ĐẠI HỌC ĐÀ NẴNG
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG VIỆT - HÀN
VIETNAM - KOREA UNIVERSITY OF INFORMATION AND COMMUNICATION TECHNOLOGY
한-베정보통신기술대학교
Nhân bản – Phụng sự – Khai phóng

Chapter 1

Overview of Machine Learning

Machine Learning

CONTENTs

- Introduction to ML
- Types of ML Systems
- Challenges of ML
- Testing & Validating

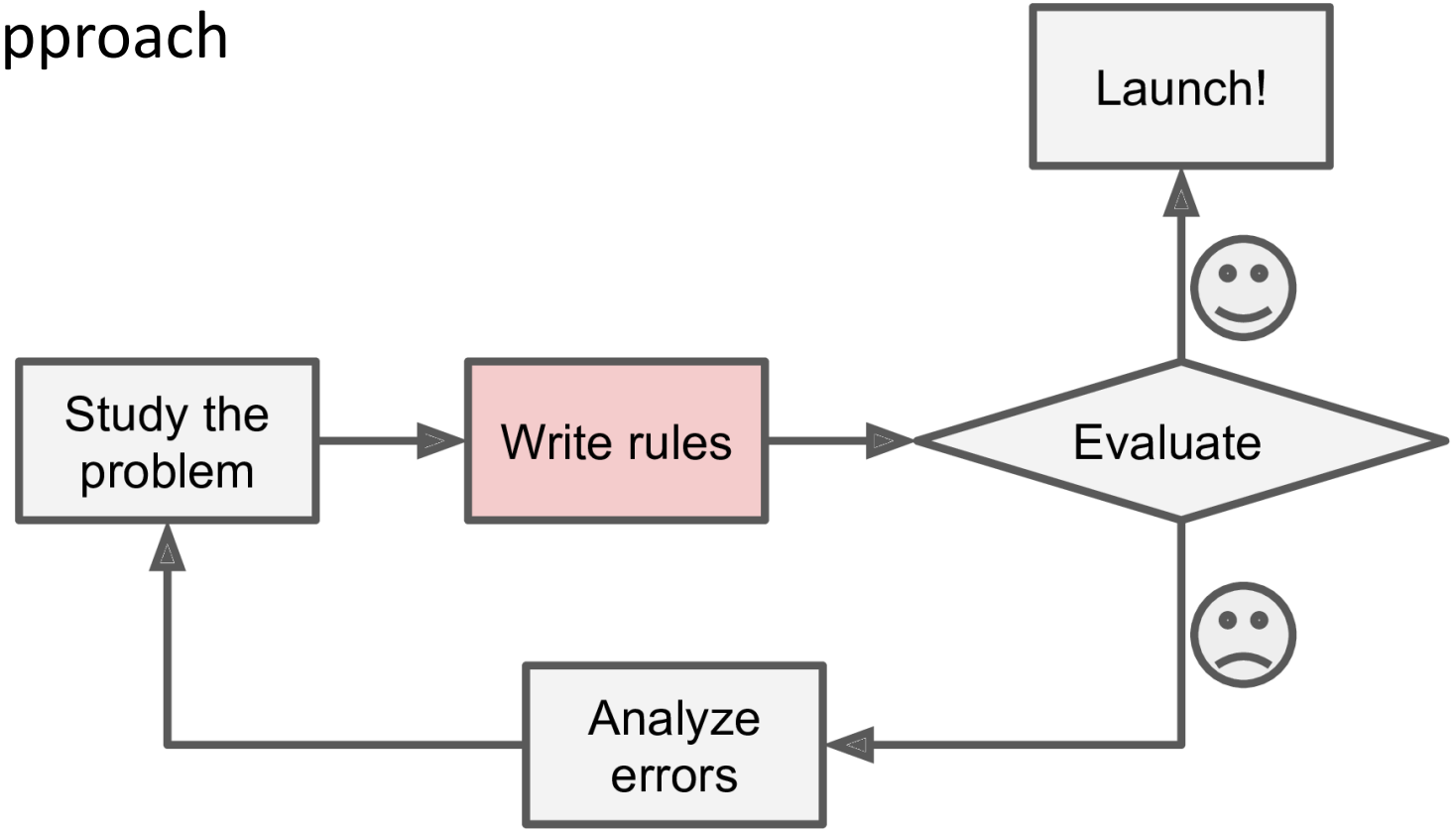
CONTENTs

- **Introduction**
- Types of ML Systems
- Challenges of ML
- Testing & Validating

