterns the machine identifies

TYPES OF PROBLEMS TO WHICH IT'S SUITED

CLUSTERING

Identifying similarities in groups

For Example: Are there patterns in the data to indicate certain patients will respond better to this treatment than others?

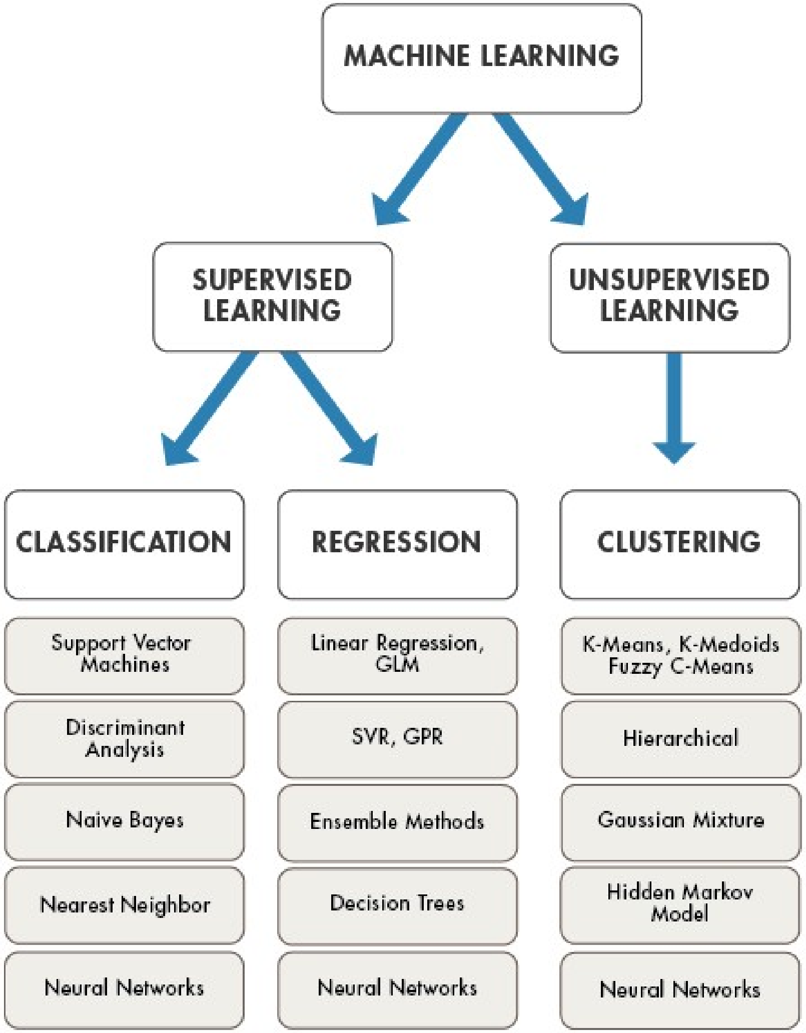
ANOMALY DETECTION

Identifying abnormalities in data

For Example: Is a hacker intruding in our network?

.com

Typical Algorithms in Machine Learning



MACHINE LEARNING

SUPERVISED LEARNING

UNSUPERVISED LEARNING

CLASSIFICATION	REGRESSION	CLUSTERING
Support Vector Machines	Linear Regression, GLM	K-Means, K-Medoids Fuzzy C-Means
Discriminant Analysis	SVR, GPR	Hierarchical
Naive Bayes	Ensemble Methods	Gaussian Mixture
Nearest Neighbor	Decision Trees	Hidden Markov Model
Neural Networks	Neural Networks	Neural Networks

Spark MLlib

- MLlib is nothing but a machine learning (ML) library of Apache Spark.
- Basically, it helps to make practical machine learning scalable and easy. Moreover, it provides the following ML Algorithms:
  - Basic statistics
  - Classification and Regression
  - Clustering
  - Collaborative filtering

Spark MLlib

- Apache Spark offers a Machine Learning API called MLlib. PySpark has this machine learning API in Python as well. It supports different kind of algorithms, which are mentioned below –
- mllib.classification – The spark.mllib package supports various methods for binary classification, multiclass classification and regression analysis. Some of the most popular algorithms in classification are Random Forest, Naive Bayes, Decision Tree, etc.
- mllib.clustering – Clustering is an unsupervised learning problem, whereby you aim to group subsets of entities with one another based on some notion of similarity.

