CS 540 Introduction to Artificial Intelligence

Machine Learning Overview

University of Wisconsin-Madison

Spring 2026 Sections 1 & 2

Announcements

- **HW2** due on **Wednesday February 11th at 11:59 PM**

- HW3 will also be released on Wednesday

- Class roadmap:

ML Introduction
ML Unsupervised I
ML Unsupervised II
ML: Linear Regression

Machine Learning

Outline

- NLP Review
- What is machine learning?
- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
  - Clustering
  - Self-Supervised Learning
- Reinforcement Learning

Review: Language Models

- Basic idea: use probabilistic models to **assign a probability to a sentence W**

$$P(W) = P(w_1, w_2, \ldots, w_n) \text{ or } P(w_{\text{next}}|w_1, w_2 \ldots)$$

- Recall the chain rule of probability:

$$P(w_1, w_2, \ldots, w_n) = P(w_1)P(w_2|w_1) \ldots P(w_n|w_{n-1} \ldots w_1)$$

- Markov assumption with shorter history:

$$P(w_i|w_{i-1}w_{i-2} \ldots w_1) = P(w_i|w_{i-1}w_{i-2} \ldots w_{i-k})$$

Review: **n**-gram Model

- **k=0 Unigram Model:** Full independence assumption:
  - (Present doesn't depend on the past)

$$P(w_1, w_2, \ldots, w_n) = P(w_1)P(w_2)\ldots P(w_n)$$

- **k = 1 Bigram Model: Markov Assumption:**
  - (Present depends on immediate past)

$$P(w_1, w_2, \ldots, w_n) = P(w_1)P(w_2|w_1)P(w_3|w_2)\ldots P(w_n|w_{n-1})$$

- Can do trigrams, 4-grams, and so on
  - More expressive as *n* goes up
  - Harder to estimate

Review: Bag of Words (BOW)

Document

Madison is the capital city of Wisconsin. The city is known for its beautiful lakes and the university.

$count(w_i)$ →

$\frac{count(w_i)}{\sum_{i=1}^{n} count(w_i)}$ → normalized

Word	$count(w_i)$	normalized
and	1	0.056
are	0	0
beautiful	1	0.056
between	0	0
capital	1	0.056
city	2	0.111
for	1	0.056
in	0	0
is	2	0.111
Its	1	0.056
known	1	0.056
lakes	1	0.056
Madison	1	0.056
of	1	0.056
popular	0	0
sits	0	0
summer	0	0
the	3	0.167
two	0	0
university	1	0.056
Wisconsin	1	0.056

Review: Term Frequency - Inversed Document Frequency (TF-IDF)

- Term Frequency (TF_{ij}) : How many times the term i appears in the document j (normalized over the total number of terms in the document j)
  - Bag of Words (BOW)

- Inversed Term Frequency (IDF_{ij}) : How rare a term is in a set of documents.

$$IDF_{ij} = \log(\frac{N}{df_i})$$

← Total number of documents in the corpus (N)

↑ number of documents containing the term w_i (df_i)

- TF-IDF$_{ij}$: How important is the term w_i for the document j

$$TF\text{-}IDF_{ij} = TF_{ij} \times IDF_{ij} = TF_{ij} \times \log(\frac{N}{df_i})$$

Review: Representing Words

Traditional representation: **one-hot vectors**

$$\text{dog} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Dense vectors:

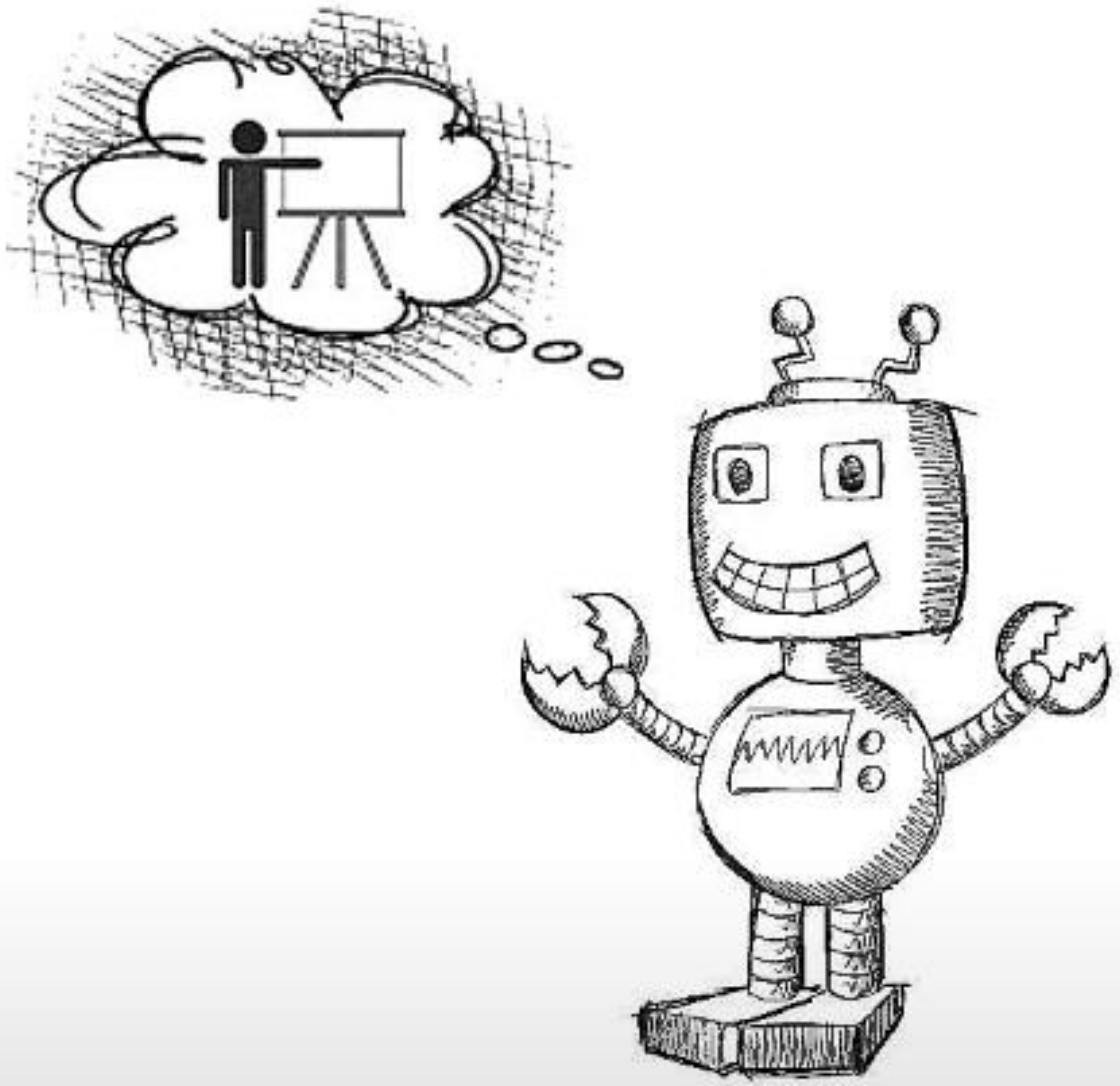
$$\text{dog} = \begin{bmatrix} 0.13 & 0.87 & -0.23 & 0.46 & 0.87 & -0.31 \end{bmatrix}^T$$

$$\text{cat} = \begin{bmatrix} 0.07 & 1.03 & -0.43 & -0.21 & 1.11 & -0.34 \end{bmatrix}^T$$

AKA **word embeddings**

Part I: What is machine learning?

HUMANS LEARN FROM PAST EXPERIENCES

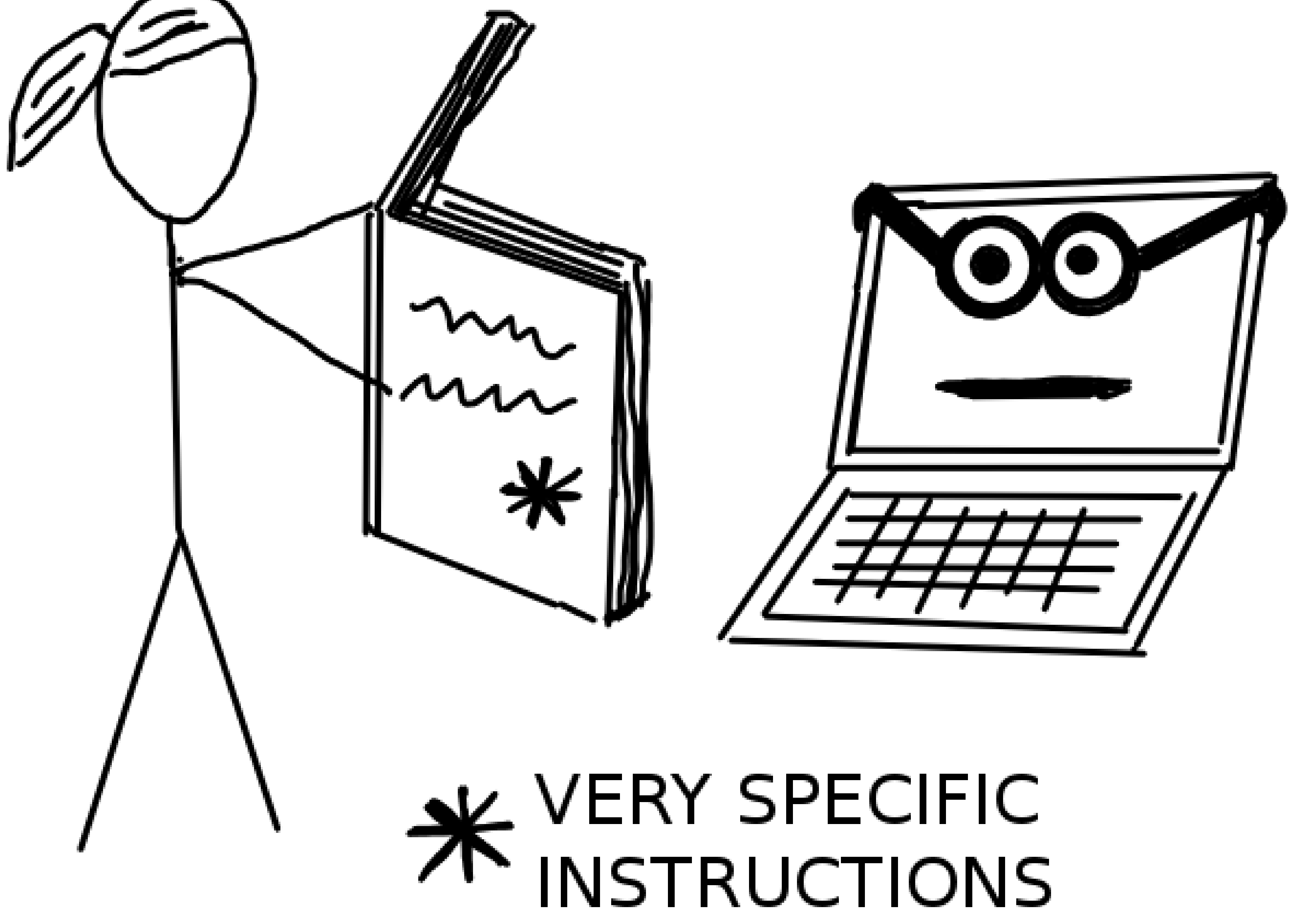


MACHINES FOLLOW INSTRUCTIONS GIVEN BY HUMANS

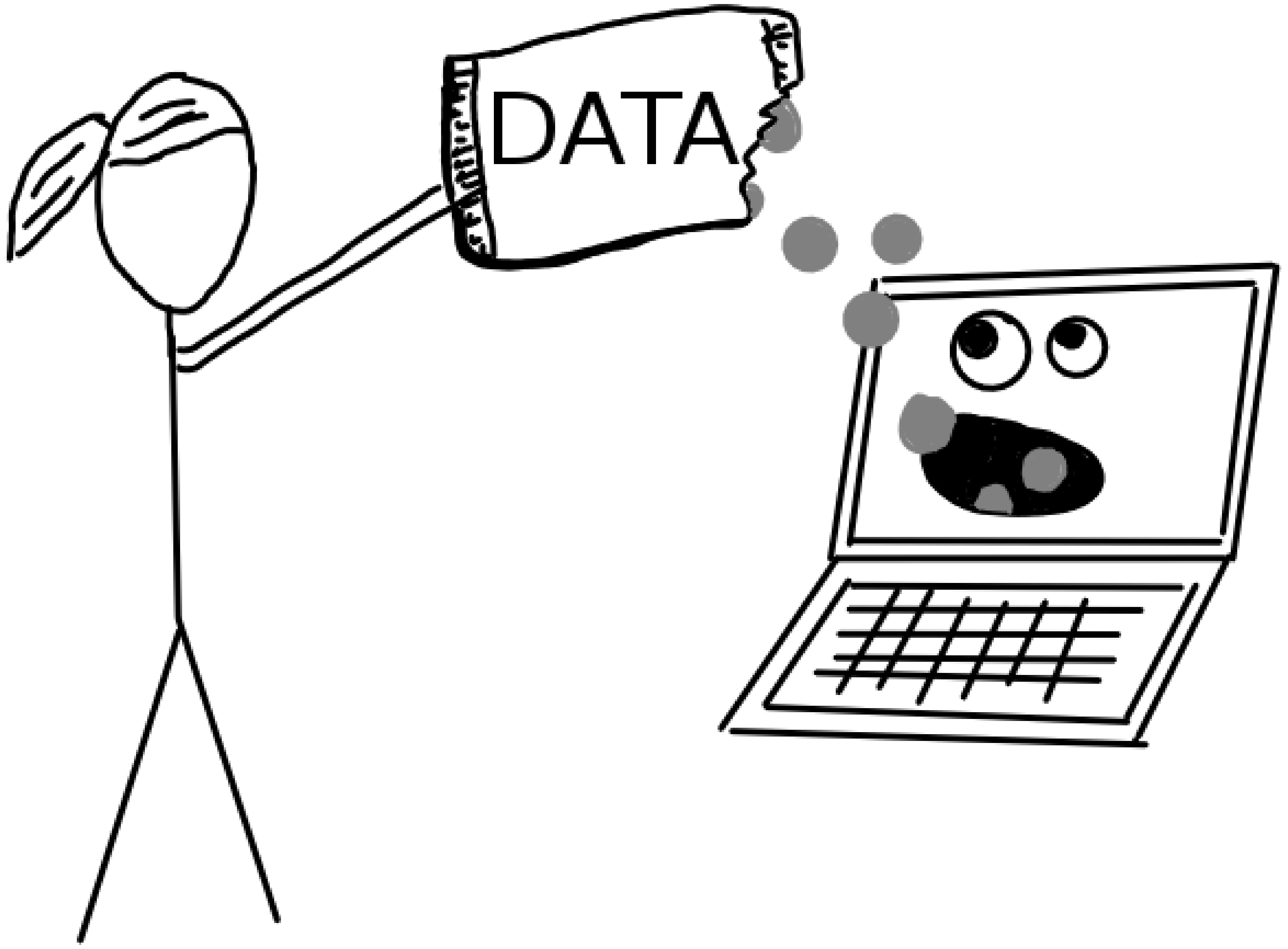
What is machine learning?

- Arthur Samuel (1959): Machine learning is the field of study that gives the computer the ability to learn **without being explicitly programmed**.

