

Announcements

- **HW2** due on **Friday Sep 25th at 11:59 PM**
- 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**

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

- Recall the chain rule of probability:

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

- Markov assumption with shorter history:

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

Review: n-gram Model

- **k=0 Unigram Model:** Full independence assumption:

- (Present doesn't depend on the past)

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

- **k = 1 Bigram Model: Markov Assumption:**

- (Present depends on immediate past)

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

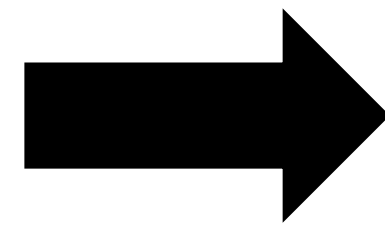
- Can do trigrams, 4-grams, and so on
 - More expressive as n goes up
 - Harder to estimate

Review: Bag of Words (BOW)

Document

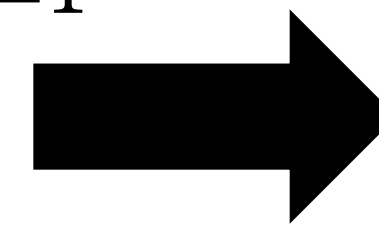
Madison is the capital city of Wisconsin. The city is known for its beautiful lakes and the university.

$count(w_i)$



1	and
0	are
1	beautiful
0	between
1	capital
2	city
1	for
0	in
2	is
1	Its
1	known
1	lakes
1	Madison
1	of
0	popular
0	sits
0	summer
3	the
0	two
1	university
1	Wisconsin

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



normalized

0.056	and
0	are
0.056	beautiful
0	between
0.056	capital
0.111	city
0.056	for
0	in
0.111	is
0.056	Its
0.056	known
0.056	lakes
0.056	Madison
0.056	of
0	popular
0	sits
0	summer
0.167	the
0	two
0.056	university
0.056	Wisconsin

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

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

← Total number of documents in the corpus

↑
number of documents containing the term w_i

- TF-IDF_{ij}: How important is the term w_i for the document j

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

Review: Representing Words

Traditional representation: **one-hot vectors**

$$\text{dog} = [0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0]$$

Dense vectors:

$$\text{dog} = [0.13 \quad 0.87 \quad -0.23 \quad 0.46 \quad 0.87 \quad -0.31]^T$$

$$\text{cat} = [0.07 \quad 1.03 \quad -0.43 \quad -0.21 \quad 1.11 \quad -0.34]^T$$

AKA word embeddings

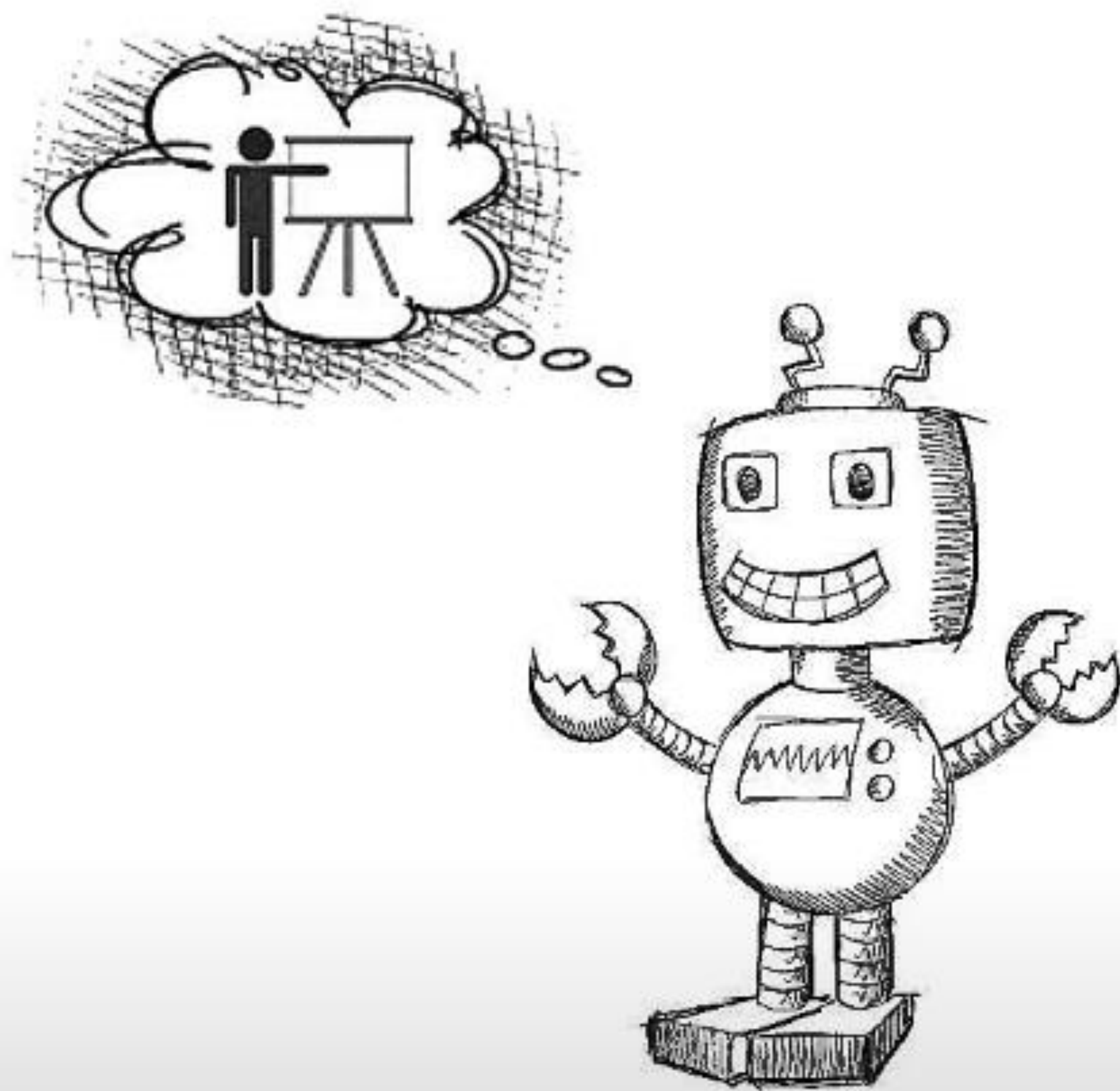




Part I: What is machine learning?



**HUMANS LEARN FROM
PAST EXPERIENCES**



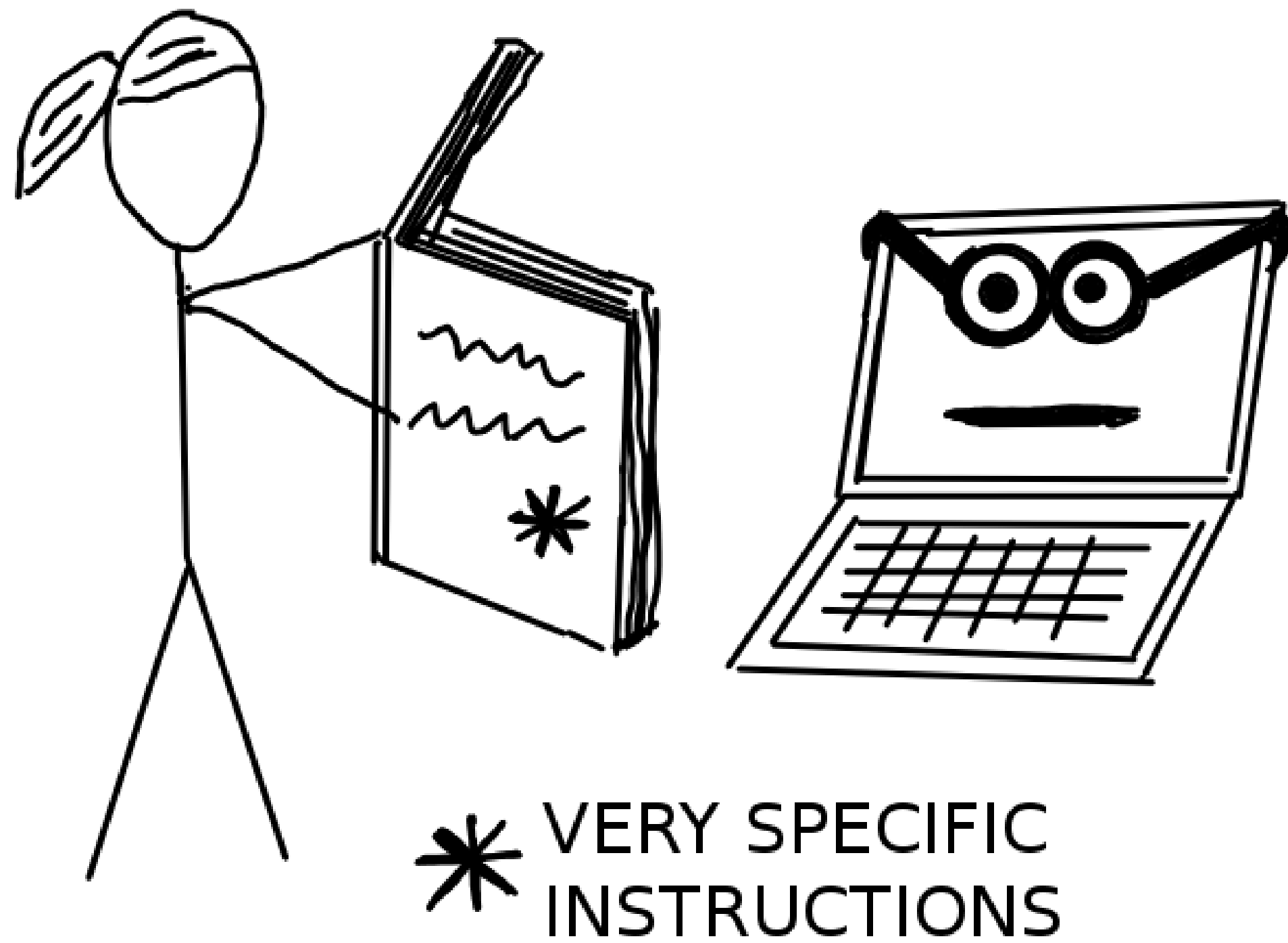
**MACHINES FOLLOW INSTRUCTIONS
GIVEN BY HUMANS**

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

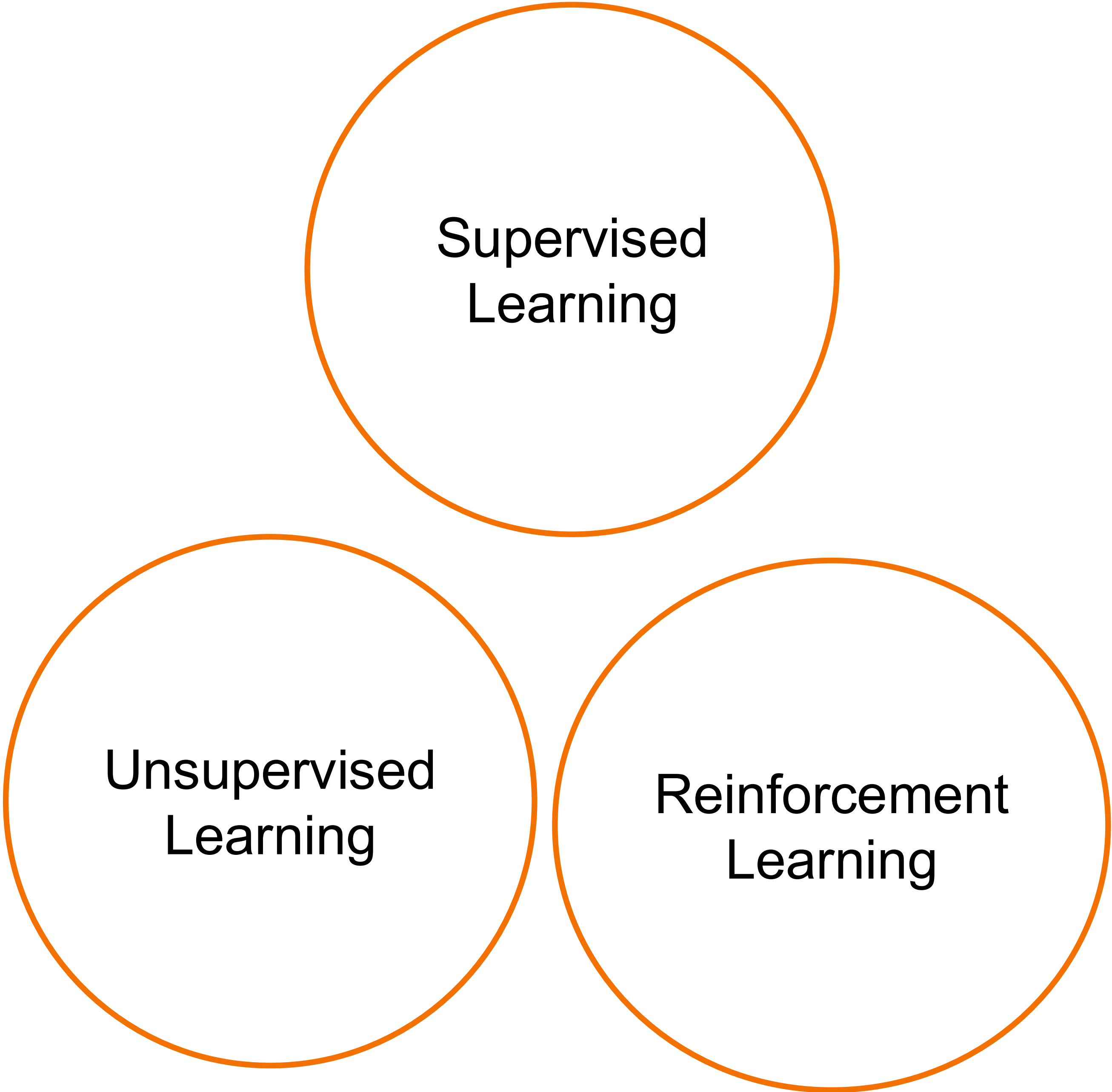


What is **machine learning**?

- 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.



An incomplete Taxonomy of ML



Supervised
Learning

The diagram consists of three orange-outlined circles arranged in a triangular pattern. The top circle is labeled 'Supervised Learning'. The bottom-left circle is labeled 'Unsupervised Learning'. The bottom-right circle is labeled 'Reinforcement Learning'. The circles are positioned such that they appear to be part of a larger, incomplete set, as indicated by the title 'An incomplete Taxonomy of ML'.

