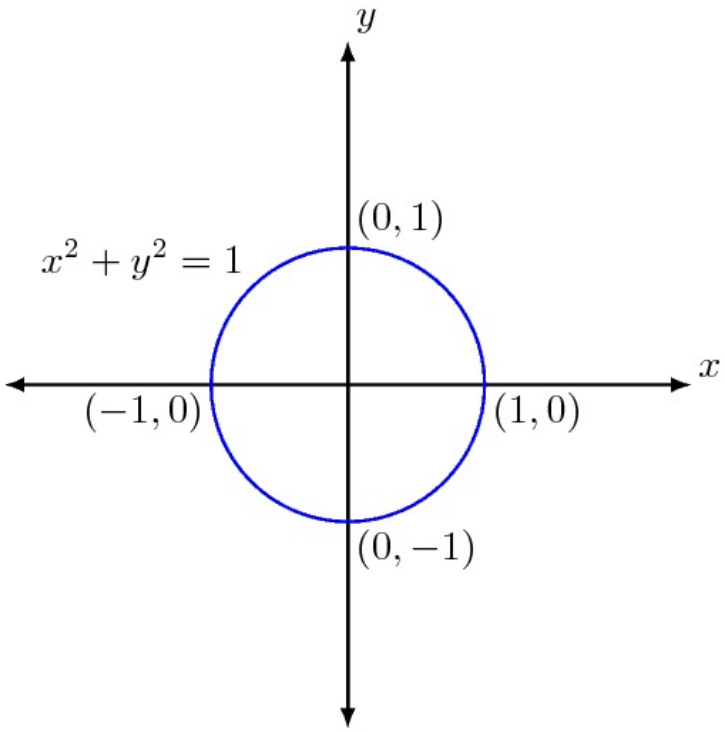
IOWA STATE UNIVERSITY
Department of Computer Science

COM S 573: Machine Learning

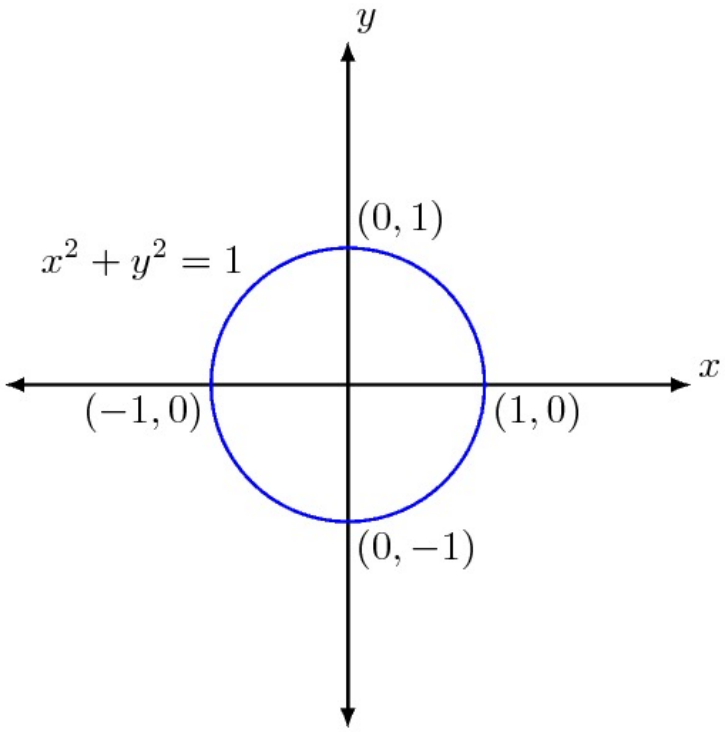
Lecture 2: Machine Learning Introduction

How to Define a Circle?



How to Define a Circle?

