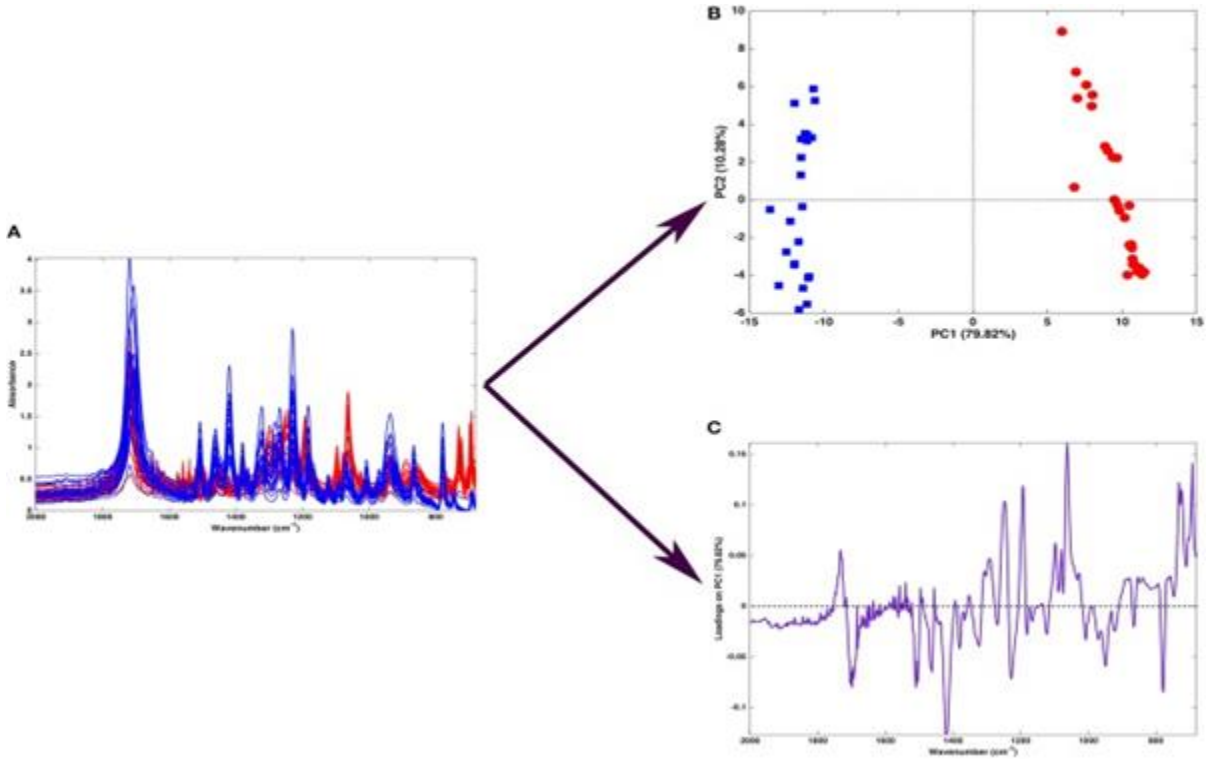
Chemometrics: Definition

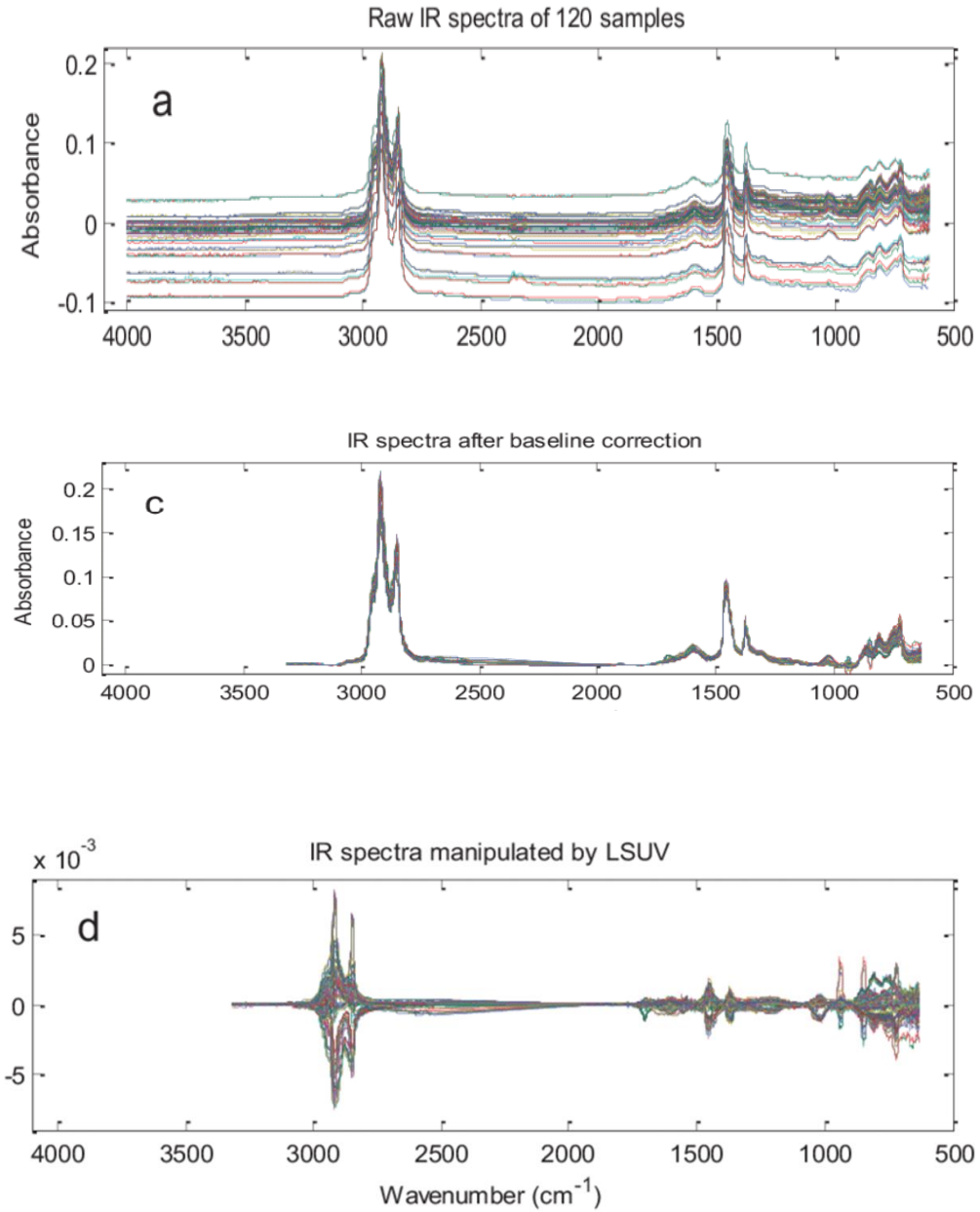
Chemometrics is the field of chemistry that deals with the application of mathematical and statistical methods to chemical data analysis. Chemometric techniques are used to extract useful information from complex chemical data sets and to develop models for understanding and predicting chemical behavior. These techniques can be applied to a wide range of chemical data, including spectroscopic, chromatographic, and mass spectrometric data, as well as data from other analytical techniques. Chemometrics is widely used in areas such as process control, quality assurance, environmental monitoring, and materials science.



Background for understanding chemometrics (1): The needle in the haystack

The expression "finding the needle in a haystack" means to locate or identify something that is difficult to find or hidden among a large number of similar things. This expression can be translated to chemometrics as finding the relevant information or signal from a large amount of data or noise. In chemometrics, the goal is to extract useful information from complex data sets, where there may be many variables measured on each sample, and the signal of interest may be obscured by noise or other irrelevant variables. By reducing the dimensionality of the data and focusing on the most important variables, chemometric methods can help to identify the relevant signal and extract useful information from complex data sets.

