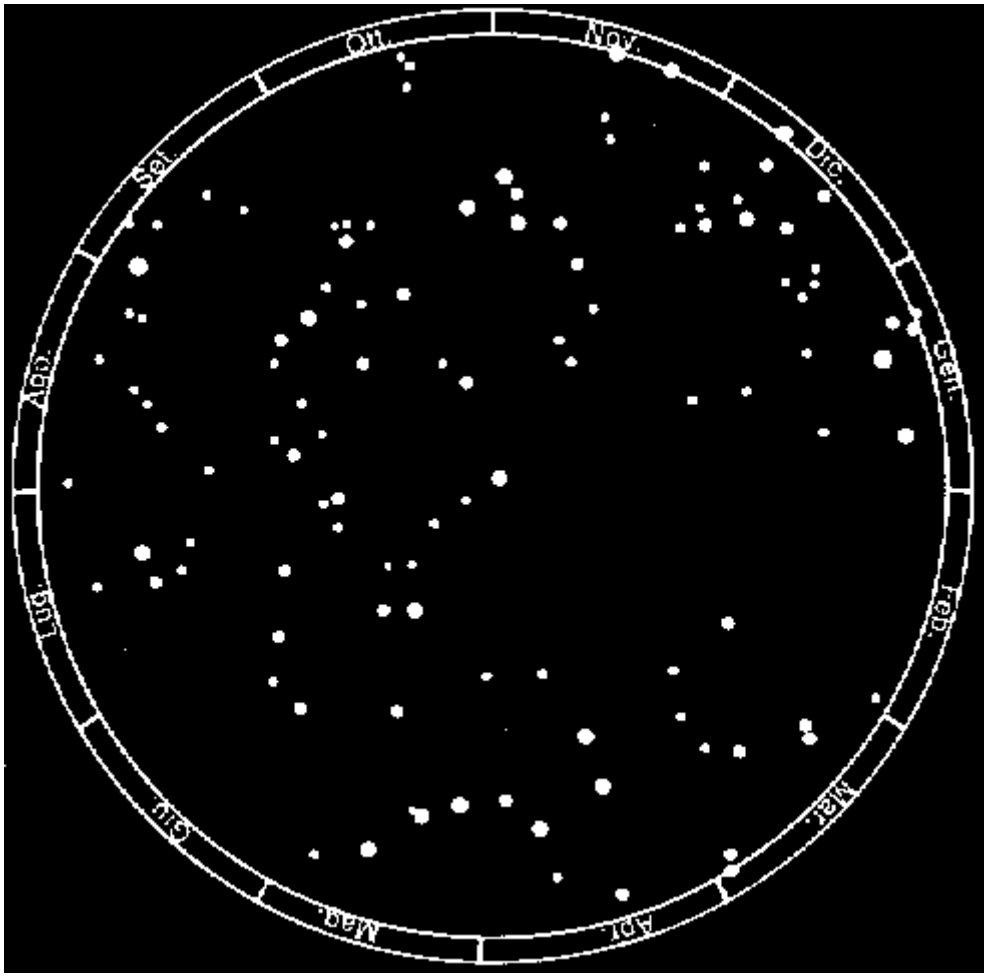
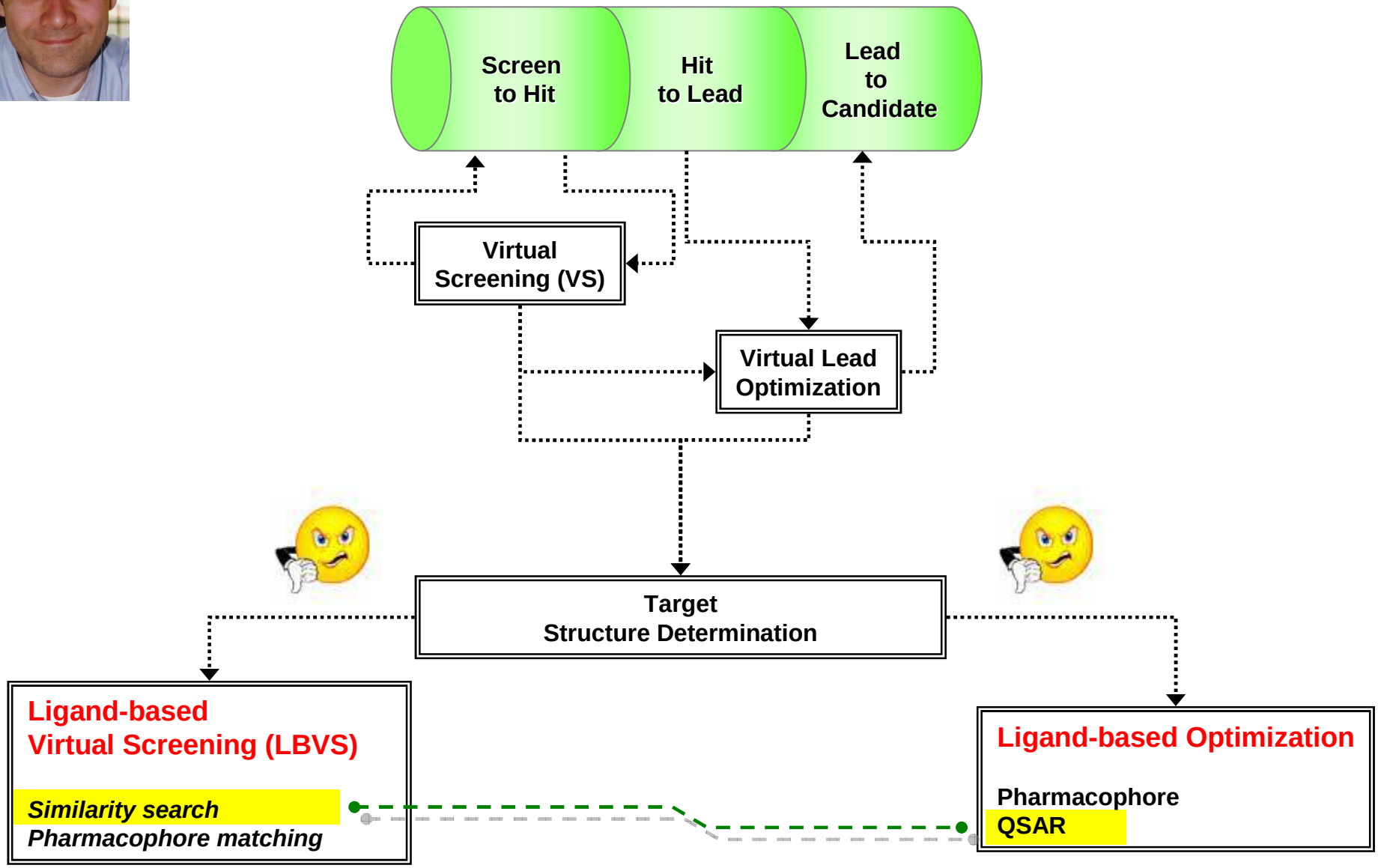
In and around.... (Q)SAR



here we are again:



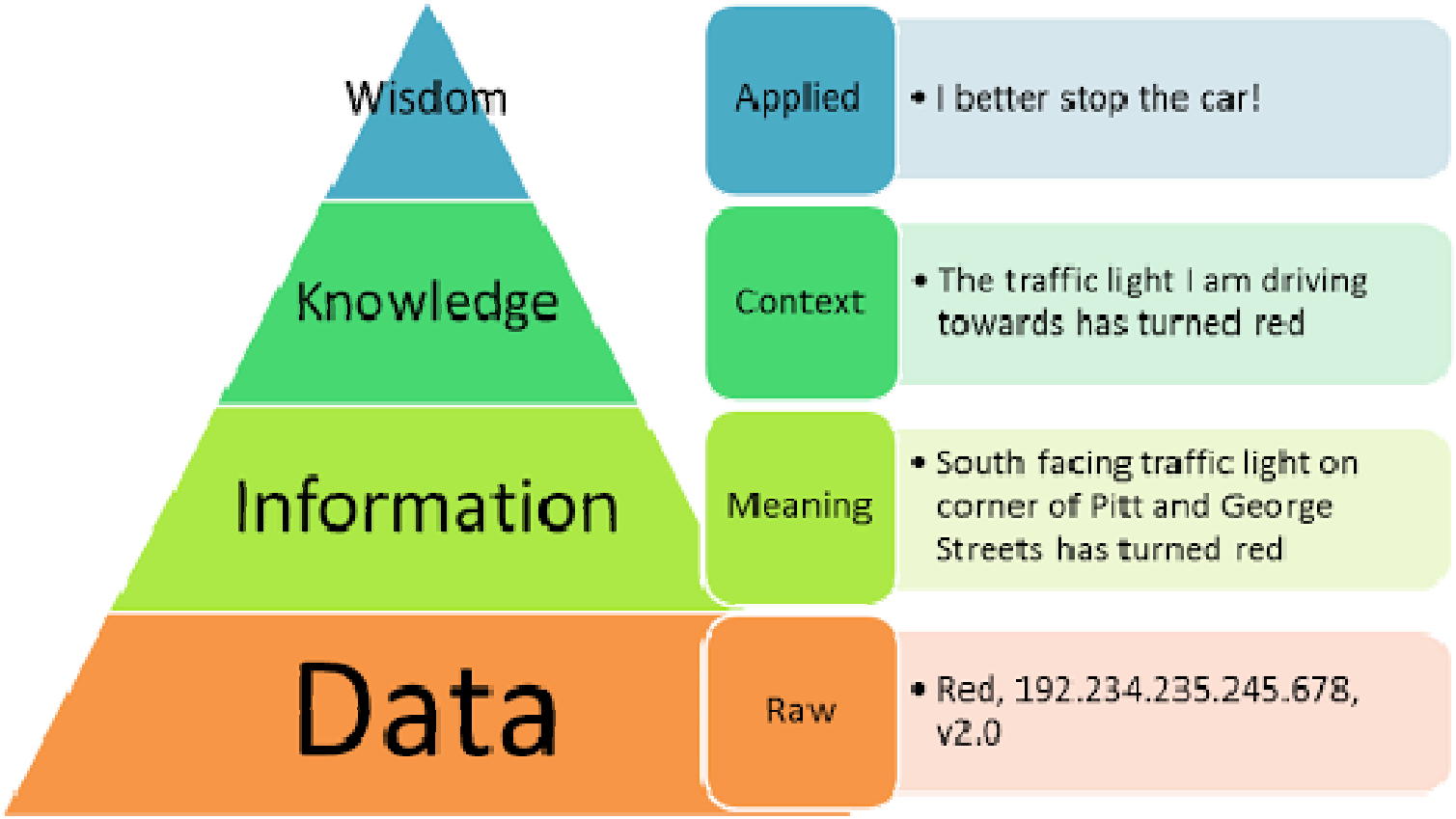
… do you surely remember:

Biological Space

Chemical Space	TargetA	TargetB	TargetC	TargetD	TargetN
Comp.1	$K_1(A)$				
Comp.2	$K_2(A)$				
Comp.3	$K_3(A)$				
Comp.4	$K_4(A)$				
Comp.n	$K_n(A)$				

Screening

… and surely also this concept:



Here is another wonderful example of data → information transformation:

Biological Space

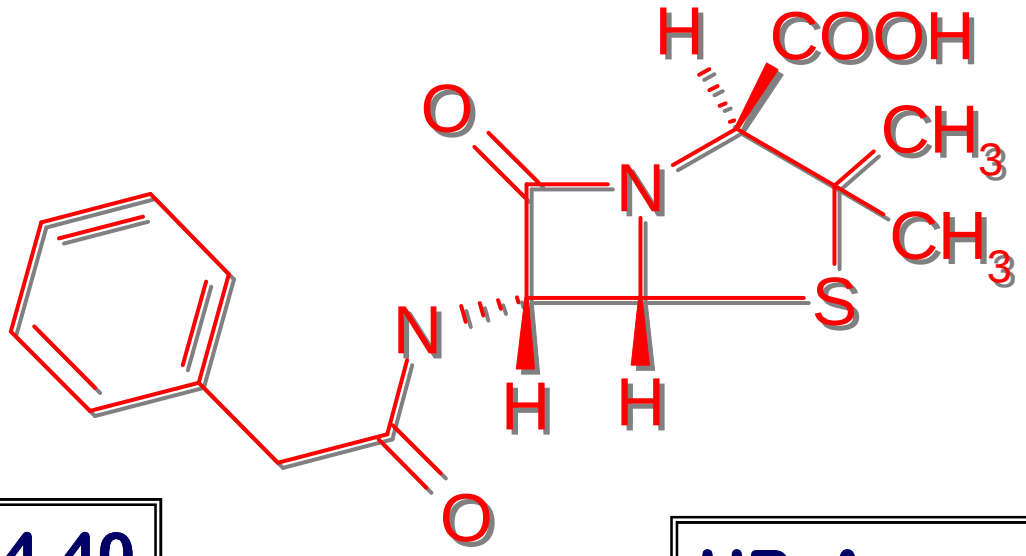
Chemical Space

	TargetA (EC_{50}, μM)	TargetB	TargetC	TargetD	TargetN
Comp.4	0.5±0.1				
Comp.3	1.1±0.1	SORT!			
Comp.1	2.0±0.1				
Comp.5	3.3±0.1				
Comp.2	8.3±0.1				

… please: do not start any quantitative structure-activity relationship if you're not confident of 'reproducibility' of your data of biological activity!!!

EC_{50} = 1.0 μM ≠
1.0 ± 0.1 μM ≠
1.0 ± 0.5 μM ≠
1.0 ± 1.0 μM

NUMB3RS!!!



MW = 334,40

HB_Acc = 4

pKa = 2.7

Volume = 302,37

nC = 16

logP = 0.8

…

We can start filling this table in an other way:

Chemical Space

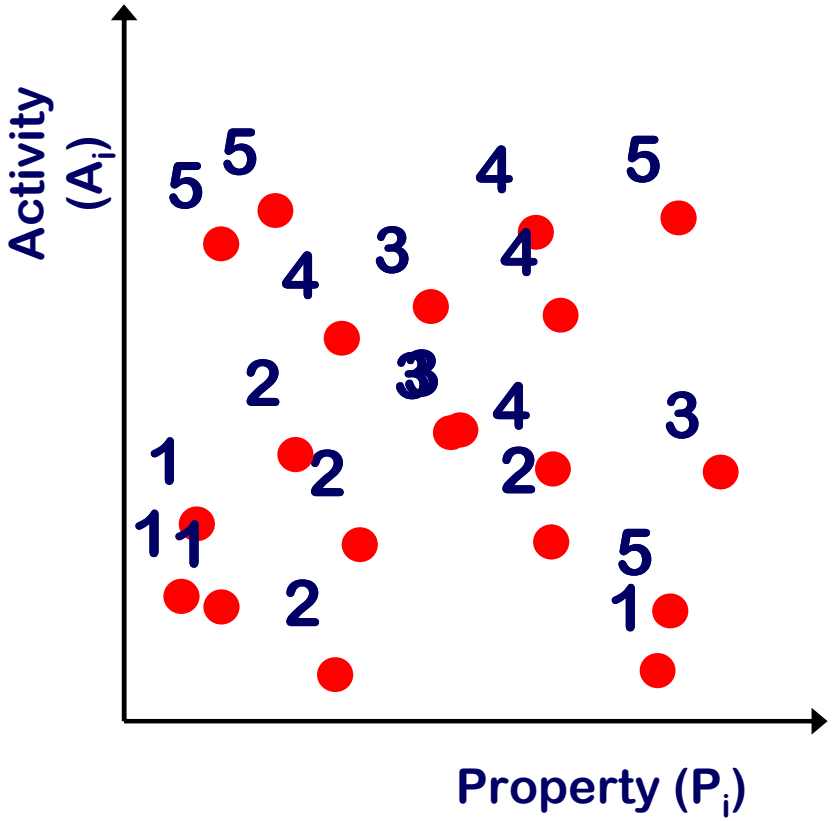
	TargetA (EC_{50}, μM)	Descriptor A	Descriptor B	Descriptor C	Descriptor n
Comp.4	0.5±0.1				
Comp.3	1.1±0.1				
Comp.1	2.0±0.1				
Comp.5	3.3±0.1				
Comp.2	8.3±0.1				

… how many molecular descriptor can be calculated?

DRAGON 6.0 is able to calculate 4885 molecular descriptors!

http://www.talete.mi.it/index.htm

and so....



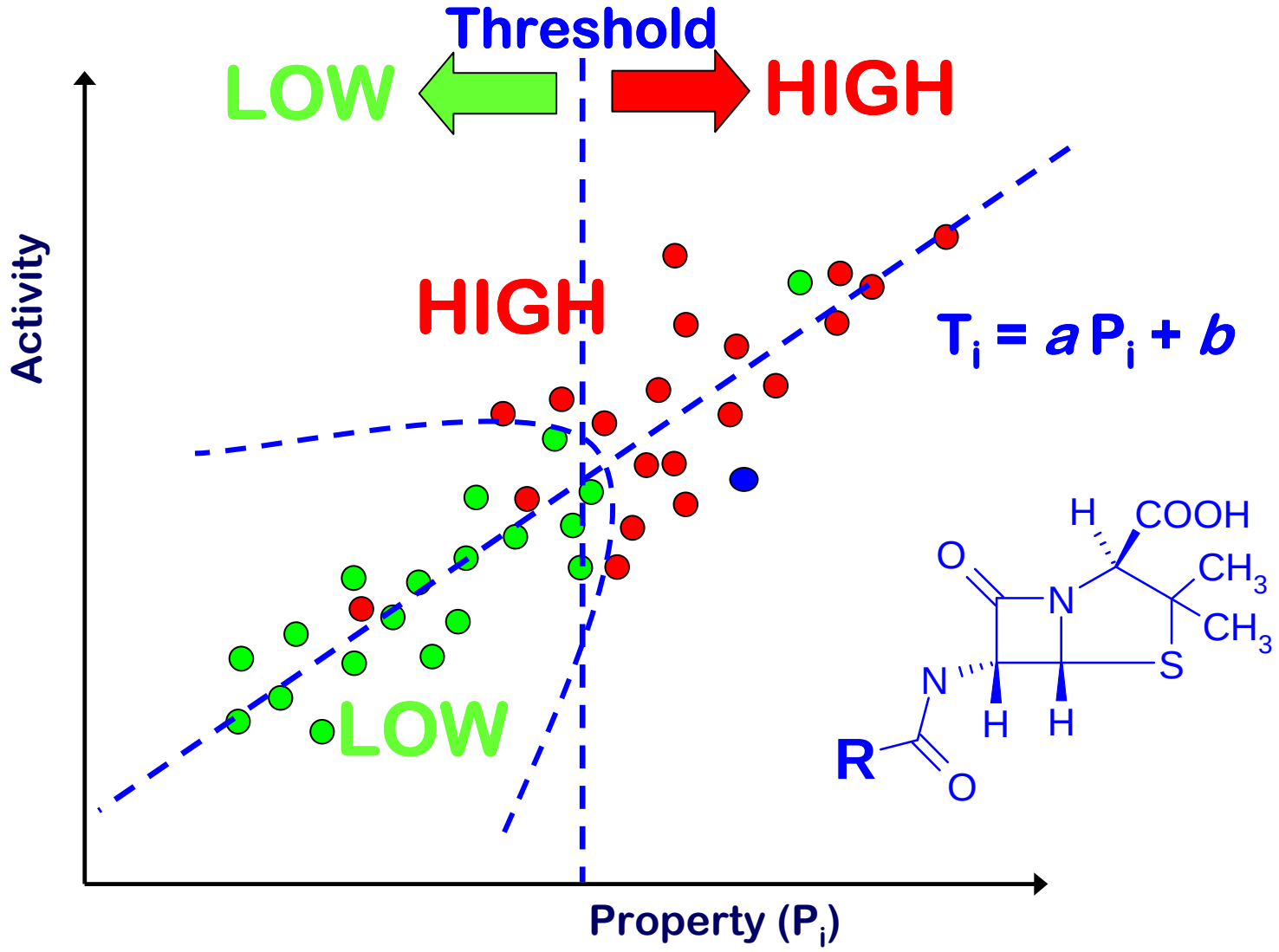
Scatterplot... an interesting place where scouting for patterns!!!

How can we select the good "*molecular descriptor(s)*"?

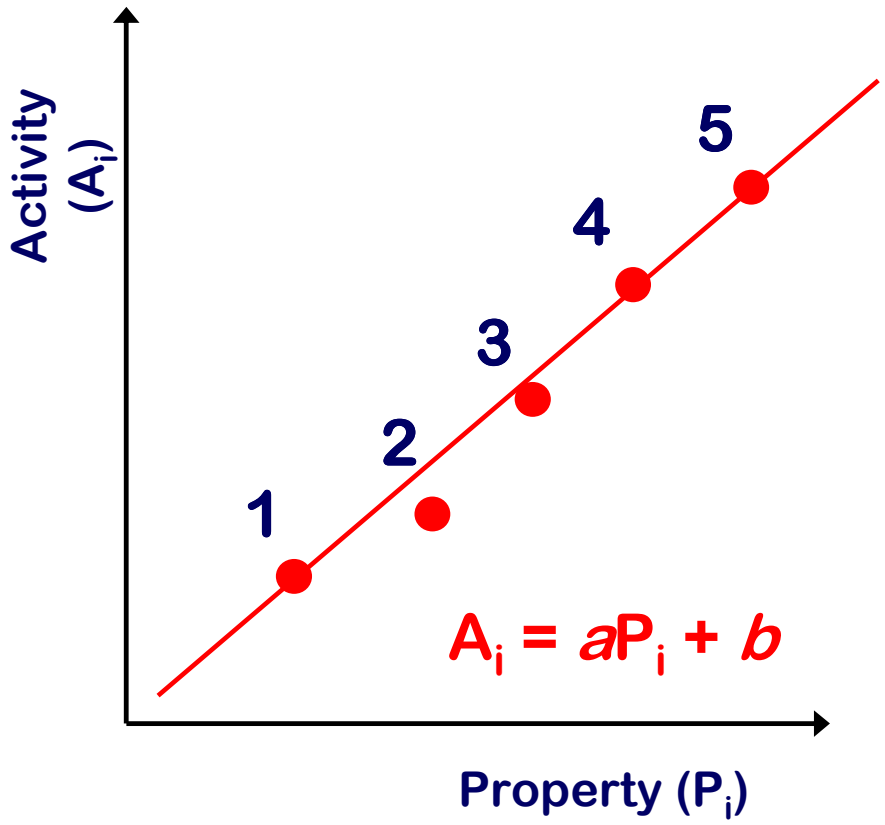
$$f(x)$$

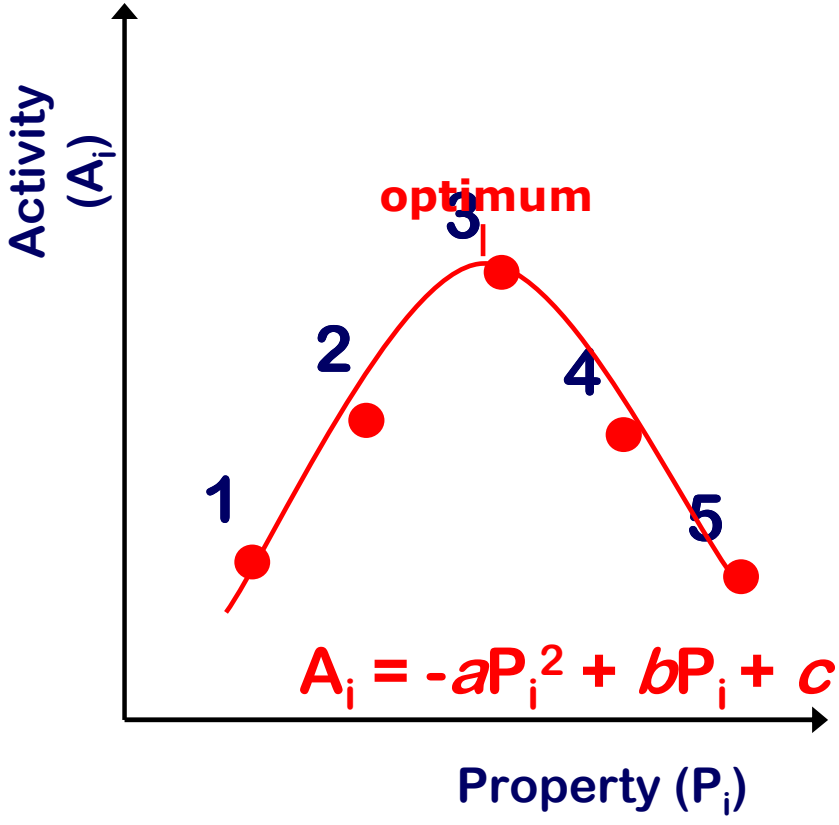
Yes, we can look for "*regularity*" (pattern) between the variability of molecular descriptors and the corresponding variability of experimental activities.

The scatter plot: the best place where explore (Q)SAR.

