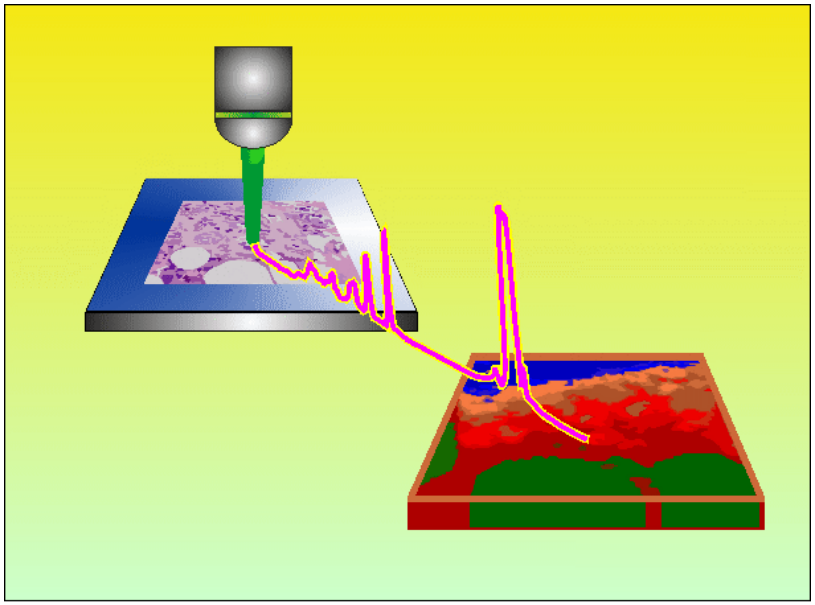
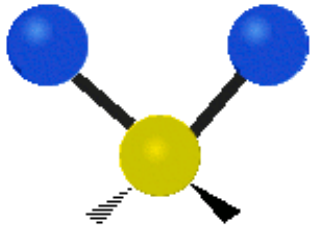
High resolution Raman imaging of human digestive tract

Chemometric analysis of Raman imaging data

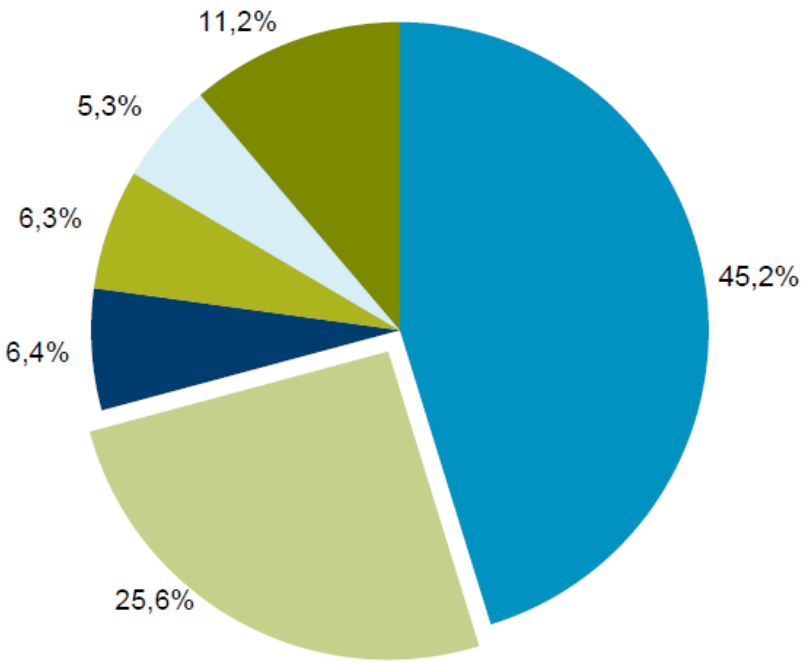


Beata Brożek-Płuska

PhD, DSc
Associate Professor
Laboratory of Laser Molecular Spectroscopy



Cancer



The magnitude of this problem concerns both the epidemiological and the diagnostic aspects. In 2023, 1,958,310 new cancer cases and 609,820 cancer deaths are projected to occur in the United States.

Causes of deaths

Cardiovascular disease

Cancers

Incorrect clinical trial results

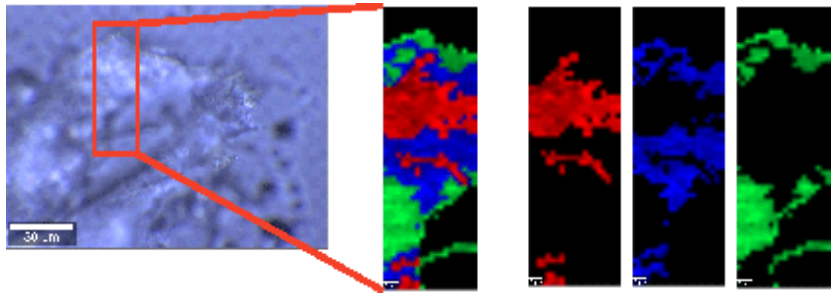
External causes of illness and death

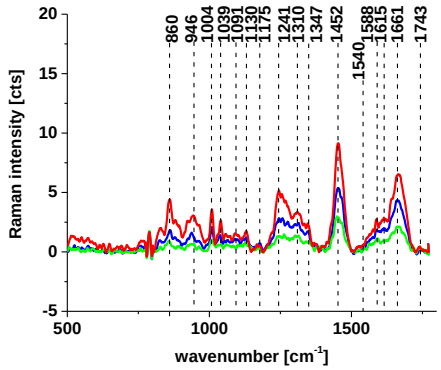
Diseases of the respiratory system (PL) / digestive system (EU)

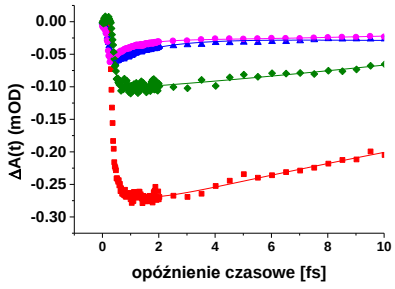
others

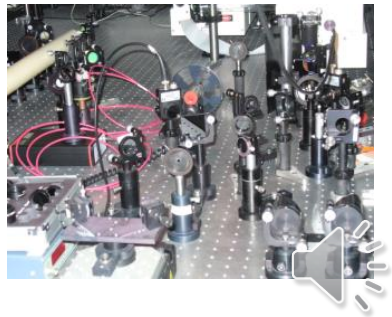
GOALS

- Demonstrate the usefulness of Raman spectroscopy, Raman imaging to identify human colon cancer.
- Identification of these classes of compounds that can be used as markers of cancer changes by Raman spectroscopy and imaging.
- Estimate the sensitivity and specificity of Raman spectroscopy, and thus to estimate the reliability of spectroscopic method by using chemometric algorithms.
- Explanation of the mechanisms of energy dissipation in the cancerous and the noncancerous tissues.

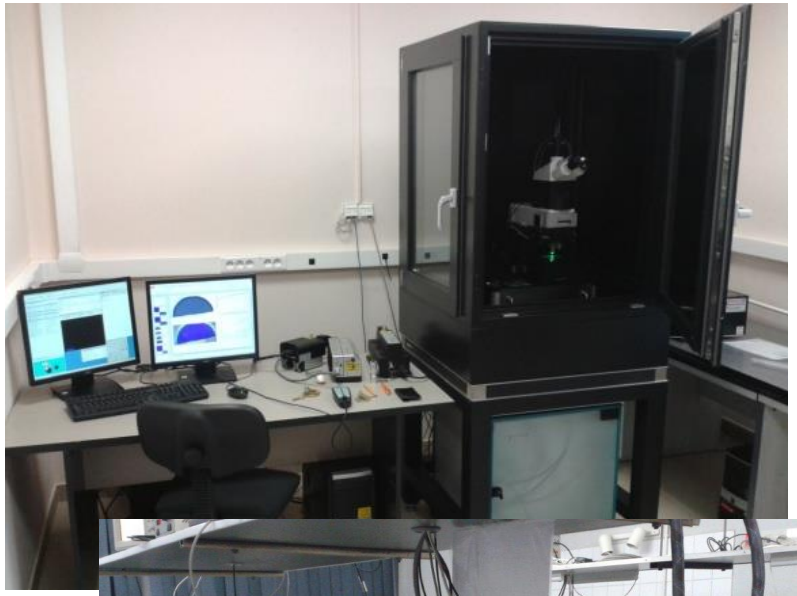


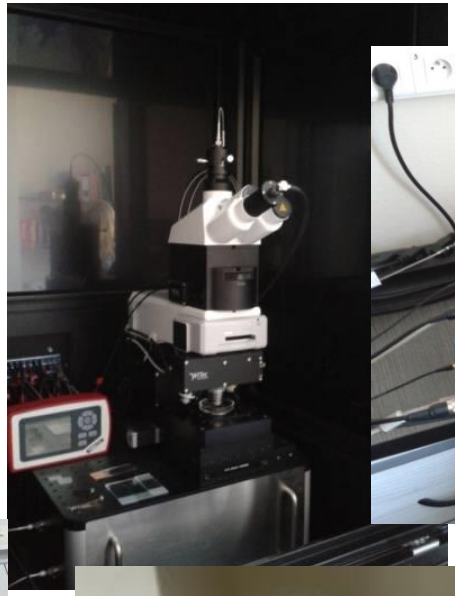


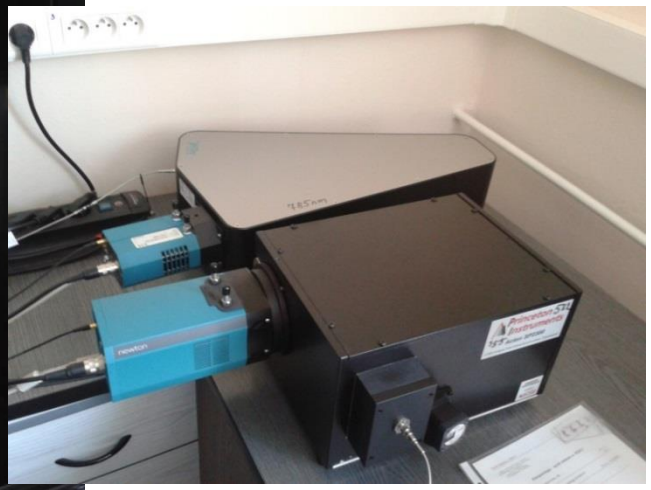


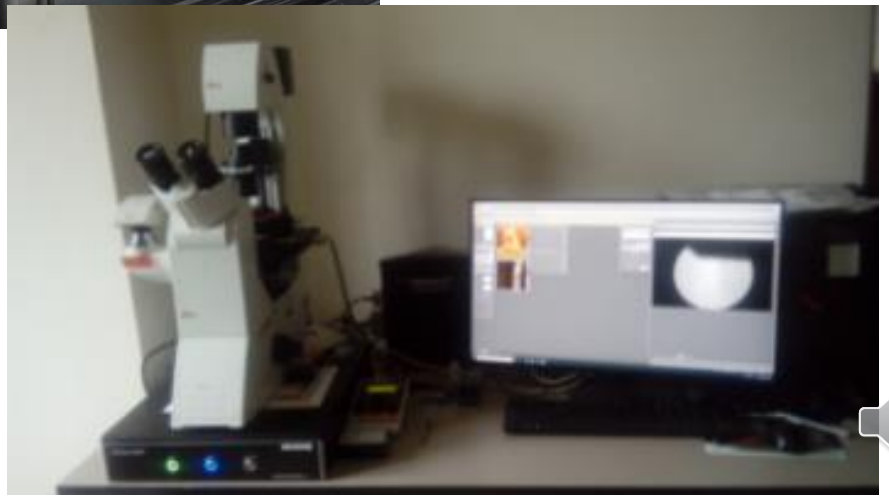


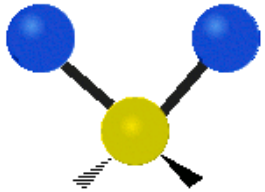
Raman spectroscopy and imaging, femtosecond spectroscopy, AFM



