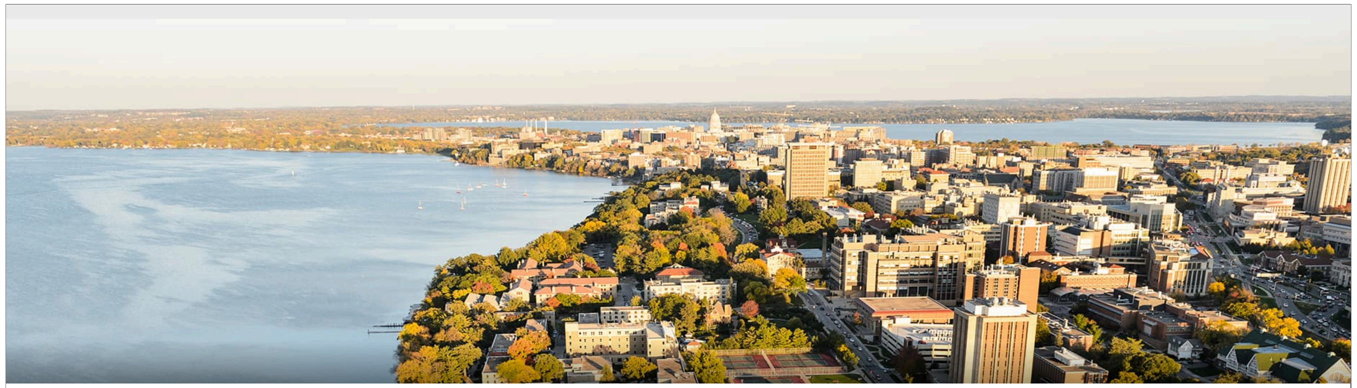


CS 540 Introduction to Artificial Intelligence

Machine Learning Overview

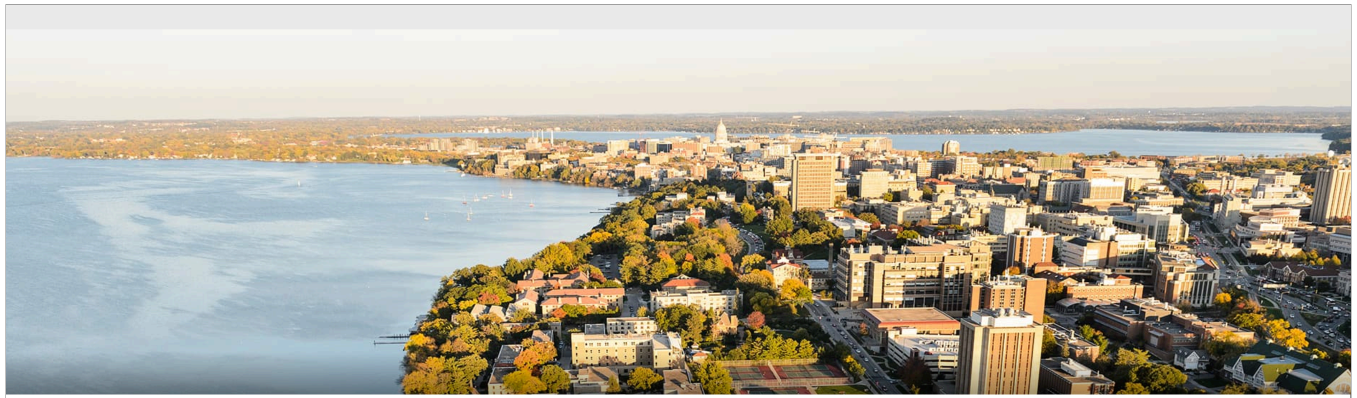
Yingyu Liang
University of Wisconsin-Madison

Sept 30

Slides created by Sharon Li [modified by Yingyu Liang]

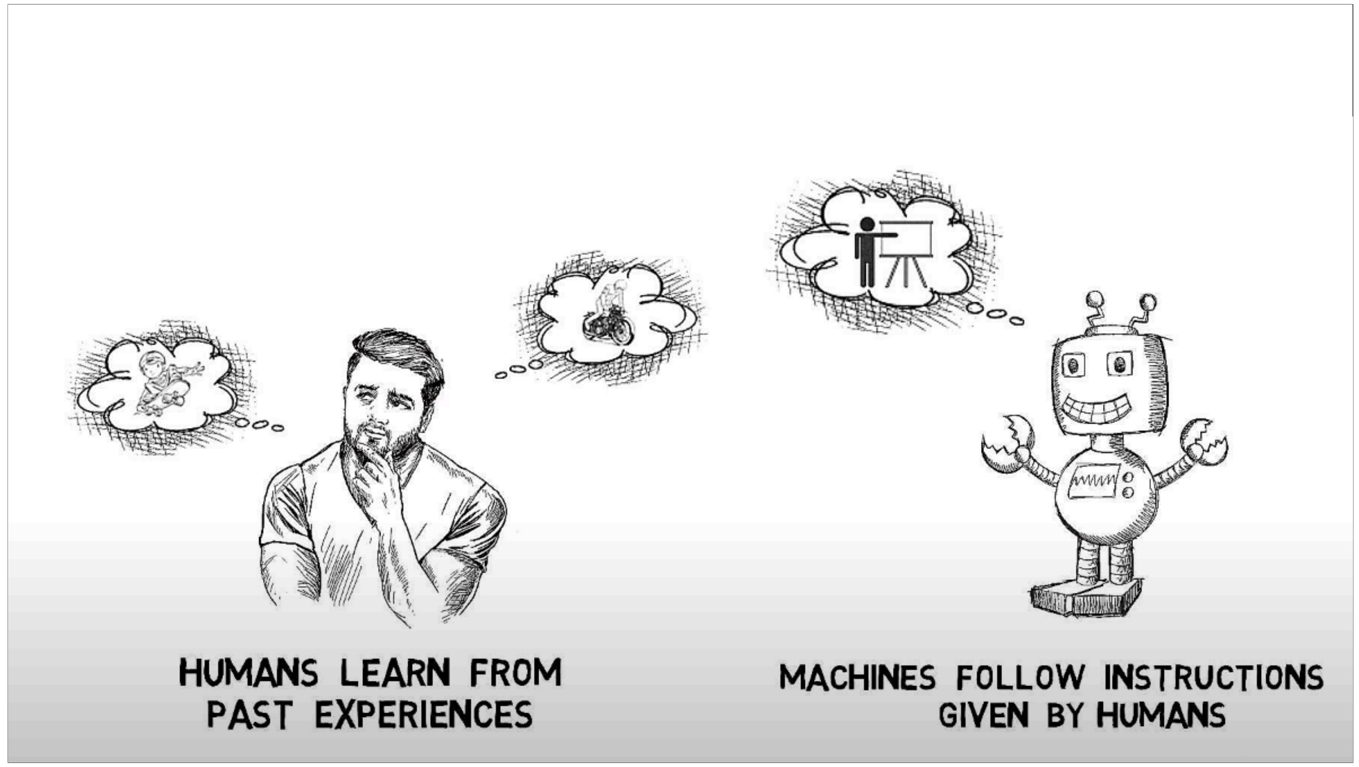
Today's outline

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



Part I: What is machine learning?

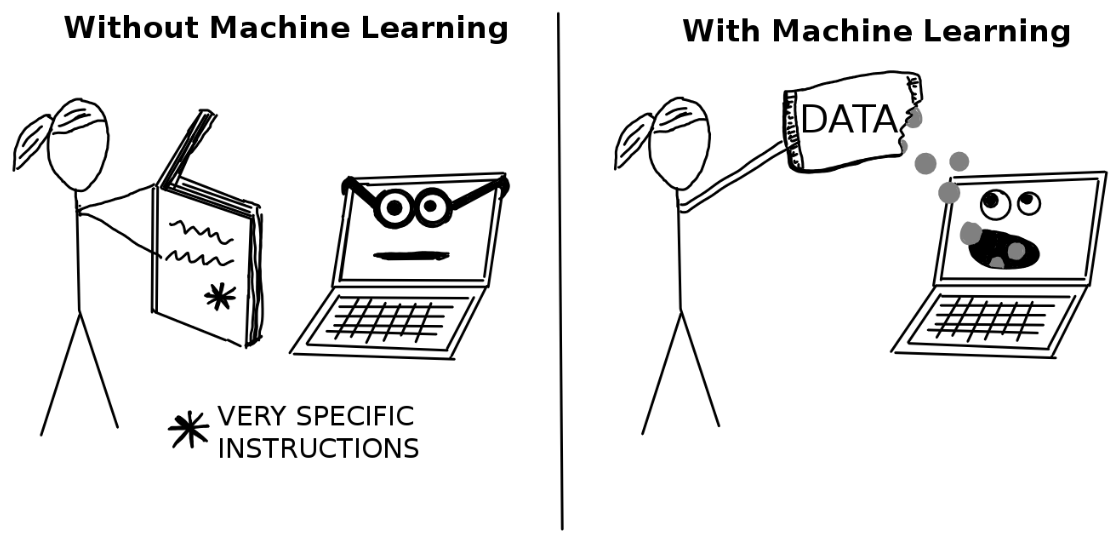
Now Let’s dive into the course overview.



