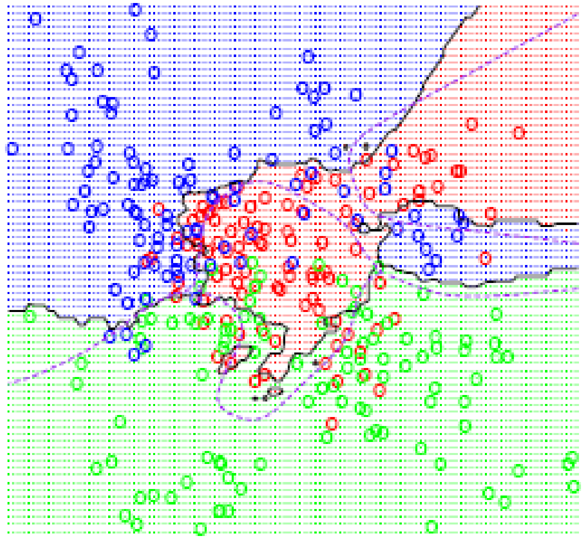
Machine Learning

Lecture 1: Introduction to Machine Learning



Hesam Montazeri
Department of Bioinformatics, IBB, University of Tehran

Shahrivar 31, 1398

University of TEHRAN

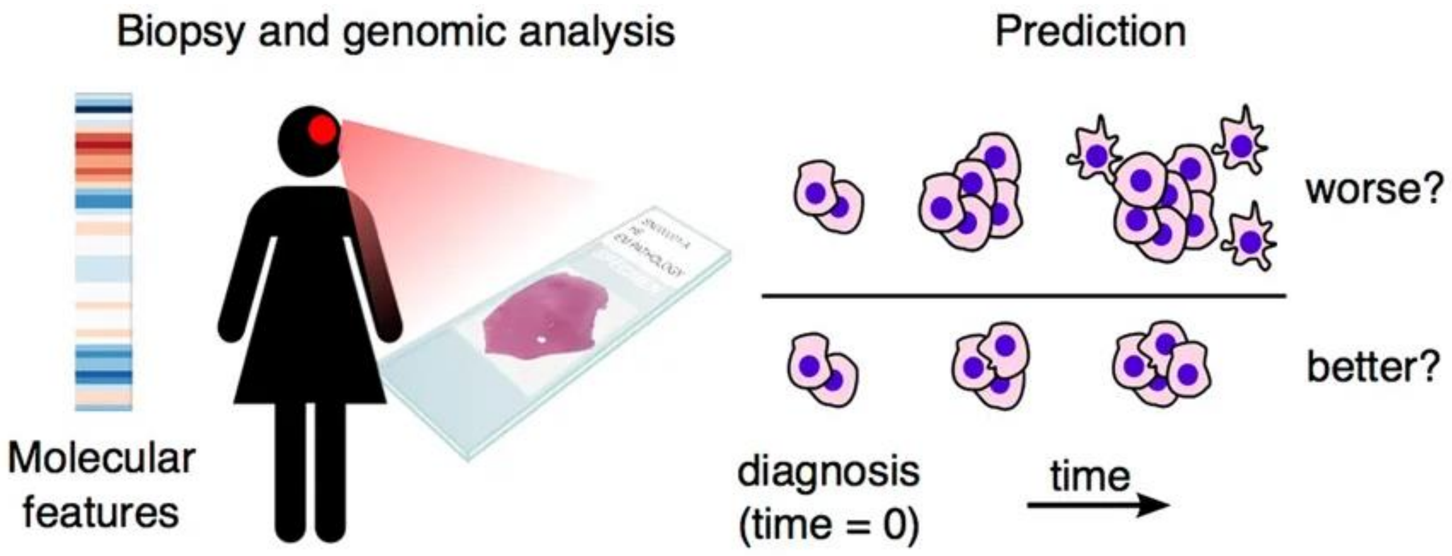
Definition

- “Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead.” (Wikipedia)

- Another definition by Tom Mitchell
  - *“Machine learning is the study of computer algorithms that allow computer programs to automatically improve through experience”*

Example: prediction of clinical outcomes

- Prediction of clinical outcomes from high-dimensional molecular data
- It can be done based on patterns in existing data.



From S. Yousefi, Scientific Reports, 2017

Machine learning problems in Bioinformatics

- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Predict HIV drug resistance from genotype data
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Protein structure prediction
- Identify the risk factors for prostate cancer.
- Gene prediction: determine the location of protein-encoding genes within a given DNA sequence
- Finding regulatory motifs

Other applications

- Speech recognition
- Computer vision
- Automatic translation
- Product recommendation
- Spam detection
- Game playing
- And many many more!

An overview of machine learning

- Machine learning tools can be classified to
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning

Supervised learning- definition

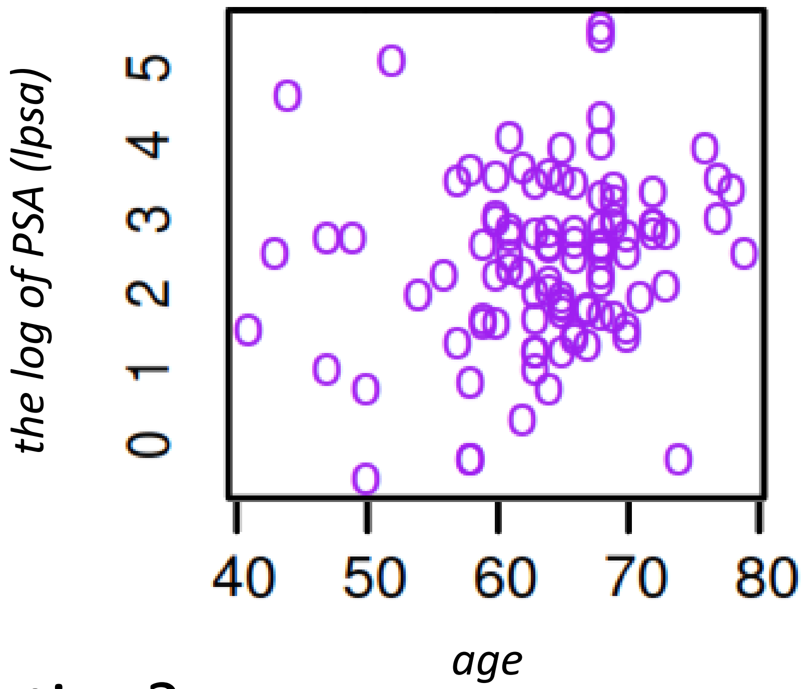
- Supervised learning is the task of learning a function that maps an input to an output based on a set of example input-output pairs.
- Supervised learning uses labeled data.