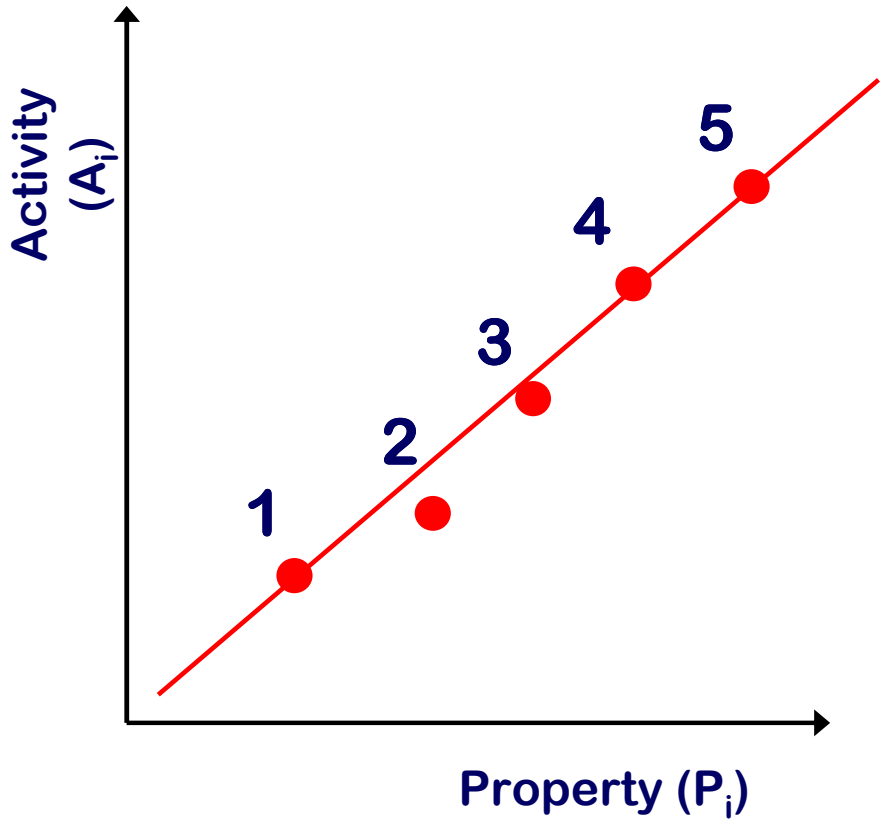


This is the base of a quantitative structure-activity relationship (QSAR): find *patterns*!





The beauty of mathematics:



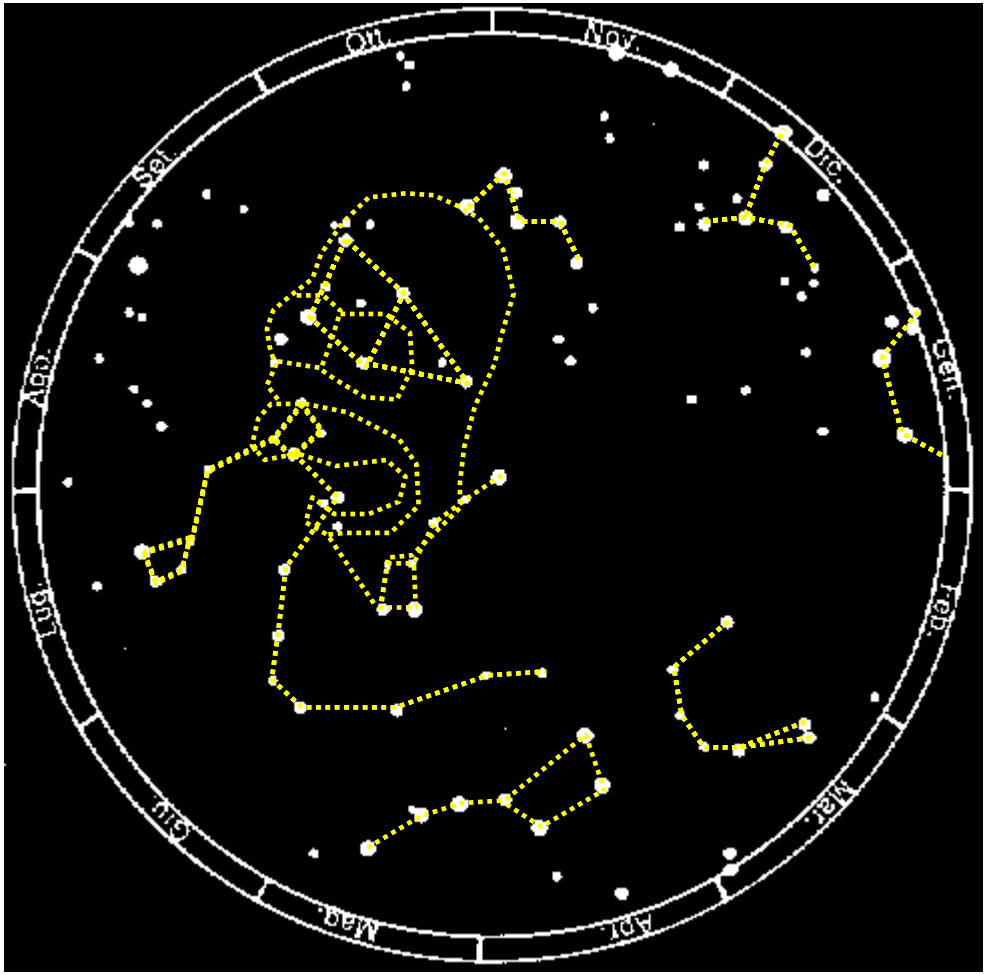
$$A_i = a P_i + b$$

Discrete (few *x-y* correspondences)
Continuum (∞ *x-y* correspondences)

Patterns are beautiful:

- Patterns can be mathematically condensed in equations;
- Pattern can be used to describe relationships among variables;
- Patterns can be used to predict new data;
- Patterns can be used to verify exiting data;

sometimes too beautiful…



The two statistical gold rules do build up linear models:

- For each independent variable (*molecular descriptor*) you need at least five (5) dependent variable values (*a*
- Th nt lues (*a* st b and *pr*

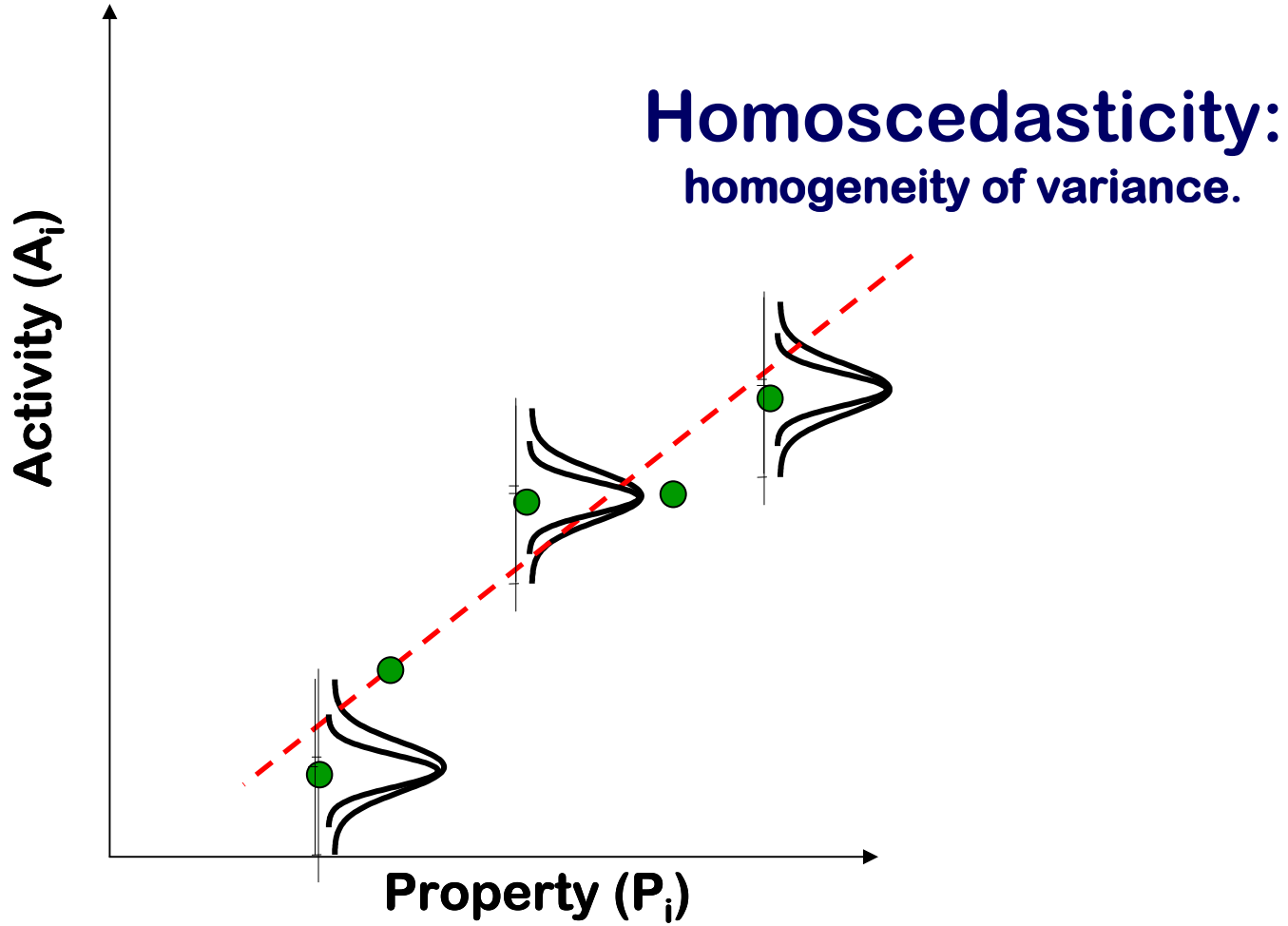
High accuracy, but low precision

High precision, but low accuracy

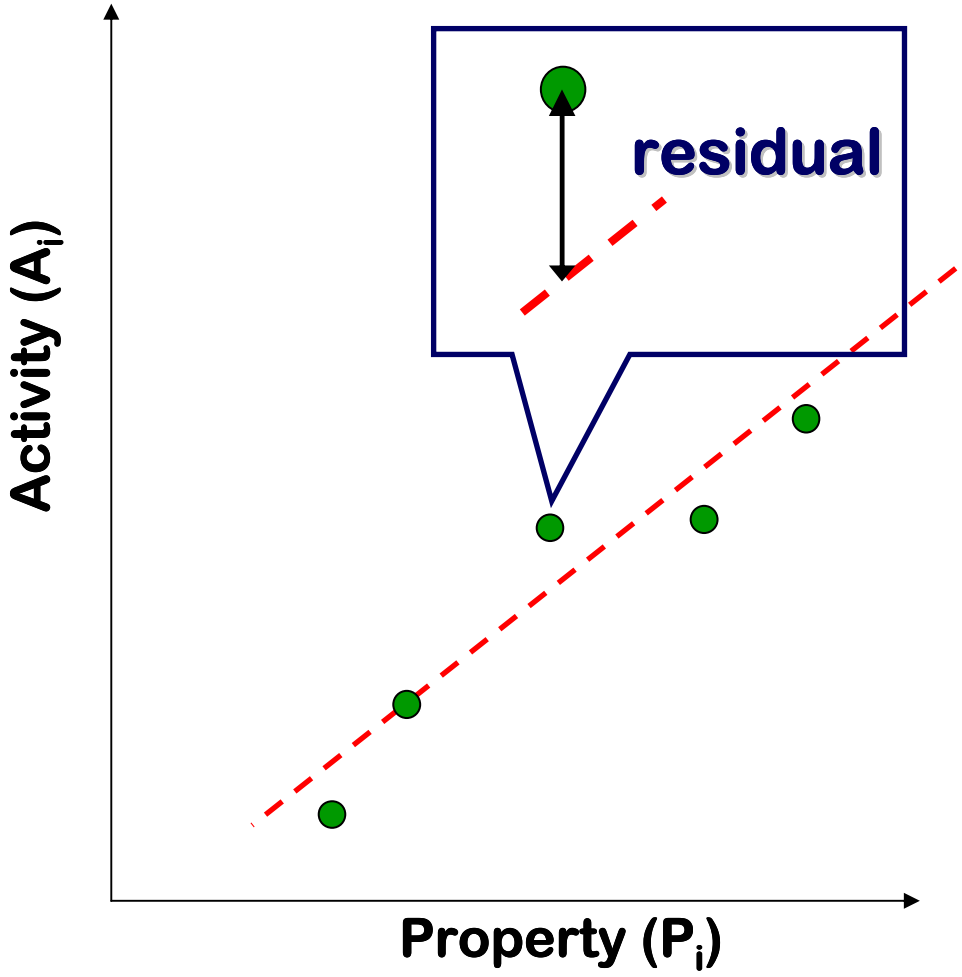
And how can we select the "*good*" linear model:

Homoscedasticity:

homogeneity of variance.



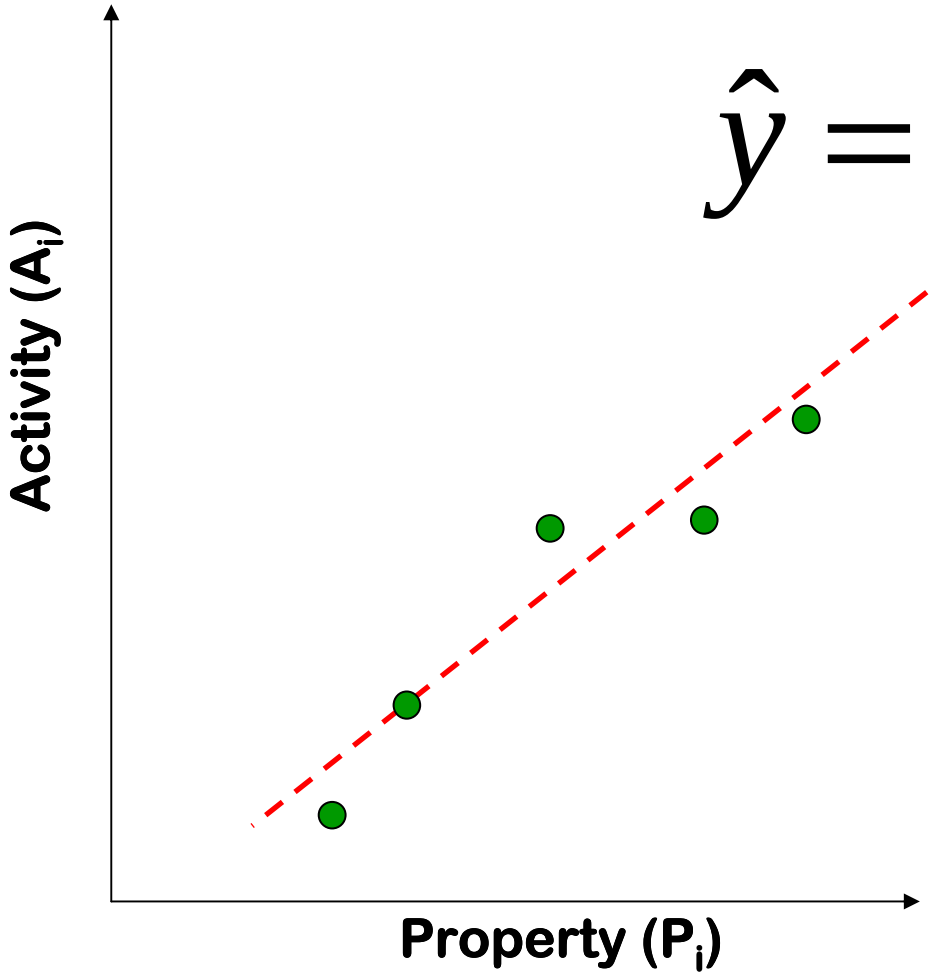
Do you remember… Least Squares Analysis?



LSA is a method for linear regression that determines the values of unknown quantities in a statistical model by minimizing the sum of the residuals, the difference between the predicted (ŷ) and observed values (y) squared.

$$e = \hat{y} - y$$

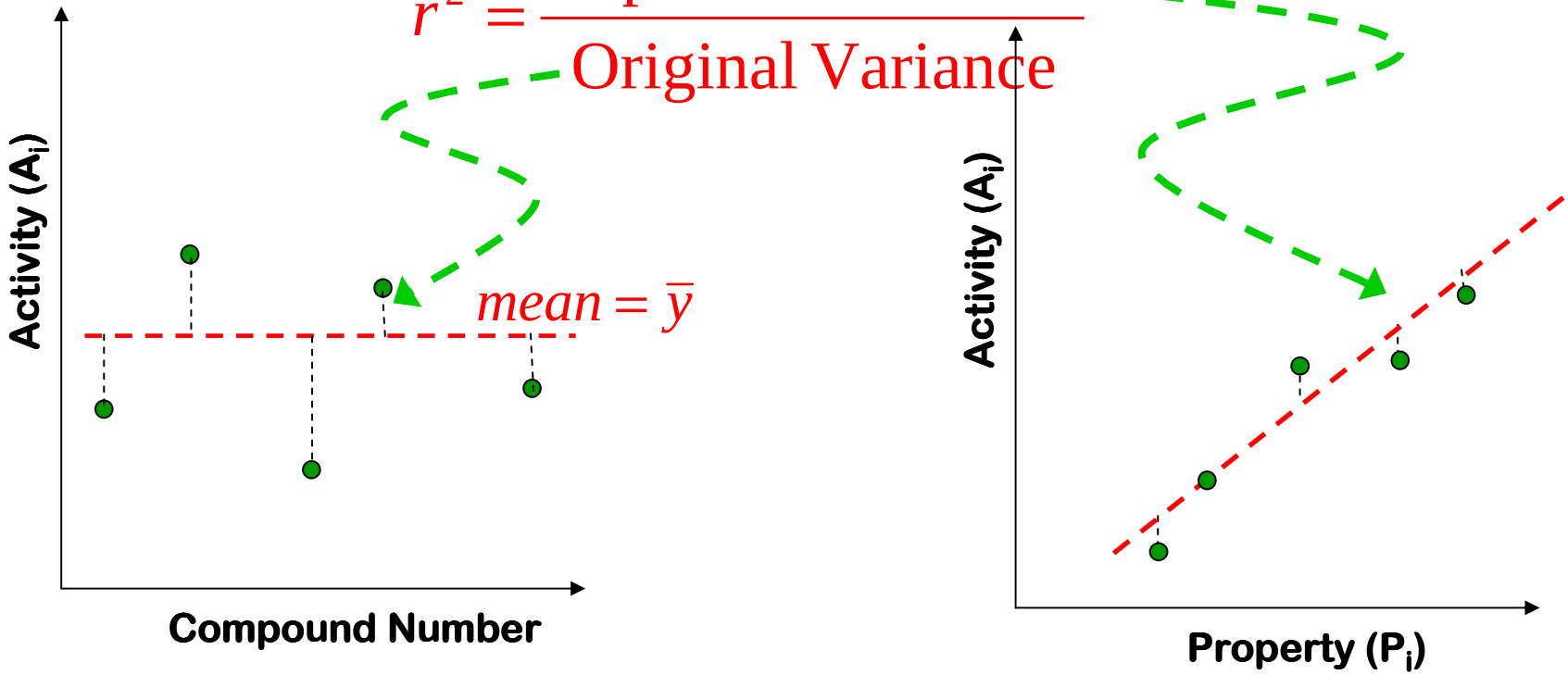
$$\hat{y} = a + bx$$



$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} \qquad a = \frac{\sum y}{n} - b\frac{\sum x}{n}$$

Goodness of fit: variation in the data is quantified by the *coefficient of determination (r^2)* which measures how closely the observed data tracks the fitted regression line. Errors in either the model or in the data will lead to a bad fit. This indicator of fit to the regression line is calculated as:

$$r^2 = \frac{\text{Explained Variance}}{\text{Original Variance}}$$



Original variance = Explained variance (*i.e.*, variance explained by the equation) + Unexplained variance (*i.e.*, residual variance around regression line)

Calculating r^2

- **Original variance:** $$TSS = \sum_{i=1}^{N} (y_i - \bar{y})^2$$
- **Explained variance:** $$ESS = \sum_{i=1}^{N} (y_{i,calc} - \bar{y})^2$$
- **Variance around regression line:** $$RSS = \sum_{i=1}^{N} (y_i - y_{calc,i})^2$$

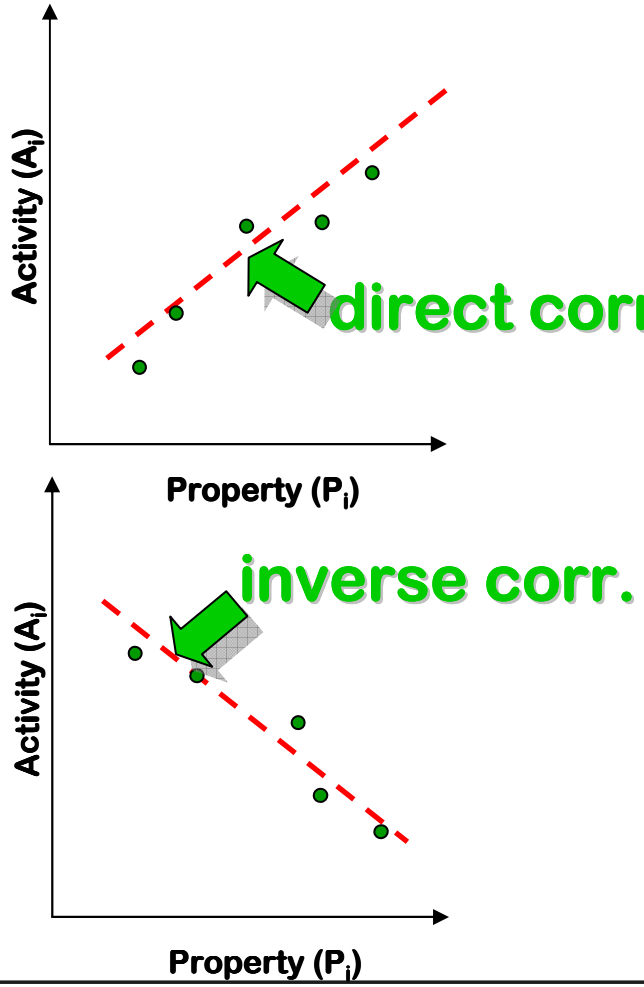
$$r^2 = \frac{ESS}{TSS} \equiv \frac{TSS - RSS}{TSS} \equiv 1 - \frac{RSS}{TSS}$$

$$\mathbf{0 < r^2 < 1}$$

Possible values reported for r^2 fall between 0 and 1. For example: with r^2 of 0.83, you can say that 83% of the variability in activity can be explained by the different value of the selected molecular property. The remaining 17% of variability is due to other unexplained factors.

Goodness of fit: the *Pearson correlation coefficient*

(r) **is the square root of r^2 expressed as a decimal. Its *size* is always between 0 and 1. The *sign* of the correlation coefficient depends on the slope of the regression line:**

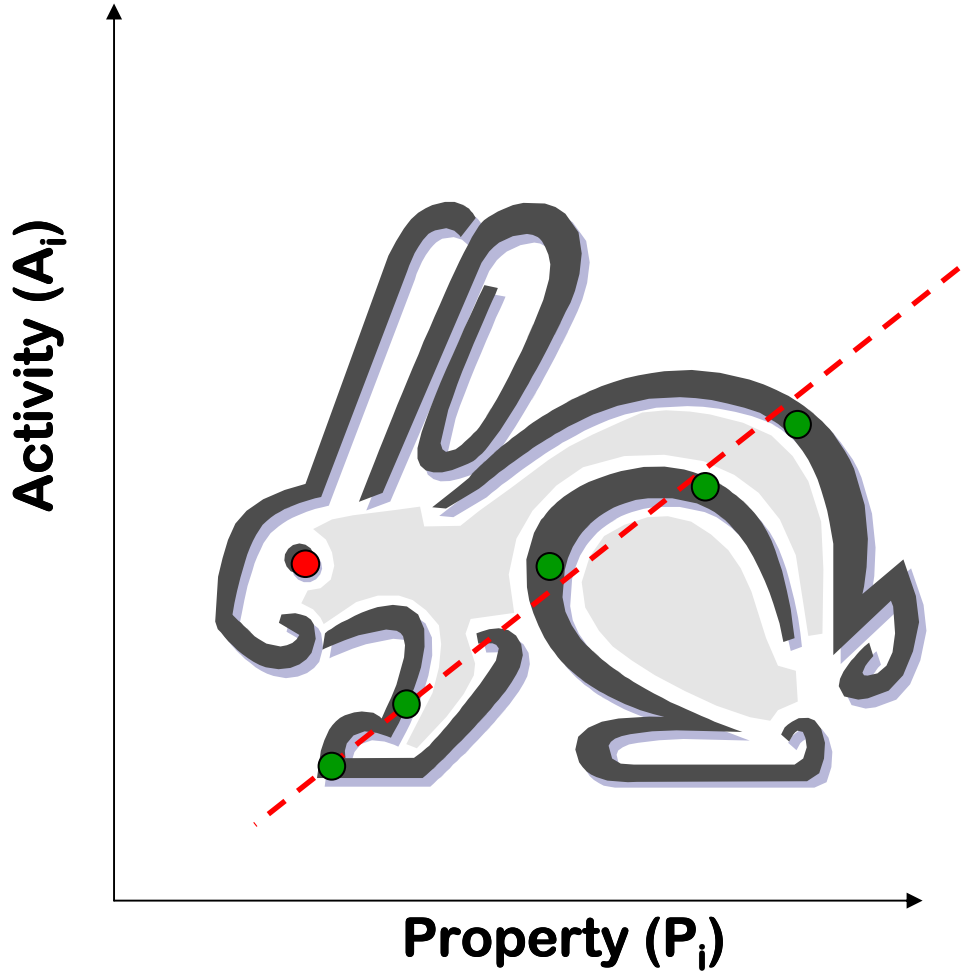


$$r^2 = \frac{ESS}{TSS} \qquad r = \sqrt{\frac{ESS}{TSS}}$$

$$0 < r < 1$$

A perfect correlation of ± 1 occurs only when the data points all lie exactly on a straight line. A correlation greater than 0.8 would be described as strong, whereas a correlation less than 0.5 would be described as weak.