We know humans learn from their past experiences and machines follow instructions given by humans but what if humans can Turing the machines to learn from the past data and to what humans can do act much faster well that’s called machine learning but it’s a lot more than just learning it’s also about understanding and reasoning, so today we will learn about the basics of 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.**



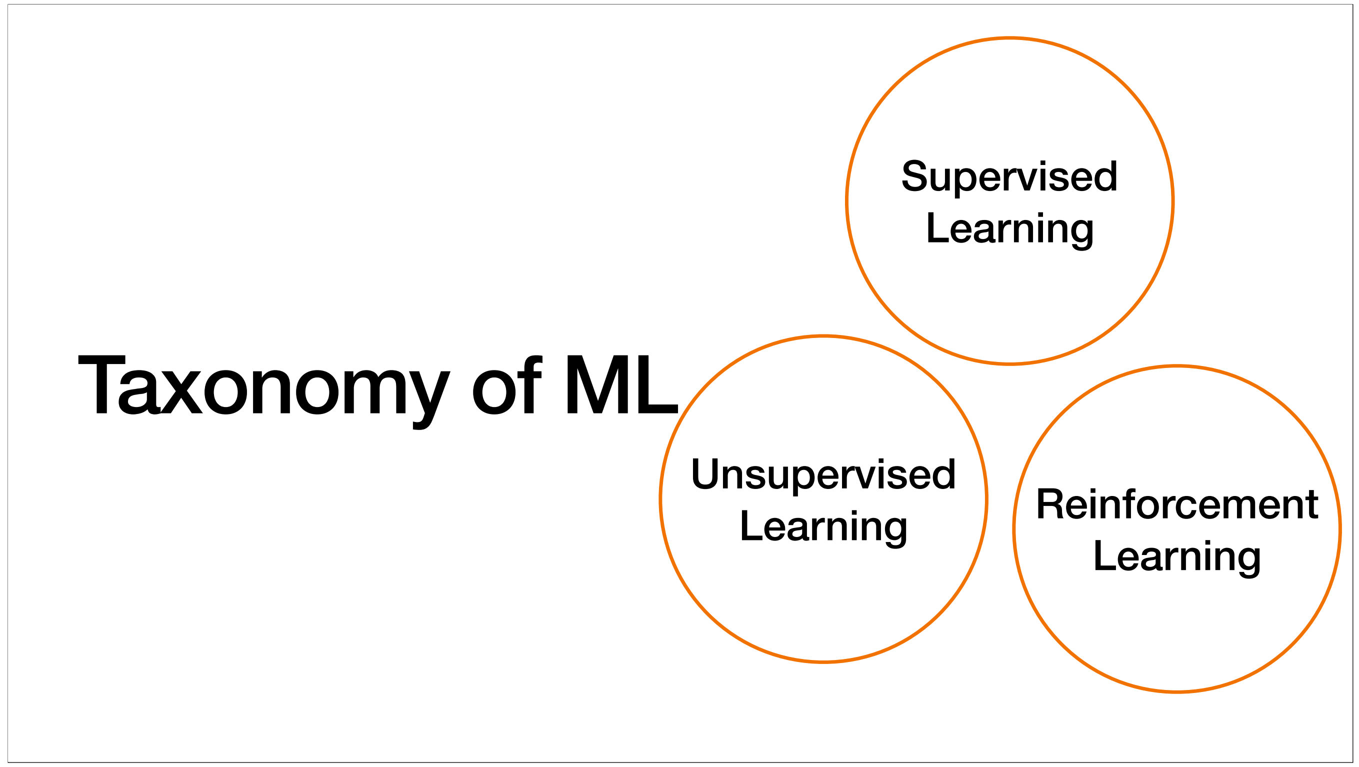
https://tung-dn.github.io/programming.html

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.

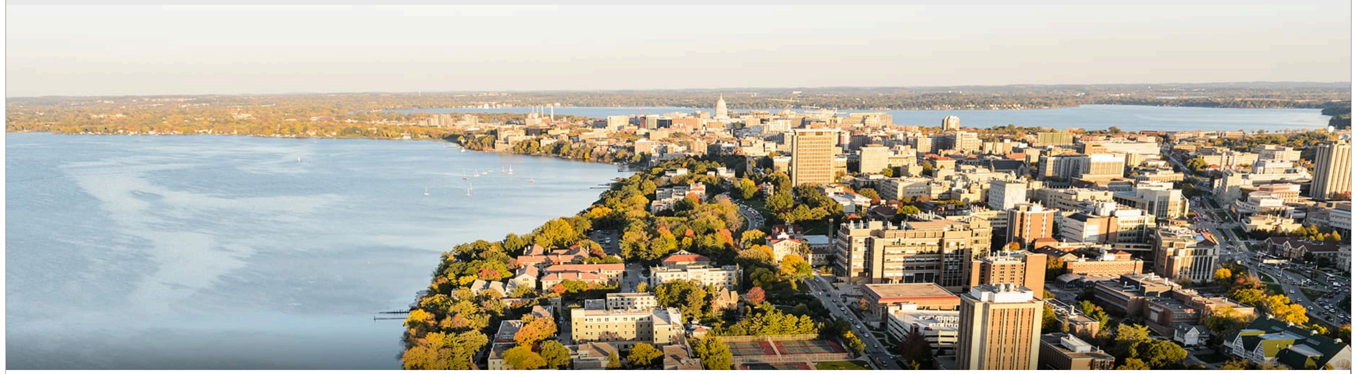
Three key elements of ML by Tom Mitchell:

1. Experience (nowadays called data)
2. Tasks
3. Performance measure



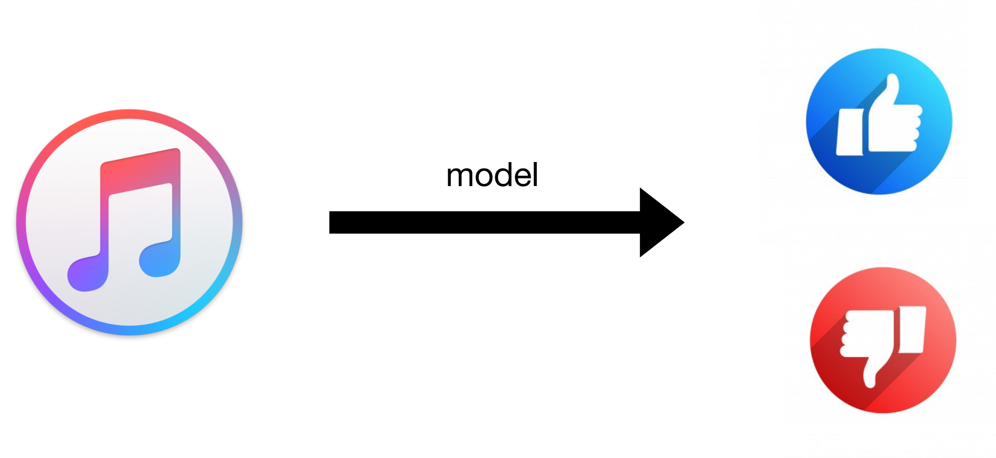
Three main types of machine learning.

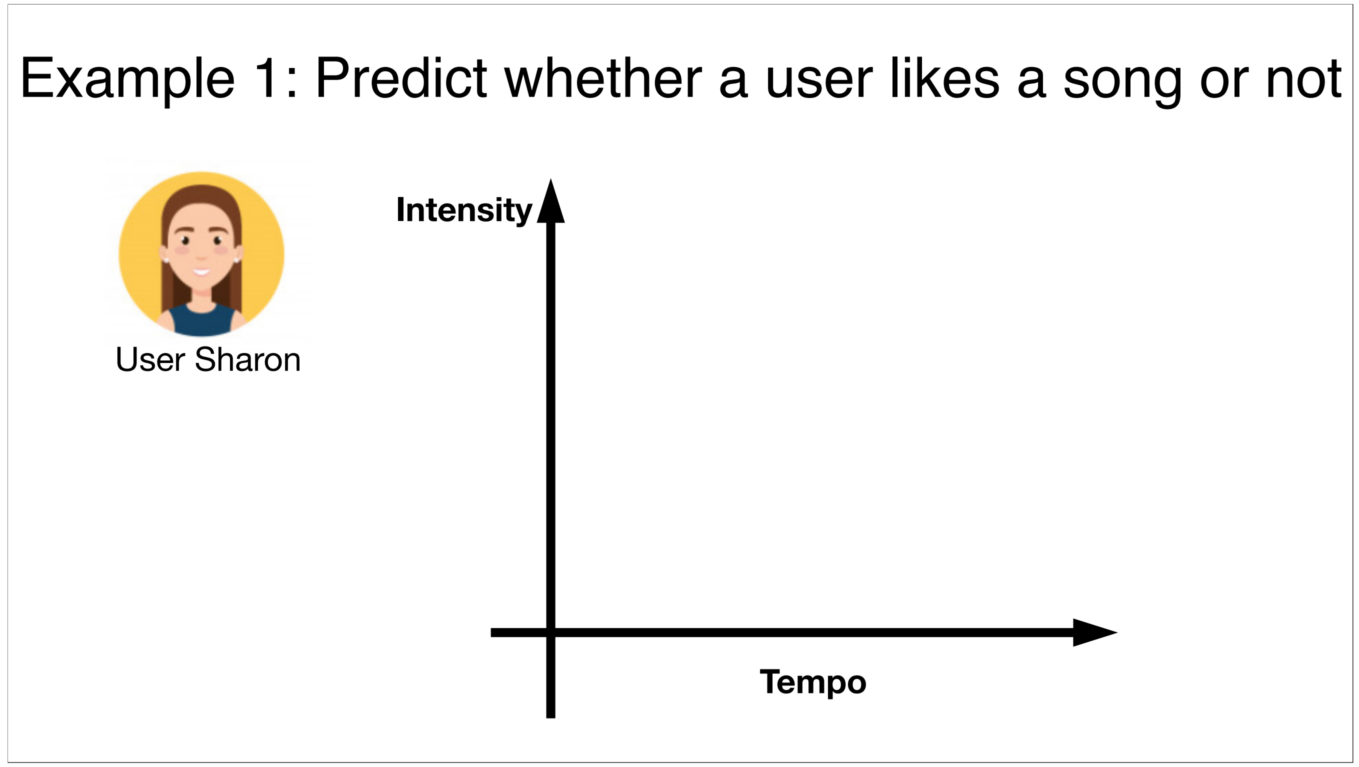
There are other types of machine learning, e.g., semi-supervised learning, active learning. These can be learned in subsequent courses.