Unsupervised
Learning

Reinforcement
Learning



Part II: Supervised Learning

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



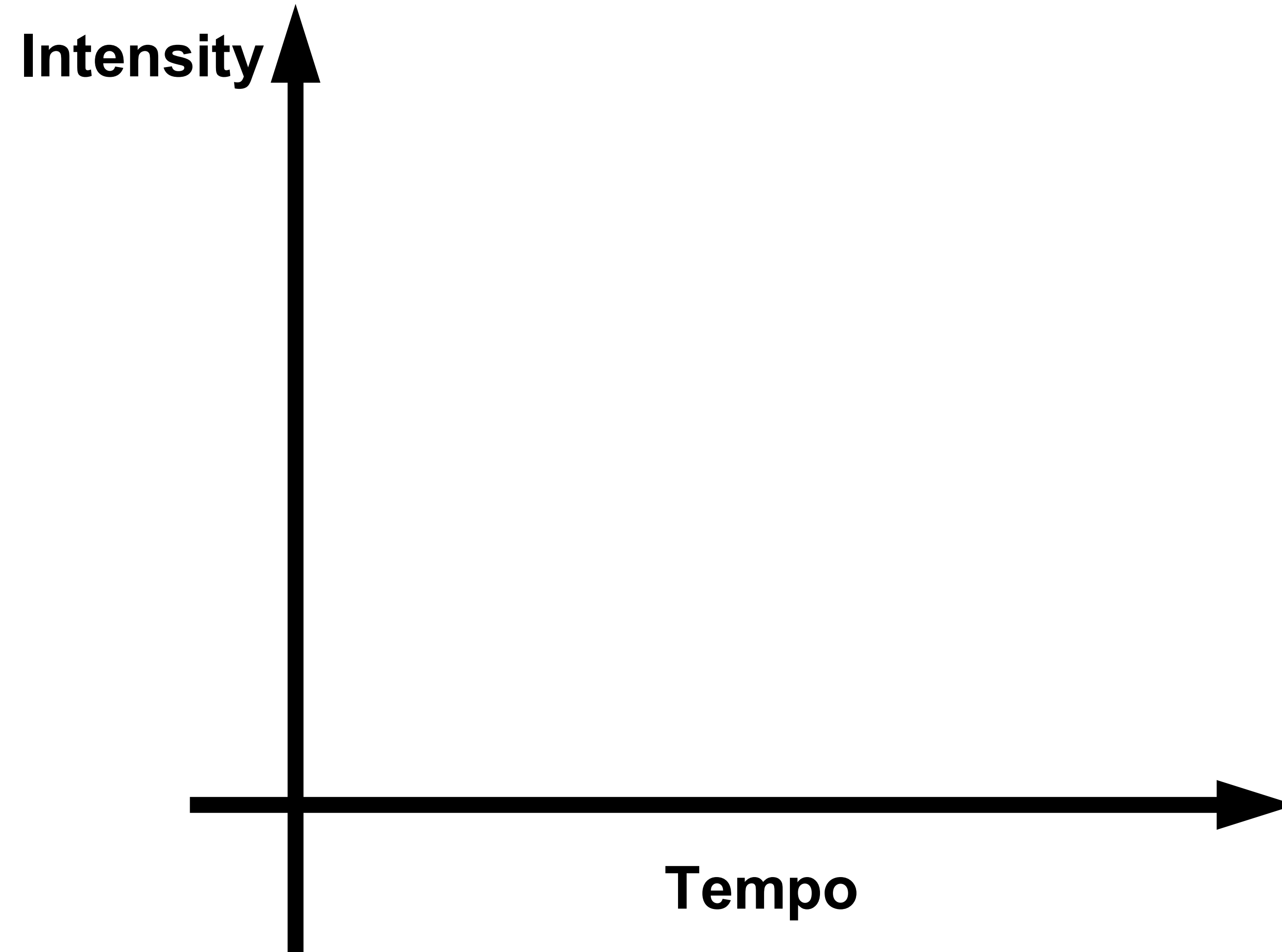
model



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



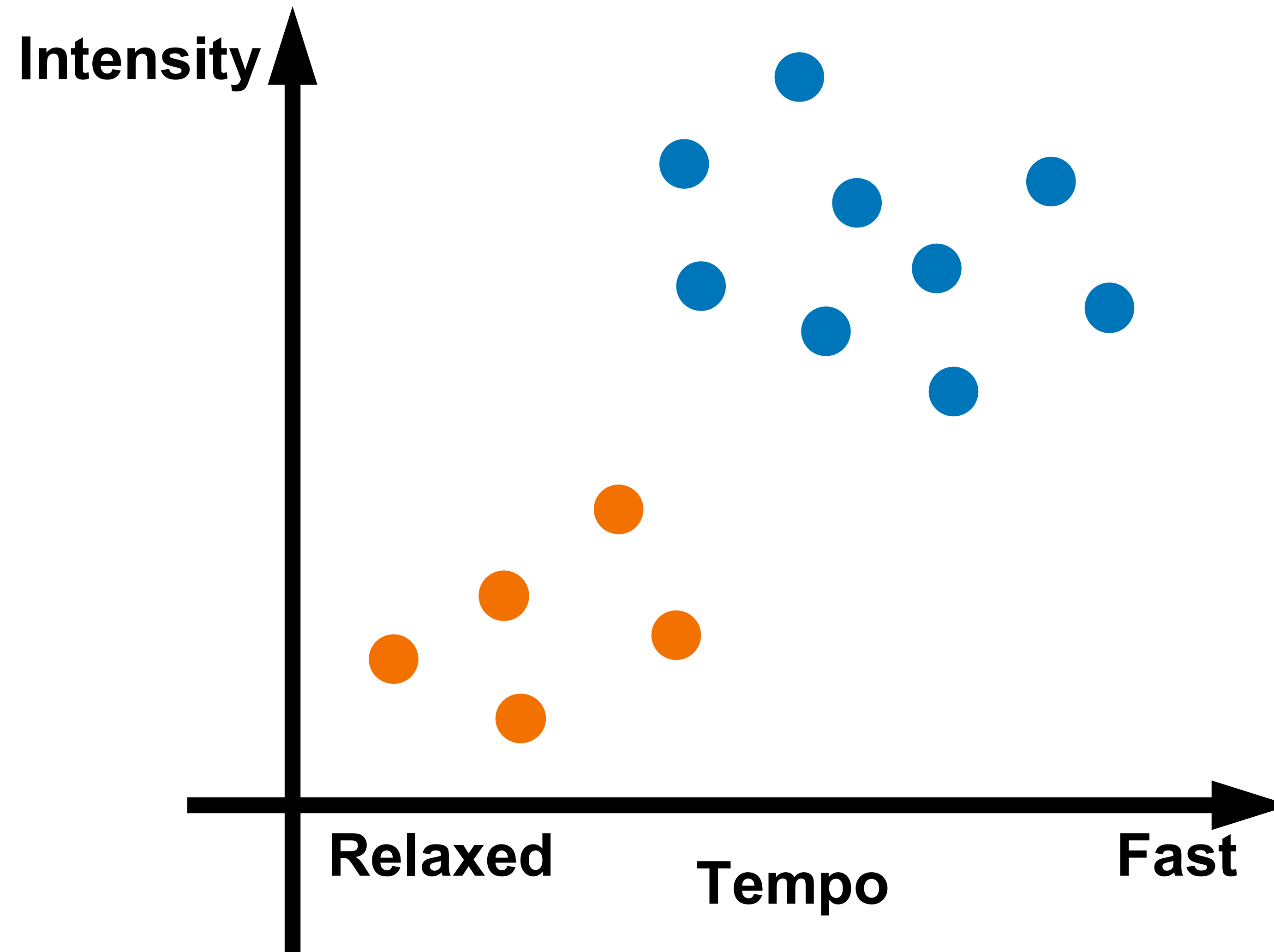
User Sharon



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



User Sharon



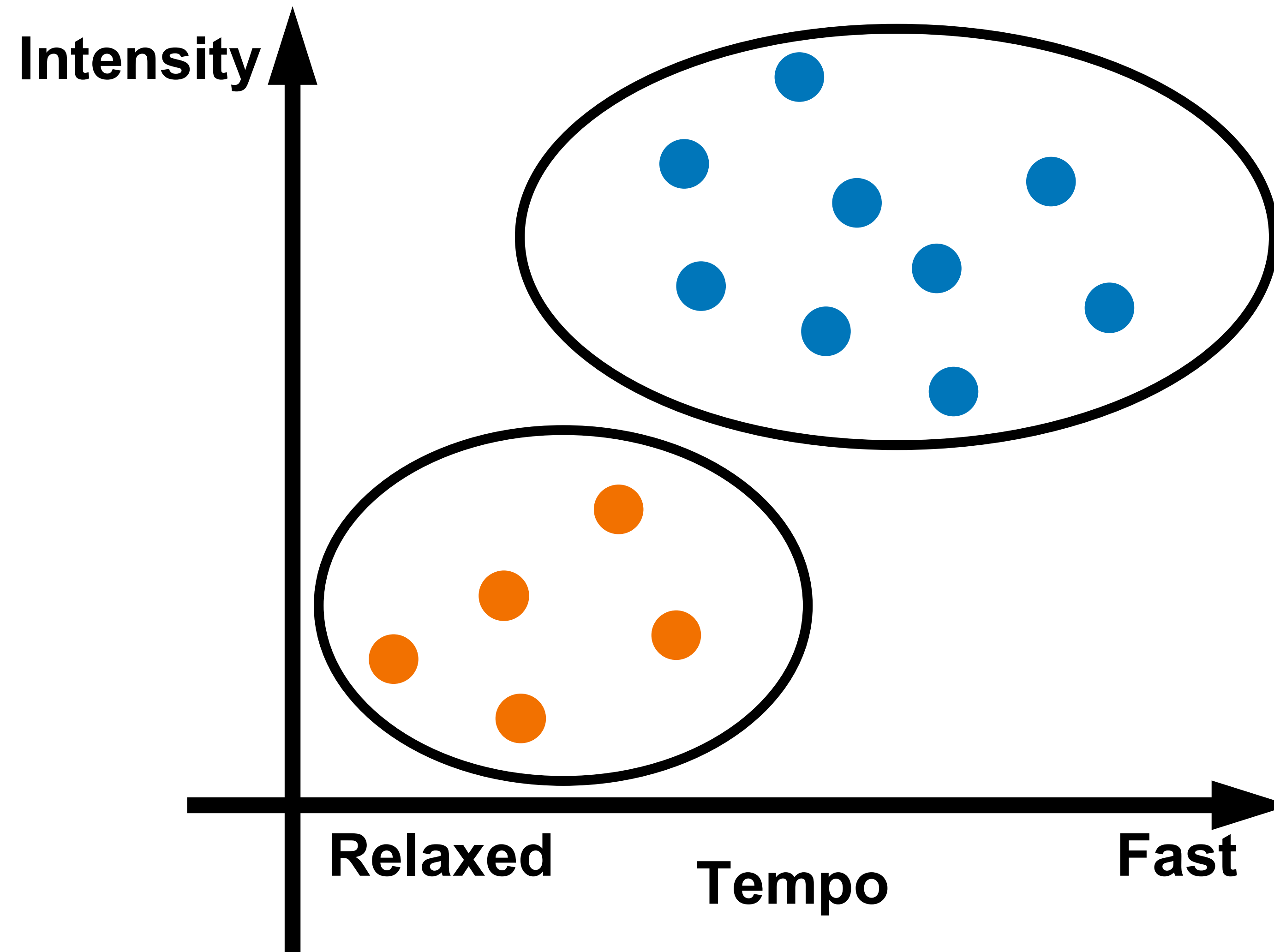
Example 1: Predict whether a user likes a song or not



User Sharon

 Dislike

 Like



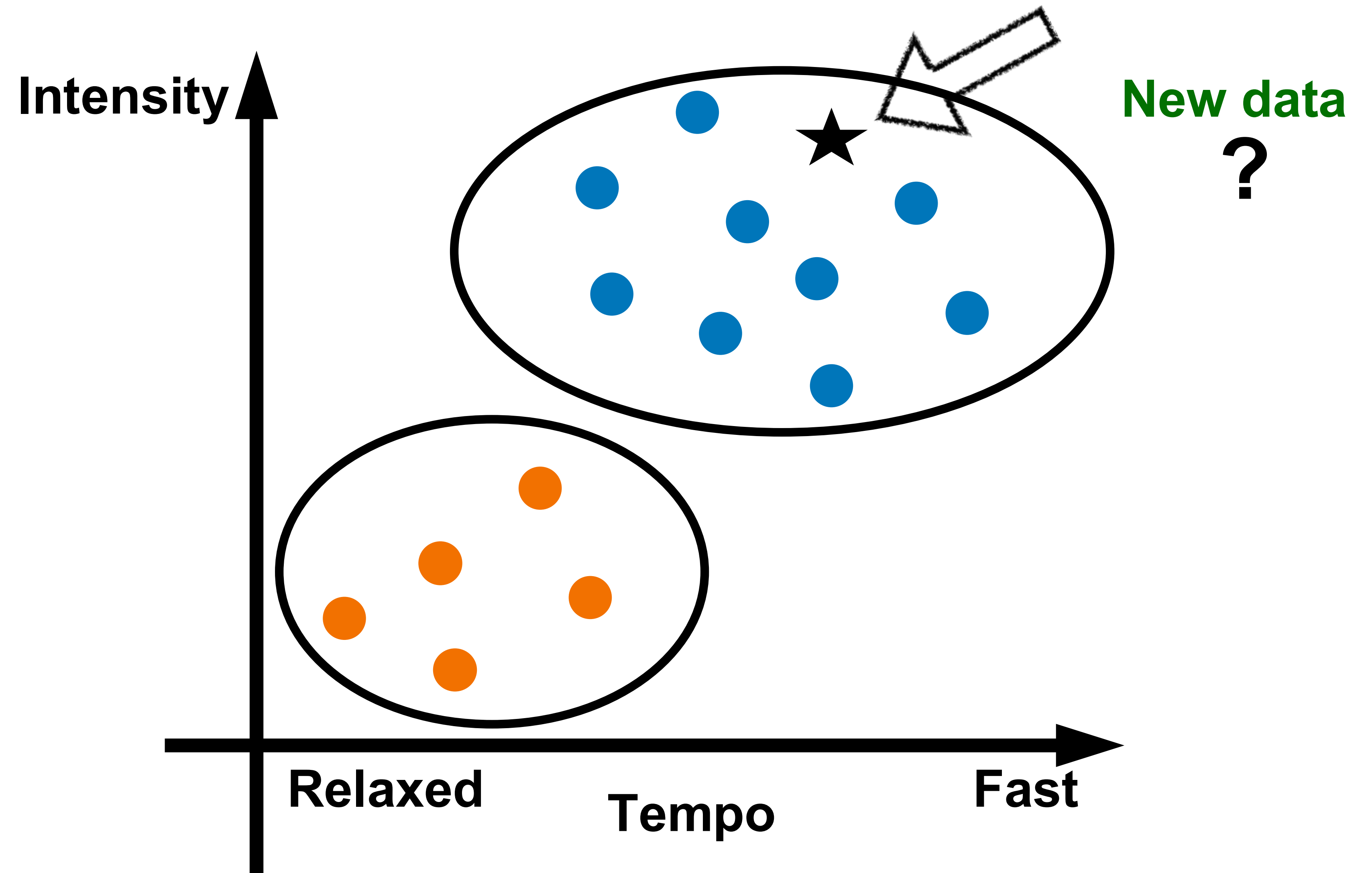
Example 1: Predict whether a user likes a song or not