Without Machine Learning



With Machine Learning

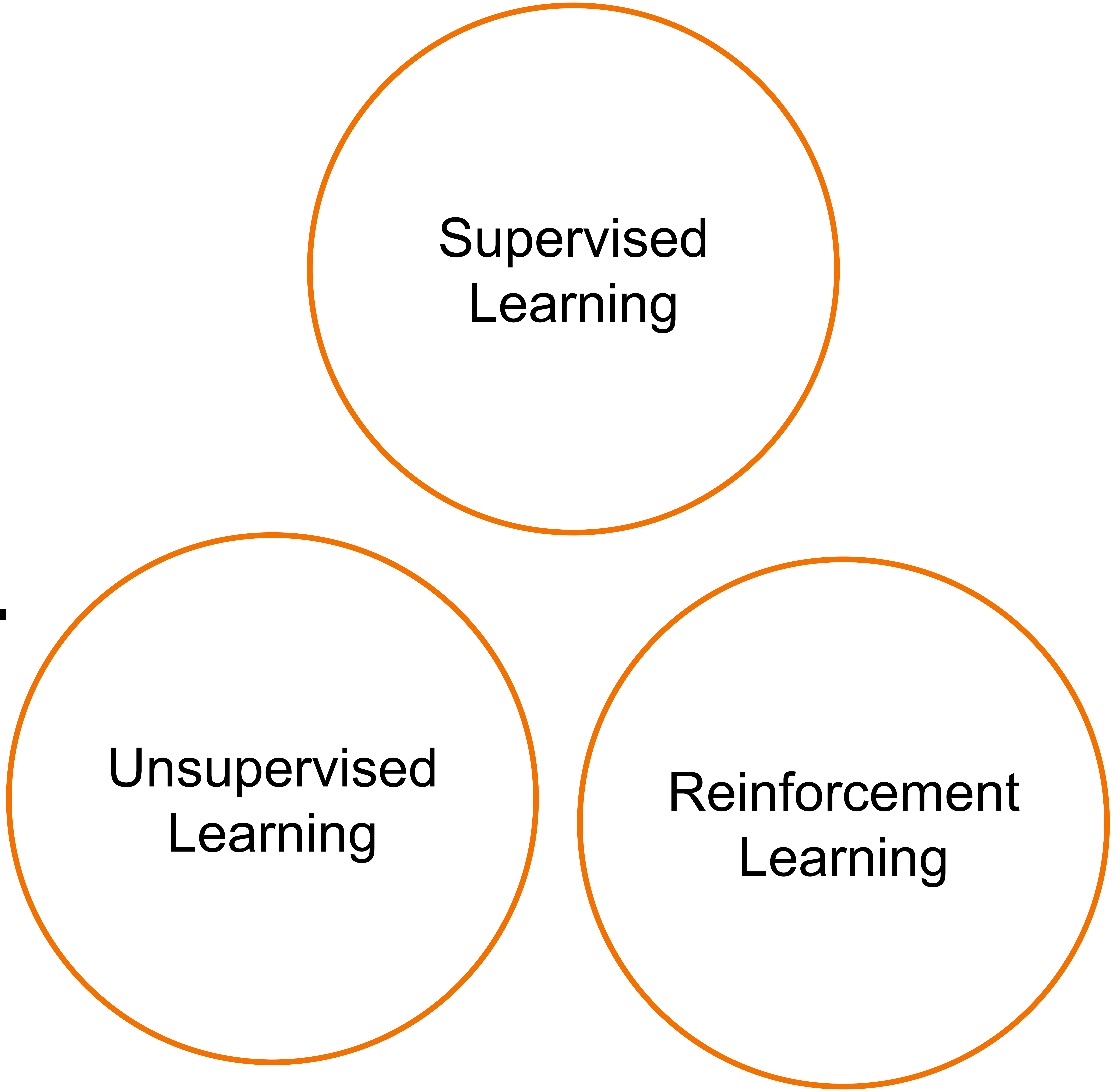


https://tung-dn.github.io/programming.html

What is machine learning?

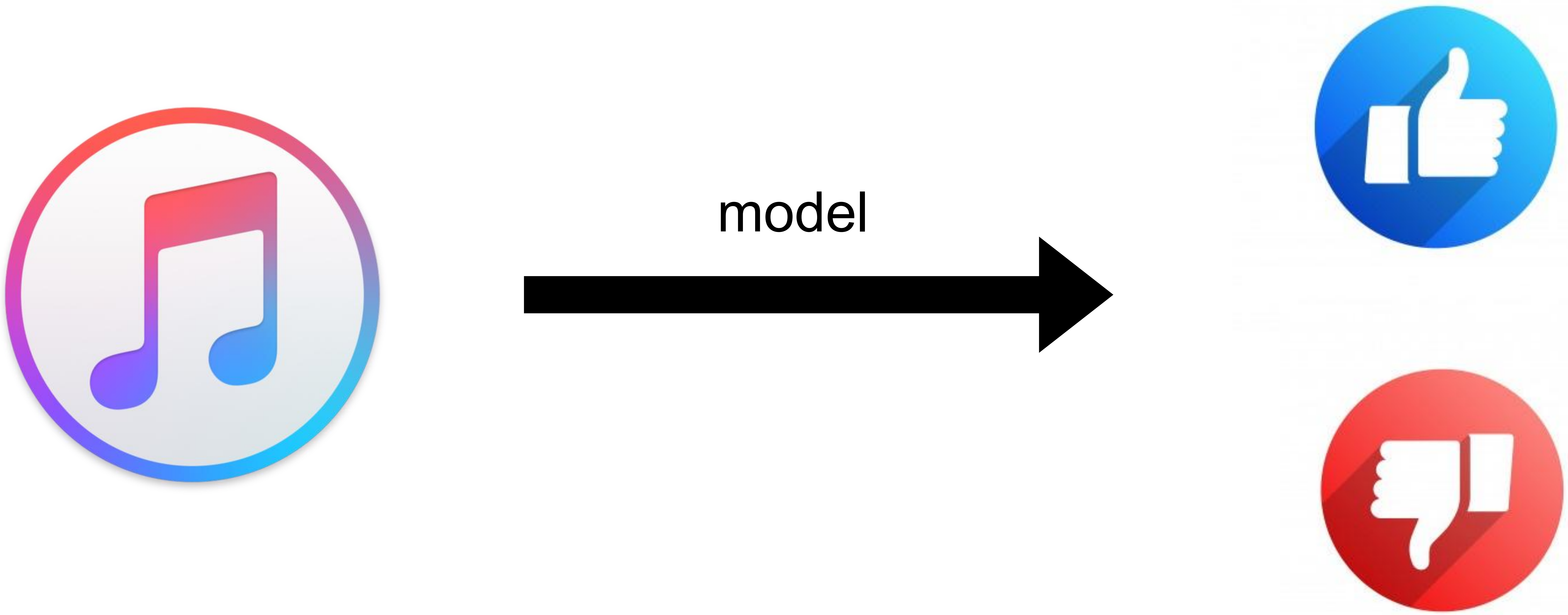
- Arthur Samuel (1959): Machine learning is the field of study that gives the computer the ability to learn **without being explicitly programmed**.
- Tom Mitchell (1997): A computer program is said to learn from **experience E** with respect to some class of **tasks T** and **performance measure P**, if its performance at tasks in T as measured by P, improves with experience E.

