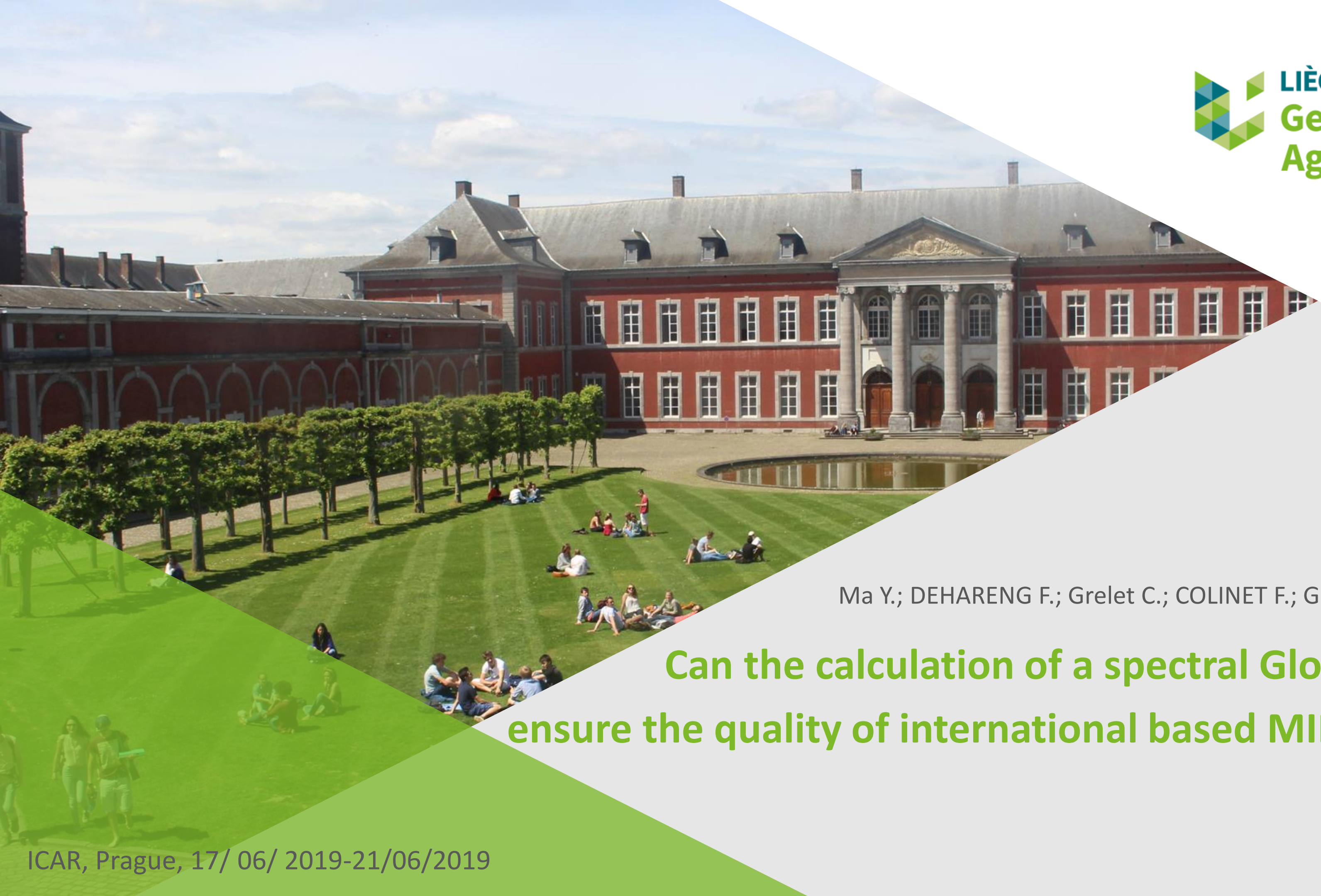


LIÈGE université
Gembloux
Agro-Bio Tech

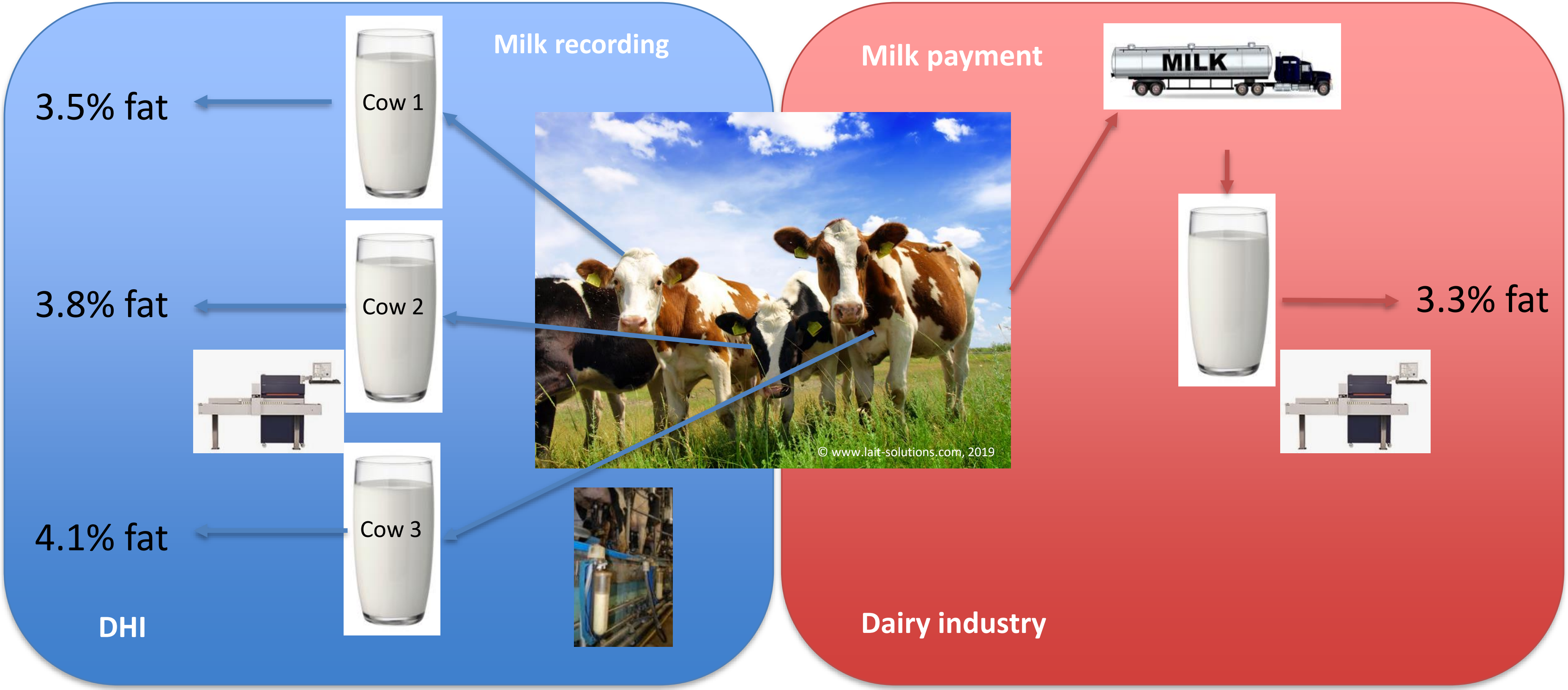
Lei ZHANG

Ma Y.; DEHARENG F.; Grelet C.; COLINET F.; GENGLER N.; SOYEURT H.

Can the calculation of a spectral Global H distance ensure the quality of international based MIR predictions?

