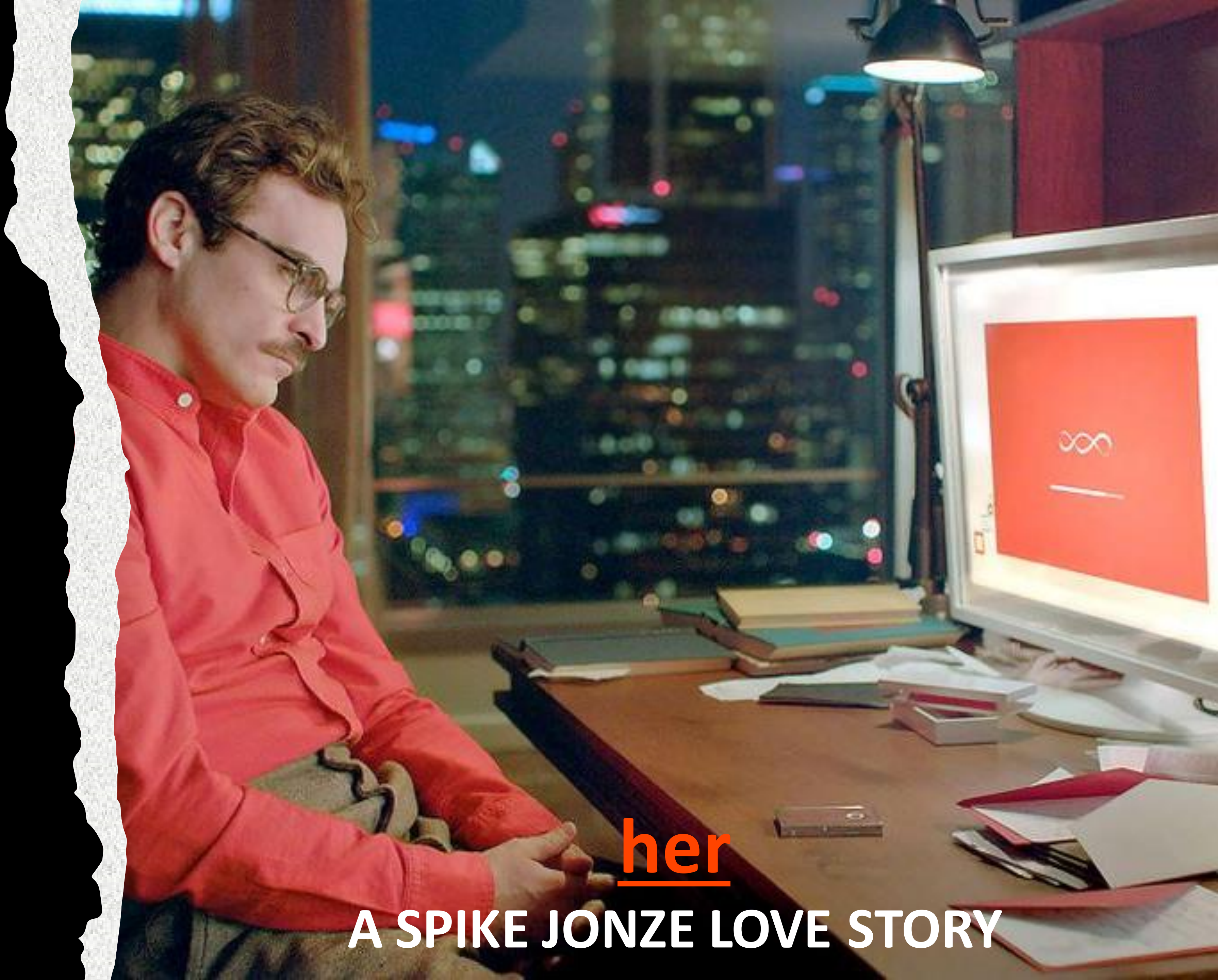
# Introduction to ML

Week 2




Alireza Fazeli
afazeli2@uwo.ca
WSC-272

# Terminology

**Imagine a dataset with columns such as mileage, age, brand, model etc., and you want to predict the price of a car**

---

**Attribute:** A data type, *e.g.*, mileage

**Feature:** Attribute plus its value, *e.g.*, "mileage = 8000"

**Predictor:** Features chosen to train the ML model

**Label:** Desired solution, *i.e.*, price

**Feature Extraction:** Merge correlated features into one feature,

*e.g.*, mileage+age = wear&tear

**Target:** What to be predicted, *i.e.*, price

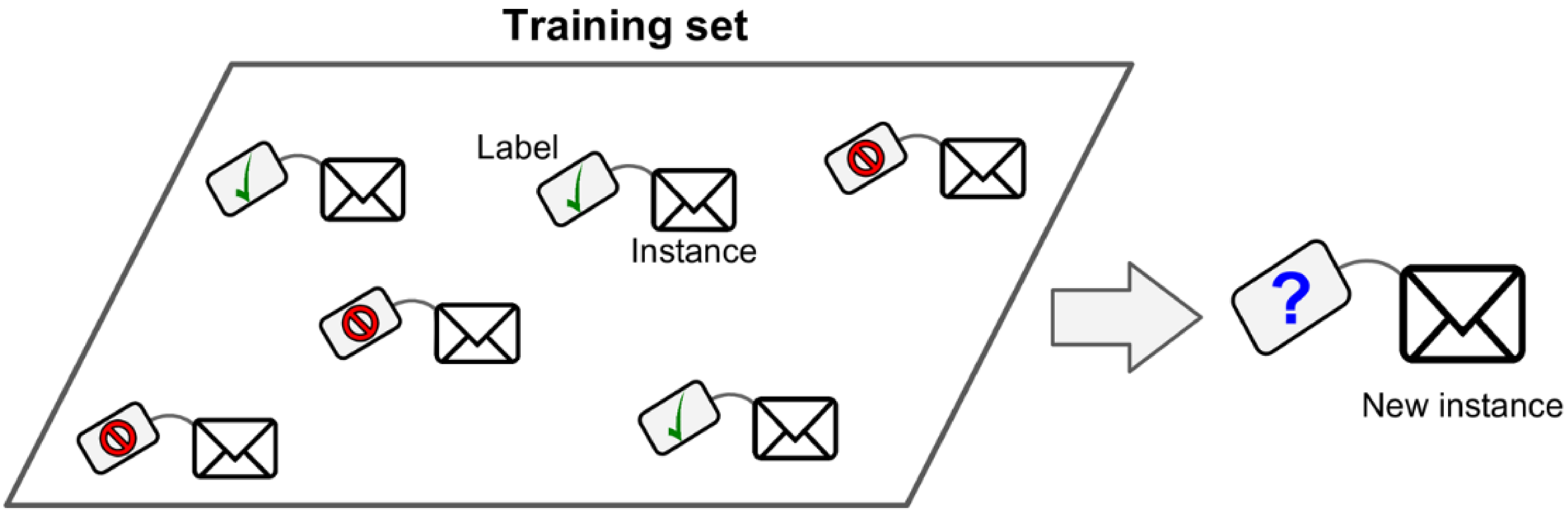
# Types of Machine Learning Sytems

**According to the amount and type of supervision they get during training:**

- **Supervised Learning**
- **Unsupervised Learning**
- **Semi-Supervised Learning**
- **Reinforcement Learning**