Taxonomy of ML



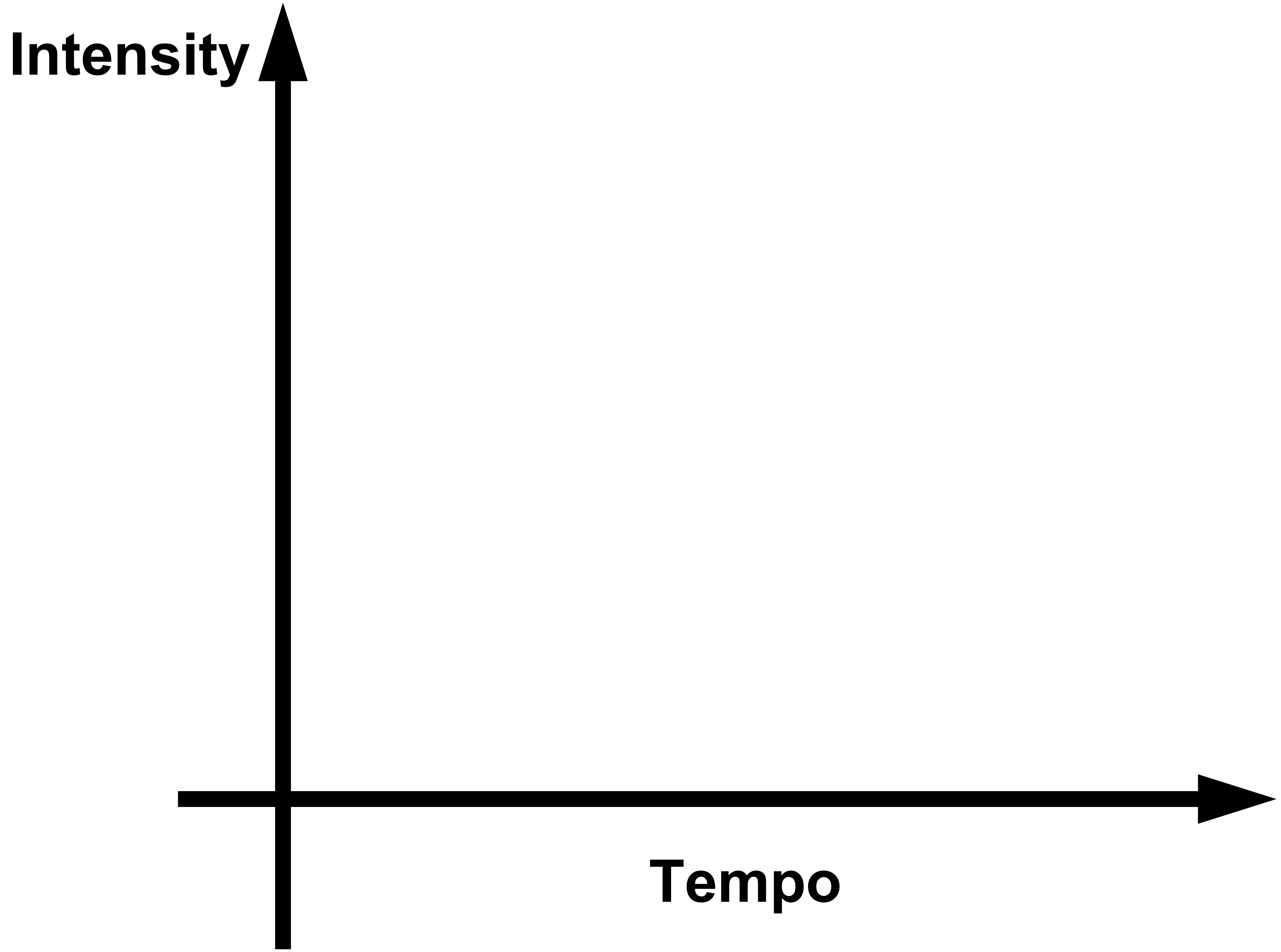
Part II: Supervised Learning

Example 1: Predict whether a user likes a song or not



Example 1: Predict whether a user likes a song or not

User Sharon

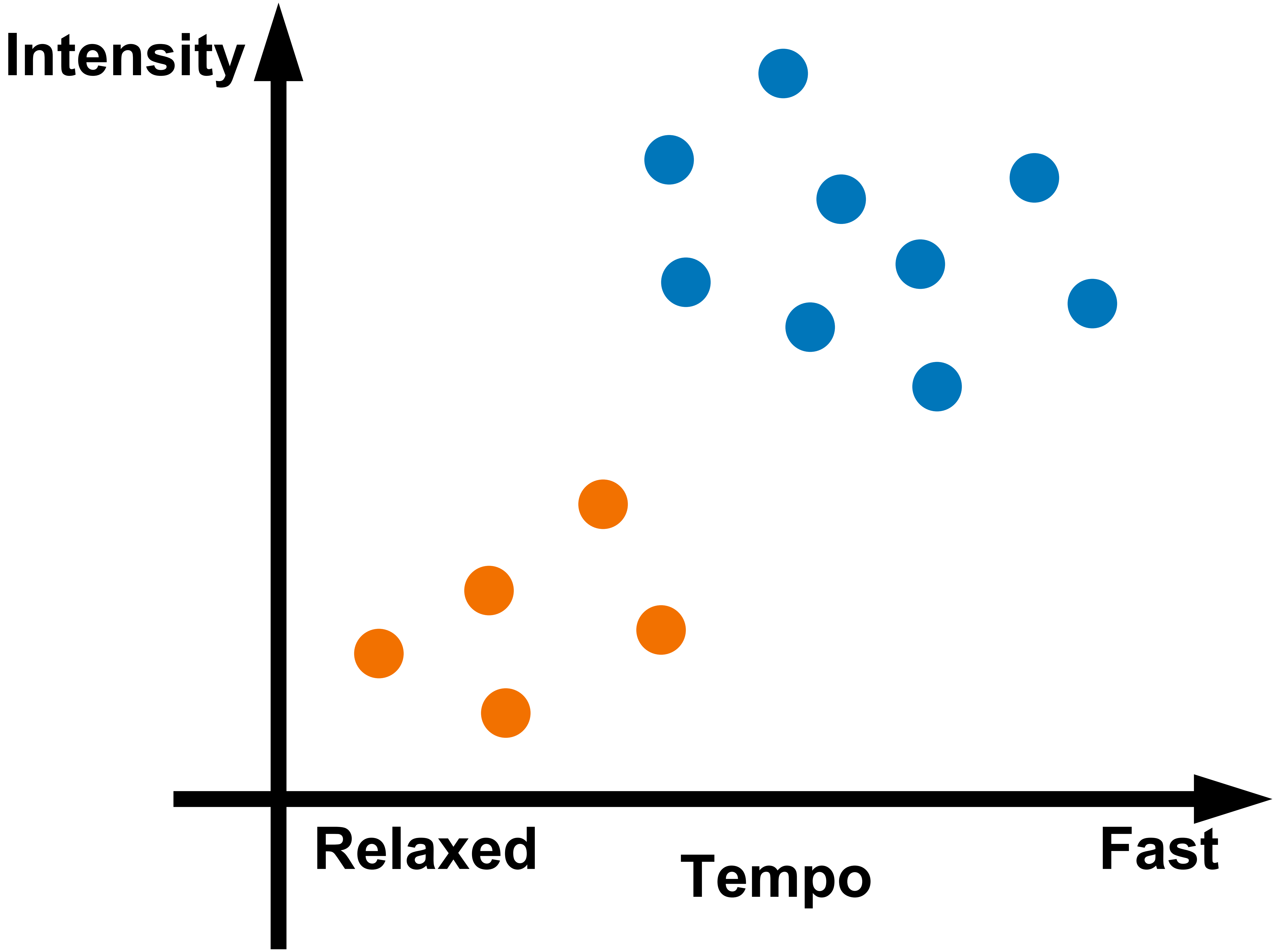


Example 1: Predict whether a user likes a song or not

User Sharon

Dislike

Like

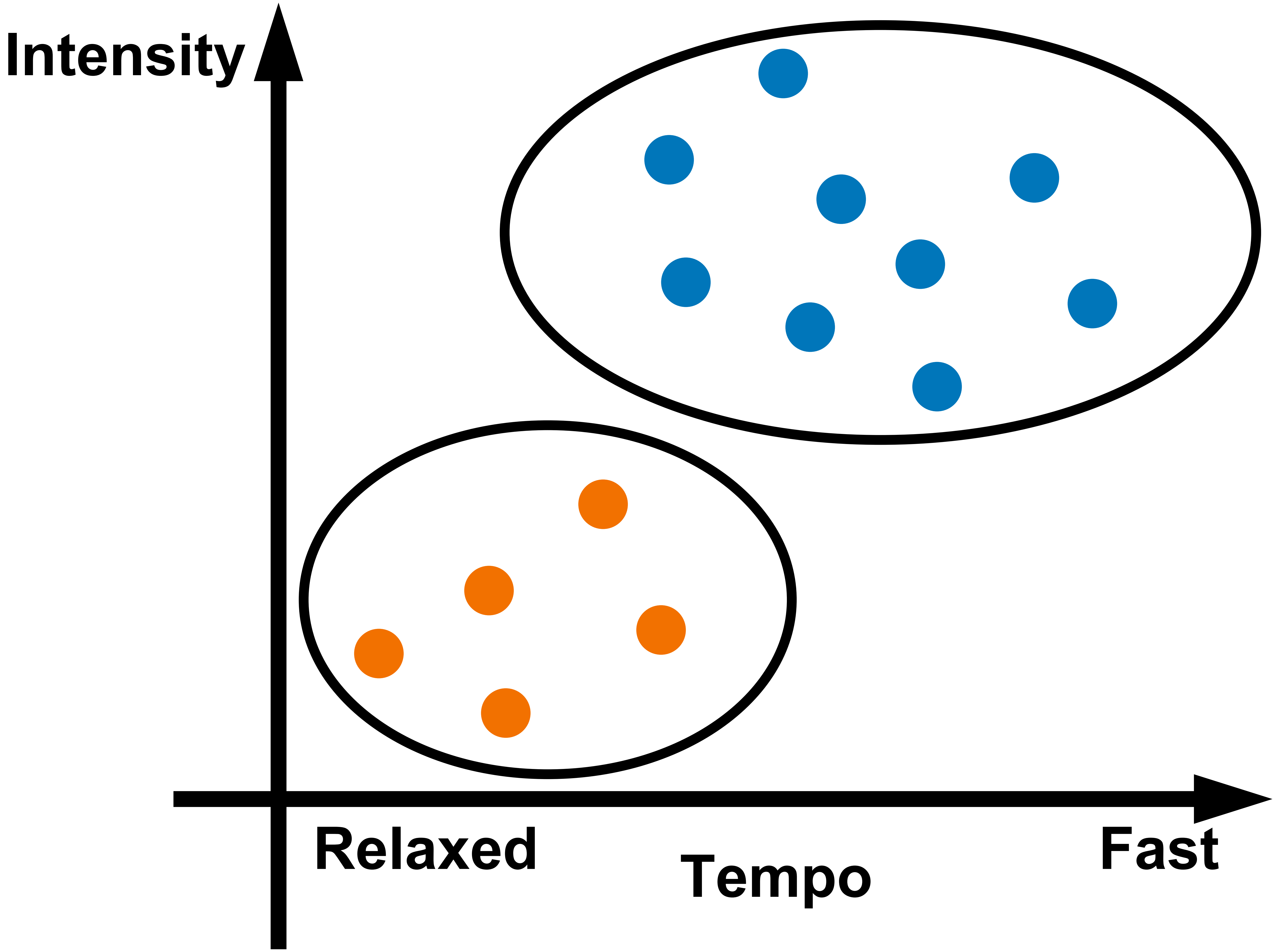


Example 1: Predict whether a user likes a song or not

User Sharon

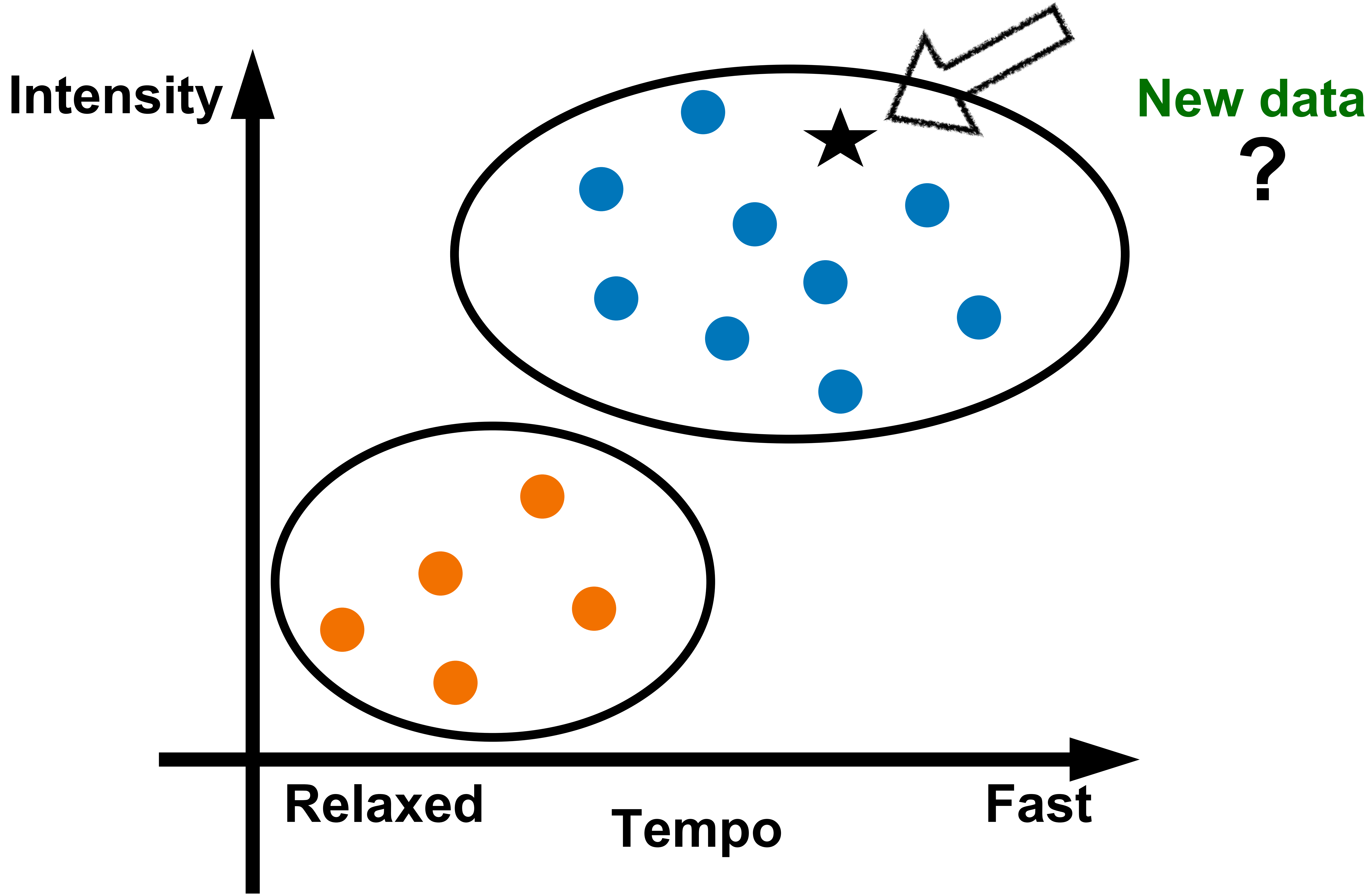
Dislike

Like



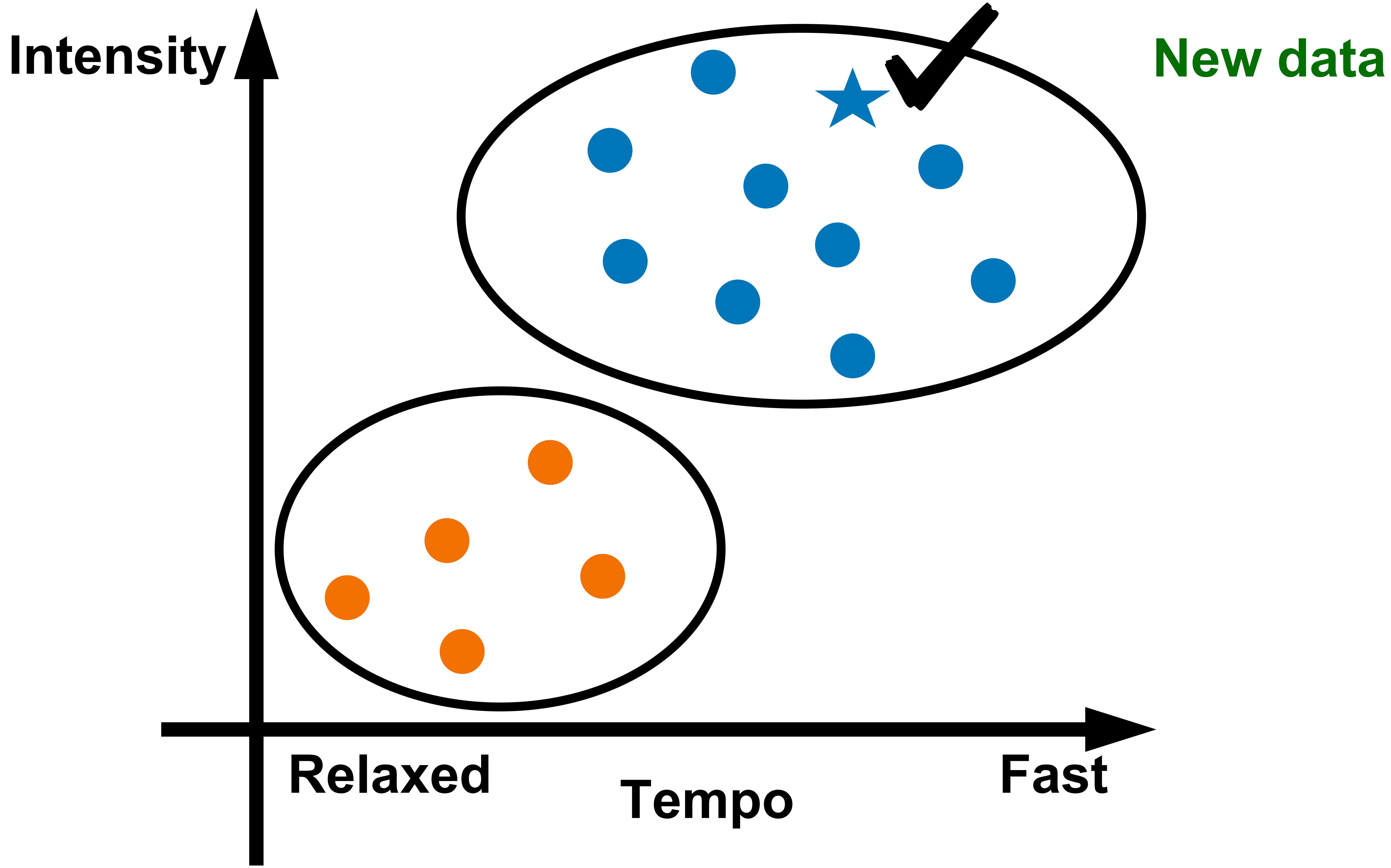
Example 1: Predict whether a user likes a song or not

