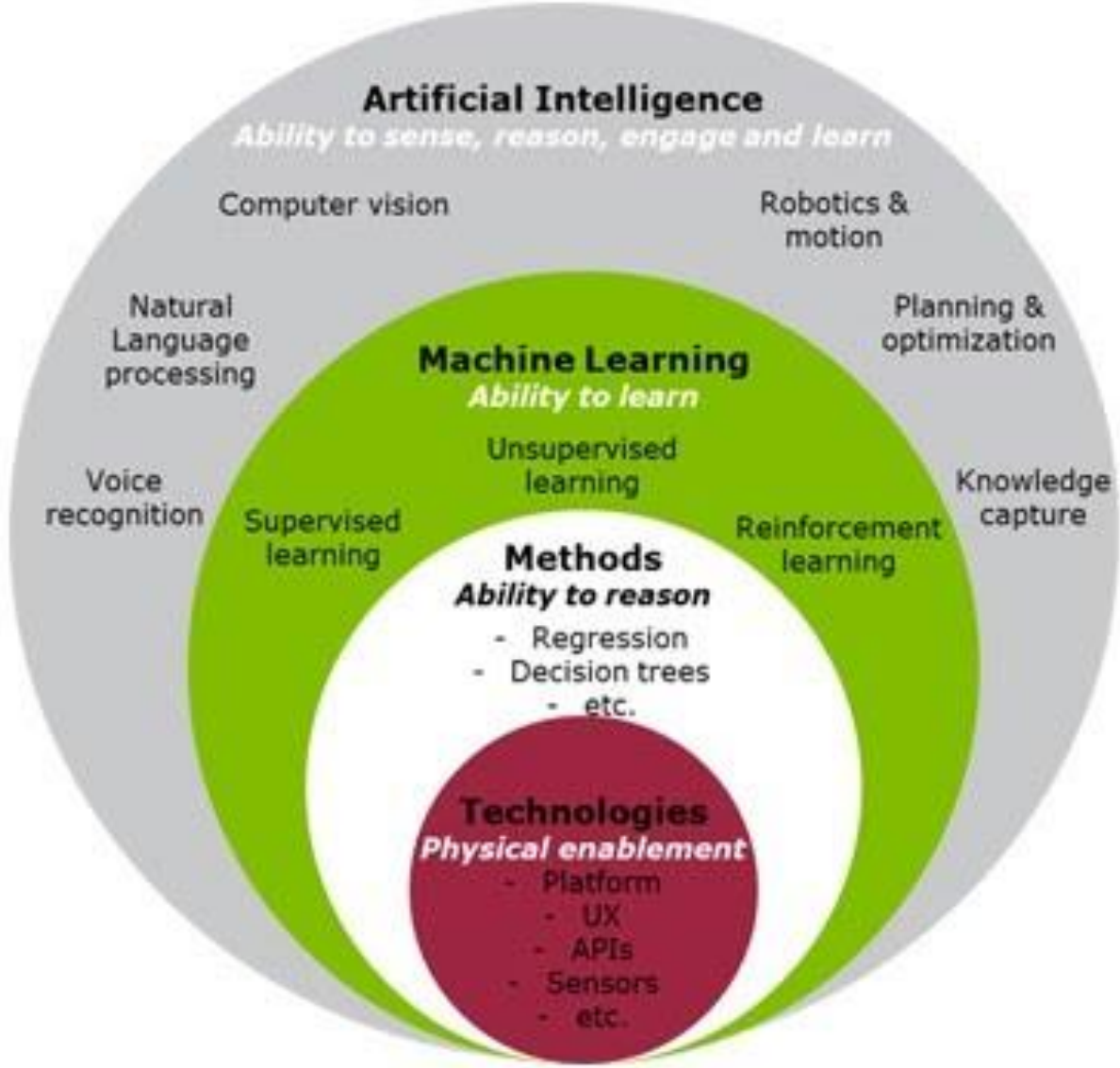
Machine Learning Introduction

By

Amritansh


Artificial Intelligence
Ability to sense, reason, engage and learn
Computer vision
Robotics & motion
Natural Language processing
Planning & optimization
Machine Learning
Ability to learn
Unsupervised learning
Voice recognition
Supervised learning
Reinforcement learning
Knowledge capture
Methods
Ability to reason
- Regression
- Decision trees
- etc.
Technologies
Physical enablement
- Platform
- UX
- APIs
- Sensors
- etc.

