NBE-4070 : Basics of Biomedical Data Analysis

Stéphane Deny
Prof. in Neuroscience and Biomedial Engineering and Computer Science
Aalto University

Lecture 4: Clustering Methods

Quiz 3

Question 1

Edit question | Flag question

What is a correct description of PCA? Check all answers that apply.

Select one or more:

- a. PCA is the decomposition of the covariance matrix of the data into the product of 3 matrices, two of which contain the eigenvectors and one of which contains the eigenvalues.
- b. PCA consists in finding a change of basis such that the data covariance matrix is diagonal in that new basis.
- c. PCA consists in finding a change of basis such that the data covariance matrix is null in that new basis.

Question 2

What can PCA be used for? Check all answers that apply.

Select one or more:

- a. Expanding the dimensionality of a low-dimensional dataset
- b. Denoising data
- c. Reducing dimensionality of a high-dimensional dataset

Question 3

What are some risks when interpreting PCA? Check all answers that apply.

Select one or more:

- a. PCA always yields a decomposition, even on random data, and the results of PCA can thus be overinterpreted.
- b. Some important signals might only be present in the low-variance dimensions, which are typically discarded in PCA.

Question 4

Edit question

What are eigenvalues?

- a. Eigenvalues are the vectors onto which to project the data to obtain the PC scores.
- b. Eigenvalues populate the diagonal matrix of the PCA decomposition and correspond to the variance of the data projected onto the corresponding eigenvectors.

Explain to your neighbour (use sketches)

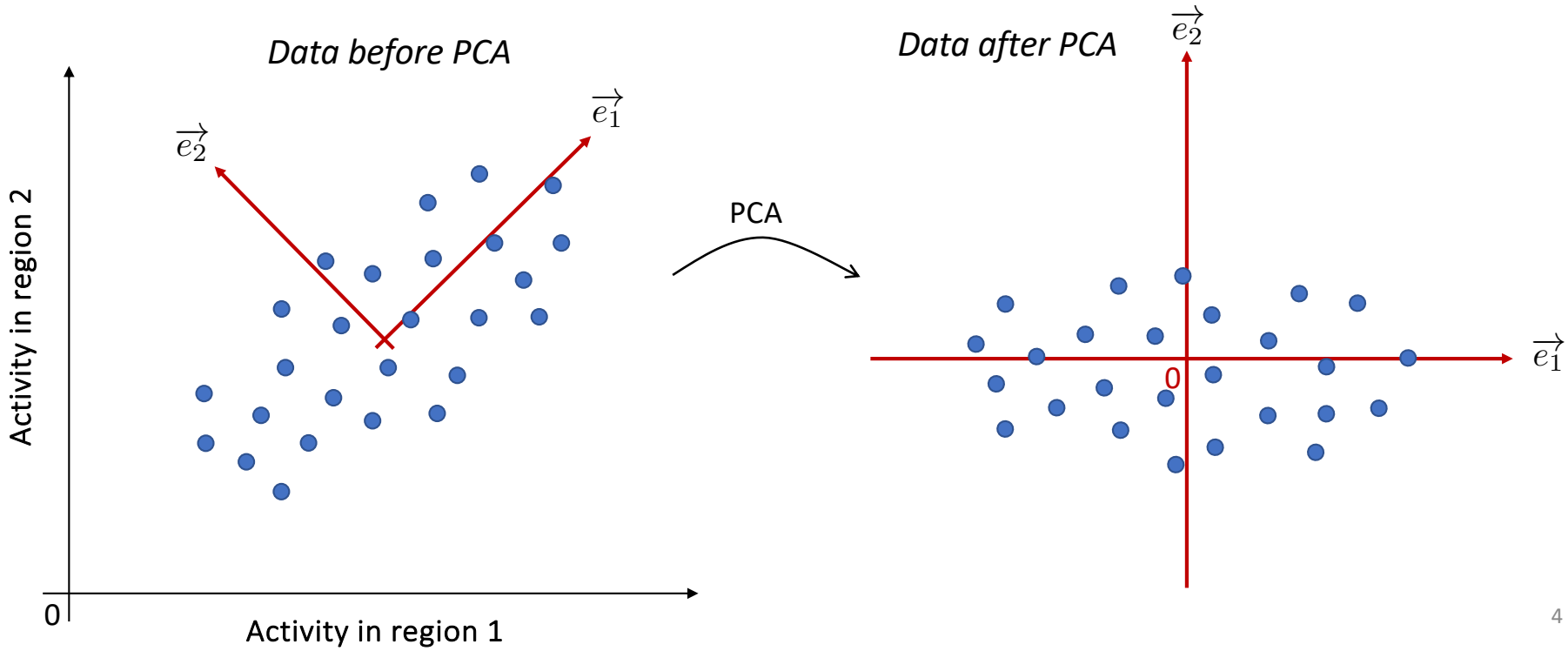
- What is PCA? What can it be used for? (3 min)

  (switch roles)

- What is the Fourier transform? What can it be used for? (3 min)

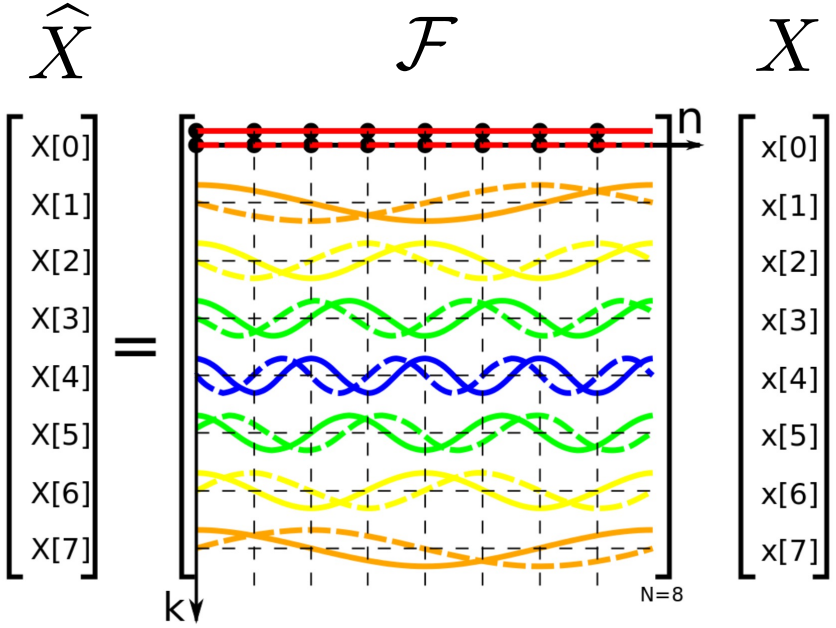
Connection between PCA and Fourier transform

PCA consists in finding a new basis such that the *covariance* between variables is null in that new basis.



Connection between PCA and Fourier transform

Fourier transform is a change of basis, which re-expresses an input sequence X into a new basis of sine and cosine functions (i.e. *frequency domain*):



The real part (cosine wave) is denoted by a solid line, and the imaginary part (sine wave) by a dashed line.

where $\hat{X}$ is the Fourier representation

$\mathcal{F}$ is the Fourier transform

X is the input sequence

Connection between PCA and Fourier transform

A stationary process is a process whose joint probability distribution does not change when shifted in time:

$$\text{let } \{x_t\}_{t\in\mathbb{R}} \text{ be a stationary process}$$

then

$$p(x_{t_1+\tau}, x_{t_2+\tau}, ..., x_{t_n+\tau}) = p(x_{t_1}, x_{t_2}, ..., x_{t_n})$$

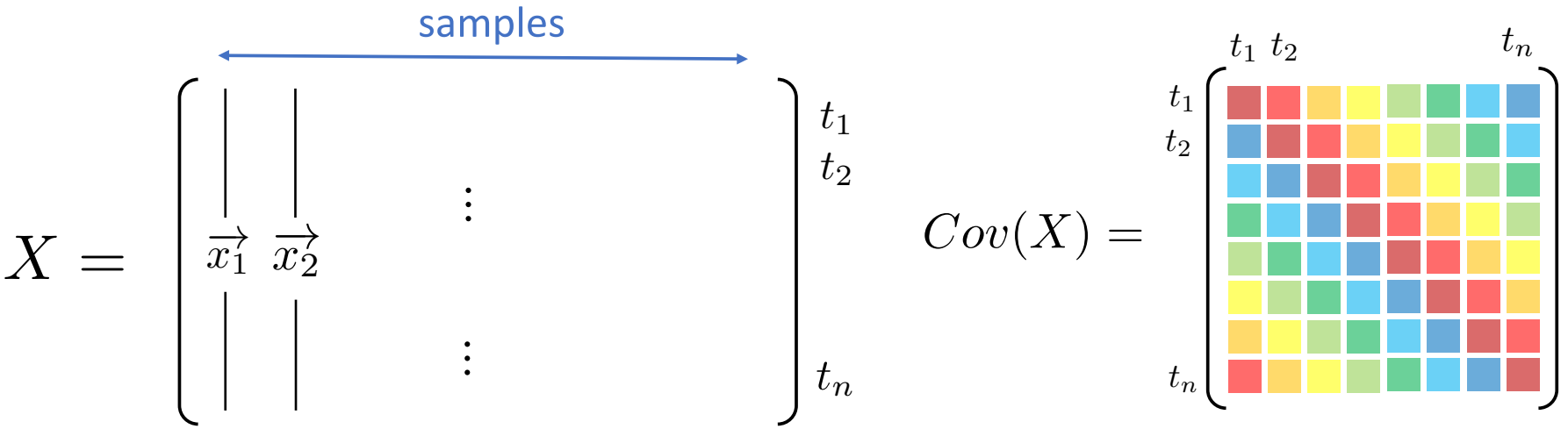
$$\text{for all } \tau \in \mathbb{R}$$

$$\text{for all } n \in \mathbb{N}$$

source: https://en.wikipedia.org/wiki/Stationary_process

Connection between PCA and Fourier transform

The covariance of a dataset of samples drawn from a (periodic) stationary process is a circulant matrix:



