Pattern recognition

Bioinformatics · Universitat Politècnica de Catalunya

Alexandre Perera-Lluna

Introduction to Pattern Recognition: Example 1

10 Patients

- 6 Patients Cancer type A
- 4 Patients Cancer tupe B

4 Genes

a,b,c,d

What is the role of each gene in the cancer type?

Data

Patient#	1	2	3	4	5	6	7	8	9	10
Tumor	A	A	A	A	A	A	B	B	B	B
r(a)	5.00	1.33	3.45	3.05	4.22	2.09	0.33	0.65	0.22	0.12
r(b)	0.98	0.87	1.04	1.10	?	2.11	1.23	1.32	0.85	0.77
r(c)	0.33	1.40	0.42	0.55	0.24	0.60	2.44	?	3.00	2.22
r(d)	0.89	0.90	1.00	0.92	0.66	1.05	1.32	1.01	0.97	0.87

Note the missing values…

Data

Patient#	1	2	3	4	5	6	7	8	9	10
Tumor	A	A	A	A	A	A	B	B	B	B
r(a)	5.00	1.33	3.45	3.05	4.22	2.09	0.33	0.65	0.22	0.12
r(b)	0.98	0.87	1.04	1.10	?	2.11	1.23	1.32	0.85	0.77
r(c)	0.33	1.40	0.42	0.55	0.24	0.60	2.44	?	3.00	2.22
r(d)	0.89	0.90	1.00	0.92	0.66	1.05	1.32	1.01	0.97	0.87

Data

Patient#	1	2	3	4	5	6	7	8	9	10
Tumor	A	A	A	A	A	A	B	B	B	B
r(a)	5.00	1.33	3.45	3.05	4.22	2.09	0.33	0.65	0.22	0.12
r(b)	0.98	0.87	1.04	1.10	?	2.11	1.23	1.32	0.85	0.77
r(c)	0.33	1.40	0.42	0.55	0.24	0.60	2.44	?	3.00	2.22
r(d)	0.89	0.90	1.00	0.92	0.66	1.05	1.32	1.01	0.97	0.87

Data

Patient#	1	2	3	4	5	6	7	8	9	10
Tumor	A	A	A	A	A	A	B	B	B	B
r(a)	5.00	1.33	3.45	3.05	4.22	2.09	0.33	0.65	0.22	0.12
r(b)	0.98	0.87	1.04	1.10	?	2.11	1.23	1.32	0.85	0.77
r(c)	0.33	1.40	0.42	0.55	0.24	0.60	2.44	?	3.00	2.22
r(d)	0.89	0.90	1.00	0.92	0.66	1.05	1.32	1.01	0.97	0.87

Note that r(a) and r(c) are both negatively co-regulated…

Data and Goals

Patient#	1	2	3	4	5	6	7	8	9	10
Tumor	A	A	A	A	A	A	B	B	B	B
r(a)	5.00	1.33	3.45	3.05	4.22	2.09	0.33	0.65	0.22	0.12
r(b)	0.98	0.87	1.04	1.10	?	2.11	1.23	1.32	0.85	0.77
r(c)	0.33	1.40	0.42	0.55	0.24	0.60	2.44	?	3.00	2.22
r(d)	0.89	0.90	1.00	0.92	0.66	1.05	1.32	1.01	0.97	0.87

- Could we guess the most involved genes in the Cancer process differentiation?
- Given a readout with the expression values for a,b,c,d genes, can we predict whether a new patient has tumor A or B?

Pattern recognition

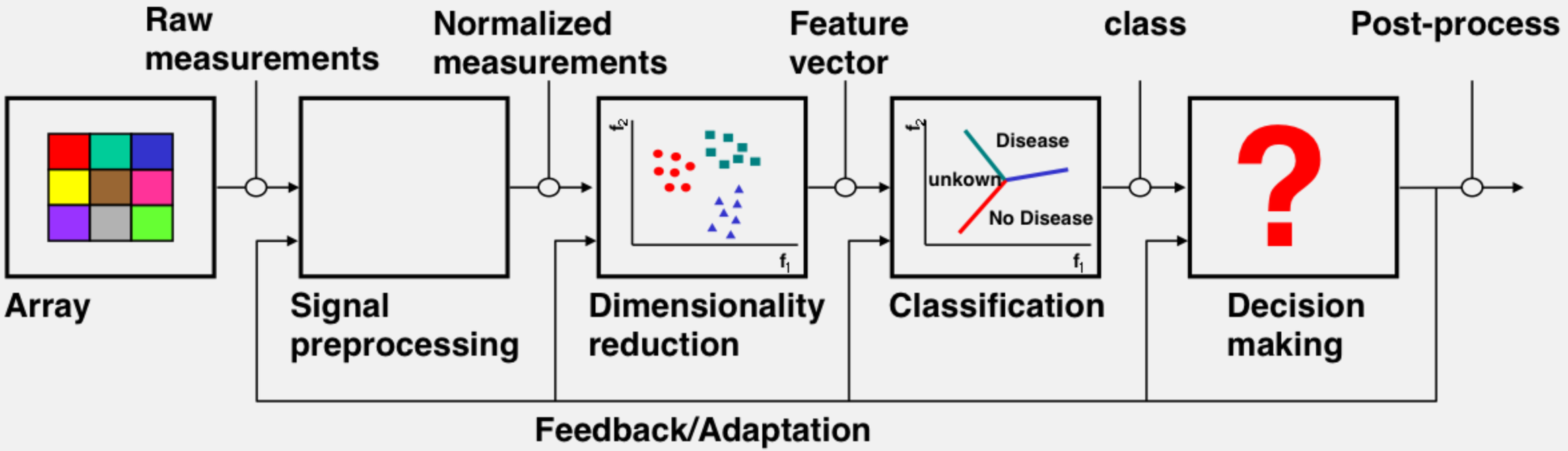
- **Classification:** predict a categorical label from labelled examples.
- **Regression:** predict a numerical response.
- **Clustering:** group observations without using known class labels.

The representation, target, and evaluation method depend on the scientific question.

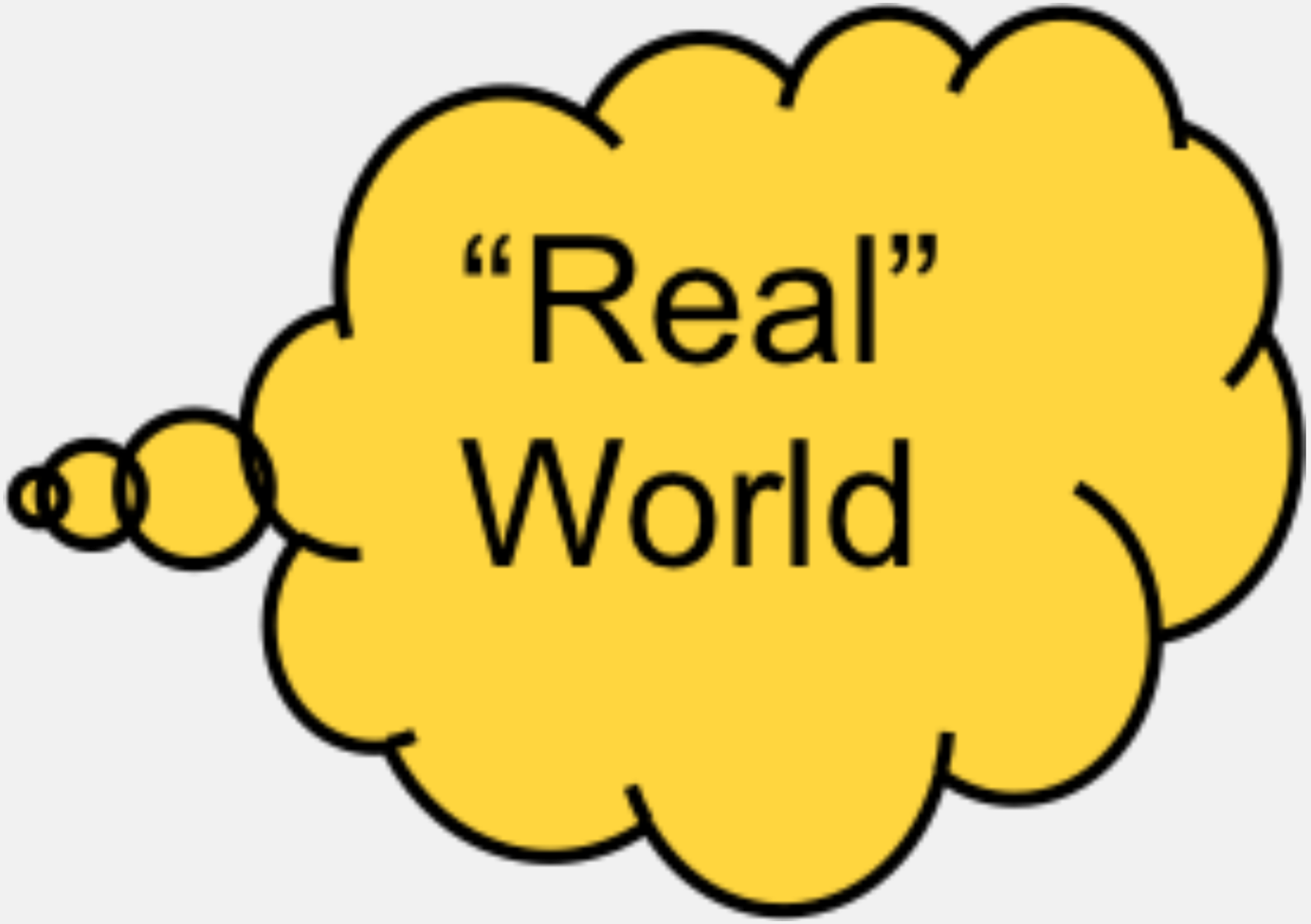
Pattern Recognition Scheme in classification (1/3)



Pattern Recognition Scheme in classification (2/3)



Pattern Recognition Scheme in classification (3/3)



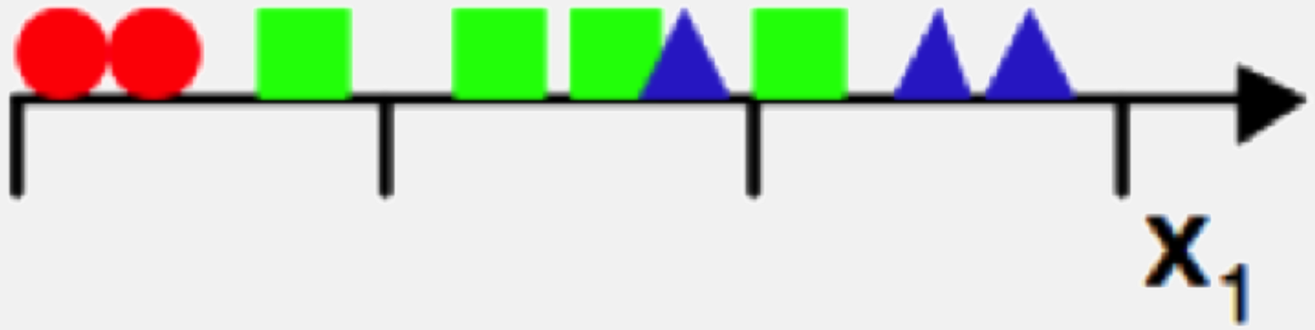
The curse of Dimensionality (1/2)

Term coined by Bellman (1961)

Describes a problem related with the relationship of multivariate data and number of samples

Consider a 3-class pattern recognition

- One feature or variable.



The curse of Dimensionality (2/2)

- The Easiest classifier: three bins / three samples per bin, then, vote

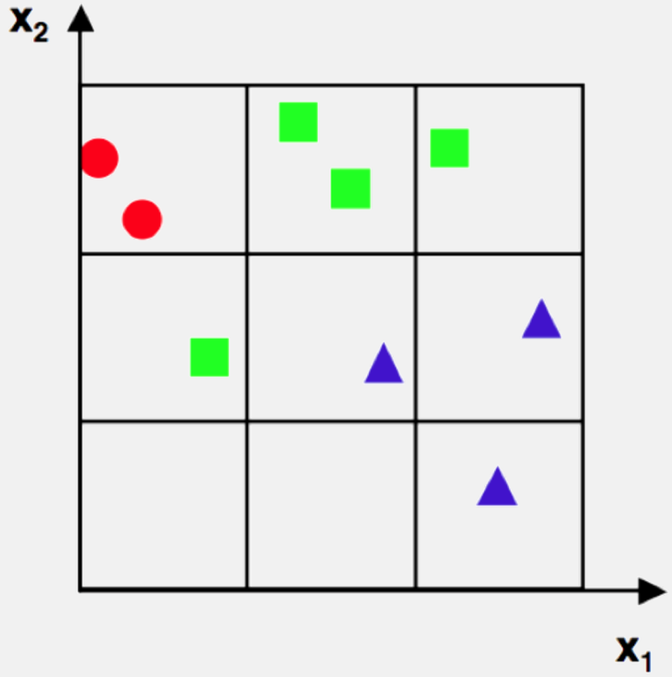


- Can we improve the classification performance adding a second feature?

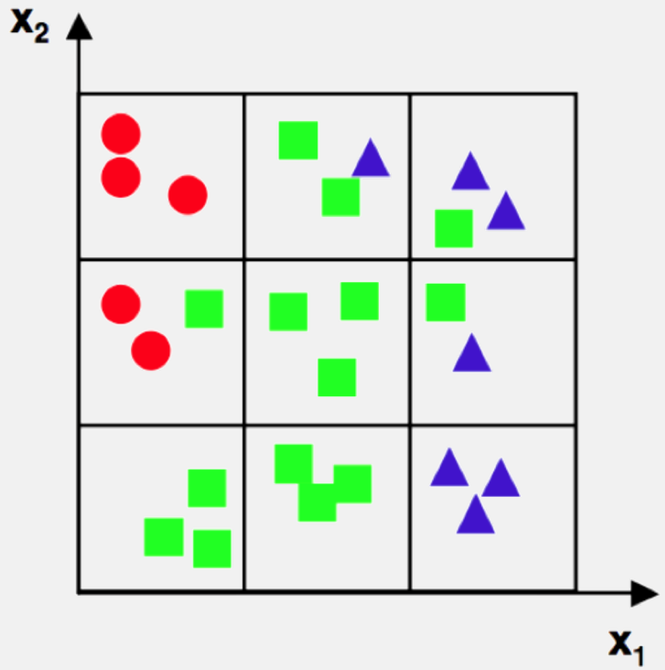
Adapted from Gutierrez-Osuna

We shall maintain the density of examples we used on the 1D case?

Constant # examples



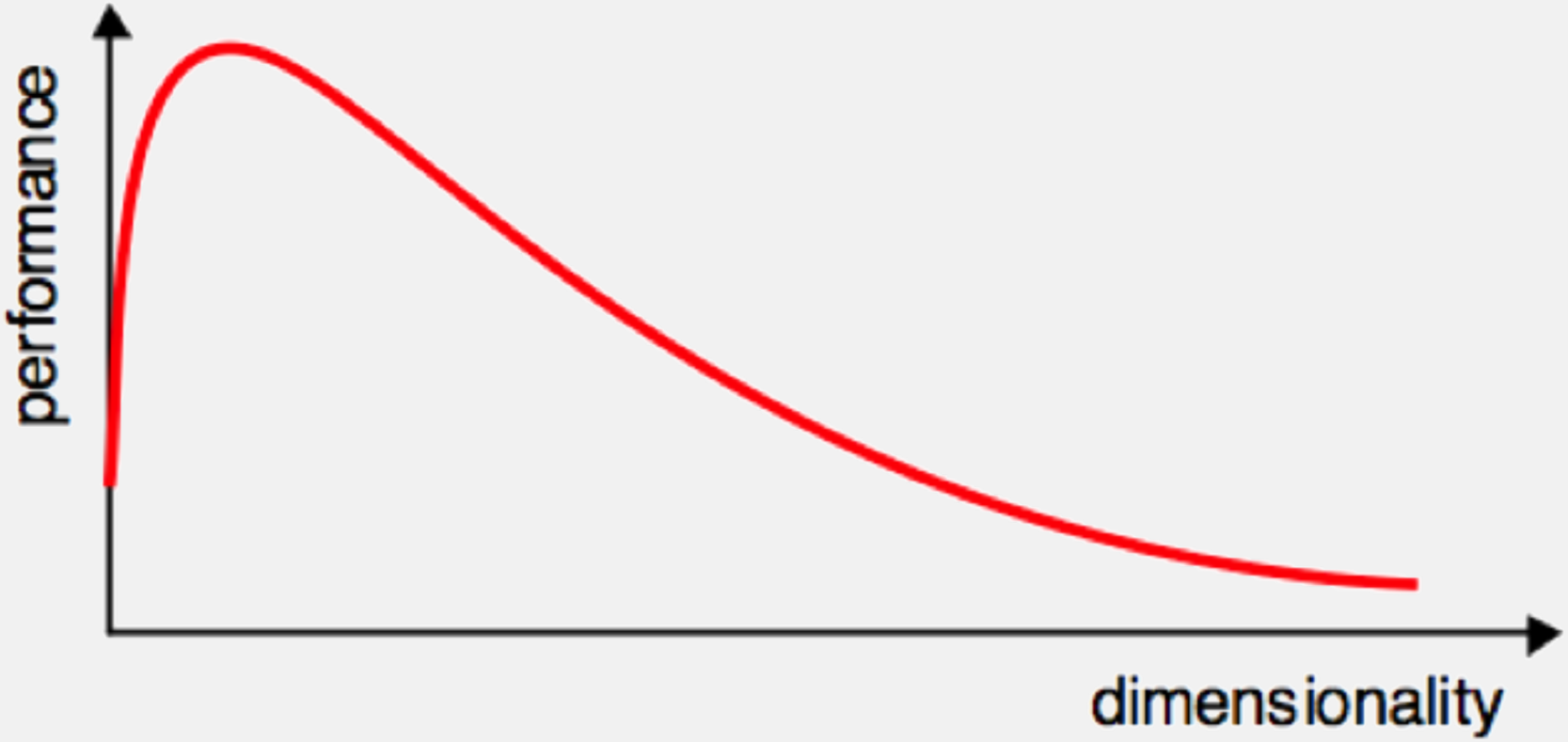
Constant density



- Bins are now 9 -> we will need 27 samples to maintain density.
- on 3d, 27bins, 81 samples…

Adapted from Gutierrez-Osuna

The Curse of Dimensionality



Houston! We…

How could we avoid the curse of dimensionality?

- Using prior knowledge.
- Smoothing the target function.
- Reducing the dimensionality.

Feature Selection vs Feature Extraction

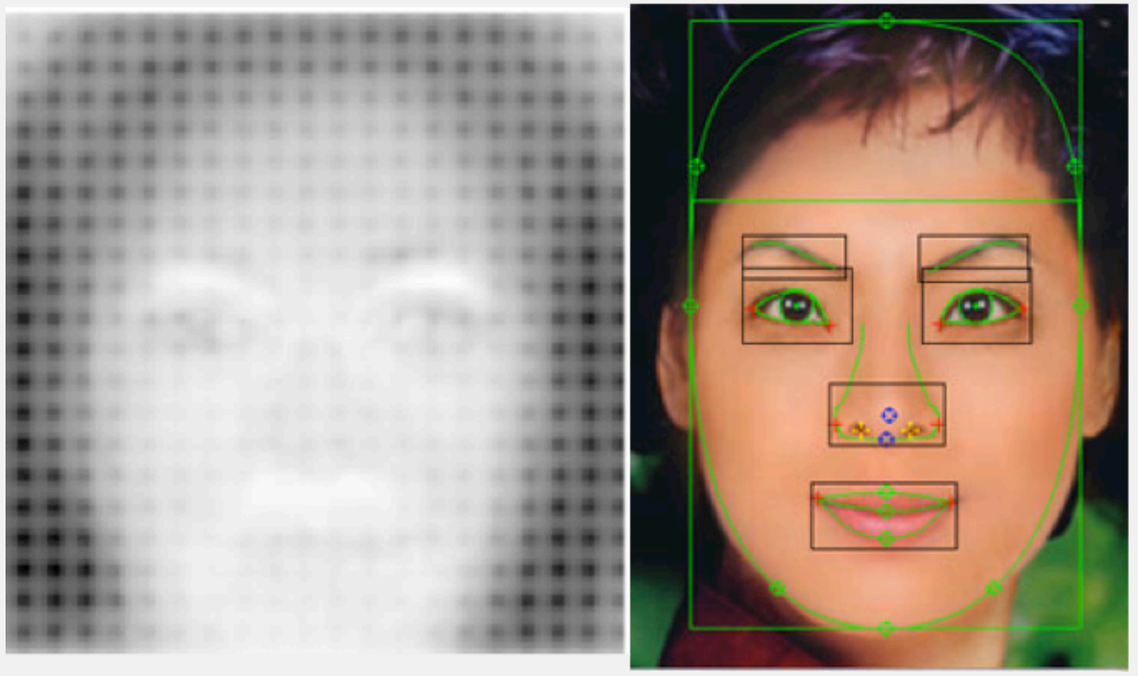
Feature Selection

From a set of input N variables, select those which contain the most informative to the problem



Feature Extraction

From a set of input N variables, could we define a set of functions or transformations that contains relevant information for our problem?



Aim

- Choose a subset of input variables by eliminating features through a given criteria (e.g. with little or no predictive information).
- Feature selection can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points.

Types

Supervised

Using a class/target label. Main goal is to find a feature subset that produces higher classification/regression accuracy (e.g. linear filtering with gene expression data).

Unsupervised

Without using class/target label (e.g. Variability filtering analysis in gene expression data).

Feature ranking

For samples (x_k, y_k), assign each feature i a score using its observed values and the target:

$$S_i = f(\{x_{k,i}, y_k\}_{k=1}^{m}).$$

Rank features by a score appropriate to the question. Estimate supervised rankings inside each training fold to avoid leaking test information.

Pearson Correlation Coefficient

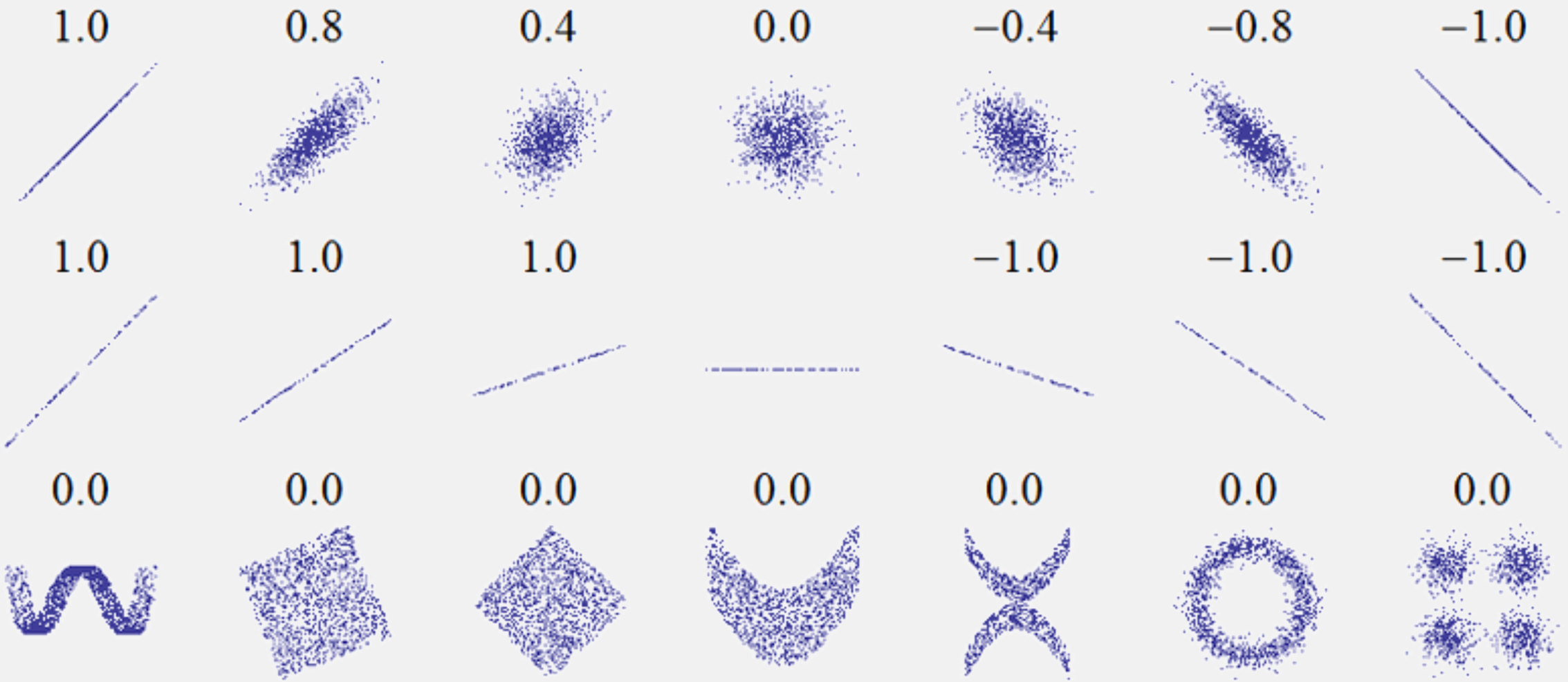
$$R(i) = \frac{cov(Xi, Y)}{\sqrt{var(X_i)var(Y)}}$$

Estimate of $R(i)$

$$\Re(i) = \frac{\sum_{k=1}^{m}(x_{k,i} - \bar{x}_i)(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{m}(x_{k,i} - \bar{x}_i)^2 \sum_{k=1}^{m}(y_k - \bar{y})^2}}$$

.. eq. to a goodness of linear fit of individual variables.

