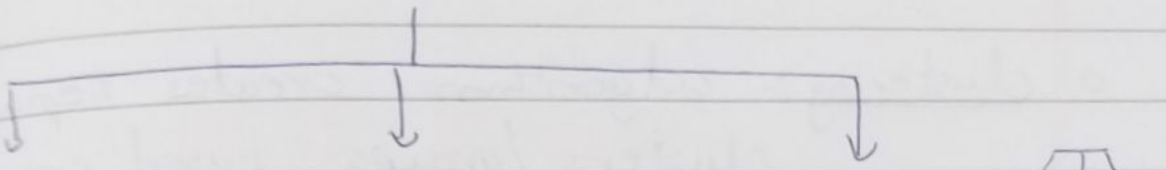
Types of machine learning

3 broad categories



Human Supervision	On the Fly	Comparing or detecting patterns
1. Supervised	1. Online	1. instance-based
2. Unsupervised	2. batch	2. model-based
3. Semi supervised		
4. Reinforcement		

Supervised (Predictive)

- Labeled training dataset
  inputs => Features (or attributes)
  outputs (correct ones) => Labels

| Results depend on quality of training data.

- Two types :-
  1. classification - (spam filter) when o/p variables are categorical.
  2. Regression - Algorithm has to detect relationship b/w 2 or more variables.
     ↳ numerical Features

* Unsupervised
-> The system tries to learn without a teacher.

- clustering :- algorithm creates separate clusters / groups based on common characteristics of the data.

- dimensionality reduction :- (also Feature extraction)

- Anomaly detection :-
Based on normal instance data identifies abnormal.
novelty detection.

- Association rule :-
digs into large amt of data to discover interesting relations b/w attributes.

* Semisupervised (Google photos is ex.)
Clustering.
Both of the above's combination.

* Reinforcement
agent
learns based on rewards and punishment parameters.

* Batch
→ It is not capable of learning incrementally. It must be first trained and then keeps ~~of~~ on working on that learning until updated.

→ Lots of computational power is required.

* Online
→ This can be trained incrementally.
→ data instances are fed sequentially. individually or in mini-batches (grp)

→ Fast & cheap
→ It can be used to train systems on huge data sets that can't fit in the machine's RAM (main memory) this is called out-of-core learning. (its done offline not on the live system)

→ setting up a learning rate is a problem.
→ Big challenge is bad data
↓
→ That's why needs close monitoring.

* Intance-based (lazy learning)

→ Machine learns by heart.
→ They predict based on finding similar examples in the training data.

* Model-based (eager / structure based)

→ This constructs models based on training data.

→ It learns underlying relationships and patterns in the data by creating mathematical representation.

Main challenges of Machine Learning

* Insufficient Quantity of Training data
→ The unreasonable effectiveness of data.
→ Somehow the amount of data turns out to be more important than the algorithm.
→ But data isn't easy or cheap to get.

* Non-representative data
→ Sampling bias issues
→ No response bias.

* Poor Quality data
→ Most of a data scientist's time goes in cleaning the collected data.
→ missing values, errors, incorrect data, etc.

* Irrelevant Features
Feature engineering
↳ Feature selection
↳ Feature extraction (unsupervised learning)
↳ Creating new Feature

Training

* overfitting data
  → Works good on training data but not on testing
  → happens when model is too complex.
  → "Regularization"

* underfitting the Training data
  → opposite of above
  → happens when model is too simple

Testing and Validating

→ Testing is the only practical way to see how good a model is.

→ General practice. 80% of data for training & 20% for testing.
However, it largely depends on the size of dataset.