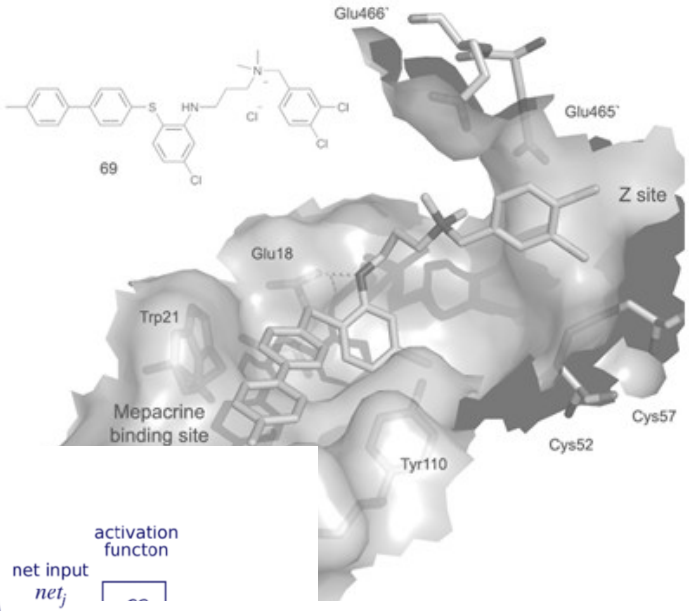
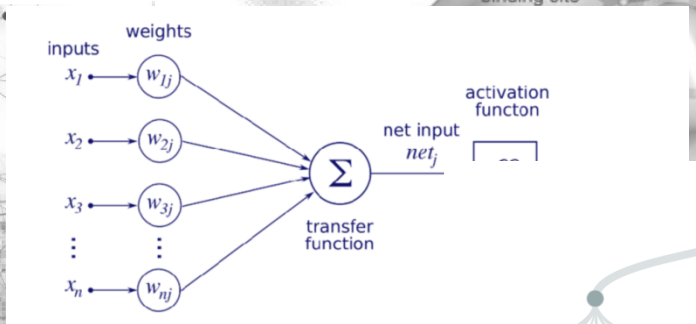
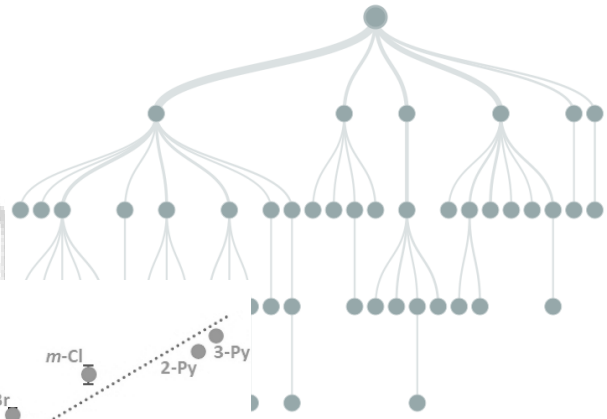
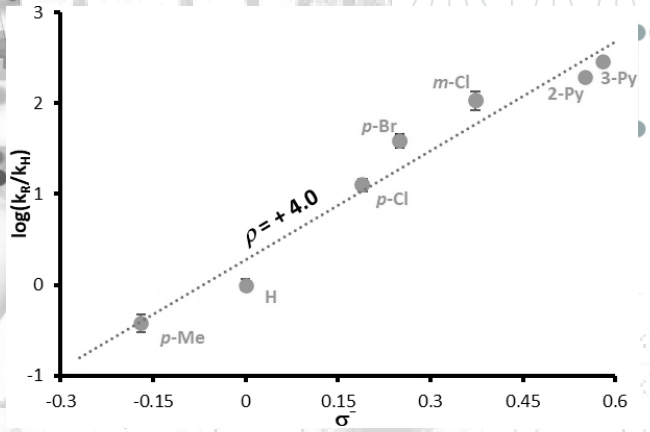
Introduzione ai metodi di Intelligenza Artificiale...

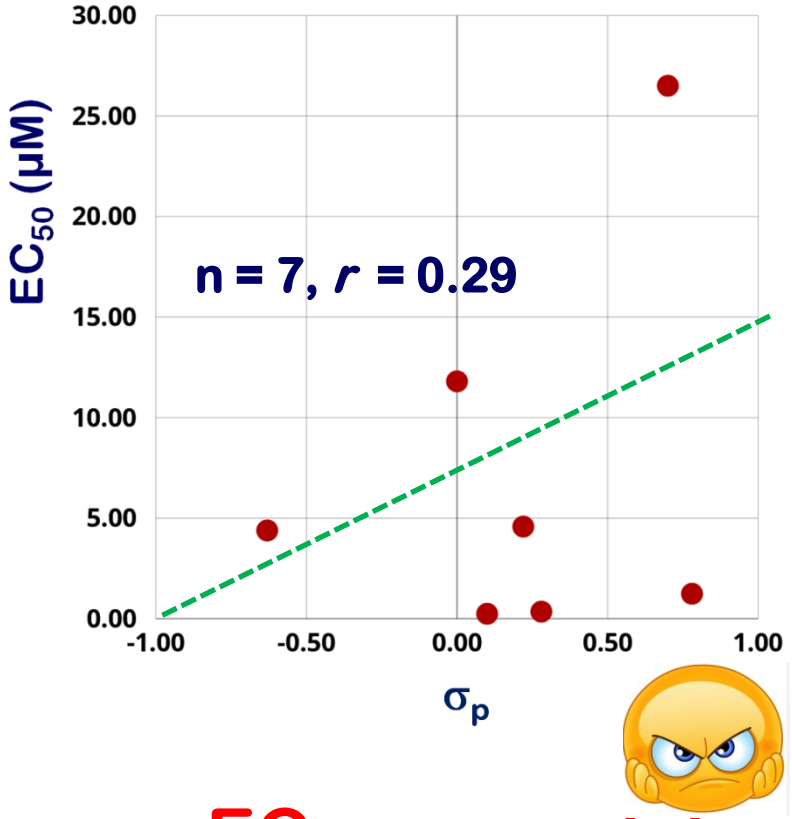


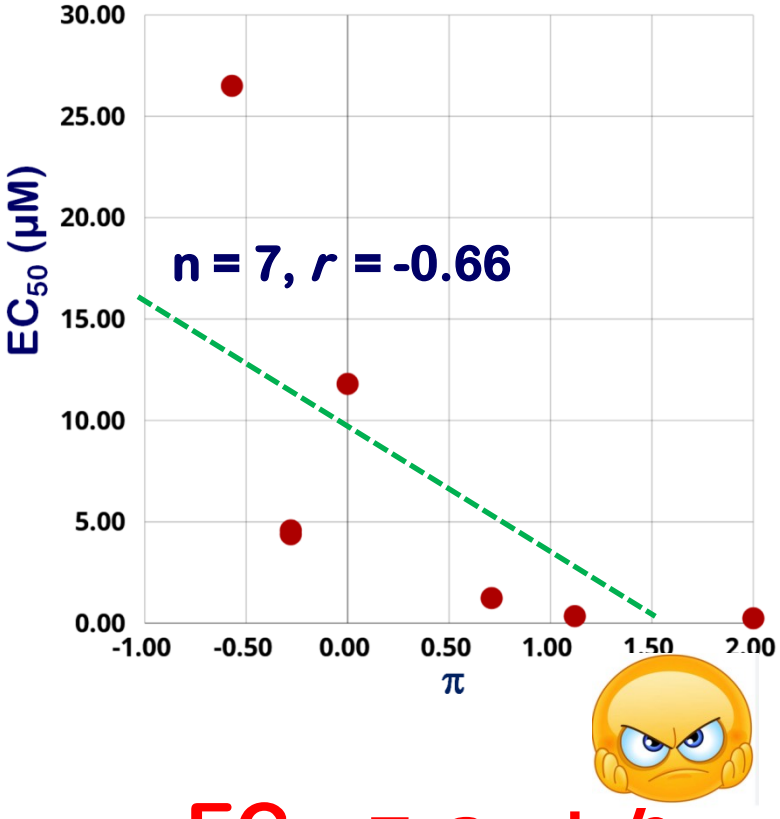






(Q)SAR: a strong and understandable temptation…





$EC_{50} = a\,\sigma_p + b$

$EC_{50} = 5.79\,\sigma_p + 5.81$

$EC_{50} = a\,\pi + b$

$EC_{50} = -6.69\,\pi + 9.59$

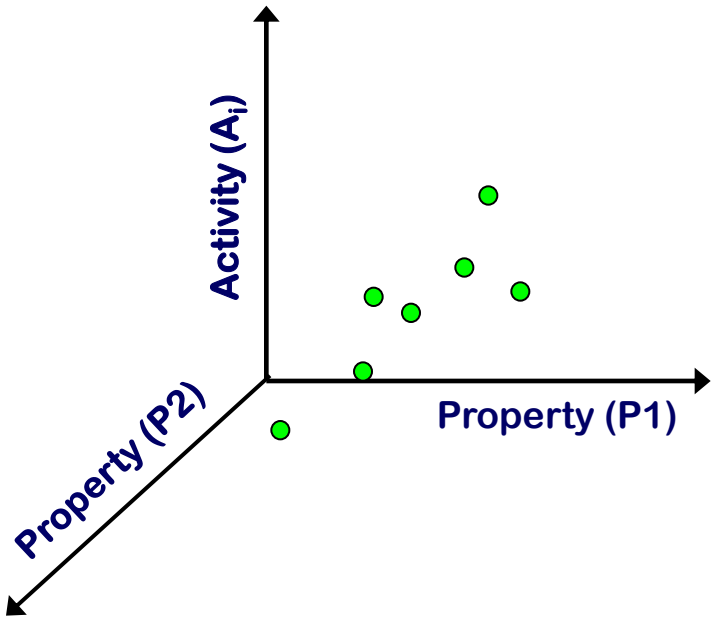
(Q)SAR: a strong and understandable temptation…

$EC_{50} = a\,\sigma_p + b$ $\qquad$ $EC_{50} = a\,\pi + b$

$$EC_{50} = a\,\sigma_p + b\,\pi + c$$

(Q)SAR: multiple regression analysis (MRA)

and if the experimental activity depended on more than one descriptor?



$$y_1 = a_1x_1 + b_1x_2 + z_1$$

$$\ldots$$

$$y_n = a_nx_n + b_nx_n + z_n$$

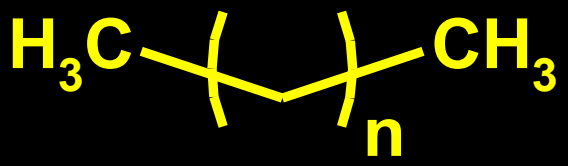
Multiple Regression Analysis (MRA)

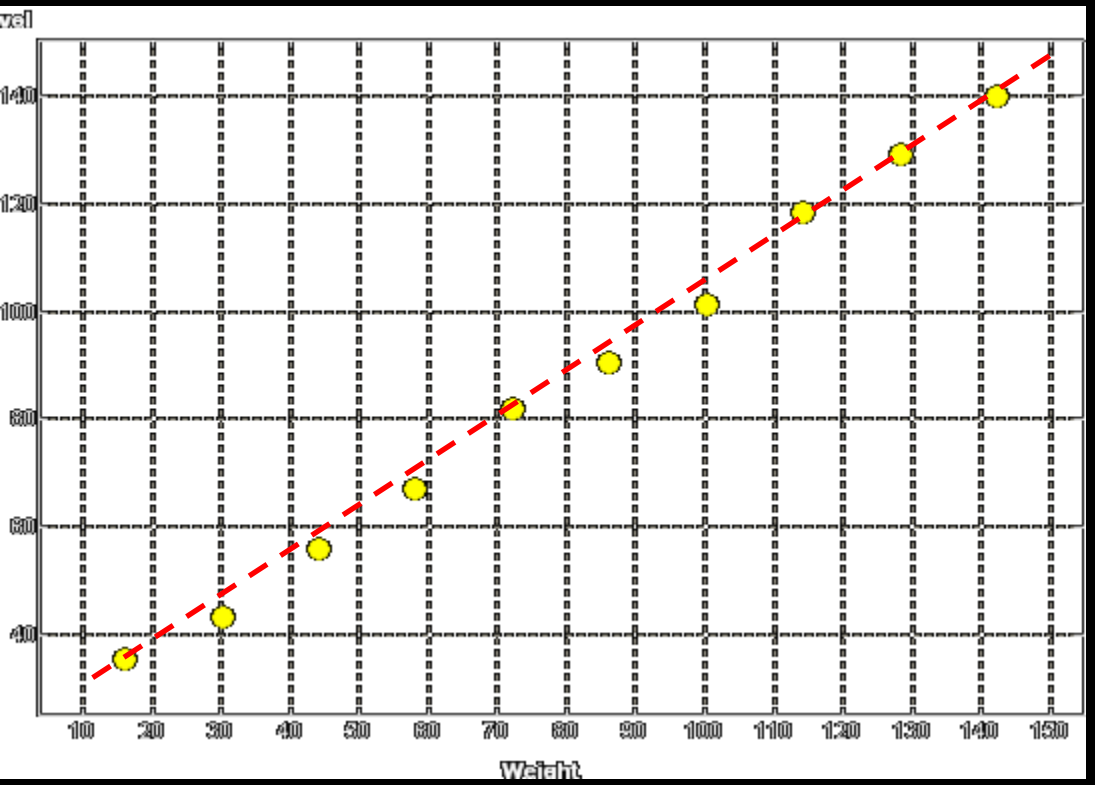
Requirements:

- There should be at least 5 times more samples than descriptors.
- Total number of descriptors should not exceed ~10 (looks the number of compounds you need!!!)
- Descriptors should be *uncorrelated*.

