Introduction to Machine Learning (ML) and applications in hydraulic and hydromorphology

Tuesdays (8:00 – 11:00)

Join the Zoom-Meeting
https://tum-conf.zoom.us/j/66893758998
Meeting-ID: 668 9375 8998
Password: 395789

PD. Dr.-Ing. habil. Dipl.-Math. Minh Duc Bui
Tel: 089.289.23925
E-Mail: bui@tum.de
Room: 0506.01.614

Intended Learning Outcomes

- Understanding the basic of “how ML models work”,
- Acquiring first-hand experience with using ML and MATLAB to develop simple models for hydraulic and hydromorphology, and
- Having hands-on experience implementing ML models in real cases.

Content

- Basics of Machine Learning (1)
- Basic mathematical prerequisites and MATLAB programming (1)
- Regression and Classification (2)
- Basics of Artificial Neural Networks (ANN) (1)
- ANN applications for hydromorphology (1)
- Computer lab: MATLAB and ANN for sediment transport (1)
- ANN applications for hydraulic (1)
- Computer lab: MATLAB and ANN for modelling flooding (1)
- Project work: Applying ML to solve real problems (2)
- Report submission and project presentation (1)

Course assessment

- Homework assignment: 40%
- Project report: 30%
- Project presentation: 30%

Reading list

1. Hagan et al. Neural Network Design, 2nd edition, 2014;
   Online version: https://hagan.okstate.edu/nnd.html
2. Abu-Mostafa et al. Learning from Data, a Short Course, 2012.
3. Mathworks, Neural Network Toolbox User‘s Guide, 2017.

Basics of Machine Learning

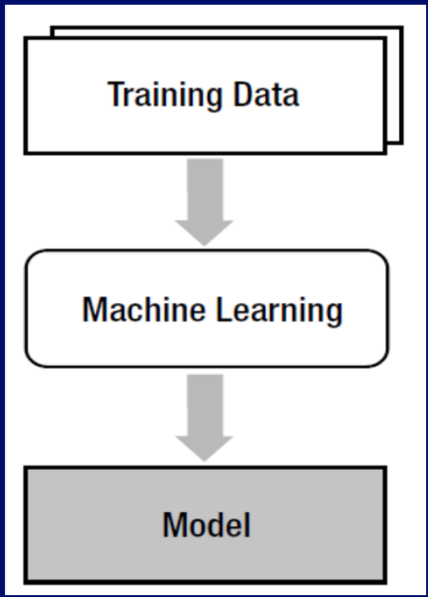
- What is Machine Learning?
- Machine Learning vs. Traditional Programming
- Why do we use Machine Learning?
- How does Machine Learning work?
- Types of Machine Learning
- Challenges of Machine Learning

What is Machine Learning?

- Machine Learning is a modeling technique that figures out the “model” out of “data”.
  $\Rightarrow$ ML receives data as inputs and uses an algorithm to formulate answers.
  - Data: information (tables, documents, audio, images, etc.).
  - Model: ML’s final product (answers / rules).

- A suitable computer algorithm can analyze the data and find the model through self-improvement without being explicitly coded by a programmer.
  - We call it “learning” because the process resembles being trained with the data to solve the problem of finding a model.
  - The data that ML uses in the modeling process is called “training” data.

- What happens in the ML process?



Kim, 2017

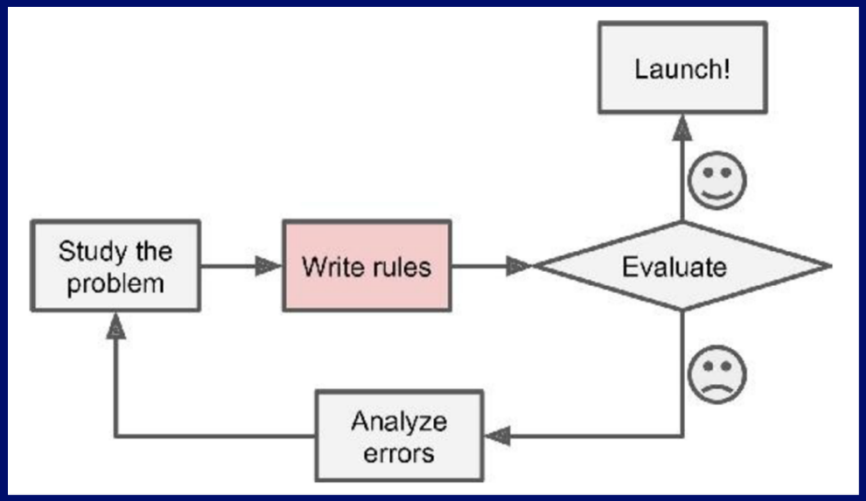
Machine Learning vs. Traditional Programming

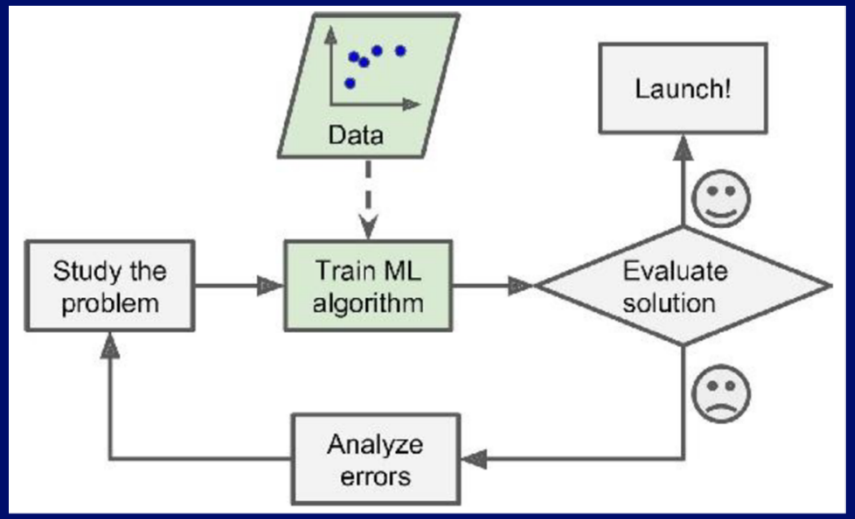
Traditional programming

- A programmer code all the rules. Each rule is based on a logical foundation. The computer program will execute outputs / answers based on input data following the logical statements.
- Problems for complex systems: more rules needed (some times impossible)

Machine learning

- The machine learns from the training data how inputs and outputs / answers are correlated and it find the rules. The programmers do not need to know any rule before.
- The trained rules can be applied for the unseen data.

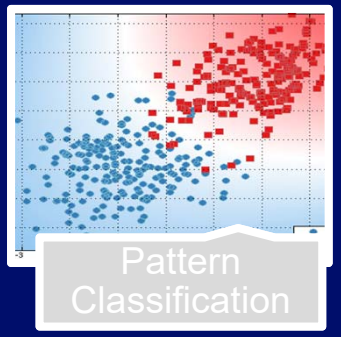




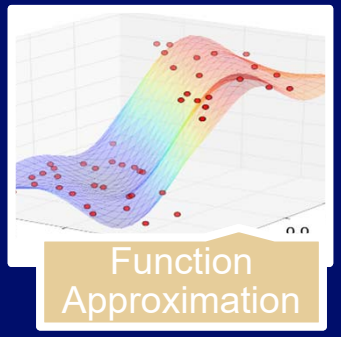
Geron, 2017

Why do we use Machine Learning?