Background for understanding chemometrics (2): Deconvolution



Cleaning and deconvoluting complex matrices is important for several reasons:

Accuracy: **Complex matrices often contain multiple components or interference**s that can affect the accuracy of analytical measurements. By removing these interferences, the accuracy of the analysis can be improved.

Sensitivity: Removing interferences can increase the sensitivity of the analysis, allowing for the detection of lower concentrations of analytes.

Specificity: Deconvoluting complex matrices can help **identify and separate individual components**, allowing for more specific and targeted analysis.

Reproducibility: Cleaning and deconvoluting complex matrices can improve the reproducibility of analytical measurements, ensuring that results are consistent across multiple analyses.

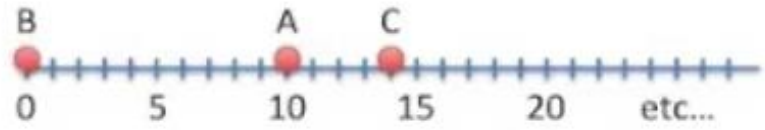
Dimensions

The following is an example of how to graphically represent dimensions. One dimension equals one line; two, a plane. Three will provide volume. However, starting at four, we run out of ways to represent them, since we live in a three-dimensional world. Therefore, dimensions must be reduced. When talking about dimension reduction in multivariate analyses, it actually mean reducing variables