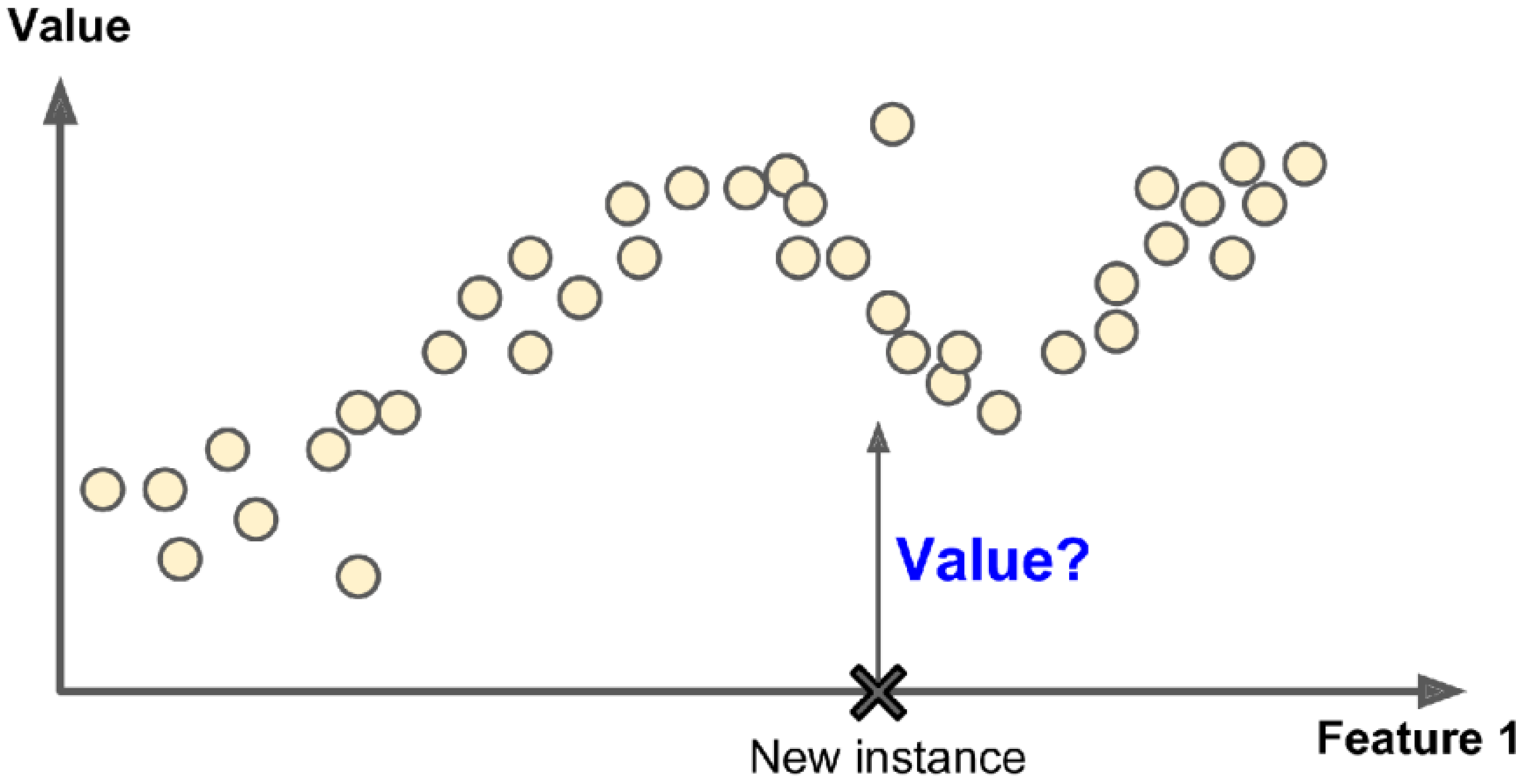
# Supervised Learning

**Training data is labeled.**
**A typical supervised learning task is classification e.g., spam email filter:**



# Supervised Learning

**Another example is regression that is to predict a target numeric value e.g., car price predictor:**



# Supervised Learning

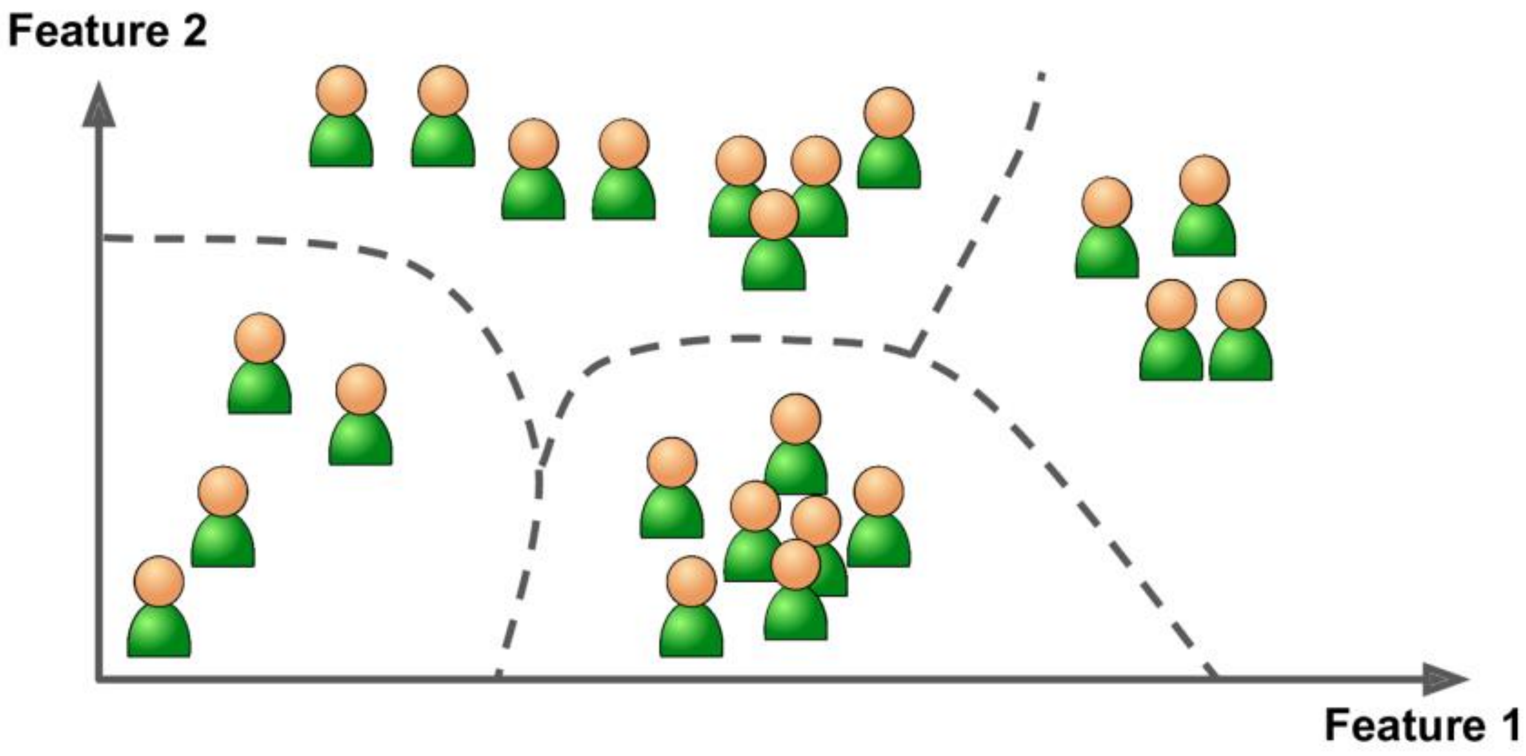
**Some of the most important supervised learning algorithms:**

- **k-Nearest Neighbors**
- **Linear Regression**
- **Logistic Regression**
- **Support Vector Machines (SVMs)**
- **Decision Trees and Random Forests**
- **Neural networks,** which can also be unsupervised such as autoencoders, or semisupervised such as in deep belief networks (DBN), or unsupervised pretraining.

