Machine Learning

Introduction

Parcours Progis

Etudes, Medias, communication, Marketing

UGA
Université
Grenoble Alpes

Sciences Po
Grenoble UGA

Course Information

- **Bahareh Afshinpour**
- **Email:** bahareh.afshinpour@univ-grenoble-alpes.fr
- **Home page:** http://afshinpour.com/
- **Office:** IMAG building -246
- **Course time**
  - **1h30 TD**
  - **1h30 TP**

Textbook and References

- **Ethem Alpaydin, ”Introduction to Machine Learning”, MIT Press, Prentice Hall of India, 3rd 2014.**
- **Neural Networks for Pattern Recognition by C. Bishop**
- **Python Machine Learning by S. Raschka and V. Mirjalili (highly recommended).**
- **YouTube videos.**

Skills you will acquire

After completing this course, you will be able to:

✓ explain, compare, and contrast various machine learning topics and concepts like supervised learning, unsupervised learning, classification, regression, and clustering.

✓ You will also be able to describe how the various machine learning algorithms work.

✓ And finally, you will learn how to apply these machine learning algorithms in Python using various Python libraries.

Grading

- TP (Homeworks- Report) 10%
- Mid-Term Exams 25%
- Course Project 25%
  - To enable the students to get hands-on experience in the design, implementation and evaluation of machine learning algorithms.
  - Teams: 1-2 students
  - Solve a practical machine learning problem of your choice.
  - Use Python language.
  - Good projects involve using multiple learning algorithms and evaluating their performance in solving the problem. And should use preprocessing and feature extraction and selection.
- Final Exam 40%

Policies

- Attendance is required.
- All submitted work must be yours.

- Protect your efforts! Don’t let others see your codes, don’t give others your results.
  - Lending your codes to others or allowing others to copy your work will be considered as collusion, thus receiving the same punishment as the plagiarist.

- Your codes need to be able to generate the results.
- This course requires significant effort.

Important Dates

Février	**Midterm Exam**
Avril/Mai	**TP Report Due**
Juin	**Final Exam**
Juin	**Project Final Report Due**

Computing needs

- **Basic Desktop or Laptop**

- **Ability to :**
  - **Install and run Python + notepad++**
    - » (Python offers a large number of packages that make it simple to get started and build applications of all complexity levels.)
  - **Jupyter Notebook**
    - » **(**The Jupyter Notebook is a simple interactive environment for programming with Python, which makes it really easy to share your results.**)**

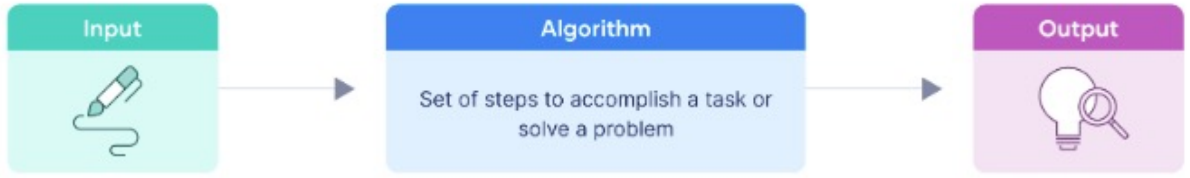
Introduction

- **This is the age of "big data"**
  - **we all became producers of data.**
    - ✓ Every time we buy a product, write a blog, or post on the social media, we are generating data.
  - **Each of us is not only a generator but also a consumer of data.**
    - ✓ We want to have products and services specialized for us.
    - ✓ We want our needs to be understood and interests to be predicted.
    - ✓ There are certain patterns in the data.

https://www.lebigdata.fr/definition-big-data

Introduction- Algorithm

- To solve a problem on a computer, we need an algorithm.



https://www.scribbr.com/ai-tools/what-is-an-algorithm/

- An algorithm is a sequence of instructions that should be carried out to transform the input to output.
  - ✓ For example, one can devise an algorithm for sorting. The input is a set of numbers and the output is their ordered list.

- For some tasks, however, we do not have an algorithm.
  - ✓ Predicting customer behavior is one;
  - ✓ another is to tell spam emails from legitimate ones.
    - what we want is to “learn” what constitutes spam from them.
    - In other words, we would like the computer(machine) to extract automatically the algorithm for this task.

any piece of software that will consume training examples,
in order to make decisions over **unseen** data
without explicit programming is considered
learning.

Introduction- Machine Learning application

- **Machine Learning application areas are abundant:**
  - **In finance banks**
    - Fraud detection
    - Credit approval
    - Price and market prediction
  - **In manufacturing,**
    - learning models are used for optimization, control, and troubleshooting.
  - **In medicine,**
    - drug discovery
    - computational genomics (analysis and design)
    - medical imaging and diagnosis
  - **In telecommunications,**
    - call patterns are analyzed for network optimization and maximizing the quality of service.
  - **Computer vision and robotics:**
    - detection, recognition and categorization of objects
    - face recognition
    - tracking objects (rigid and articulated) in video
  - **In science,**
    - large amounts of data in physics, astronomy, and biology can only be analyzed fast enough by computers.

Introduction- What Is Machine Learning?

- Machine learning is

  the process of teaching computers how to improve performance by using example data or information from the past.

The real question is what is learning?

Using past experiences to improve future performance.

For a machine, experiences come in the form of data.

Introduction- What Is Machine Learning?

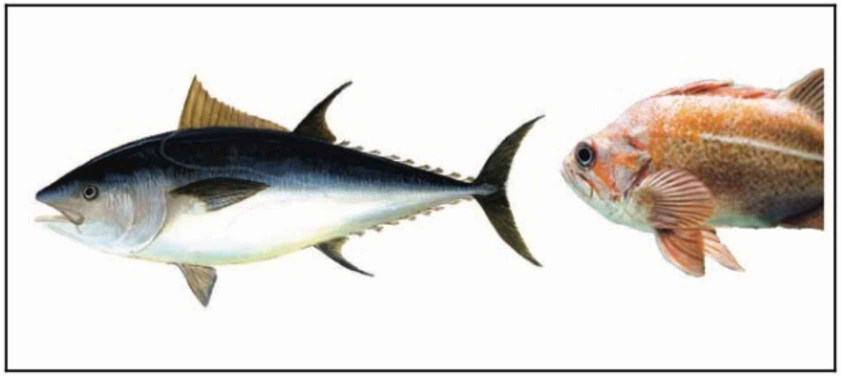
- We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience.
- The model may be predictive to make predictions in the future, or descriptive to gain knowledge from data, or both.