ICAR, Prague, 17/ 06/ 2019-21/06/2019

Milk Recording Scheme



Milk recording

3.5% fat

Cow 1

3.8% fat

Cow 2

4.1% fat

Cow 3

DHI

Milk payment

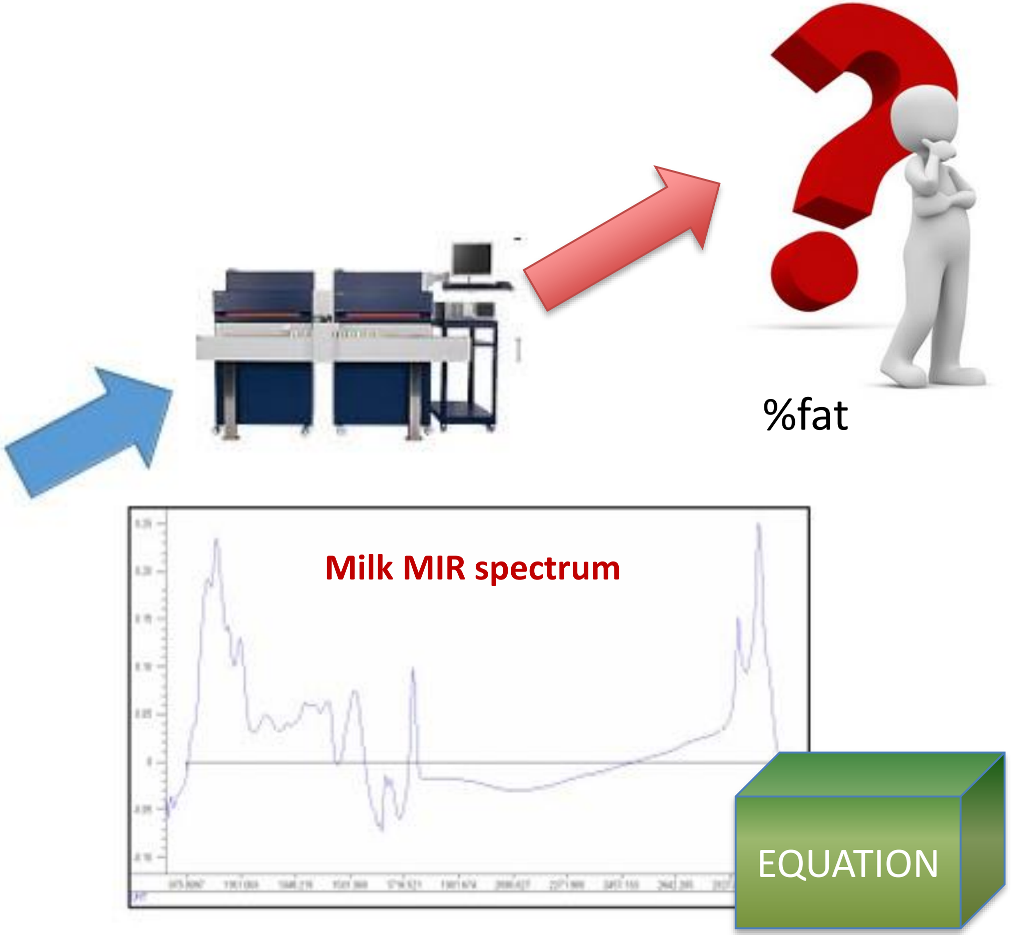
3.3% fat

Dairy industry

© www.lait-solutions.com, 2019

Milk Analysis

Milk recording
(About 1 month for each cow)