1 dimension= One line. Example…

Gen	Cell 1
A	10
B	0
C	14

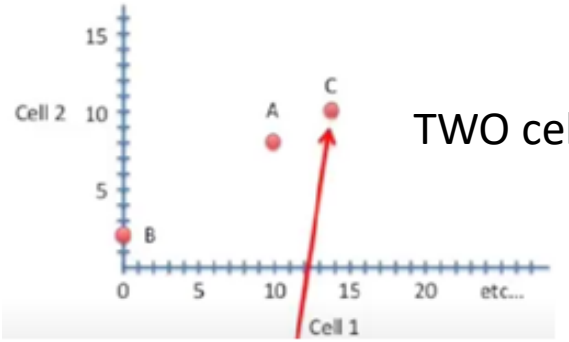
ONE cell



2 dimensions= Un plane. Example…

Gen	Cell 1	Cell 2
A	10	8
B	0	2
C	14	10

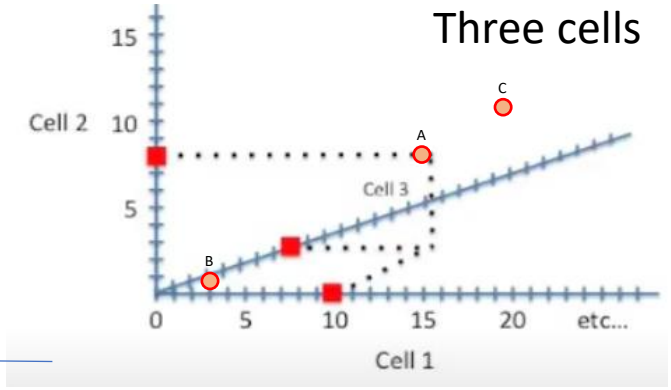
TWO cells



3 dimensions= One space. Example…

Gen	Cell 1	Cell 2	Cell 3
A	10	8	8
B	0	2	4
C	14	10	12

Three cells



What if we have 4 cells?... What if we have 87 cells?... **87 Dimensions? Impossible!**

Reducing dimensions. Importance

It is often necessary to reduce dimensions in data analysis because of the problem of "curse of dimensionality." As the number of dimensions (i.e., variables) increases, the amount of data required to accurately represent the data also increases exponentially. This can lead to issues with overfitting, increased computational complexity, and difficulty in interpreting the data. By reducing the number of dimensions, the data can be more easily analyzed, visualized, and interpreted. These are some reasons why it is important:

Improved computational efficiency: High-dimensional data requires more computational resources to process, analyze, and visualize. By reducing the dimensions, we can simplify the data and reduce the computational burden.

Improved accuracy: High-dimensional data often contains noise, redundancy, and irrelevant features that can negatively impact the accuracy of machine learning models. By reducing dimensions, we can eliminate these factors and improve the accuracy of our models.

Improved interpretability: High-dimensional data can be difficult to interpret and visualize. By reducing the dimensions, we can create more intuitive and understandable representations of the data.

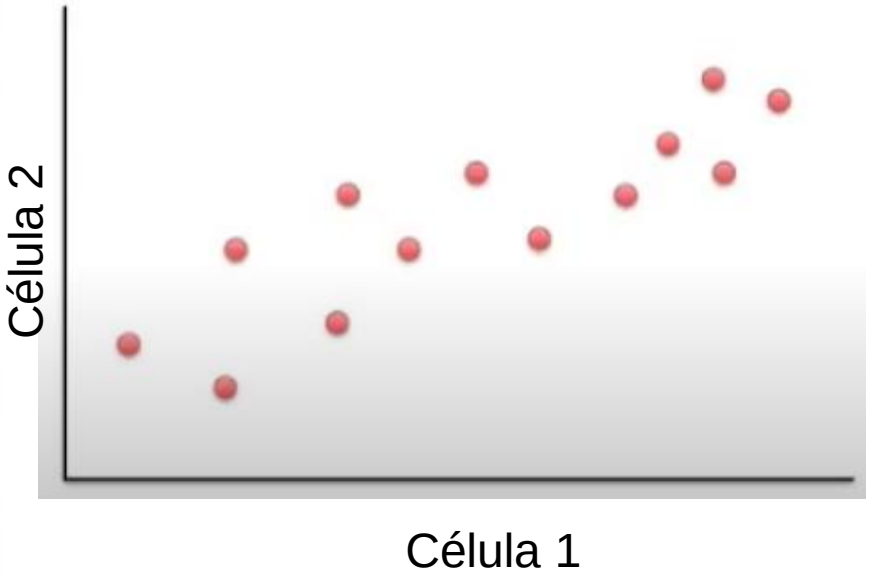
Improved generalization: High-dimensional data can lead to overfitting, where models perform well on the training data but poorly on new, unseen data. By reducing dimensions, we can reduce the risk of overfitting and improve the generalization of our models.