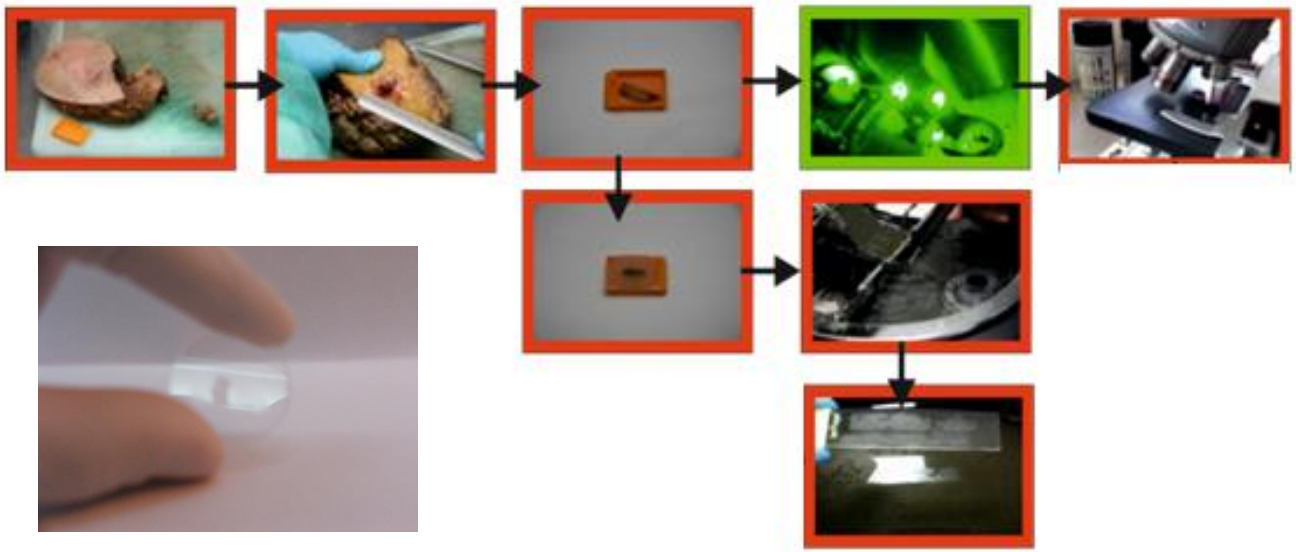


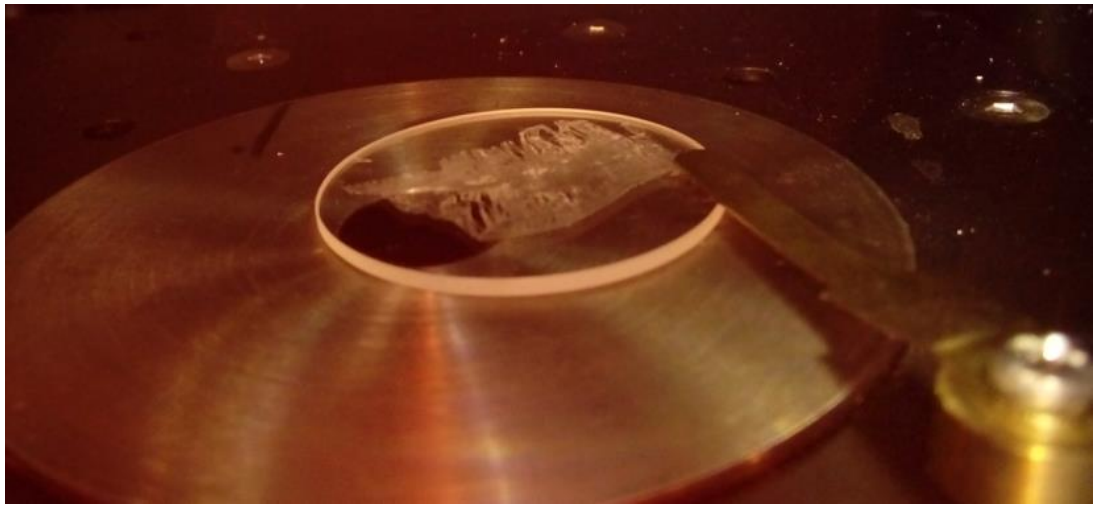






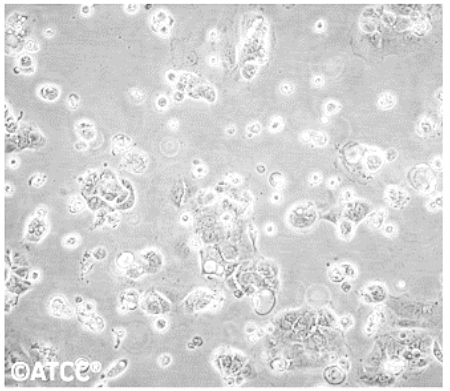
TISSUES



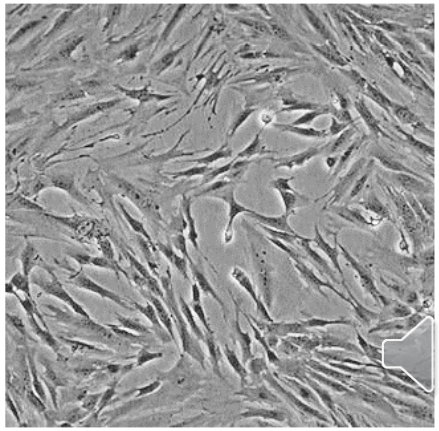


CELLS

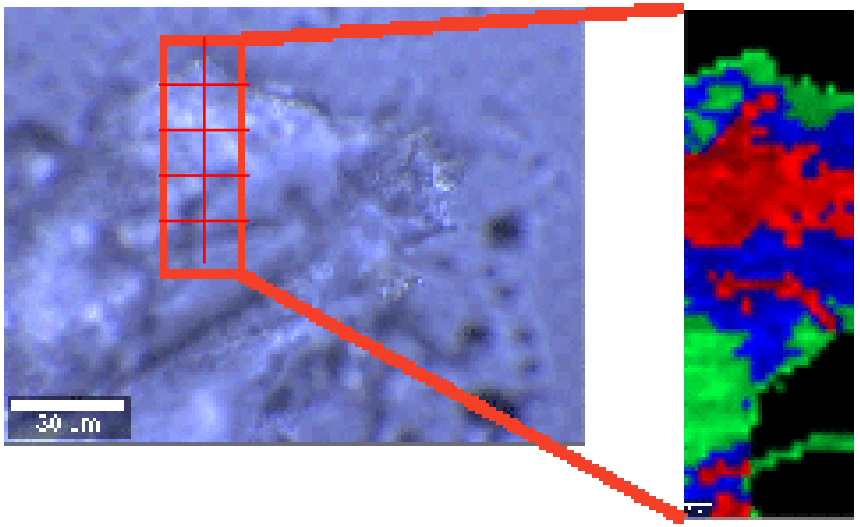
CaCo-2, cancerous

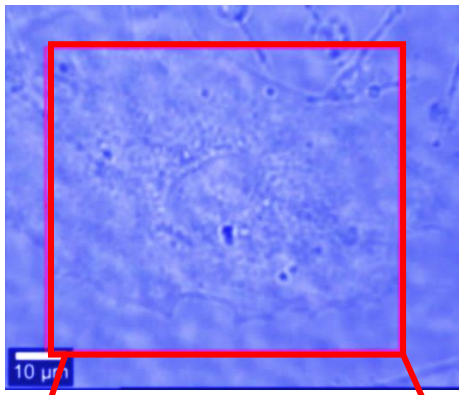


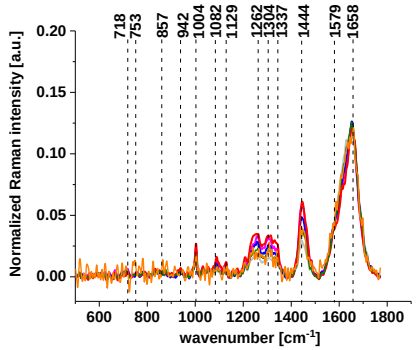
CCD-18Co, normal

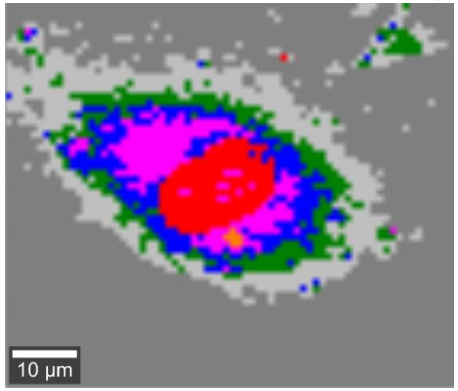


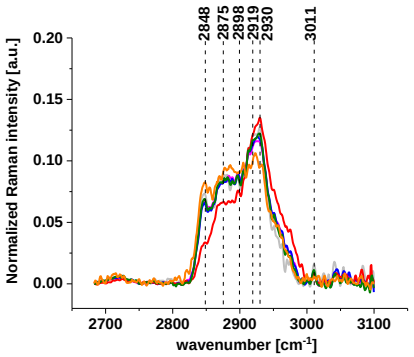
Raman spectroscopy and imaging

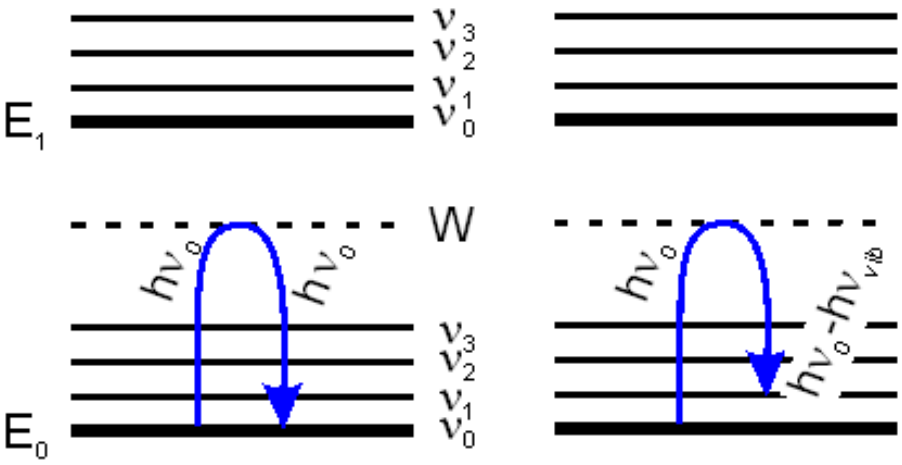


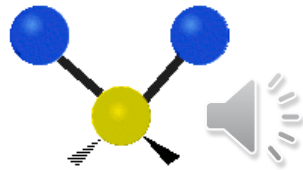




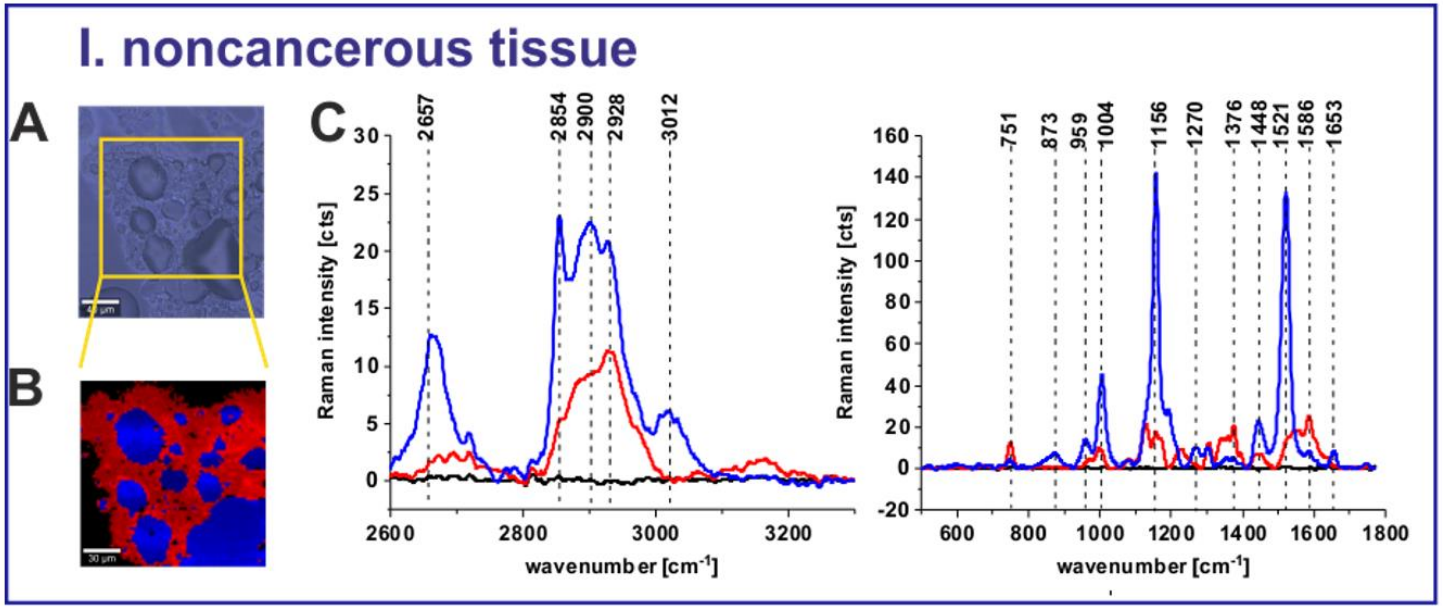








Raman spectroscopy and imaging

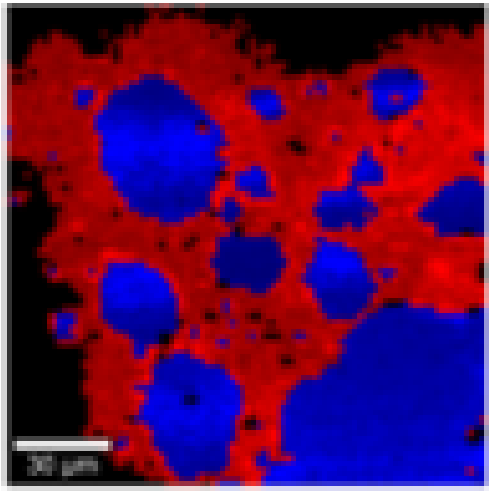


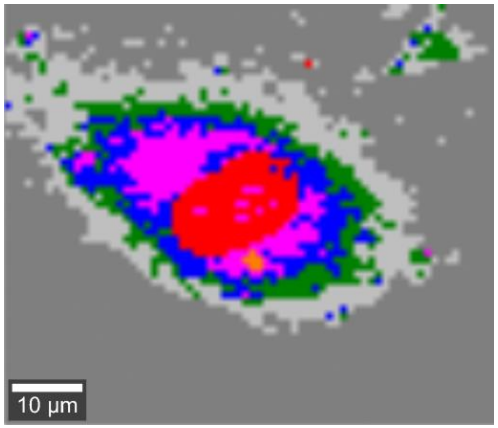
Journal of Molecular Liquids 303 (2020) 112676
Contents lists available at ScienceDirect
Journal of Molecular Liquids
journal homepage: www.elsevier.com/locate/molliq

Virtual spectral histopathology of colon cancer - biomedical applications of Raman spectroscopy and imaging

Beata Brozek-Pluska [a,*], Adam Dziki [b], Halina Abramczyk [a]
[a] Laboratory of Laser Molecular Spectroscopy, Institute of Applied Radiation Chemistry, Lodz University of Technology, Wroblewskiego 15, 93-590 Lodz, Poland
[b] Department of General and Colorectal Surgery, Medical University of Lodz, University Clinical Hospital No. 1, Military Medical Academy, Central Veterans' Hospital, Heller Square 1, 90-647 Lodz, Poland

Raman spectroscopy and imaging

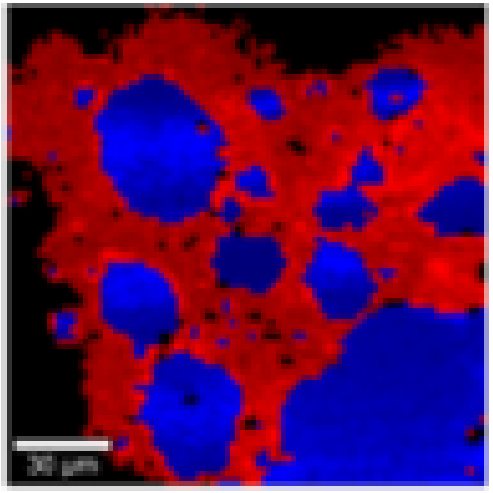


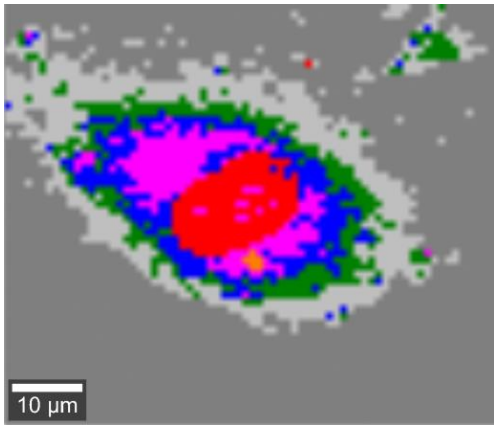


Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (**called a cluster**) are more similar (in some sense) to each other than to those in other groups (clusters).

It is a main task of exploratory data analysis, and a common technique for statistical data analysis, used in many fields, including pattern recognition, **image analysis**, information retrieval, bioinformatics, data compression, computer graphics and machine learning.

Raman spectroscopy and imaging

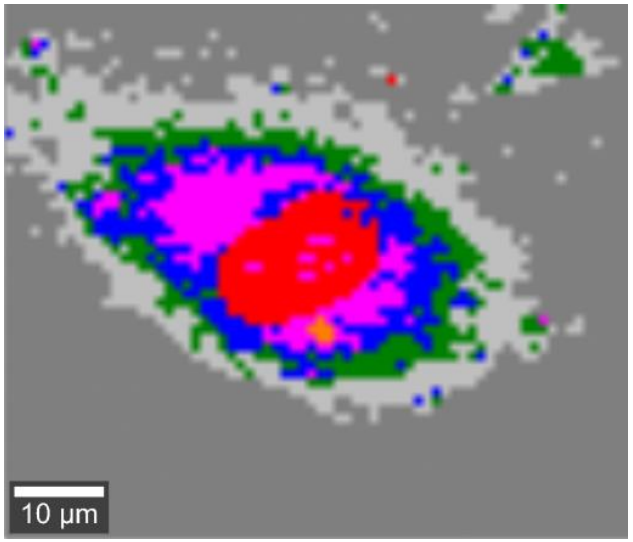


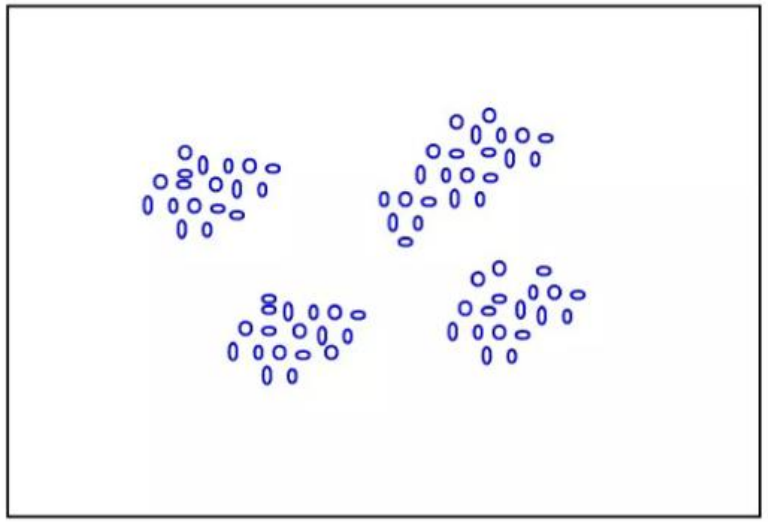


Cluster analysis itself is not one specific algorithm, but the general task to be solved.

Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings depend on the individual data set and intended use of the results. **Cluster analysis as such is not an automatic task.**

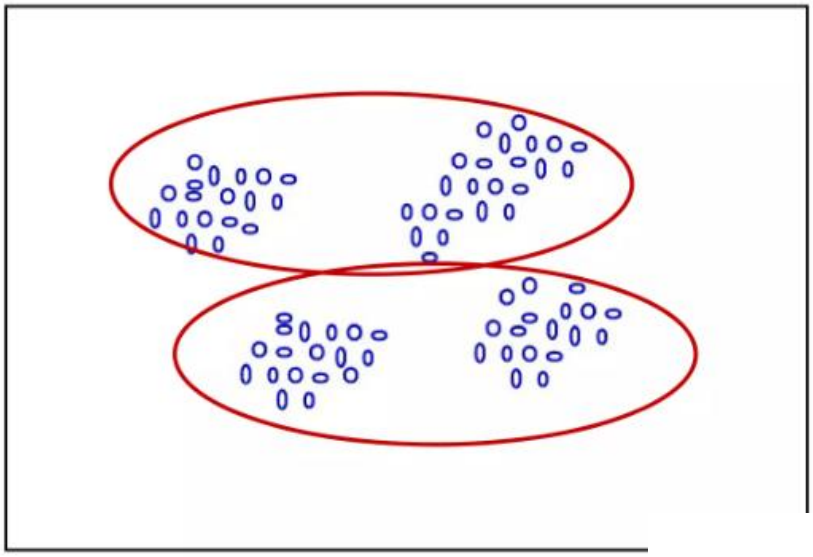
Cluster analysis



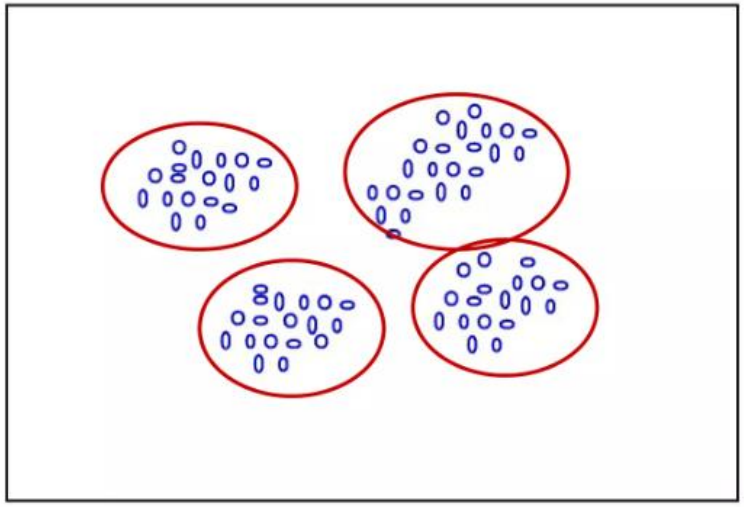


in cluster analysis a group of objects is split up into a number of more or less homogeneous subgroups on the basis of an often subjectively chosen measure of similarity (i.e., chosen subjectively based on its ability to create "interesting" clusters), such that the similarity between objects within a subgroup is larger than the similarity between objects belonging to different subgroups

Cluster analysis

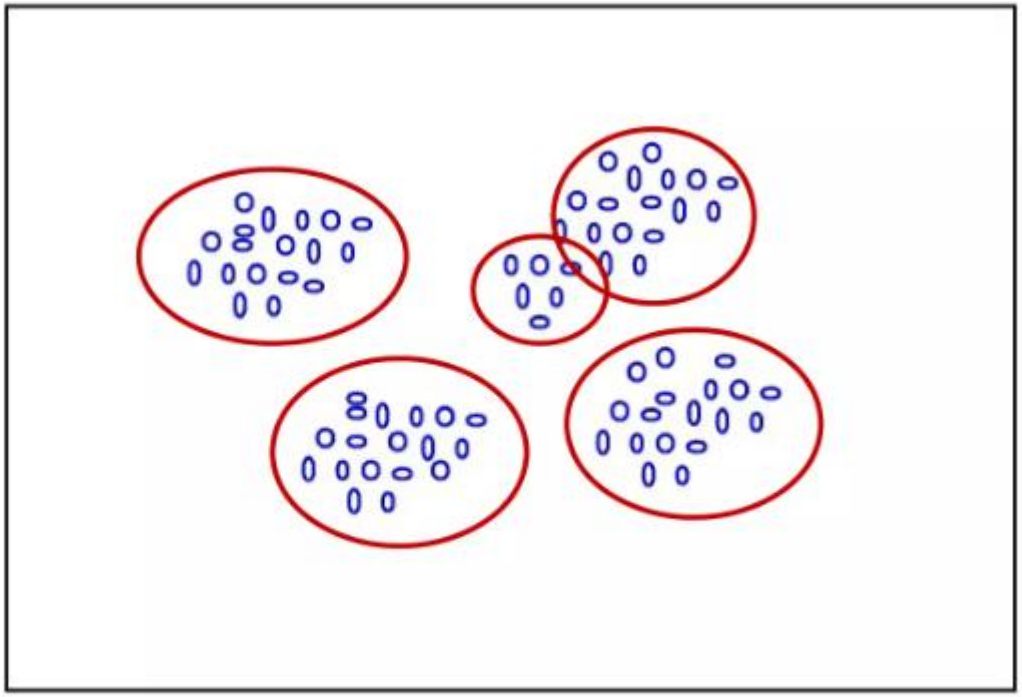


- Do you thing this is good?



- Do you thing this is better?

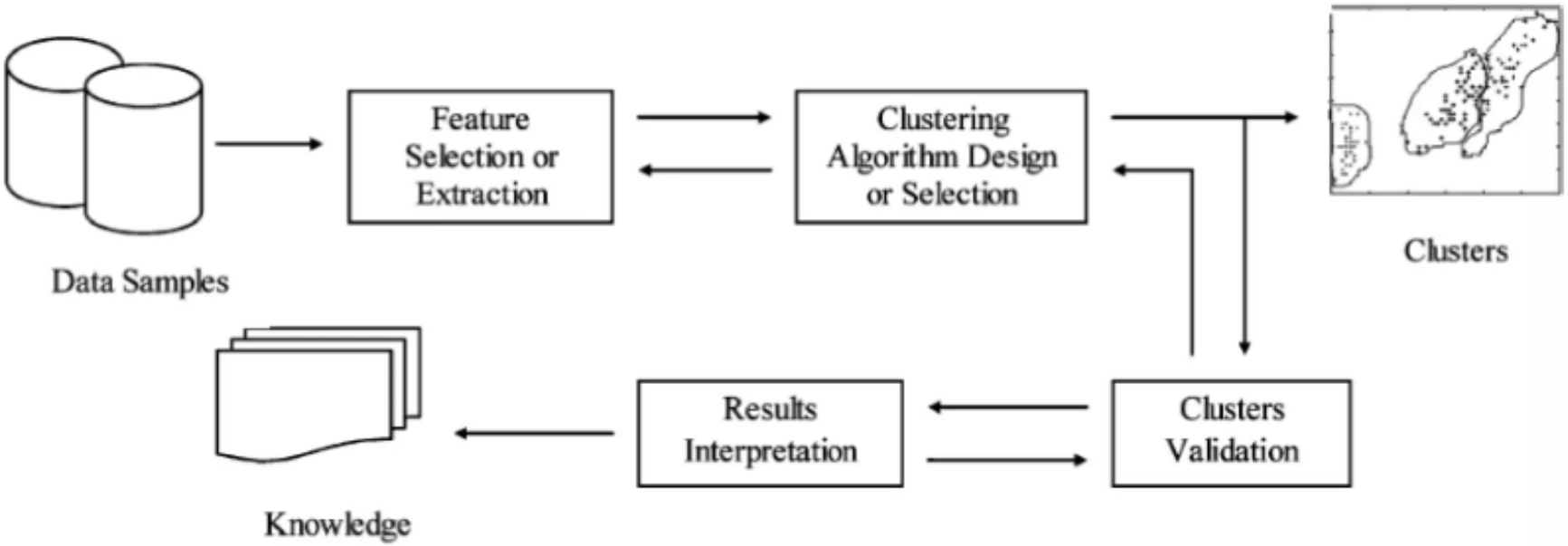
Cluster analysis



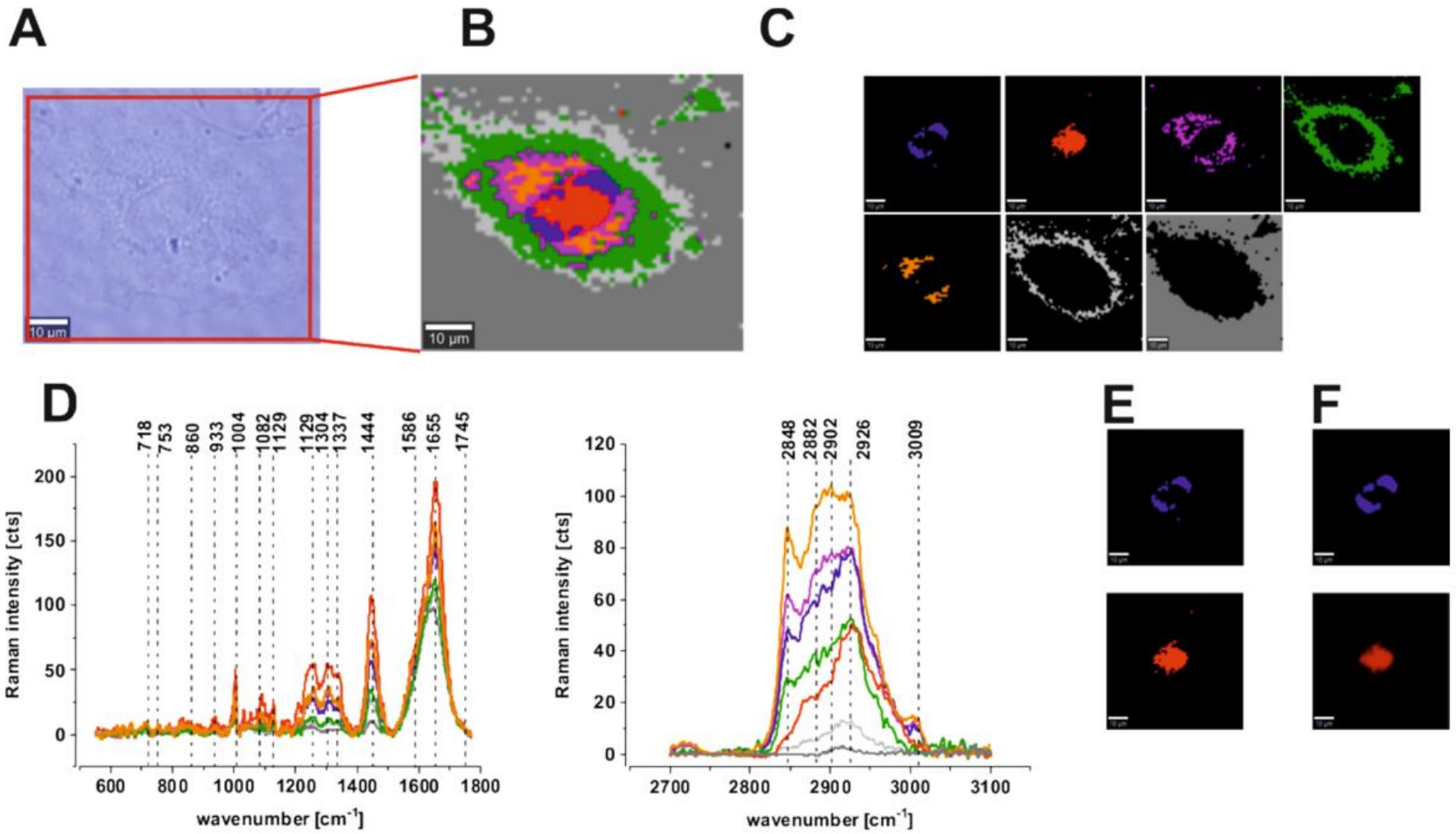
- **Do you thing this is better?**

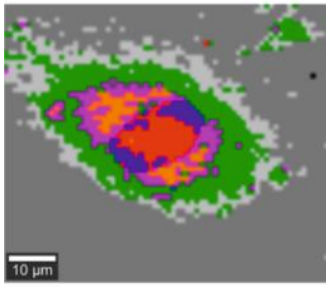
- **Homogeneity and separation principles**
  - Homogeneity: Elements within a cluster are close to each other
  - Separation: Elements in different clusters are further apart from each other

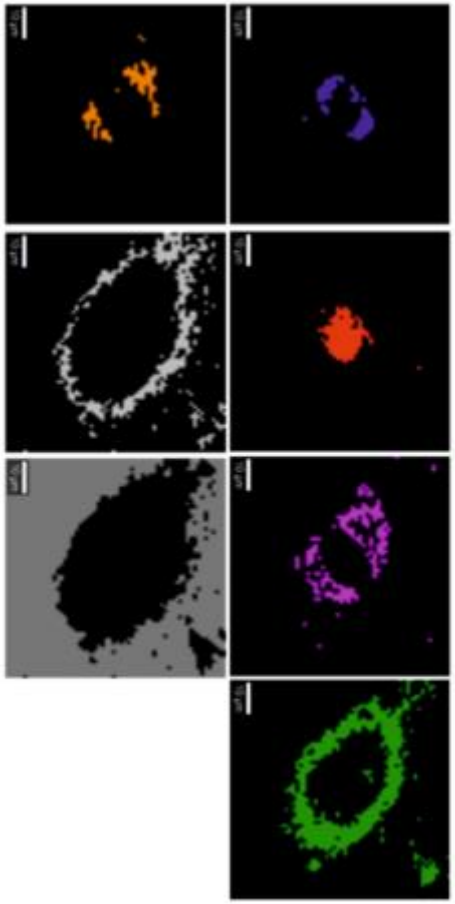
Cluster analysis



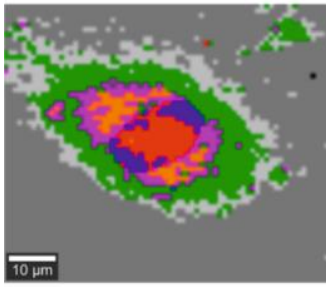
Raman spectroscopy and imaging

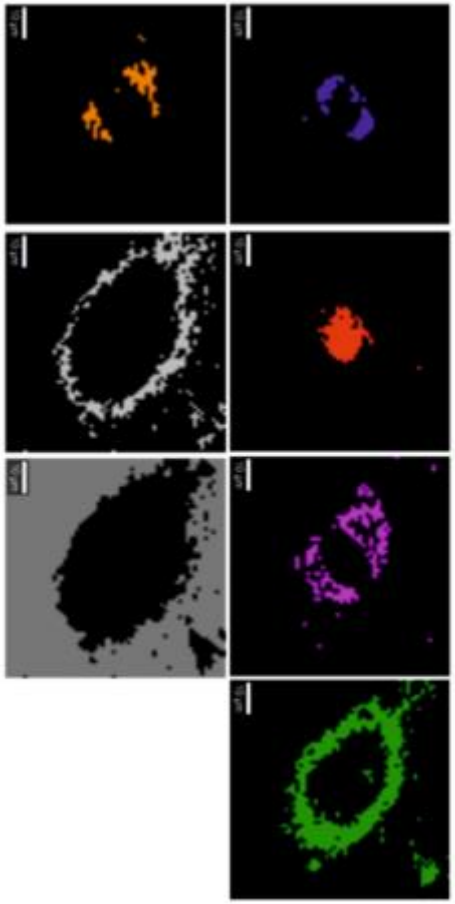




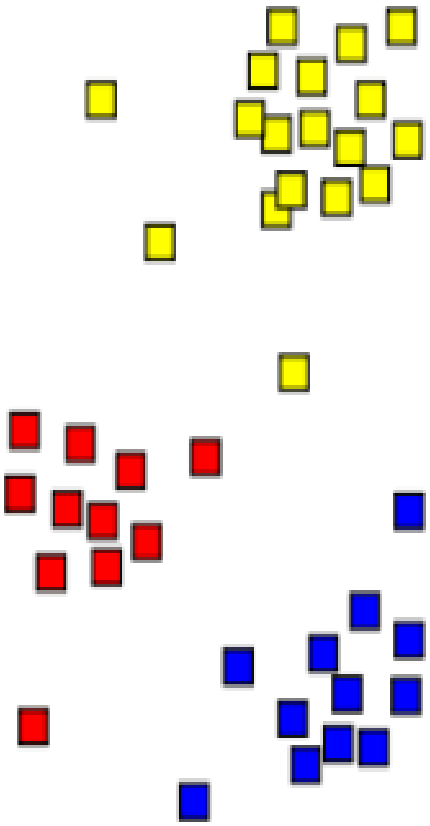


- *Connectivity models*: for example, hierarchical clustering builds models based on distance connectivity.
- *Centroid models*: for example, the k-means algorithm represents each cluster by a single mean vector.
- *Distribution models*: clusters are modeled using statistical distributions, such as multivariate normal distributions used by the expectation-maximization algorithm.