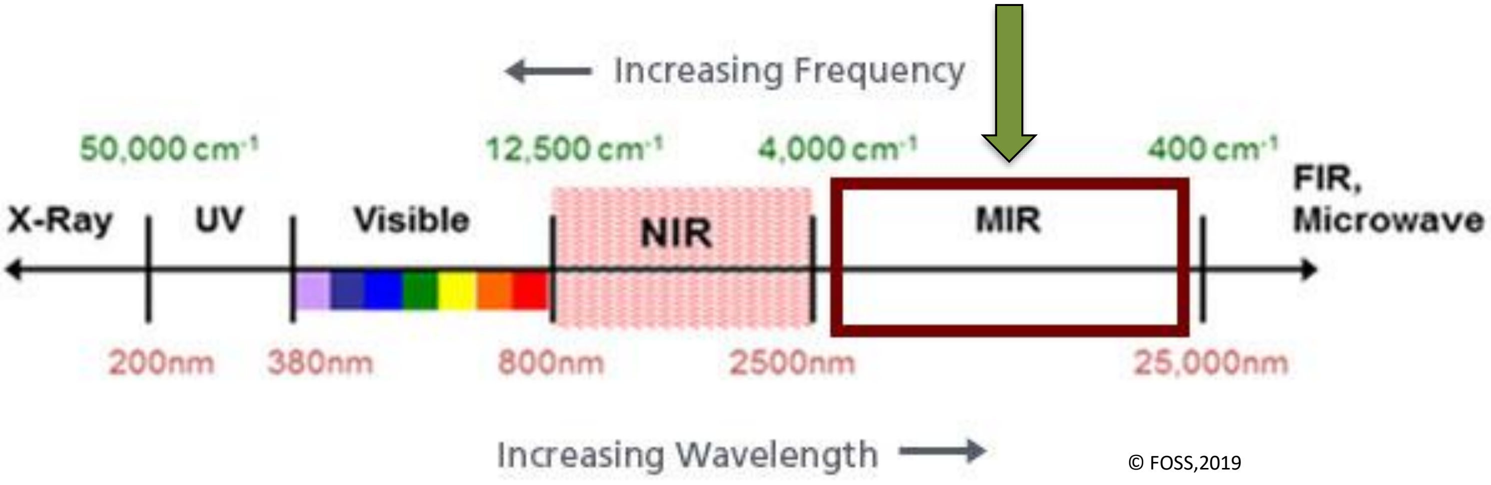
%fat

Milk MIR spectrum

EQUATION

What is Mid-infrared spectrum?

- Approximately 2,500-25,000nm (4,000-400 cm^{-1})



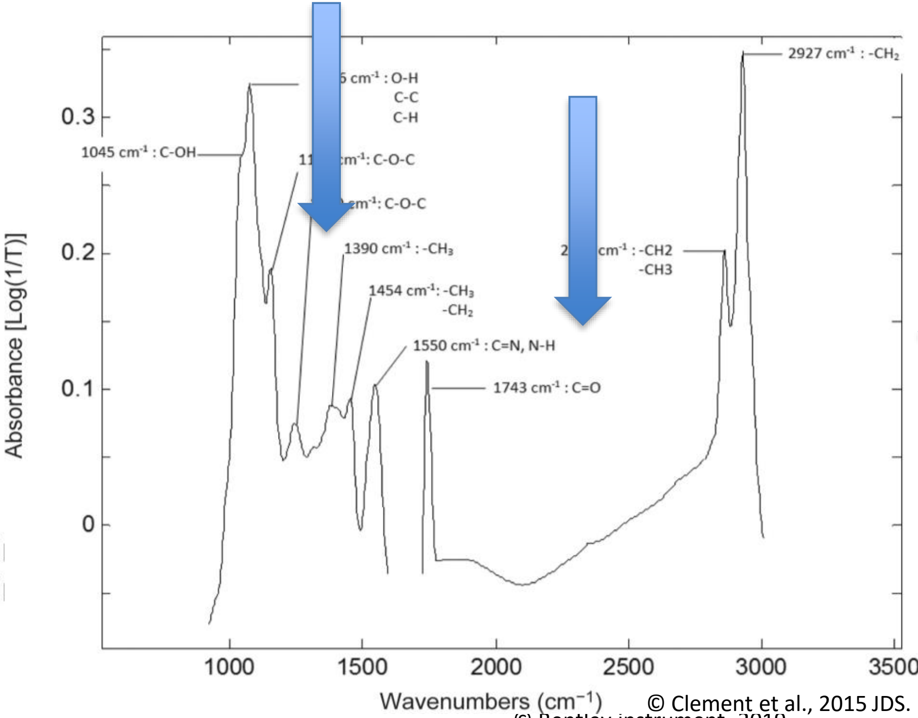
© FOSS,2019

Frequency = 1 / wavelength

Principle of MIR spectrometry

I

I



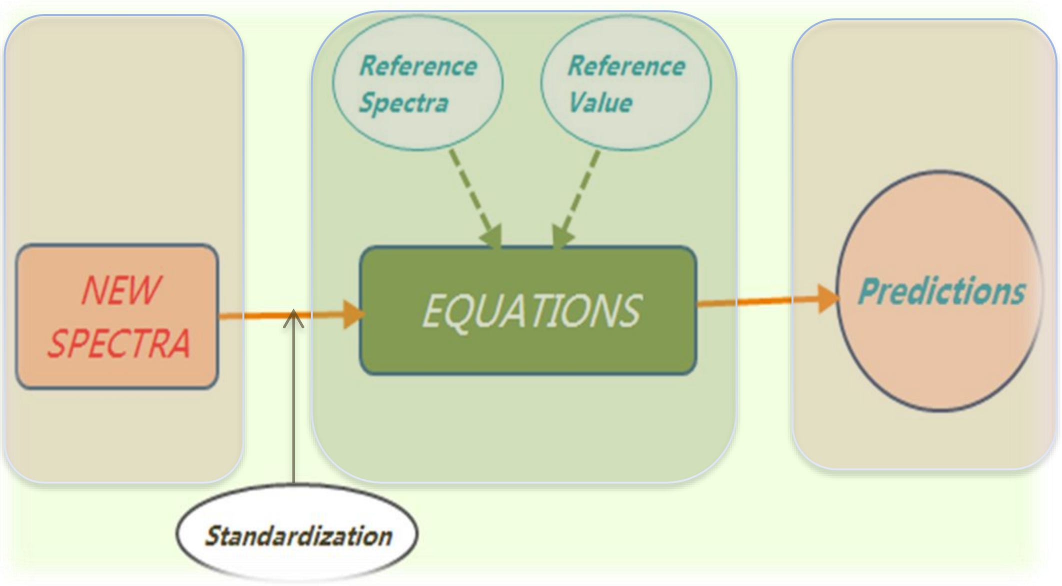
bance

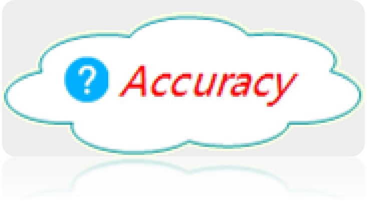
rum

© Clement et al., 2015 JDS.