User Sharon

● Dislike

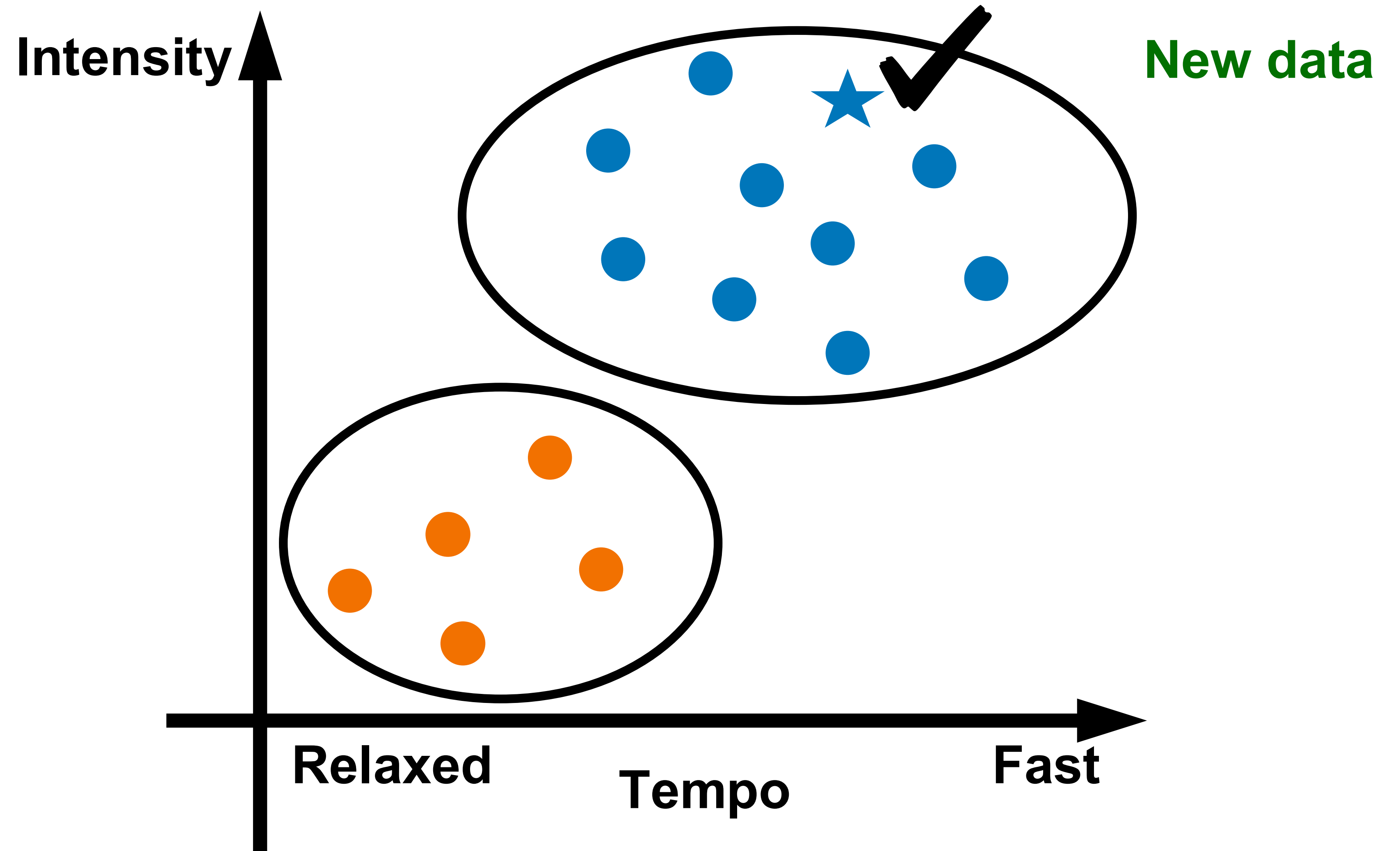
● Like



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



User Sharon

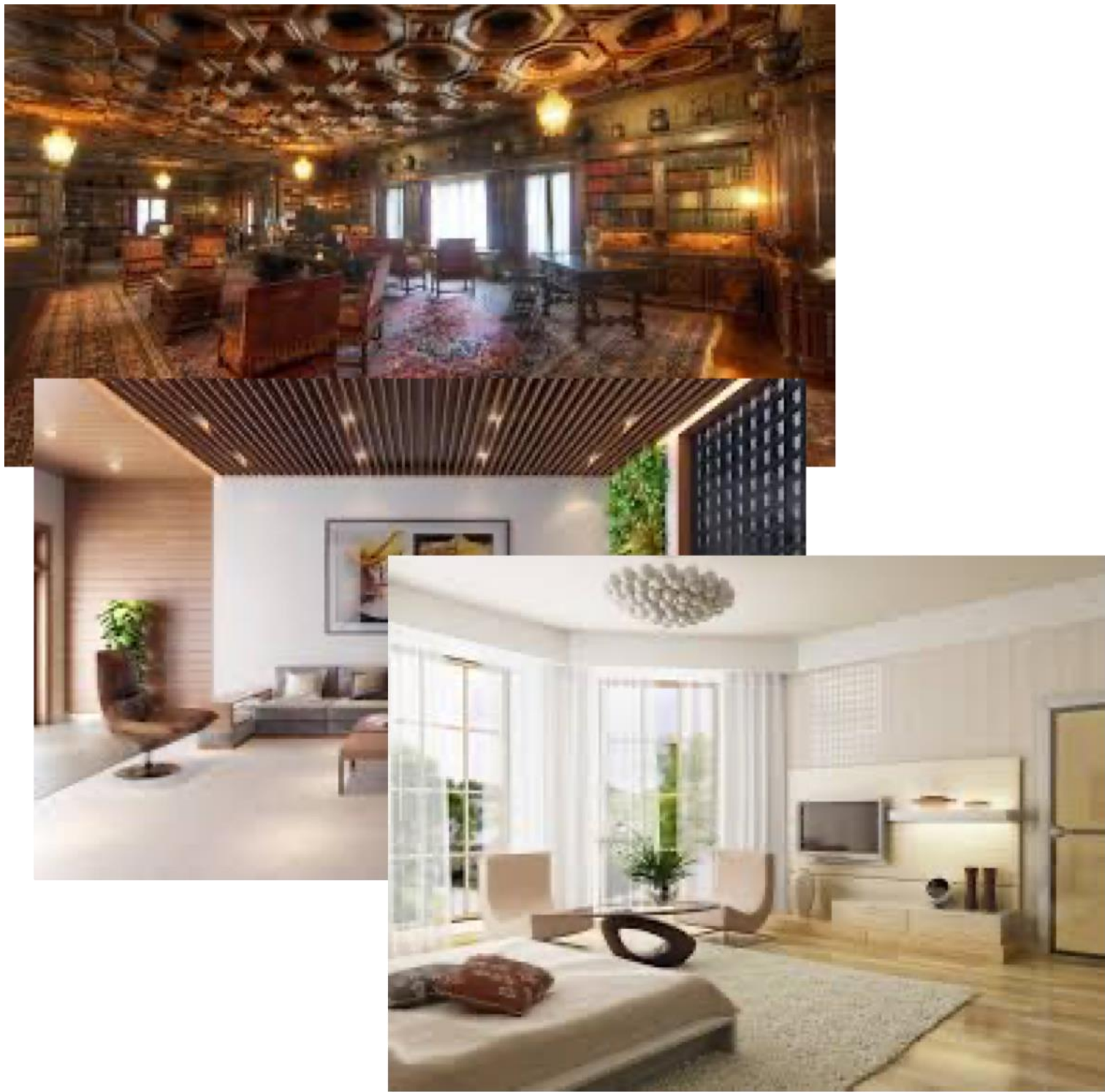


Example 2: Classify Images

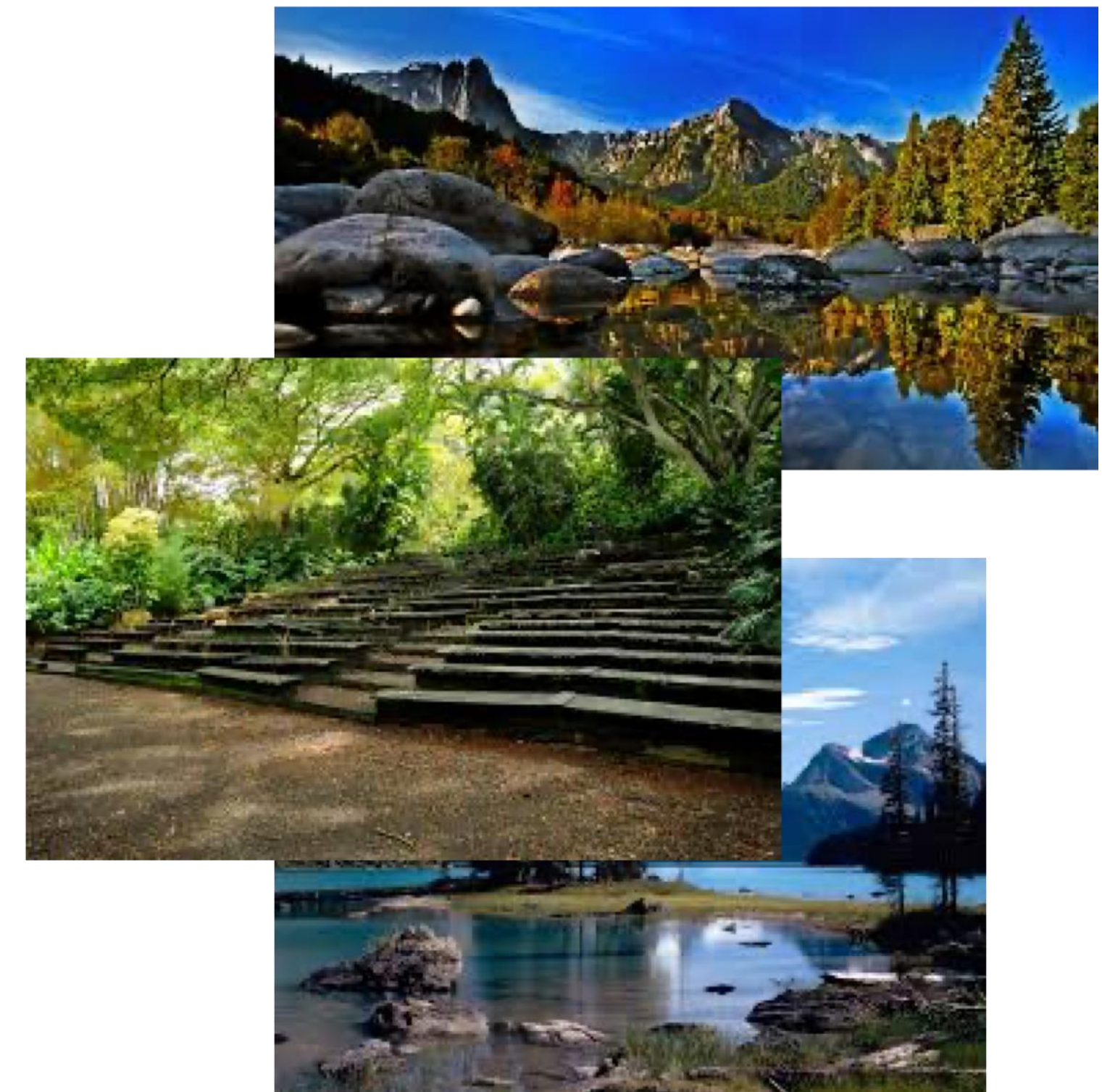
<http://www.image-net.org/>



Example 2: Classify Images

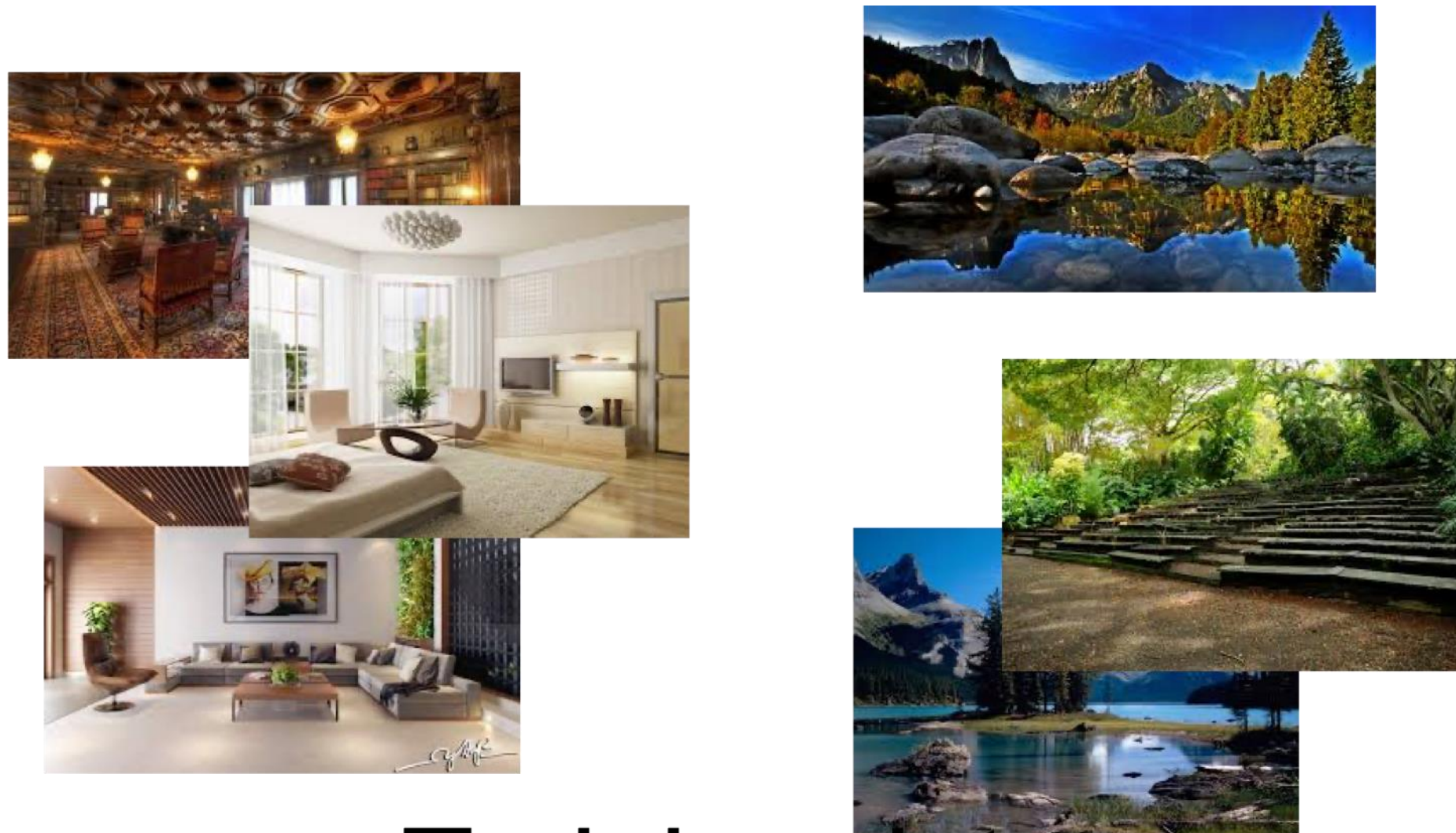


indoor



outdoor

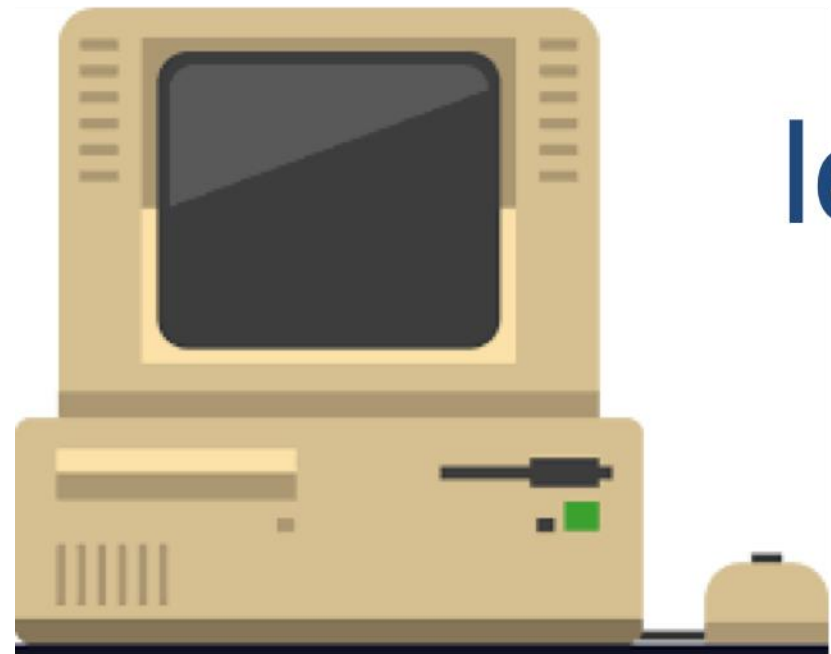
Example 2: Classify Images

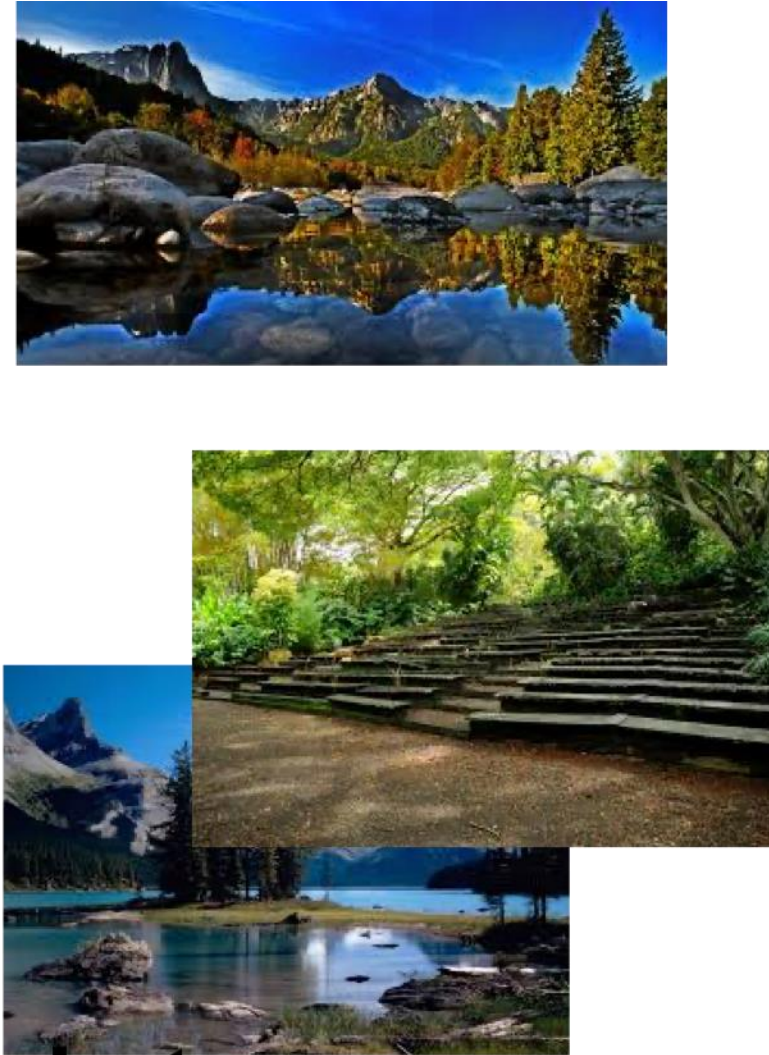
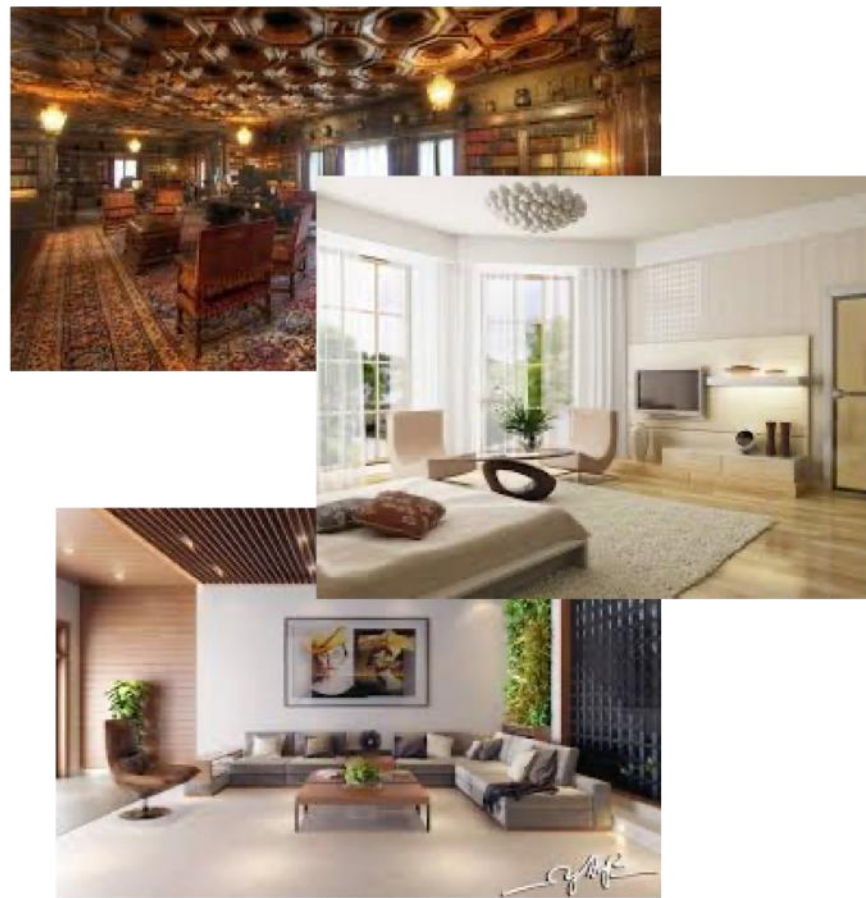