# Unsupervised Learning

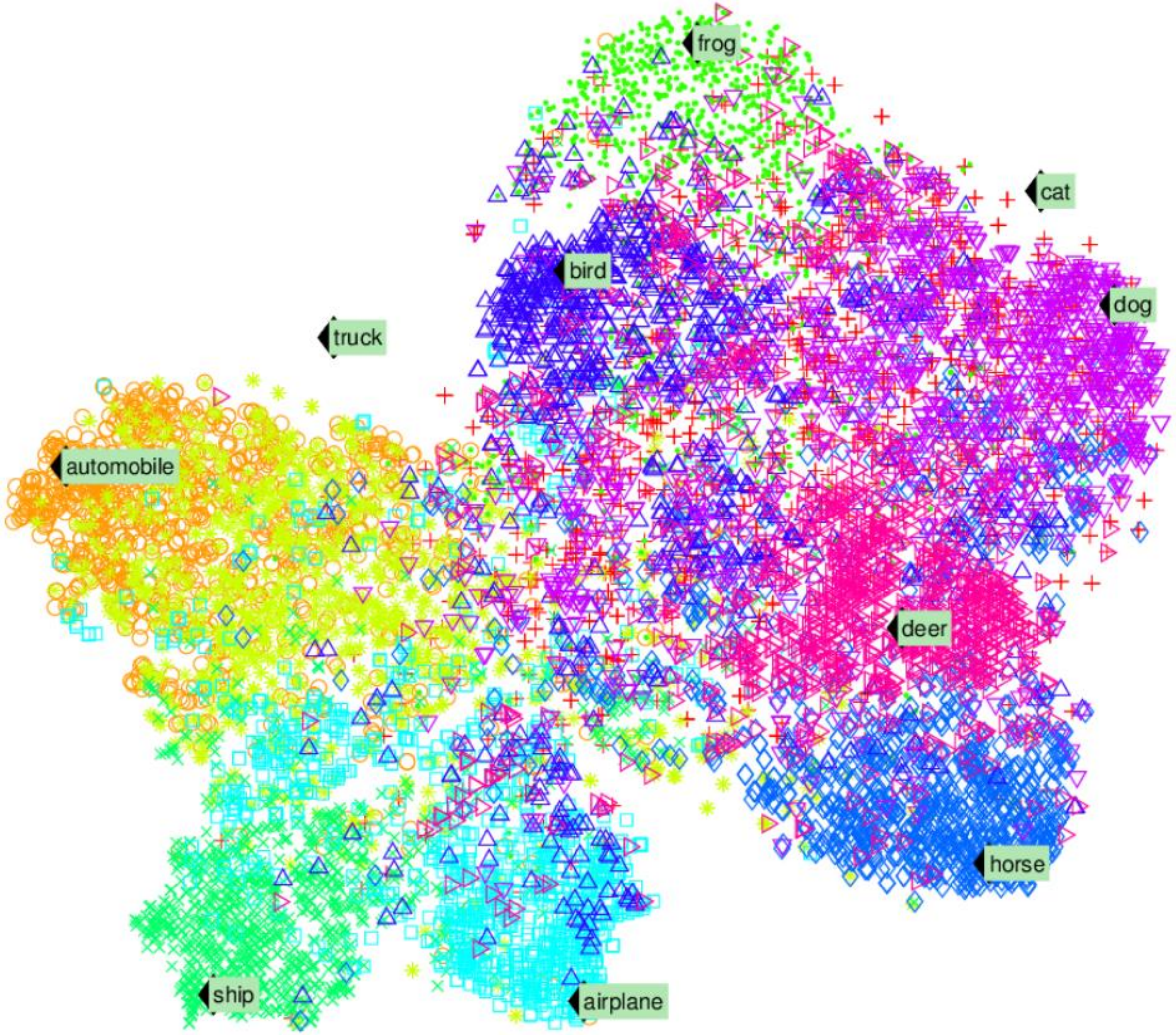
**Training data is unlabeled.**

**A typical unsupervised learning task is clustering, *e.g.*, detect groups of similar visitors to a website** (without telling the algorithm which group a visitor belongs to):



