Machine Learning
CS 6830

Lecture 01

Razvan C. Bunescu

School of Electrical Engineering and Computer Science

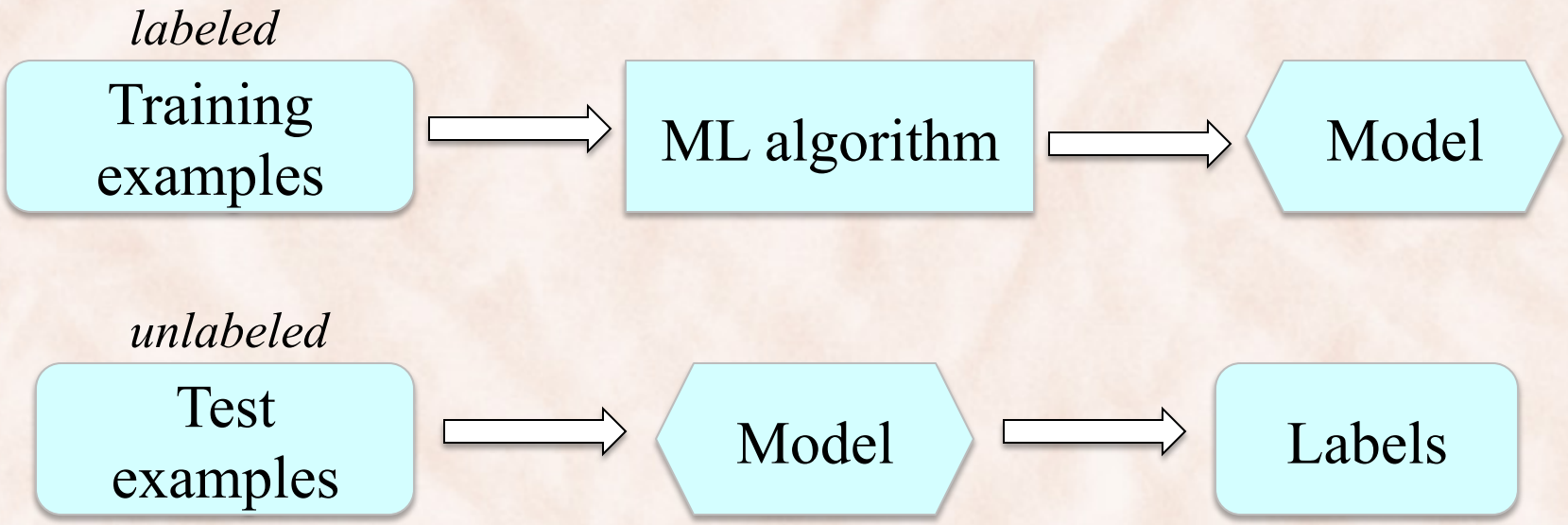
bunescu@ohio.edu

What is (Human) Learning?

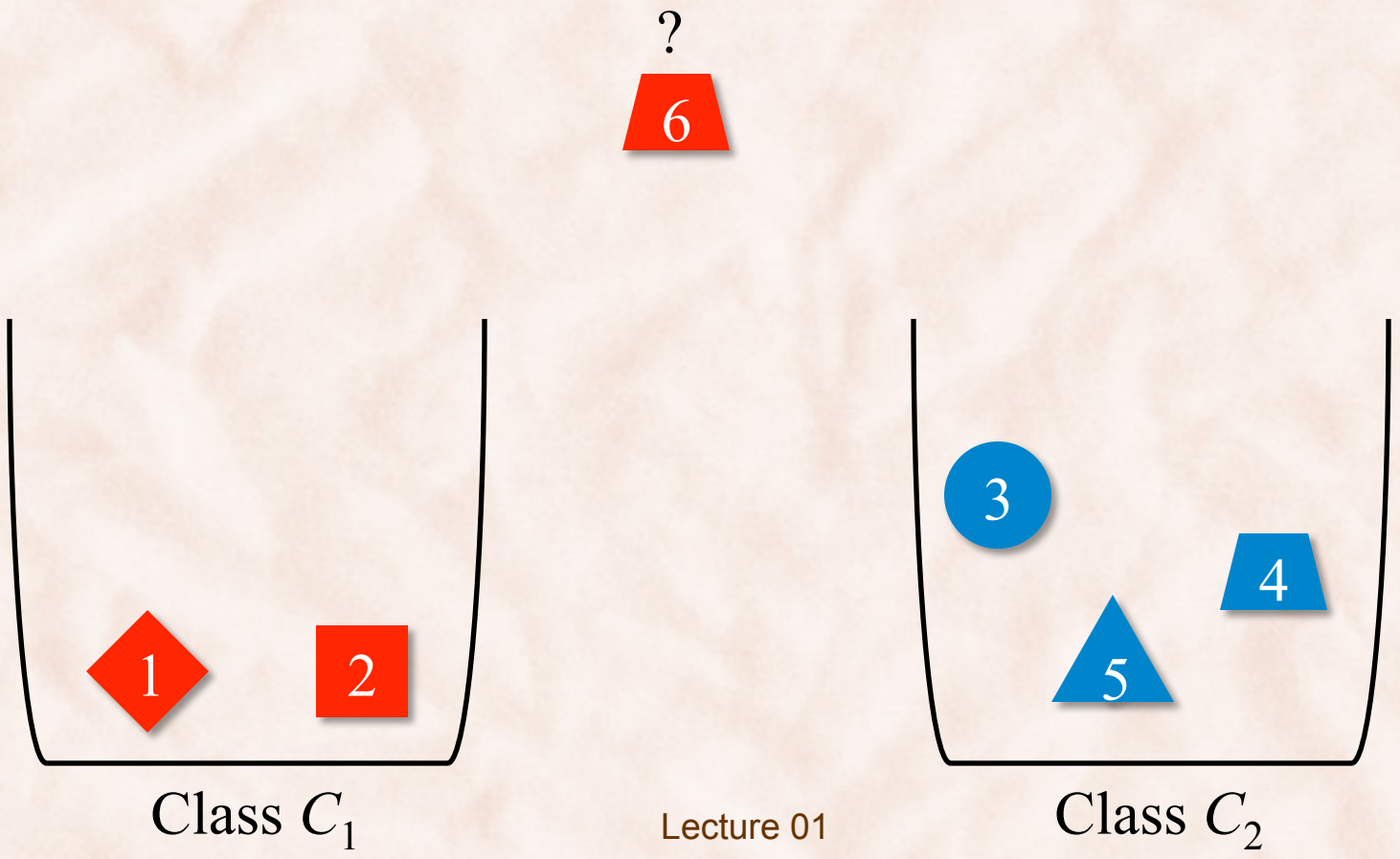
- Merriam-Webster:
  - ***learn*** = to acquire knowledge, understanding, or skill … by study, instruction, or ***experience***.
- Why do we learn?
  - to ***improve performance*** on a given task.
- What (tasks) do we learn:
  - read, translate, write, speak.
  - walk, play backgammon, ride bikes, drive cars, fly helicopters.
  - categorize email, recognize faces, diagnose diseases, …

What is Machine Learning?