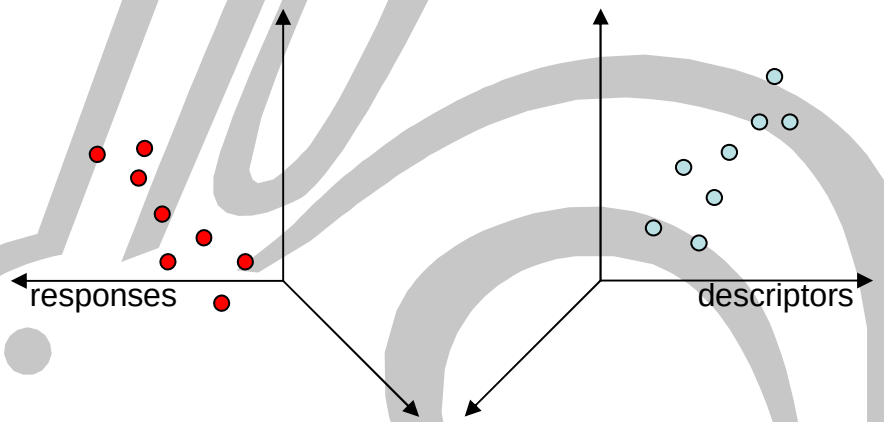
Outliers: an outlier is an observation that is numerically distant from the rest of the data.



Be carefull... the rabbit is out there!!!

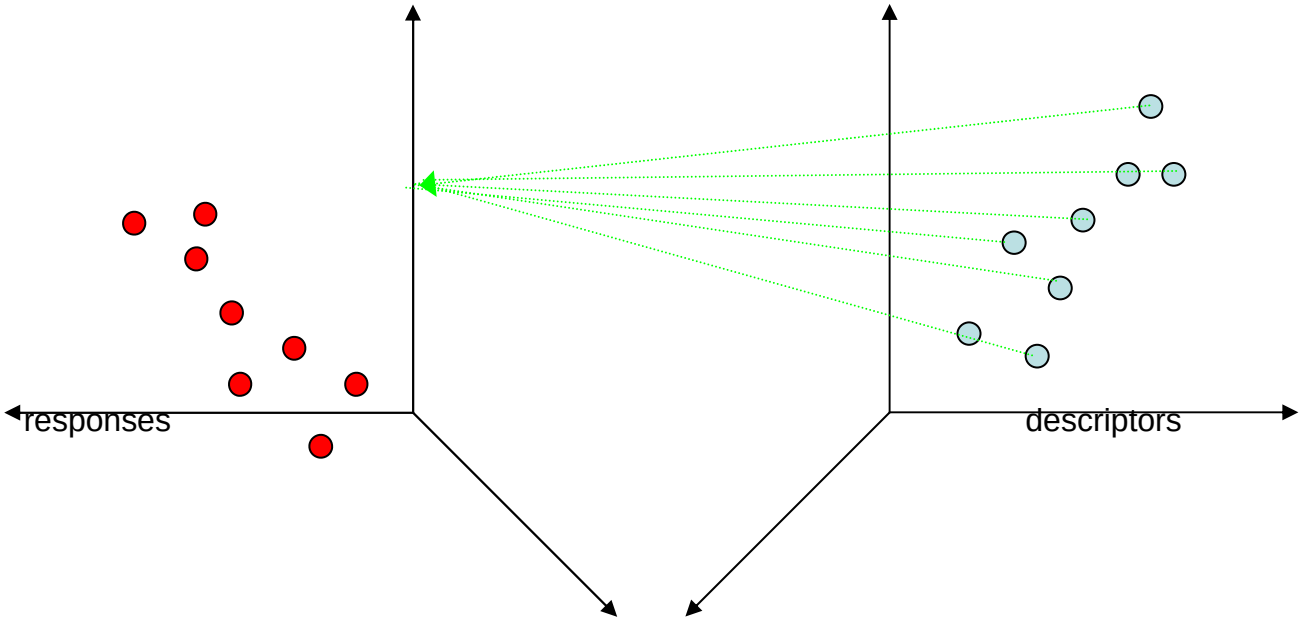
But usually in medchem, we have much more molecular descriptors than activity data. How can we statistically manage this situation?

Multiple Regression Analysis



$$y_1 = a_1x_1 + b_1x_2 + \ldots + m_1x_m + z_1$$
...
$$y_n = a_nx_n + b_nx_n + \ldots + m_nx_m + z_n$$

Multiple Regression Analysis (MRA)



$$y_1 = a_1x_1 + b_1x_2 + \ldots + m_1x_m + z_1$$

$$\ldots$$

$$y_n = a_nx_n + b_nx_n + \ldots + m_nx_m + z_n$$

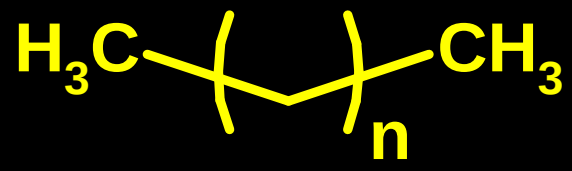
Multiple Regression Analysis (MRA)

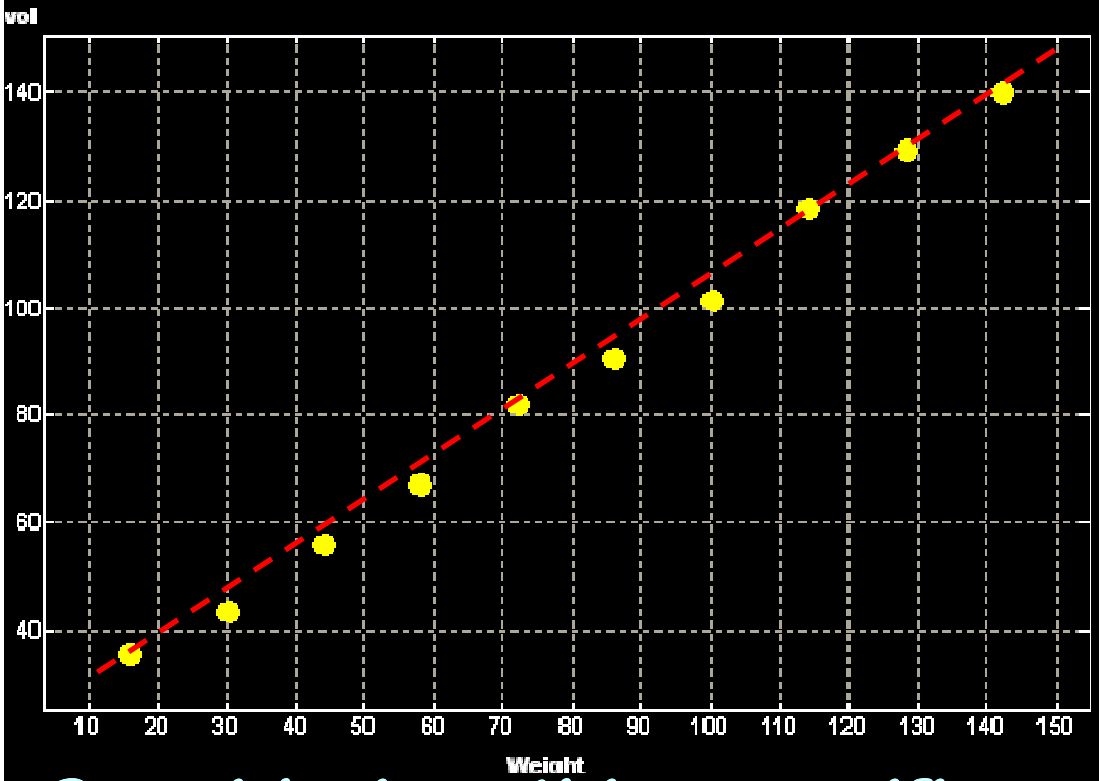
Requirements:

- 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



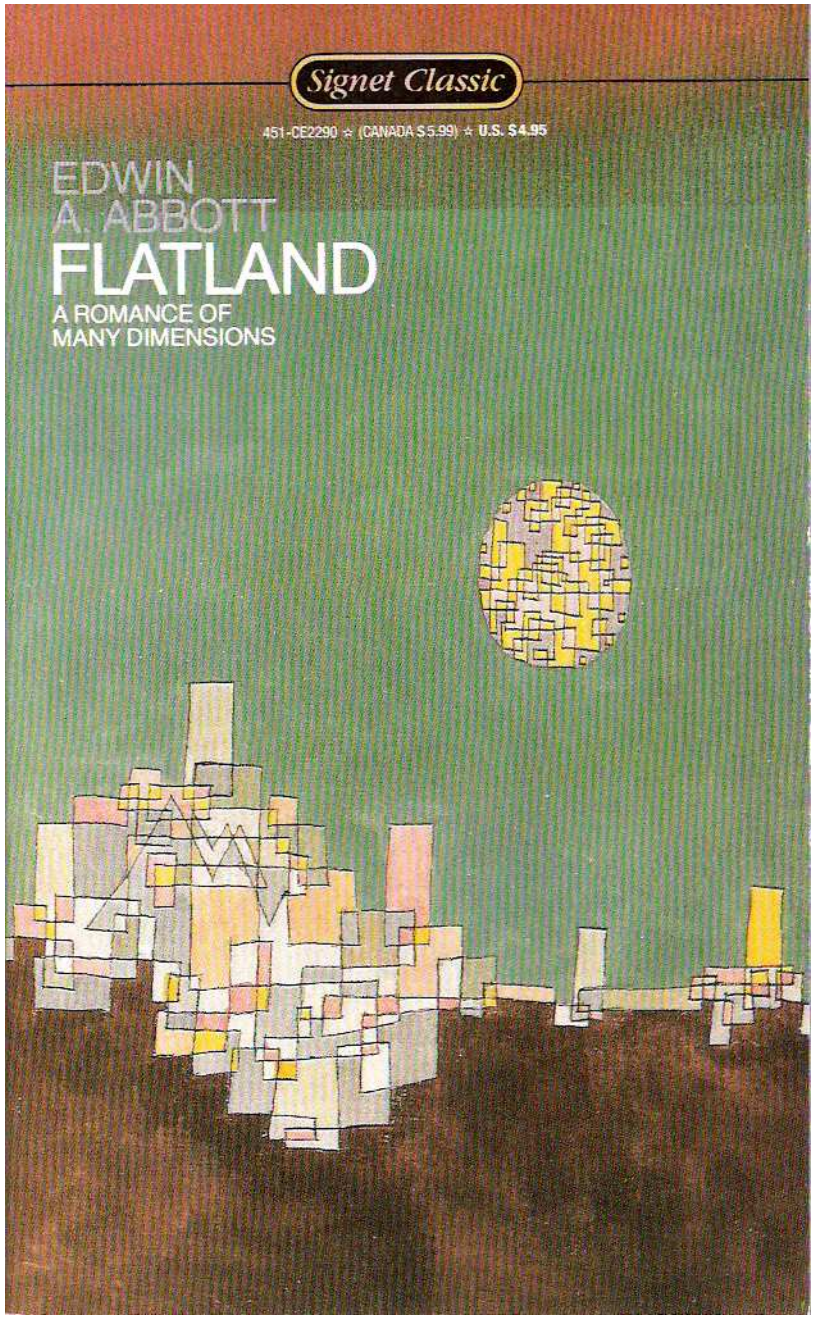


Descriptors:

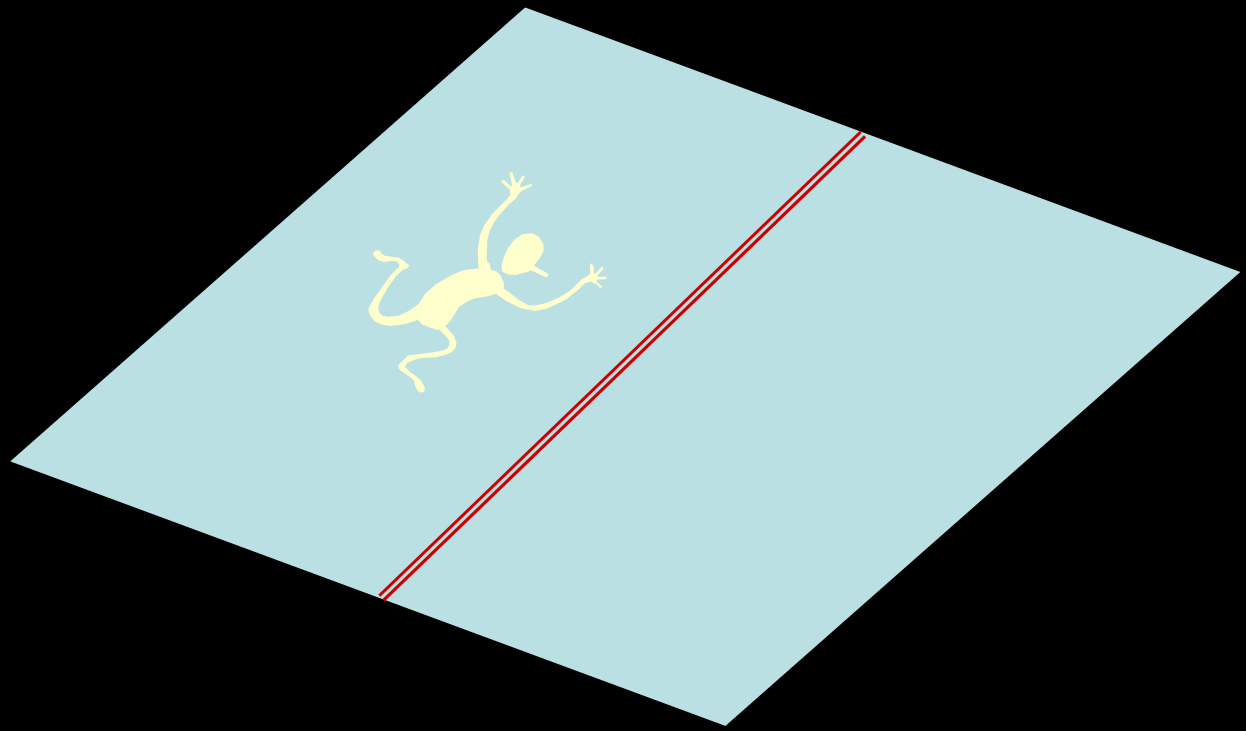
- Molecular Volume
- Molecular Weight

$r^2 = 0.9973$

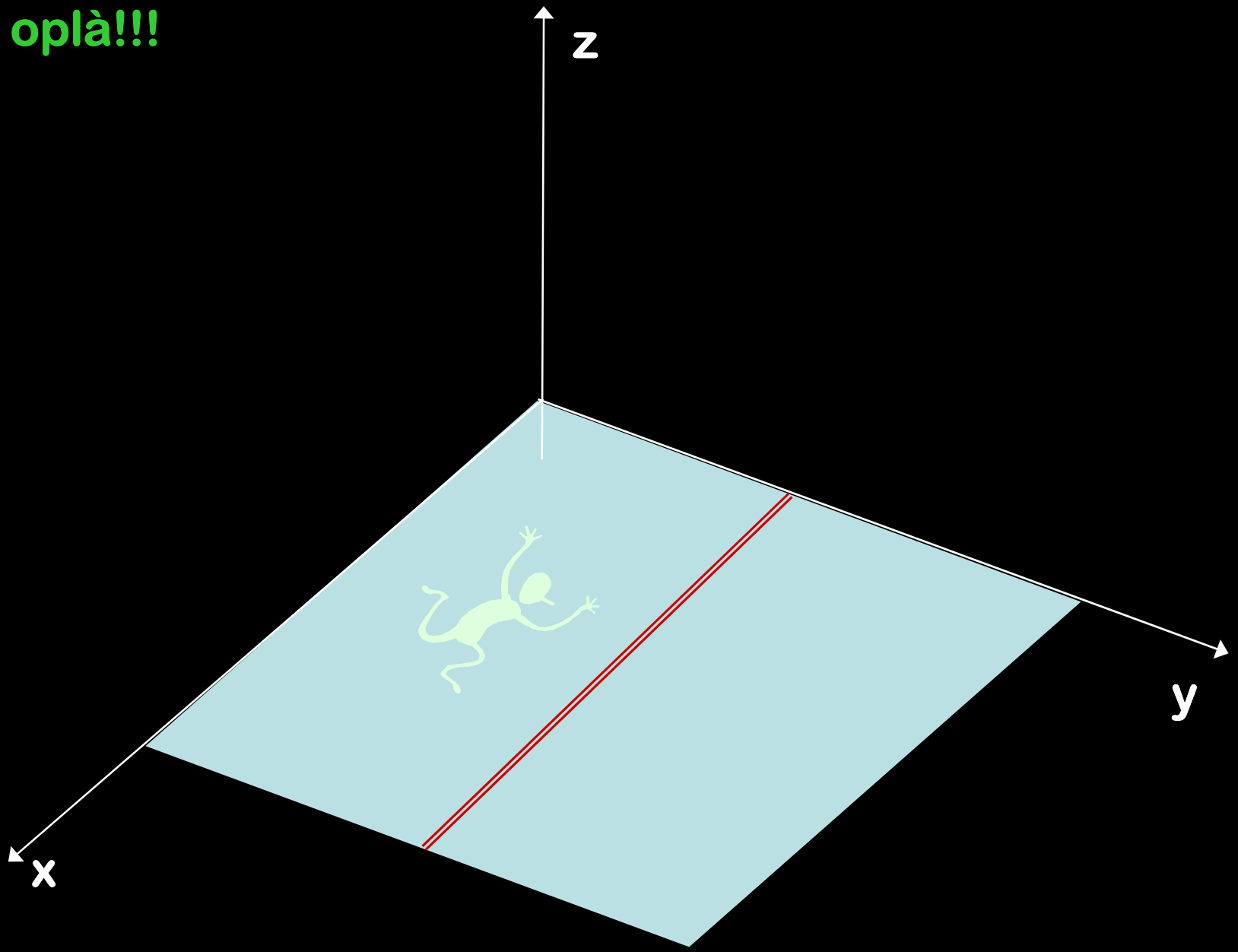
Considering this specific combination of dataset (aliphatic hydrocarbons and molecular descriptors) molecular volume and molecular weight are strongly correlate thus redundant!


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

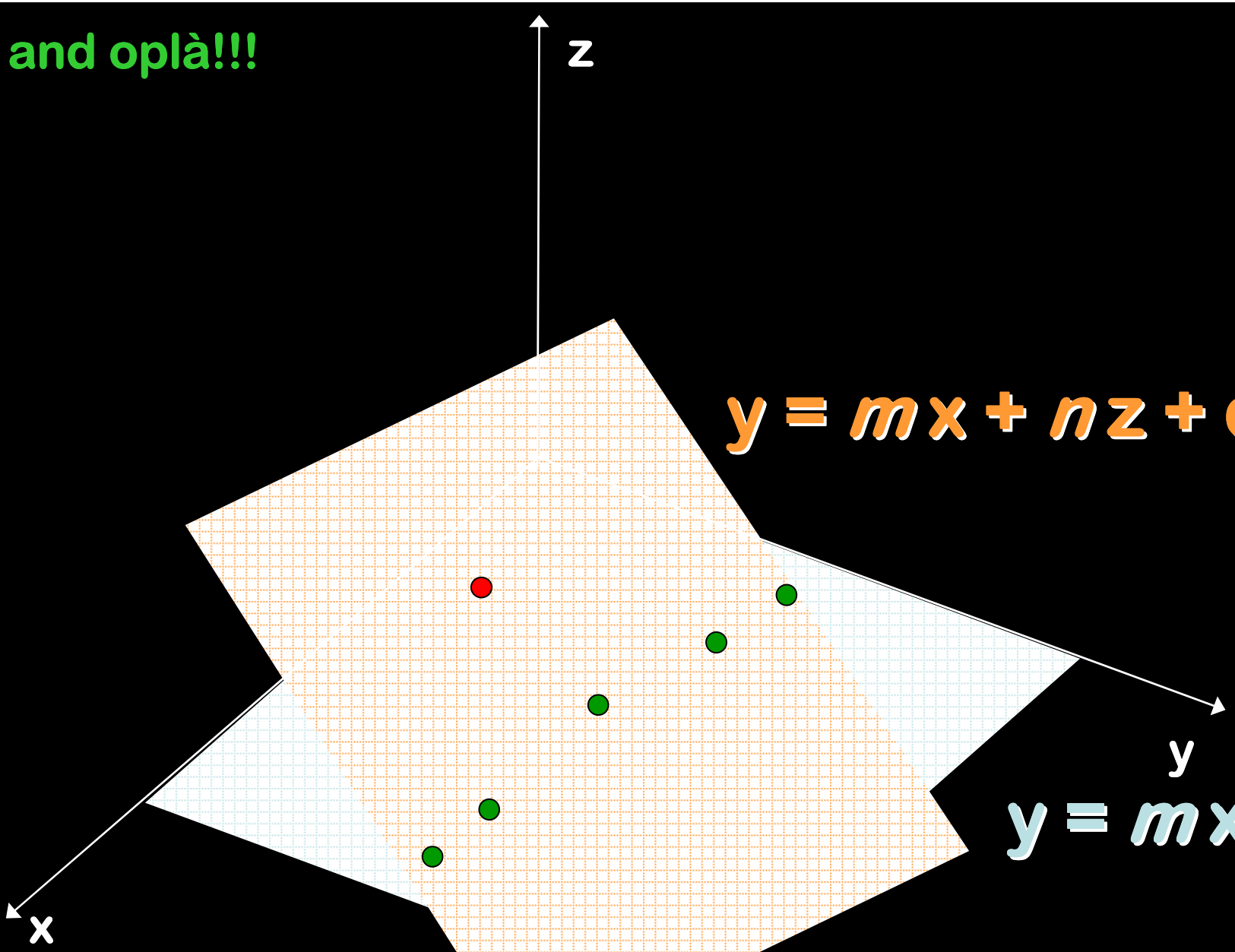
MRA approaches can transform the life in Flatland!



… oplà!!!



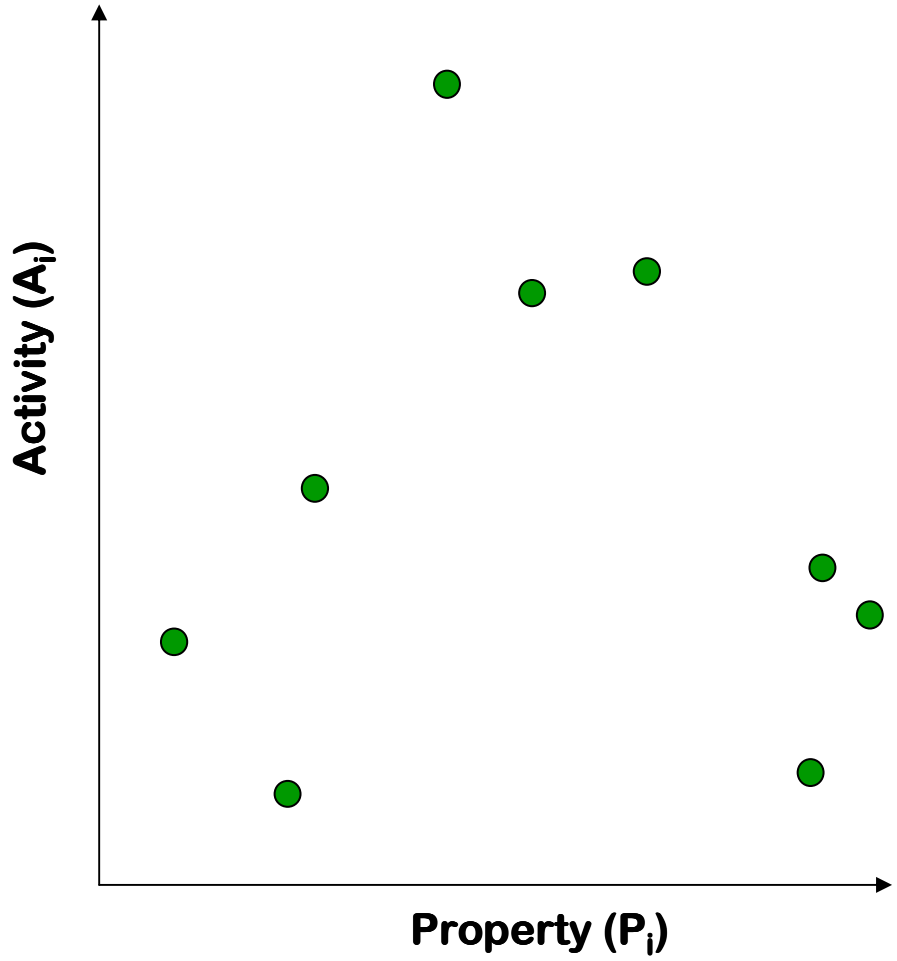
… and oplà!!!



Here is the MRA nightmare:

r^2 will always increase as new descriptors are added.

Cross-validation (CV) for detecting and preventing overfitting!