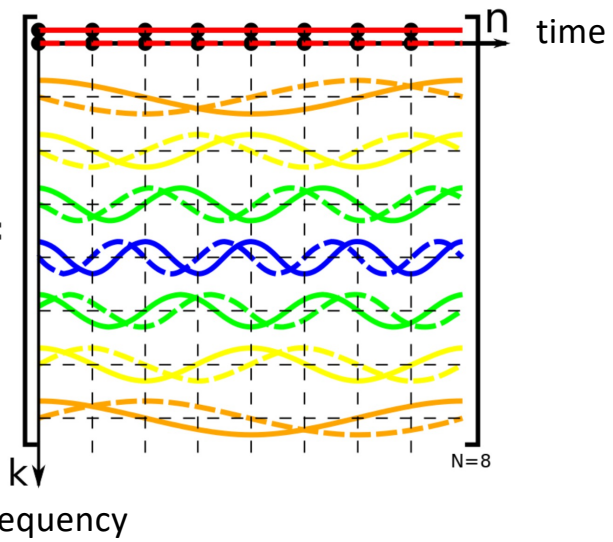
Connection between PCA and Fourier transform

When PCA decomposition is computed over a circulant covariance matrix, the resulting eigenvector basis *is* the Fourier basis:

$$Cov(X) = \mathcal{F}^{t*} D \mathcal{F}$$

where

$$\mathcal{F} =$$



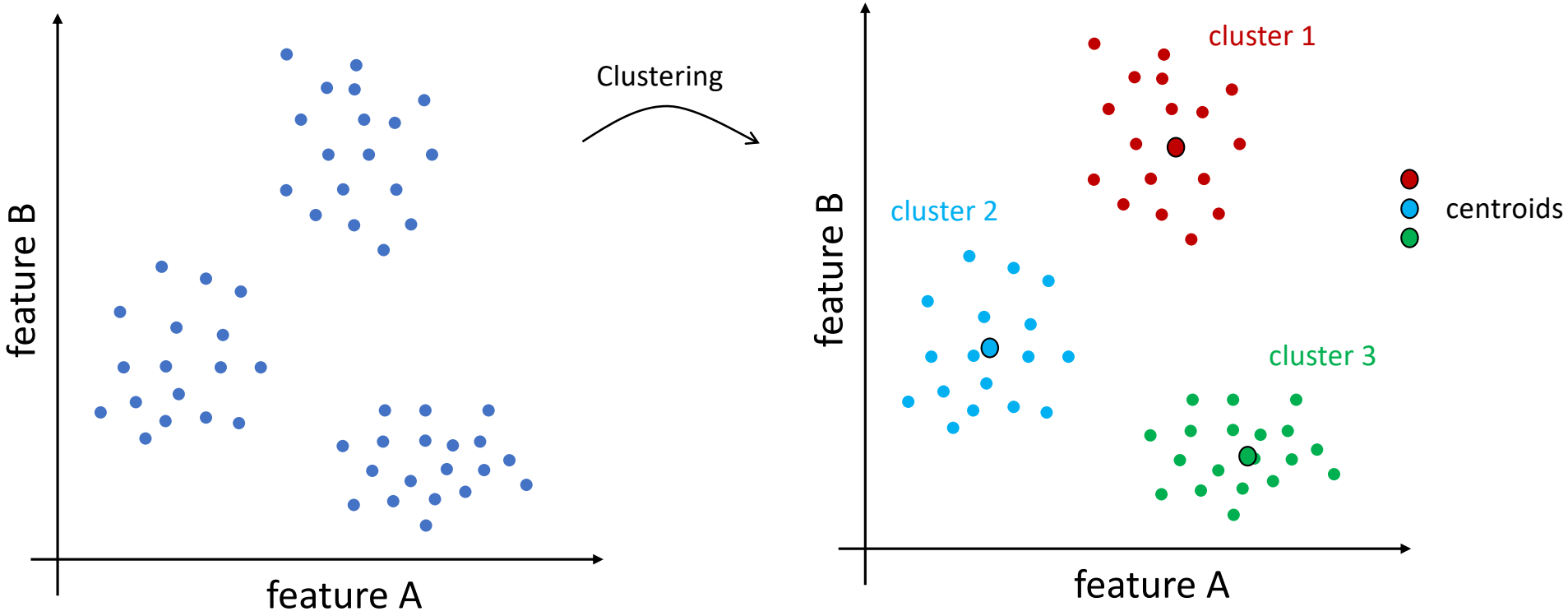
Conclusion: For any stationary process, performing Fourier transform *is equivalent* to performing PCA (except that in Fourier transform the basis is known and does need to be computed).

Outline of the course

1. Mean, Standard Deviation, Standard Error, Confidence Intervals, T-test
2. Fourier Transform, Wavelet Transforms, Spectrograms, High-pass, Low-pass filters
3. Principal Component Analysis (PCA) and Singular Value Decomposition (SVD)
4. Clustering Methods
5. Linear Regression / Logistic Regression
6. Non-linear Methods: Independant Component Analysis, t-Stochastic Neighbour Embedding, Random Forests, Deep Networks
7. Invited lectures from the biomedical industry

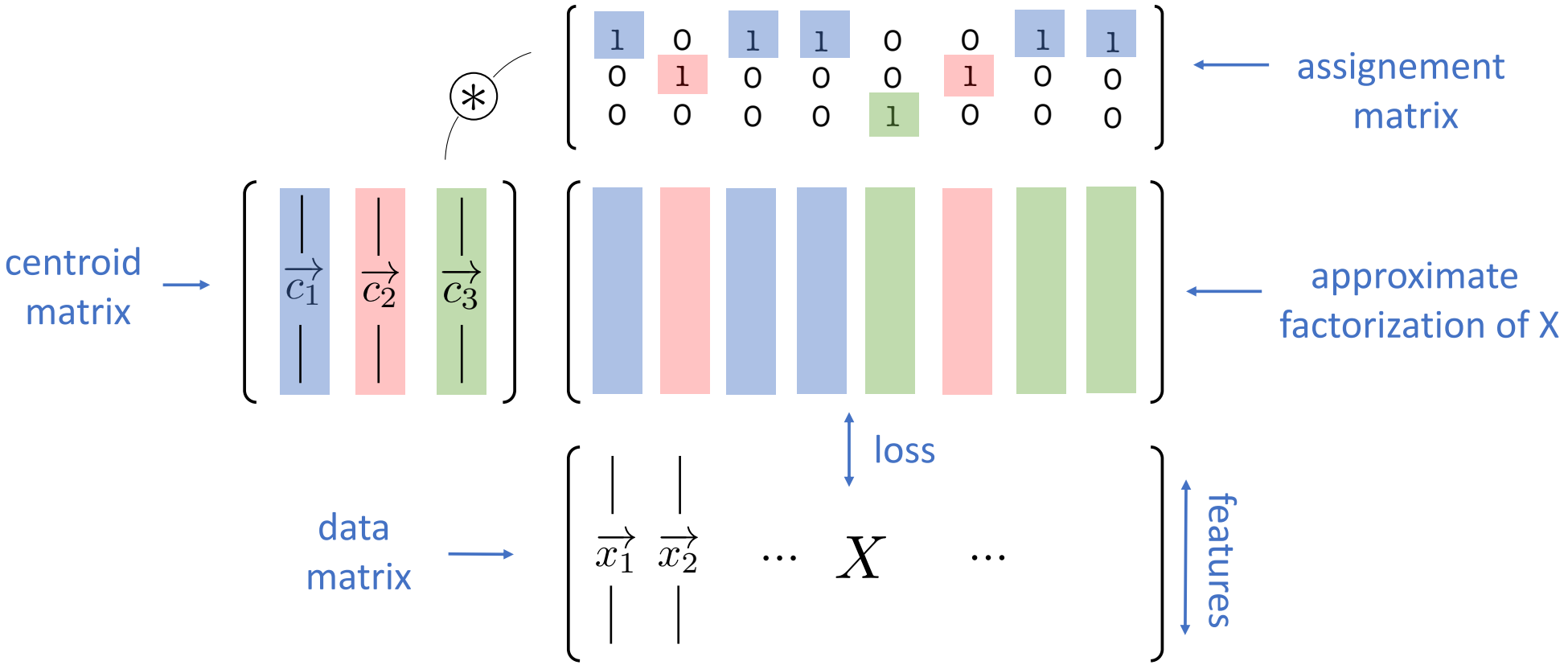
Definition of clustering

Clustering is grouping a collection of data points into subsets or “clusters”, such that the points within each cluster are closer to one another than points assigned to different clusters.