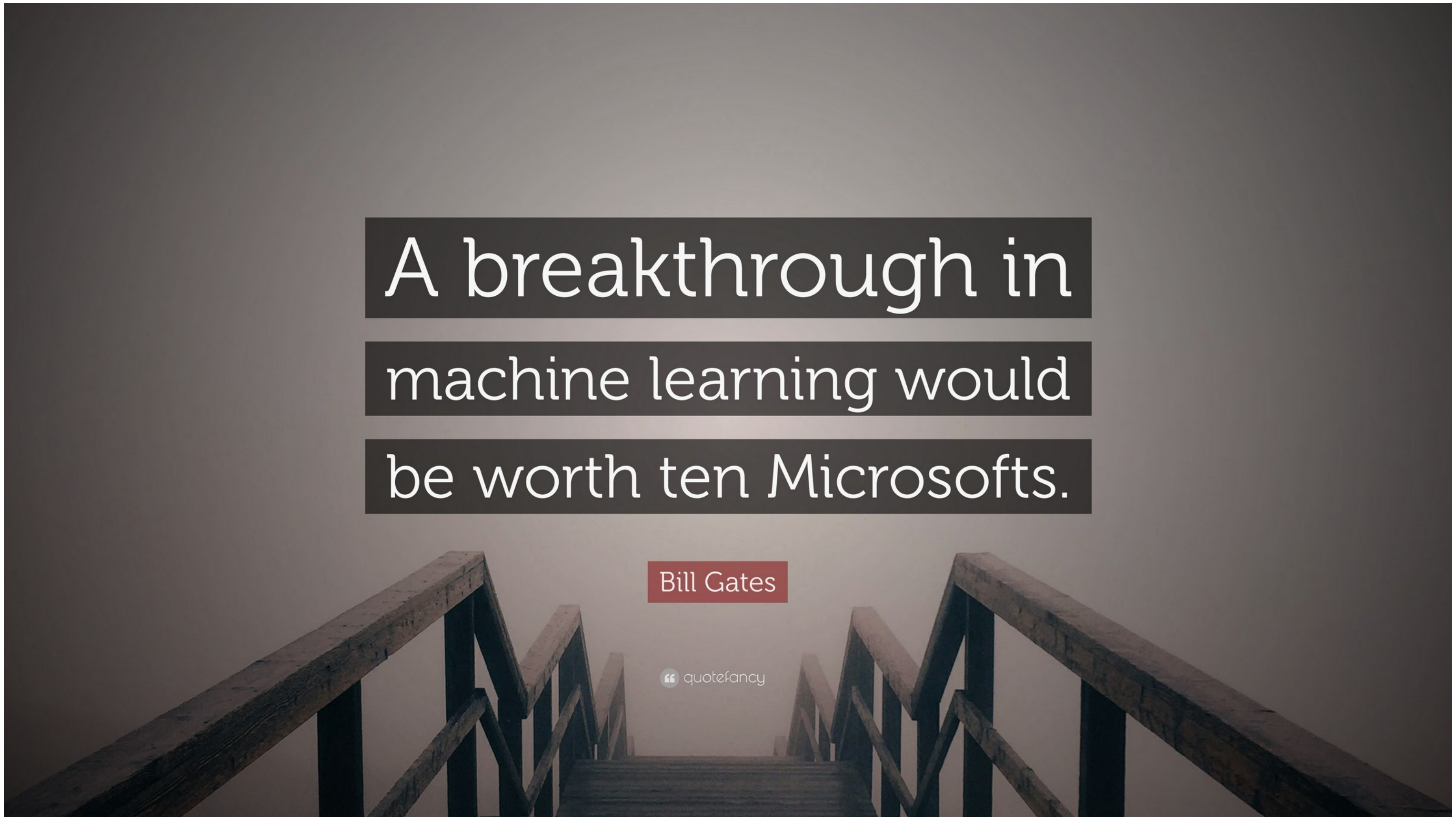
Spark MLlib

- spark.mllib – It ¬currently supports model-based collaborative filtering, in which users and products are described by a small set of latent factors that can be used to predict missing entries. spark.mllib uses the Alternating Least Squares (ALS) algorithm to learn these latent factors.
- mllib.regression – Linear regression belongs to the family of regression algorithms. The goal of regression is to find relationships and dependencies between variables. The interface for working with linear regression models and model summaries is similar to the logistic regression case.

Spark MLlib

- mllib.fpm – Frequent pattern matching is mining frequent items, itemsets, subsequences or other substructures that are usually among the first steps to analyze a large-scale dataset. This has been an active research topic in data mining for years.
- mllib.linalg – MLlib utilities for linear algebra.
- mllib.recommendation – Collaborative filtering is commonly used for recommender systems. These techniques aim to fill in the missing entries of a user item association matrix.

Conclusion



A breakthrough in machine learning would be worth ten Microsofts.

Bill Gates

Useful web resources

- www.mitu.co.in
- www.pythonprogramminglanguage.com
- www.scikit-learn.org
- www.towardsdatascience.com
- www.medium.com
- www.analyticsvidhya.com
- www.kaggle.com
- www.stephacking.com
- www.github.com

Thank you

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com