- Naming convention:
  - Input variables:
    - features, predictors, independent variables, or just variables
    - Usually is denoted by the symbol X
  - Output variable:
    - target, response or dependent variable
    - Usually is denoted by the symbol Y

- Prediction task is a
  - regression task when the output variable is quantitative (or continuous).
  - classification task when the output variable is qualitative (categorical).

Example: Prostate cancer

- Goal: to examine the correlation between the level of prostate antigen (PSA) and age



- Regression/classification?

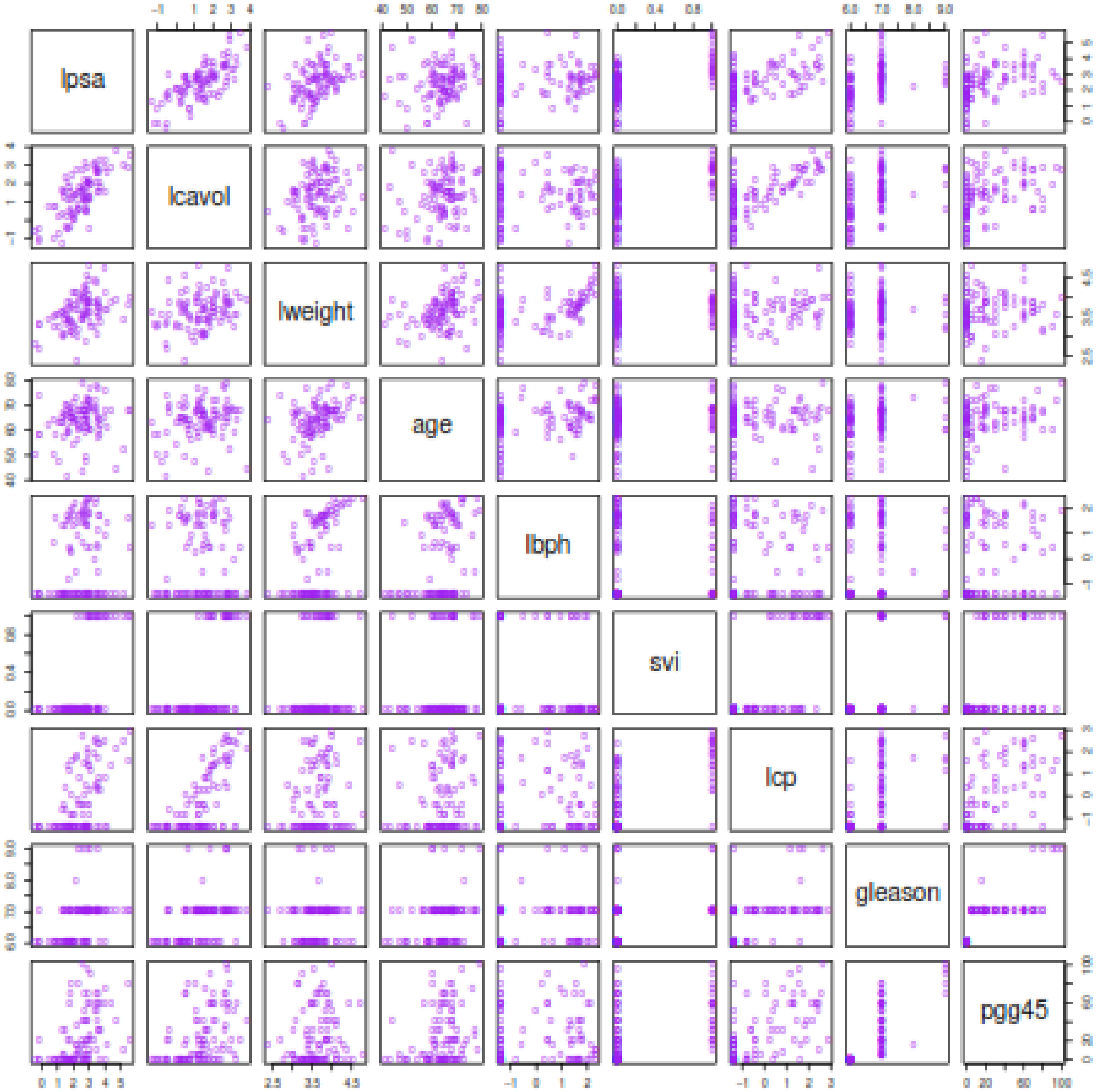
Example: Prostate cancer

- Goal: to examine the correlation between the level of prostate antigen (PSA) and **a number of clinical measures**
- Data: 97 men who were about to receive a radical prostatectomy
- Output: the log of PSA (lpsa)
- Inputs:
  - Log cancer volume (*lcavol*)
  - Log prostate weight (lweight)
  - Age
  - Log of benign prostatic hyperplasia amount (lbph)
  - Seminal vesicle invasion (svi)
  - Log of capsular penetration (lcp)
  - Gleason score (gleason)
  - Percent of Gleason scores 4 and 5 (pgg45)