Training data



learning (i.e., training)





Label: outdoor

Label: indoor

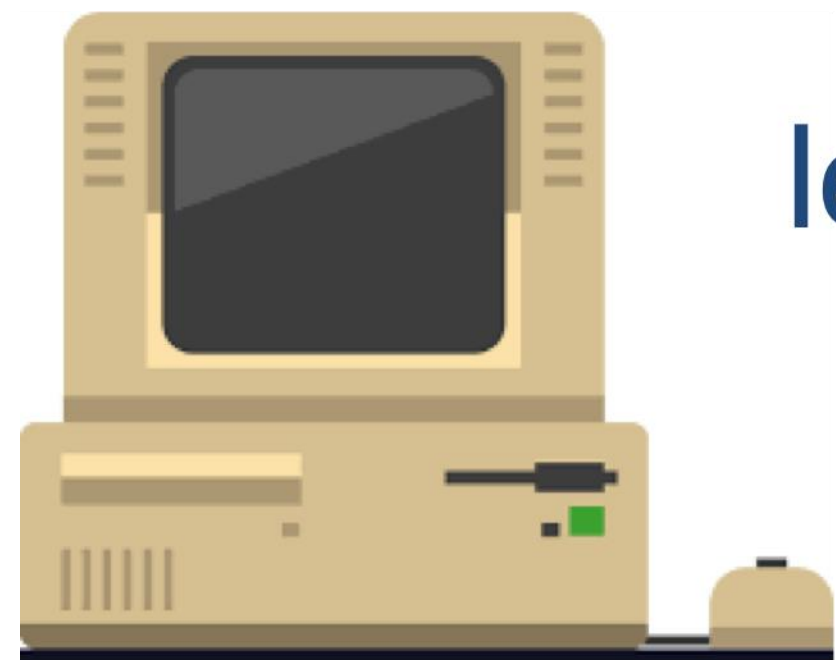
Training data

Test data

learning (i.e., training)

testing

performance



How to represent data?

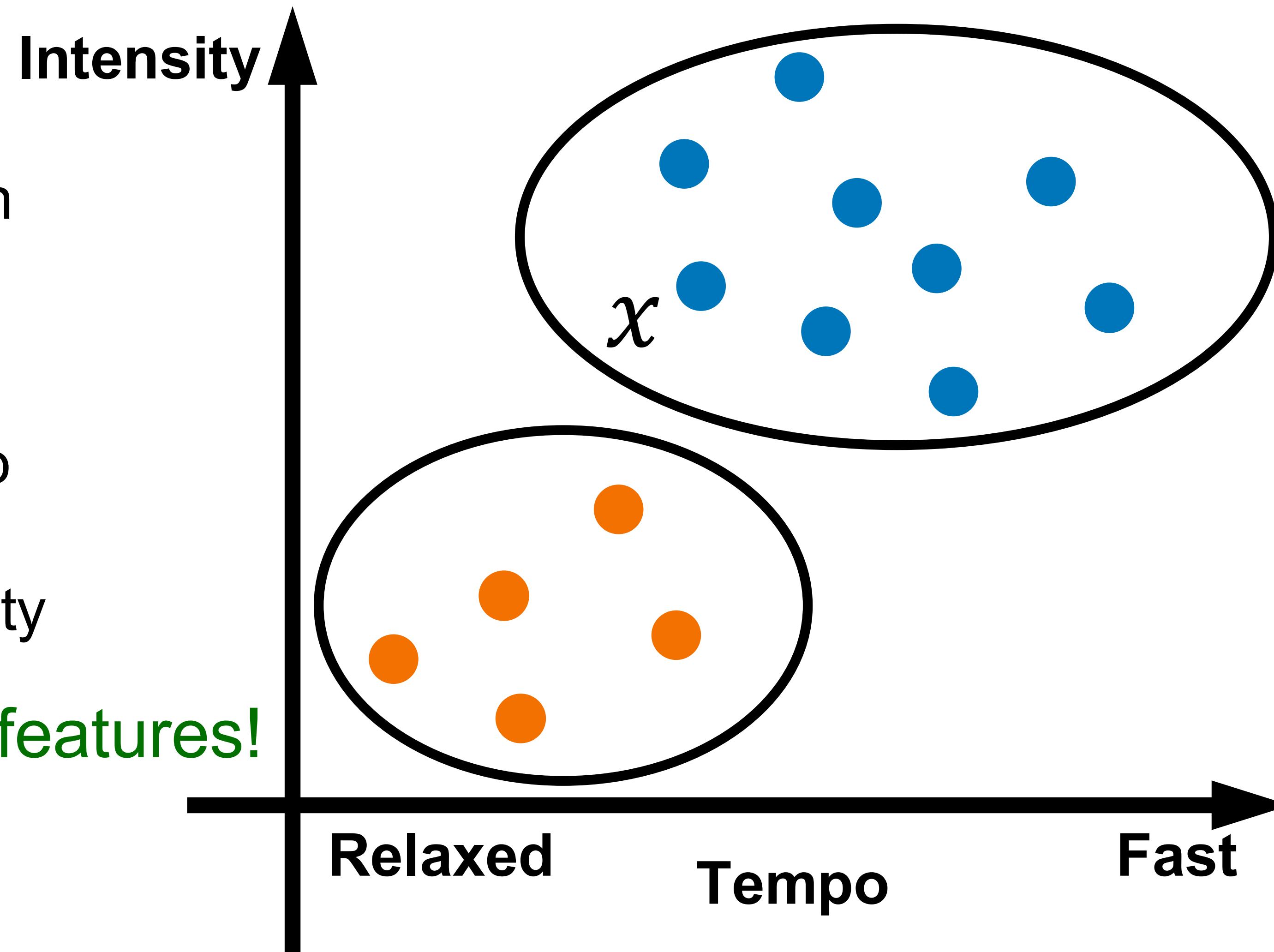
input data

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

d : feature dimension

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

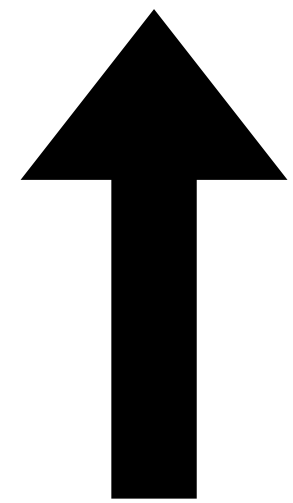
There can be many features!



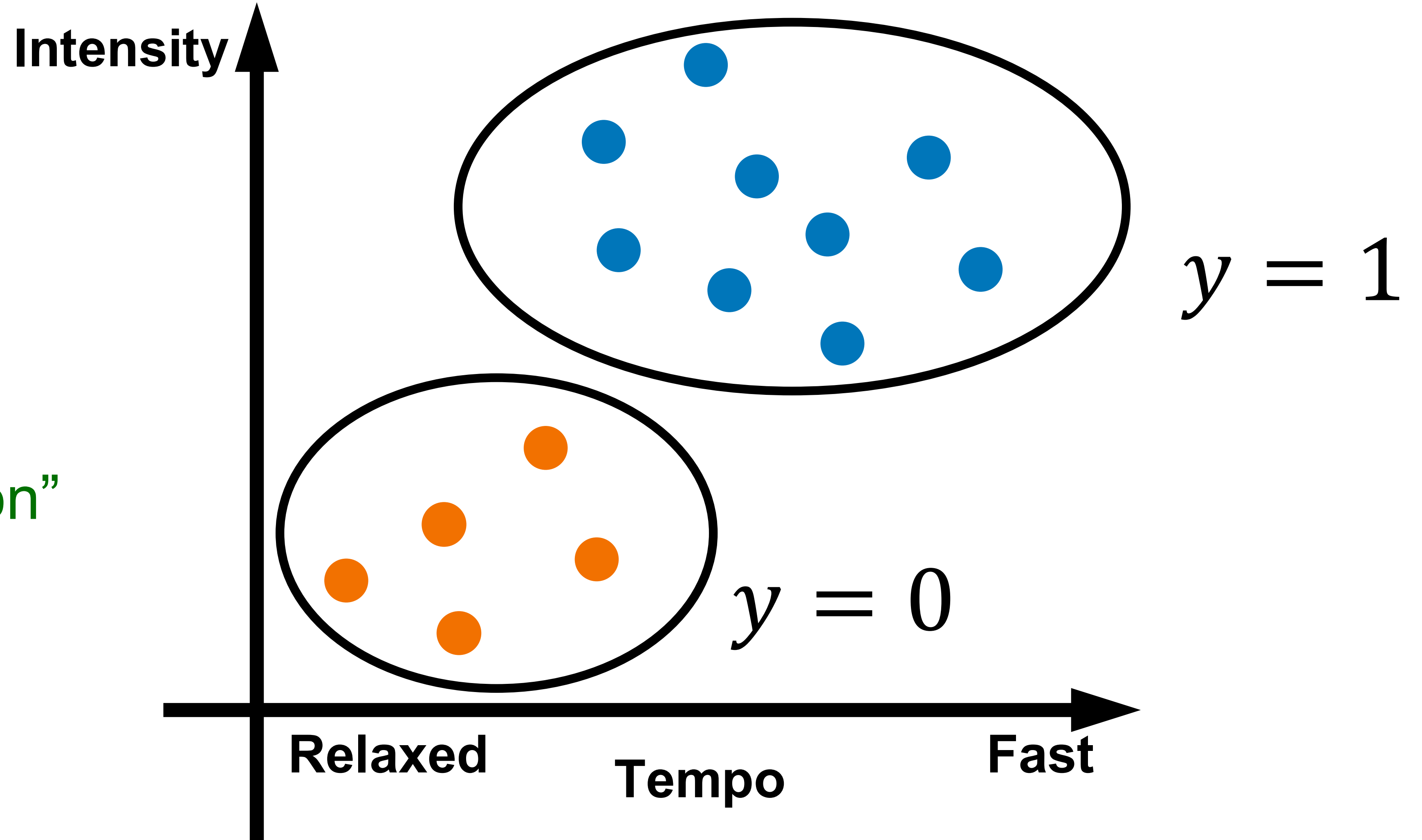
How to represent data?