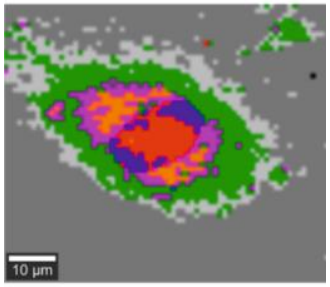


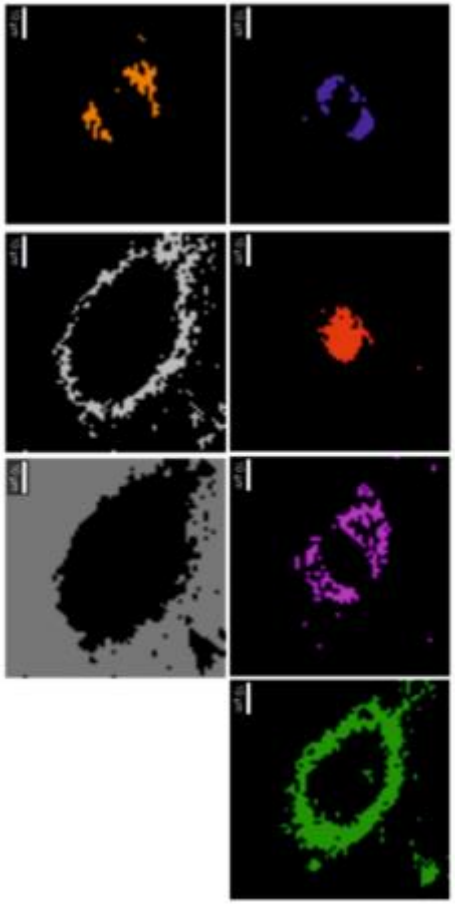


•*Neural models: the most well known unsupervised neural network is the self-organizing map and these models can usually be characterized as similar to one or more of the above models, and including subspace models when neural networks implement a form of Principal Component Analysis or Independent Component Analysis.*



•*Density models*: for example, DBSCAN and OPTICS defines clusters as connected dense regions in the data space.
•*Subspace models*: in biclustering (also known as co-clustering or two-mode-clustering), clusters are modeled with both cluster members and relevant attributes.
•*Group models*: some algorithms do not provide a refined model for their results and just provide the grouping information.
•*Graph-based models*: a clique, that is, a subset of nodes in a graph such that every two nodes in the subset are connected by an edge can be considered as a prototypical form of cluster. Relaxations of the complete connectivity requirement (a fraction of the edges can be missing) are known as quasi-cliques, as in the HCS clustering algorithm.



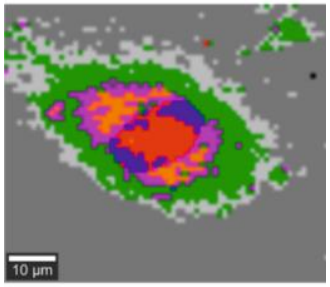


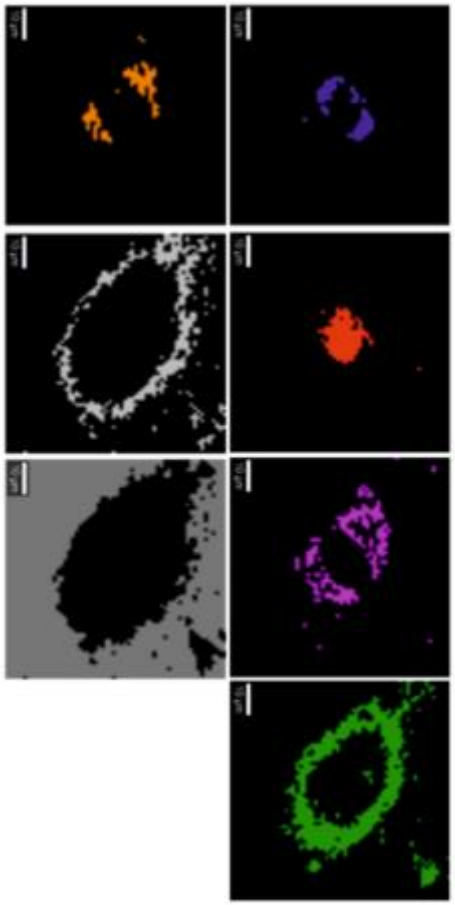
•*Centroid models*: for example, the k-means algorithm represents each cluster by a single mean vector.

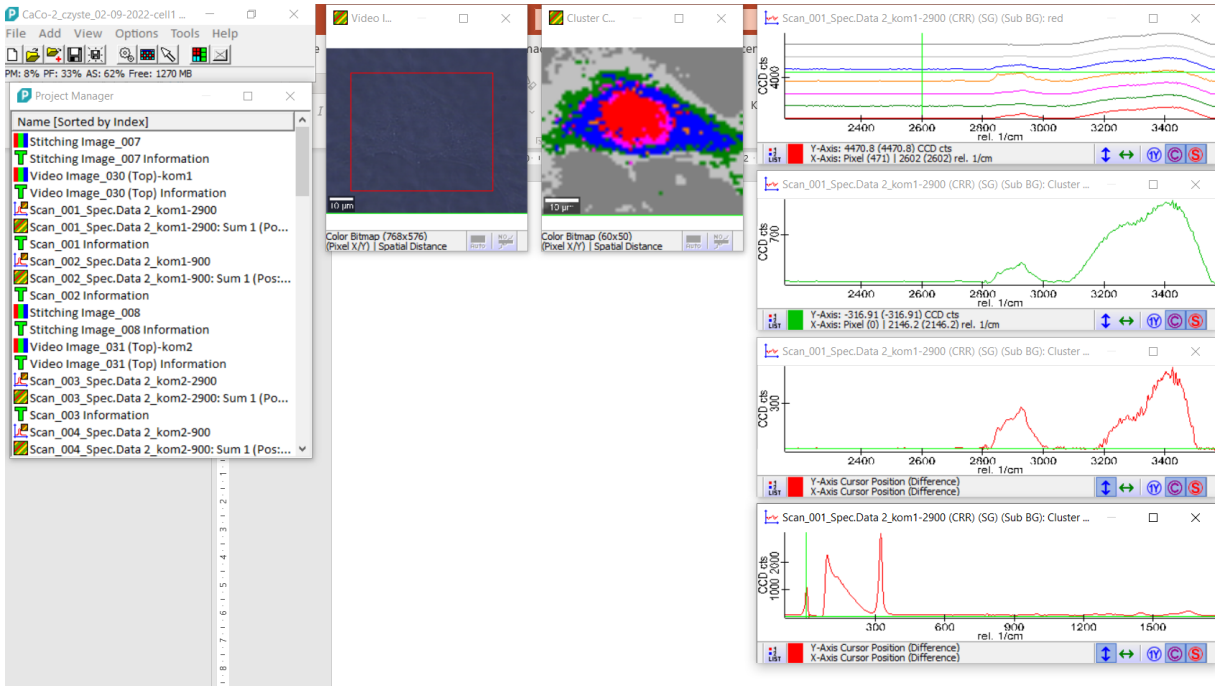
•*k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which* ***each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid)****, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using k-medians and k-medoids.*

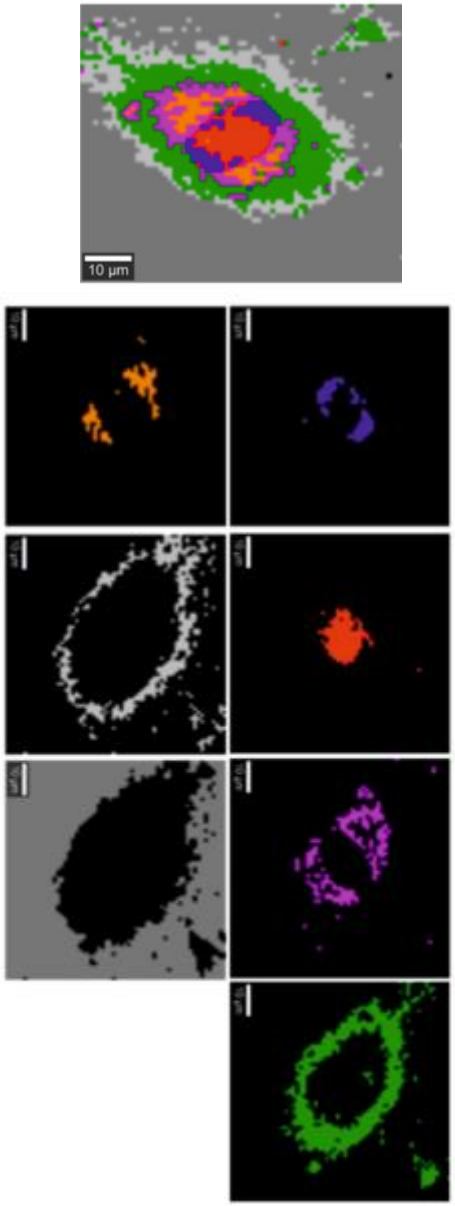
EXAMPLE

Based on LLMS data base

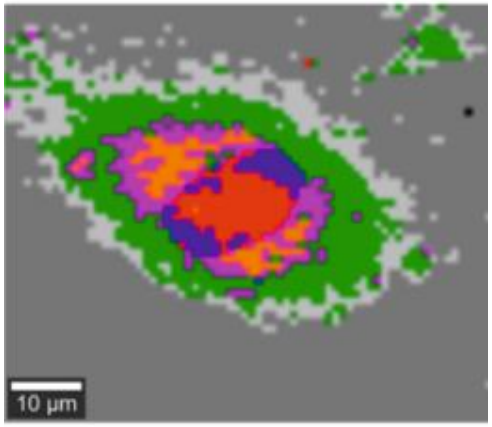


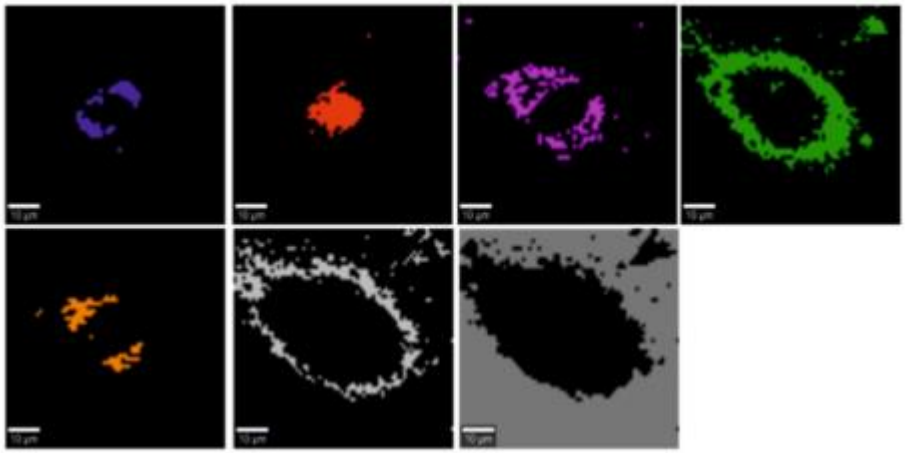




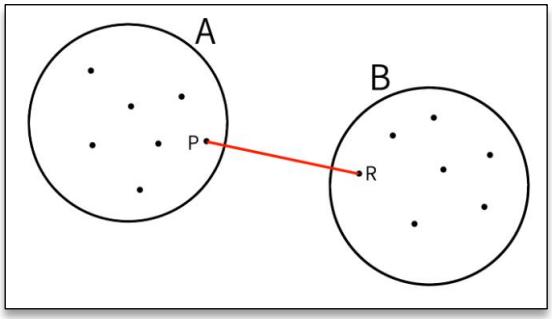


- ***Strict partitioning clustering***: each object belongs to exactly one cluster
- ***Strict partitioning clustering with outliers***: objects can also belong to no cluster, in which case they are considered outliers
- ***Overlapping clustering*** (also: *alternative clustering*, *multi-view clustering*): objects may belong to more than one cluster; usually involving hard clusters
- ***Hierarchical clustering***: objects that belong to a child cluster also belong to the parent cluster
- ***Subspace clustering***: while an overlapping clustering, within a uniquely defined subspace, clusters are not expected to overlap