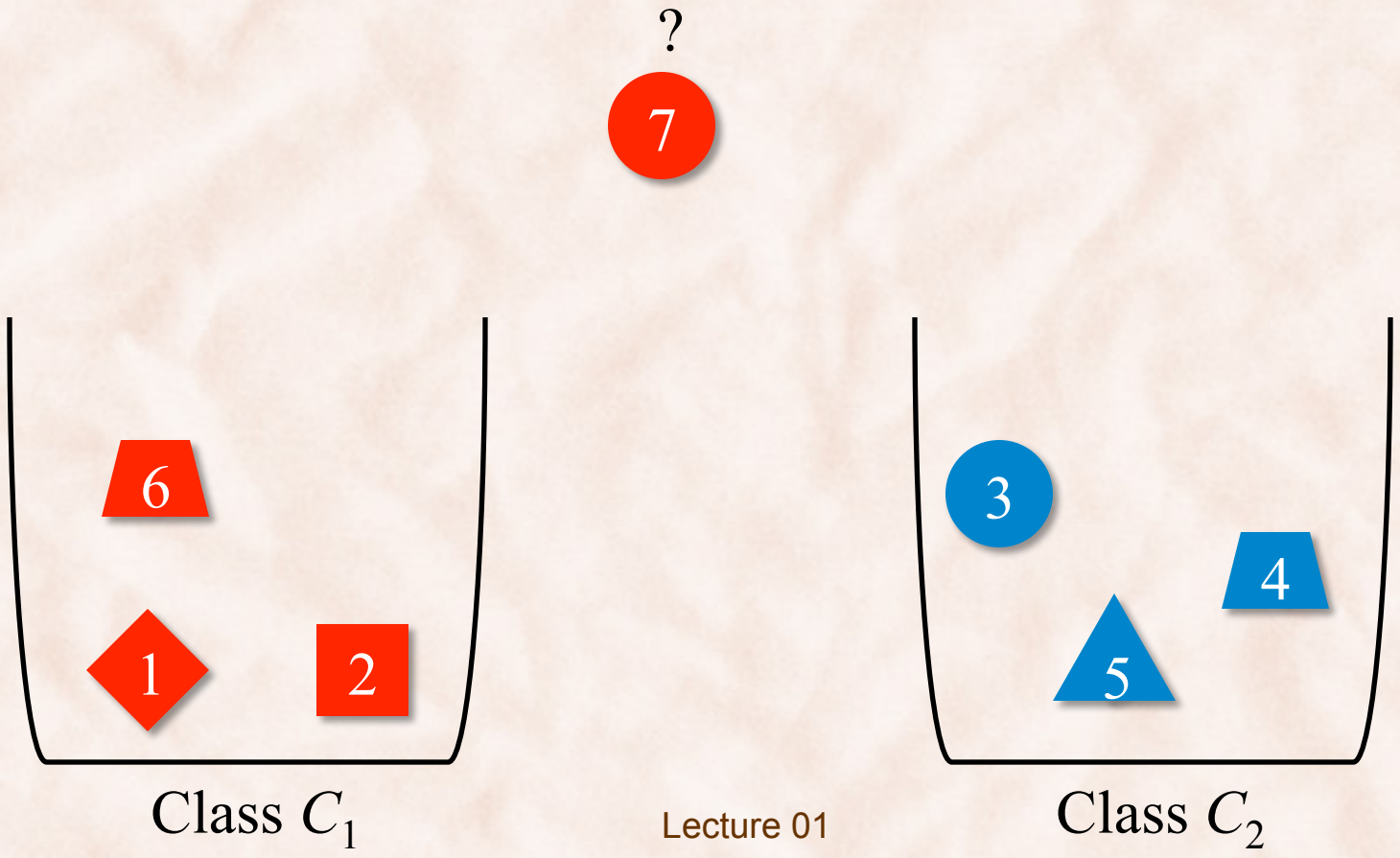
- **Machine Learning** = constructing computer programs that *automatically improve with experience*:
  - Supervised Learning i.e. learning from labeled examples.