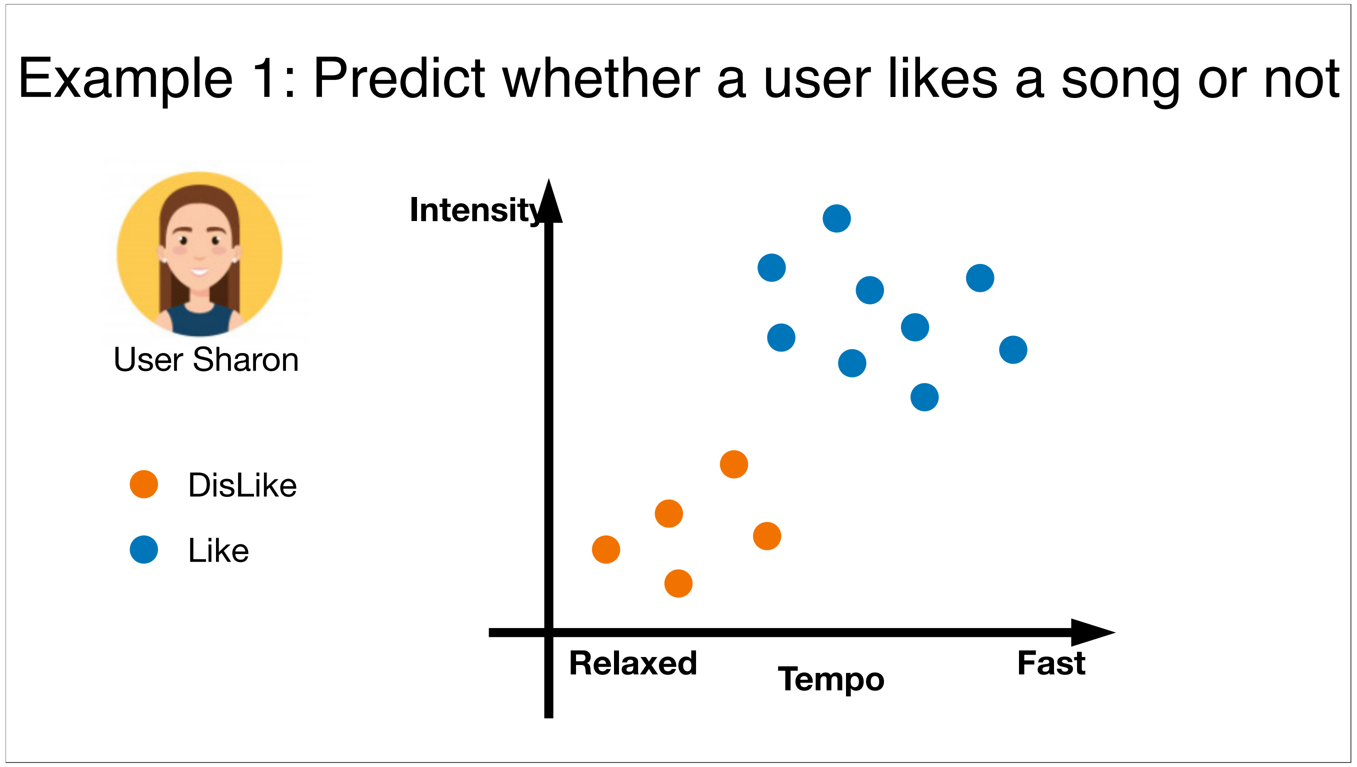
Part II: Supervised Learning

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



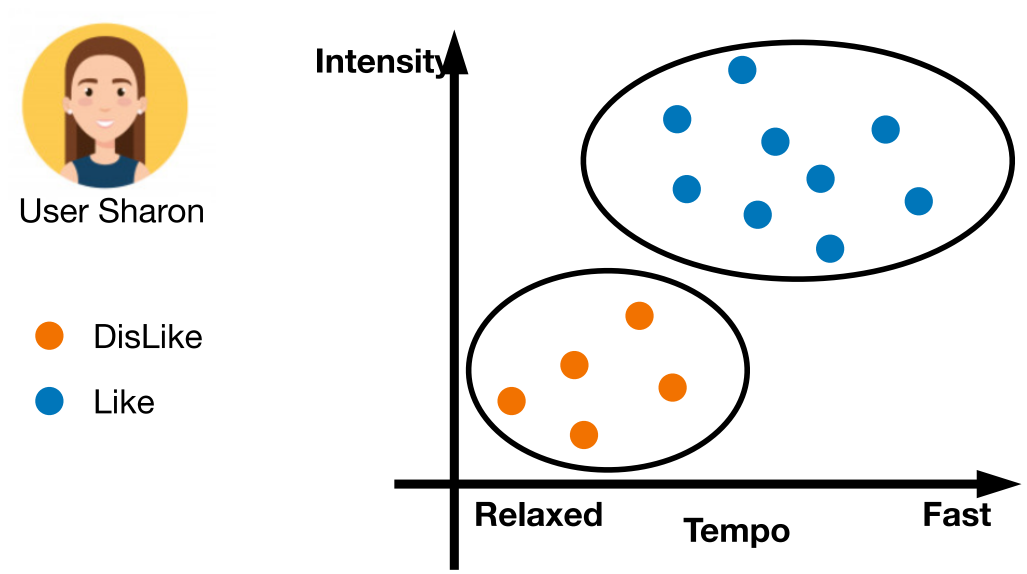


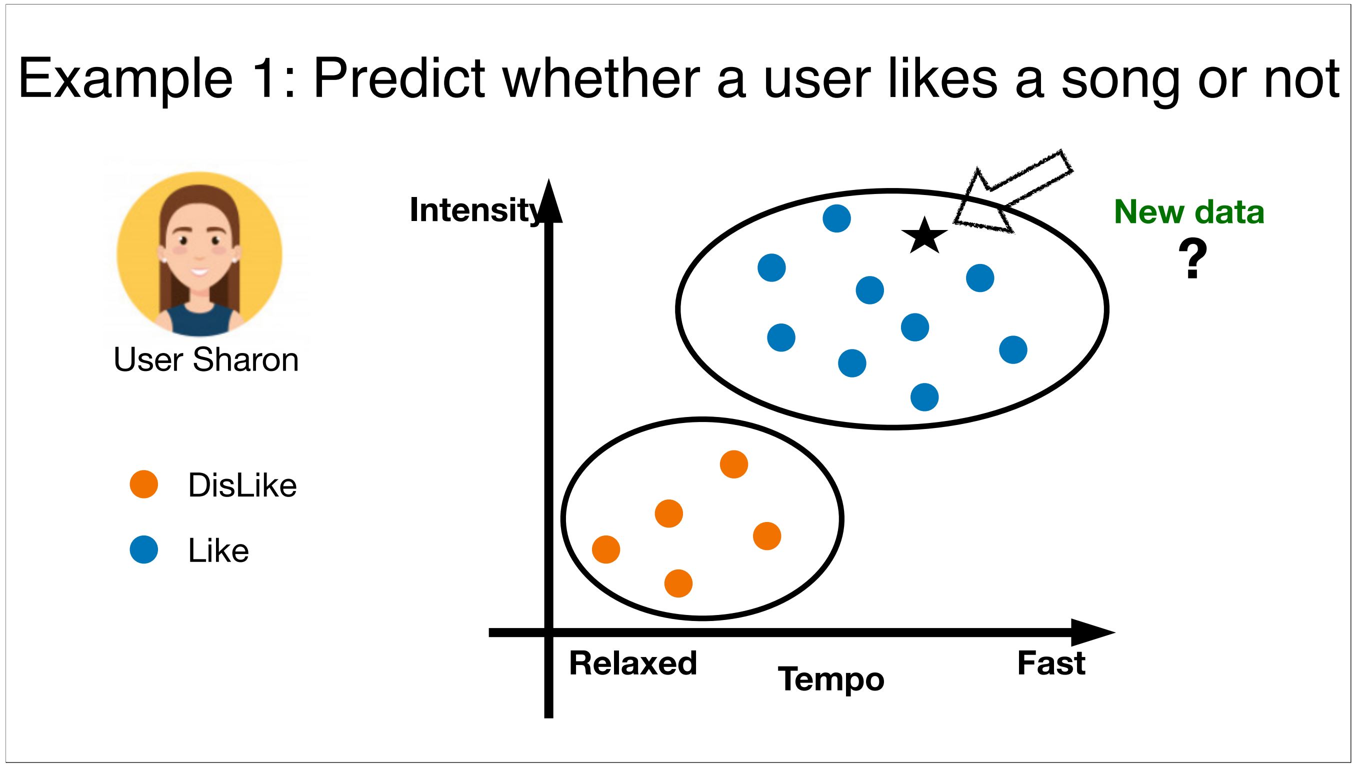
In this running example, we represent each song as a two-dimensional vector. This is called the feature vector, whose dimensions are called features. Here we have two features: intensity and tempo.