Algebraic view of clustering

Clustering consists in approximating a data matrix by the product of a *centroid matrix* and an *assignment matrix*, which is a matrix with only one non-zero element per column:



Clustering algorithms

- The number of possible partitions (i.e. grouping of points in clusters) grows very quickly with the number of points. For example, for 4 clusters of 19 points, there are more than 10^{10} potential partitions.

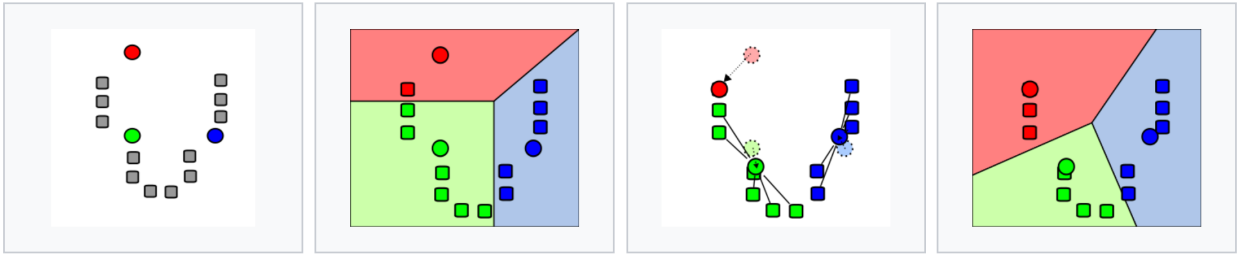
- Clustering algorithms are able to examine only a very small fraction of all possible partitions. The goal of clustering algorithms is to identify a small subset that is likely to contain the optimal one, or at least a good suboptimal partition.

- Such feasible strategies are based on iterative greedy descent: (1) an initial partition is specified; (2) at each iterative step, the cluster assignments are changed in such a way that the value of the criterion is improved from its previous value.

- There are many clustering algorithms, each with their advantages. In this lecture we will see two commonly used algorithms: *K-means* and *hierarchical clustering*.

K-means clustering algorithm

K-means clustering starts with guesses for the 'K' cluster centers. Then it alternates the following steps until convergence:
1) for each data point, the closest cluster center (in Euclidean distance) is identified;
2) each cluster center is replaced by the coordinate-wise average of all data points that are closest to it (i.e., center of mass).



1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

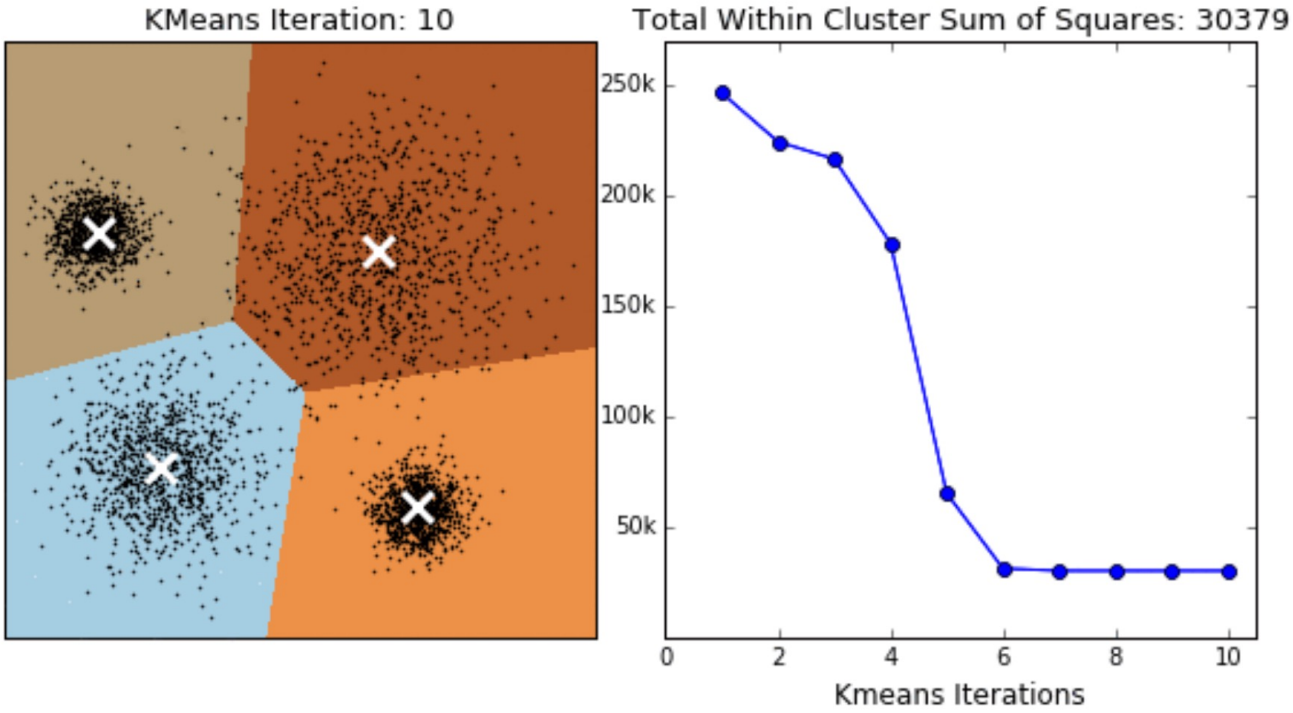
2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

source: https://en.wikipedia.org/wiki/K-means_clustering

K-means clustering in action



source: https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/

How to choose the number of clusters ‘K’ in K-means?

- Sometimes the number of centroids ‘K’ is imposed by the constraints of your problem: e.g., you need to divide a dataset in ‘K’ streams for parallel processing.

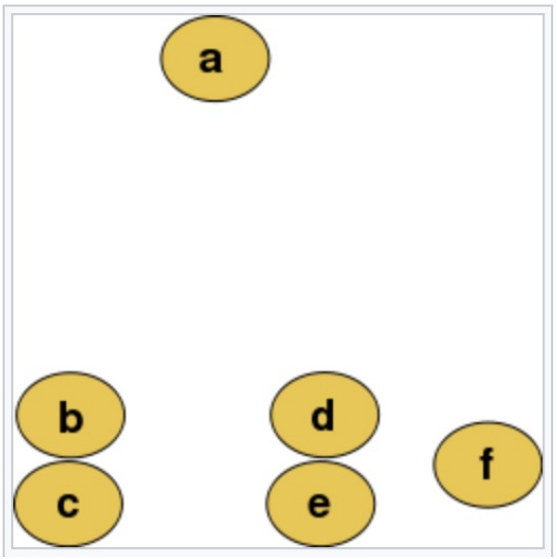
- Sometimes there is a natural number of clusters in your data. You can find it by inspecting the clusters visually and deciding whether you are subdividing your data into too many or too few clusters.

- Sometimes natural clusters appear, sometimes not. In all cases, a clustering analysis can be useful.

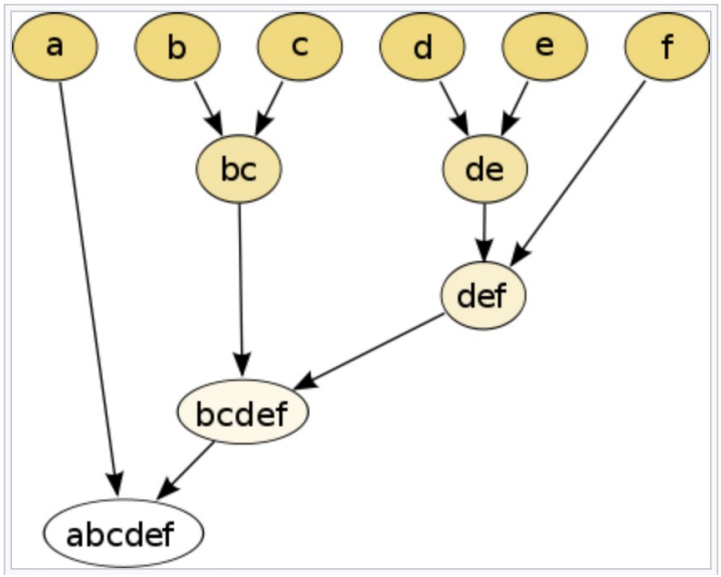
Hierarchical clustering algorithm

Hierarchical clustering recursively merges a selected pair of clusters into a single cluster. This produces a grouping at the next higher level with one less cluster. The pair chosen for merging consists of the two groups with the smallest intergroup dissimilarity.

