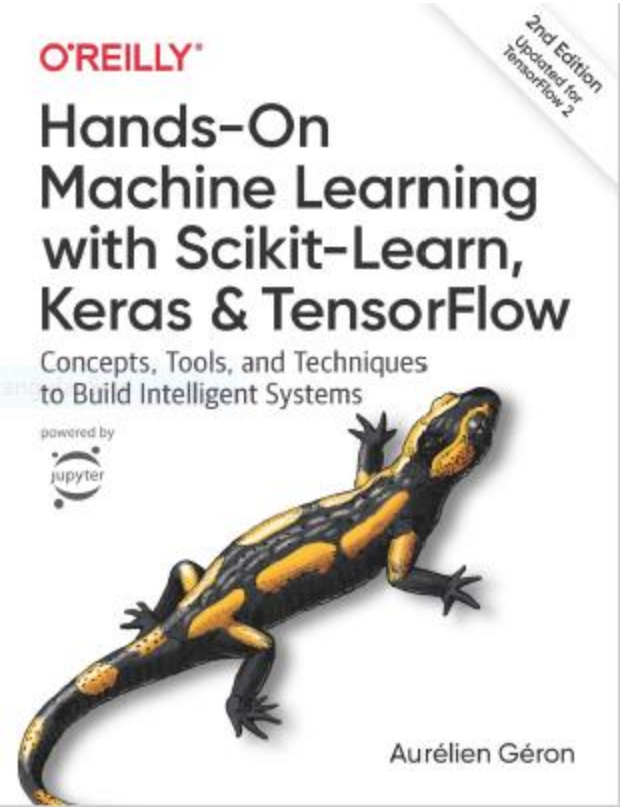
Machine Learning Introduction

Prof. Gheith Abandah

Reference

- Chapter 1: **The Machine Learning Landscape**



- Aurélien Géron, **Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow**, O'Reilly, 2nd Edition, 2019
  - Material: https://github.com/ageron/handson-ml2

Outline

- The Machine Learning Tsunami
- What Is Machine Learning?
- Why Use Machine Learning?
- Types of Machine Learning Systems
- Main Challenges of Machine Learning
- Testing and Validating
- Summary
- Exercises

The Machine Learning Tsunami

- YouTube Video: **From Artificial Intelligence to Superintelligence: Nick Bostrom on AI & The Future of Humanity** From Science Time

https://youtu.be/Kktn6BPg1sI

The Machine Learning Tsunami

- In **2006**, **Geoffrey Hinton** *et al*. published a paper showing how to **train a deep neural network** capable of recognizing handwritten digits with state-of-the-art precision (>98%). They branded this technique **Deep Learning**.
- Training a deep neural net was widely considered impossible at the time, and most researchers had abandoned the idea since the 1990s.
- Fast-forward 10 years and **ML has conquered the industry**: it is now at the heart of much of the magic in today's high-tech products.

Outline

- ✓The Machine Learning Tsunami
- What Is Machine Learning?
- Why Use Machine Learning?
- Types of Machine Learning Systems
- Main Challenges of Machine Learning
- Testing and Validating
- Summary
- Exercises

What Is Machine Learning?

- YouTube Video: **What is Machine Learning?** from Google Cloud Platform

https://youtu.be/HcqpanDadyQ

What Is Machine Learning?

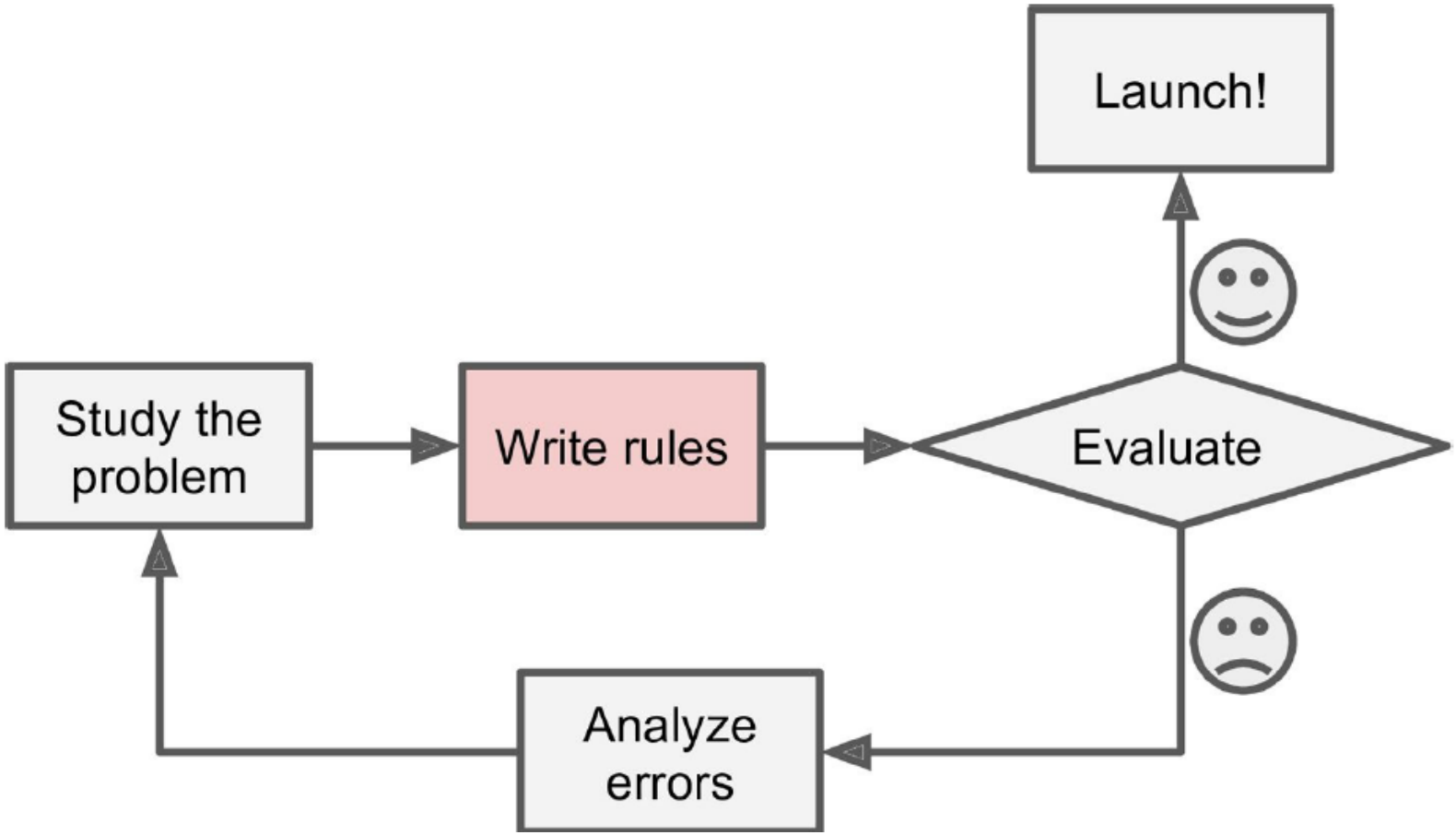
- The science (and art) of programming computers so they can **learn from data**.
- The field of study that gives computers the ability to **learn without being explicitly programmed**. Arthur Samuel, *1959*
- A computer program is said to learn from **experience E** with respect to some **task T** and some **performance** measure **P**, if its performance on T, as measured by P, **improves with experience E**. Tom Mitchell, *1997*
  - **E**: **Training set** made of **training instances** (**samples**)
  - **T**: **Test set**
  - **P**: Such as **accuracy**

Outline

- ✓The Machine Learning Tsunami
- ✓What Is Machine Learning?
- Why Use Machine Learning?
- Types of Machine Learning Systems
- Main Challenges of Machine Learning
- Testing and Validating
- Summary
- Exercises