- We need the center, radius

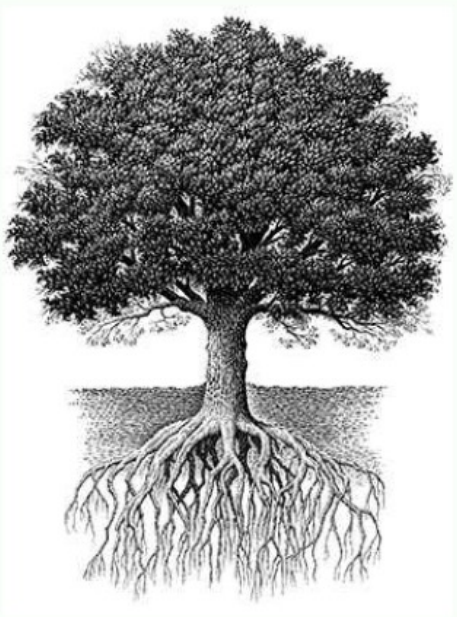


How to Define a Tree?

- A brown trunk moving upwards and branching with leaves …

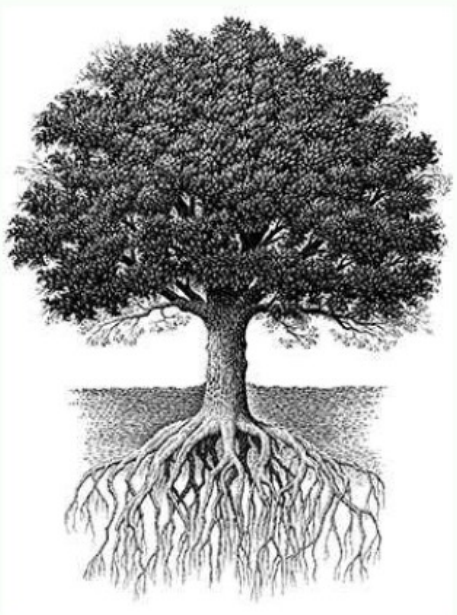
Defining is Hard; Recognizing is Easy

- Hard to give a mathematical definition of a tree
- Even a 3-year-old can tell a tree from a non-tree
- Why?



Defining is Hard; Recognizing is Easy

- Hard to give a mathematical definition of a tree
- Even a 3-year-old can tell a tree from a non-tree
- The 3-year-old has learned from data



What is Machine Learning?

- The goal of the field of machine learning is to build computer systems that learn from experience and that are capable to adapt to their environments.

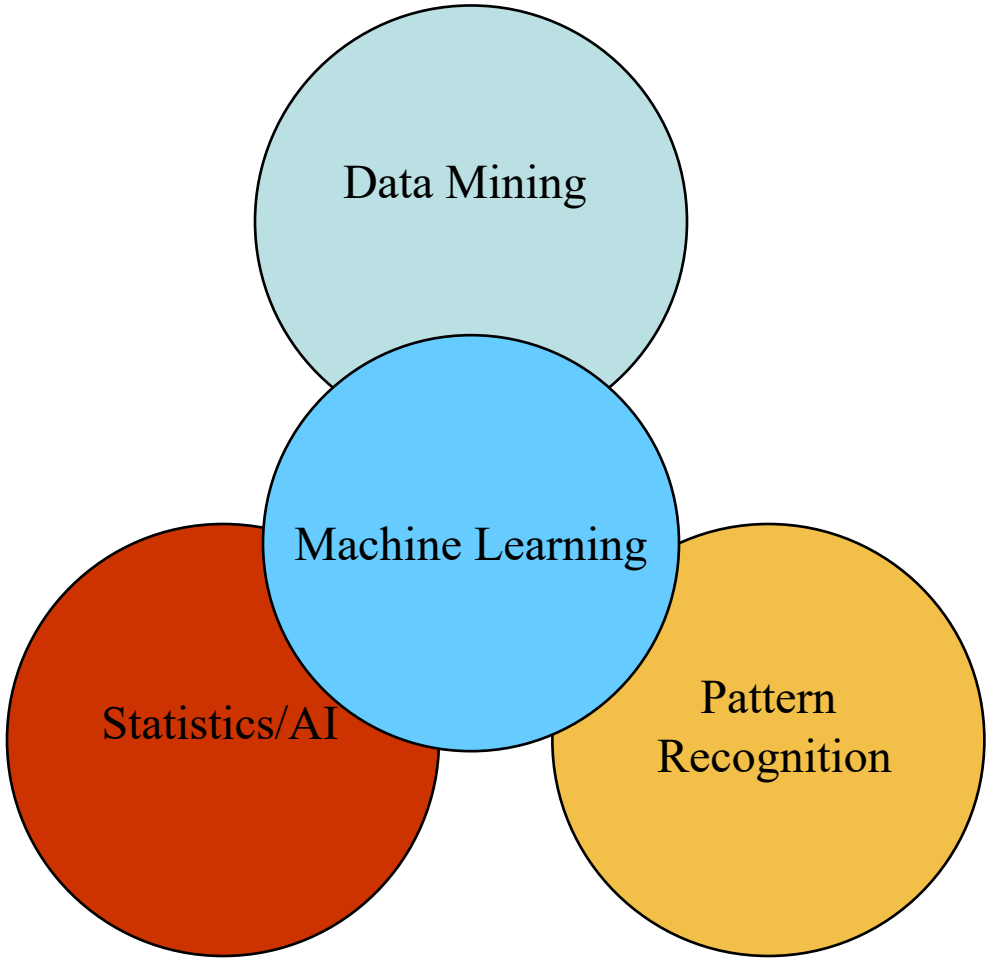
- Learning techniques and methods developed in this field have been successfully applied to a variety of applications.