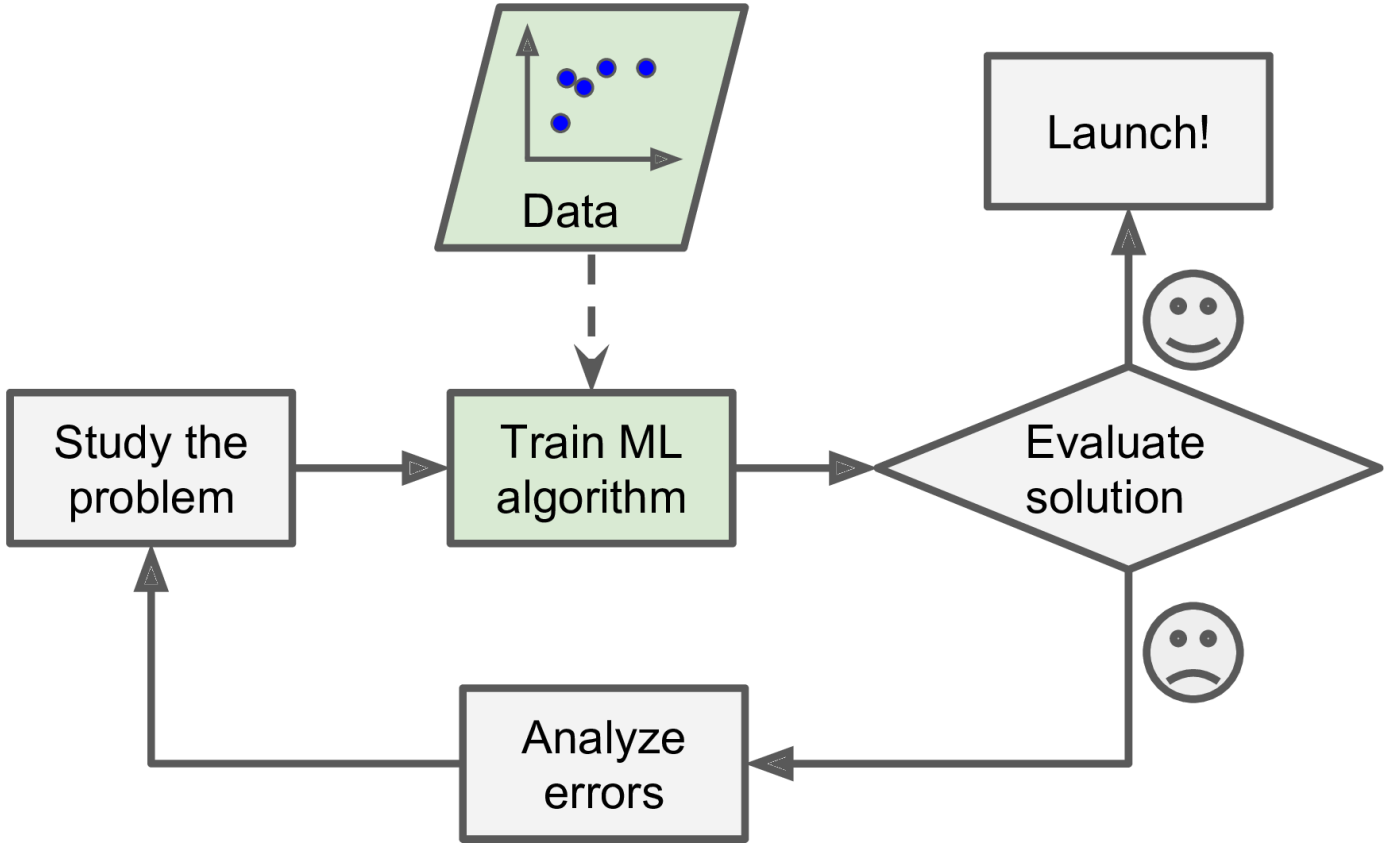
- **What is ML?**
  - ML is the science (& art) of programming computers so they can learn from data.
  - "ML is the field of study that gives computers the ability to learn without being explicitly programmed" - [Arthur Samuel, 1959]
  - "**A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E**" – [Tom Mitchell, 1997]

    e.g: **spam filter**
    - task T ⇔ flag spam for new emails
    - experience E ⇔ training data
    - performance measure P ⇔ ratio of correctly classified emails (*accuracy*)

