CUNY School of Professional Studies

SPS.CUNY.EDU

2020 Spring Data-622
Introduction to Machine Learning and Big Data
Raman Kannan

Instructor Email Address: Raman.Kannan@sps.cuny.edu

Acknowledgements:
Generous support from IBM Power Systems Academic Initiative
IBM PSAI provides computing infrastructure for free

Subject Matter of the Course

Machine Learning: Birds Eye View
Supervised Learning:Classifiers
Strategies for Supervised Learning
Analytical Foundation of Classifiers
- – Density Estimation:MLE
- – Optimization:Loss Function
- – Convergence: GD, SGD

Comparative Analysis of Classifiers
Sources of Error Estimation
Sources of: Bias vs Variance
Error Reduction Strategies
Resampling Methods
Penalization Methods
Cross Validation
Ensemble Methods
Combining Like Classifiers
Combining Unlike Classifiers
Brief Introduction to Big Data
Opportunities of Parallelism

Commitment is assumed.

How will this class run?

100% Online
Weekly One hour Call – atleast one hour call biweekly
Weekly Reading Assignment/Contribute Discussion Forum
Instructor Led Clinical session
- 90% of the work in R programming language
- 10% shell scripting (bash) and remote computing

What will learners do	How will learners be evaluated
Read the syllabus	
2 individual modeling assignments	30%
Engaging and Participation	12%
2 individual with performance improvement	40%
Open book/open notes test – t1	9%
Open book/open notes test – t2	9%
follow process discipline&on time	Late submission not graded

Agreements

Instructor is obligated to help you understand the material covered in the course, to assist you in clinical aspects.

Students agree to make the effort to read,integrate outside material and experiment with tools and techniques.

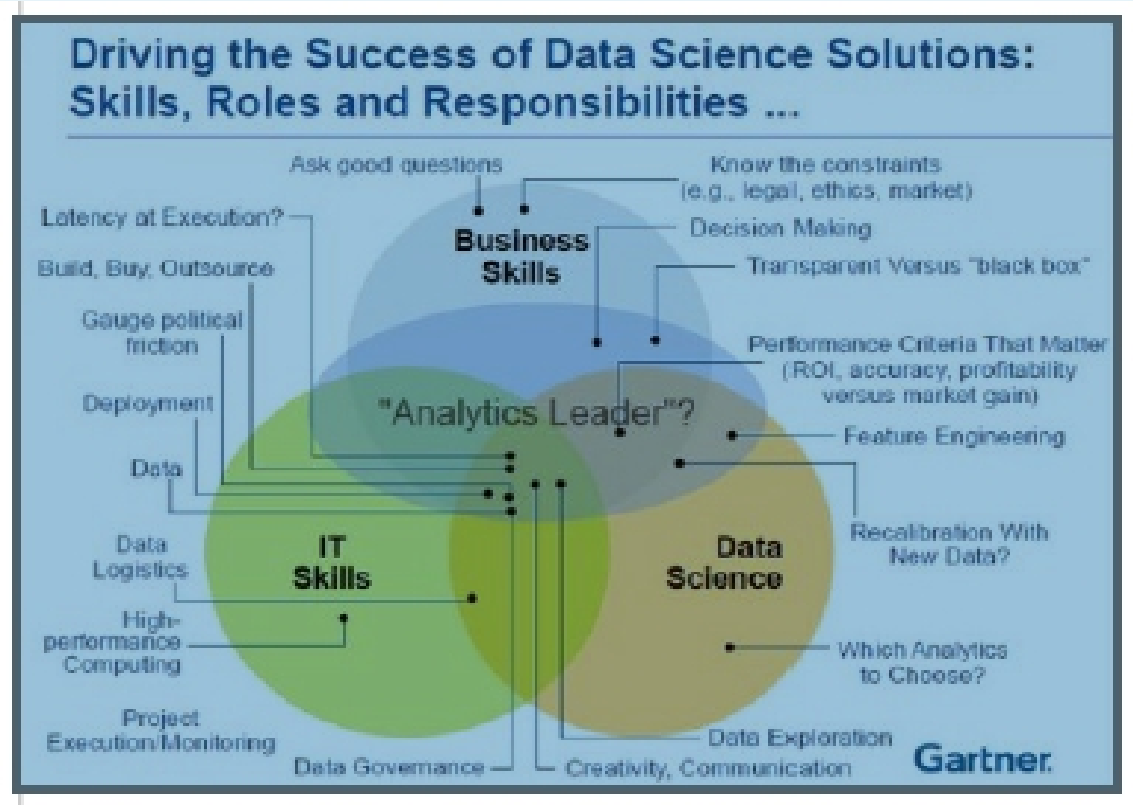
3 to 4 hours a week of effort is required to assimilate the material. Bill Gates opines 10000 hours to master new skills. Steep hill! Let us get started.