Raw data



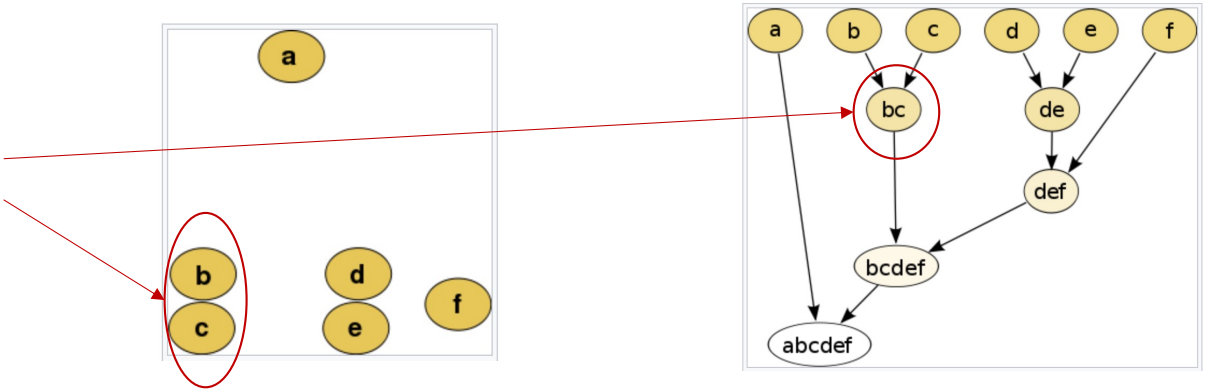
Hierarchical clusters



source: https://en.wikipedia.org/wiki/Hierarchical_clustering

Hierarchical clustering: linkage

Where is group ‘bc’ located?



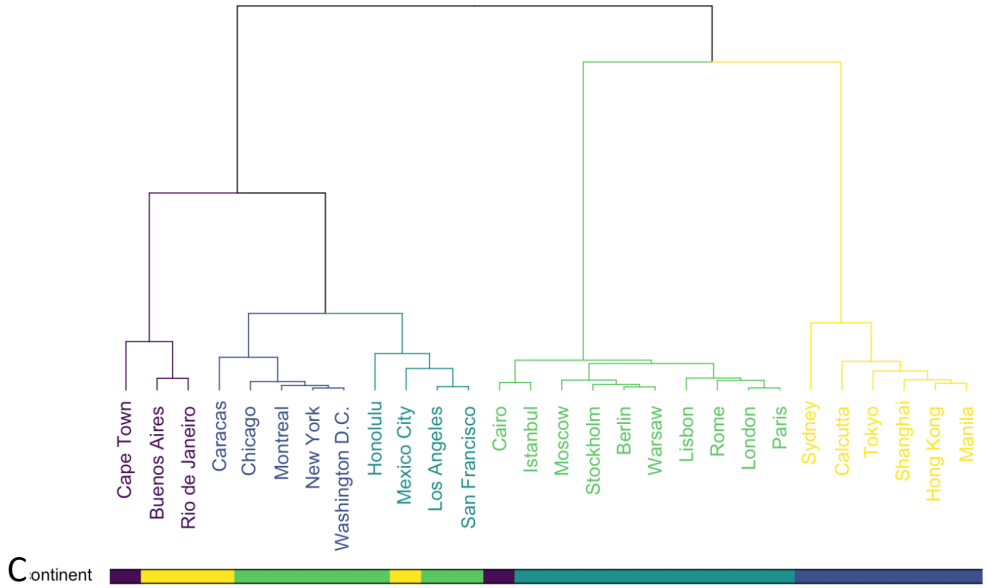
- *Single linkage* agglomerative clustering takes the intergroup dissimilarity to be that of the closest (least dissimilar) pair.

- *Complete linkage* agglomerative clustering (furthest-neighbor technique) takes the intergroup dissimilarity to be that of the furthest (most dissimilar) pair

- *Group average linkage* clustering uses the average dissimilarity between the groups

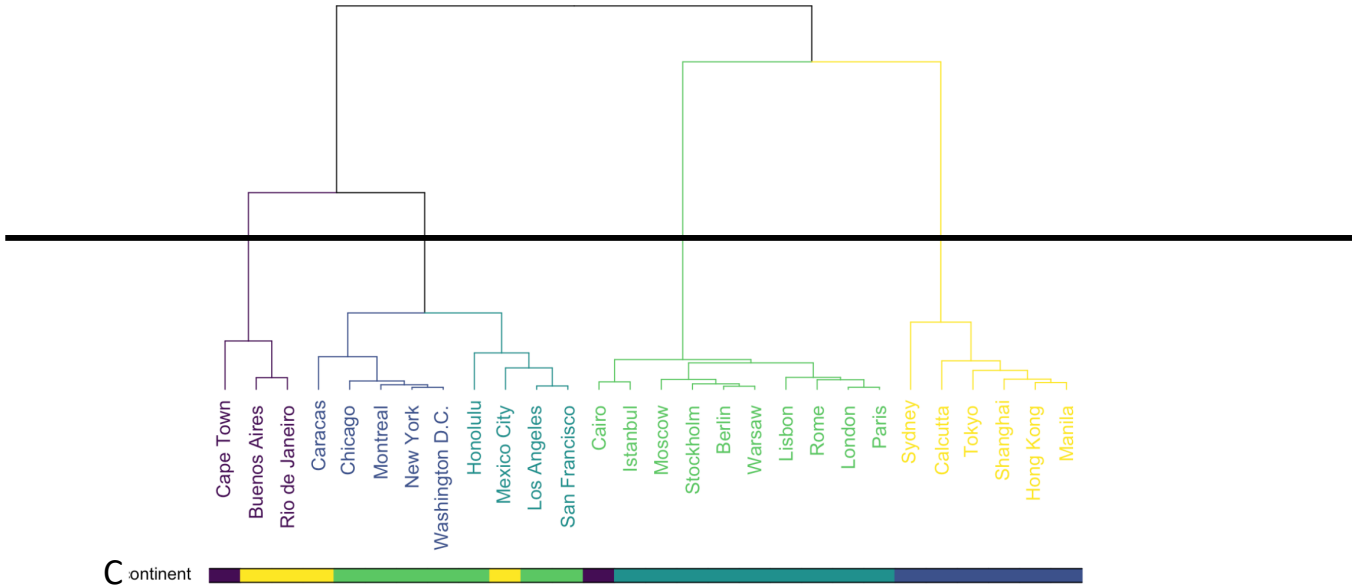
Dendogram: definition

A dendogram is a binary tree representing the hierarchical clusters, such that the height of each node is proportional to the value of the intergroup dissimilarity between its two daughters.

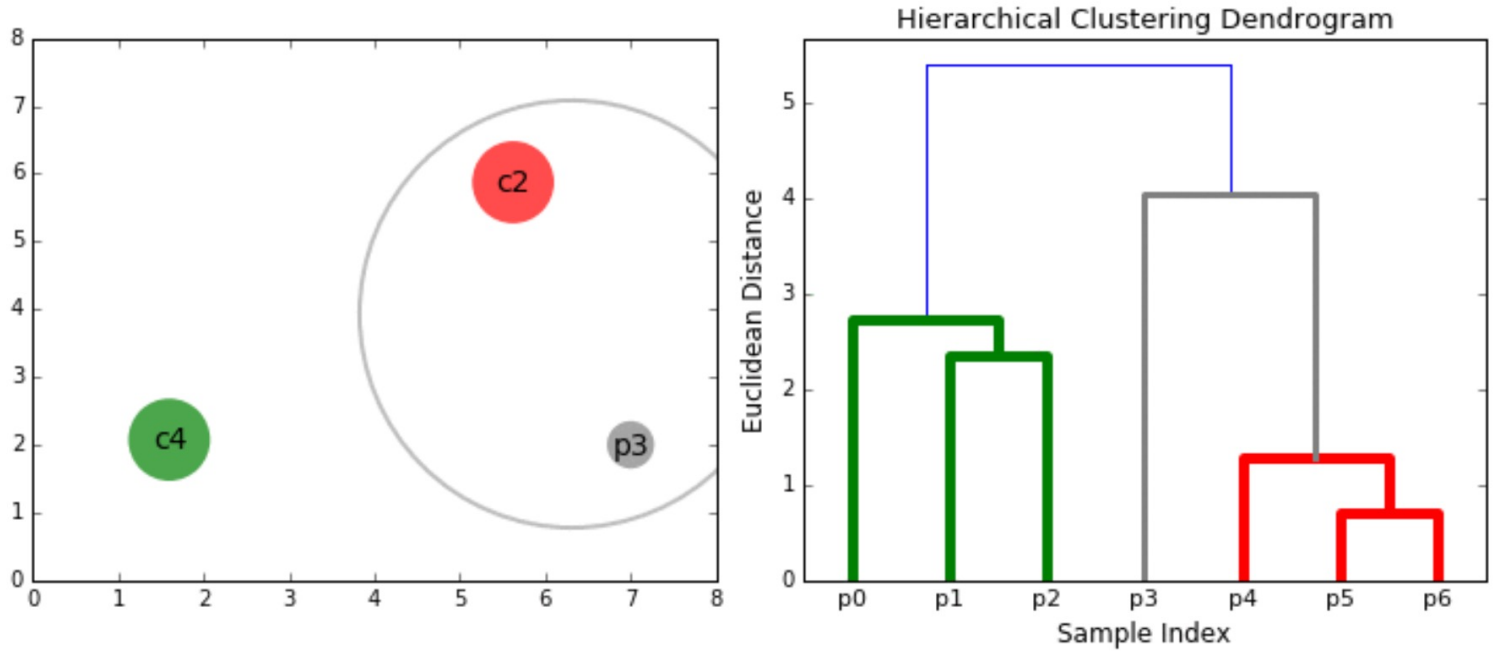


How to ‘cut’ a dendogram?