User Sharon



Example 1: Predict whether a user likes a song or not

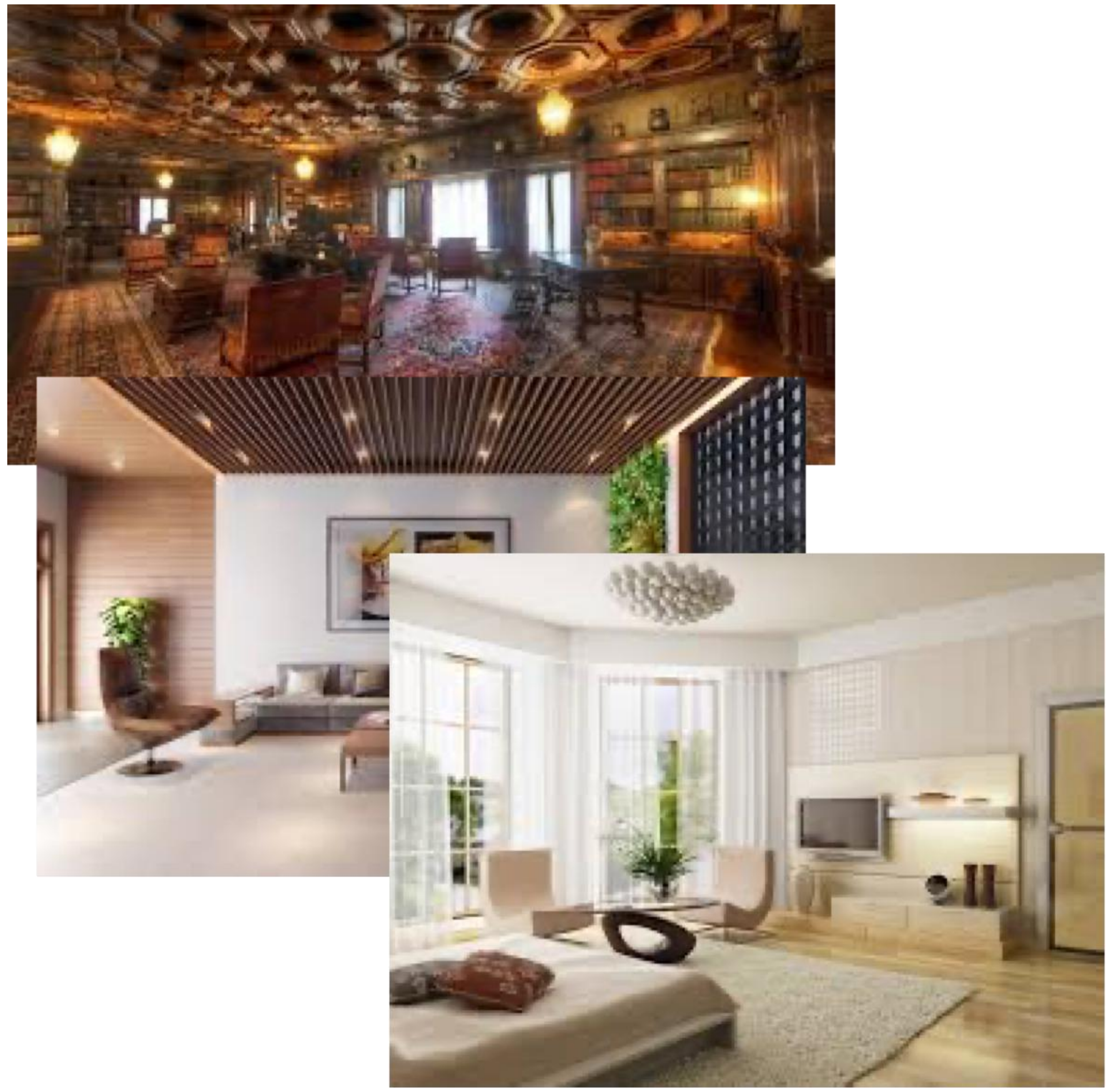
User Sharon



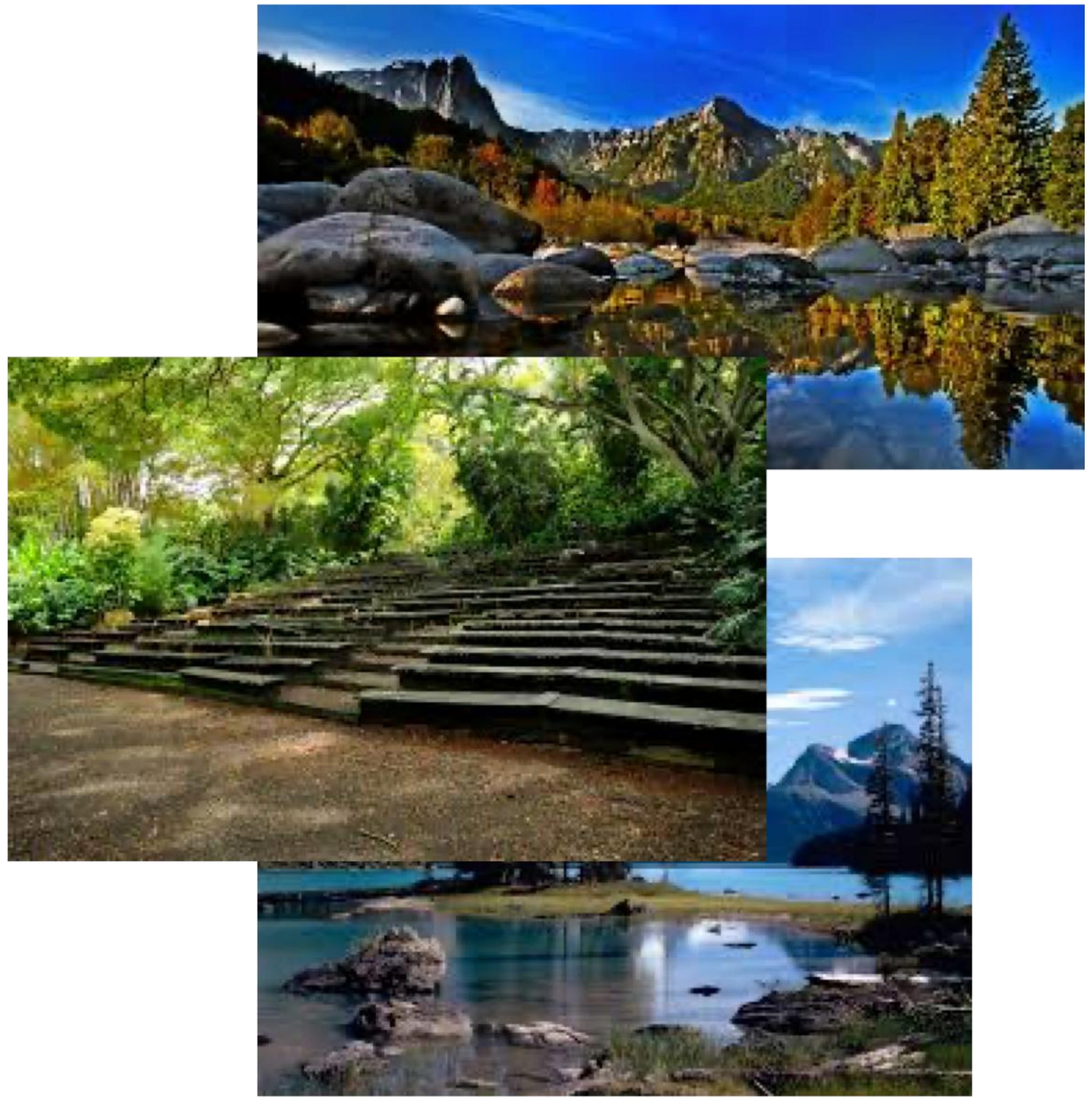
Example 2: Classify Images

http://www.image-net.org/

Example 2: Classify Images

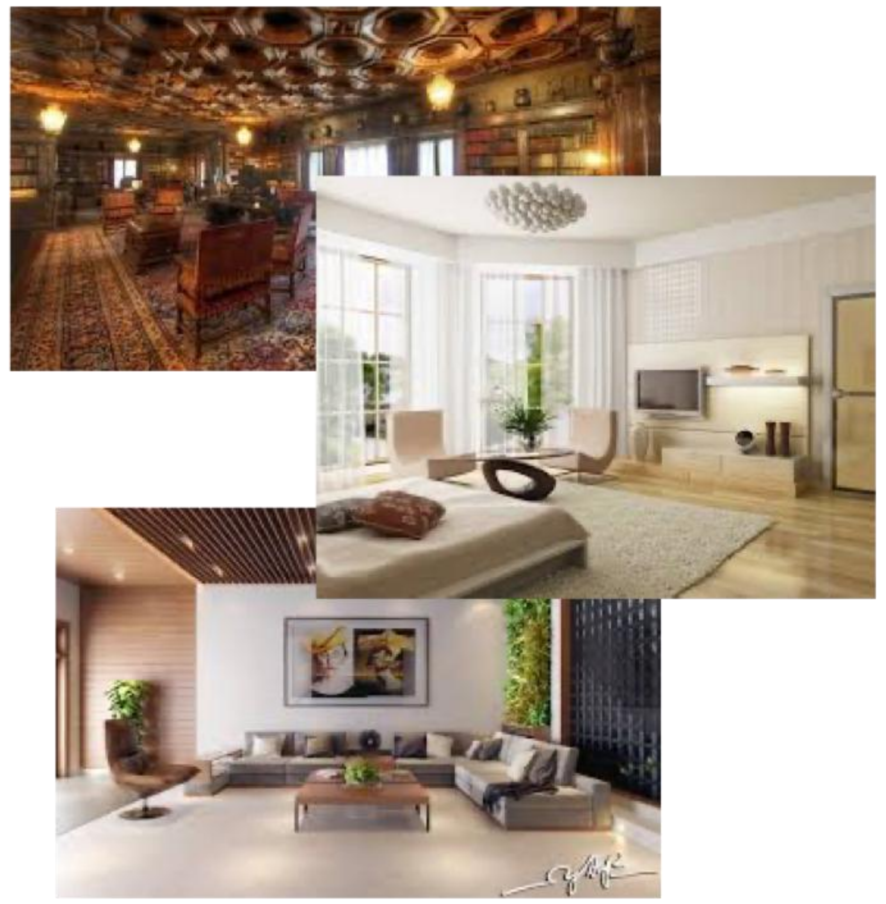


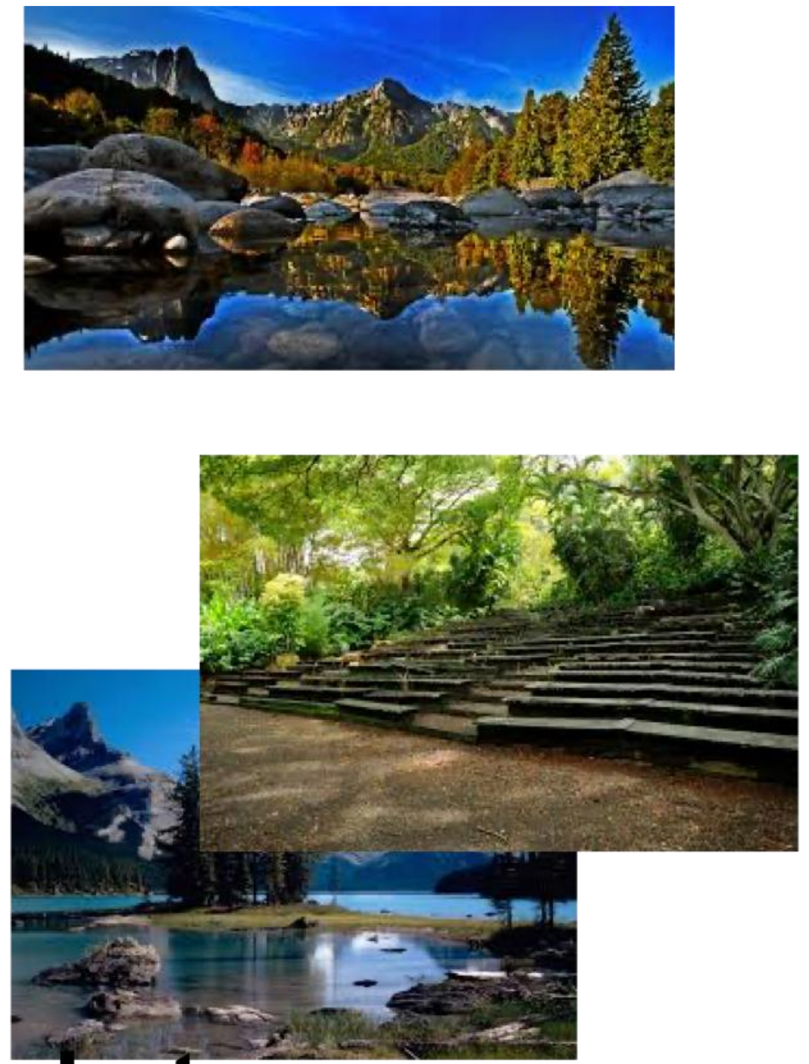
indoor



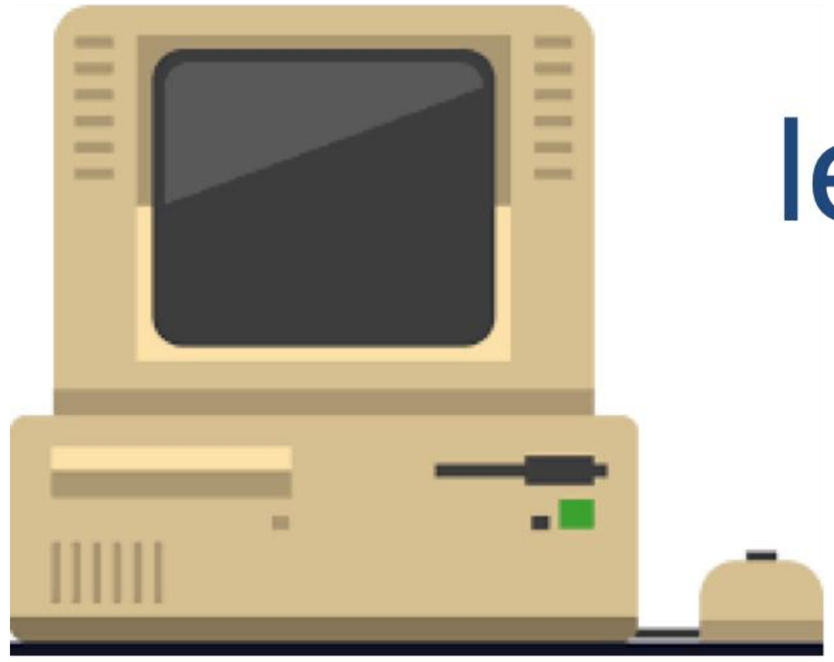
outdoor

Example 2: Classify Images

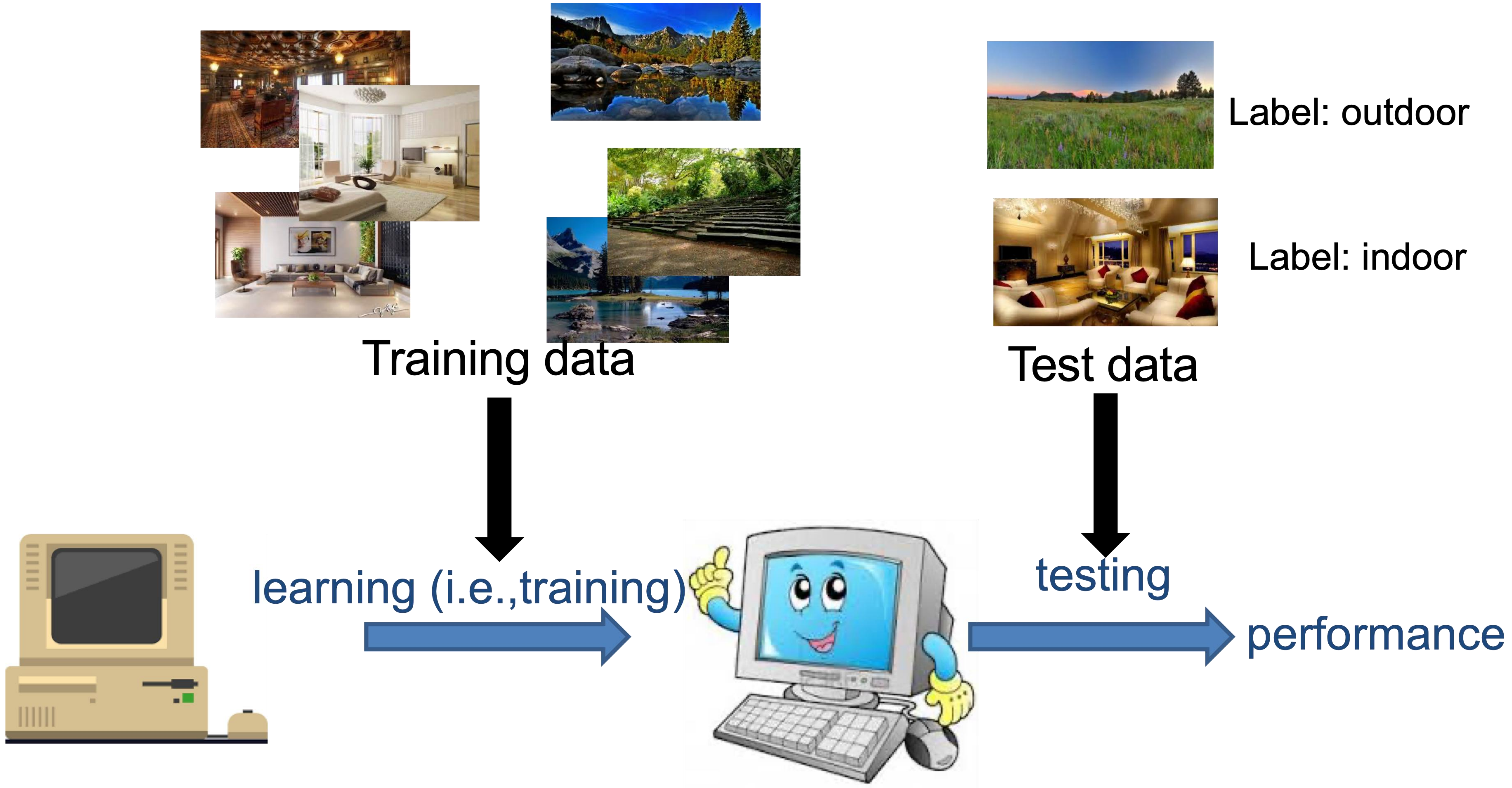




Training data



learning (i.e.,training)


Label: outdoor
Label: indoor
Training data
Test data
learning (i.e.,training)
testing
performance

How to represent data?

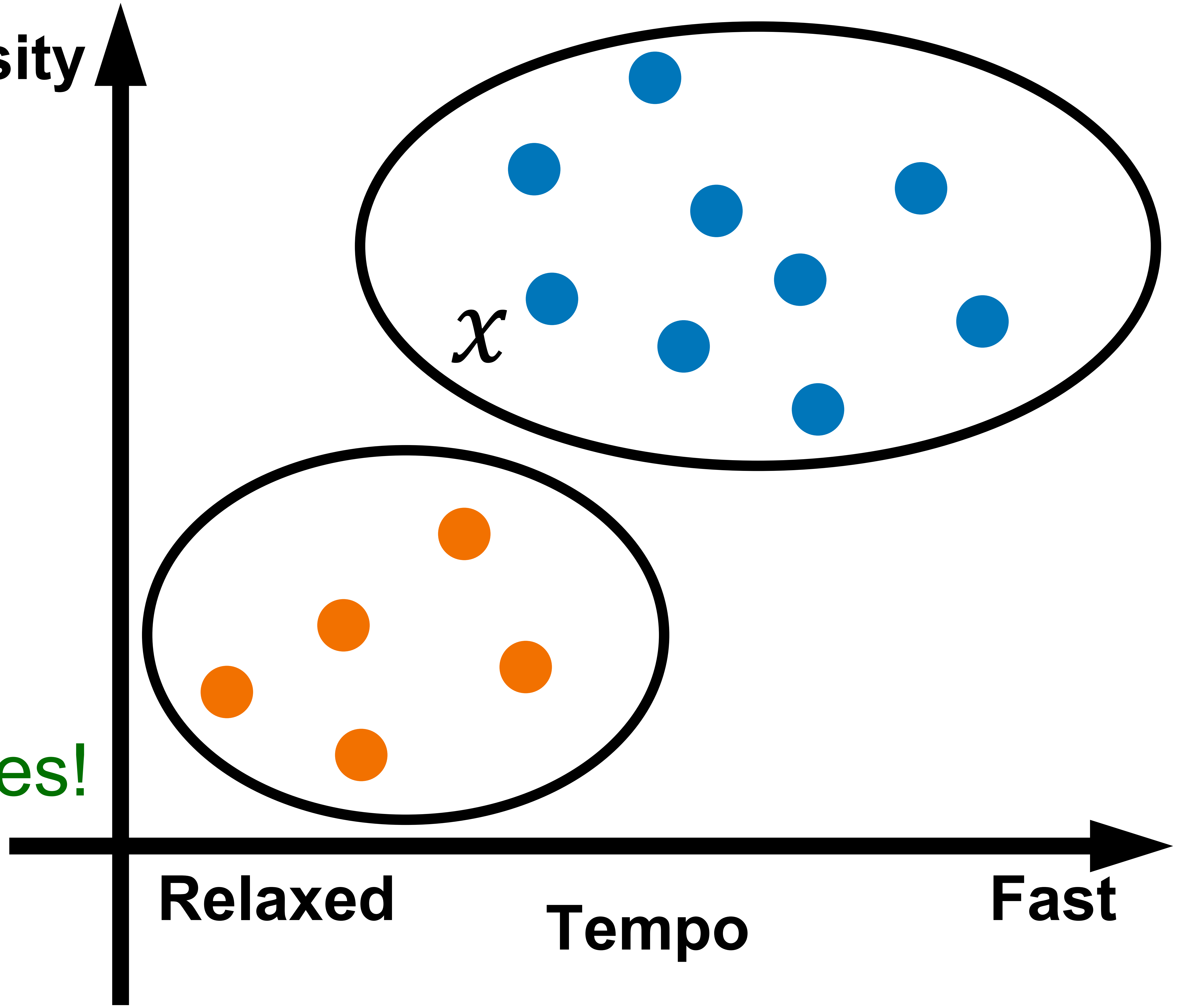
input data

$$x \in \mathbb{R}^d$$

d : feature dimension

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$ Tempo / Intensity

There can be many features!