Machine Learning

- Machine: automation/objective/endurance
  - patently human, we seek to automate
- Learning: Observe/infer/improve/feedback
  - Learning is how we evolve/improve in life.

M/L is a sub-discipline of AI, confluence of Computing & Statistics
M/L shares numerous topics with data mining, pattern recognition
M/L, therefore, shares topics from Data Science and Statistical Learning
M/L employs data science techniques, linear algebra,calculus and probability and statistics.
To be masterful in M/L one has to be a competent/efficient software engineer –able to gather,prepare, process, analyze data and most importantly convey the findings to a broader audience.

What is M/L ?



M/L is amorphous boundary.
It is part of AI, always has been.
Many other related areas such as Data Mining/Data Science have evolved that today the distinction is further blurred.

M/L has co-opted Neural Nets and NLP.

But there is no debate on the need to integrate Business, technology and Data Science to achieve M/L.

Difference between Machine Learning, Data Science, AI, Deep Learning, and Statistics – Vincent Granville
https://www.datasciencecentral.com/profiles/blogs/difference-between-machine-learning-data-science-ai-deep-learning
https://mlplatform.nl/what-is-machine-learning/
https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/
https://en.wikipedia.org/wiki/Timeline_of_machine_learning

Historical Perspective

> Over the past 50 years the study of Machine Learning has grown from the efforts of a handful of computer engineers exploring whether computers could learn to play games, and a field of Statistics that largely ignored computational considerations, to a broad discipline that has produced fundamental statistical-computational theories of learning processes, has designed learning algorithms that are routinely used in commercial systems for speech recognition, computer vision, and a variety of other tasks, and has spun off an industry in data mining to discover hidden regularities in the growing volumes of online data.
>
> – Tom Mitchell, CMU Machine Learning Department

To be more precise, we say that a machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E. Depending on how we specify T, P, and E, the learning task might also be called by names such as data mining, autonomous discovery, database updating, programming by example, etc. – also credited to Tom Mitchell

"The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience."
-- Tom Mitchell

https://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf

Definition of M/L

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.-- *https://en.wikipedia.org/wiki/Machine_learning*

A scientific field is best defined by the central question it studies. The field of Machine Learning seeks to answer the question
“How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?”

M/L Tasks

https://en.wikipedia.org/wiki/Machine_learning#Types_of_learning_algorithms

Supervised learning,Unsupervised learning,Semi-supervised learning
Reinforcement learning

Our focus however is

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.[1] It infers a function from *labeled training data* consisting of a set of *training examples*.[2] In supervised learning, each example is a *pair* consisting of an input object (typically a vector) and a desired output value (also called the *supervisory signal*). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way (see inductive bias).

The parallel task in human and animal psychology is often referred to as concept learning.

Course focus

Our objective develop vocabulary around critical concepts central to practicing Supervised Learning, aka Classifiers or classification algorithms.

We will develop required expertise to:

- prepare dataset,
  - Load
  - Clean (missing values, range, outlier, unbalanced)
  - Exploratory Analysis (discriptive statistics)
  - Feature selection/scale/transform
  - run one or more classifiers
    - training phase
    - test/validation phase
- write a report on summary findings.

https://www.r-bloggers.com/the-real-prerequisite-for-machine-learning-isnt-math-its-data-analysis/

Inventory of Classifiers

https://www.datasciencecentral.com/profiles/blogs/top-10-machine-learning-algorithms