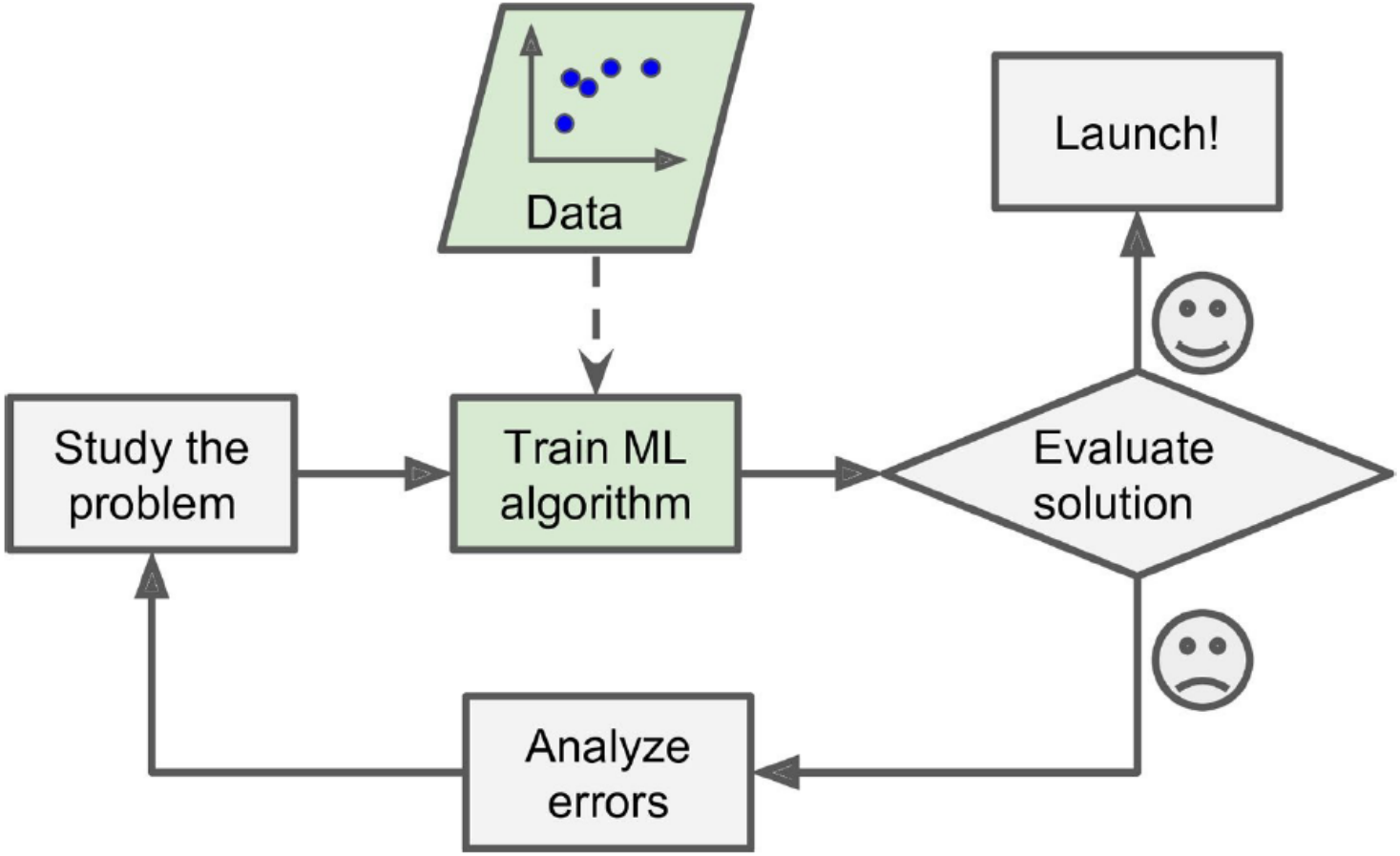
Why Use Machine Learning?

Spam filter using traditional programming techniques



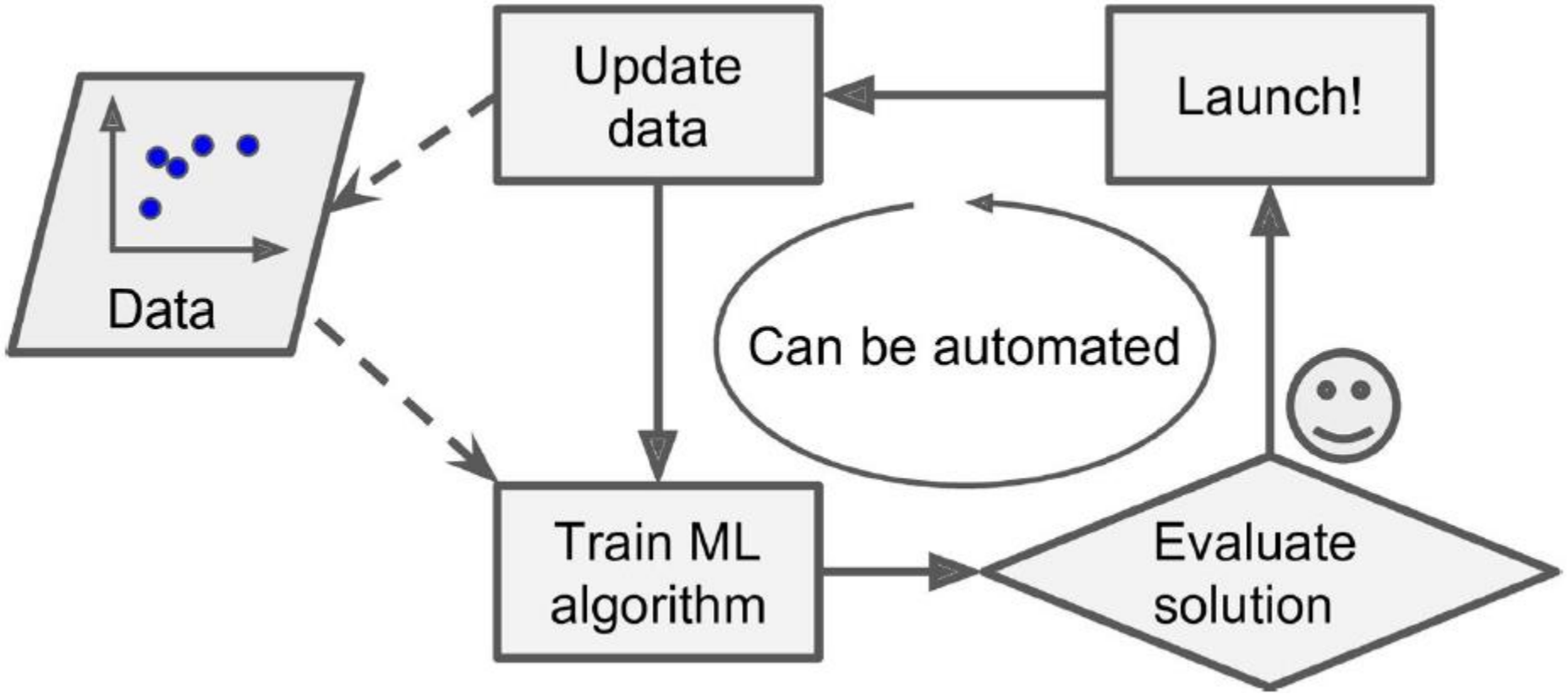
Why Use Machine Learning?

Spam filter using machine learning techniques 1/2



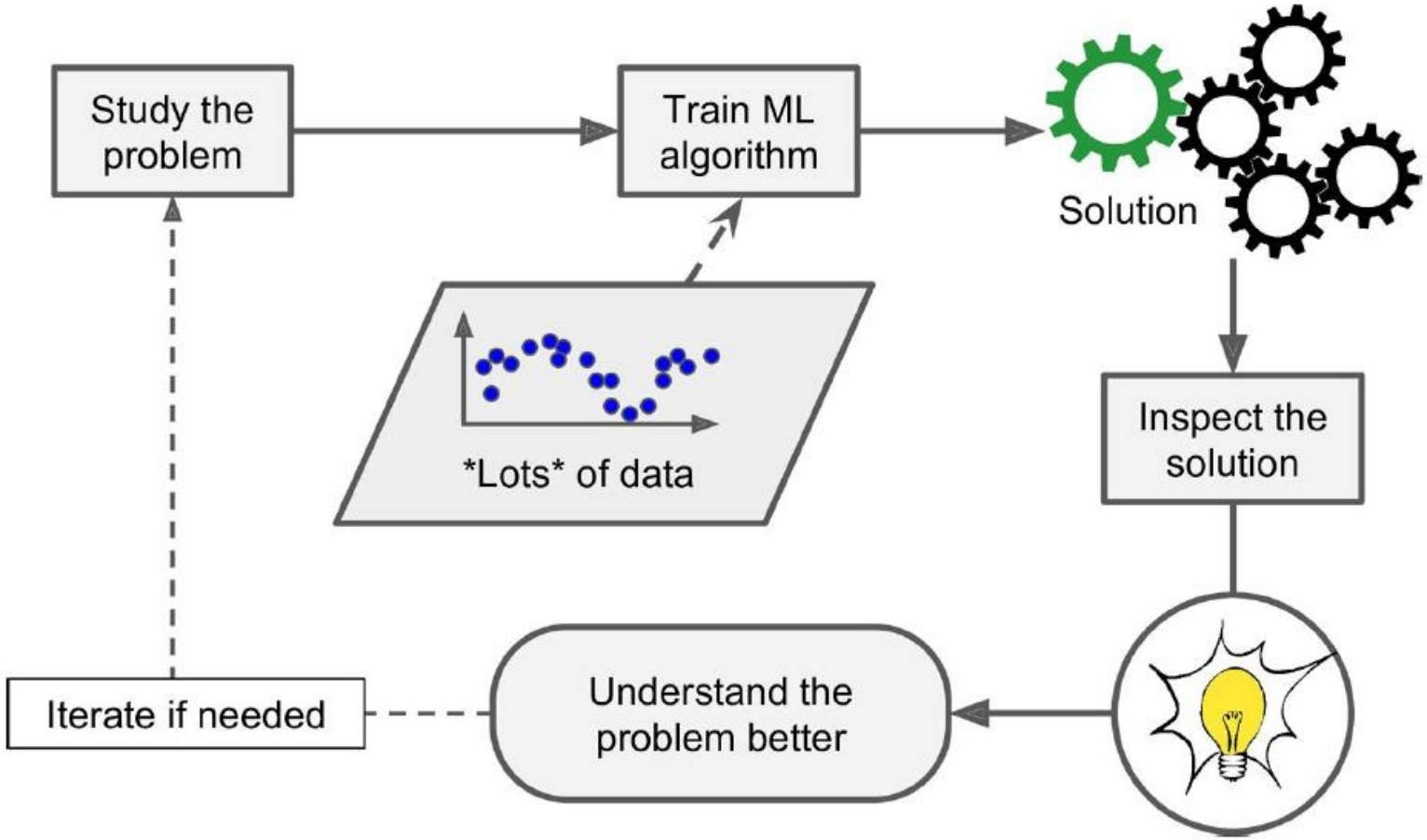
Why Use Machine Learning?