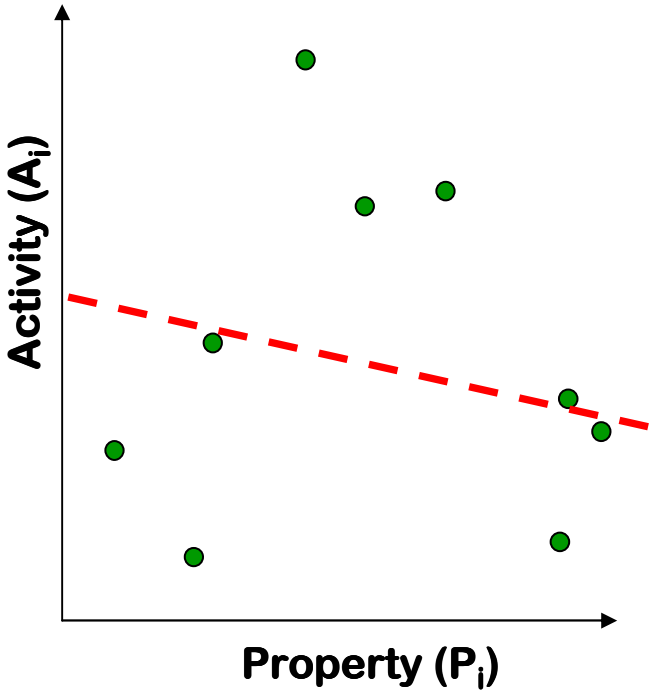


$$y = f(x) + noise$$

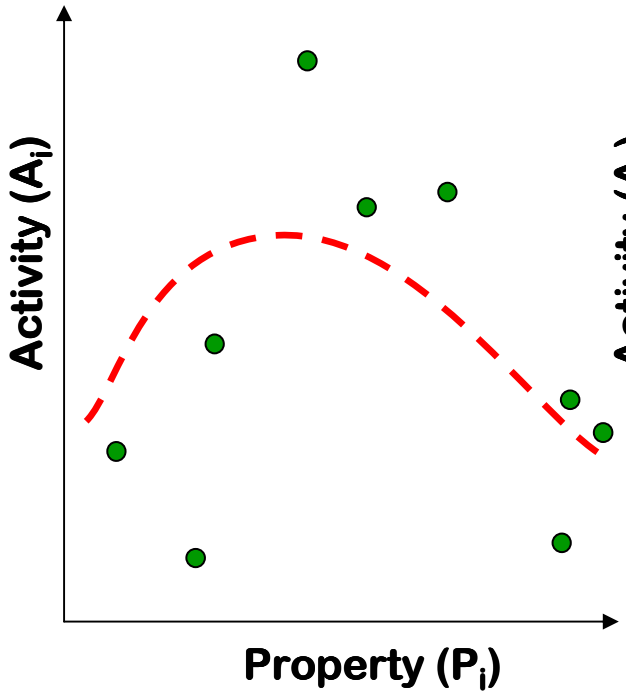
How we can deal with these data?

Let's consider three different methods…

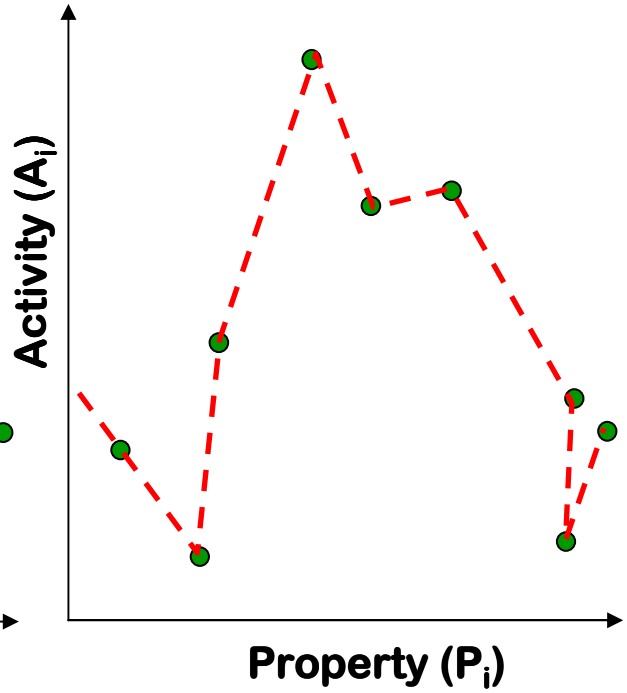
Linear



Quadratic



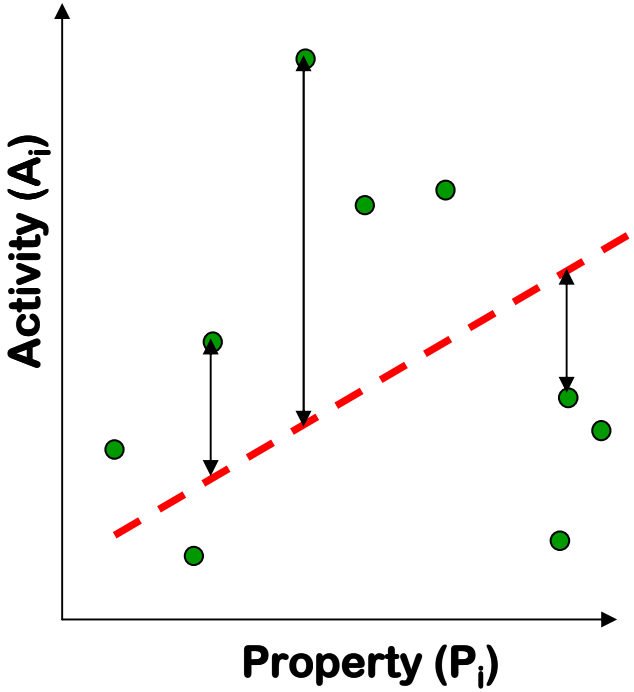
Joint-the-dots



Also known as *piecewise linear non parametric regression*... if that makes you feel better!!!

"How well are you going to predict future data drawn from the same distribution?"

The "*test set*" method...looks this:



1. Randomly choose 30% of the data to be in a *test set*;

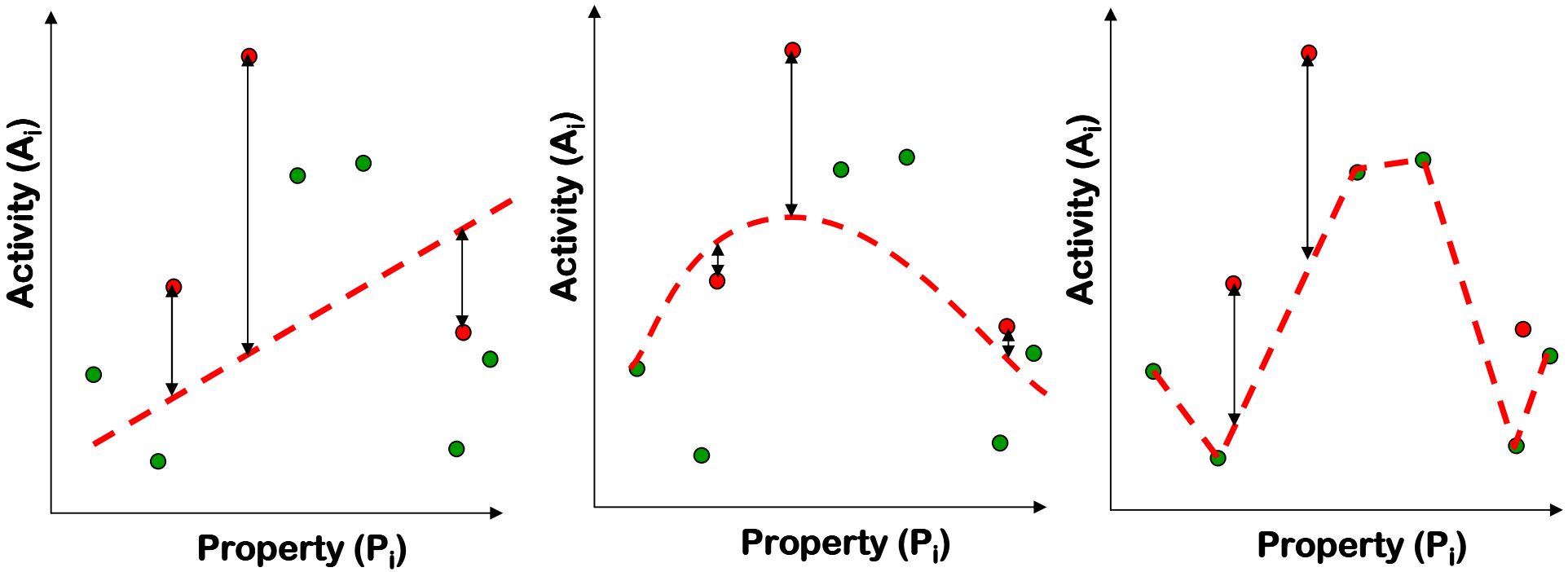
2. The remainder is a *training set*;

3. Perform your regression on the *training set*;

4. Estimate your future performance with the *test set*.

Mean Squared Error (MSE)

$$MSE = \frac{\sum (x_i - \bar{x})^2}{n}$$



MSE = 2.4

0.9

2.2

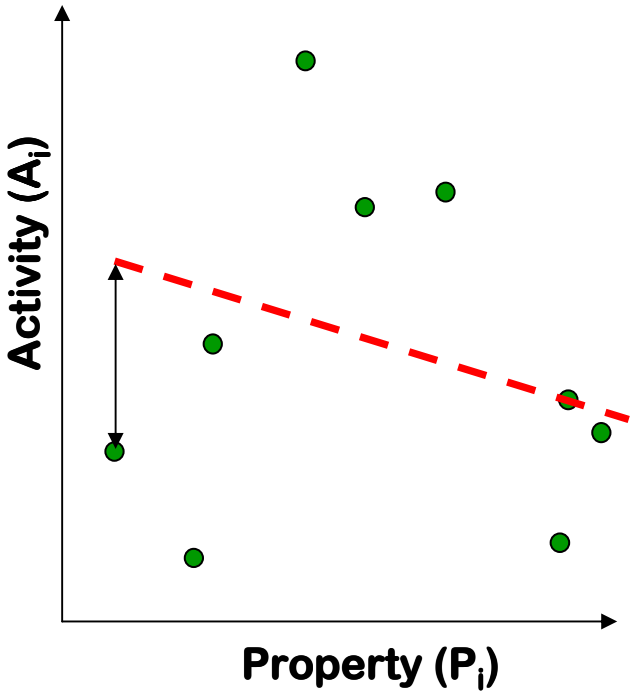
Good news:

- Very very simple;
- Can then simply choose the method with the best "*test set*" score.

Bad news:

- Wastes data: we get an estimate of the best method to apply to 30% less data;
- If we don't have much data, our test-set might just be lucky or unlucky.

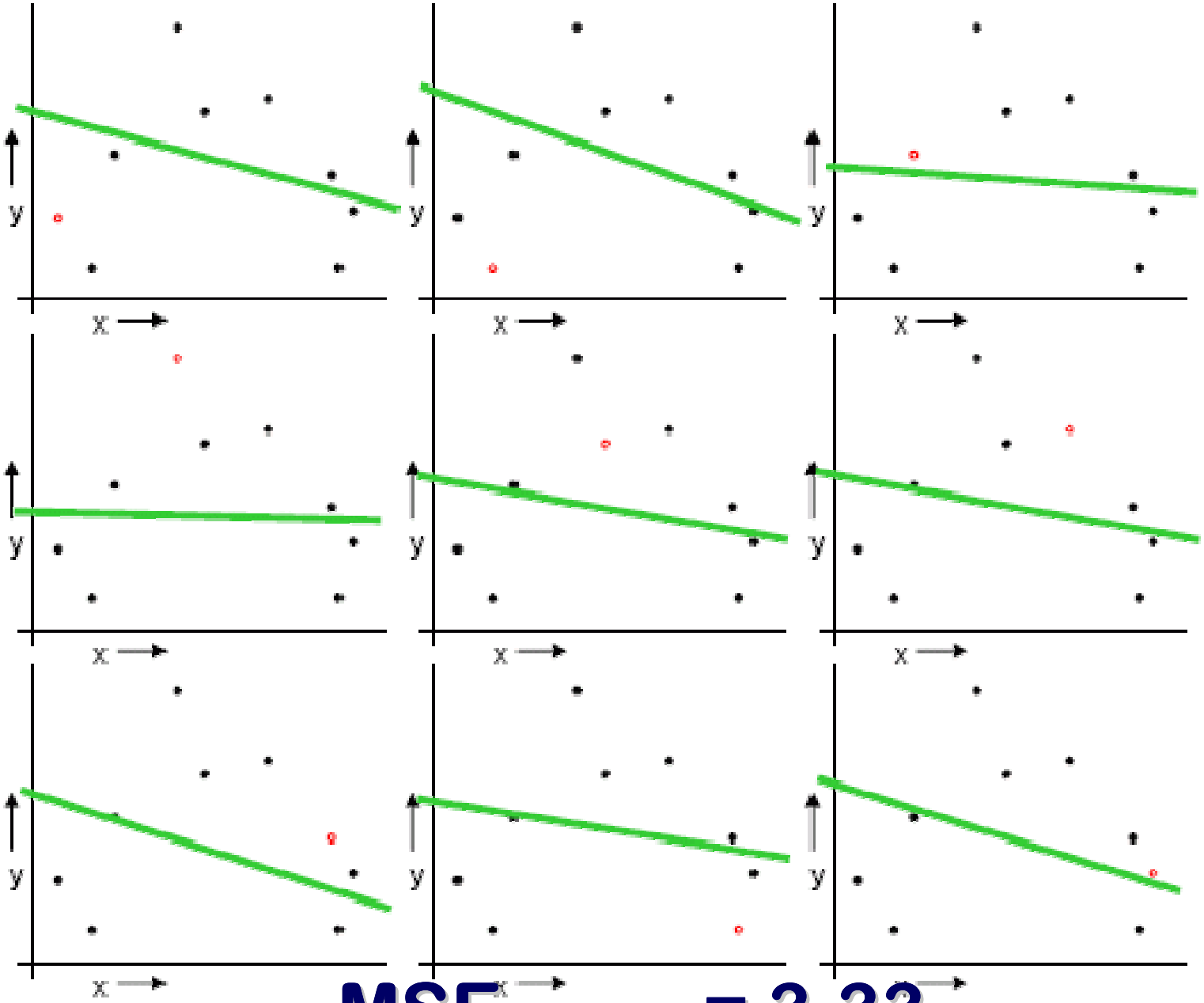
or "*LOOCV*" (Leave-One-Out Cross Validation) method... looks this:



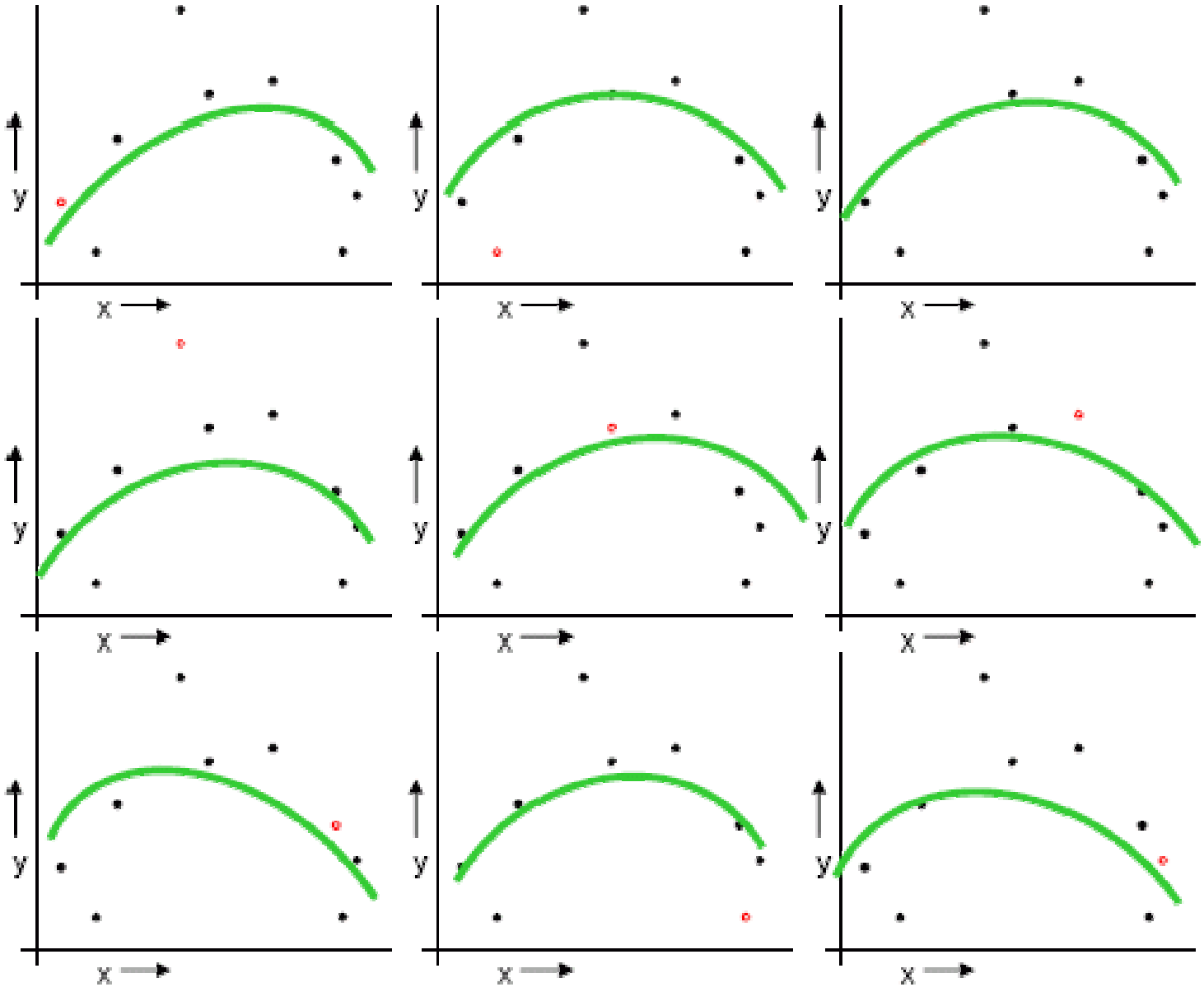
For each data consider this loop:

1. Select the first (x_i, y_i) data;
2. Temporary remove (x_i, y_i) from the data set;
3. Train on the remaining n-1 datapoints;
4. Note your error (x_i, y_i);

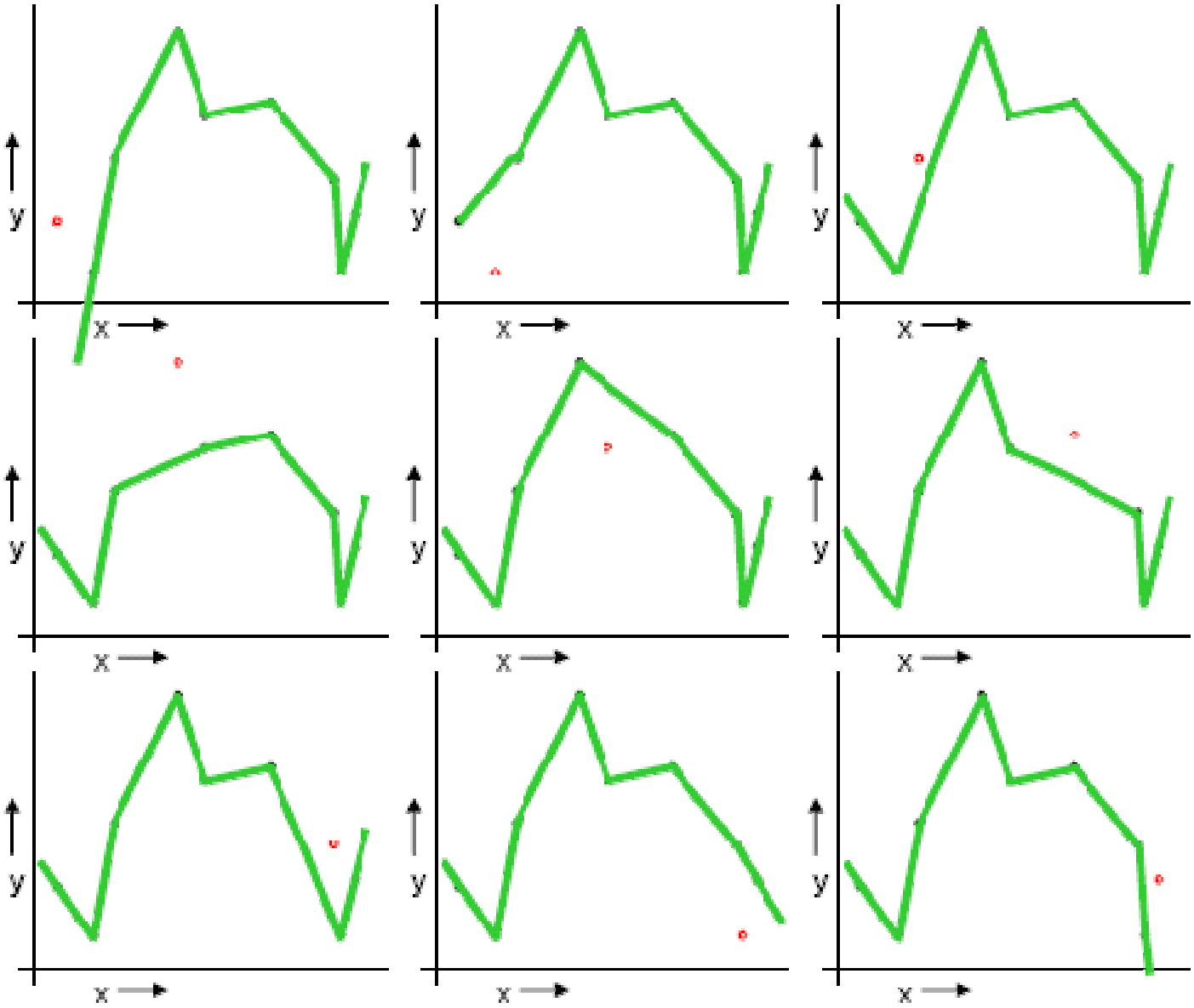
When you've done all points, report the mean squared errors (MSE).



$$MSE_{LOOCV} = 3.33$$



$MSE_{LOOCV} = 0.96$



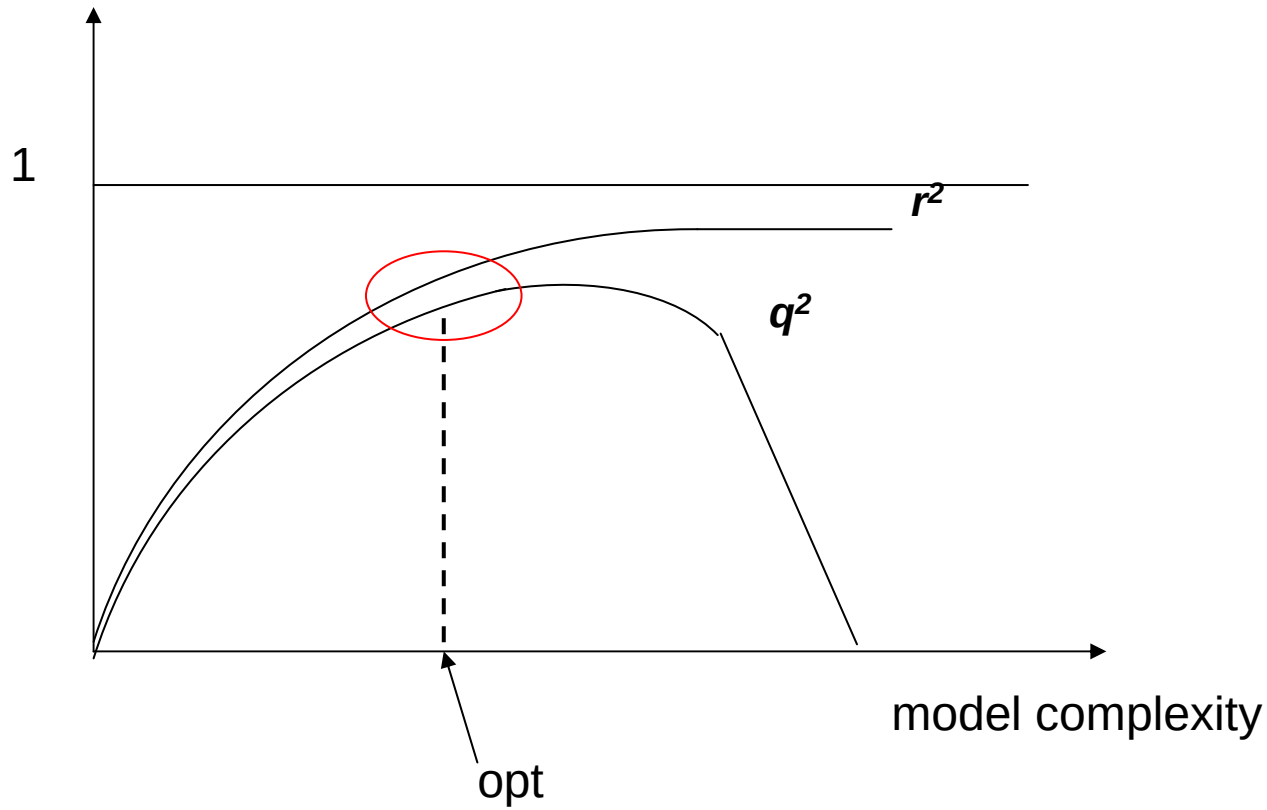
$$MSE_{LOOCV} = 2.12$$

Cross-validation coefficient (Q^2)

$$Q^2 = 1 - \frac{PRESS}{\sum_{i=1}^{N}(y_i - \bar{y})^2}; \qquad PRESS = \sum_{i=1}^{N}\left(y_{pred,i} - y_i\right)^2$$

$$r^2 = 1 - \frac{RSS}{\sum_{i=1}^{N}(y_i - \bar{y})^2}; \qquad RSS = \sum_{i=1}^{N}\left(y_{calc,i} - y_i\right)^2$$

Q^2 initially increases as more parameters are added but then starts to decrease indicating data over fitting. Thus Q^2 is a better indicator of the model quality.


1
r²
q²
model complexity
opt

So.... which kind of validation?