Pearson



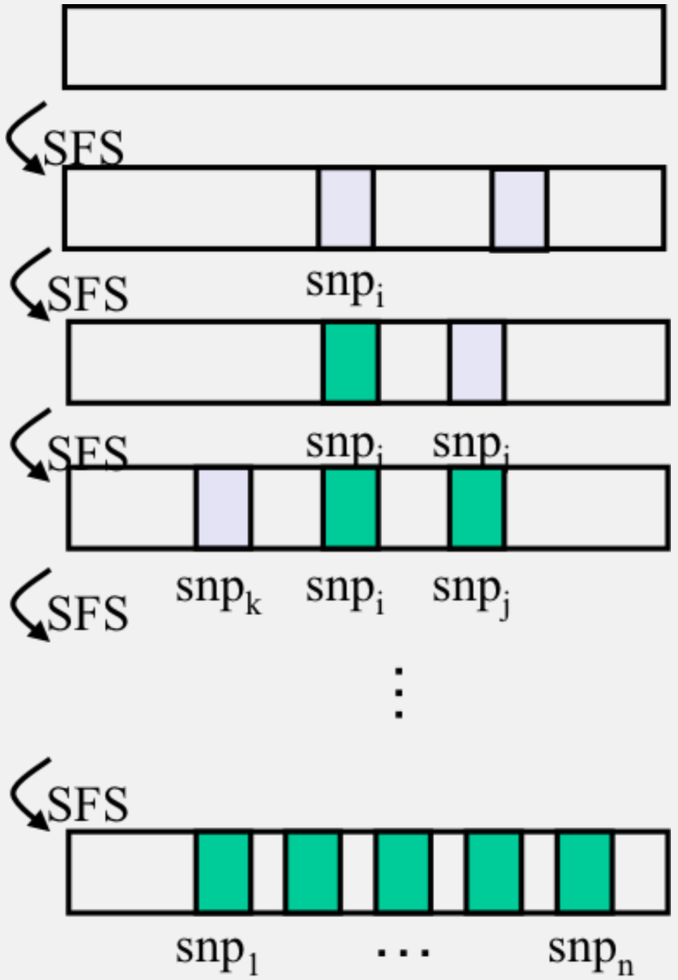
Mutual Information

$$I(i) = MI(x_i, y) = \int_{x_i} \int_{y} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)p(y)} \mathrm{d}x\mathrm{d}y$$

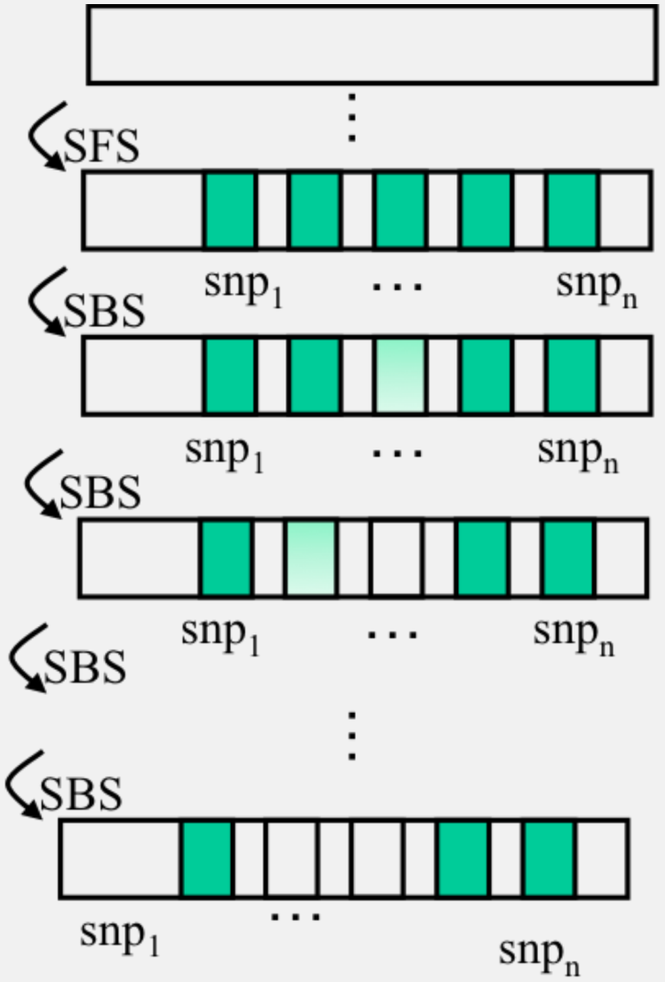
, where $p(x_i)$ and p(y) are the probability densities of x_i and y, and $p(x_i, y))$ is the joint density.

$$\mathfrak{I}(i) = \sum_{x_i} \sum_{y} P(X = x_i, Y = y) \log \frac{P(X = x_i, Y = y)}{P(X = x_i)P(Y = y)}$$

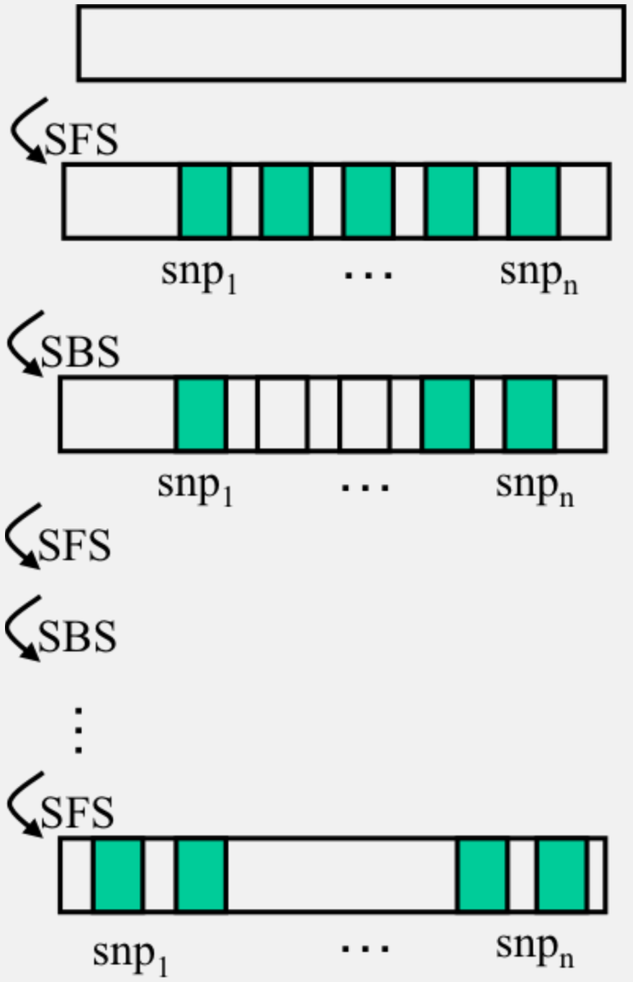
Sequential Selection Algorithms (1/3)



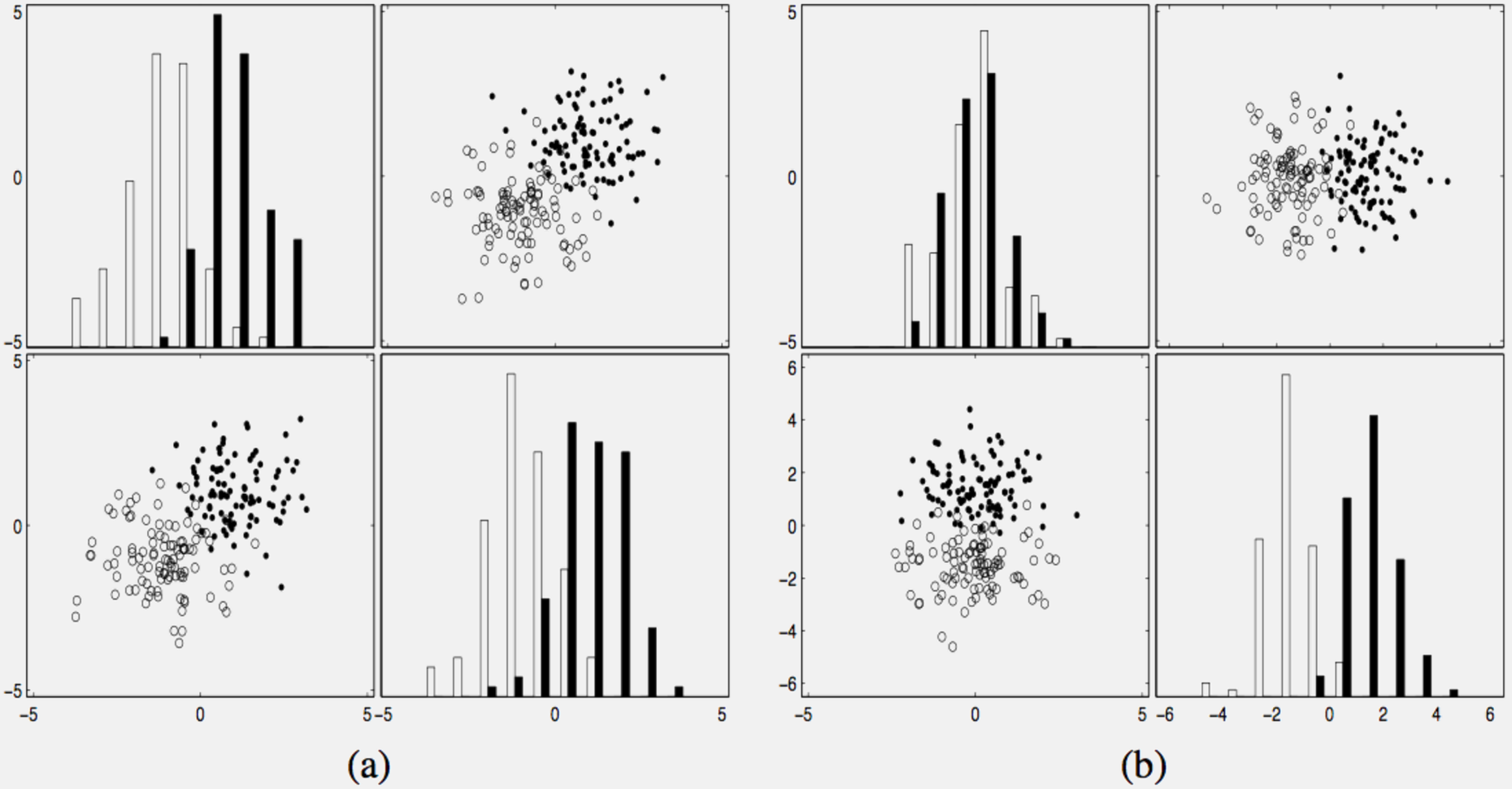
Sequential Selection Algorithms (2/3)



Sequential Selection Algorithms (3/3)

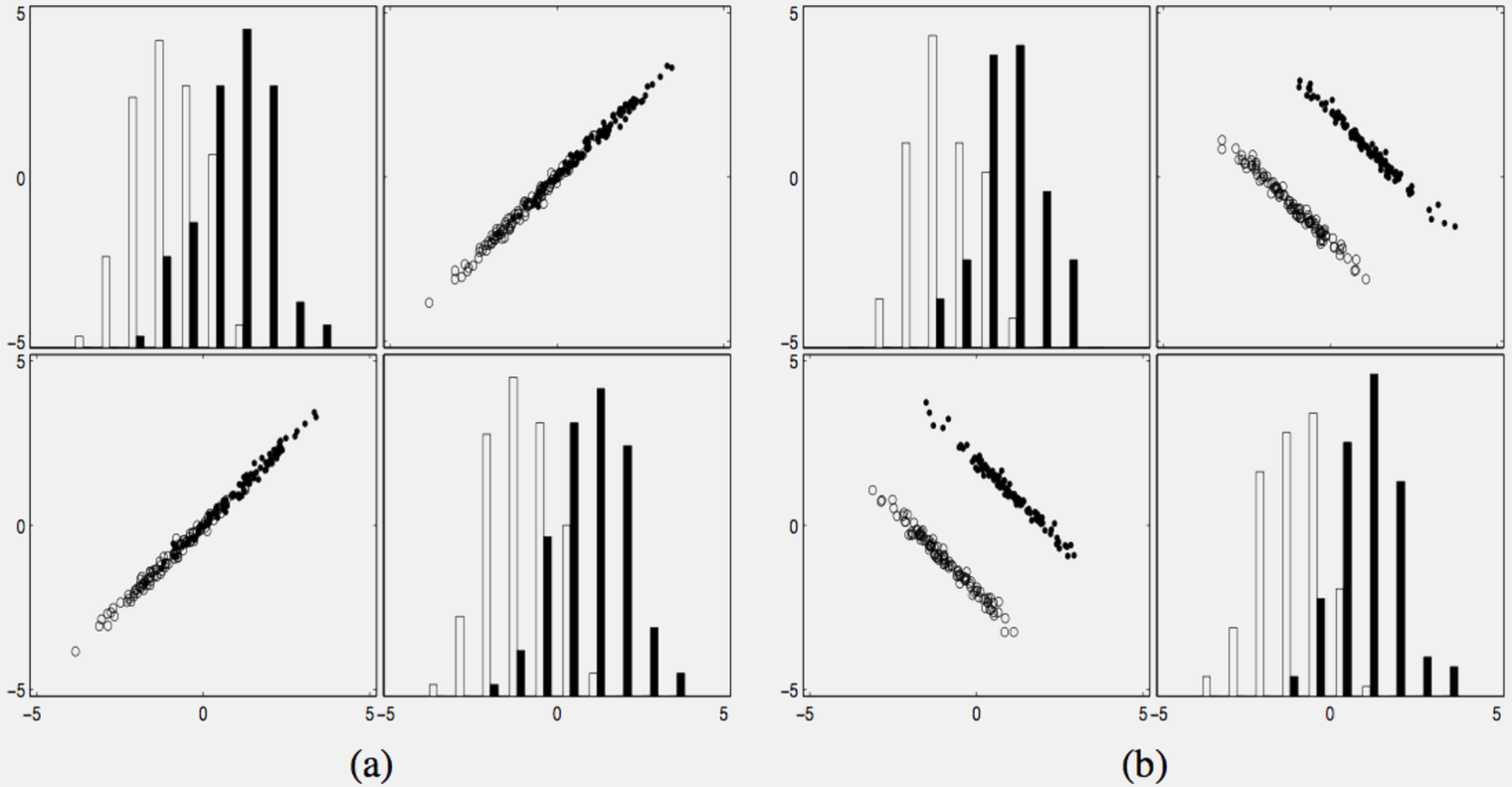


Can redundant variables help each other?



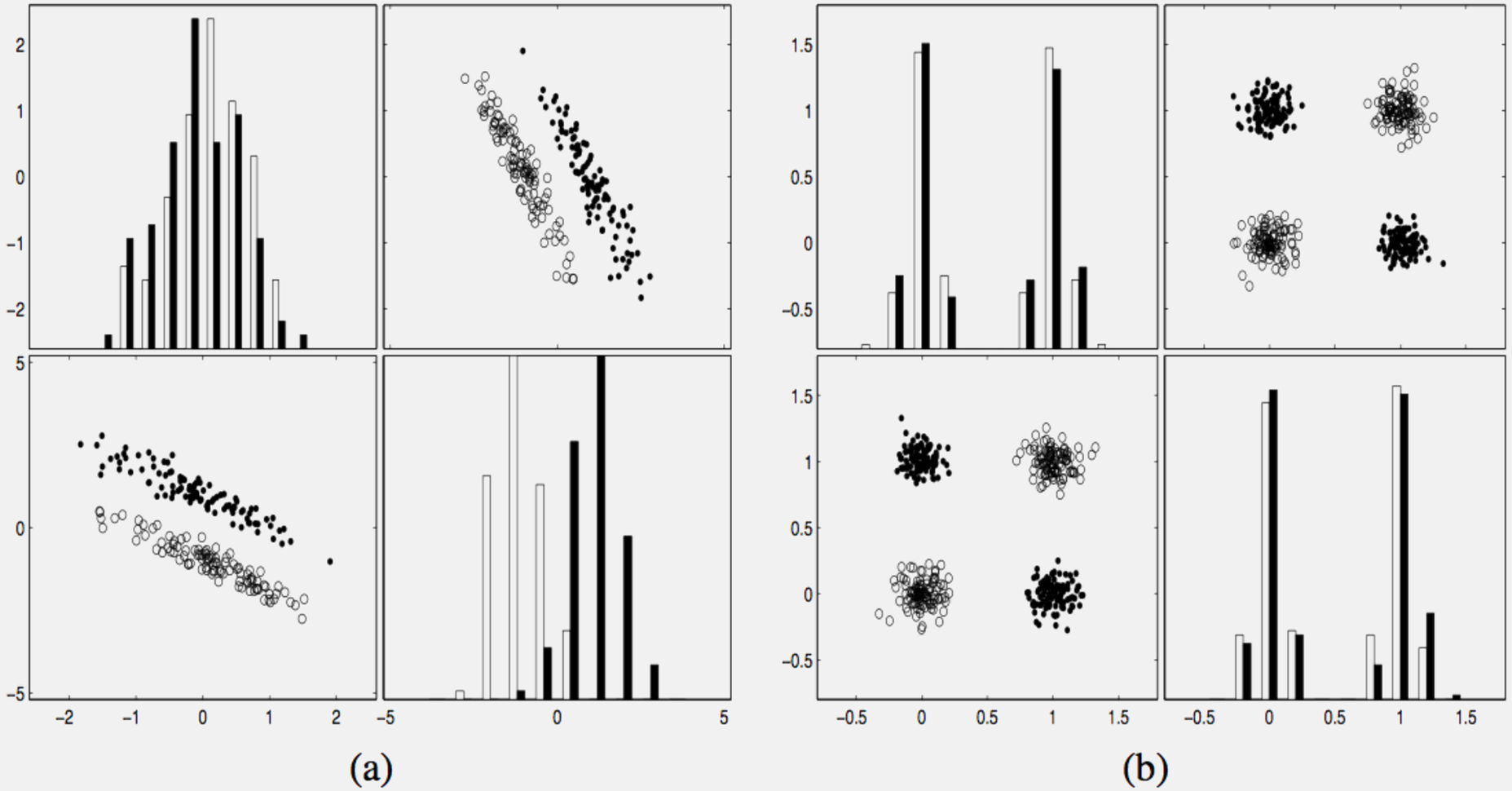
(a): The presumably redundant variables, (b) set after a 45 degree rotation from Guyon 2003

Intra-Class covariance



(a) The class conditional distributions have a high covariance aligned with two class centers. (b) The class conditional distributions have a high covariance in the direction perpendicular to the line of the two class centers.
from Guyon 2003

Is covariance useful?



from Guyon 2003

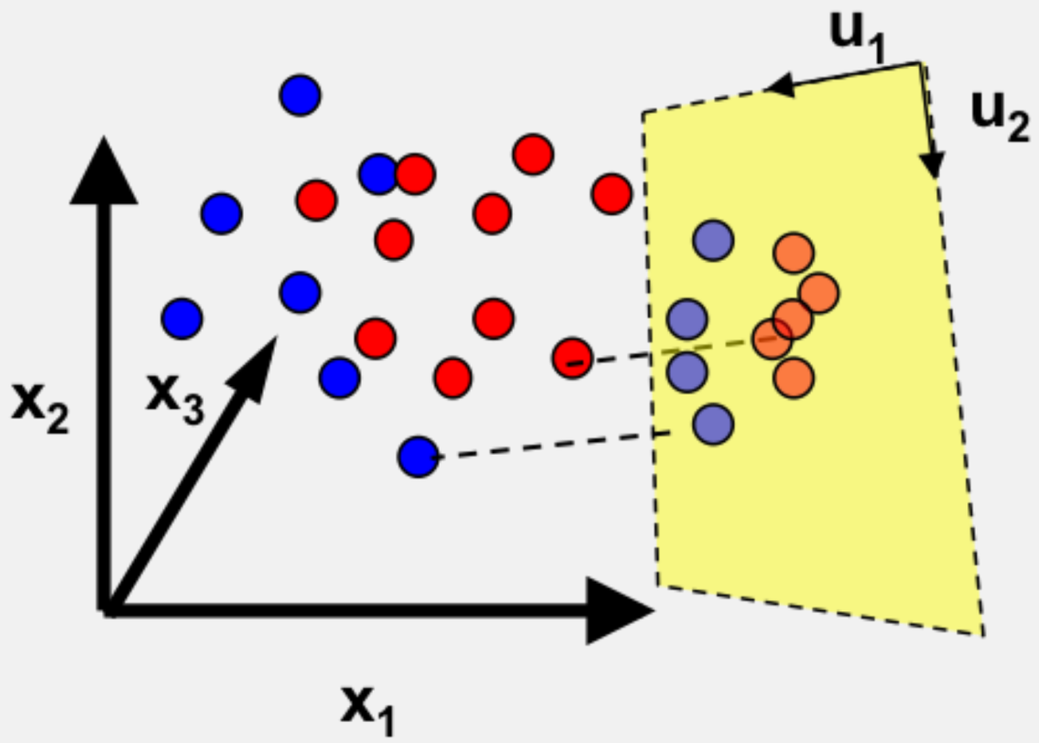
Feature extraction: subspace methods

Project observations into a lower-dimensional representation that retains information relevant to the task.

Examples include:

- Principal component analysis (PCA)
- Linear discriminant analysis (LDA)
- Independent component analysis (ICA)

Each method uses a different criterion; dimension reduction does not automatically preserve predictive information.



Principal Component Analysis

Most popular and commonly used methods. Sometimes included in the “Factor Analysis” methods.

- It’s the second most main tool for visualization
  - (First is always to plot the signals!)
- Main goal of PCA is to capture main directions of variance in input space.
- Dimensionality reduction:
  - Allows for projecting the dataset onto a low dimensional subspace.
  - Form of compression, or data modeling.
  - Filtering.

The mean as a least-squares representation

For observations $x_1, \ldots, x_n$, choose a single representative vector a:

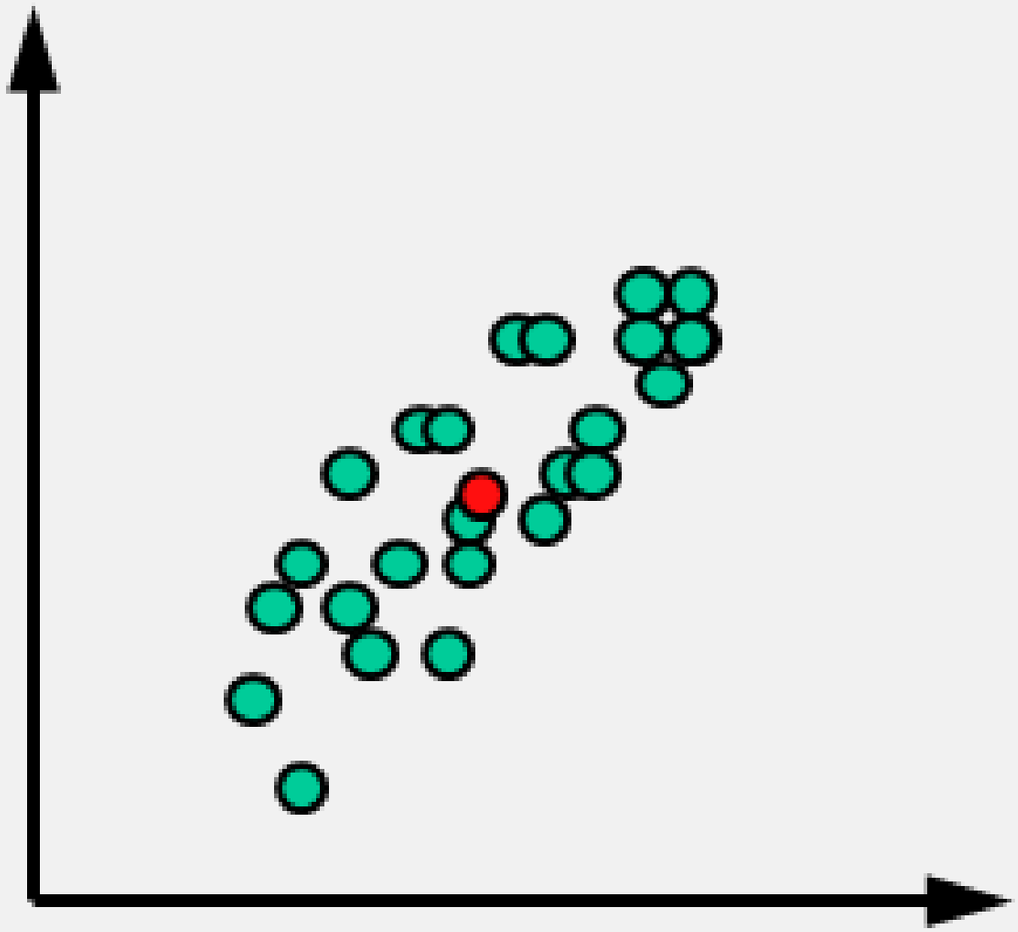
$$J(a) = \sum_{k=1}^{n} \|x_k - a\|^2.$$

With $\bar{x} = n^{-1} \sum_k x_k$,

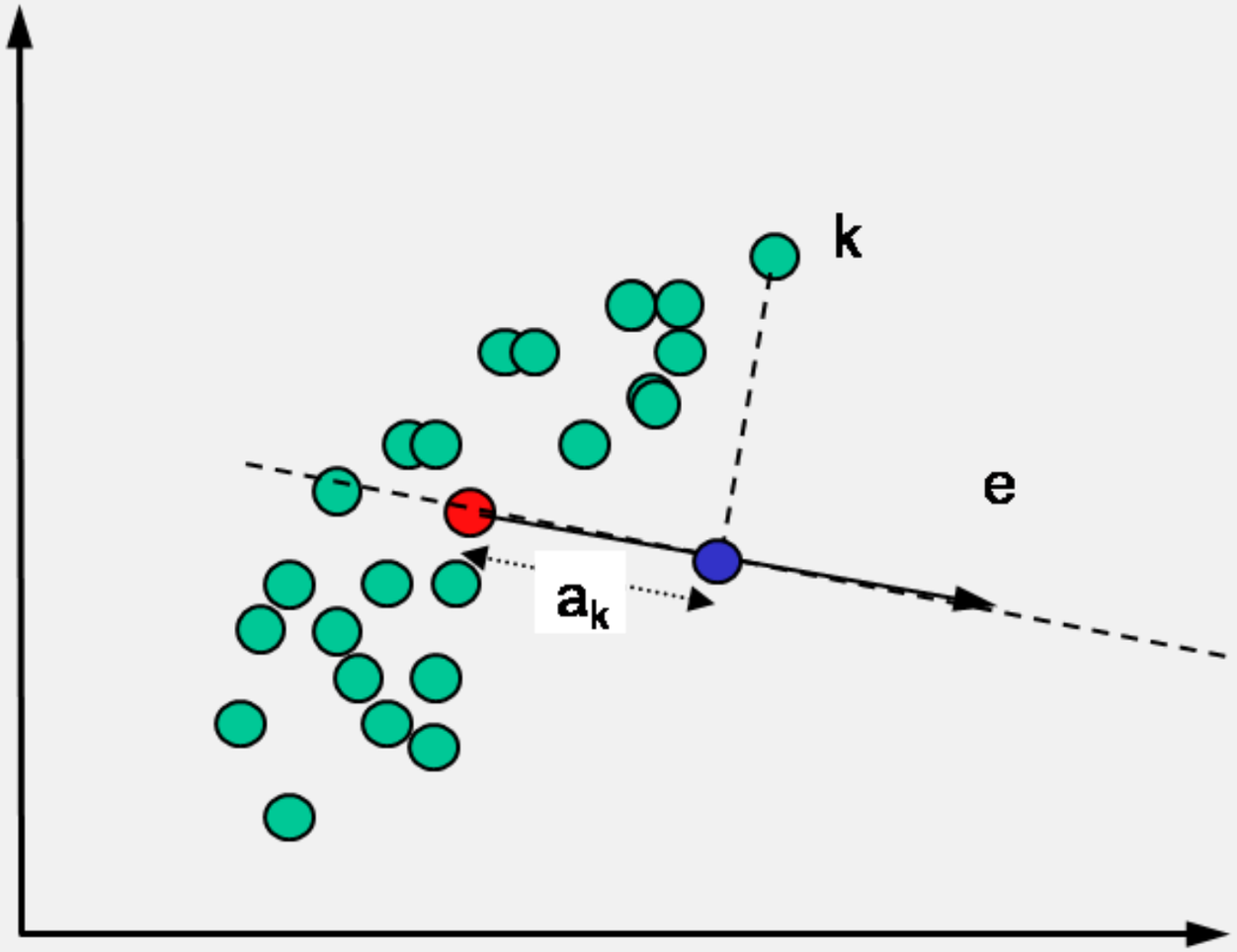
$$J(a) = \sum_{k} \|x_k - \bar{x}\|^2 + n\|a - \bar{x}\|^2.$$

The minimum occurs at $a = \bar{x}$.

Principal Component Analysis



Principal Component Analysis



Projection onto one direction

For a unit vector e, approximate each observation by

$$\hat{x}_k = \bar{x} + a_k e.$$

The least-squares coefficient is

$$a_k = e^\top (x_k - \bar{x}).$$

This minimises $\sum_k \|x_k - \bar{x} - a_k e\|^2$ for a fixed direction. PCA then chooses the direction with the smallest reconstruction error.