Consider a real example

- The Nature Conservancy is working with other fishing companies and partners to monitor fishing activities and preserve fisheries for the future.
- So they are looking to use cameras in the future to scale up this monitoring process.
- The amount of data that will be produced from the deployment of these cameras will be cumbersome and very expensive to process manually. So the conservancy wants to develop a learning algorithm to automatically detect and classify different species of fish to speed up the video reviewing process.

Consider a real example

- So our aim in this example is to separate different species such as tunas, sharks, and more that fishing boats catch. As an illustrative example, we can limit the problem to only two classes, tuna and opah.



- Now, we can take a sample of some random images from our collection and start to note some physical differences between the two types. For example, consider the following physical differences:
  - Length: You can see that compared to the opah fish, the tuna fish is longer
  - Width: Opah is wider than tuna
  - Color: You can see that the opah fish tends to be more red while the tuna fish tends to be blue and white, and so on
- We can use these physical differences as features that can help our learning algorithm(classifier) to differentiate between these two types of fish.

Consider a real example

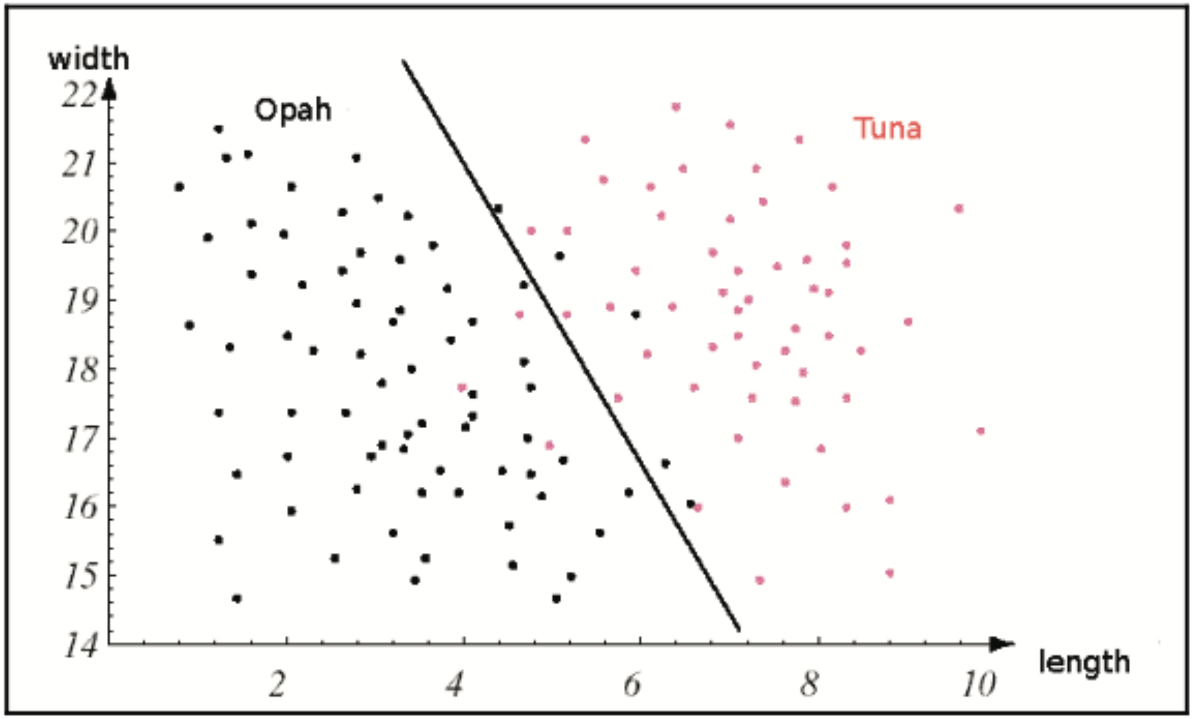
- In this case, the collection of tuna and opah fish will act as the **knowledge base** for our classifier.

- Initially, the knowledge base (training samples) will be labeled/tagged, and for each image, you will know beforehand whether it's tuna or opah fish.
- So the classifier will use these training samples to model the different types of fish, and then we can use the output of the training phase to automatically label unlabeled/untagged fish that the classifier didn't see during the training phase.

- This kind of unlabeled data is often called unseen data.

Consider a real example

- if we combine both features, we might get something that looks like the following graph:



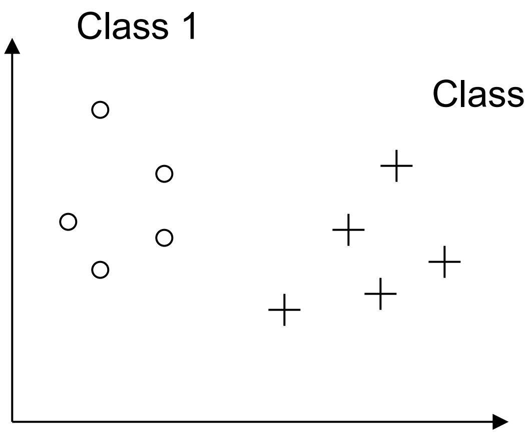
Feature

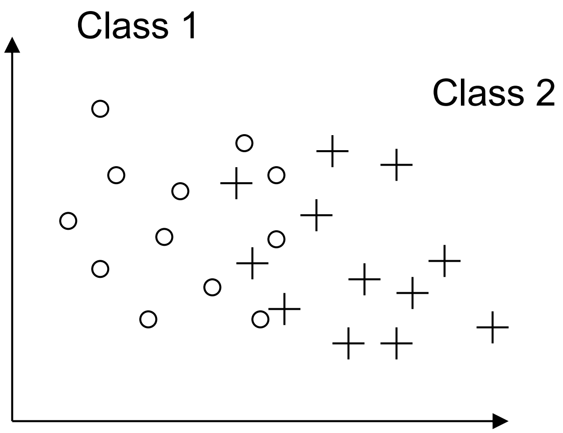
- Features are numerically expressed properties of the signal.

- The set of features used for pattern recognition is called feature vector.