	Downside	Upside
Test-set	Variance: unreliable estimate of future performance	Time cheap
Leave-one-out	Time expensive. Has some weird behaviour	Doesn't waste data

Another important consideration using MRA technique:

#	MR	logP	Volume	PM	Surface	density	n. X atoms
CH2Cl2	1,62959	1,30436	67,1359	84,933	176,1169	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

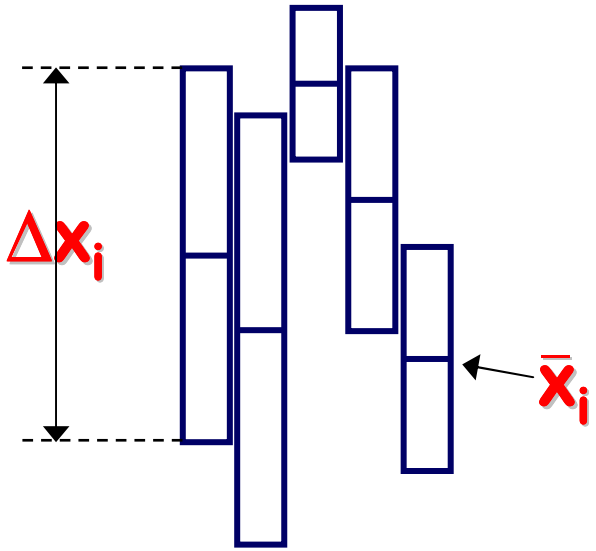
$\bar{x}_i =$ 2.37 2.28 93.77 145.80 211.13 1.99 5.10

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

data scaling and data centering

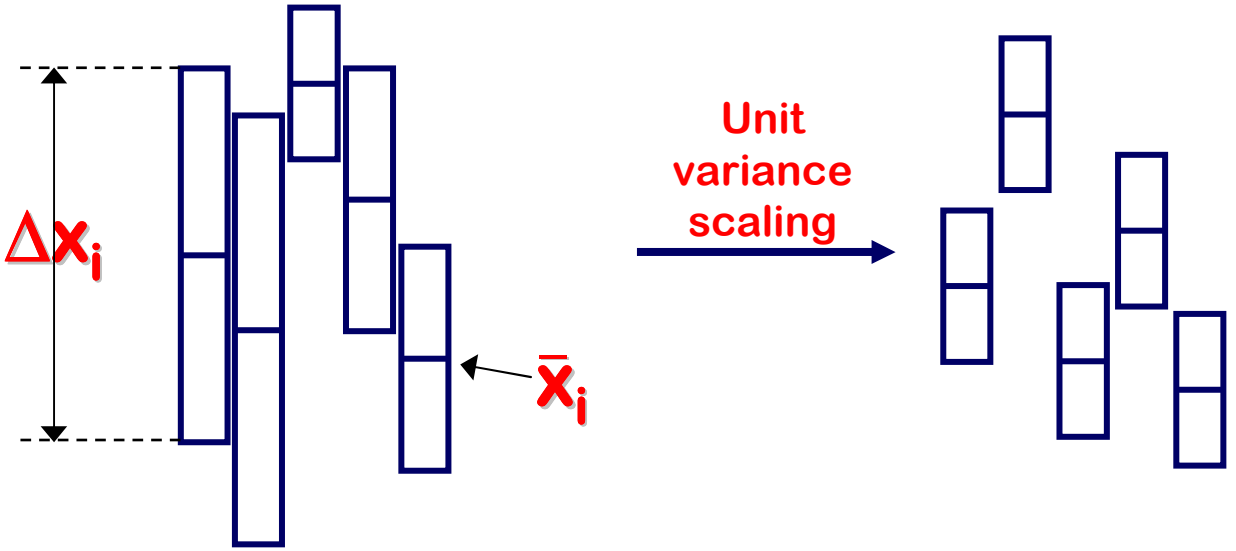
data scaling and data centering

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



data scaling and data centering

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



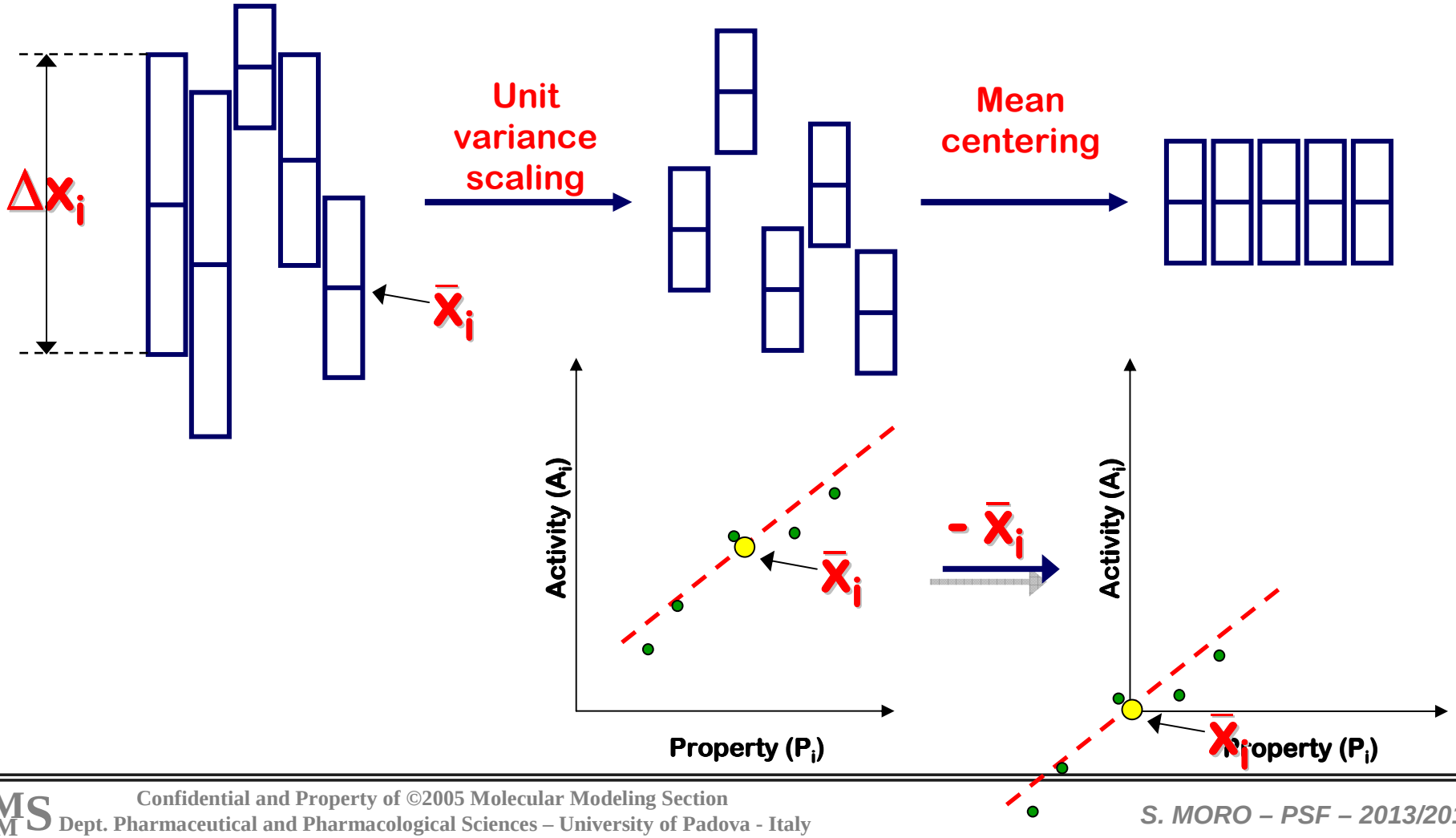
$$\sigma_i = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$ where *n* is the number of data taken

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

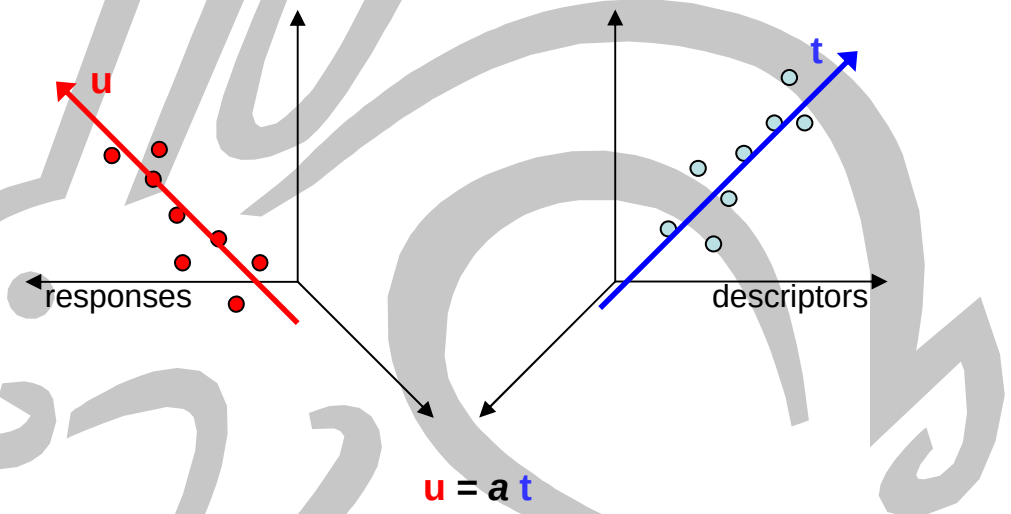
- ***Mean centering***: subtract from each column its average value.



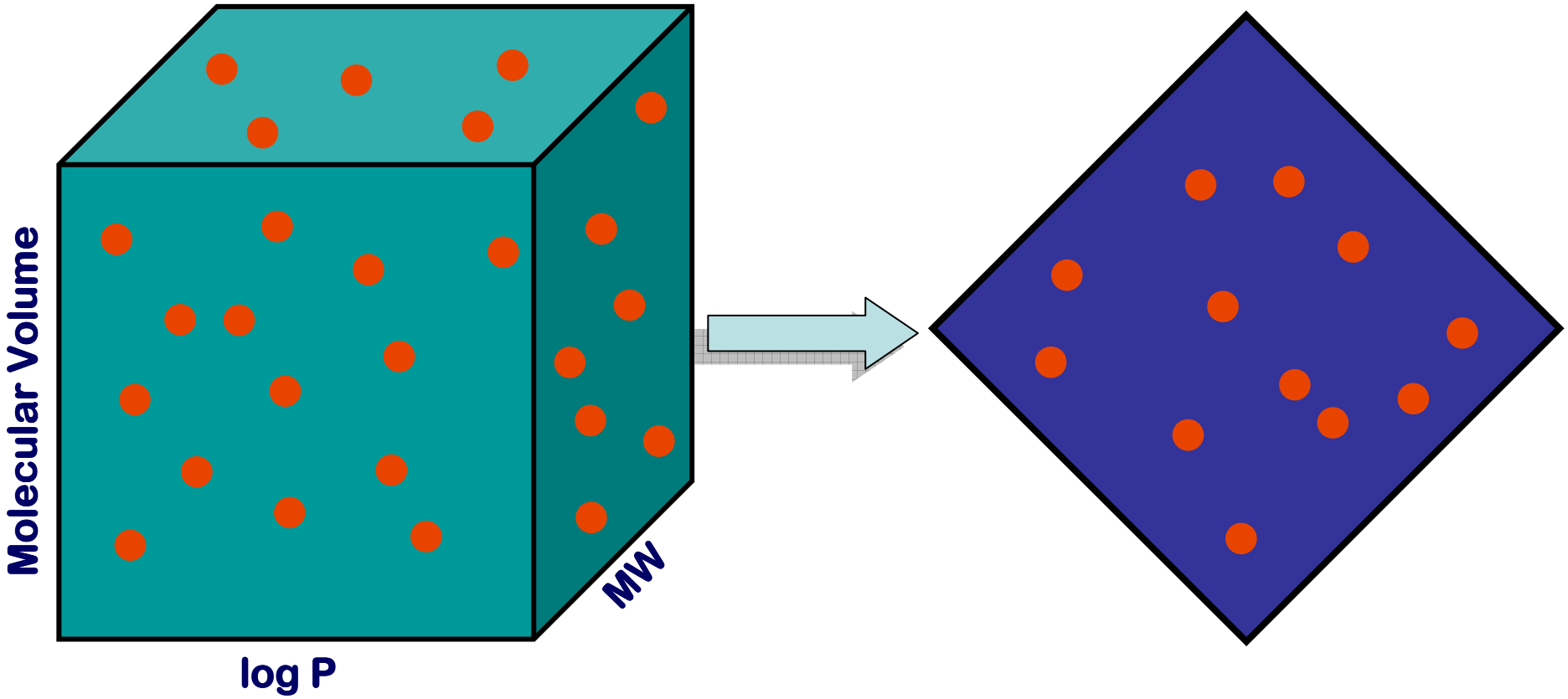
MRA should be a suitable tool only if these criteria are respected:

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.

Principal Component Regression



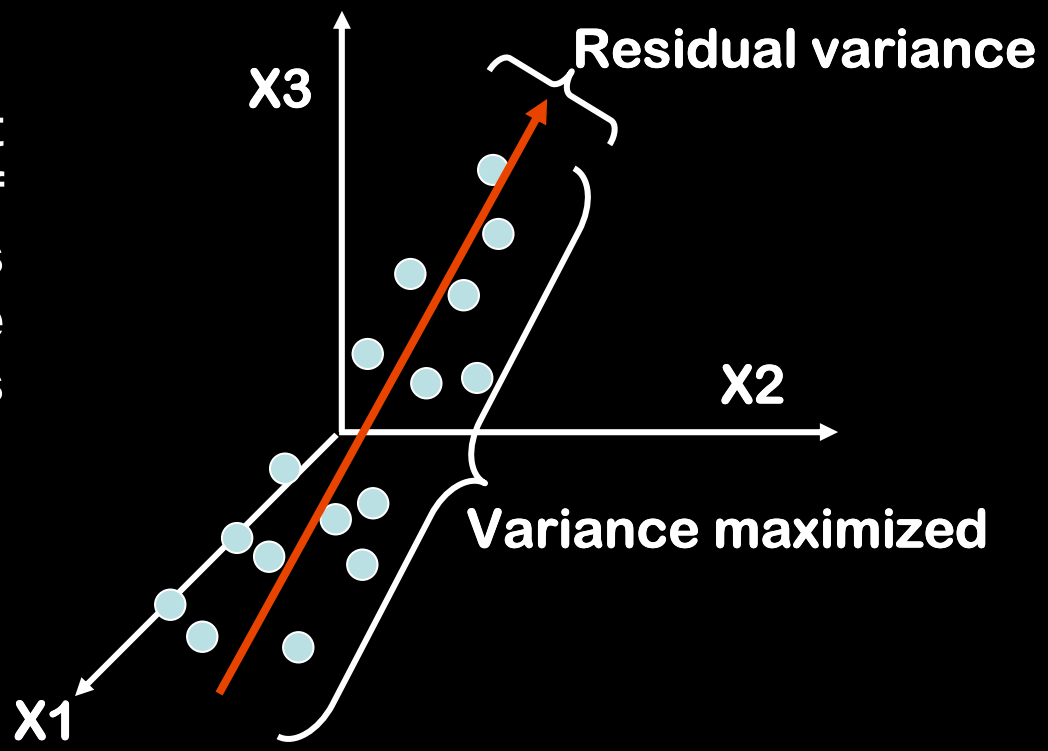
Data Presentation: Property Space



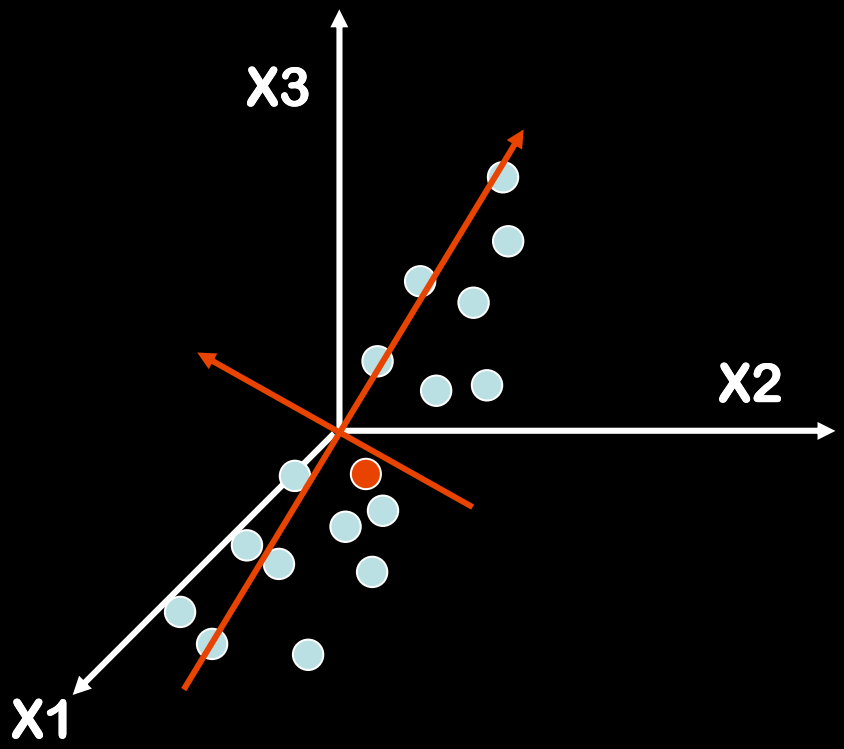
Principle Component Analysis (PCA)

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.



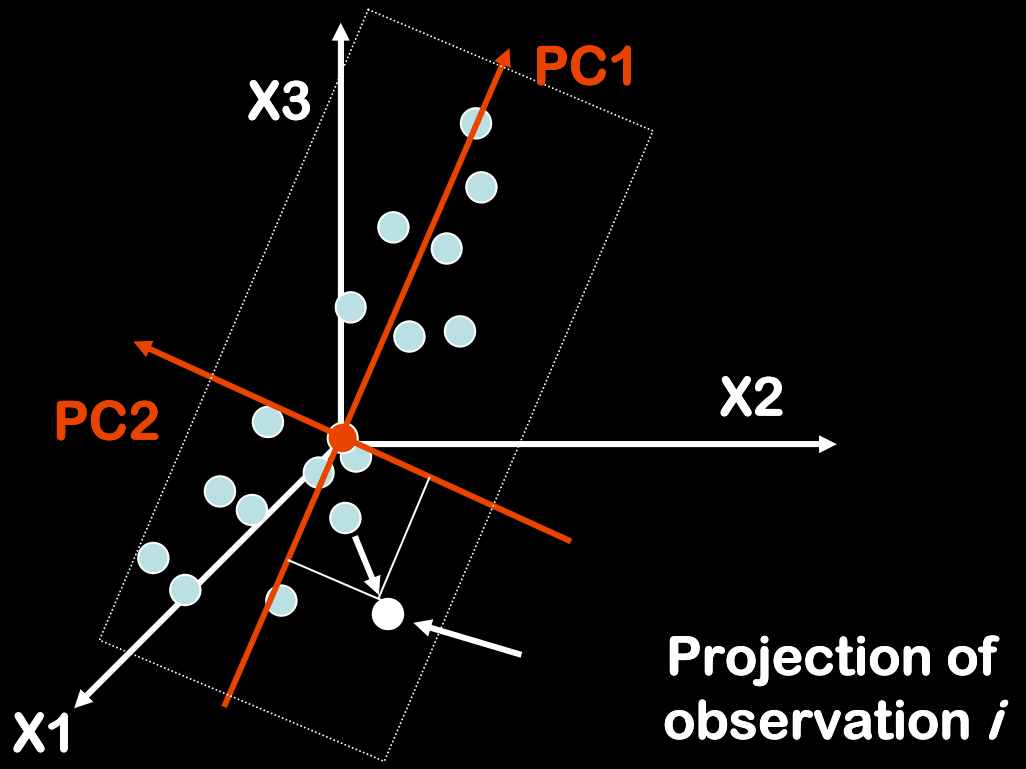
A Geometrical Interpretation of PCA



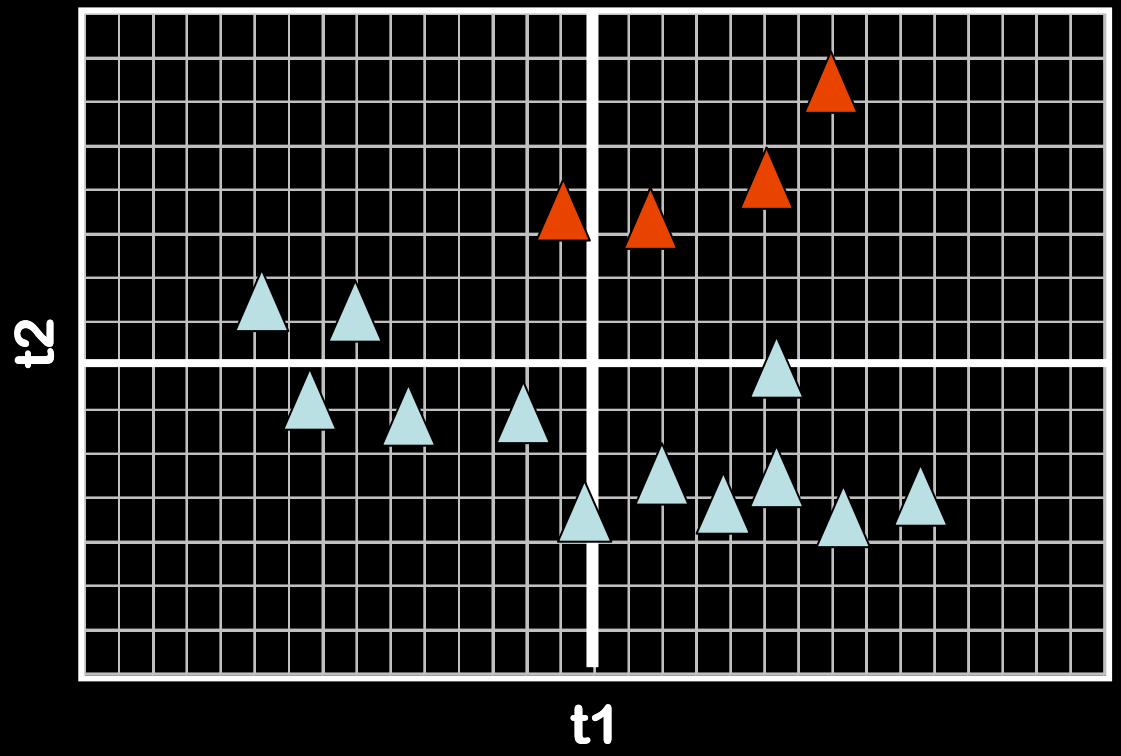
1. Set up k-dimensional space;
2. Plot point;
3. Plot vector of averages at the center of gravity;
4. Mean-center the data;
5. **Generate the first PC:**
   Passes through the origin
   Best approximates the data in a least squares sense
6. **Generate the second PC:**
   Passes through the origin and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

A Geometrical Interpretation of PCA

- First 2 PC's define a plane in the original K-dimensions space.
- By projecting all data points into this plane it is possible to visualize the structure of the data set.
- ***Scores (t)*** are the coordinates of the original data points in this plane.



Scoring Plot



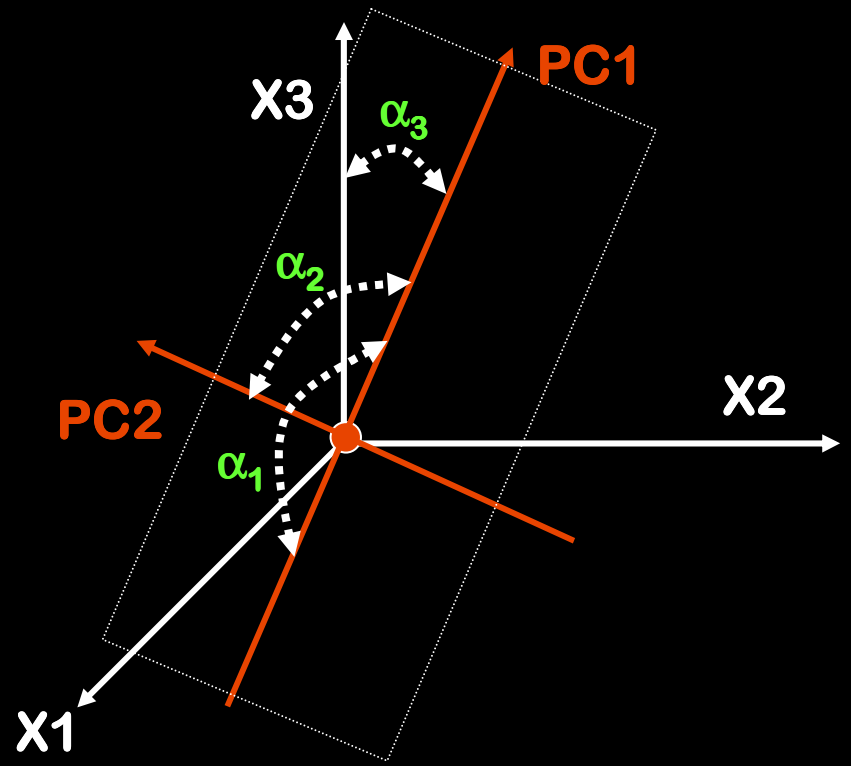
On the score plot,
"Sit together": similar behavior between descriptors

The Geometrical Meaning of Loading

The loadings express the orientation of the model plane in the original K-dimensional variables space.

The direction of PC1 in relation to the original coordinates is given by the cosine of α_1, α_2 and α_3.