Covariance matrix

For a random vector $X \in \mathbb{R}^p$ with mean μ:

$$\Sigma = \mathrm{E}[(X - \mu)(X - \mu)^\top].$$

Its (i, j) entry is $\mathrm{Cov}(X_i, X_j)$. Diagonal entries are variances.

The usual unbiased sample estimate is

$$\widehat{\Sigma} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(x_k - \bar{x})^\top.$$

Scaling features changes the covariance matrix and can change the principal components.

The first principal component

Let $S = \sum_k (x_k - \bar{x})(x_k - \bar{x})^\top$ be the scatter matrix.

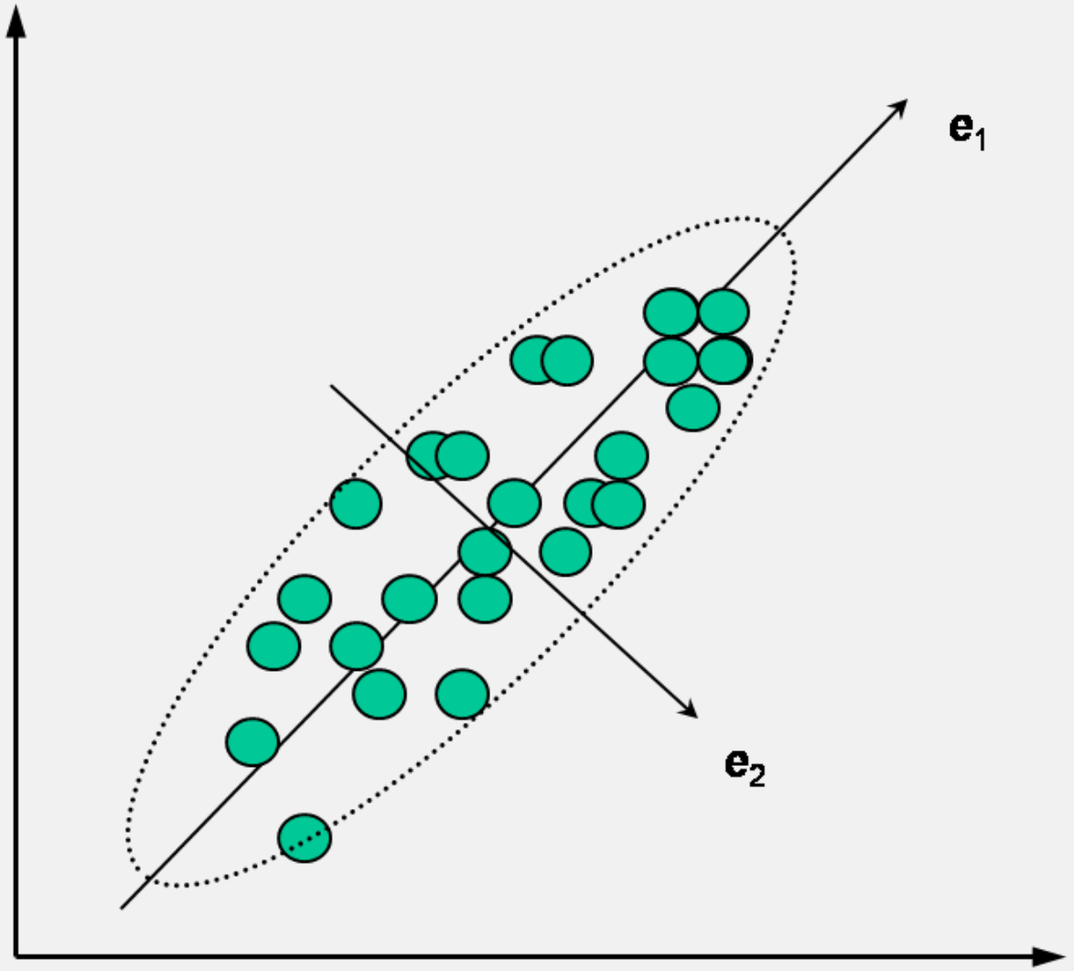
For a unit vector e, reconstruction error is

$$J(e) = \sum_k \|x_k - \bar{x}\|^2 - e^\top S e.$$

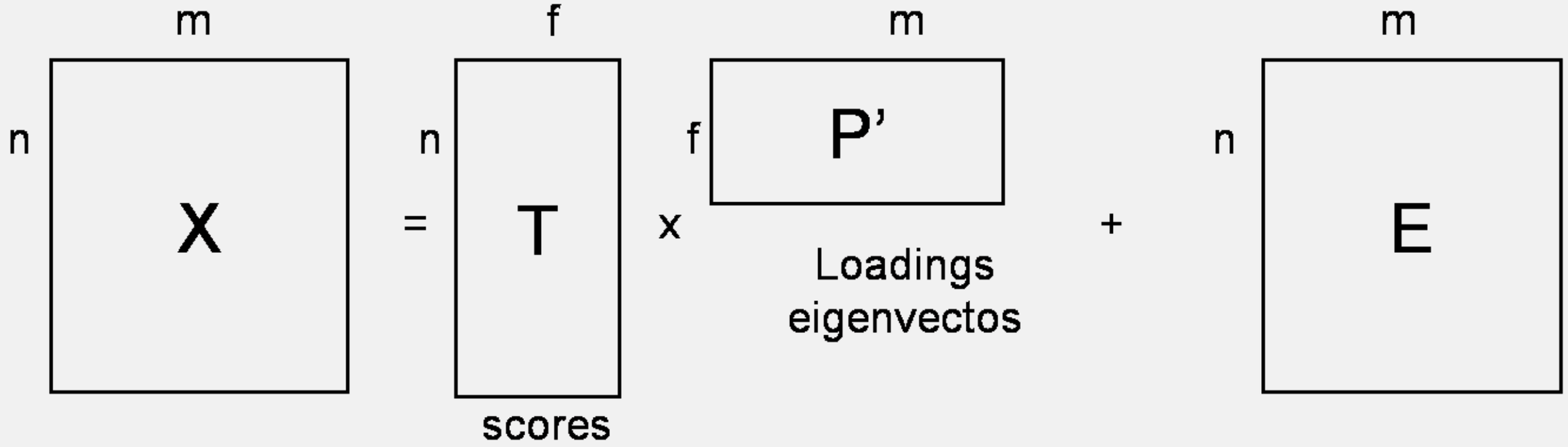
Minimising J means maximising $e^\top S e$ subject to $e^\top e = 1$. The solution is an eigenvector associated with the largest eigenvalue:

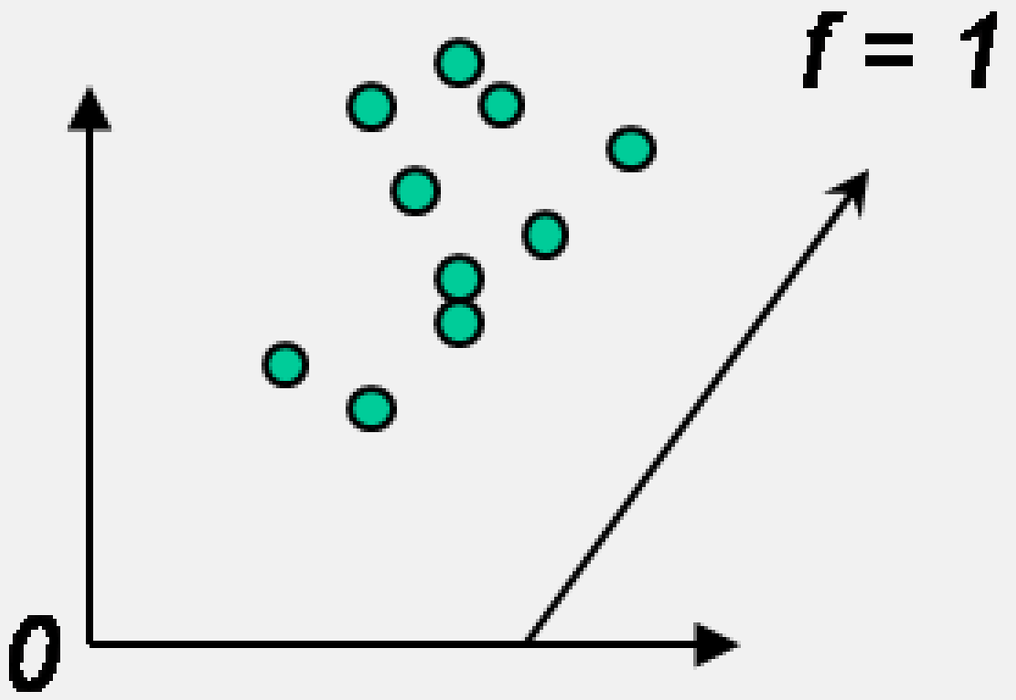
$$Se = \lambda e.$$

PCA



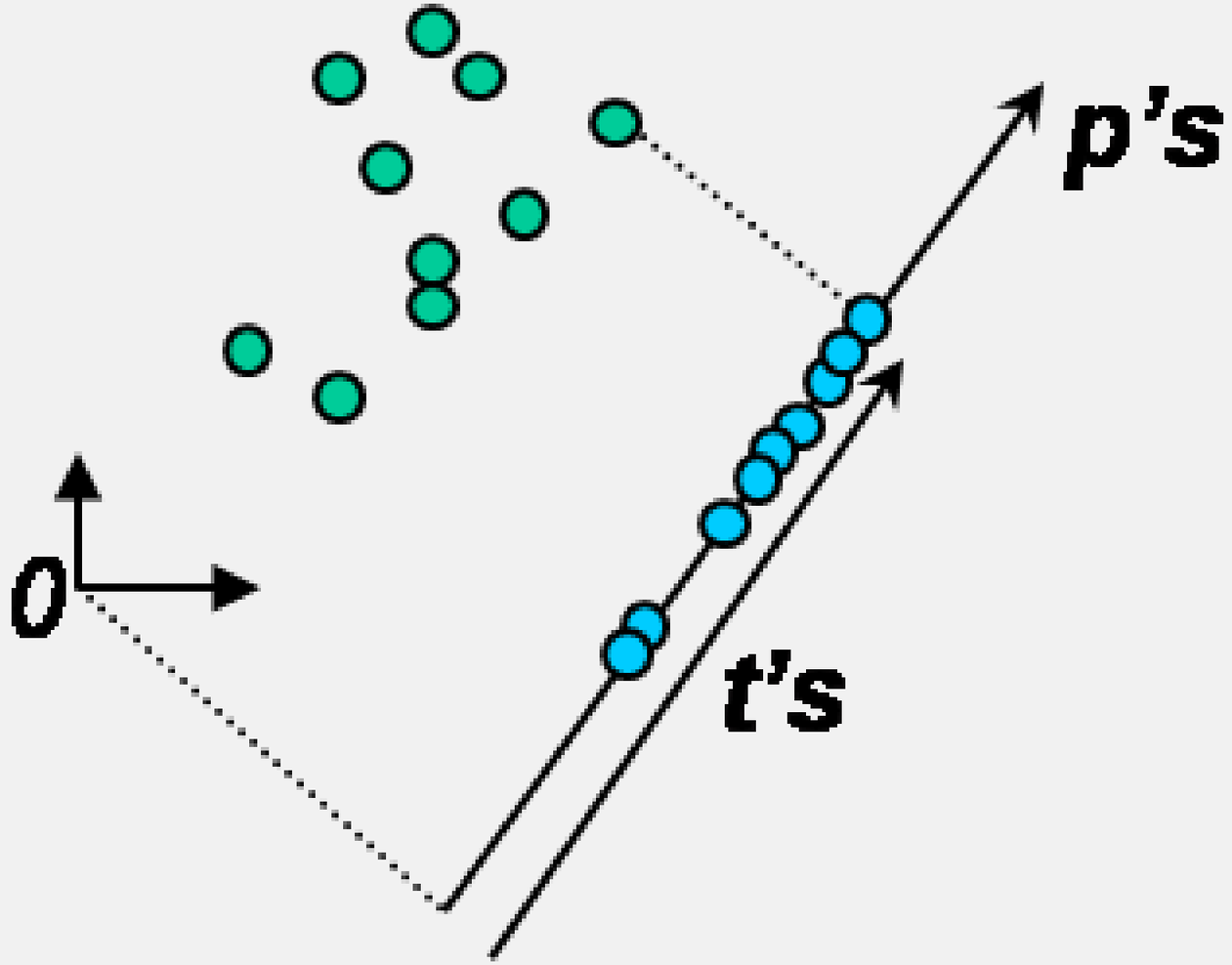
PCA (1/4)



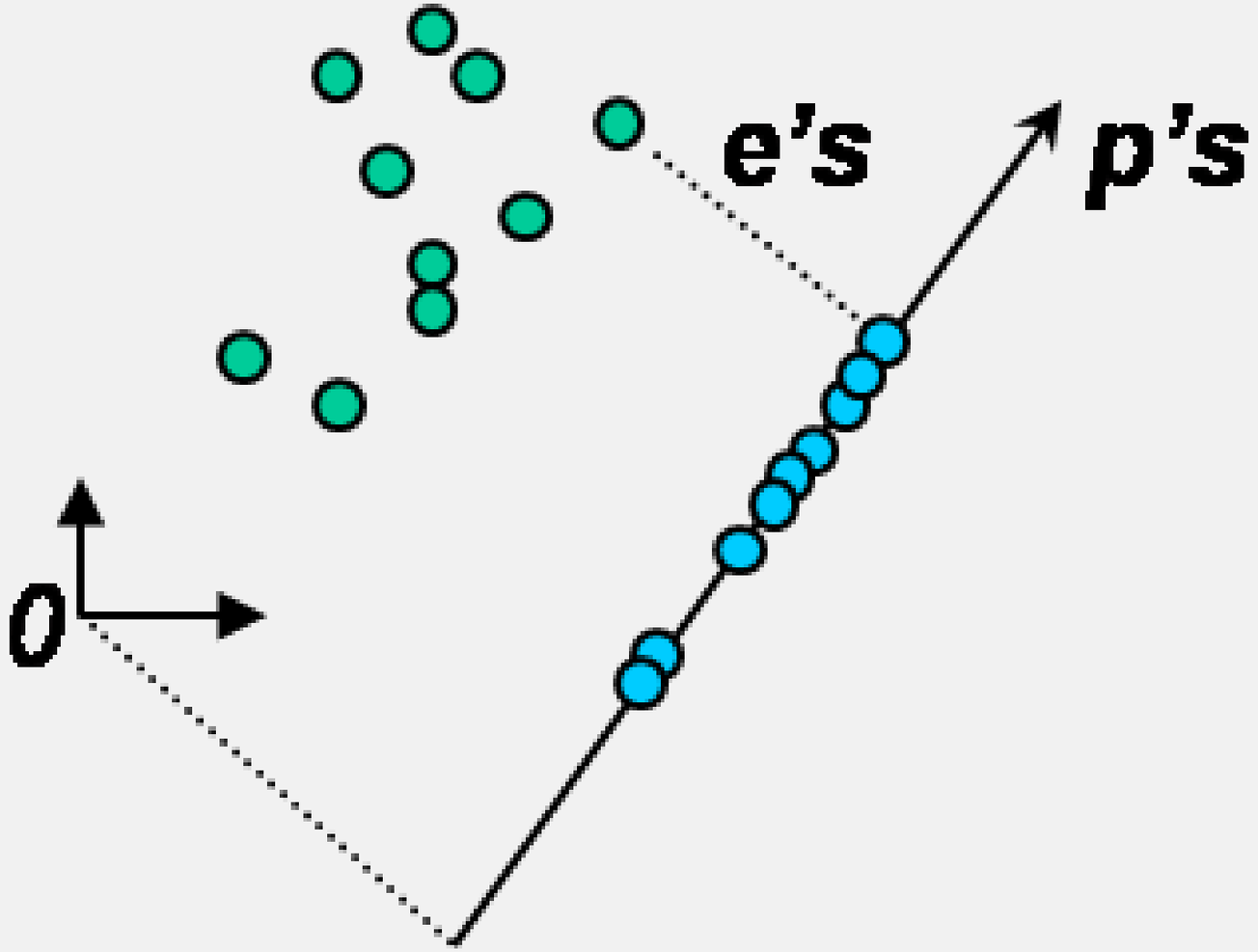


$m = 2$

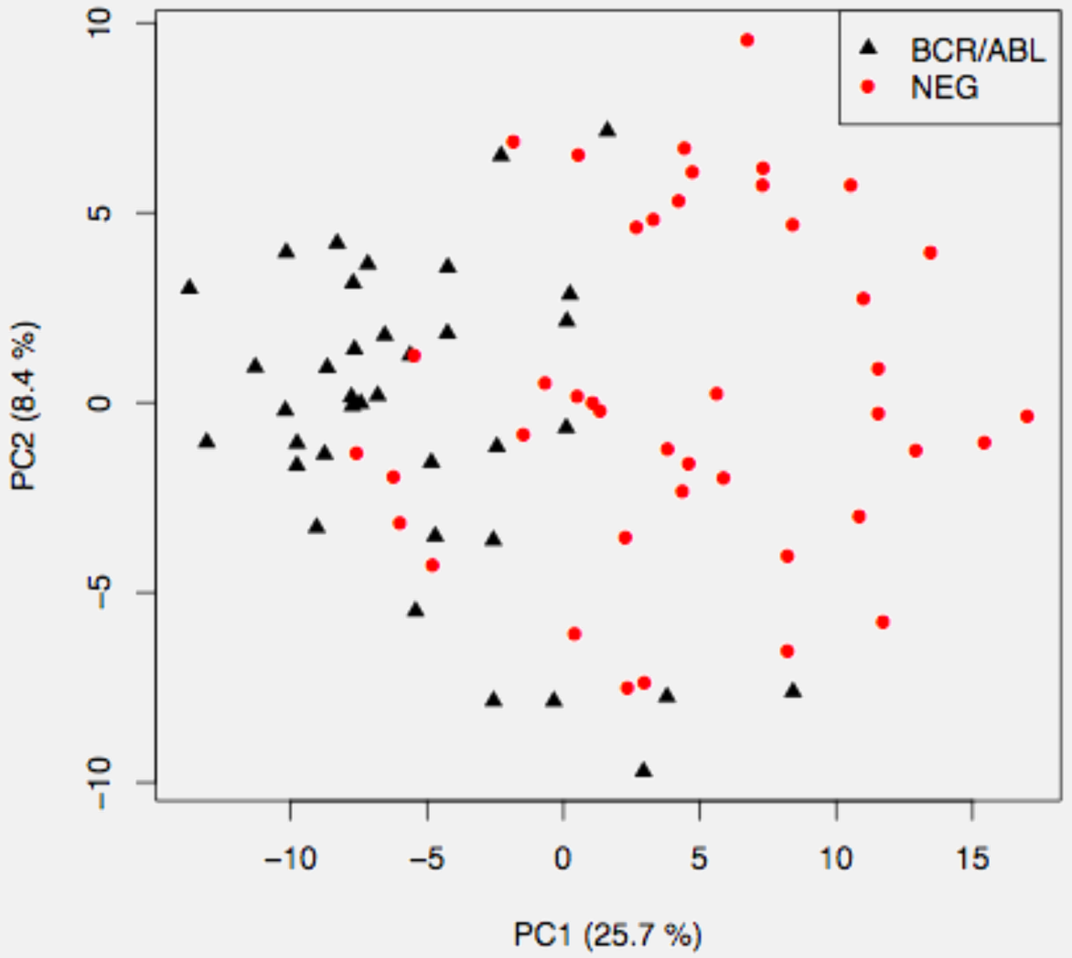
PCA (3/4)



PCA (4/4)



PCA



Linear Discriminant Analysis (1/3)

LDA looks for the projection

$$y = w^\top x$$

So that maximizes the Fisher Criterion

$$J(W) = \frac{w^\top S_B w}{w^\top S_W w}$$

where,

S_B: Between Scatter Matrix

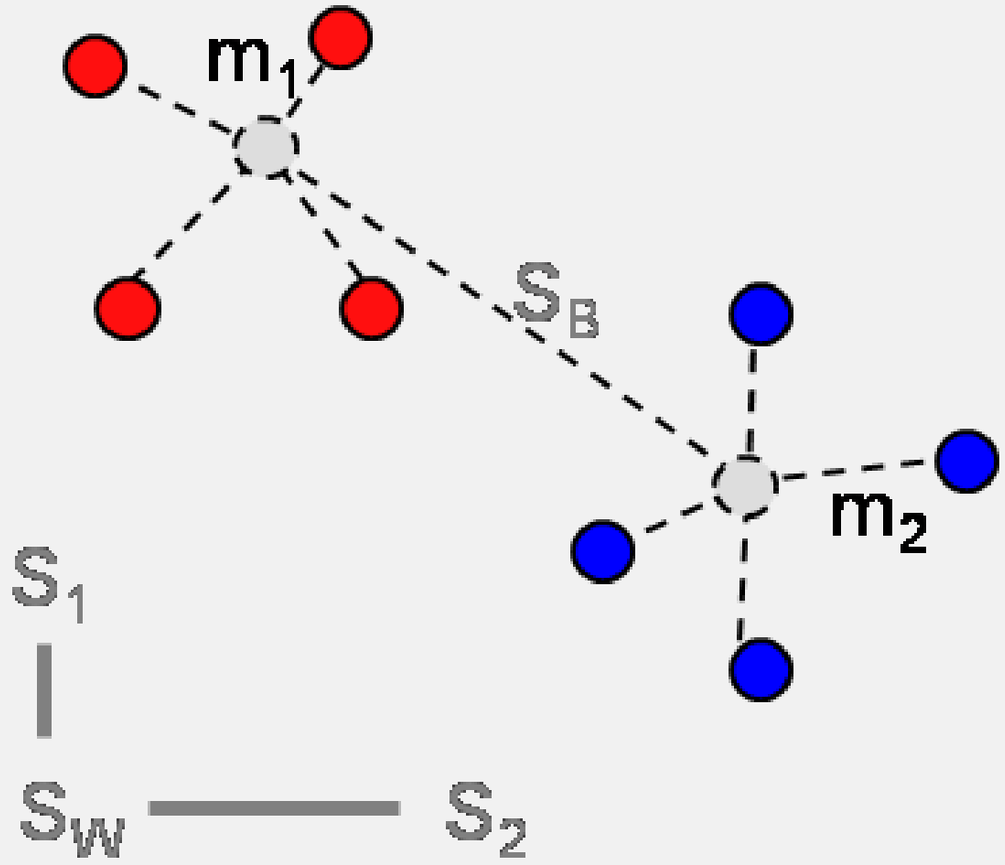
$$S_B = (m_1 - m_2)(m_1 - m_2)^\top$$

S_w: Within Scatter Matrix

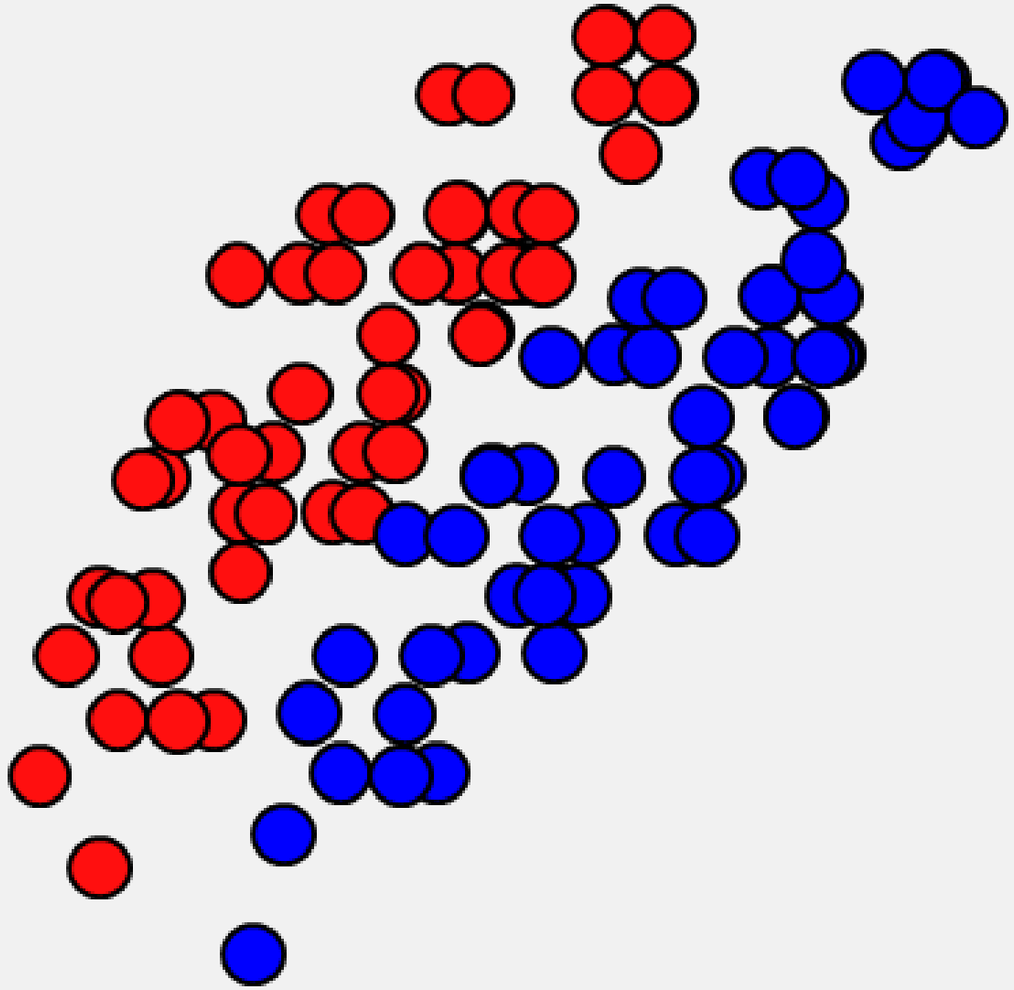
$$S_W = S_1 + S_2, \text{where } S_i = \sum_{x \in C_i} (x - m_i)(x - m_i)^\top$$

LDA minimizes the within scatter matrix while maximizes the between scatter matrix.