The second statistical gold rules do build up linear models:

- Having more the one molecular descriptors, the internal correlation (*cross-correlation*) between them has to be lower than 0.5





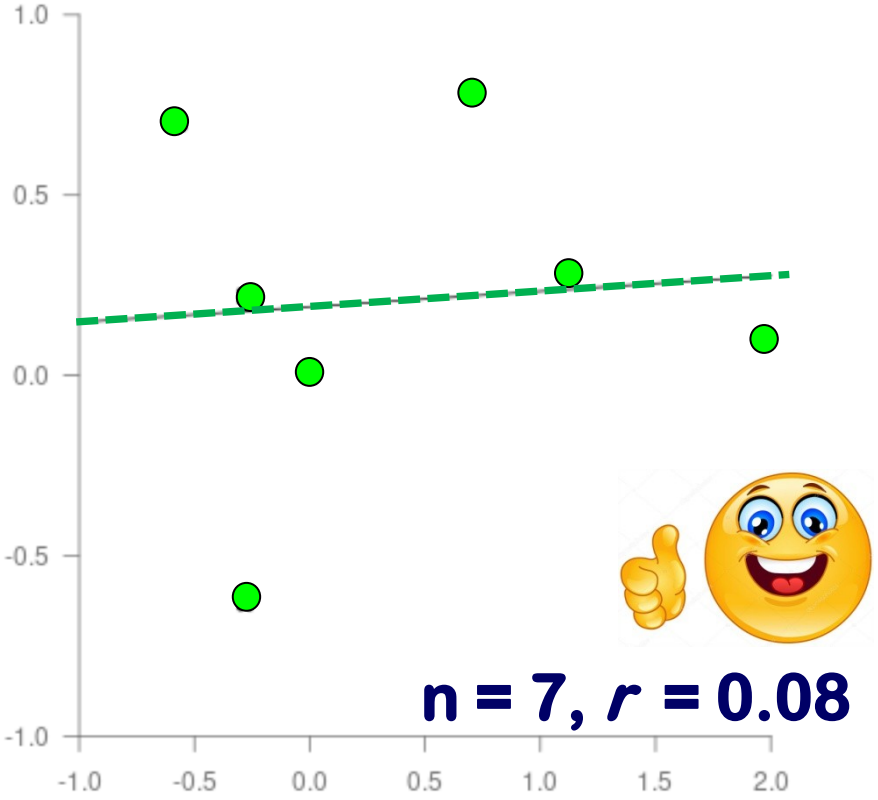
Descriptors:

- Molecular Volume
- Molecular Weight

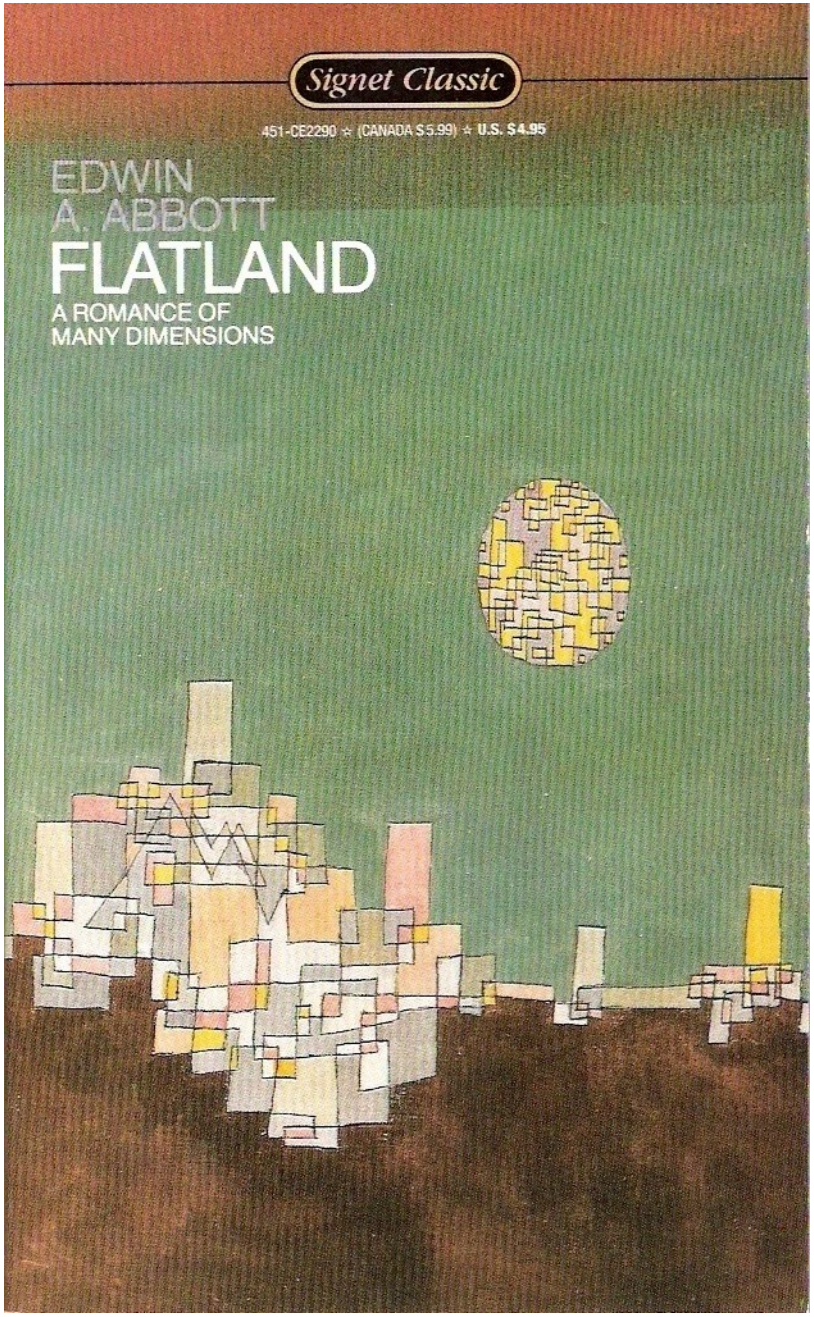
$r^2 = 0.9973$

Considering this specific combination of dataset (aliphatic hydrocarbons and molecular descriptors) molecular volume and molecular weight are strongly correlate thus redundant!

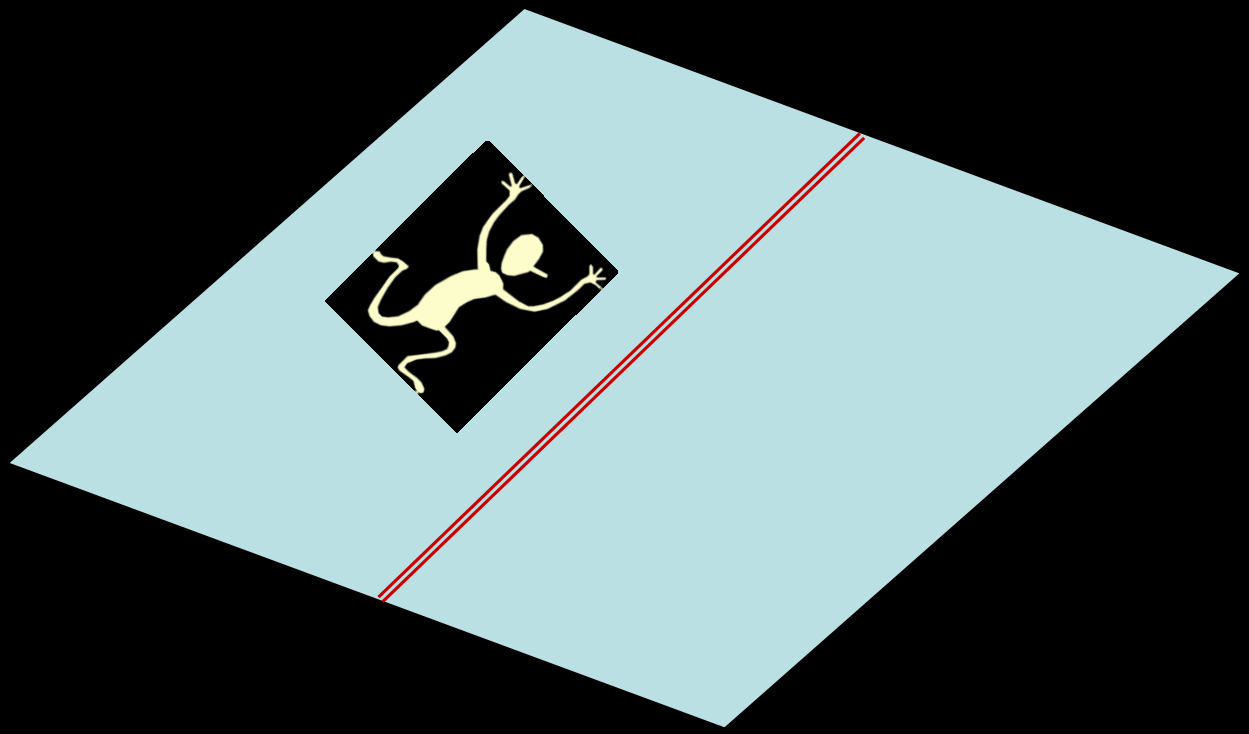
… are "σ_p" and "π" correlated?



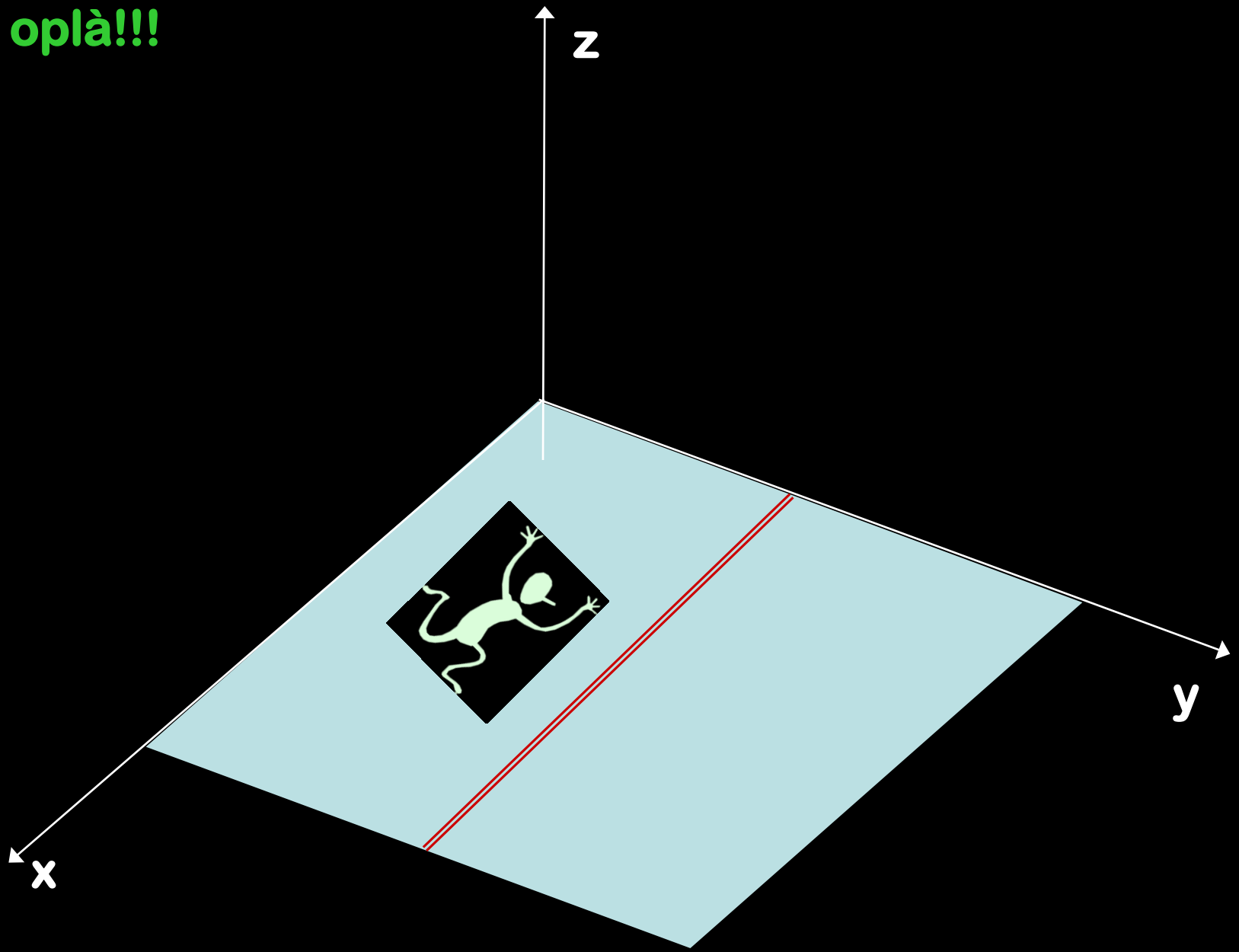
σ_p	π
0	0
0.78	0.71
0.22	- 0.28
0.70	- 0.57
0.10	2.00
-0.63	- 0.28
0.28	1.12
0.12	- 0.70