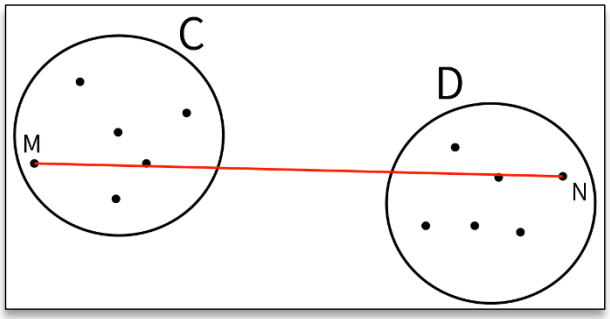


definition of distance



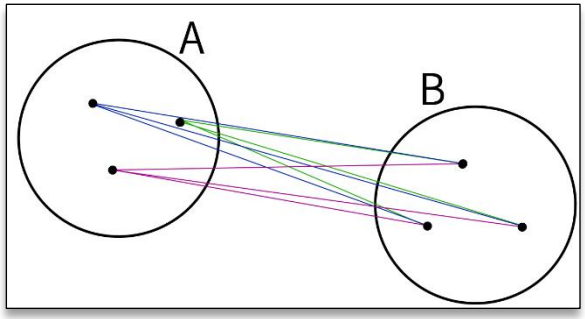
single linkage algorithm

$$d = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

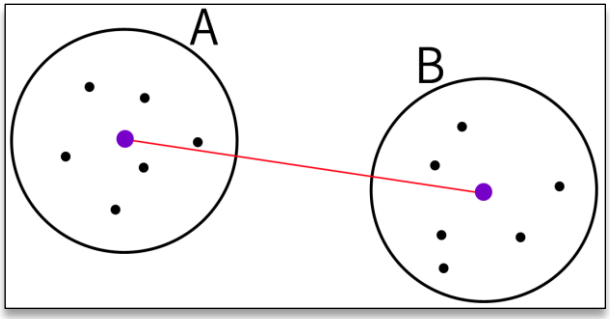


complete linkage

$$d = \sum_{i=1}^{n}(q_i - p_i)^2$$

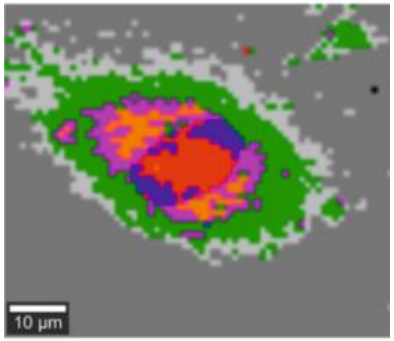


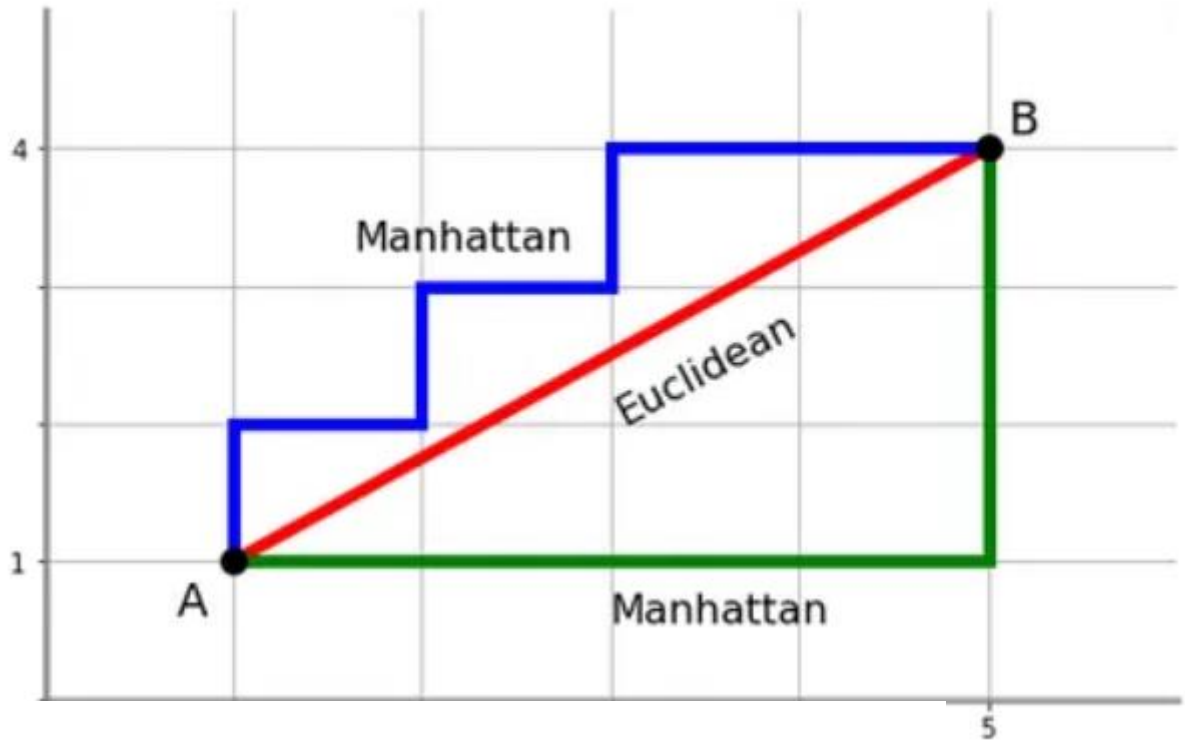
group average



distance between centroids

$$d = \sum_{i=1}^{n}|q_x - p_x| + |q_y - p_y|$$



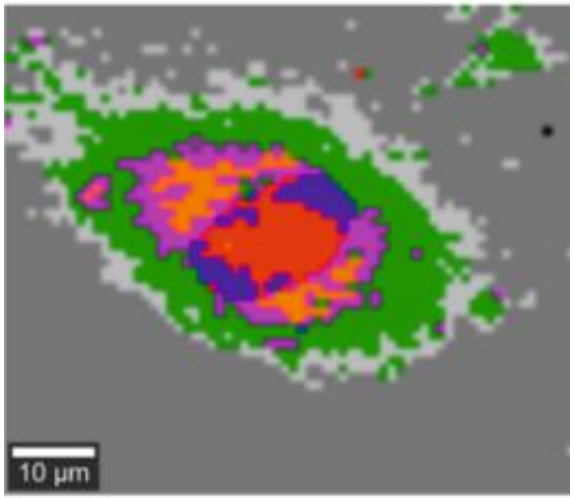


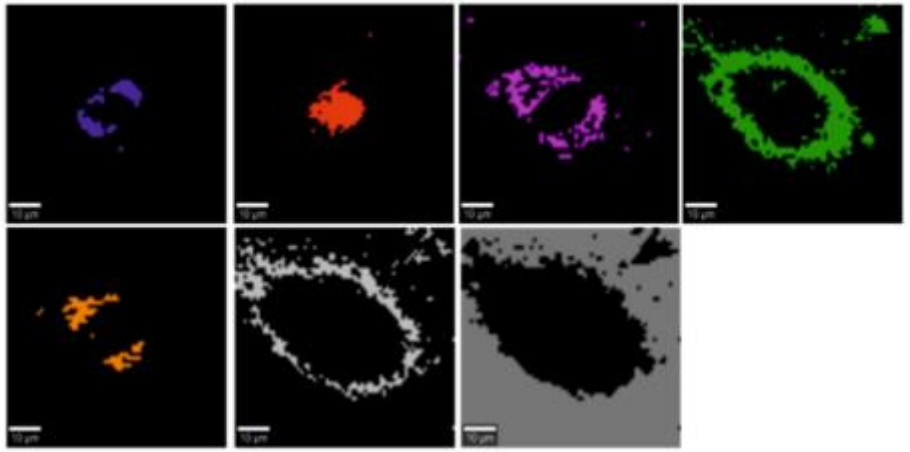
$$d = \sum_{i=1}^{n} |q_x - p_x| + |q_y - p_y|$$

The manhattan distance is a different way of measuring distance. It is named after the grid shape of streets in Manhattan. If there are two points, (x_1, y_1) and (x_2, y_2), the manhattan distance between the two points is $|x_1 - x_2| + |y_1 - y_2|$.

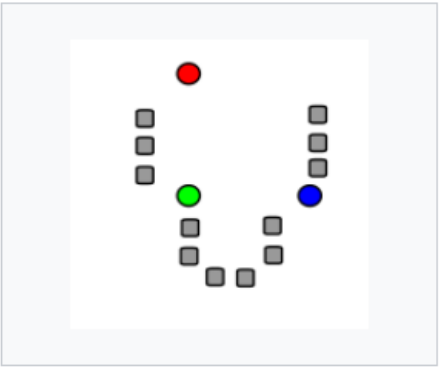
This distance can be imagined as the length needed to move between two points in a grid where you can only move up, down, left or right.

BB-P

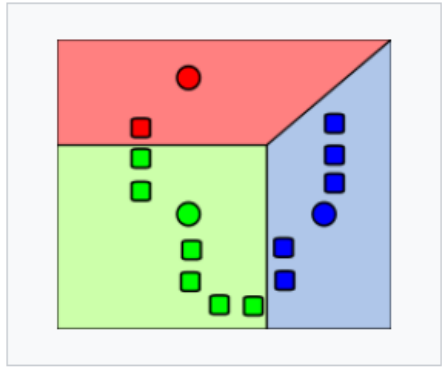




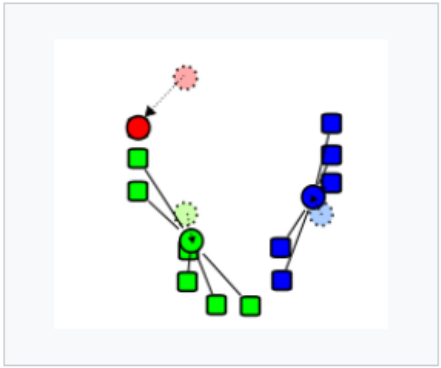
Demonstration of the standard algorithm



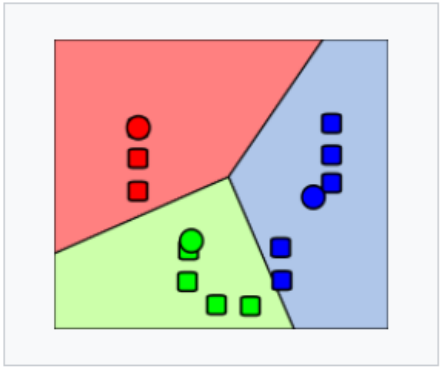
1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).



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

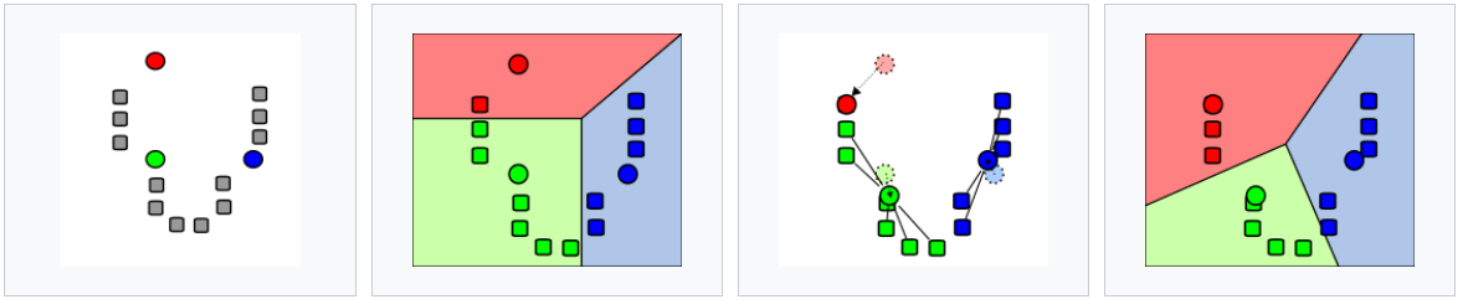


3. The centroid of each of the *k* clusters becomes the new mean.



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

Demonstration of the standard algorithm



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.

K-means algorithm iteratively **minimizes** the distances between every data point and its centroid in order to find the most optimal solution for all the data points.

1.k random points of the data set are chosen to be centroids.

2.Distances between every data point and the *k* centroids are calculated and stored.

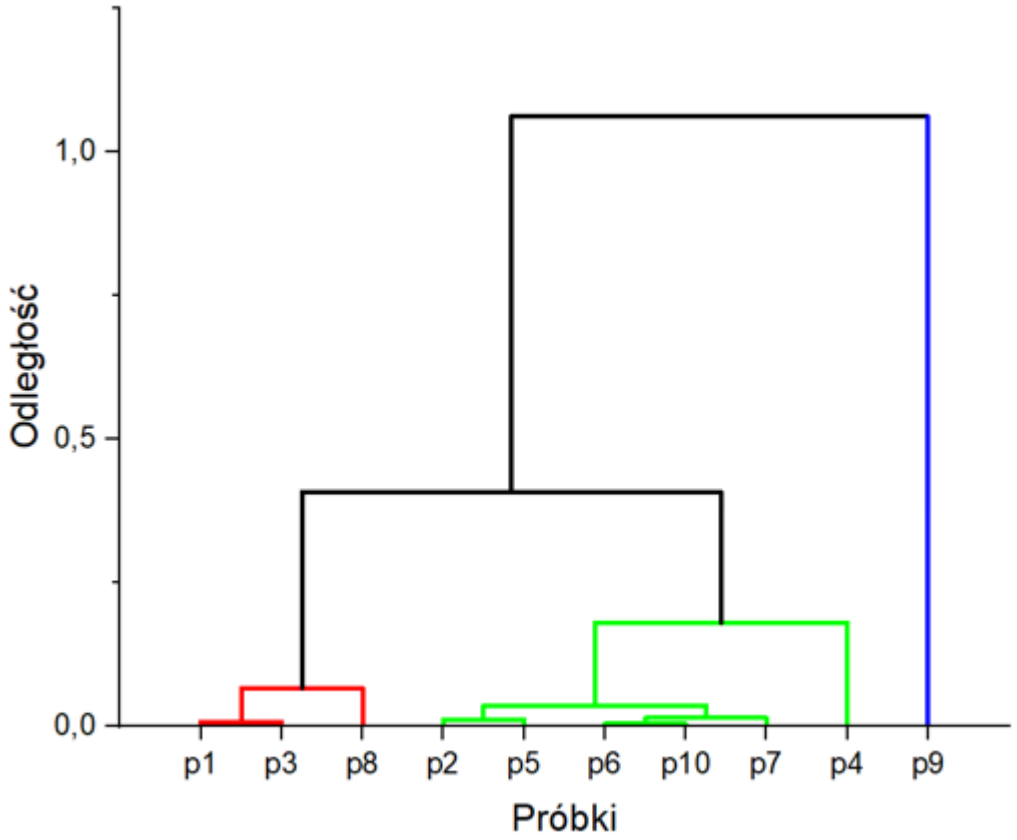
3.Based on distance calculates, each point is assigned to the nearest cluster

4.New cluster centroid positions are updated: similar to finding a mean in the point locations

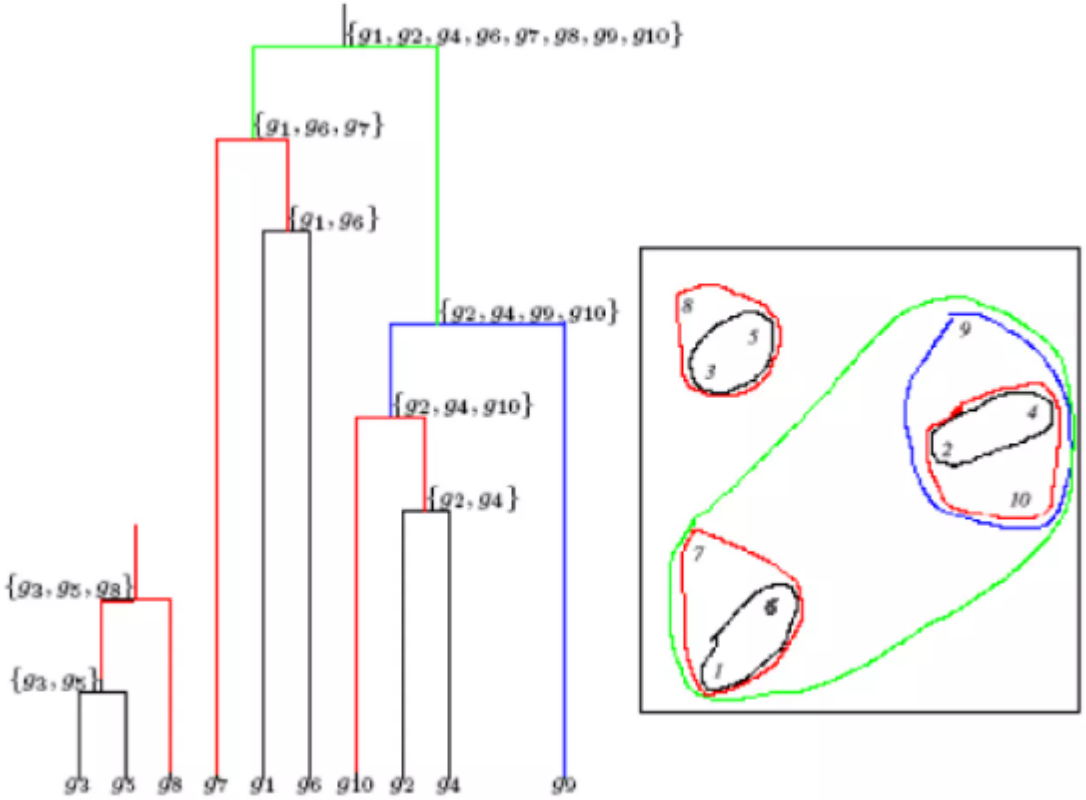
5.If the centroid locations changed, the process repeats from step 2, until the calculated new center stays the same, which signals that the clusters' members and centroids are now set.

BB-P

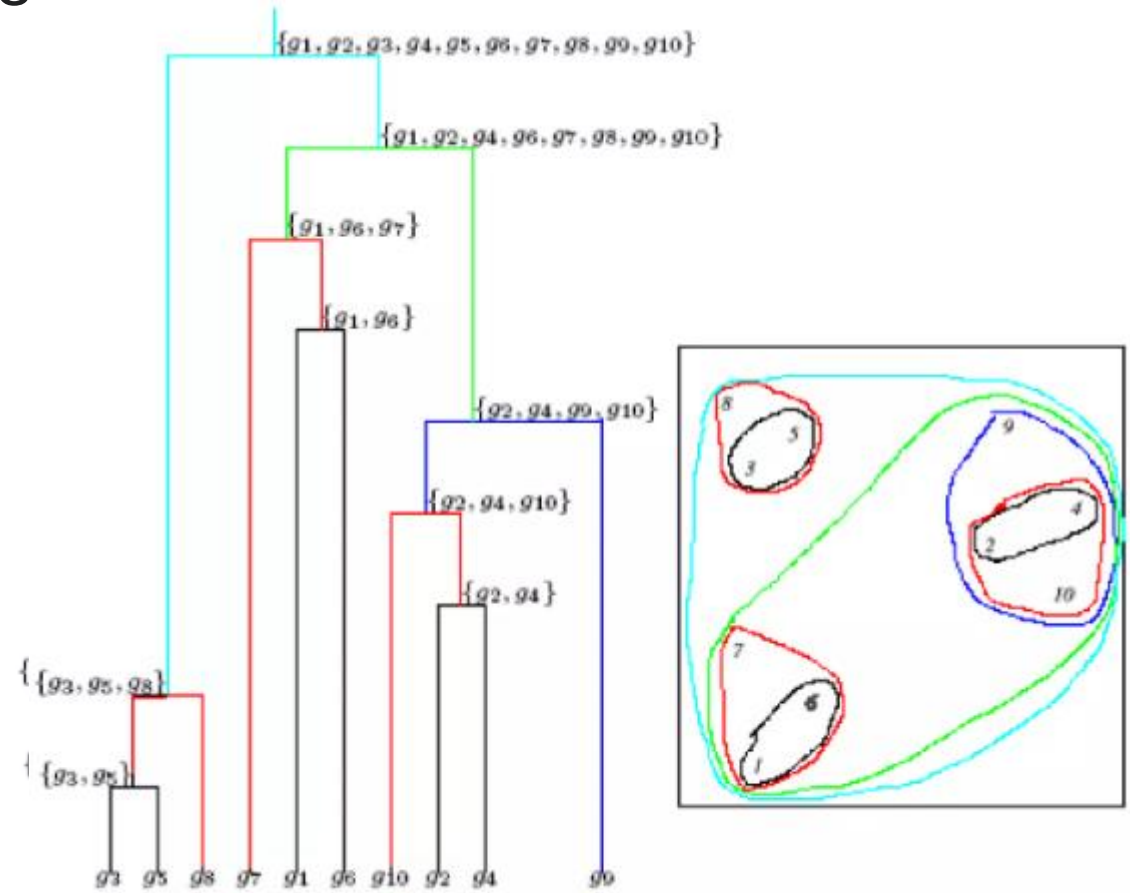
•***Hierarchical clustering***: objects that belong to a child cluster also belong to the parent cluster



- ***Hierarchical clustering***: objects that belong to a child cluster also belong to the parent cluster



- ***Hierarchical clustering***: objects that belong to a child cluster also belong to the parent cluster

