CS340 Machine learning
Lecture 2

What is machine learning?

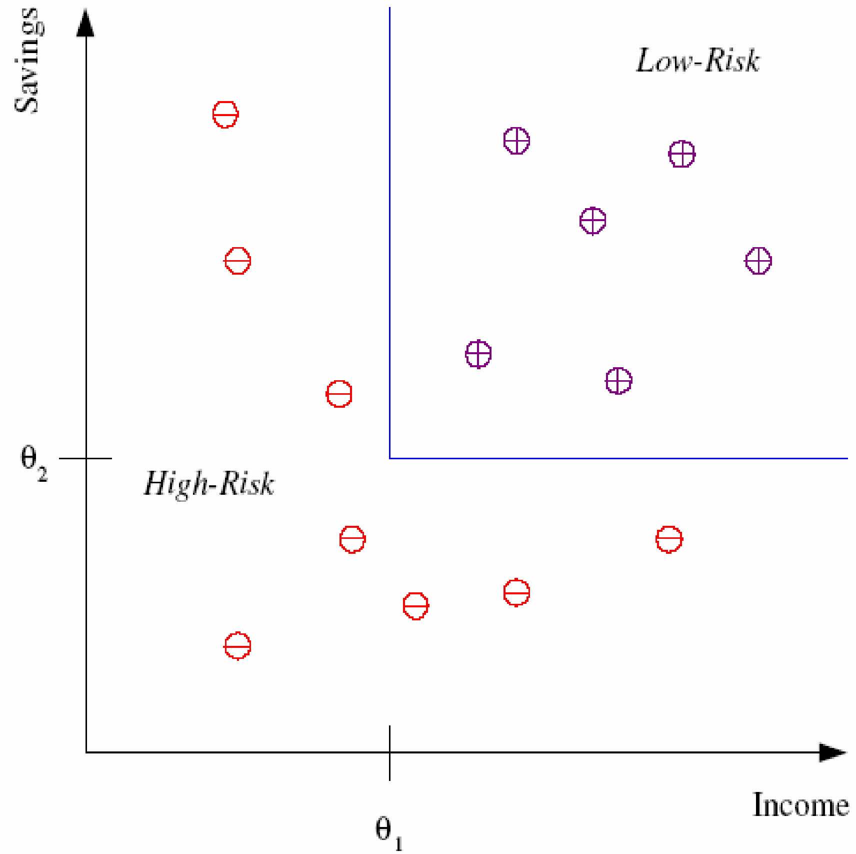
- ``Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the task or tasks drawn from the same population more efficiently and more effectively the next time.'' -- Herbert Simon
- Closely related to
  - Statistics (fitting models to data and testing them)
  - Data mining/ exploratory data analysis (discovering models)
  - Adaptive control theory
  - AI (building intelligent machines by hand)

Types of machine learning

- Supervised Learning
  - Classification (pattern recognition)
  - Regression
- Unsupervised Learning
- Reinforcement Learning

Classification

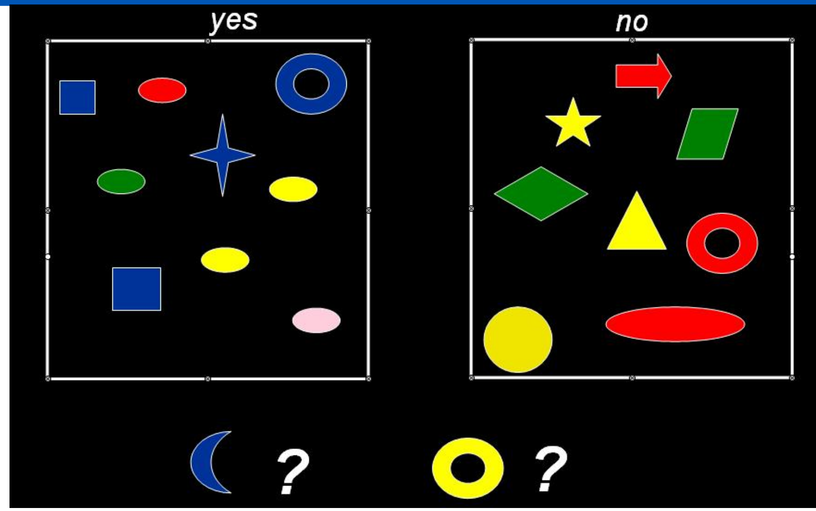
- Example: Credit scoring
- Differentiating between low-risk and high-risk customers from their *income* and *savings*



Discriminant: IF *income* > θ_1 AND *savings* > θ_2
THEN low-risk ELSE high-risk

Input data is two dimensional, output is binary}

Classification



p features (attributes)

Training set:
X: n §p
y: n § 1

n cases

Color	Shape	Size
Blue	Square	Small
Red	Ellipse	Small
Red	Ellipse	Large

Label
Yes
Yes
No

Test set

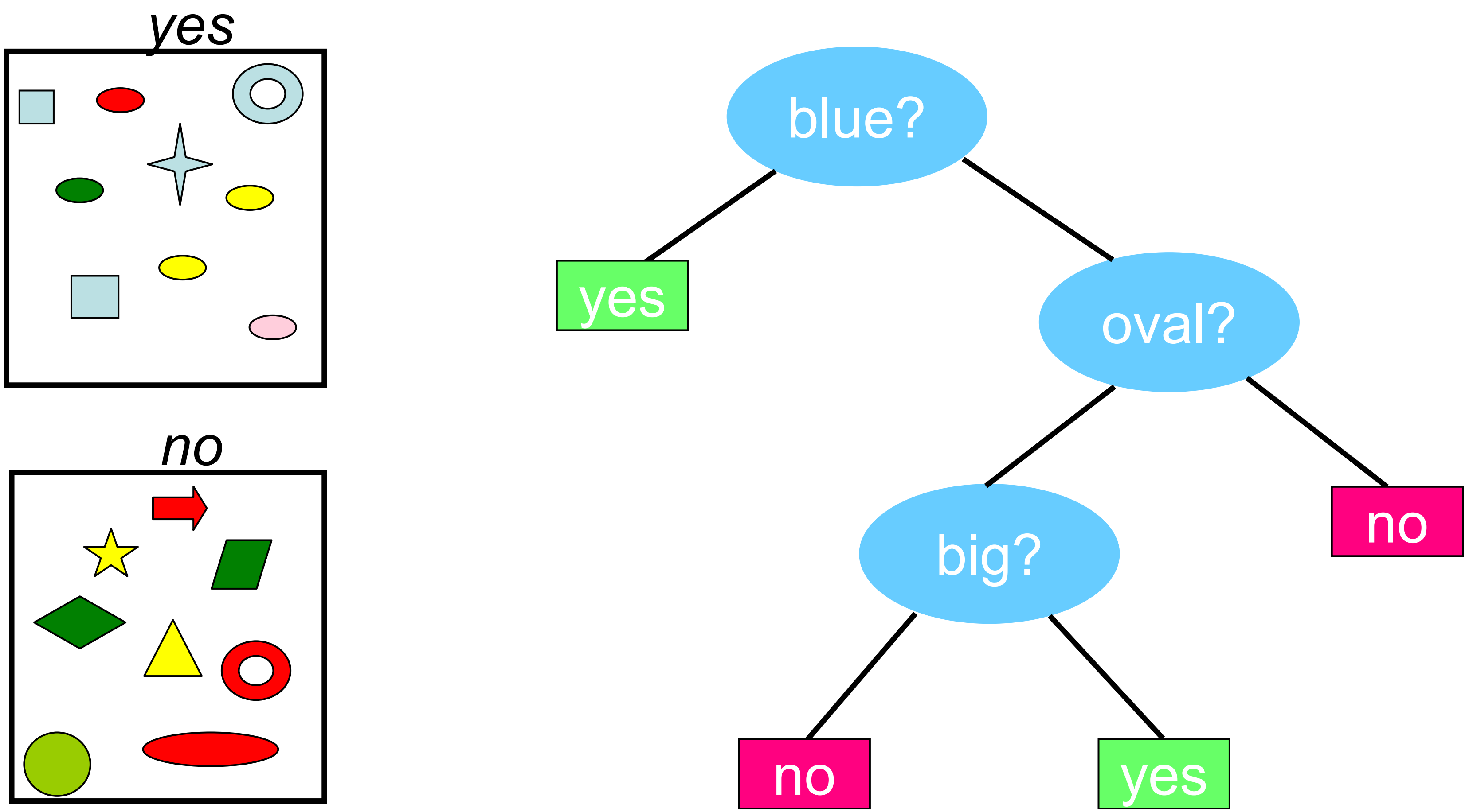
Blue	Crescent	Small
Yellow	Ring	Small

?
?

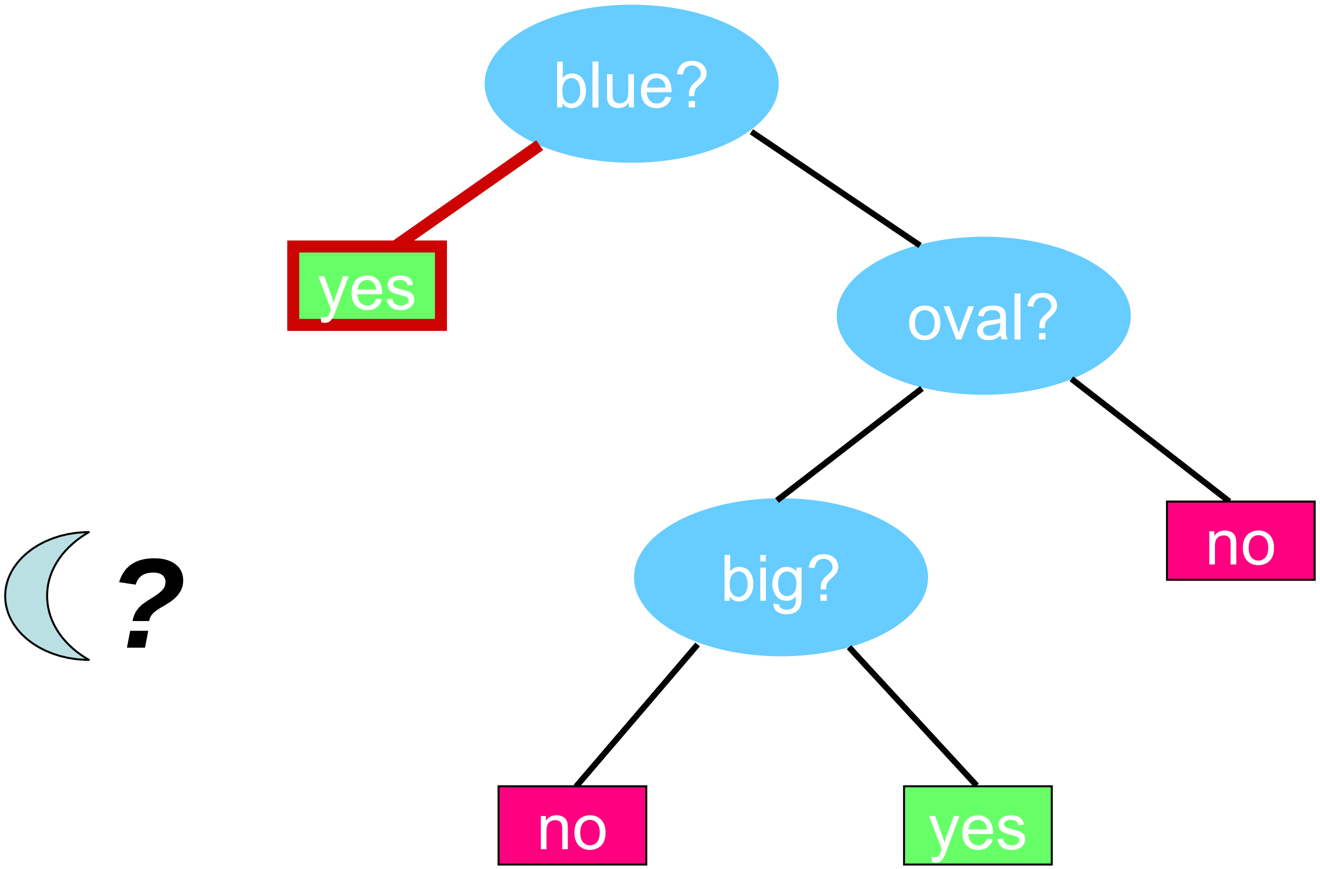
Notation

- Alpaydin book uses x^t (d-dimensional) to denote t'th training input, and r^t to denote t'th training output (response), for t=1:N
- Bishop book uses x_n (d-dimensional) for n'th input, and t_n for n'th output (target), n=1:N
- Hastie book uses x_i (p-dimensional) for i'th covariate, and y_i for i'th output, i=1:n
- We will often omit vector notation **x**_i
- Please do not let notation obscure the ideas!

