Chapter 6: Machine Learning with Big data

Basanta Joshi, PhD
Asst. Prof., Depart of Electronics and Computer Engineering
Program Coordinator, MSc in Information and Communication Engineering
Member, Laboratory for ICT Research and Development (LICT)
Institute of Engineering
basanta@ioe.edu.np
http://www.basantajoshi.com.np

What is Learning?

- Learning is one of those everyday terms which is broadly and vaguely used in the English language
  - Learning is making useful changes in our minds
  - Learning is constructing or modifying representations of what is being experienced
  - Learning is the phenomenon of knowledge acquisition in the absence of explicit programming
- **Herbert Simon, 1983**
  - **Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the same task or tasks drawn from the same population more efficiently and more effectively next time.**

Implications

Learning involves 3 factors:

changes

Learning changes the learner: for machine learning the problem is determining the nature of these changes and how to best represent them

generalization

Learning leads to generalization: performance must improve not only on the same task but on similar tasks

improvement

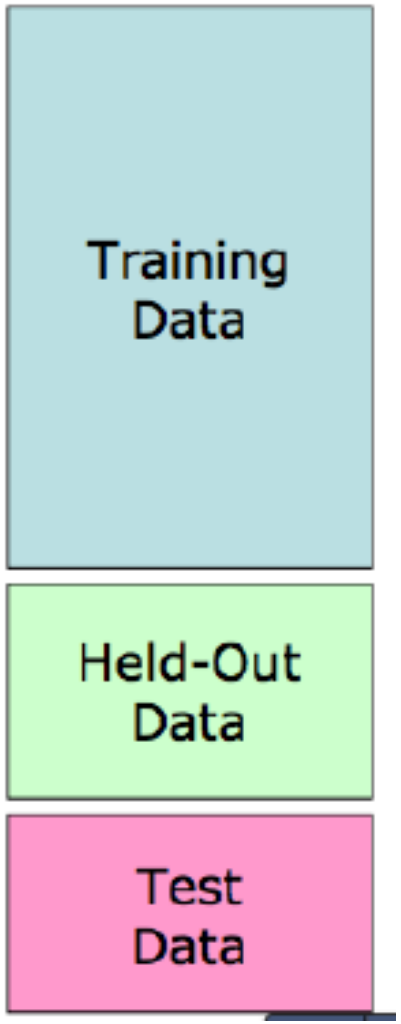
Learning leads to improvements: machine learning must address the possibility that changes may degrade performance and find ways to prevent it.

Areas of Influence for Machine Learning

- *Statistics:* How best to use samples drawn from unknown probability distributions to help decide from which distribution some new sample is drawn?
- *Brain Models:* Non-linear elements with weighted inputs (Artificial Neural Networks) have been suggested as simple models of biological neurons.
- *Adaptive Control Theory:* How to deal with controlling a process having unknown parameters that must be estimated during operation?