- The number of used features is the dimensionality of the feature vector.

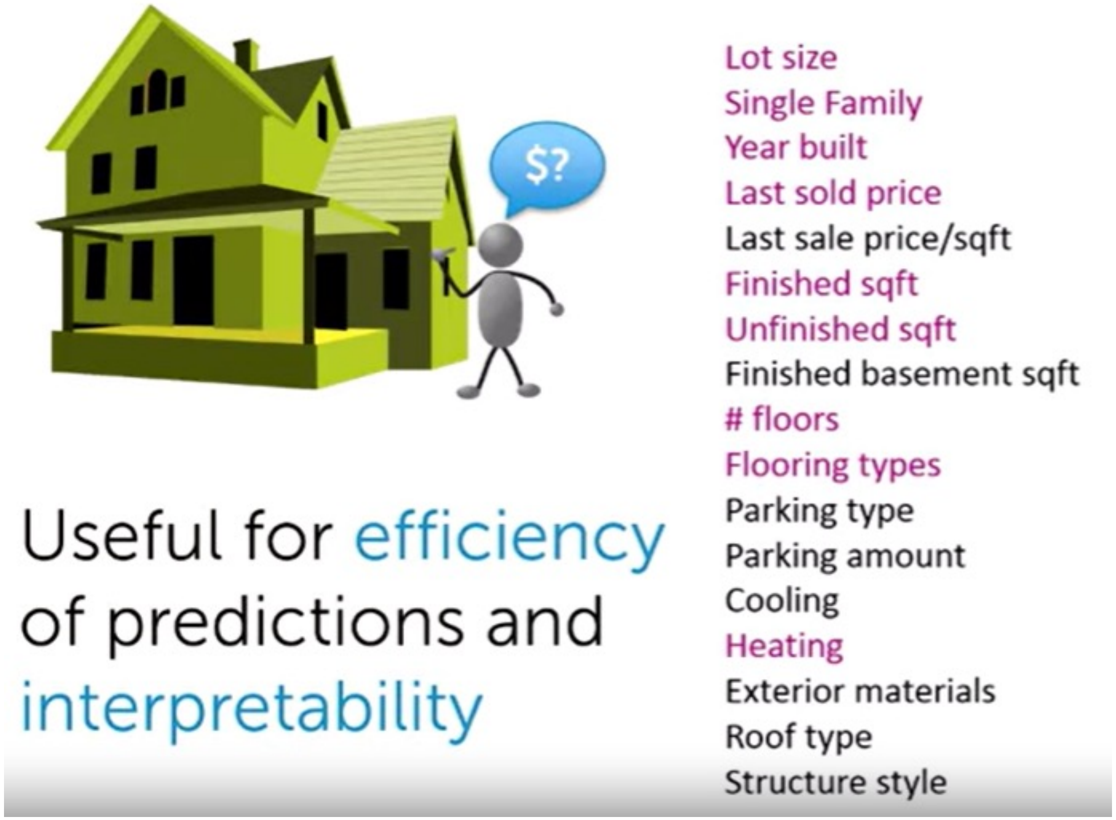
- n-dimensional feature vectors can be represented as points in n-dimensional feature space.





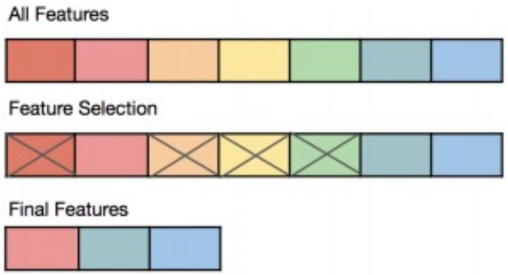
Feature selection

- For example **having macrowave** is not very help ful feature for buying house



Useful for efficiency of predictions and interpretability

Lot size
Single Family
Year built
Last sold price
Last sale price/sqft
Finished sqft
Unfinished sqft
Finished basement sqft
# floors
Flooring types
Parking type
Parking amount
Cooling
Heating
Exterior materials
Roof type
Structure style



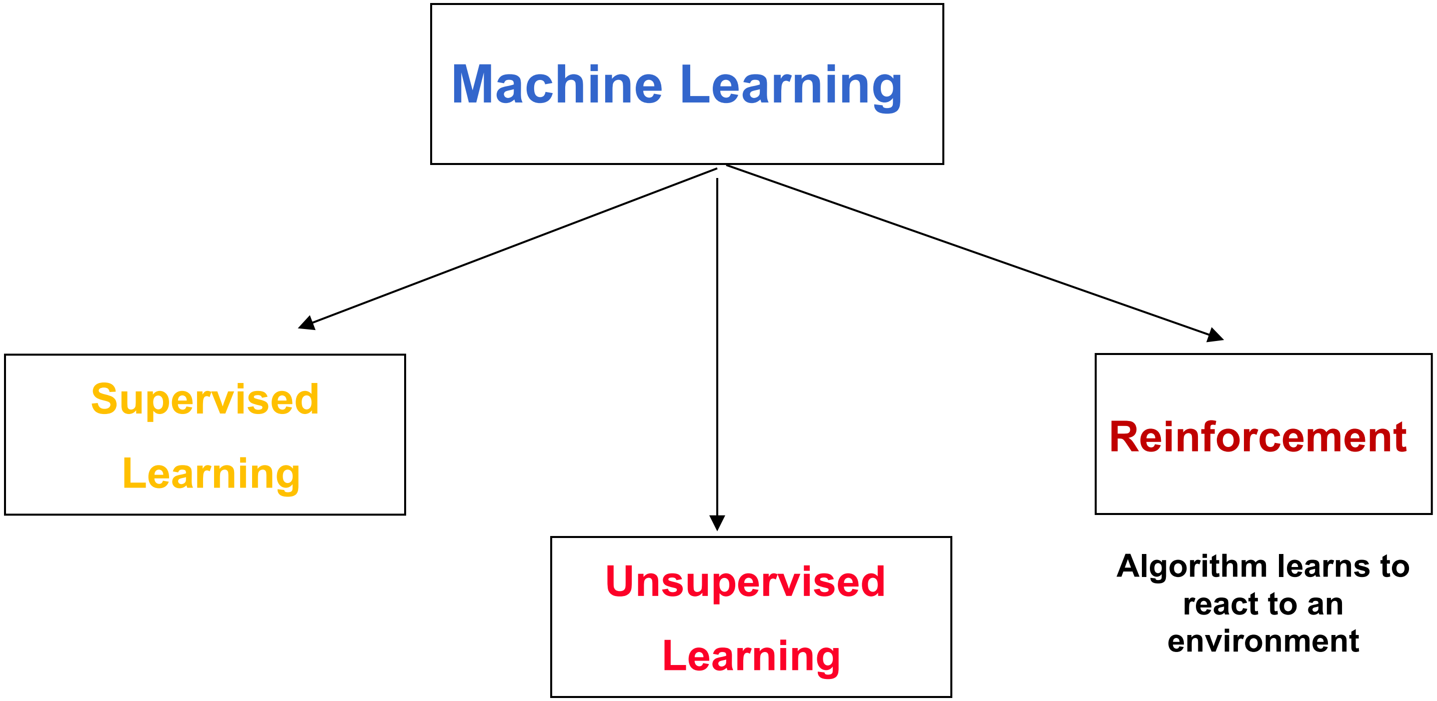
https://vitalflux.com/machine-learning-feature-selection-feature-extraction/