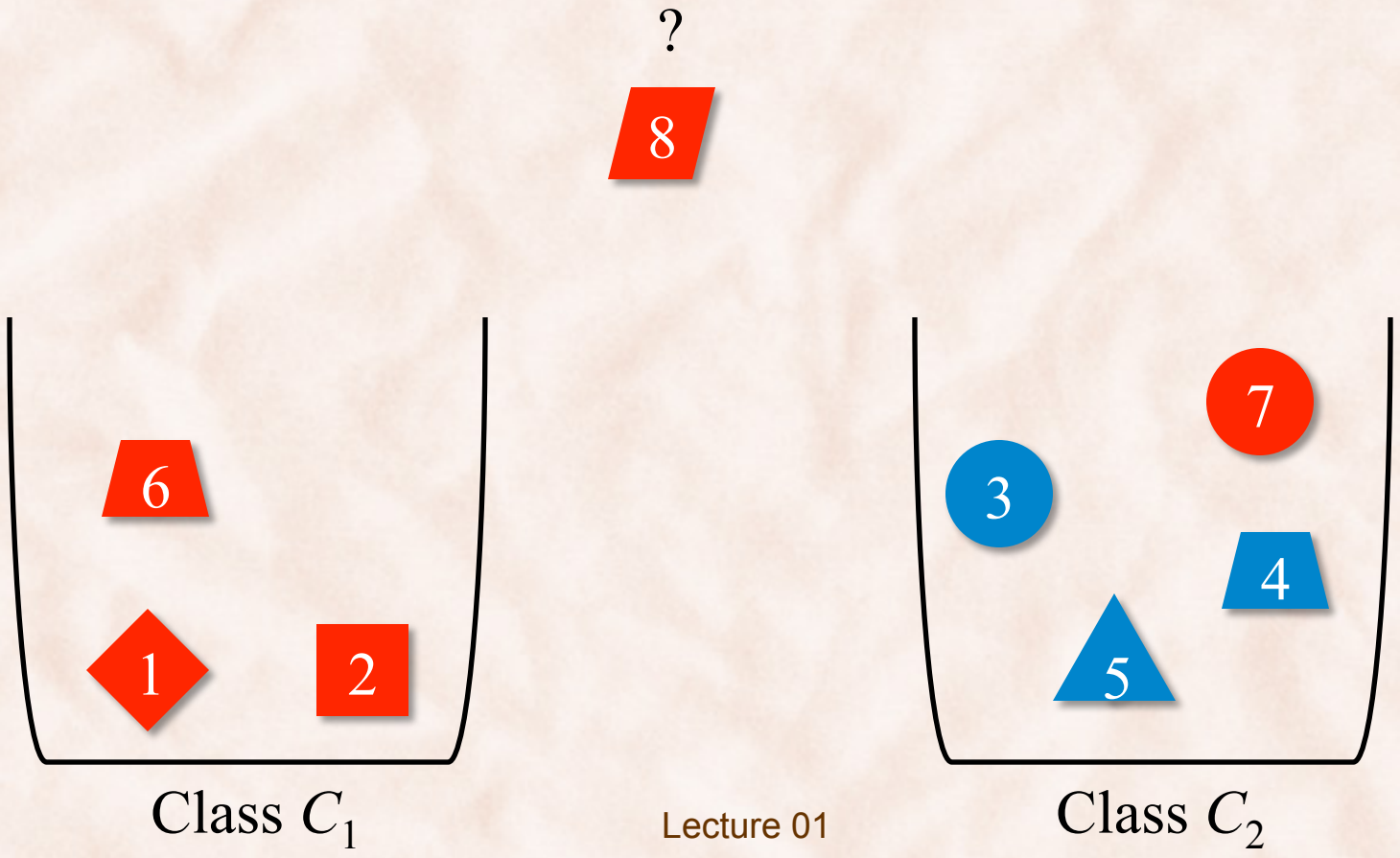
What is Learning?



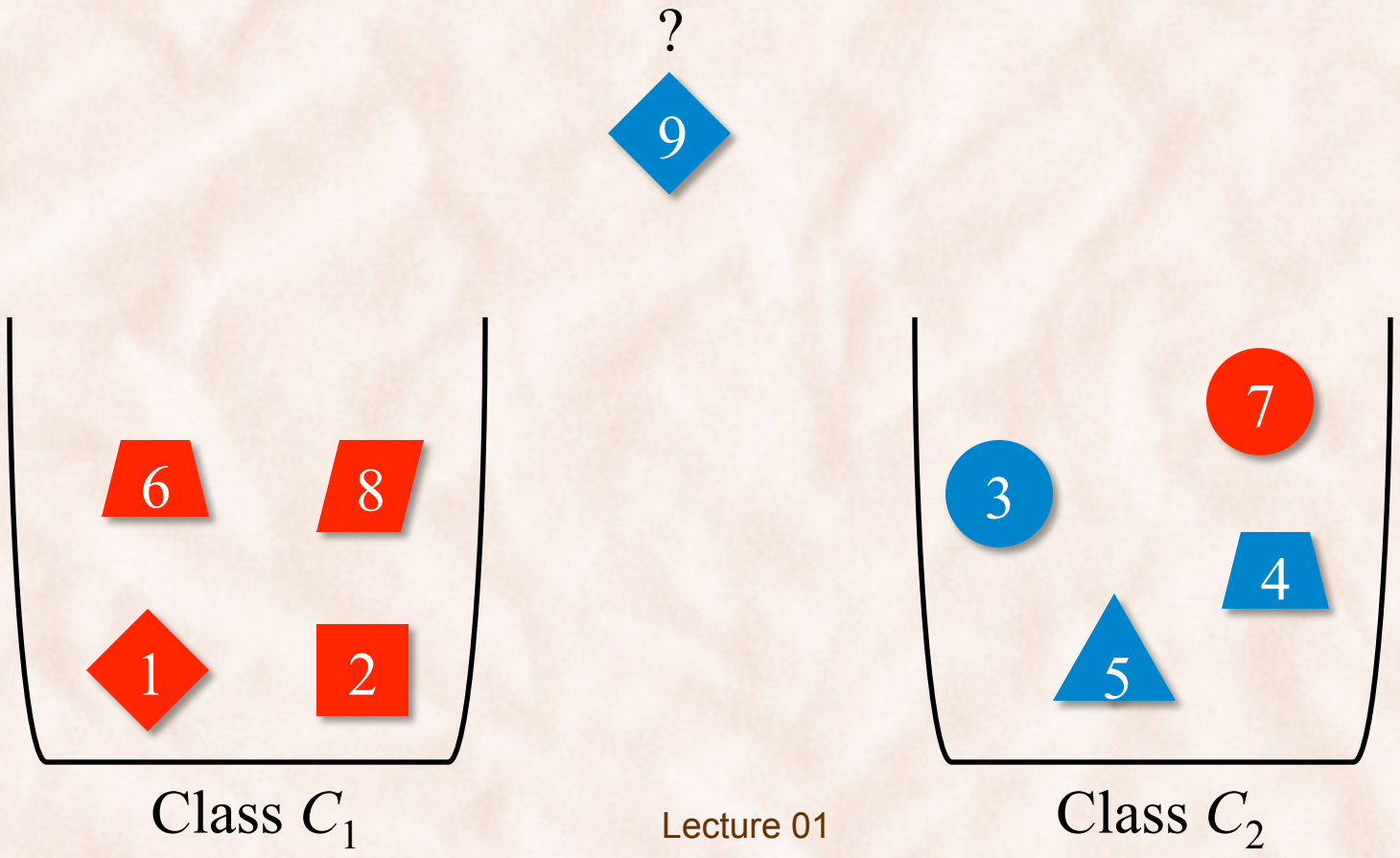
What is Learning?