- Text classification, protein function prediction, financial forecasting, credit card fraud detection.

Machine Learning Applications

- Computer vision
- Natural language processing
- Search engines
- Medical diagnosis
- Bioinformatics
- Detecting credit card fraud
- Speech and handwriting recognition
- …

Related Fields



Machine Learning Tasks

- Supervised learning: (x, y)
  - Classification
  - Regression
- Unsupervised learning (x)
  - Clustering
  - Density estimation
  - Visualization (dimensionality reduction)
- Semi-supervised learning
- …

Classification: Definition

- Given a collection of records (training set)
  - Each record contains a set of attributes, one of the attributes is the class.
- Find a model for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A test set is used to determine the accuracy of the model.

Classification Example

Tid	Refund (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Cheat (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

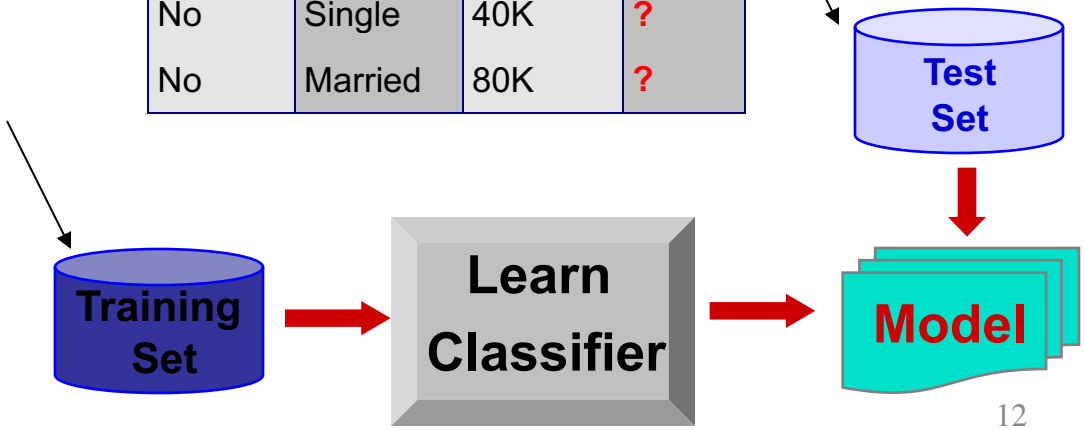
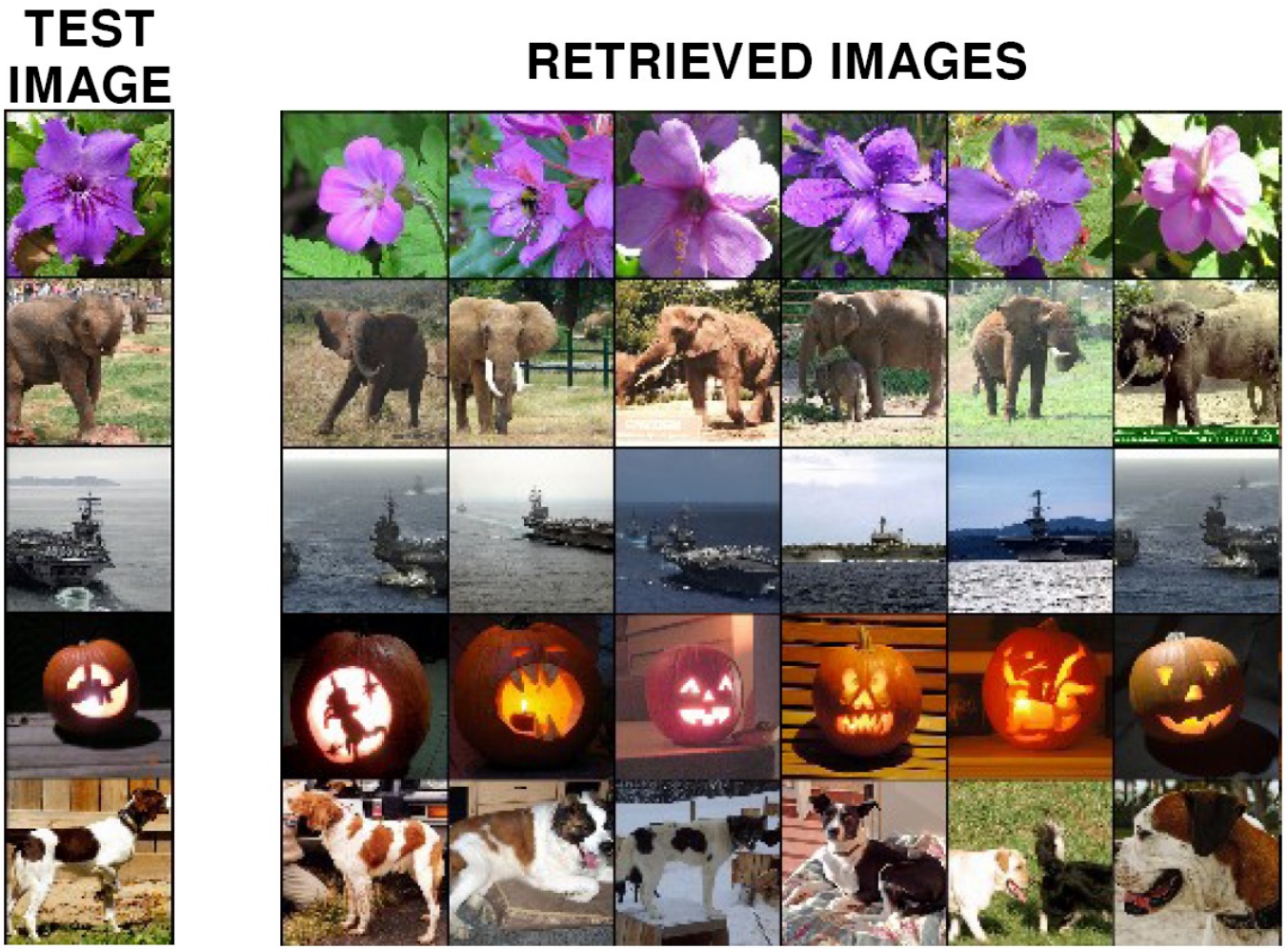


Image Retrieval

Classification Problem?

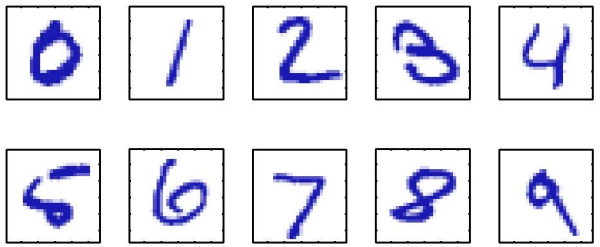


Face/Pedestrian Detection

Classification Problem?

Handwritten Digit Recognition

Classification Problem?