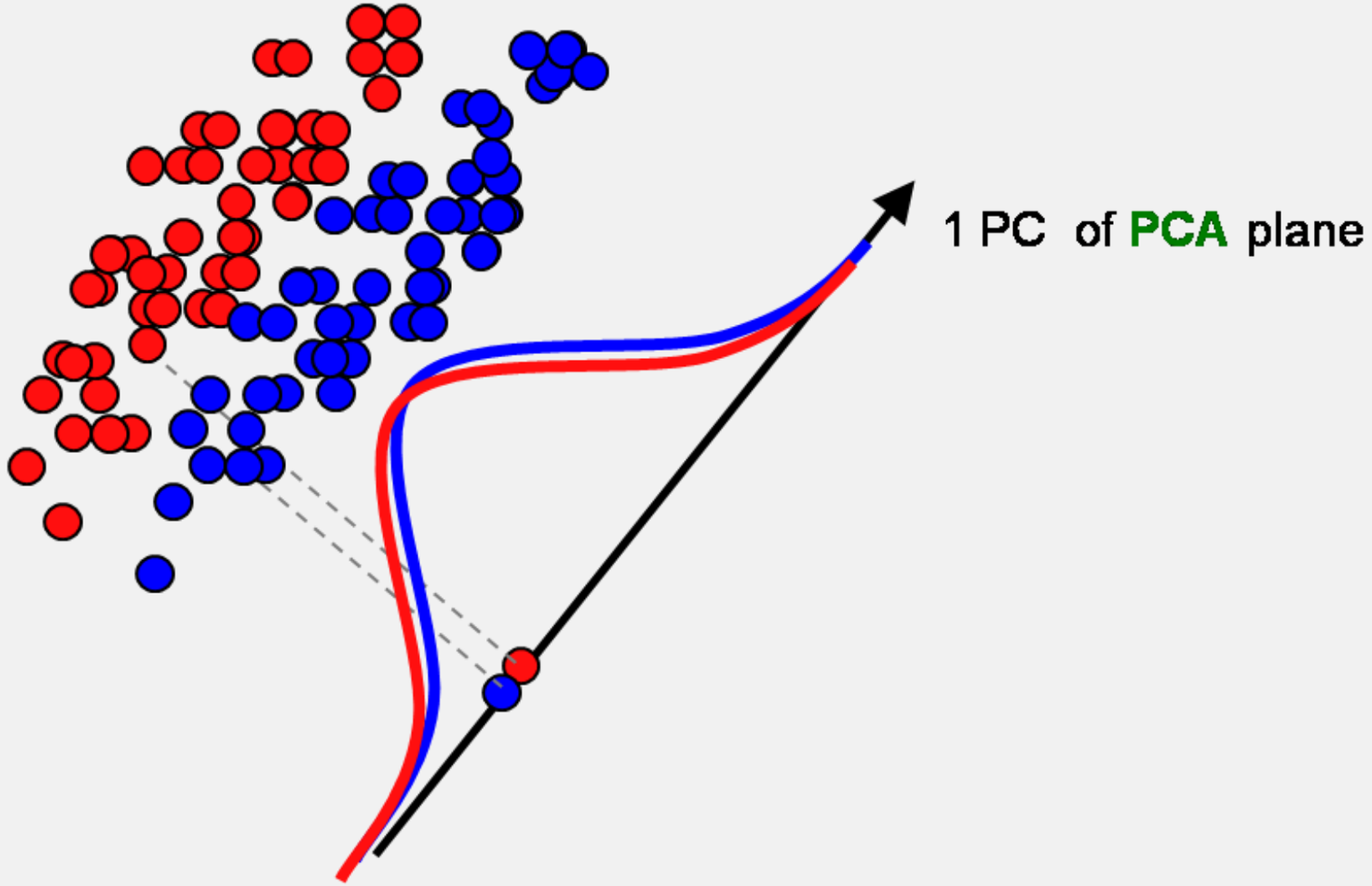
It can be extended to Multiple classes (Multiple Discriminant Analysis)



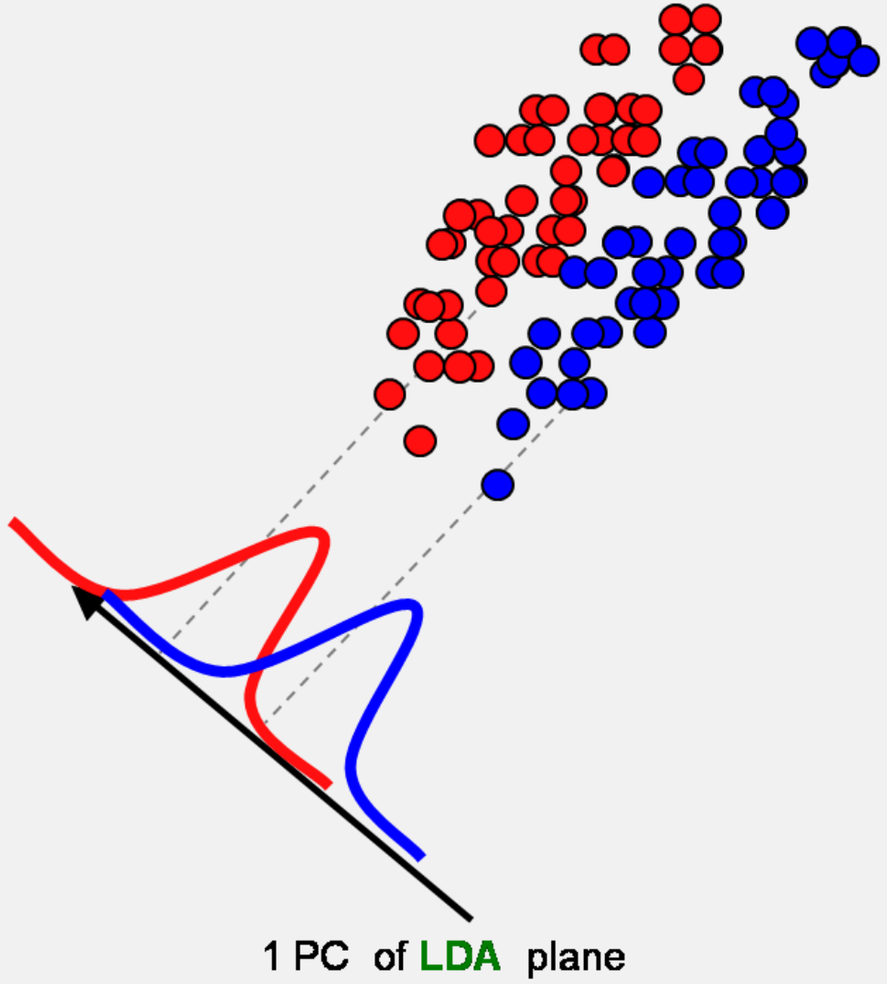
LDA vs PCA (1/4)



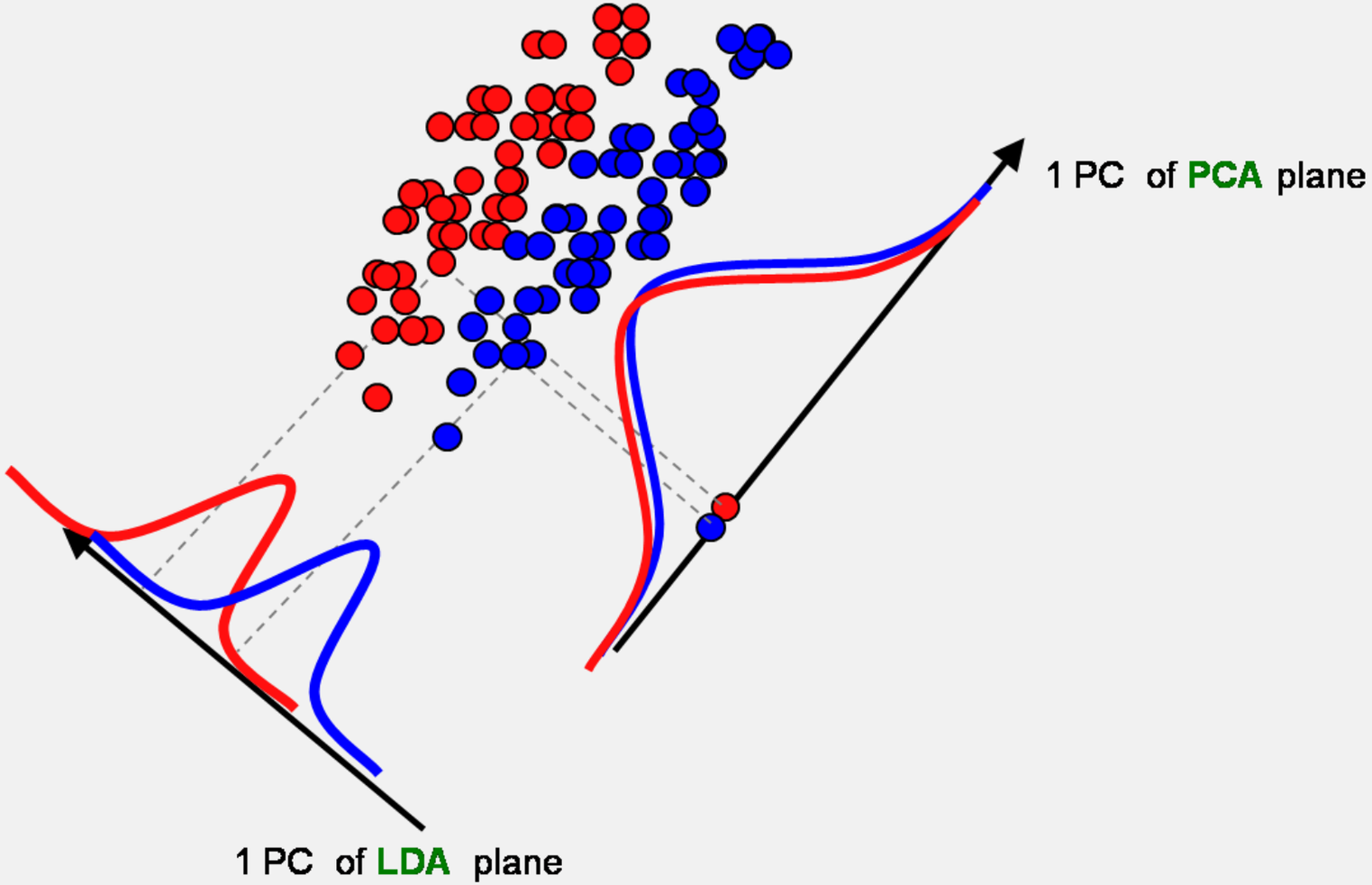
LDA vs PCA (2/4)



LDA vs PCA (3/4)



LDA vs PCA (4/4)

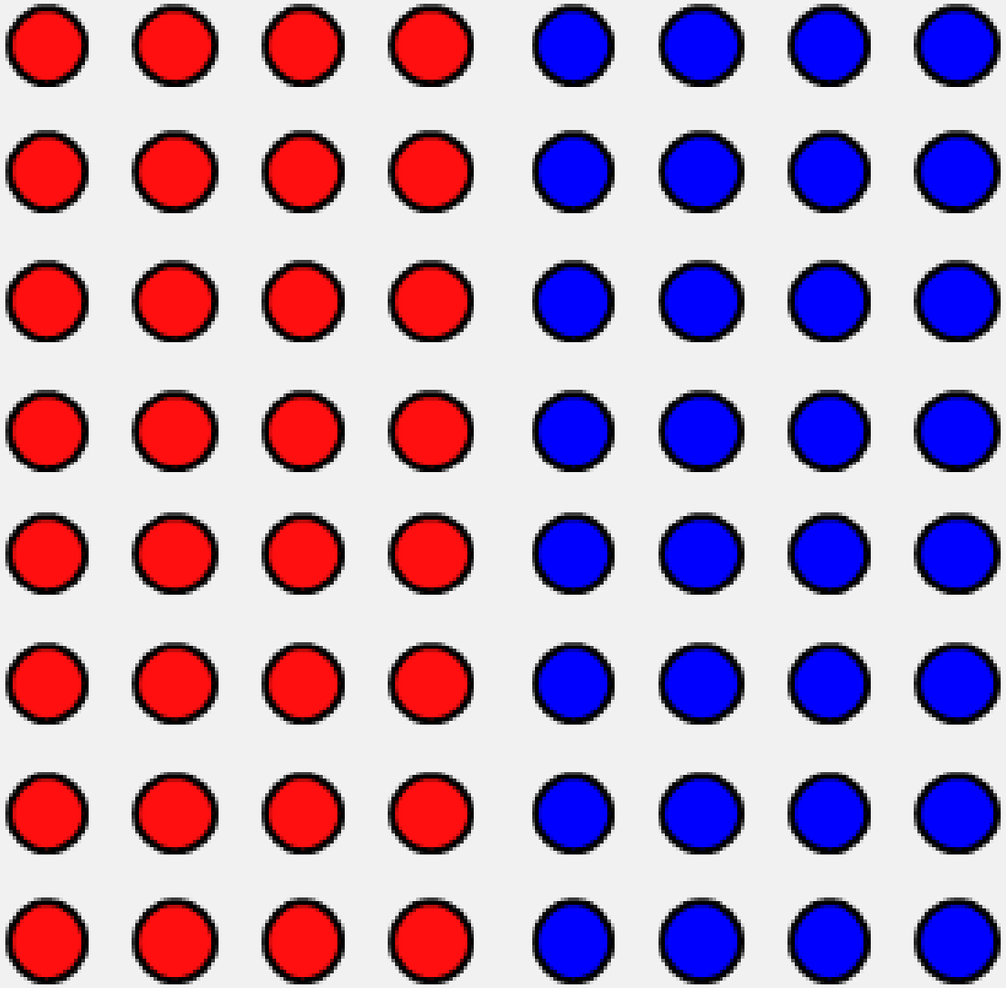


ICA models observations as mixtures of latent sources:

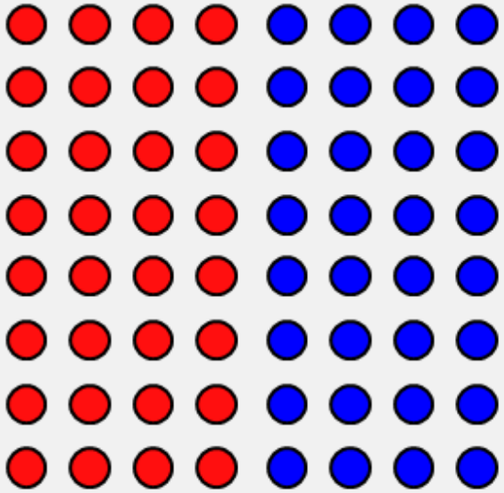
$$X = AS.$$

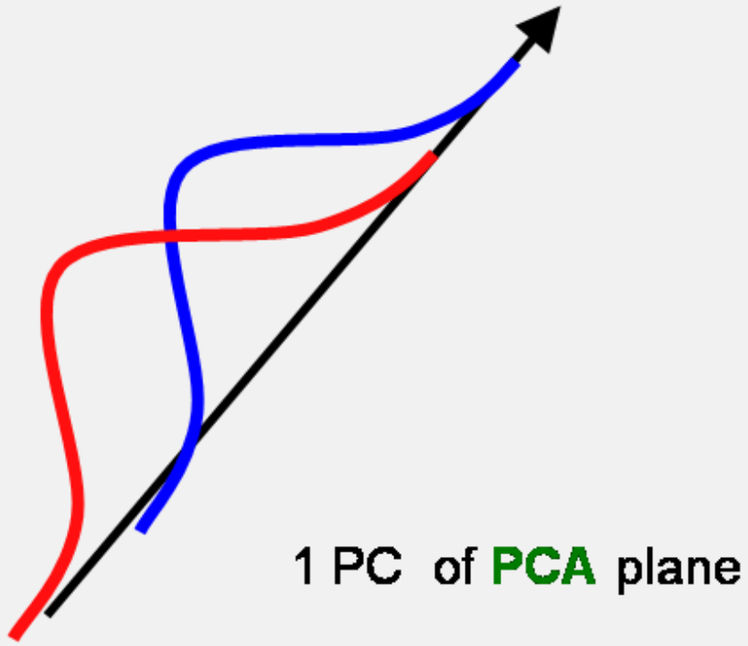
- The aim is to recover statistically independent source components.
- Non-Gaussianity is central to common ICA identifiability assumptions.
- In a square invertible model, the unmixing matrix is $W = A^{-1}$; more general models need other treatment.
- PCA removes correlations. Uncorrelated components need not be independent.

LDA vs PCA vs ICA (1/4)

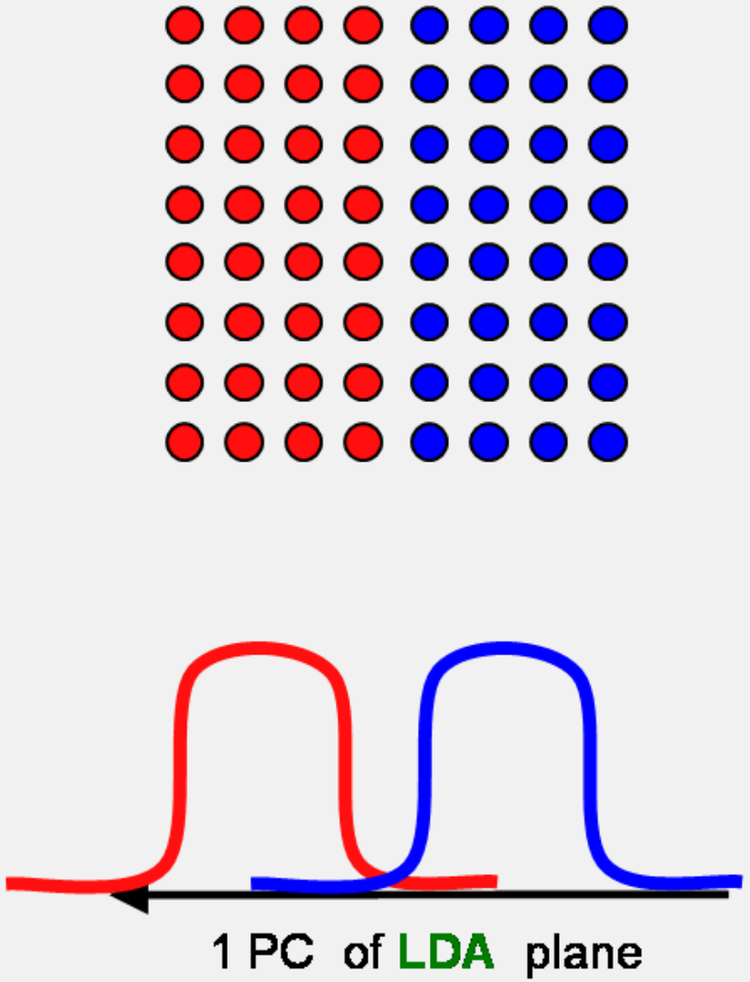


LDA vs PCA vs ICA (2/4)

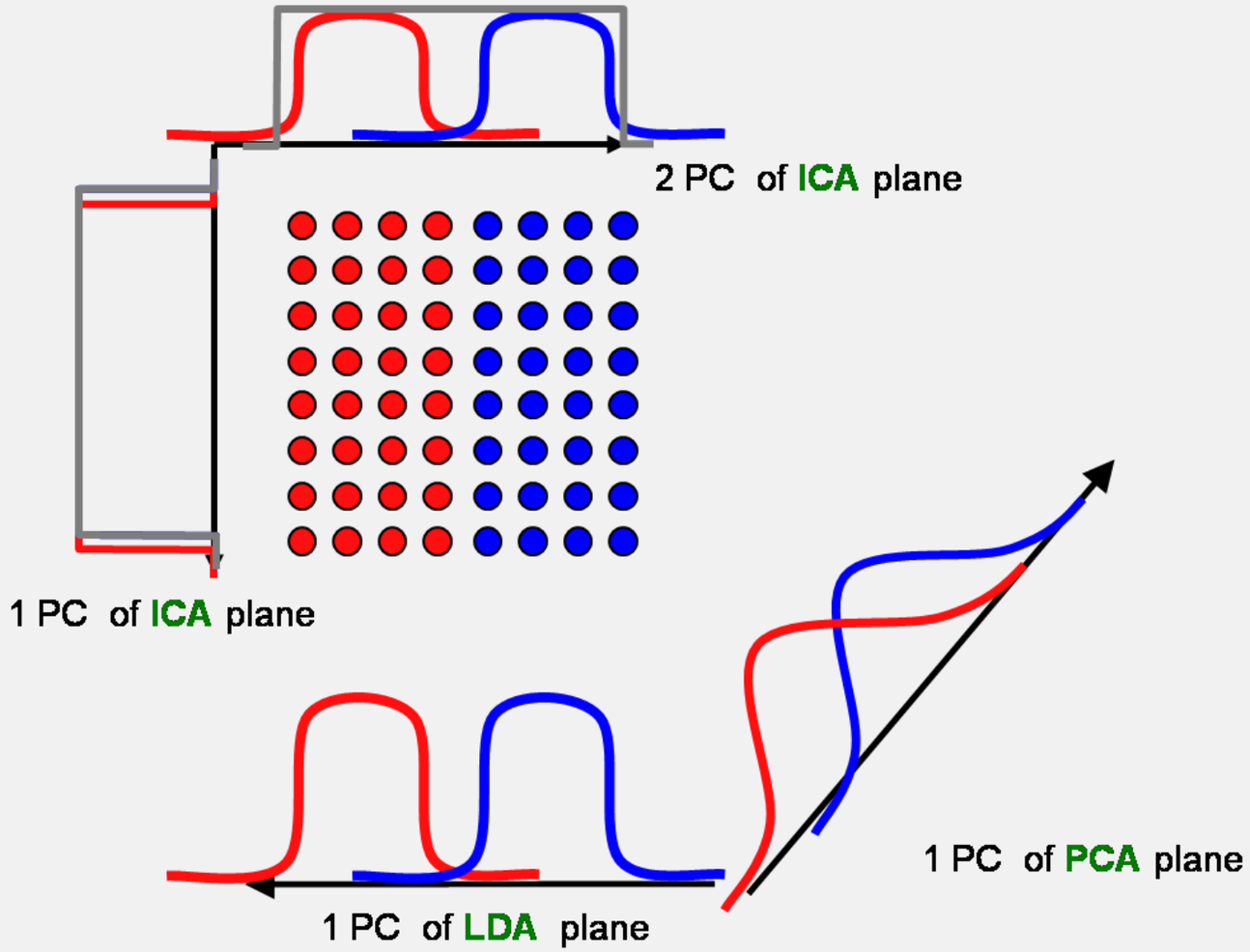




LDA vs PCA vs ICA (3/4)



LDA vs PCA vs ICA (4/4)

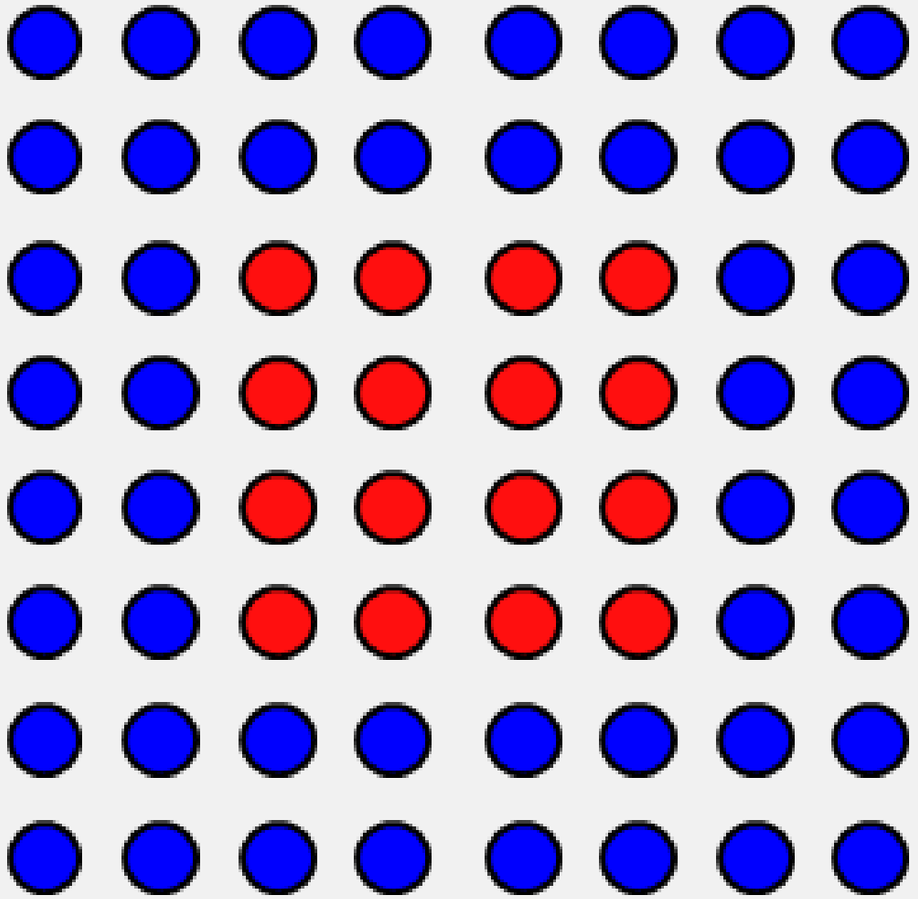


Nonlinear class boundaries

There are non-linear problems:

- Can be linearized?
- Can we change the features measured?
- Representation change?

If “no”, then there are non-linear alternatives.