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

- the label is a **discrete** variable

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

Regression

- the label is a **continuous** variable

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

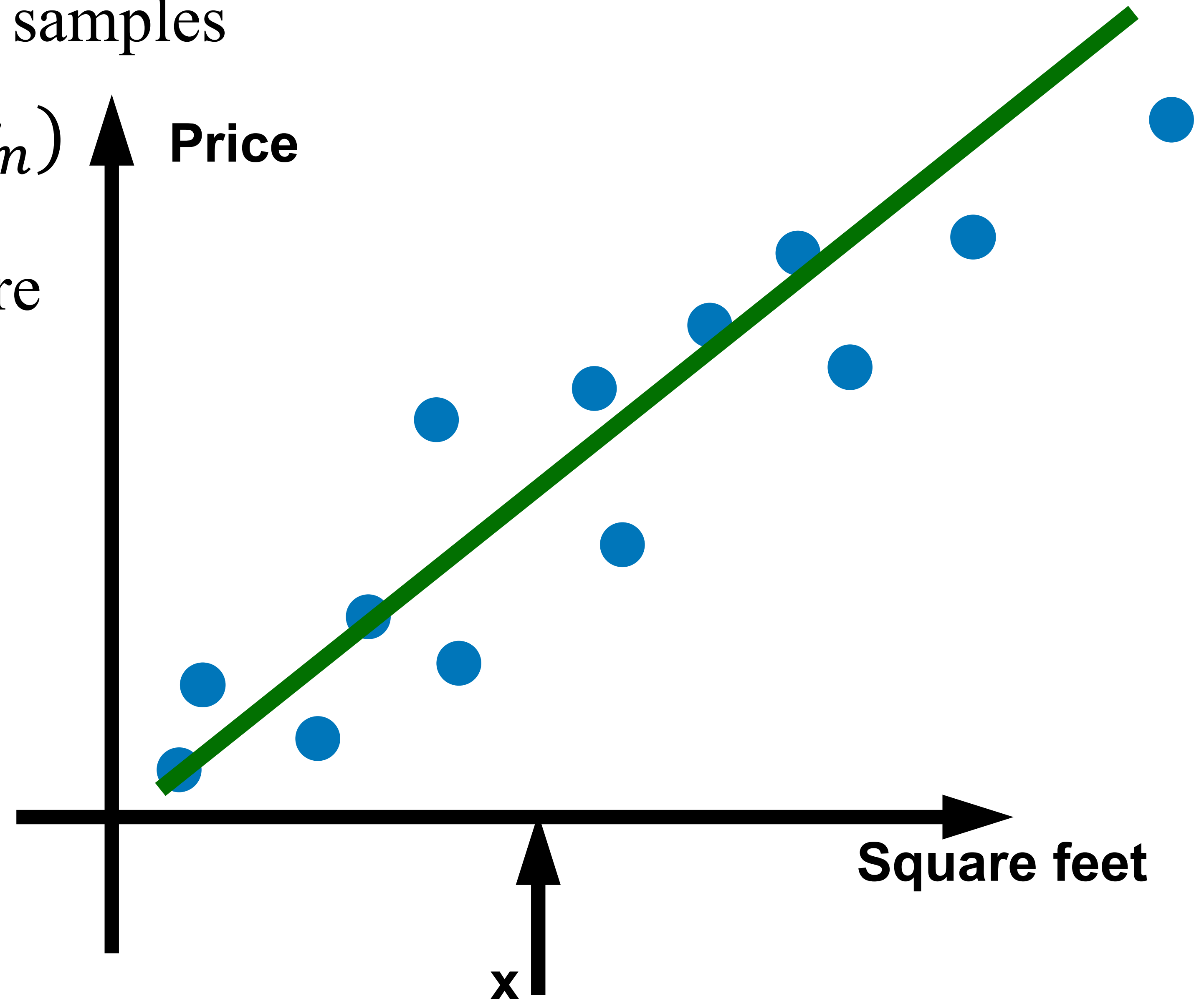
Example of regression: housing price prediction

Given: a dataset that contains n samples

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

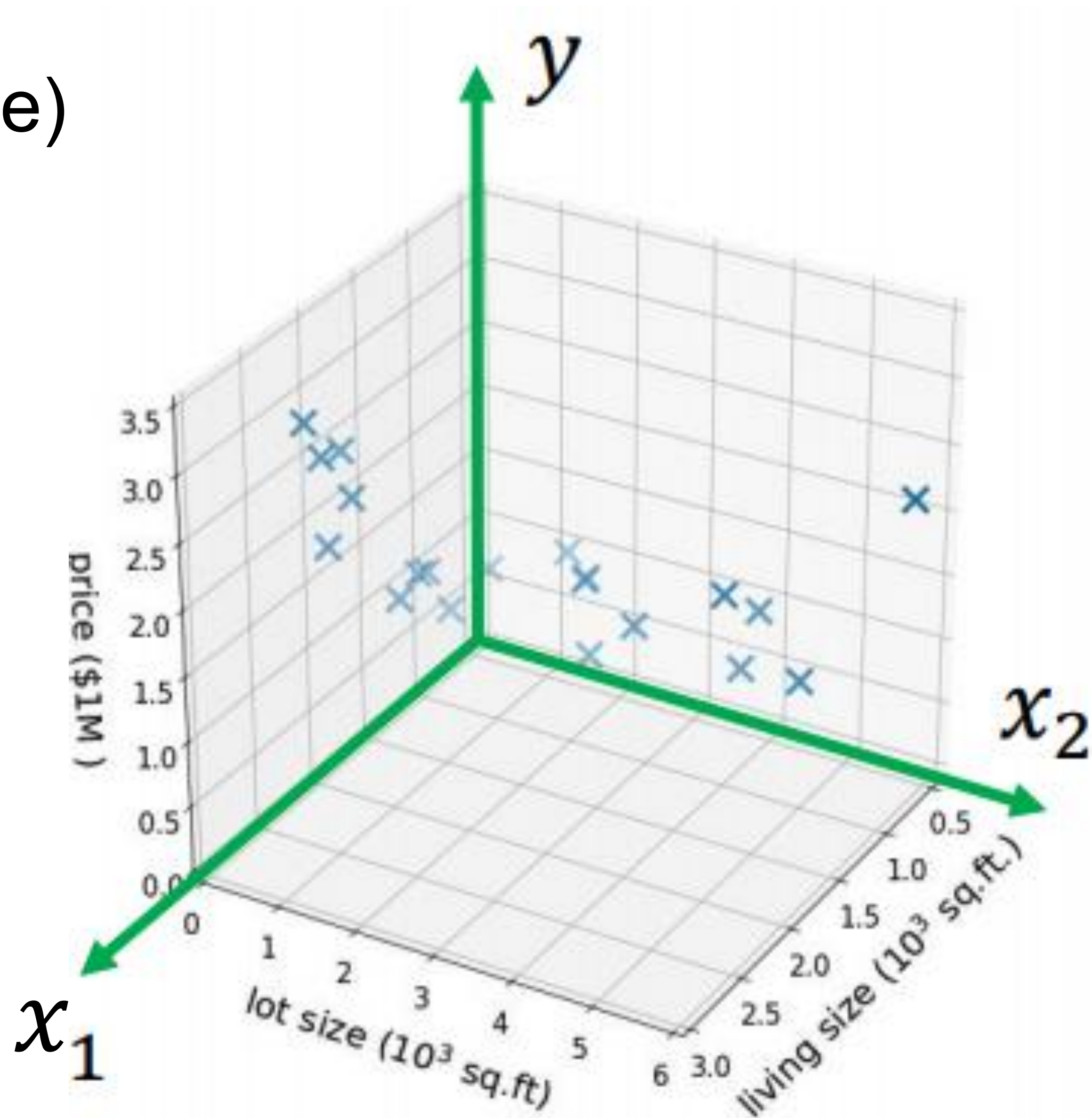
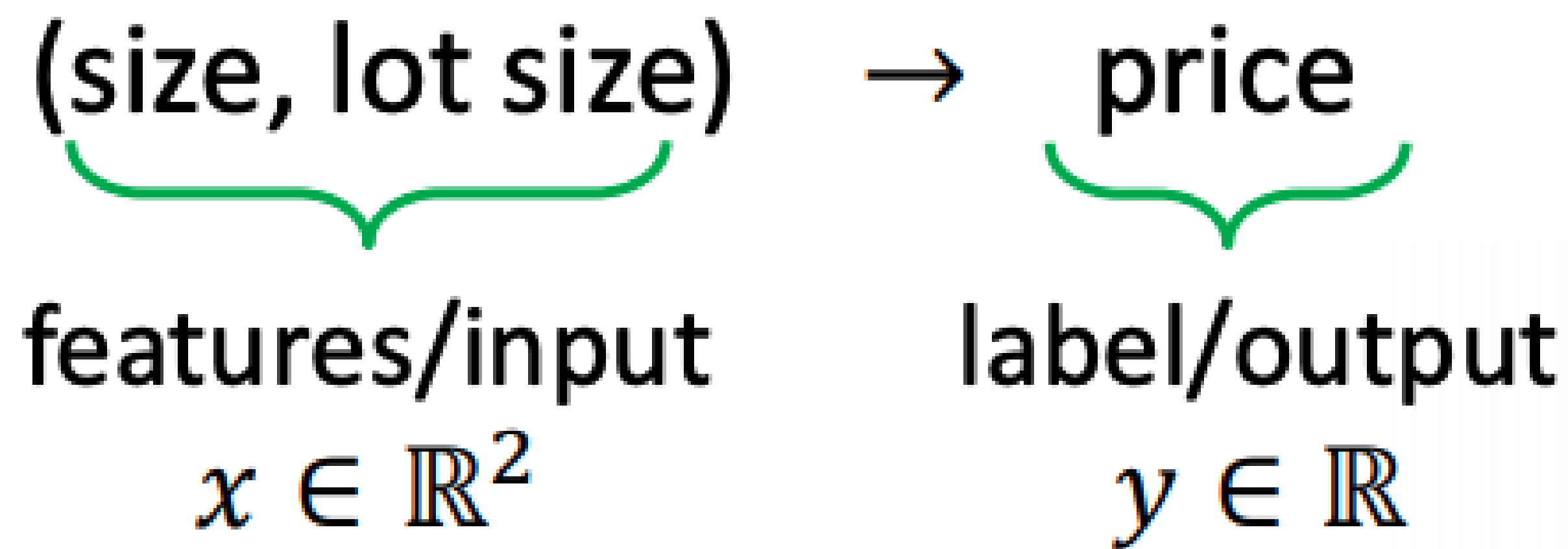
Task: if a residence has x square feet, predict the price?

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



Example of regression: housing price prediction

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

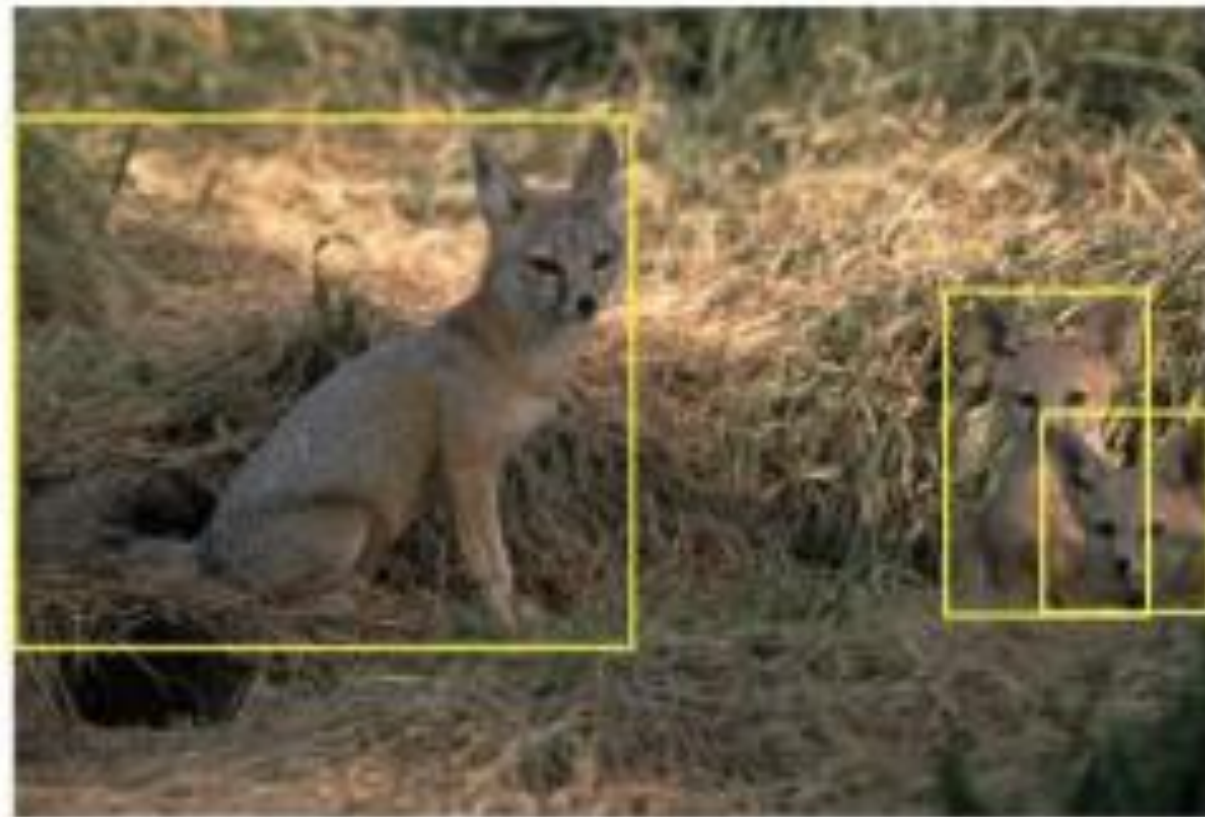


(credit: stanford CS229)