Herbert Simon

Machine Learning

- Learning is any process by which a system improves performance from past experiences.
  ~ Herbert Simon

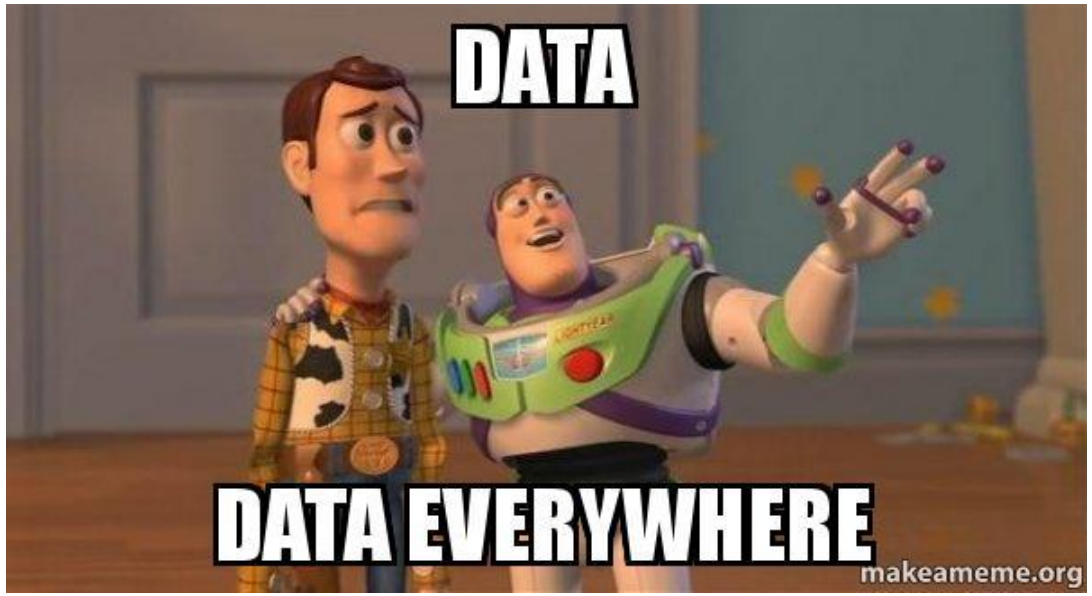
- Machine Learning corresponds to computer programs or algorithms which improve their performance through experience without being explicitly programmed.

Why?

- Develop systems which can automatically adapt and customize themselves.
- Discover new knowledge from large datasets (Data mining)
- Ability to mimic human and replace monotonous tasks. Eg. OCR, handwriting recognition.
- To develop systems that are too difficult and expensive to construct manually.
- Speed up the innovation and analytics
- Make more sense of chaotic world around us

Why now?

- Surge of Big data in our lives : We create new data in amounts of petabytes everyday through our calls, sms, chats, selfies, videos, emails.
- Increasing Computational power (faster CPUs and GPU cores)
- Growing collaboration among researchers in academia and industry.
- Profit margin for corporates



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

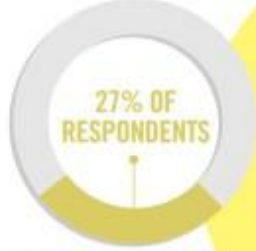
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions



in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Challenges with Big Data

- Volume

  The sheer size of data we are handling is increasing exponentially everyday

- Velocity

  The rate at which new data is being gathered by our systems and sensors

- Variety

  Diverse formats and sensor type result in different data point to represent

- Veracity

  The data quality is really bad in most cases with little or no structure in them

Concept of learning in ML

We have a task T, performance measure P, experience E.

We say a system is learning when while performing task T, the performance measure P improving as it goes through more experiences E.