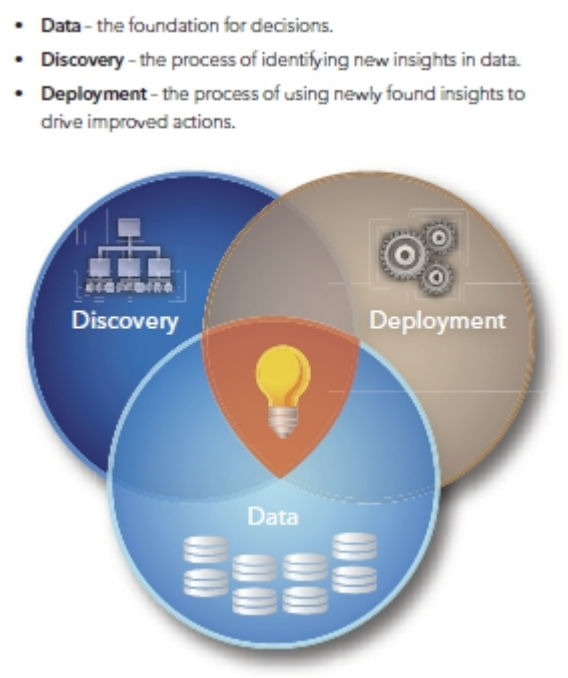
Brief overview of Linear Regression
Logistic Regression
Descision Tree
Linear Discriminant Analysis LDA
Quadratic Discriminat Analysis QDA
Naive Bayes
Support Vector Machine
KNN –Nearest Neighbor
Random Forest

We will prescribe and adopt a process and repeat the process
Using each of these algorithms – over the same dataset in R
We will adopt how we may measure the performance
Compare each of the above classifiers using that measure

Reading on Introduction

https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf
https://amueller.github.io/COMS4995-s19/slides/aml-01-introduction/#p43
https://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf
https://www.r-bloggers.com/the-real-prerequisite-for-machine-learning-isnt-math-its-data-analysis/
https://www.datasciencecentral.com/profiles/blogs/top-10-machine-learning-algorithms
https://ciml.org
https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/data-mining-from-a-z-104937.pdf

Process



Data – the foundation for decisions.
• Discovery – the process of identifying new insights in data.
• Deployment – the process of using newly found insights to drive improved actions

https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/data-mining-from-a-z-104937.pdf

Know your data!

- Data
  - Set of observations, one record
  - Each observation is a set of
    - Attribute,fields,variables
    - Most are independent
    - One dependent variable

If the variates are numerical/continuous → Regression
If the variates are categorical/nominal → Supervised Learning

When all the data relating to observation is in one record, dataset is said to be in wide format.The algorithms we will consider require wide format.

Structure of data

- Wide format – each row is an observation
- One or more dimensions (attributes or features)
- Domains:categorical and numerical

What all can we do with a dataset?

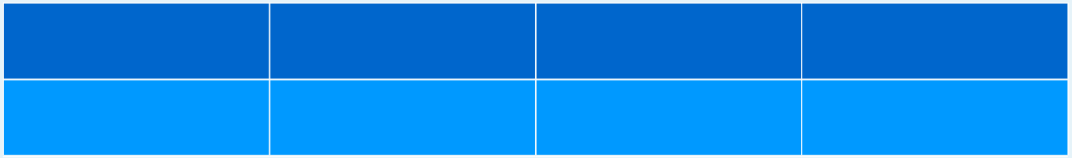
Broad tasks

- Visualize
- Unearth patterns & relationships
  - Group them, descriptive analytics
  - Categorize, predict, classify
  - Rank/Order
  - Associations

M/L: Goals and types

Machine Learning Goals

- unearth “hidden” structures (unsupervised)
- predict/forecast (supervised)



Classifier – Lingo

web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf

1 Introduction

Consider the standard supervised learning problem. A learning program is given training examples of the form $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ for some unknown function $y = f(\mathbf{x})$. The $\mathbf{x}_i$ values are typically vectors of the form $\langle x_{i,1}, x_{i,2}, \ldots, x_{i,n} \rangle$ whose components are discrete- or real-valued such as height, weight, color, age, and so on. These are also called the *features* of $\mathbf{x}_i$. Let us use the notation x_{ij} to refer to the j-th feature of $\mathbf{x}_i$. In some situations, we will drop the i subscript when it is implied by the context.