© Bentley instrument, 2019

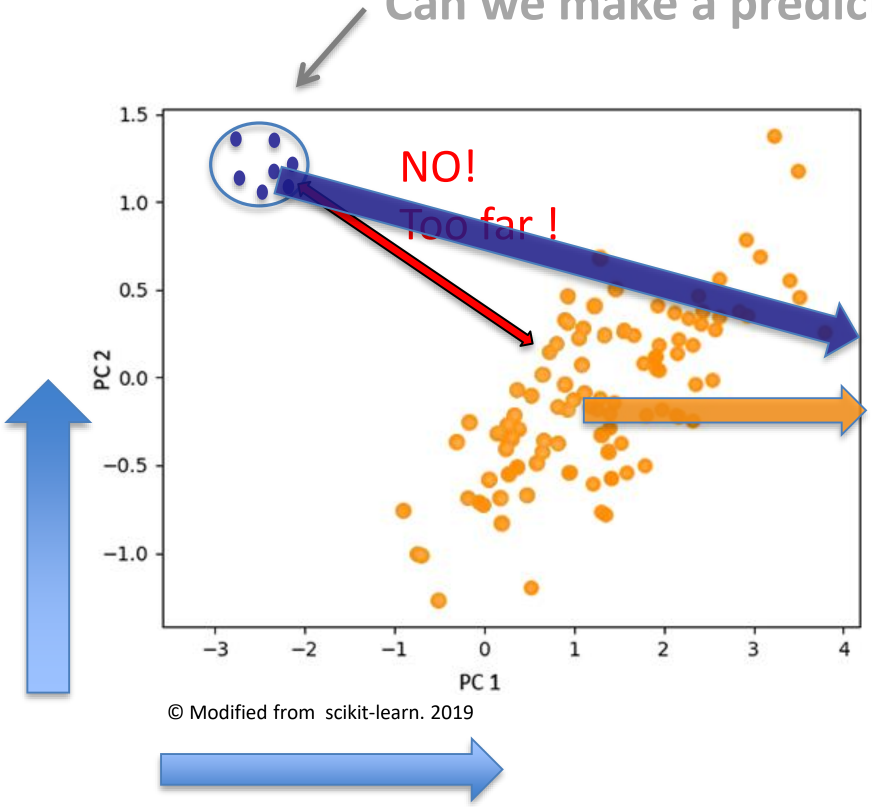
How can we make a prediction ?



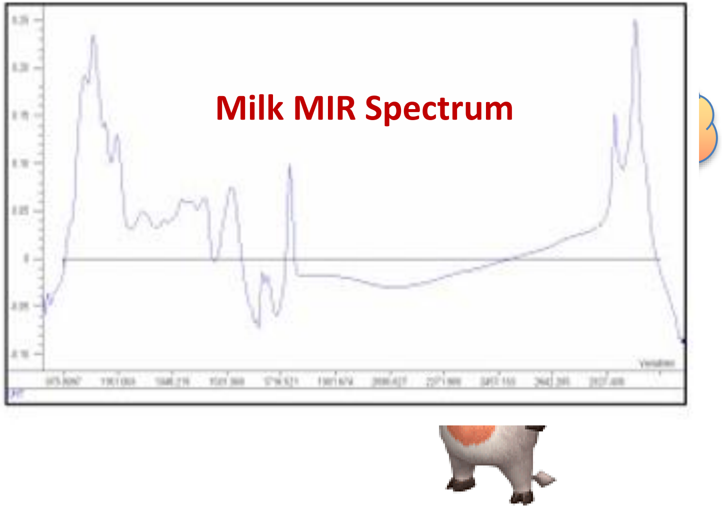


Can we make a prediction for all spectra ?

Can we make a prediction from those spectra?

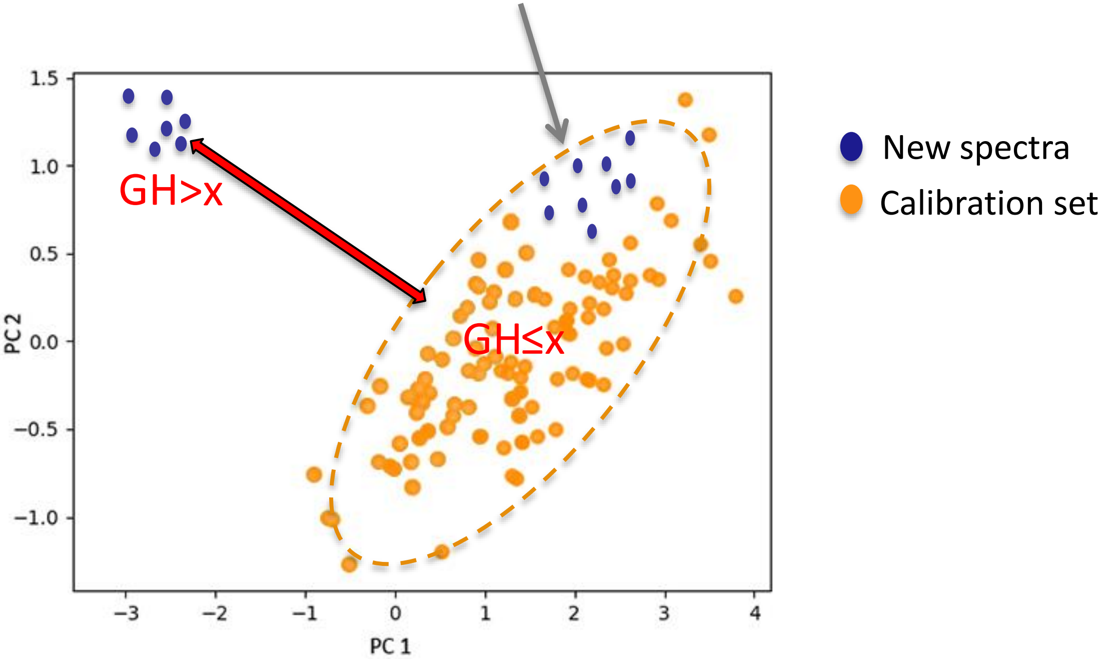


© Modified from scikit-learn. 2019



Can we make a prediction for all spectra ?

Can I make a prediction from those spectra?