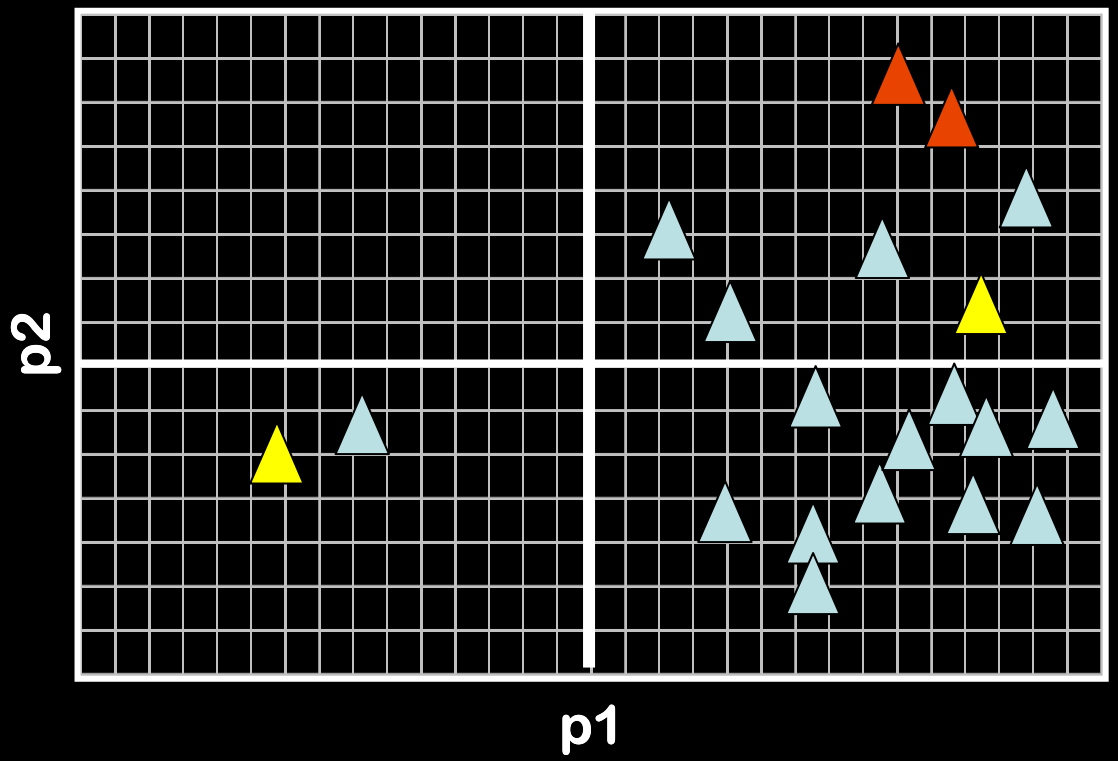
With 2 PC's and 3 original variables, 6 loading values (cosine of angles) are needed to specify how the model is positioned in the K-space.



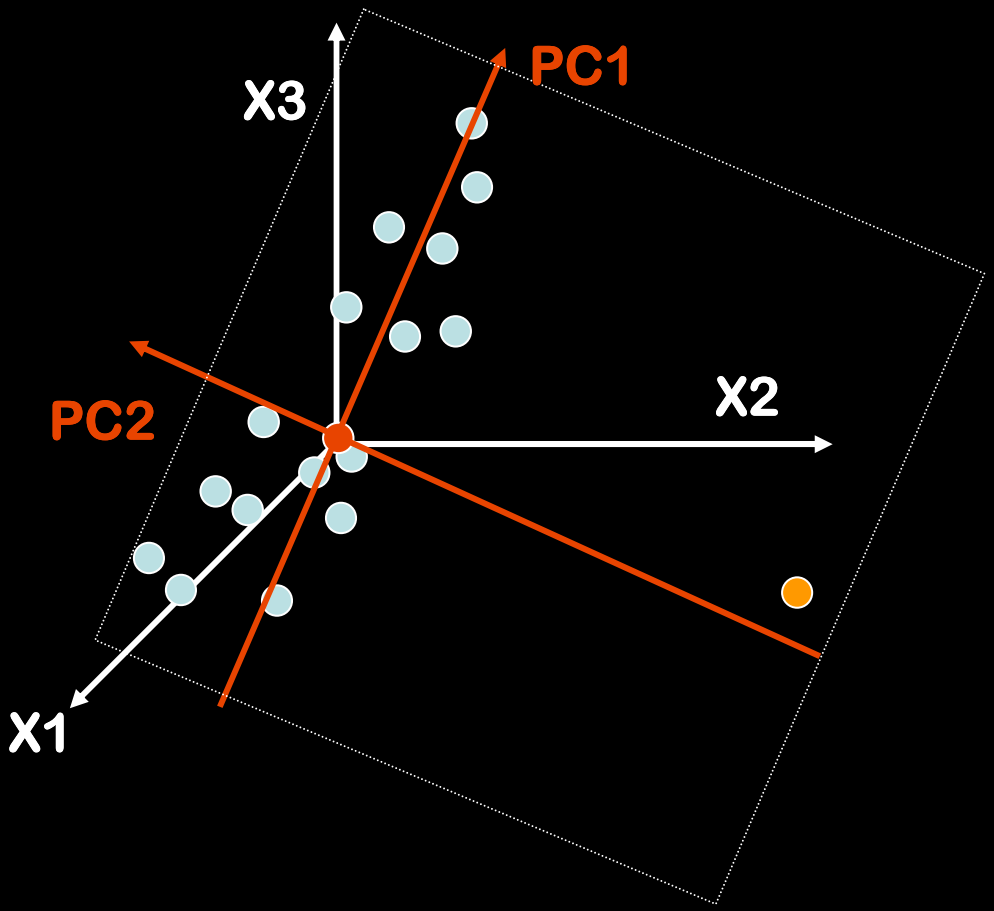
Loading Plot

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

Variables correlations:
- Positively correlated
- Negatively correlated



Outliers Detection



Back to the real case:

#	LD25	MR	logP	Volume	PM	Surface	density
CH2Cl2	0.96	1,62959	1,3044	67,136	84,933	176,1169	1,63677
CHCl3	1.45	2,01731	1,7381	75,751	119,378	186,8237	2,03576
CCl4	1.53	2,35508	2,4212	83,27	153,823	194,0565	2,3224
CF3CHBrCl	1.31	2,35642	2,3611	88,098	197,381	206,6438	2,85284
CHCl2CHCl2	2.42	2,92829	2,4947	94,238	167,85	215,3294	2,26061
Cl2C=CHCl	2.26	2,46705	2,2884	115,83	131,389	241,6985	1,42863
CCl2=CCl2	2.26	2,82835	3,3747	132,11	165,834	257,2367	1,46129

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

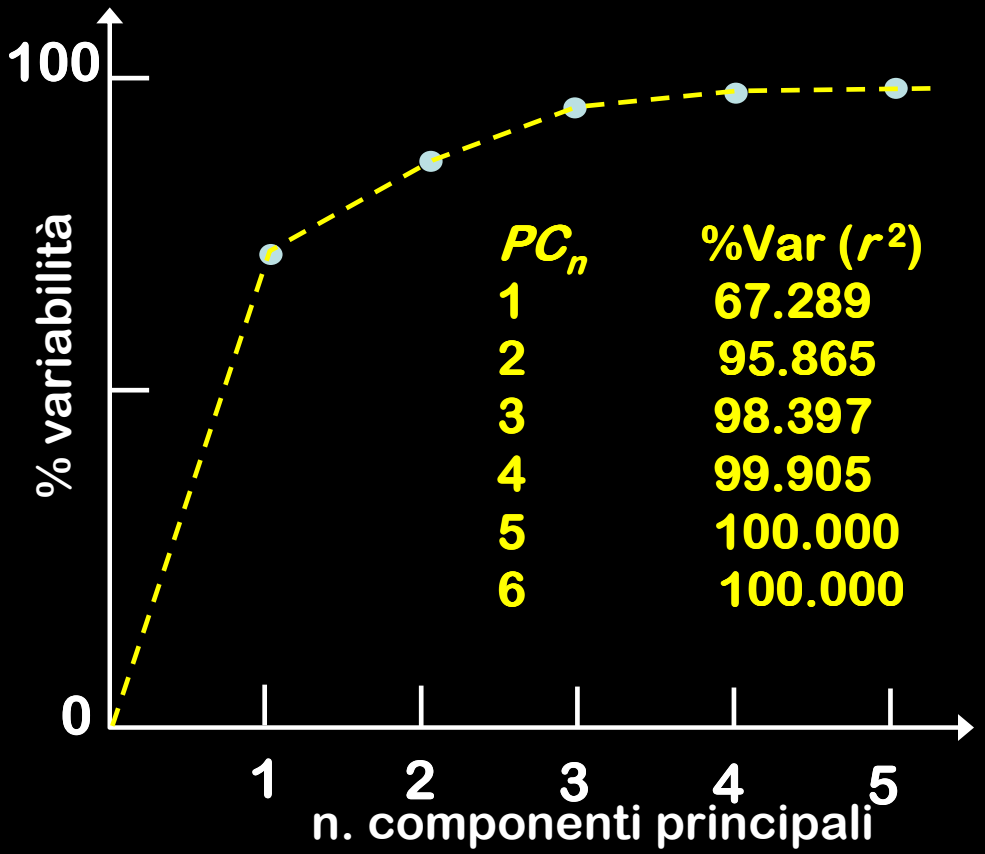
or

PC_1 = 0.47 MR + 0.34 logP + 0.008 Vol + 0.005 PM + 0.006 Sup +0.02 dens + 0.15 atom

PC_2 = -0.01 MR + 0.08 logP + 0.01 Vol - 0.008 PM + 0.008 Sup – 1.02 dens -0.18 g_2 atom

PC_3 = -4.68 MR - 0.67 logP + 0.03 Vol + 0.01 PM + 0.02 Sup + 0.08 dens + 0.62 g_2 atom

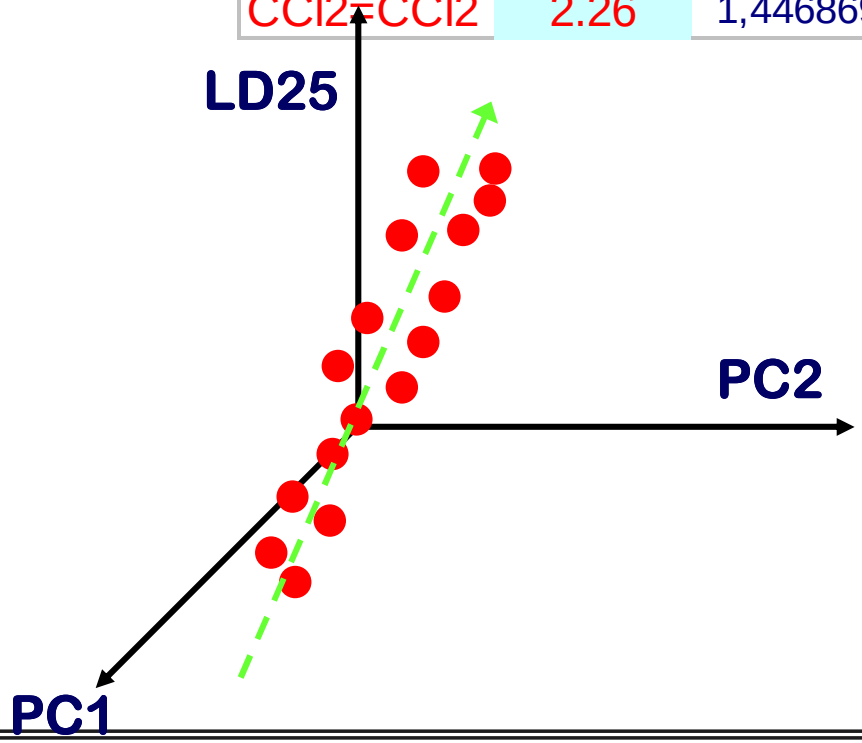
…



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

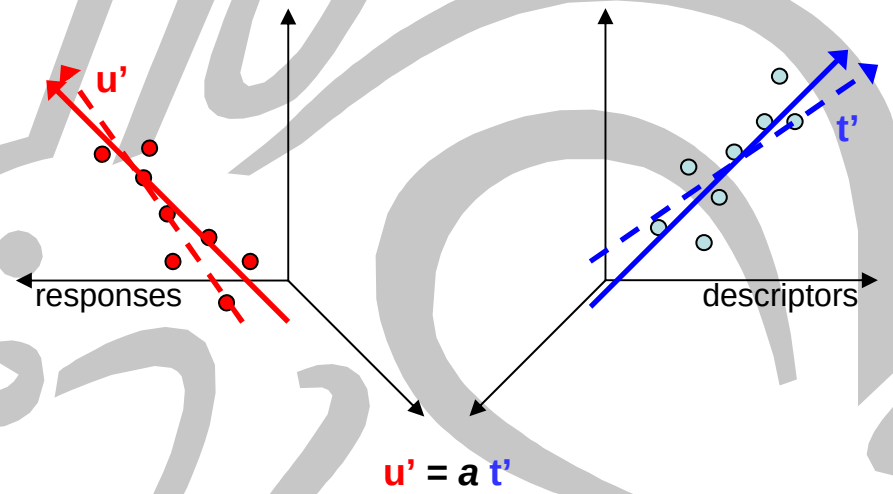
and finally....

#	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

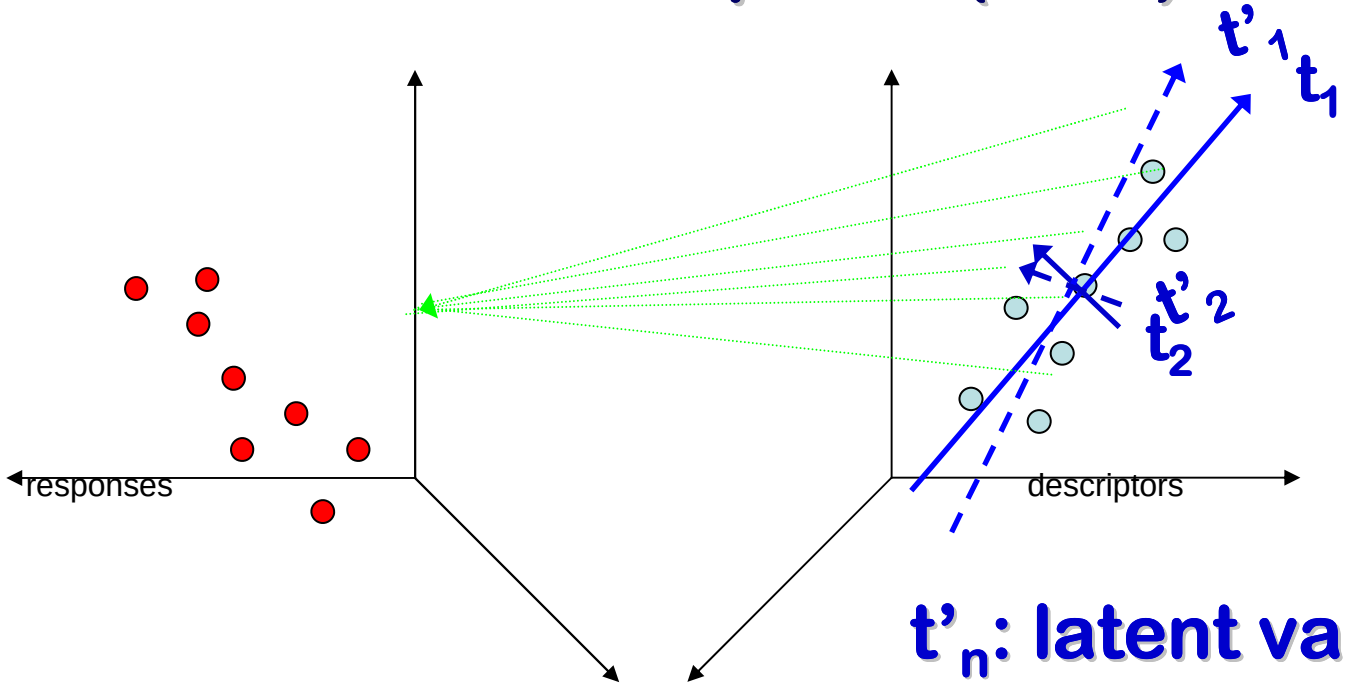


$$\log (1/LD_{25}) = m\,PC1 + n\,PC2$$

Partial Least Square Regression



Partial Least Square (PLS)



t'$_n$: latent variables

$$t'_1 = a_1x_1 + b_1x_2 + \ldots + m_1x_m$$

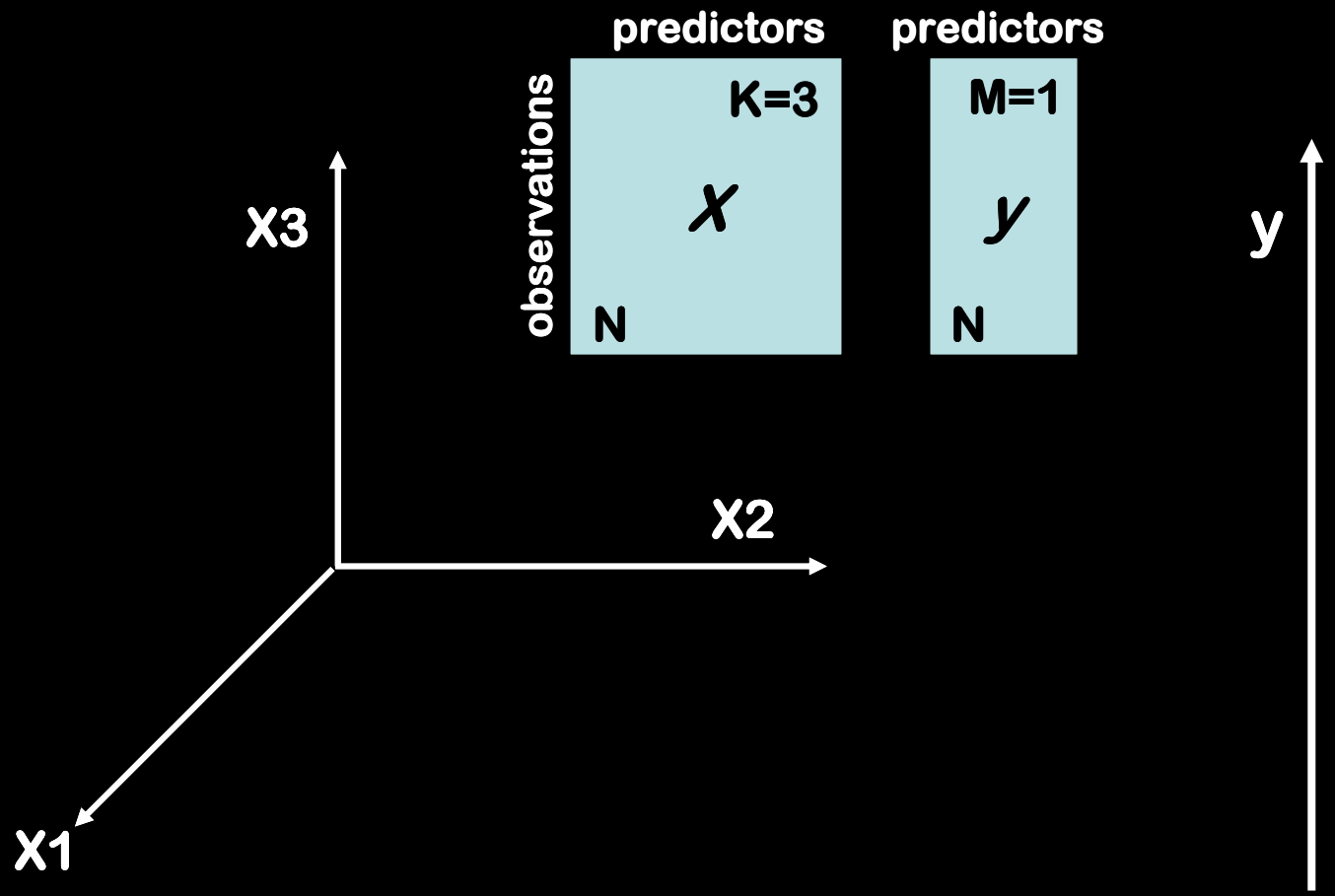
$$t'_2 = a_2x_1 + b_2x_2 + \ldots + m_2x_m$$

$$y_1 = a_1t'_1 + b_1t'_2 + \ldots + m_1t'_m$$

$$\ldots$$

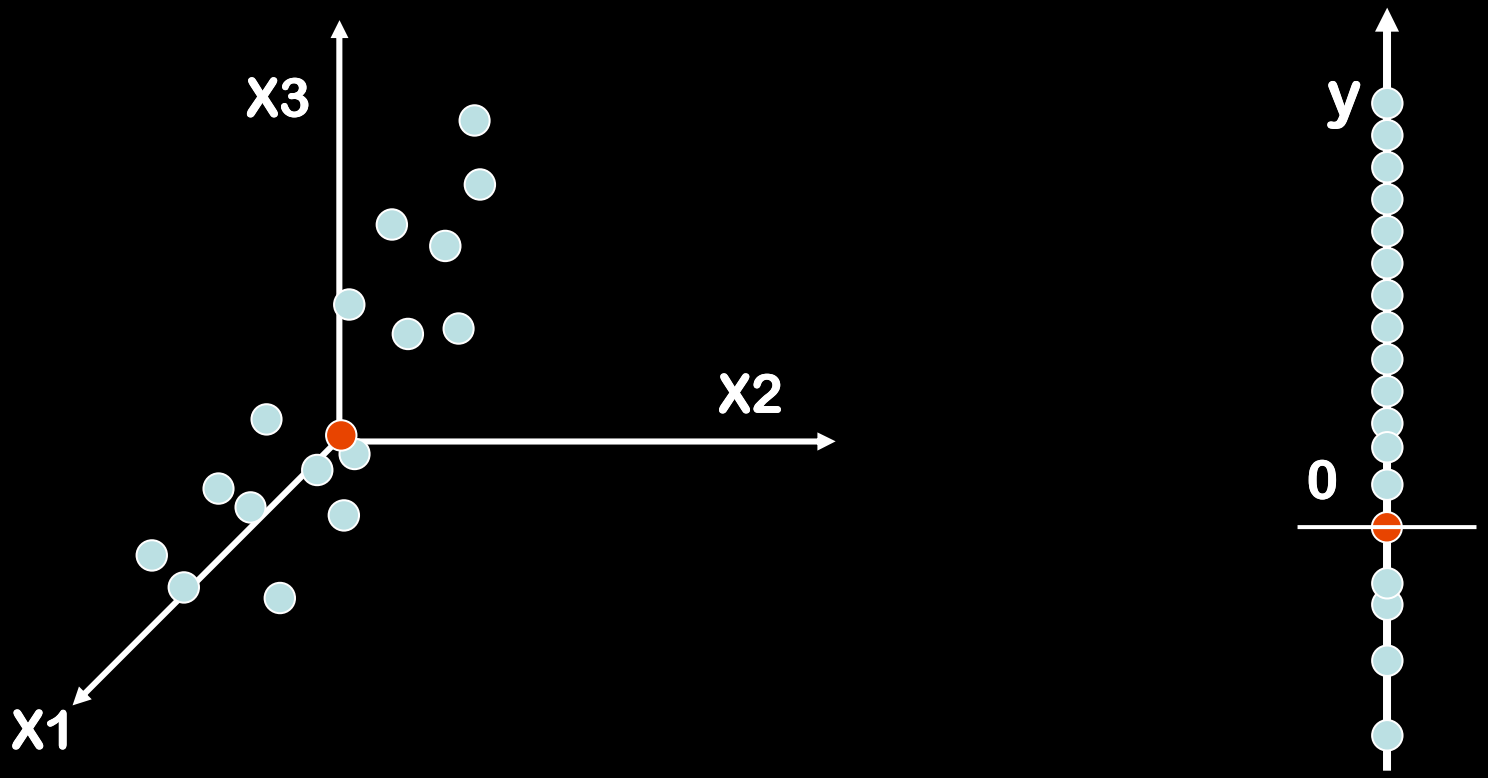
$$y_n = a_nt'_n + b_nt'_n + \ldots + m_nt'_m$$

A geometrical Interpretation of PLS



A geometrical Interpretation of PLS

PLS describes the relationship between the (mean centered) positions of the observations in the predictor space (*x*) and their (mean centered) positions in the response space (*y*).