The y values are typically drawn from a discrete set of classes $\{1, \ldots, K\}$ in the case of *classification* or from the real line in the case of *regression*. In this chapter, we will consider only classification. The training examples may be corrupted by some random noise.

Given a set S of training examples, a learning algorithm outputs a *classifier*. The classifier is an hypothesis about the true function f. Given new $\mathbf{x}$ values, it predicts the corresponding y values. I will denote classifiers by $h_1, \ldots, h_L$.

ML Workflow

https://mapr.com/ebooks/spark/08-recommendation-engine-spark.html

ISL | Statistics | How to Stop Skype Fr | Attributes: Defining C | Fundry & Yesco - Firm

Typical Machine Learning Workflow

A typical machine learning workflow is shown below.

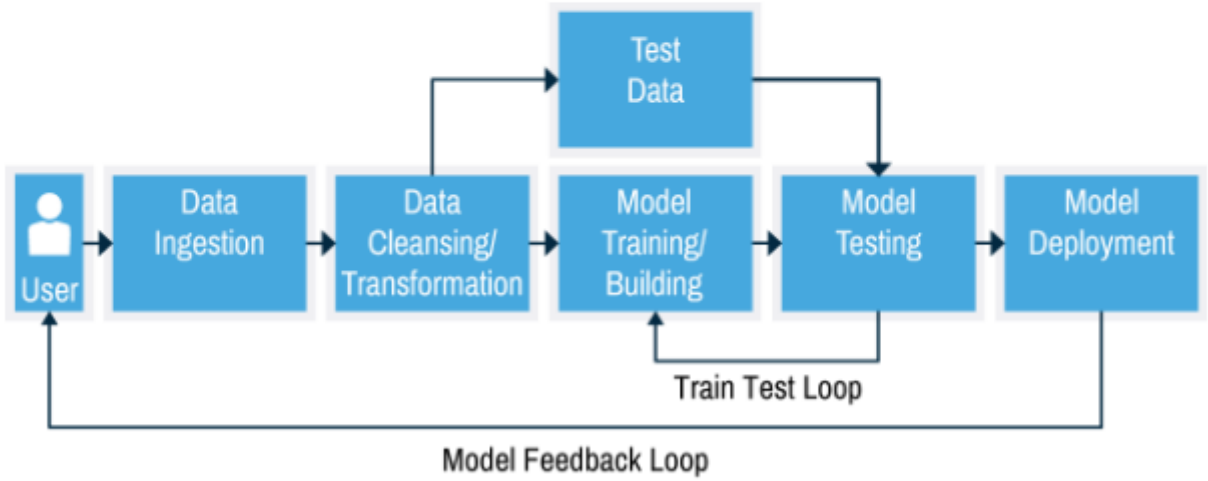
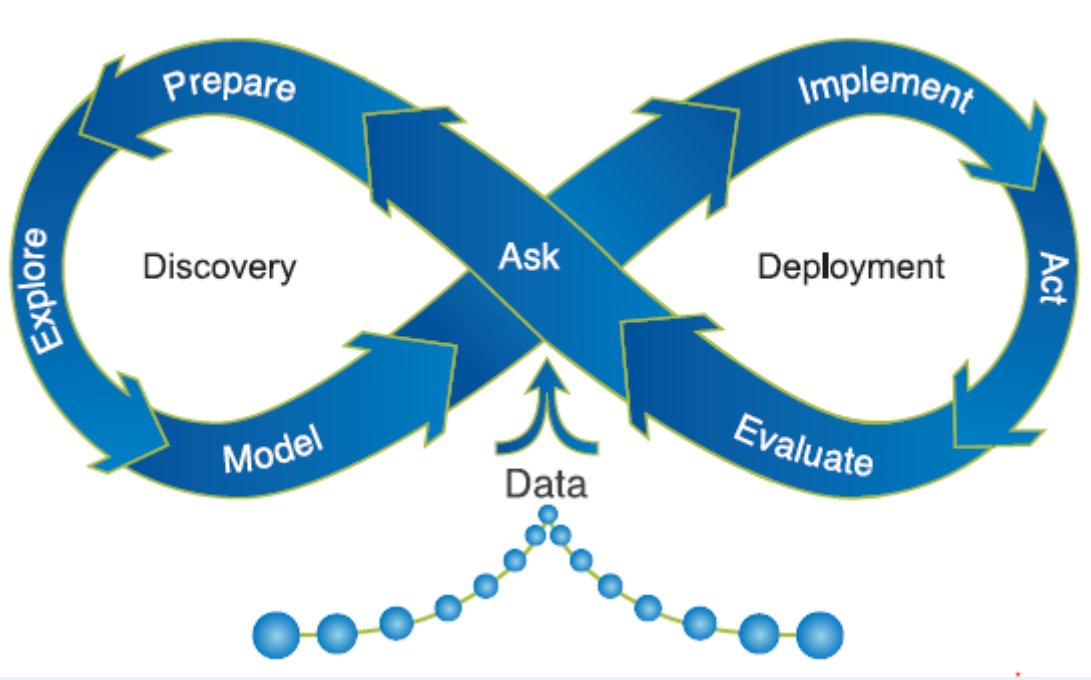