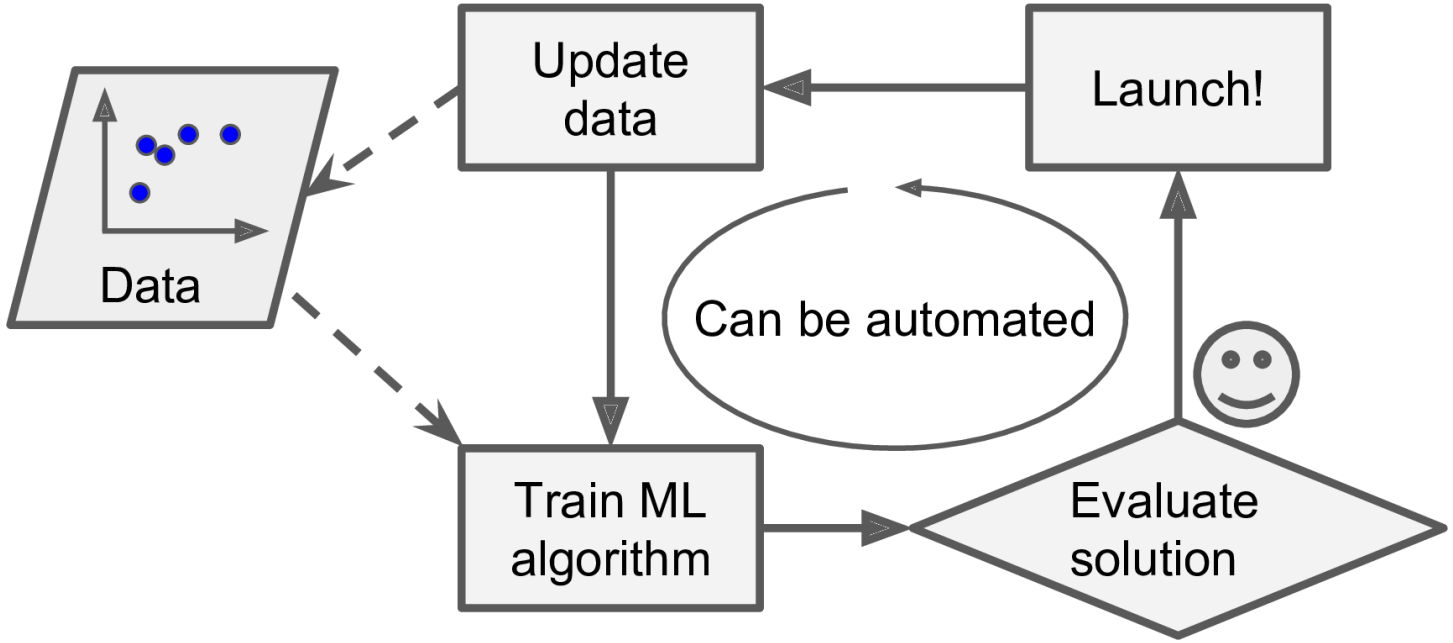
- **Why use ML?**
  - Traditional approach



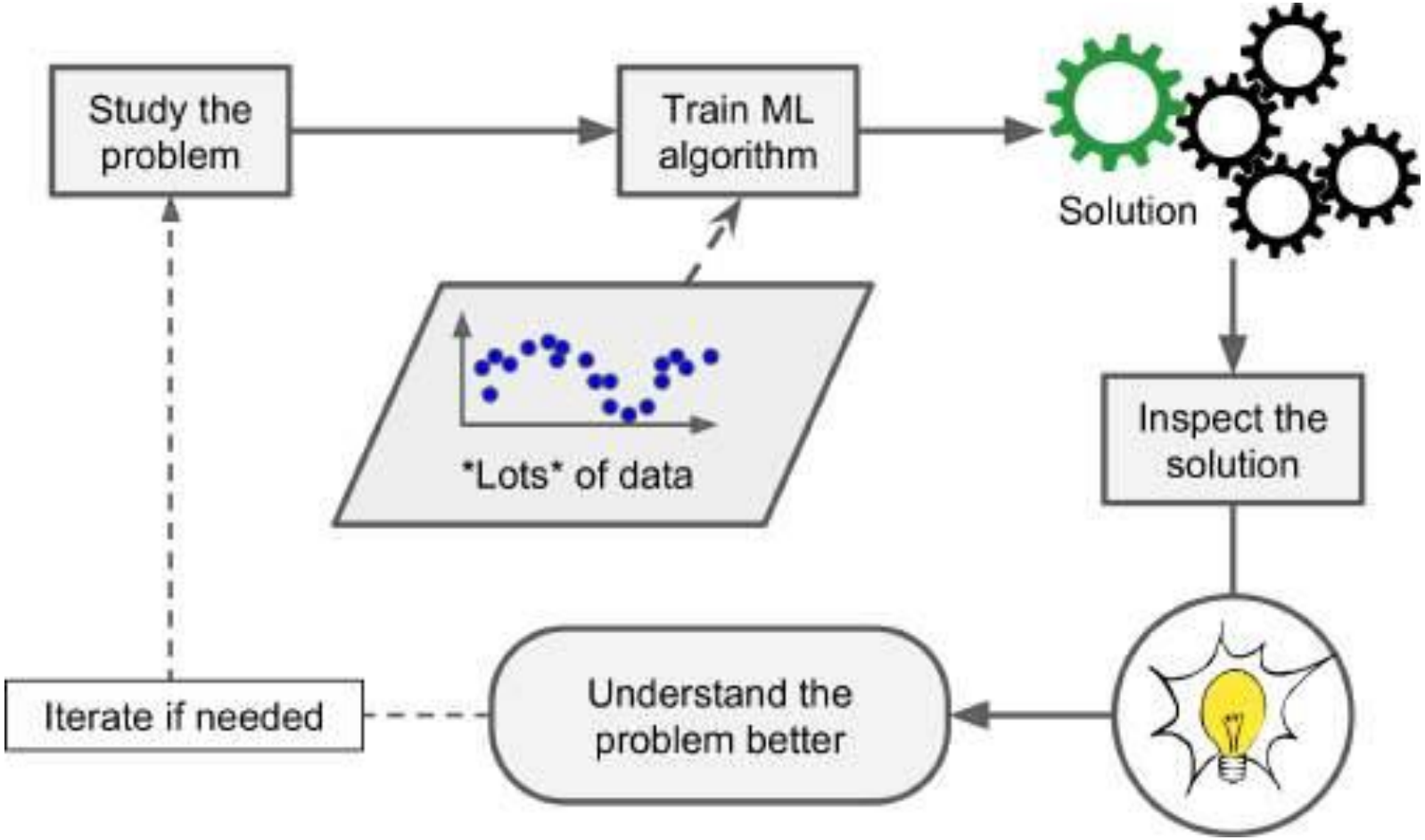
- **Why use ML?**
  - Machine Learning approach