Signet Classic
451-CE2290 ☆ (CANADA $5.99) ☆ U.S. $4.95
EDWIN
A. ABBOTT
FLATLAND
A ROMANCE OF
MANY DIMENSIONS

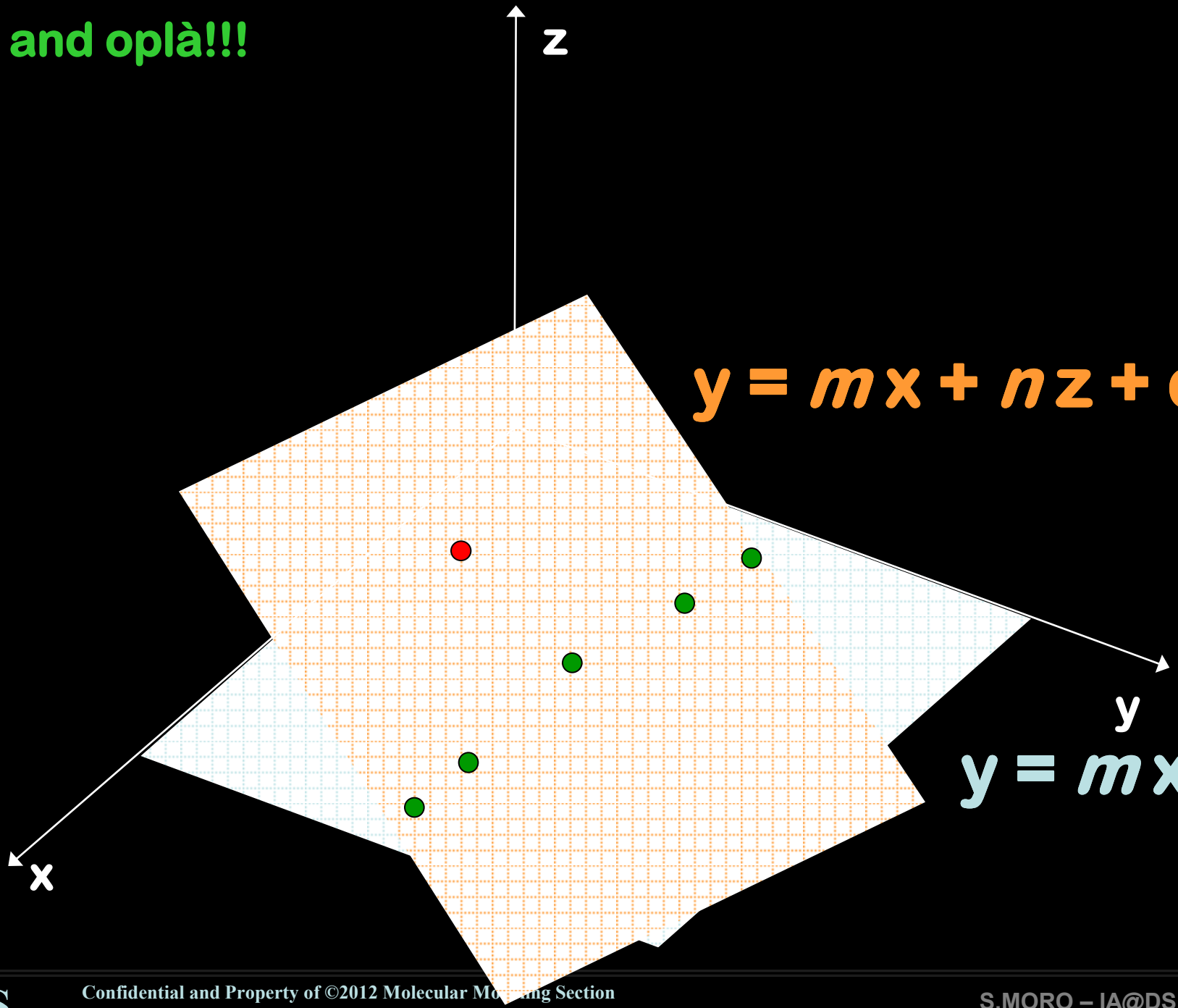
MRA approaches can transform the life in Flatland!



… oplà!!!



… and oplà!!!



$$y = mx + nz + q$$

$$y = mx + q$$

The nightmare in using MRA technique: OVERFITTING!!!

Another important consideration using MRA technique:

#	MR	logP	Volume	PM	Surface	density	n. X atoms
CH2Cl2	1,62959	1,30436	67,1359	84,933	176,117	1,63677	3
CHCl3	2,01731	1,73808	75,7514	119,378	186,824	2,03576	4
CCl4	2,35508	2,42116	83,2702	153,823	194,057	2,3224	5
CF3CHBrCl	2,35642	2,36112	88,098	197,381	206,644	2,85284	7
CHCl2CHCl2	2,92829	2,49472	94,2376	167,85	215,329	2,26061	6
Cl2C=CHCl	2,46705	2,28836	115,831	131,389	241,699	1,42863	5
CCl2=CCl2	2,82835	3,37472	132,106	165,834	257,237	1,46129	6

n = 7

cross-correlation?

credits: anonymous

Another important consideration using MRA technique:

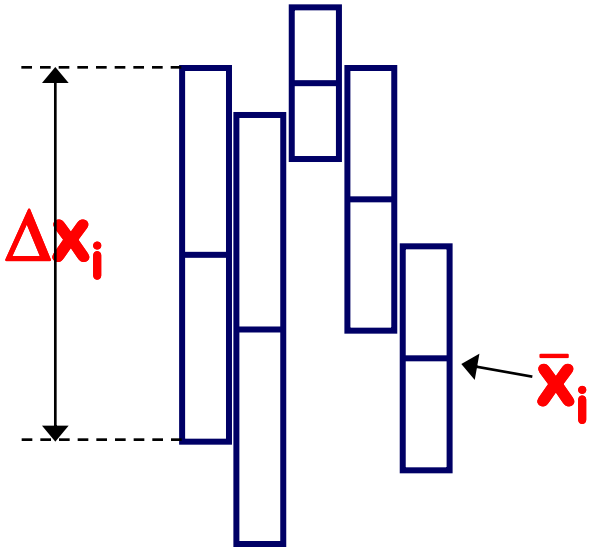
#	MR	logP	Volume	PM	Surface	density	n. X atoms
CH2Cl2	1,62959	1,30436	67,1359	84,933	176,117	1,63677	3
CHCl3	2,01731	1,73808	75,7514	119,378	186,824	2,03576	4
CCl4	2,35508	2,42116	83,2702	153,823	194,057	2,3224	5
CF3CHBrCl	2,35642	2,36112	88,098	197,381	206,644	2,85284	7
CHCl2CHCl2	2,92829	2,49472	94,2376	167,85	215,329	2,26061	6
Cl2C=CHCl	2,46705	2,28836	115,831	131,389	241,699	1,42863	5
CCl2=CCl2	2,82835	3,37472	132,106	165,834	257,237	1,46129	6

$\Delta x_i =$ 1.30 2.07 44.00 112.45 81.12 1.42 4.00

data scaling and data centering

data scaling and data centering

- Each indipendent variable influences the model according to its variance.
- Thus scaling corresponds to the assumption that all variables are *a priori* equally important.



Another important consideration using MRA technique:

#	MR	logP	Volume	PM	Surface	density	n. X atoms
CH2Cl2	1,62959	1,30436	67,1359	84,933	176,117	1,63677	3
CHCl3	2,01731	1,73808	75,7514	119,378	186,824	2,03576	4
CCl4	2,35508	2,42116	83,2702	153,823	194,057	2,3224	5
CF3CHBrCl	2,35642	2,36112	88,098	197,381	206,644	2,85284	7
CHCl2CHCl2	2,92829	2,49472	94,2376	167,85	215,329	2,26061	6
Cl2C=CHCl	2,46705	2,28836	115,831	131,389	241,699	1,42863	5
CCl2=CCl2	2,82835	3,37472	132,106	165,834	257,237	1,46129	6

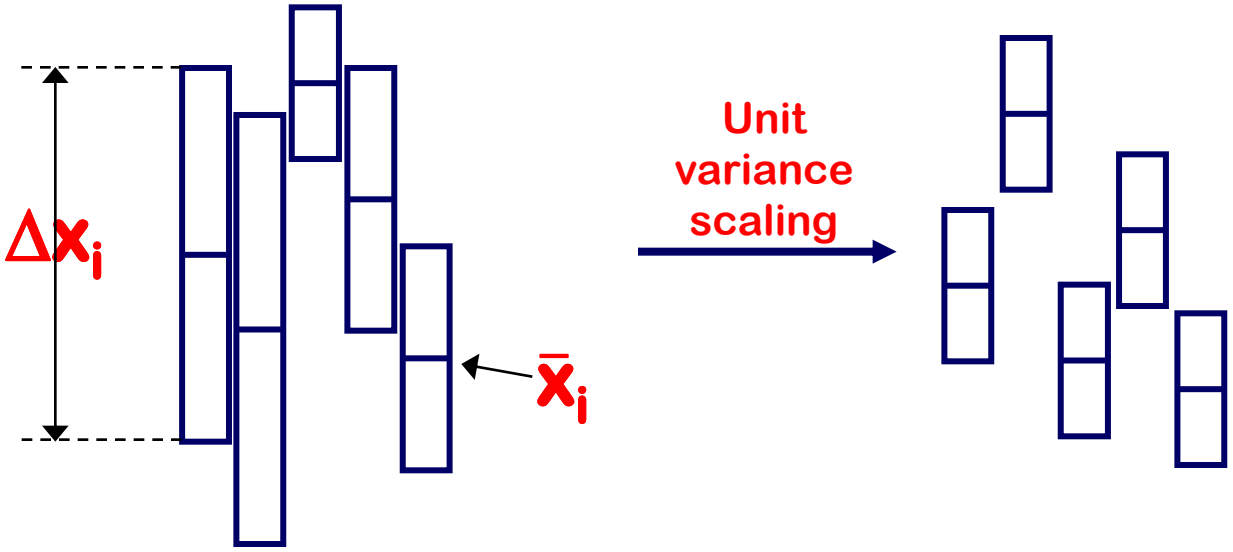
$\bar{x}_i$ = 2.37 2.28 93.77 145.80 211.13 1.99 5.10

Δx_i = 1.30 2.07 44.00 112.45 81.12 1.42 4.00

data scaling and data centering

data scaling and data centering

- ***Unit variance scaling***: **multiply each column by $1/\sigma_i$, σ_i being the standard deviation.**



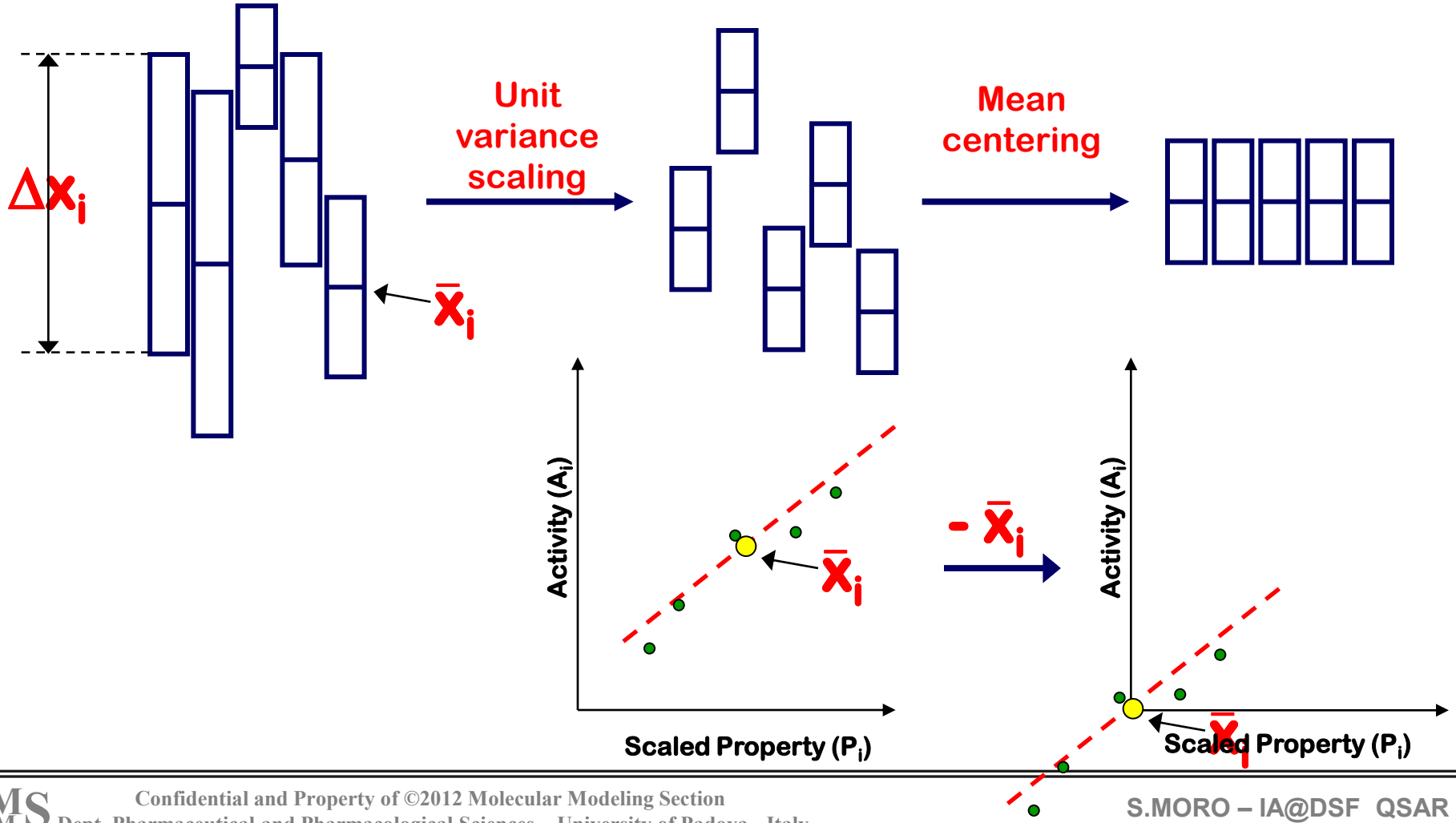
$$\sigma_i = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$ where *n* is the number of data taken

Back to the real case:

#	MR	logP	Volume	PM	Surface	density	n. X atoms
CH2Cl2	1,62959	1,30436	67,1359	84,933	176,1169	1,63677	3
CHCl3	2,01731	1,73808	75,7514	119,378	186,8237	2,03576	4
CCl4	2,35508	2,42116	83,2702	153,823	194,0565	2,3224	5
CF3CHBrCl	2,35642	2,36112	88,098	197,381	206,6438	2,85284	7
CHCl2CHCl2	2,92829	2,49472	94,2376	167,85	215,3294	2,26061	6
Cl2C=CHCl	2,46705	2,28836	115,831	131,389	241,6985	1,42863	5
CCl2=CCl2	2,82835	3,37472	132,106	165,834	257,2367	1,46129	6
$\bar{x}_i$ =	2.37	2.28	93.77	145.80	211.13	1.99	5.10
Δx_i =	1.30	2.07	44.00	112.45	81.12	1.42	4.00
σ_i =	0.45	0.65	22.85	37.02	29.46	0.52	1.34
$\Delta x_i/\sigma_i$ =	2.89	3.18	1.92	3.04	2.75	2.73	2.98

data scaling and data centering

- ***Mean centering***: **subtract from each column its average value.**



MRA should be a suitable tool only if these criteria are respected:

1. Good ratio between independent and dependent variables;

2. Statistical significance of the regression coefficient;

3. The magnitude of the typical effect "$b_i x_i$";

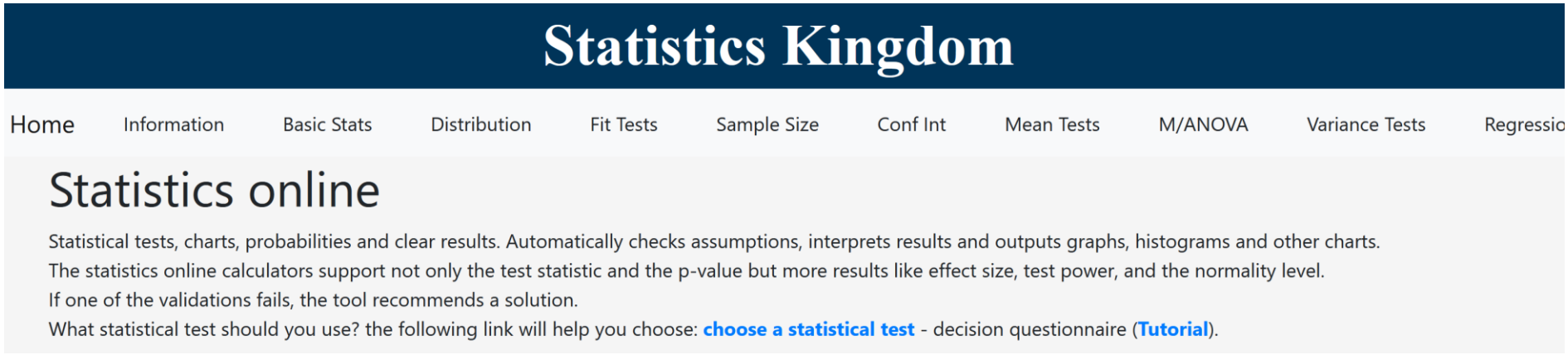
4. Any cross-correlation with other terms.