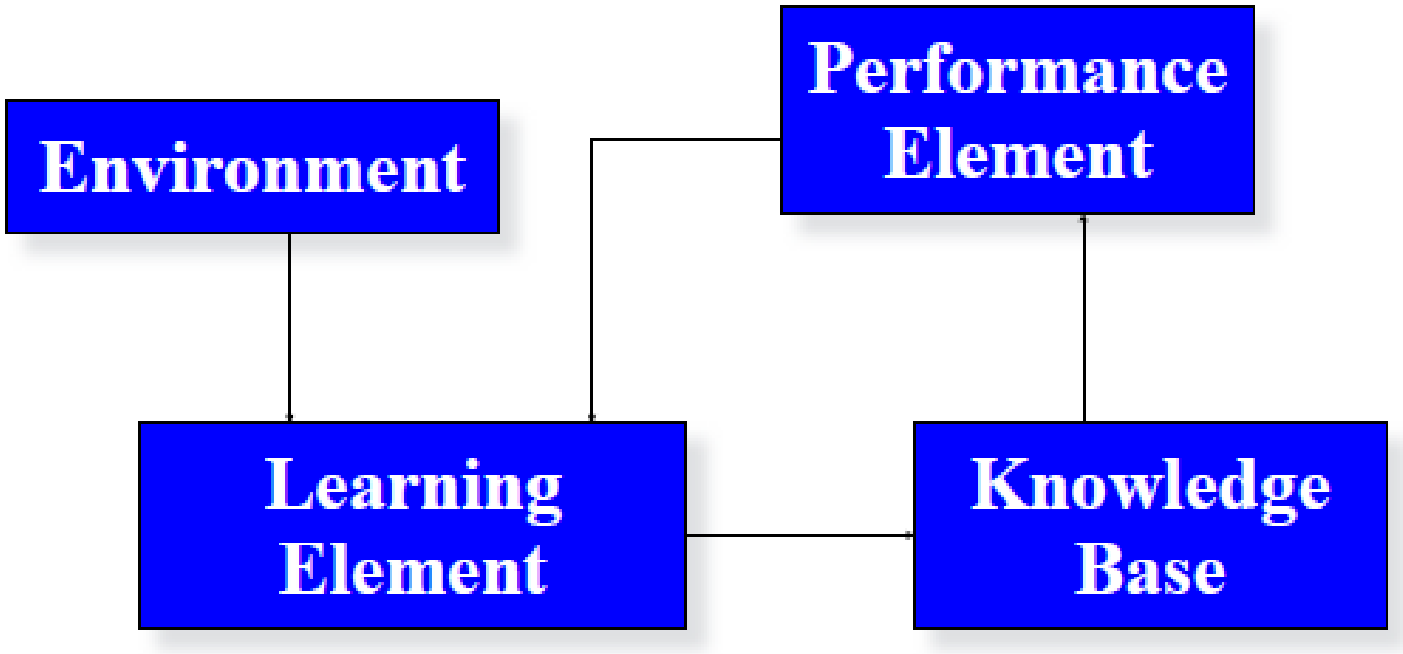
Important Concepts

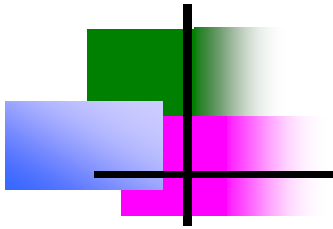
- Data: labeled instances, e.g. emails marked spam/ham
  - Training set
  - Held out set
  - Test set
- Features: attribute-value pairs which characterize each x
- Experimentation cycle
  - Learn parameters (e.g. model probabilities) on training set
  - (Tune hyperparameters on held-out set)
  - Compute accuracy of test set
  - Very important: never "peek" at the test set!
- Evaluation
  - Accuracy : fraction of instances predicted correctly
- Overfitting and generalization
  - Want a classifier which does well on *test* data
  - Overfitting: fitting the training data very closely, but not generalizing well



Learning Framework

- There are four major components in a learning system:





The Environment

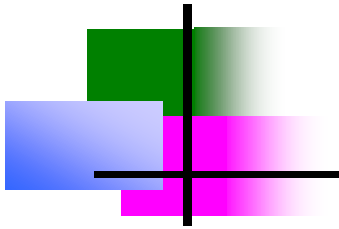
- The environment refers the nature and quality of information given to the learning element

- The nature of information depends on its level (the degree of generality wrt the performance element)
  - high level information is abstract, it deals with a broad class of problems
  - low level information is detailed, it deals with a single problem.

- The quality of information involves
  - noise free
  - reliable
  - ordered

Learning Elements

- **Four Learning situations**
- Rote Learning
  - environment provides information at the required level
- Learning by being told
  - information is too abstract, the learning element must hypothesize missing data
- Learning by example
  - information is too specific, the learning element must hypothesize more general rules
- Learning by analogy
  - information provided is relevant only to an analogous task, the learning element must discover the analogy

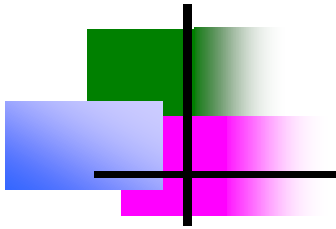


The Knowledge Base

- Expressive
  - the representation contains the relevant knowledge in an easy to get to fashion
- Modifiable
  - it must be easy to change the data in the knowledge base
- Extendibility
  - the knowledge base must contain meta-knowledge (knowledge on how the data base is structured) so the system can change its structure

The Performance Element

- Complexity
  - for lear ning, the simplest task is classif ication based on a single rule while the most complex task requires the application of multiple rules in sequenc e
- Feedback
  - the performance element must send information to the learning system to be used to evaluate the overall
  - performance
- Transparency
  - the learning element should have access to all the internal actions of the performance element



What is Machine Learning?