University of TEHRAN

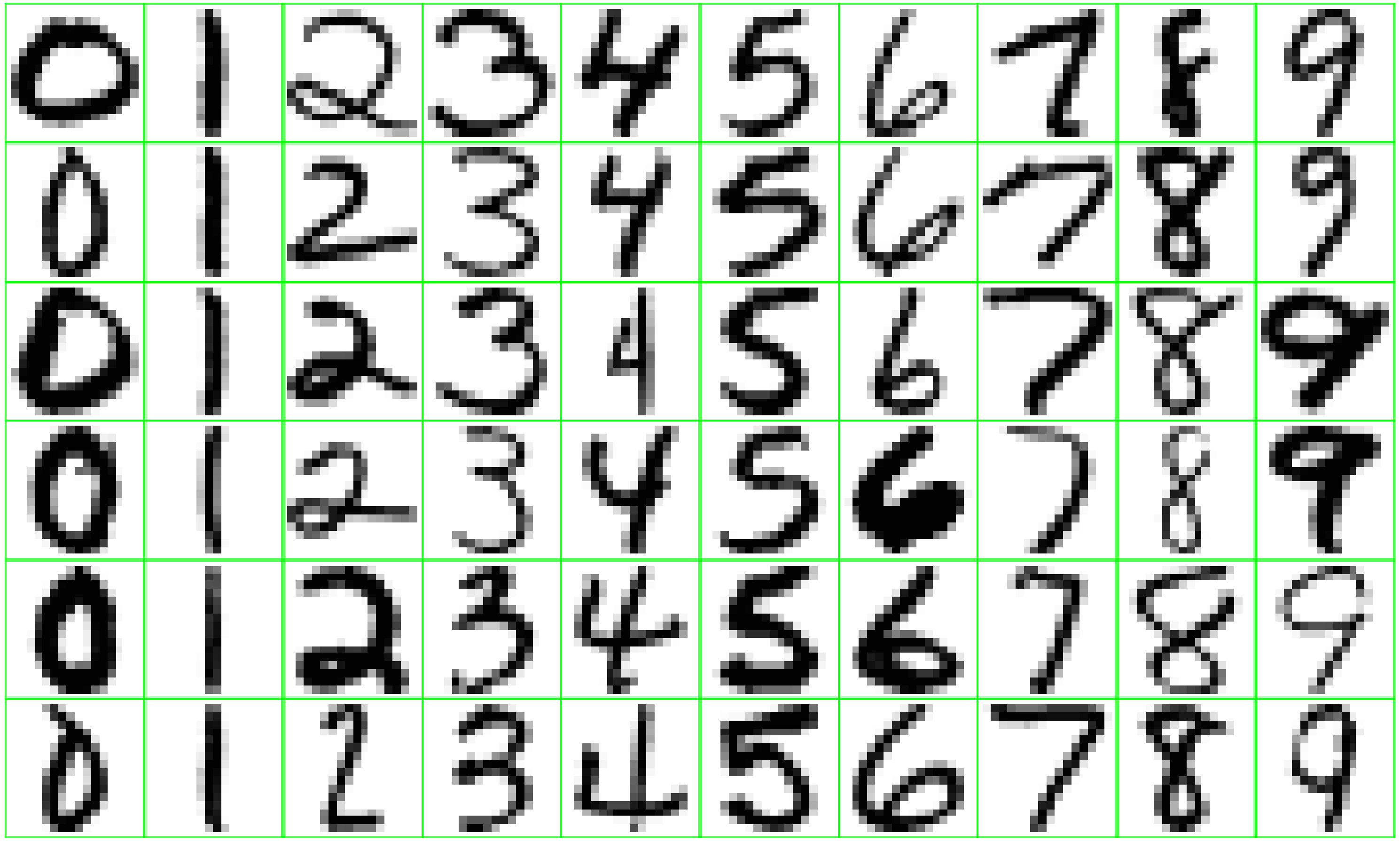
Example: Prostate cancer

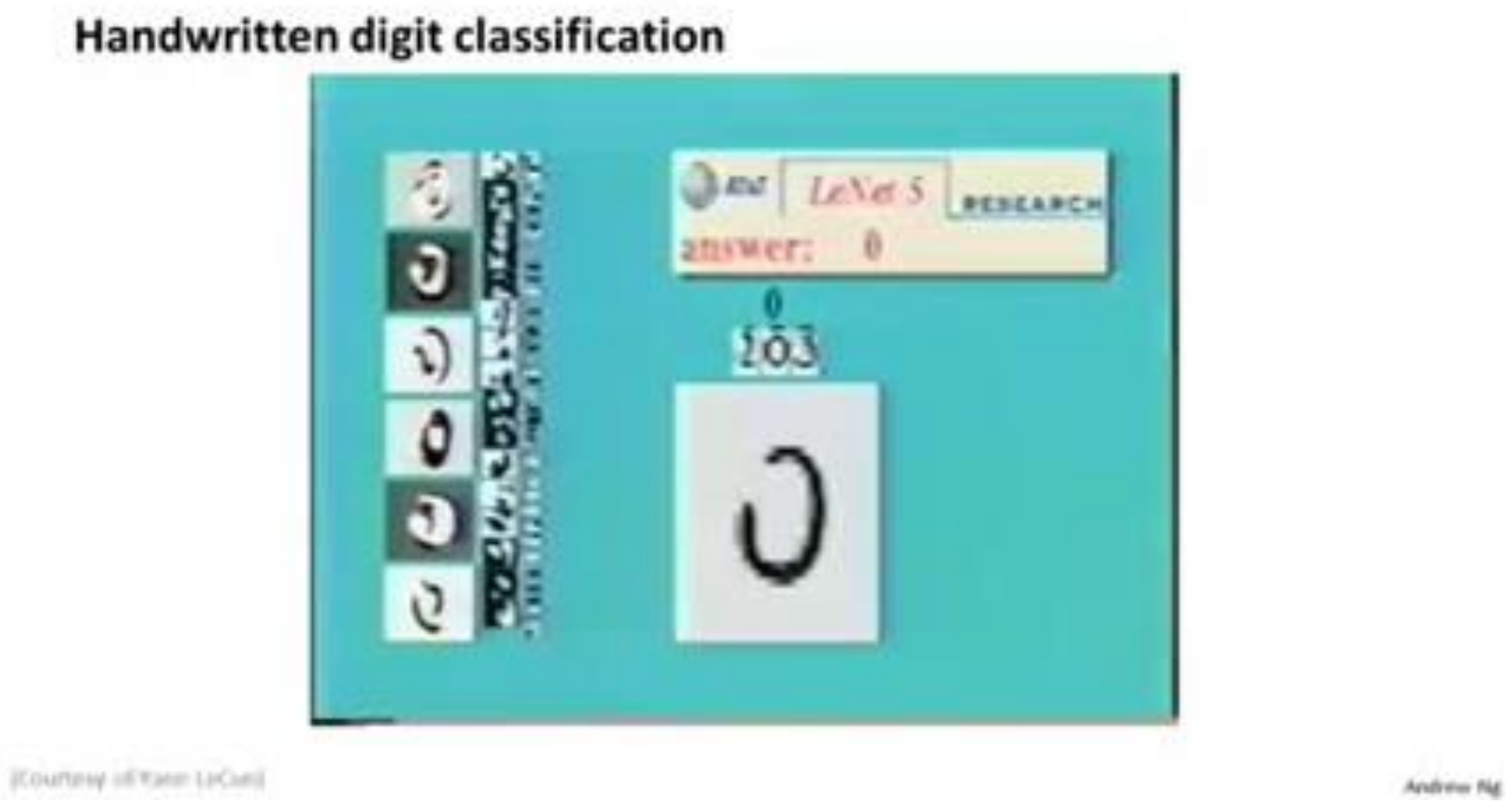
- *Gleason* and *svi* are categorical.
- Some correlations are evident from the scatter plot.