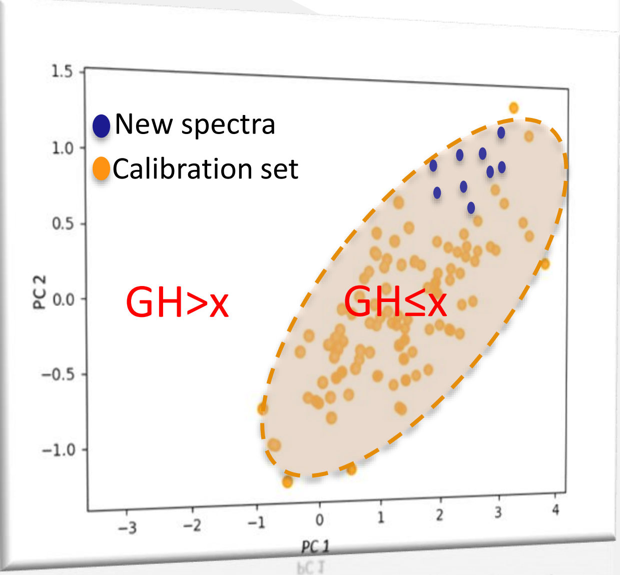
© Modified from scikit-learn. 2019

© Wikipedia
P.C. Mahalaobis
(1893 – 1972)

Mahalanobis Distance:

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}$$

Where： $\vec{x}$ is PC scores of one spectrum;
$\vec{\mu}$ is the mean of PC scores of spectra in the calibration set;
S is covariance matrix between PC scores of the calibration spectra



© Modified from scikit-learn. 2019

Global H which is the Standardized Mahalanobis Distance

GH=$\frac{DM}{nPCs}$

LIÈGE université
Gembloux
Agro-Bio Tech

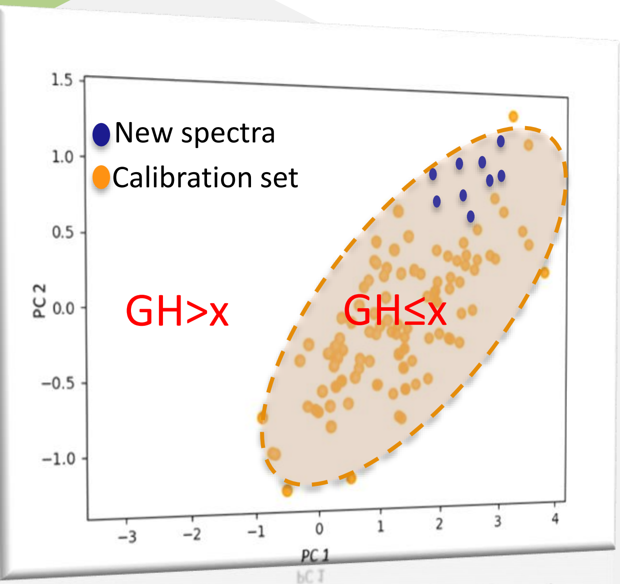
Mahalanobis Distance:

$$D_M(\vec{x}) = \sqrt{(\vec{x}-\vec{\mu})^T S^{-1}(\vec{x}-\vec{\mu})}$$

GH: Global H which is the Standardized Mahalanobis Distance

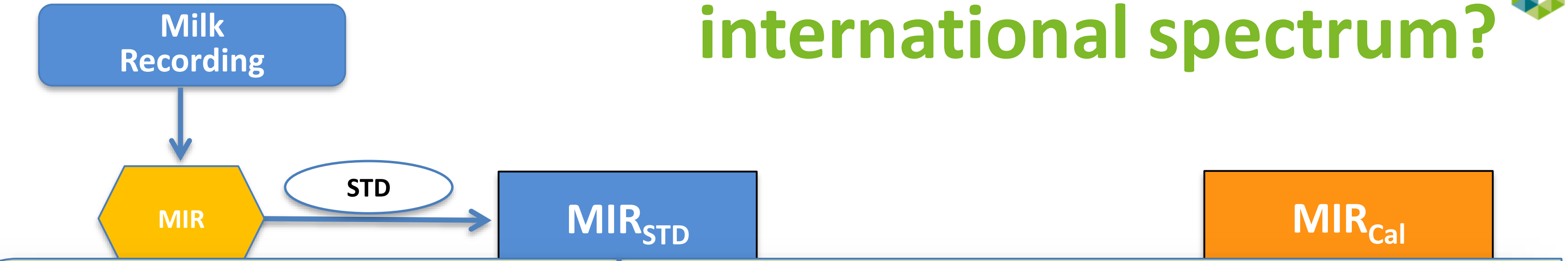
$$\text{GH}=\frac{DM}{nPCs}$$

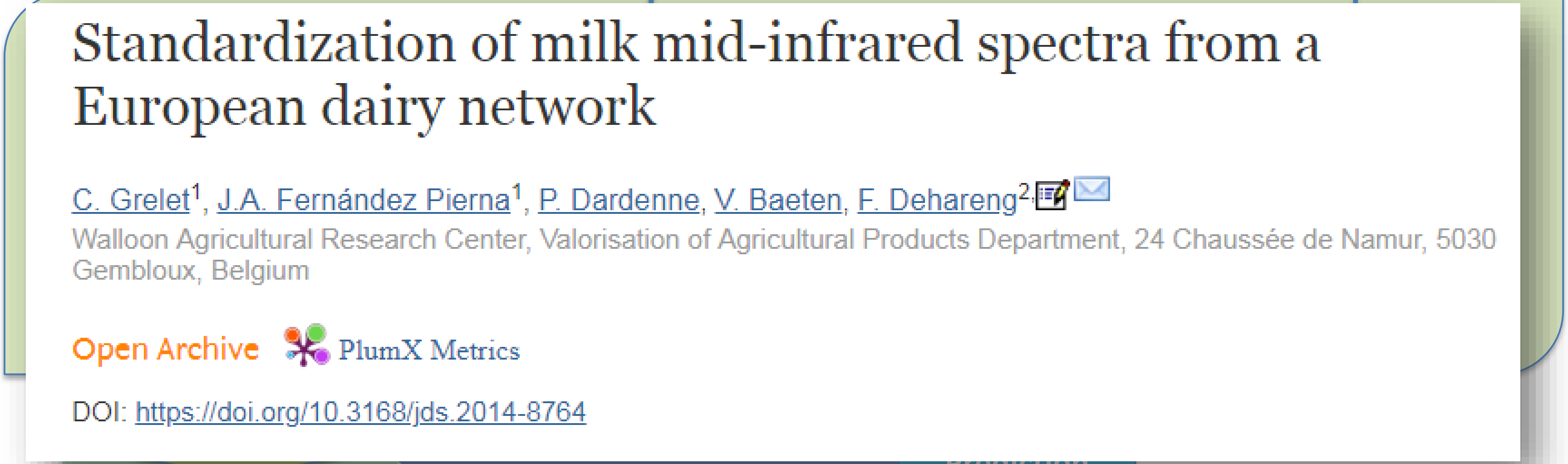
Where:
DM is the distance calculated from the fomular;
nPCs is the number of the principal components from PCA



© Modified from scikit-learn. 2019

What is the accuracy of prediction of international spectrum?