Reducing dimensions. Principal Component Analysis (PCA)

First point to consider: Some dimensions are more important than others (for example, that's why 3D TVs failed... Seeing everything on a flat screen is good enough for almost everyone). The key is to **identify which dimensions present the greatest variation in the system as a whole.**

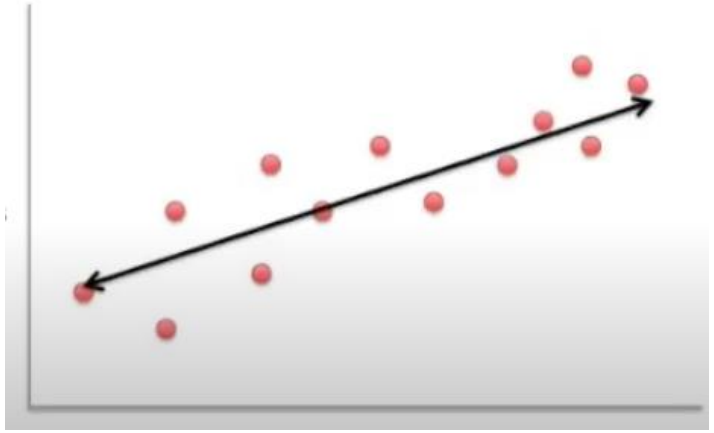
Second point to consider: Therefore, we should **focus on maximizing the differences** between the data among different elements of the system. In other words, we should seek diversity rather than homogenization.

Gene	Cell1 reads	Cell2 reads
a	10	8
b	0	2
c	14	10
d	33	45
e	50	42
f	80	72
g	95	90
h	44	50
i	60	50
... (etc)	... (etc)	... (etc)

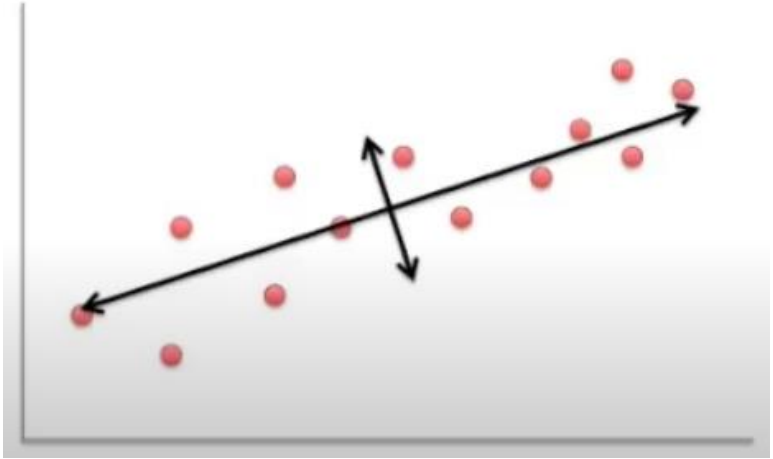


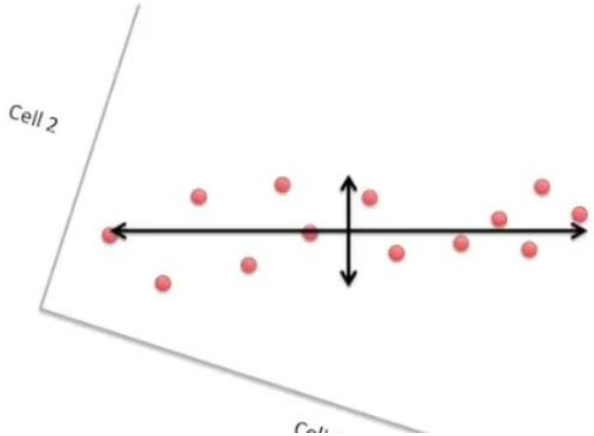
Reducing dimensions. Principal Component Analysis (PCA)

1. The points are arranged around a diagonal line, between whose vertices there is the maximum variation.



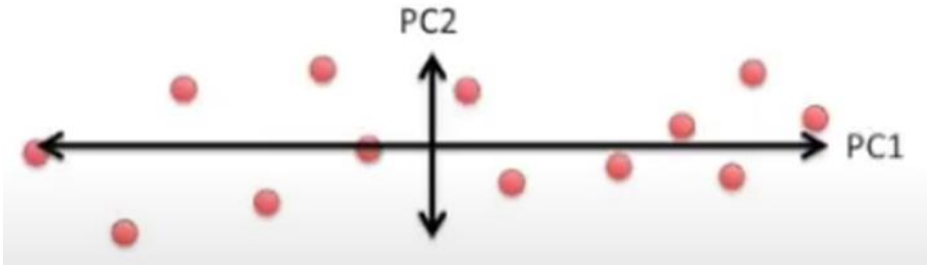
2. The points are also distributed above and below this line, with maximum variation also at the limits of said line.