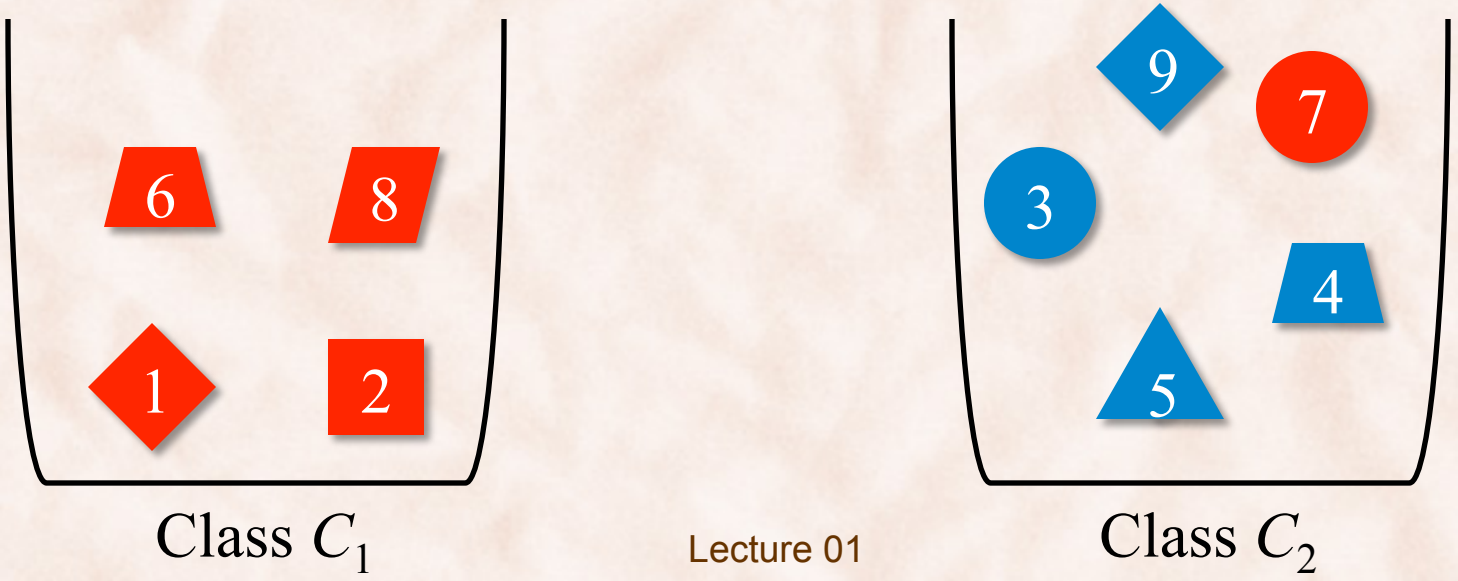
What is Learning?



What is Learning?



What is Learning?



Class C_1

Class C_2

What is Machine Learning?

- **Machine Learning** = constructing computer programs that *automatically improve with experience*:
  - Supervised Learning i.e. learning from labeled examples:
    - Classification
    - Regression
  - Unsupervised Learning i.e. learning from unlabeled examples:
    - Clustering.
    - Dimensionality reduction (vizualization).
    - Density estimation.
  - Reinforcement Learning i.e. learning with delayed feedback.

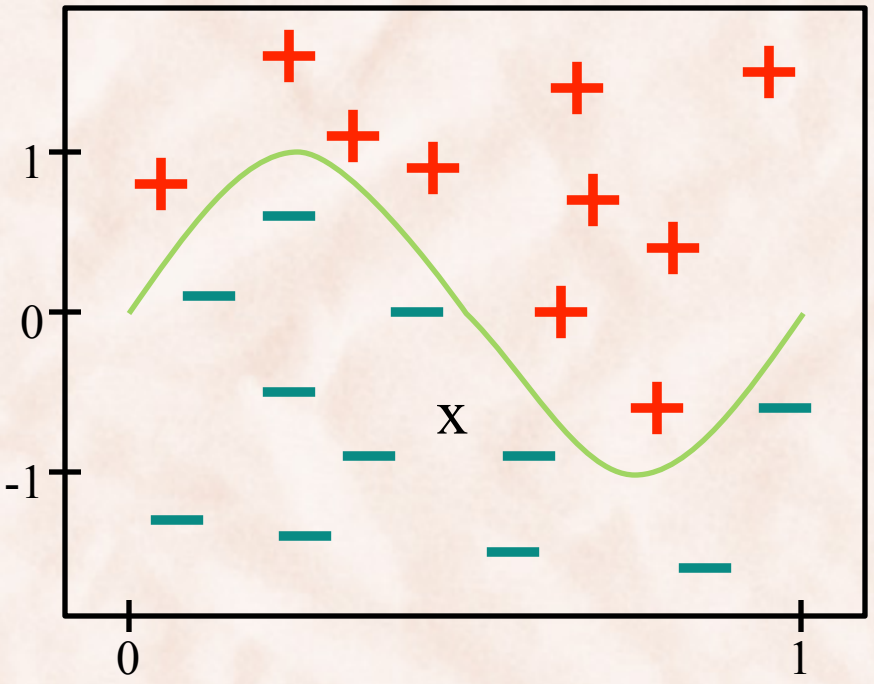
Supervised Learning

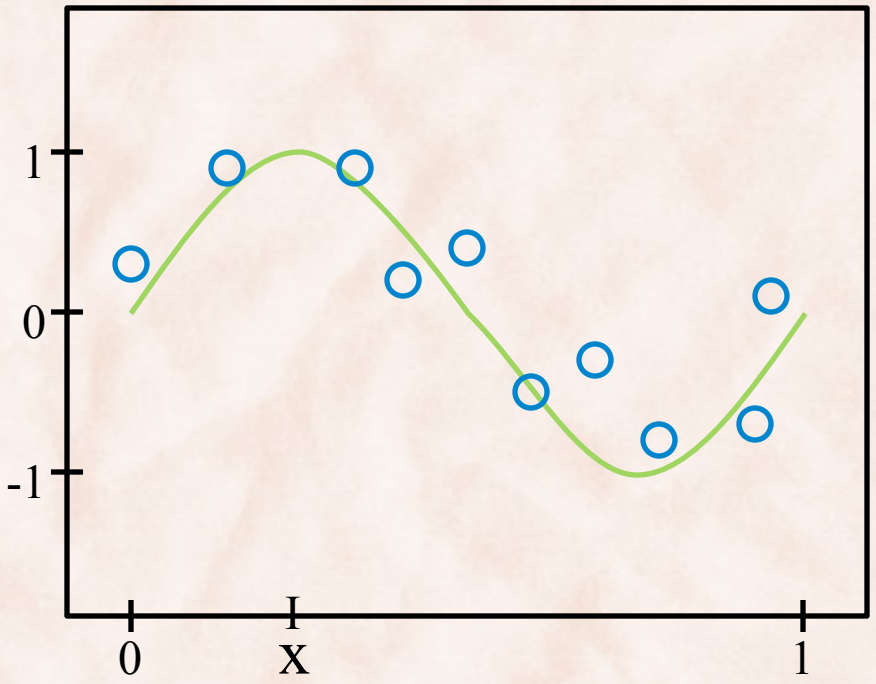
- Task = learn a function $f : X \rightarrow T$ that maps input instances $x \in X$ to output targets $t \in T$:
  - Classification:
    - The output $t \in T$ is one of a finite set of discrete categories.
  - Regression:
    - The output $t \in T$ is continuous, or has a continuous component.