Automatically adapting to change 2/2



Why Use Machine Learning?

ML can help humans learn (Data mining)



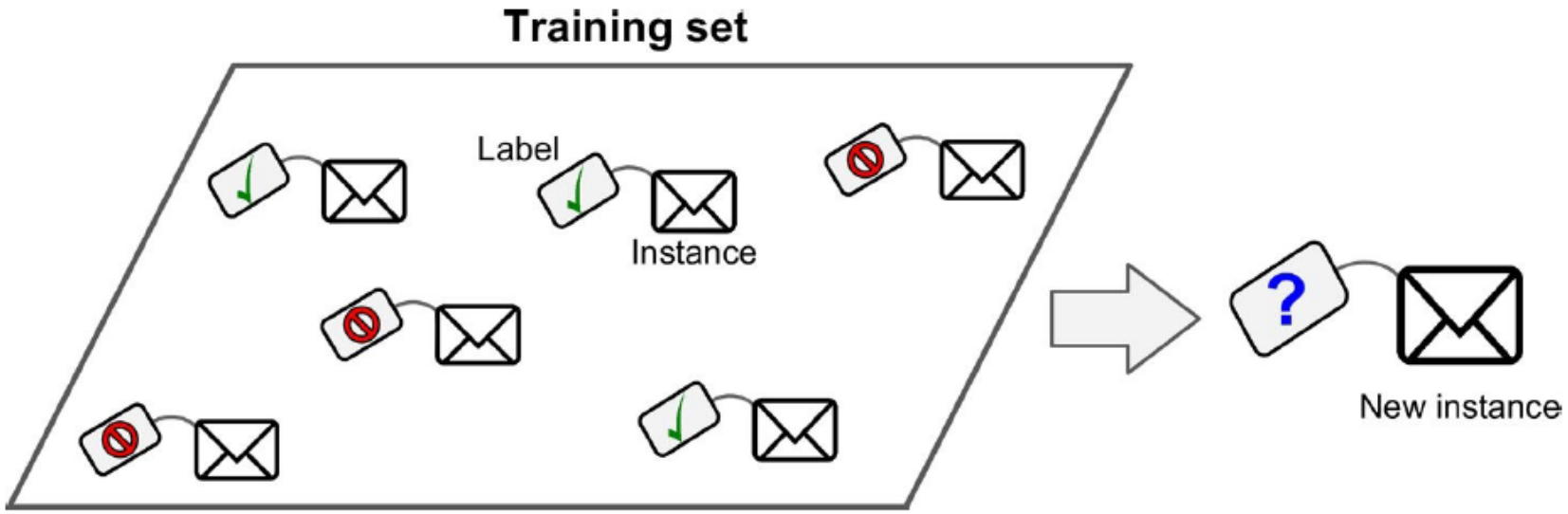
Outline

- ✓The Machine Learning Tsunami
- ✓What Is Machine Learning?
- ✓Why Use Machine Learning?
- Types of Machine Learning Systems
- Main Challenges of Machine Learning
- Testing and Validating
- Summary
- Exercises

Types of Machine Learning Systems

- **Involves human supervision?**
  1. Supervised learning
  2. Unsupervised learning
  3. Semi-supervised learning
  4. Reinforcement learning
- **Learns incrementally?**
  1. Batch learning
  2. Online learning
- **Generalization approach**
  1. Instance-based learning
  2. Model-based learning

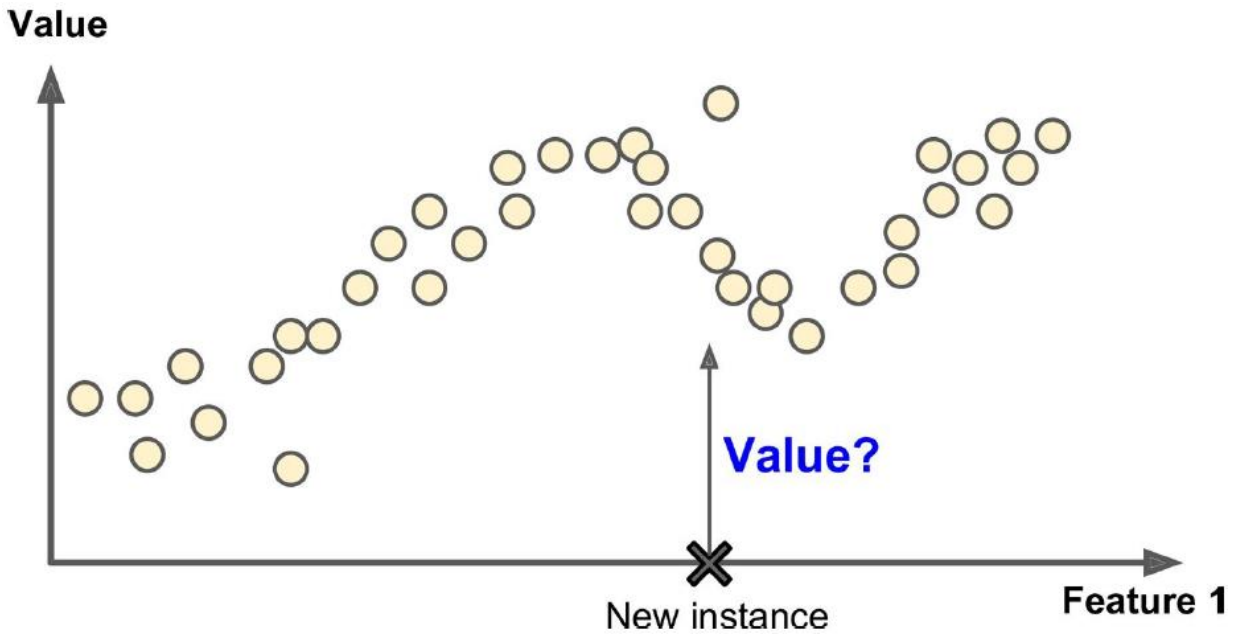
1. Supervised Learning



The training data you feed to the algorithm includes the desired solutions, called **labels**

Classification: finds the class, e.g., email type (spam or ham)

1. Supervised Learning

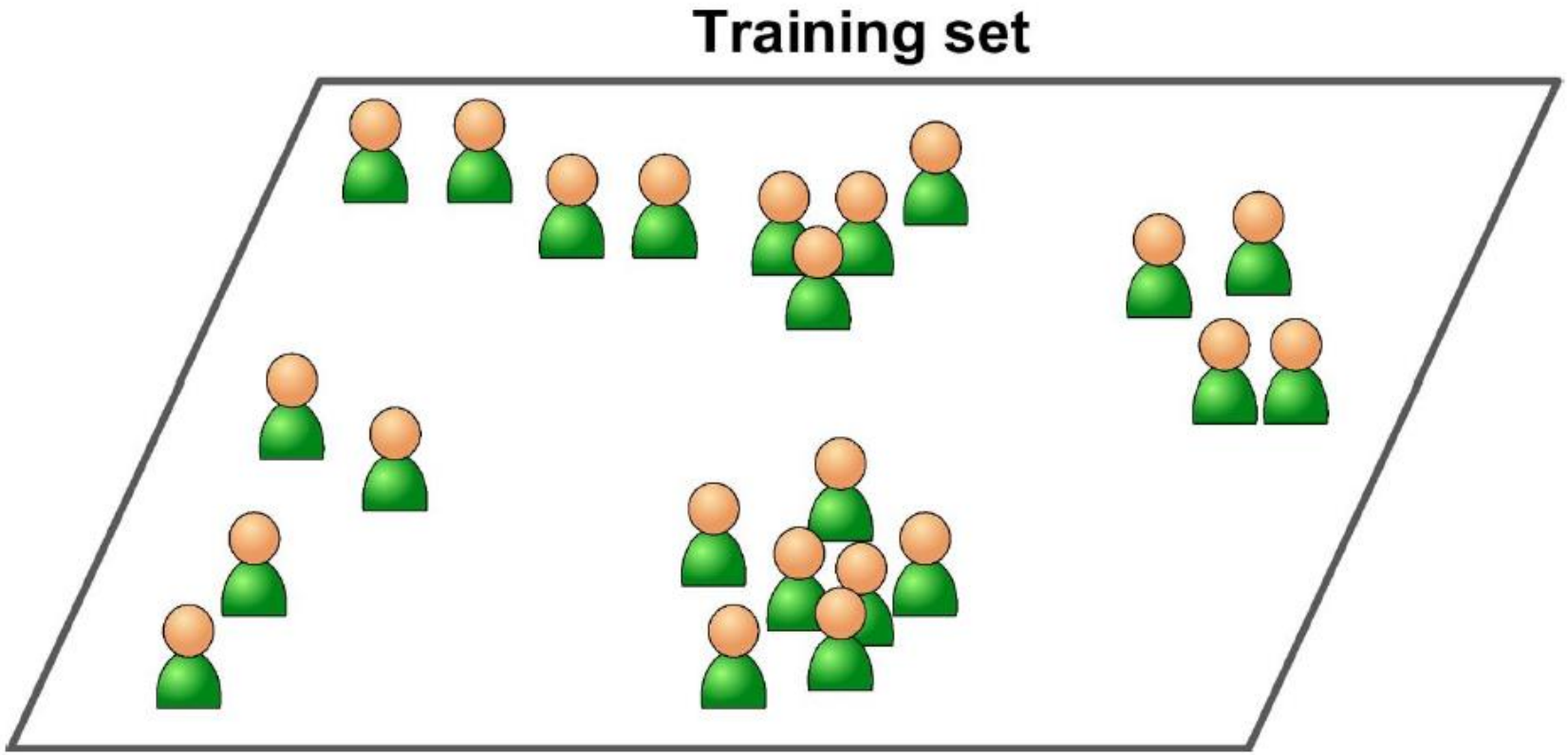


Regression: finds the value, e.g., car price

1. Supervised learning algorithms

Algorithm	Type
k-Nearest Neighbors	Both
Linear Regression	Regression
Logistic Regression	Classification
Support Vector Machines (SVMs)	Both
Decision Trees	Both
Random Forests	Both
Neural Networks	Both