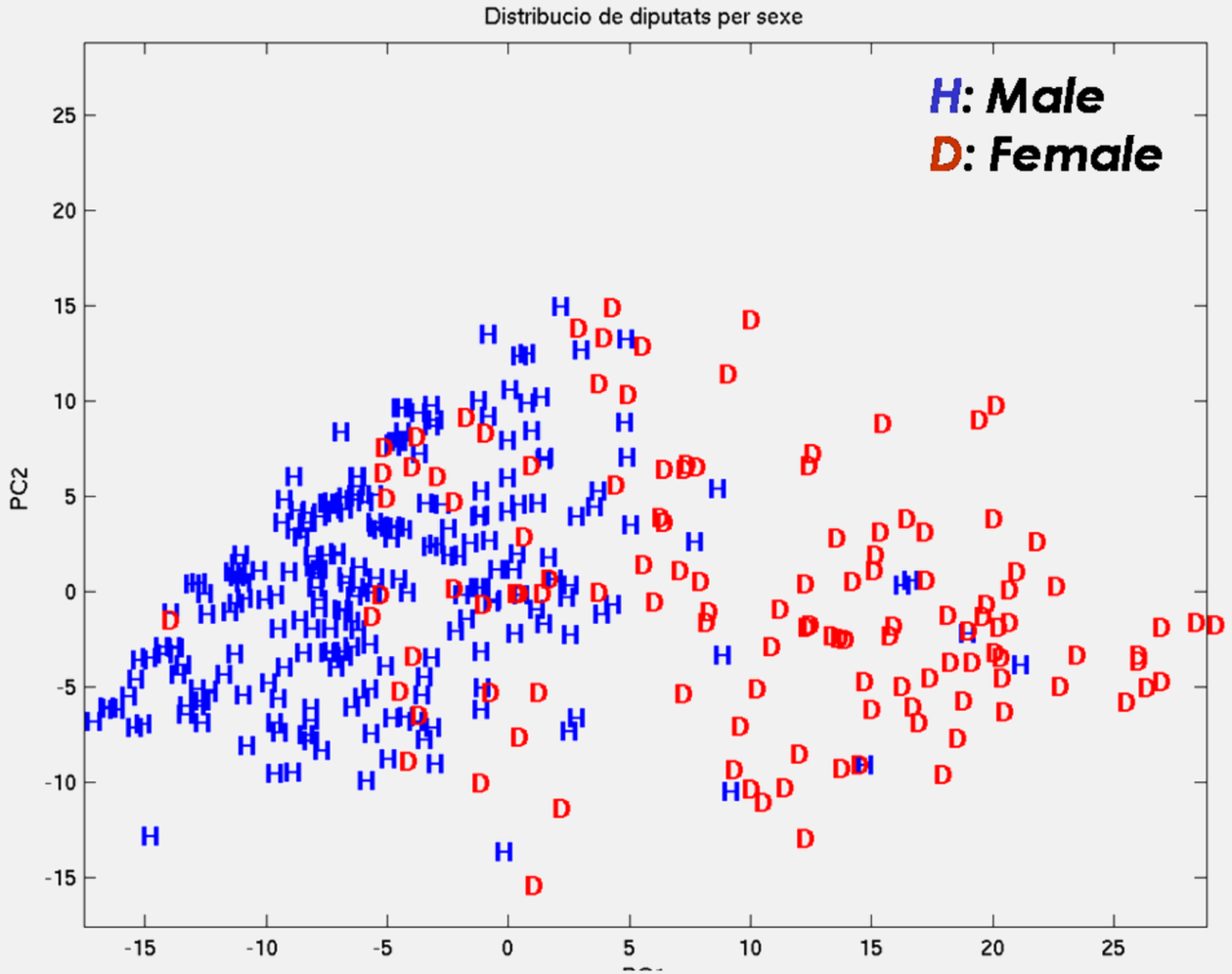


k-nearest neighbours

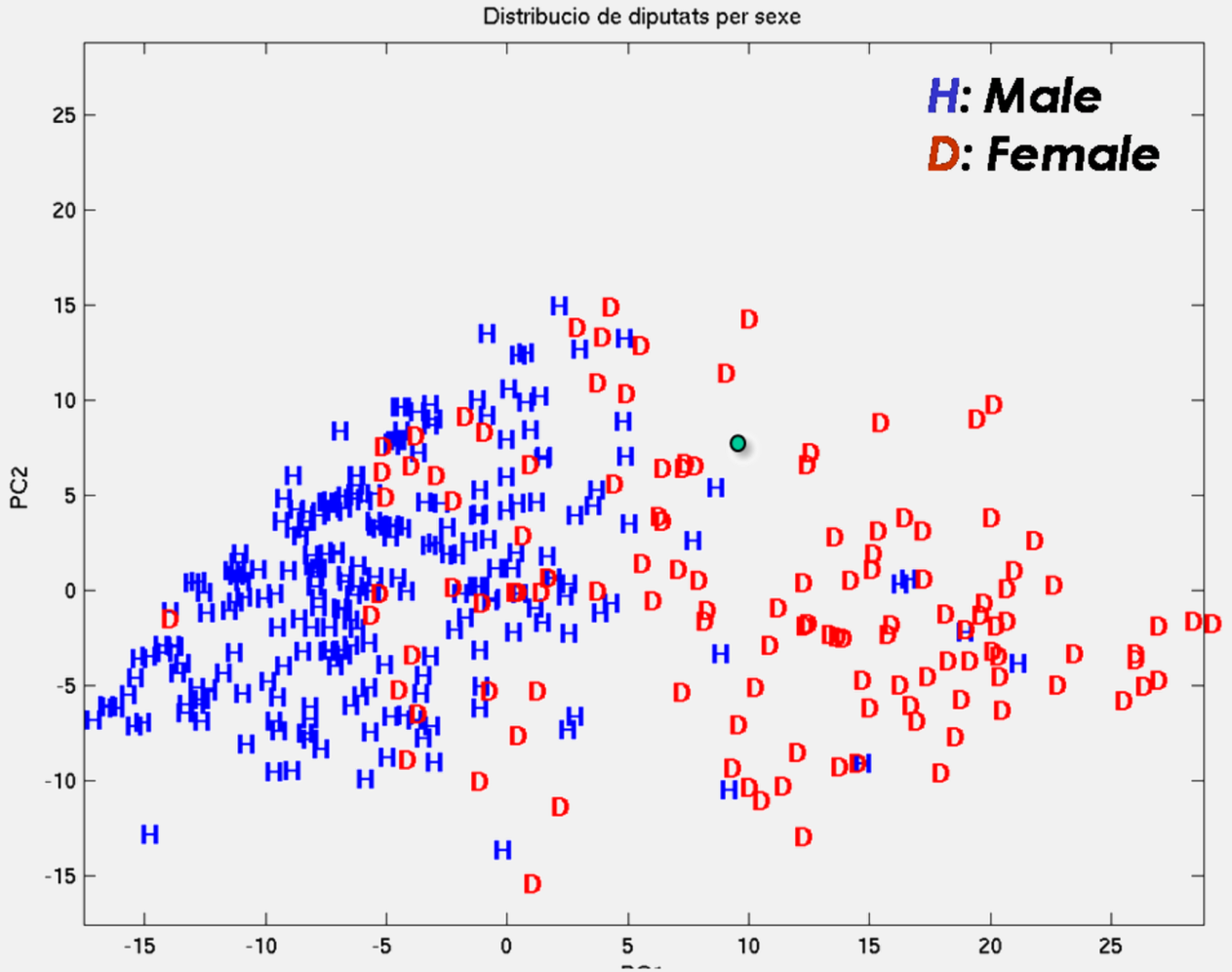
- Choose a distance measure and a value of **k**.
- Find the k nearest training observations to the new observation.
- For classification, combine their labels, commonly by majority vote.
- For regression, combine their numerical responses, commonly by averaging.

Scale features appropriately. Select k using training-only validation; prediction cost can be substantial for large datasets.

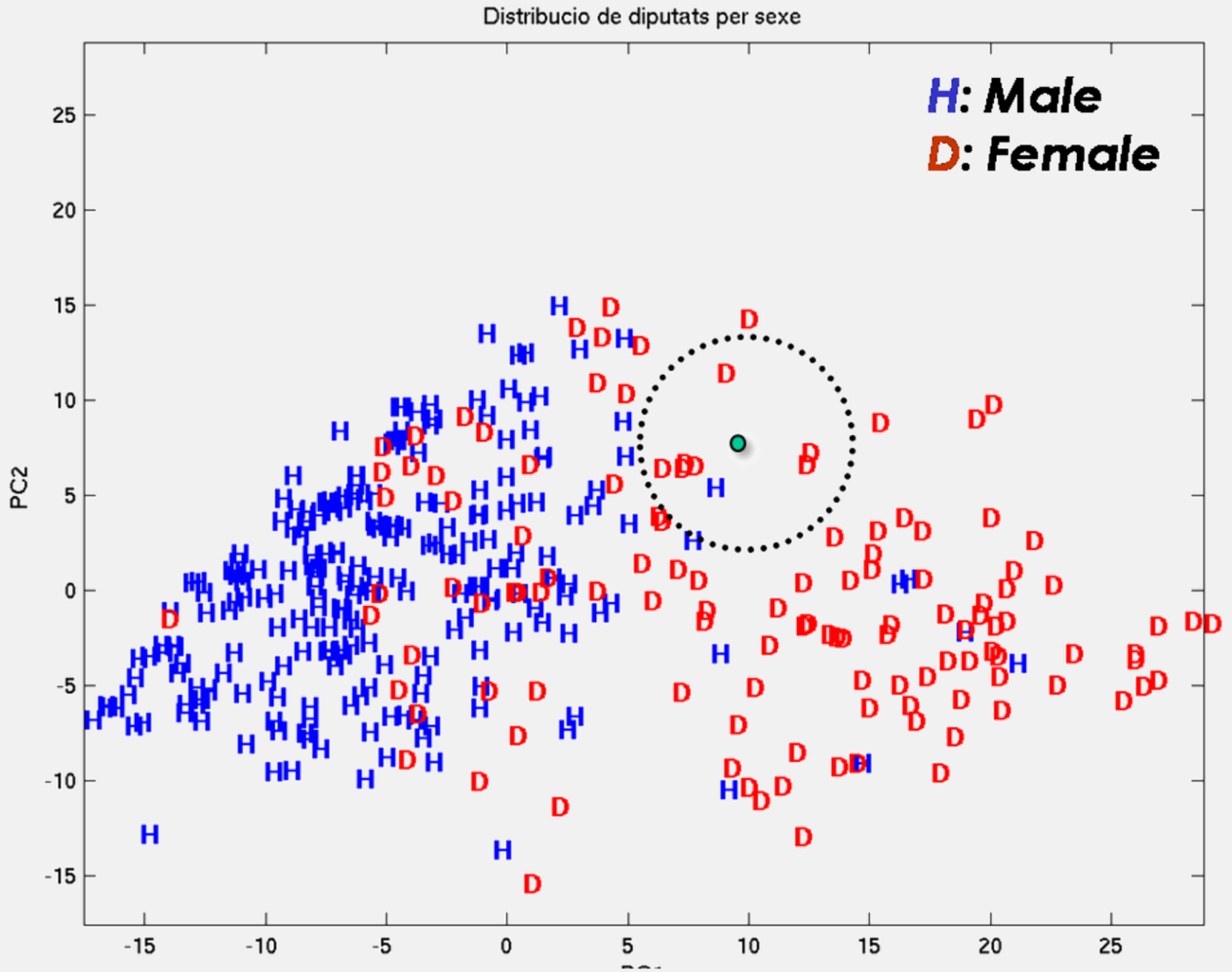
k-NN (1/4)


Distribucio de diputats per sexe
H: Male
D: Female
PC2
25
20
15
10
5
0
-5
-10
-15
-15
-10
-5
0
5
10
15
20
25

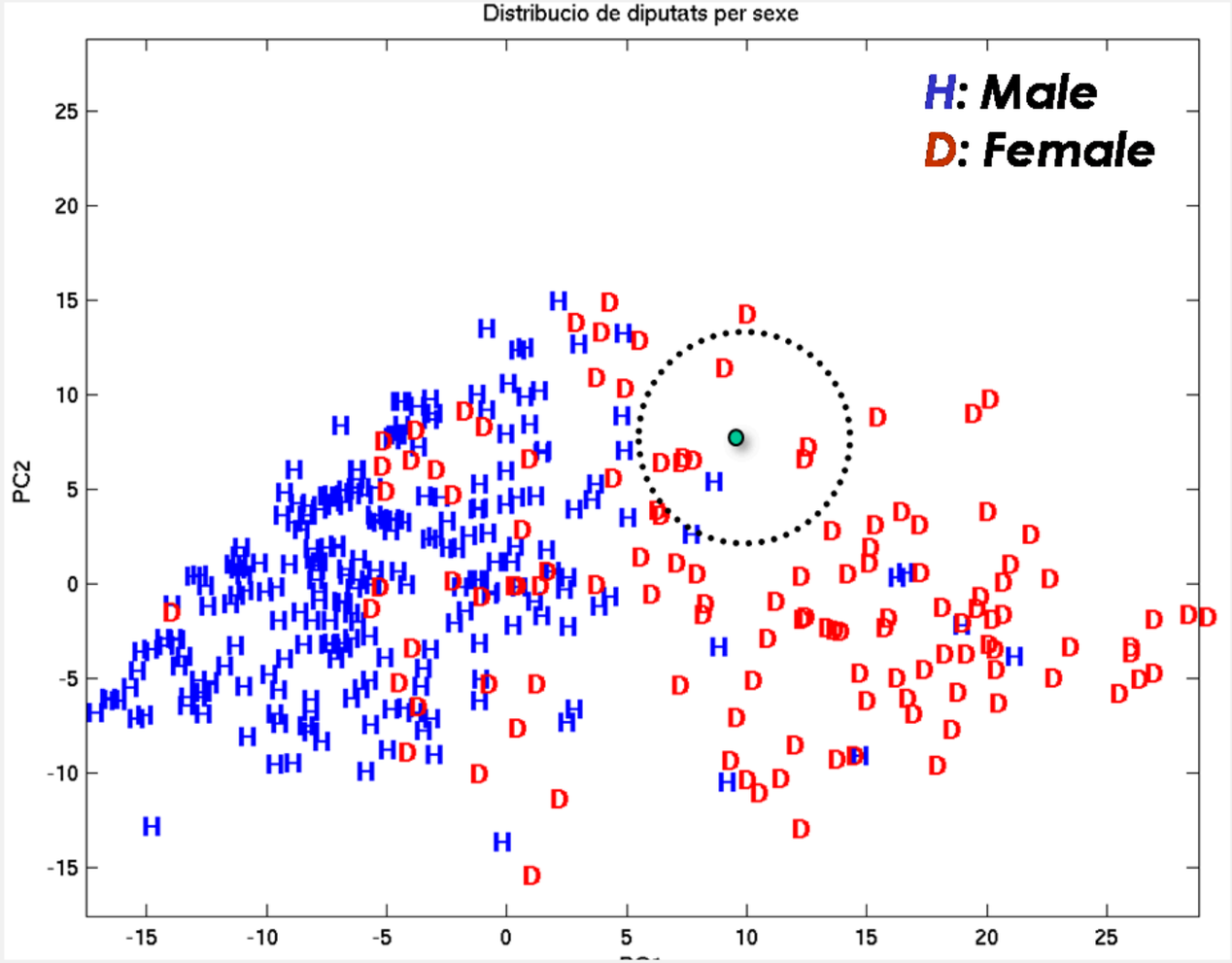
k-NN (2/4)


Distribucio de diputats per sexe
H: Male
D: Female
PC2
25
20
15
10
5
0
-5
-10
-15
-15
-10
-5
0
5
10
15
20
25

k-NN (3/4)

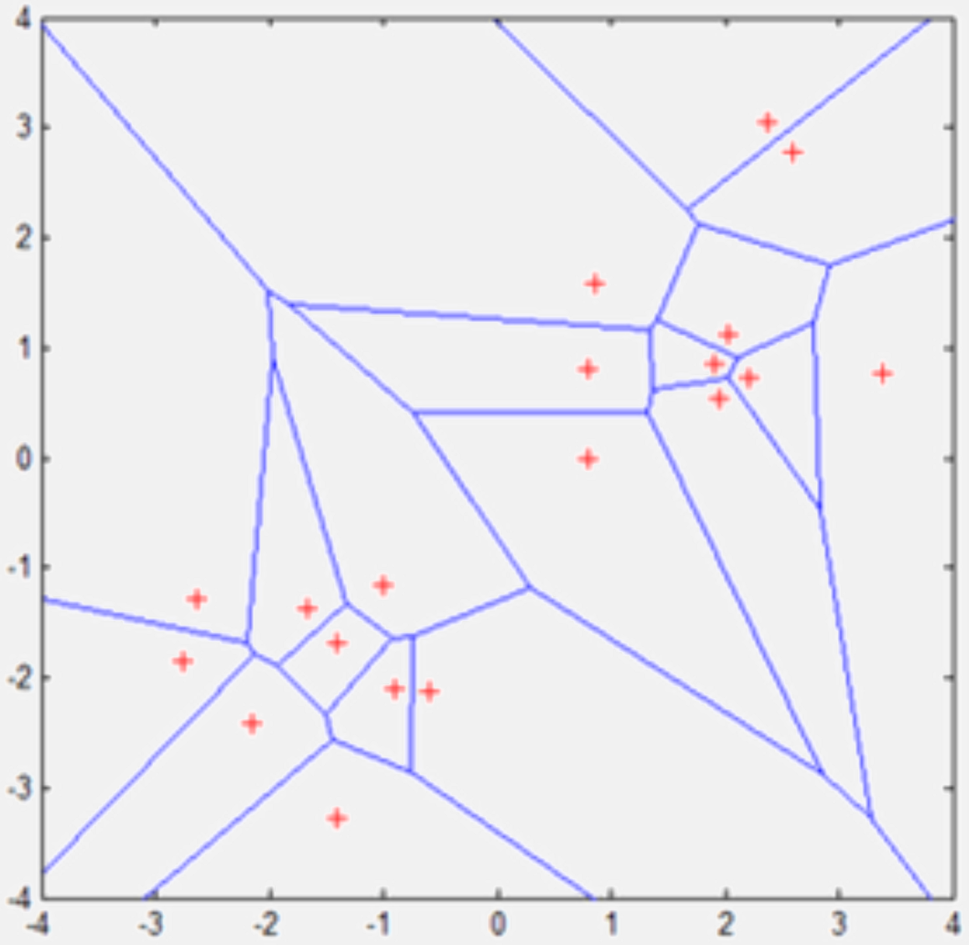

Distribucio de diputats per sexe
H: Male
D: Female
PC2
25
20
15
10
5
0
-5
-10
-15
-15
-10
-5
0
5
10
15
20
25

k-NN (4/4)

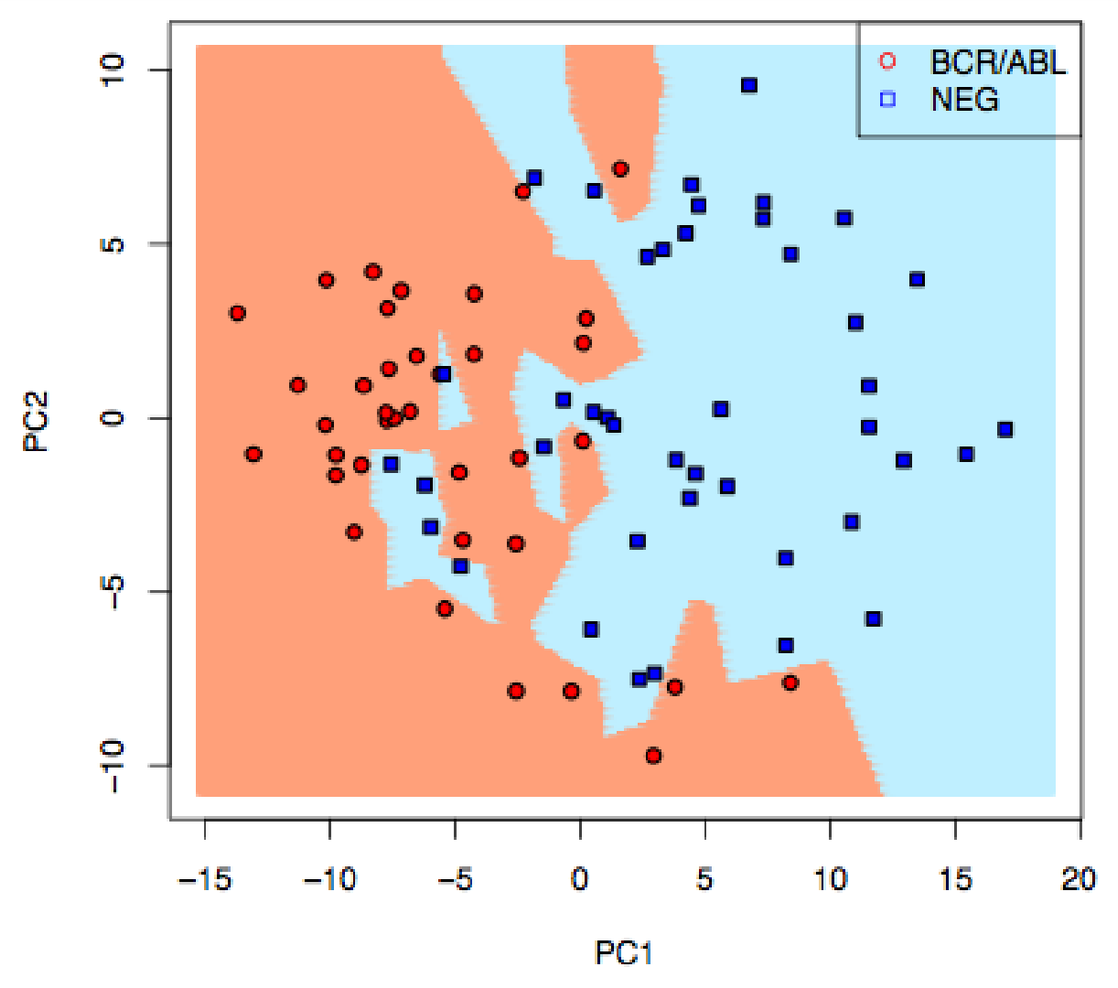


DDDDDDH = D

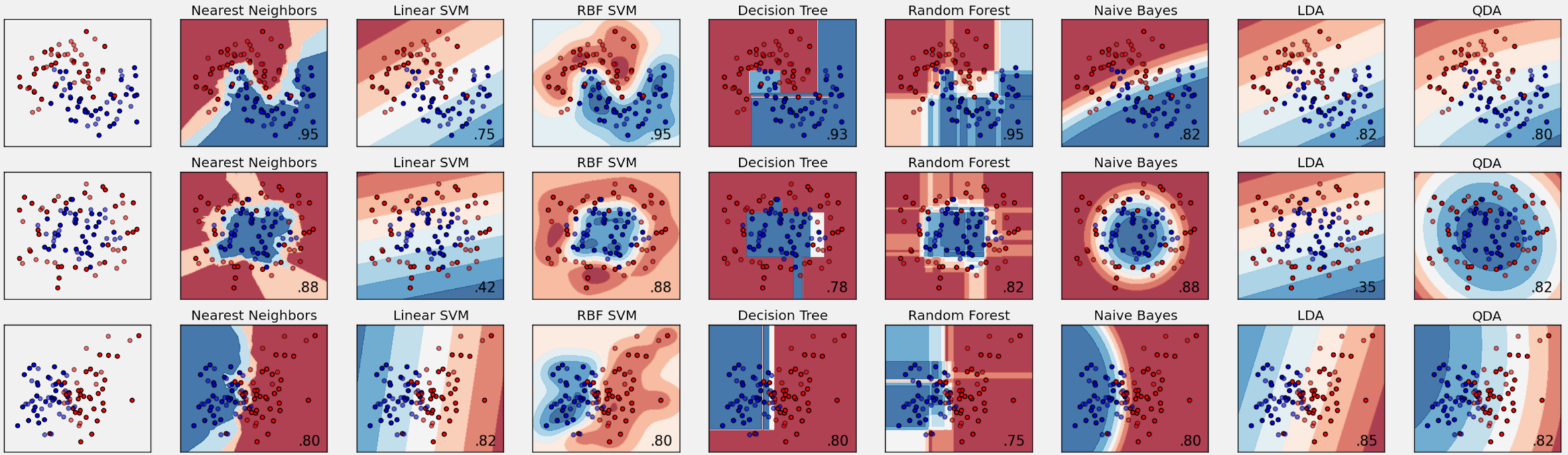
Voronoi Tessellation



Voronoi Tessellation



Classification algorithms in sklearn



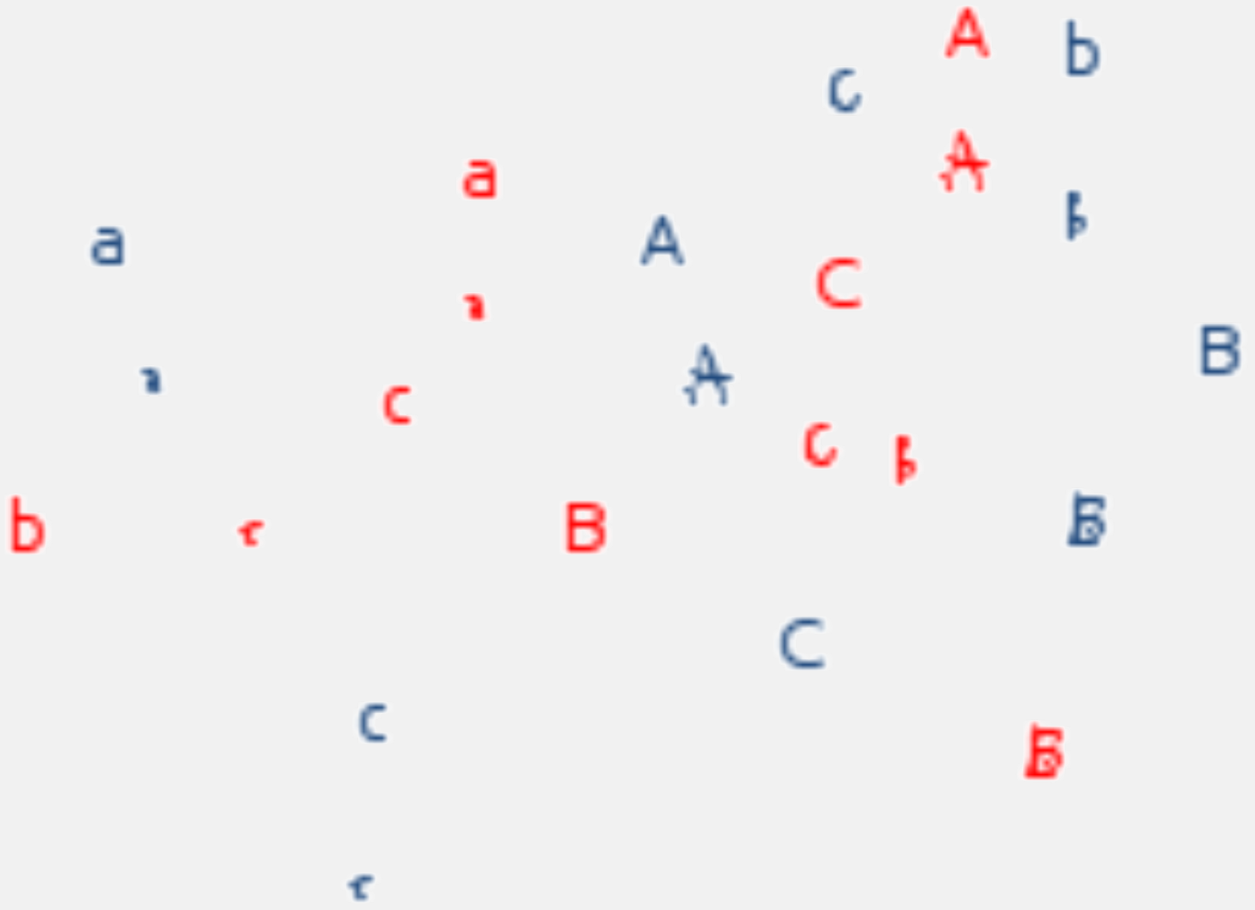
Clustering

Partitioning of a data set into subsets (clusters). Individuals in each subset share some common trait according to some defined distance measure The most typical trait is proximity

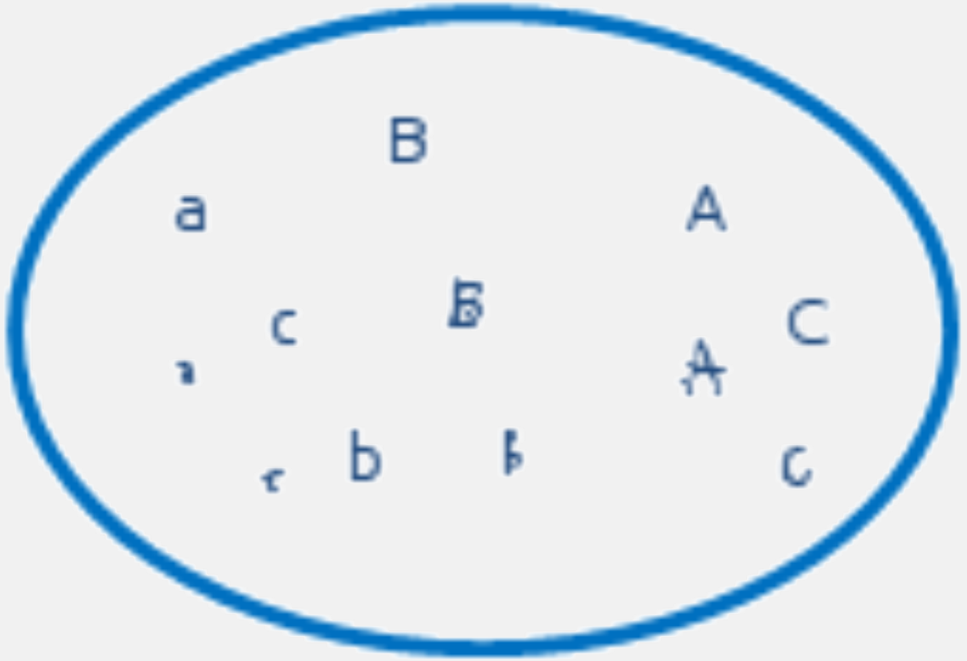
Data clustering is a common technique for statistical data analysis. There are a number of techniques available