We have a few songs together with the label whether the user likes or dislike the song.

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



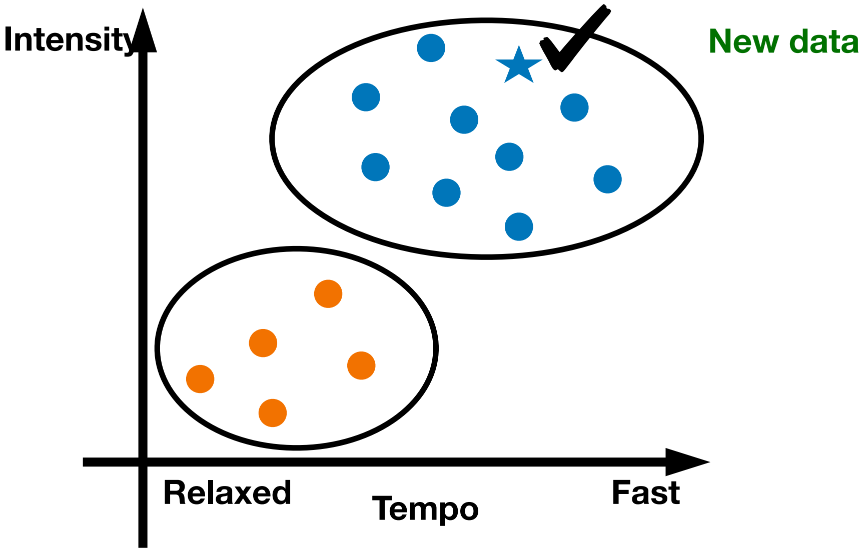


Suppose in the future, the AI system sees a new song. What should be the predicted label (like vs dislike)?

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

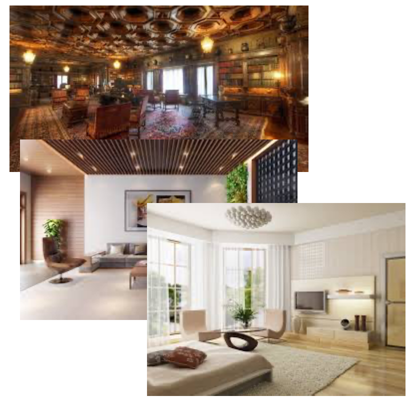
User Sharon

- DisLike
- Like

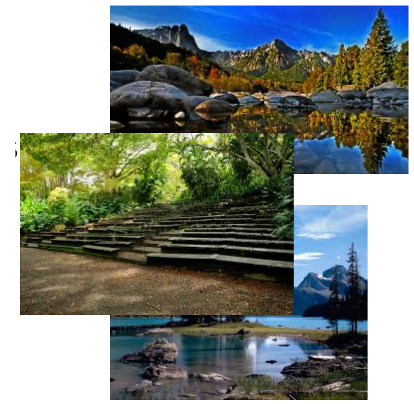


Many of you probably heard about ImageNet, which has become most popular and accepted benchmark dataset for image classification. Which contains over 1M images across over 1k categories. This was collected and released by Fei-fei Li's group at Stanford in 2009.

Example 2: Classify Images

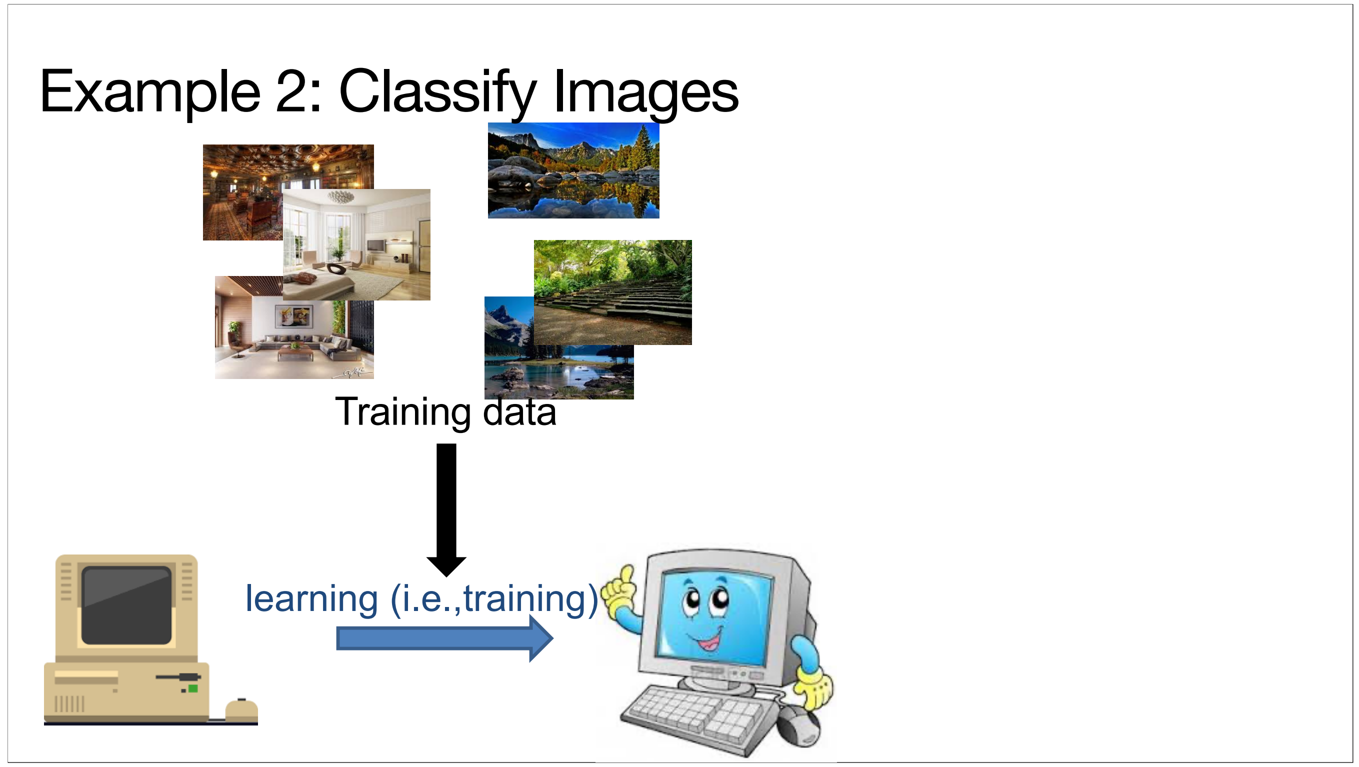


indoor

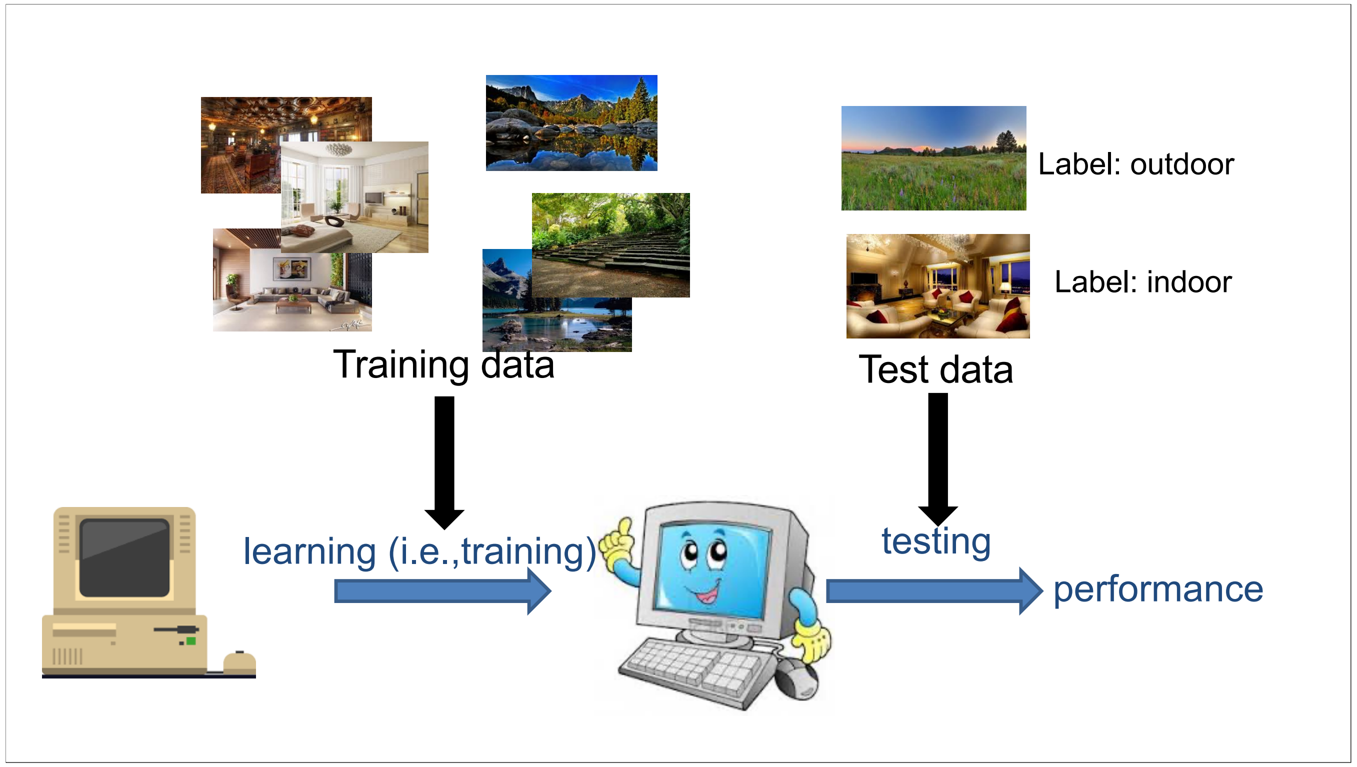


outdoor

Suppose we are given some indoor or outdoor images together with their indoor/outdoor labels.