- To derive meaning from complicated or imprecise data,
- To extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques,
- To provide projections given new situations of interest and answer "what if" questions.
- Problems of interest:

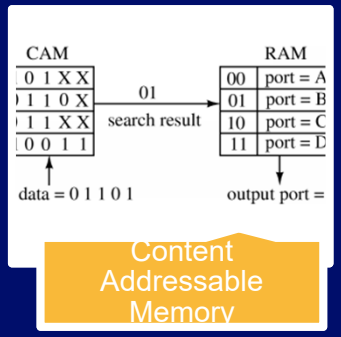

Pattern Classification


Clustering/ Categorization

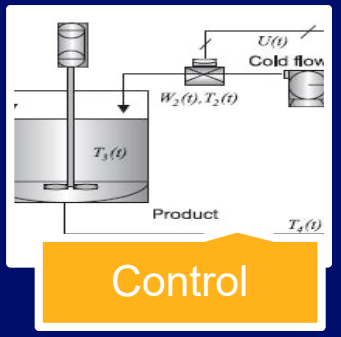

Function Approximation


Prediction/ Forecasting


Optimization

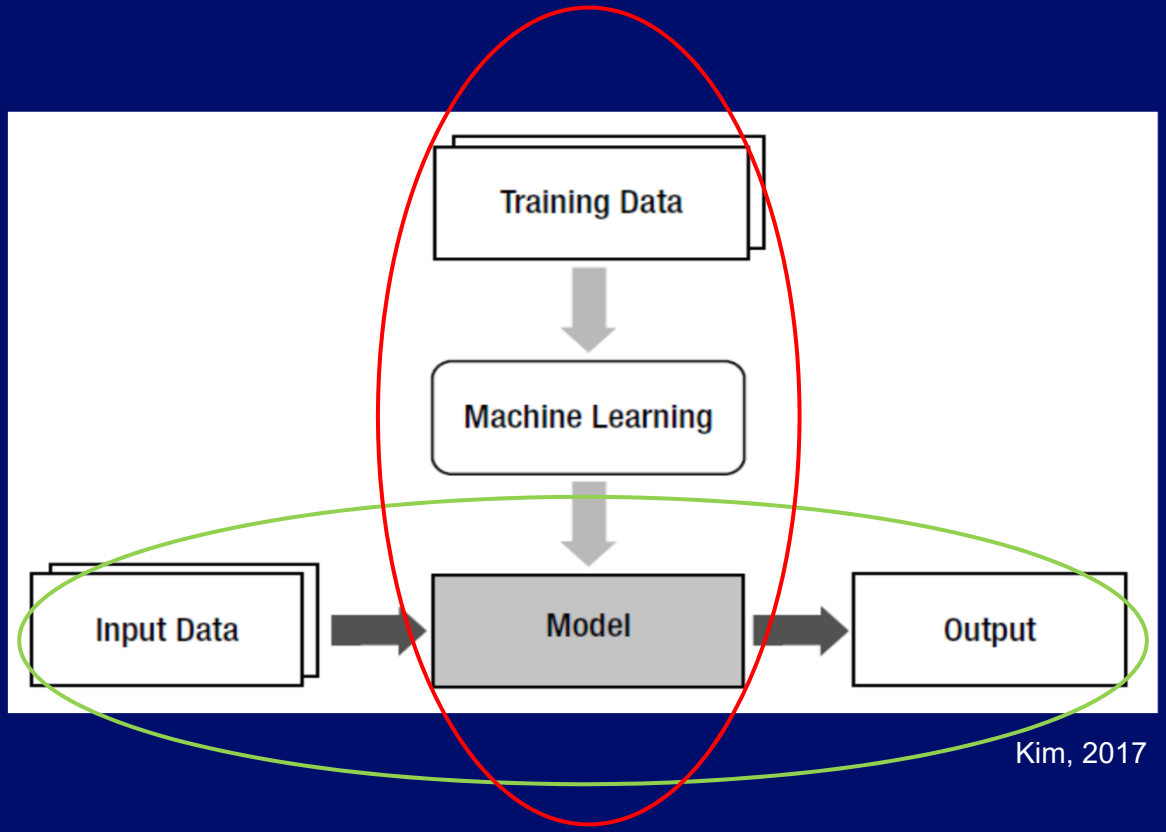


Content Addressable Memory



Control

How does Machine Learning work?

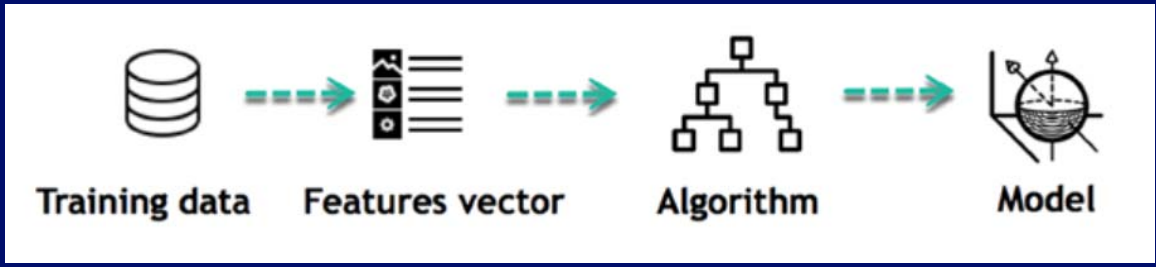


Kim, 2017

The core objective of machine learning is the learning and inferring.

Learning / training

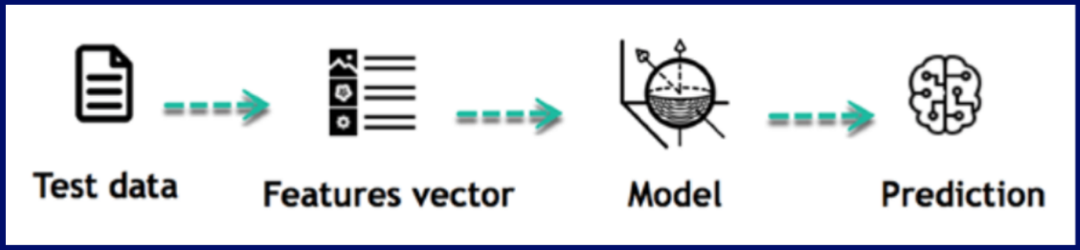
- The machine uses some fancy algorithms to simplify the reality and transform this discovery into a model (called learning phase).
- Therefore, the learning stage is used to describe the data and summarize it into a model.



Inferring

- When the model is built, it is possible to test how powerful it is on never-seen-before data (test data). The new data are transformed into a features vector, go through the model and give a prediction.
- There is no need to update the rules or train again the model

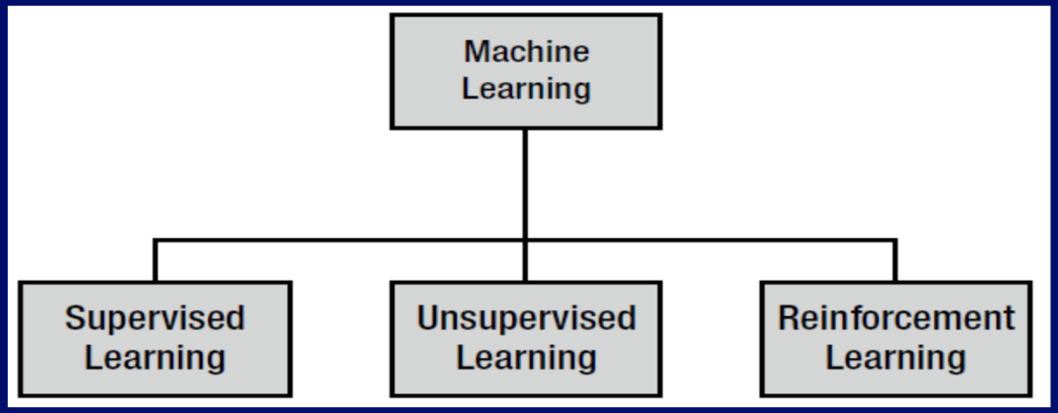
  ⇒ You can use the trained model to make inference on other new data.