- Cutting the dendrogram horizontally at a particular height partitions the data into disjoint clusters represented by the vertical lines that intersect it. These are the clusters that would be produced by terminating the procedure when the optimal intergroup dissimilarity exceeds that threshold cut value. Groups that merge at high values, relative to the merger values of the subgroups contained within them lower in the tree, are candidates for natural clusters. Note that this may occur at several different levels, indicating a clustering hierarchy: that is, clusters nested within clusters.



Hierarchical clustering in action



source: https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/

Hierarchical Clustering vs. K-means

- The results of applying the K-means clustering algorithm depends on the choice for the number of clusters to be searched and a starting configuration assignment. In contrast, hierarchical clustering methods do not require such specifications.

- The goal is sometimes to arrange the clusters into a natural hierarchy. Hierarchical clustering involves successively grouping the clusters themselves so that at each level of the hierarchy, clusters within the same group are more similar to each other than those in different groups. It is up to the user to decide which level (if any) actually represents a “natural” clustering in the sense that observations within each of its groups are sufficiently more similar to each other than to observations assigned to different groups at that level.

Steps to perform clustering on a computer

1. Prepare your data into a matrix of size ‘samples’ x ‘features’.

2. Feed this matrix into a clustering algorithm (e.g., K-means or hierarchical clustering)

3. Vizualise the clusters obtained (e.g., with a dendogram for hierarchical clustering, or by superposing all elements of each cluster for K-means)

4. Adjust the clustering parameters according to your specific criteria (i.e., readjust ‘K’ in K-means, or adjust the vertical cut of the dendogram in hierarchical clustering)

Case study 1: genetic analysis of tumor patients

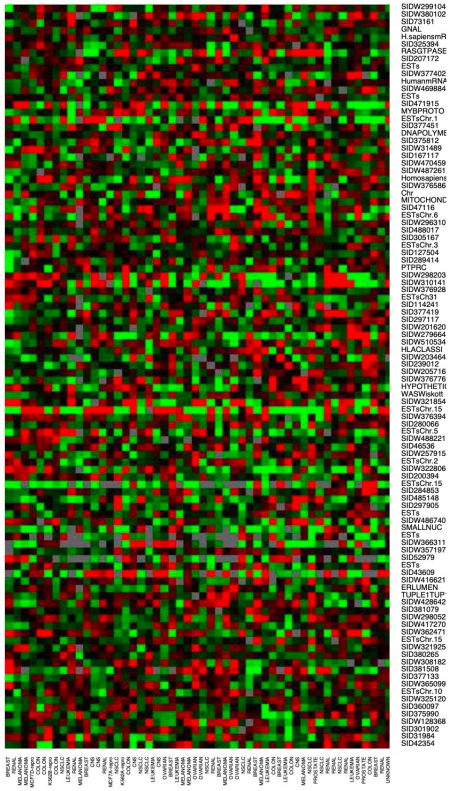


FIGURE 1.3. *DNA microarray data: expression matrix of* 6830 *genes (rows) and* 64 *samples (columns), for the human tumor data. Only a random sample of* 100 *rows are shown. The display is a heat map, ranging from bright green (negative, under expressed) to bright red (positive, over expressed). Missing values are gray. The rows and columns are displayed in a randomly chosen order.*

source: The Elements of Statistical Learning, Hastie, Tibshirani, Friedman (2017)

Example 4: DNA Expression Microarrays

DNA stands for deoxyribonucleic acid, and is the basic material that makes up human chromosomes. DNA microarrays measure the expression of a gene in a cell by measuring the amount of mRNA (messenger ribonucleic acid) present for that gene. Microarrays are considered a breakthrough technology in biology, facilitating the quantitative study of thousands of genes simultaneously from a single sample of cells.

Here is how a DNA microarray works. The nucleotide sequences for a few thousand genes are printed on a glass slide. A target sample and a reference sample are labeled with red and green dyes, and each are hybridized with the DNA on the slide. Through fluoroscopy, the log (red/green) intensities of RNA hybridizing at each site is measured. The result is a few thousand numbers, typically ranging from say -6 to 6, measuring the expression level of each gene in the target relative to the reference sample. Positive values indicate higher expression in the target versus the reference, and vice versa for negative values.

A gene expression dataset collects together the expression values from a series of DNA microarray experiments, with each column representing an experiment. There are therefore several thousand rows representing individual genes, and tens of columns representing samples: in the particular example of Figure 1.3 there are 6830 genes (rows) and 64 samples (columns), although for clarity only a random sample of 100 rows are shown. The figure displays the data set as a heat map, ranging from green (negative) to red (positive). The samples are 64 cancer tumors from different patients.

The challenge here is to understand how the genes and samples are organized. Typical questions include the following:

(a) which samples are most similar to each other, in terms of their expression profiles across genes?

(b) which genes are most similar to each other, in terms of their expression profiles across samples?

(c) do certain genes show very high (or low) expression for certain cancer samples?

We could view this task as a regression problem, with two categorical predictor variables—genes and samples—with the response variable being the level of expression. However, it is probably more useful to view it as *unsupervised learning* problem. For example, for question (a) above, we think of the samples as points in 6830–dimensional space, which we want to *cluster* together in some way.

Case study 1: genetic analysis of tumor patients

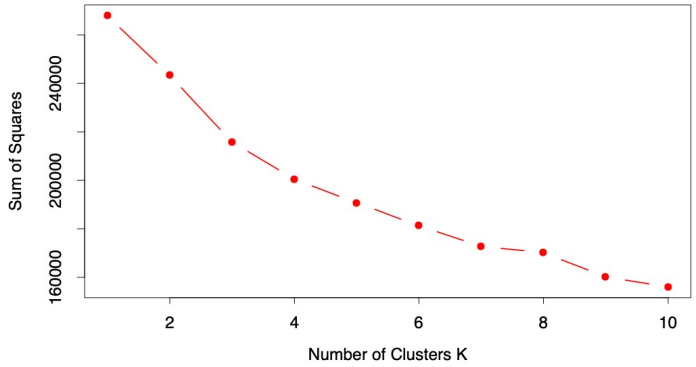


FIGURE 14.8. *Total within-cluster sum of squares for K-means clustering applied to the human tumor microarray data.*

The data are a 6830×64 matrix of real numbers, each representing an expression measurement for a gene (row) and sample (column). Here we cluster the samples, each of which is a vector of length 6830, corresponding to expression values for the 6830 genes. Each sample has a label such as `breast` (for breast cancer), `melanoma`, and so on; we don't use these labels in the clustering, but will examine *posthoc* which labels fall into which clusters.

We applied K-means clustering with K running from 1 to 10, and computed the total within-sum of squares for each clustering, shown in Figure 14.8. Typically one looks for a kink in the sum of squares curve (or its logarithm) to locate the optimal number of clusters (see Section 14.3.11). Here there is no clear indication: for illustration we chose $K = 3$ giving the three clusters shown in Table 14.2.

TABLE 14.2. *Human tumor data: number of cancer cases of each type, in each of the three clusters from K-means clustering.*

Cluster	Breast	CNS	Colon	K562	Leukemia	MCF7
1	3	5	0	0	0	0
2	2	0	0	2	6	2
3	2	0	7	0	0	0
Cluster	Melanoma	NSCLC	Ovarian	Prostate	Renal	Unknown
1	1	7	6	2	9	1
2	7	2	0	0	0	0
3	0	0	0	0	0	0

We see that the procedure is successful at grouping together samples of the same cancer. In fact, the two breast cancers in the second cluster were later found to be misdiagnosed and were melanomas that had metastasized. However, K-means clustering has shortcomings in this application. For one, it does not give a linear ordering of objects within a cluster: we have simply listed them in alphabetic order above. Secondly, as the number of clusters K is changed, the cluster memberships can change in arbitrary ways. That is, with say four clusters, the clusters need not be nested within the three clusters above. For these reasons, hierarchical clustering (described later), is probably preferable for this application.