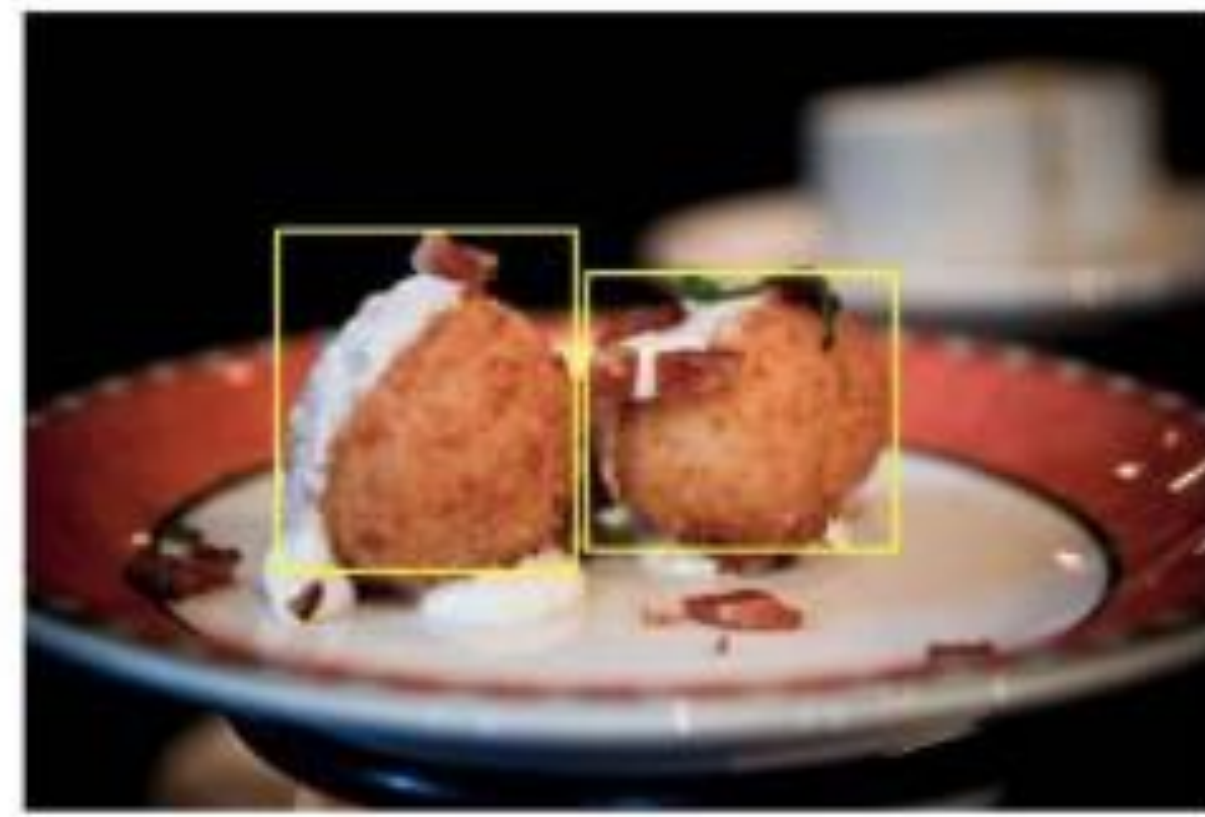
Supervised Learning: More examples

x = raw pixels of the image

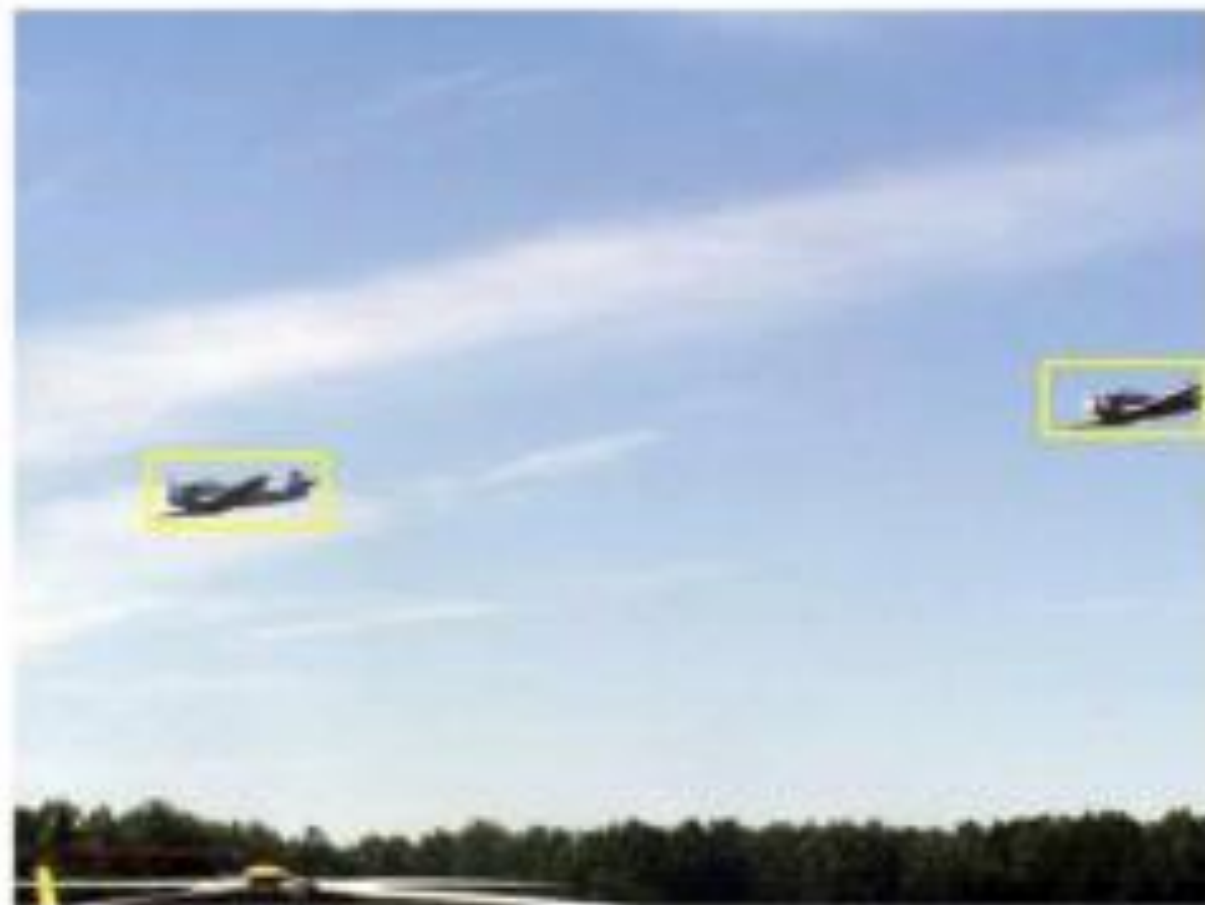
y = bounding boxes



kit fox



croquette



airplane



frog

Training Data for Supervised Learning

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

$$(\text{input}, \text{label}), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$$

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

Goal of Supervised Learning

Given training data

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

Learn a function mapping $f: X \rightarrow Y$, 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 :)