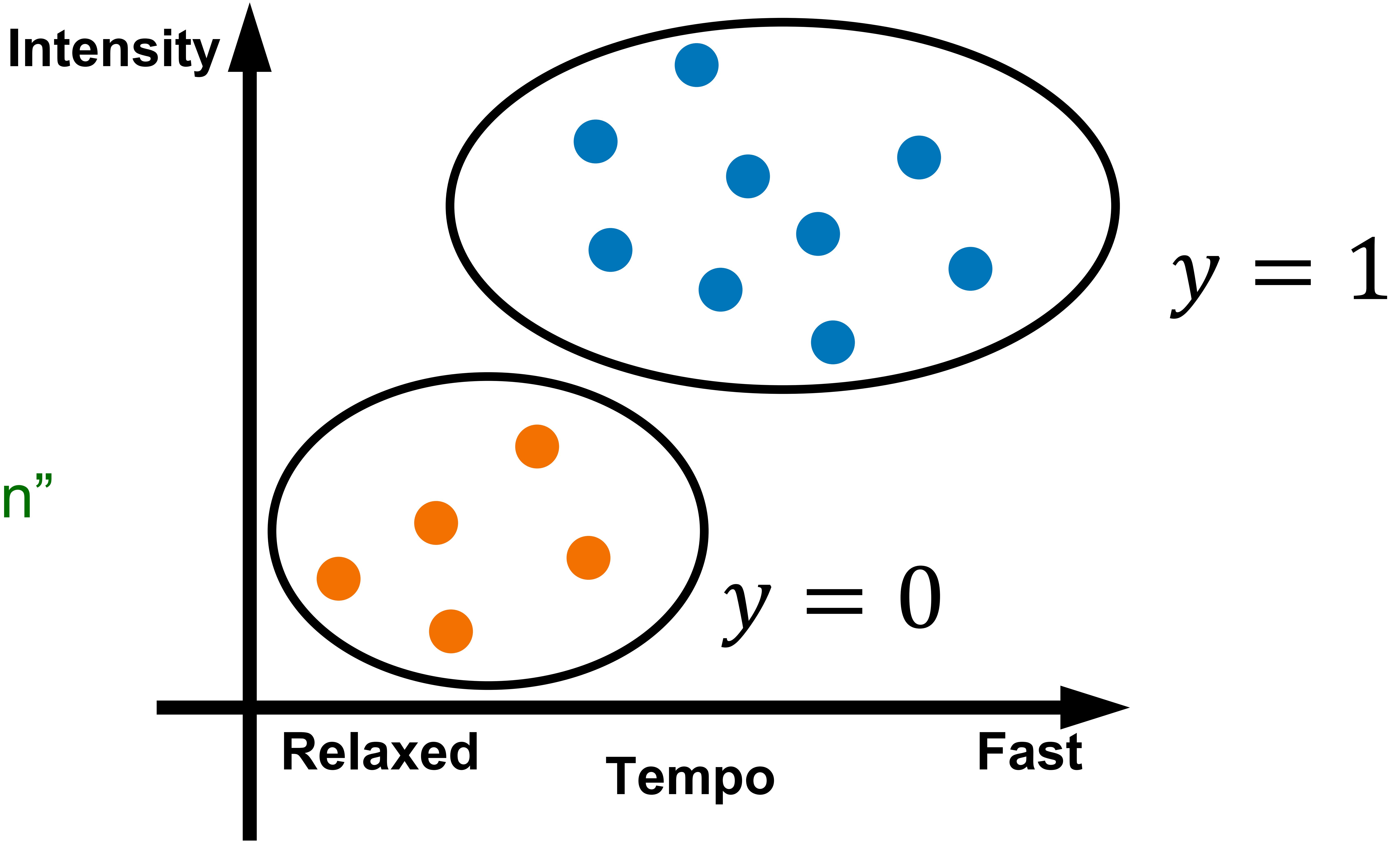


How to represent data?

Label

$y \in \{0,1\}$

Where "supervision" comes from



Represent various types of data

- Image
  - Pixel values

- Bank account
  - Credit rating, balance, # deposits in last day, week, month, year, #withdrawals

Two Types of Supervised Learning Algorithms

Classification

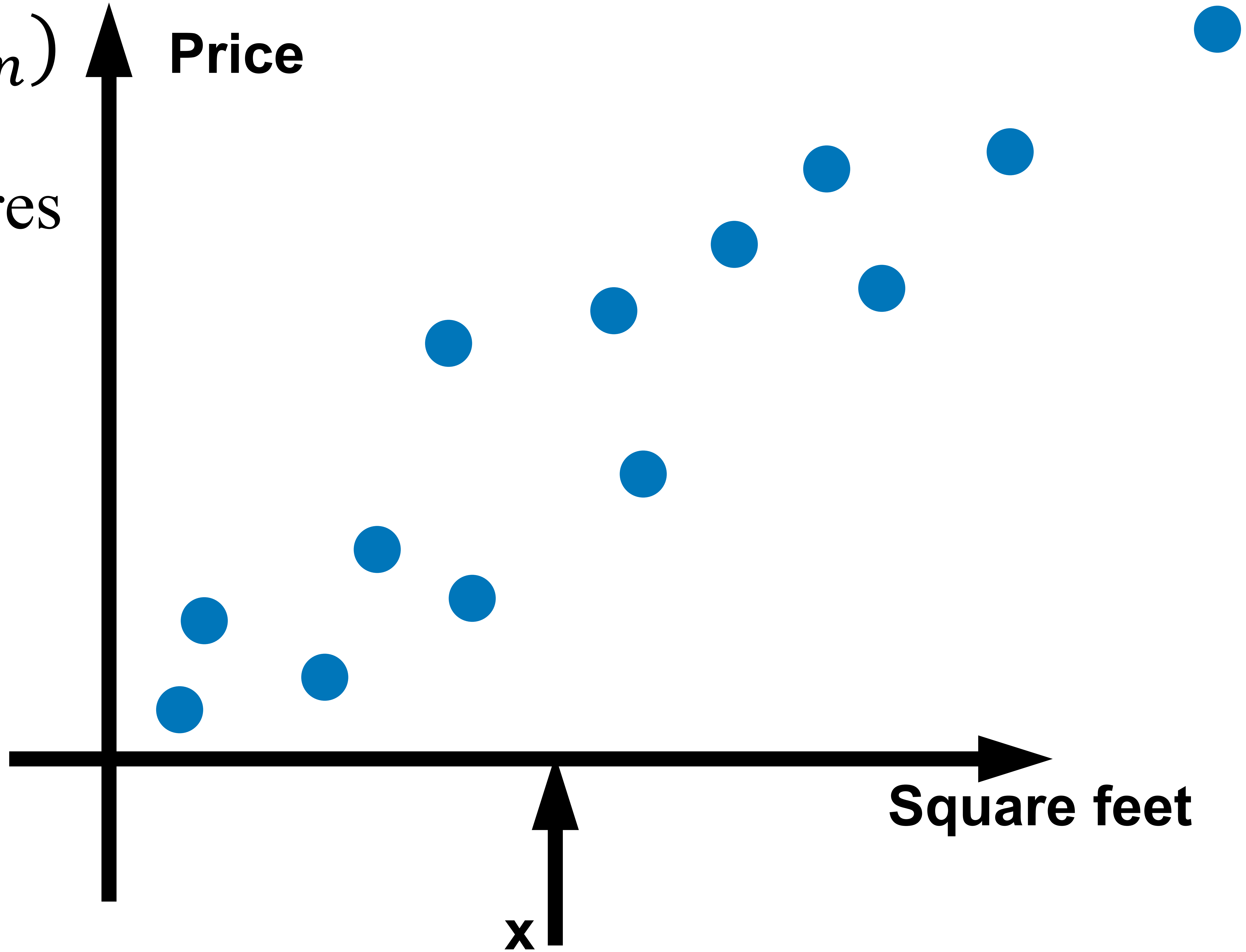
Regression

Example of regression: housing price prediction

Given: a dataset that contains n samples

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

Task: if a residence has **x** squares feet, predict the price?



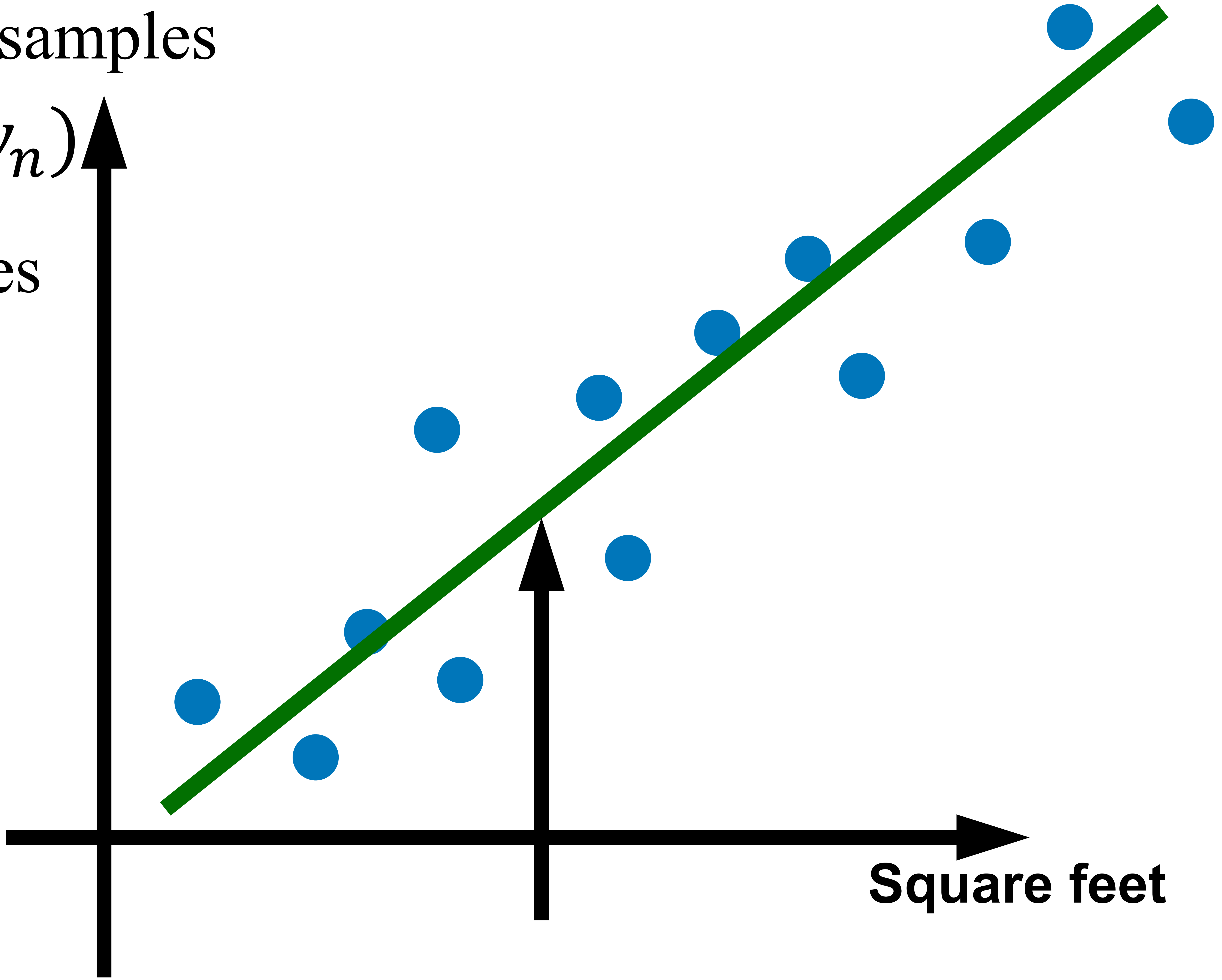
Example of regression: housing price prediction

Given: a dataset that contains n samples

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

Task: if a residence has **x** squares feet, predict the price?

$$y \in \mathbb{R}$$



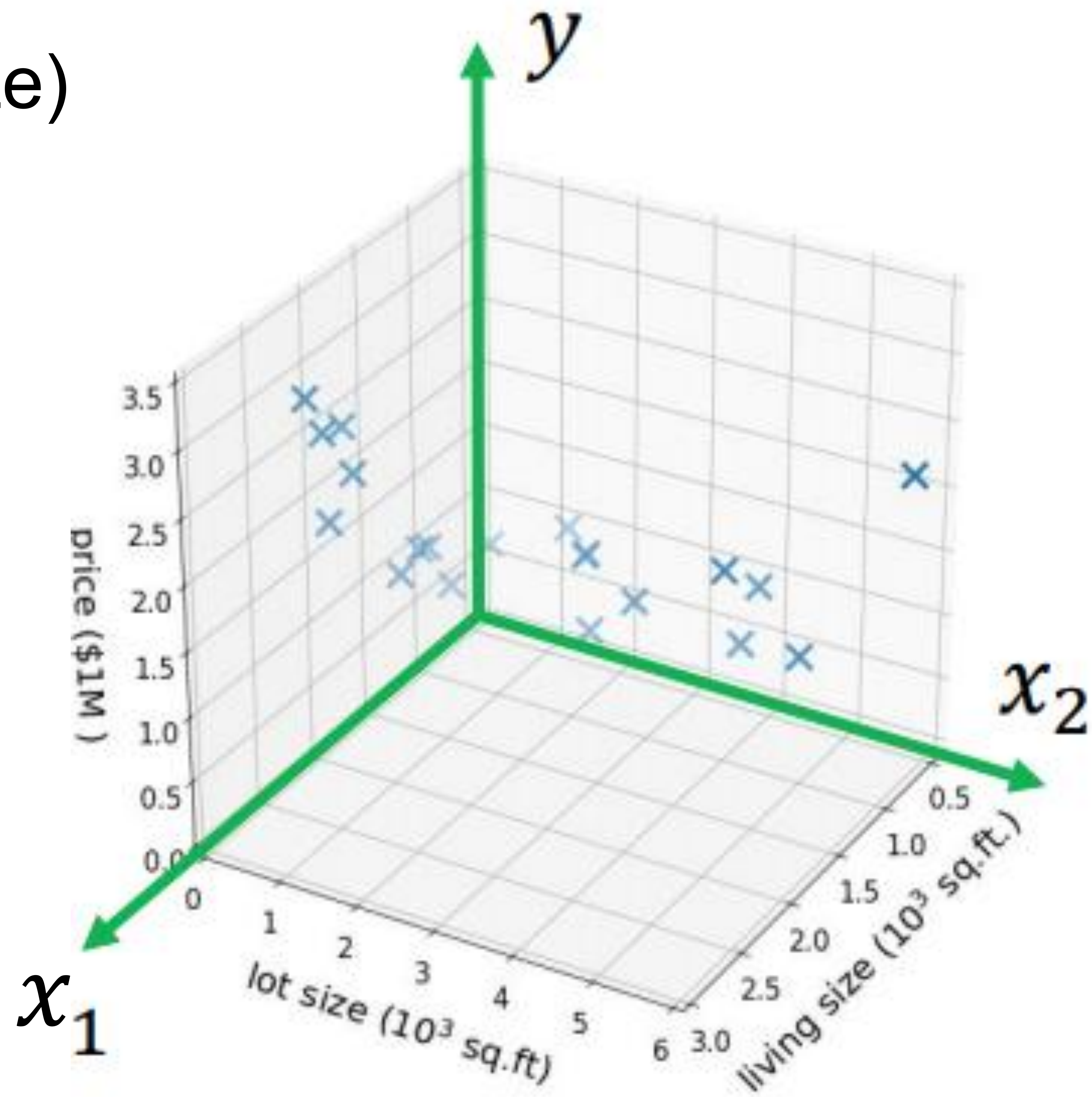
Example of regression: housing price prediction