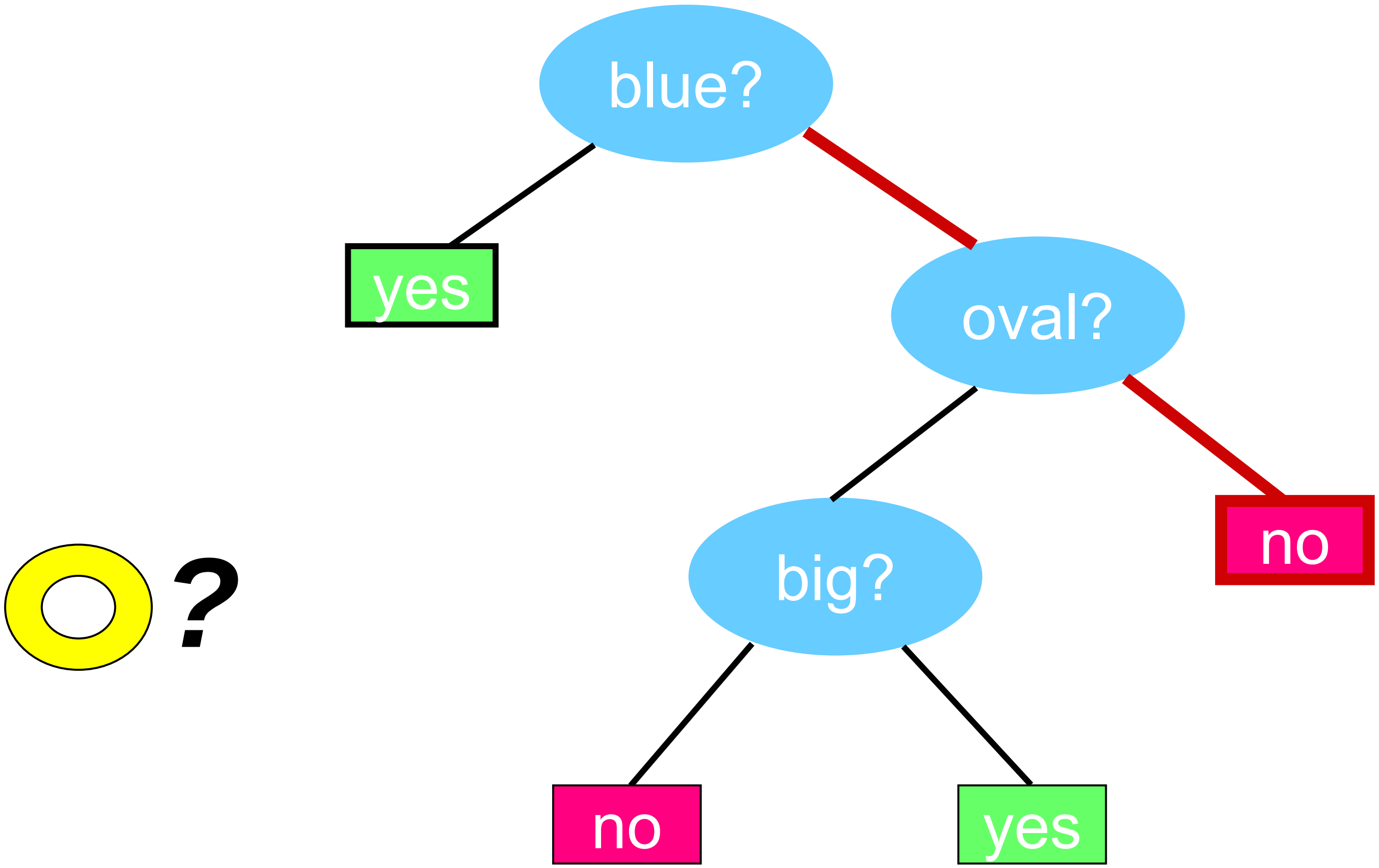
Hypothesis (decision tree)



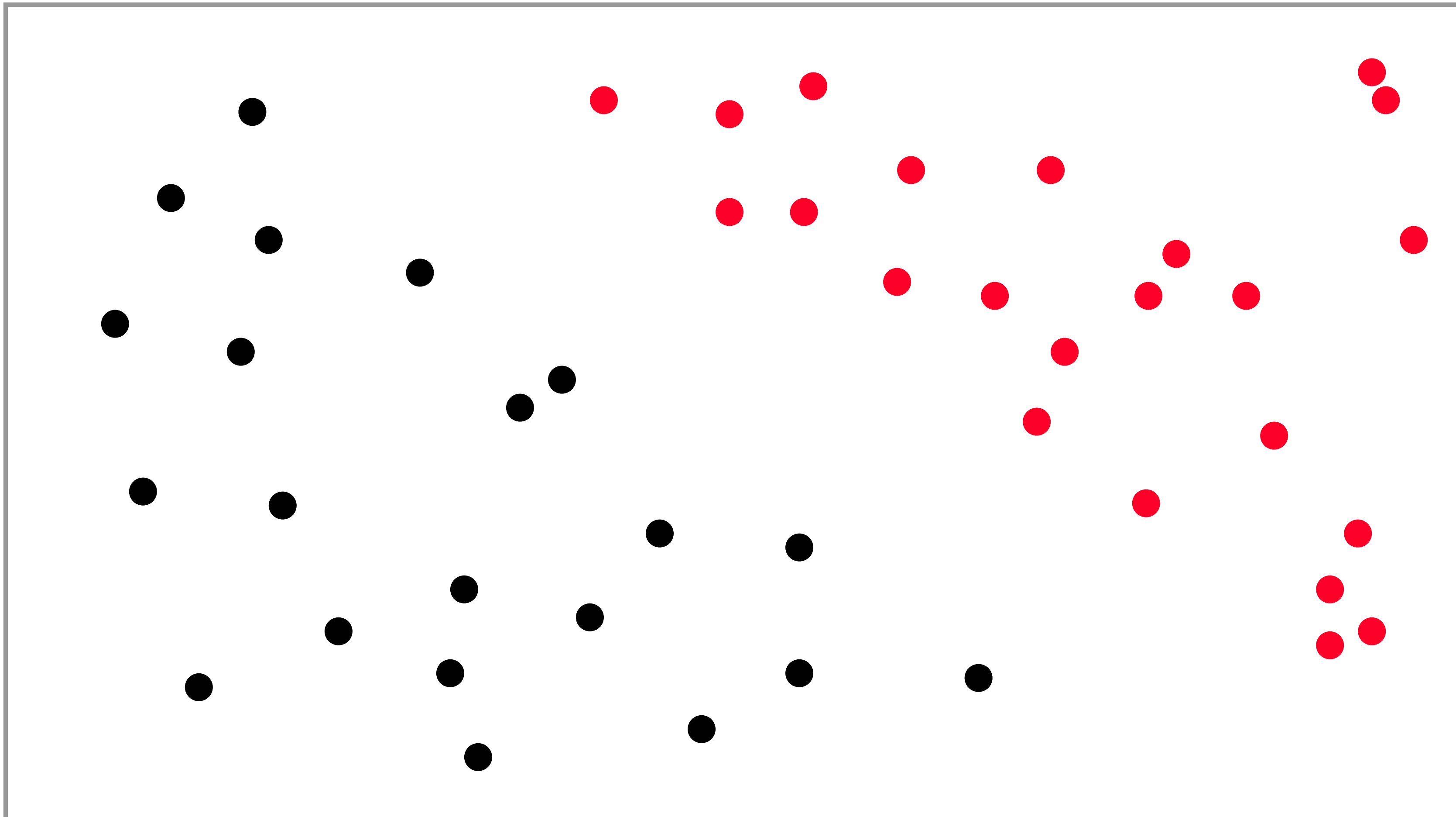
Decision Tree



Decision Tree

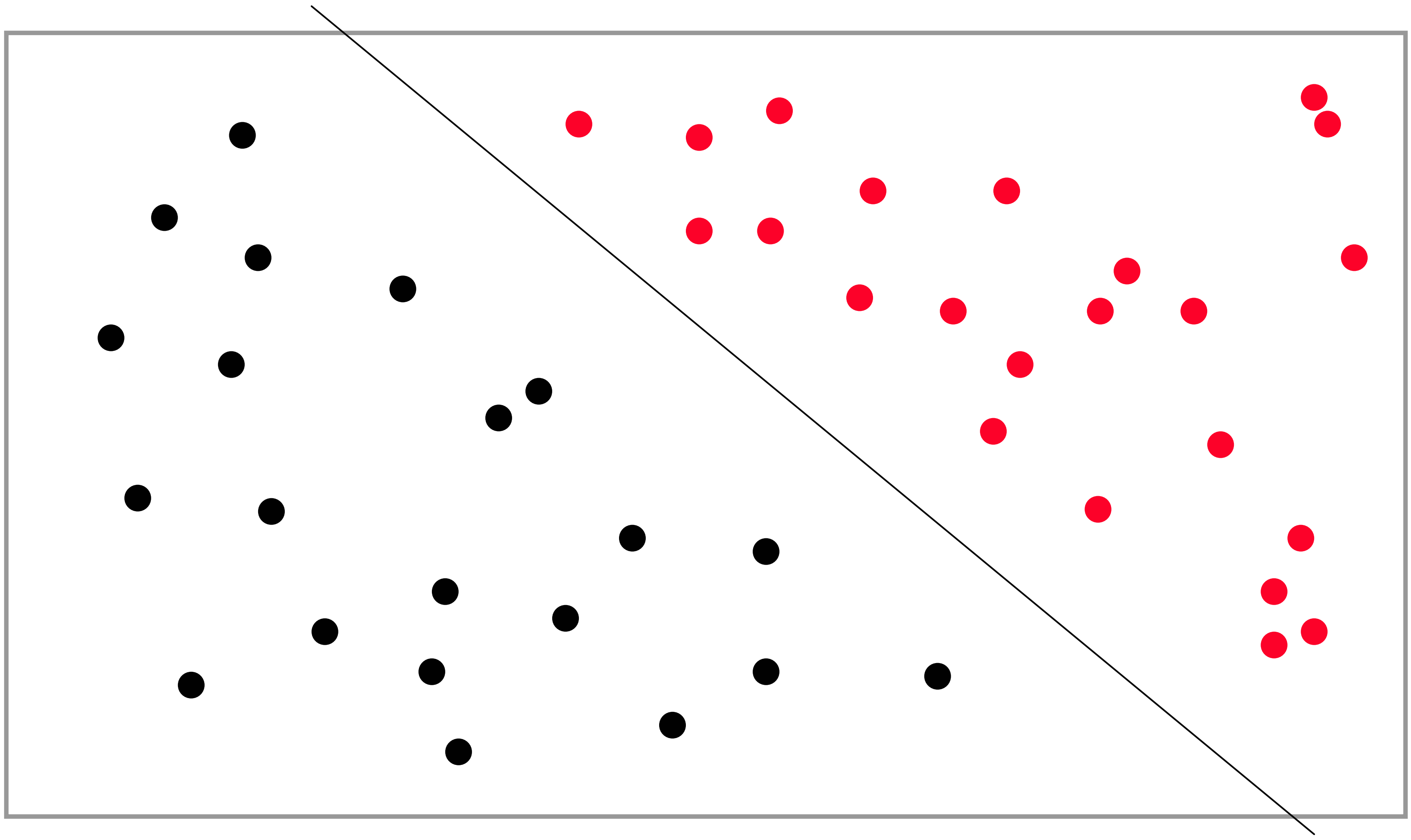


What's the right hypothesis?