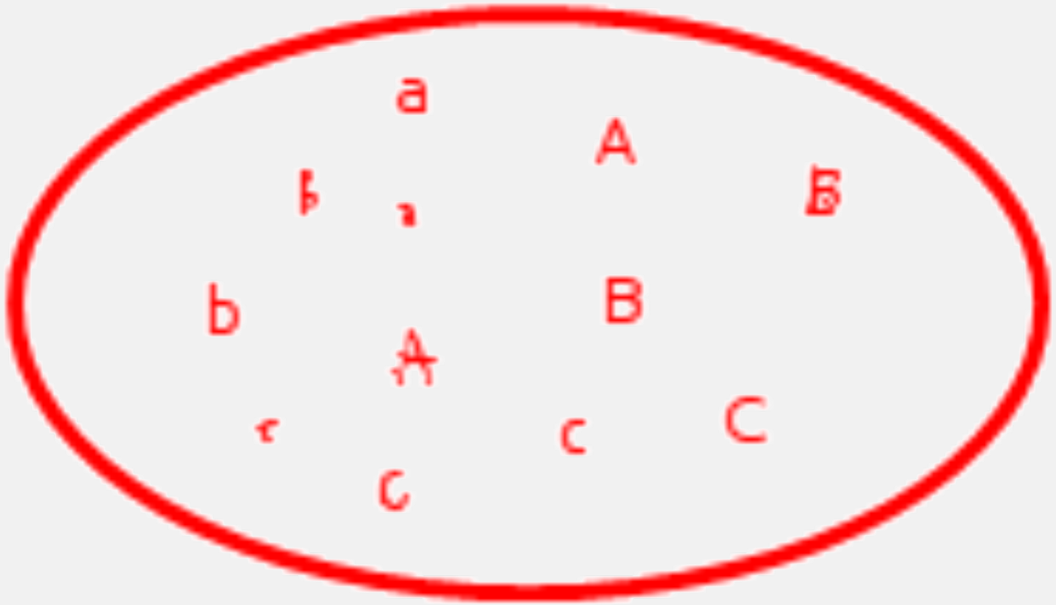
- Hierarchical (e.g. agglomerative clustering)
- Partitional (e.g. k-means clustering)
- Graph base (Spectral)

Clustering (1/5)

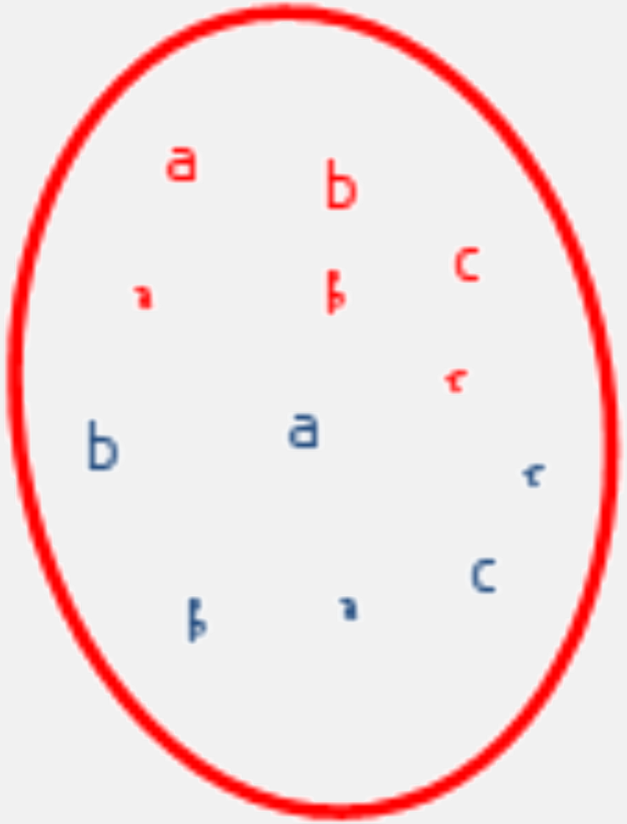


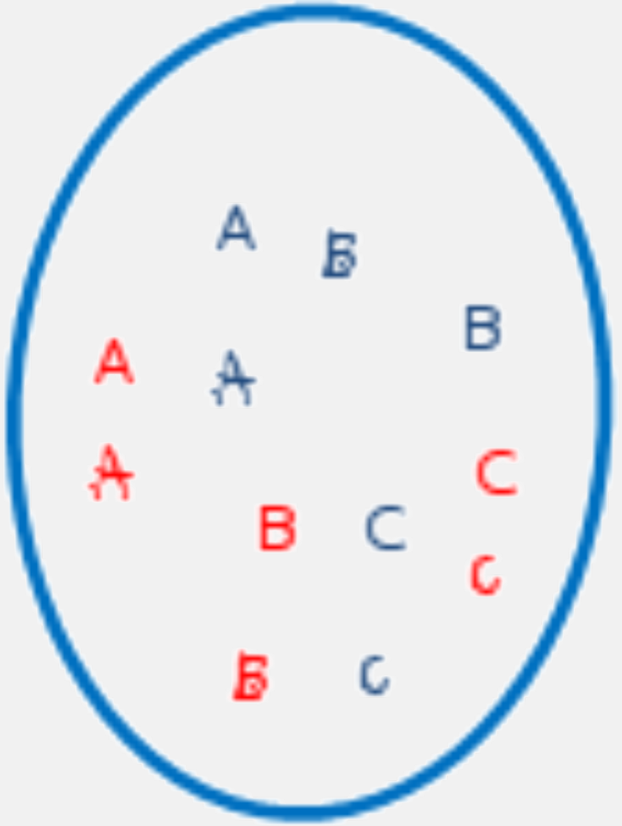
Clustering (2/5)



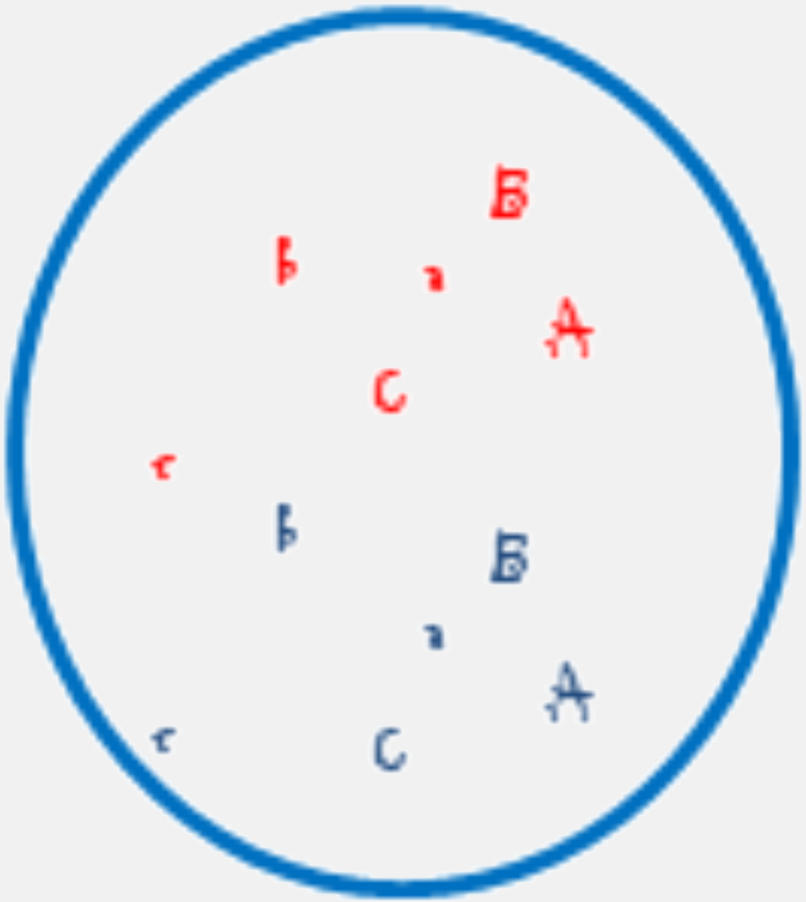


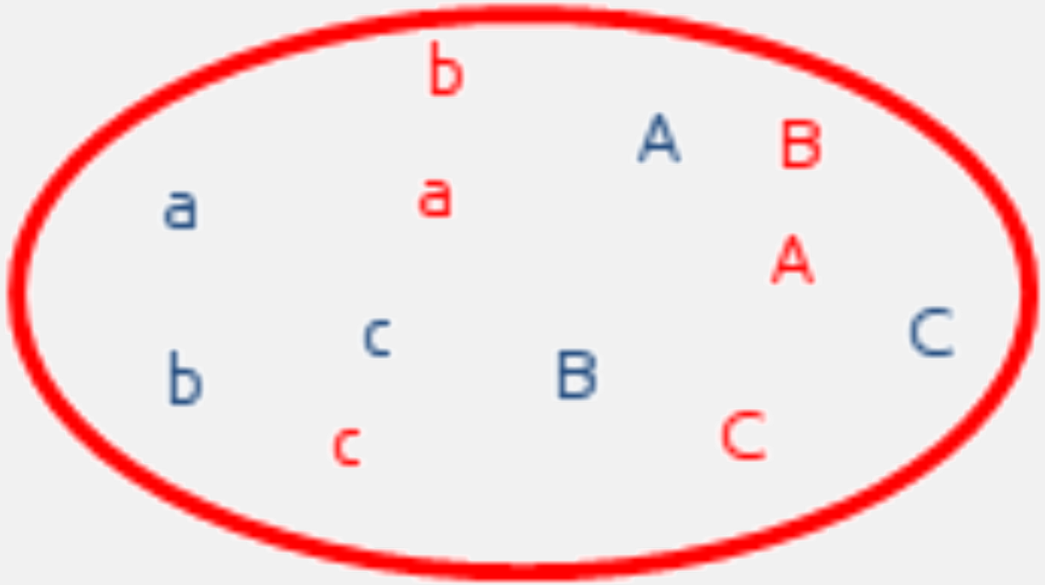
Clustering (3/5)



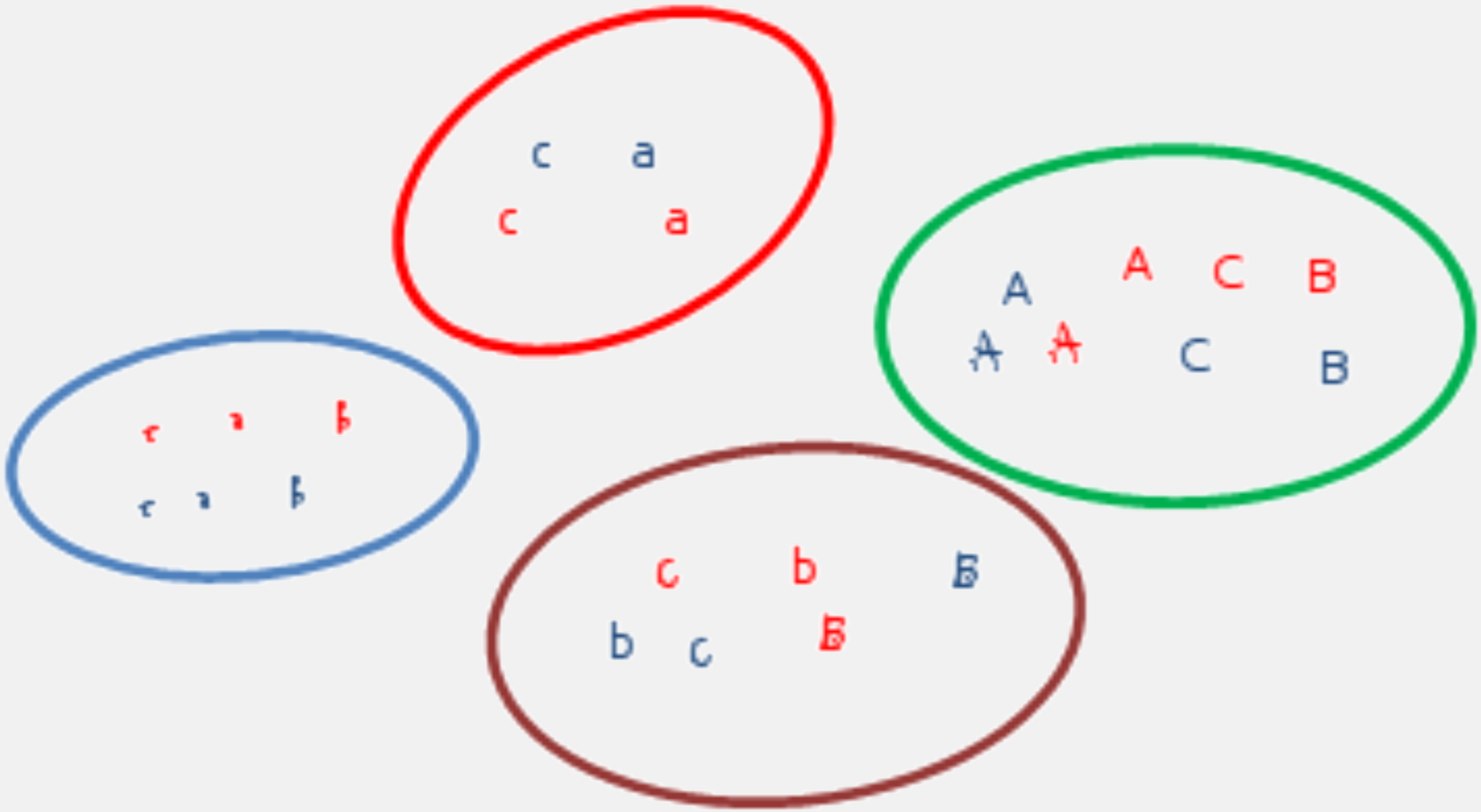


Clustering (4/5)

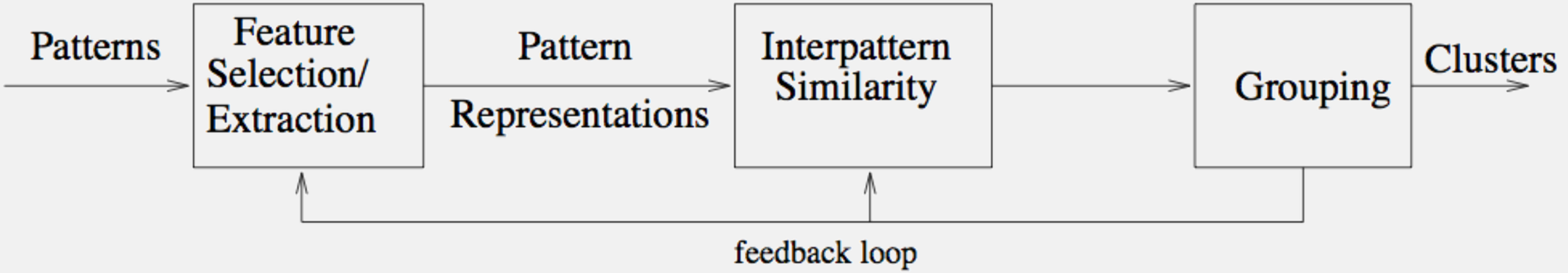




Clustering (5/5)



Clustering


Patterns
Feature Selection/ Extraction
Pattern Representations
Interpattern Similarity
Grouping
Clusters
feedback loop

A metric satisfies non-negativity, identity, symmetry, and the triangle inequality:

$$d(x,y) = 0 \iff x = y, \qquad d(x,y) = d(y,x),$$

$$d(x,y) \le d(x,z) + d(z,y).$$

For $p \ge 1$, a common norm-induced distance is

$$d_p(x,y) = \left(\sum_i |x_i - y_i|^p \right)^{1/p}.$$

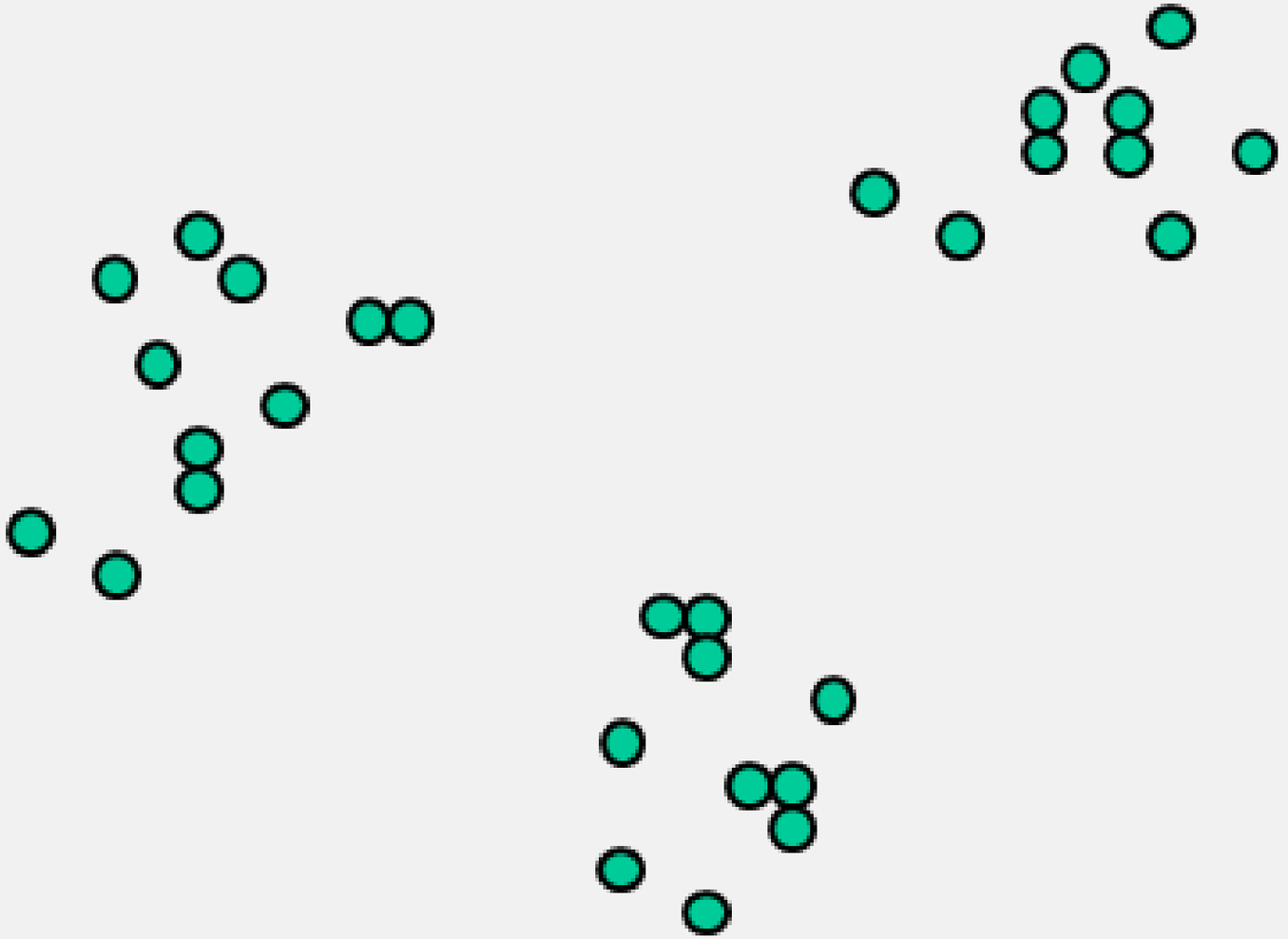
$p = 2$ gives Euclidean distance. Not every metric is induced by a norm.

k-means

The most common clustering algorithm, (J. B. MacQueen, 1967)

1. Ask user how many clusters are desired.
2. Randomly guess k cluster center locations.
3. Each pattern is assigned to the closest center.
4. A New center is assigned as the centroid of the assigned patterns.
5. Repeat from (3).

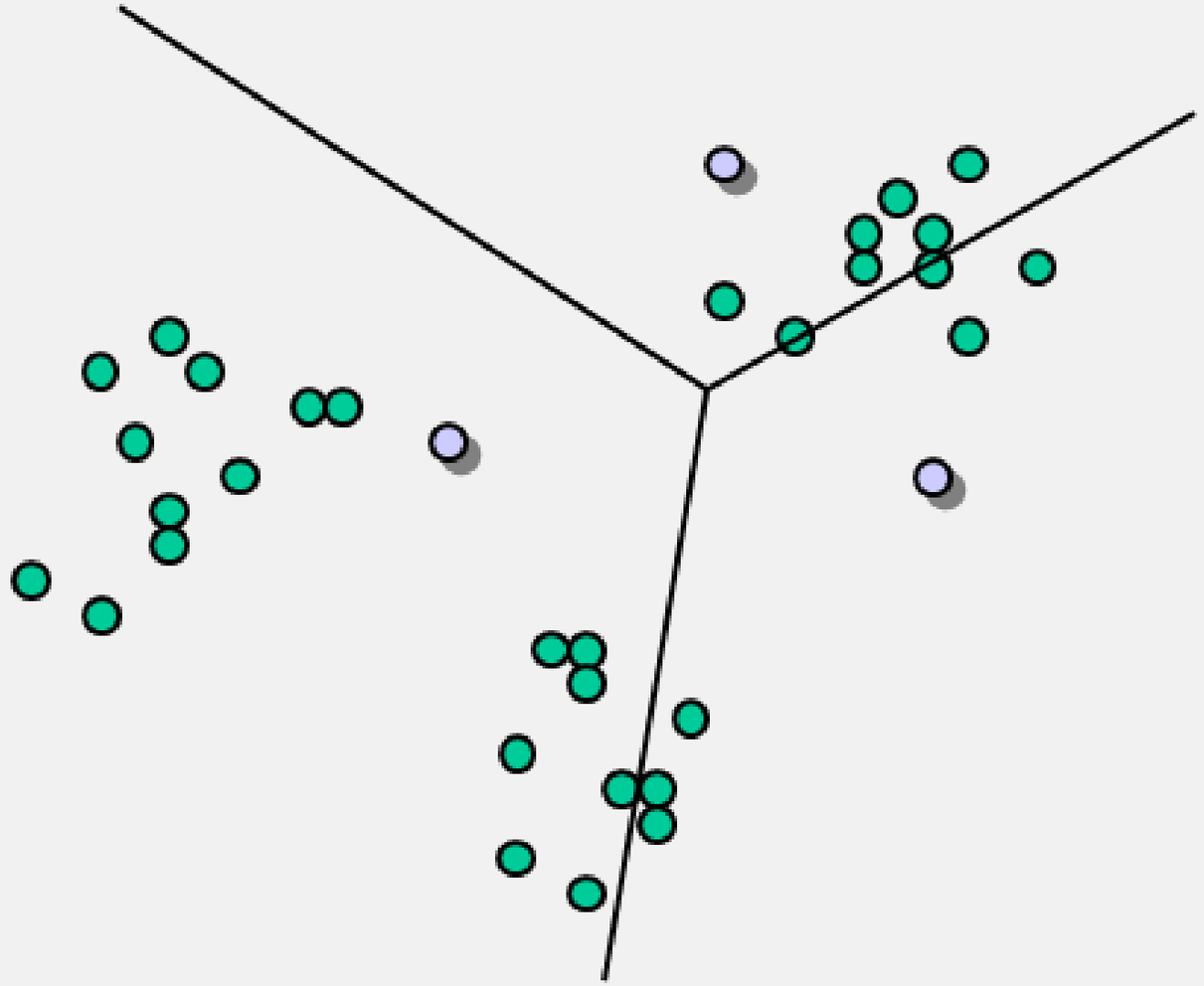
kmeans (1/10)



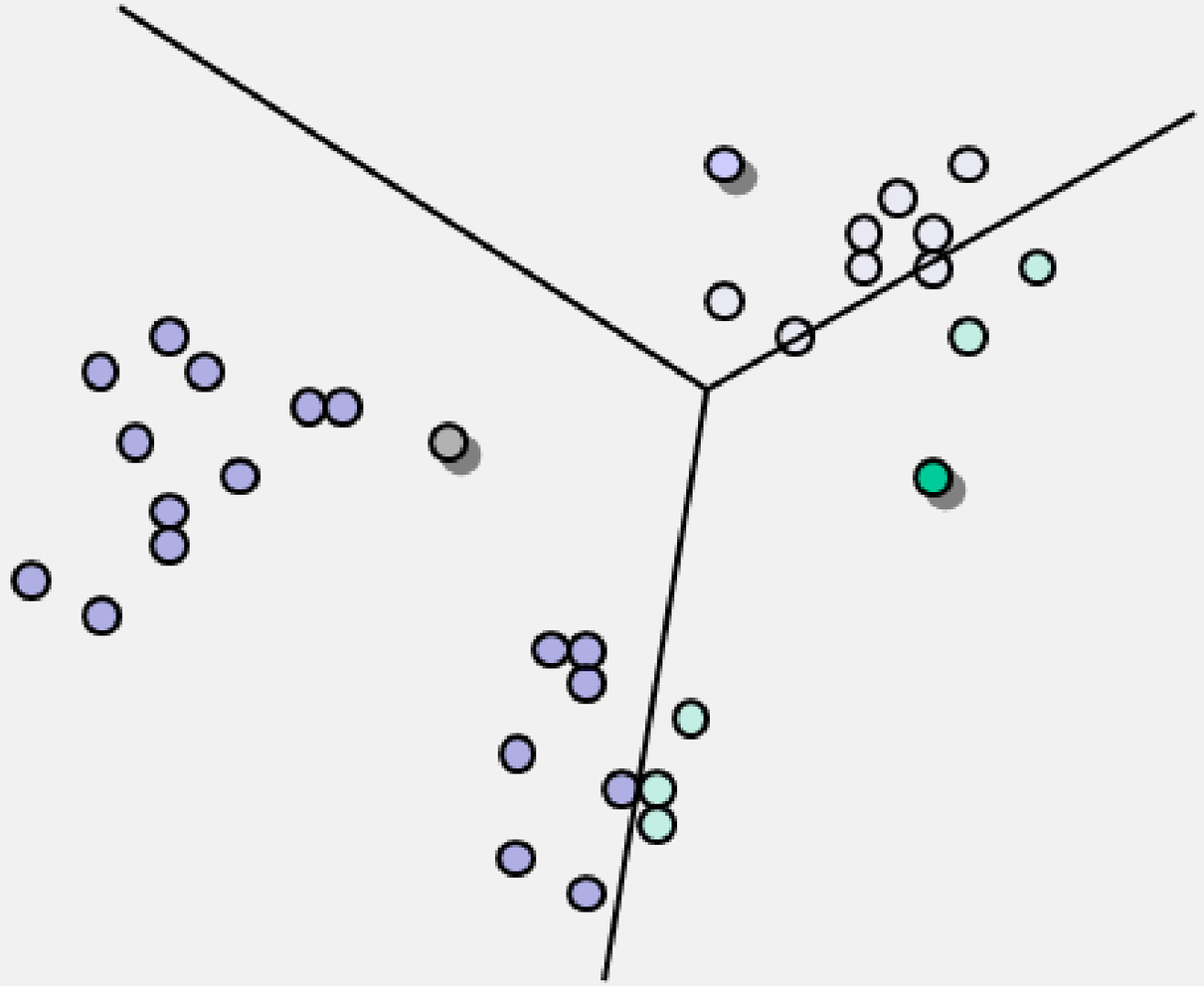
kmeans (2/10)