- Humans and animals: learning from experience.
- Machine learning: computational algorithms which enable machine to "learn" from experience (a dataset) to perform tasks without following static program instructions.
- Performance increases as number of samples increases.
- Machines performance better than humans in certain tasks e.g. cancer detection, character recognition

Machine Learning and Artificial Intelligence

"Evolved from the study of pattern recognition and computational learning theory in artificial intelligence, machine learning explores the study and construction of algorithms that can learn from and make predictions on data – such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs."

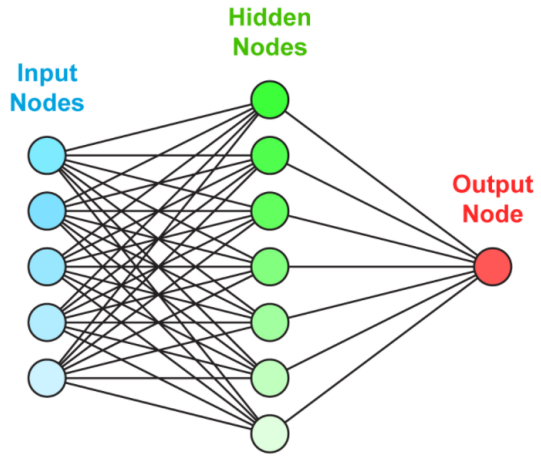
by Arthur Samuel (pioneer in machine learning at IBM in 1959)

Reference: https://en.wikipedia.org/wiki/Machine_learning

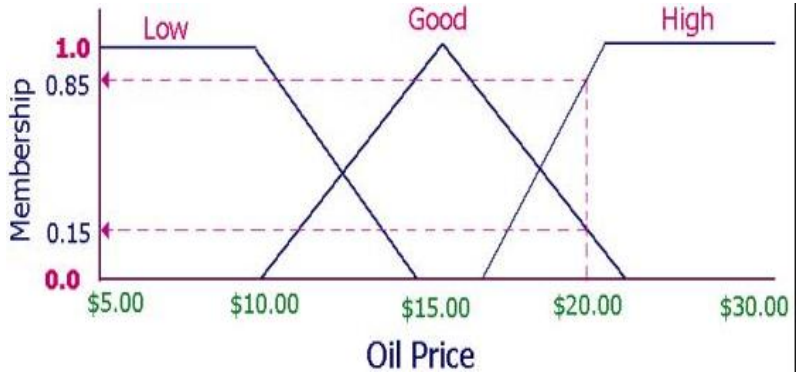
Artificial Intelligence (A.I.)

Intelligence exhibited by machines

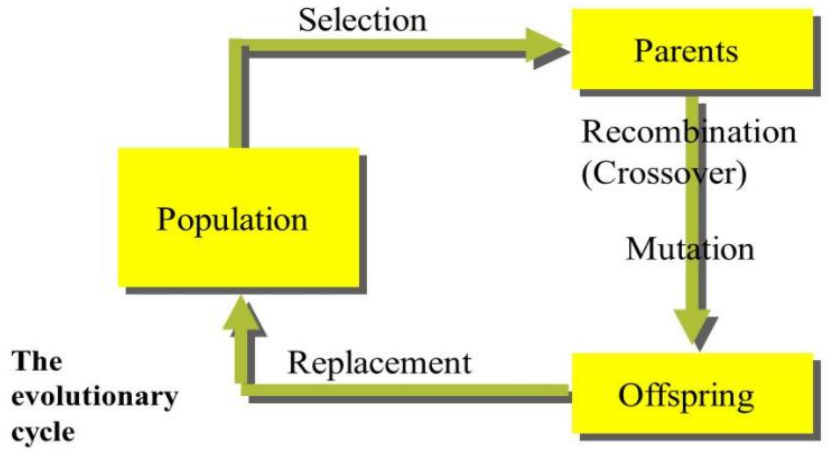
Machine Learning
(e.g. neural networks)