Input with more features (e.g., lot size)

(size, lot size) → price

features/input $x \in \mathbb{R}^2$

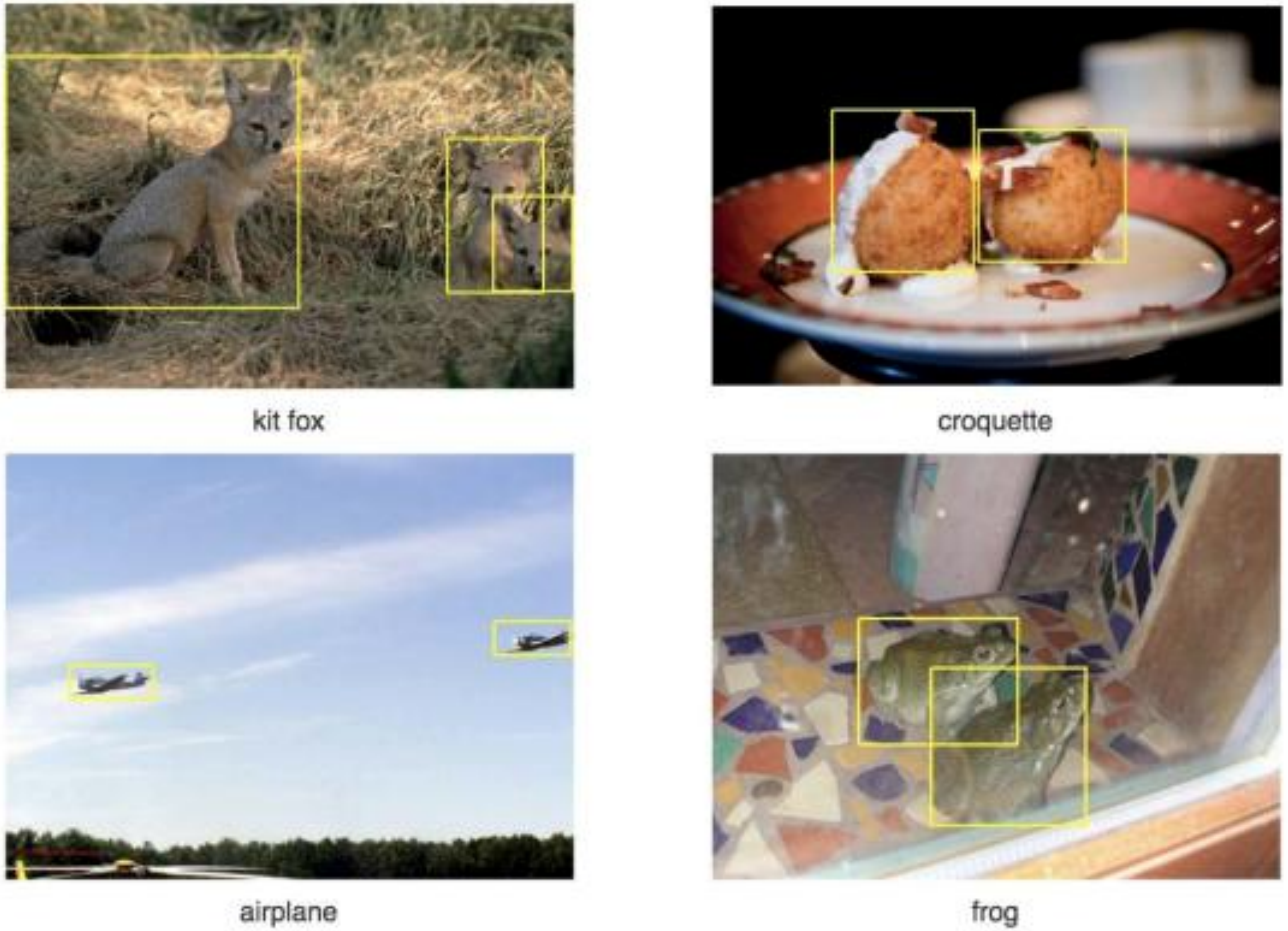
label/output $y \in \mathbb{R}$



(credit: stanford CS229)

Supervised Learning: More examples

x = raw pixels of the image y = bounding boxes



Russakovsky et al. 2015

Two Types of Supervised Learning Algorithms

Classification

- the label is a **discrete** variable

$$y \in \{1,2,3,\ldots,K\}$$

Regression

- the label is a **continuous** variable

$$y \in \mathbb{R}$$

Training Data for Supervised Learning

Training data is a collection of input instances to the learning algorithm:

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$$

input label

The training data is the "**experience**" given to a learning algorithm

Goal of Supervised Learning

Given training data

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$$

Learn a function mapping $f: X \to Y$, such that $f(x)$ predicts the label y on **future** data x (not in training data)

Goal of Supervised Learning

Training set error

- 0-1 loss for classification $\ell = \frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i) \neq y_i)$
- Squared loss for regression: $\ell = \frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i) - y_i)^2$

A learning algorithm optimizes the training objective

$$f^* = \operatorname{argmin} \mathbb{E}_{(x,y)}\ell(f(x), y)$$

Details in upcoming lectures :)

Quiz Break

Q1-1: Which is true about feature vectors?

A. Feature vectors can have at most 10 dimensions
B. Feature vectors have only numeric values
C. The raw image can also be used as the feature vector
D. Text data don’t have feature vectors

Quiz Break

Q1-2: Which of the following is not a common task of supervised learning?

A. Object detection (predicting bounding box from raw images)
B. Classification
C. Regression
D. Dimensionality reduction

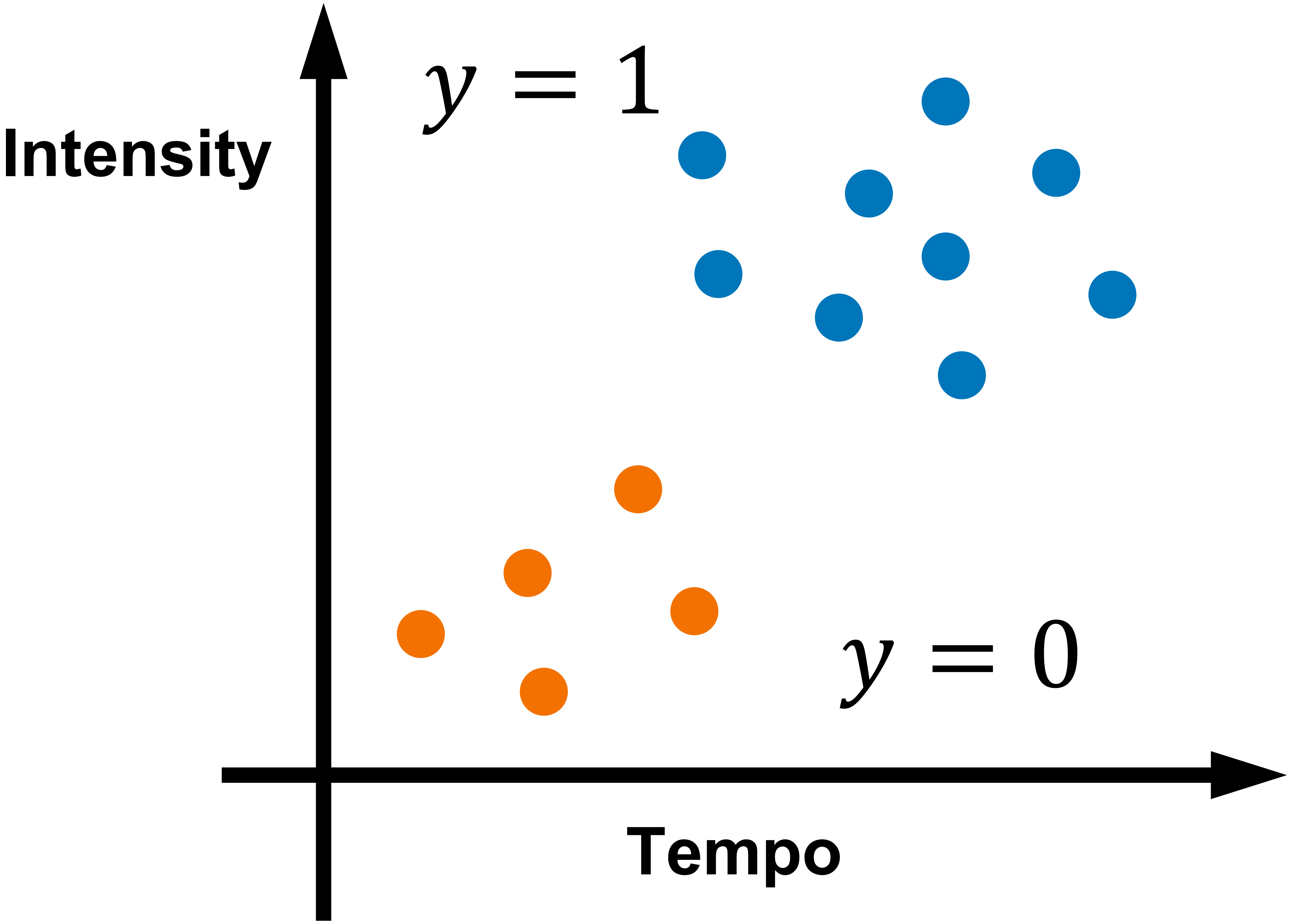
Part II: Unsupervised Learning
(no teacher)

Unsupervised Learning

- Given: dataset contains **no label** $x_1, x_2, \ldots, x_n$
- **Goal:** discover interesting patterns and structures in the data

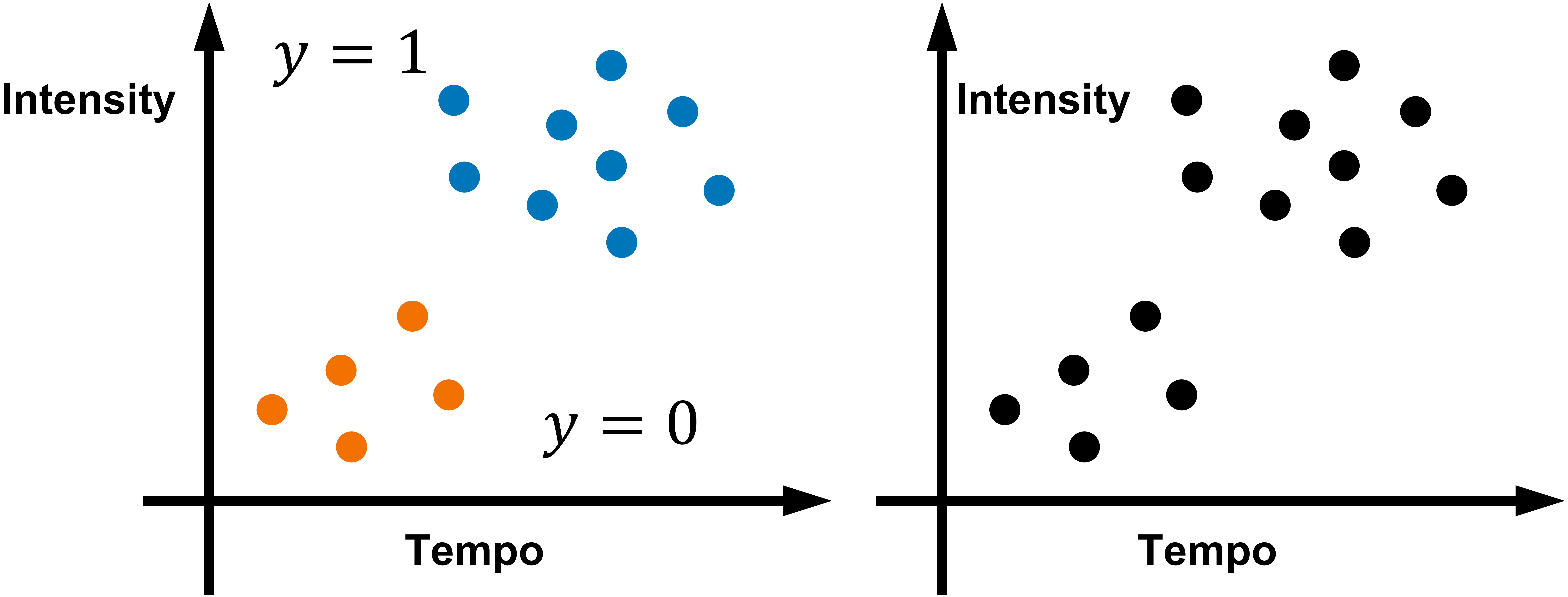
Unsupervised Learning

- Given: dataset contains **no label** $x_1, x_2, \ldots, x_n$
- **Goal:** discover interesting patterns and structures in the data



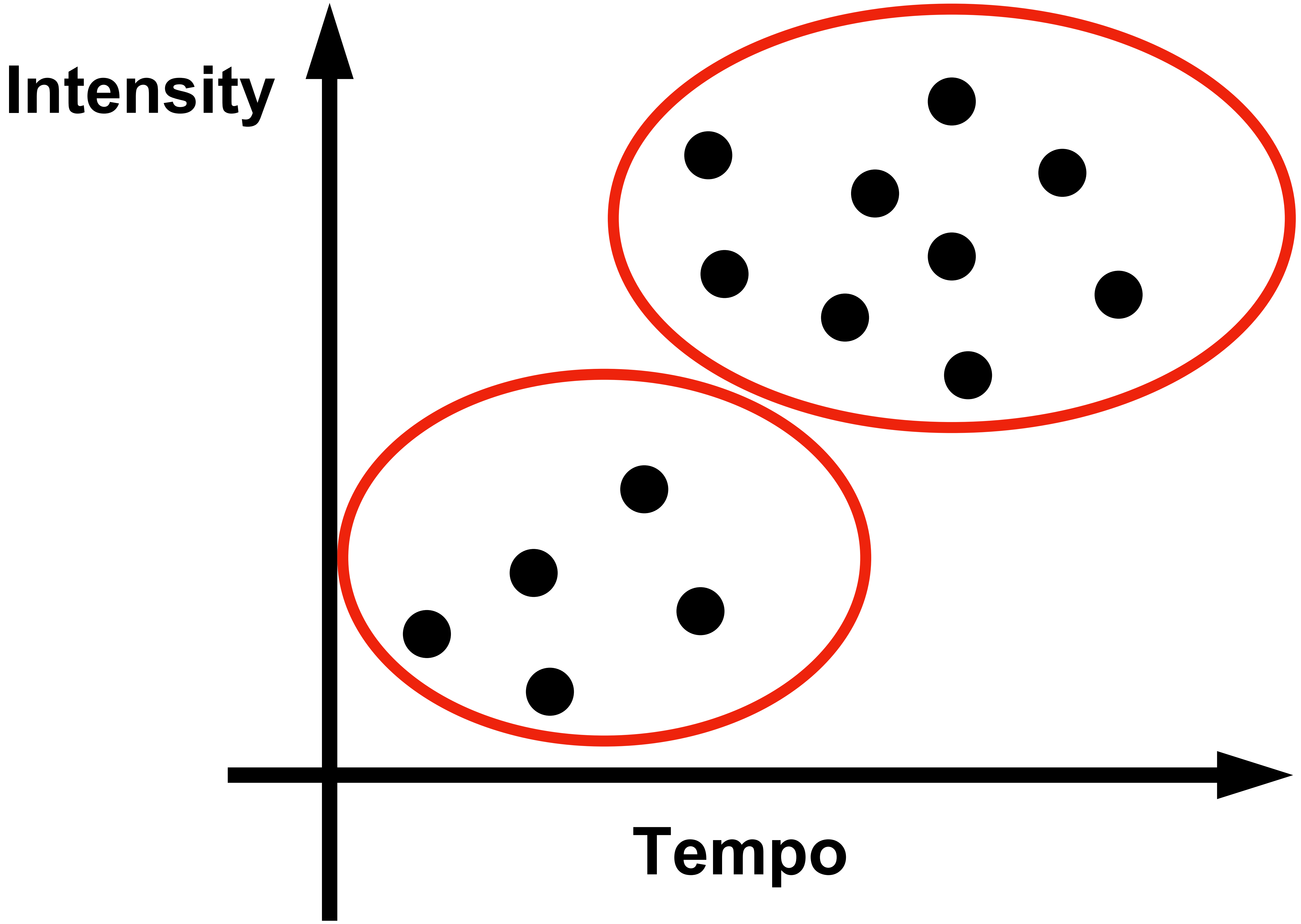
Unsupervised Learning

- Given: dataset contains **no label** $x_1, x_2, \ldots, x_n$
- **Goal:** discover interesting patterns and structures in the data