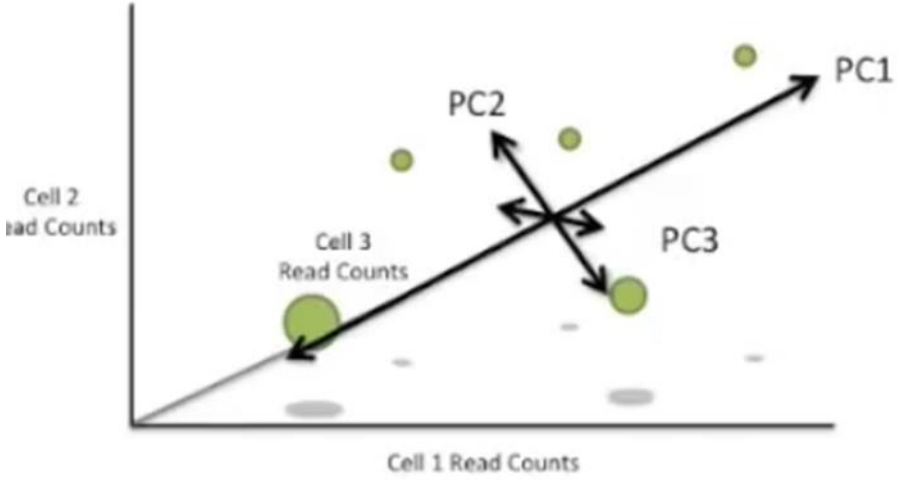
3. If we rotate the graph so that these two lines where the maximum variation occurs become the new axes, we will have this new graph.

Reducing dimensions. Principal Component Analysis (PCA)



These two new axes of this rotated graph, which represent the maximum variability among the sample points, are called "principal components" (PC for Principal Component).

What if we have 3 cells? (However, this is also relevant for 4,5, 6, 87 …. n cells



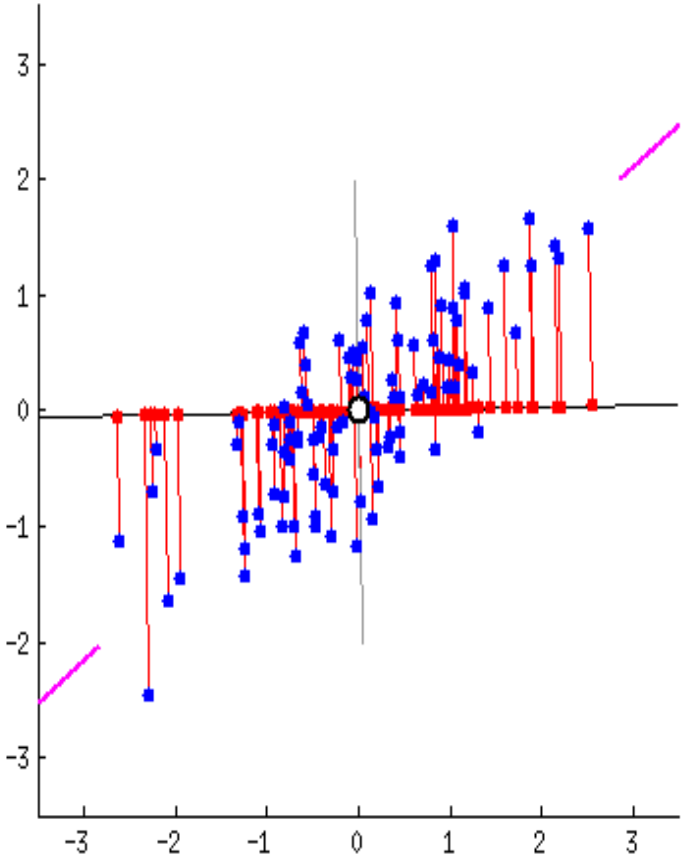
- We will have PCn where n is the number of cells (or values).
- It doesn't matter how many dimensions there are because it is calculated based on the formula.
- PC1 is the axis with the highest variation, PC2 is the next highest variation, PC3 is the next... PC87 is the axis with the 87th highest variation... PCn is the axis with the lowest variation.

Reducing dimensions. Principal Component Analysis (PCA)

How is the calculation of PC represented?

- The information is arranged around the arithmetic mean of the total data.
- The axis is shifted to this arithmetic mean.
- The line that passes through this new origin is sought where the sum of the squares of the distances from the points to their projection on that line is **minimized (PC1**).
- The line that passes through this new origin is sought where the sum of the squares of the distances from the points to their projection on that line is **maximized (PC2).**

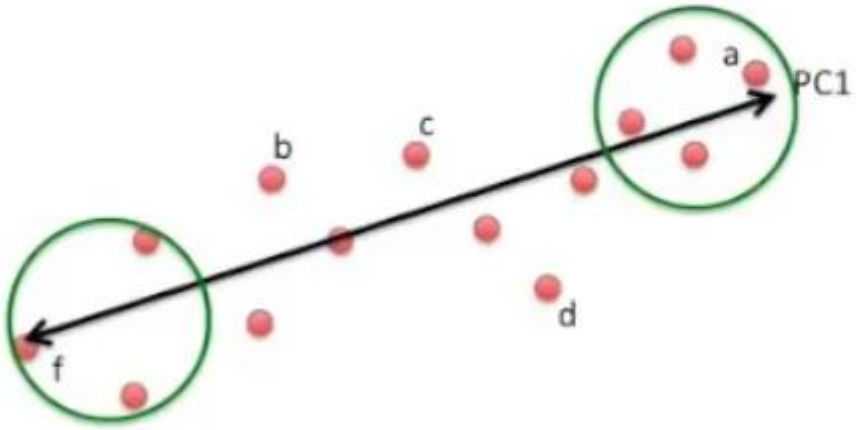
Eigenvalues: The sum of the square of the distances from each dot to the corresponding PC. They are a set of scalars associated with a linear system of equations or a matrix. In the context of chemometrics, eigenvalues are used to quantify the variation in the data captured by principal components. For example, in principal component analysis (PCA), eigenvalues represent the variance of each principal component, indicating how much of the total variation in the data is explained by each component. The eigenvalues are calculated by solving the characteristic equation of the matrix, and the resulting eigenvectors provide the direction and magnitude of the principal components.