- The Machine Learning process is straightforward and can be summarized in the following steps:
    1. Define a problem
    2. Collect data
    3. Discover and visualize data
    4. Prepare data for ML algorithms
    5. **Select a model and train it**
    6. Test the model
    7. Collect feedback
    8. Refine the algorithm
    9. Loop 4-7 until the results are satisfying
    10. Use the model to make a prediction
- Once the algorithm gets good at drawing / statistic analyzing the right answers, it applies that knowledge to new sets of data.

Types of Machine Learning

- Many different types of ML techniques have been developed to solve problems in various fields.
- These techniques can be classified into three types depending on the training method.

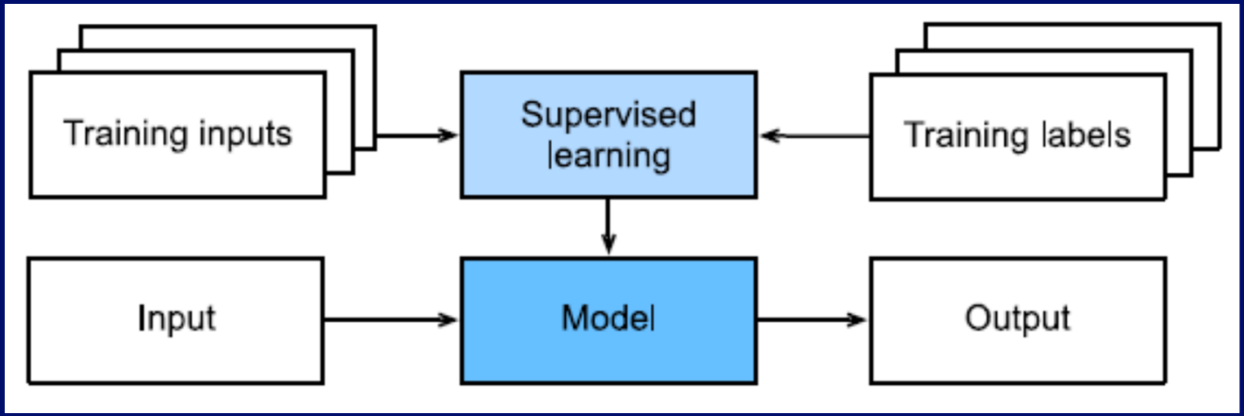


- The choice of the ML algorithm is based on the study objective.

Supervised learning

- In supervised learning, each training dataset should consist of input(s) (*feature*) and correct output (*label*) pairs (*called labeled data*):

  { input, correct output }
- An algorithm uses training data and feedback from humans to learn the relationship of given inputs to a given output (*target*).
- Learning is the series of revisions of a model to reduce the difference between the correct output (*target*) and the output from the model for the same input.
- If a model is perfectly trained, it will produce a correct output that corresponds to the input from the training data.



- There are two typical tasks of supervised learning: Classification and Regression

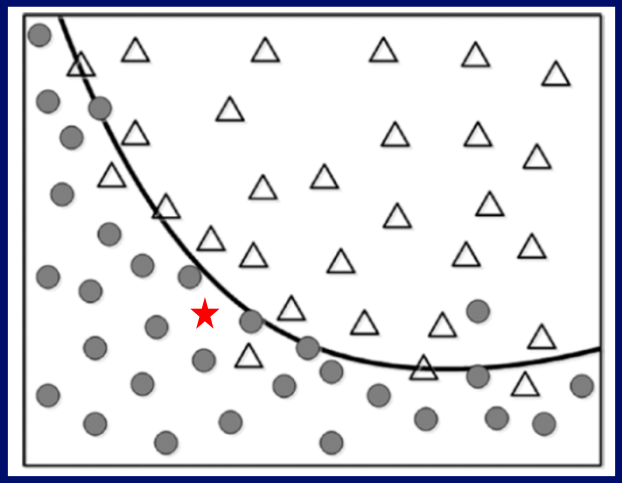
Classification

- The classification problem focuses on literally finding the classes to which the data belongs ⇒ the data pair has the class in place of the correct output corresponding to the input:

  { input, categorical output }

- The label can be of two or more classes.

$\{X_1,\ Y_1, \Delta\}$
$\{X_2,\ Y_2, \bullet\}$
…
$\{X_N,\ Y_N, \bullet\}$



Example: input with two numeric features (X, Y) and label with two classes (• , Δ)

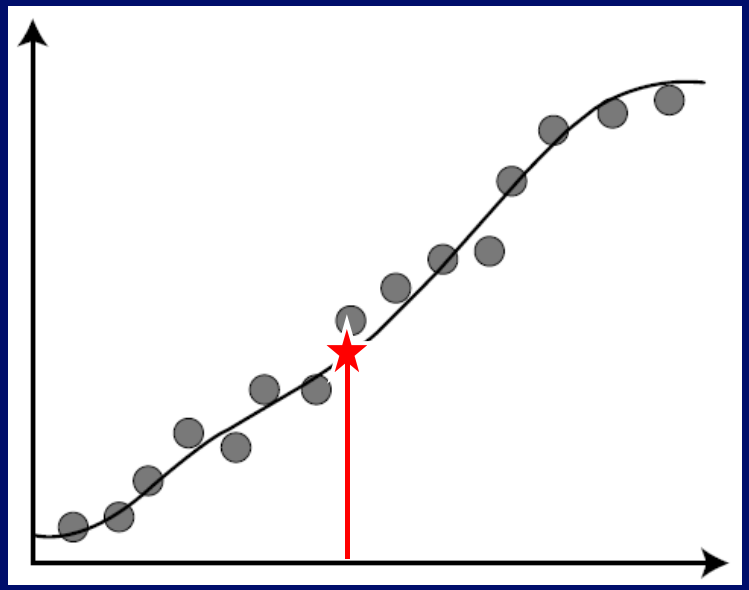
Regression

- When the output is a continuous value, the task is a regression.