Data appeared in Stamey et al. , 1989

Classification problem: Identify digits in a handwritten zip code

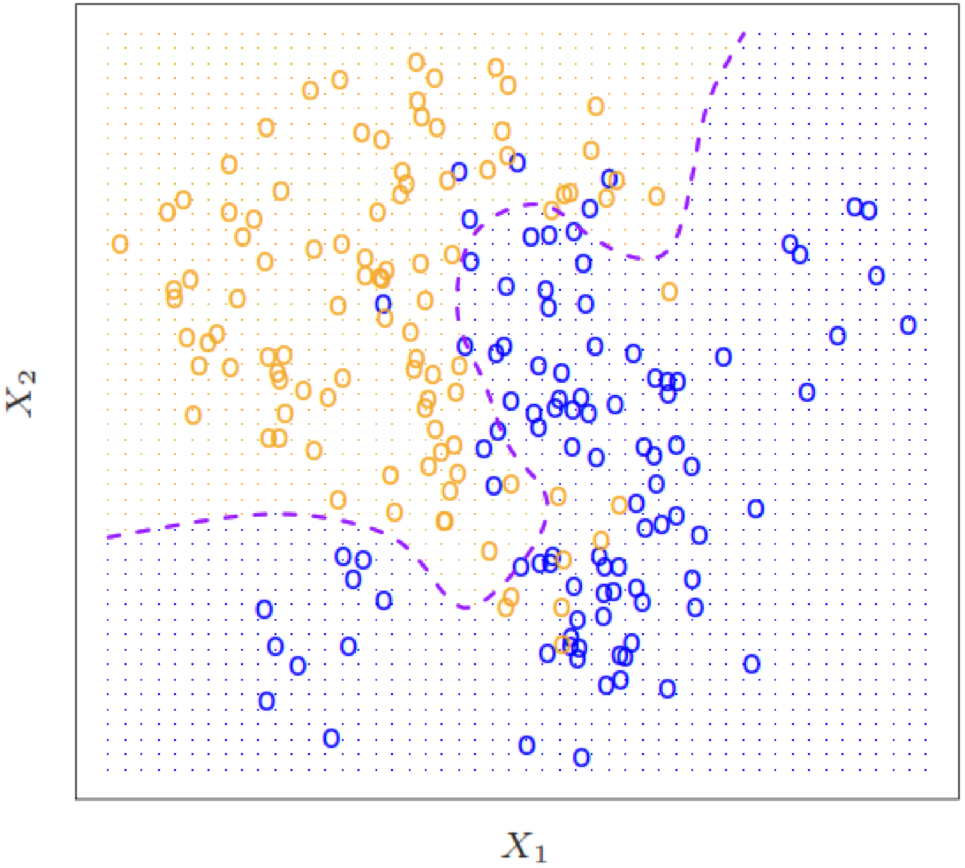




Handwritten Digit Classification - Yann Lecun
https://www.youtube.com/watch?v=yxuRnBEczUU