- **Why use ML?**
  - Automatically adapting to change

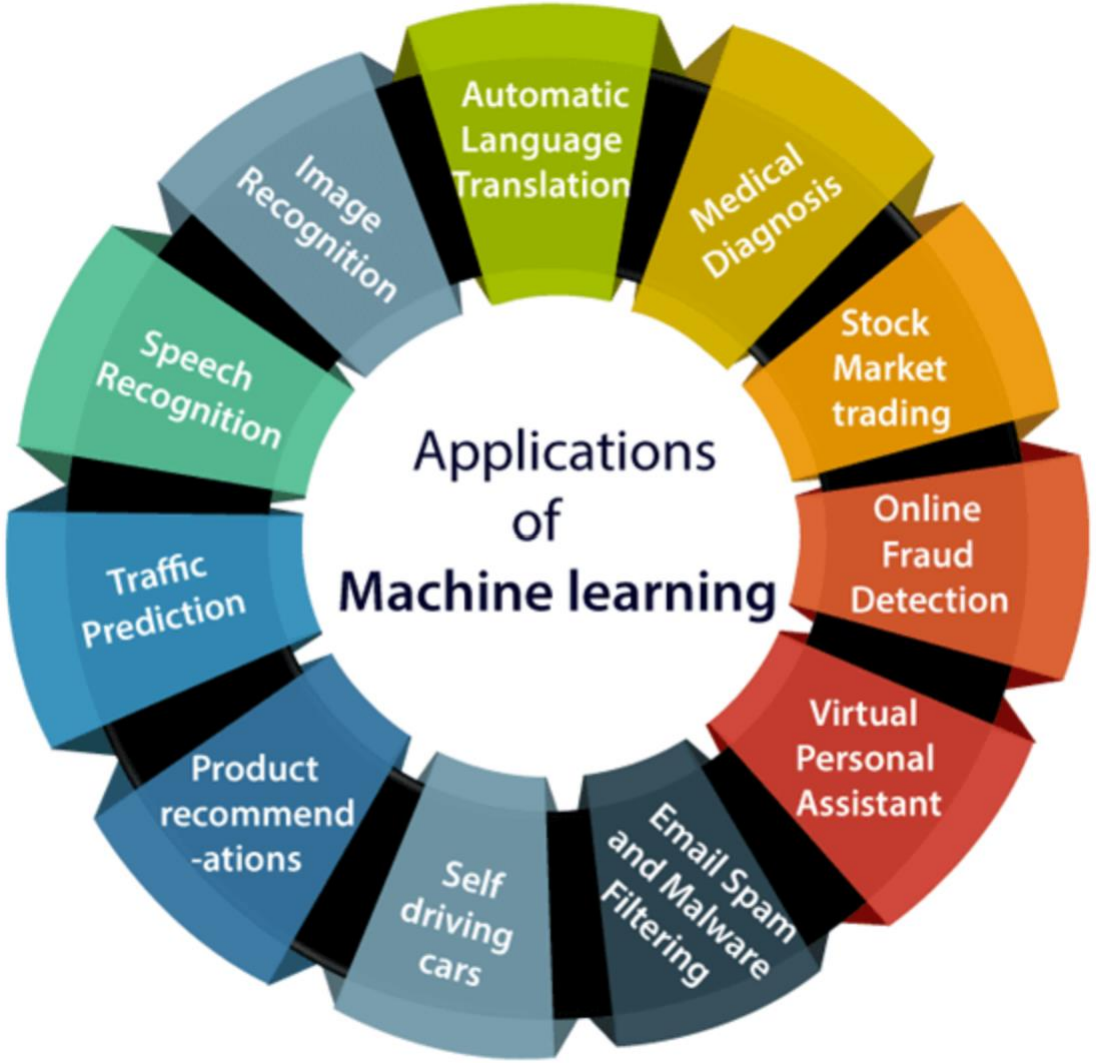


- **Why use ML?**
  - Machine Learning can help humans learn



- **Machine Learning is great for**
  - Problems require a lot of hand-tuning or long lists of rules
    ⇨ ML algorithm can often simplify code & perform better.
  - Complex problems, no good solution by using a traditional approach
    ⇨ ML techniques can find a solution.
  - Fluctuating environments ⇨ ML system can adapt to new data.
  - Getting insights about complex problems and large amounts of data.