PCA. What can we do with it?

1. We can check the contribution that each element brings to a given PC

The length and direction of PC1 is mostly determined by the circled genes.



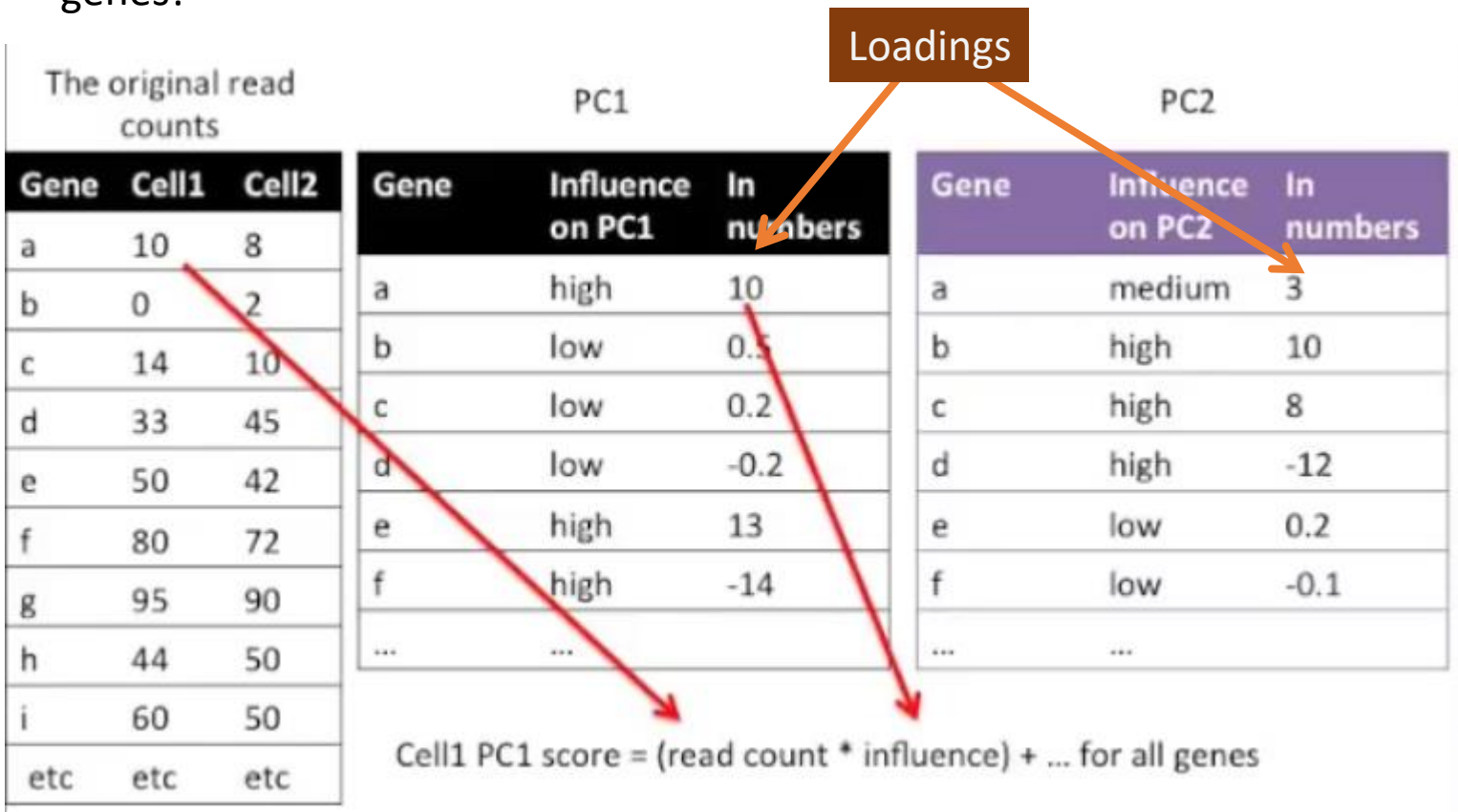
Some genes have more influence on PC1 than others.