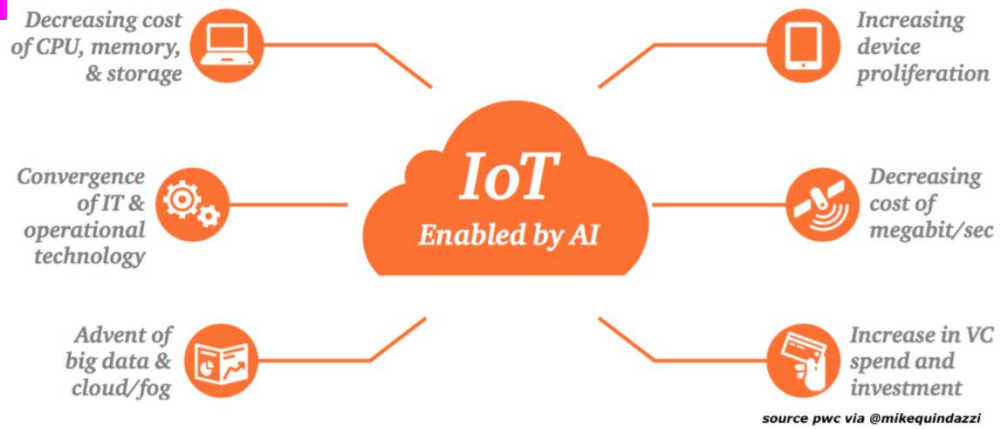
Fuzzy Logic
(modelling human vague concepts)

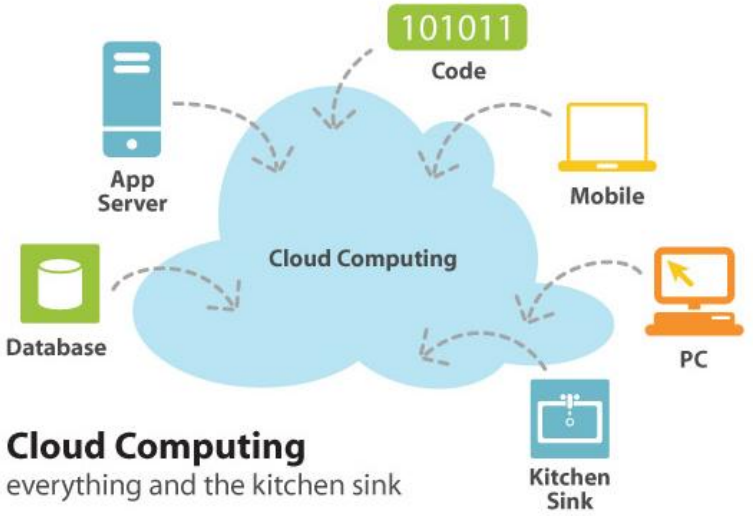


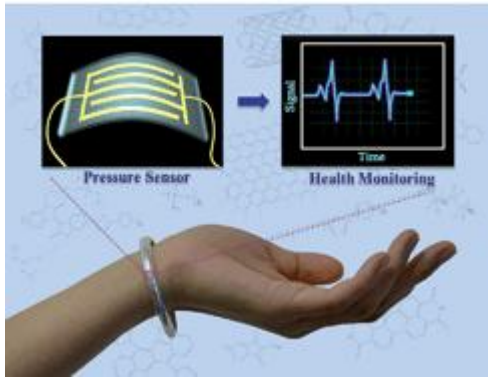
Evolutionary optimisation algorithms
(near-optimal solutions)



Applications of A.I.