Feature extraction

But Feature **extraction** is about extracting/deriving information from the original features set to create a new features subspace.

-Dimensionality Reduction

About the course



- We'll cover the most widely used machine learning methods such as
  - ✓ Regression method
  - ✓ Classification methods
  - ✓ Clustering methods
  - ✓ random forests
  - ✓ neural nets, and so on.

Supervised Learning Problem

Supervised Learning Problem

How do we supervise a machine learning model?

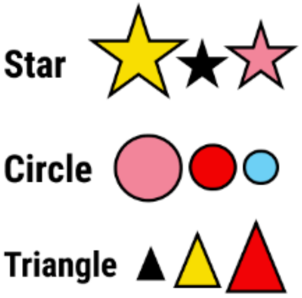
We do this by teaching the model

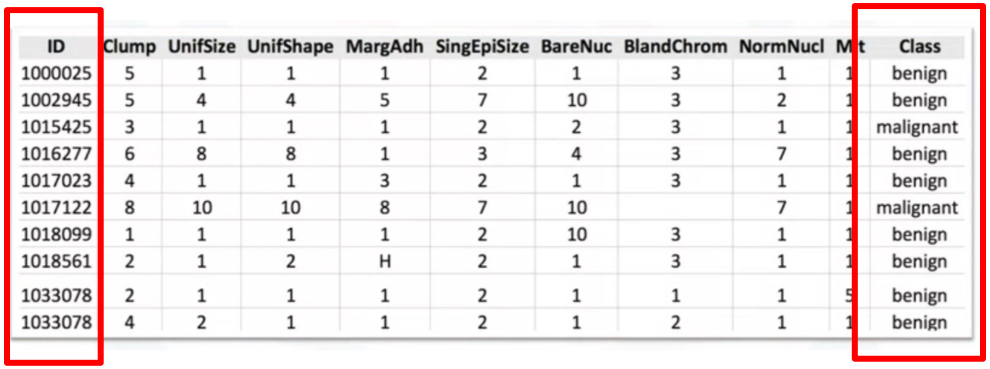
How exactly do we teach a model?

We teach the model by training it with some data from a labeled dataset

what does a labeled dataset look like?

It could look something like this





ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	M t	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	malignant
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		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

Supervised Learning Problem

- In supervised learning,
  - we are given a data set and already know what our right output should be
  - with the assumption that there is a relationship between the input and the result
  - you have to study in relation with a different set of features called independent features.
- Two types of supervised learning tasks:
  - "Regression"
  - "classification“

Regression	Classification
Target: Continuous data	Target: Categorical data
Aim: Predict the value	Aim: Predict class/class probability

How does a supervised problem look?

X1	X2	X3	X4	X5	Y1	Y2
7	1	3	1	2	50	0
2	4	7	12	9	24	1
5	9	1	10	6	16	0
8	11	2	3	4	62	1
1	5	5	7	1	12	0

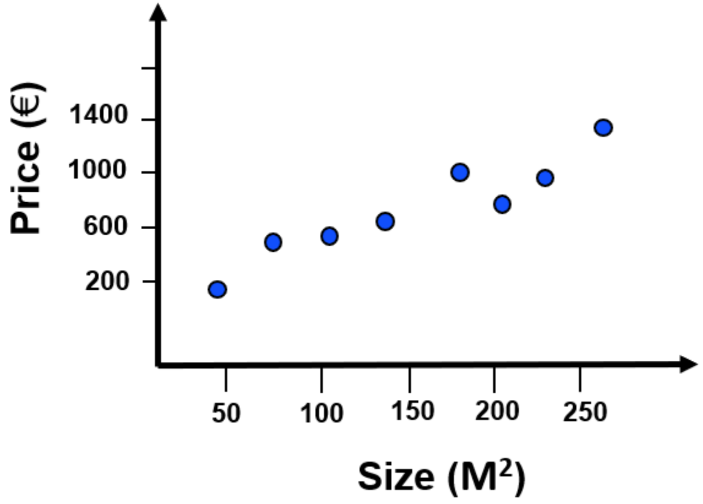
- x1,x2,x3,x4 are independent features
- Y1 and Y2 are dependent features
- Y1 is continuous
- Y2 is categorical

The outcome of a supervised learning problem is to have a dependent feature defined as a function of independent feature.

For example Y1=f(X1,X2,X3,X4)

Example of a Supervised Learning algorithm

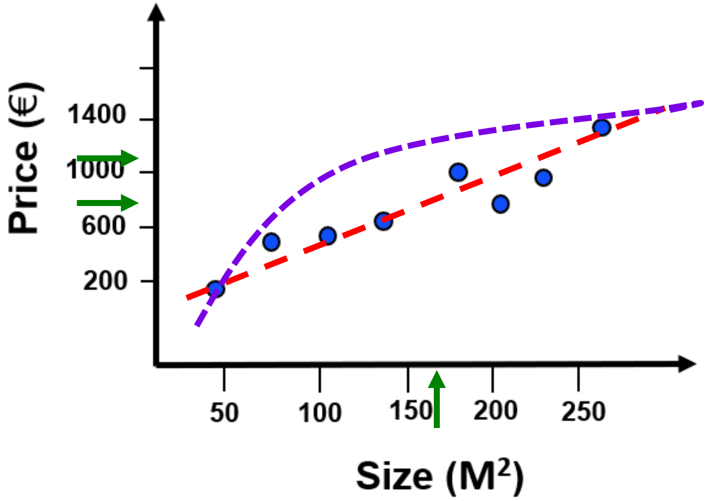
- Given data about the size of houses on the real estate market, try to predict their price.
- Price as a function of size is a **continuous** output, so this is a regression problem.