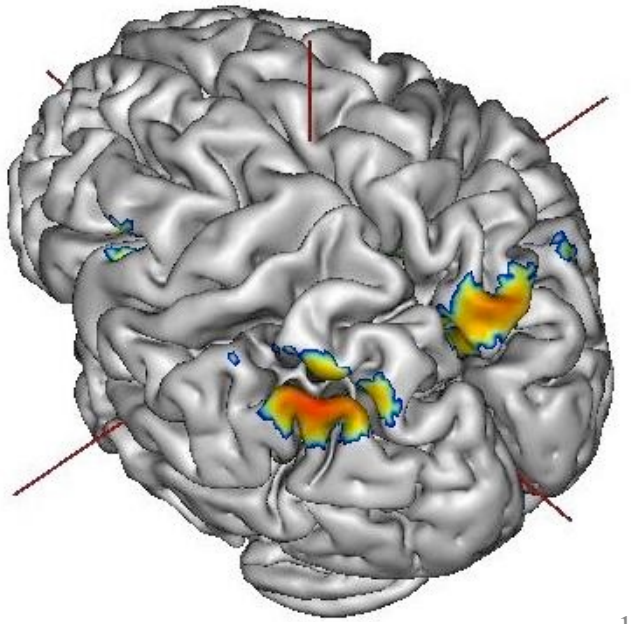
Handwritten Digit Recognition

Classification Problem?

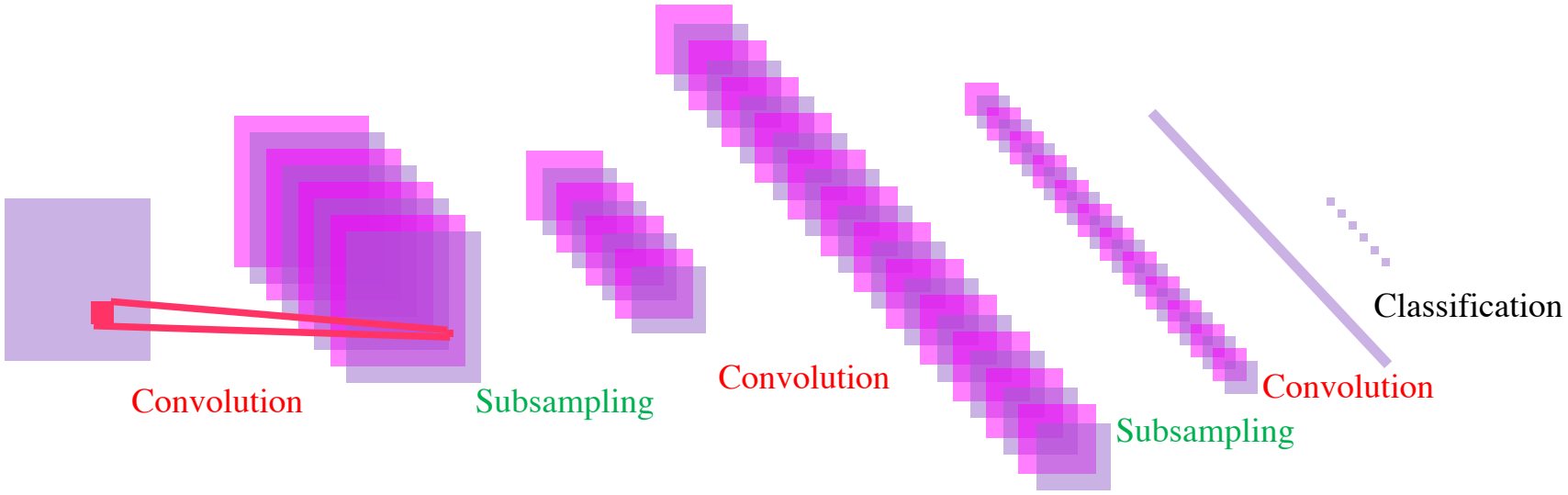
Medical Informatics

- Alzheimer's Disease Detection
  - Goal: To predict class (AD or normal) of a sample (person), based on neuroimaging data such as MRI and PET

Reduced gray matter volume (colored areas) detected by MRI voxel-based morphometry in AD patients compared to normal healthy controls.