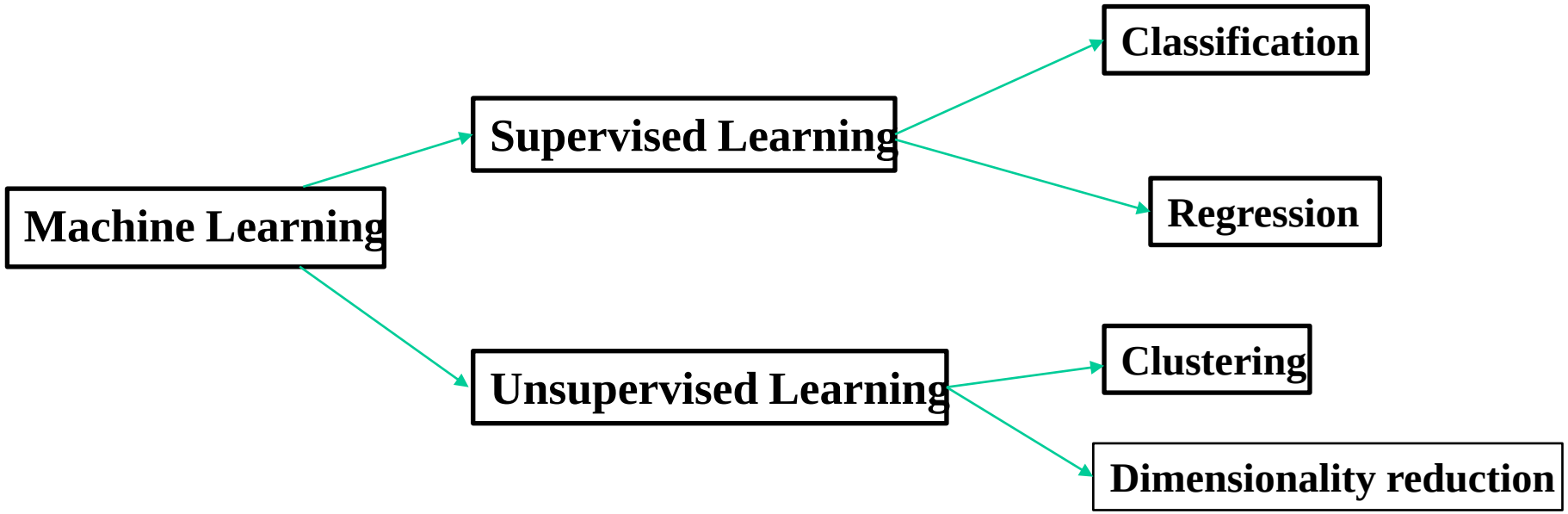
Images source: Google

When to use Machine Learning?

- Complex tasks involving a large amount of data, but no formula (static rules) of performing the tasks can be defined.

- Applications:
  - Sales forecasting
  - Face recognition (identification of users)
  - Speech recognition (customer services)
  - Stock market forecasting
  - Medical diagnosis e.g. cancer detection
  - Energy usage forecasting e.g. British Gas
  - Automatic recommendations on web

Types of Machine Learning Problems



- **Supervised Learning:** develop predictive models from labelled data (i.e. data with classes or targets)
- **Unsupervised learning:** describe hidden structure of unlabelled data
  - Clustering: Group similar data into categories (clusters) based only on input data
  - Dimensionality reduction: Reduce input variables of a dataset to a smaller set of variables (structure of dataset)

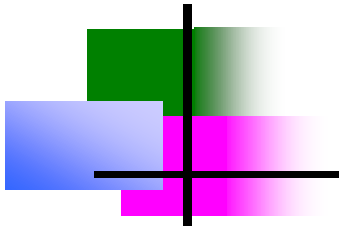
Supervised Learning

- Aim: build a model (e.g. neural network) to makes predictions by optimizing the parameters of the model

- The process of supervised learning:
  - **Phase 1 (Training):** Train a model using a training set (optimizing the parameters of the model)
  - **Phase 2 (Testing):** Evaluate the performance of the mode using a test set

- The training set: a dataset with known output (classes or targets).
- The test set: a dataset with known output and no common data with the training set

- Supervised learning problems:
  1. Classification
  2. Regression



Classification

- Aim: Build a model to predict the classes or categories of input data.
- Examples of classification problems:
  - Customer credit rating: 3 classes (high, medium, low)
  - Energy consumption rating of a household
  - Email spam detection: 2 classes (spam and non-spam)
  - Tumour detection: 2 classes (has tumour and has no tumour)
  - Detection of different types of aeroplanes
  - Websites categorization and recommendation

Regression

- Aim: Build a model to predict a continuous real value (any value in a range)

- Examples of regression problems:
  - Sales of products (millions of pounds)
  - House price (thousands of pounds)
  - Weather forecast (temperature degree, wind speed and direction)
  - Electric load forecasting (MegaWatts)

Machine Learning algorithms

– Supervised Learning Algorithms:
  - **Neural networks**
  - Logistic regression
  - Support Vector Machines
  - Decision tree
  - Naïve Bayes
  - Bayesian networks
  - Nearest neighbours
  - …
– Unsupervised Learning Algorithms:
  - Principal Components Analysis (PCA)
  - K-means clustering
  - Self-organizing map (SOM)
  - …