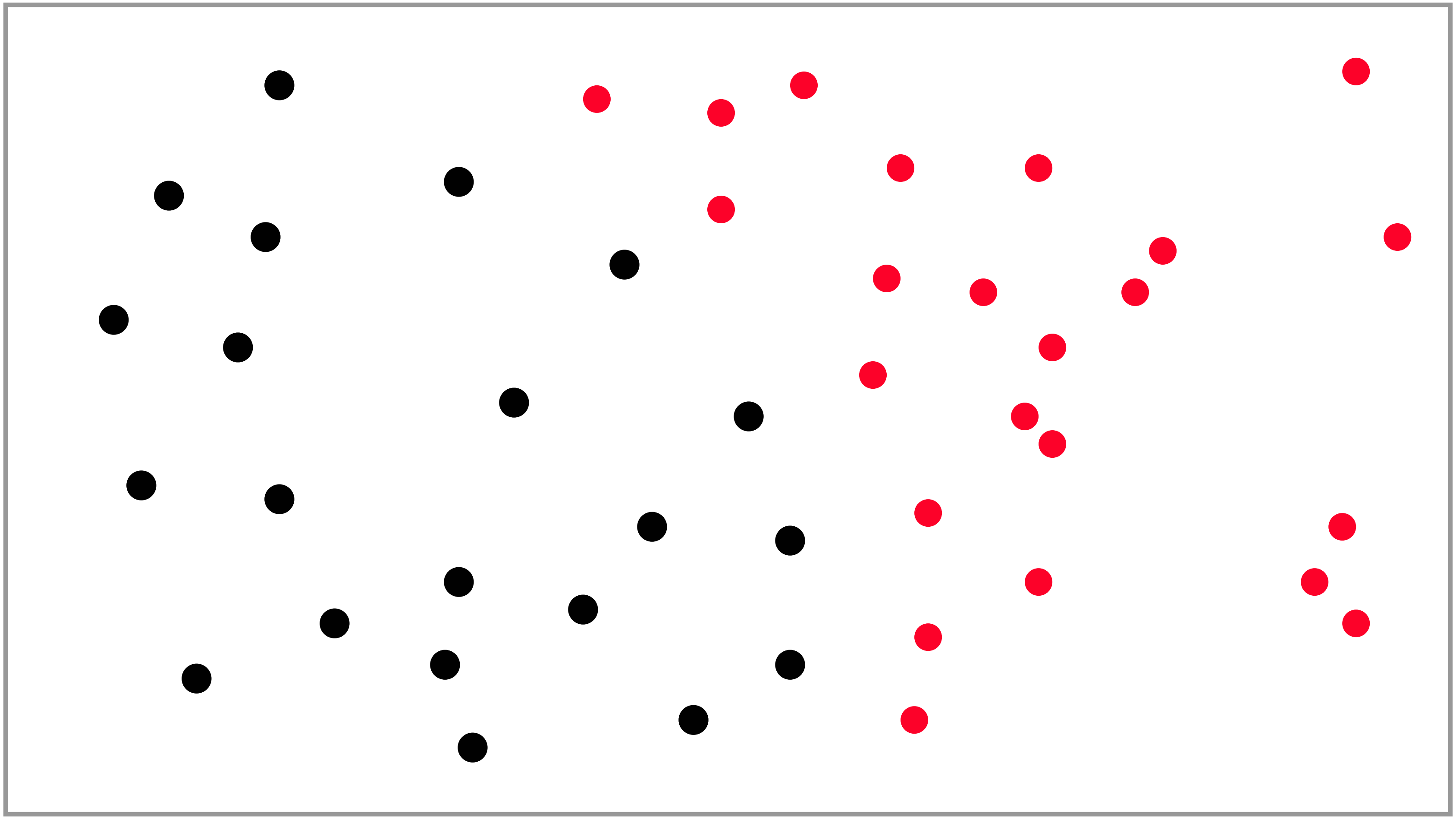


What's the right hypothesis?

- Linearly separable data

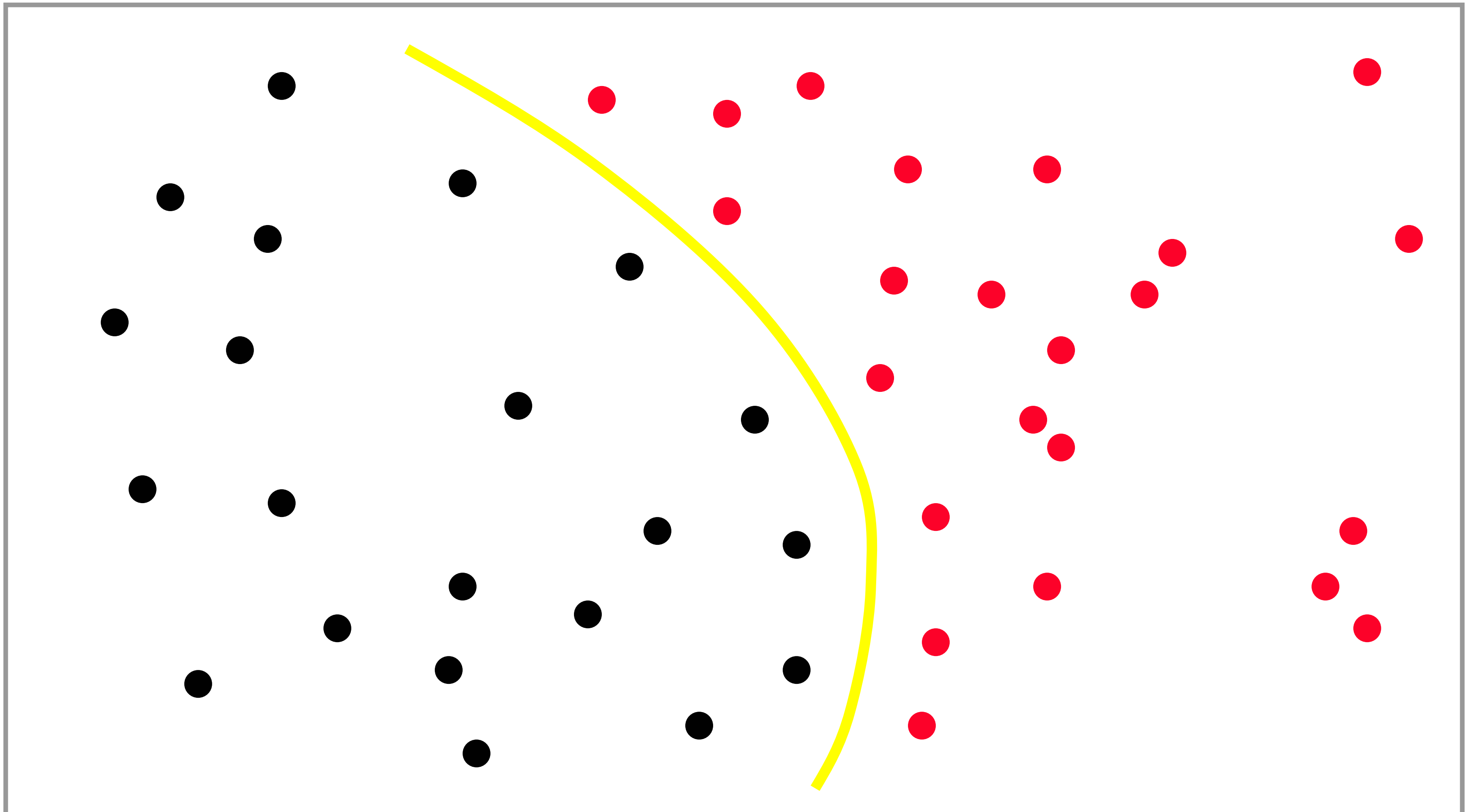


How about now?

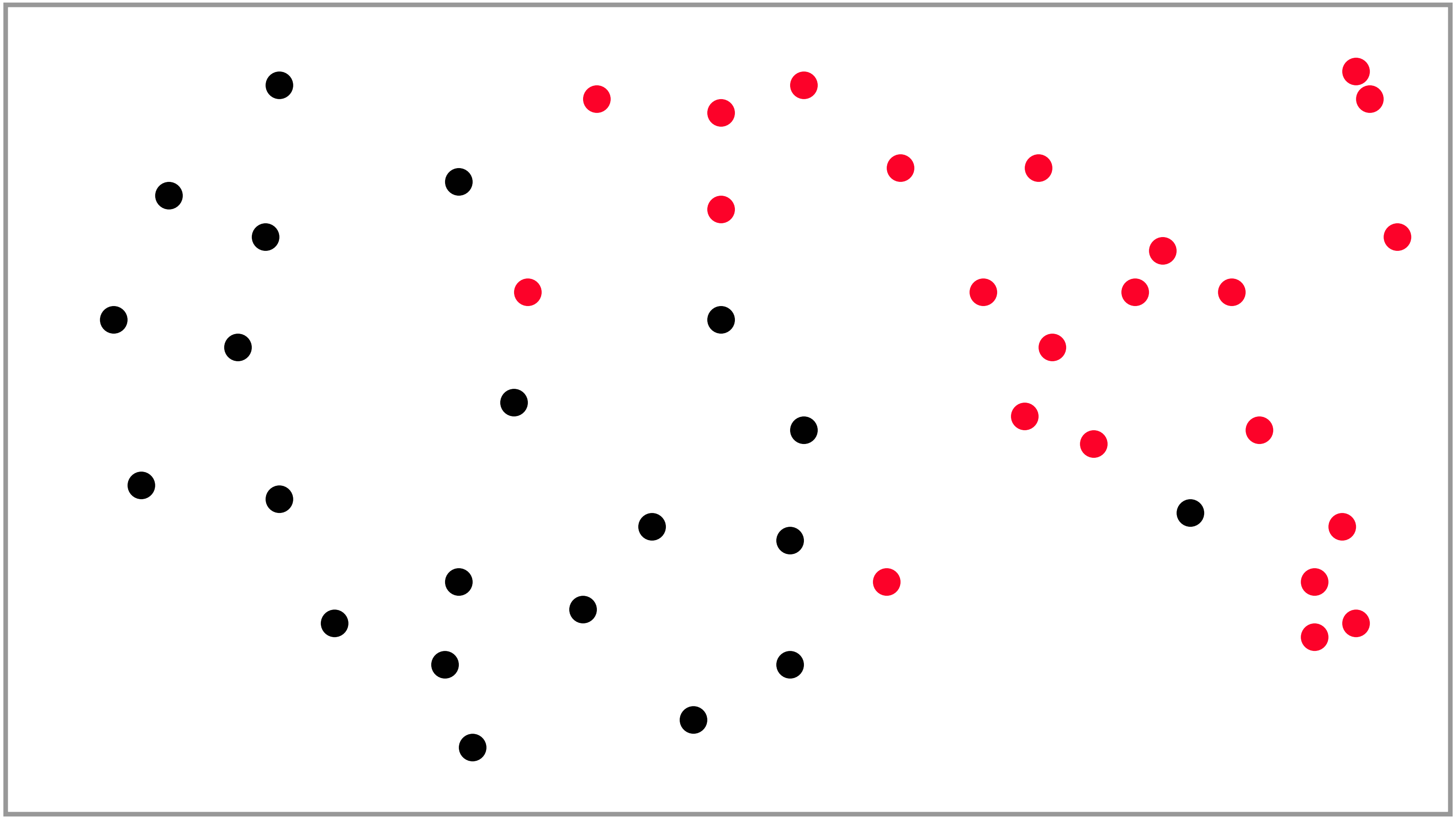


How about now?