(Q)SAR: multiple regression analysis (MRA) back to oue example

Cmpd number	Cmpd name	X	EC_{50} (μM)	σ_p	π
1	6a	H	11.80 ± 1.90	0	0
2	6b	Cl	1.24 ± 0.11	0.78	0.71
3	6d	NO_2	4.58 ± 0.29	0.22	- 0.28
4	6e	CN	26.50 ± 5.87	0.70	- 0.57
5	6f	C_6H_5	0.24 ± 0.30	0.10	2.00
6	6g	$N(CH_3)_2$	4.39 ± 0.67	-0.63	- 0.28
7	6h	I	0.35 ±0.05	0.28	1.12
8	6i	NHCHO	???	0.12	- 0.70

$$EC_{50} = a\,\sigma_p + b\,\pi + c$$

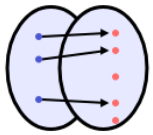
(Q)SAR: multiple regression analysis (MRA) please visit:



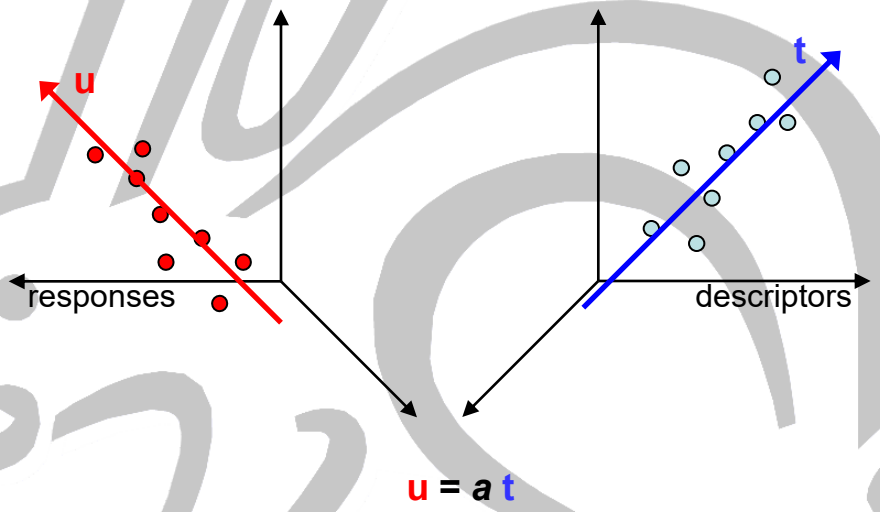
https://www.statskingdom.com/index.html

(Q)SAR: multiple regression analysis (MRA) please visit:

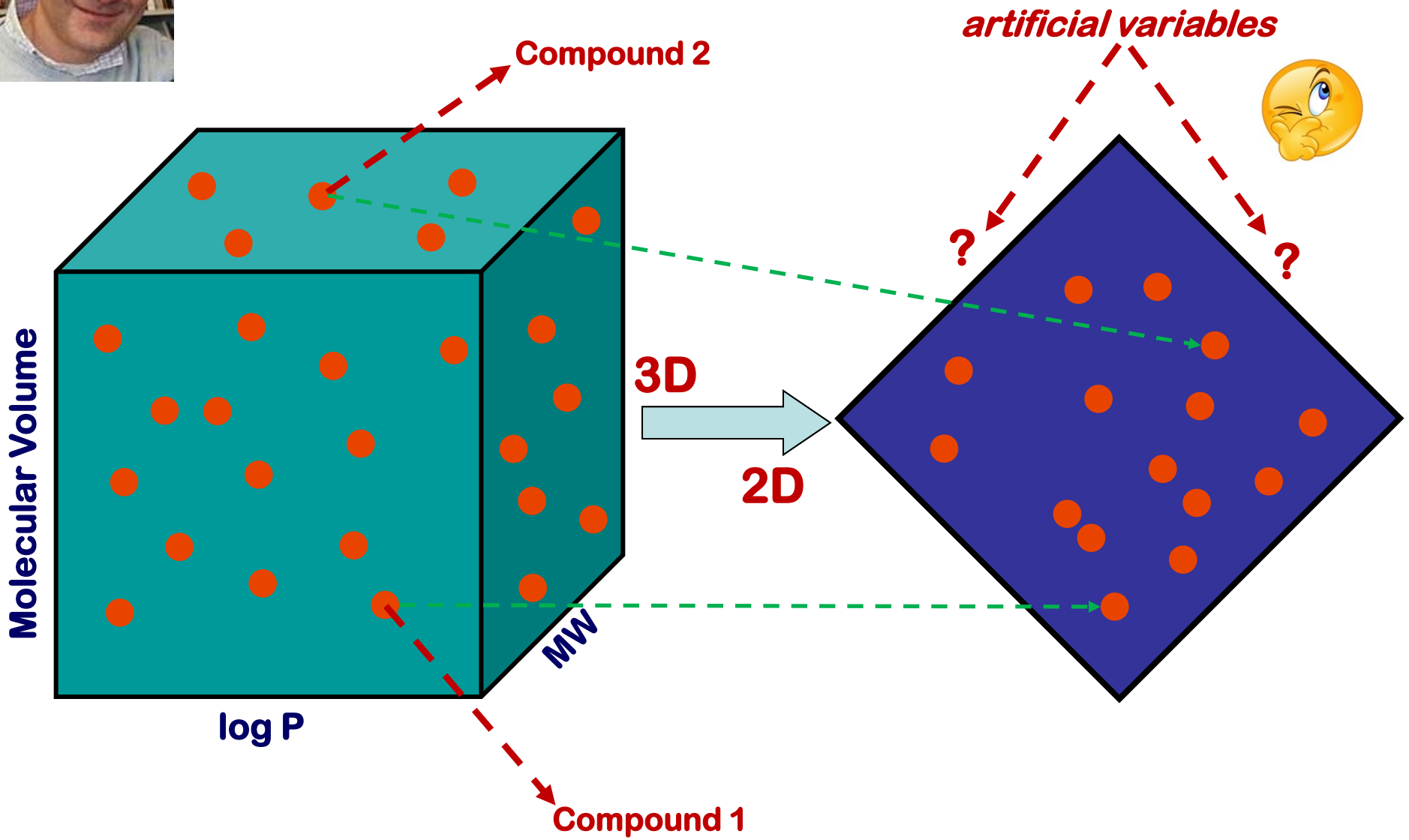
Regression calculator

#	Regression	Statistic
1	Simple Linear Regression	$F = \frac{\mathrm{MS}_{\mathrm{Regression}}}{\mathrm{MS}_{\mathrm{Error}}}$
2	Multiple Linear Regression	$F = \frac{\mathrm{MS}_{\mathrm{Regression}}}{\mathrm{MS}_{\mathrm{Error}}}$
3	*Bulk Linear Regression	$F = \frac{\mathrm{MS}_{\mathrm{Regression}}}{\mathrm{MS}_{\mathrm{Error}}}$
4	Binary Logistic Regression	$\chi 2 = 2(LL_1\text{-}LL_0)$
5	Multinomial Logistic Regression	$\chi 2 = 2(LL_1\text{-}LL_0)$
6	Propensity Score Matching	

Principal Component Analysis and Regression



Data Presentation: Property Space



Data Presentation: Property Space

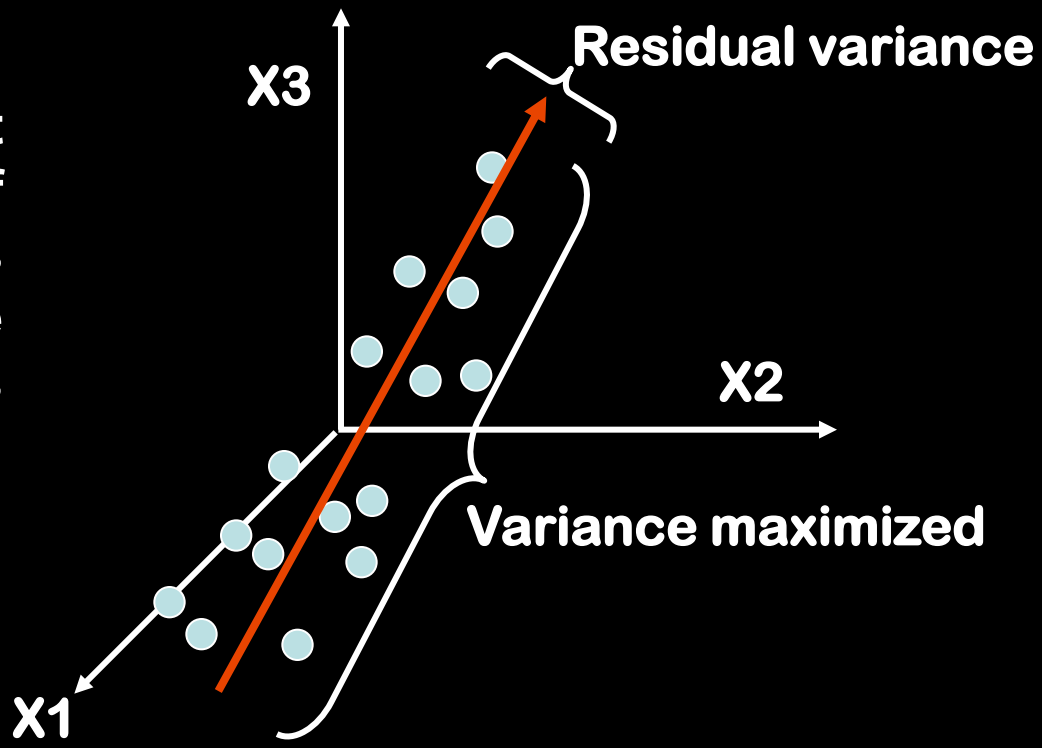
Principal component analysis, or PCA, is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Hotelling H (1933) "*Analysis of a complex of statistical variables into principal components*". J Educ Psychol 24:417–441, 498–520

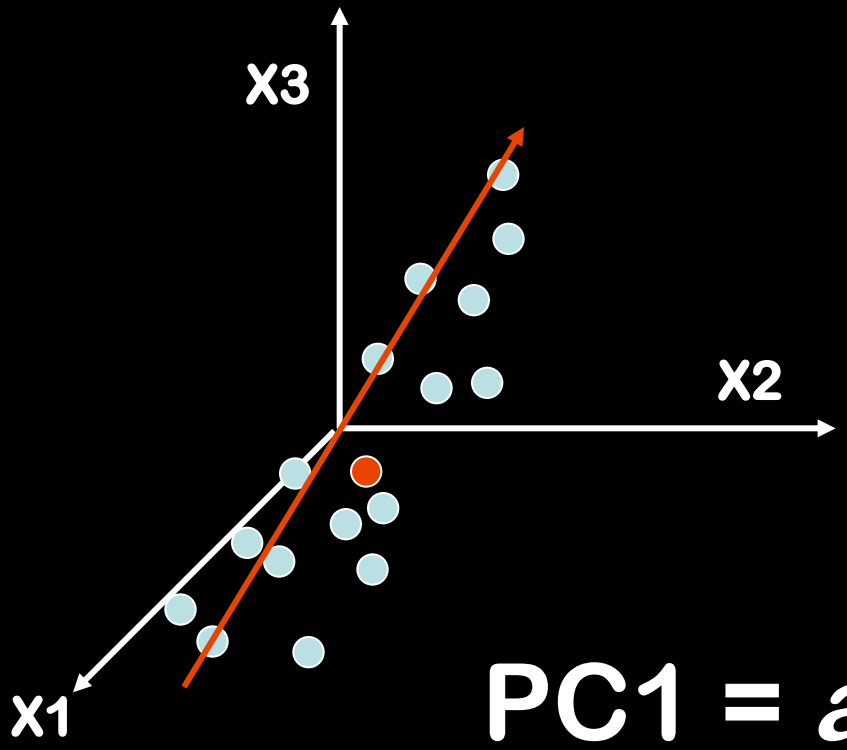
Principle Component Analysis (PCA)

PCA finds lines, planes and hyperplanes in the originally K-dimensions space that approximate the data as well as possible in the least square sense. In such a case, the variance in the original data is maximized.