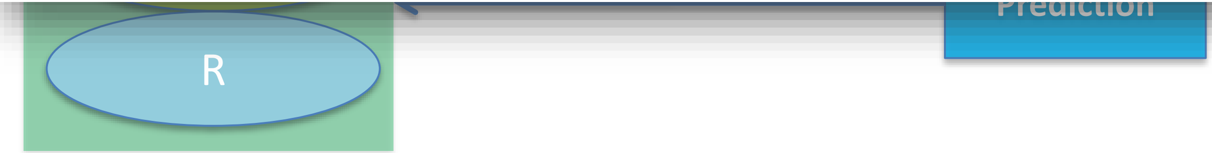
Standardization of milk mid-infrared spectra from a European dairy network

C. Grelet[1], J.A. Fernández Pierna[1], P. Dardenne, V. Baeten, F. Dehareng[2]

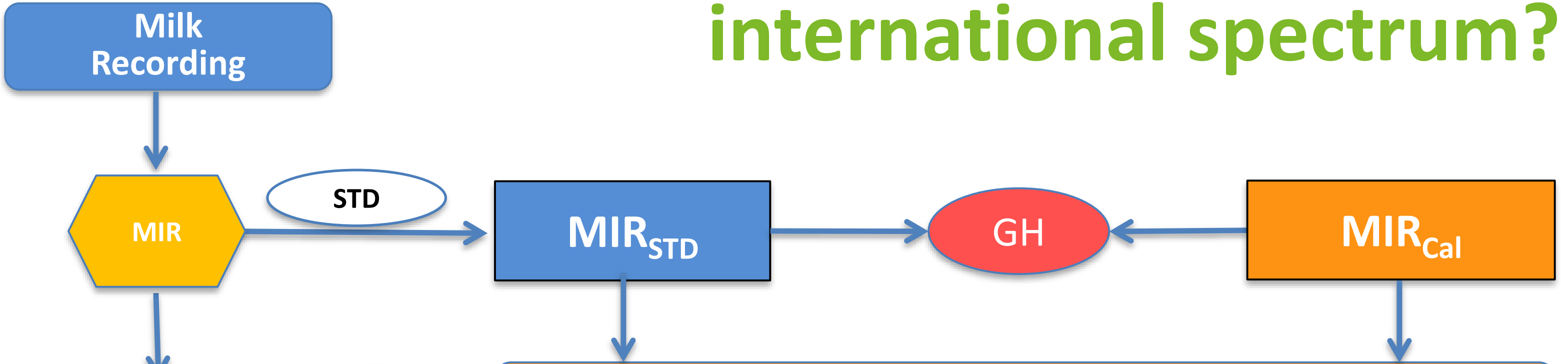
Walloon Agricultural Research Center, Valorisation of Agricultural Products Department, 24 Chaussée de Namur, 5030 Gembloux, Belgium

Open Archive PlumX Metrics

DOI: https://doi.org/10.3168/jds.2014-8764



What is the accuracy of prediction of international spectrum?



Moderat
GH limit

R

GH	Mean	SD	Minimum	Maximum
	1.93	2.43	0.00	475.00

GH ≤3

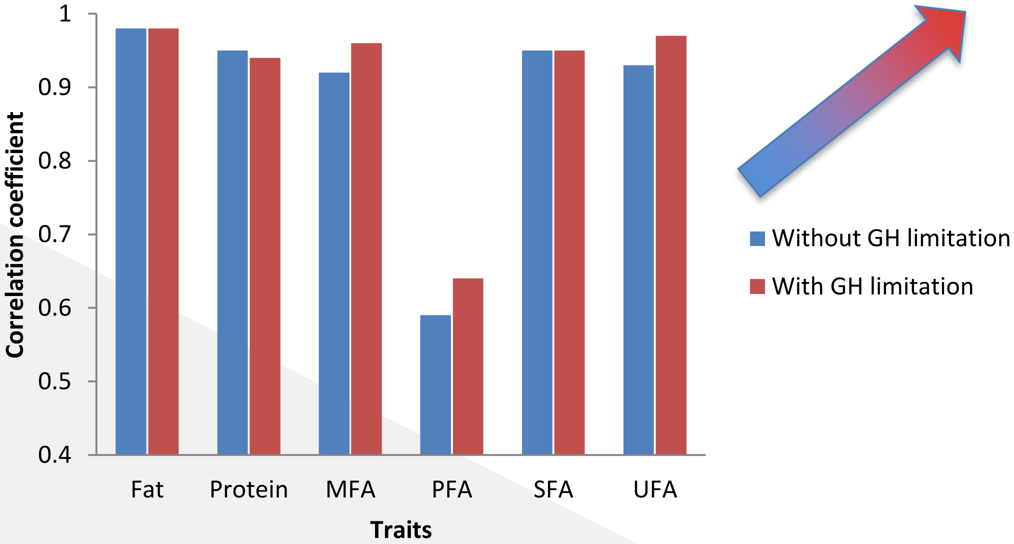
N records	GH	198,394	198,394	198,394	198,394	198,394	198,394
	With GH	172,547	174,062	159,651	159,509	174,825	159,467
Percent	%	*13.03*	*12.26*	*19.53*	*19.60*	*11.88*	*19.62*

Descriptive statistics

Table 1. Descriptive statistics of predicted value

Traits g/dL	Reference value		Predicted value		Predicted value (GH <= 3)	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Fat	3.97	0.95	3.99	0.95	3.90	0.86
Protein	3.43	0.40	3.53	0.46	3.52	0.39
MFA	0.86	0.27	1.15	0.36		0.31
PFA	0.07	0.04	0.15	0.05	0.15	0.04
SFA	2.62	0.67	2.64	0.68	2.59	0.62
UFA	0.93	0.31	1.29	0.39	1.25	0.34