- **Applications of ML**
  - Image Recognition
  - Speech Recognition
  - Sentiment Analysis
  - Traffic prediction
  - Product recommendations
  - Self-driving cars
  - Email Spam and Malware Filtering
  - Medical Diagnosis
  - Automatic Language Translation
  - …



CONTENTs

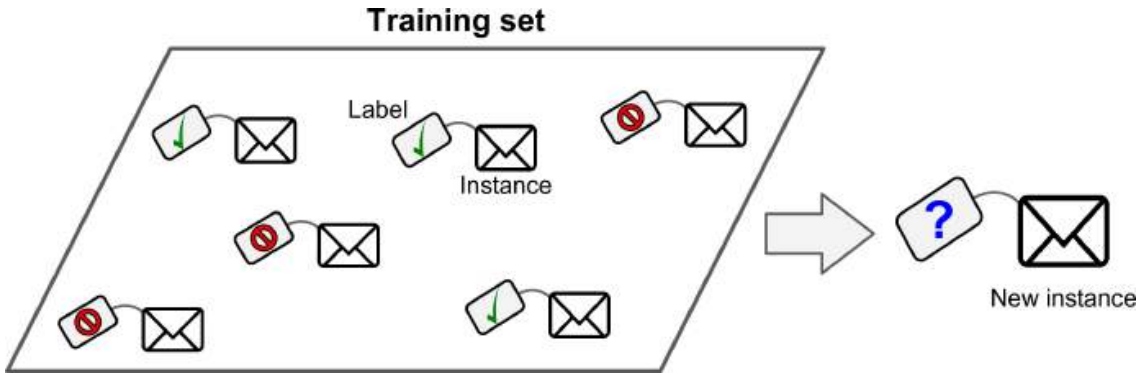
Types of ML Systems

Many different types of ML systems, classify them in categories based on:

- **Supervised** >< **Unsupervised Learning**
- **Batch** >< **Online Learning**
- **Instance-Based** >< **Model-Based Learning**

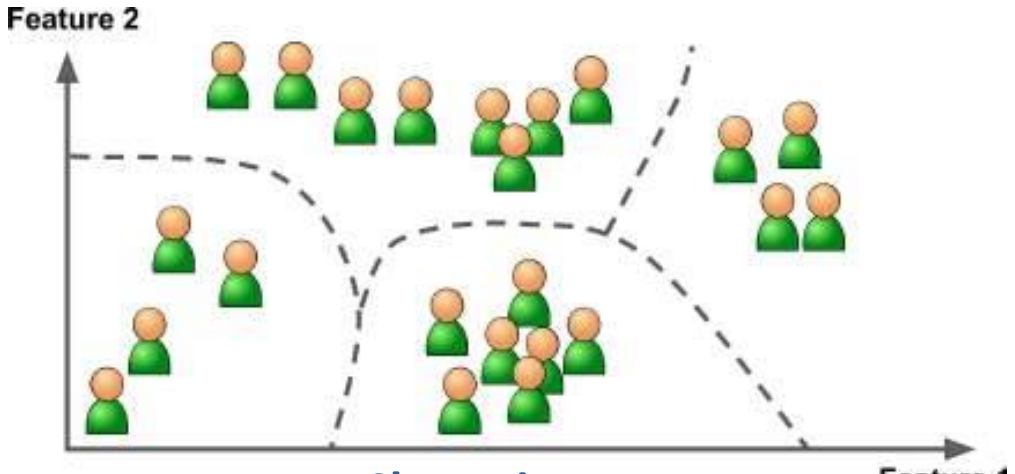
- **Supervised >< Unsupervised Learning**
  - ML systems can be classified according to the amount and type of supervision they get during training.
  - There are four major categories:
    - Supervised learning
    - Unsupervised learning
    - Semisupervised learning
    - Reinforcement Learning

- **Supervised**
  - Supervised learning: the training data you feed to the algorithm includes the desired solutions, called labels
  - 2 types of supervised learning:
    - Regression
    - Classification
  - Important supervised learning algorithms:
    - k-Nearest Neighbors
    - Linear Regression
    - Logistic Regression
    - Support Vector Machine (SVM)
    - Decision Trees and Random Forests
    - Neural networks