2. Unsupervised Learning

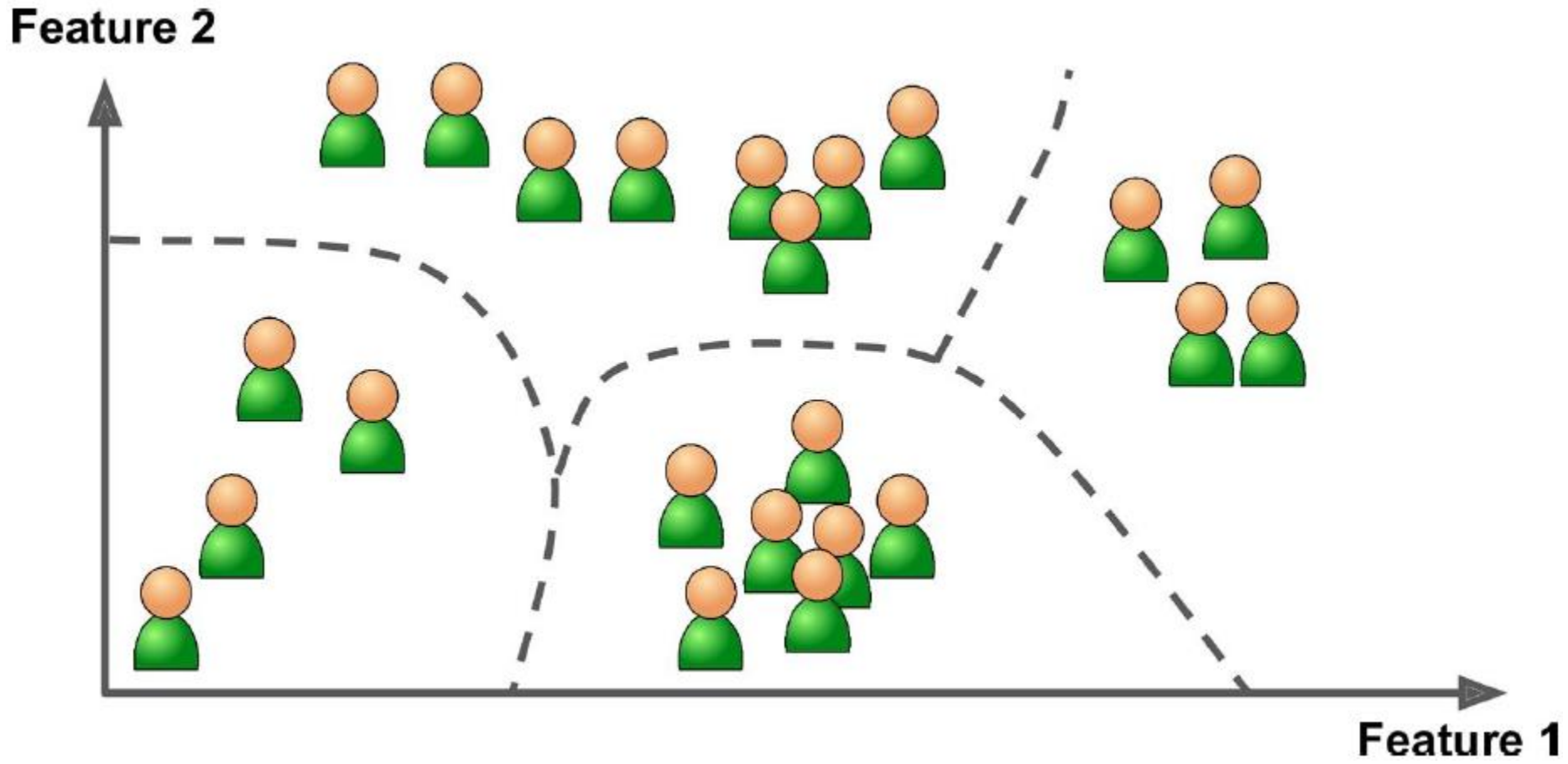


The training data is **unlabeled**.

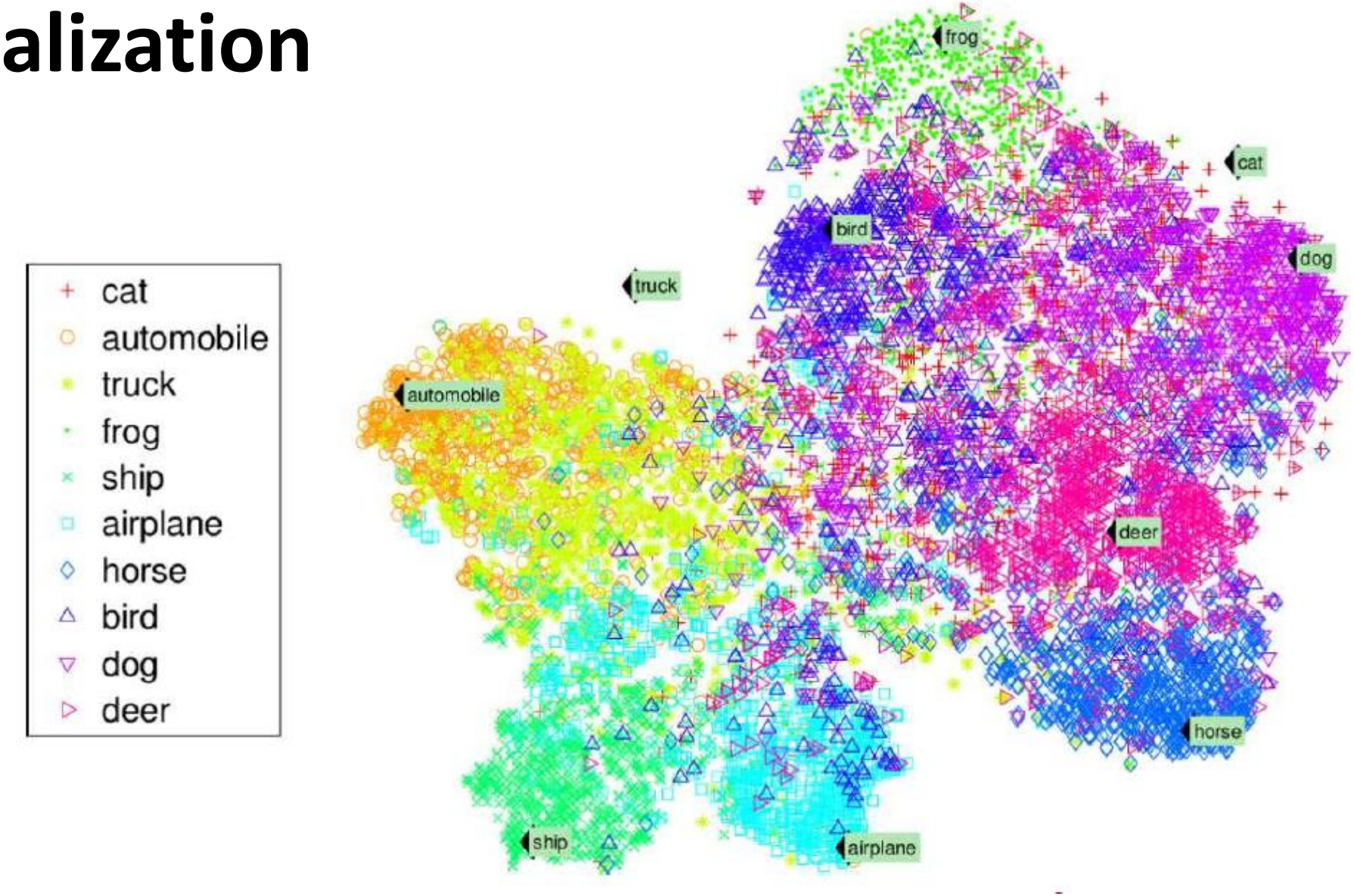
2. Unsupervised learning algorithms

- Clustering
  - k-Means
  - Hierarchical Cluster Analysis (HCA)
  - Expectation Maximization
- 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