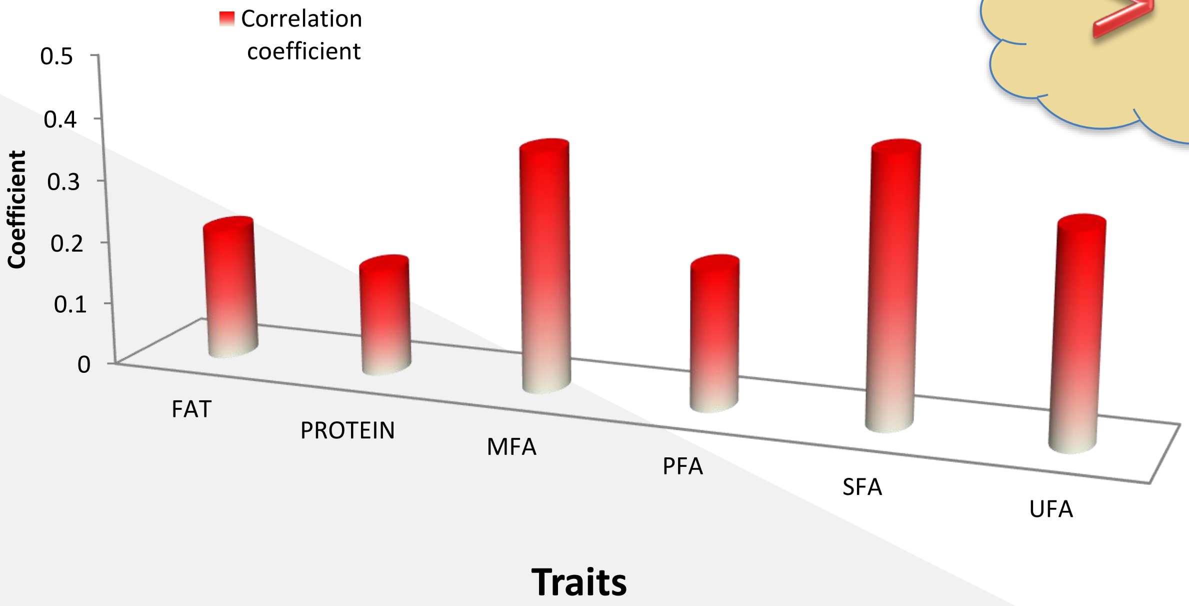
The correlation coefficient



Why?

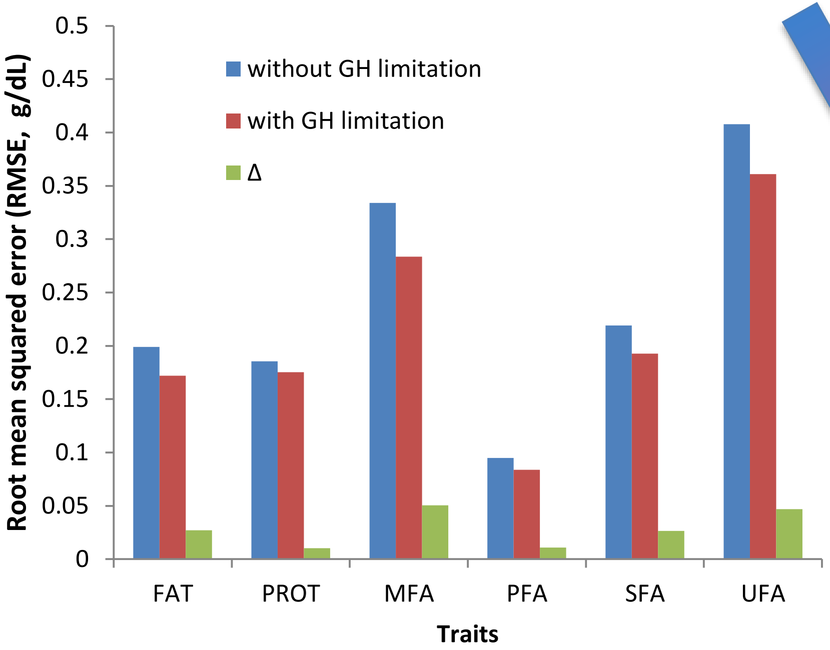
LIÈGE université
Gembloux
Agro-Bio Tech

Squared residual and GH



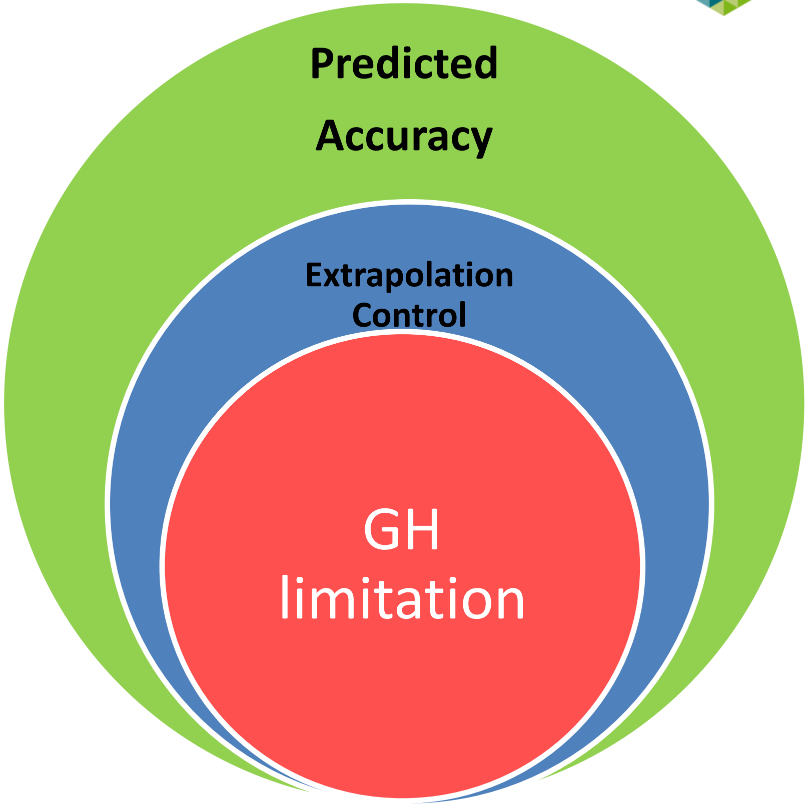
> 0

GH limitation decreased RMSE for most traits



Conclusion:

- GH limitation helps to **ensure the quality** of the MIR predictions
- It allows **avoiding spectral extrapolation**
- More work needed to be done to get **more accurate** predictions…



LIÈGE université
Gembloux
Agro-Bio Tech

Thanks for your attention!