University of TEHRAN

An abstract example in two dimensions



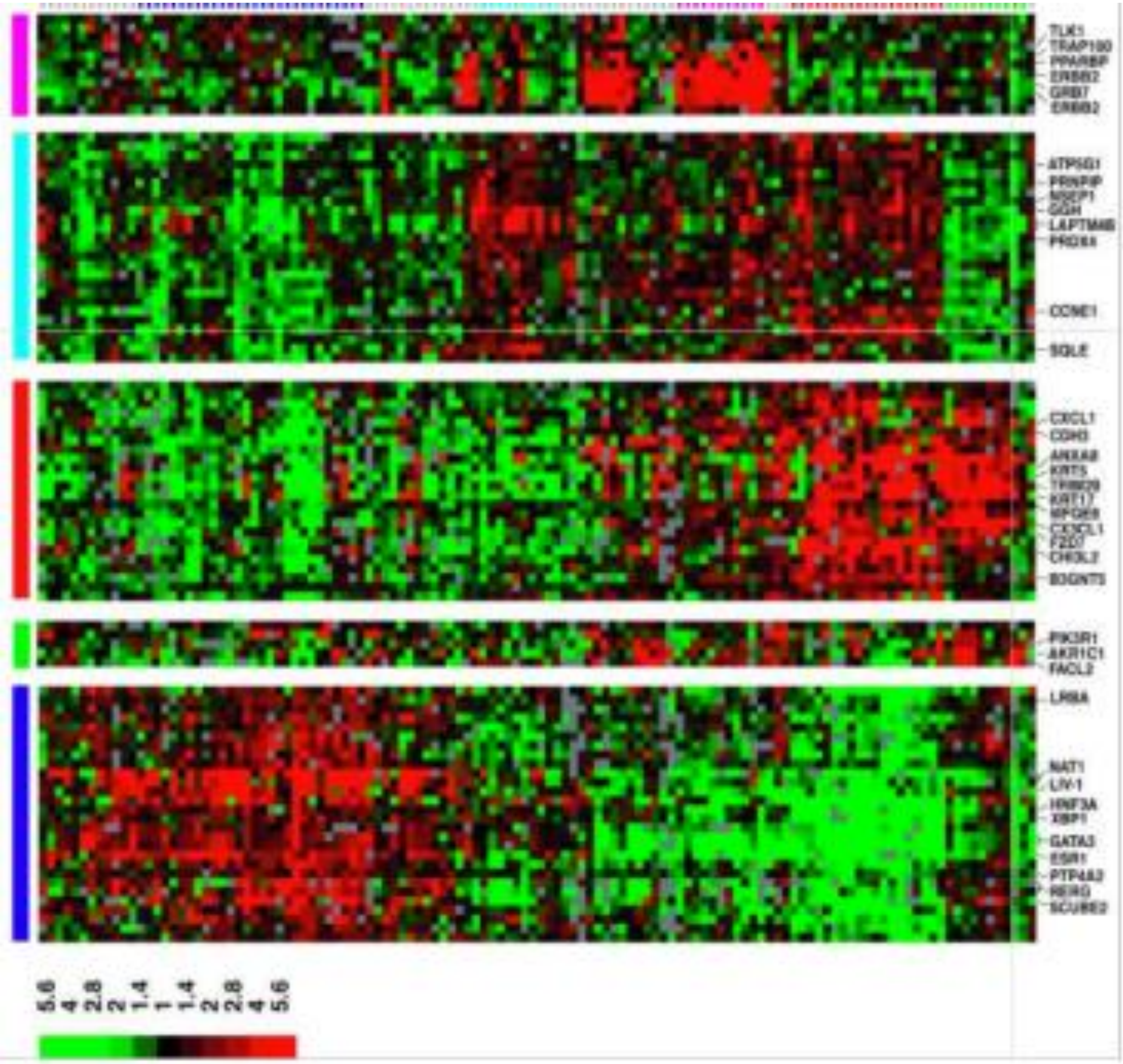
Unsupervised learning

- A set of statistical tools intended to the setting in which we have only a set of features $X_1, X_2, \ldots, X_p$ measured on *n* observations.

- The goal is to discover interesting things about the measurements. For example,
  - Can we discover subgroups among the variables or among the observations?
  - Is there an informative way to visualize the date?

- Difficult to know how well you are doing.

Finding interesting gene sets from a gene expression profile



University of TEHRAN

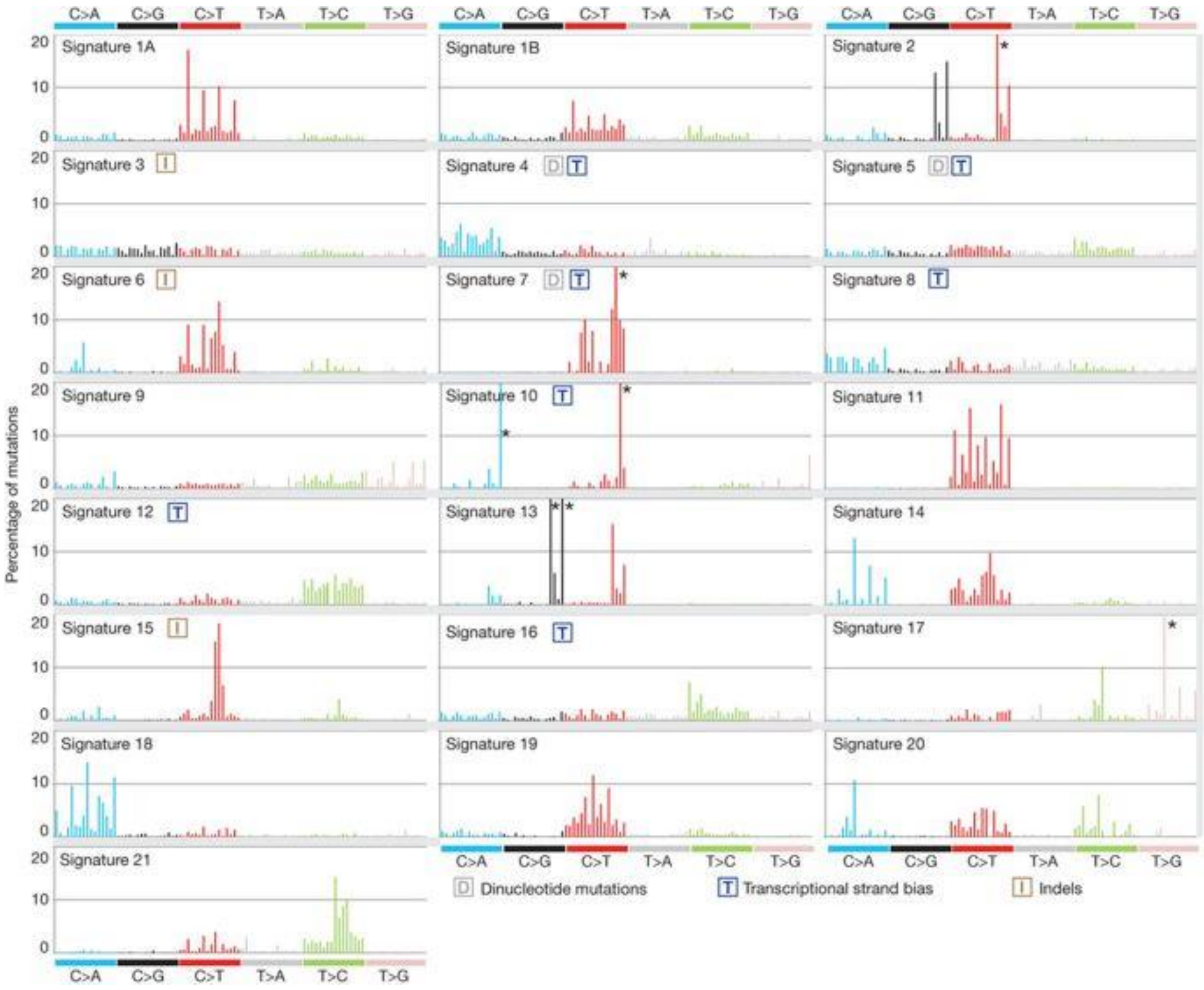
Unsupervised learning- blind source separation