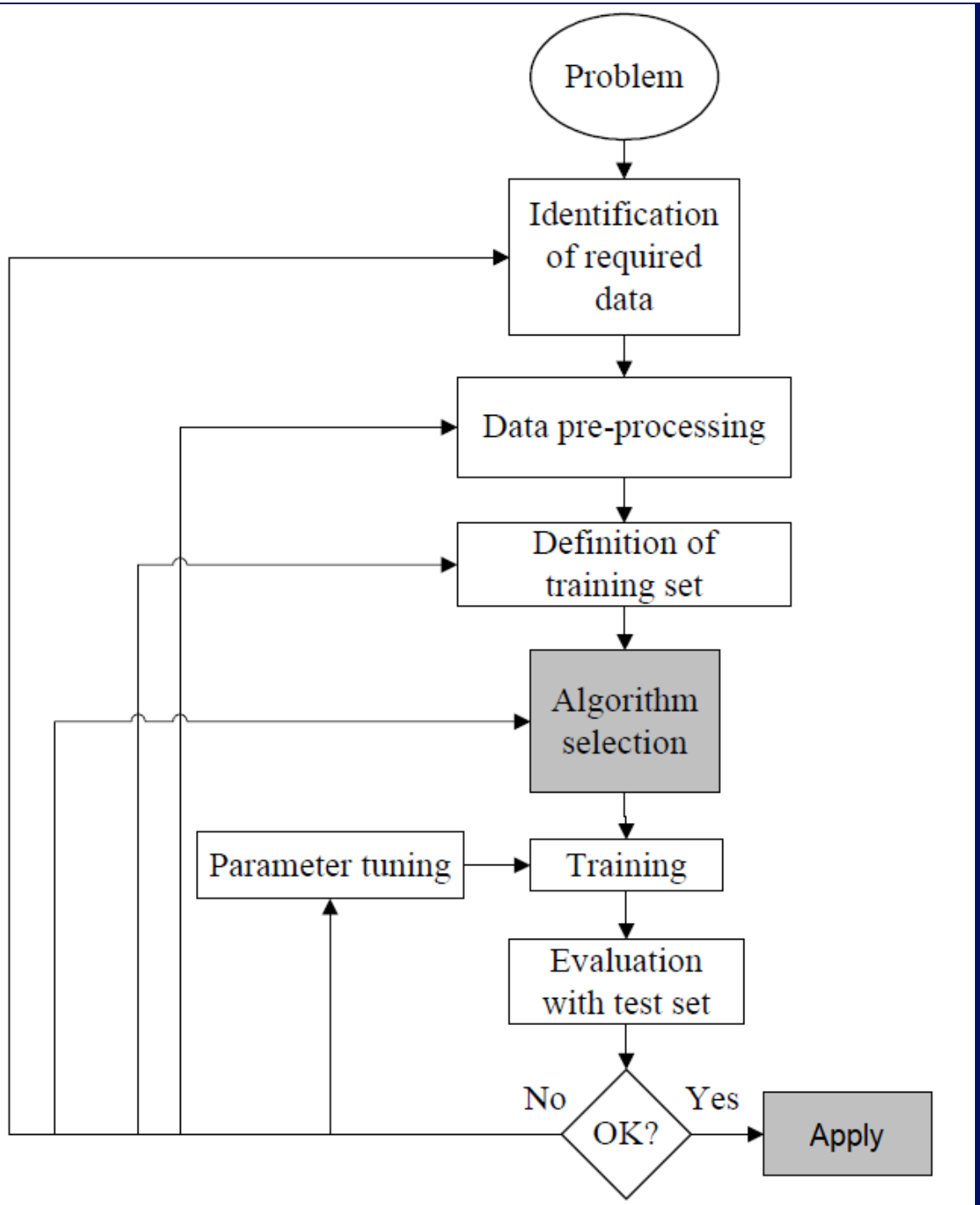
  { input, numeric output }

- The system will be trained to estimate Y with the lowest possible error.

$\{X_1, Y_1\}$
$\{X_2, Y_2\}$
...
$\{X_N, Y_N\}$



Example: input with one feature X and label with value Y



The process of supervised ML

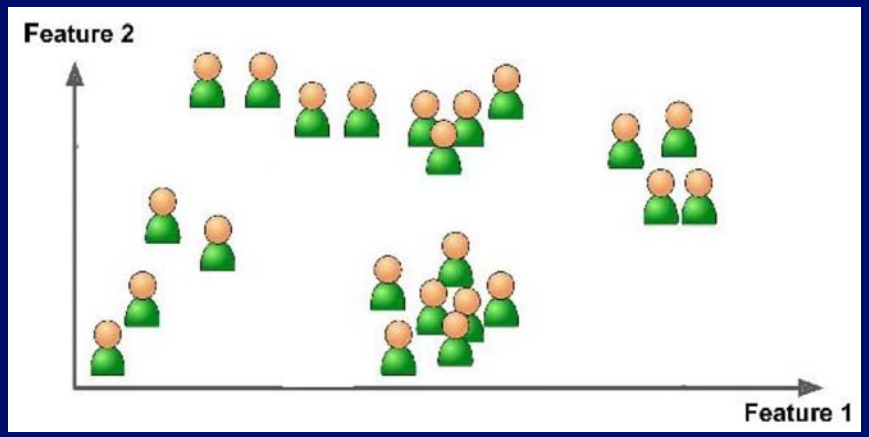
Algorithm	Description	Type
Linear/nonlinear regression	Finds a way to correlate each feature to the output to help predict future values.	Regression
Logistic regression	Extension of linear regression that's used for classification tasks. The output variable is binary (e.g., only black or white) rather than continuous (e.g., an infinite list of potential colors)	Classification
Decision tree	Highly interpretable classification or regression model that splits data-feature values into branches at decision nodes (e.g., if a feature is a color, each possible color becomes a new branch) until a final decision output is made	Regression Classification
Support vector machine	is typically used for the classification task. It finds a hyperplane that optimally divided the classes. It is best used with a non-linear solver.	Regression (not very common) Classification
Random forest	is built upon a decision tree to improve the accuracy drastically. It generates many times simple decision trees and uses the 'majority vote' method to decide on which label to return. For the classification task, the final prediction will be the one with the most vote; while for the regression task, the average prediction of all the trees is the final prediction.	Regression Classification

Unsupervised learning

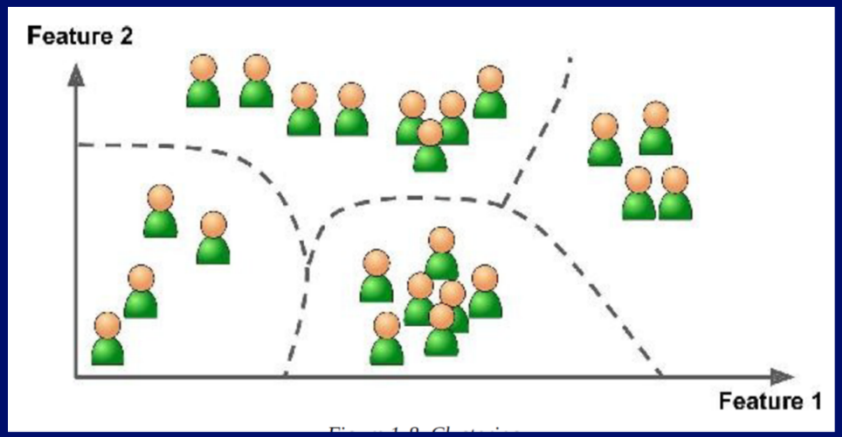
- In unsupervised learning, an algorithm explores input data without being given an explicit output.

{ input }

- It finds data patterns and then classify the data.
- There are two categories of unsupervised learning:
  - ✓ Clustering task
  - ✓ Dimension Reduction task



Unlabeled Data



Clustering

Algorithm	Description	Type
K-means clustering	Puts data into some groups (k) that each contains data with similar characteristics (as determined by the model, not in advance by humans)	Clustering
Gaussian mixture model	A generalization of k-means clustering that provides more flexibility in the size and shape of groups (clusters)	Clustering
Hierarchical clustering	Splits clusters along a hierarchical tree to form a classification system. Can be used for Cluster loyalty-card customer	Clustering
Recommender system	Helps to define the relevant data for making a recommendation.	Clustering
PCA/T-SNE	Mostly used to decrease the dimensionality of the data. The algorithms reduce the number of features to 3 or 4 vectors with the highest variances.	Dimension Reduction