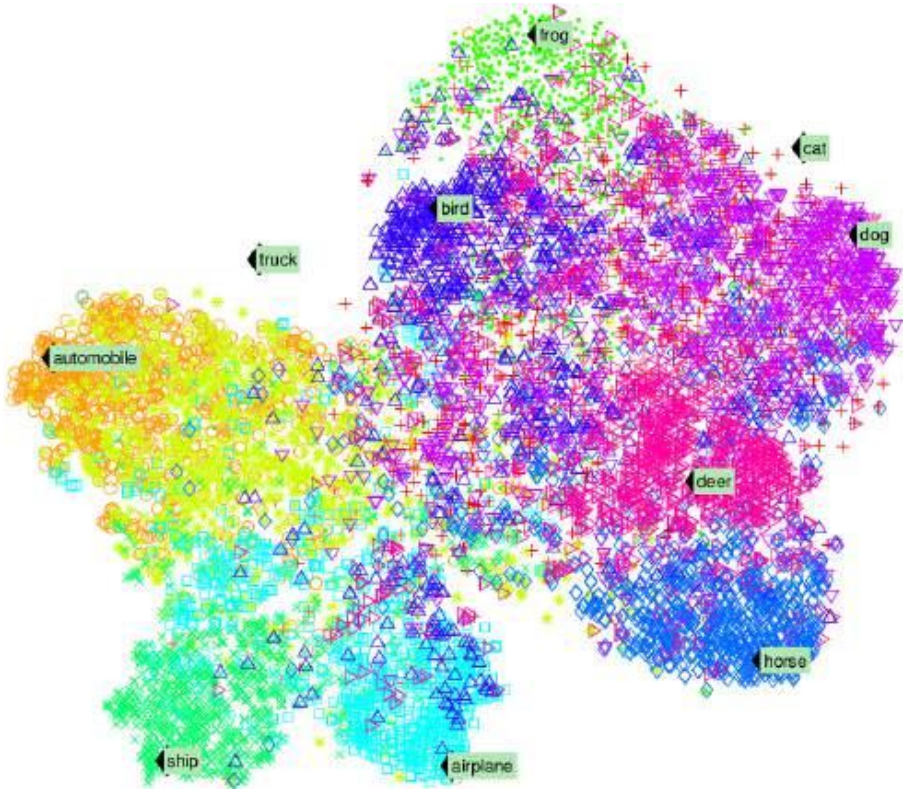
- **Unsupervised Learning**
  - Unsupervised learning: the training data is unlabeled (the system tries to learn without a teacher).
  - Important Unsupervised learning algorithms:
    - Clustering: K-Means, DBSCAN, Hierarchical Cluster Analysis (HCA)
    - Anomaly detection and novelty detection: One-class SVM, Isolation Forest
    - Visualization and dimensionality reduction: Principal Component Analysis (PCA), Kernel PCA, Locally-Linear Embedding (LLE), t-distributed Stochastic Neighbor Embedding (t-SNE)
    - Association rule learning: Apriori, Eclat.

- Unsupervised Learning



Clustering

Anomaly detection



t-distributed Stochastic Neighbor Embedding (t-SNE)

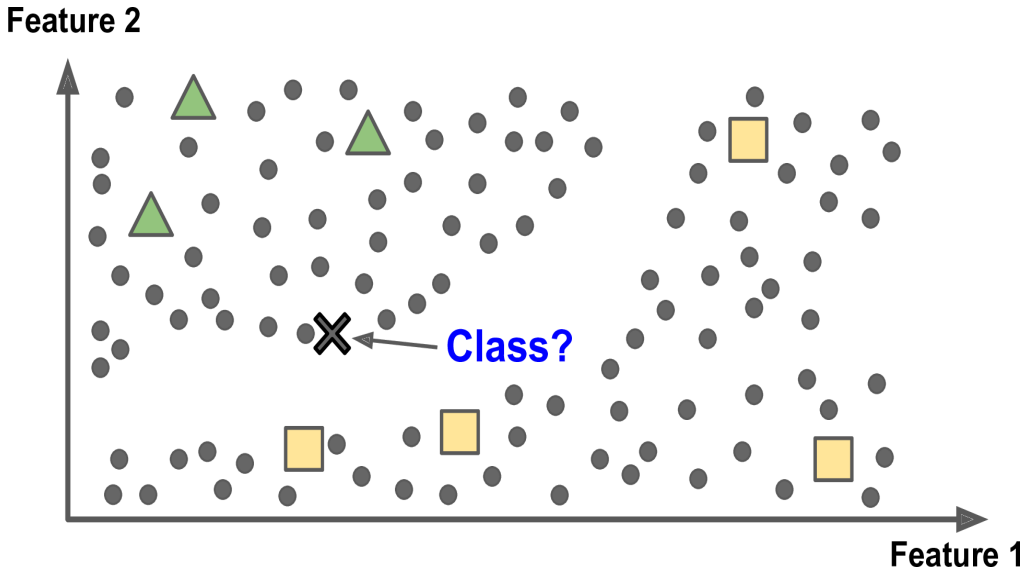
- **Semisupervised learning**

  - Semisupervised learning: deal with partially labeled training data (usually a lot of unlabeled data & a little bit of labeled data)