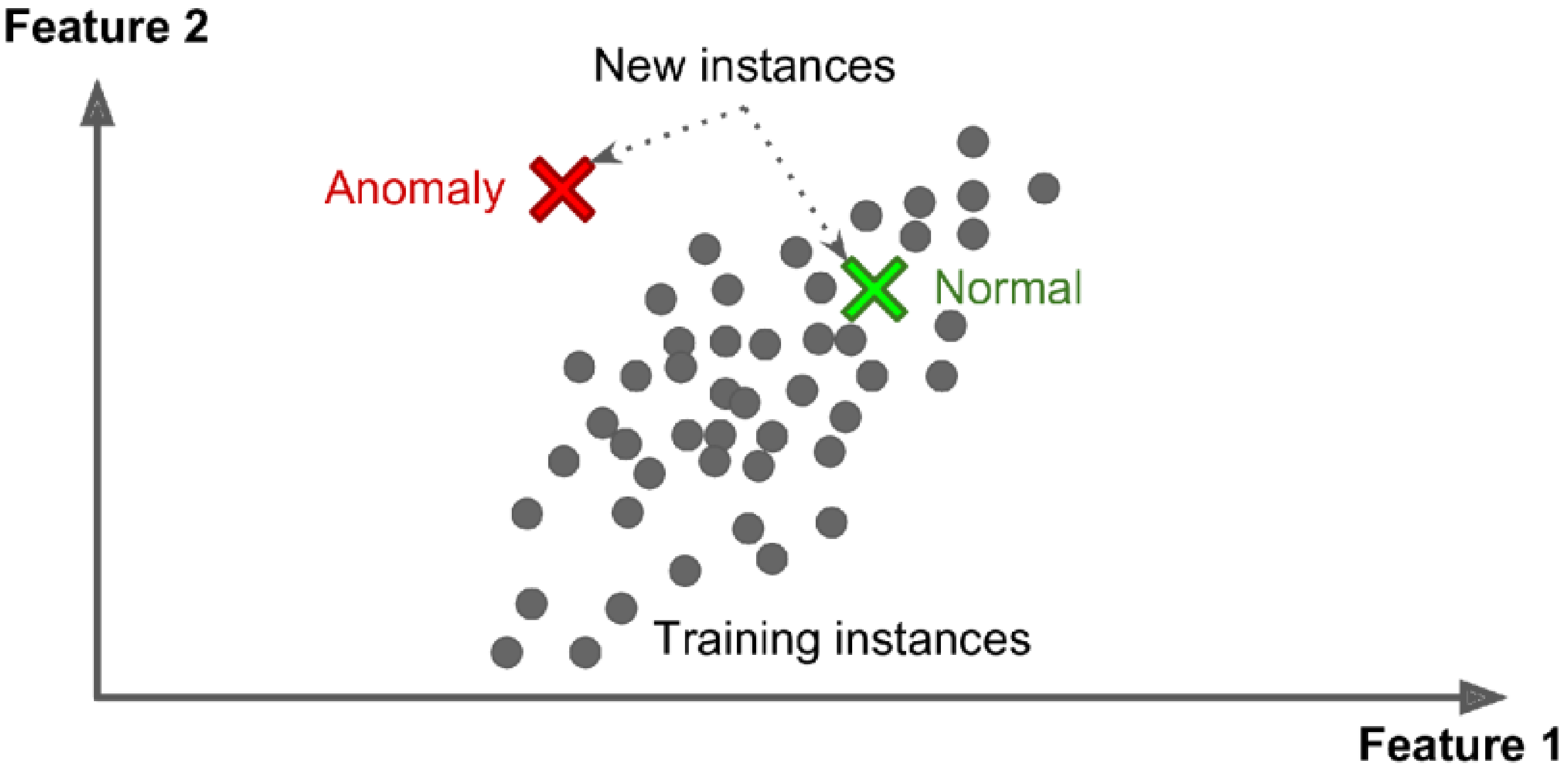
# Unsupervised Learning

**Another example is visualization algorithms such as t-SNE for clustering:**



# Unsupervised Learning

**Another example is anomaly/novelty detection, *e.g.*, catching manufacturing defects:**



# Unsupervised Learning

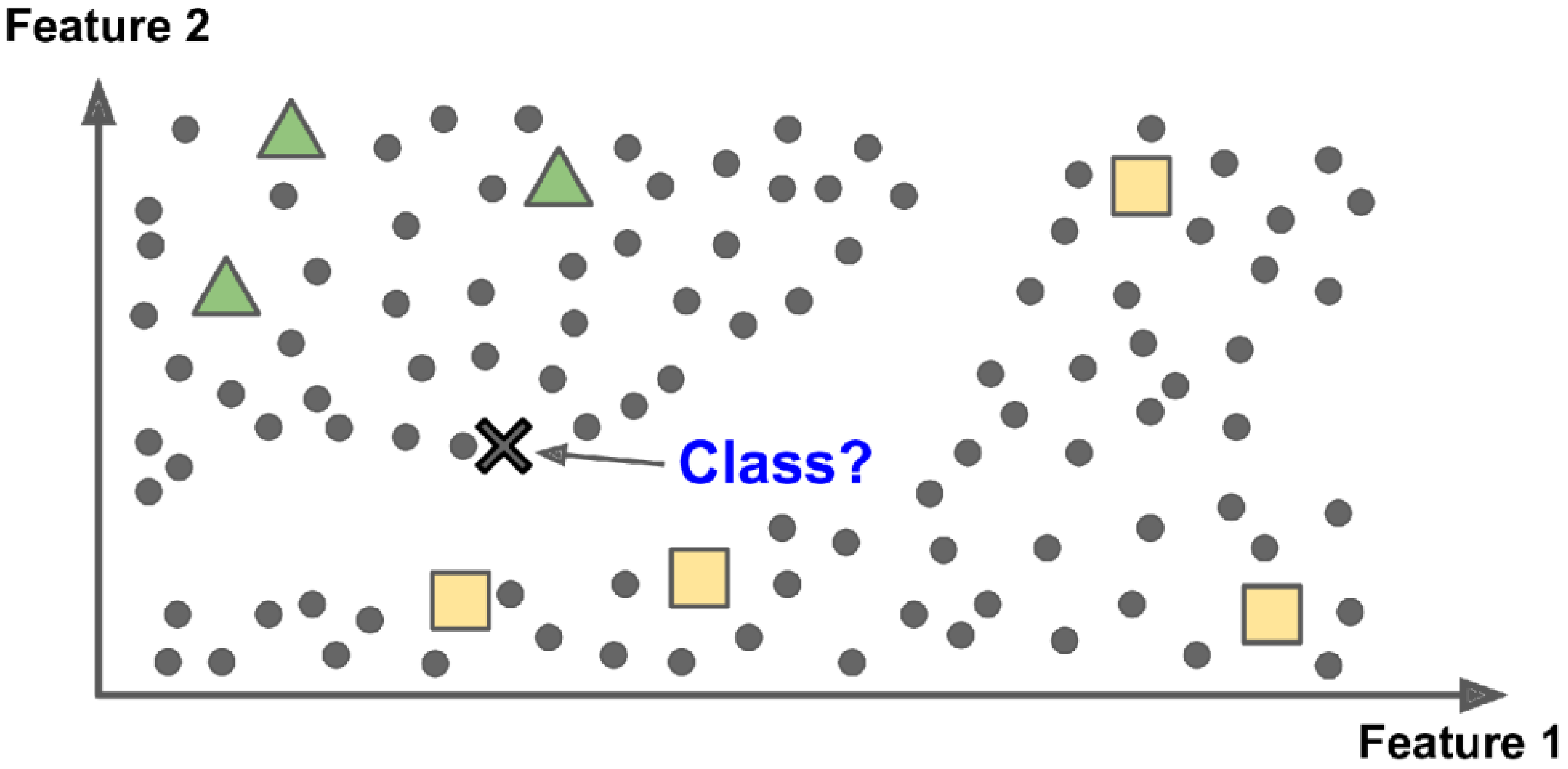
**Some of the most important unsupervised learning algorithms:**

- **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** (*i.e.,* rule-based machine learning)**:** Apriori, Eclat

# Semisupervised Learning

**Training data is partially labeled.**

**Most semisupervised learning algorithms are combinations of unsupervised and supervised algorithms, *e.g.*, Google Photos.**



# Reinforcement Learning

**The learning system, called an agent, can observe the environment, select and perform actions, and get rewards in return (or penalties). 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.**

Examples:

- Robots learn how to walk
- DeepMind’s AlphaGo program beat the world champion at the game of Go

# Types of Machine Learning Sytems

**Whether or not the system can learn incrementally from a stream of incoming data:**

- **Batch learning (or offline learning):** Train using all the available data and retrain once data is updated.

- **Online learning:** Train the system incrementally by feeding it data instances sequentially, in small groups called mini-batches. An example is a system that predicts stock prices.

Some relevant terms:

Out-of-Core Learning; Learning Rate; Inertia

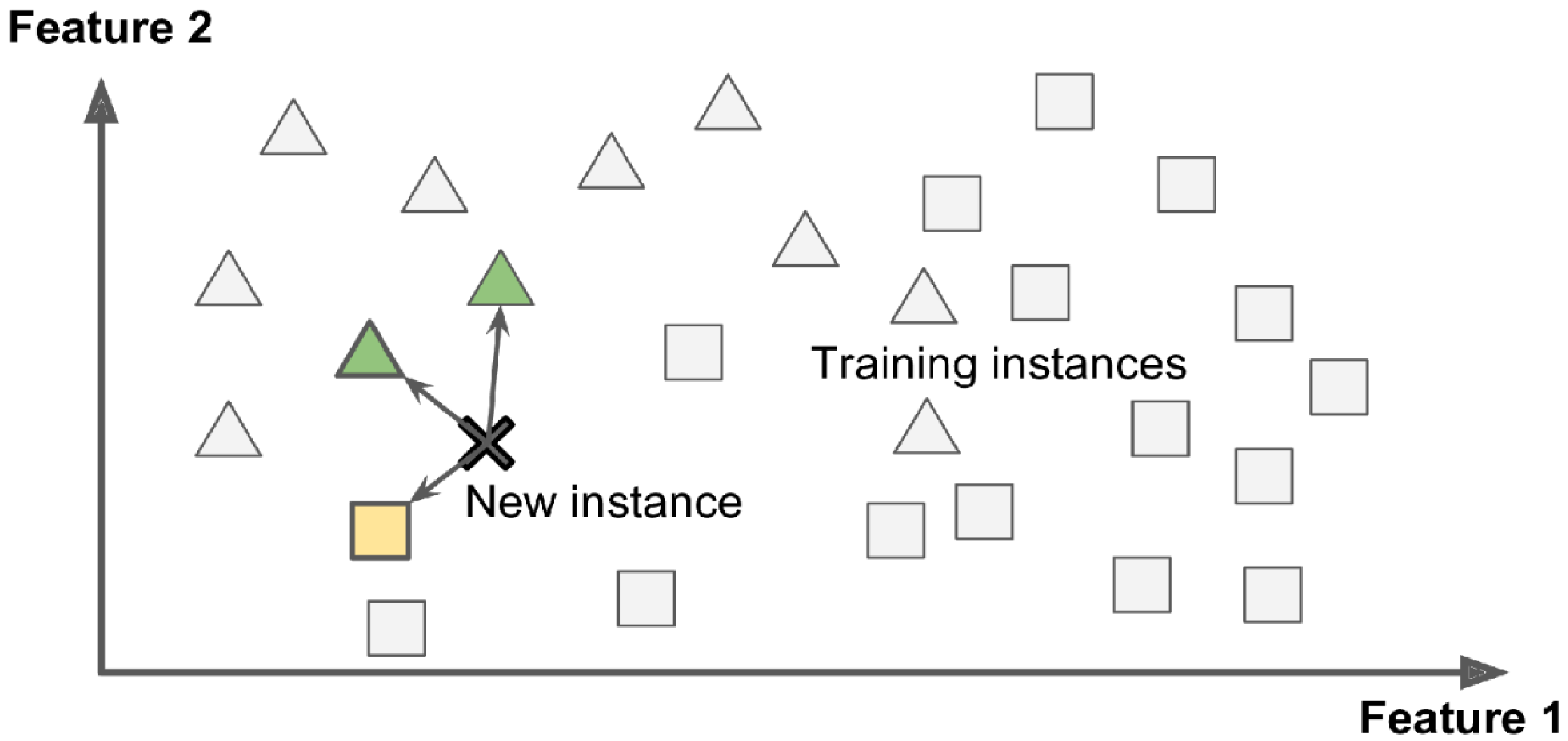
# Types of Machine Learning Sytems

**By how they generalize to (*i.e.*, make predictions for) examples not seen before:**

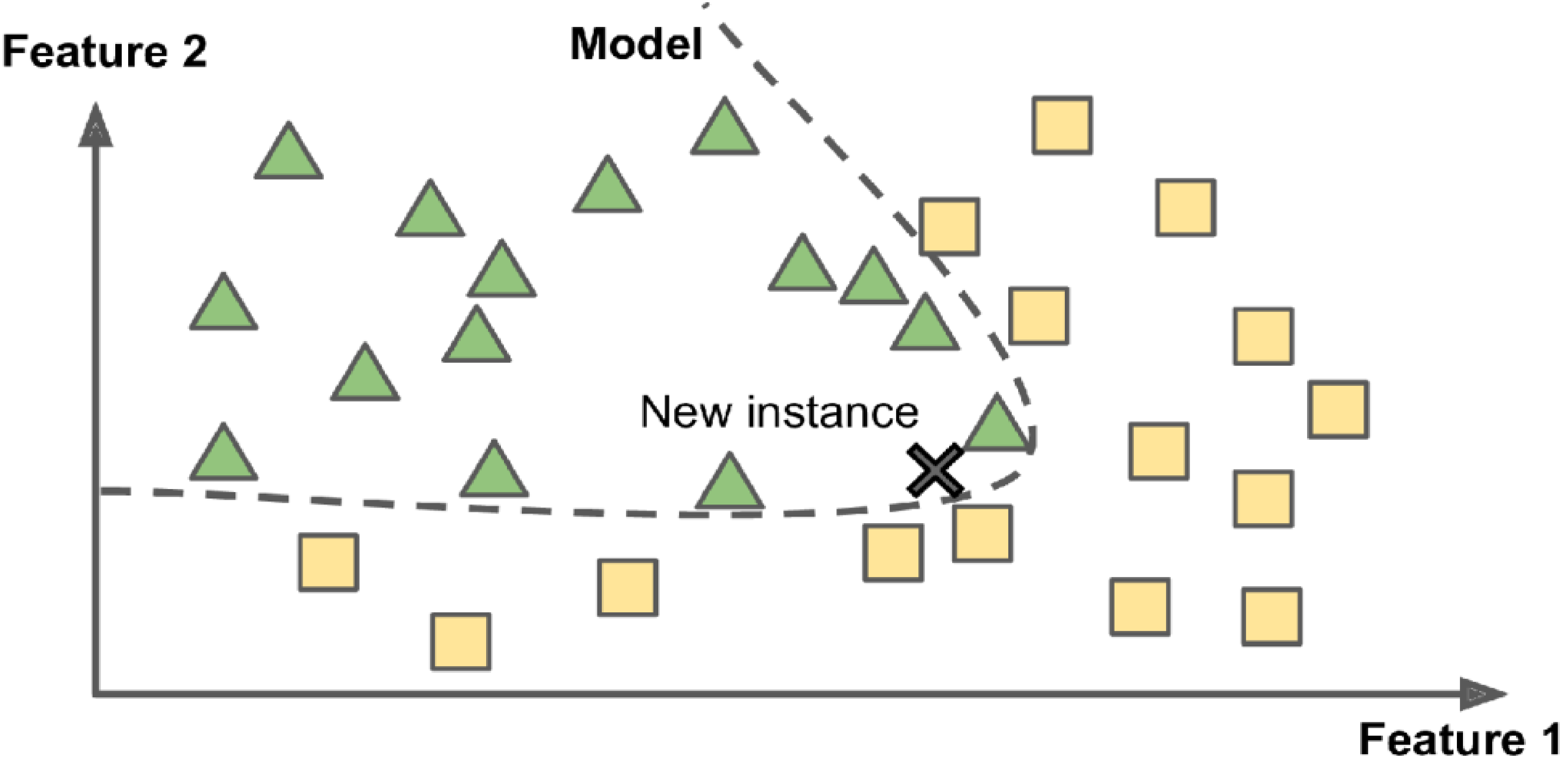
- **Instance-Based:** The system learns the examples by heart, then generalizes to new cases by using a similarity measure to compare them to the learned examples.

- **Model-Based:** Creates a (mathematical) model and uses that model to make predictions.