We given these training data to the AI system, which gets improved.



In the test time, we get some new images (with hidden ground-truth labels indoor/outdoor). The AI system gives predictions on these test images. If we have the ground-truth labels, then we can compare the predictions to the ground-truth labels (i.e., evaluate the performance).

How to represent data?

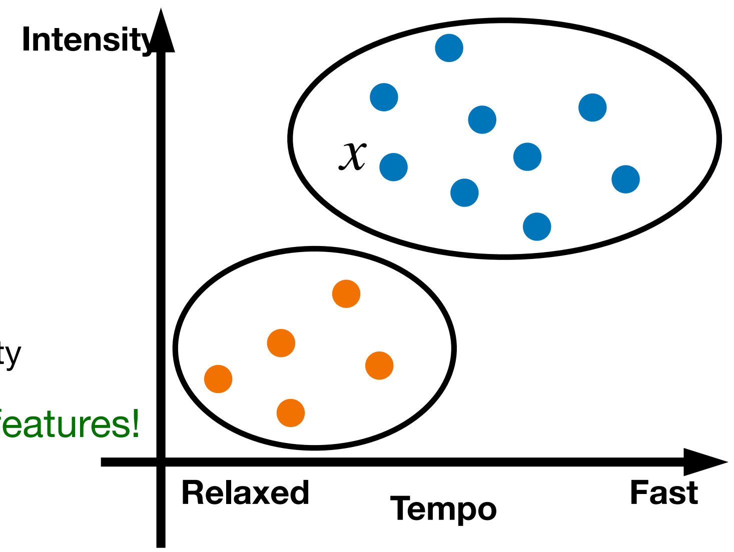
input data

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

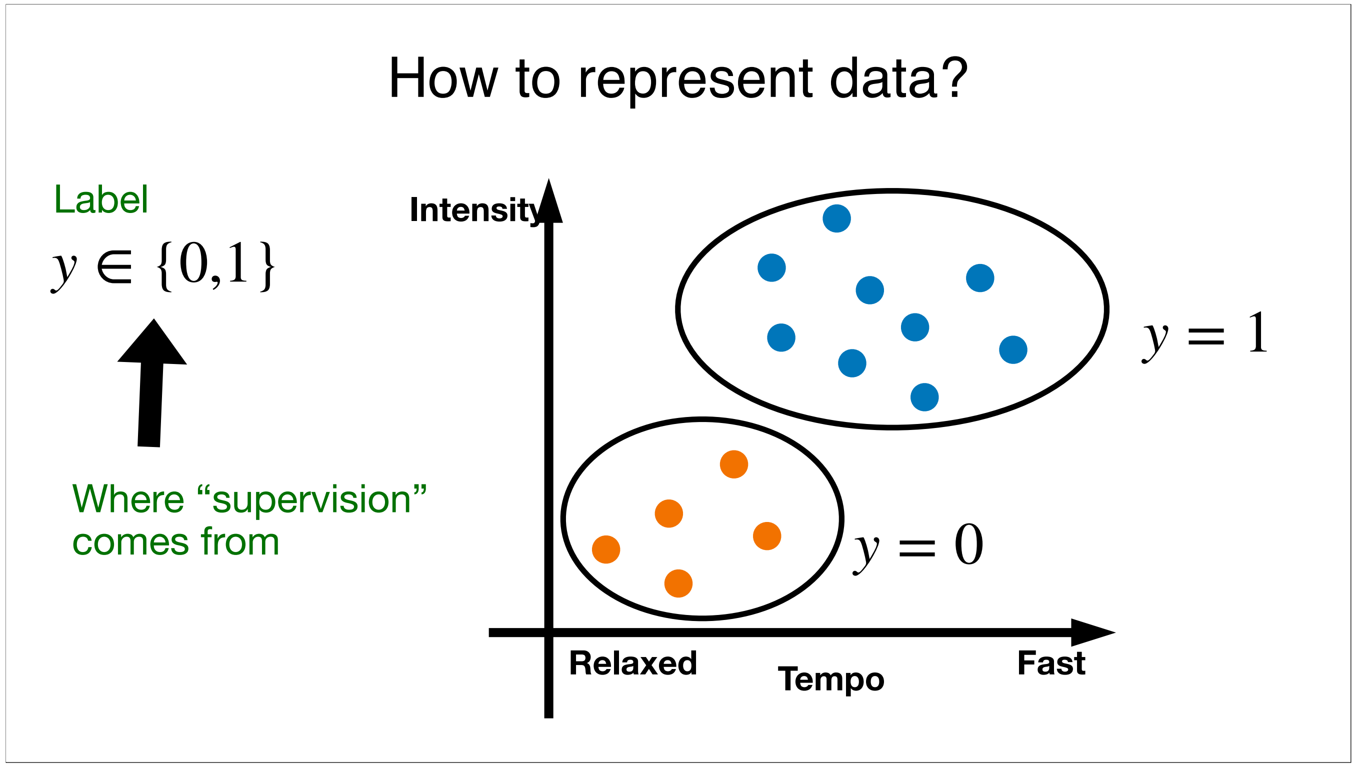
d: feature dimension

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

There can be many features!



We typically represent the input data as feature vectors. Each dimension is called a feature.



The label is also represented as some numbers. Usually, for two classes (e.g., like and dislike), we represent the label as {0,1} (sometimes as {-1, +1}). For multiple classes, we use {1,2,…,K} where K is the total number of classes. If the label has continuous values, we simply use those numeric values.

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**

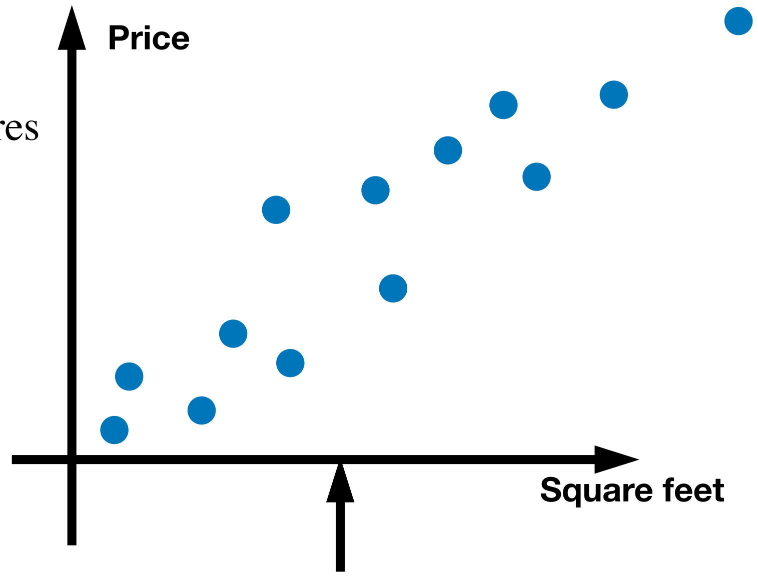
Classification: discrete finite values for the label
Regression: continuous values for the label

Example of regression: housing price prediction

Given: a dataset that contains *n* samples

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

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



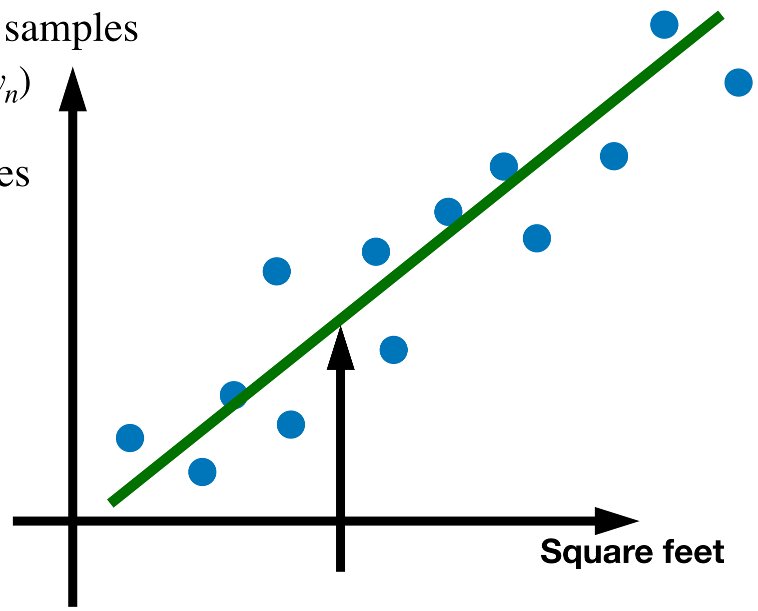
Example of regression: housing price prediction

Given: a dataset that contains n samples

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

Task: if a residence has $\mathbf{x}$ squares feet, predict the price?