Email: lei.zhang@doct.uliege.be

LIÈGE université
Gembloux
Agro-Bio Tech

Ma Y.; DEHARENG F.; Grelet C.; COLINET F.; GENGLER N.; SOYEURT H.,&

Lei ZHANG

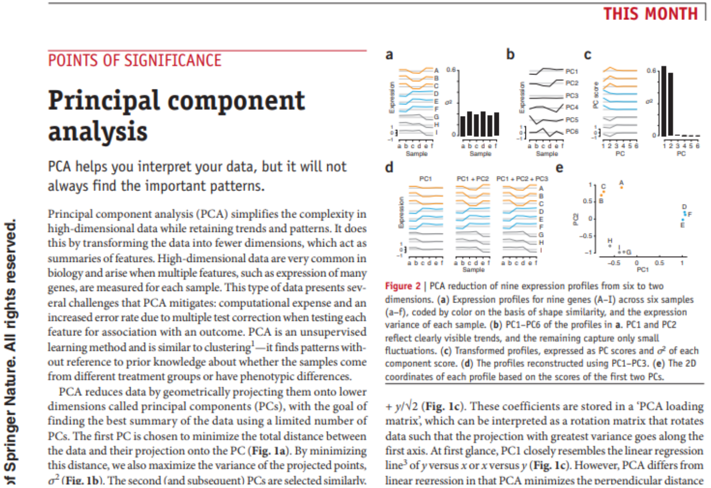
Email: lei.zhang@doct.uliege.be

ICAR, Prague, 17/ 06/ 2019-21/06/2019

Additional information:

Why do PCA?

- To decrease the dimensionality of the raw data
- To make it easy for calculating the inverse of the covariance matrix



THIS MONTH

POINTS OF SIGNIFICANCE

Principal component analysis

PCA helps you interpret your data, but it will not always find the important patterns.

Principal component analysis (PCA) simplifies the complexity in high-dimensional data while retaining trends and patterns. It does this by transforming the data into fewer dimensions, which act as summaries of features. High-dimensional data are very common in biology and arise when multiple features, such as expression of many genes, are measured for each sample. This type of data presents several challenges that PCA mitigates: computational expense and an increased error rate due to multiple test correction when testing each feature for association with an outcome. PCA is an unsupervised learning method and is similar to clustering[1]—it finds patterns without reference to prior knowledge about whether the samples come from different treatment groups or have phenotypic differences.

PCA reduces data by geometrically projecting them onto lower dimensions called principal components (PCs), with the goal of finding the best summary of the data using a limited number of PCs. The first PC is chosen to minimize the total distance between the data and their projection onto the PC (**Fig. 1a**). By minimizing this distance, we also maximize the variance of the projected points, σ^2 (**Fig. 1b**). The second (and subsequent) PCs are selected similarly.

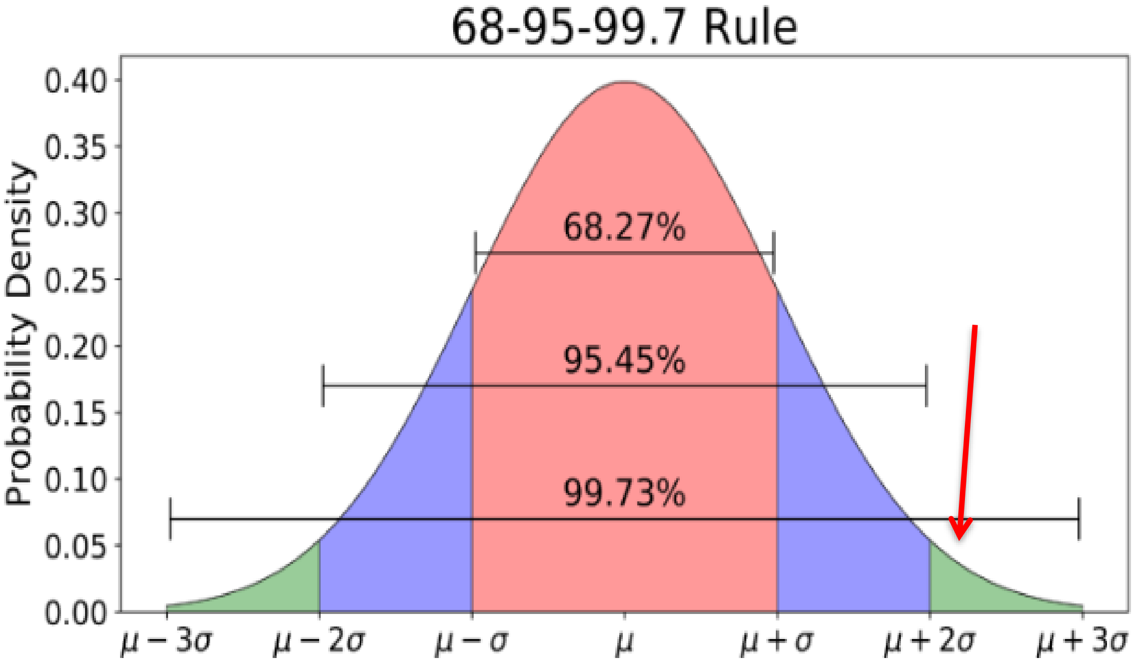
Figure 2 | PCA reduction of nine expression profiles from six to two dimensions. (**a**) Expression profiles for nine genes (A–I) across six samples (a–f), coded by color on the basis of shape similarity, and the expression variance of each sample. (**b**) PC1–PC6 of the profiles in **a**. PC1 and PC2 reflect clearly visible trends, and the remaining capture only small fluctuations. (**c**) Transformed profiles, expressed as PC scores and σ^2 of each component score. (**d**) The profiles reconstructed using PC1–PC3. (**e**) The 2D coordinates of each profile based on the scores of the first two PCs.

$+ y/\sqrt{2}$ (**Fig. 1c**). These coefficients are stored in a 'PCA loading matrix', which can be interpreted as a rotation matrix that rotates data such that the projection with greatest variance goes along the first axis. At first glance, PC1 closely resembles the linear regression line[3] of y versus x or x versus y (**Fig. 1c**). However, PCA differs from linear regression in that PCA minimizes the perpendicular distance

Lever et al., 2017 Nature Method 2017

Additional information:

Why GH ≤ 3?


68-95-99.7 Rule
Probability Density
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$