# Instance-Based Learning



# Model-Based Learning



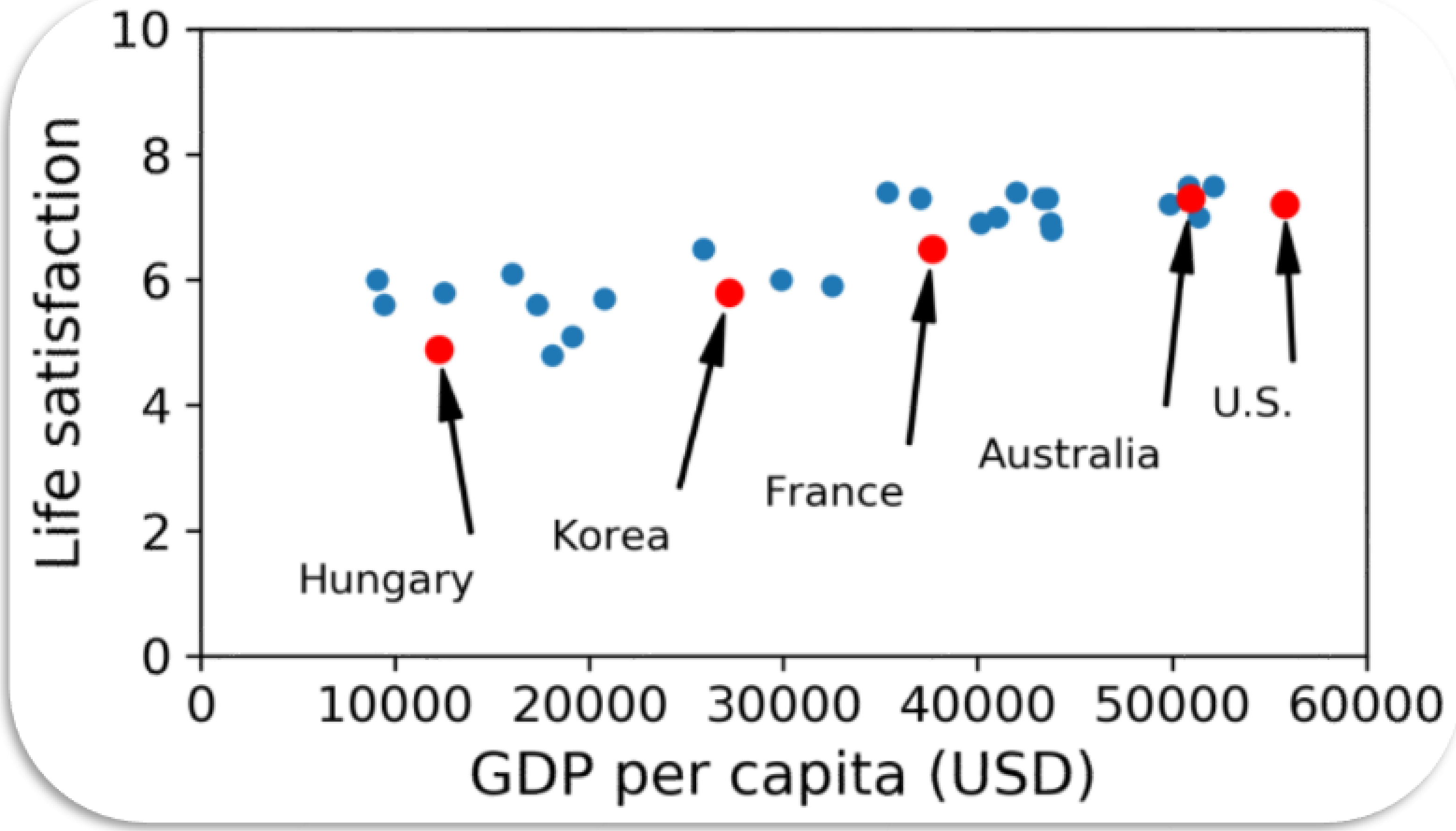
# Linear Regression

**Define the model:** Choose the "class" of functions that relates the inputs (x) to the output (y) and apply to data.

For example:

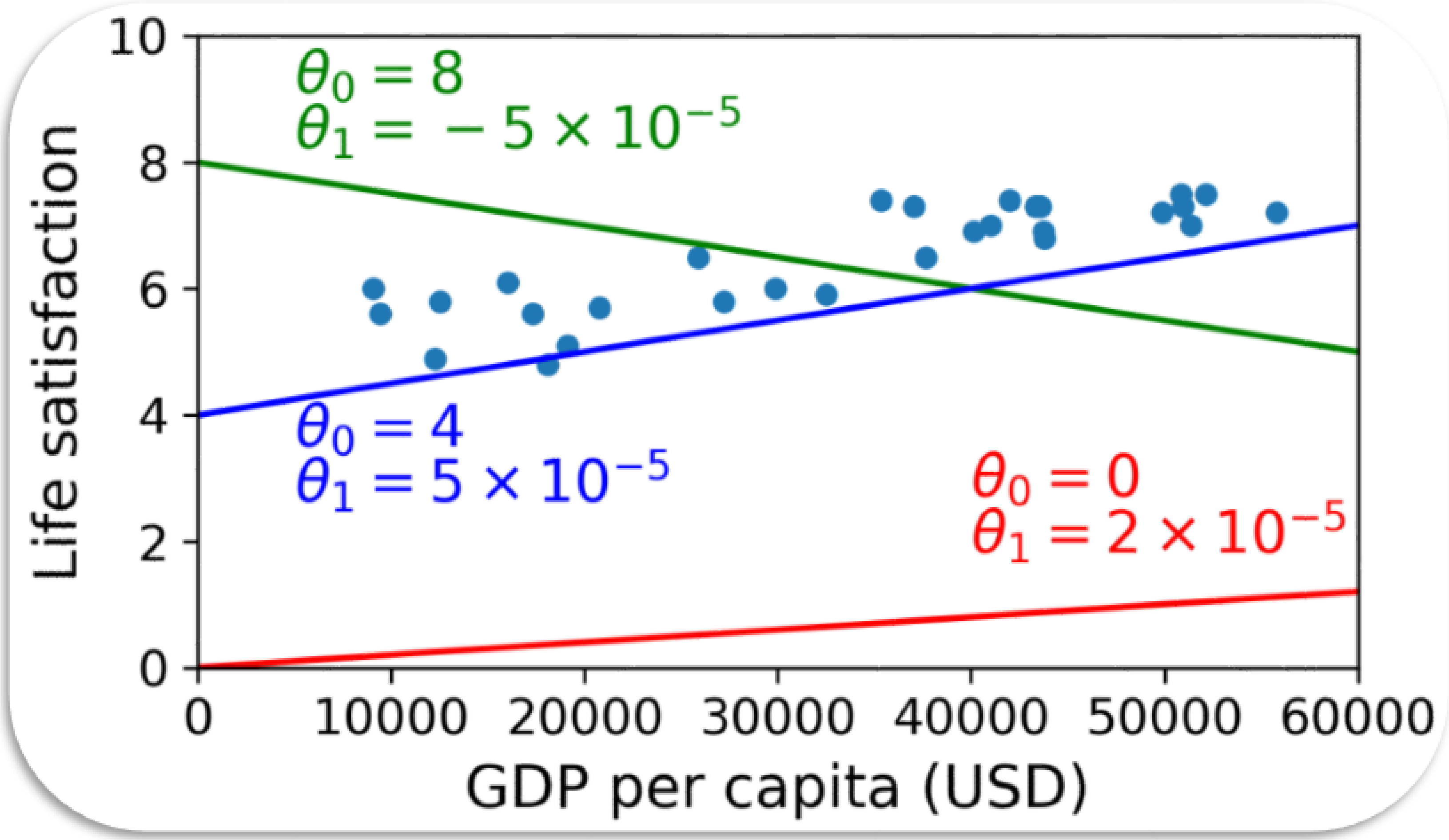
$$\hat{y} = b_0 + b_1 x$$

which is a model with two model parameters.



# Linear Regression

**Tweaking the model parameters gives different possible linear models:**



# Linear Regression

- **Model:** Can refer to a type of model (*e.g.*, Linear Regression), to a fully specified model architecture (*e.g.*, Linear Regression with one input and one output), or to the final trained model ready to be used for prediction (*e.g.*, Linear Regression with one input and one output, using $b_0 = 4.85$ and $b_1 = 4.91 \times 10^{-5}$ ).

- **Model Selection:** Choosing the type of model and fully specifying its architecture.

- **Model Training:** running an algorithm to find the model parameters that will make it best fit the training data (and hopefully make good predictions on new data)