Case study 1: genetic analysis of tumor patients

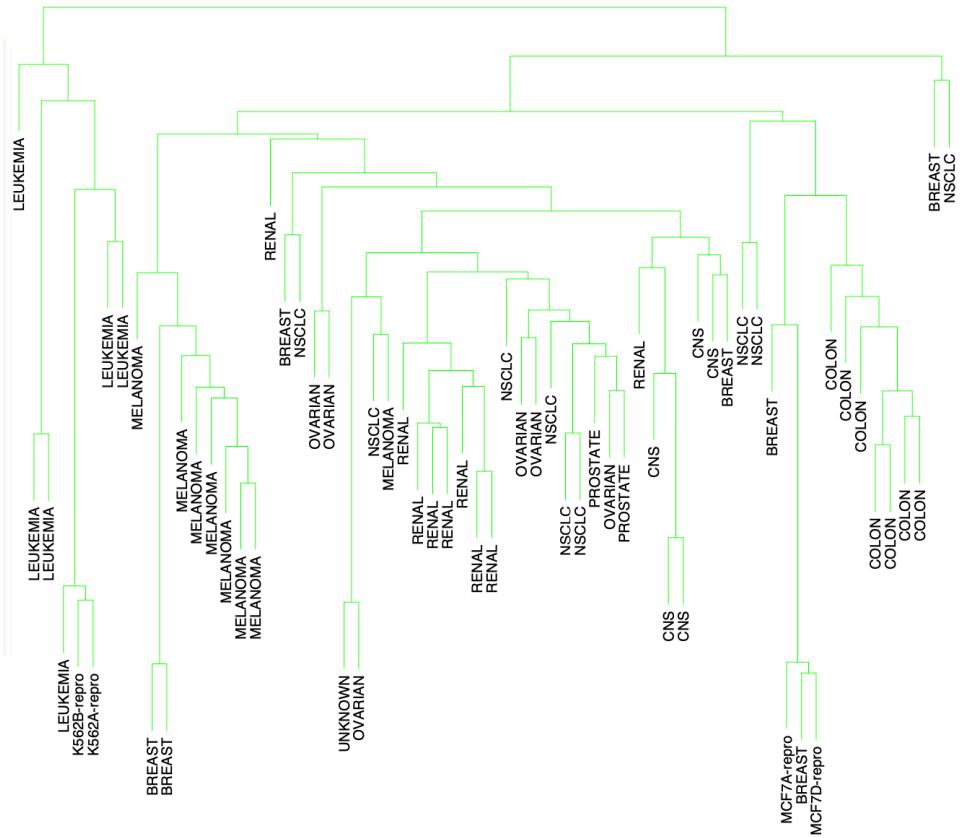


FIGURE 14.12. *Dendrogram from agglomerative hierarchical clustering with average linkage to the human tumor microarray data.*

Like K-means clustering, hierarchical clustering is successful at clustering simple cancers together. However it has other nice features. By cutting off the dendrogram at various heights, different numbers of clusters emerge, and the sets of clusters are nested within one another.

Case study 1: genetic analysis of tumor patients

Secondly, it gives some partial ordering information about the samples. In Figure 14.14, we have arranged the genes (rows) and samples (columns) of the expression matrix in orderings derived from hierarchical clustering.

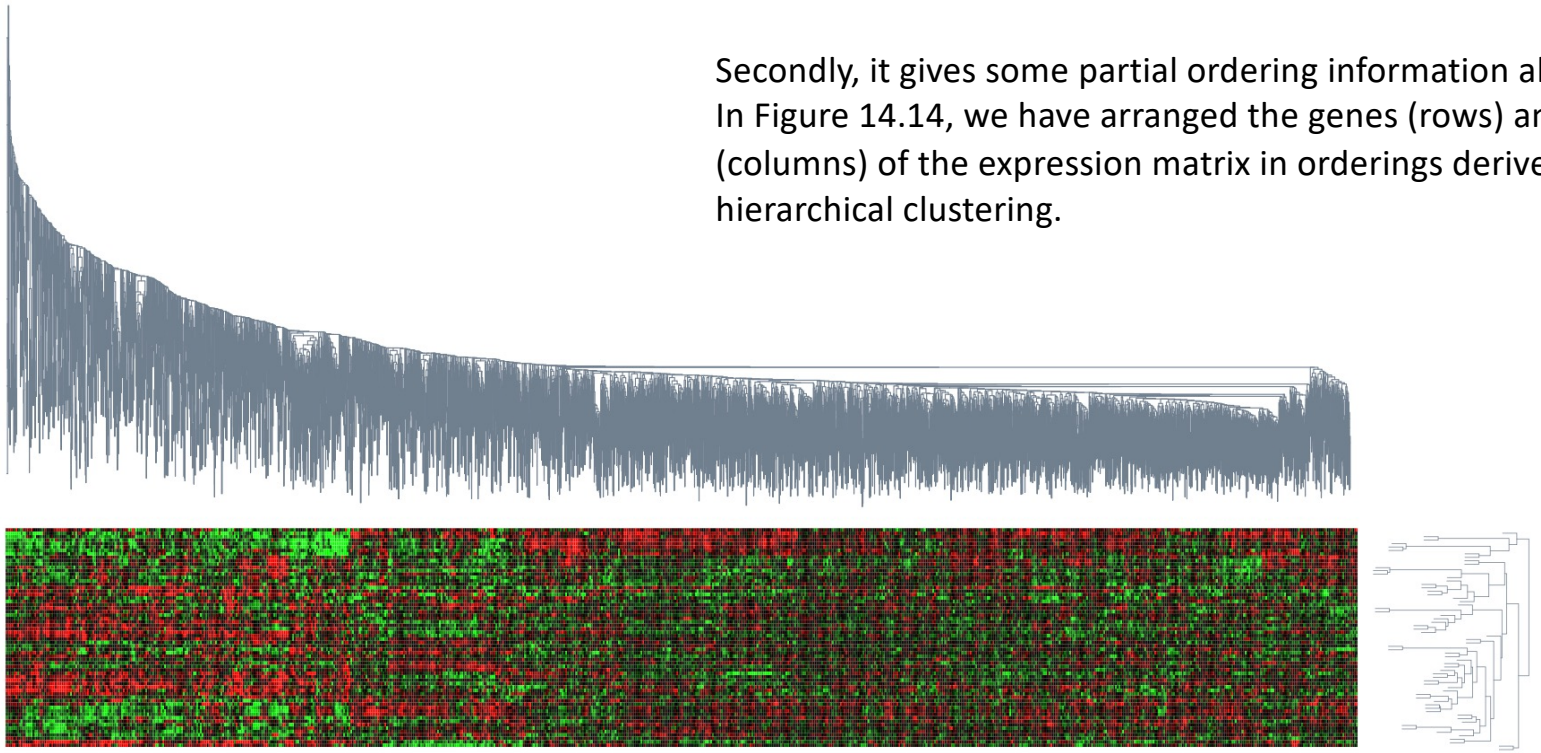
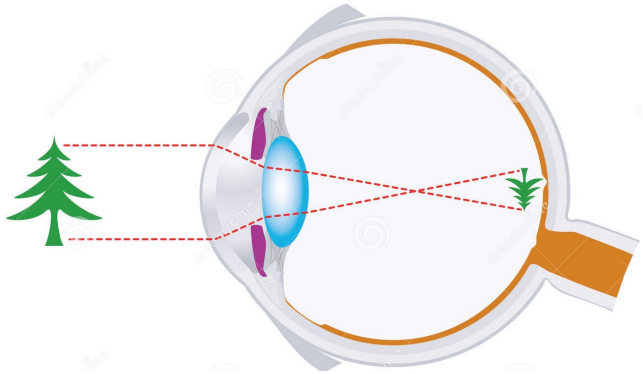
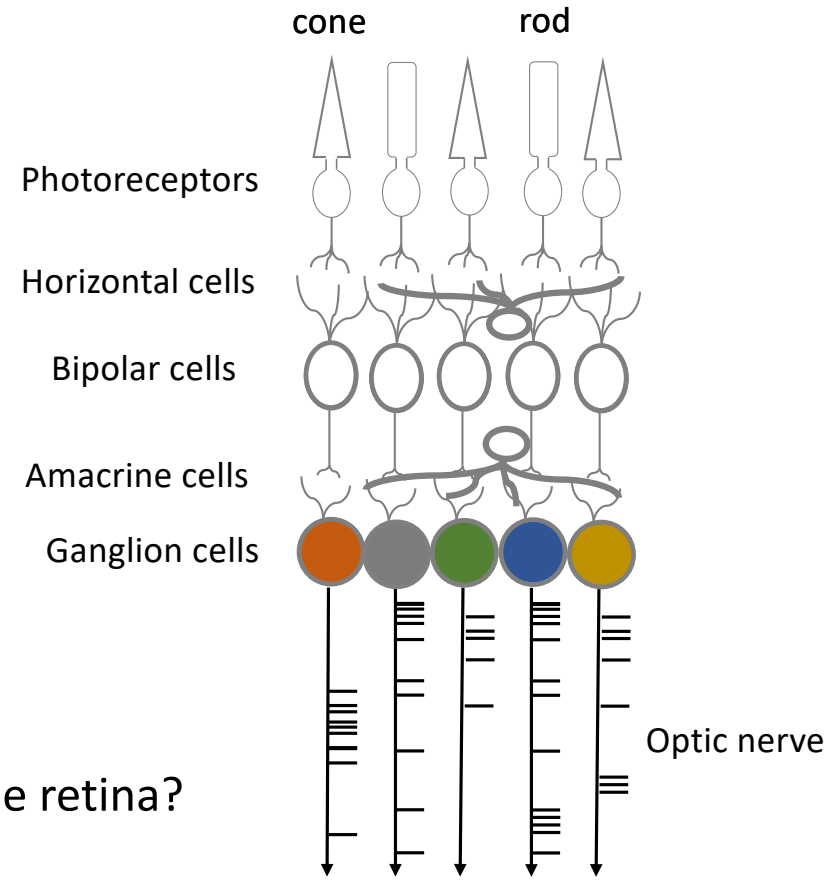