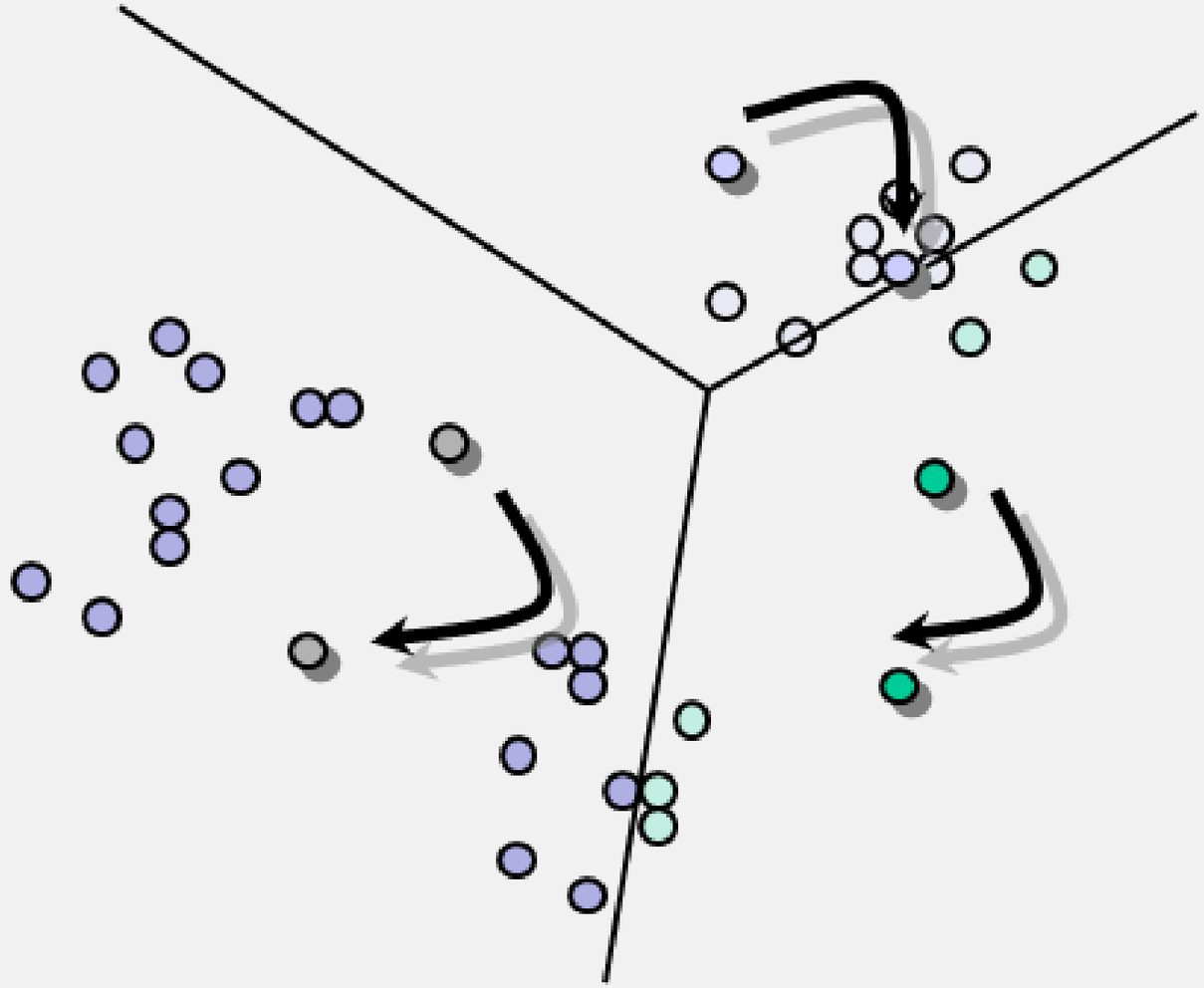
kmeans (3/10)



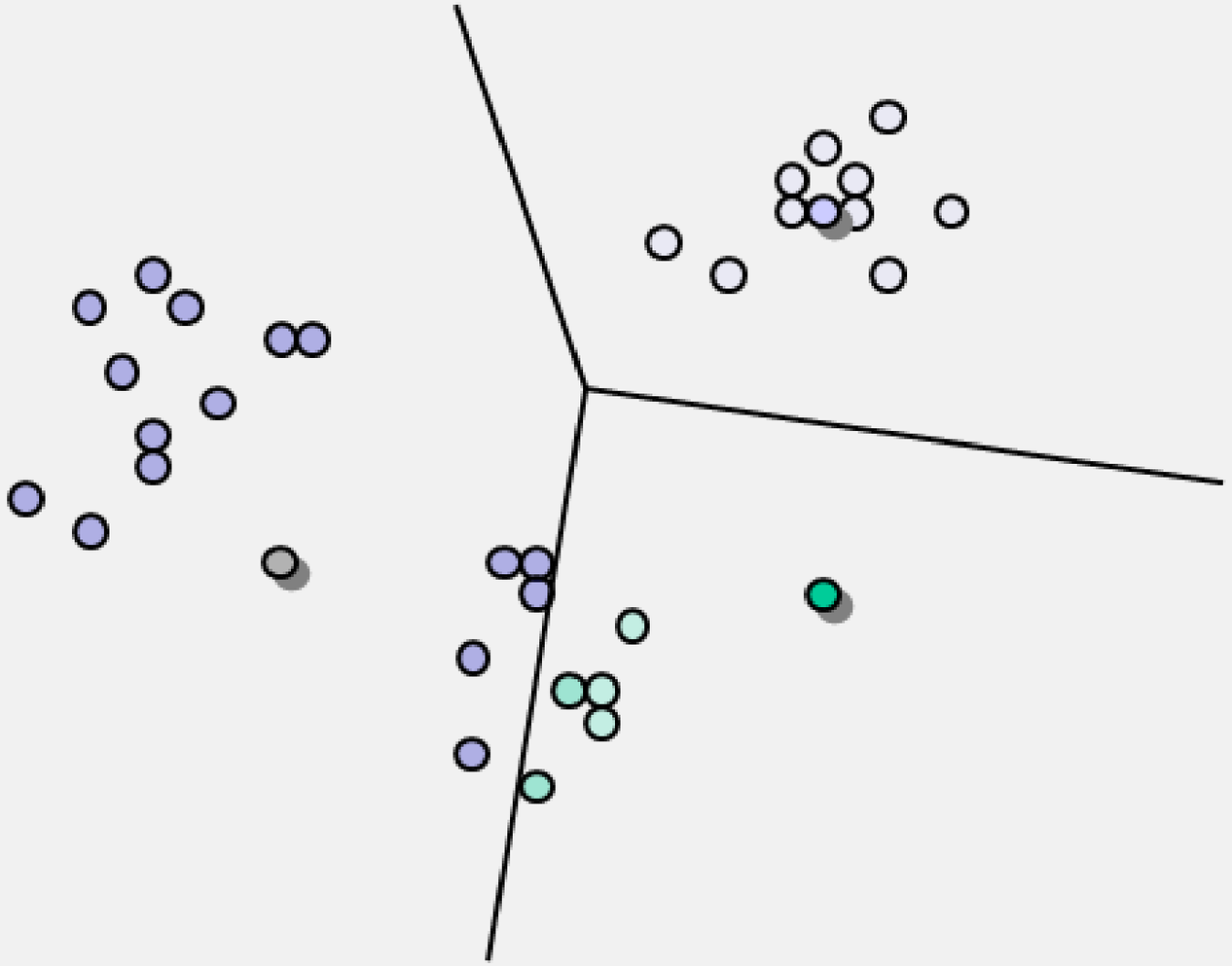
kmeans (4/10)



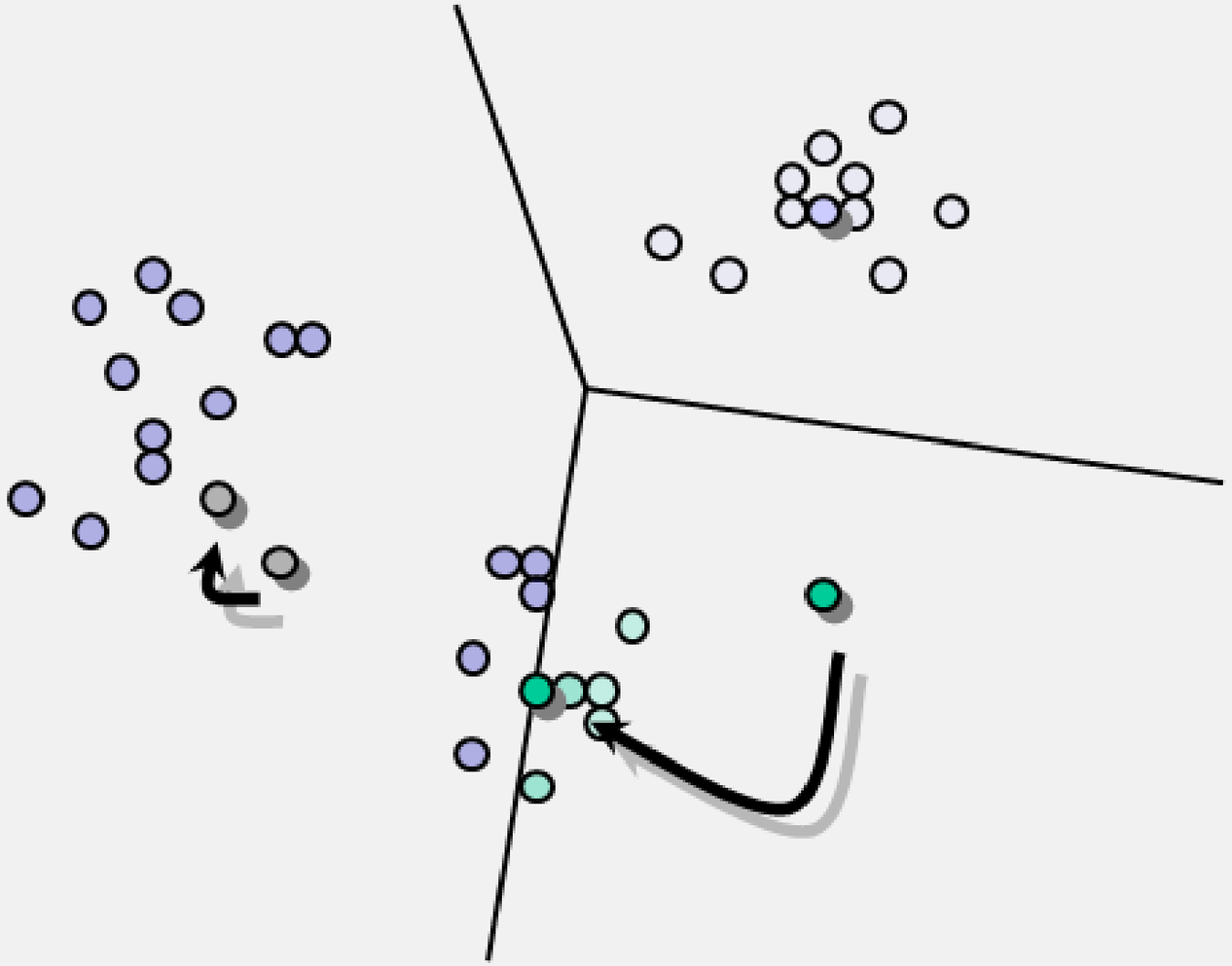
kmeans (5/10)



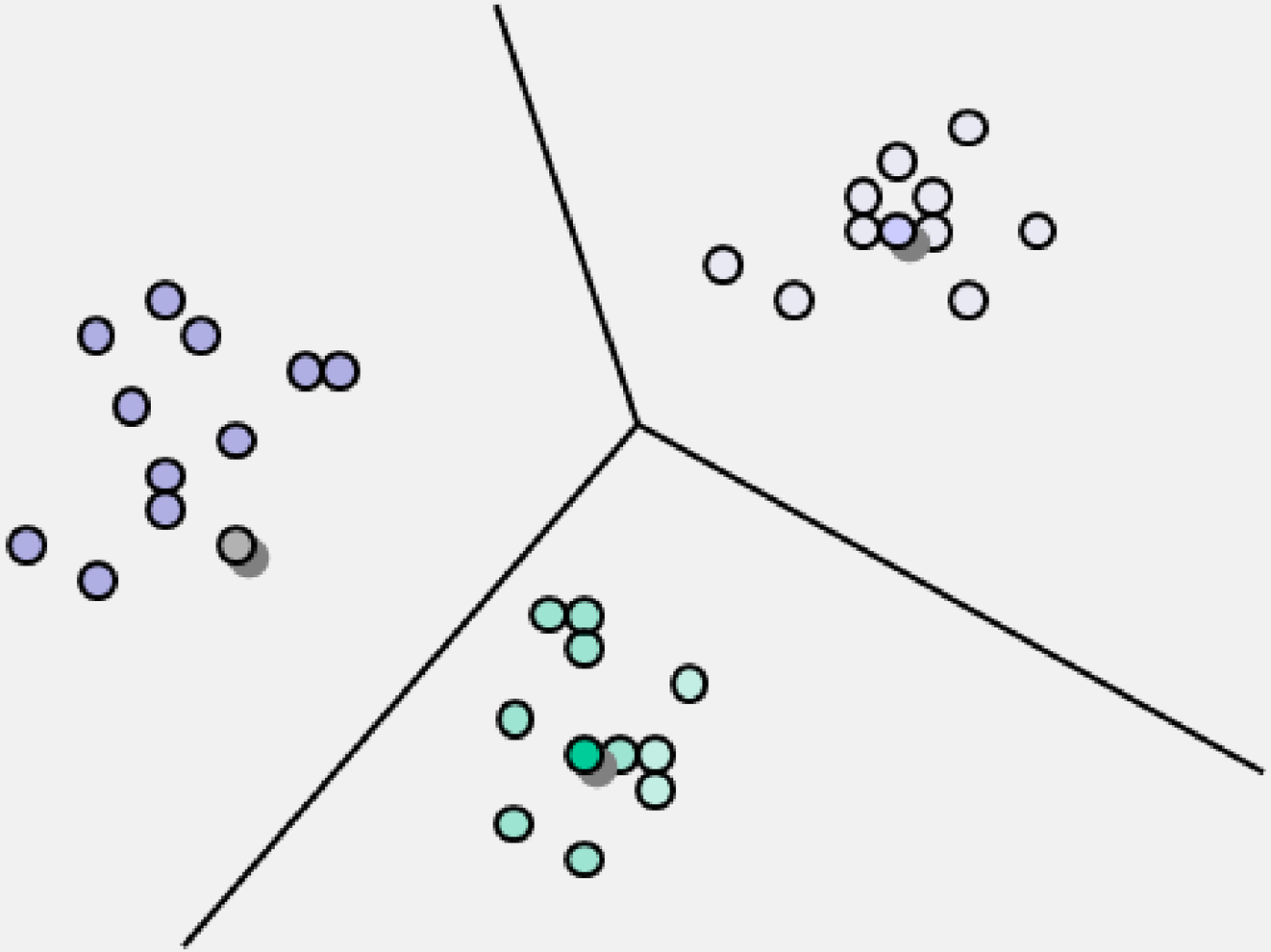
kmeans (6/10)



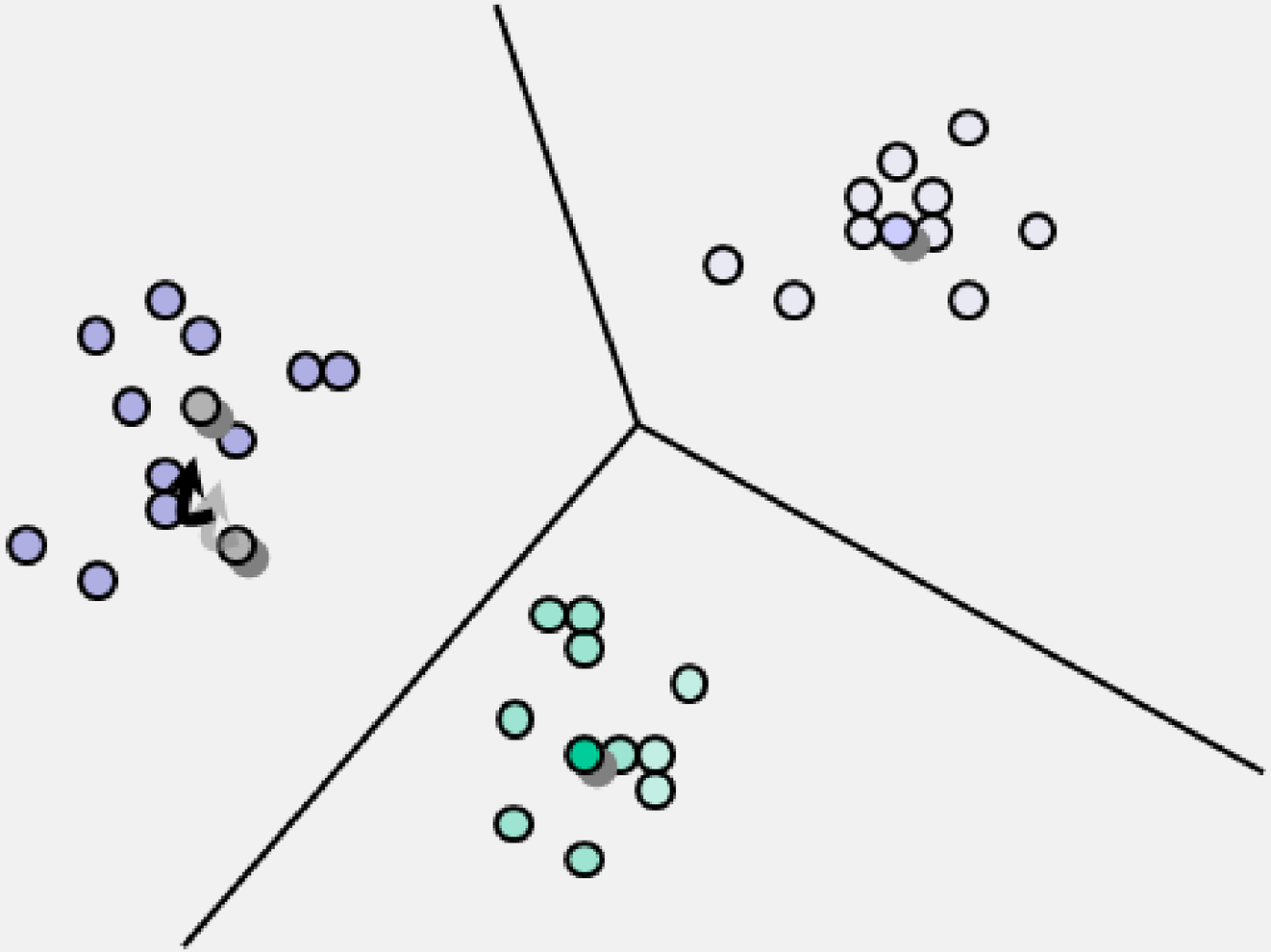
kmeans (7/10)



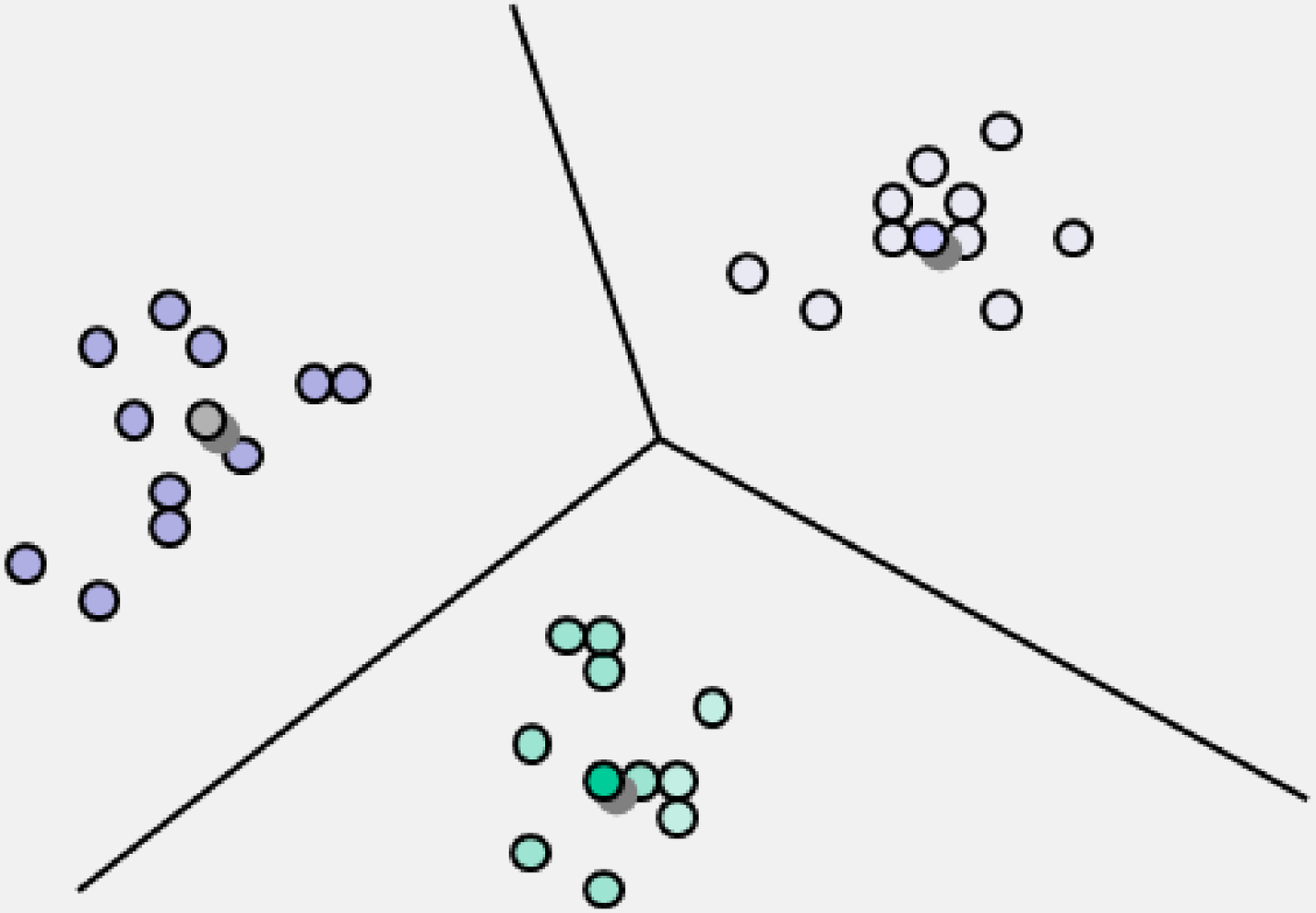
kmeans (8/10)



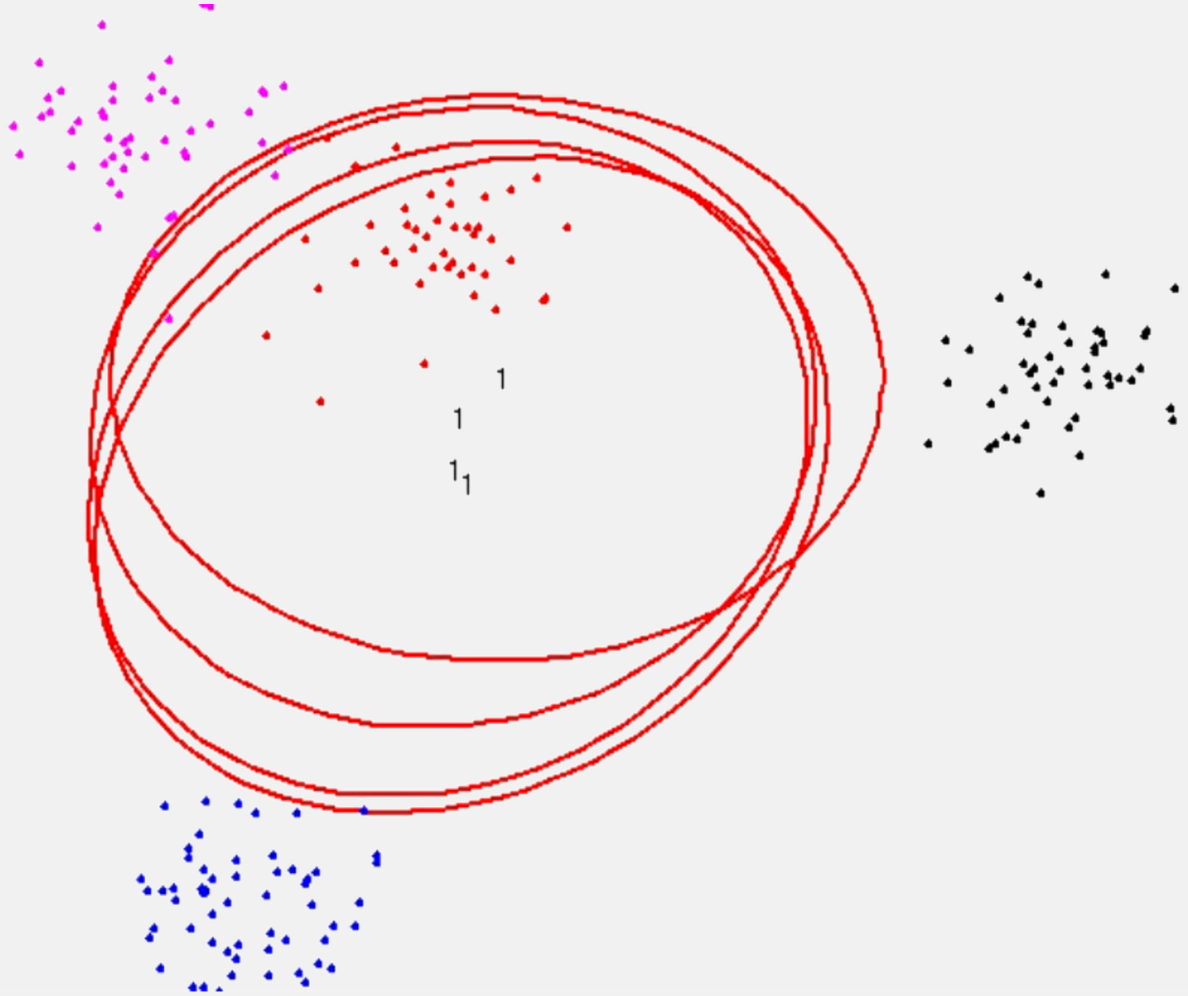
kmeans (9/10)



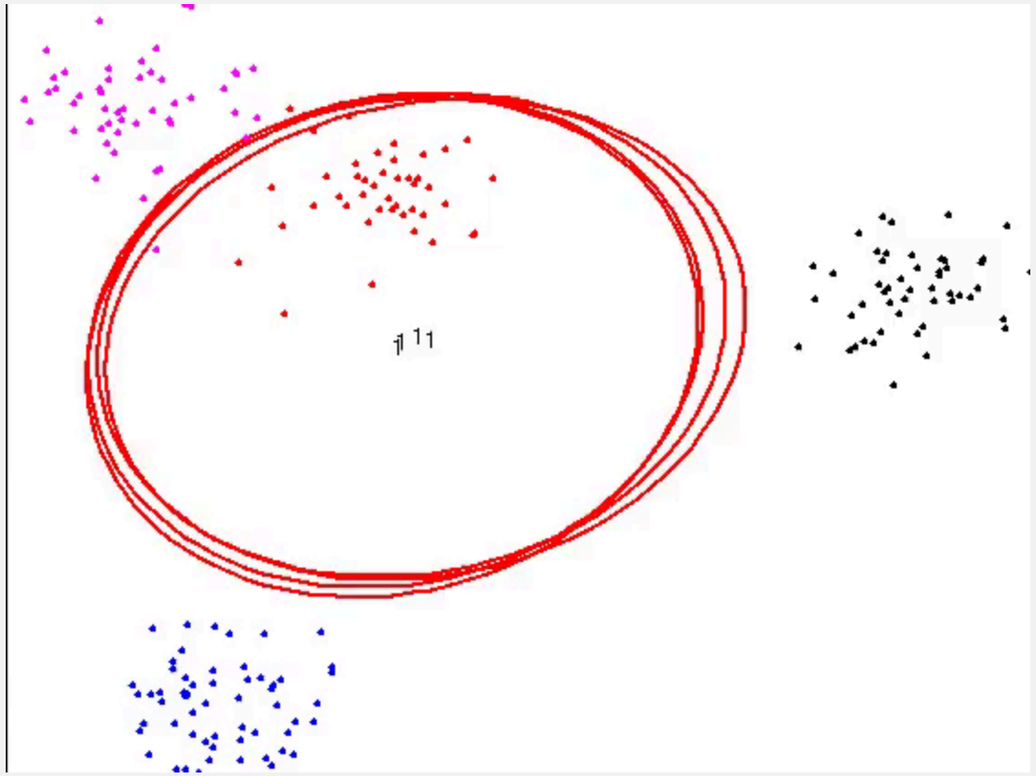
kmeans (10/10)