- Quadratically separable data

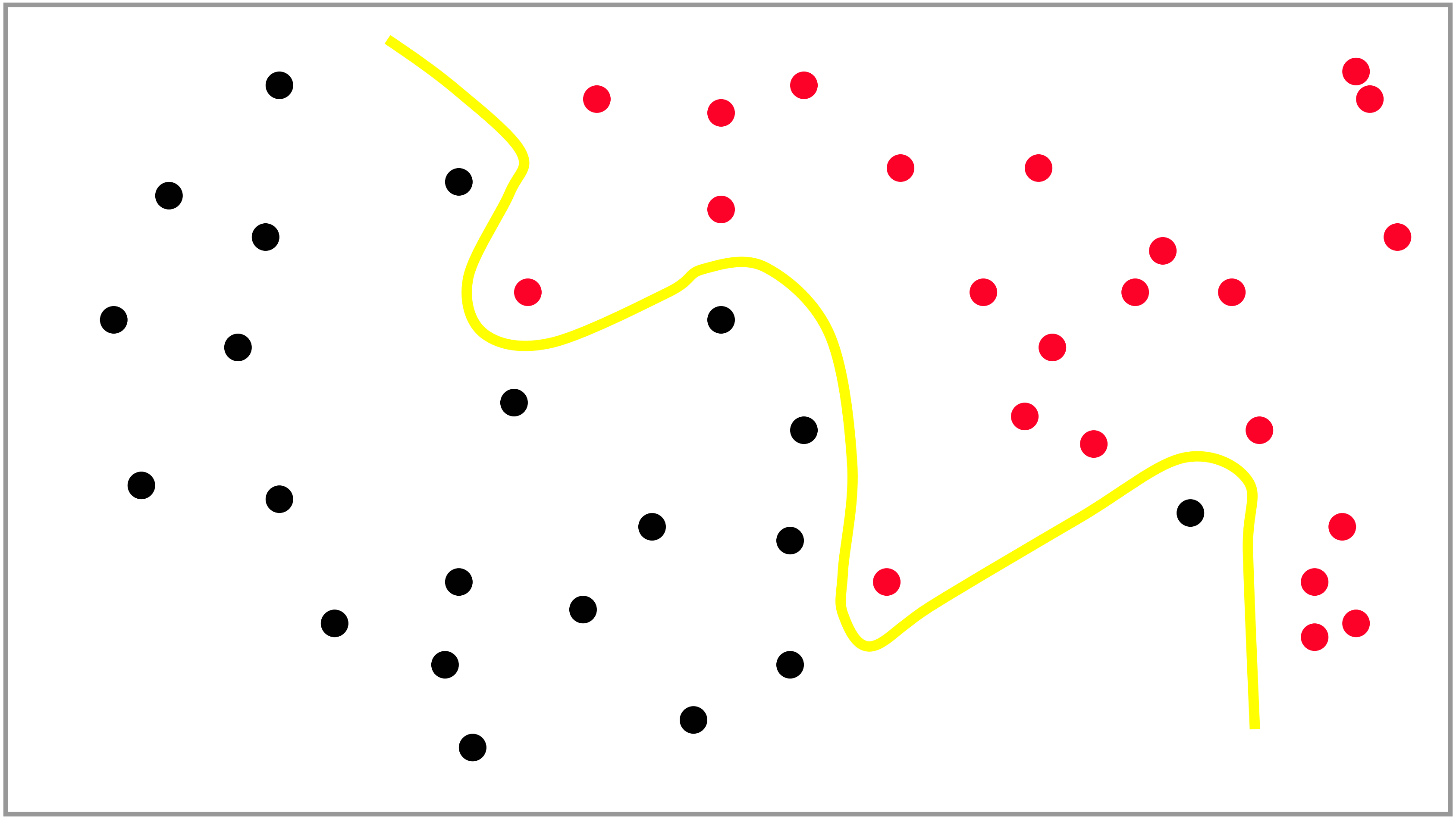


Noisy/ mislabeled data



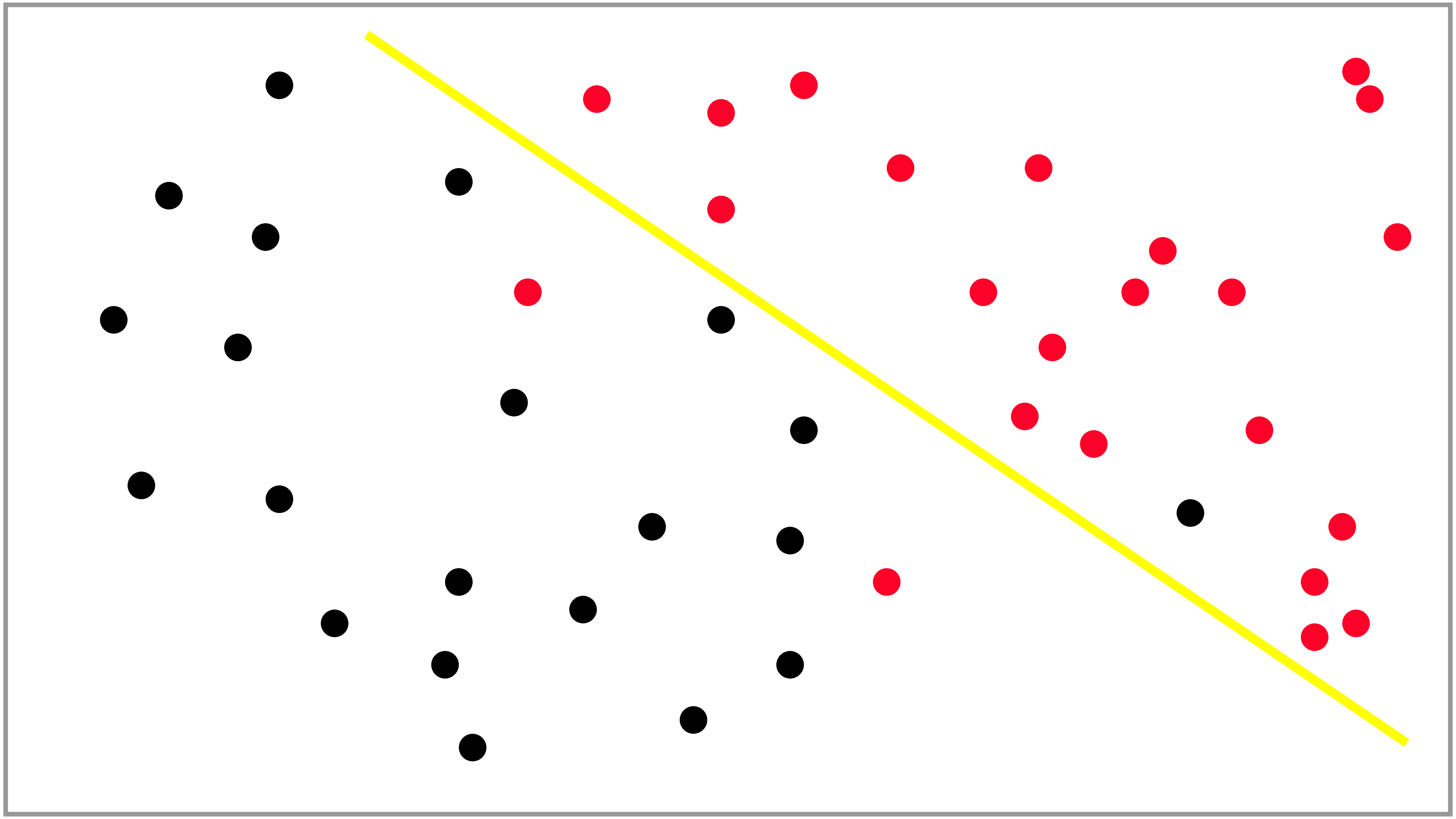
Overfitting

- Memorizes irrelevant details of training set

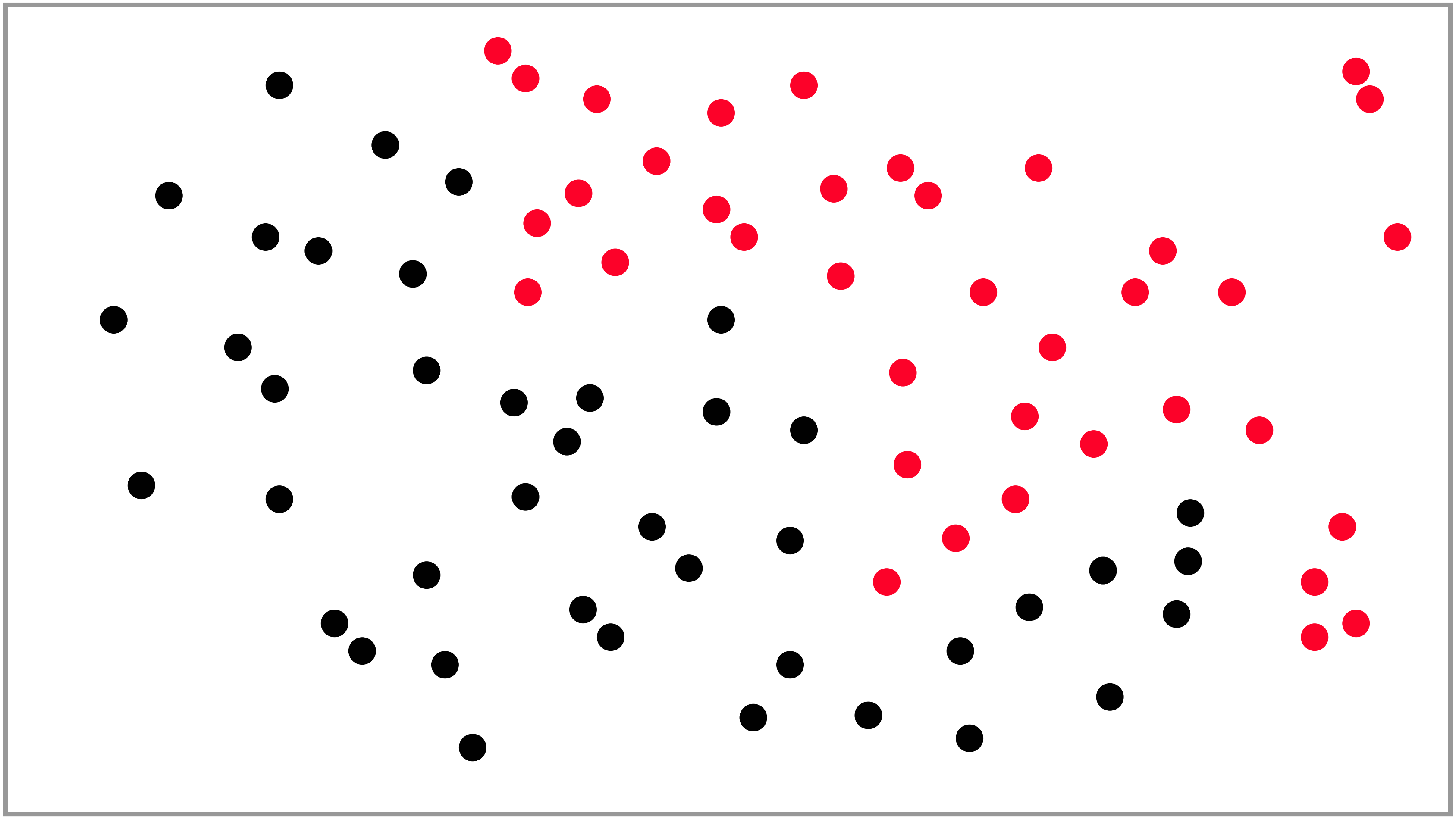


Underfitting

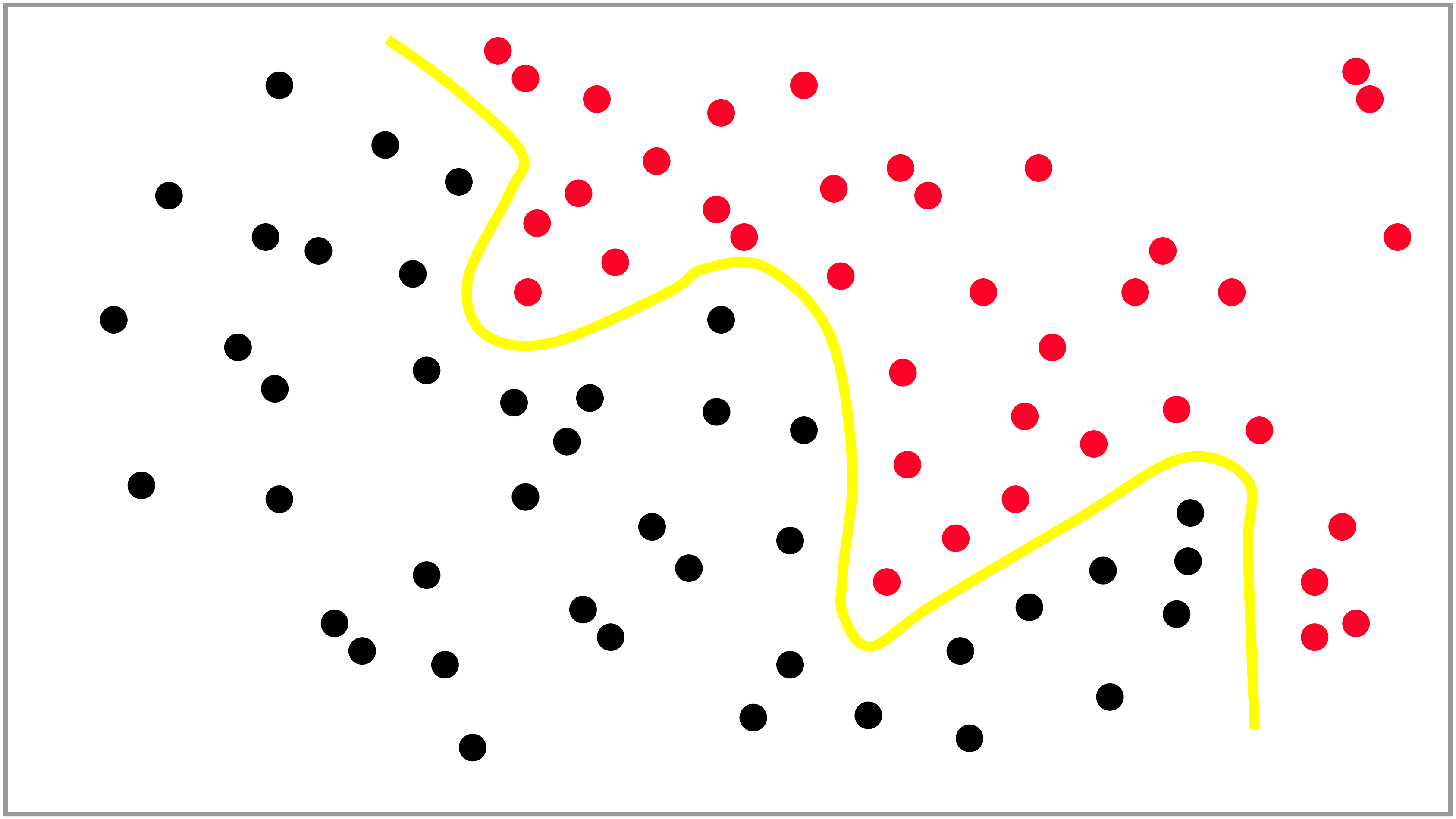
- Ignores essential details of training set



Larger data set



Now more complex hypothesis is ok



No free lunch theorem

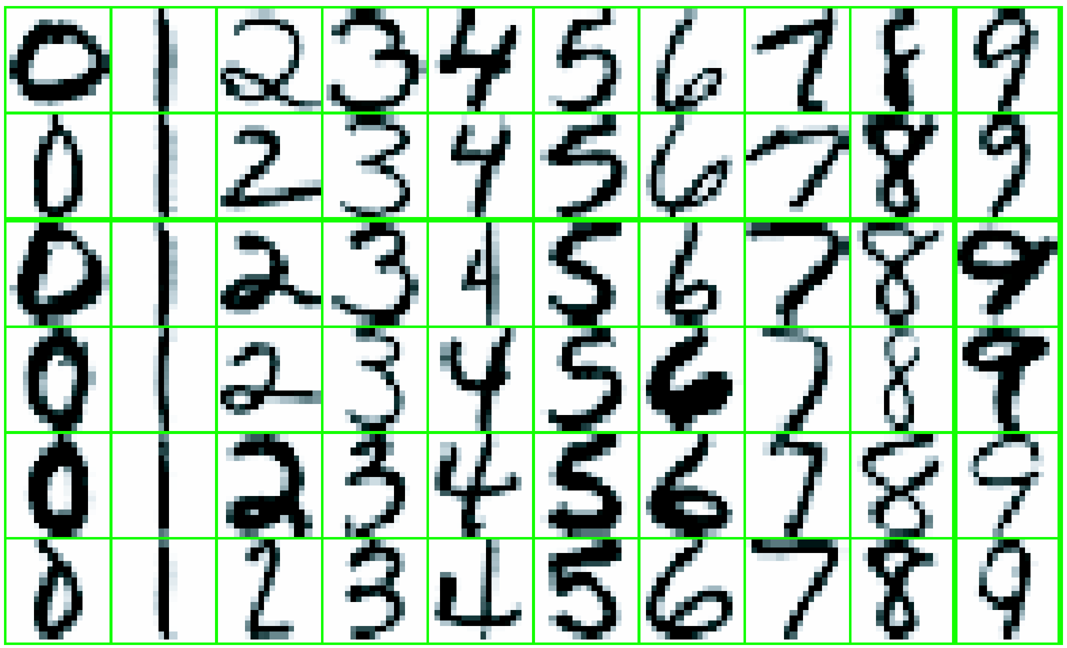
- Unless you know something about the distribution of problems your learning algorithm will encounter, ***any hypothesis that agrees with all your data is as good as any other***.
- You have to make *assumptions* about the underlying future.
- These assumptions are implicit in the choice of hypothesis space (and maybe the algorithm).
- Hence learning is inductive, not deductive.

Supervised learning methods

- Methods differ in terms of
  - The form of hypothesis space they use
  - The method they use to find the best hypothesis given data
- There are many successful approaches
  - Neural networks
  - Decision trees