A line that is the least squares approximation of a set of data points makes the variance of the coordinates on the line as large as possible.



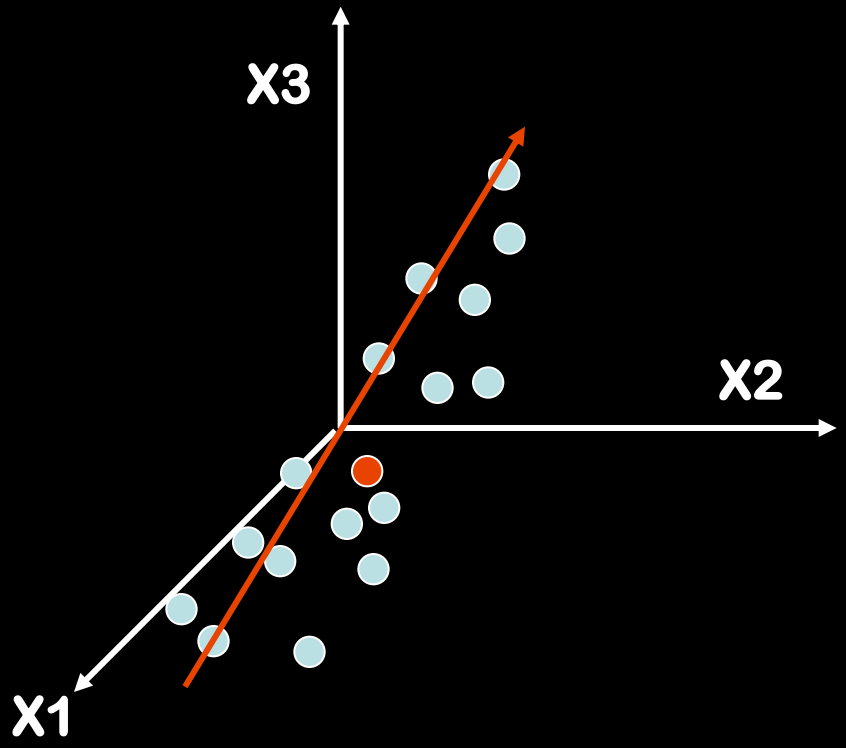
A Geometrical Interpretation of PCA

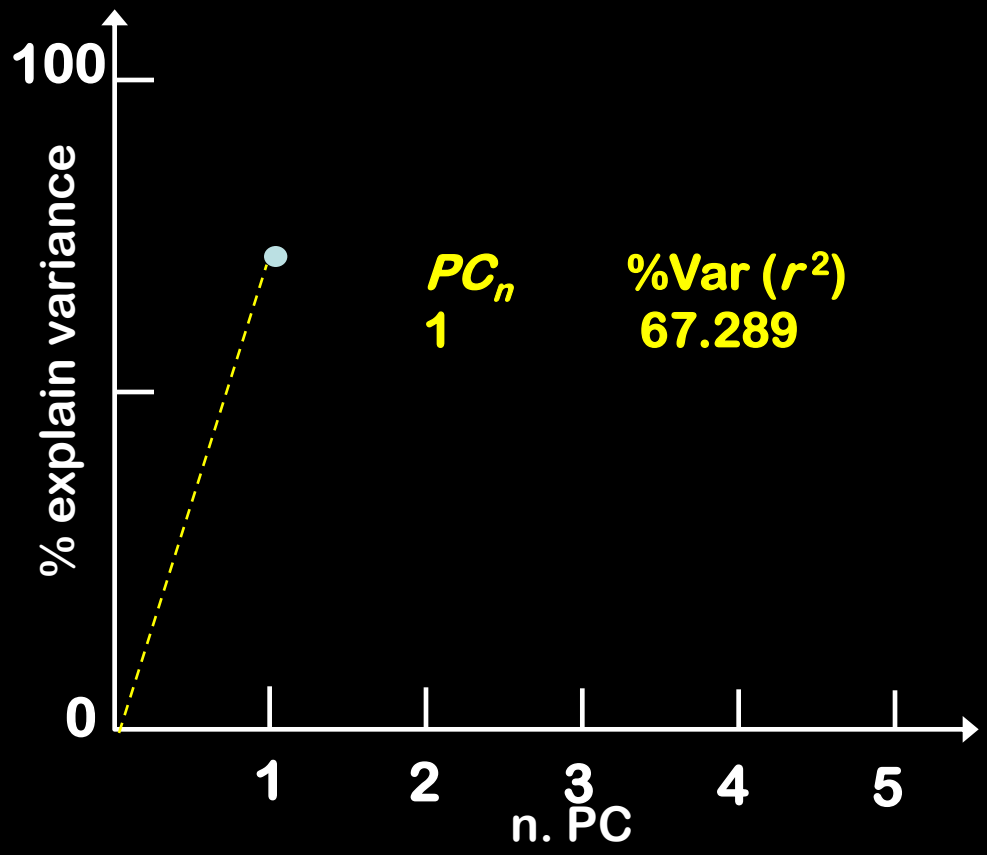


1. Set up k-dimensional space;
2. Plot point;
3. Plot vector of averages at the center of gravity;
4. Mean-center the data;
5. **Generate the first PC:**
   Passes through the origin
   Best approximates the data in a least squares sense

$$PC1 = a_1 X1 + b_1 X2 + c_1 X3$$

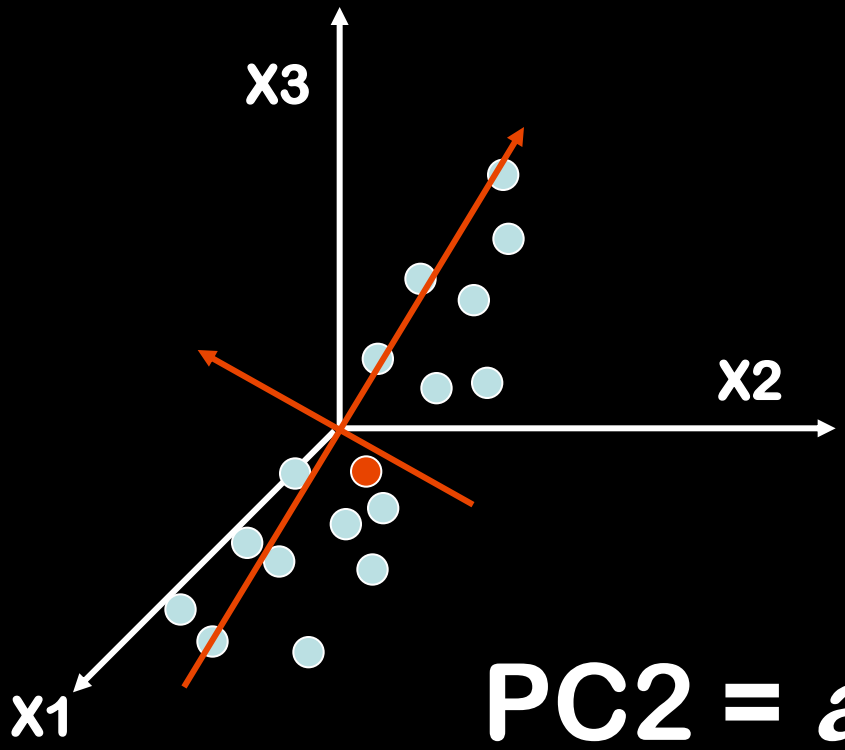
A Geometrical Interpretation of PCA: The scree plot





$$PC1 = a_1 X1 + b_1 X2 + c_1 X3$$

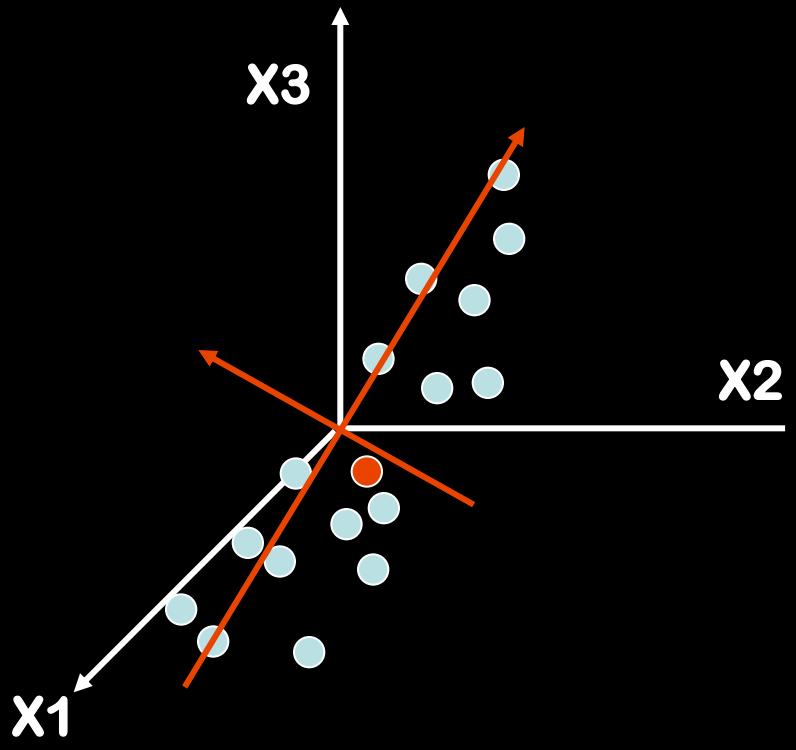
A Geometrical Interpretation of PCA

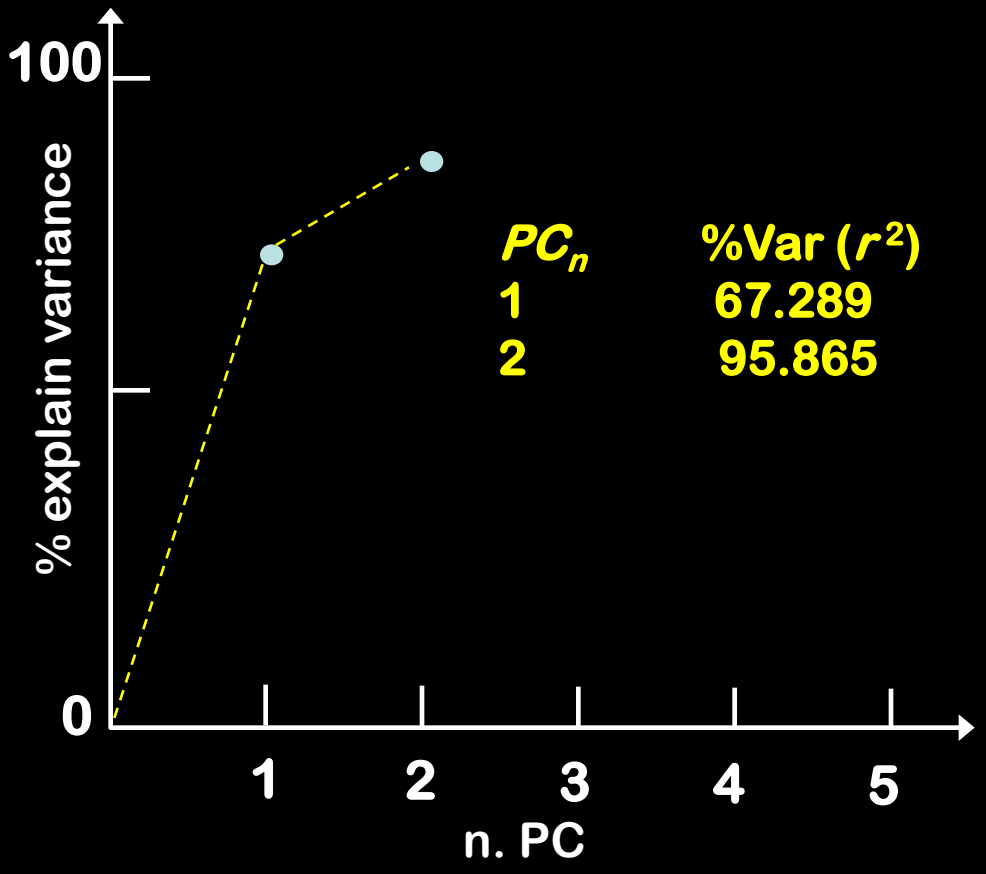


6. **Generate the second PC:**
   Passes through the origin and orthogonal to first PC
   Maximally improves the approximation of the X-data