FIGURE 14.14. *DNA microarray data: average linkage hierarchical clustering has been applied independently to the rows (genes) and columns (samples), determining the ordering of the rows and columns (see text). The colors range from bright green (negative, under-expressed) to bright red (positive, over-expressed).*

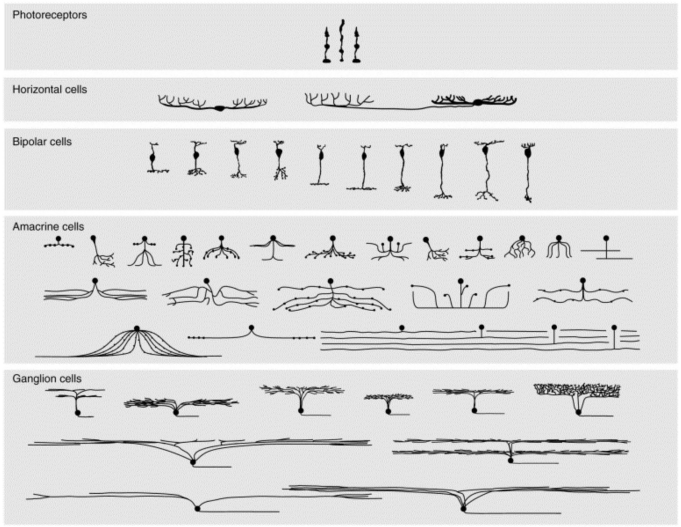
Case study 2: Understanding retinal coding



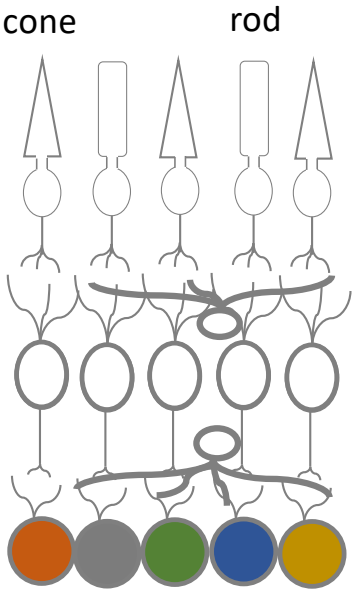


How is visual information encoded at the output of the retina?

Case study 2: Understanding retinal coding



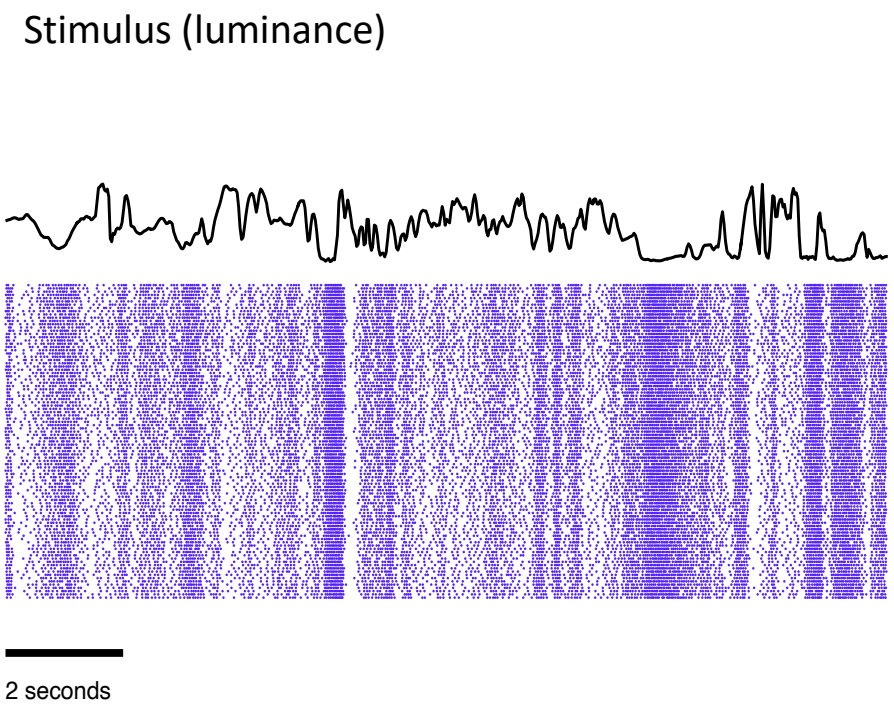
~ 100 types of cells in the retina



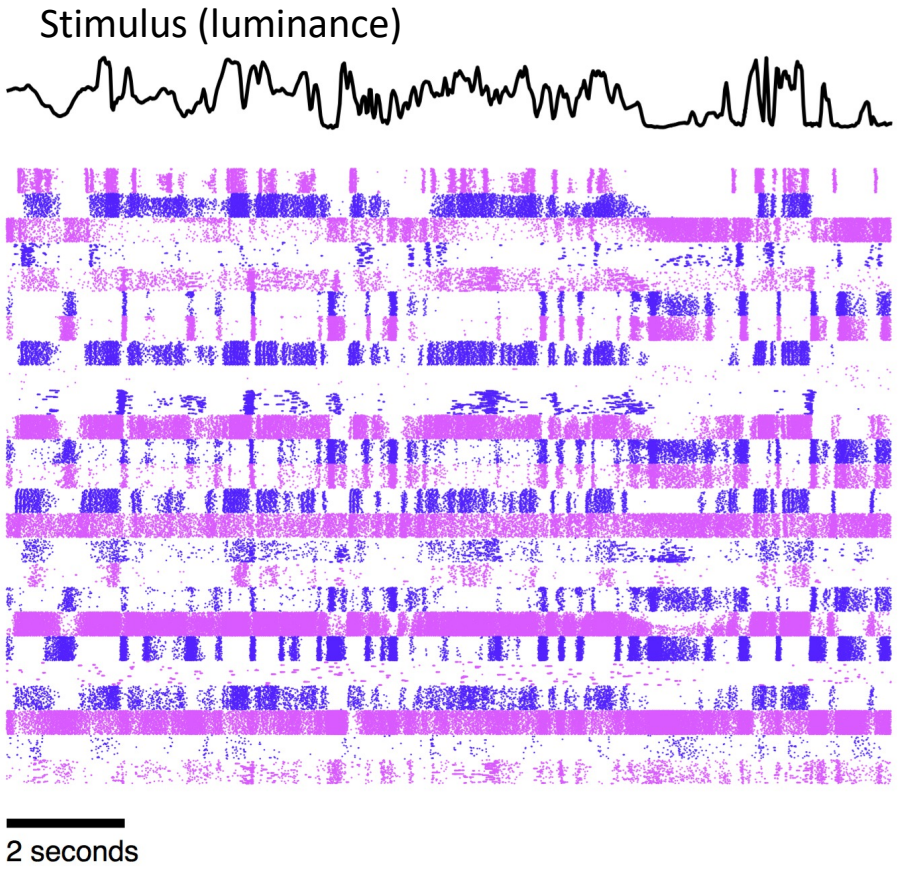
~ 30 types of ganglion cells

- Multiple stages of processing in the retina
- A non-local processing
- A complex encoding of visual information