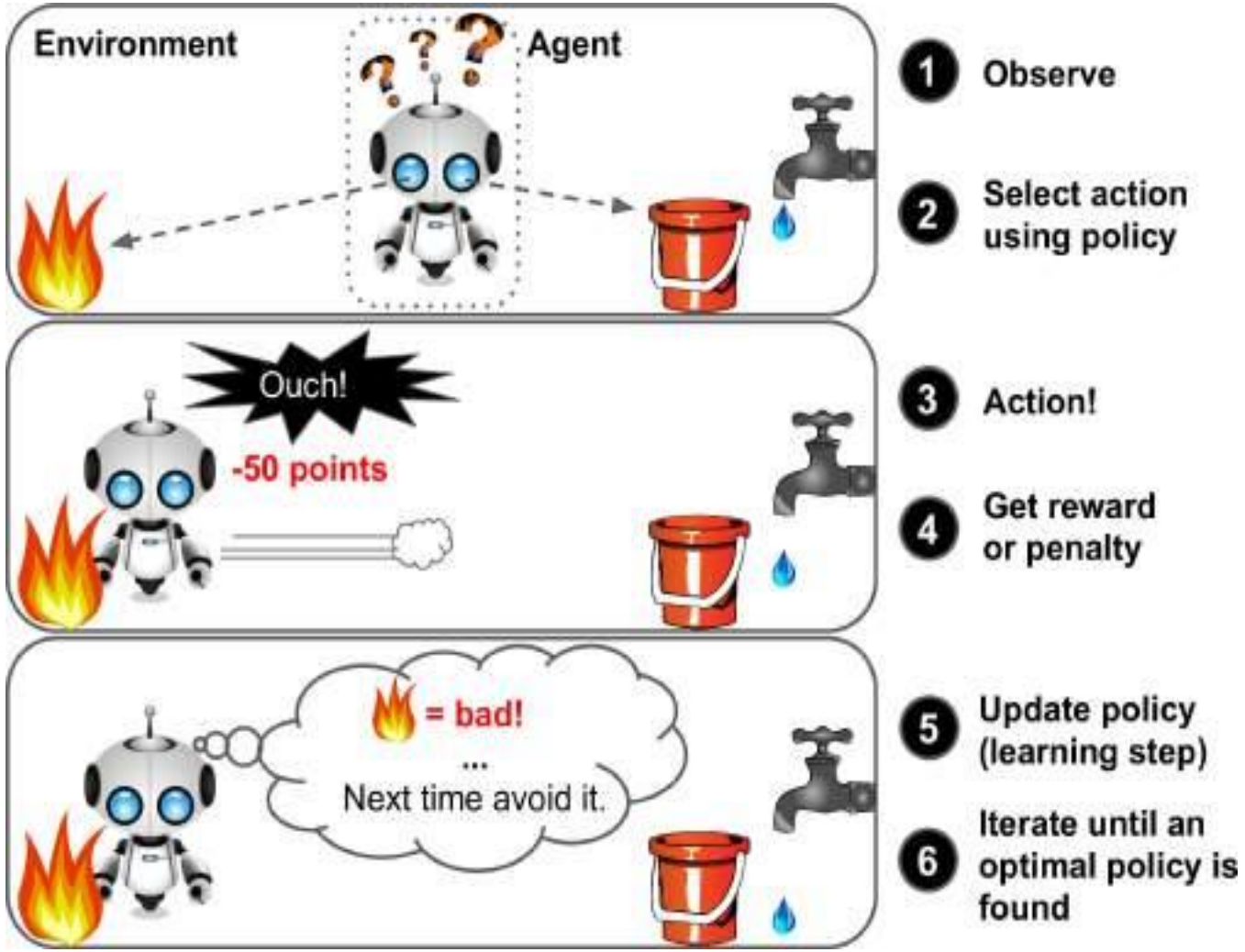
  - Most semisupervised learning algorithms are combinations of unsupervised & supervised algorithms.

    For example:

    - Deep Belief Networks (DBNs) are based on unsupervised components called Restricted Boltzmann Machines (RBMs).
    - RBMs are trained sequentially in an unsupervised manner, and then the whole system is fine-tuned using supervised learning techniques.