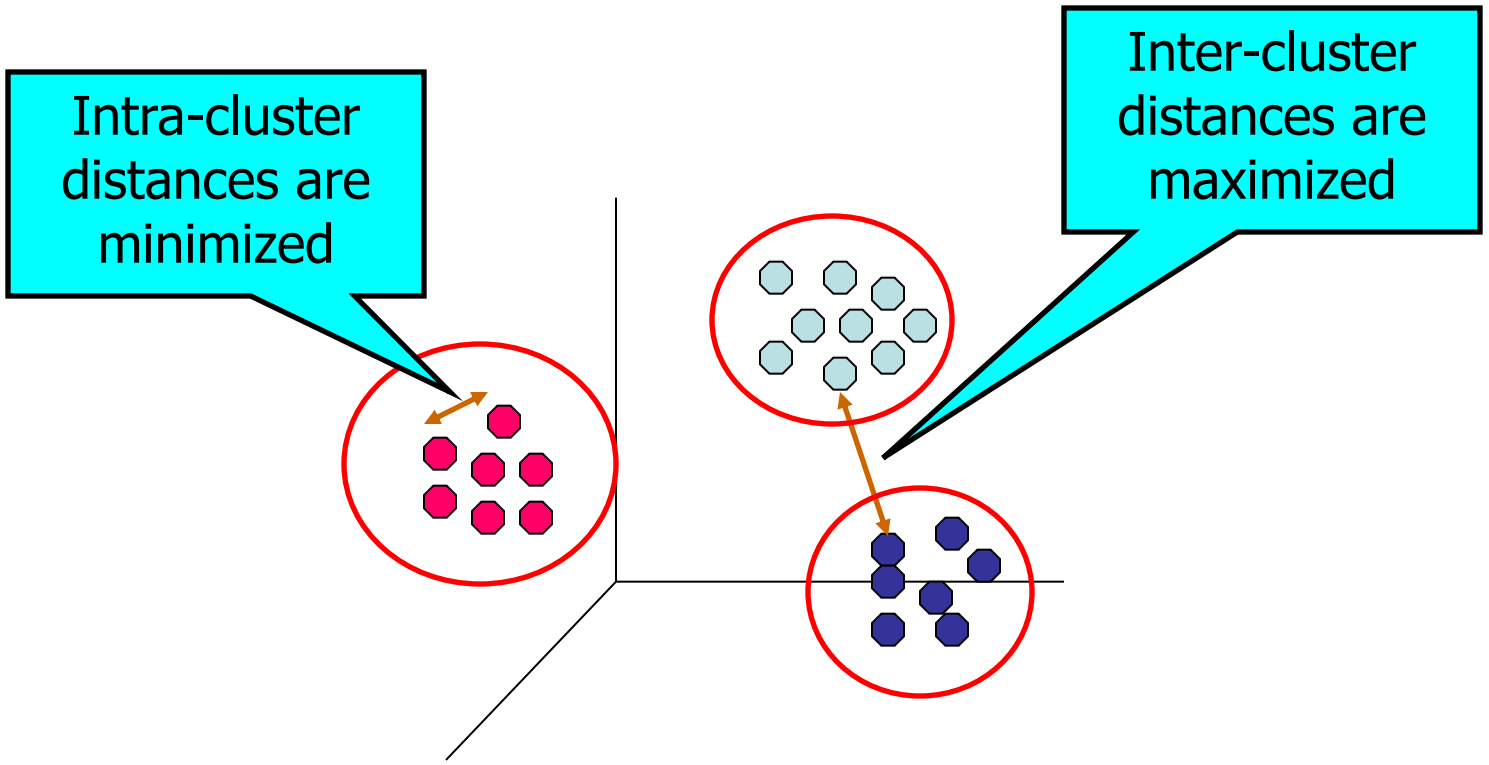
Deep learning



Clustering Definition