- Imagine you have a friend who owns a house that is say 170 square meter, and they are hoping to rent the house, and they want to know how much they can get for the house.
- how can the learning algorithm help them?

Example

- Put a straight line through the data, also fit a straight line to the data. It looks like maybe their house can be sold for about $700. But maybe this isn't the only learning algorithm you can use, and there might be a better one.



- Instead of fitting a straight line to the data, we might decide that it's better to fit a second-order polynomial to this data. then it looks like, well, maybe they can sell the house 800.

What is regression?

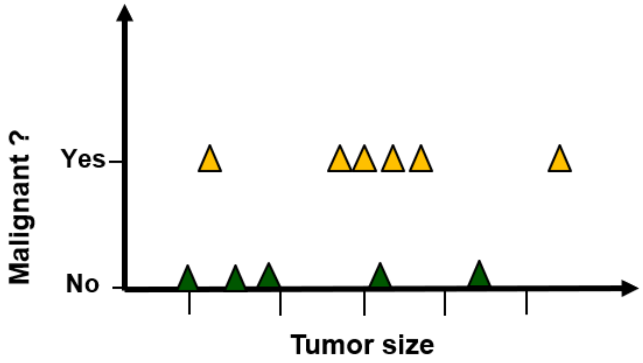
Regression is the process of predicting continuous values.

What is Classification?

Is the process of predicting discrete class labels or categories.

Examples

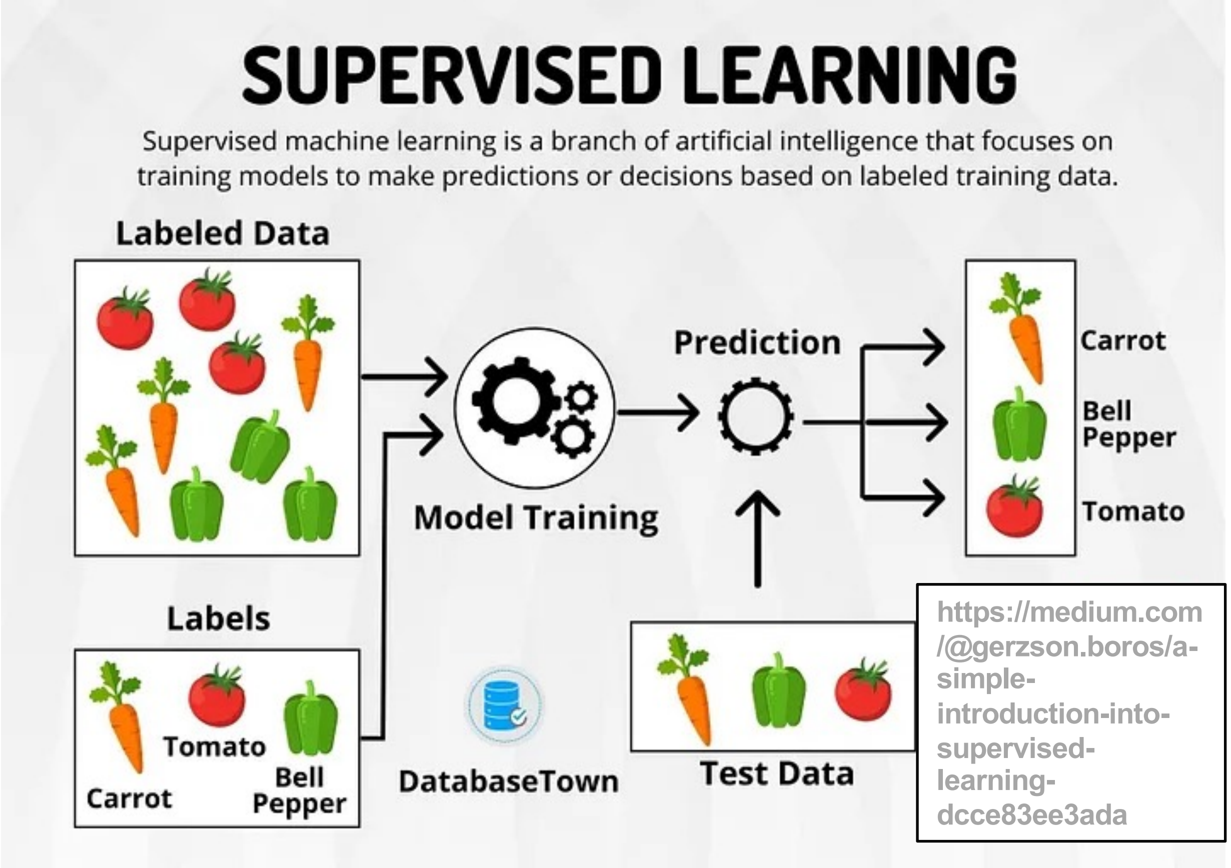
- Given a patient with a tumor, we have to predict whether the tumor is malignant or benign.
- a malignant tumor
  - is a tumor that is harmful and dangerous
- a benign tumor
  - is a tumor that is harmless



The Machine Learning question is, can you estimate what is the probability, what's the chance that tumor X as malignant versus benign?

- For some learning problems what you really want is not to use like two or three features
  - For example, the age of the patient and the size of the tumor are two features
  - instead you want to use an infinite number of features

How do you deal with an infinite number of features?



SUPERVISED LEARNING

Supervised machine learning is a branch of artificial intelligence that focuses on training models to make predictions or decisions based on labeled training data.

https://medium.com/@gerzson.boros/a-simple-introduction-into-supervised-learning-dcce83ee3ada

Example of Supervised Learning

Suppose you're running a company and you want to develop learning algorithms to address each of two problems.

1.In the first problem, you have a large inventory of identical items. So, imagine that you have thousands of copies of some identical items to sell, and you want to predict how many of these items you sell over the next three months.