  - Support vector machines (SVMs)
  - Gaussian processes
  - Boosting
  - etc

Handwritten digit recognition

- x^t 〉\ $R^{16 \text{ § } 16}$, y^t 〉\ {0,...,9}



Face Recognition

Training examples of a person

Test images

AT&T Laboratories, Cambridge UK
http://www.uk.research.att.com/facedatabase.html

Linear regression

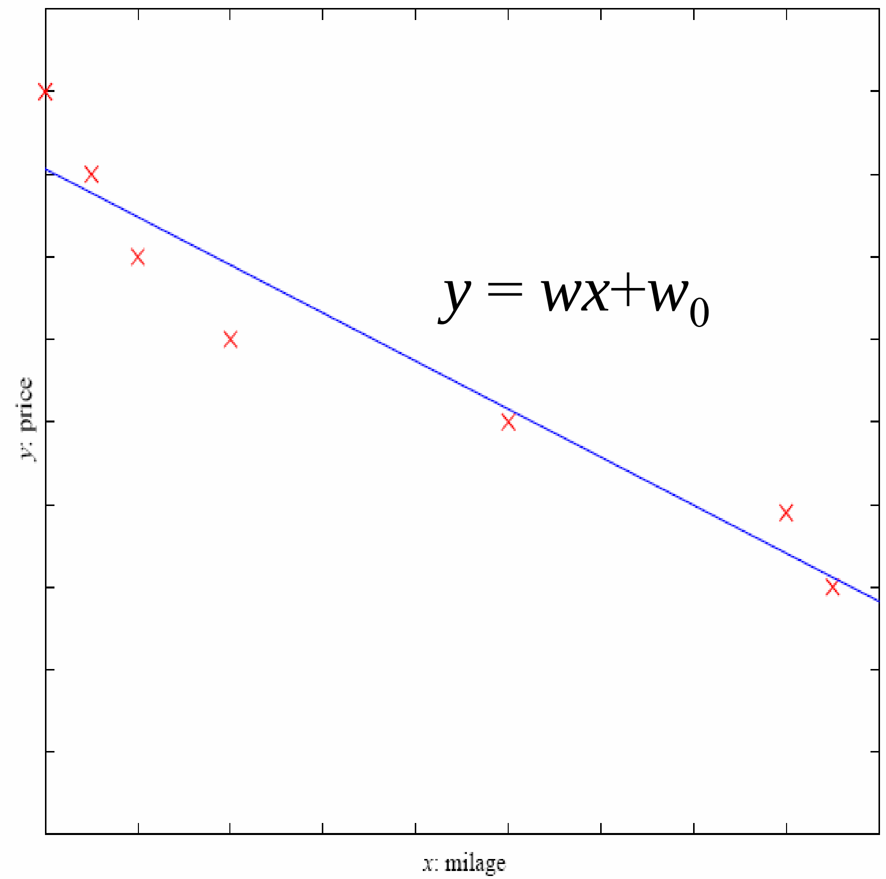
- Example: Price of a used car
- x : car attributes

  y : price

  $$y = g\,(x \mid \theta\,)$$

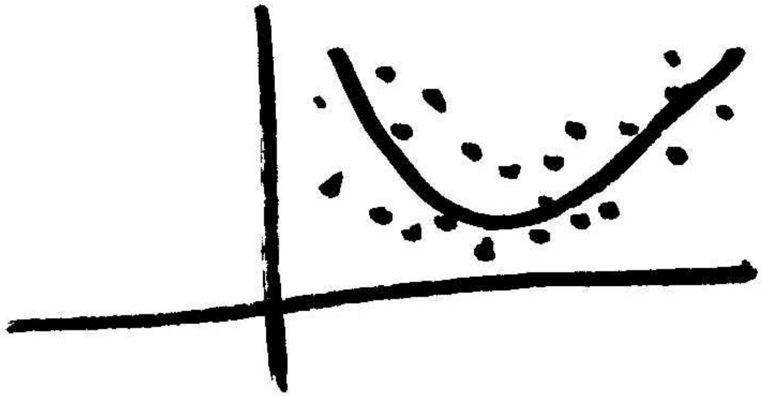
  g ( ) model,

  $\theta = (w, w_0)$ parameters



Regression is like classification except the output is a real-valued scalar

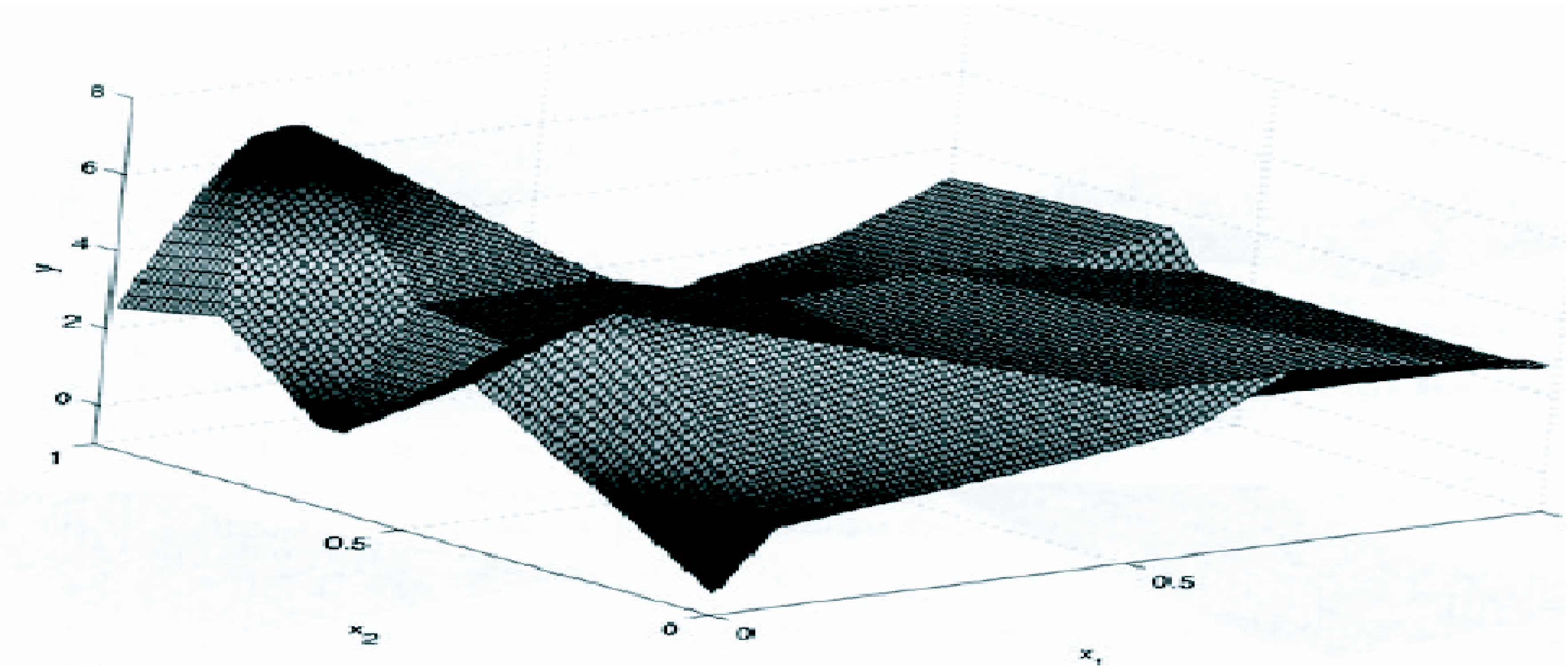
Polynomial regression



$$\begin{aligned} y &= w_0 + w_1 x + w_2 x^2 \\ &= w^T [1, x, x^2] \\ &= w^T \phi(x) \end{aligned}$$

Polynomial regression is linear regression with polynomial basis functions

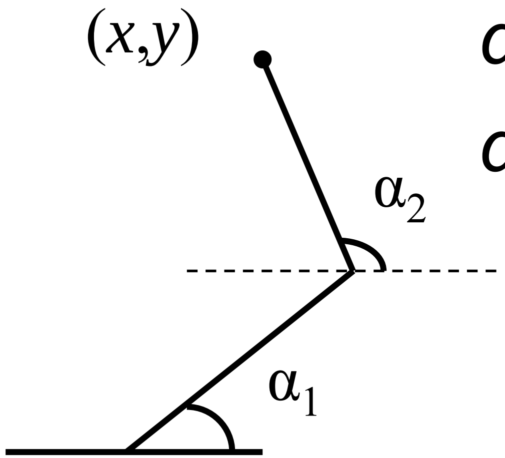
Piecewise linear 2D regression



Now the basis functions $\phi(x_1,x_2)$ must be learned from data:
how many pieces? where the put them? flat or curved?
Much harder problem!