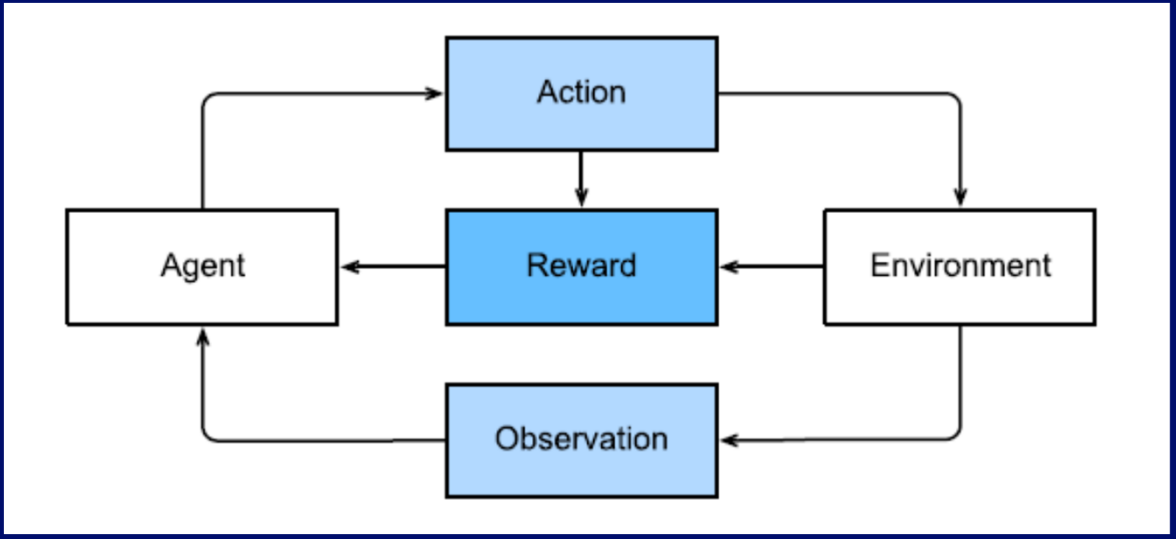
Reinforcement learning

- RL is a very different beast. The learning system, called an agent , can observe the environment, select and perform actions, and get rewards in return (or penalties in the form of negative rewards).
- It must then learn by itself what is the best strategy, called a policy, to get the most reward over time.
- A policy defines what action the agent should choose when it is in a given situation.
- RL employs sets of input, some output, and grade as training data:

  { input, some output, grade for this output }

- It is generally used when optimal interaction is required, such as control and game plays, robotics, fish behavior modelling ...

The interaction between reinforcement learning and an environment

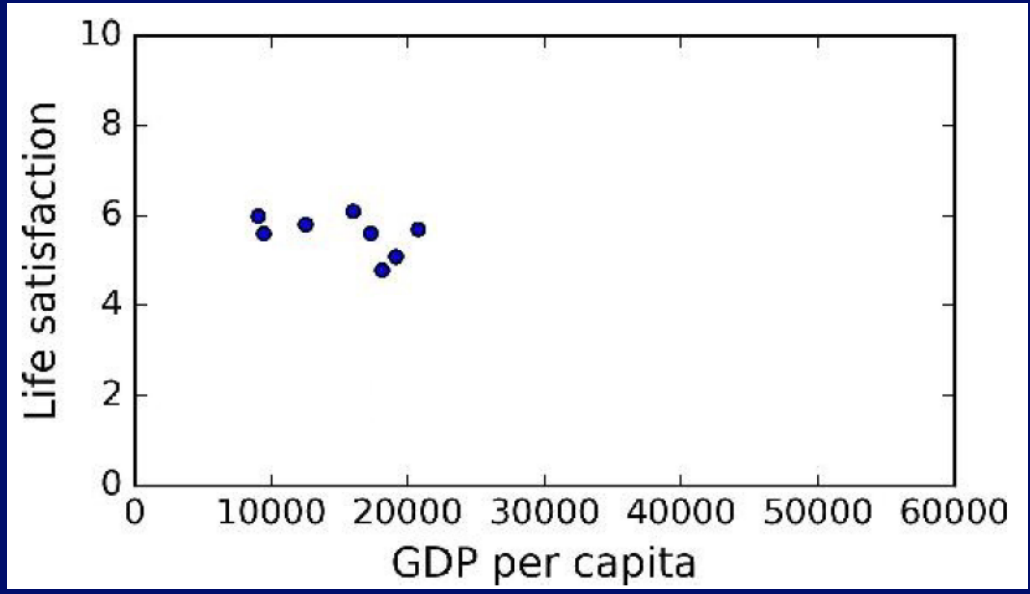


Main Challenges of Machine Learning

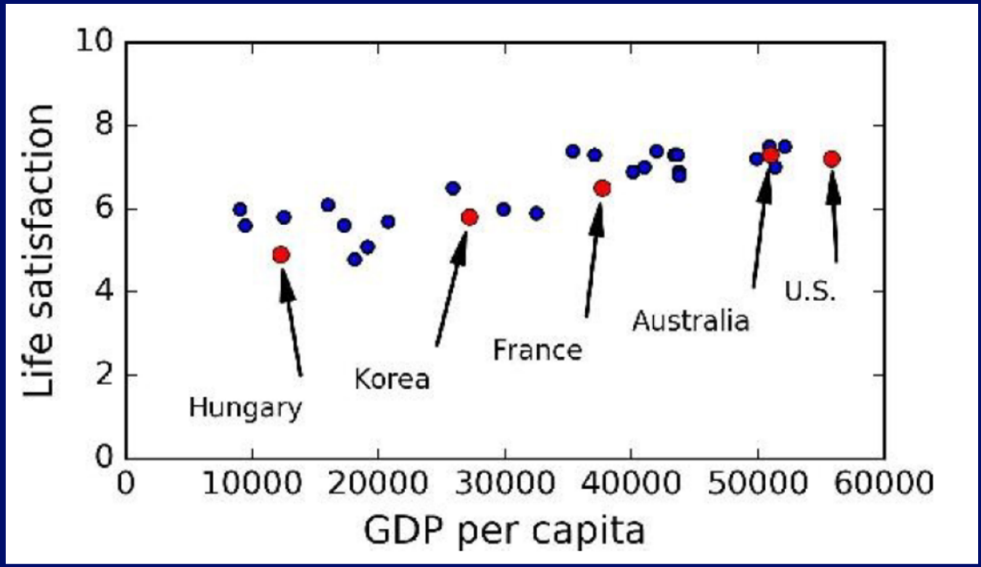
Data

1. **The primary challenge of machine learning is the insufficient quantity of training data.**

Example: we want to know if money makes people happy.