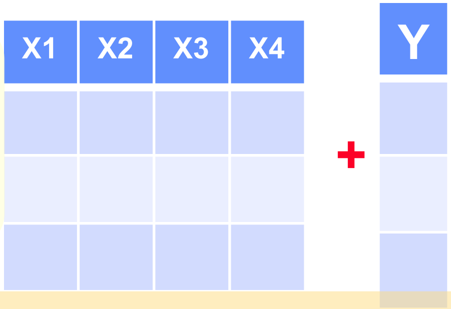
2.In the second problem, problem two, you have lots of users, and you want to write software to examine each individual of your customer's accounts, so each one of your customer's accounts. For each account, decide whether or not the account has been hacked or compromised.

Should you treat these as classification or as regression problems?

- For problem one, I would treat this as a regression problem because if I have thousands of items, I would probably just treat this as a real value, as a continuous value.
- For the second problem, I would treat that as a classification problem, because I might say set the value I want to predict with zero to denote the account has not been hacked, and set the value one to denote an account that has been hacked into. So, I might set this be zero or one depending on whether it's been hacked, and have an algorithm try to predict each one of these two discrete values.

Implementing Supervised Learning Model

1. Extract Data
2. Join data with Target



3. Data cleaning

4. Univariate and Bi-Variate Analysis

5. Select the Bests features

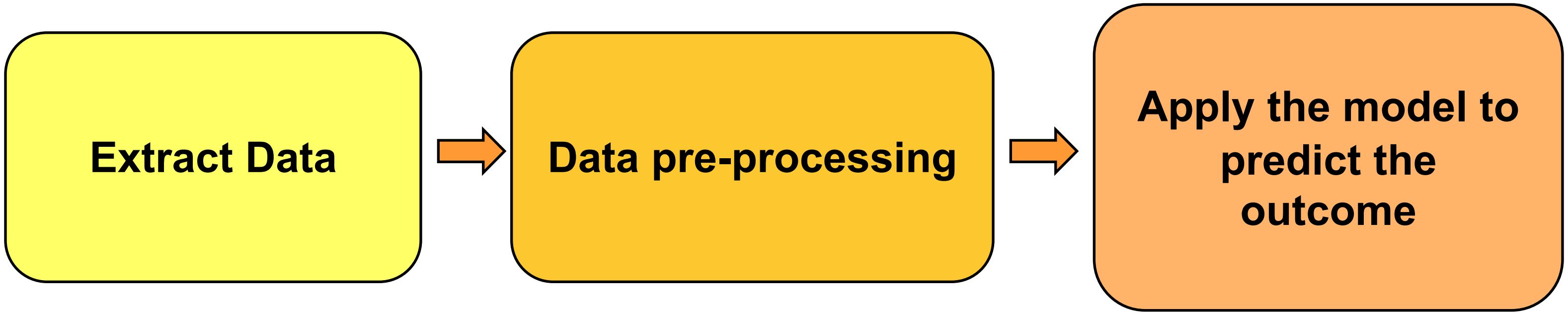
6. Data Split (Building training and testing Samples)

7. Apply appropriate Supervised Algorithm

8. Validate model results

Implementing Supervised Learning Model

- So in general:



Extract Data

Data pre-processing

Apply the model to predict the outcome

What we learned in Supervised learning

- In classification problems the task is to assign new inputs to one of a number of discrete classes or categories.
- However, there are many other tasks, which we shall refer to as *regression* problems, in which the outputs represent the values of continuous variables.

classification

-K-nearest Nieghbors
-Support vector machine
-Naive bayes classifier
-MLP
-Random forest

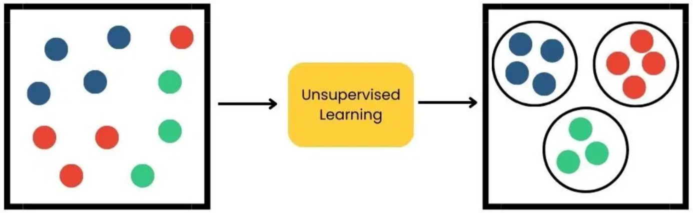
Regression

-Linear regression
-decision tree regression
-logistic regression

Unsupervised Learning Problem

Unsupervised Learning

- Here, we let the model work on its own to discover information that may not be visible to the human eye.
- It means, the unsupervised algorithm trains on the dataset, and draws conclusions on unlabeled data.
- unsupervised learning has more difficult algorithms than supervised learning since we know little to no information about the data.
- Dimension reduction and clustering are the most widely used unsupervised machine learning techniques.

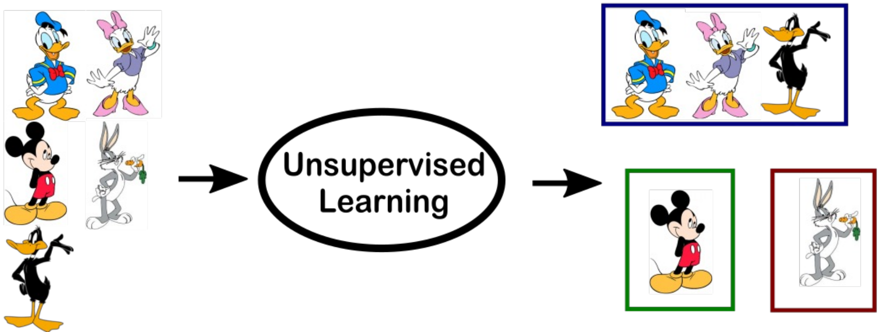


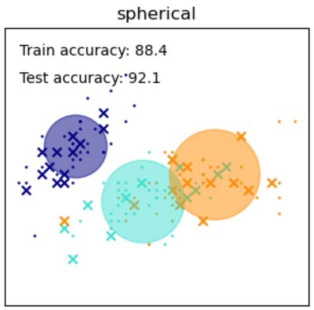
https://www.bombaysoftwares.com/blog/introduction-to-unsupervised-learning

Unsupervised Learning

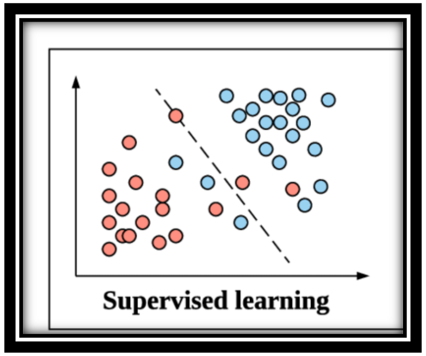
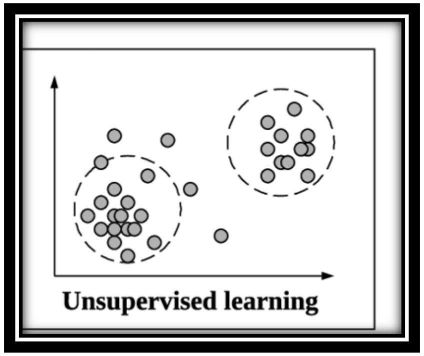
- Here we only talk about grouping and there is no prediction aim.
- Unsupervised learning algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points.
- The lack of the target variable defines an unsupervised problem.