Case in point: Task = To make better route decisions

Performance = Less travel time

Experience = Travelling through different routes and time

If the system is learning it should start making decisions about selecting routes as it experiences more routes and time associated.

Depending on various features learning can go in either direction.

Take an example of object detection

The problem statement is to detect object present in a scene.

Now consider any natural image as one presented here. It can have multiple features, brightness level and color difference through all pixels it contains.

The way computer sees the data of just 4 x 4 pixel at base level:
x0011FD x0011FF x0011FD x0012FF
x0011FD x0011FF x0011ED x0011FF
x0011BD x0011FF x0011FE x0011FE
x0011ED x0011FF x0011FA x0011AB

The idea is to output the coordinates of pixels which form closest rectangle around the object, label of the class and percentage.

If the system is learning ideally it should return correct classes with closest coordinates around the object and with higher confidence value.

What are the Challenges?

- To create systems which perform with accuracy and precision of human intelligence and can leverage machine's innate architecture to accelerate and keep up with big data (Scales of terabytes or petabytes of data)
- To allow for faster and better decision making through machines wherever possible
- To convert large amounts of raw data into useful analytics
- To help forecast the future trends and correct our estimates based on the analytic output

Terminology

Population : The population is any specific collection of objects of interest.

Sample : Sample is any subset of the population.

Measurement : A measurement is a number or attribute computed for each member of a population or of a sample.

Parameter: Parameter or feature is a numerical value that summarizes some aspects of the whole population.

Inference: Any key knowledge about the data sample or population as a whole from its attributes and properties

Features?