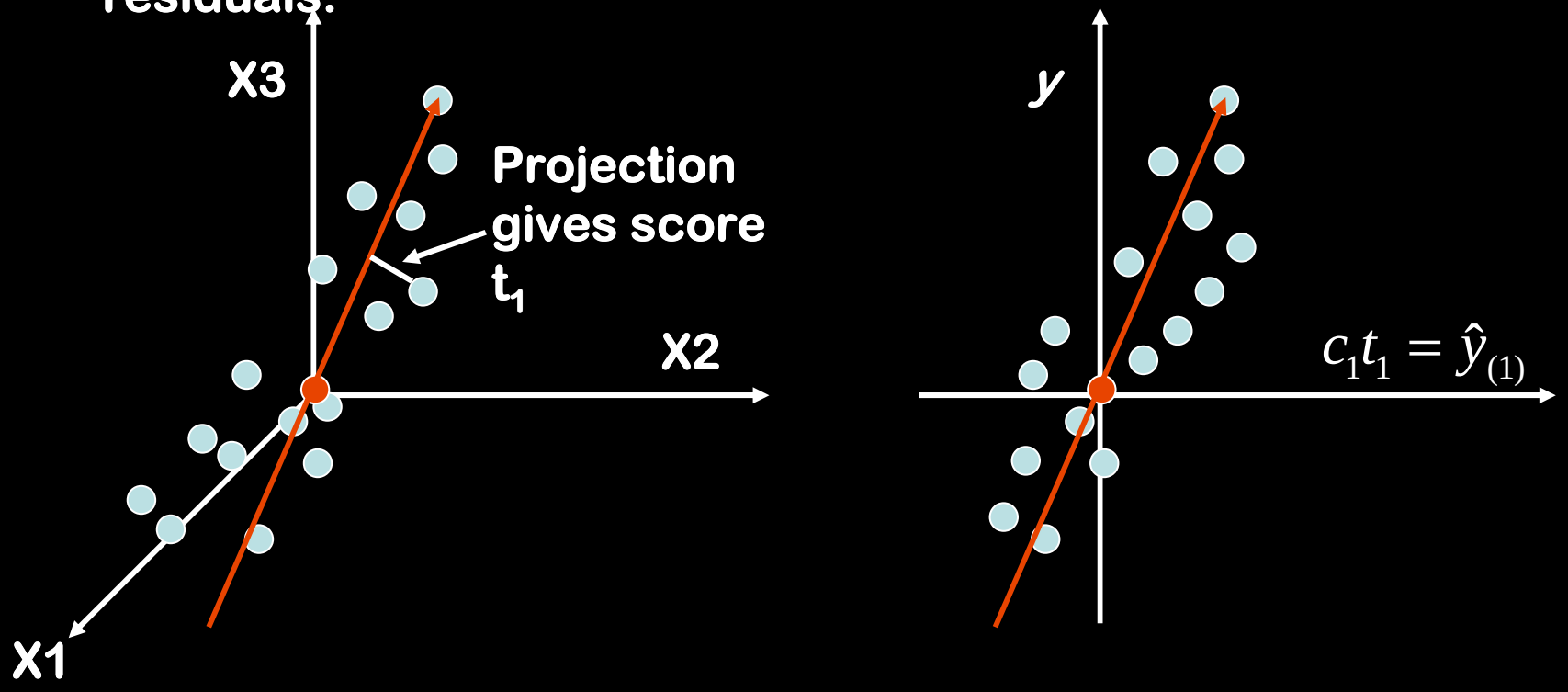
The First PLS Component

Remember our goal:

- Predict *y* from X.
- Since there are too many X's, generate a new set of variables in a PCA-manner.
- Predict *y* from a new variable *t*, *t* being a linear combination of the original X's.
- PLS assumes *y* predicted from *t* in a linear fashion, y = a*t*+b.

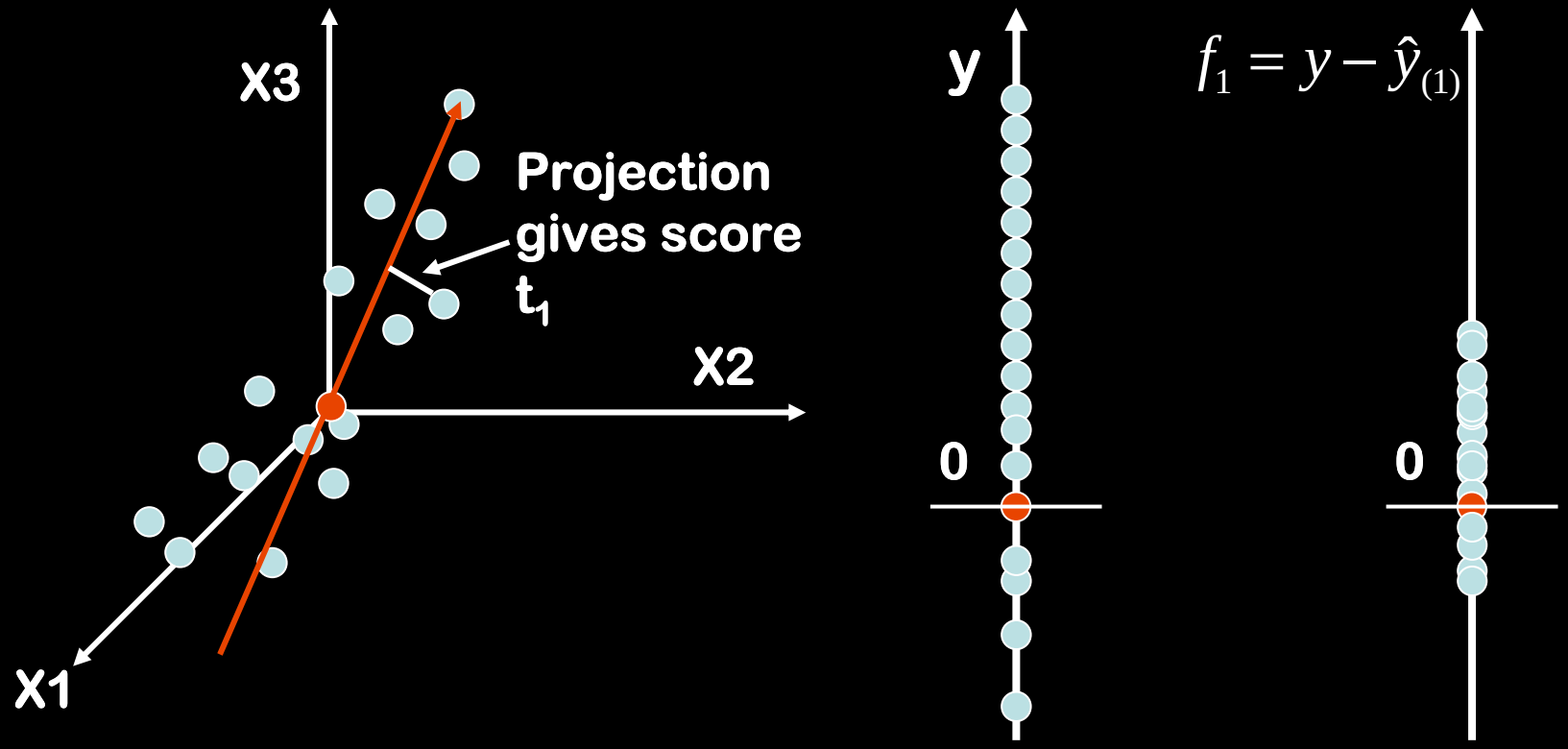
The First PLS Component

- Line in the X-space which well approximates the points in a least squares manner and at the same time provides a good correlation with the y vector.
- Degree of correlation with the y vector is judges from the residuals.



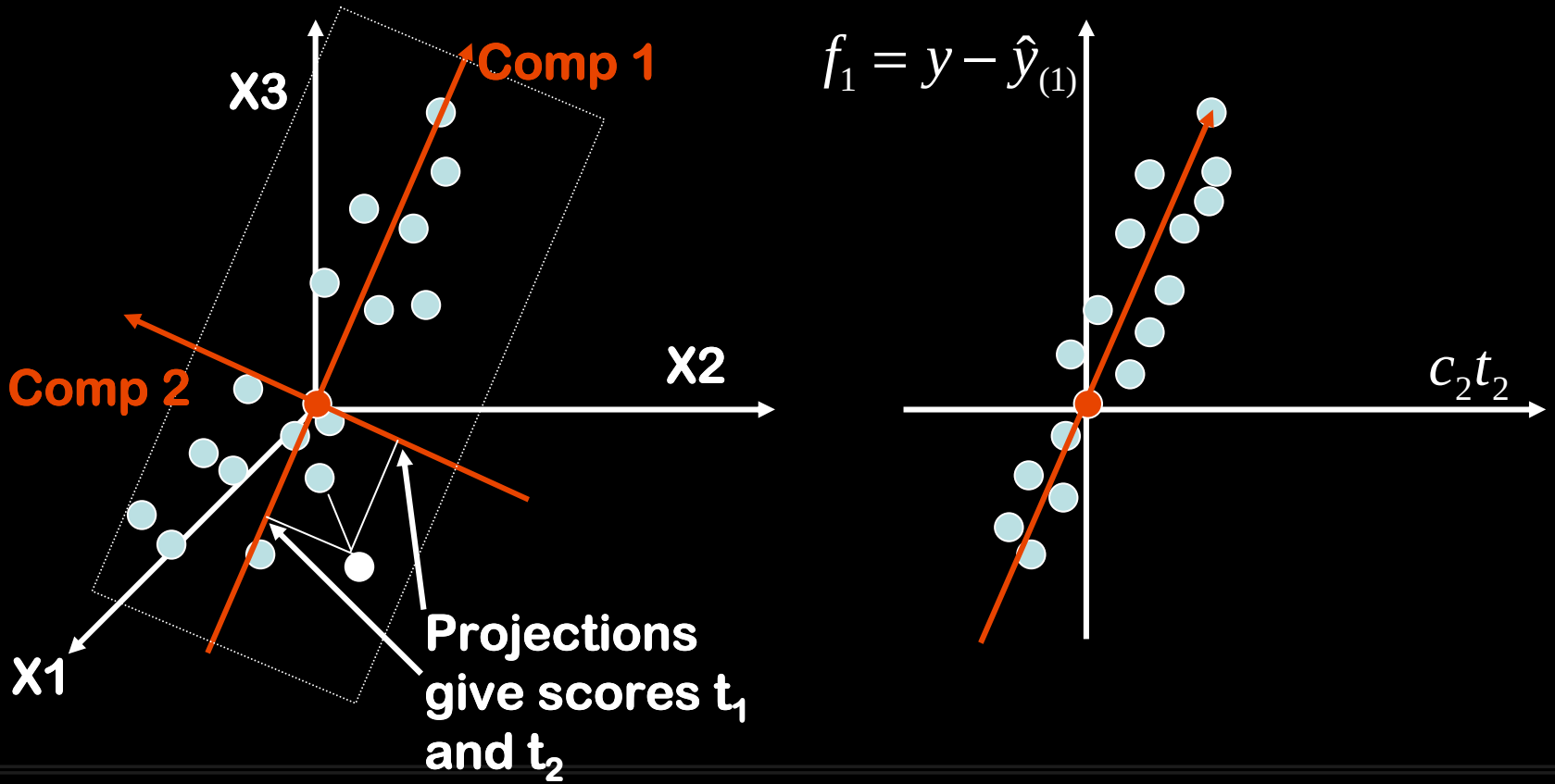
Residual of the First PLS Component

- The residual vector is shorter than the original y vector indicating that the first PLS component (first latent variable) accounted for a large part of the variation in *y*.



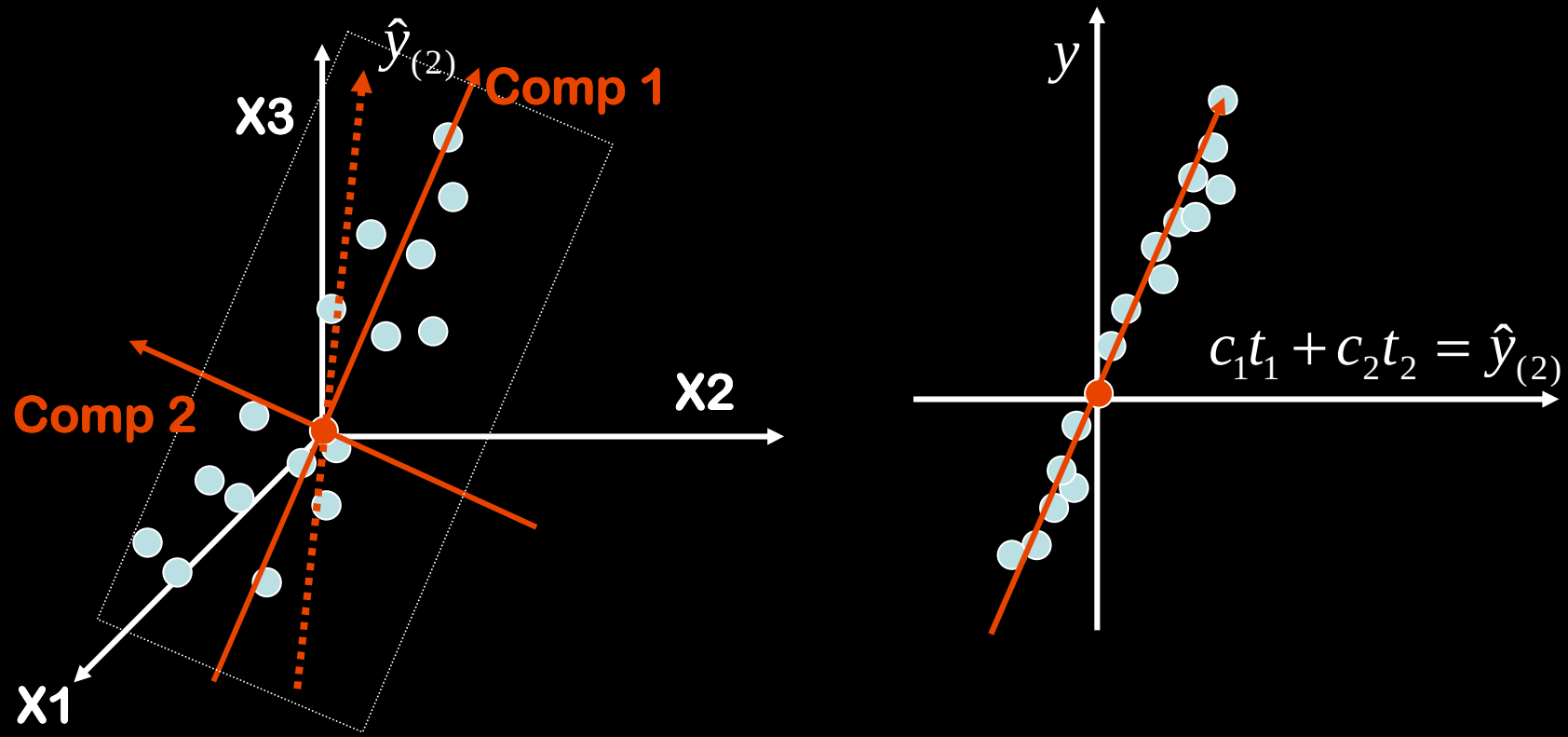
The Second PLS Component

- Line in the X-space which improves the description of the X data and at the same time provides a good correlation with the f_1 vector.
- Orthogonal to the first PLS component



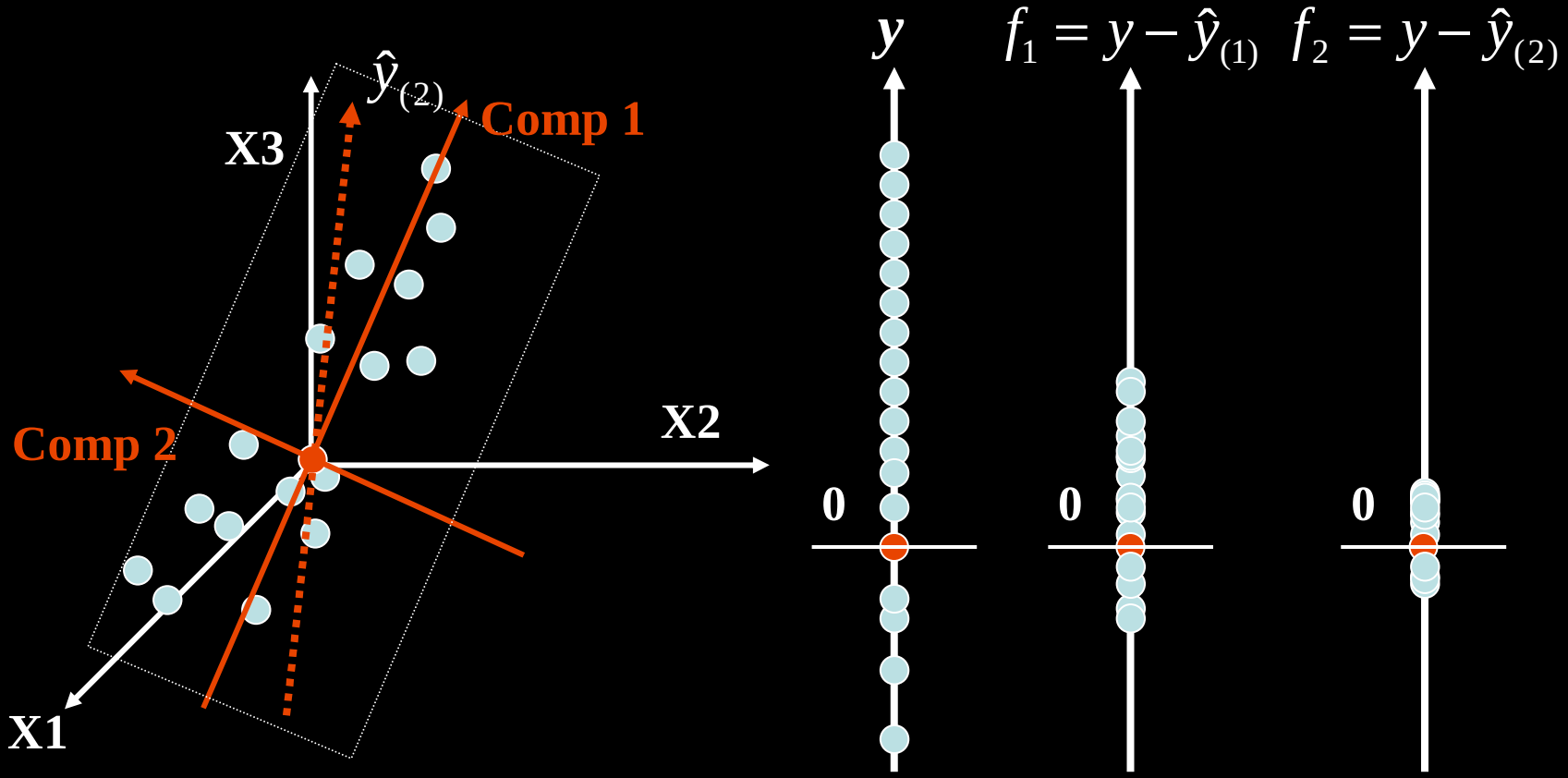
The Cumulative Effect of two PLS Components

- $y_{(2)}$ is a vector addition of the two first components in the X space.
- $y_{(2)}$ is a better predictor of the *y* values.



Explanatory Power of PLS

- $y_{(2)}$ is a vector addition of the two first components in the X space.
- $y_{(2)}$ is a better predictor of the *y* values.



Linear Methods:

m descriptors

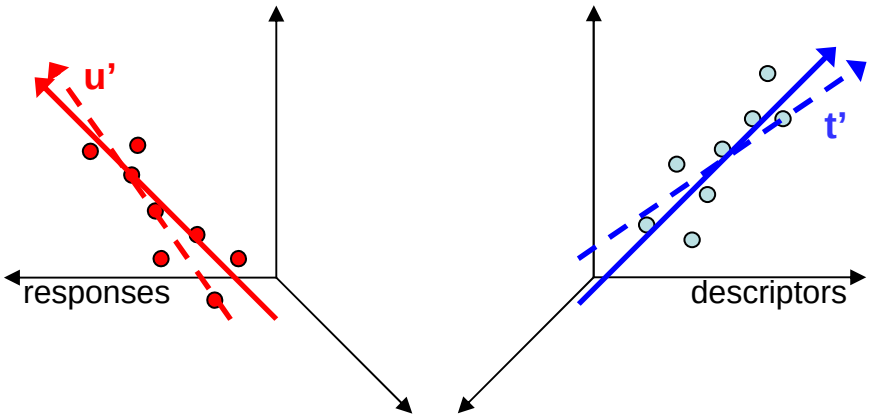
q responses

***n* molecules** | **X-matrix** | **Y-matrix** | ***n* molecules**

Input data

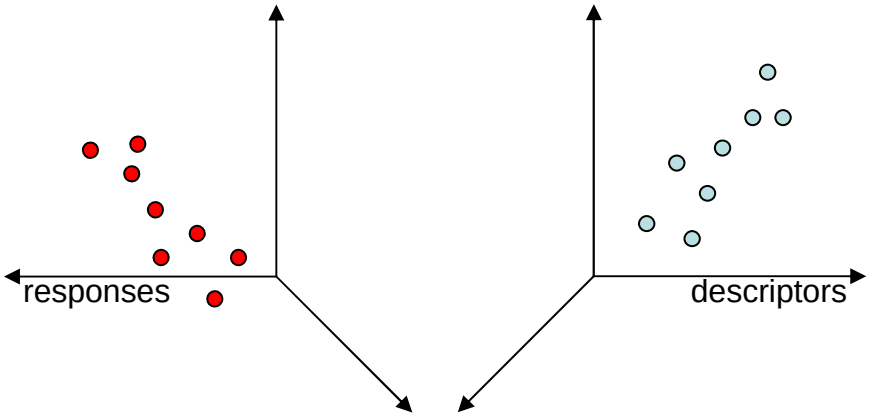
Output data

Partial Least Square (PLS)



u' = *a* t'

Multiple Regression Analysis (MRS)

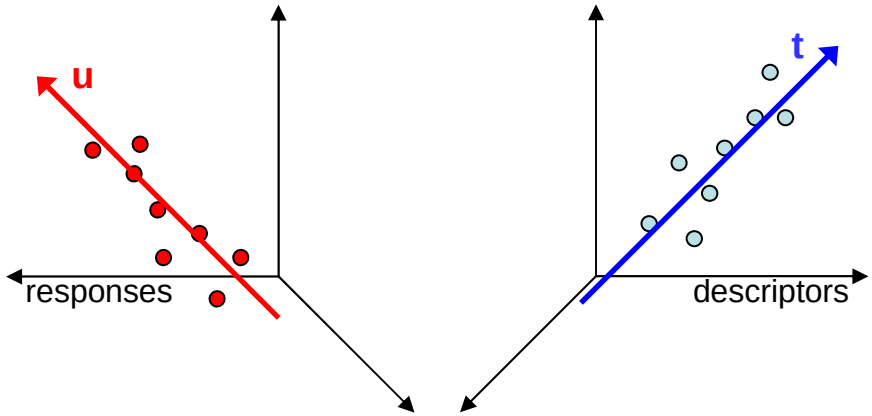


$y_1 = a_1x_1 + b_1x_2 + \ldots + m_1x_m$

...

$y_n = a_nx_n + b_nx_n + \ldots + m_nx_m$

Principle Components Regression (PCR)



u = *a* t

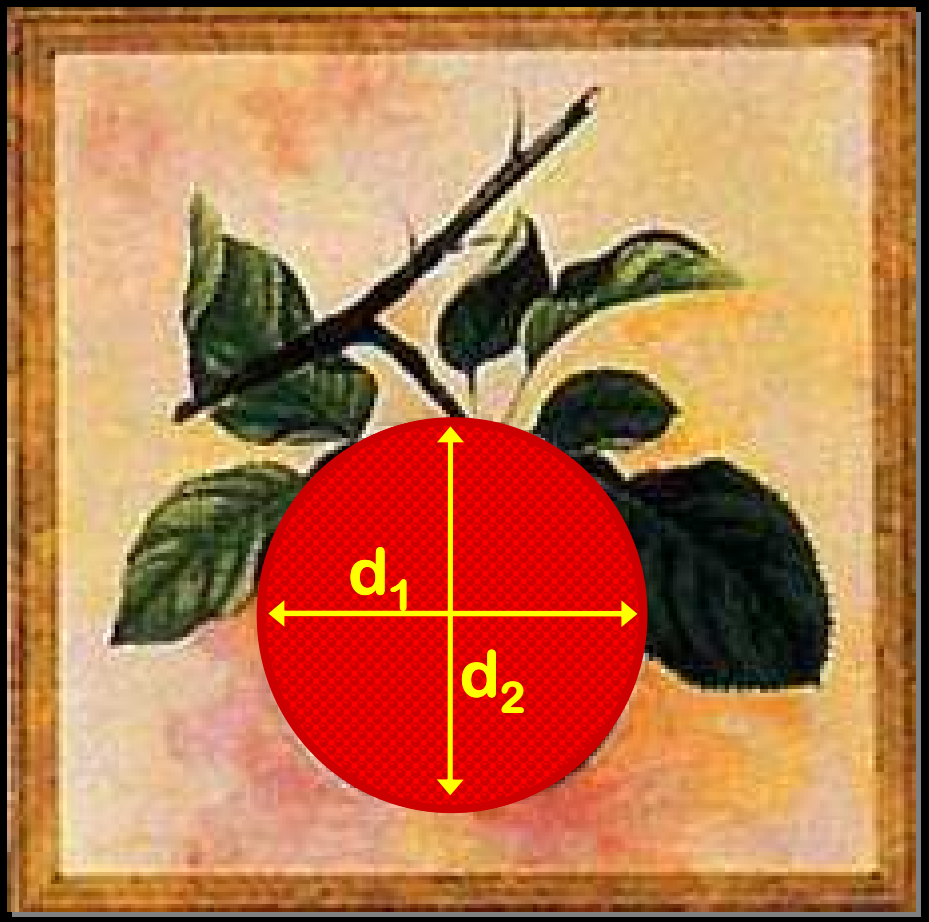
Artificial Neural Networks (ANN): una soluzioni alternativa alla forzata linearità delle QSARs.

Artificial Neural Networks (ANN)

How we can distinguish these two objects?