X1	X2	X3	X4	X5	Cluster ID
7	1	3	1	2	1
2	4	7	12	9	1
5	9	1	10	6	2
8	11	2	3	4	3
1	5	5	7	1	3



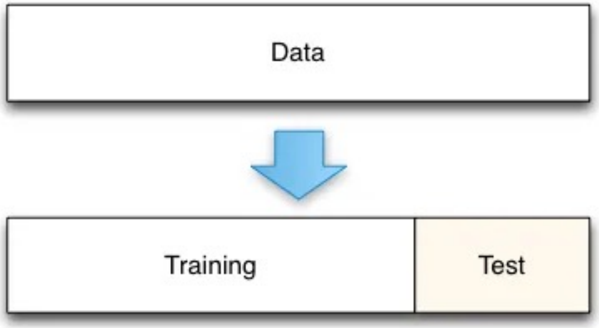


Supervised Vs Unsupervised

Supervised Learning	Unsupervised Learning
	
Our model should effectively classify observations into either success or failure classes based on the features that we have.	Good clusters should capture similar observations within and capture different observations across clusters.

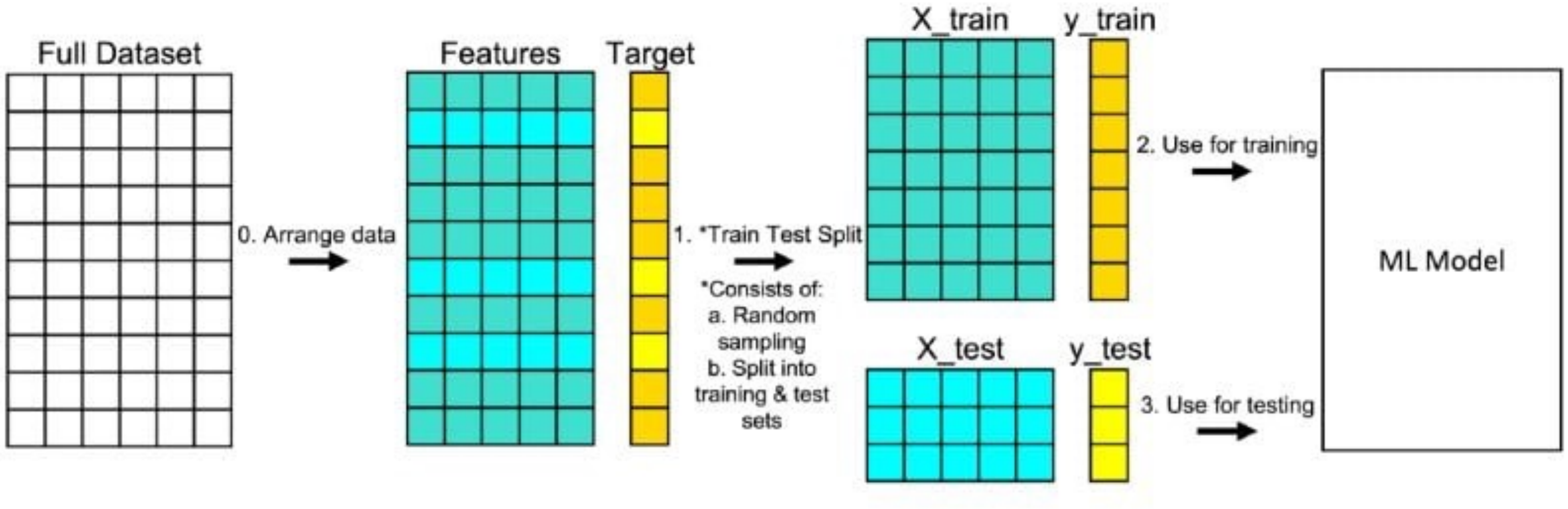
Training and testing

- Universal set population
- Training set
  - Build a model only on a portion of the population.
- Testing set
  - How the model actually behaves on unobserved data.



https://medium.com/analytics-vidhya/only-train-and-test-set-is-not-enough-for-generalizing-ml-model-significance-of-validation-set-cf68bb26881a

Three parts to machine learning
First is get the data, second is build your model on the training set and third is to validate the performance of the model on the testing set.

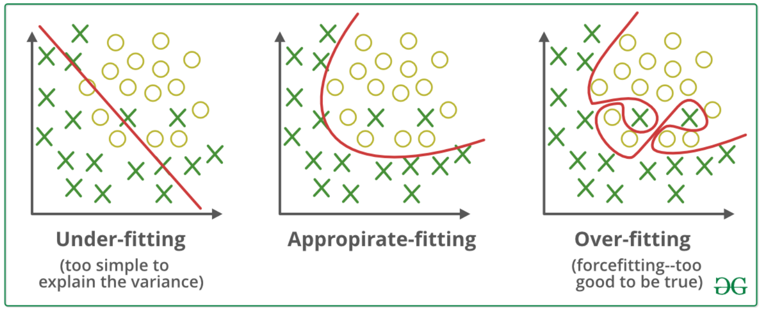


https://builtin.com/data-science/train-test-split

Problem with Model Fit

- Underfitting
  - Underfitting indicates that the model isn't working as expected.
  - The Model does not know enough about the training data.
  - We do not get strong performance in training data.

- Overfitting
  - The model knows almost everything that there is to know about the training data.
  - The model shows excellent results on the training data
  - Level of complexity will be Extremely high.
  - It displays poor results during the testing process.