Color: Red

Type: Fruit

Weight: 100 gm

Price: 140/kg

Availability: Yes

Sweet: Yes

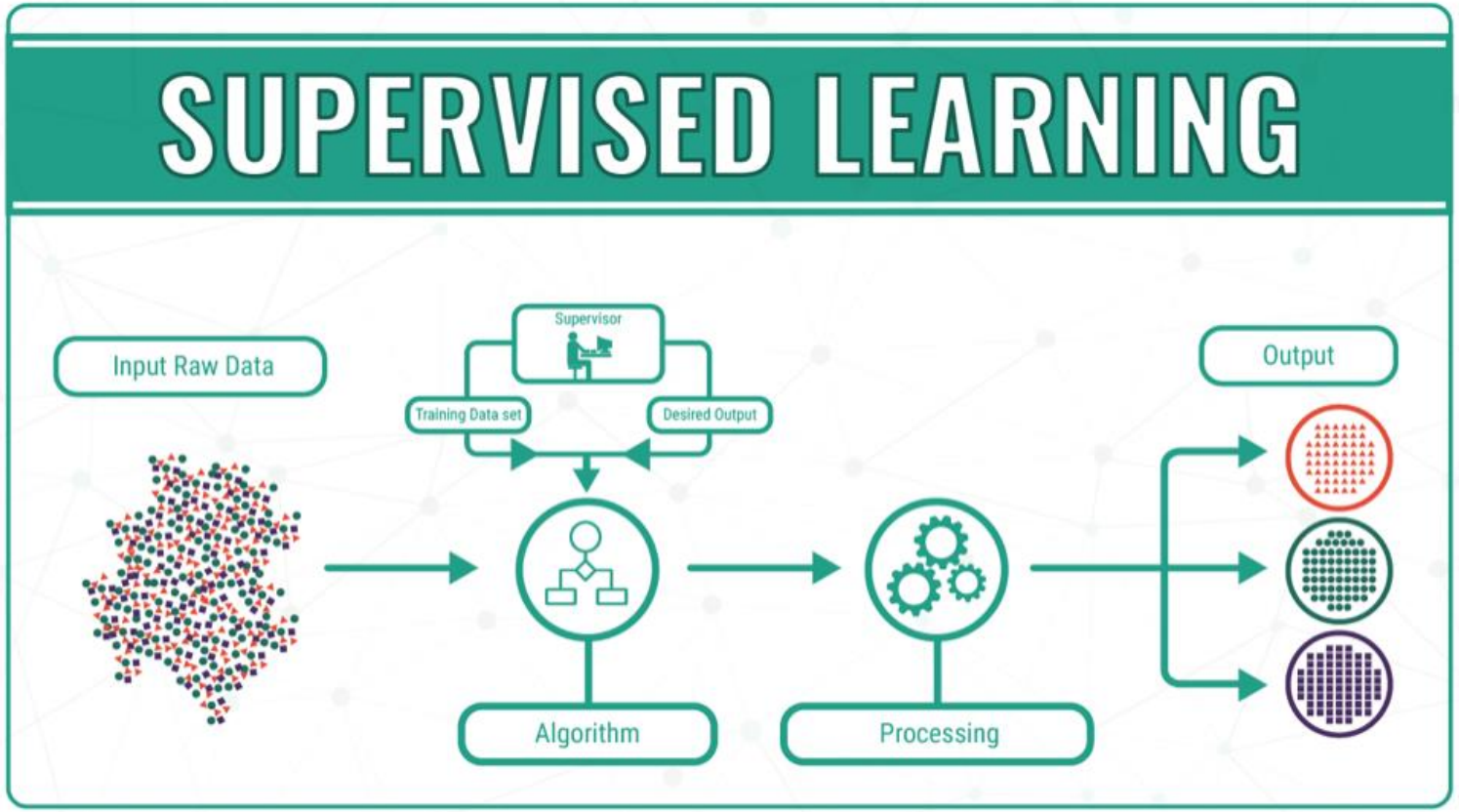
Organic: No

Types of Machine Learning

- Supervised Learning (We provide output labels/tags with input data)
  - Regression/Prediction
  - Classification
- Unsupervised Learning (No output labels are given)
- Reinforcement Learning (Works on reward policy feedback system)