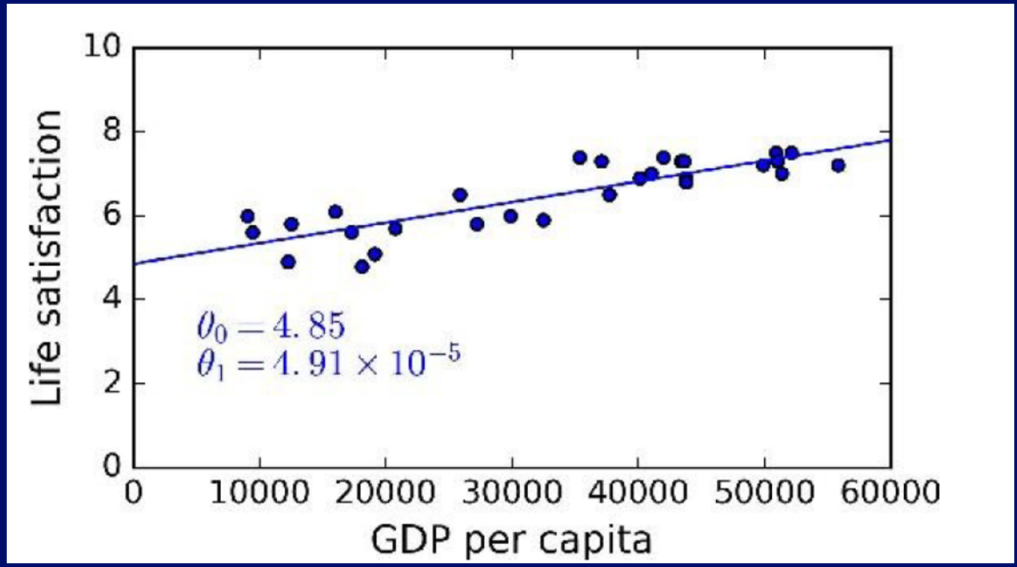


Very few training data

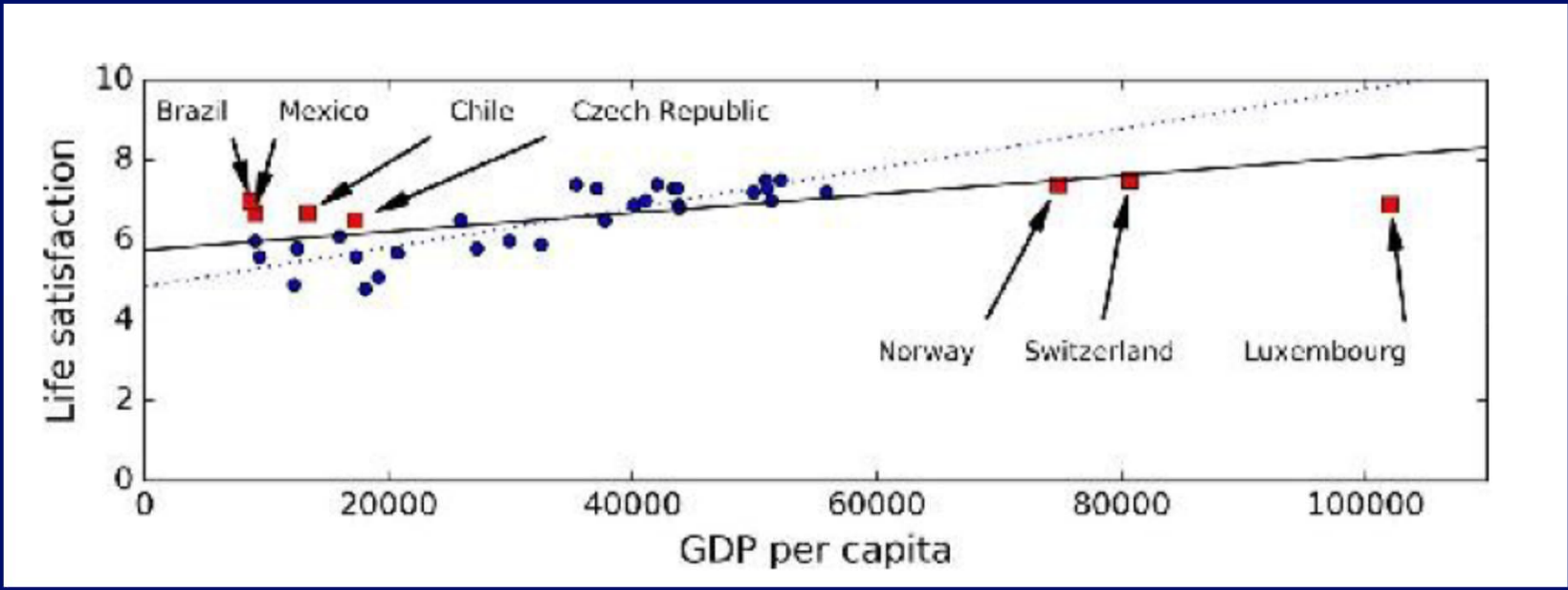


Country	GDP per capita (USD)	Life satisfaction
Hungary	12,240	4.9
Korea	27,195	5.8
France	37,675	6.5
Australia	50,962	7.3
United States	55,805	7.2

More training data



2. Non-representative training data



A more representative training sample

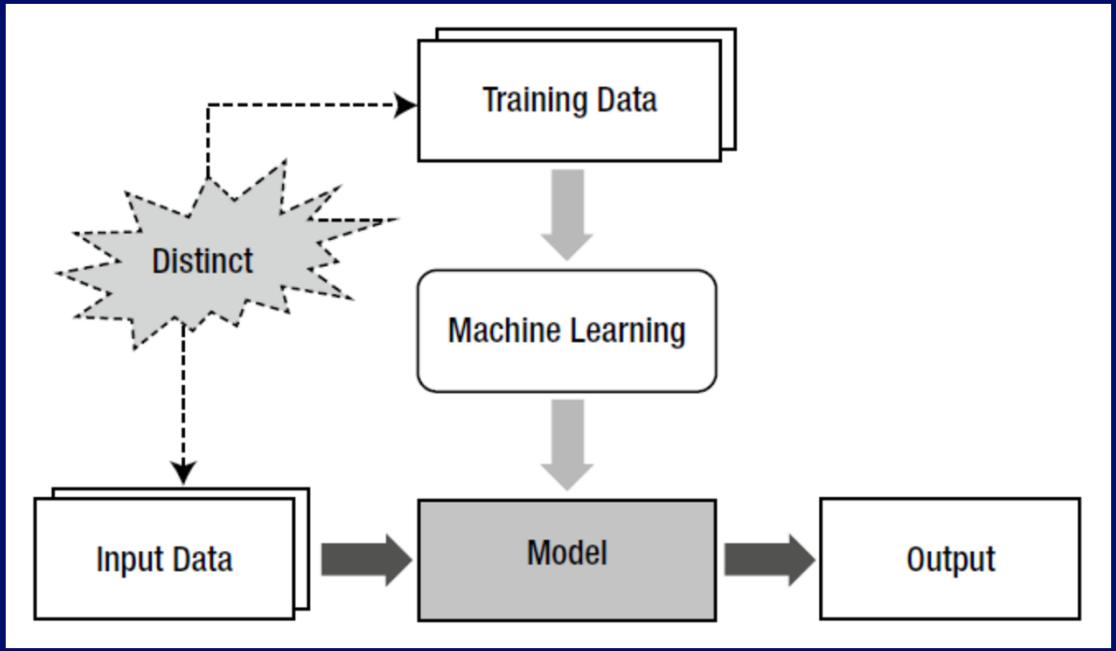
3. Poor-Quality Data

- As training data is full of errors, outliers, and noise (e.g., due to poor-quality measurements), it will make difficulty for the model to detect the underlying patterns $\Rightarrow$ reduce the model performance.
  - ➢ First, clean up the training data.
- If some samples are missing a few features, we must decide whether we want to ignore this attribute altogether, ignore these samples, fill in the missing values (e.g., with the median value), or train one model with the feature and one model without it, and so on.

4. Irrelevant Features

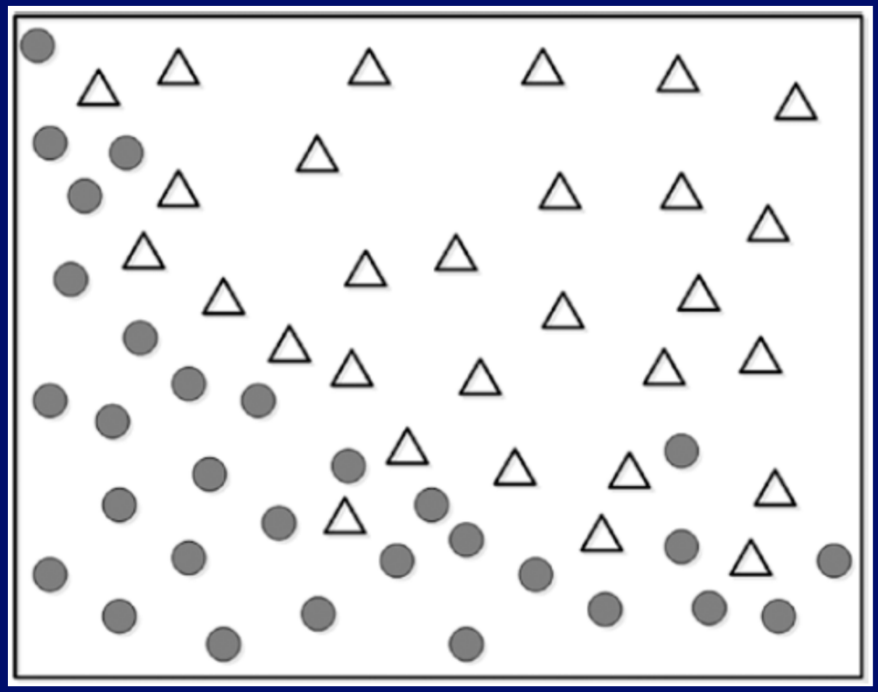
- System can learn if the training data contains enough relevant features and not too many irrelevant ones.
- A critical part of the success of a Machine Learning project is coming up with a good set of features to train on. This process, called feature engineering, involves:
  - Feature selection: selecting the most useful features to train on among existing features.
  - Feature extraction: combining existing features to produce a more useful one (dimensionality reduction can help).
  - Creating new features by gathering new data.