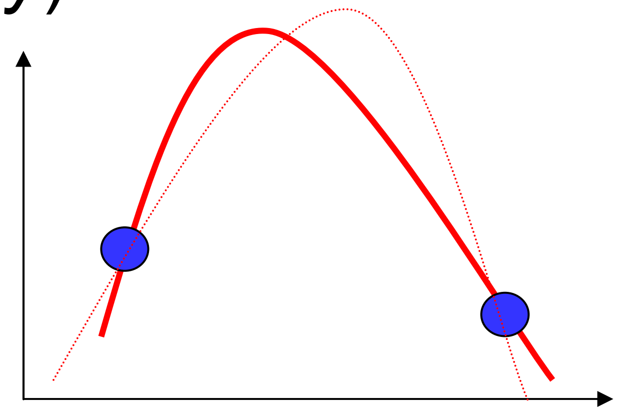
Regression Applications

- Navigating a car: Angle of the steering wheel (CMU NavLab)
- Kinematics of a robot arm



$$\alpha_1 = g_1(x,y)$$
$$\alpha_2 = g_2(x,y)$$

- Response surface design



Supervised Learning: Uses

- Prediction of future cases: Use the rule to predict the output for future inputs
- Knowledge extraction: The rule is easy to understand
- Compression: The rule is simpler than the data it explains
- Outlier detection: Exceptions that are not covered by the rule, e.g., fraud

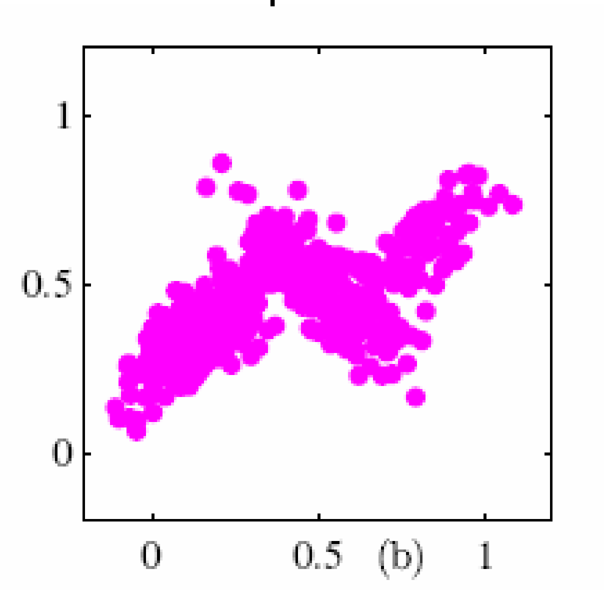
Unsupervised Learning

- Learning “what normally happens”
- No output
- Can be formalized in terms of probability density estimation
- Examples:
  - clustering
  - dimensionality reduction
  - abnormality detection
  - latent variable estimation

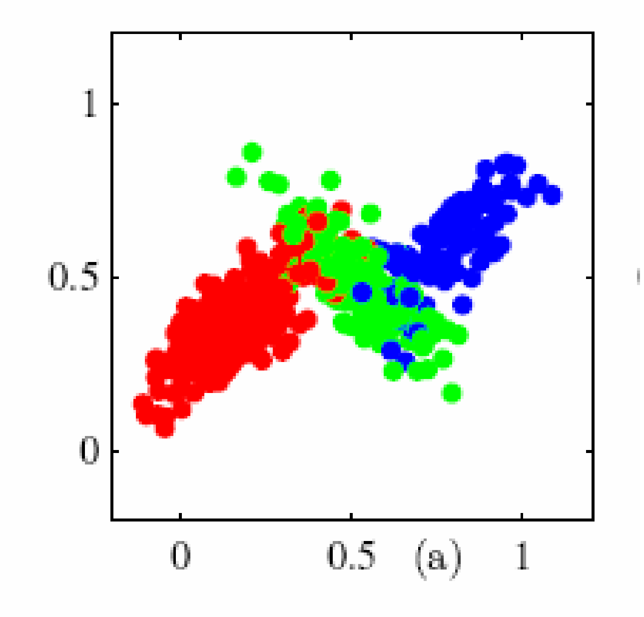
K-means clustering

Desired output

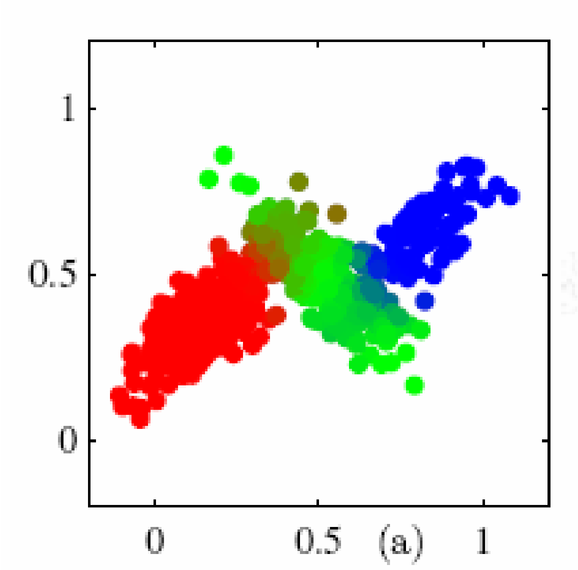
Input



Hard labeling



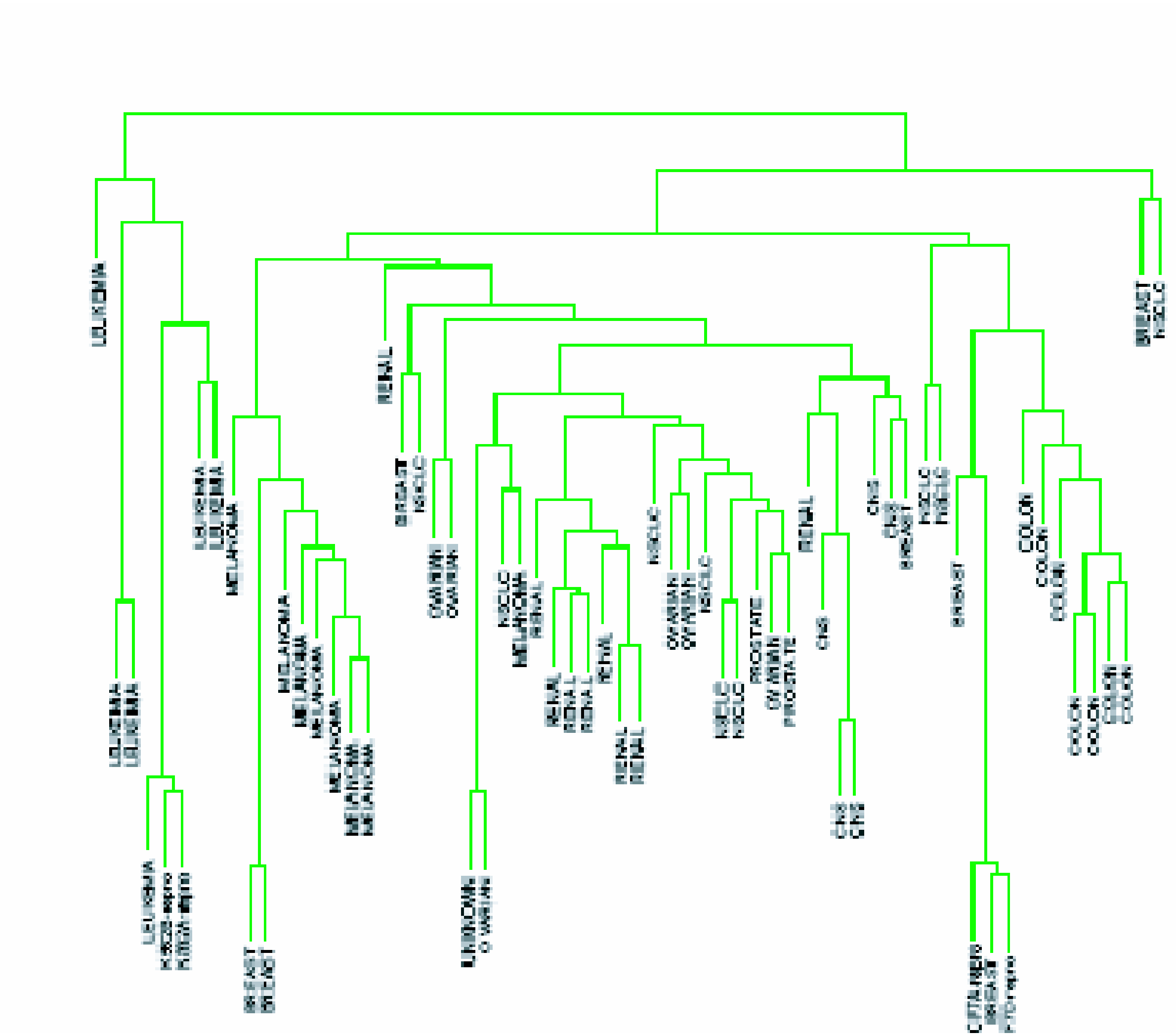
Soft labeling



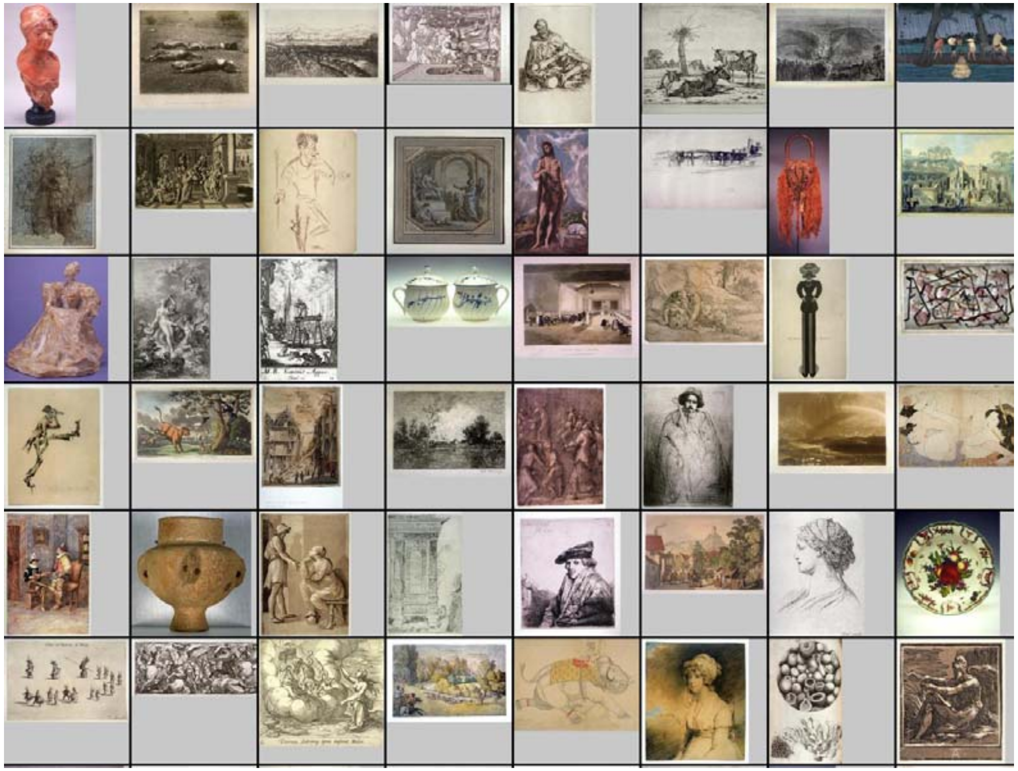
K=3 is the number of clusters, here chosen by hand