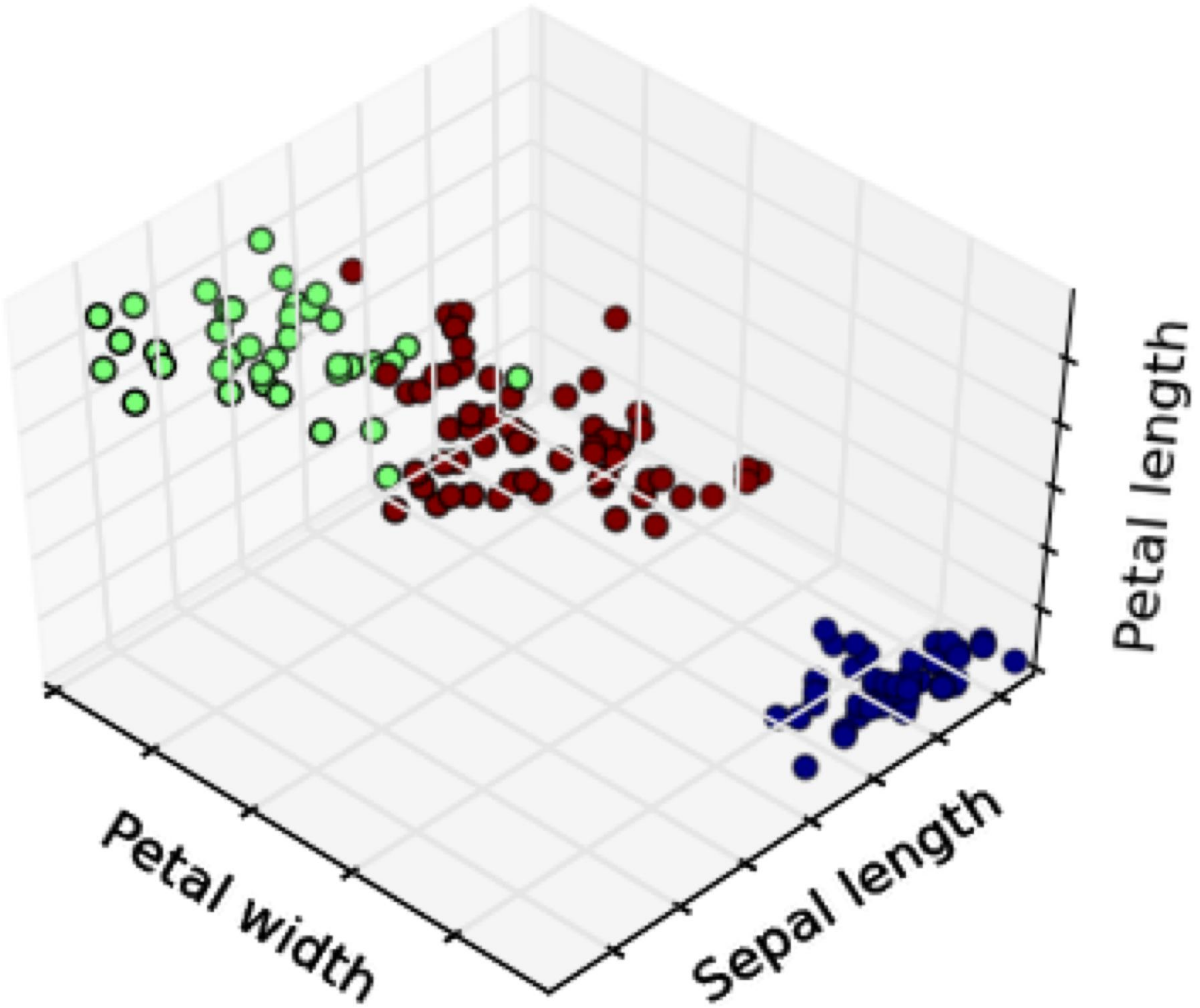


Clustering

- Given: dataset contains **no label** $x_1, x_2, ..., x_n$
- **Output:** divides the data into clusters such that there are intra-cluster similarity and inter-cluster dissimilarity



Clustering

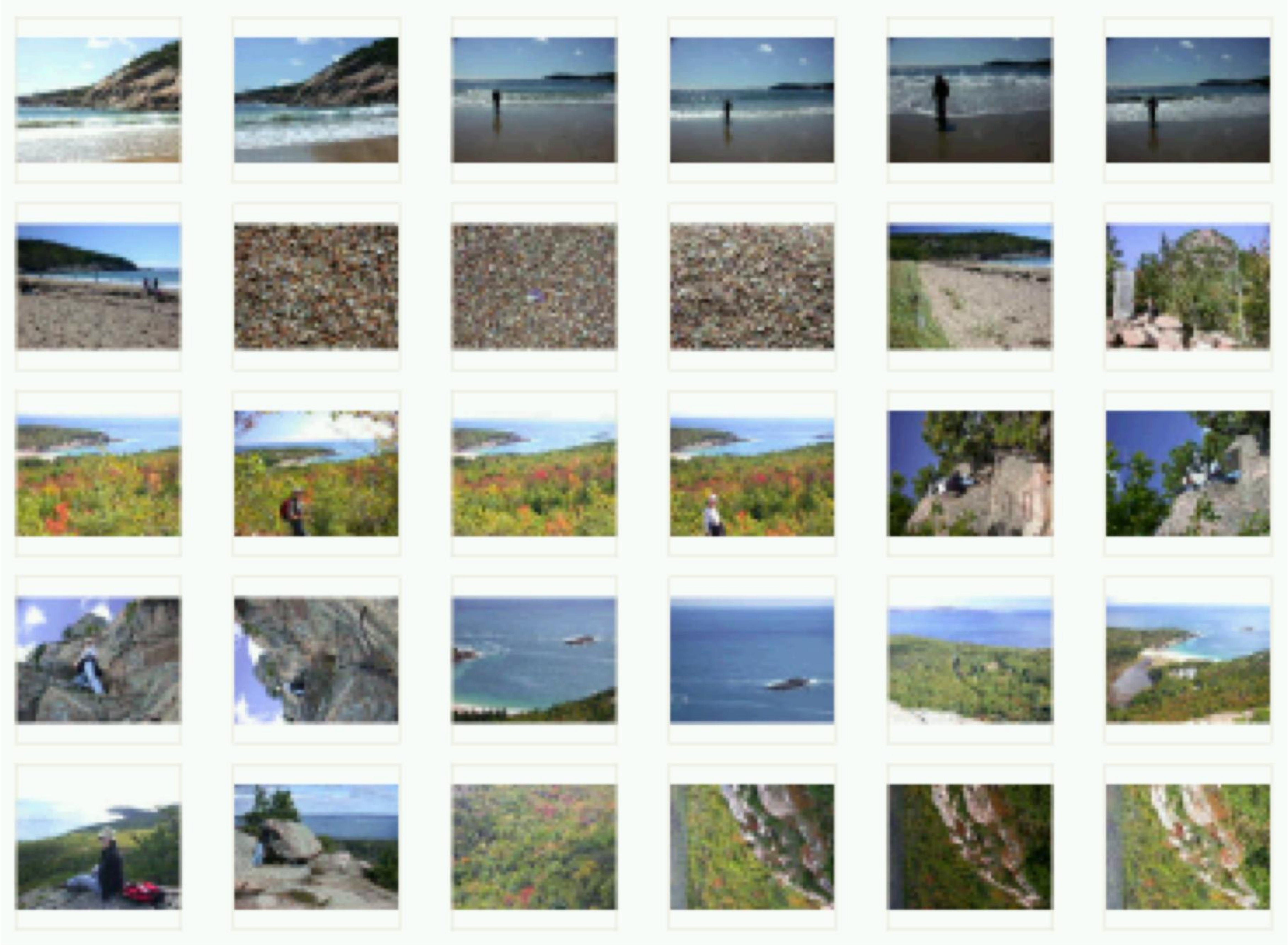


Clustering Irises using three different features

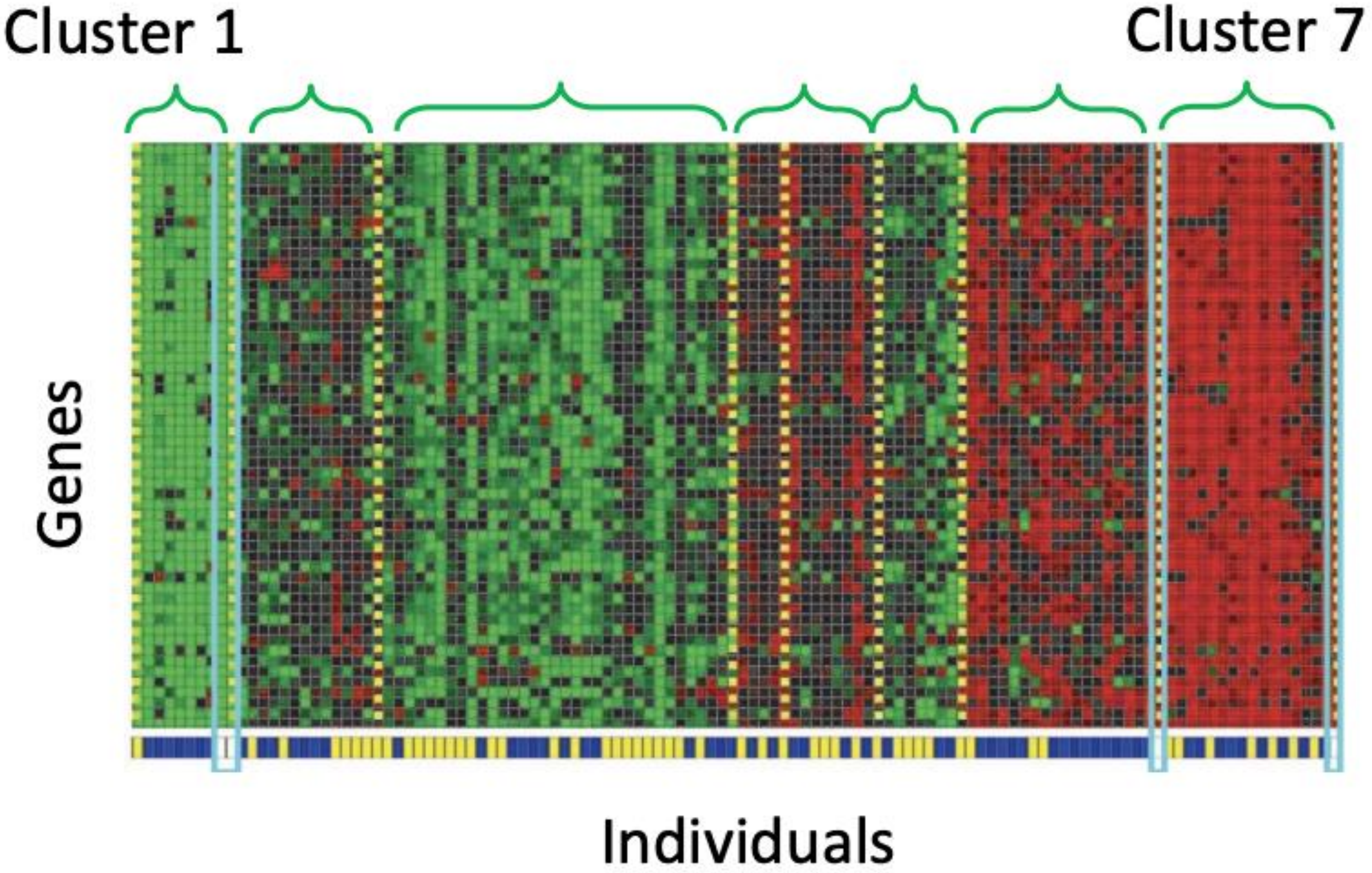
The colors represent clusters identified by the algorithm, **not** y’s provided as input

Clustering

- You probably have >1000 digital photos stored on your phone
- After this class you will be able to organize them better (based on visual similarity)

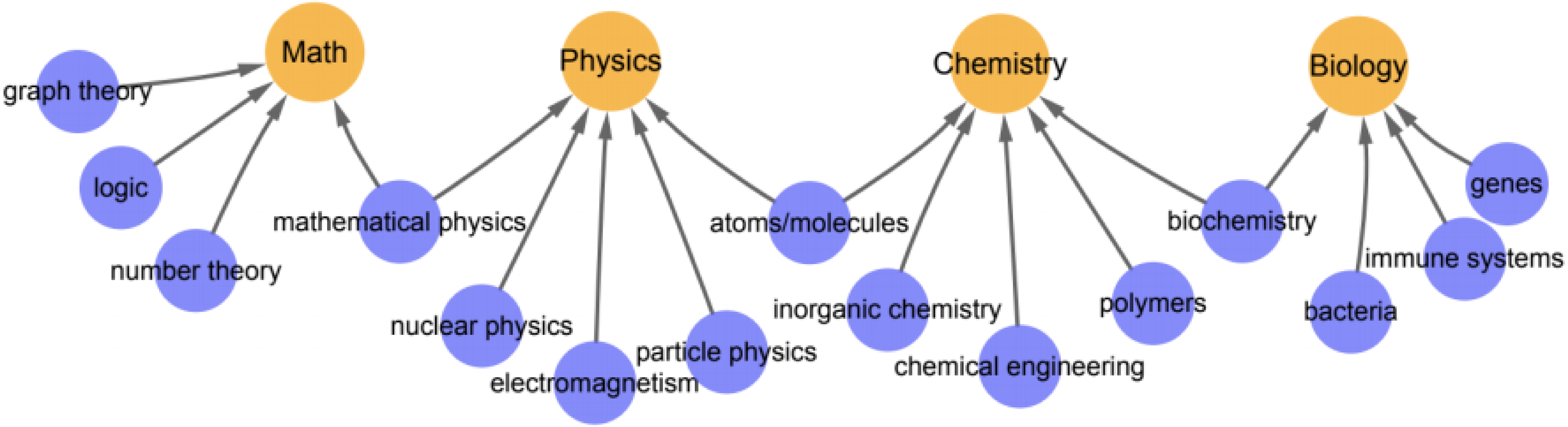


Clustering Genes



Identifying Regulatory Mechanisms using Individual Variation Reveals Key Role for Chromatin Modification. [Su-In Lee, Dana Pe'er, Aimee M. Dudley, George M. Church and Daphne Koller. '06]

Clustering Words with Similar Meanings



[Arora-Li-Liang-Ma-Risteski, TACL’17,18]

How do we perform clustering?

Many clustering algorithms.

We will look at the two most frequently used ones:

- *K-means clustering*: we specify the desired number of clusters, and use an iterative algorithm to find them
- *Hierarchical clustering*: we build a binary tree over the dataset

Quiz Break

Q2-1: Which is true about machine learning?

A. The process doesn’t involve human inputs
B. The machine is given the training and test data for learning
C. In clustering, the training data also have labels for learning
D. Supervised learning involves labeled data

Quiz Break

Q2-2: Which is true about unsupervised learning?