- Supervision = set of training examples:

$$(x_1, t_1), (x_2, t_2), \ldots (x_n, t_n)$$

Classification vs. Regression





Classification: Junk Email Filtering

[Sahami, Dumais & Heckerman, AAAI'98]

From: Tammy Jordan
jordant@oak.cats.ohiou.edu
Subject: Spring 2015 Course

CS690: Machine Learning

Instructor: Razvan Bunescu
Email: bunescu@ohio.edu
Time and Location: Tue, Thu 9:00 AM , ARC 101
Website: http://ace.cs.ohio.edu/~razvan/courses/ml6830

Course description:
Machine Learning is concerned with the design and analysis of algorithms that enable computers to automatically find patterns in the data. This introductory course will give an overview …

From: UK National Lottery
edreyes@uknational.co.uk
Subject: Award Winning Notice

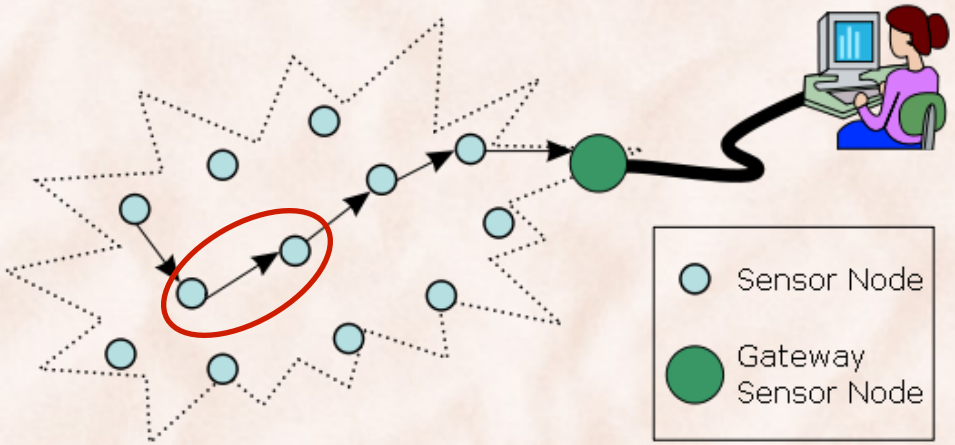
UK NATIONAL LOTTERY. GOVERNMENT ACCREDITED LICENSED LOTTERY. REGISTERED UNDER THE UNITED KINGDOM DATA PROTECTION ACT;

We happily announce to you the draws of (UK NATIONAL LOTTERY PROMOTION) International programs held in London , England Your email address attached to ticket number :3456 with serial number : 7576/06 drew the lucky number 4-2-274, which subsequently won you the lottery in the first category …

- Email filtering:
  - Provide emails labeled as *{Spam, Ham}*.
  - Train *Naïve Bayes* model to discriminate between the two.

Classification: Routing in Wireless Sensor Networks

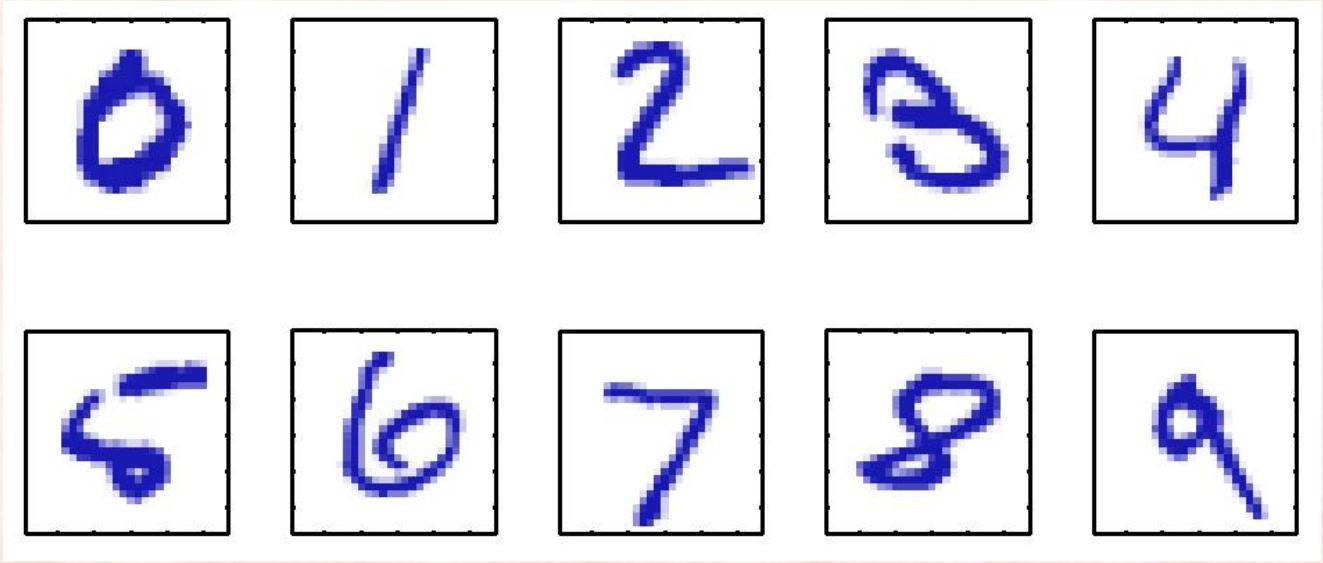
[Wang, Martonosi & Peh, SECON'06]



- Link quality prediction:
  - Provide a set of training links:
    - received signal strength, send/forward buffer sizes
    - node depth from base station, forward/backward probability
    - LQI = Link Quality Indication, binarized as {Good, Bad}
  - Train *Decision Trees* model to predict LQ using runtime features.