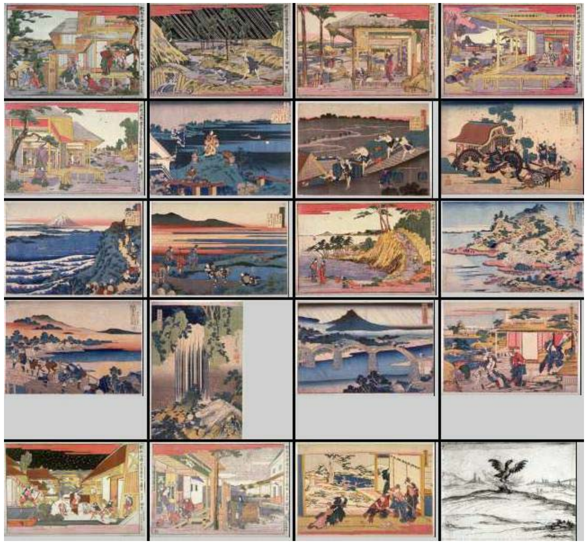
Hierarchical agglomerative clustering

- Greedily build a dendogram



Clustering art

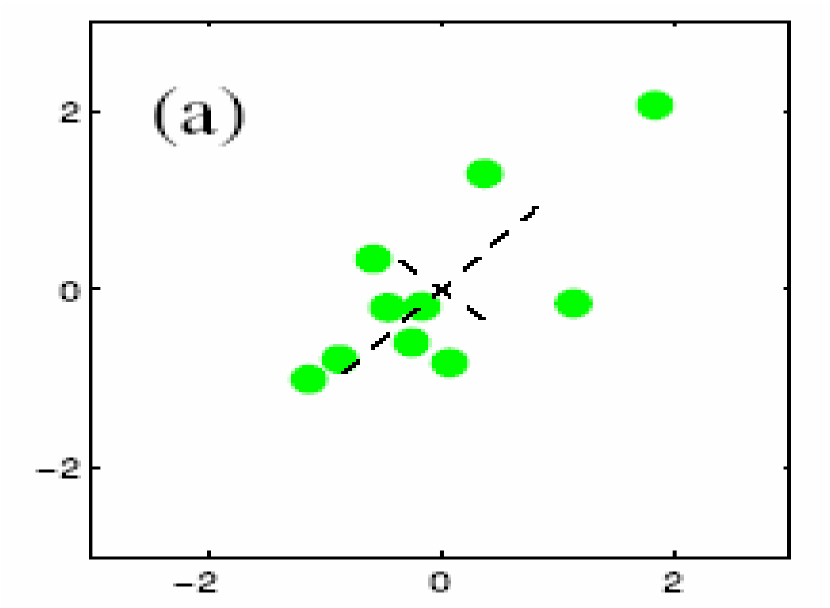




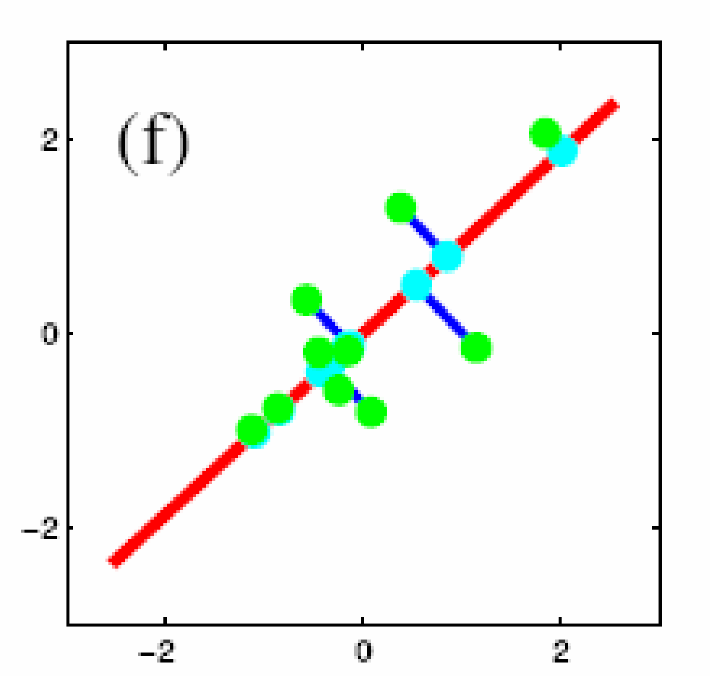
Principal components analysis (PCA)

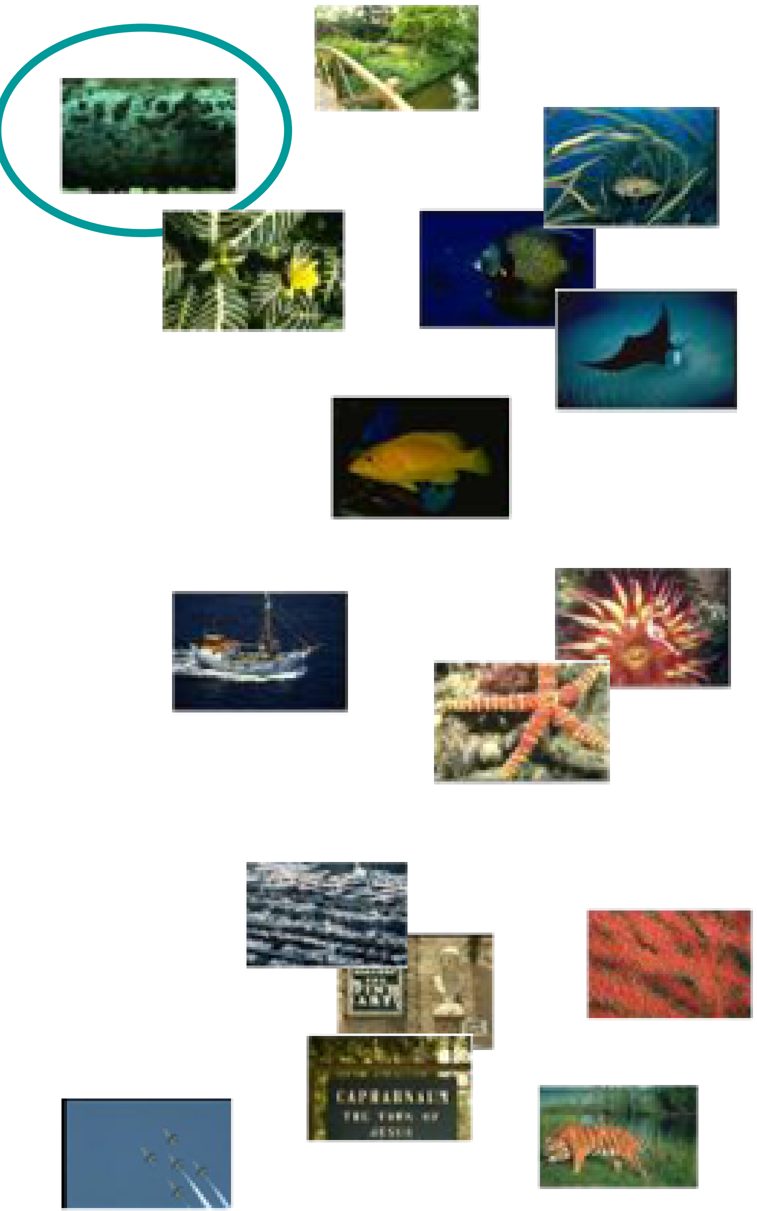
Project high dimensional data into a linear subspace which captures most of the variance of the data

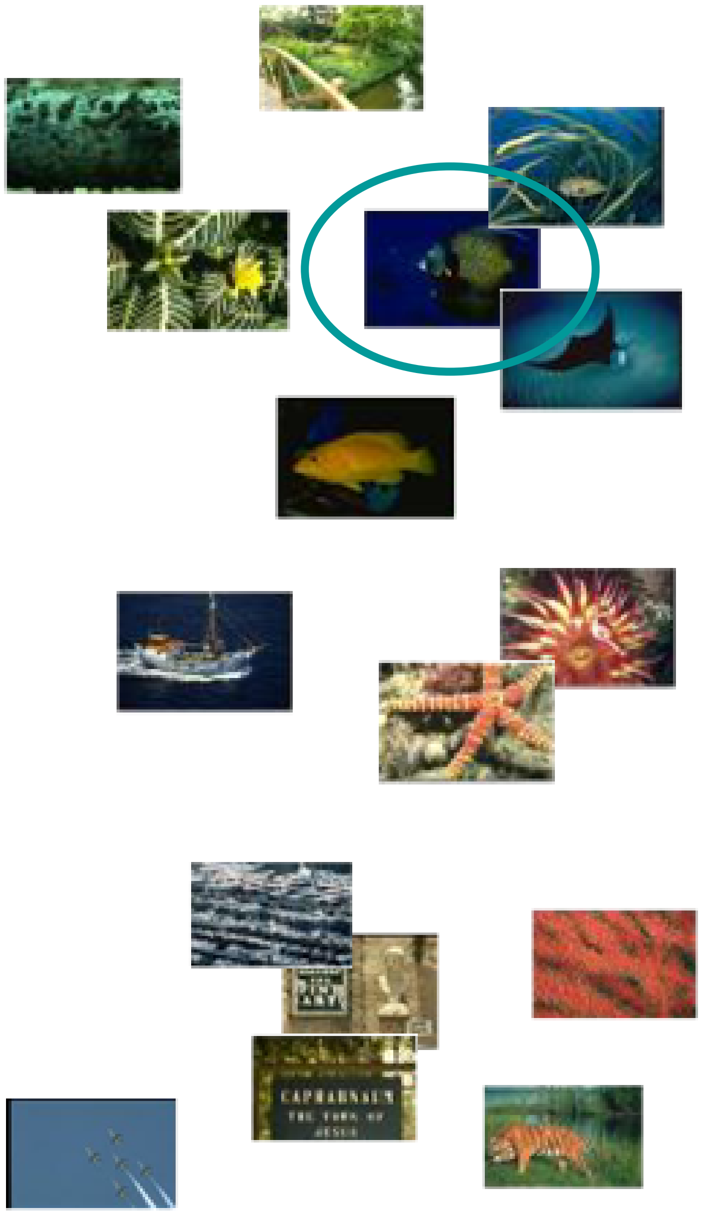
Input



Output

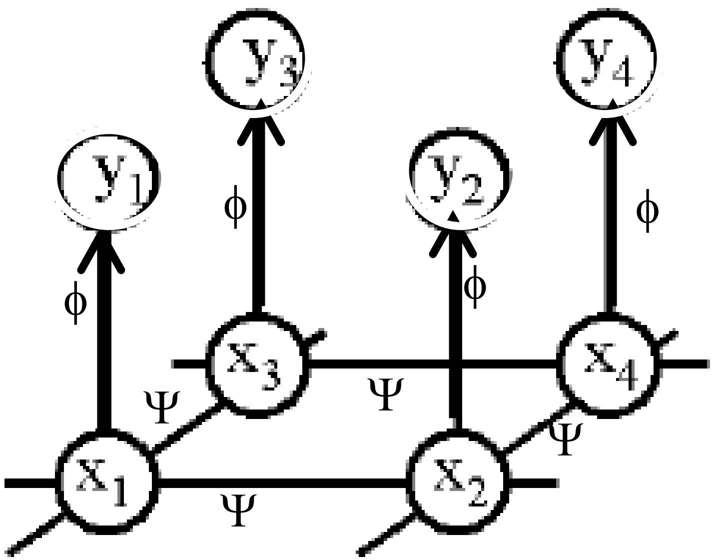


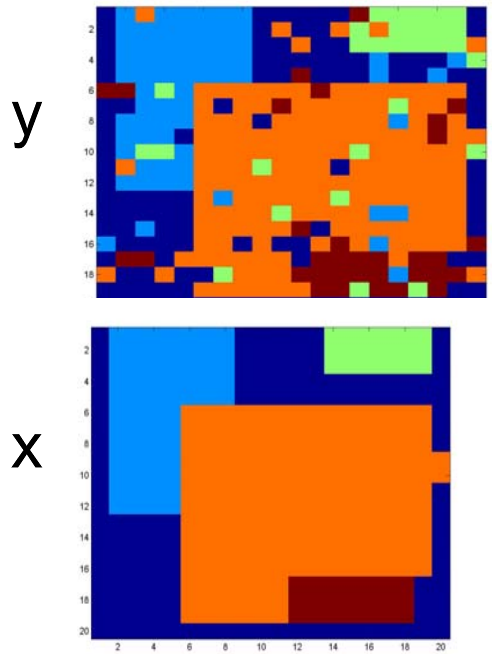

CAPHARNAUM


CAPHARNAUM


CAPHARNAUM

Image denoising with Markov random fields





Popular in:

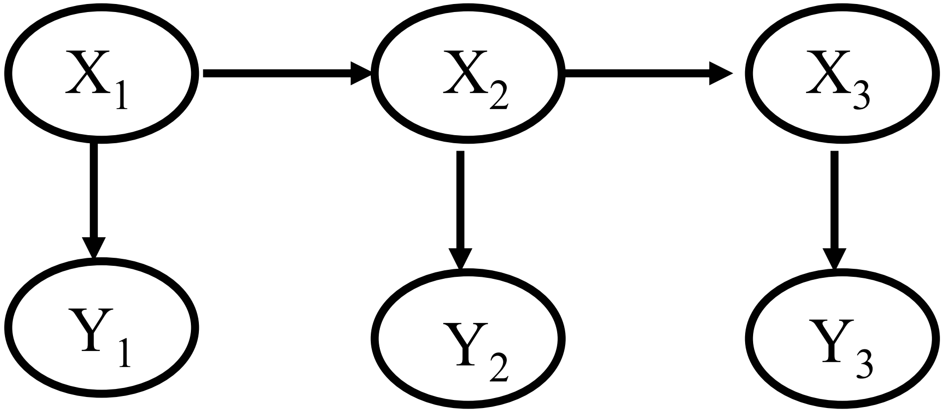
- Computer vision
- Language modeling
- Information extraction
- Sequence prediction
- Graphics

$$P(x,y) = \left[\frac{1}{Z}\prod_{j\in N_i}\psi(x_i,x_j)\right]\prod_i P_\phi(y_i|x_i)$$

Compatibility with neighbors

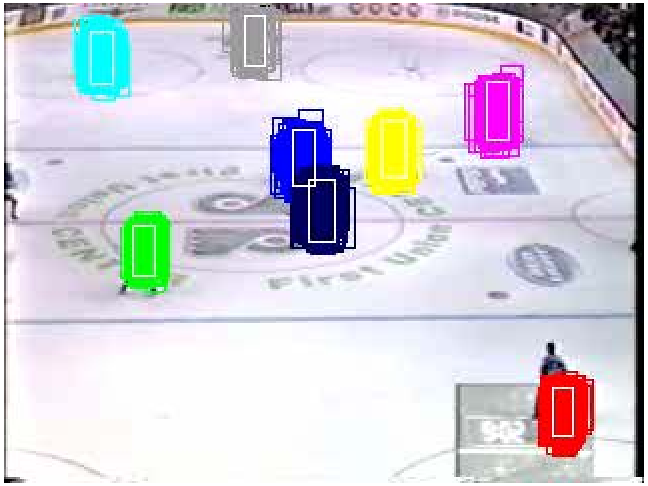
Local evidence (compatibility with image)

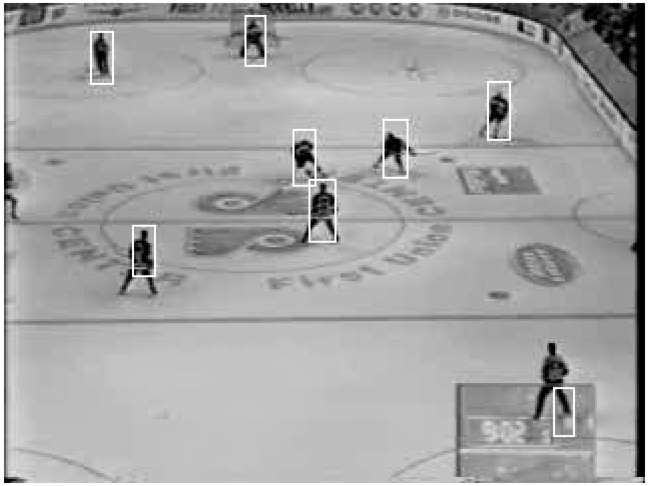
People tracking



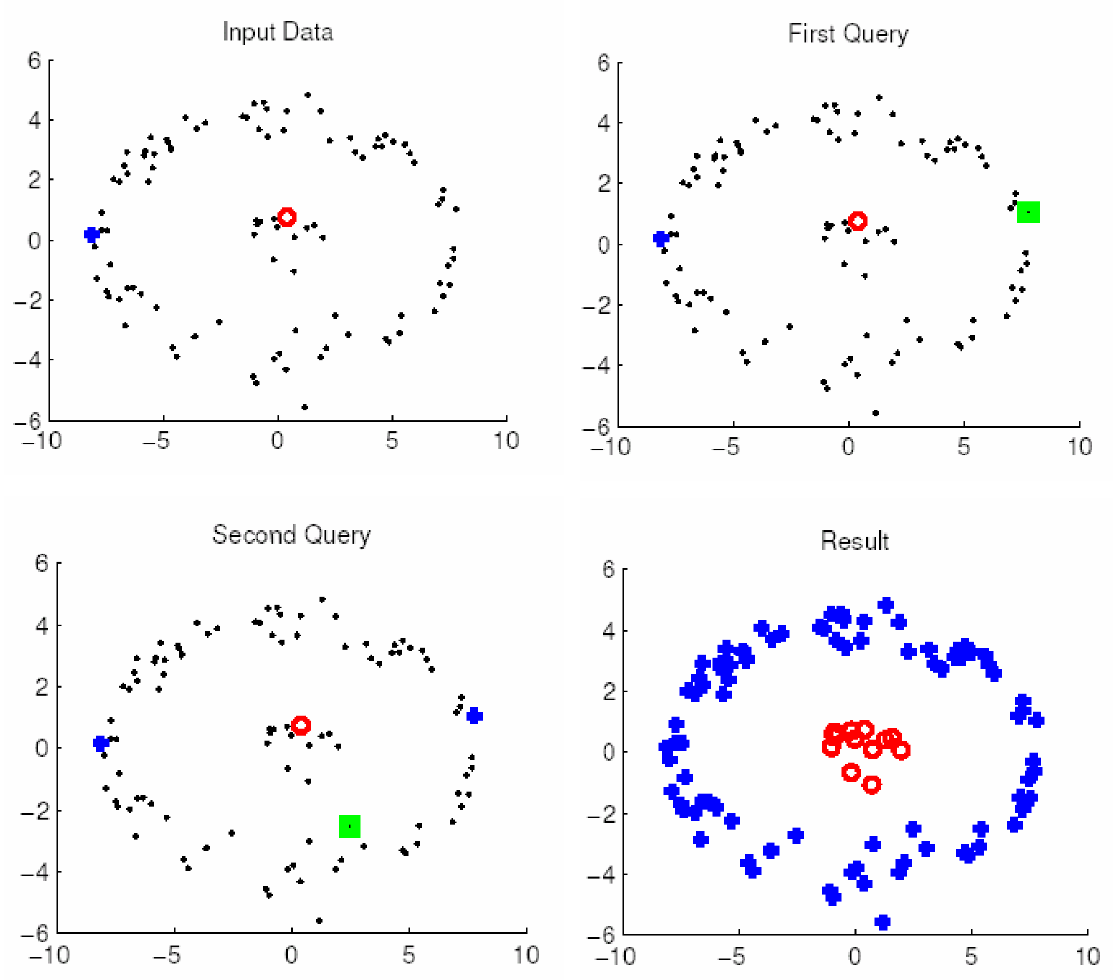
Unknown player location

Observed video frames

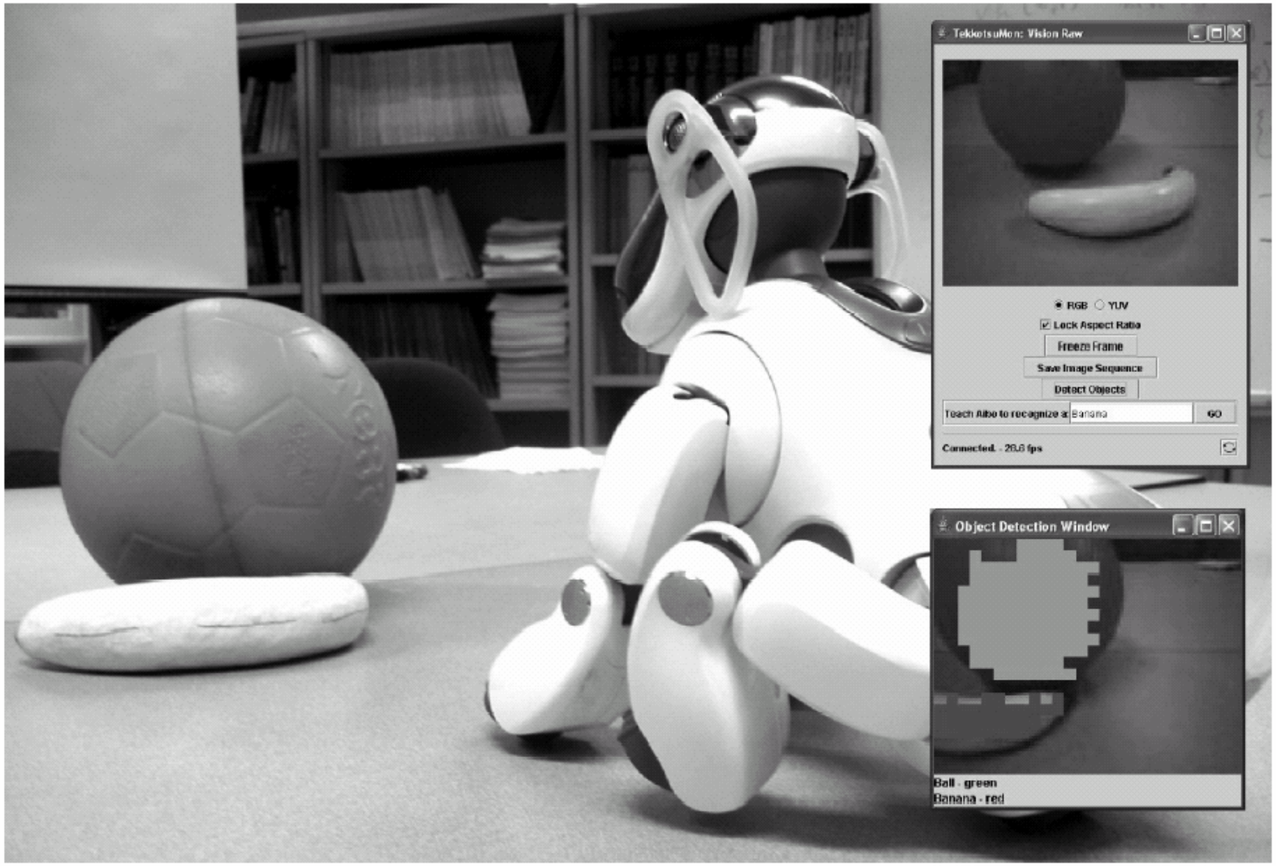




Active learning: asking the right questions



Robots that ask questions and learn



Reinforcement Learning

- Learning a policy: A sequence of outputs
- No supervised output, but delayed reward
- Credit assignment problem: which action led to me winning the game of chess?
- This is covered in CS422 (AI II), not in CS340.