2.a Clustering



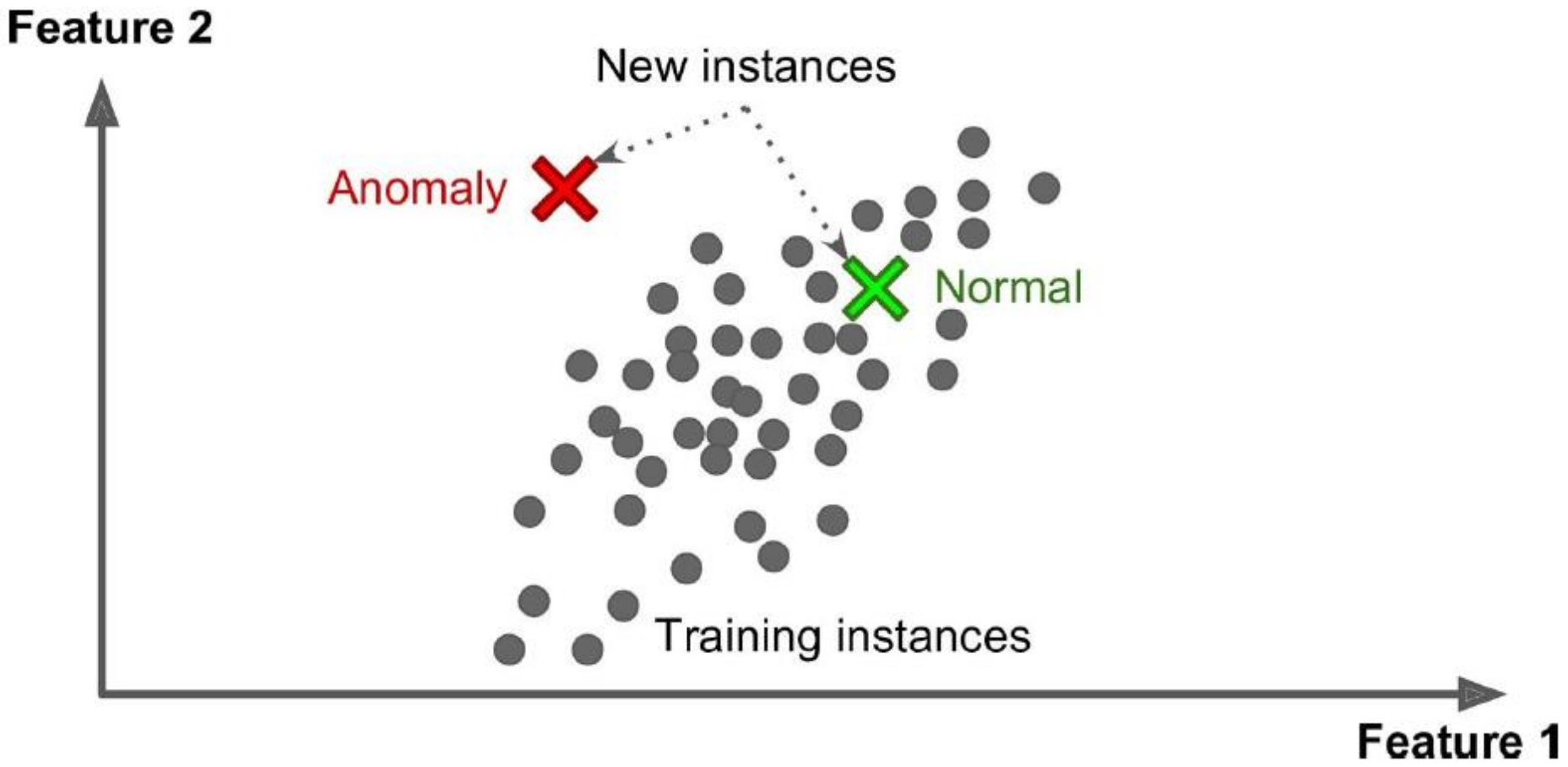
2.b Visualization



2.c Dimensionality Reduction

- The goal is to **simplify the data** without losing too much information.
- One way to do this is to **merge** several **correlated features** into one. For example, a car's mileage may be very correlated with its age, so the dimensionality reduction algorithm will merge them into one feature that represents the car's wear and tear.
- Also called **feature extraction**.

2.d Anomaly Detection

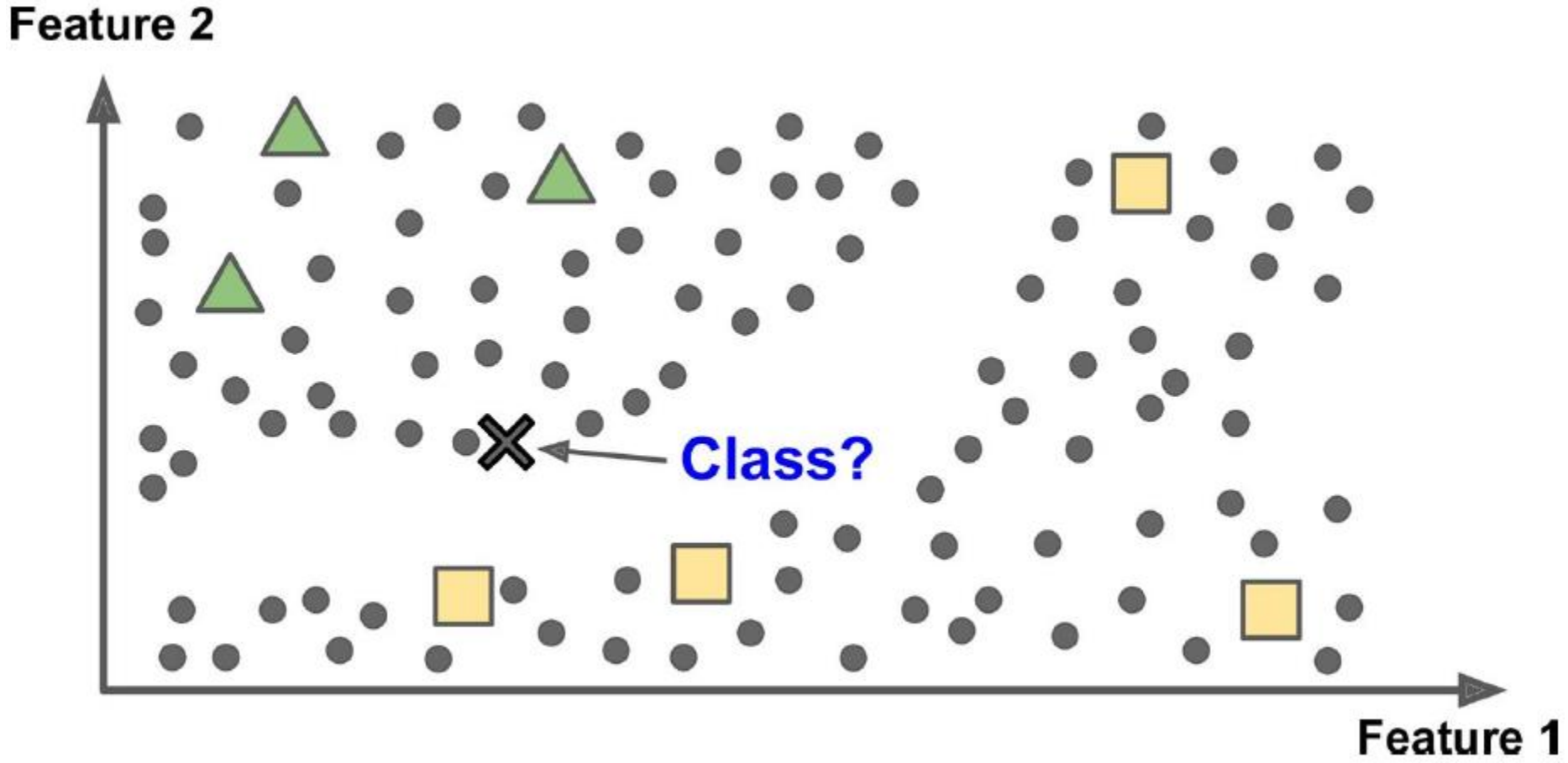


2.e Association Rule Learning

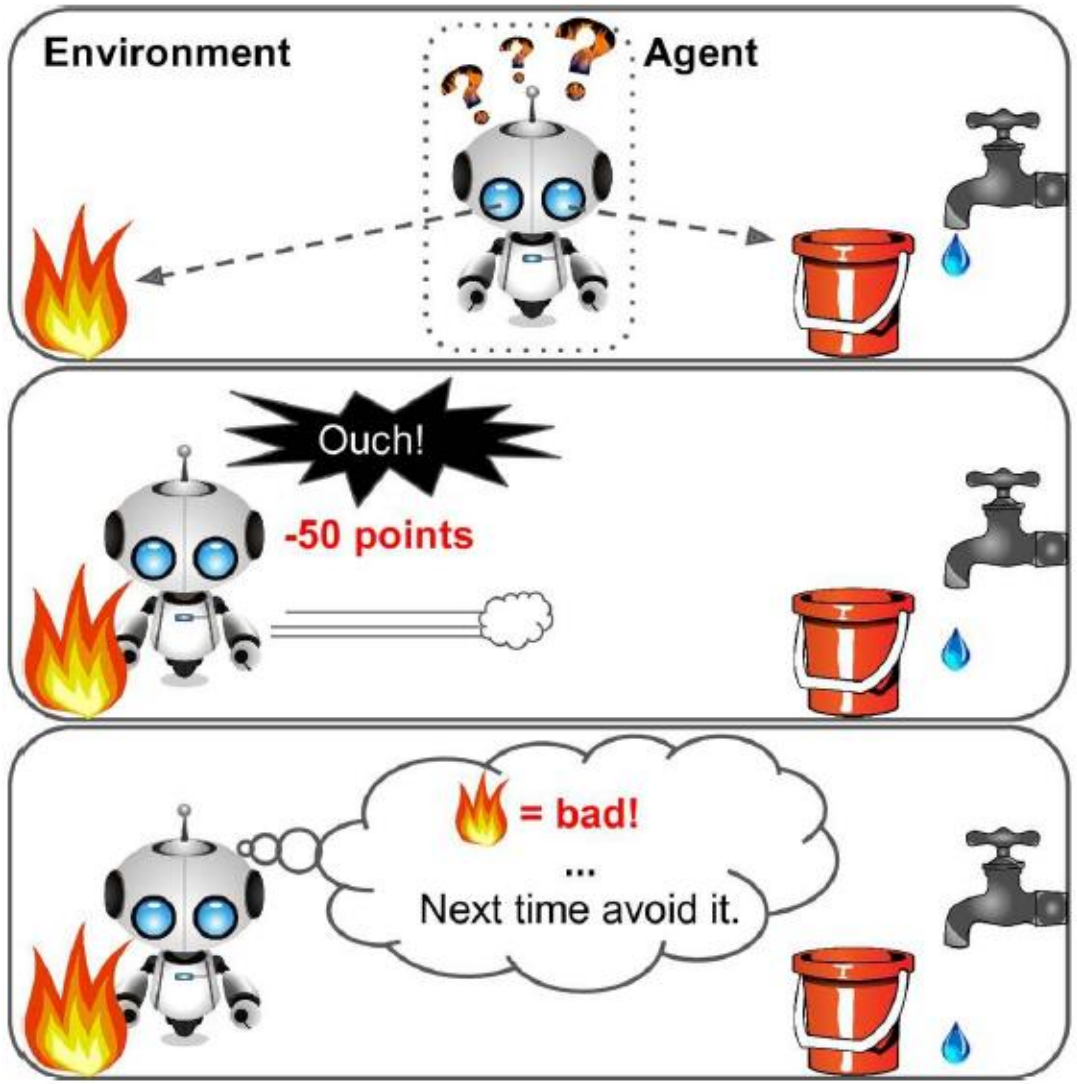
- The **goal is to dig into large amounts** of data and **discover interesting relations** between attributes.
- For example, suppose you own a supermarket. Running an association rule on your sales logs may reveal that people who purchase **barbecue sauce** and **potato chips** also tend to buy steak. Thus, you may want to place these items close to each other.