Classification: Handwritten Zip Code Recognition

[Le Cun et al., Neural Computation '89]



- Handwritten digit recognition:
  - Provide images of handwritten digits, labeled as {0, 1, …, 9}.
  - Train *Neural Network* model to recognize digits from input images.

Classification: Medical Diagnosis

[Krishnapuram et al., GENSIPS'02]

- Cancer diagnosis from gene expression signatures:
  - Create database of gene expression profiles (X) from tissues of known cancer status (Y):
    - Human accute leukemia dataset:
      - http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi
    - Colon cancer microarray data:
      - http://microarray.princeton.edu/oncology
  - Train *Logistic Regression* / *SVM* / *RVM* model to classify the gene expression of a tissue of unknown cancer status.

Classification: Other Examples

- Handwritten letter recognition
- Face recognition
- Credit card applications/transactions
- Recommender systems: books, music, …
- Fraud detection in e-commerce
- Worm detection in network packets
- Tone recognition
- Chord Recognition
- Named Entity Recognition

Regression: Examples

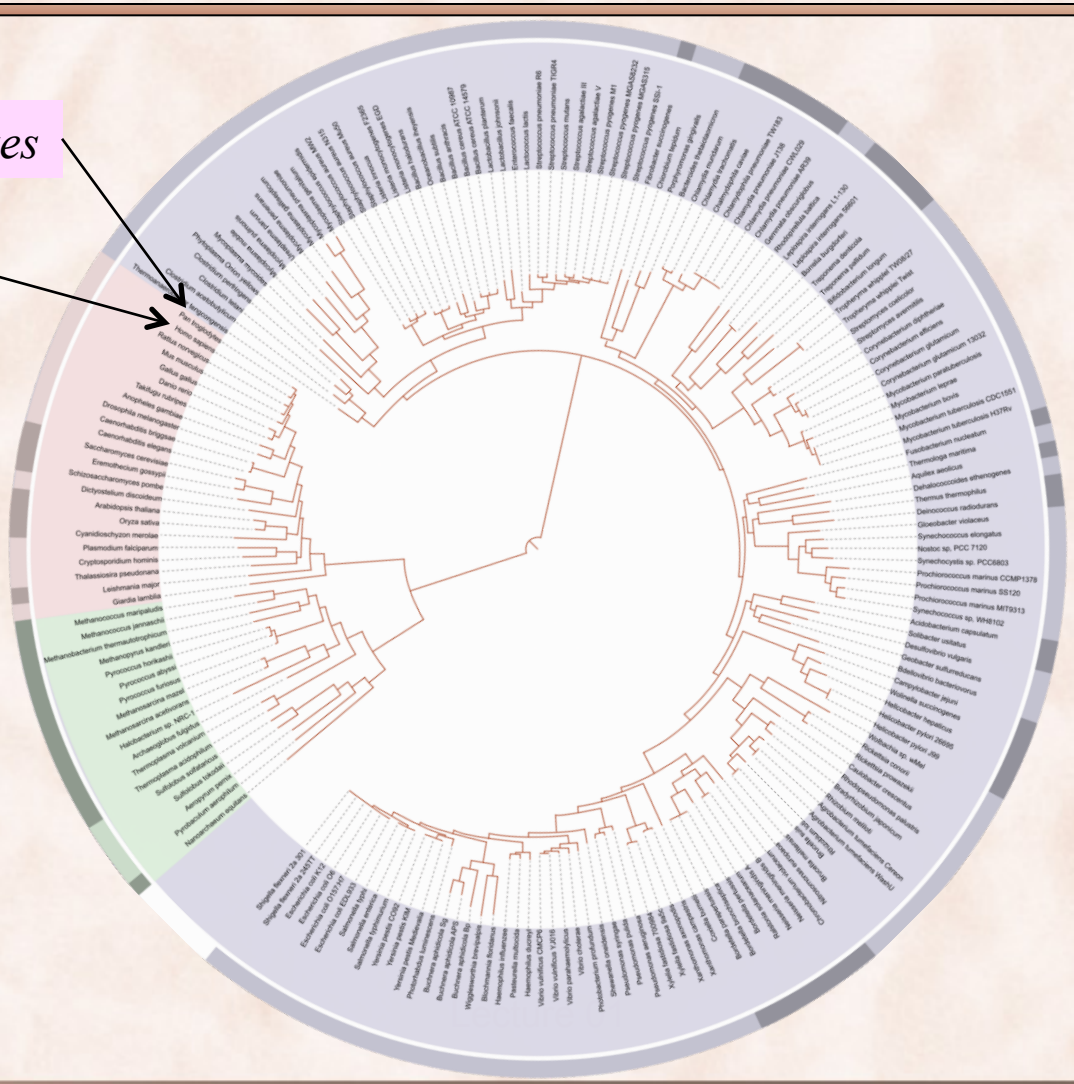
1. Stock market prediction:
   - Use the current stock market conditions ($x \in X$) to predict tomorrow's value of a particular stock ($t \in T$).
2. Oil price, GDP, income prediction.
3. Chemical processes:
   - Predict the yield in a chemical process based on the concentrations of reactants, temperature and pressure.