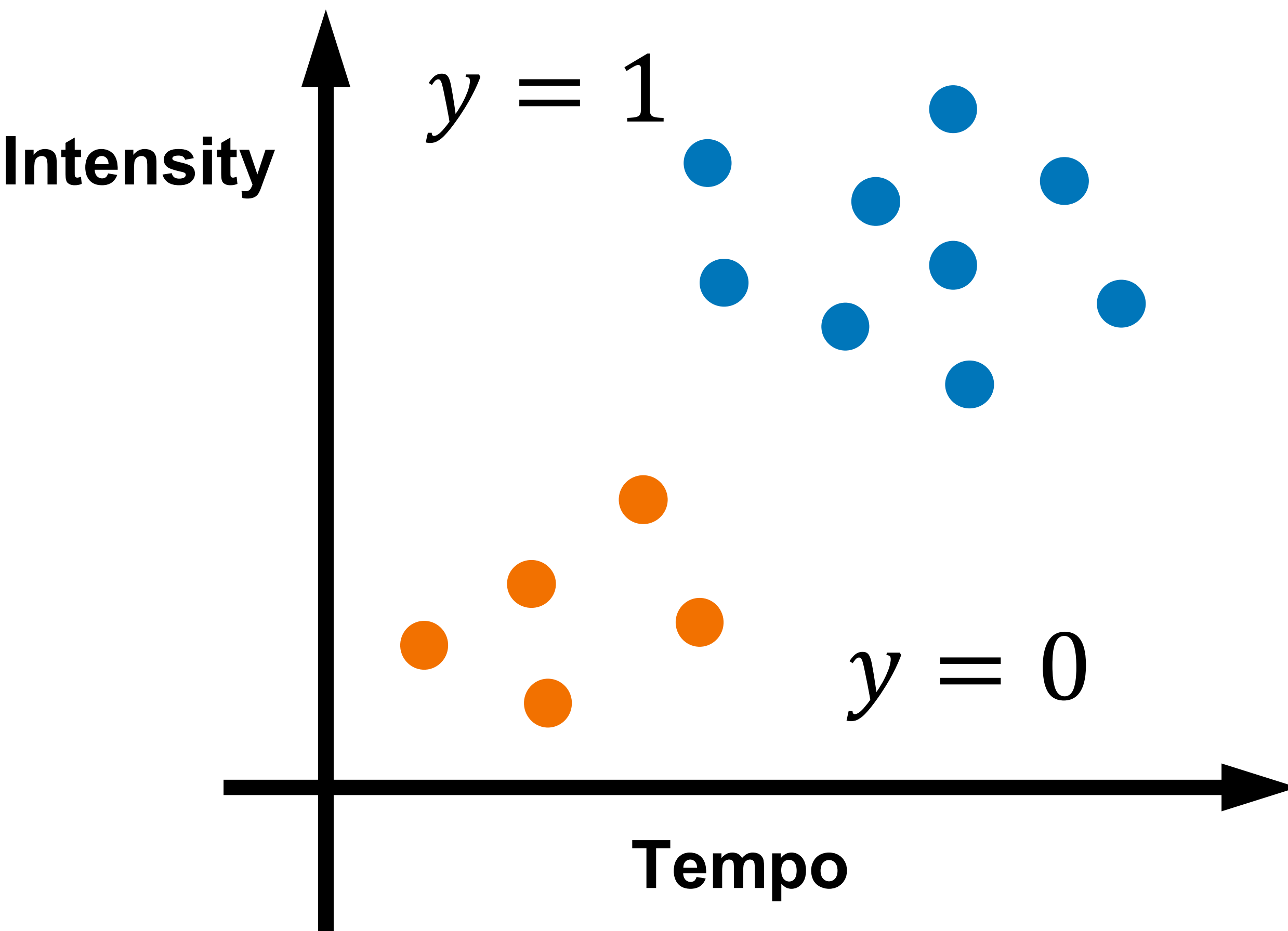
Part II: Unsupervised Learning

Unsupervised Learning

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

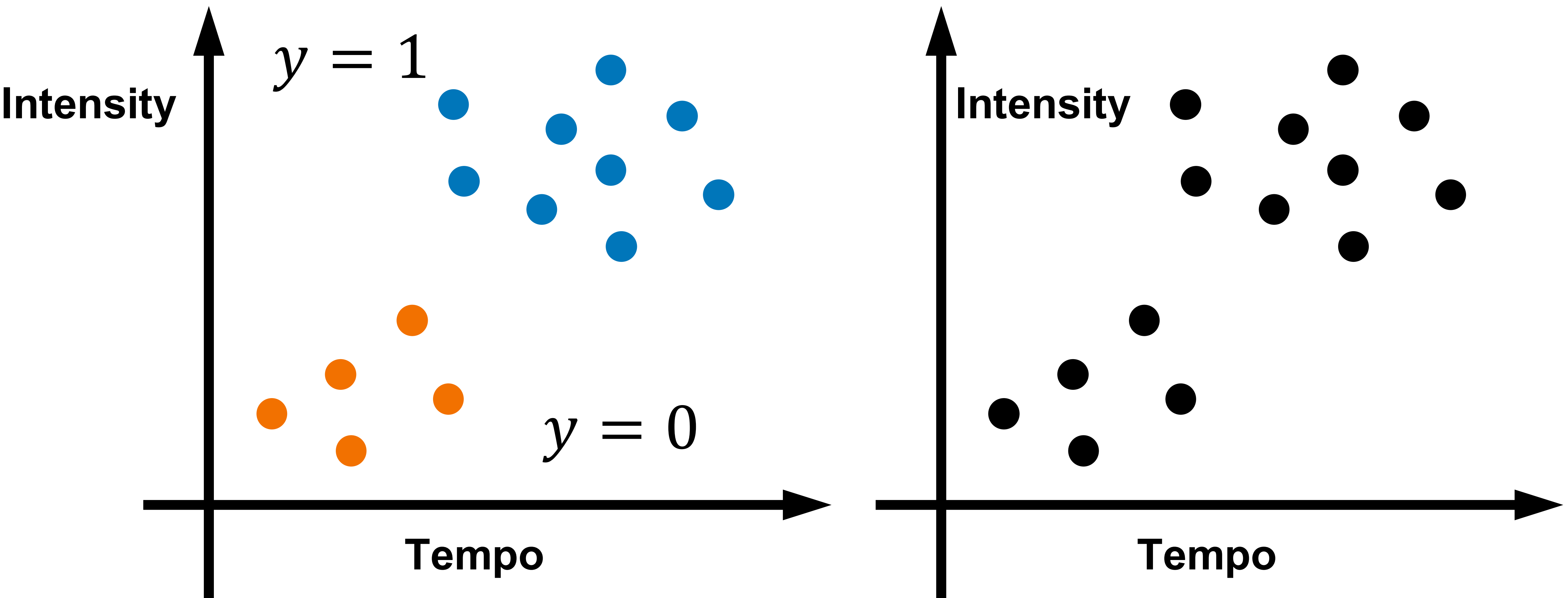
Unsupervised Learning

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



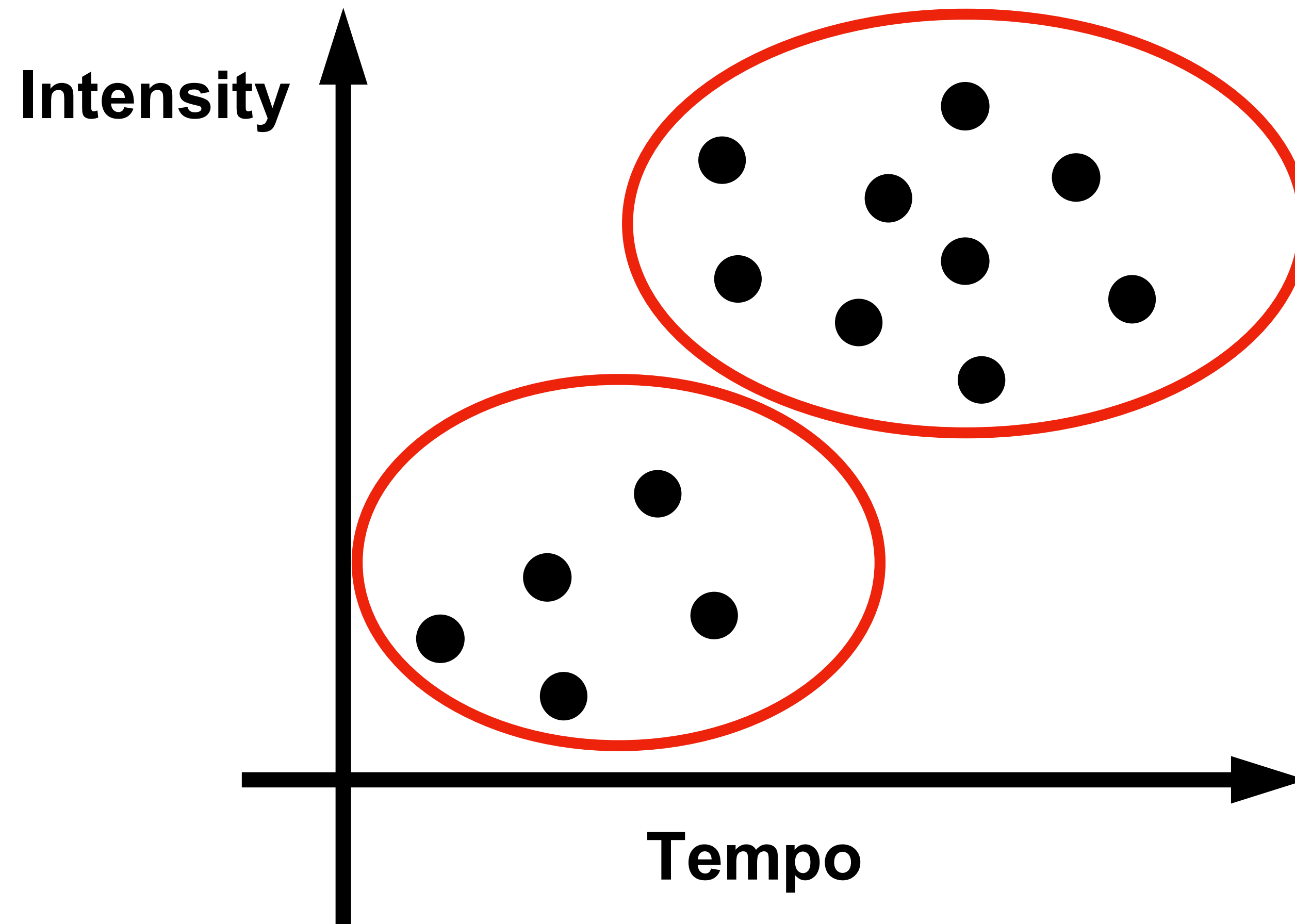
Unsupervised Learning

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

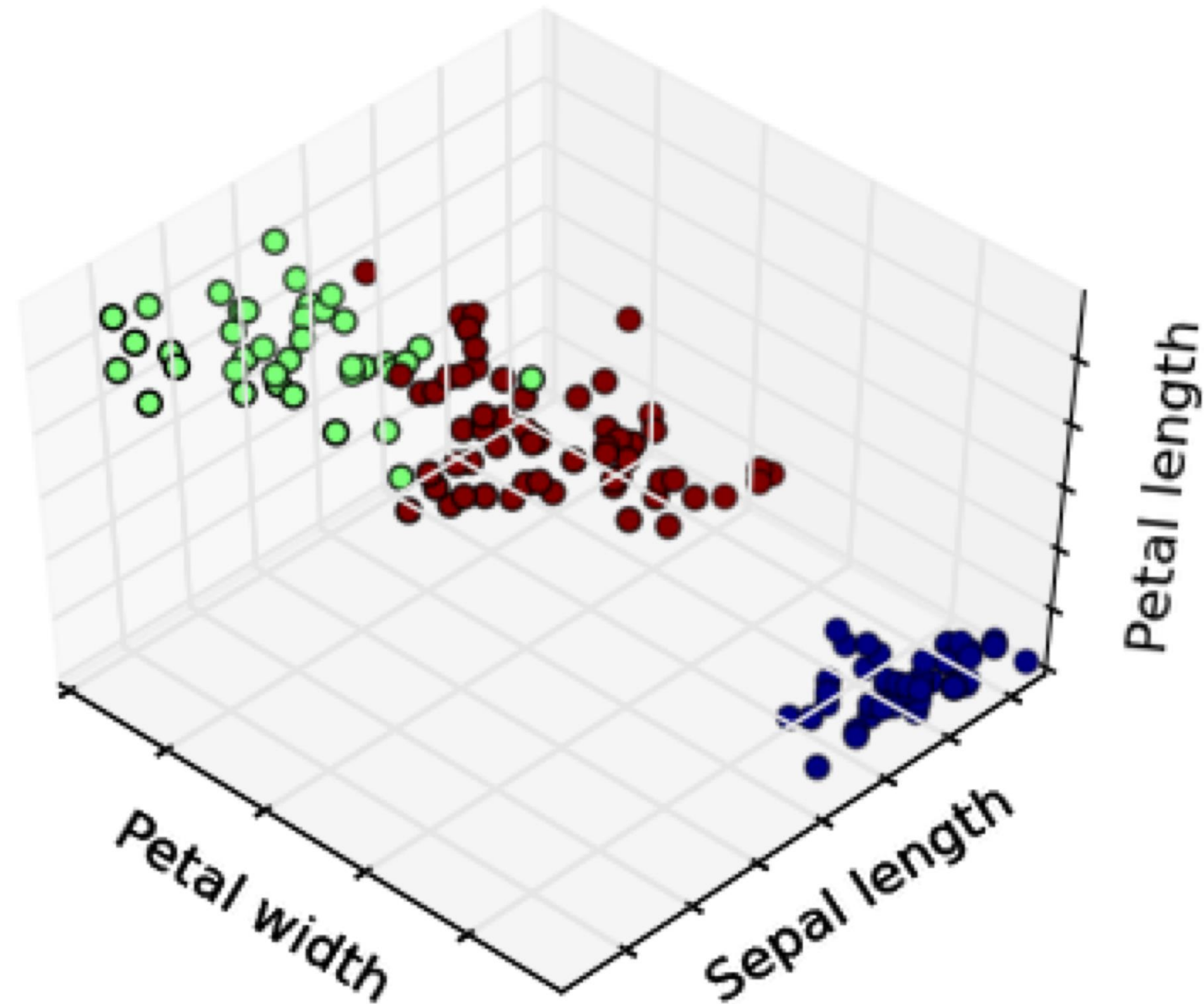


Clustering

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



Clustering

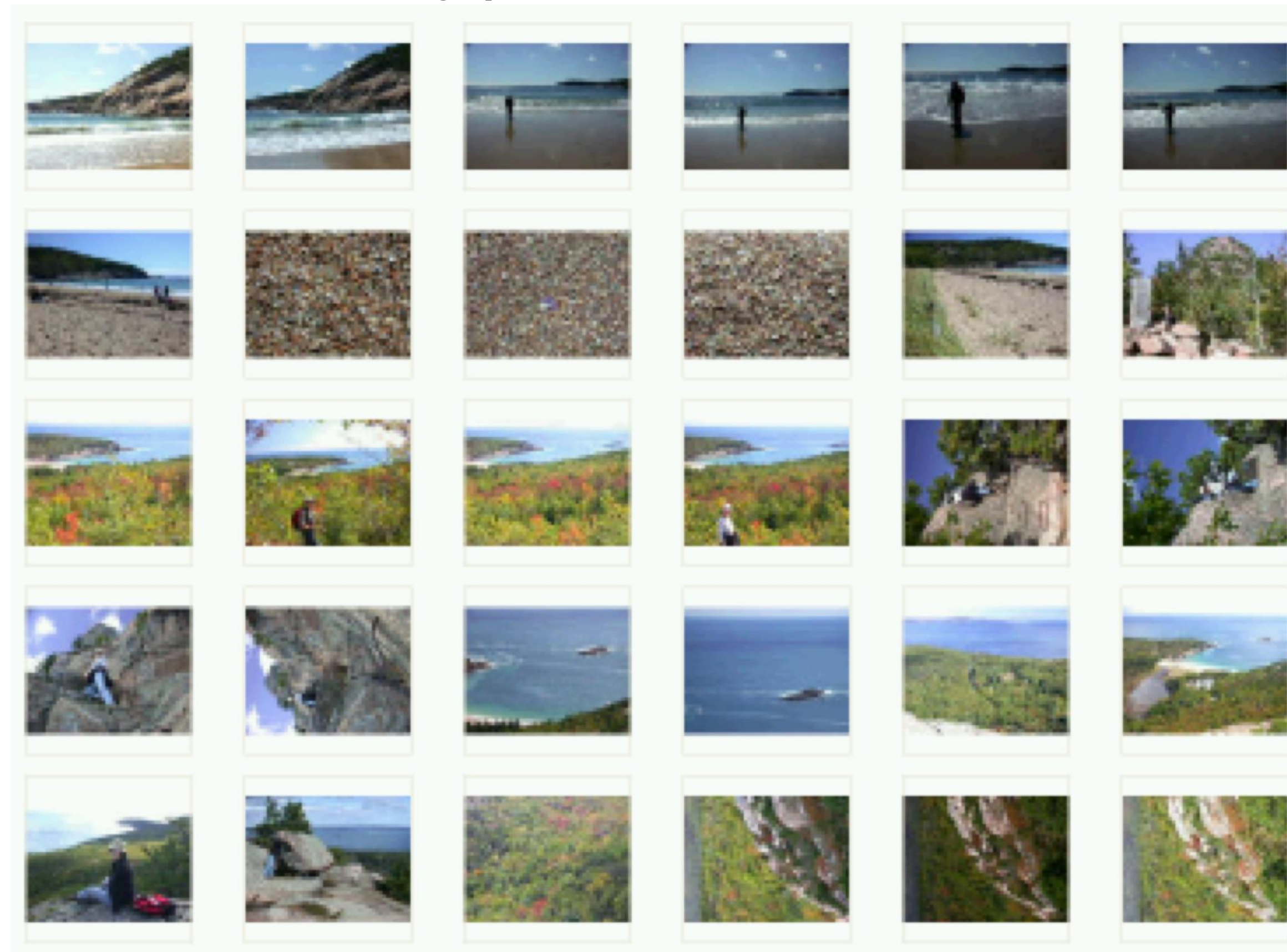


Clustering Irises using three different features

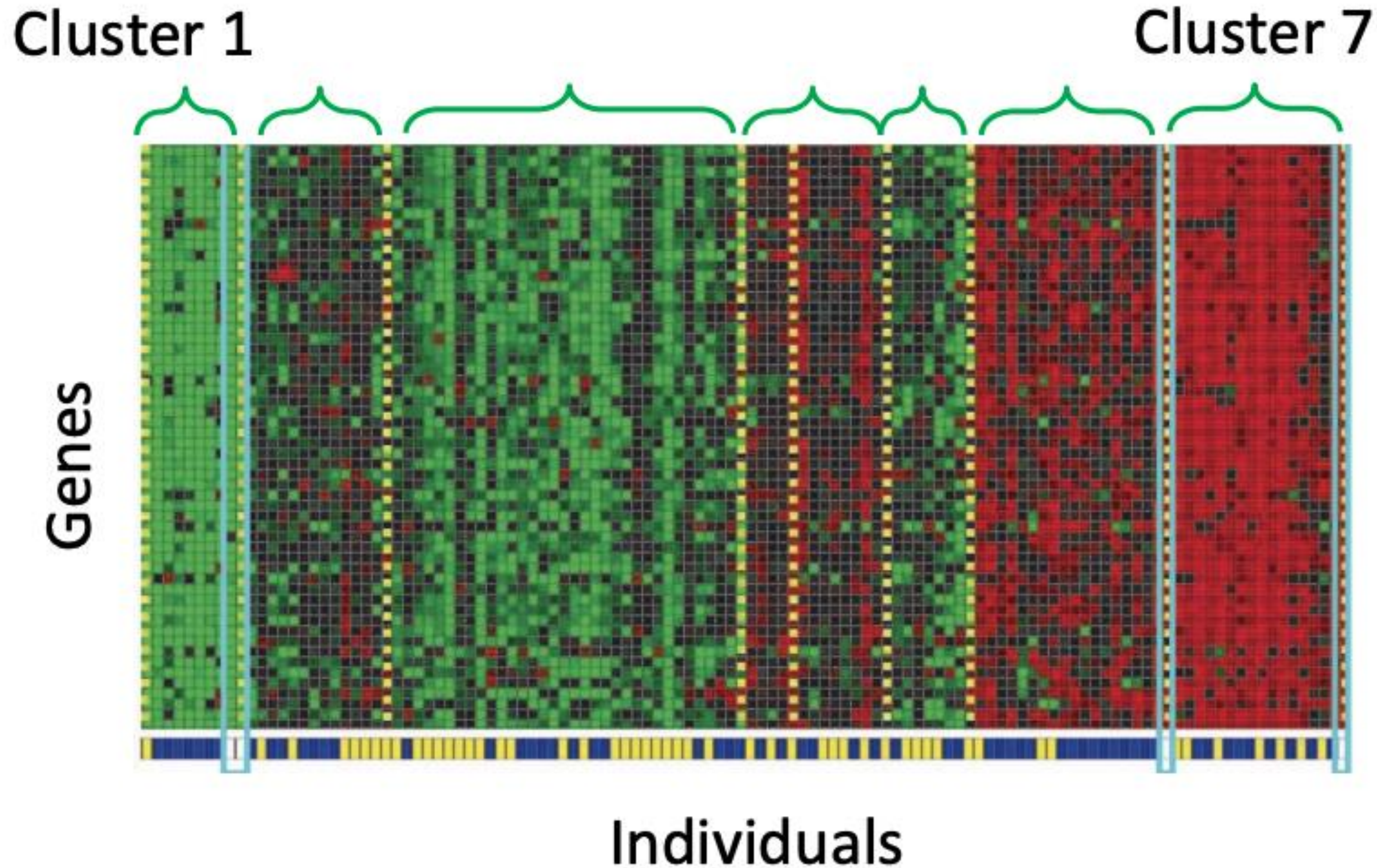
The colors represent clusters identified by the algorithm, **not** y's provided as input

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)

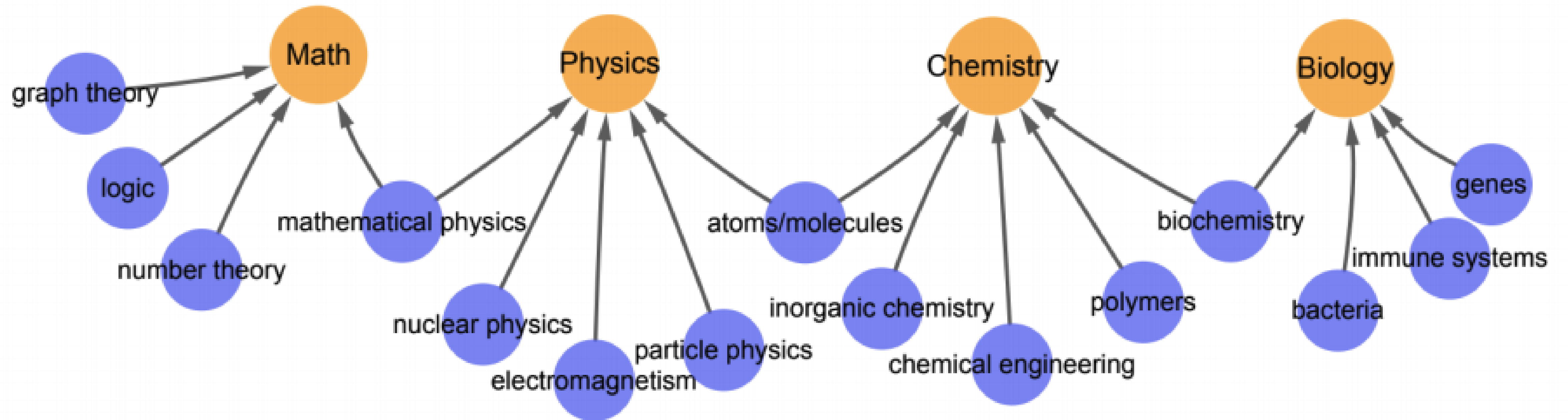


Clustering Genes



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



How do we perform clustering?

Many clustering algorithms.