Figure 12. Maching Learning Workflow

Process is – Iterative



This process applies to learning, data mining and M/L and it is iterative – not a single-pass sequence of steps.

https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/data-mining-from-a-z-104937.pdf

2 Phase Learning

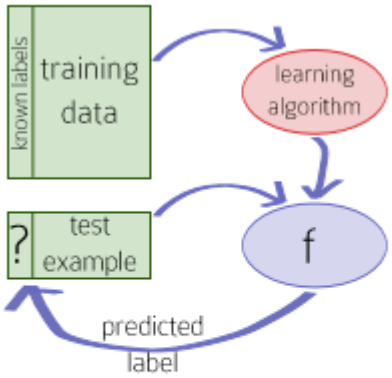


Figure 1.1: The general supervised approach to machine learning: a learning algorithm reads in training data and computes a learned function f. This function can then automatically label future text examples.

The general framework of induction.
We are given training data on which our algorithm is expected to learn.
Based on this training data, our learning algorithm induces a function f that will map a new example to a corresponding prediction.

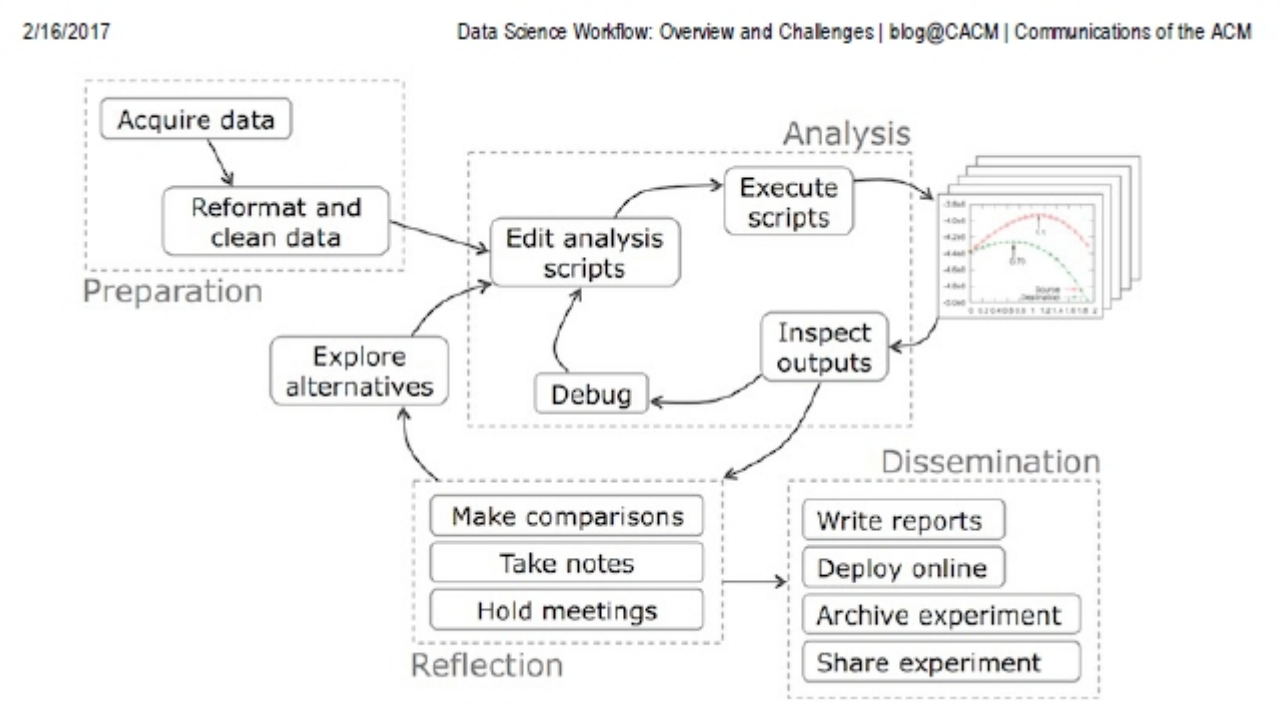
We refer to the collection of examples on which we will evaluate our algorithm as the test set. The test set is a closely guarded secret. Never seen before data.