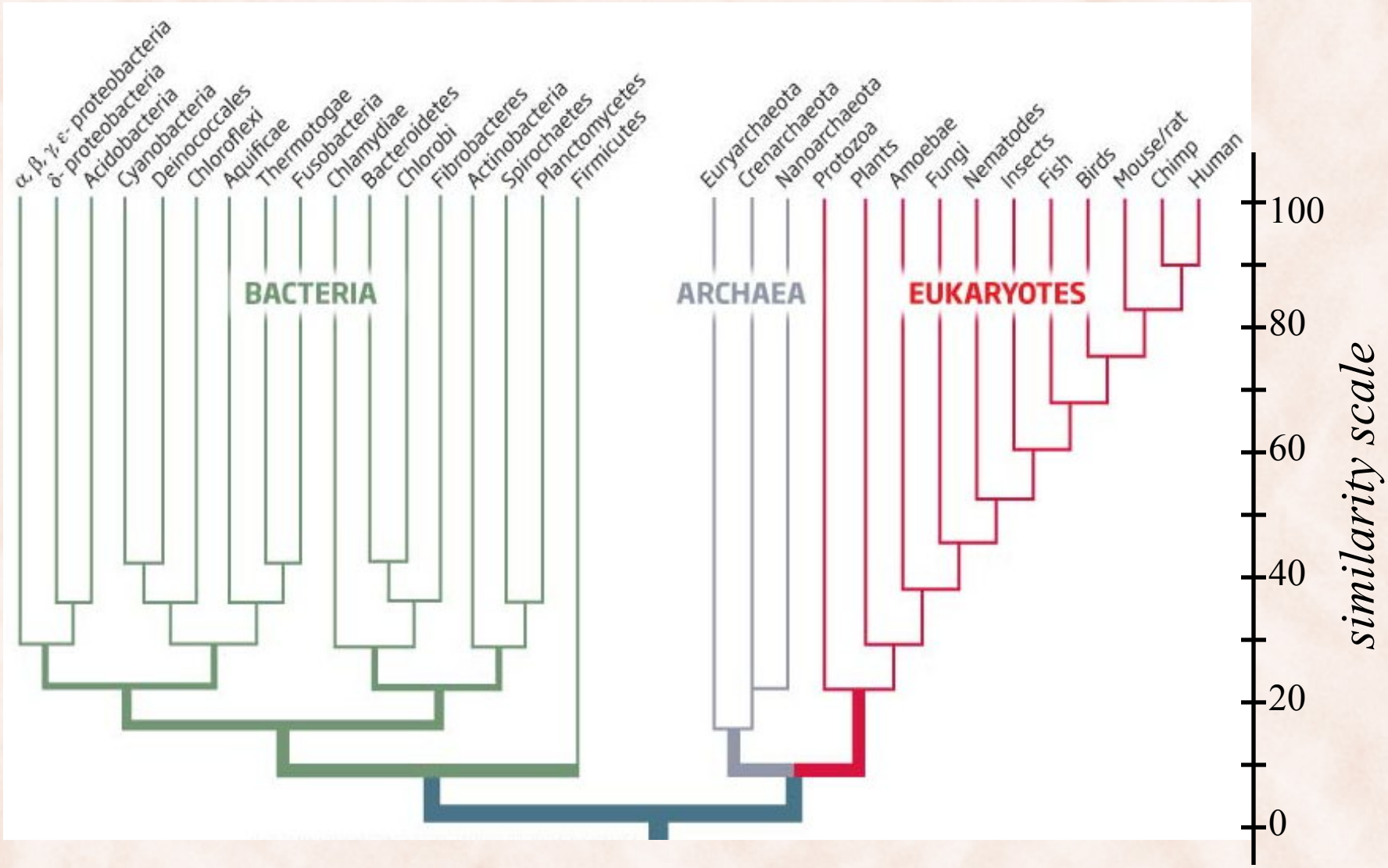
- Algorithms:
  - *Linear Regression*, *Neural Networks*, *Support Vector Machines*, …

Unsupervised Learning: Hierarchical Clustering

Pan Troglodytes

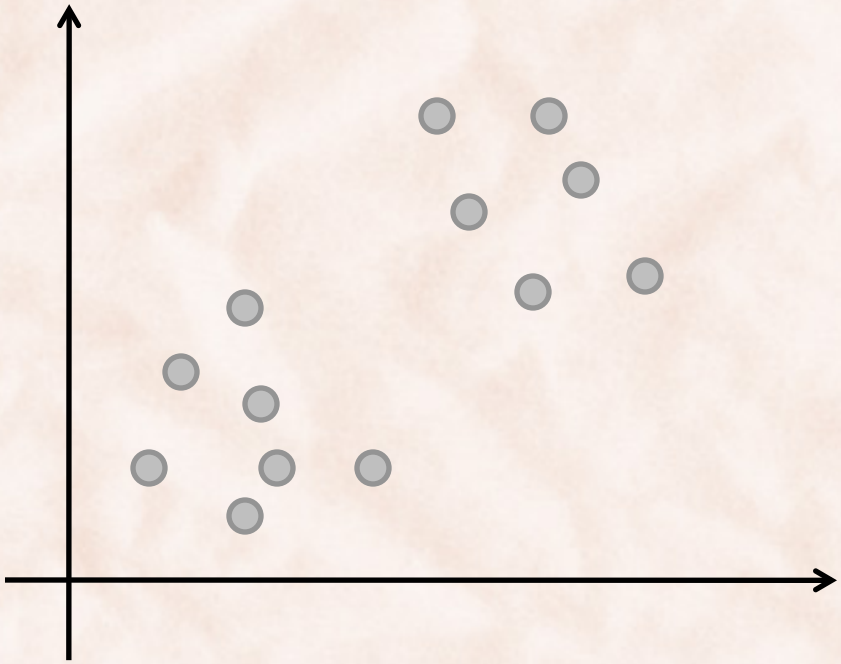
Homo Sapiens




α, β, γ, ε- proteobacteria
δ- proteobacteria
Acidobacteria
Cyanobacteria
Deinococcales
Chloroflexi
Aquificae
Thermotogae
Fusobacteria
Chlamydiae
Bacteroidetes
Chlorobi
Fibrobacteres
Actinobacteria
Spirochaetes
Planctomycetes
Firmicutes
Euryarchaeota
Crenarchaeota
Nanoarchaeota
Protozoa
Plants
Amoebae
Fungi
Nematodes
Insects
Fish
Birds
Mouse/rat
Chimp
Human
BACTERIA
ARCHAEA
EUKARYOTES
100
80
60
40
20
0
similarity scale