5. **Very large distinctness of the training data and application input-data**



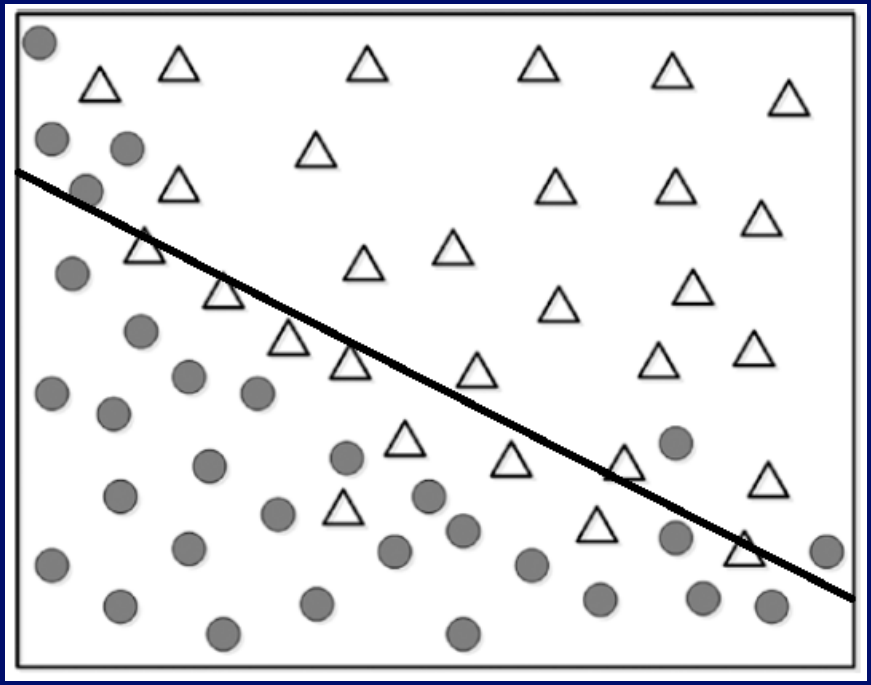
Kim, 2017

Algorithm

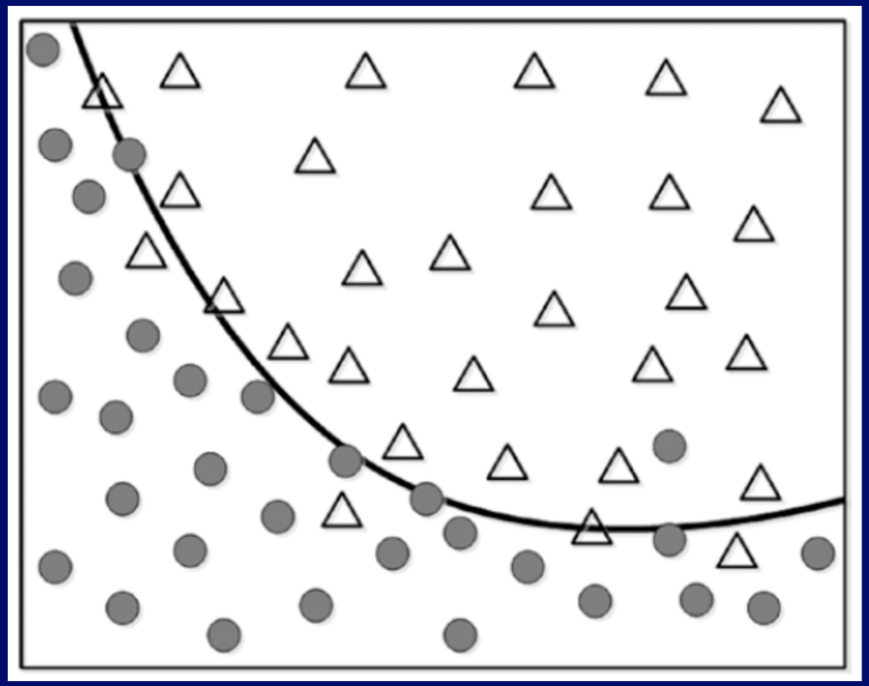


Training data to determine a curve to divide two groups of data

1. **Underfitting the training data**

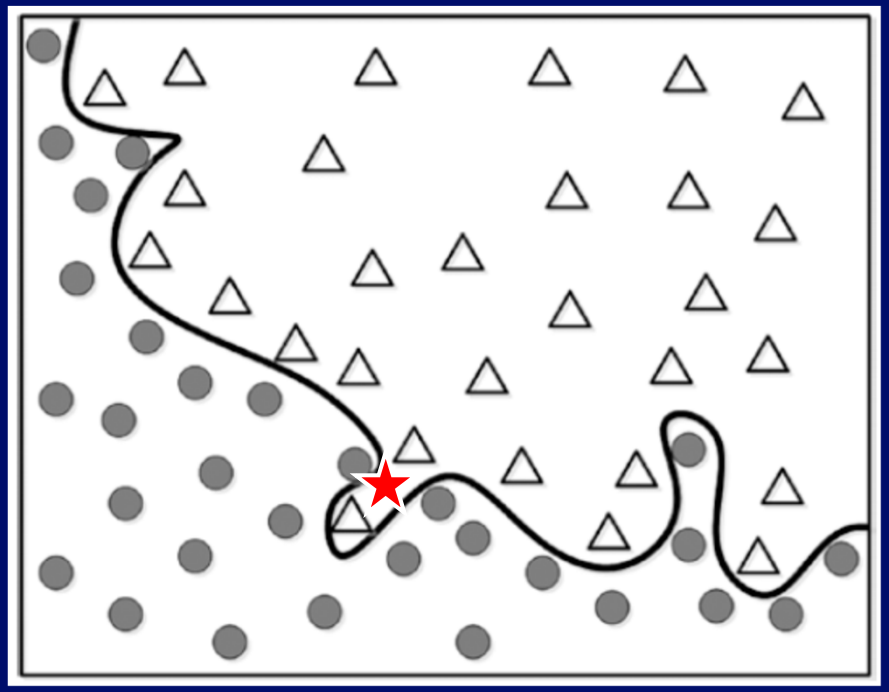


Model is too simple to learn the underlying structure of the data



Curve to differentiate between two types of data (normal fitting)

2. Overfitting the training data



A complex model strongly overfits the training data.

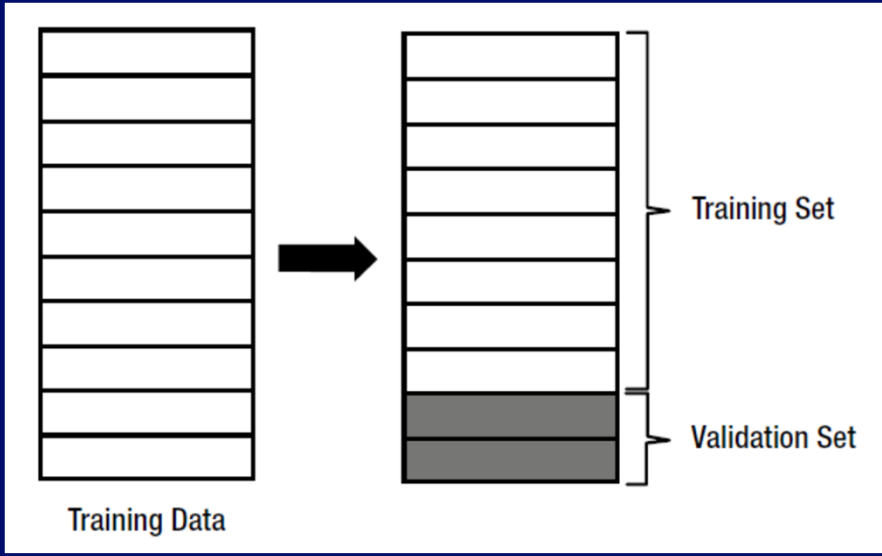
Confronting Overfitting

Overfitting happens when the model is too complex relative to the amount and noisiness of the training data. The possible solutions are:

- **Data preprocessing**:
  - To gather more training data
  - To reduce the noise in the training data (e.g., fix data errors and remove outliers)

- **Regularization:** a numerical method attempts to construct a model structure as simple as possible. The simplified model can avoid the effects of overfitting at the small cost of performance.
  - To simplify the model by selecting one with fewer parameters (e.g., a linear model rather than a high-degree polynomial model), by reducing the number of attributes in the training data or by constraining the model

- **Validation**: a process reserves a part of the training data and uses it to monitor the performance. The validation set is not used for the training process. Because the modeling error of the training data fails to indicate overfitting, we use some of the training data to check if the model is overfitted.