The goal of inductive machine learning is to take some training data and use it to induce a function f. This function f will be evaluated on the test data. The machine learning algorithm has succeeded if its performance on the test data is high.

What question comes to your mind?

ciml.info/dl/v0_99/ciml-v0_99-ch01.pdf

Closer look inside



http://cacm.acm.org/blogs/blog-cacm/169199datascienceworkflowoverviewandchallenges/fulltext

Process Discipline

Process Discipline
→ Process Script
→ Repeatable/Reproducible

Supervised-Learning-[SB-Kotsiantis].pdf - Adobe Acrobat Reader DC

File Edit View Window Help

Home Tools ROCintro.pdf Performance_Evalu...

2 / 20

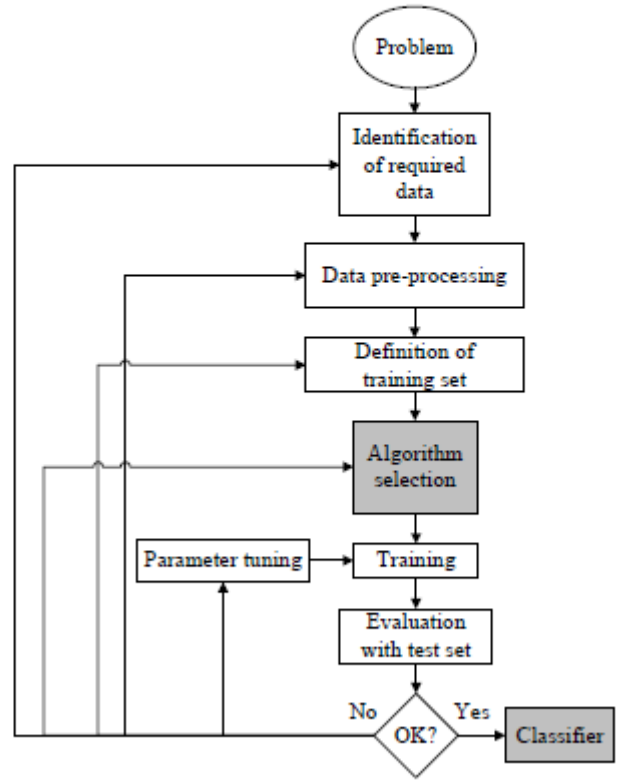


Figure 1. The process of supervised ML

Formalizing the Learning

- We must ask *What is performance?*

- The performance of the learning algorithm should be measured on unseen "test" data.
- The way in which we measure performance should depend on the problem we are trying to solve.
- There should be a strong relationship between the data that our algorithm sees at training time and the data it sees at test time.

In order to accomplish this, let's assume that someone gives us a **loss function**, $\ell(\cdot, \cdot)$, of two arguments. The job of ℓ is to tell us how "bad" a system's prediction is in comparison to the truth. In particular, if y is the truth and $\hat{y}$ is the system's prediction, then $\ell(y, \hat{y})$ is a measure of error.

WHAT

WHY

HOW

Never seen before data – Generalize