Gene	Influence on PC1	In numbers
a	high	10
b	low	0.5
c	low	0.2
d	low	-0.2
e	high	13
f	high	-14
...	...	

Genes with little influence on PC1 get values close to zero, and genes with more influence get numbers further from zero.

PCA. What can we do with it?

2. If we consider the influence that each element (in this case genes) has on a main component, and **multiply that value by the number of times that element is manifested in a variable** (in this case cell), we can, in our example... Plot cells, not genes!



Eigenvector: An arrangement of loadings

In chemometrics, "**loadings**" are a set of coefficients that show the contribution of each variable (or predictor) in a multivariate model. They represent the correlation between the original variables and the principal components (PCs) extracted from the data. Loadings are calculated during principal component analysis (PCA) or partial least squares (PLS) regression, which are common methods in chemometrics. Loadings can be used to reduce the number of variables in a dataset by identifying the most important variables that contribute the most to the variance of the data. By selecting the variables with the highest loadings, a subset of variables can be chosen that still captures the majority of the variance in the original dataset. This process is known as "variable selection" or "variable reduction" and can help simplify the model and improve its interpretability. It can also help to reduce the risk of overfitting the model and improve its predictive performance.

PCA

Principal component analysis (PCA) is an important technique in data analysis and machine learning. It is used to reduce the dimensionality of large data sets by identifying the most important variables or features that explain the majority of the variance in the data. **Being a supervised algorithm, it aims to MAXIMIZING DISPERSON among datasets**

The importance of PCA can be summarized as follows:

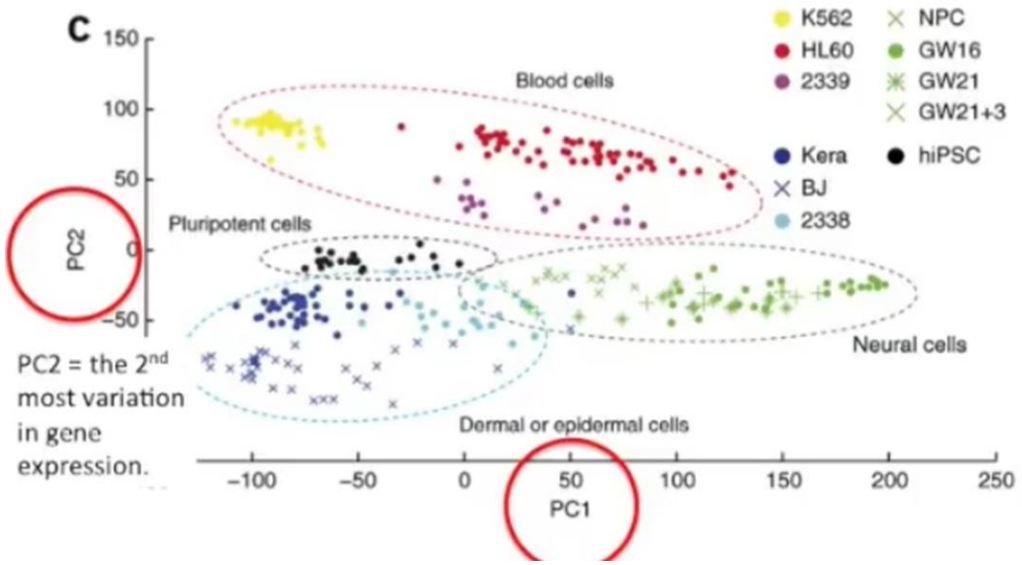
Dimensionality reduction: PCA helps to reduce the number of variables in a data set while retaining the most important information. This makes it easier to analyze and visualize complex data sets.

Feature selection: PCA helps to identify the most important features or variables that contribute to the variability in the data. This can be useful in feature selection for machine learning models.

Data visualization: PCA can be used to visualize high-dimensional data in two or three dimensions, making it easier to interpret and understand.