$$PC2 = a_2X1 + b_2X2 + c_2X3$$

A Geometrical Interpretation of PCA: The scree plot

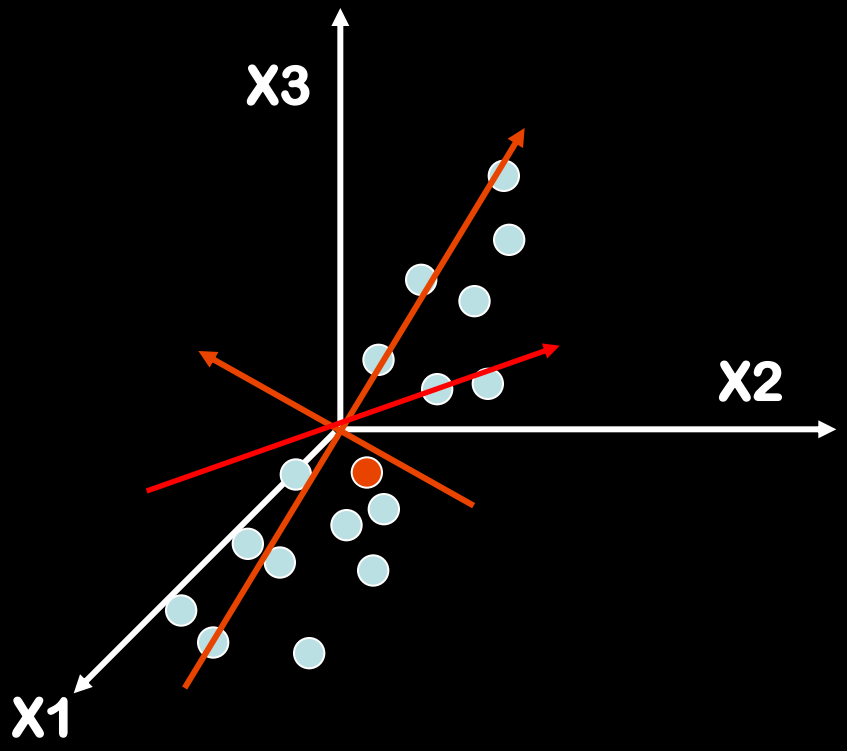


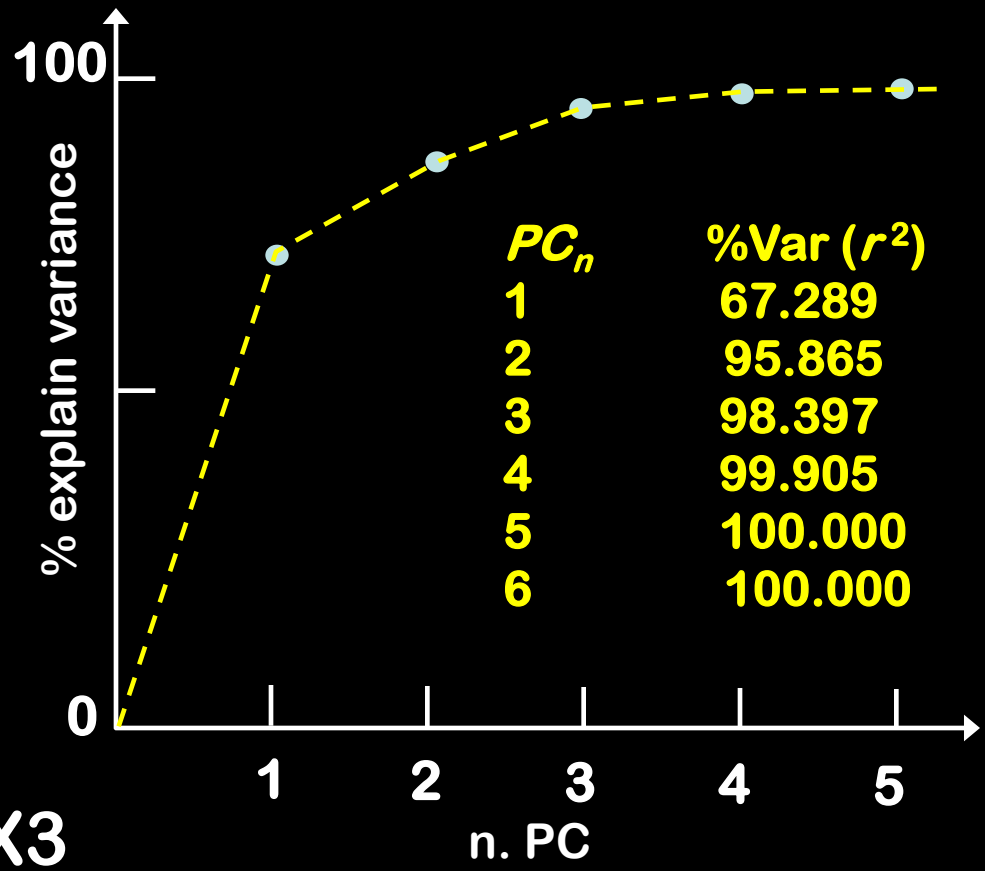


PC_n	%Var (r^2)
1	67.289
2	95.865

$$PC2 = a_2X1 + b_2X2 + c_2X3$$

A Geometrical Interpretation of PCA: The scree plot





PC_n	%Var (r^2)
1	67.289
2	95.865
3	98.397
4	99.905
5	100.000
6	100.000

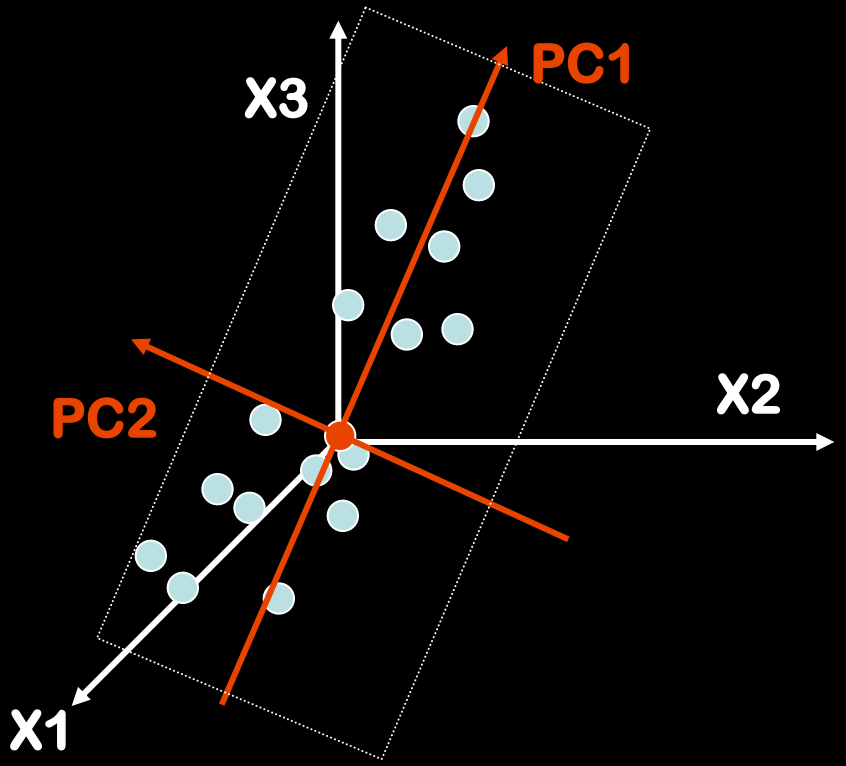
$$PC3 = a_3X1 + b_3X2 + c_3X3$$

…

$$PCn = a_nX1 + b_nX2 + c_nX3$$

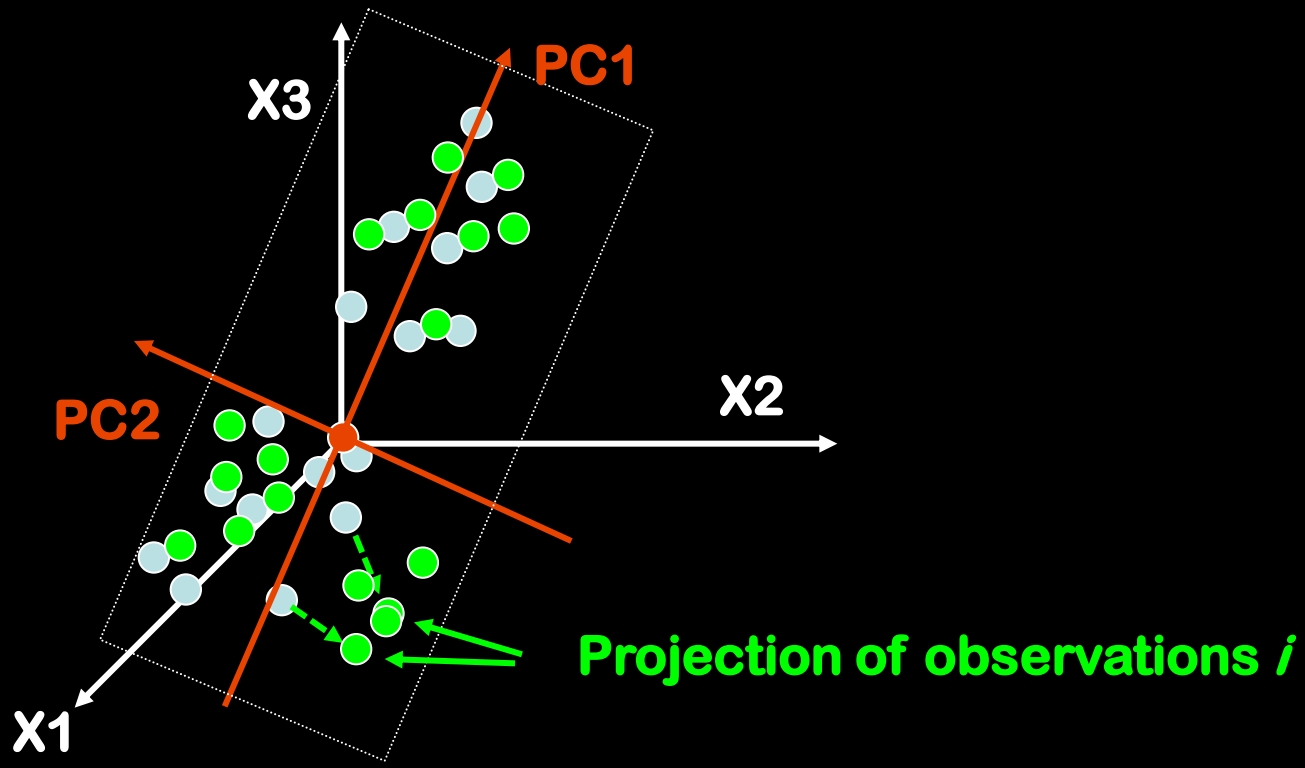
A Geometrical Interpretation of PCA

- First 2 PC's define a plane in the original K-dimensions space.



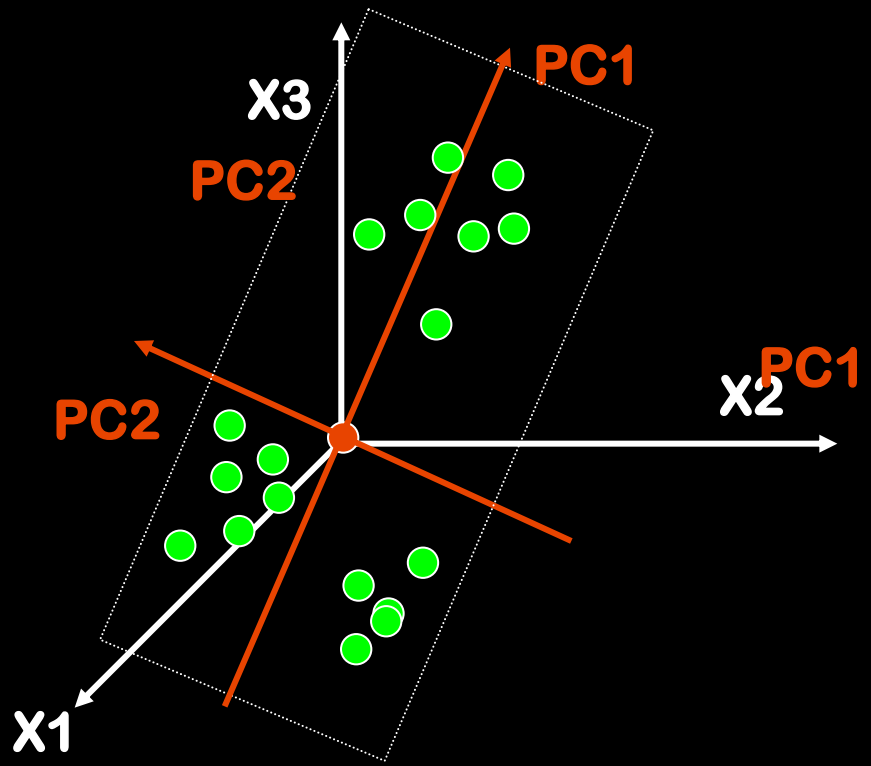
A Geometrical Interpretation of PCA

- By projecting all data points into this plane it is possible to visualize the structure of the data set.



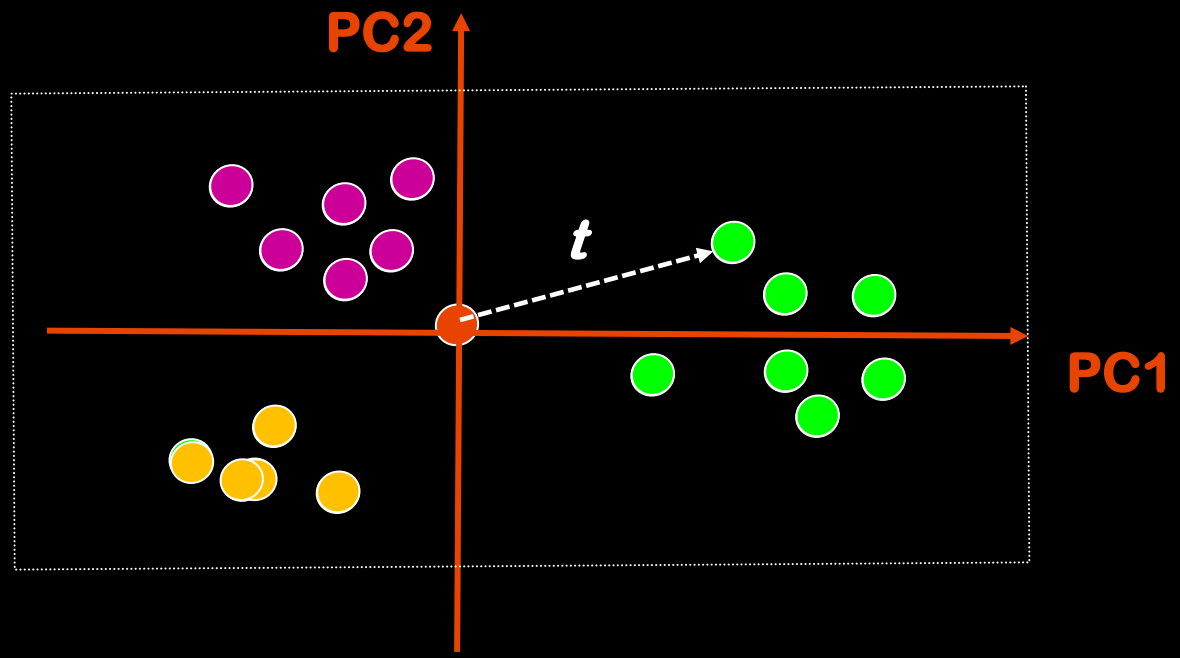
A Geometrical Interpretation of PCA

- By projecting all data points into this plane it is possible to visualize the structure of the data set.