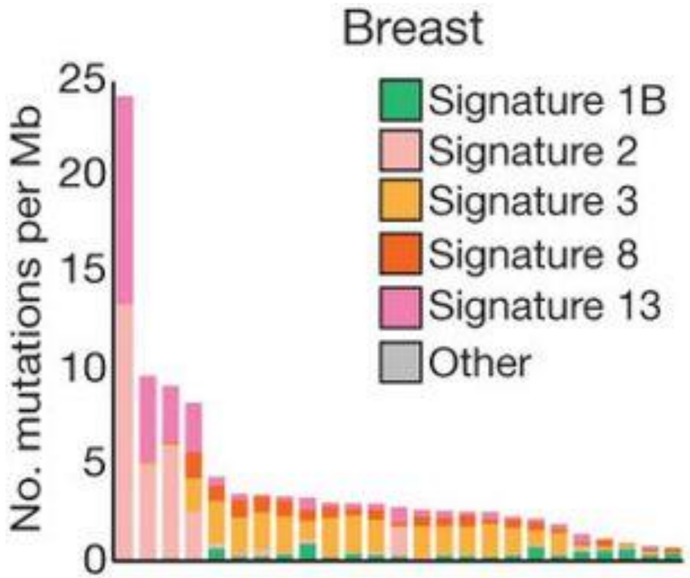
Mixed

Separated

slide from CS229, Stanford

Unsupervised learning- signatures of mutational processes in human cancer



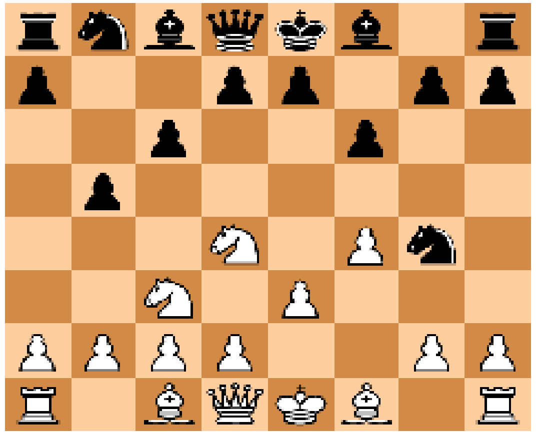


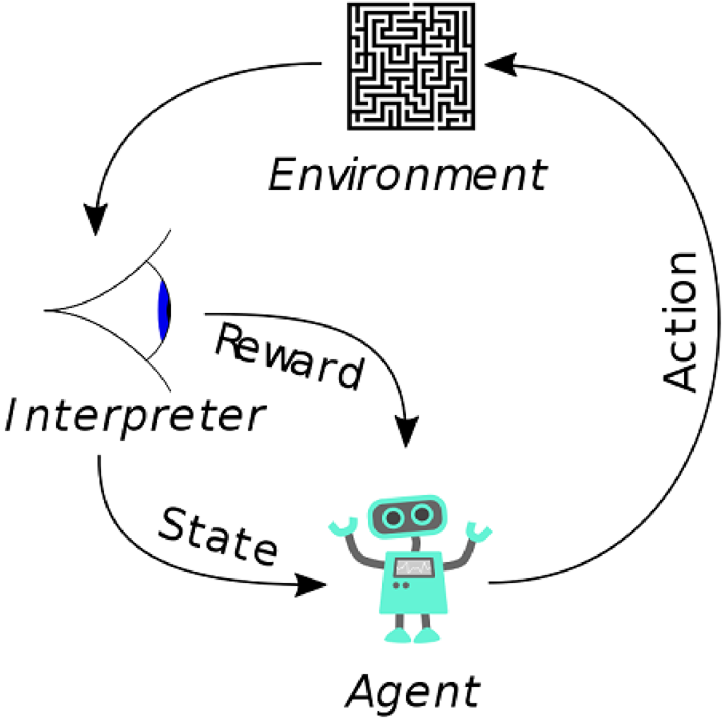
Ludmil B. Alexandrov, Nature, 2013

Reinforcement learning

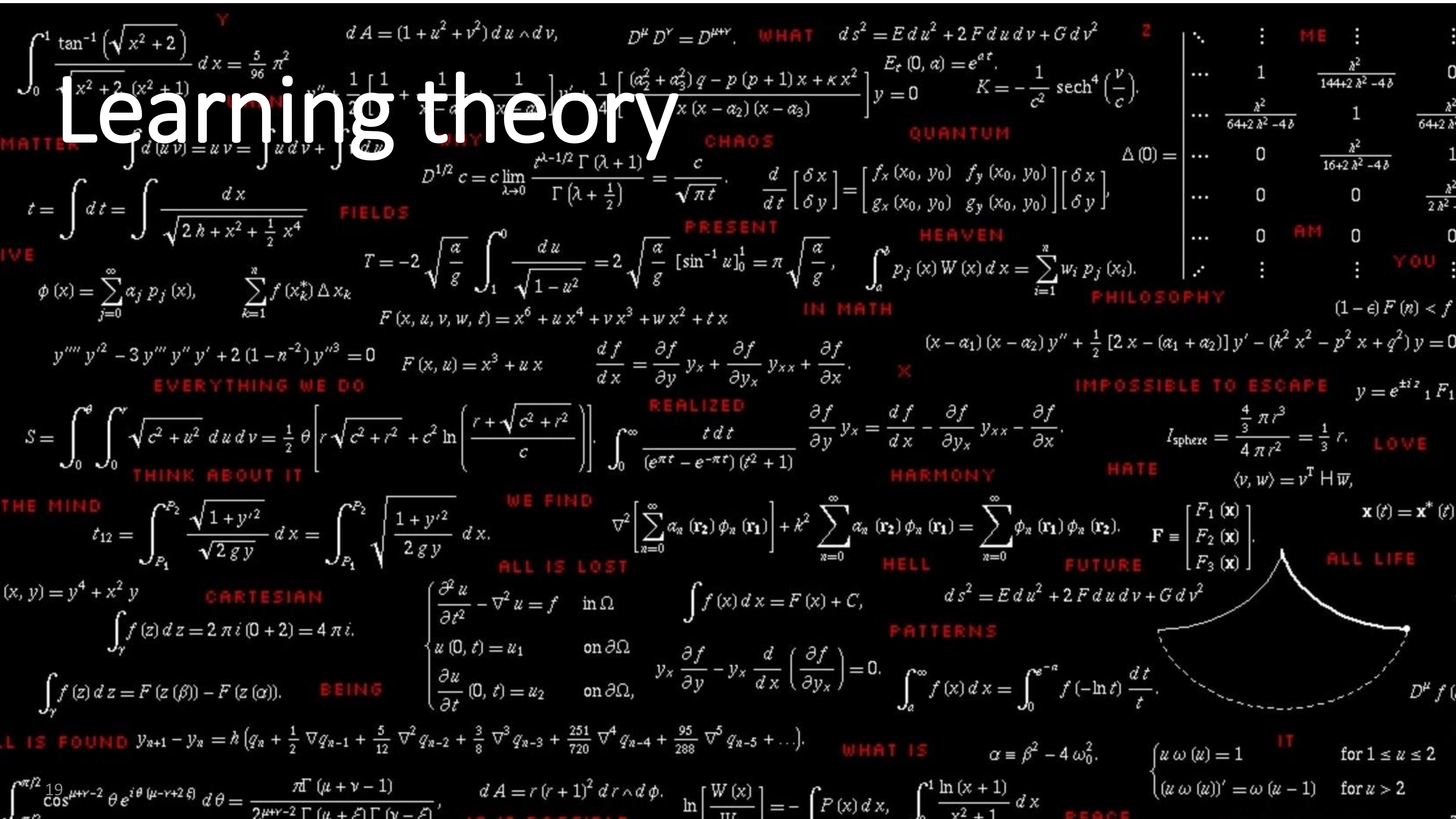
- Reinforcement learning (RL) is another paradigm in machine learning.