3. Semi-supervised Learning

Partially labeled training data, usually a lot of unlabeled data and a little bit of labeled data. E.g., Google Photos.



4. Reinforcement Learning



1. Observe
2. Select action using policy
3. Action!
4. Get reward or penalty
5. Update policy (learning step)
6. Iterate until an optimal policy is found

Types of Machine Learning Systems

✓**Involves human supervision?**

1. Supervised learning
2. Unsupervised learning
3. Semi-supervised learning
4. Reinforcement learning

- **Learns incrementally?**
  1. Batch learning
  2. Online learning

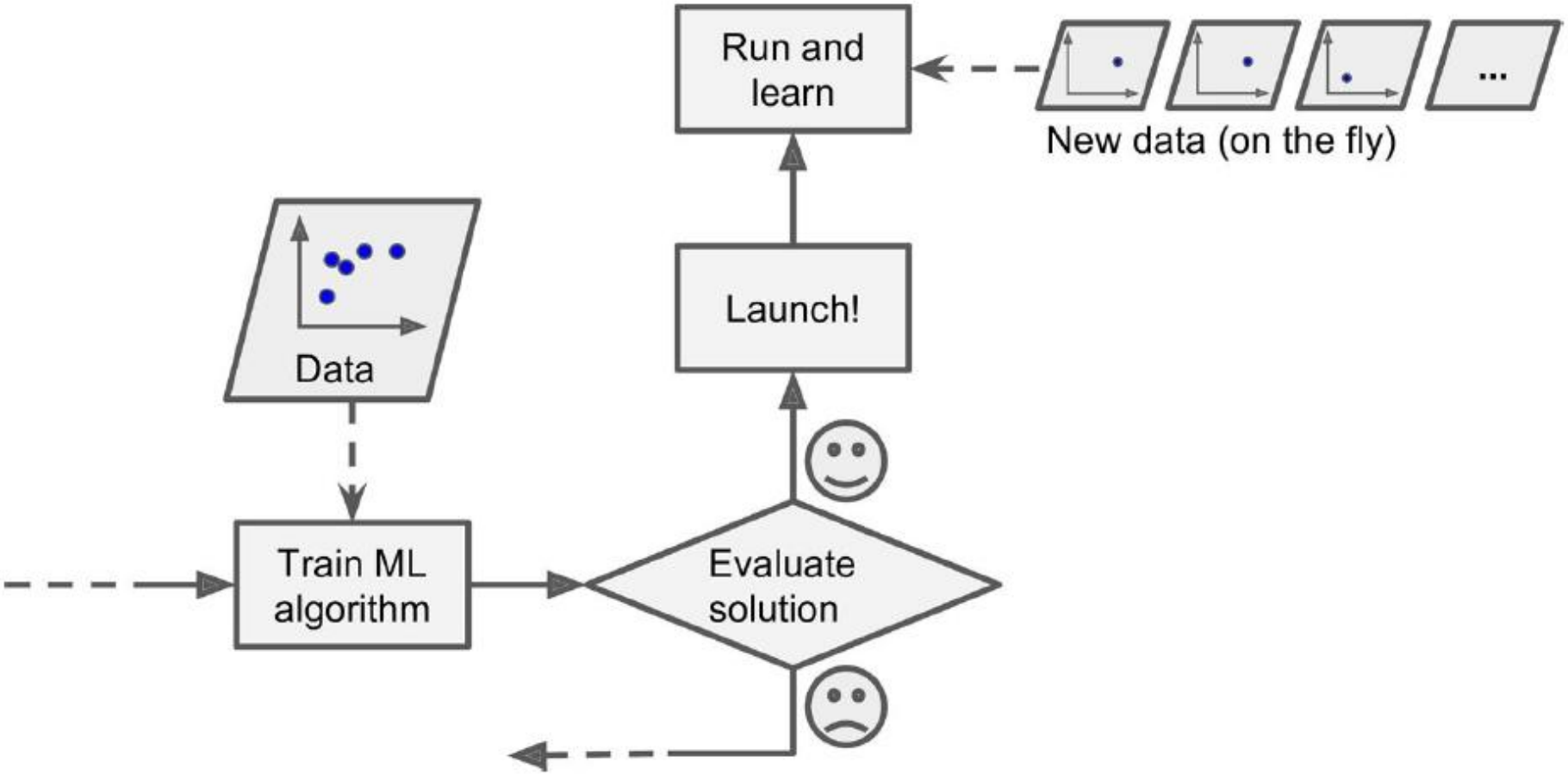
- **Generalization approach**
  1. Instance-based learning
  2. Model-based learning

1. Batch (offline) Learning

- Must be **trained** using **all the available data**.
- This will generally take a **lot of time** and computing resources, so it is typically done **offline**.
- First the system is **trained**, and **then** it is **launched** into production and runs without learning anymore; it just applies what it has learned.

2. Online Learning

Examples: Stock prices, huge data



Types of Machine Learning Systems

✓**Involves human supervision?**

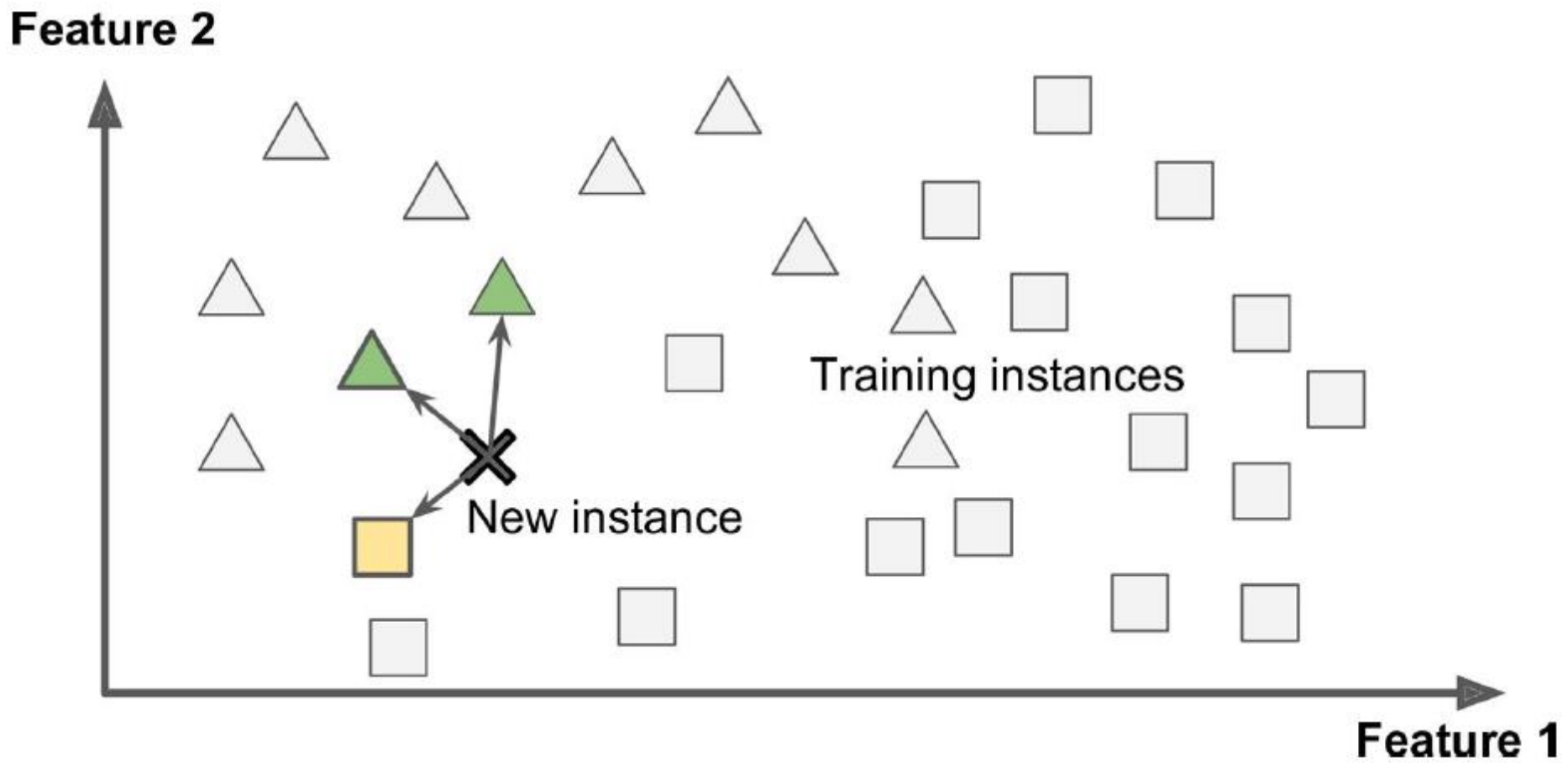
1. Supervised learning
2. Unsupervised learning
3. Semi-supervised learning
4. Reinforcement learning