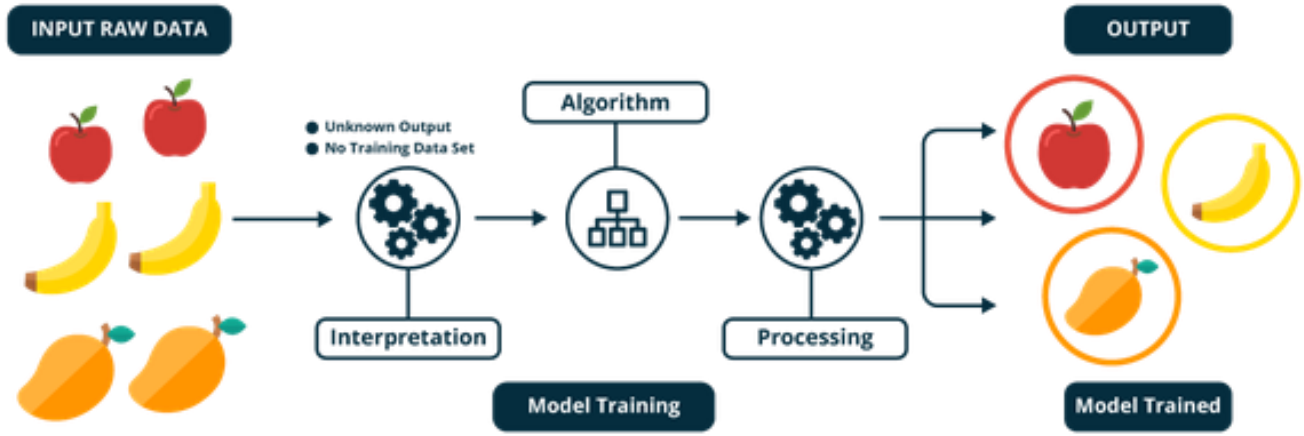
Supervised learning

We provide human annotated data to the model as ground truth



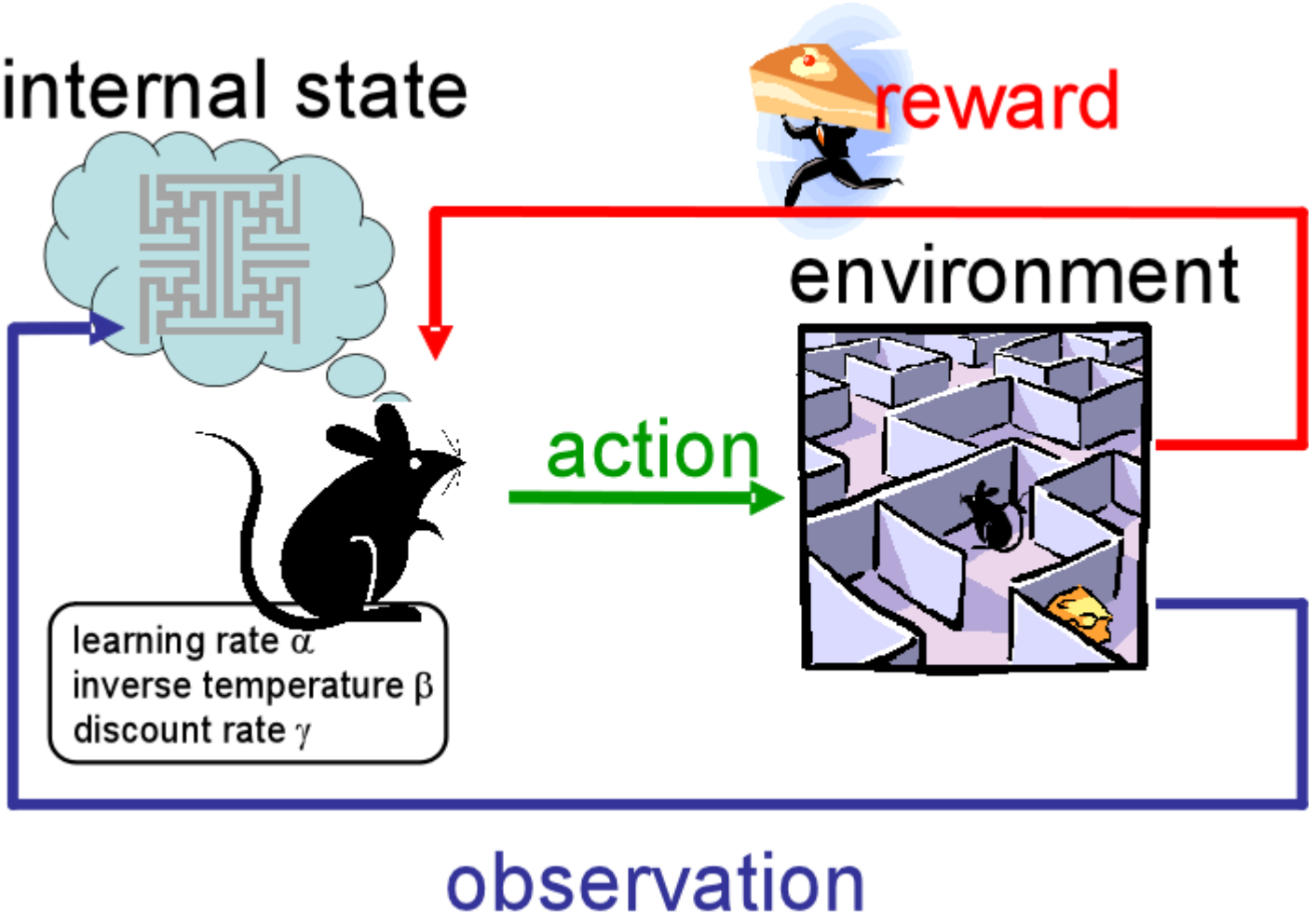
Unsupervised learning

To infer a function that describes the structure of "unlabeled" data (i.e. data that has not been classified or categorized)



Reinforcement learning

A bot/software ought to take *actions* in an *environment* so as to maximize some notion of cumulative *reward*