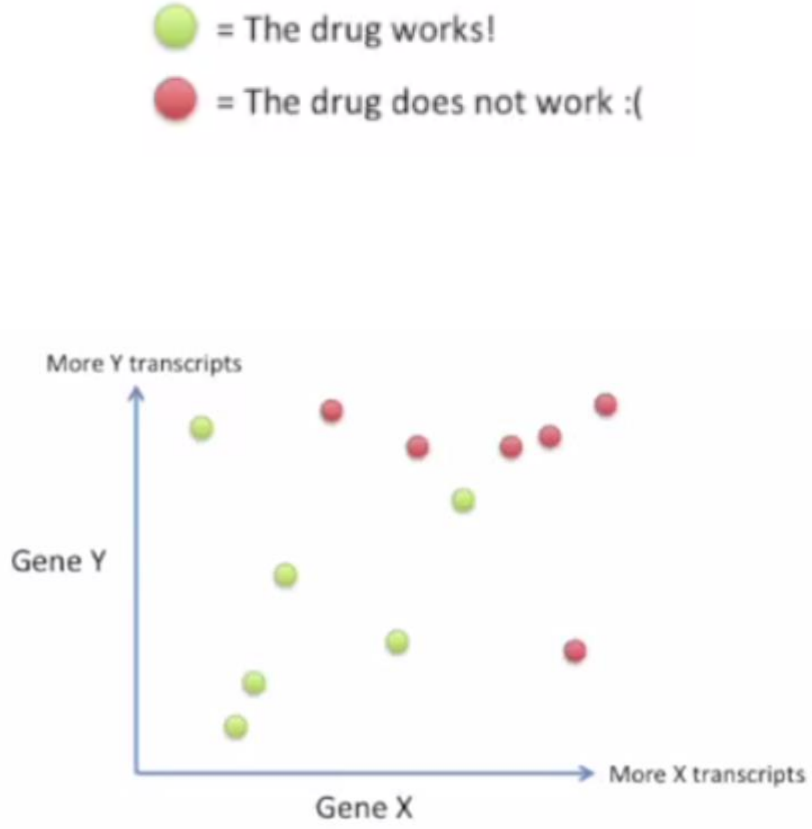
Noise reduction: PCA can filter out noise and irrelevant variables, which can improve the accuracy of machine learning models.

Clustering: PCA can be used to cluster similar data points together, which can be useful in identifying patterns and trends in the data.



Supervised algorithm

In chemometrics, a supervised algorithm is a type of machine learning algorithm that requires a labeled dataset to train a model. This means that the dataset is already classified or labeled with known outcomes, and the algorithm uses this information to learn how to classify new data. Supervised algorithms are commonly used in chemometrics for tasks such as classification, regression, and prediction. Examples of supervised algorithms in chemometrics include linear regression, logistic regression, support vector machines, and random forests.

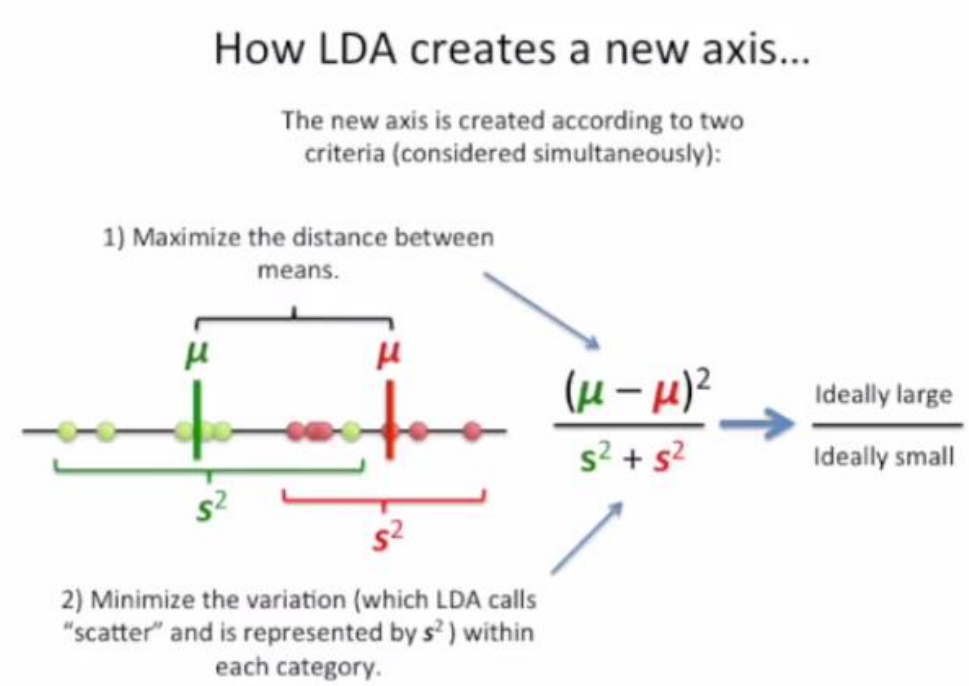


Linear Discriminant Analysis (LDA)

Two criteria are considered for LDA:

Maximizing the distance between the means of both groups of previously labeled data (μ of each data set)

Minimizing the dispersion between the different points of each previously labeled data set.



Supervised vs Unsupervised algorithms

A supervised algorithm is a machine learning algorithm that is trained on a labeled dataset where the outcome variable is known. In a supervised learning approach, the algorithm uses the labeled data to learn the relationship between the input variables (also known as features or predictors) and the outcome variable. Once the algorithm has learned this relationship, it can be used to predict the outcome variable for new data points based on their input variables. Supervised algorithms differ from unsupervised algorithms in that unsupervised algorithms do not require labeled data. Instead, unsupervised algorithms try to find patterns or relationships within the data itself. This can be useful for tasks such as clustering or dimensionality reduction. However, unsupervised algorithms cannot be used for prediction tasks because they do not have a known outcome variable to train on.