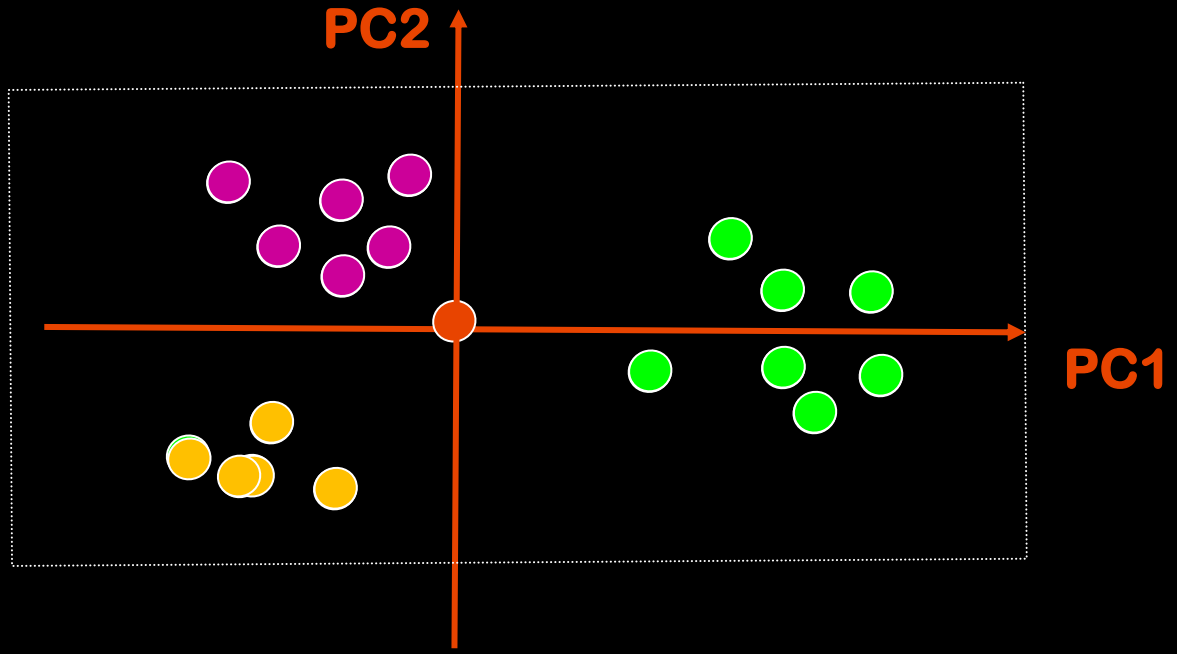
A Geometrical Interpretation of PCA

- ***Scores (t)*** **are the coordinates of the original data points in this plane.**



On the score plot,
"Sit together": similar behavior between descriptors

PCA as clustering method



As seen on the plot each group forms a distinct cluster indicating that there are differences between the groups in the data.
The groups could just as well overlap on the plot. Small overlaps indicate that groups are slightly similar in the data and that we are not able to distinguish between groups in extreme cases. Large or complete overlaps indicate that we do not see any difference between groups in the data.

PCA and Loading Plot

The contribution (*loading*) of each original variable to each PC.

Which variables are responsible for the pattern observed.

PC's can be associated with certain dataset characteristics.

$$PC1 = a_1 X1 + b_1 X2 + c_1 X3$$
$$PC2 = a_2 X1 + b_2 X2 + c_2 X3$$
$$PC3 = a_3 X1 + b_3 X2 + c_3 X3$$
$$\ldots$$
$$PCn = a_n X1 + b_n X2 + c_n X3$$

PCA and Loading Plot

In the loading plot of a PCA, each point represents an independent variable (i.e., one of the original variables in your dataset). So:

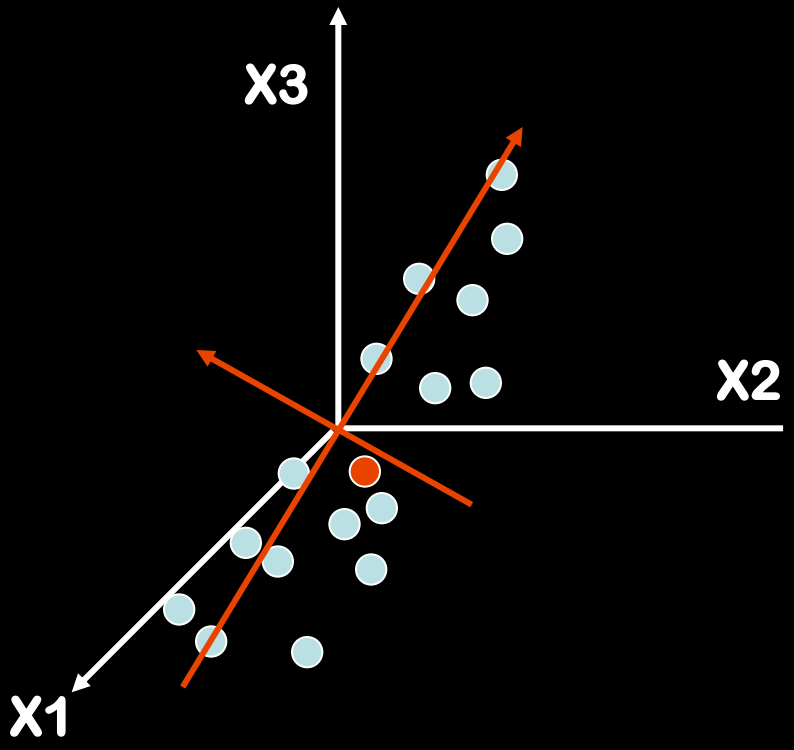
If you have, for example, 10 independent variables, you'll see 10 points in the loading plot.

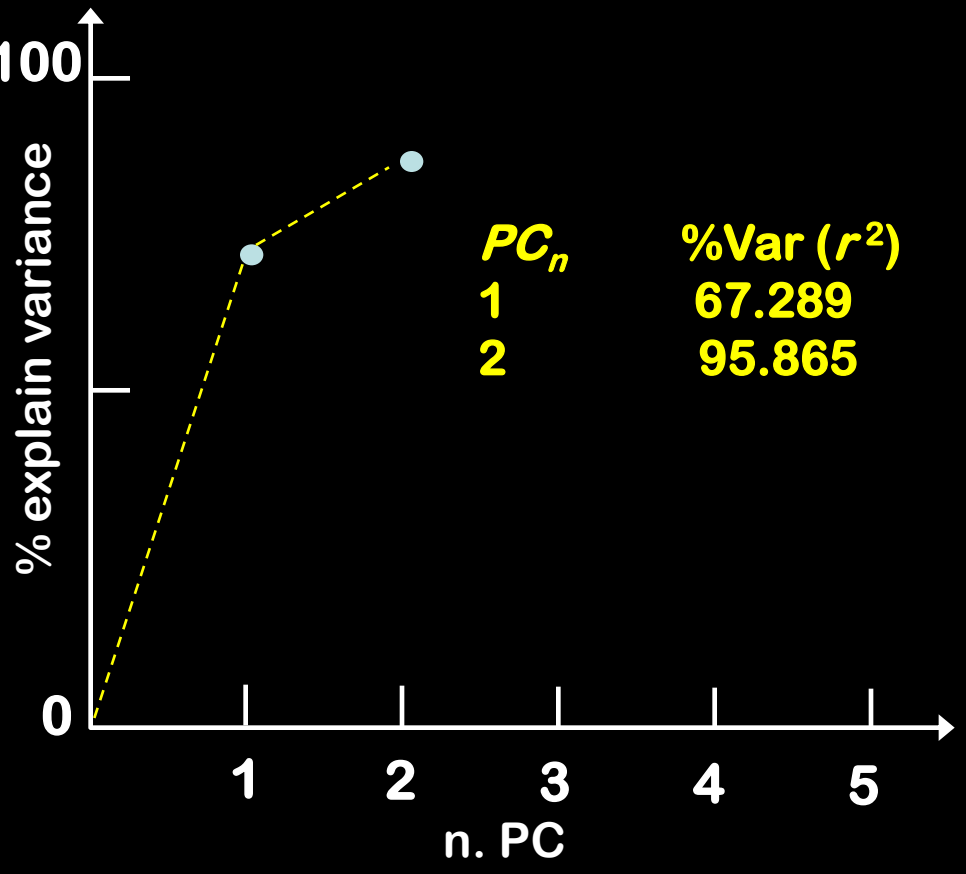
The position of each point in the plot shows how much and in which direction that variable contributes to the principal components (usually PC1 and PC2). In this case:

$$PC1 = a_1 X1 + b_1 X2 + c_1 X3$$
$$PC2 = a_2 X1 + b_2 X2 + c_2 X3$$
$$PC3 = a_3 X1 + b_3 X2 + c_3 X3$$
$$\dots$$
$$PCn = a_n X1 + b_n X2 + c_n X3$$

PCA and Loading Plot

As an example:





PC_n	%Var (r^2)
1	67.289
2	95.865

PC1 = *0.45* X1 + *0.33* X2 - *0.1* X3

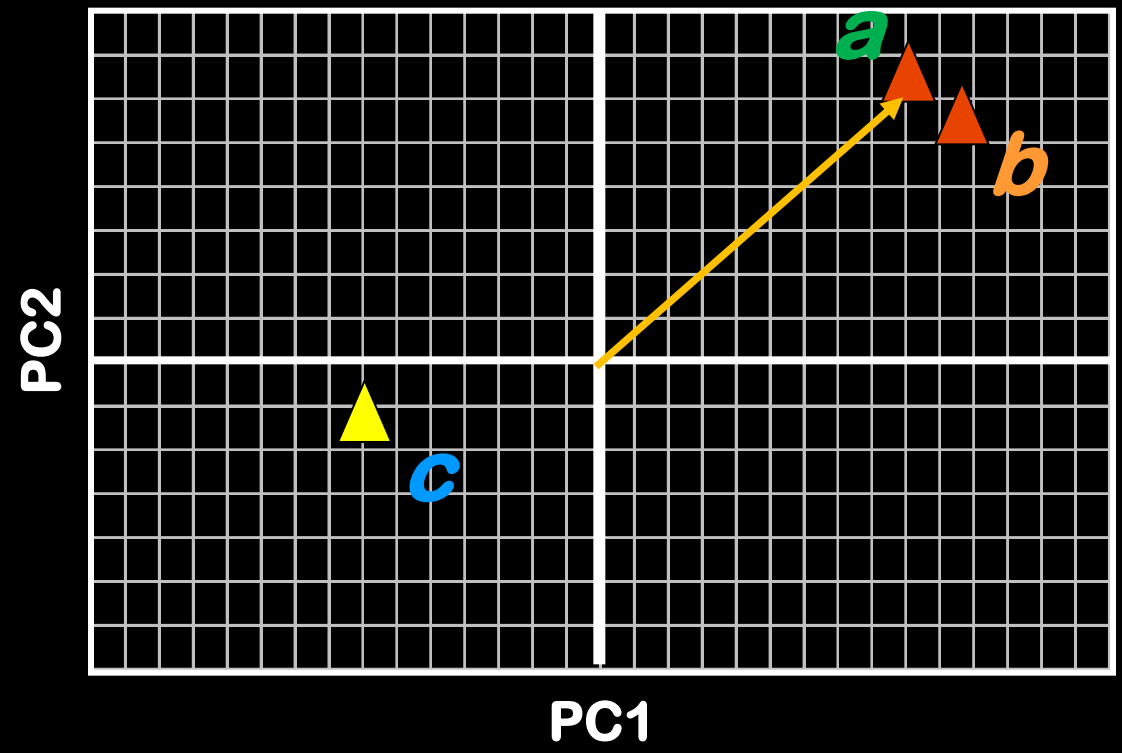
PC2 = *0.32* X1 + *0.47* X2 - *0.42* X3

PCA and Loading Plot

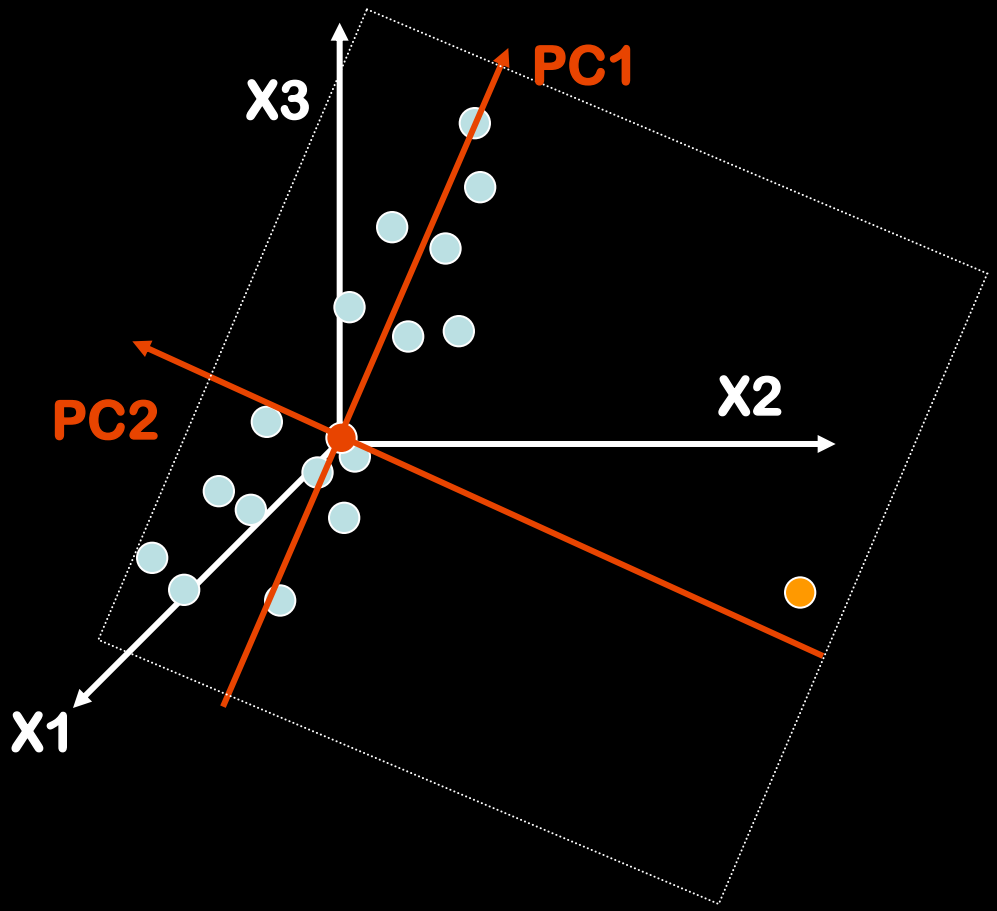
The further away from the origin a variable lies, the stronger impact it has on the model.