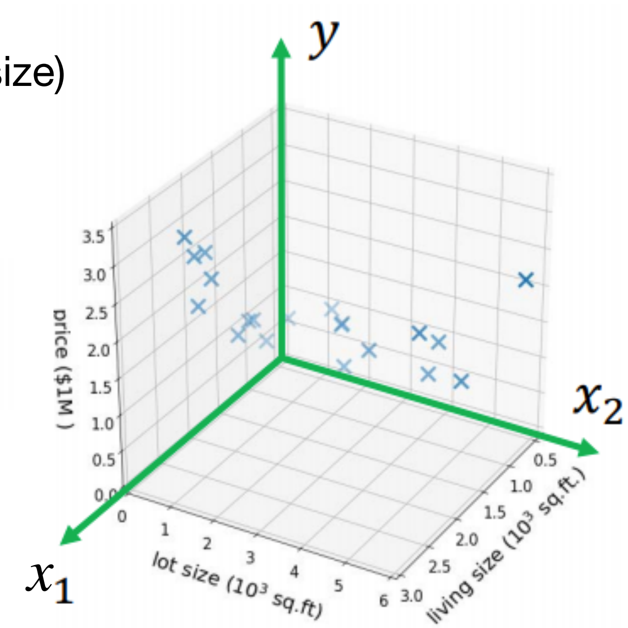
$y \in \mathbb{R}$



Example of regression: housing price prediction

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

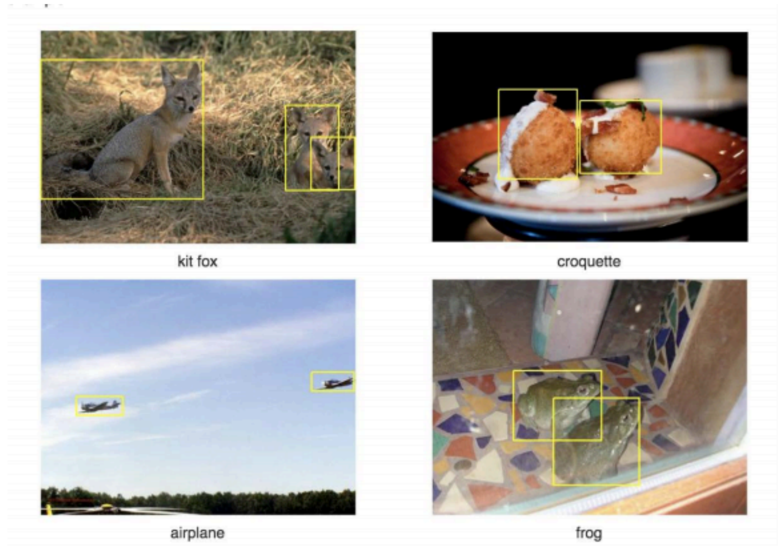
$$\underbrace{\text{(size, lot size)}}_{\substack{\text{features/input} \\ x \in \mathbb{R}^2}} \rightarrow \underbrace{\text{price}}_{\substack{\text{label/output} \\ y \in \mathbb{R}}}$$



(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

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

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:

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

input label

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

Goal of Supervised Learning

Given training data

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

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

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

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

Q1-1: 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

A. Feature vectors can be in high dimen.
B. Some feature vectors can have other types of values like strings
D. Bag-of-words is a type of feature vector for text

Quiz Break

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

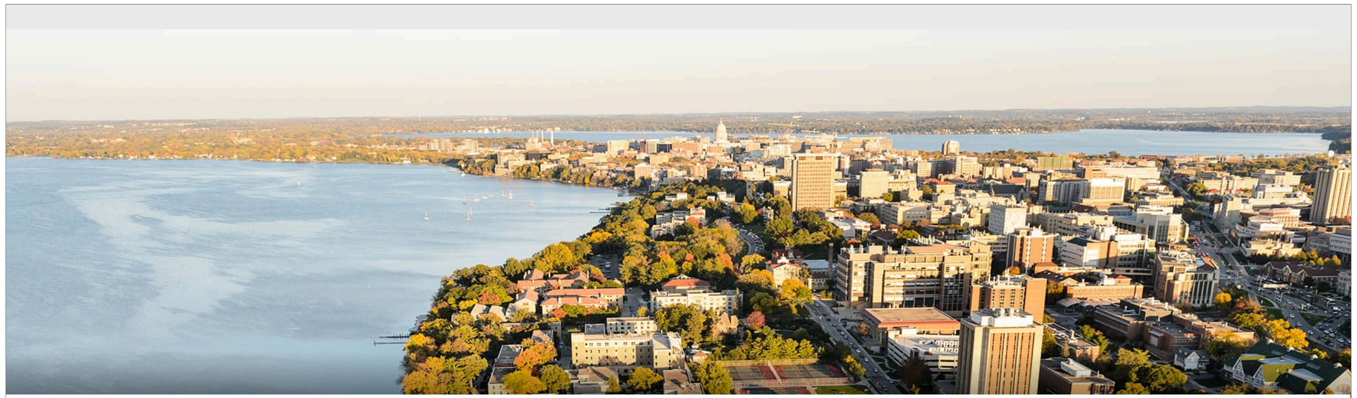
A. Object detection (predicting bounding box from raw images)
B. Classification
C. Regression
D. Dimensionality reduction

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

Dimensionality reduction is unsupervised since it has no labeling information.



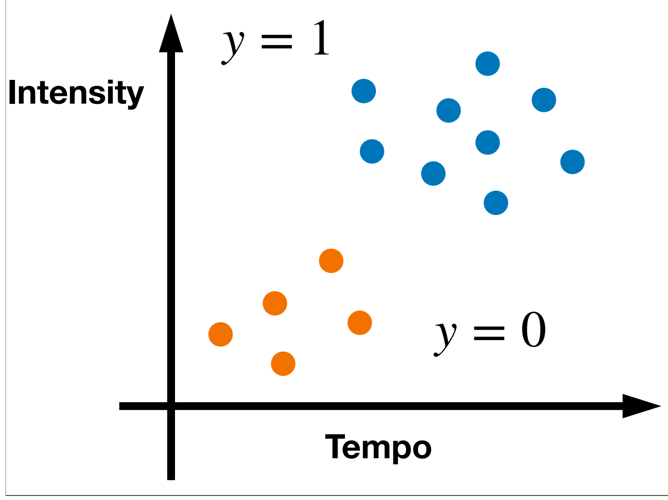
Part II: Unsupervised Learning
(no teacher)