✓**Learns incrementally?**

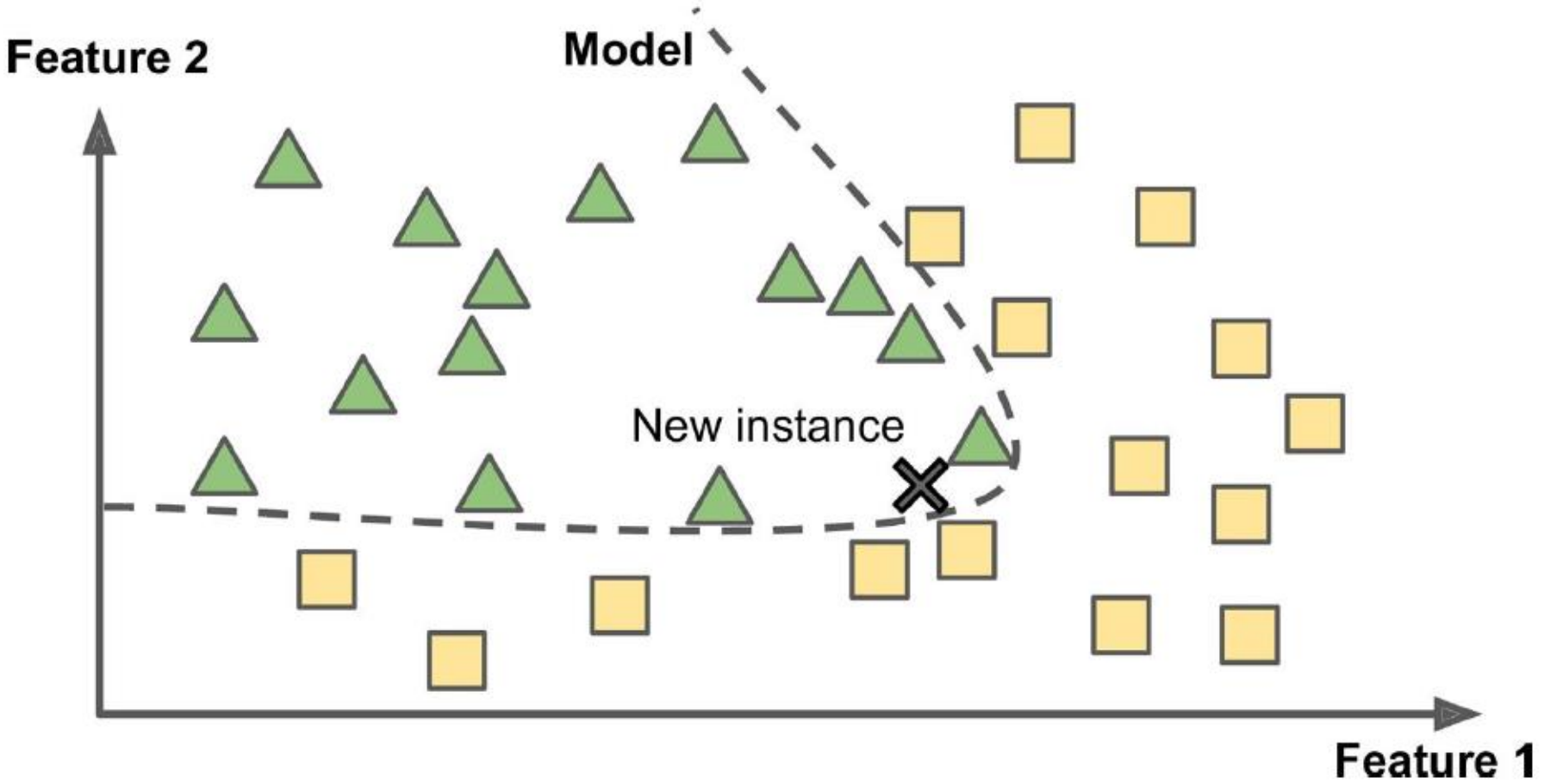
1. Batch learning
2. Online learning

- **Generalization approach**
  1. Instance-based learning
  2. Model-based learning

1. Instance-based Learning



2. Model-based Learning

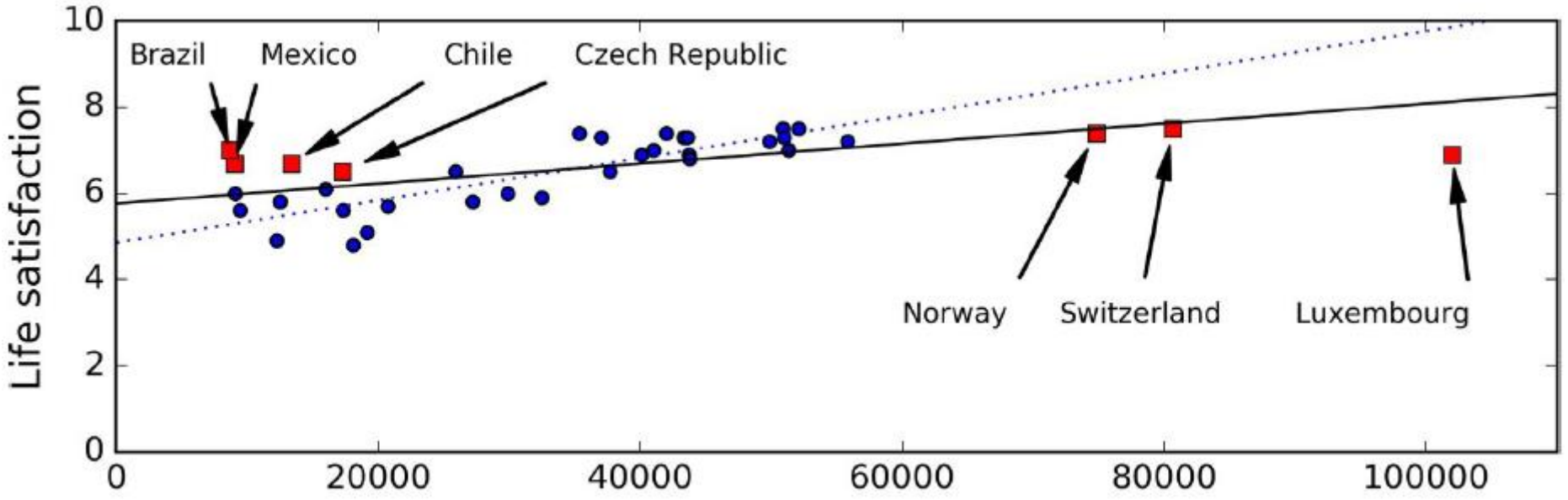


Outline

- ✓The Machine Learning Tsunami
- ✓What Is Machine Learning?
- ✓Why Use Machine Learning?
- ✓Types of Machine Learning Systems
- Main Challenges of Machine Learning
- Testing and Validating
- Summary
- Exercises

Main Challenges of Machine Learning (due to bad data)

1. **Insufficient** quantity of training data
2. **Non-representative** training data

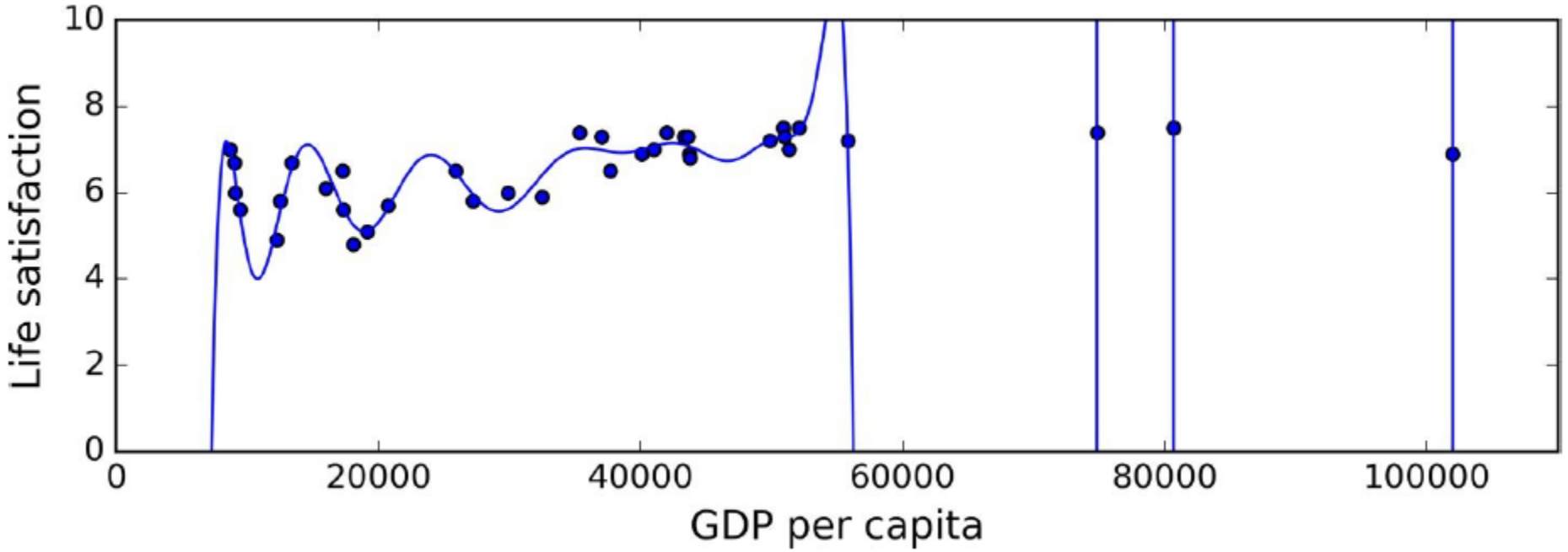


Main Challenges of Machine Learning (due to bad data)