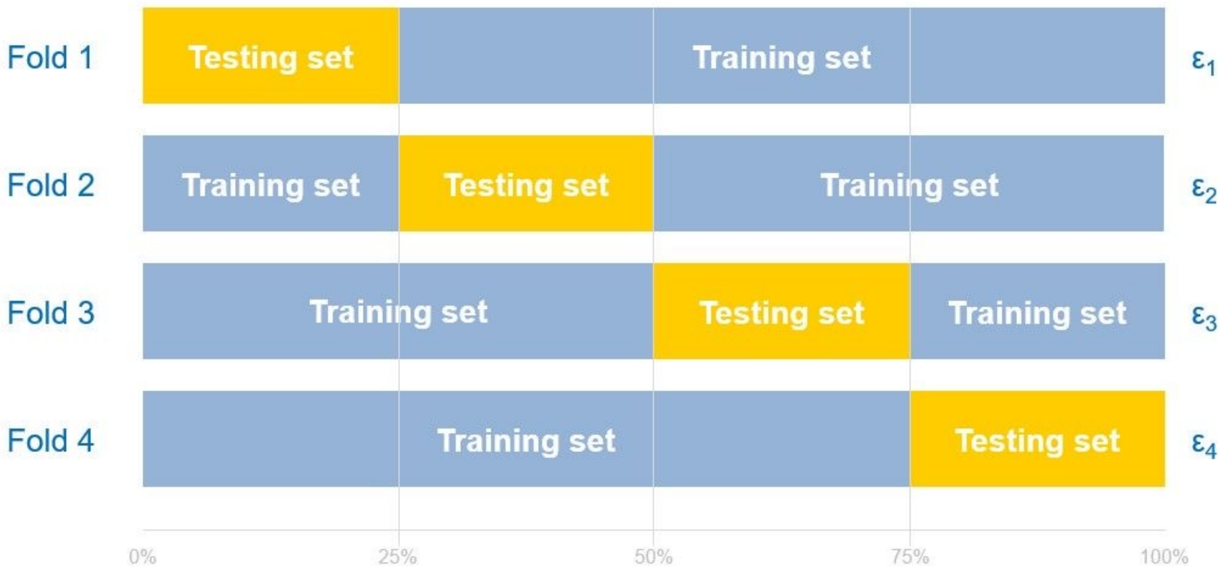

Under-fitting
(too simple to explain the variance)
Appropirate-fitting
Over-fitting
(forcefitting--too good to be true)

https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

Having trained models that work perfectly over the training samples but fail to perform well over the testing samples is called overfitting.

Cross validation

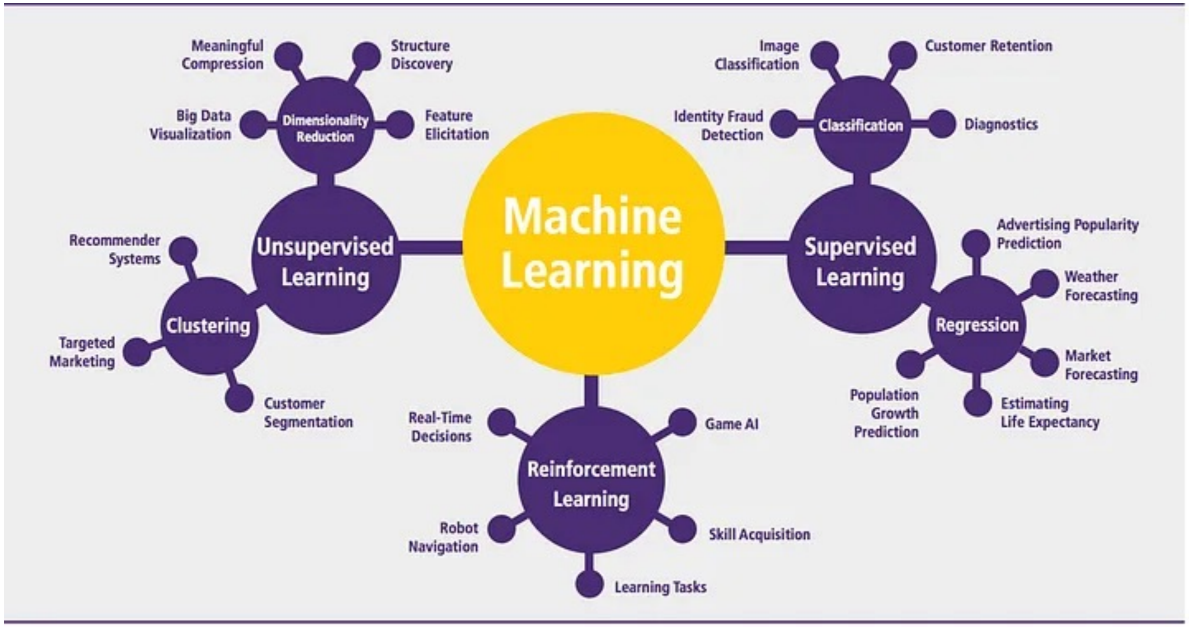
4-fold validation (k=4)



- Cross-validation is a model assessment technique used to evaluate a ML algorithm’s performance in making predictions on new datasets that it has not been trained on.
- Each round of cross-validation involves randomly partitioning the original dataset into a training set and a testing set.

What is reinforcement learning?

-Reinforcement learning (RL) is a machine learning (ML) technique that trains software to make decisions to achieve the most optimal results.
-It mimics the trial-and-error learning process that humans use to achieve their goals. Software actions that work towards your goal are reinforced, while actions that detract from the goal are ignored.



https://medium.com/analytics-vidhya/which-machine-learning-algorithm-should-you-use-by-problem-type-a53967326566

Semi-supervised learning

- Semi-supervised learning is a type of learning that sits in between supervised and unsupervised learning, where you have got training examples with input variables (X), but only some of them are labeled/tagged with the output variable (Y).
- A good example of this type of learning is Flickr, where you have got lots of images uploaded by users but only some of them are labeled (such as sunset, ocean, and dog) and the rest are unlabeled.

Flickr
https://www.flickr.com ⋮

pictures photos on Flickr

Flickr photos, groups, and tags related to the "pictures" Flickr tag.

END

Occam's razor

simpler explanations are more plausible and any unnecessary complexity should be shaved off