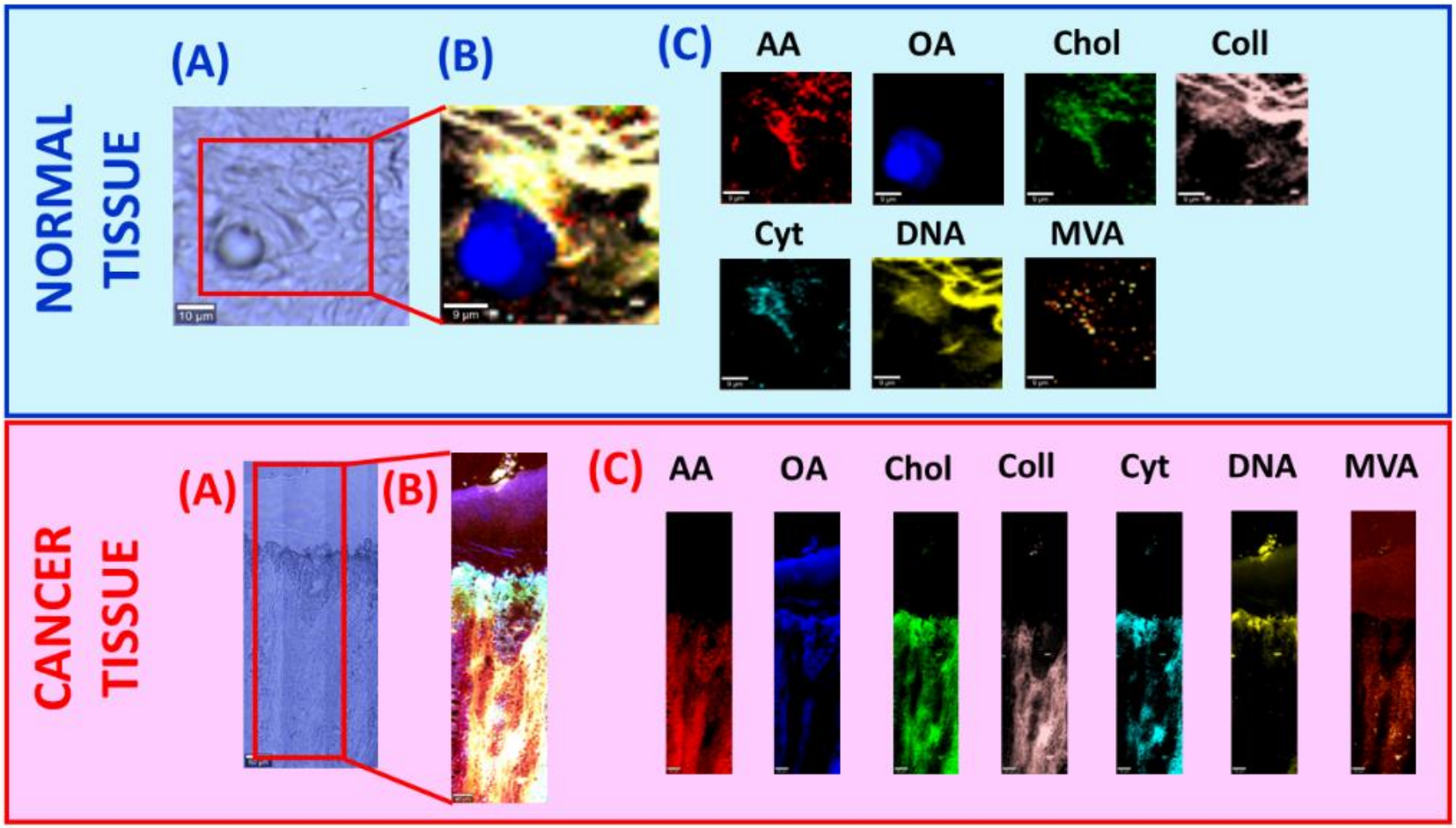


Pseudocode

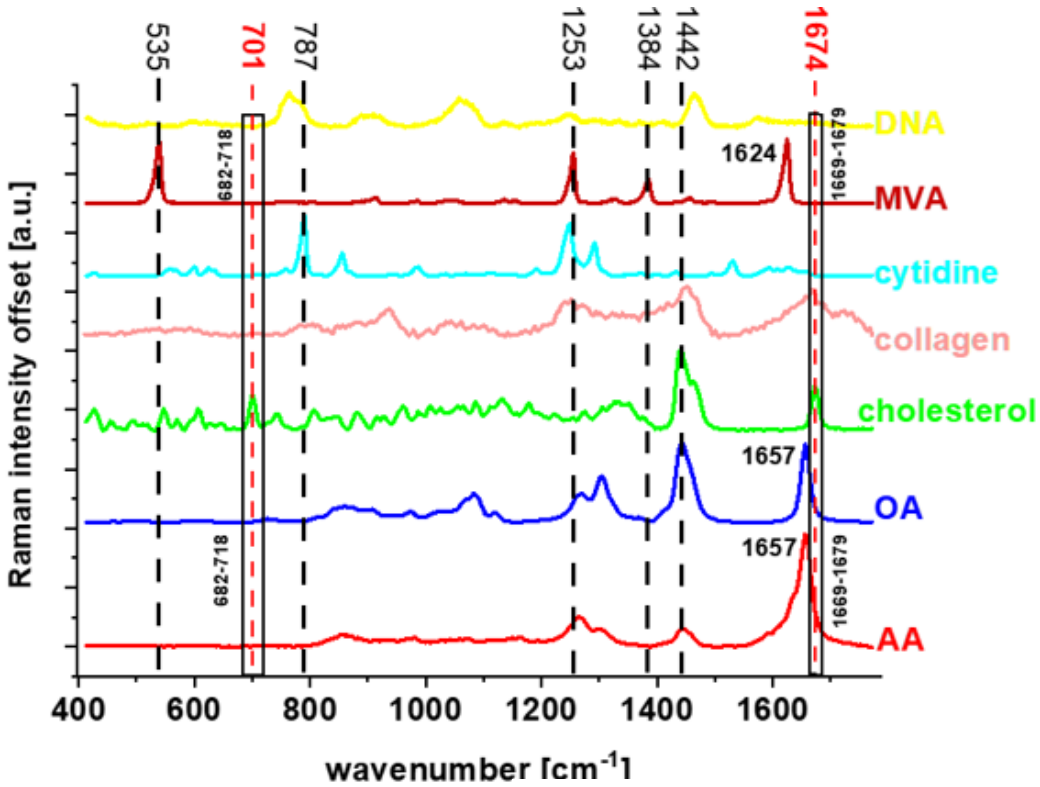
Hierarchical Clustering (d , n)

1. Form *n* clusters each with one element
2. Construct a graph *T* by assigning one vertex to each cluster
3. **while** there is more than one cluster
    1. Find the two closest clusters C1 and C2
    2. Merge C1 and C2 into new cluster C with |C1| +|C2| elements
    3. Compute distance from C to all other clusters
    4. **if** they are close
        1. Add a new vertex C to *T* and connect to vertices C1 and C2
        2. Remove rows and columns of d corresponding to C1 and C2
        3. Add a row & column to d corresponding to the new cluster C
4. **return** *T*

Basis analysis



Basis analysis

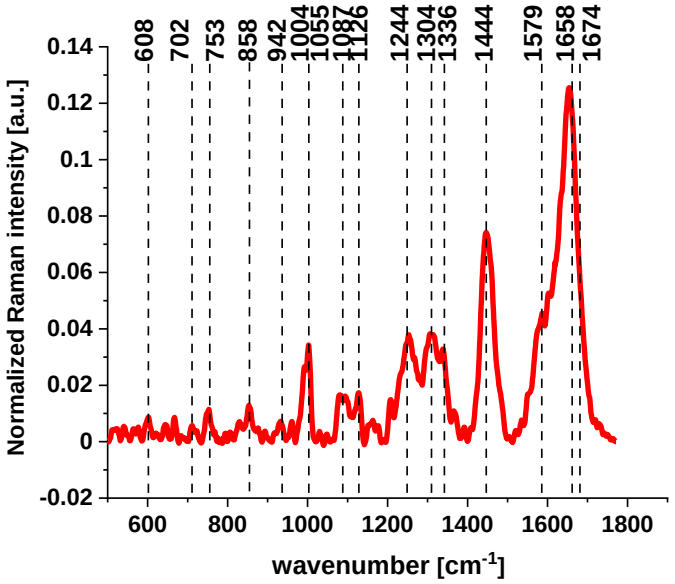


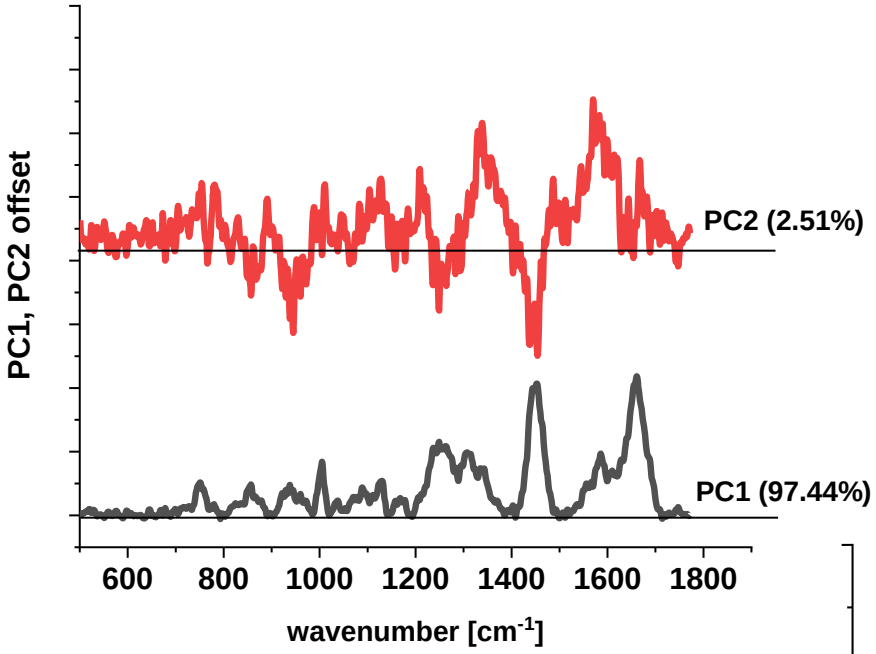
During the analysis each measured spectrum of the 2D spectral array of the analyzed human breast sample was compared to the spectra of pure chemical components mentioned above using a least square to fit each convergence to minimize the fitting error D described by equation:

$$D = \left(\overrightarrow{[\text{Recorded spectrum}]} - a \times \overrightarrow{BS_A} - b \times \overrightarrow{BS_B} - c \times \overrightarrow{BS_C} - \cdots \right)^2$$

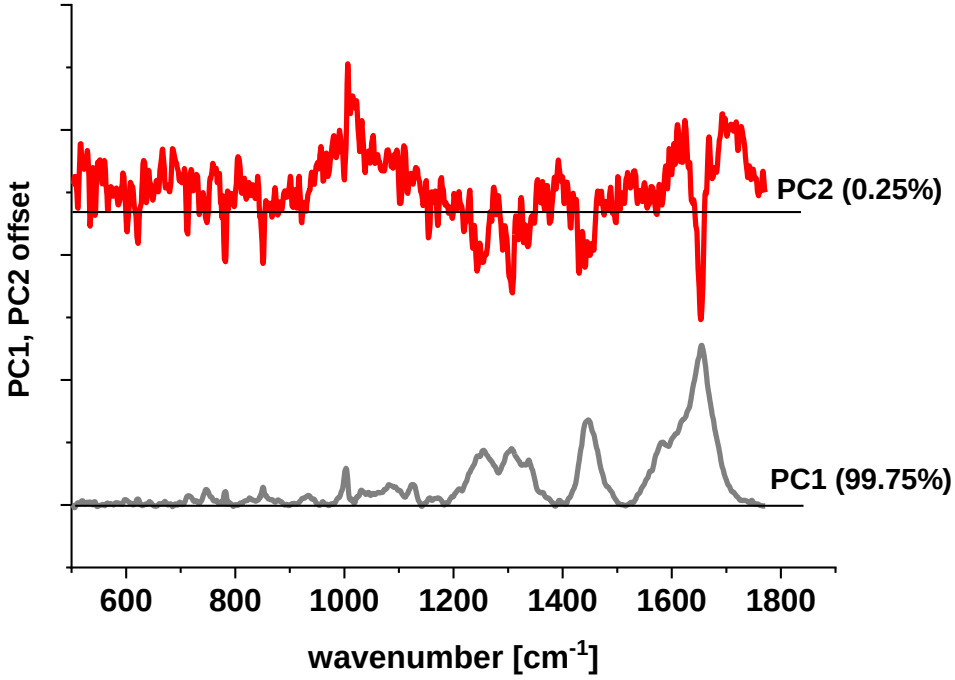
by varying the weighting factors a, b, c,… of the basis spectra $\underset{BS}{\rightarrow}$.

Principal component analysis (**PCA**) is a popular technique for analyzing large datasets containing a high number of dimensions/features per observation, increasing the interpretability of data while preserving the maximum amount of information, and enabling the visualization of multidimensional data. Formally, PCA is a statistical technique for reducing the dimensionality of a dataset.





PCA



molecules
MDPI
Article

Raman Spectroscopy and Imaging Studies of Human Digestive Tract Cells and Tissues—Impact of Vitamin C and E Supplementation

Karolina Beton-Mysur and Beata Brozek-Pluska *

PCA is defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some scalar projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

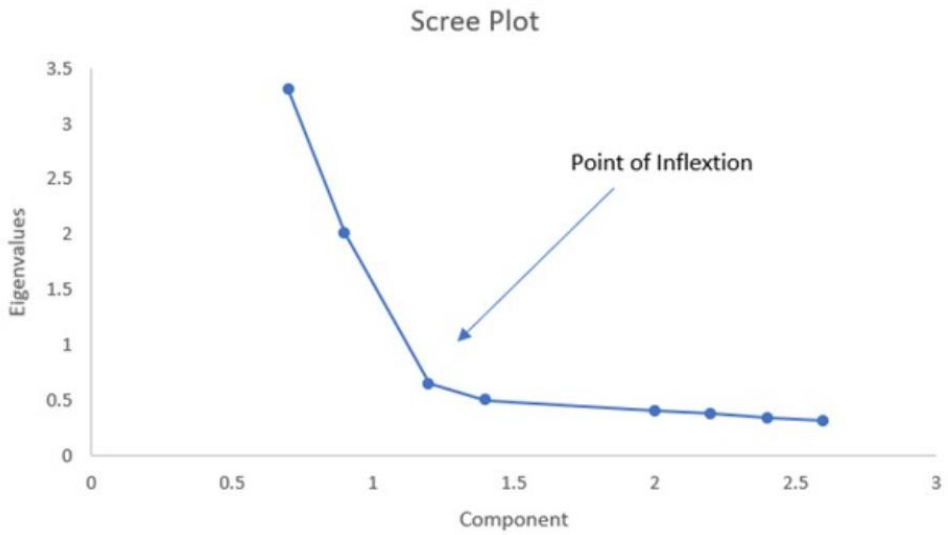
Dimensionality reduction

The transformation $\mathbf{T} = \mathbf{X}\,\mathbf{W}$ maps a data vector $\mathbf{x}_{(i)}$ from an original space of p variables to a new space of p variables which are uncorrelated over the dataset. However, not all the principal components need to be kept. Keeping only the first L principal components, produced by using only the first L eigenvectors, gives the truncated transformation

The following variable reduction criteria can be used:

1. Criterion of sufficient proportion - when the degree of explanation of variability reaches at least 75%, it can be considered that the number of principal components is sufficient
2. Kaiser's criterion - consists in the elimination of principal components with eigenvalues less than 1.
3. Scree plot – plotting