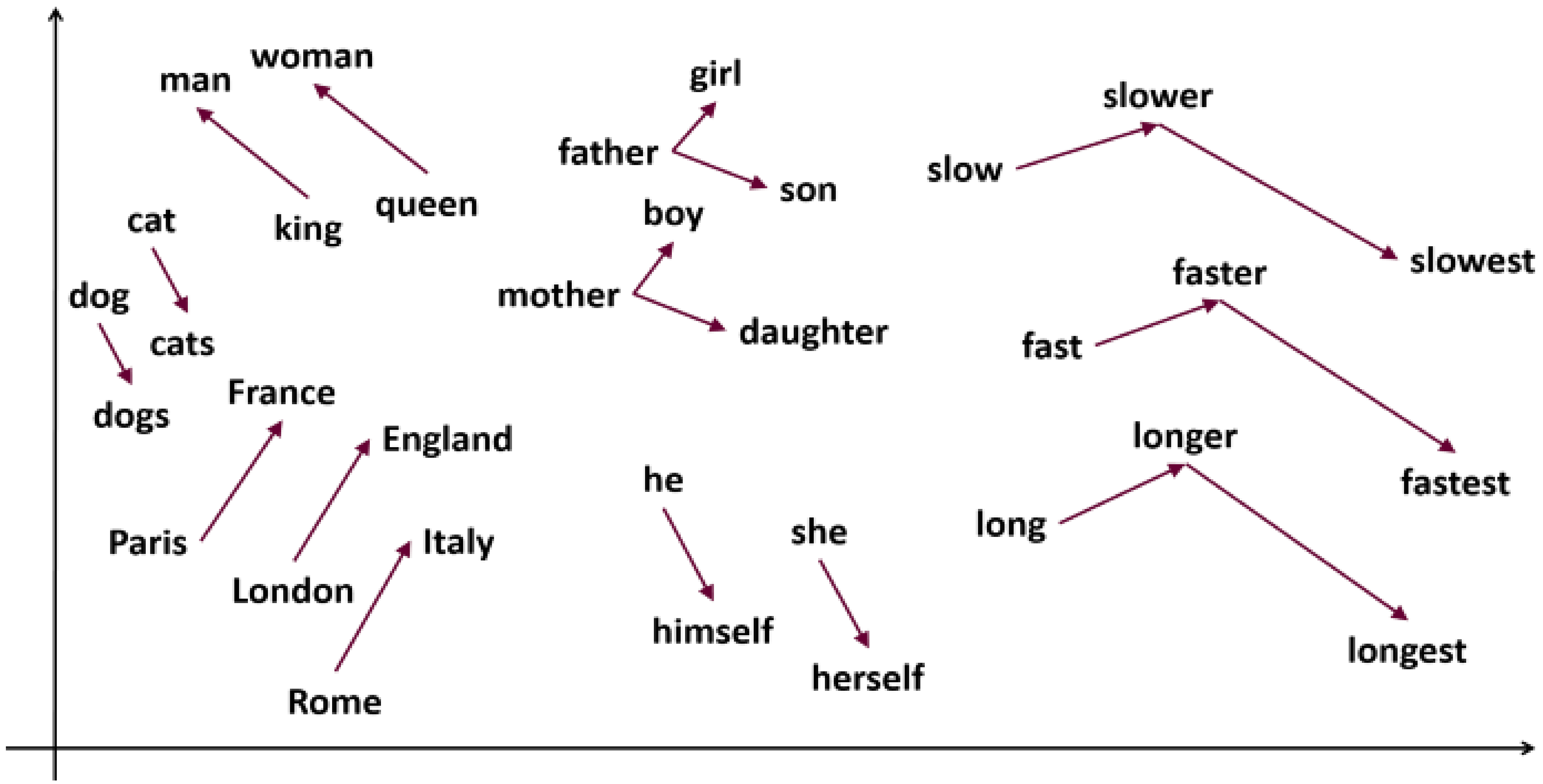
A. There are only 2 unsupervised learning algorithms
B. Kmeans clustering is a type of hierarchical clustering
C. Kmeans algorithm automatically determines the number of clusters k
D. Unsupervised learning is widely used in many applications

Self-Supervised Learning

- Given: dataset contains **no label** $x_1, x_2, \ldots, x_n$
- **Goal:** discover interesting patterns and structures in the data

- **Approach:** generate supervision signal from data.
  Solve a *pretext task*

Example: word embeddings



Saurabh Pal – Implementing Word2Vec in Tensorflow

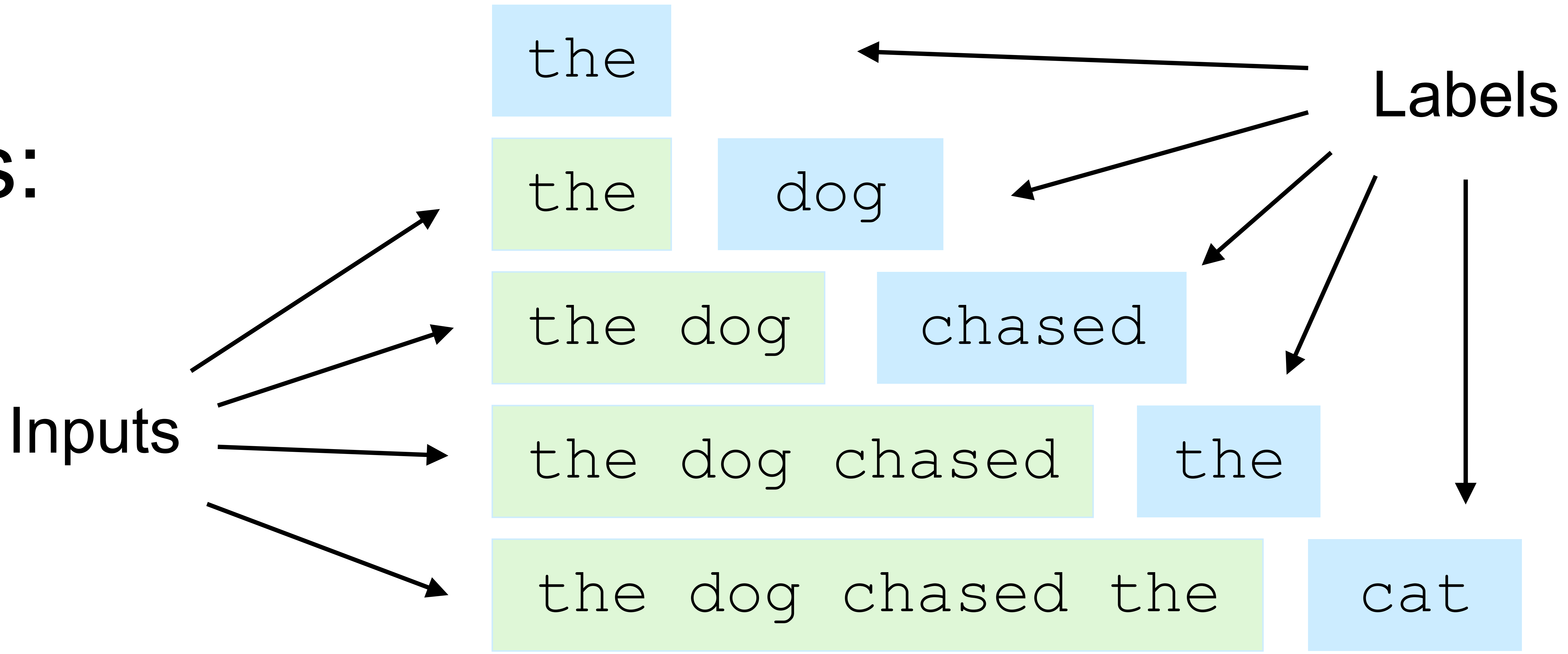
Self-Supervised Learning for LLMs

- Pretext task for large language models:

  next-word prediction

- Original text: `the dog chased the cat`

- Split into five labeled problems:

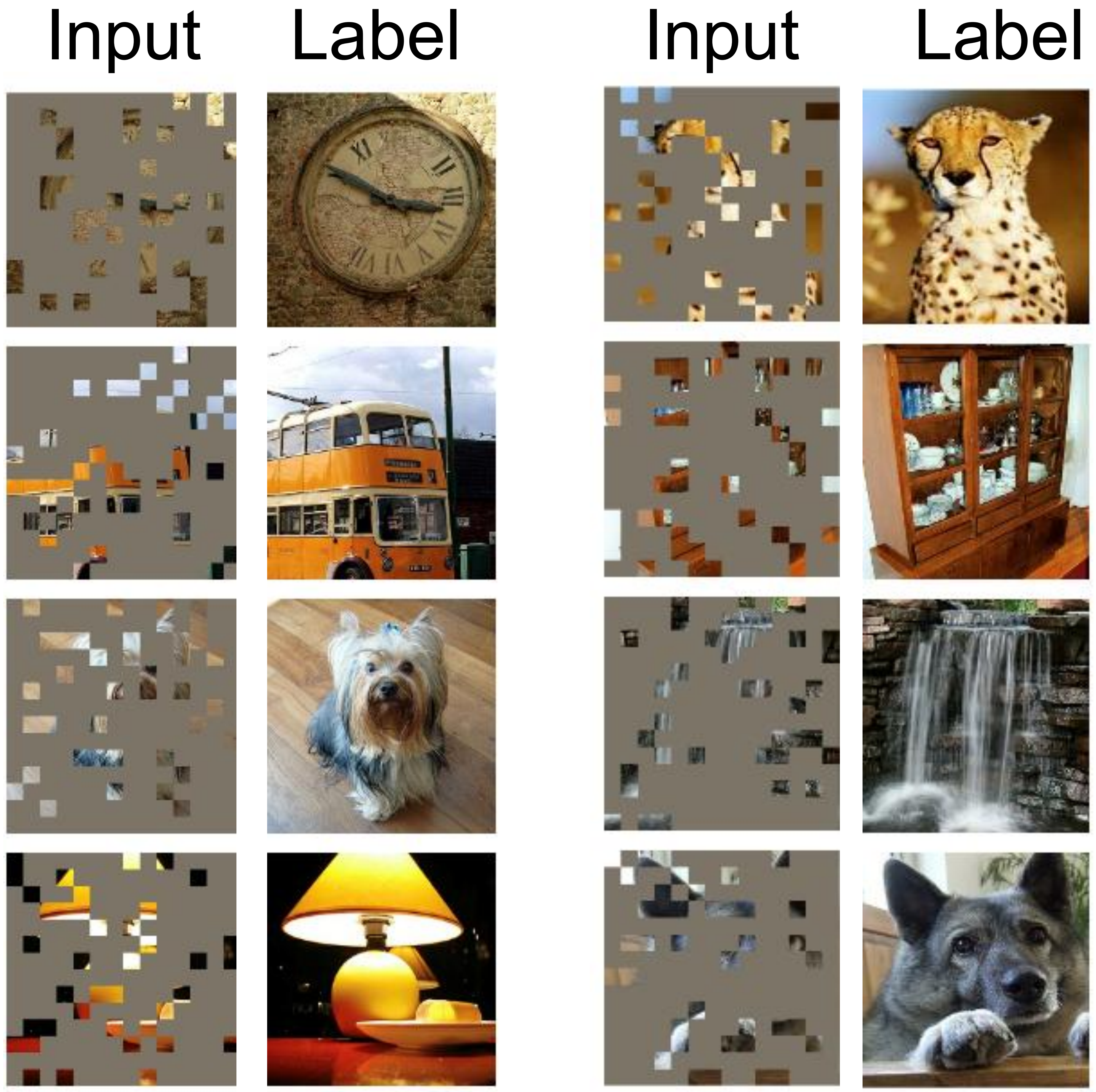


Self-Supervised Learning in Vision

- Another common pretext task: **image inpainting**

- High-dimensional label!

- Type of **autoencoder**
  - “Auto-” = “self”



He et al. Masked autoencoders are scalable vision learners. 2021.

Part III: Reinforcement Learning
(Learning from rewards)

Reinforcement Learning

- Given: an agent that can take actions and a reward function specifying how good an action is.
- **Goal:** learn to choose actions that maximize future reward total.

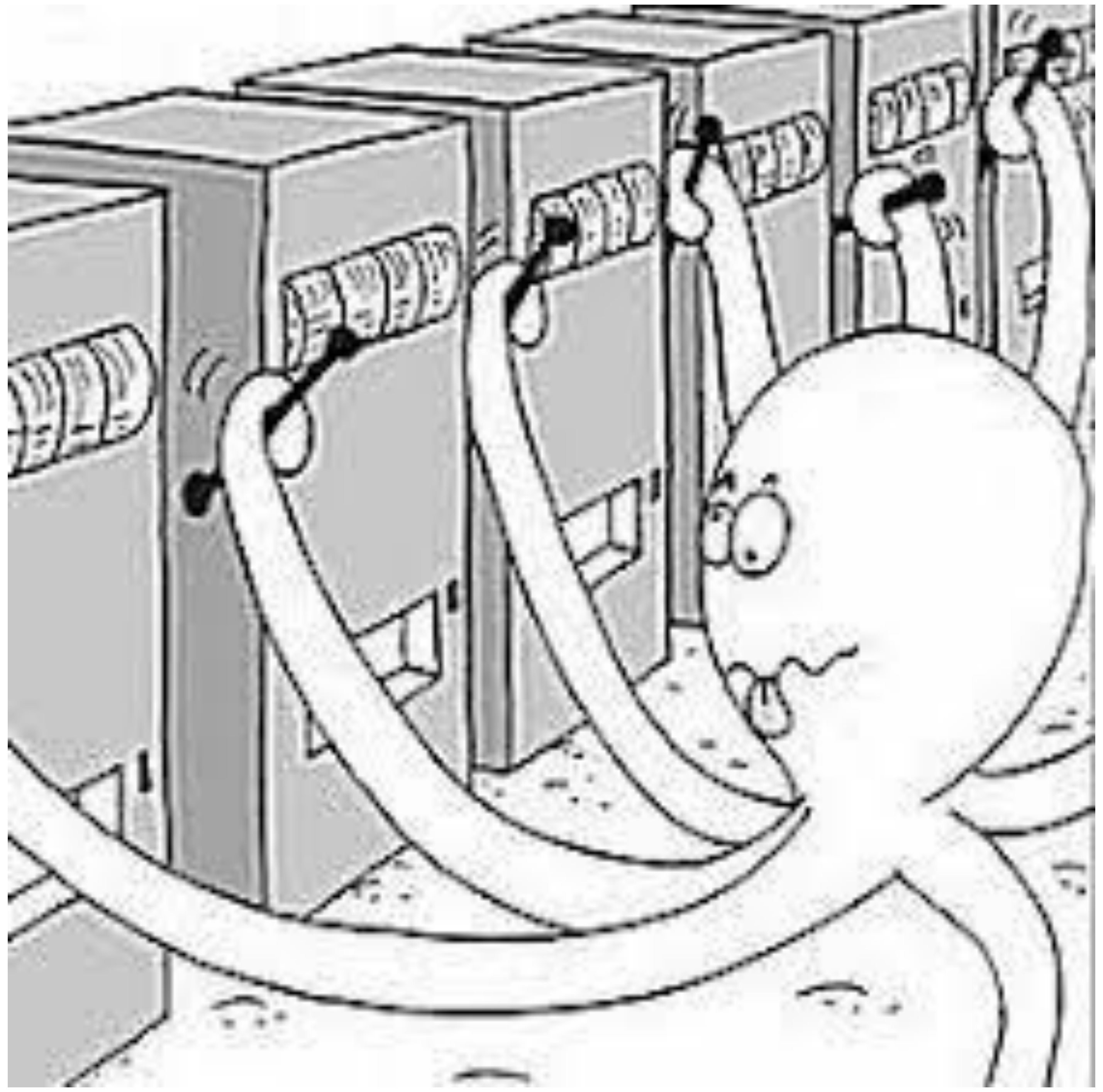
Google Deepmind

Reinforcement Learning Key Problems

1. Problem: actions may have delayed effects.
   - Requires **credit-assignment**
2. Problem: maximal reward action is unknown
   - Exploration-exploitation trade-off

"..the problem [exploration-exploitation] was proposed [by British scientist] to be dropped over Germany so that German scientists could also waste their time on it."

- Peter Whittle



Multi-armed Bandit

Today’s recap

- NLP Review
- What is machine learning?
- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
  - Clustering
  - Self-Supervised Learning
- Reinforcement Learning

Suggested Readings

- Textbook: Artificial Intelligence: A Modern Approach (4th edition). Stuart Russell and Peter Norvig. Pearson, 2020. Sections 19.1

Thanks!