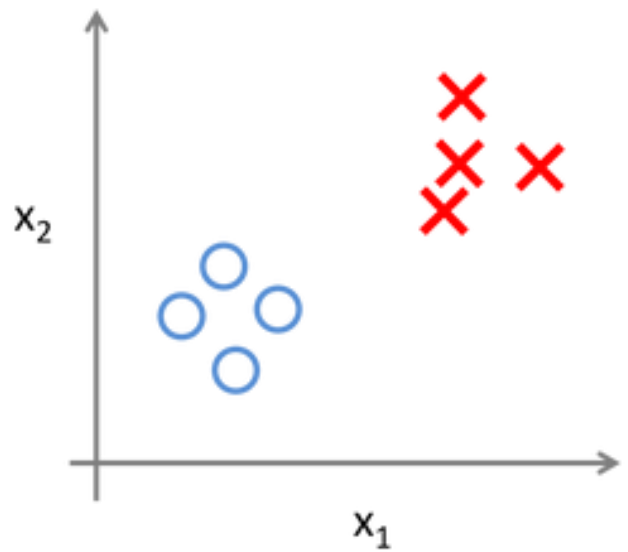
Supervised vs Unsupervised

- The procedures in **supervised learning** are well comprehensible due to their structure. It is possible to contrast different methods, to parameterize and thereby find a solution that is optimal for the application . The interpretation of the data is easier due to the given traceability than with unsupervised learning methods.
- The disadvantage, however, is often a very high manual effort in the preparation of the data.
- It require lot of man hours to prepare fully formatted data to work with supervised learning.

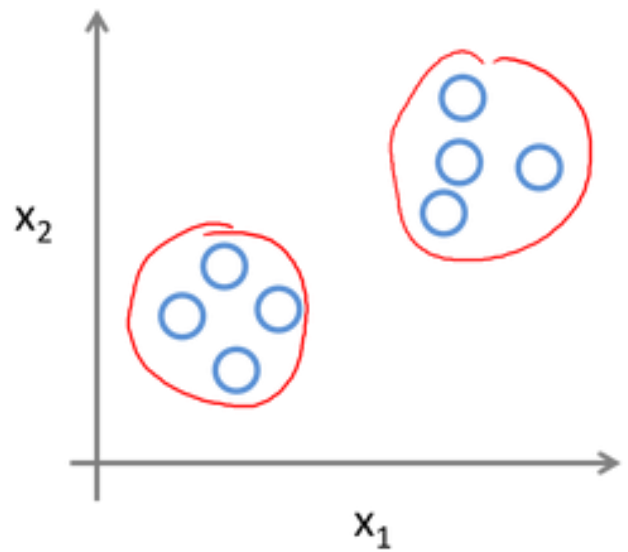
Supervised vs Unsupervised

- The advantages of **unsupervised learning** are the partially fully automated creation of models. These can produce a very good prognosis about new data or even create new content. The model learns with each new record and at the same time refines its calculations and classifications. Manual intervention is no longer necessary.
- The biggest disadvantages are there is no control over what model learns. It can start to cluster wrong type as one group
- It can give bad results amiss the output labels and can lead to lot of misclassifications.

Supervised Learning



Unsupervised Learning



Regression

- Regression refers to correlation between dependent and independent variables.