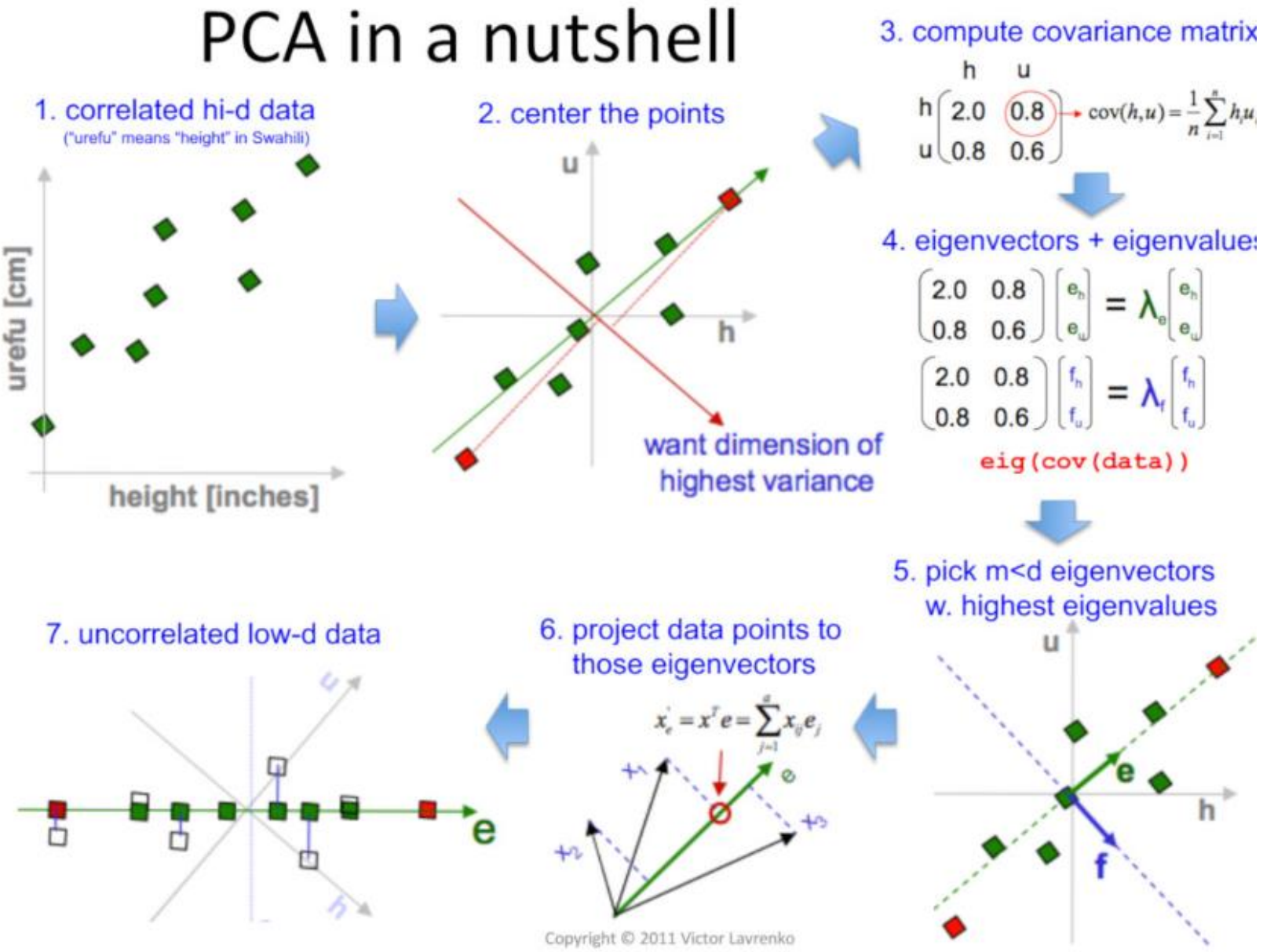
a line plot with successive eigenvalues and finding a point from which there is a slight decrease in eigenvalues to the right



When applying PCA, particular attention should be paid to the following issues:

1. Statistical distribution of data should be close to normal distribution.
2. It should be ensured that the examined sample is as numerous and as representative as possible. The group size may be smaller, the higher the correlation of the data with each other. The representativeness of the sample is based on the selection of samples that do not show excessive deviations.
3. However, if there are deviations in the data values that could significantly affect the final result of the analysis, they should be removed.

4. It is necessary to determine such a number of variables to enable their proper analysis. Limiting them too much may introduce errors in further analyzes and give too general a result, while introducing too many variables may complicate the analysis.
5. There may be a problem in the form of missing data for some variables. An appropriate approach may be to replace missing values with the mean value or to omit a factor in a given sample if it is not significant

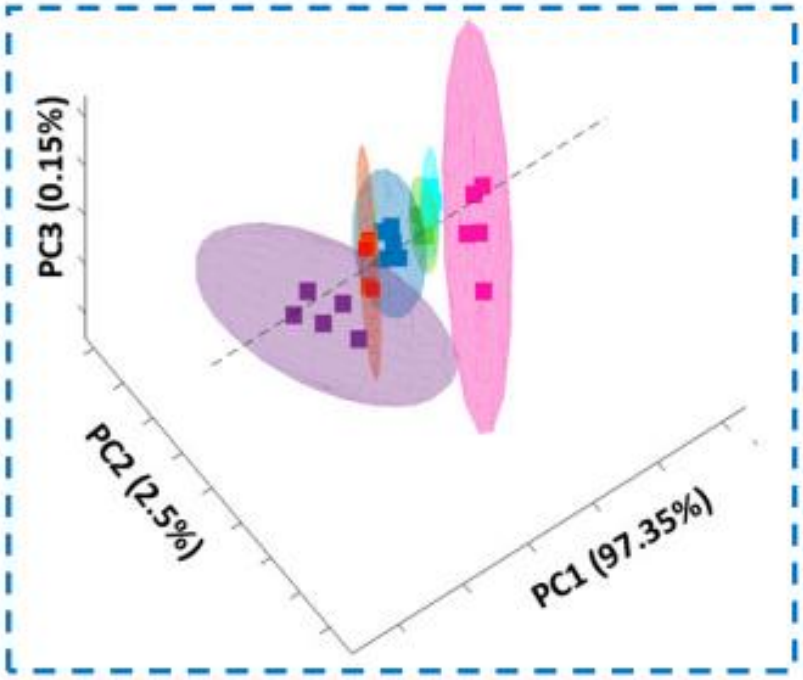


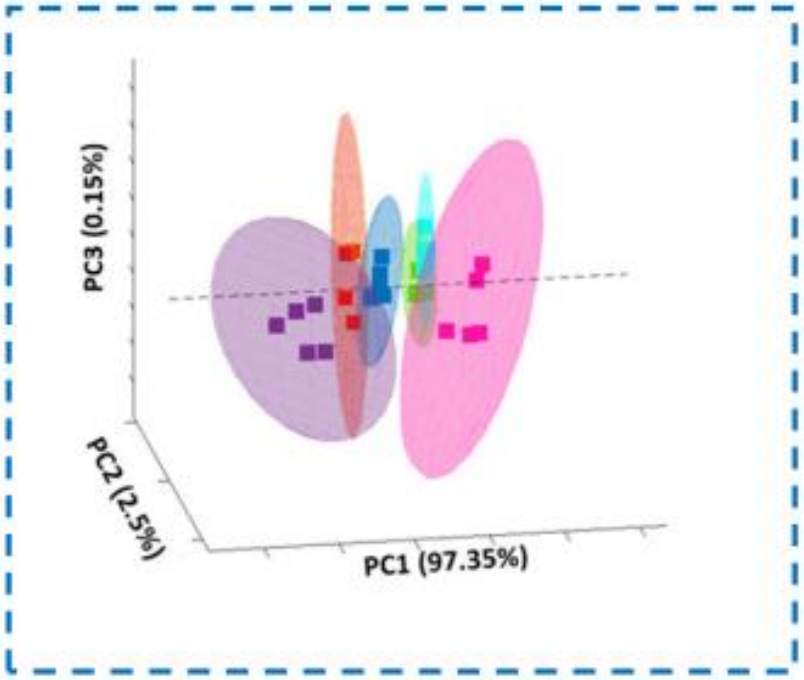
PCA in a nutshell. Source: Lavrenko and Sutton 2011, slide 13.*

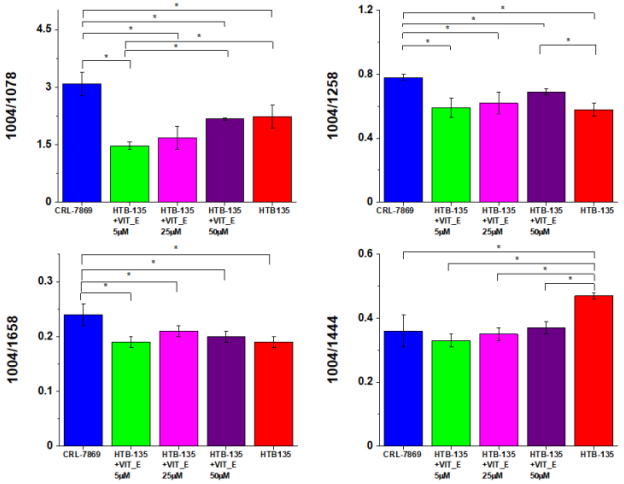
- Here are some drawbacks of PCA:

- PCA works only if the observed variables are linearly correlated. If there's no correlation, PCA will fail to capture adequate variance with fewer components.
- PCA is lossy. Information is lost when we discard insignificant components.
- Scaling of variables can yield different results. Hence, scaling that you use should be documented. Scaling should not be adjusted to match prior knowledge of data.
- Since each principal components is a linear combination of the original features, visualizations are not easy to interpret or relate to original features.

PCA







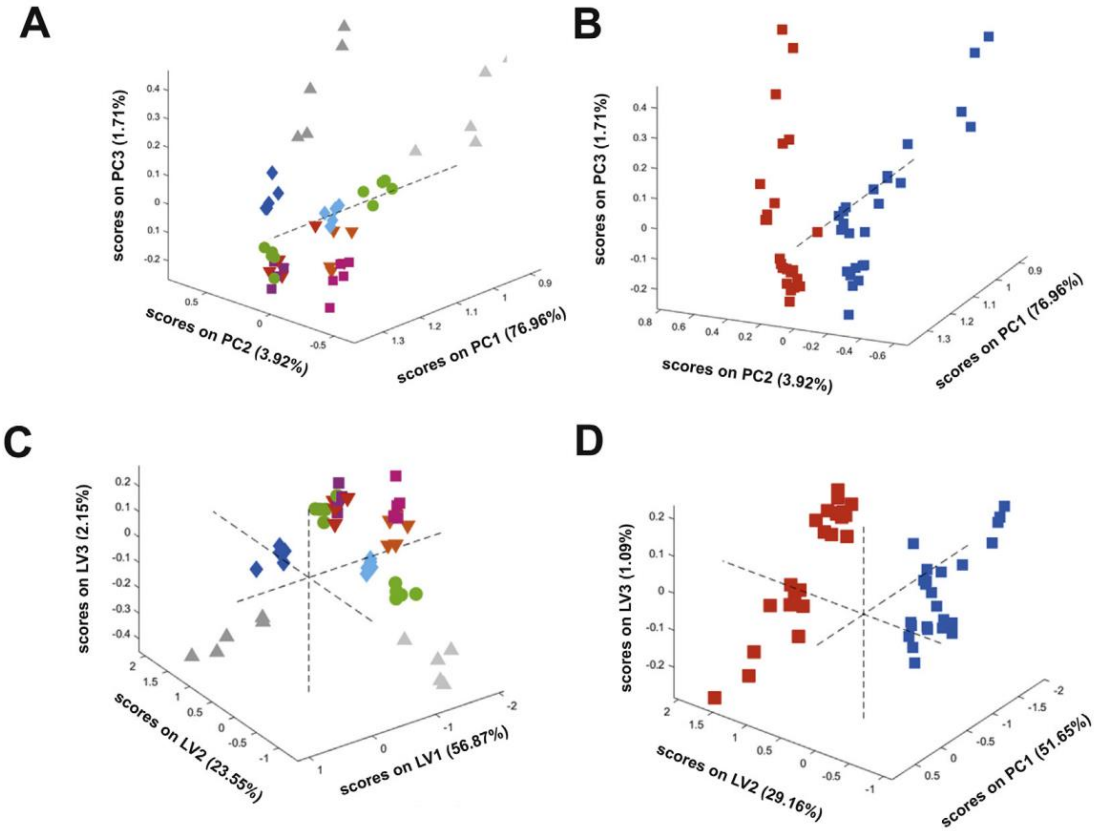
molecules
MDPI
Article

Raman Spectroscopy and Imaging Studies of Human Digestive Tract Cells and Tissues—Impact of Vitamin C and E Supplementation

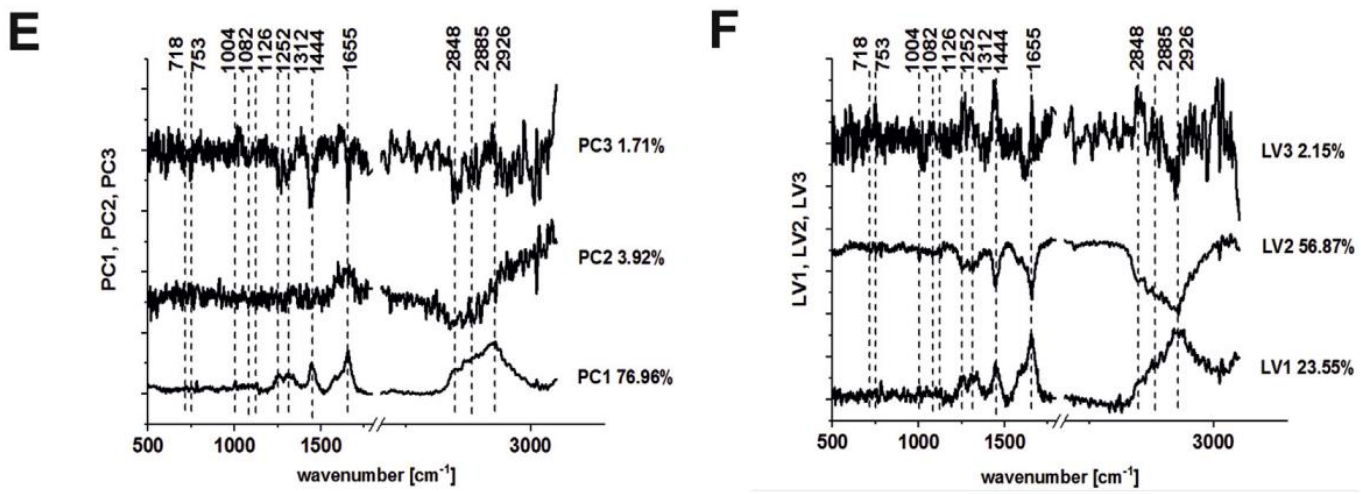
Karolina Beton-Mysur and Beata Brozek-Pluska *

PLSDA

PLS-DA combines **dimensionality reduction** and **discriminant analysis** into one algorithm and is especially applicable to modelling highdimensional data. In addition, PLS-DA does not assume the data to fit a particular distribution and thus is more flexible than other discriminant algorithms.



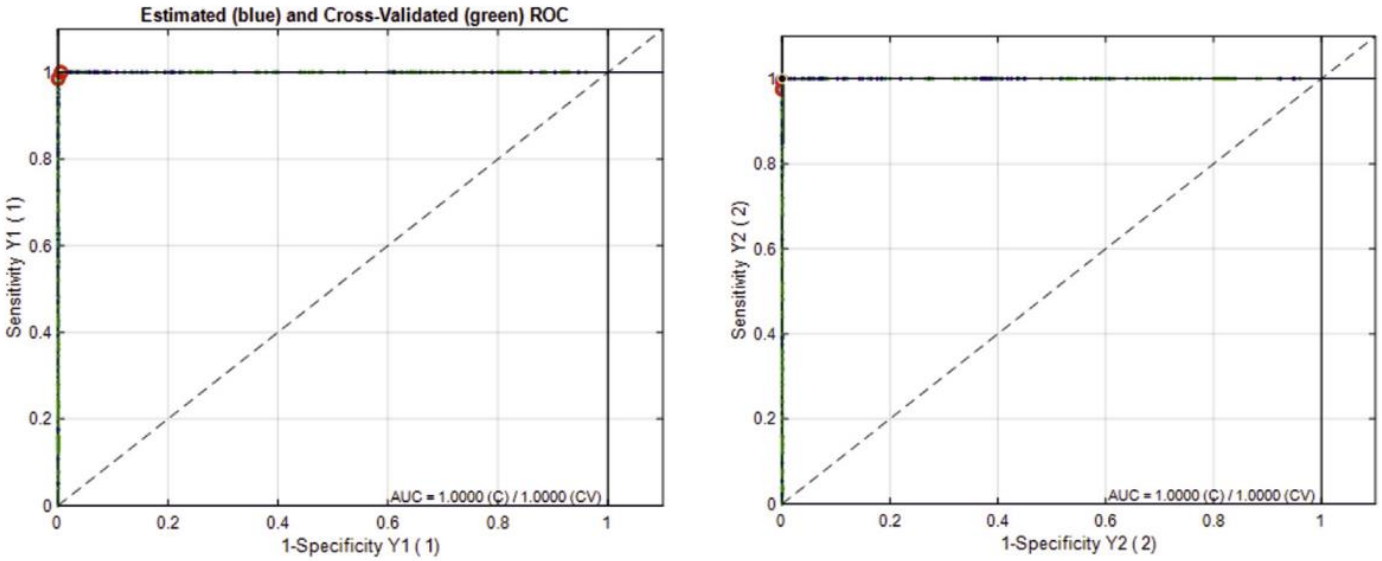
PLSDA



The sensitivity and specificity for calibration and cross validation procedure based on PLS-DA analysis for CCD18-Co and CRL-1831 normal human colon cells.