Unsupervised Learning

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

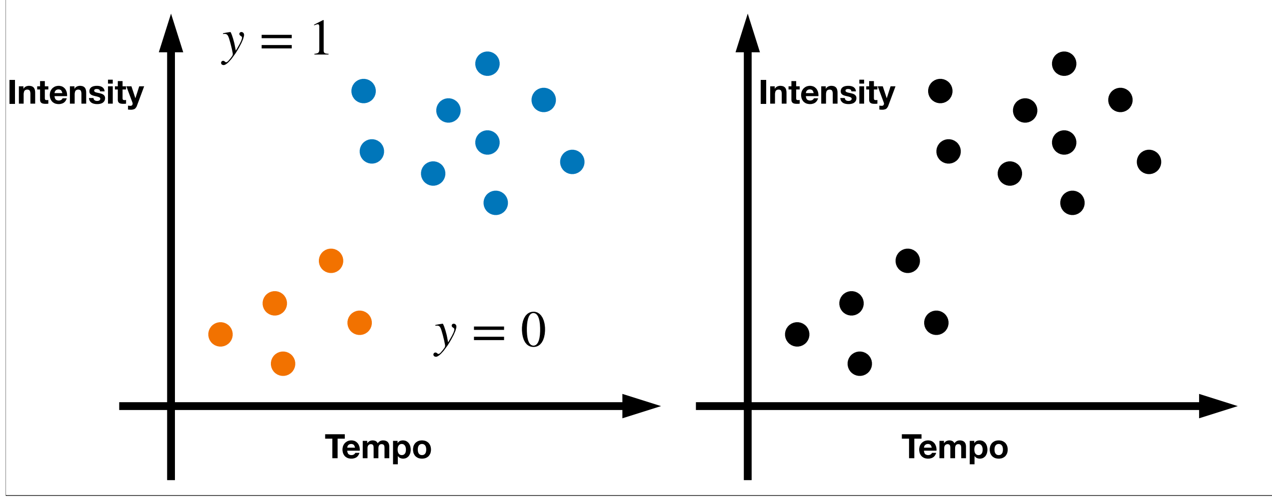
Unsupervised Learning

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



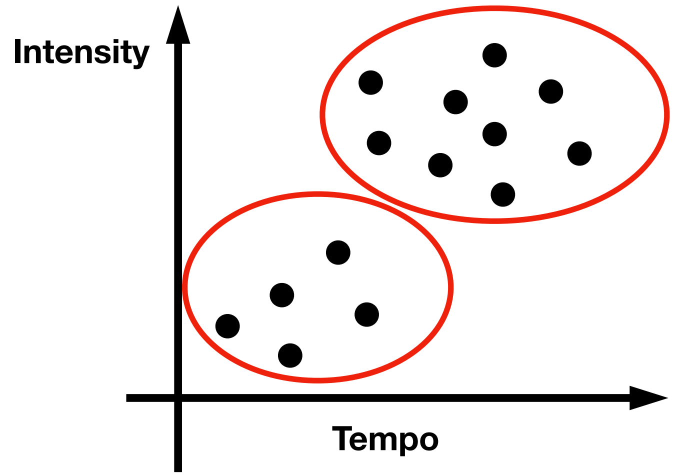
Unsupervised Learning

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

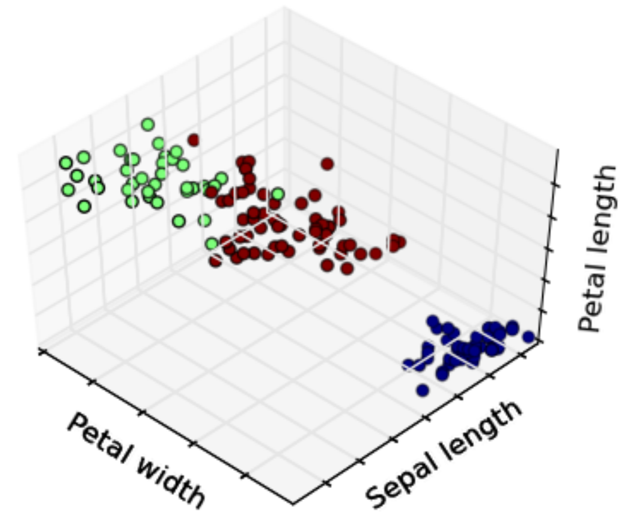


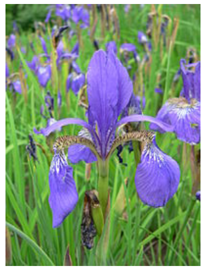
Clustering

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



Clustering



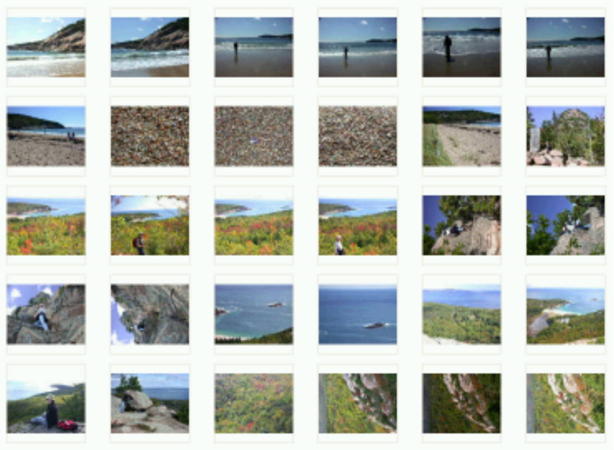


Clustering Irises using three different features

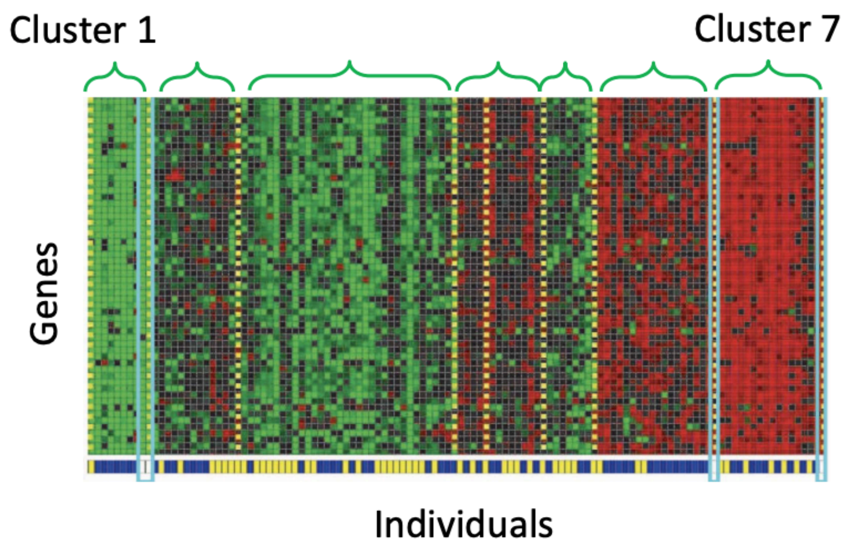
The colors represent clusters identified by the algorithm, not y's provided as input

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)

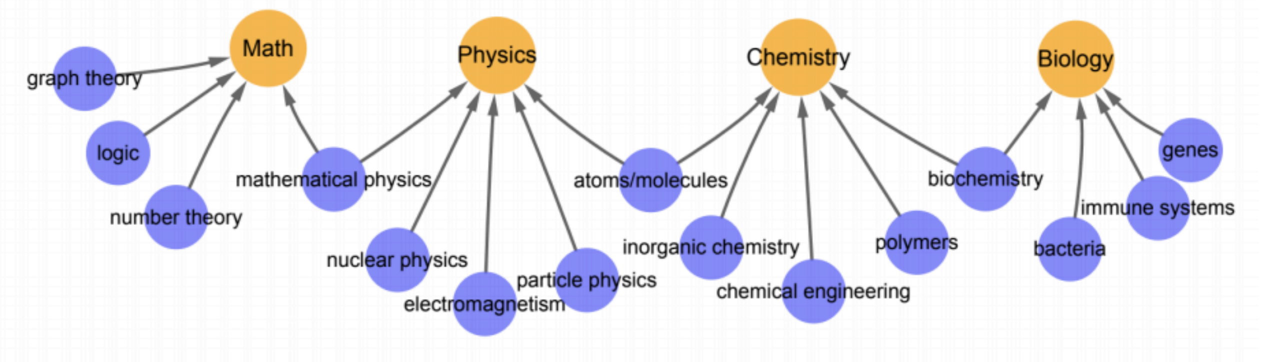


Clustering Genes



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

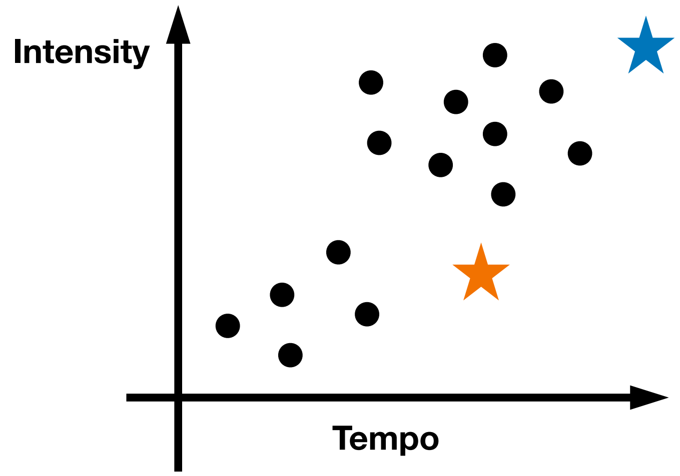
K-means clustering

- Very popular clustering method
- Don’t confuse it with k-NN classifier
- Input: a dataset $x_1, x_2, \ldots, x_n$, and assume the number of clusters **k** is given

Note that the number of clusters k is given as part of the input.

K-means clustering

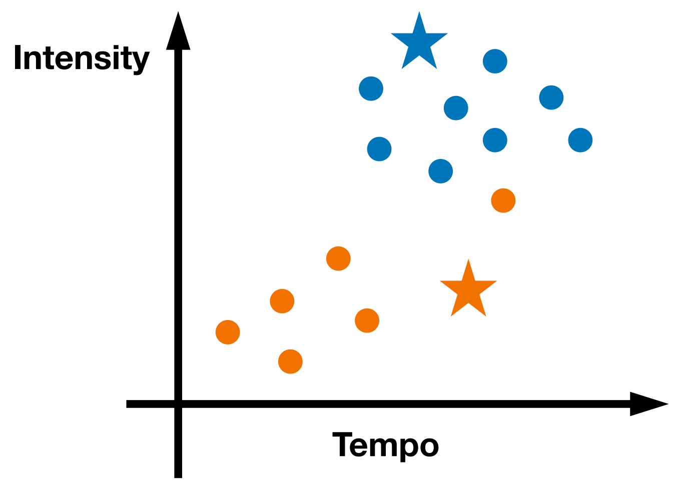
Step 1: **Randomly** picking 2 positions as initial cluster centers (not necessarily a data point)



Initialization: randomly pick initial centers.