- In this paradigm, software agents ought to take actions in an environment so as to maximize cumulative reward.
- Important concept:
  - Environment
  - Agent
  - State
  - Action
  - Reward
- Applications
  - Robotics
  - Playing games
  - Self driving cars





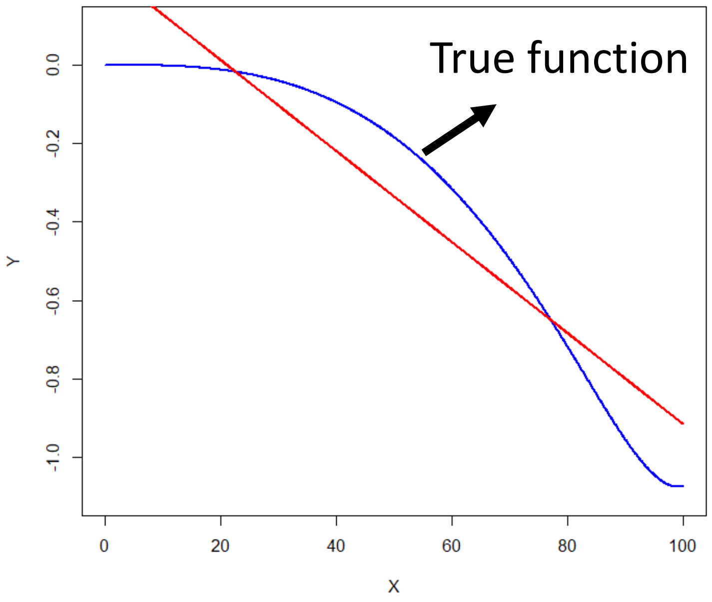
Learning theory



Linear Regression

Linear regression

- A simple approach to supervised learning.
- A useful tool for predicting a quantitative (continuous) response.
- It assumes that dependence of Y on $X_1, X_2, \ldots, X_p$ is linear.
- True regression functions are never linear!



- Linear regression is still widely used and is the basis for many modern approaches.

Simple linear regression

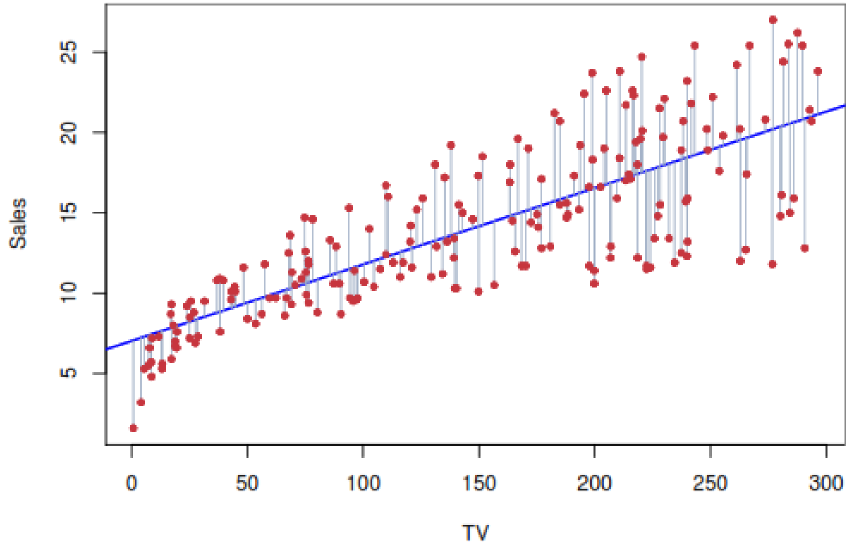
- Predicting Y based on only a single predictor variable X.
- The simple linear regression model

$$Y \approx \beta_0 + \beta_1 X$$

"approximately modelled as"

- β_0 and β_1 are known as the model *coefficients* or *parameters.*

intercept *slope*

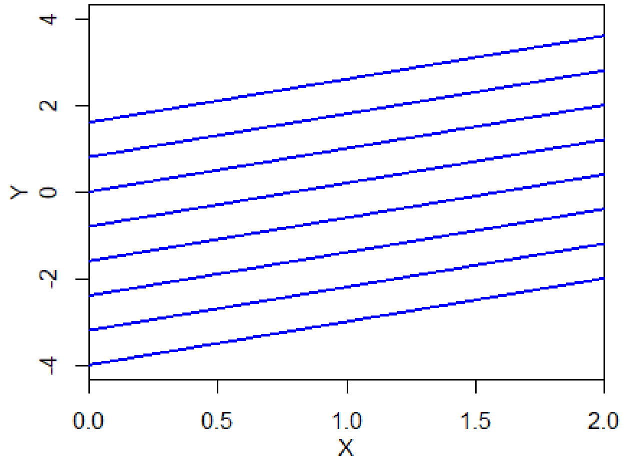


Linear relation between advertising budget on TV and sales of a particular product