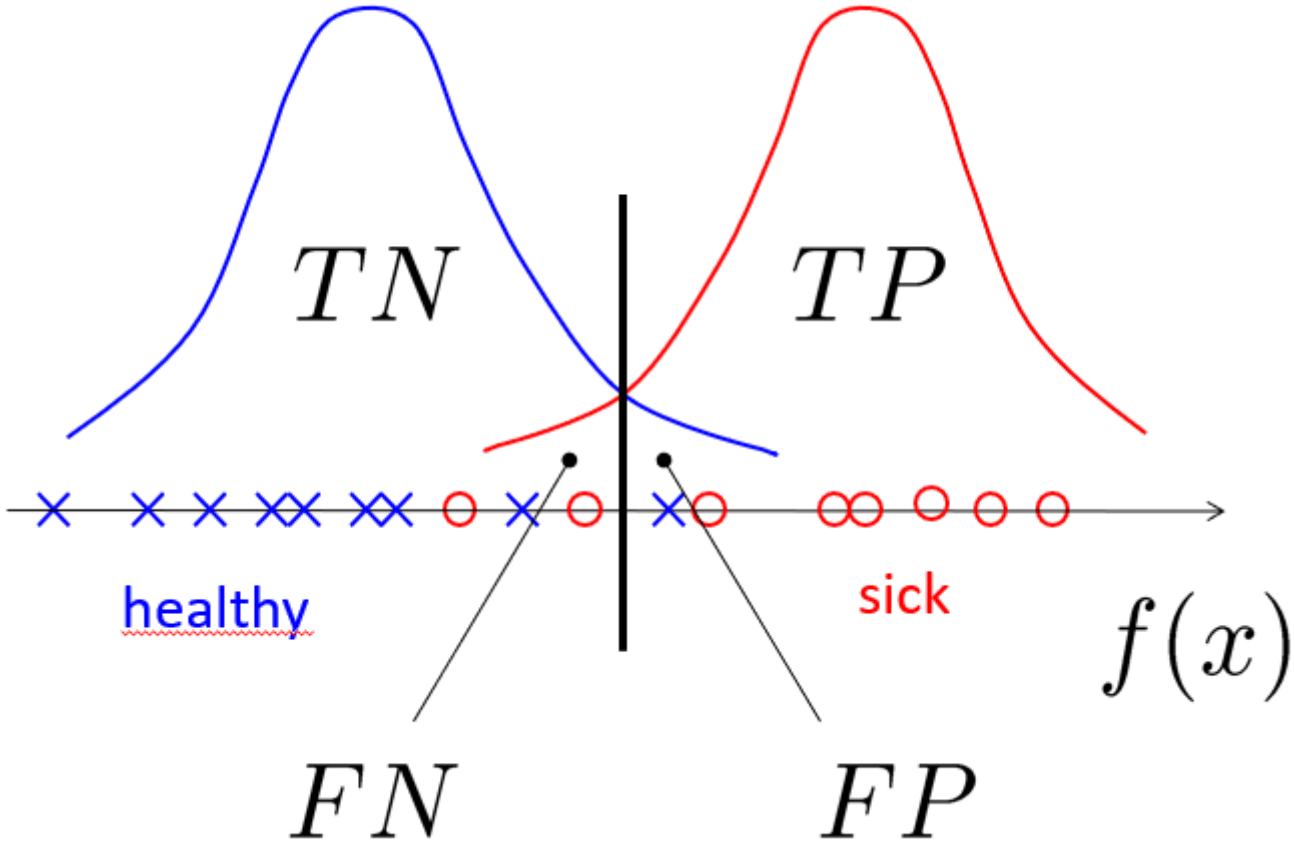
Sensitivity (calibration)	Sensitivity (cross validation)
1.0 for epithelial cells CRL-1831	1.0 for epithelial cells CRL-1831
1.0 for fibroblast cells CCD18-Co	0.8 for fibroblast cells CCD18-Co
Specificity (calibration)	**Specificity (cross validation)**
1.0 for epithelial cells CRL-1831	1.0 for epithelial cells CRL-1831
1.0 for fibroblast cells CCD18-Co	0.8 for fibroblast cells CCD18-Co

PLSDA

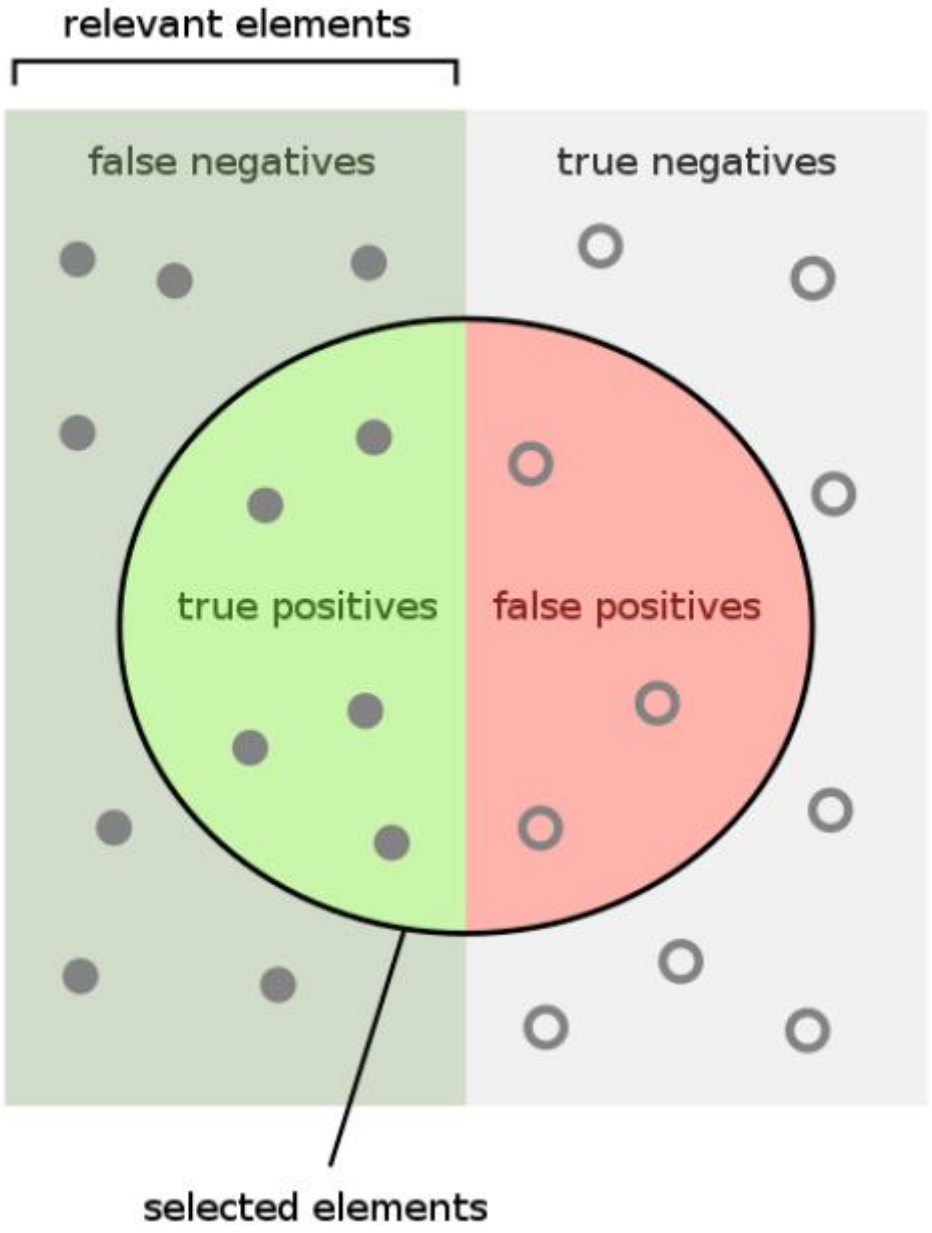


The sensitivity and specificity for calibration and cross validation procedure based on PLS-DA analysis for CCD18-Co and CRL-1831 normal human colon cells.

Sensitivity (calibration)	Sensitivity (cross validation)
1.0 for epithelial cells CRL-1831	1.0 for epithelial cells CRL-1831
1.0 for fibroblast cells CCD18-Co	0.8 for fibroblast cells CCD18-Co
Specificity (calibration)	**Specificity (cross validation)**
1.0 for epithelial cells CRL-1831	1.0 for epithelial cells CRL-1831
1.0 for fibroblast cells CCD18-Co	0.8 for fibroblast cells CCD18-Co


TN
TP
healthy
sick
f(x)
FN
FP

- **Sensitivity** (true positive rate) is the probability of a positive test result, conditioned on the individual truly being positive.
- **Specificity** (true negative rate) is the probability of a negative test result, conditioned on the individual truly being negative.



$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

$$= \frac{\text{number of true positives}}{\text{total number of sick individuals in population}}$$

$$= \text{probability of a positive test given that the patient has the disease}$$

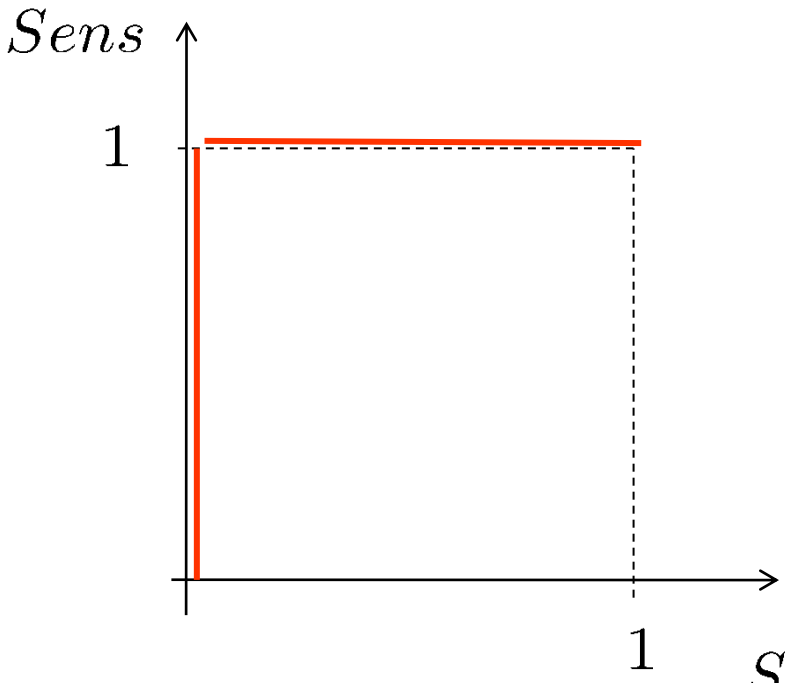
$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

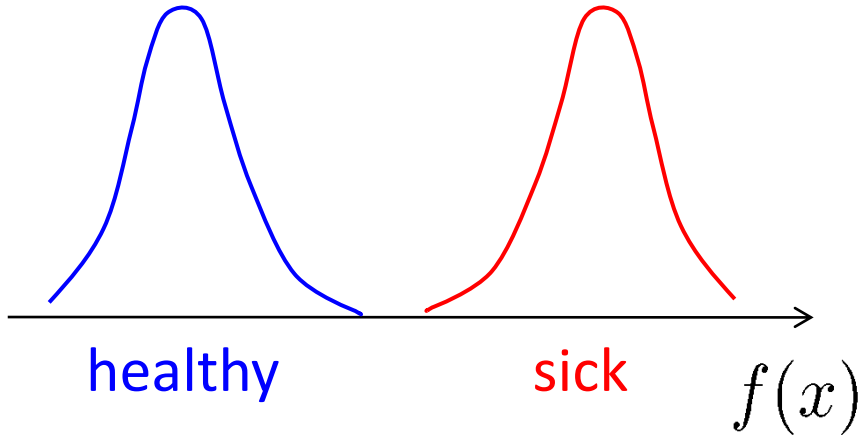
$$= \frac{\text{number of true negatives}}{\text{total number of well individuals in population}}$$

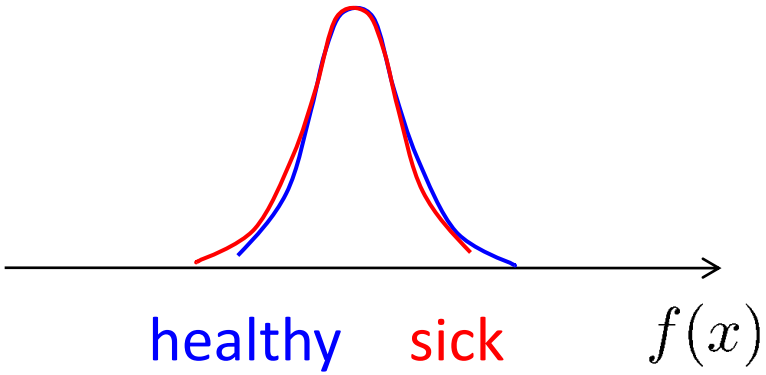
$$= \text{probability of a negative test given that the patient is well}$$

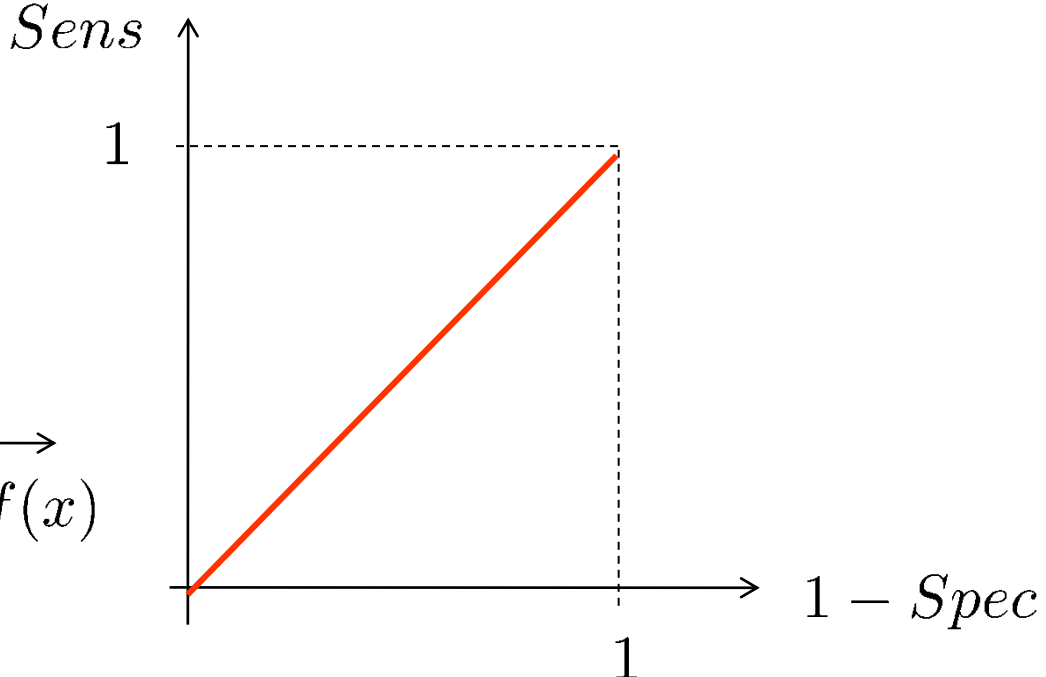
The relationship between sensitivity, specificity, and similar terms can be understood using the following table. Consider a group with **P** positive instances and **N** negative instances of some condition. The four outcomes can be formulated in a 2×2 *contingency table* or *confusion matrix*, as well as derivations of several metrics using the four outcomes, as follows:

		Predicted condition			
	Total population = P + N	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) = TPR + TNR − 1	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP),** **hit**	**False negative (FN),** **type II error, miss, underestimation**	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP),** **type I error, false alarm, overestimation**	**True negative (TN),** **correct rejection**	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P+N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN}$ $= 1 - FOR$	Markedness (MK), deltaP (Δp) = PPV + NPV − 1	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$


Sens
1
1
1 − Spec


healthy
sick
f(x)


healthy
sick
f(x)


Sens
1
1
1 − Spec