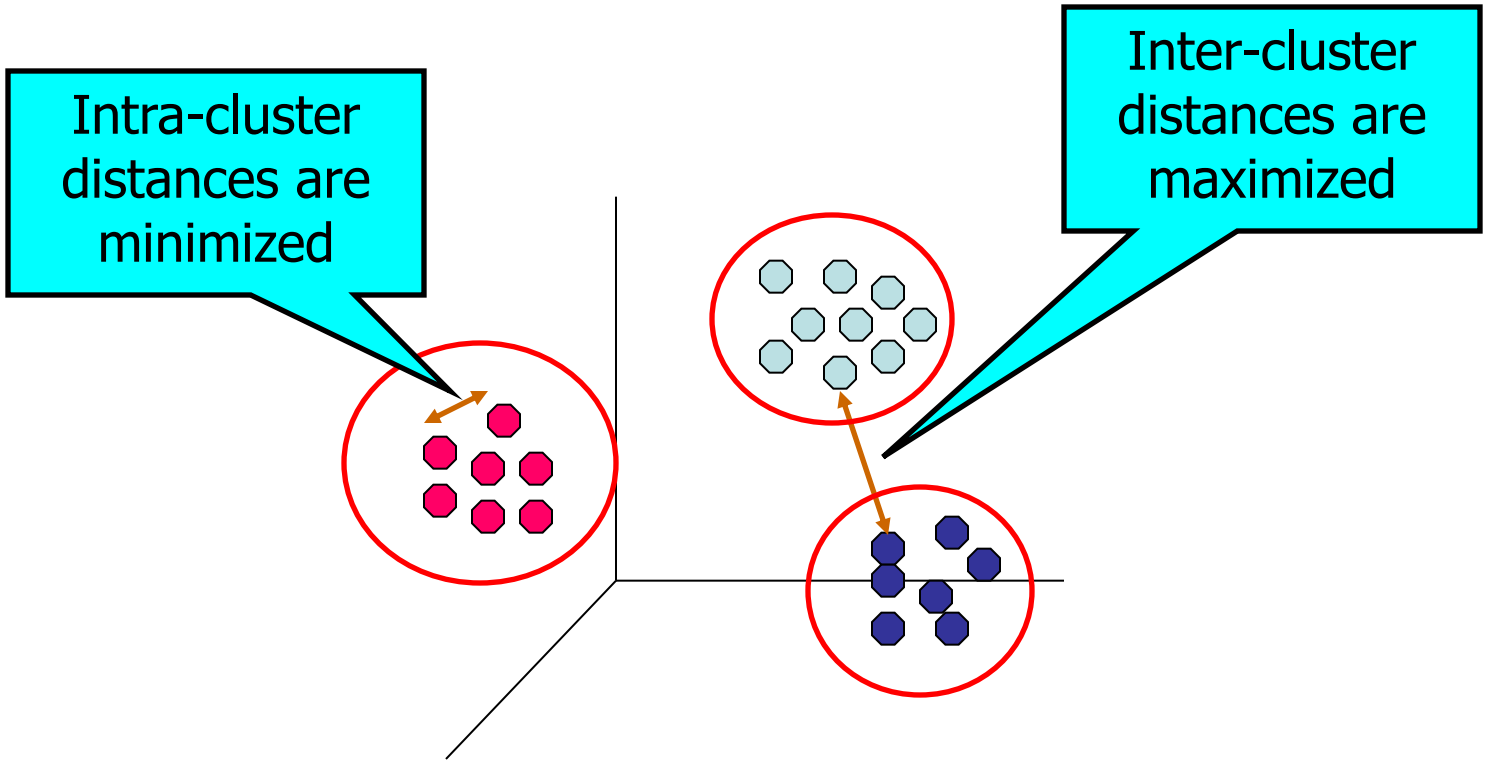
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Illustrating Clustering