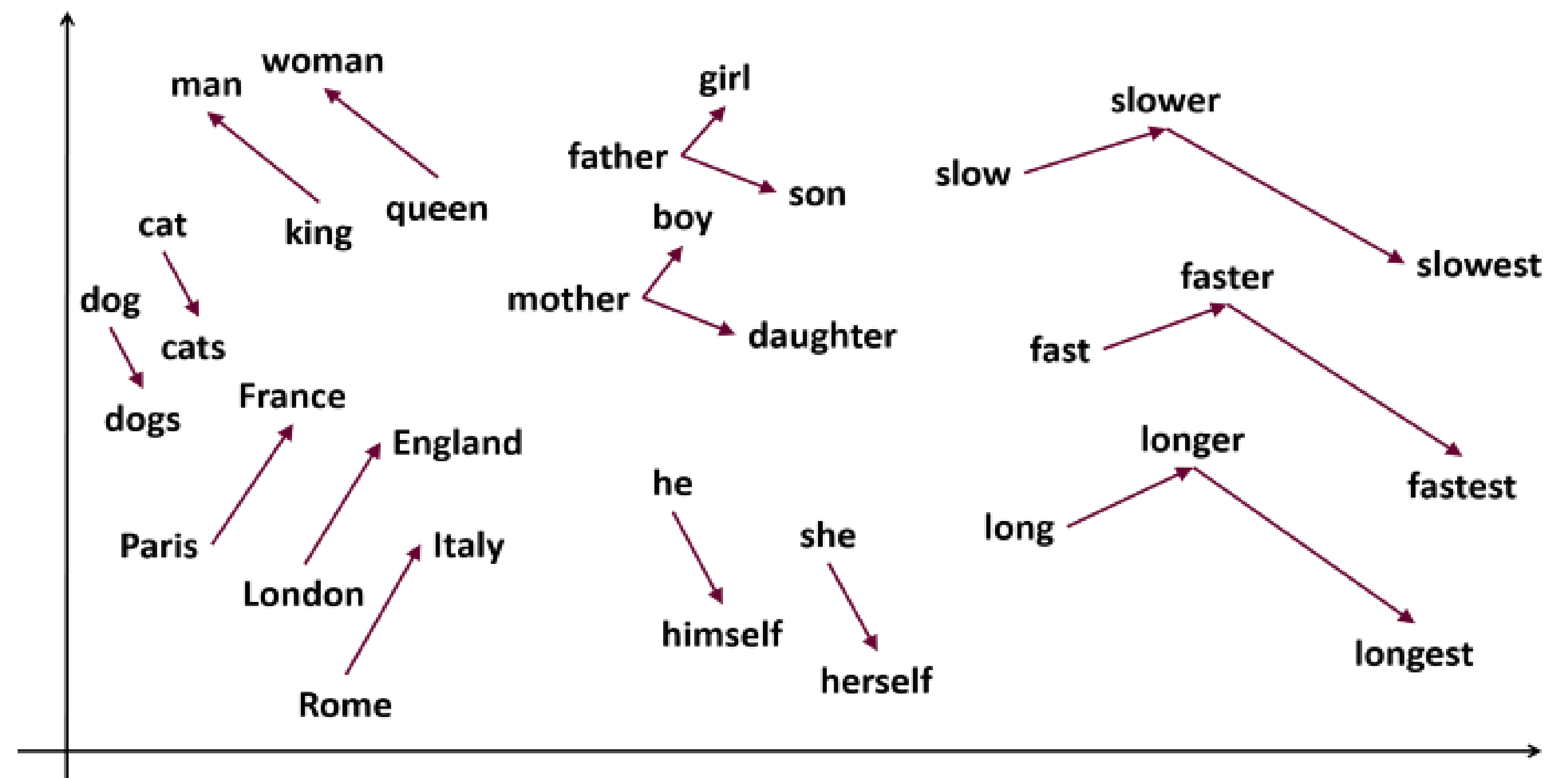
We will look at the two most frequently used ones in the coming lectures:

- *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

Self-Supervised Learning

- Given: dataset contains **no label** $x_1, x_2, \dots, x_n$
- **Goal:** discover interesting patterns and structures in the data
- **Approach:** generate supervision signal from data.
Solve a *pretext task*

Example: word embeddings



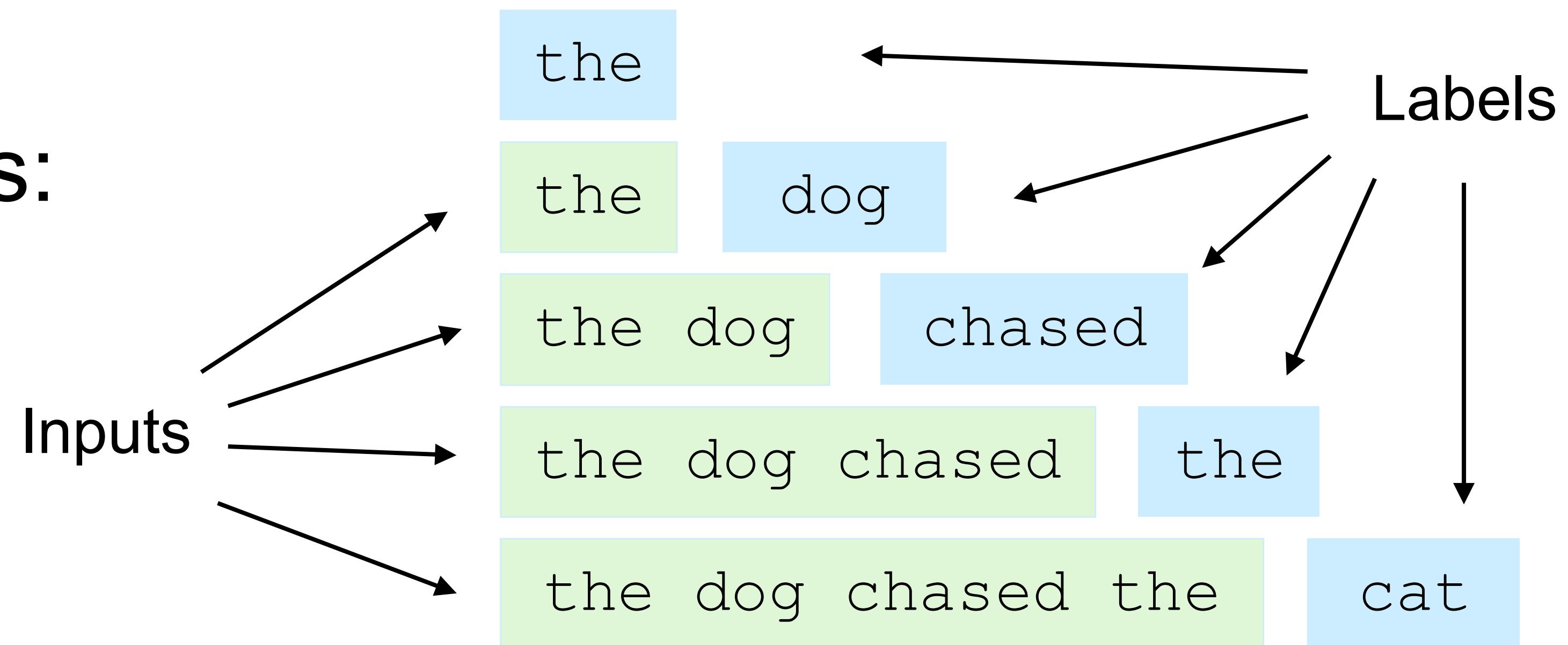
Self-Supervised Learning for LLMs

- Pretext task for large language models:
next-word prediction

- Original text:

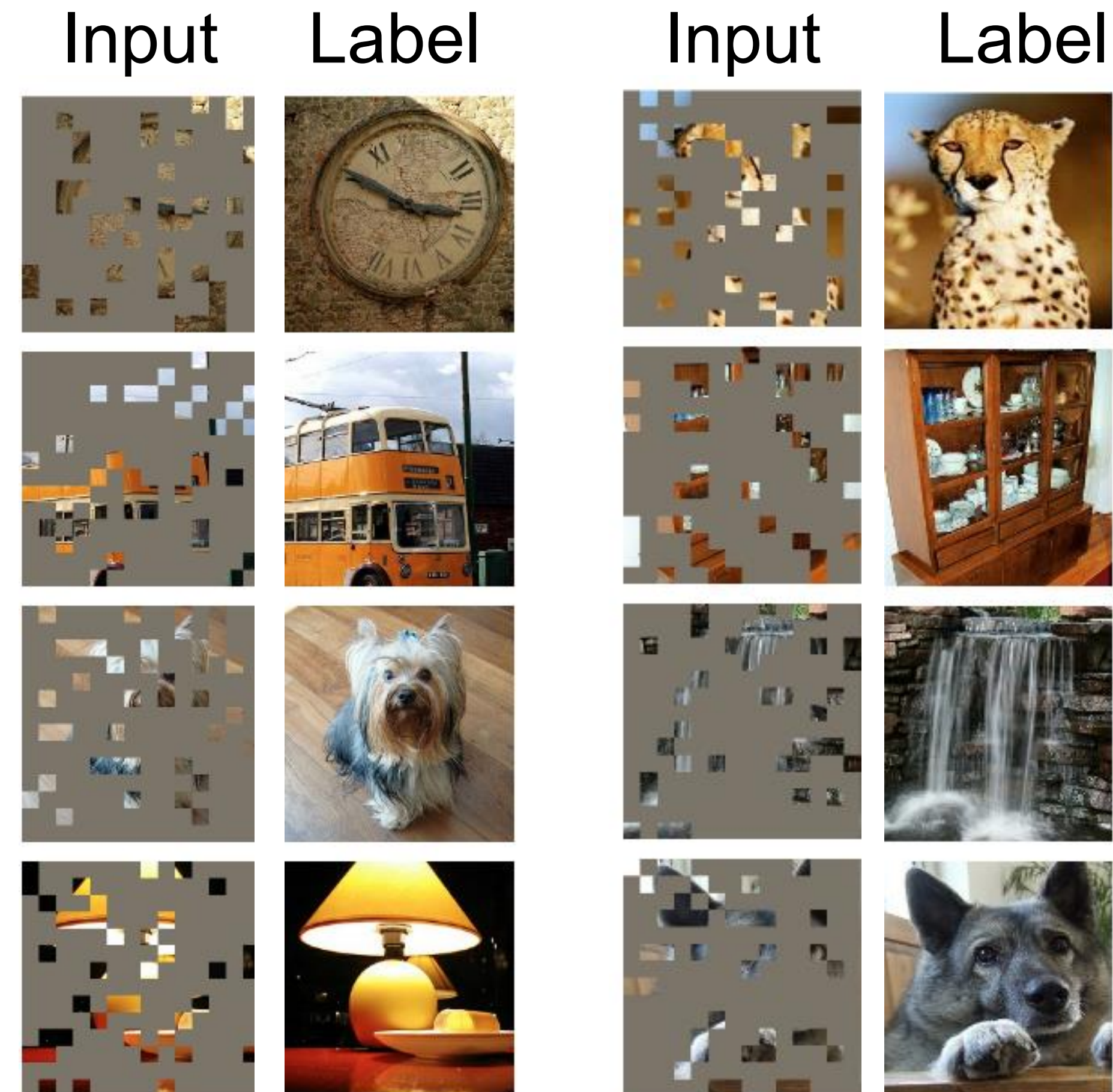
the dog chased the cat

- Split into five labeled problems:



Self-Supervised Learning in Vision

- Another common pretext task: **image inpainting**
- High-dimensional label!





Part III: Reinforcement Learning

Reinforcement Learning

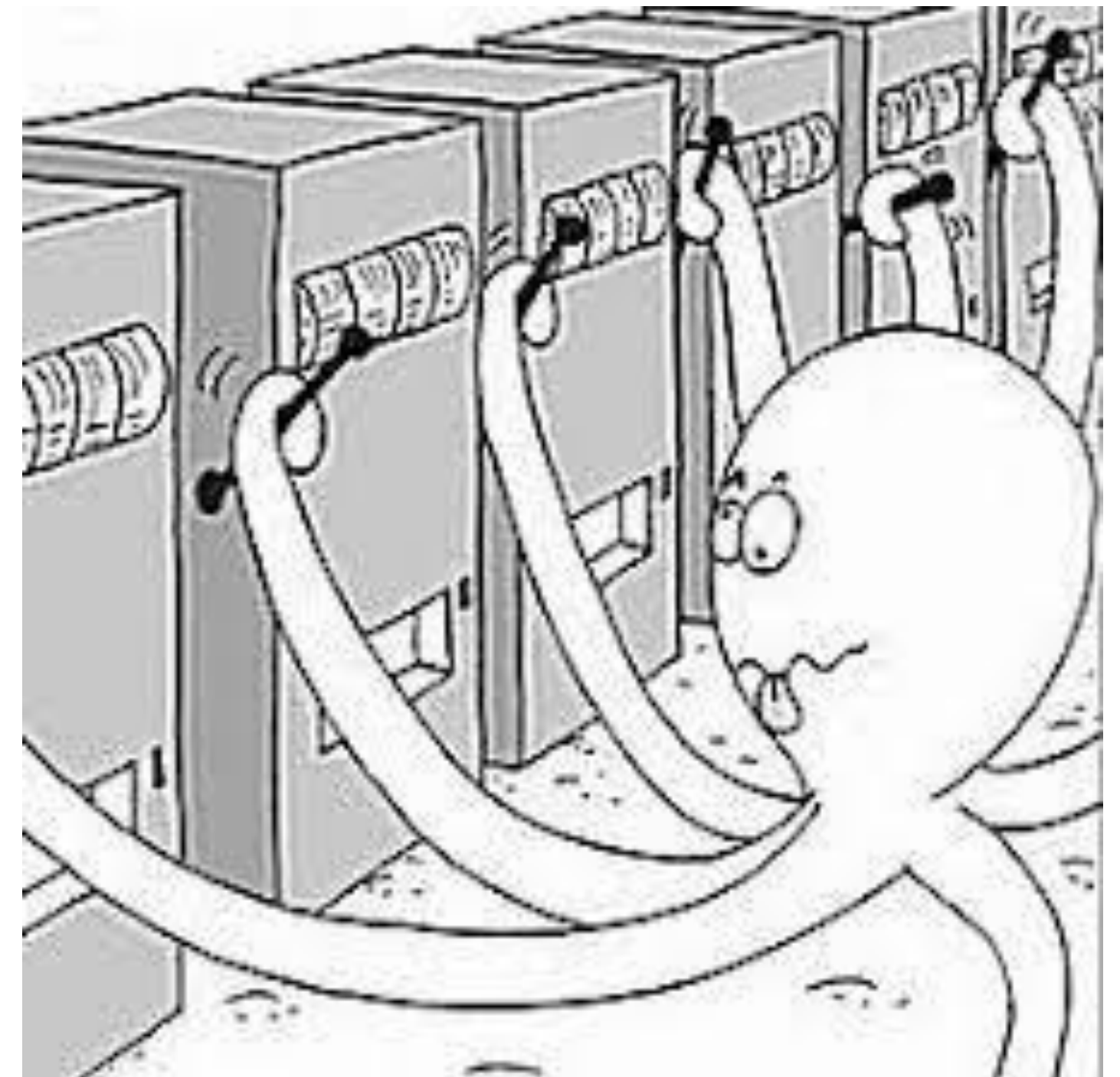
- Given: an agent that can take actions and a reward function specifying how good an action is.
- **Goal:** learn to choose actions that maximize future total reward.



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



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!