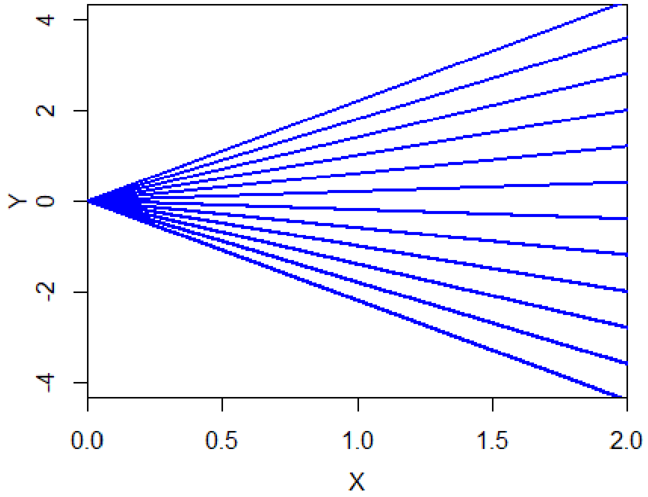
$$\text{sales} \approx \beta_0 + \beta_1 \times \text{TV}$$

Linear relationship

Changing β_0



Changing β_1



Question: which lines are associated to larger β_0 and β_1?

Estimating the coefficients

- In practice, β_0 and β_1 are unknown.
- We need training data to estimate the parameters of the model $Y \approx \beta_0 + \beta_1 X$

$$(x[1], y[1]), (x[2], y[2]), \dots, (x[n], y[n])$$

n observation pairs, each of which consists of a measurement of X and a measurement of Y.

- Using training data, we can produce estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the parameters that the linear model **fits the data well**.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- We use a hat symbol, ^ , to denote the estimated value of an unknown parameter.

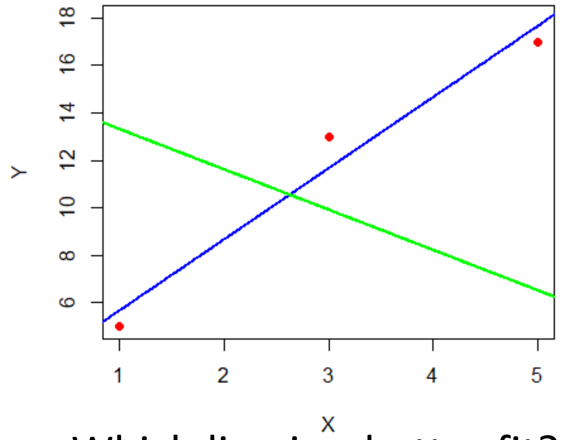
- In other words, $\hat{\beta}_0$ and $\hat{\beta}_1$ results in a line that is **as close as possible** to the observed data points.

Several ways to define the closeness

Finding a good model

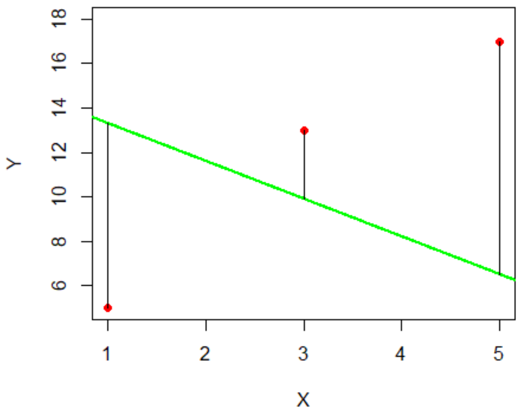
- Let's start by a synthetic data

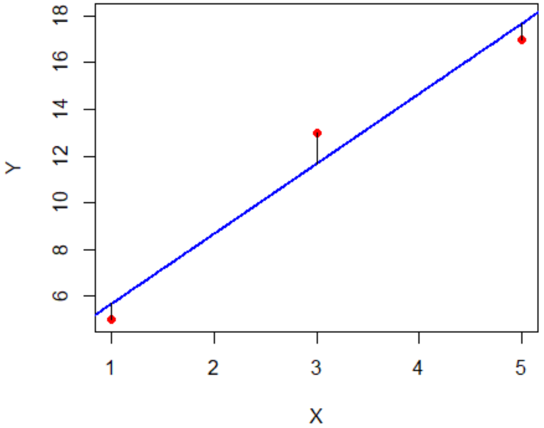
i	x_i	y_i
1	1	5
2	3	13
3	5	17



Which line is a better fit?

- Let's define quantitatively why the blue line is a better fit





- Residual for observation i
  - defined as the difference between the ith observed value ($y[i]$) and the predicted value by our linear model ($\hat{y}[i] = \hat{\beta}_0 + \hat{\beta}_1 x[i]$)

$$r[i] = y[i] - \hat{y}[i]$$

University of TEHRAN

Loss functions

- Squared loss function
  - There is an analytical closed-form solution for minimizing this loss function

$$\mathcal{L}(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y[i] - \beta_0 + \beta_1 x[i])^2$$

- Other possibility for defining the loss function, for example, the absolute loss

$$\mathcal{L}(\beta) = \frac{1}{2}\sum_{i=1}^{n}|y[i] - \beta_0 + \beta_1 x[i]|$$

- Finding the best fit can be mathematically expressed as

$$\hat{\beta}_0, \hat{\beta}_1 = \arg\min_{\beta_0, \beta_1} \mathcal{L}(\beta)$$

The term *arg min* means “find the argument that minimizes …”

University of TEHRAN

The least squares solution

- Gradient descent algorithm
  - See whiteboard notes (partly based on CS229 notes on supervised learning)

References and Acknowledgement

- References
  - An Introduction to Statistical Learning, with applications in R, 2013
  - Slide 23 is from CS229 Stanford course.

- Acknowledgement
  - Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani. Few slides are also adjusted from theirs.

University of TEHRAN