3. **Poor-quality** data that contains:
   - Errors
   - Outliers
   - Noise
4. **Irrelevant features**: Need **feature engineering**:
   - **Feature selection**: selecting the most useful features.
   - **Feature extraction**: combining existing features to produce a more useful one.
   - **Creating new features** by gathering new data.

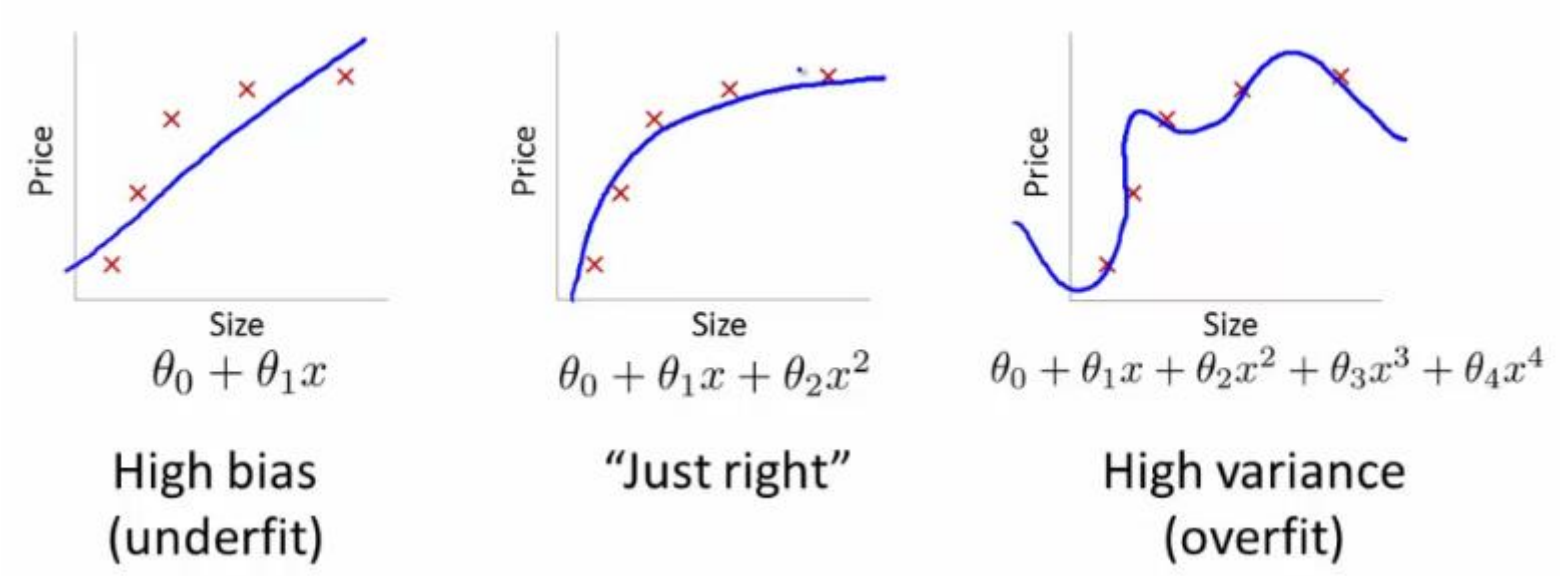
Main Challenges of Machine Learning (due to bad algorithm)

1. **Overfitting** the training data
   - **Regularization** constrains the model's **hyperparameters** to make it simpler and reduce the risk of overfitting.



Main Challenges of Machine Learning (due to bad algorithm)

2. **Under-fitting** the training data



Outline

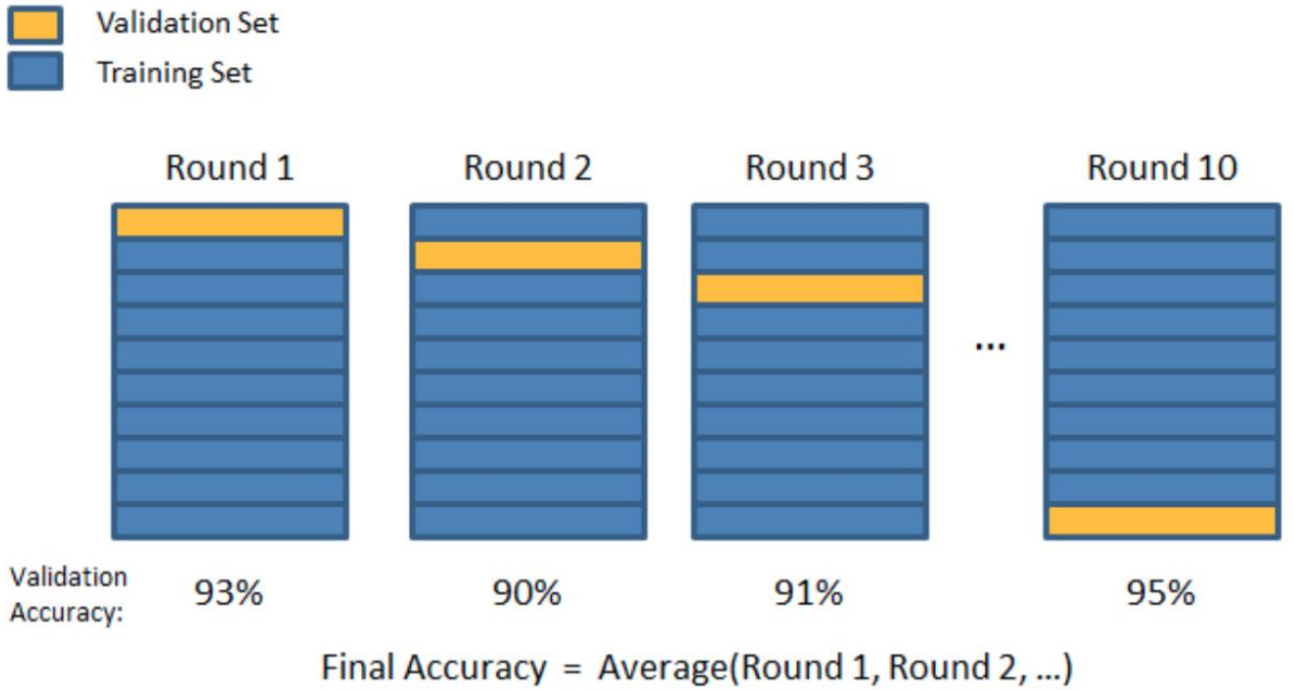
- ✓The Machine Learning Tsunami
- ✓What Is Machine Learning?
- ✓Why Use Machine Learning?
- ✓Types of Machine Learning Systems
- ✓Main Challenges of Machine Learning
- Testing and Validating
- Summary
- Exercises

Testing and Validating

- **Split** your data into two sets (**cross validation**):
  - The **training set** (80%)
  - The **test set** (20%)
- **Evaluate**:
  - The training error
  - The generalization error
- If the training error is low but the generalization error is high, it means that your model is overfitting the training data.
- When the ML algorithm is iterative, often we use a third set: **validation set**.

Cross Validation

- In **k-fold cross-validation**, the original sample is randomly partitioned into **k** equal size subsamples.



Summary

- ML is about making machines get better at some task by learning from data, instead of having to explicitly code rules.
- Types of ML systems: supervised or not, batch or online, and instance-based or model-based.
- A model-based algorithm tunes some parameters to fit the model to the training set, and then hopefully it will be able to make good predictions on new cases.
- An instance-based algorithm learns the examples by heart and uses a similarity measure to generalize to new instances.
- The system will not perform well if your training set is too small, not representative, noisy, or polluted with irrelevant features.
- Your model needs to be neither too simple (under-fit) nor too complex (over-fit).

Exercises

- How would you define Machine Learning?
- What is a labeled training set?
- Can you name four common unsupervised tasks?
- What type of Machine Learning algorithm would you use to allow a robot to walk in various unknown terrains?
- What type of algorithm would you use to segment your customers into multiple groups?
- What is an online learning system?
- What is the difference between a model parameter and a learning algorithm’s hyperparameter?
- If your model performs great on the training data but generalizes poorly to new instances, what is happening? Can you name three possible solutions?
- What is the purpose of a validation set?