Case study 2: Understanding retinal coding



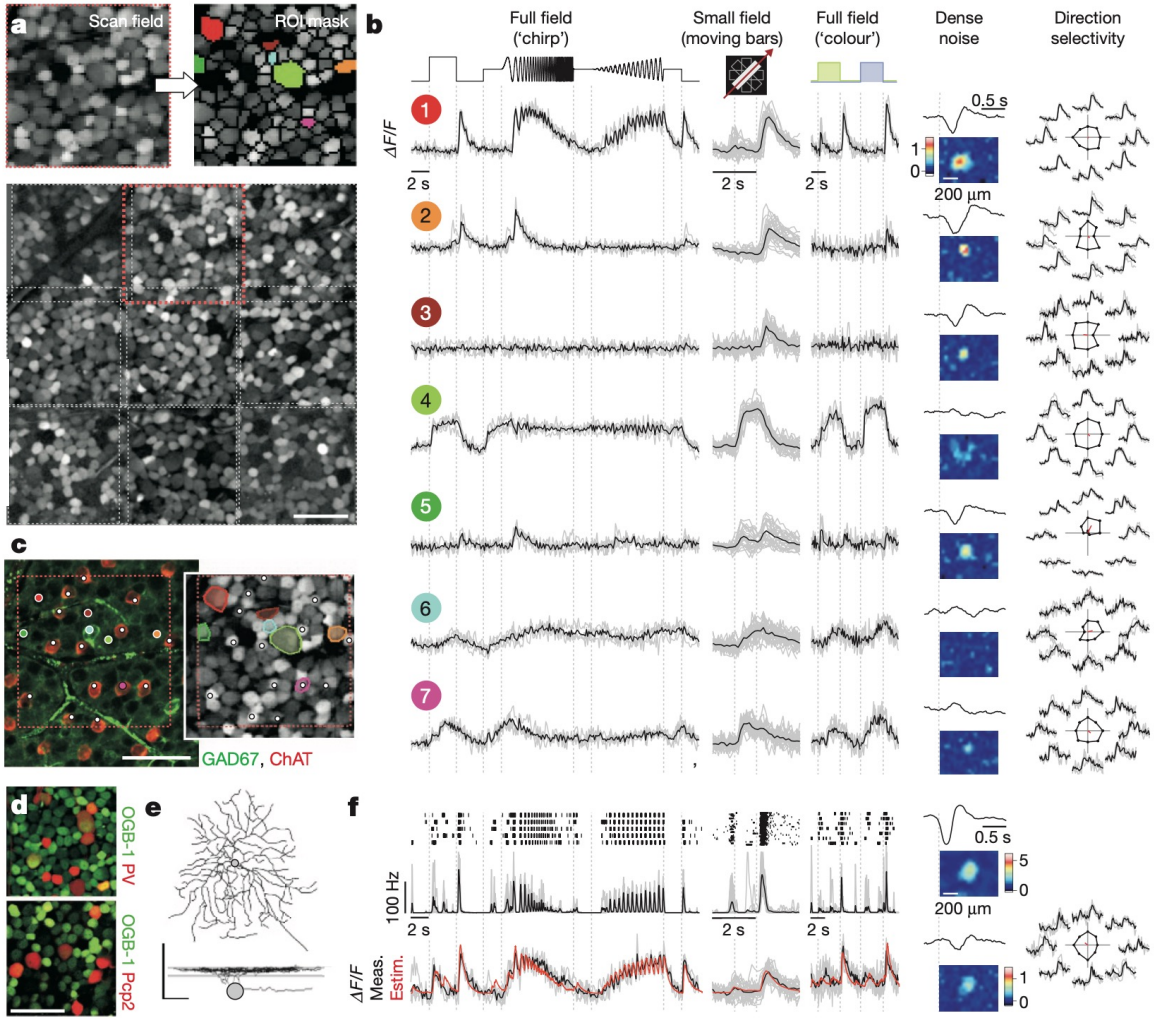
Case study 2: Understanding retinal coding



Case study 2

The functional diversity of retinal ganglion cells in the mouse

Tom Baden[1,2,3]*, Philipp Berens[1,2,3,4,5]*, Katrin Franke[1,2,3,6]*, Miroslav Román Rosón[1,2,3,6], Matthias Bethge[1,2,5,7] & Thomas Euler[1,2,3]



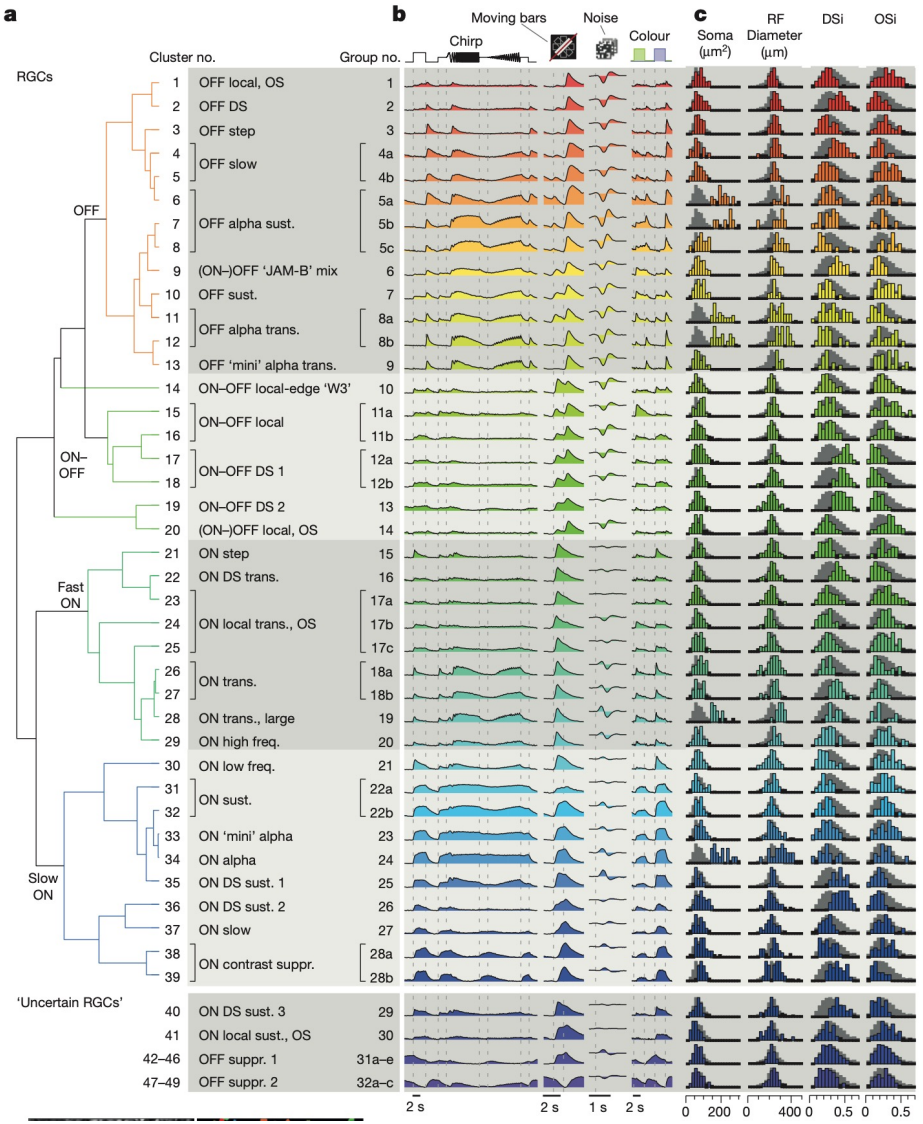
Case study 2

The functional diversity of retinal ganglion cells in the mouse

Tom Baden[1,2,3]*, Philipp Berens[1,2,3,4,5]*, Katrin Franke[1,2,3,6]*, Miroslav Román Rosón[1,2,3,6], Matthias Bethge[1,2,5,7] & Thomas Euler[1,2,3]

In the vertebrate visual system, all output of the retina is carried by retinal ganglion cells. Each type encodes distinct visual features in parallel for transmission to the brain. How many such 'output channels' exist and what each encodes are areas of intense debate. In the mouse, anatomical estimates range from 15 to 20 channels, and only a handful are functionally understood. By combining two-photon calcium imaging to obtain dense retinal recordings and unsupervised clustering of the resulting sample of more than 11,000 cells, here we show that the mouse retina harbours substantially more than 30 functional output channels. These include all known and several new ganglion cell types, as verified by genetic and anatomical criteria. Therefore, information channels from the mouse eye to the mouse brain are considerably more diverse than shown thus far by anatomical studies, suggesting an encoding strategy resembling that used in state-of-the-art artificial vision systems.

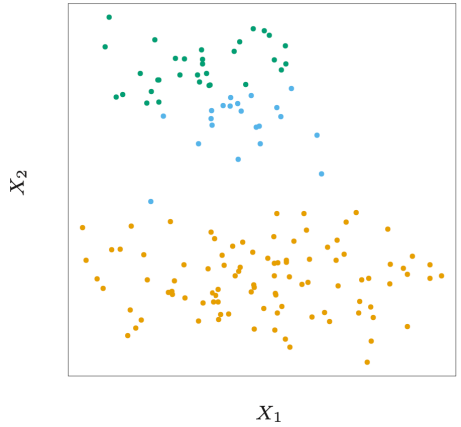
Nature 2016



Limitations and risks associated to clustering

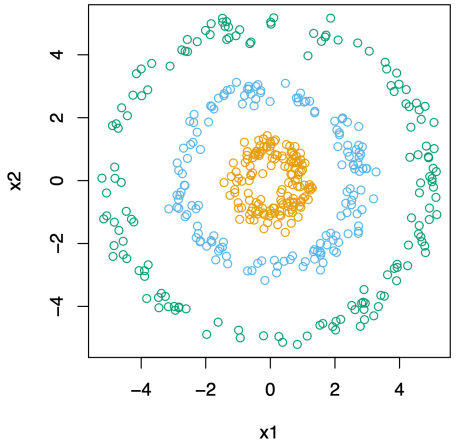
- K-means clustering will always find clusters, even if they don't exist. Similarly, hierarchical clustering imposes hierarchical structure whether or not such structure actually exists in the data.

e.g., here when K-means is applied with K = 3, it arbitrarily subdivides a group of points in two clusters



- Classical clustering methods can be inadapted to the structure of the data.

Note: advanced clustering methods exist, e.g. spectral clustering with non-euclidean kernels, t-SNE embeddings



Next lecture

- Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) (part II)