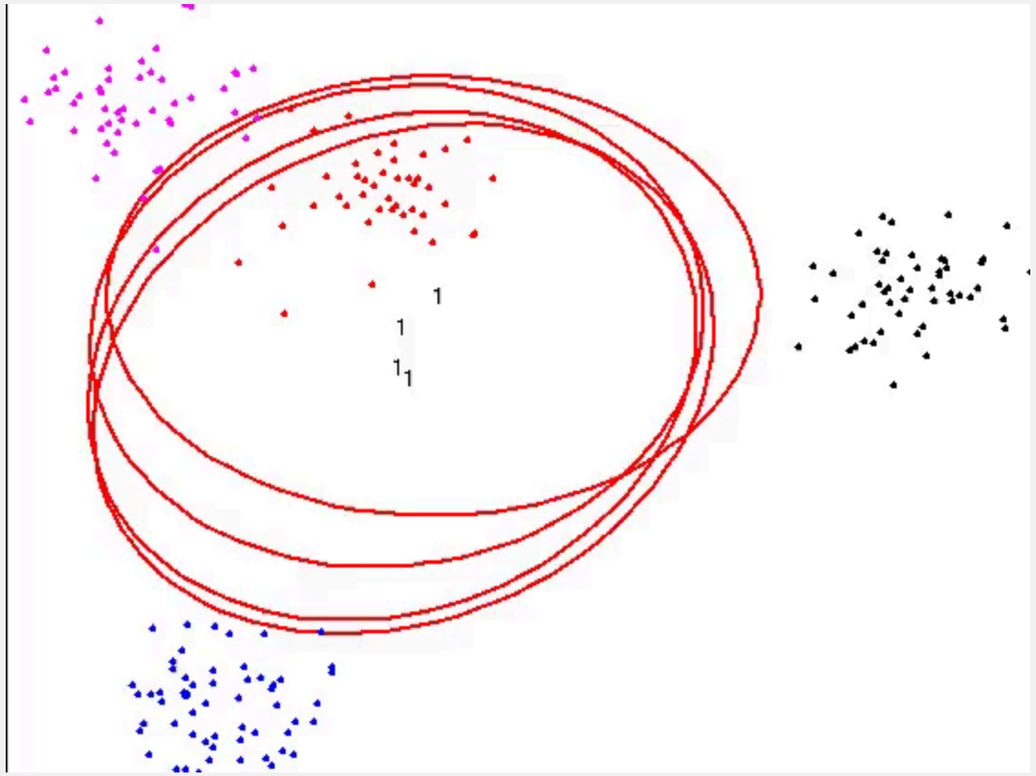
kmeans



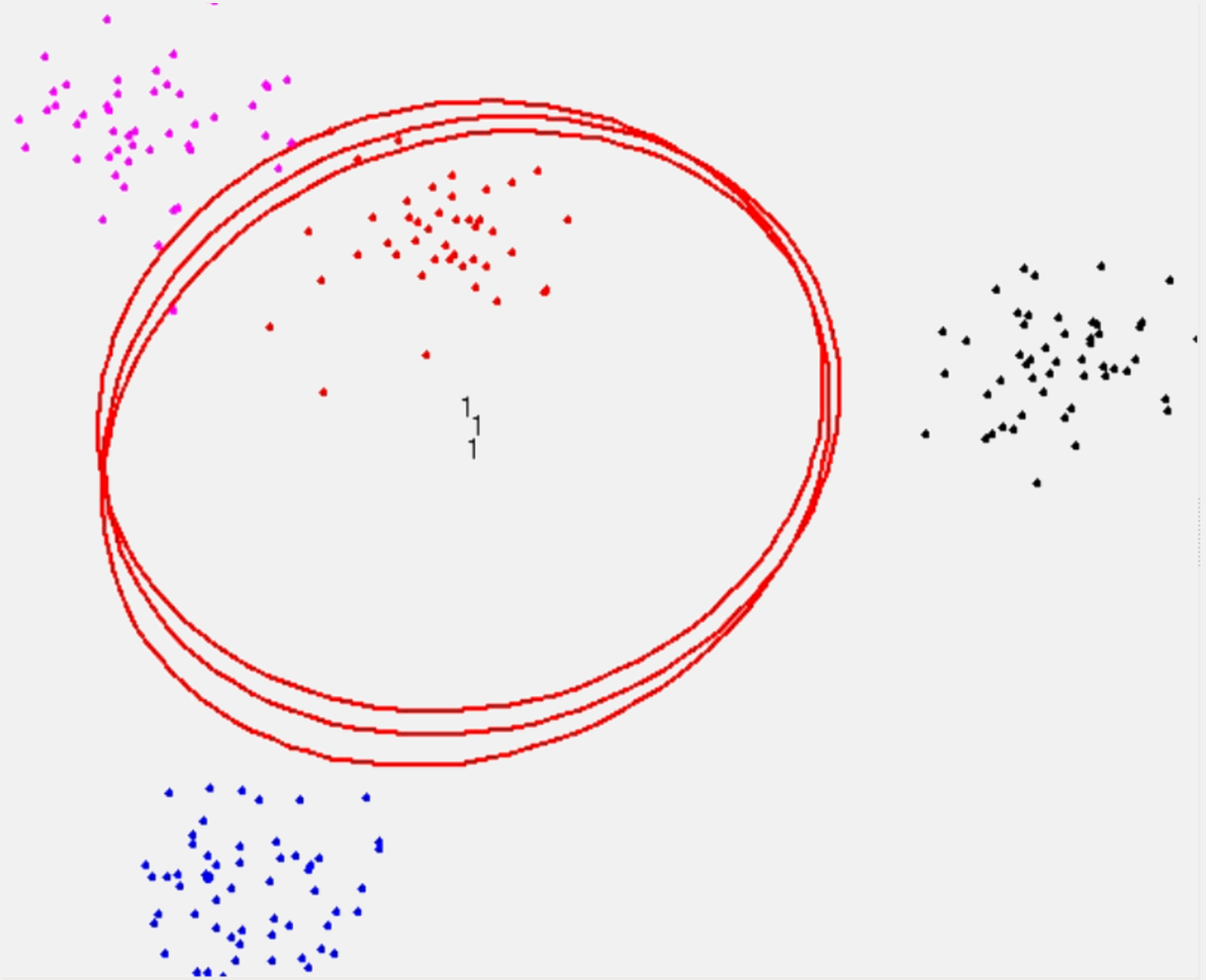
Convergence (1/2)



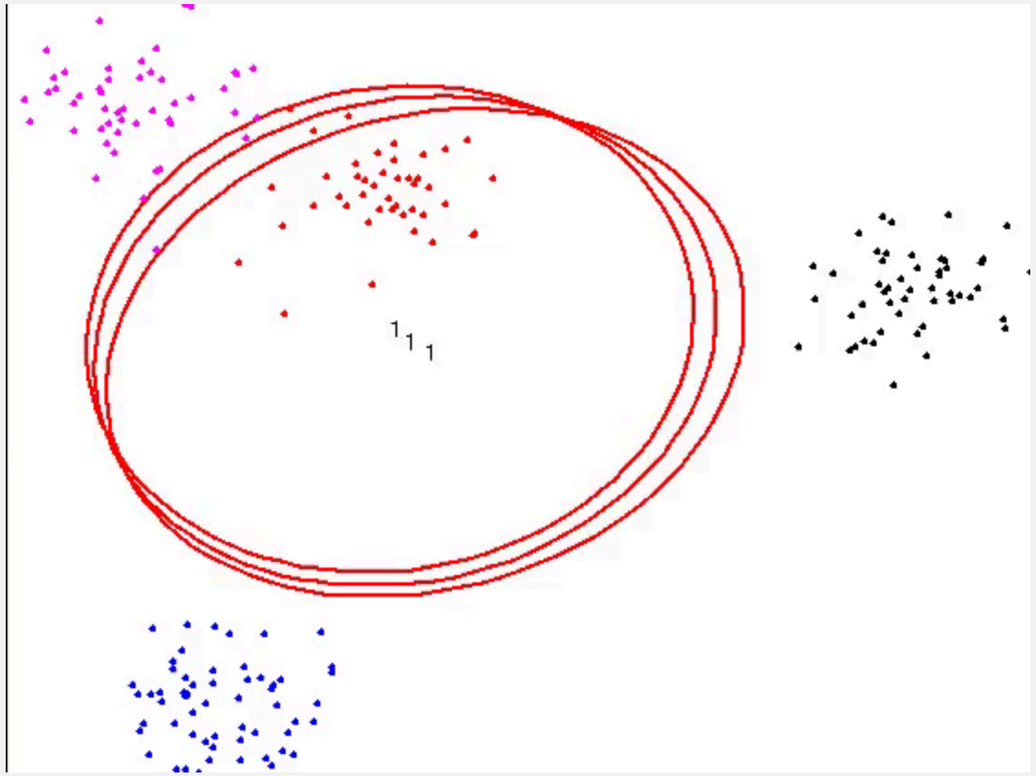
Convergence (2/2)



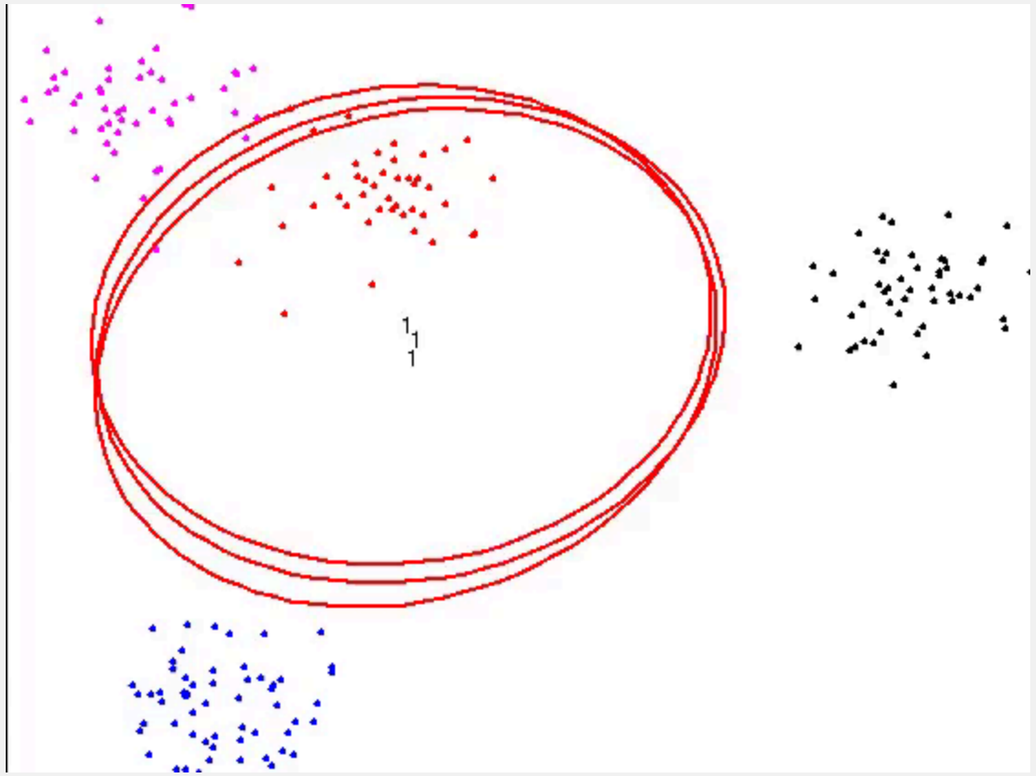
Number of Clusters



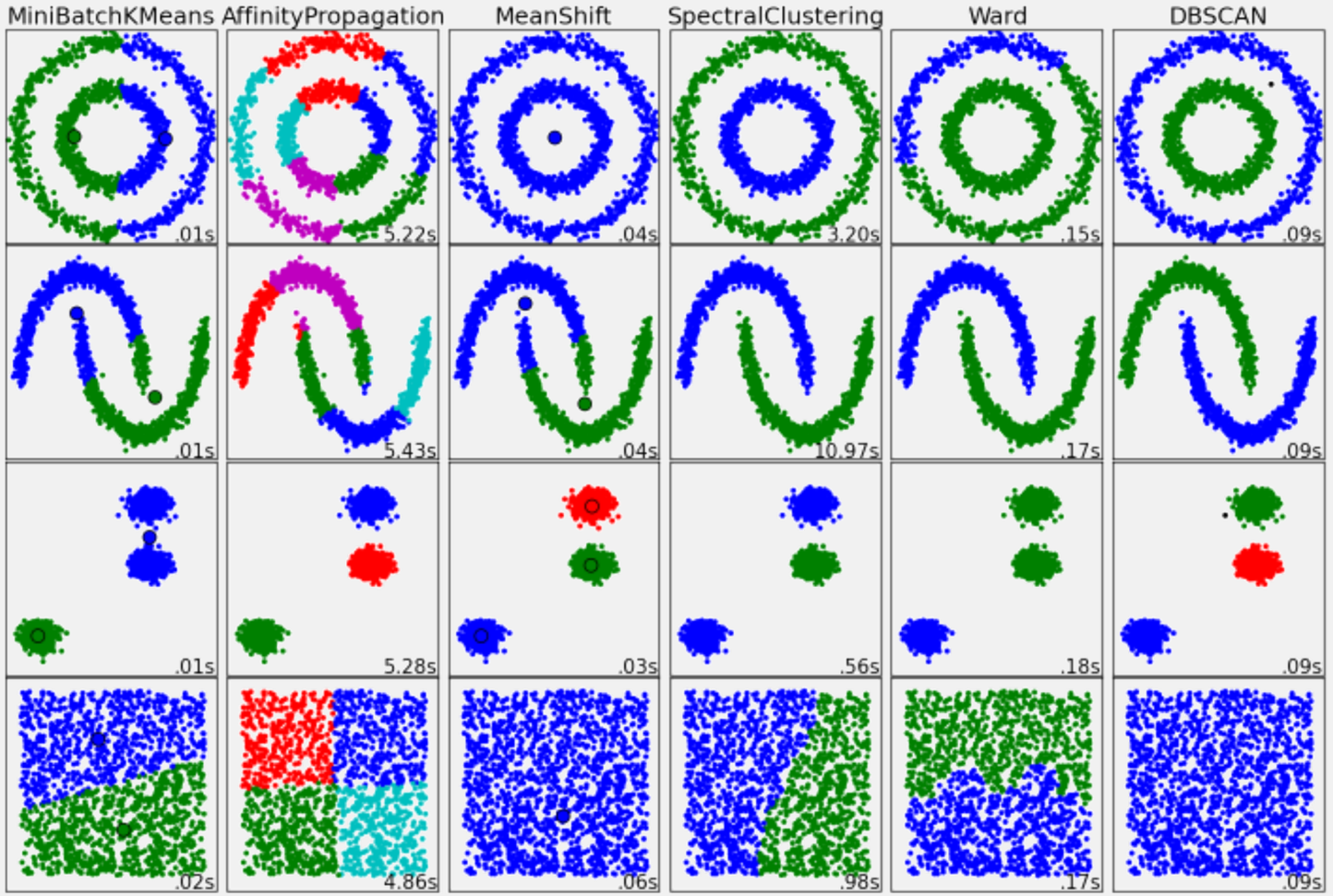
Number of Clusters (1/2)



Number of Clusters (2/2)



Clustering algorithms in sklearn



Validity and Over-Fitting

In multivariable predictive models, over-fitting occurs when a large number of predictor variables is fit to a small N of subjects. A model may “fit” well or perfectly, even if no real relationship.
Simon, JNCI 2003

Direct Consequence of Overfitting

Model performance results are not reproducible in a new set of data.

essay turning points

A meeting with Enrico Fermi

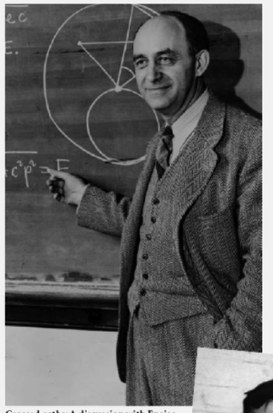
How one intuitive physicist rescued a team from fruitless research.

Freeman Dyson

One of the big turning points in my life was a meeting with Enrico Fermi in the spring of 1953. In a few minutes, Fermi politely but ruthlessly demolished a programme of research that my students and I had been pursuing for several years. He probably saved us from several more years of fruitless wandering along a road that was leading nowhere. I am eternally grateful to him for destroying our illusions and telling us the bitter truth.

Fermi was one of the great physicists of our time, outstanding both as a theorist and as an experimenter. He led the team that built the first nuclear reactor in Chicago in 1942. By 1953 he was head of the team that built the Chicago cyclotron, and was using it to explore the strong forces that hold nuclei together. He made the first accurate measurements of the scattering of mesons by protons, an experiment that gave the most direct evidence then available of the nature of the strong forces.

At that time I was a young professor of theoretical physics at Cornell University, responsible for directing the research of a small army of graduate students and postdocs. I had put them to work calculating meson–proton scattering, so that their theoretical calculations could be compared with Fermi's measurements. In 1948 and 1949 we had made similar calculations of atomic processes, using the theory of quantum electrodynamics, and found spectacular agreement between experiment and theory. Quantum electrodynamics is the theory of electrons and photons interacting through electromagnetic forces. Because the electromagnetic forces are weak, we could calculate the atomic processes precisely. By 1951, we had triumphantly finished the atomic calculations and were looking for fresh fields to conquer. We decided to use the same techniques of calculation to explore the strong nuclear forces. We began by calculating meson–proton scattering, using a theory of the strong forces known as pseudoscalar meson theory. By the spring of 1953, after heroic efforts, we had plotted theoretical graphs of meson–proton scattering. We joyfully observed that our calculated numbers agreed pretty well with Fermi's measured numbers. So I made an appointment to meet with Fermi and show him our results. Proudly, I rode the Greyhound bus from Ithaca to Chicago with a package of our theoretical graphs to show to Fermi.





Crossed paths: A discussion with Enrico Fermi (above) made Freeman Dyson (right) change his career direction.

When I arrived in Fermi's office, I handed the graphs to Fermi, but he hardly glanced at them. He invited me to sit down, and asked me in a friendly way about the health of my wife and our newborn baby son, now fifty years old. Then he delivered his verdict in a quiet, even voice. "There are two ways of doing calculations in theoretical physics", he said. "One way, and this is the way I prefer, is to have a clear physical picture of the process that you are calculating. The other way is to have a precise and self-consistent mathematical formalism. You have neither." I was slightly stunned, but ventured to ask him why he did not consider the pseudoscalar meson theory to be a self-consistent mathematical formalism. He replied, "Quantum electrodynamics is a good theory because the forces are weak, and when the formalism is ambiguous we have a clear physical picture to guide us. With the pseudoscalar meson theory there is no physical picture, and the forces are so strong that nothing converges. To reach your calculated results, you had to introduce arbitrary cut-off procedures that are not based either on solid physics or on solid mathematics."

In desperation I asked Fermi whether he was not impressed by the agreement between our calculated numbers and his measured numbers. He replied, "How many arbitrary parameters did you use for your calculations?" I thought for a moment about our cut-off procedures and said, "Four." He said, "I remember my friend Johnny von Neumann used to say, with four parameters I can fit an elephant, and with five I can make him wiggle his trunk." With that, the conversation was over. I thanked Fermi for his time and trouble, and sadly took the next bus back to Ithaca to tell the bad news to the students. Because it was important for the students to have their names on a published paper, we did not abandon our calculations immediately. We finished them and wrote a long paper that was duly published in the *Physical Review* with all our names on it. Then we dispersed to find other lines of work. I escaped to Berkeley, California, to start a new career in condensed-matter physics.

Looking back after fifty years, we can clearly see that Fermi was right. The crucial discovery that made sense of the strong forces was the quark. Mesons and protons are little bags of quarks. Before Murray Gell-Mann discovered quarks, no theory of the strong forces could possibly have been adequate. Fermi knew nothing about quarks, and died before they were discovered. But somehow he knew that something essential was missing in the meson theories of the 1950s. His physical intuition told him that the pseudoscalar meson theory could not be right. And so it was Fermi's intuition, and not any discrepancy between theory and experiment, that saved me and my students from getting stuck in a blind alley. ■

Freeman Dyson is at the Institute for Advanced Study, Einstein Drive, Princeton, New Jersey 08540, USA.

SERGIO TALAMO, COURTESY EMILIO SEGRÈ VISUAL ARCHIVES

HERA, DAVIS PHOTOGRAPHERS, COURTESY EMILIO SEGRÈ VISUAL ARCHIVES

©2004 Nature Publishing Group

over-fitting (1/3)

Fermi

“How many arbitrary parameters did you use for your calculations?”

Dyson

“I thought for a moment about our cut-off procedures and said, -Four.-"

Fermi

“I remember my friend Johnny von Neumann used to say, with four parameters I can fit an elephant, and with five I can make him wiggle his trunk.”

over-fitting (2/3)

+

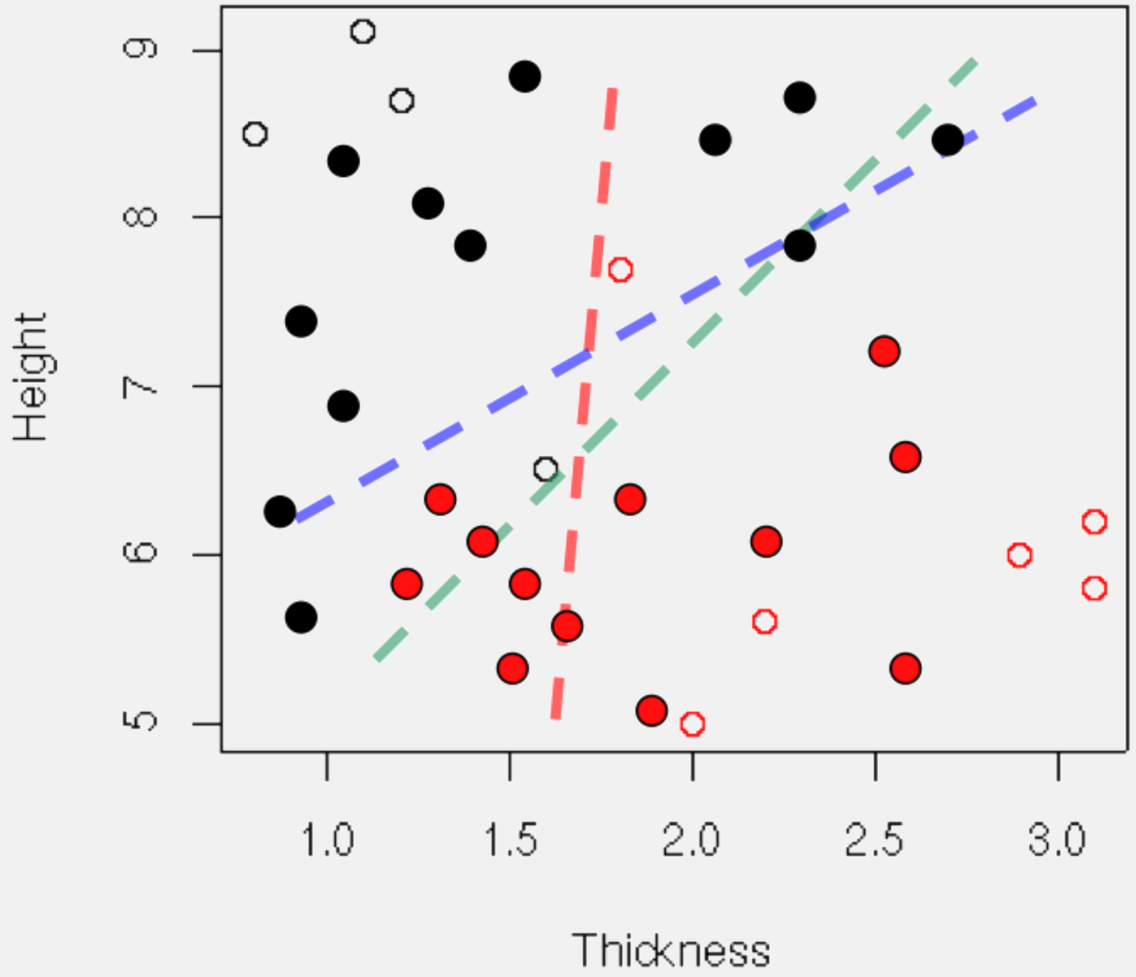


over-fitting (3/3)



From Dyson, F. A meeting with Enrico Fermi; how one intuitive physicist rescued a team from fruitless research. Nature 2004.

Over-Fitting



Ransohoff. Nat Rev Cancer 2004

Validation without leakage

Fit preprocessing, feature selection, and tuning using only the training partition. For tuned models, use nested cross-validation or an untouched test set. Keep samples from the same person together when splitting.

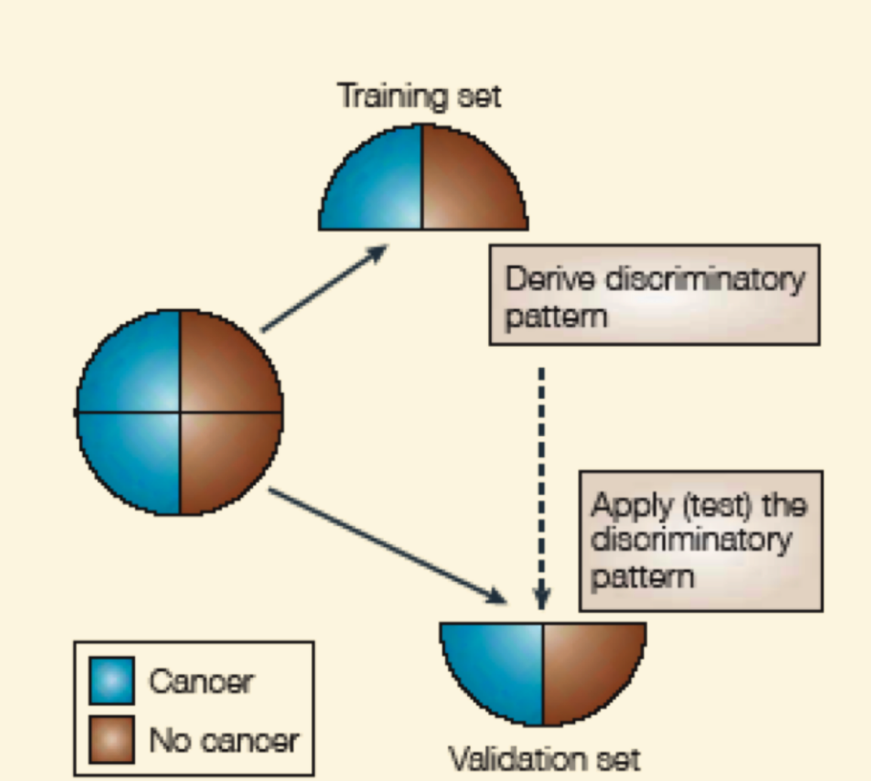


Figure 1 | **Method of dividing original sample to assess reproducibility and overfitting.**