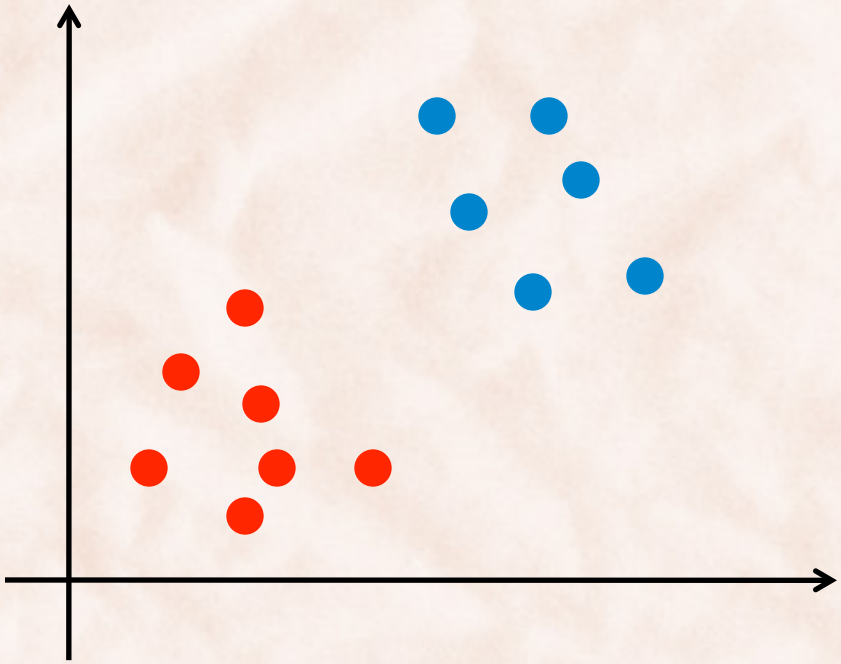
Unsupervised Learning: Clustering

- Partition unlabeled examples into disjoint clusters such that:
  - Examples in the same cluster are very similar.
  - Examples in different clusters are very different.



Unsupervised Learning: Clustering

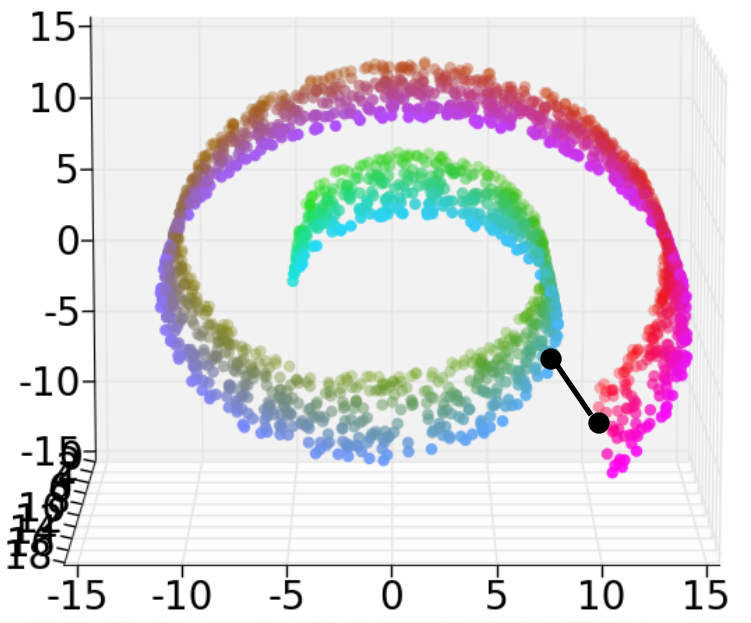
- Partition unlabeled examples into disjoint clusters such that:
  - Examples in the same cluster are very similar.
  - Examples in different clusters are very different.

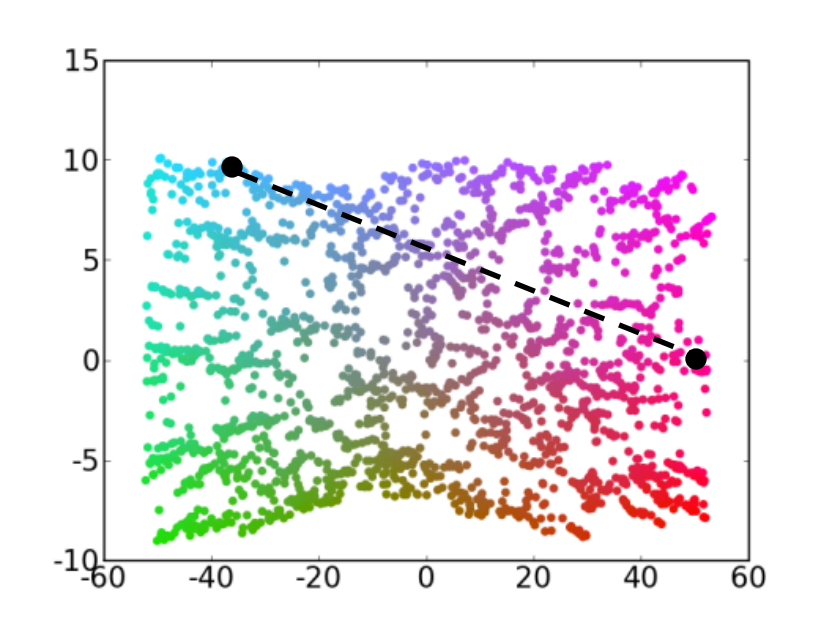


- Need to provide:
  - number of clusters ($k = 2$)
  - similarity measure (Euclidean)

Unsupervised Learning: Dimensionality Reduction

- Manifold Learning:
  - Data lies on a low-dimensional manifold embedded in a high-dimensional space.
  - Useful for *feature extraction* and *visualization*.





Reinforcement Learning

- Interaction between agent and environment modeled as a sequence of *actions* & *states*:
  - Learn *policy* for mapping states to actions in order to maximize a *reward*.
  - Reward given at the end state => *delayed reward*.
  - States may be only *partially observable*.
  - Trade-off between *exploration* and *exploitation*.
- Examples:
  - Backgammon [Tesauro, CACM'95].
  - Aerobatic helicopter flight [Abbeel, NIPS'07].
  - 49 Atari games, using deep RL [Mnih et al., Nature'15].

Reinforcement Learning: TD-Gammon

[Tesauro, CACM'95]

- Learn to play Backgammon:
  - Immediate reward:
    - +100 if win
    - -100 if lose
    - 0 for all other states
  - *Temporal Difference Learning* with a *Multilayer Perceptron*.
  - Trained by playing 1.5 million games against itself.
  - Played competitively against top-ranked players in international tournaments.

Relevant Disciplines

- Mathematics:
  - Probability & Statistics
  - Information Theory
  - Linear Algebra
  - Optimization
- Algorithms:
  - Computational Complexity
- Artificial Intelligence
- Psychology
- Neurobiology
- Philosophy

Readings

- PRML 1.2, 2.1 – 2.1.1, 2.2 – 2.2.1, 2.3 (2.3.4, 2.3.9).
- PRML Appendix B and C.