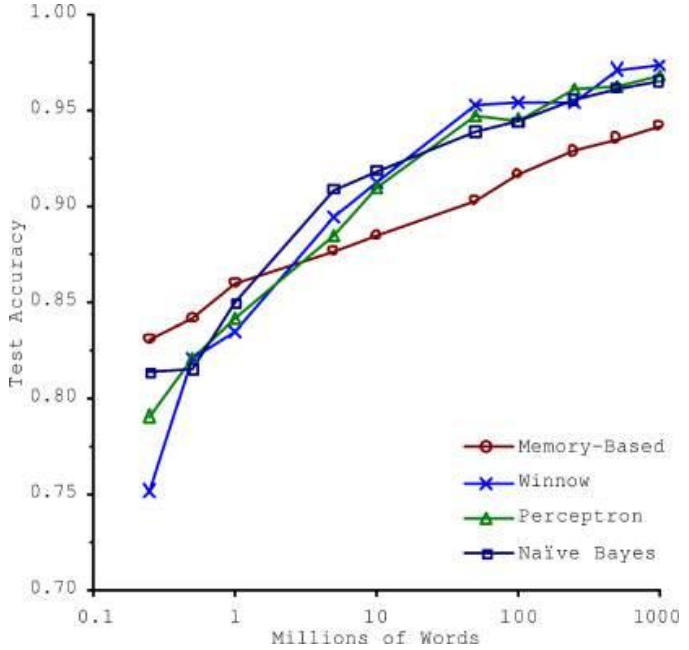
- **Reinforcement Learning**
  - Reinforcement learning:
    - can observe the environment, select and perform actions, and get rewards in return (or penalties in the form of negative rewards)
    - then, system must learn by itself what is the best strategy to get the most reward over time.



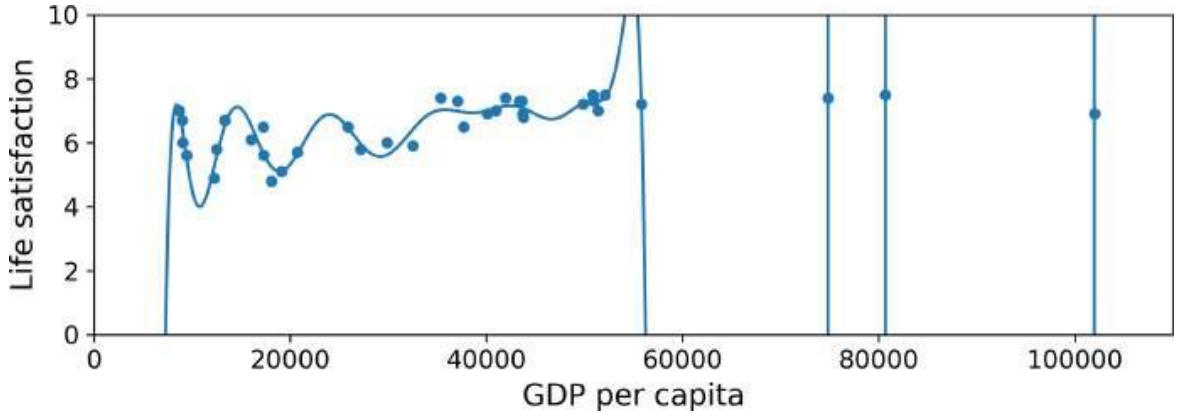
CONTENTs

- Introduction
- Types of ML Systems
- **Challenges of ML**
- Testing & Validating

- **Main Challenges of ML:** 2 problems
  - **bad data**
    - Insufficient Quantity of Training Data
    - Nonrepresentative Training Data
    - Poor-Quality Data
    - Irrelevant Features
  - **bad algorithm**
    - Overfitting the Training Data
    - Underfitting the Training Data



The importance of data versus algorithms



Overfitting the training data

CONTENTs

- Introduction
- Types of ML Systems
- Challenges of ML
- **Testing & Validating**

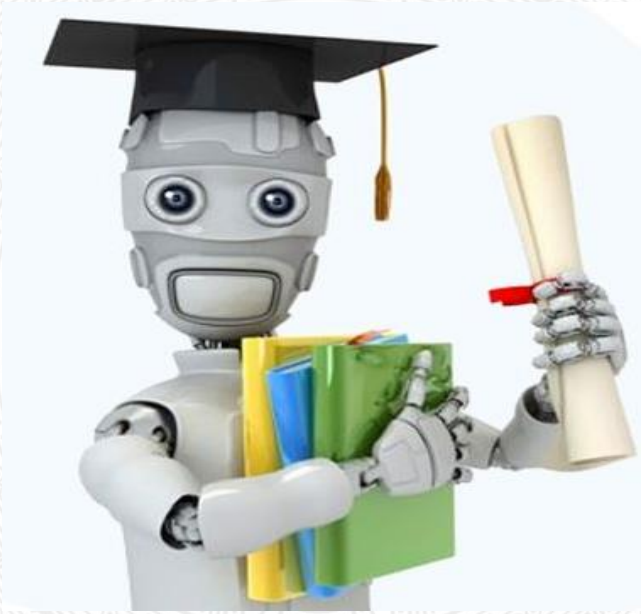
- **Testing**
  - Split data into 2 sets:
    - training set (for train model)
    - test set (for testing model)
  - Evaluating model on the test set

    ⇨ estimate of generalization error.
  - If the training error is low but the generalization error is high

    ⇨ model is overfitting the training data.
- **Validating**
  - Hold out part of the training set to evaluate several candidate models and select the best one. The new heldout set is called the **validation set**

SUMMARY

- Introduction
- Types of ML Systems
- Challenges of ML
- Testing & Validating



Enjoy the Course…!