  Consider eqn of line: **Y = mx + c**

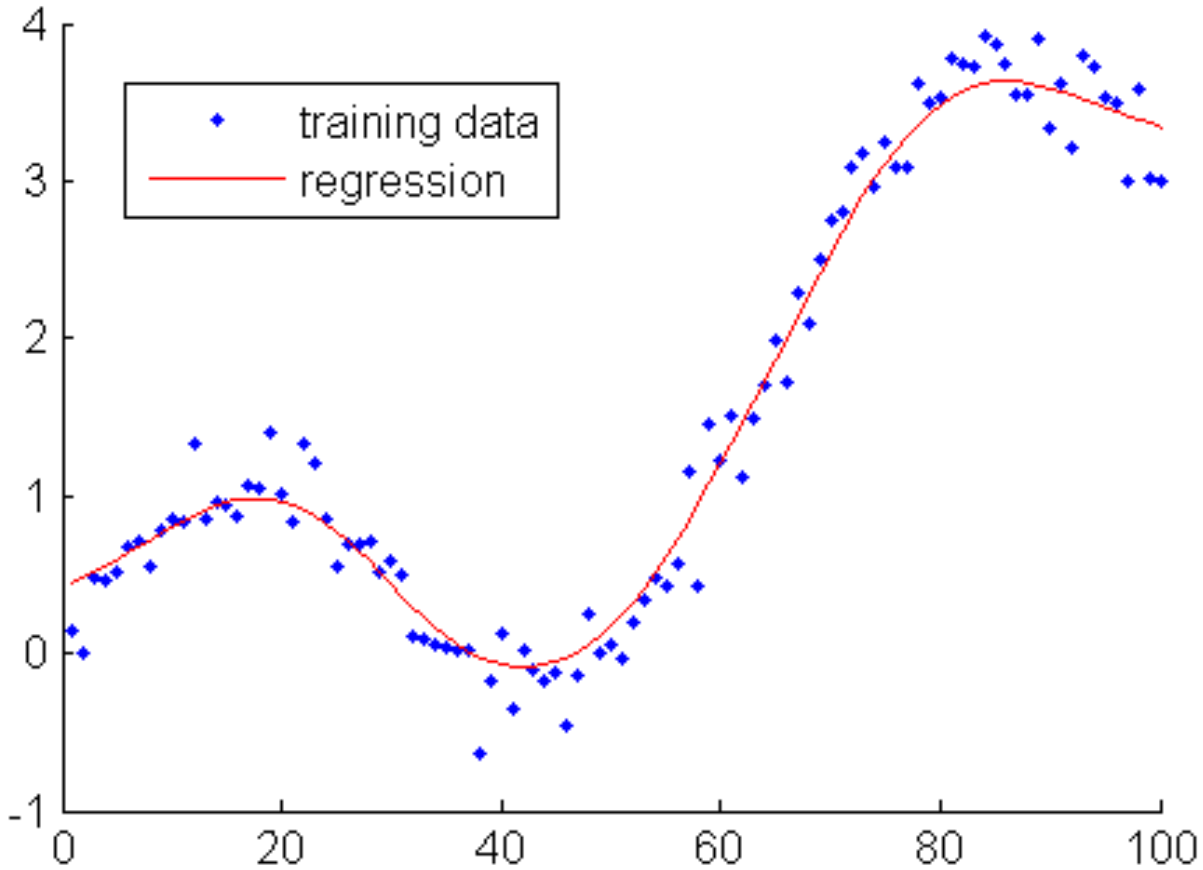
Constant/parameter

Independent variable

Dependent Variable

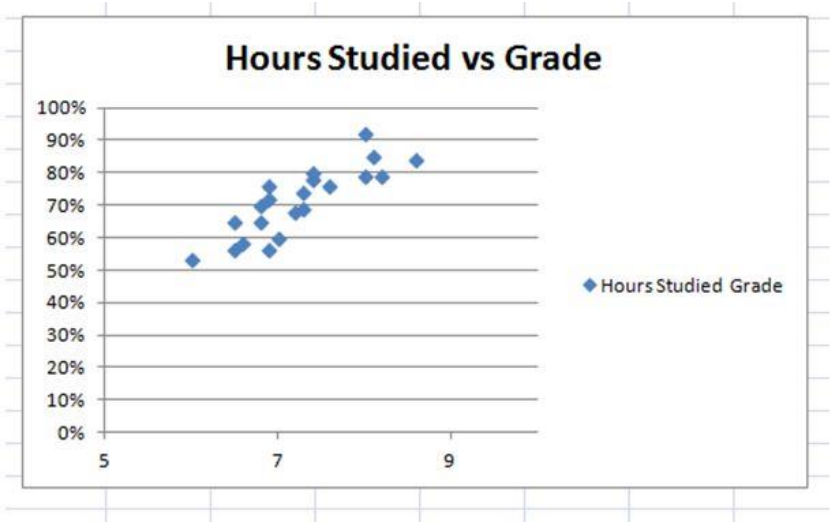
Coefficient/parameter

In statistics regression is defined as a measure of the relation between the mean value of one variable (e.g. output) and corresponding values of other variables (e.g. time and cost).