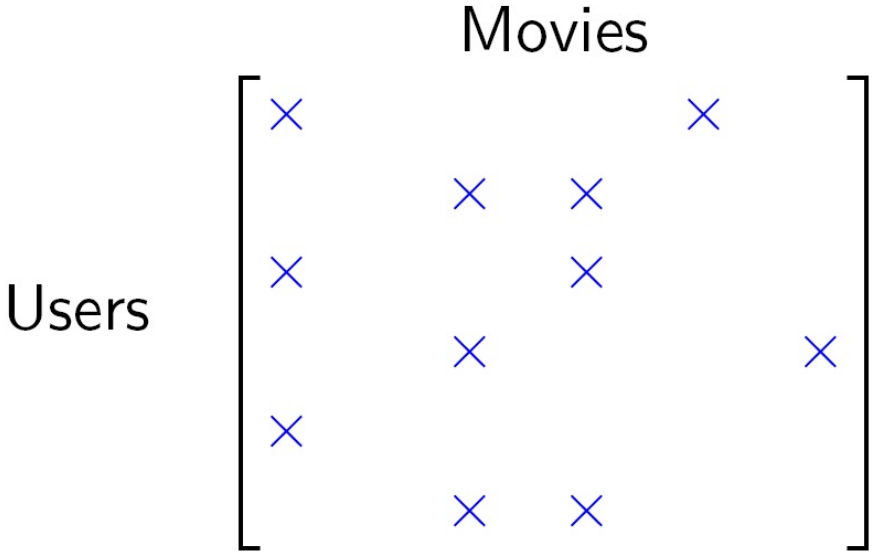
Illustrating Clustering

Why we need clustering?



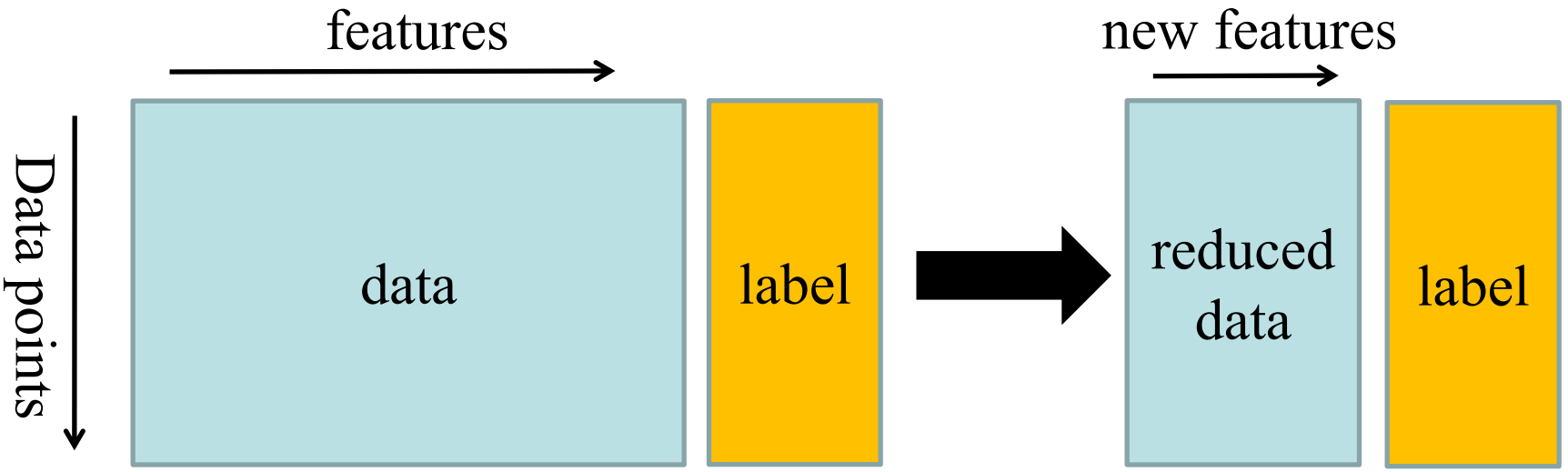
Collaborative Filtering

- Netflix database
  - About a million users
  - About 25,000 movies
- People rate movies



Dimensionality Reduction

- Dimensionality reduction extracts a small number of features by removing irrelevant, redundant, and noisy information
- Different from feature selection



Three Components of Learning

- Representation
  - Given an input x_i, representation produces a prediction $\hat{y}_i$
- Evaluation
  - Compute evaluations
  - Accuracy / Error rate, Likelihood
- Optimization
  - Optimize Representation

Three Components of Learning