K-means clustering

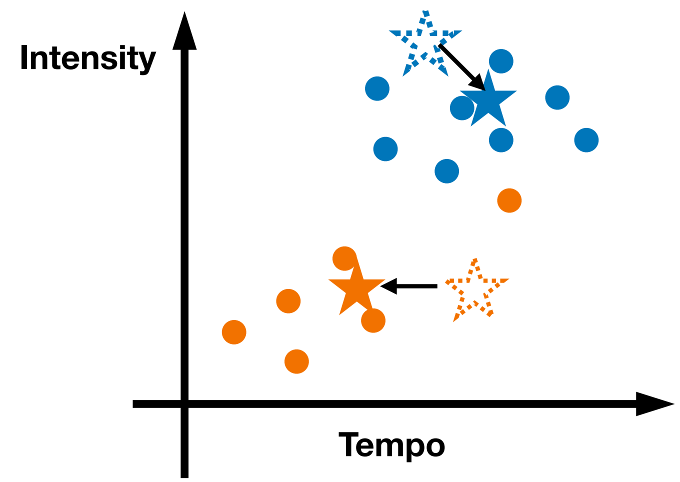
Step 2: for each point x, determine its cluster: find the closest center in Euclidean space



Then get into iterations.
In each iteration:
First assign each point to its closest cluster center.

K-means clustering

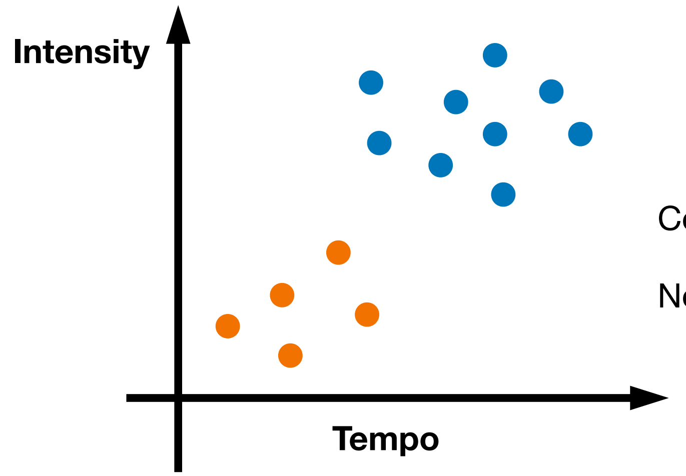
Step 3: update all cluster centers as the centroids



Then update each cluster center to the average of the points in the cluster.

K-means clustering

Repeat step 2 & 3 until convergence



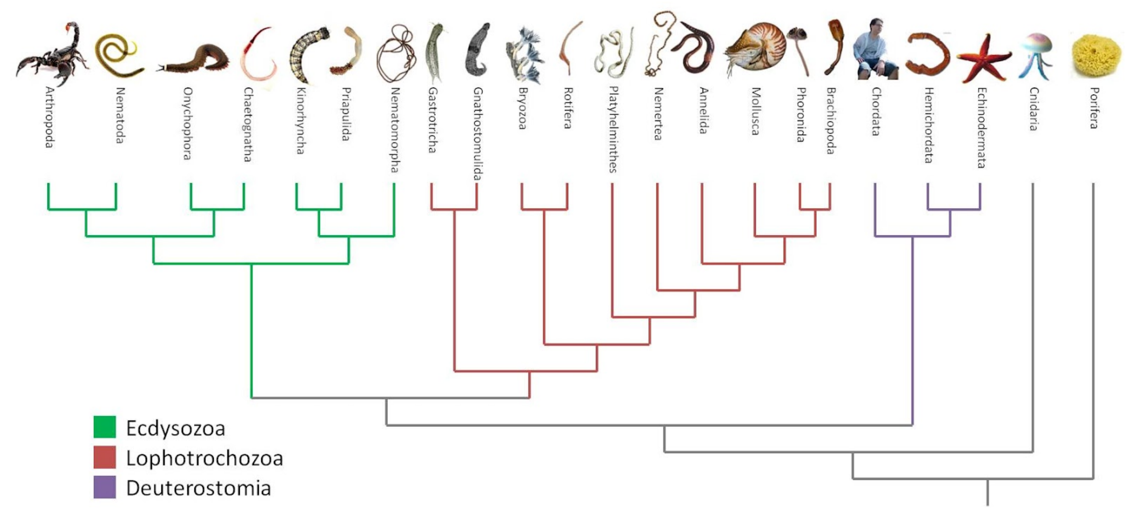
Converged solution!

No labels required!

K-means clustering: A demo

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

Hierarchical Clustering (more to follow next lecture)



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

Q2-1: 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

A. The labels are human inputs
B. The machine should not have test data for learning
C. No labels available for clustering

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

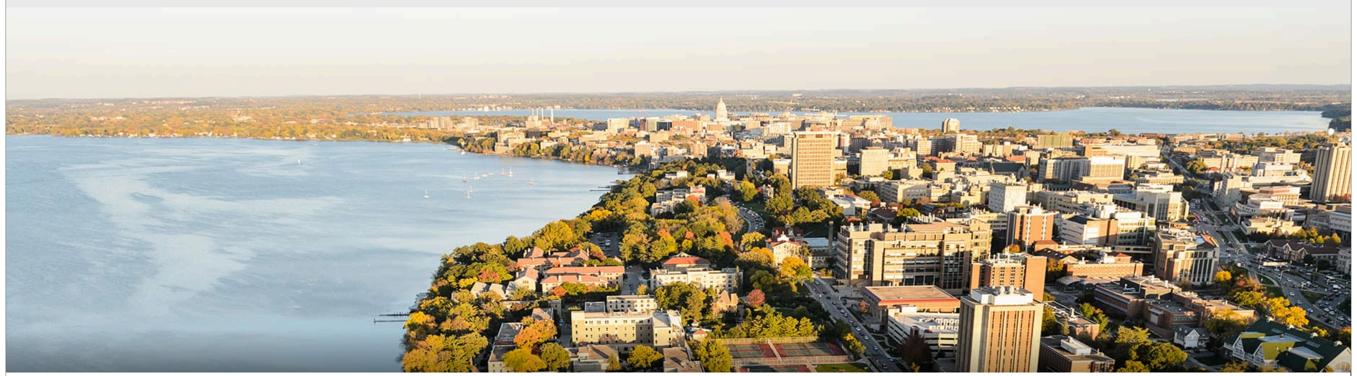
Quiz Break

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

There are many unsupervised learning methods.

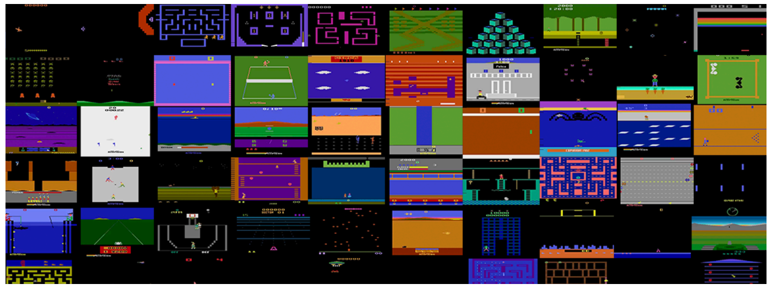
Kmeans is not producing a hierarchy but a partition of the points into clusters. It needs the number of clusters as part of the input.



Part III: Reinforcement Learning
(Learn from reward)


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.



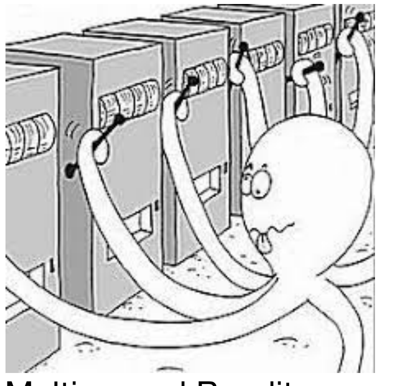
Google Deepmind

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."

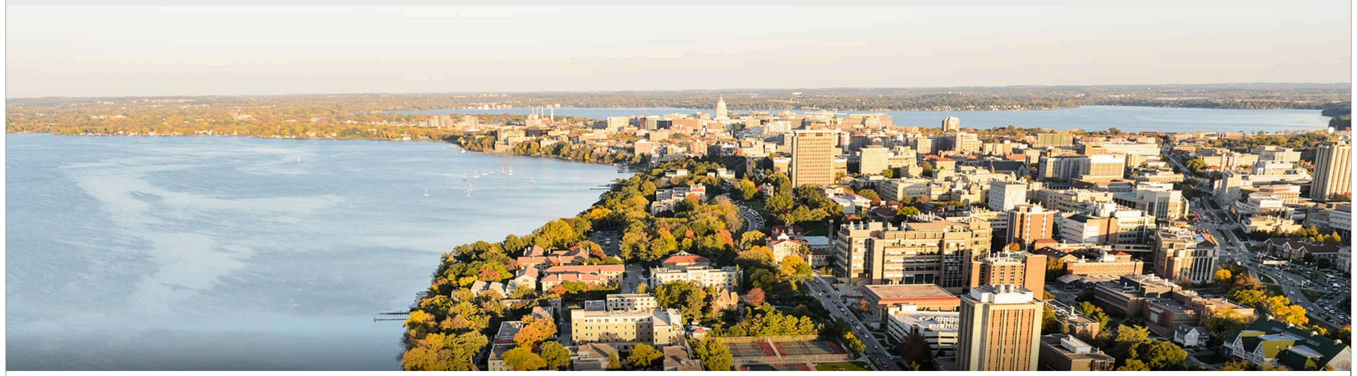
- Peter Whittle



Multi-armed Bandit

Today's recap

- What is machine learning?
- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
- Reinforcement Learning



Thanks!