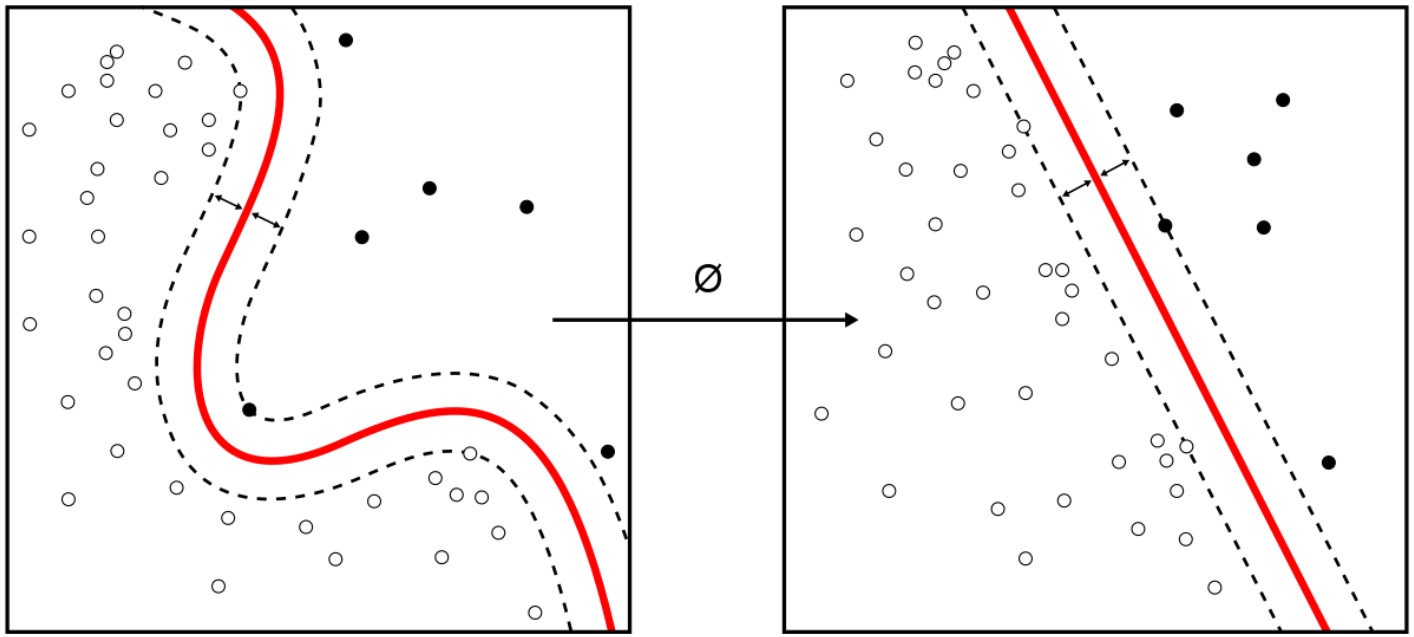
- Regression is used to fit a eqn through the given data points.
- We try to approximate the best function to fit through the data
- The core idea it to predict the points in future. We assume the data distribution will hold up and new points will be near to the function.
- It reduces uncertainty with the data and helps in correct estimation over a period of time.

- Few Cases for Regression could be:
  - Sales Prediction
  - Product Pricing
  - Expense overrun variables
  - House pricing prediction
  - Population estimates
  - Pollution particulate estimates



Honestly regression finds its usage anywhere we want to find correlations:

Temperature vs. Number of cones sold at ice cream store
Inches of rain vs. new cars sold
Daily Snowfall vs. number of skier visits



Classification

- In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

- Use cases of Classification
  - Spam Filtering
  - Cancer diagnosis
  - Text classification
  - Sentiment analysis
  - Object classification
  - Face detection

  Similar to Regression this technique can be generalized into any problem where we know there are discrete classes.

  Classification of people based on their eating habits:

  Vegetarian, Vegan, Non-Vegetarian