Variables correlations:

- Positively correlated
- Negatively correlated

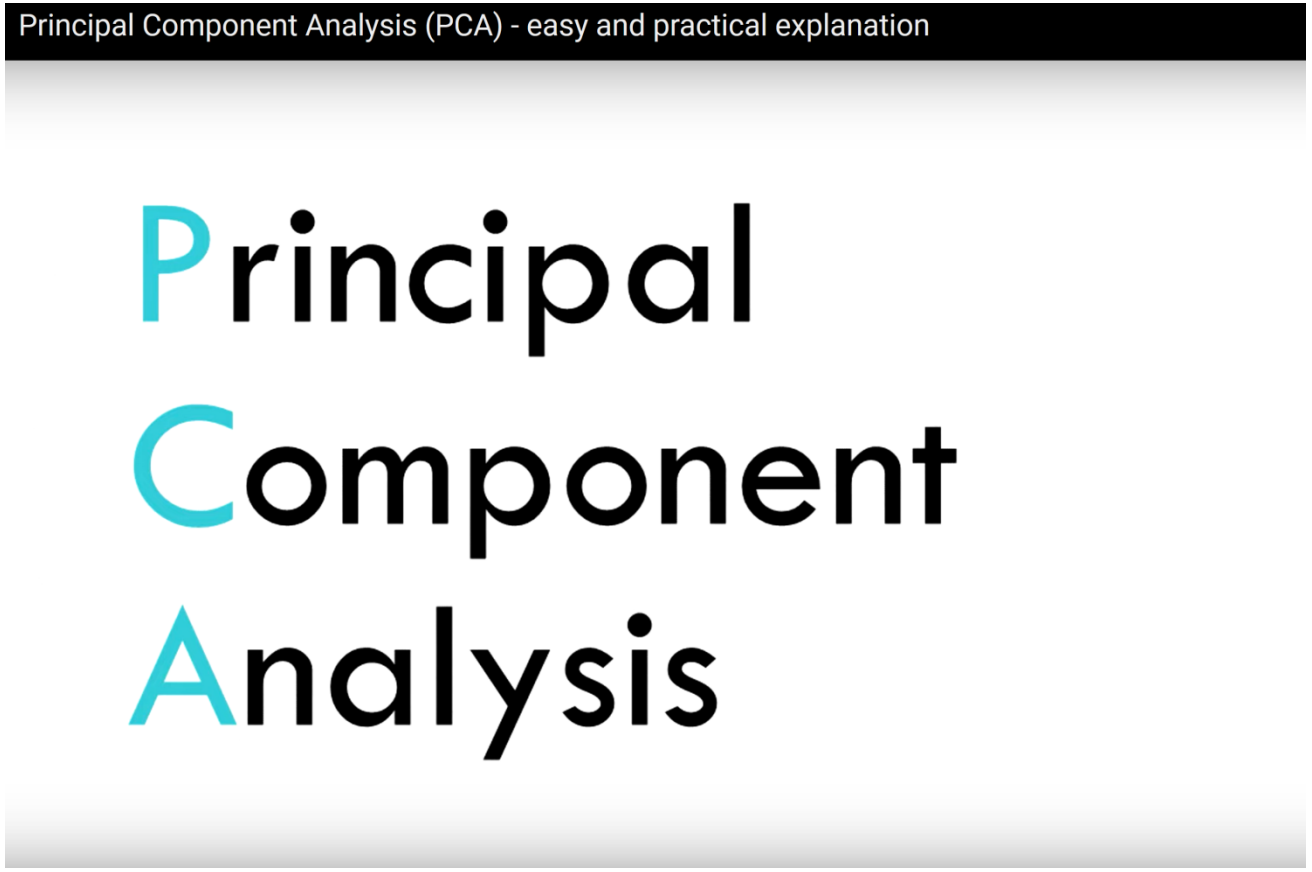


PCA and Outliers Detection



On the score plot…. easy to detect!

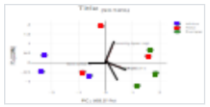
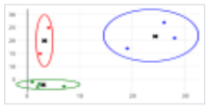
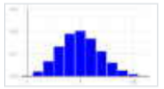
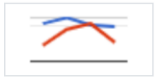
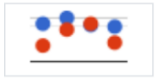
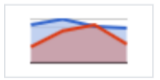
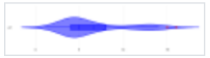
Please visit this video:



https://www.youtube.com/watch?v=5vgP05YpKdE

(Q)SAR: multiple regression analysis (MRA) please visit:

Visualization

#	Chart	Image
1	Principal component analysis	
2	Cluster analysis	
3	Histogram maker	
4	Box plot maker	
5	Bar graph maker	
6	Pie chart maker	
7	Line chart maker	
8	Scatter chart maker	
9	Area chart maker	
10	Violin plot maker	

Fron descriptor to their principal compoments:

#	MR	logP	Volume	PM	Surface	density	n. X atoms
CH2Cl2	1,62959	1,30436	67,1359	84,933	176,117	1,63677	3
CHCl3	2,01731	1,73808	75,7514	119,378	186,824	2,03576	4
CCl4	2,35508	2,42116	83,2702	153,823	194,057	2,3224	5
CF3CHBrCl	2,35642	2,36112	88,098	197,381	206,644	2,85284	7
CHCl2CHCl2	2,92829	2,49472	94,2376	167,85	215,329	2,26061	6
Cl2C=CHCl	2,46705	2,28836	115,831	131,389	241,699	1,42863	5
CCl2=CCl2	2,82835	3,37472	132,106	165,834	257,237	1,46129	6

#	PC1	PC2	PC3	PC4	PC5	
CH2Cl2	-1,78677	0,583913	0,285292	0,101483	1,469406	
CHCl3	-0,97519	-0,07003	-0,20004	0,095932	-1,21311	
CCl4	-0,13863	-0,63091	-0,75064	1,407425	-1,00089	
CF3CHBrCl	0,500215	-1,74077	1,586333	-0,26991	0,353284	
CHCl2CHCl2	0,616411	-0,55574	-1,81367	-1,19166	0,764962	
Cl2C=CHCl	0,337096	1,253576	0,681746	-1,39385	-1,0778	
CCl2=CCl2	1,446869	1,159959	0,210977	1,250575	0,704142	

Back to the real case:

#	LD25	MR	logP	Volume	PM	Surface	density
CH2Cl2	0.96	1,6296	1,3044	67,136	84,933	176,117	1,63677
CHCl3	1.45	2,0173	1,7381	75,751	119,378	186,824	2,03576
CCl4	1.53	2,3551	2,4212	83,27	153,823	194,057	2,3224
CF3CHBrCl	1.31	2,3564	2,3611	88,098	197,381	206,644	2,85284
CHCl2CHCl2	2.42	2,9283	2,4947	94,238	167,85	215,329	2,26061
Cl2C=CHCl	2.26	2,4671	2,2884	115,83	131,389	241,699	1,42863
Cl2C=CCl2	2.26	2,8284	3,3747	132,11	165,834	257,237	1,46129

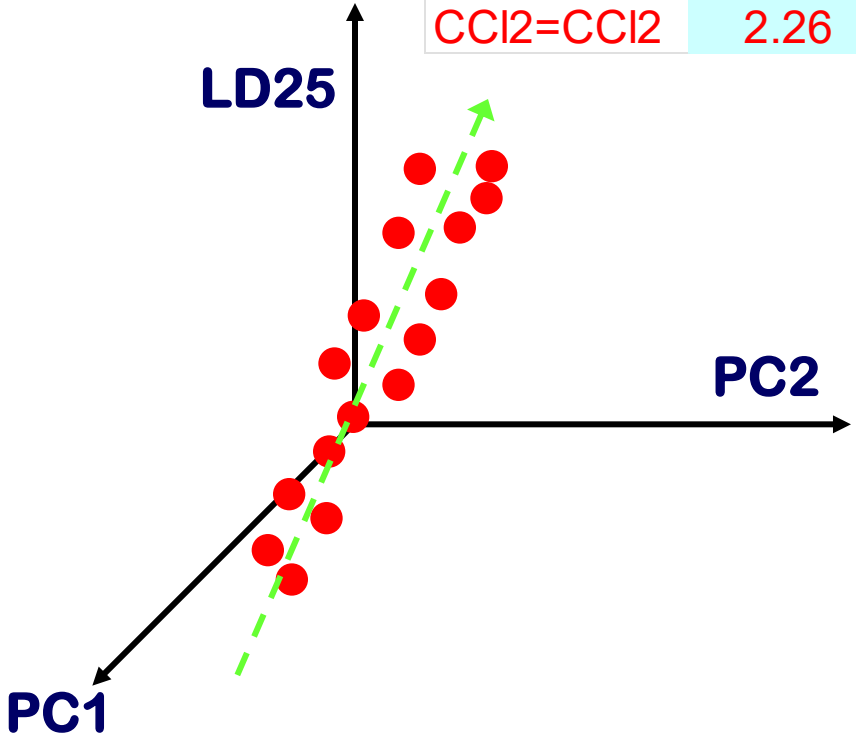
n = 7

$$\log(1/LD_{25}) = a\,MR + b\,logP + c\,Vol + d\,PM + e\,Sur + f\,dens + g\,X\text{-}atom + h$$

or

… PC regression

#	LD25	PC1	PC2	PC3	PC4	PC5
CH2Cl2	0.96	-1,78677	0,583913	0,285292	0,101483	1,469406
CHCl3	1.45	-0,97519	-0,07003	-0,20004	0,095932	-1,21311
CCl4	1.53	-0,13863	-0,63091	-0,75064	1,407425	-1,00089
CF3CHBrCl	1.31	0,500215	-1,74077	1,586333	-0,26991	0,353284
CHCl2CHCl2	2.42	0,616411	-0,55574	-1,81367	-1,19166	0,764962
Cl2C=CHCl	2.26	0,337096	1,253576	0,681746	-1,39385	-1,0778
CCl2=CCl2	2.26	1,446869	1,159959	0,210977	1,250575	0,704142

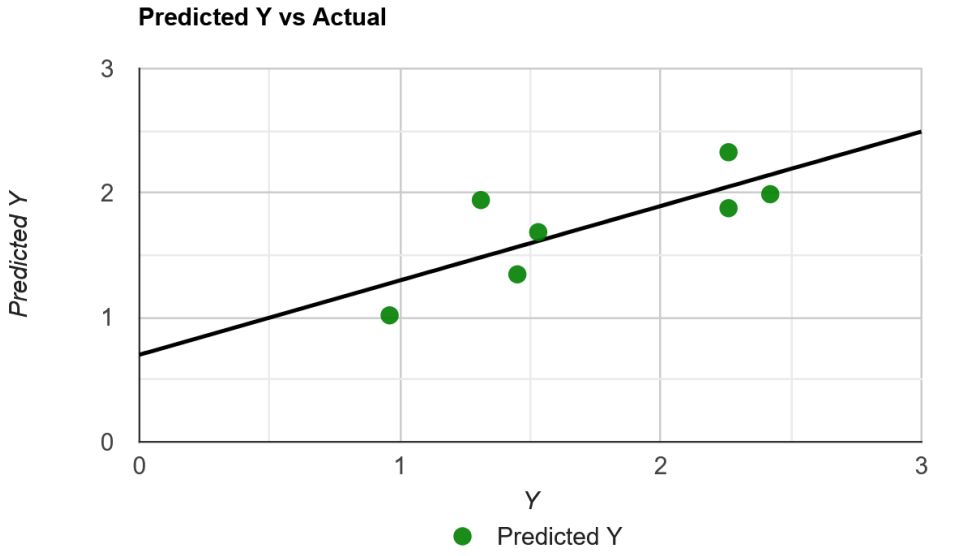


$$LD_{25} = m\,PC1 + n\,PC2$$

… PC regression

#	LD25	PC1	PC2	PC3	PC4	PC5
CH2Cl2	0.96	-1,78677	0,583913	0,285292	0,101483	1,469406
CHCl3	1.45	-0,97519	-0,07003	-0,20004	0,095932	-1,21311
CCl4	1.53	-0,13863	-0,63091	-0,75064	1,407425	-1,00089
CF3CHBrCl	1.31	0,500215	-1,74077	1,586333	-0,26991	0,353284
CHCl2CHCl2	2.42	0,616411	-0,55574	-1,81367	-1,19166	0,764962
Cl2C=CHCl	2.26	0,337096	1,253576	0,681746	-1,39385	-1,0778
CCl2=CCl2	2.26	1,446869	1,159959	0,210977	1,250575	0,704142

The coefficient of multiple correlation (R) equals 0.77.



$LD_{25} = 0.40\ PC1 + 0.18\ PC2$

PCA… and ANN!

It can use *principal components* as input for a neural network, and in many cases, it's actually a good practice, especially when:

Advantages of using PCA before a neural network:

- ***Dimensionality reduction***

If you have many variables, PCA reduces the number of inputs by removing redundancy while keeping the most important information.

- ***Better generalization***

Fewer inputs → lower risk of overfitting, especially if your dataset isn't very large.

- ***Faster training***

The network is smaller, so it trains faster.

- ***Removes correlation***

Principal components are orthogonal (uncorrelated), which can help certain models converge better.

Things to keep in mind:

- ***Interpretability:***

Principal components are not as easily interpretable as the original variables.

- ***Information loss:***

If you choose too few components, you might lose some relevant variability.

- ***Proper pipeline:***

When using PCA, make sure it's fitted only on the training data and then applied to test data using the same transformation.