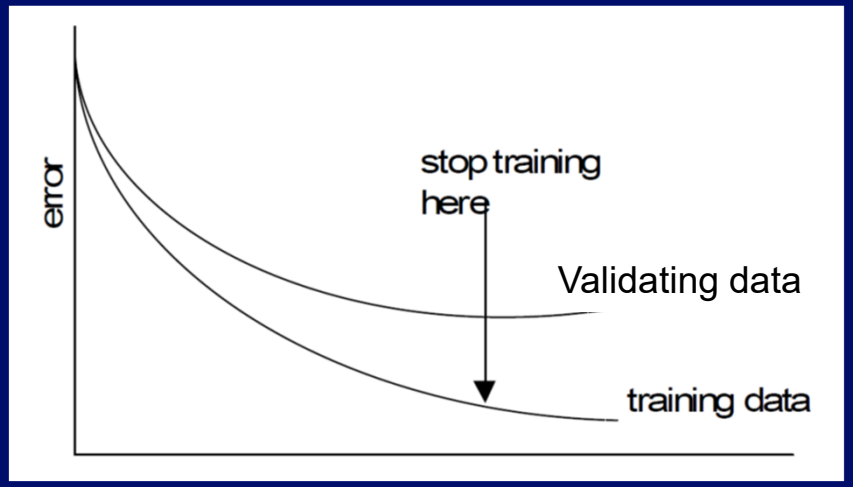


When validation is involved, the supervised training process proceeds by the following steps:

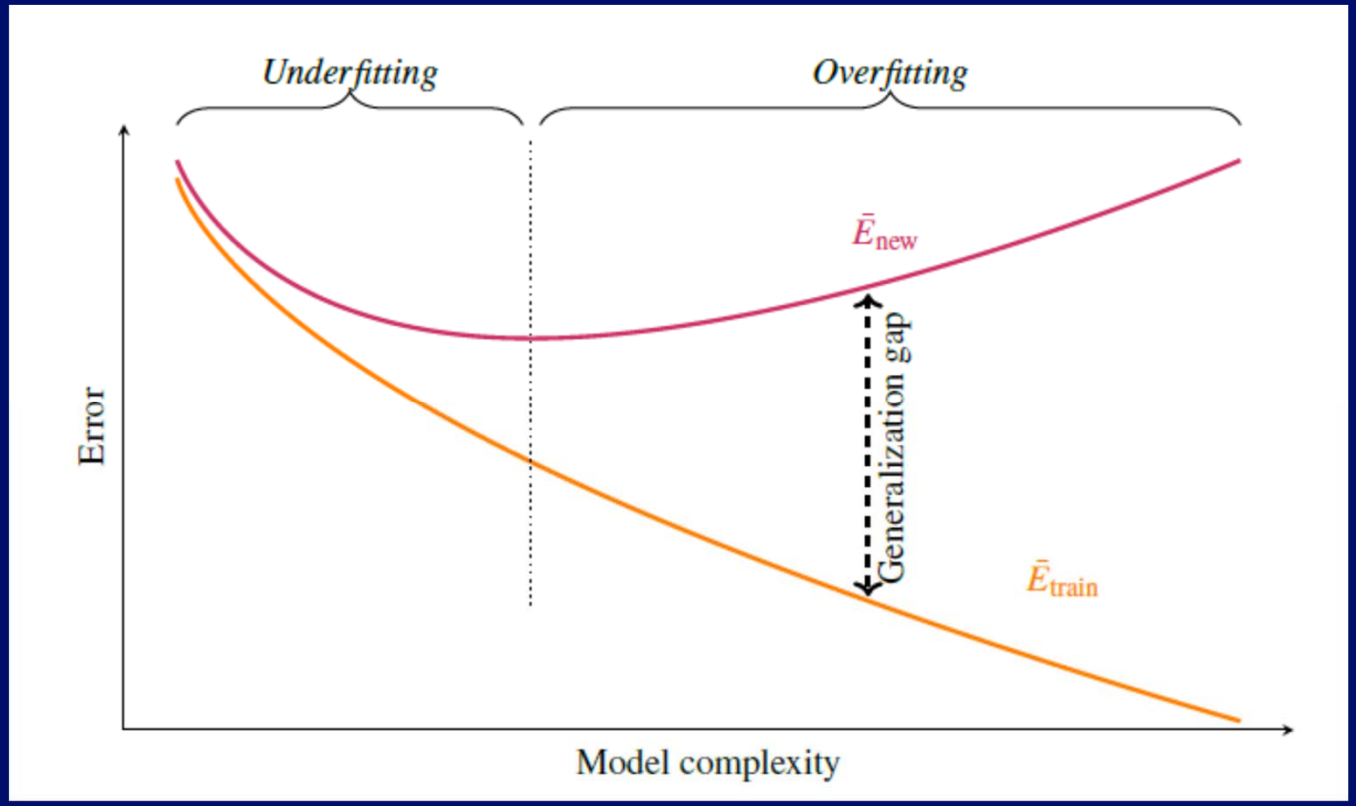
1. Divide the training data into two subsets: one for training and the other for validation.
2. Train the model with the training subset.
3. During the training we evaluate the model performance on the validation subset. The error in this set should also fall but be higher than that of the training set.

- When the error in the validation set data stops falling (perhaps even starts to rise), the point at which overtraining has started has been reached. If training is stopped at this point, overtraining can be avoided.

4. If the performance on two subsets does not produce sufficient results, we modify the model and repeat the process from Step 2.



Early Stopping



Behavior of E_{train} and E_{new} for supervised ML methods as a function of model complexity

Evaluation Metrics for Machine Learning

- Machine learning is about building a predictive and accurate model.
- Supervised learning: have information about the labels
  - Evaluation metrics RMSE, MAE, R-Squared, etc. for regression,
  - Confusion matrix and metrics drawn from it for classification.
- Unsupervised learning: don’t have the label information, have different evaluation metrics according to their outputs, e.g. Silhouette coefficient, Dunn’s index for clustering problems.

Homework 1
(deadline for submission is 10th May 2022)

Answer the following questions:

1. How would you define Machine Learning?
2. What is a labeled training set?
3. What are the two most common supervised tasks?
4. Could you use unsupervised learning algorithm to allow a robot to walk in various unknown terrains? Justify your answer.
5. What type of learning algorithm would you use to segment your hydraulic data into multiple groups?
6. Can you name seven of the main challenges in Machine Learning?
7. If your model performs great on the training data but generalizes poorly to new dataset, what is happening? Can you name three possible solutions?
8. What is a test dataset and why would you want to use it?
9. What is the purpose of a validation dataset?
10. What can you do when your model performs poorly on the training dataset ?