# Linear Regression

**How to find best-performing model parameters?**

**Specify a performance measure:**

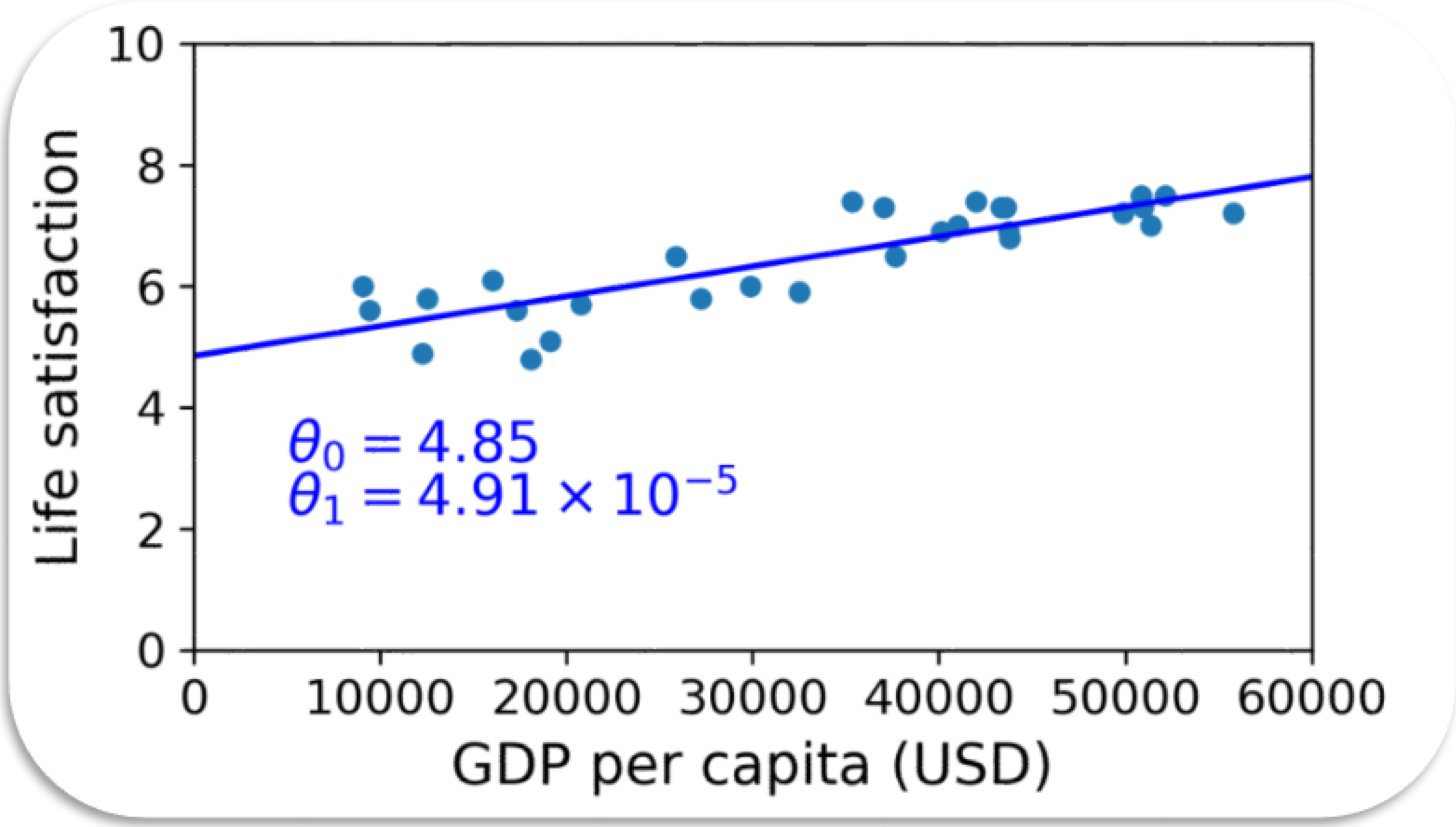
- Define a utility function (or fitness function) that measures how good your model is, or,

- Specify a cost function that measures how bad it is.

For Linear Regression problems, a cost function that measures the distance between the linear model's predictions and the training examples is typically used. The objective is to minimize this distance.

# Linear Regression

**The linear model that fits the training data best.**




**But is it representative of the new cases you want to generalize to?**

# Linear Regression

## Nonrepresentative Training Data