Experience → **Multiple Experience** → **Category**

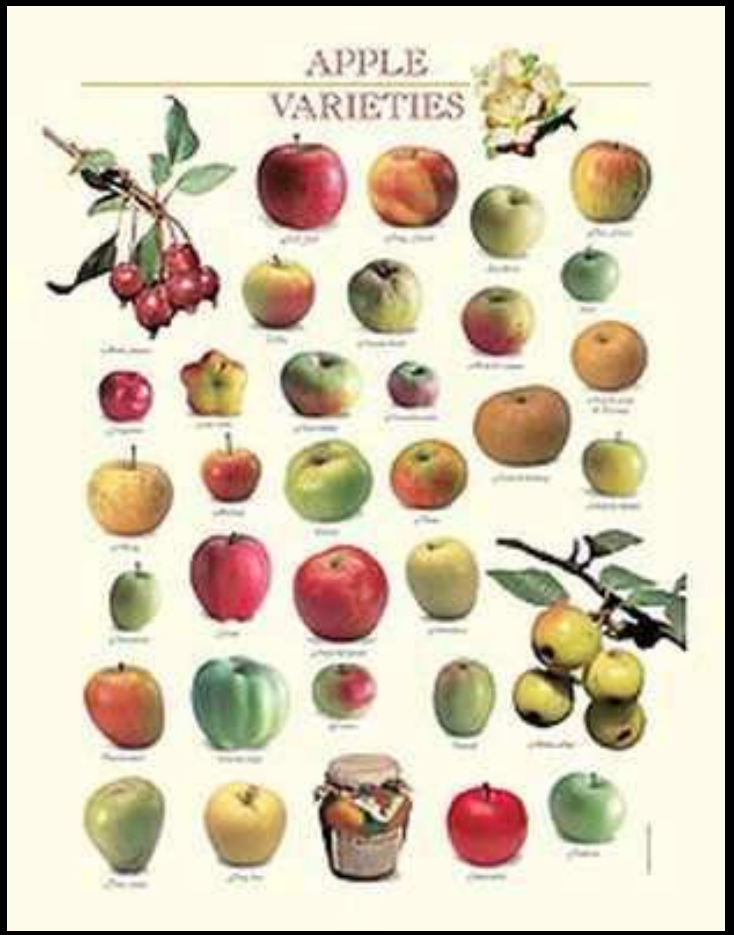
Artificial Neural Networks (ANN)

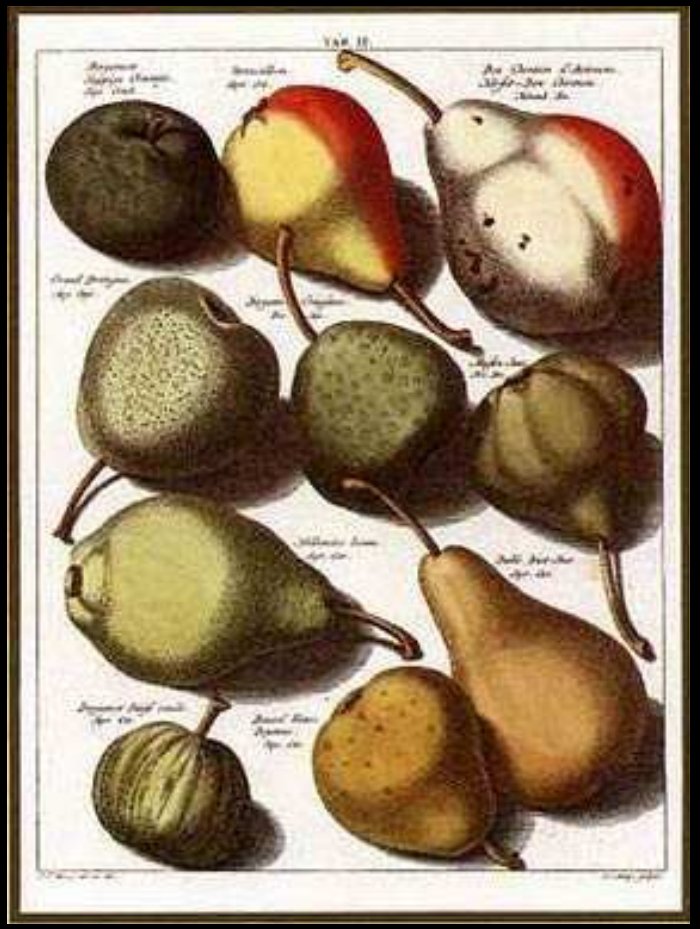


Color
Simmetry
…

Why apple?

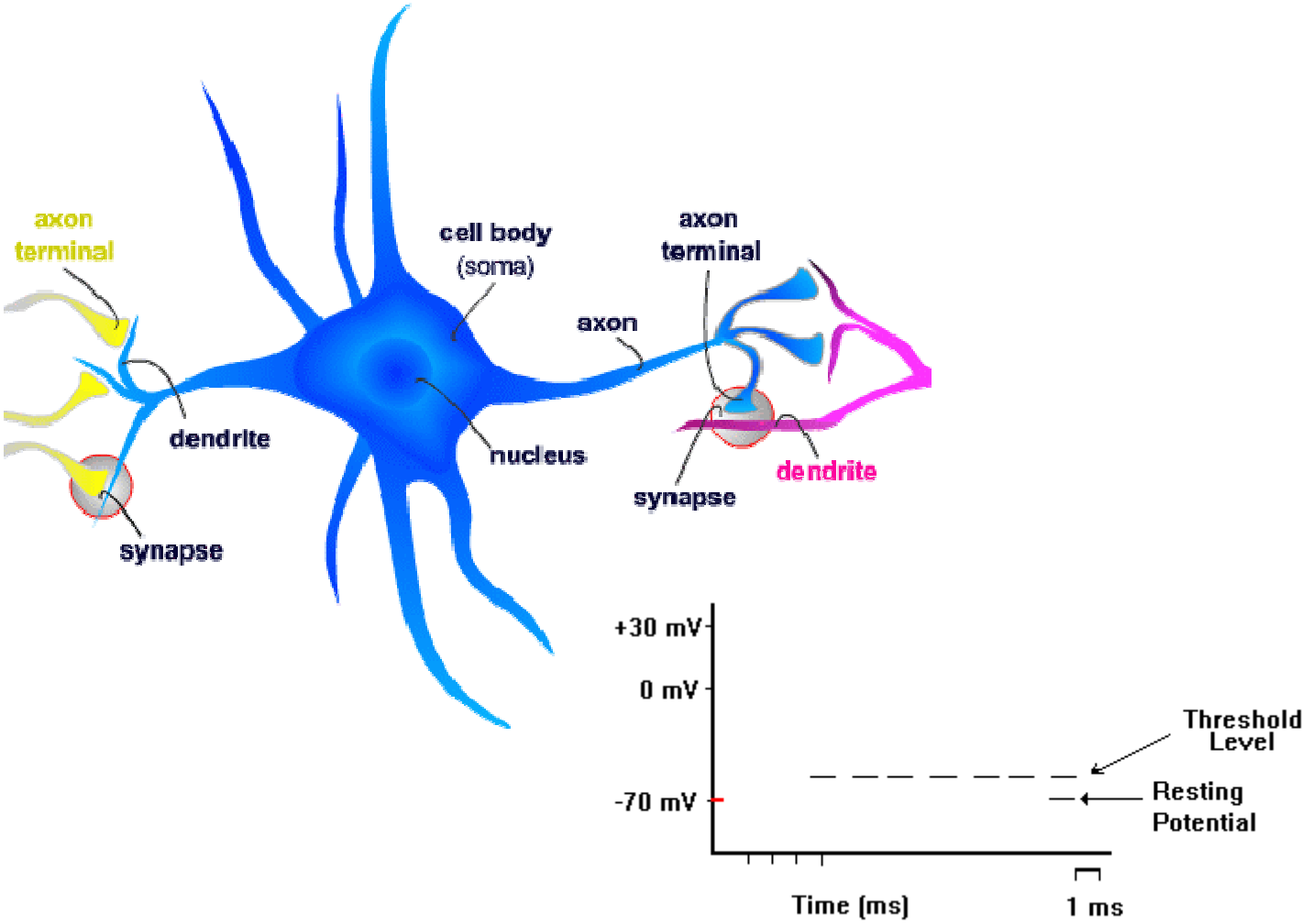
Artificial Neural Networks (ANN)



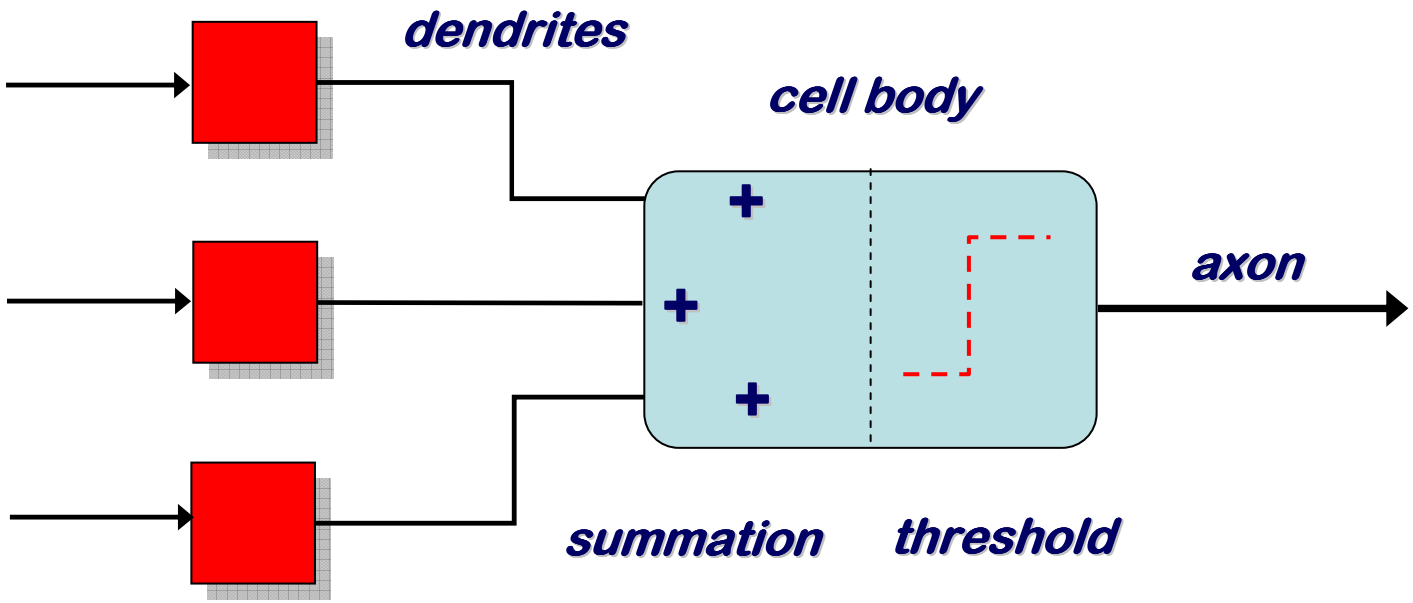


How we can distinguish apples by pears?

Do you remember the structure of neurons?

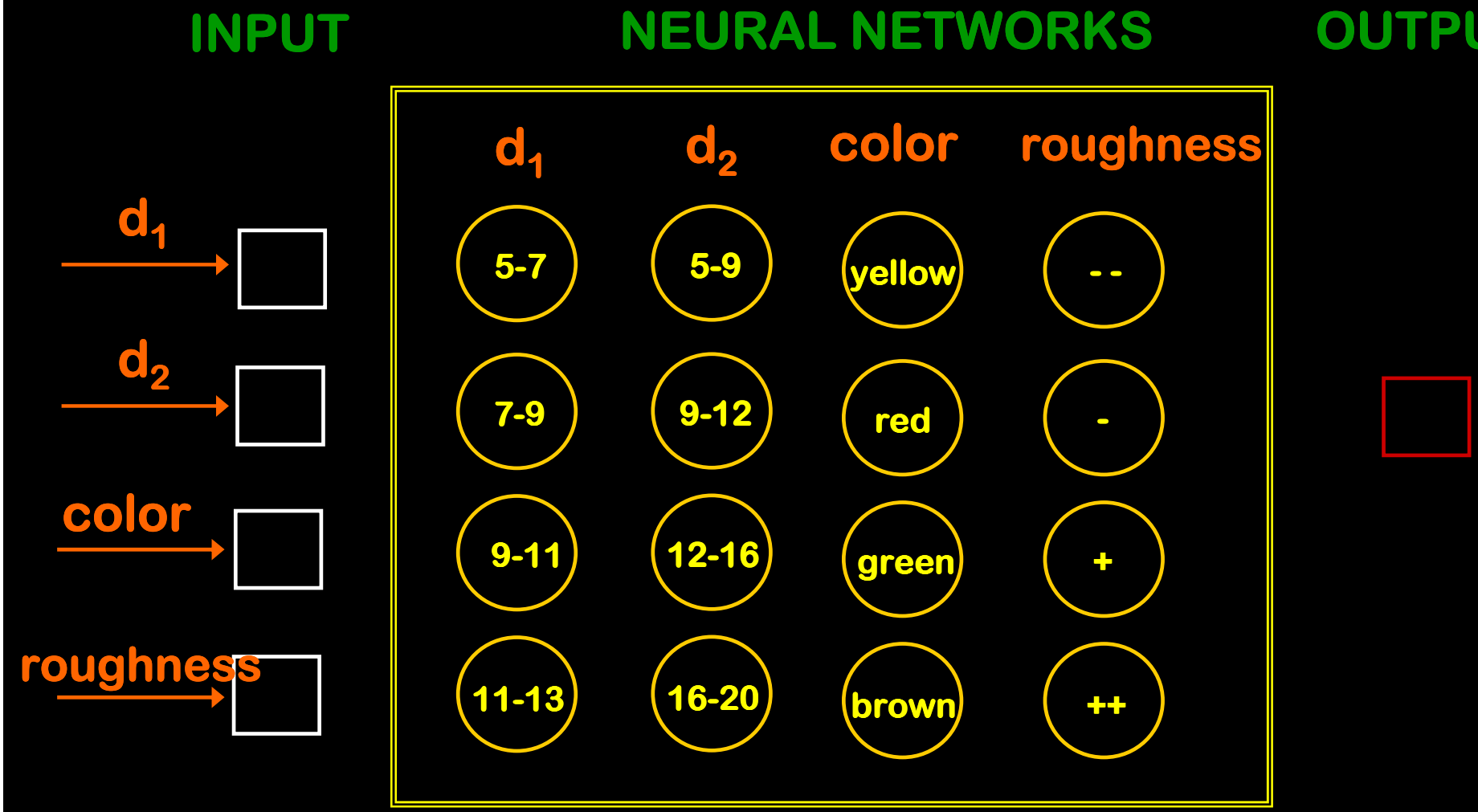


From human neurons to artificial neurons...



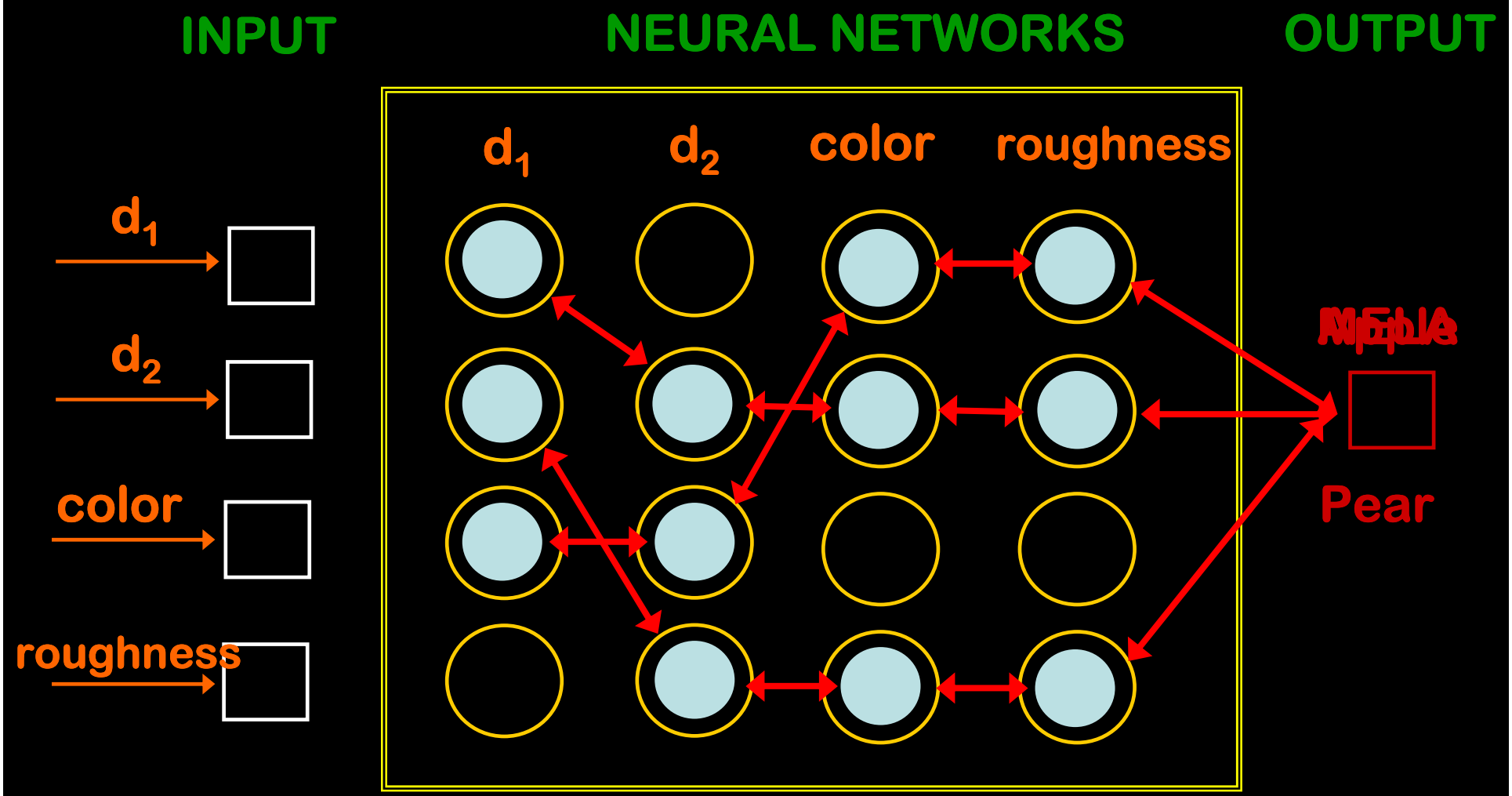
Artificial Neural Networks (ANN):

Simplify structure of an ANN



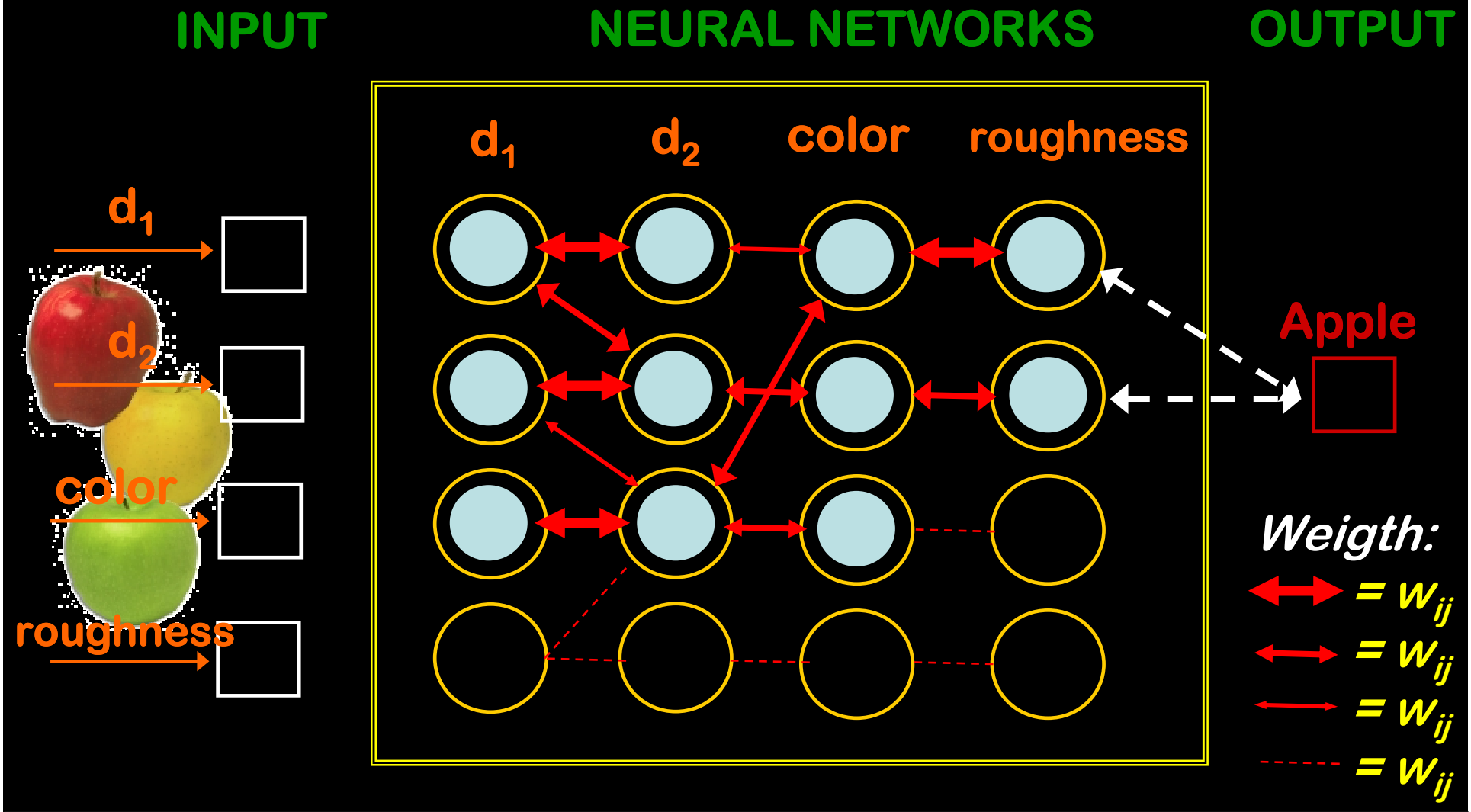
Artificial Neural Networks (ANN):

Phase 1 – Learning.



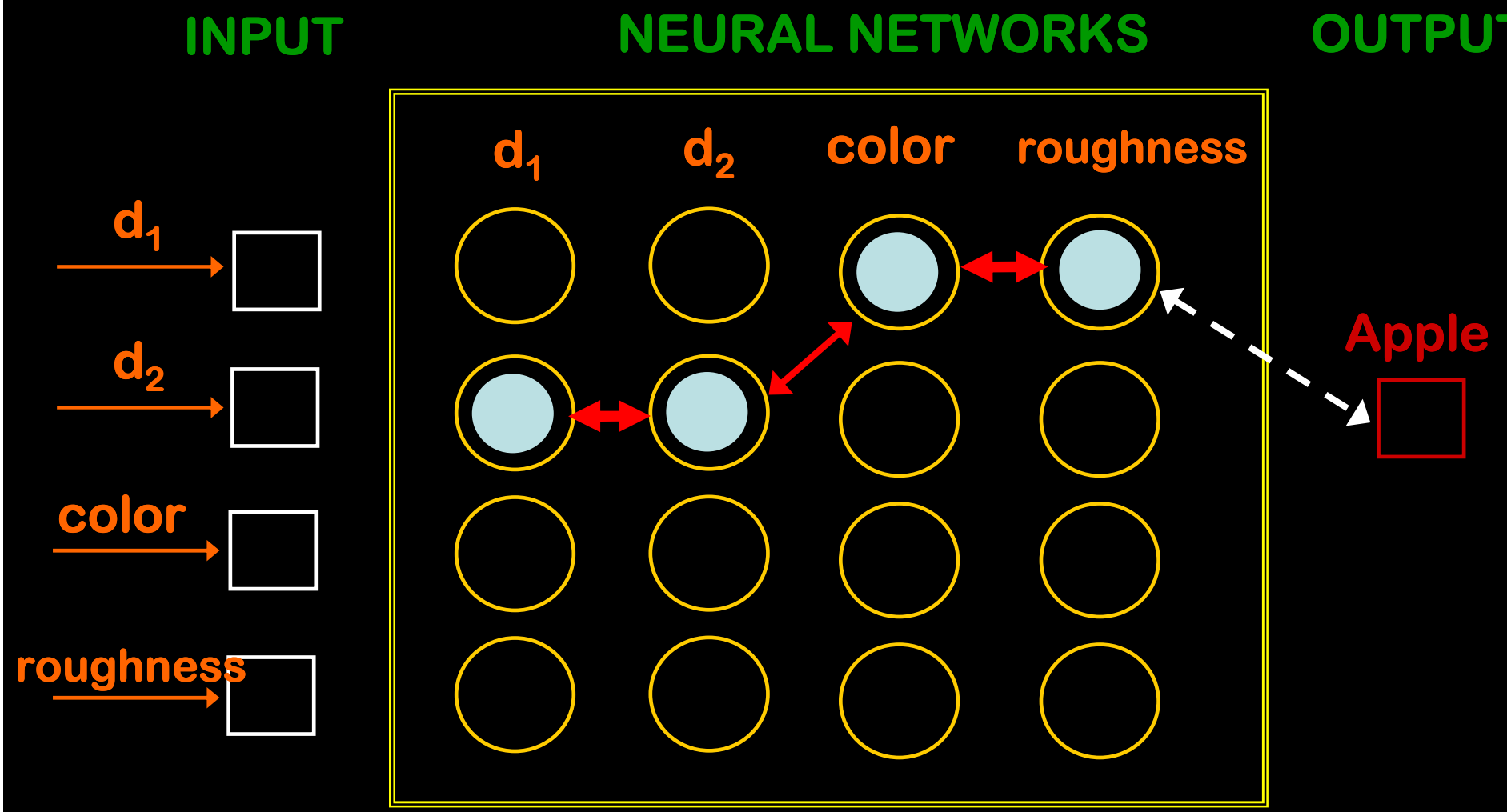
Artificial Neural Networks (ANN):

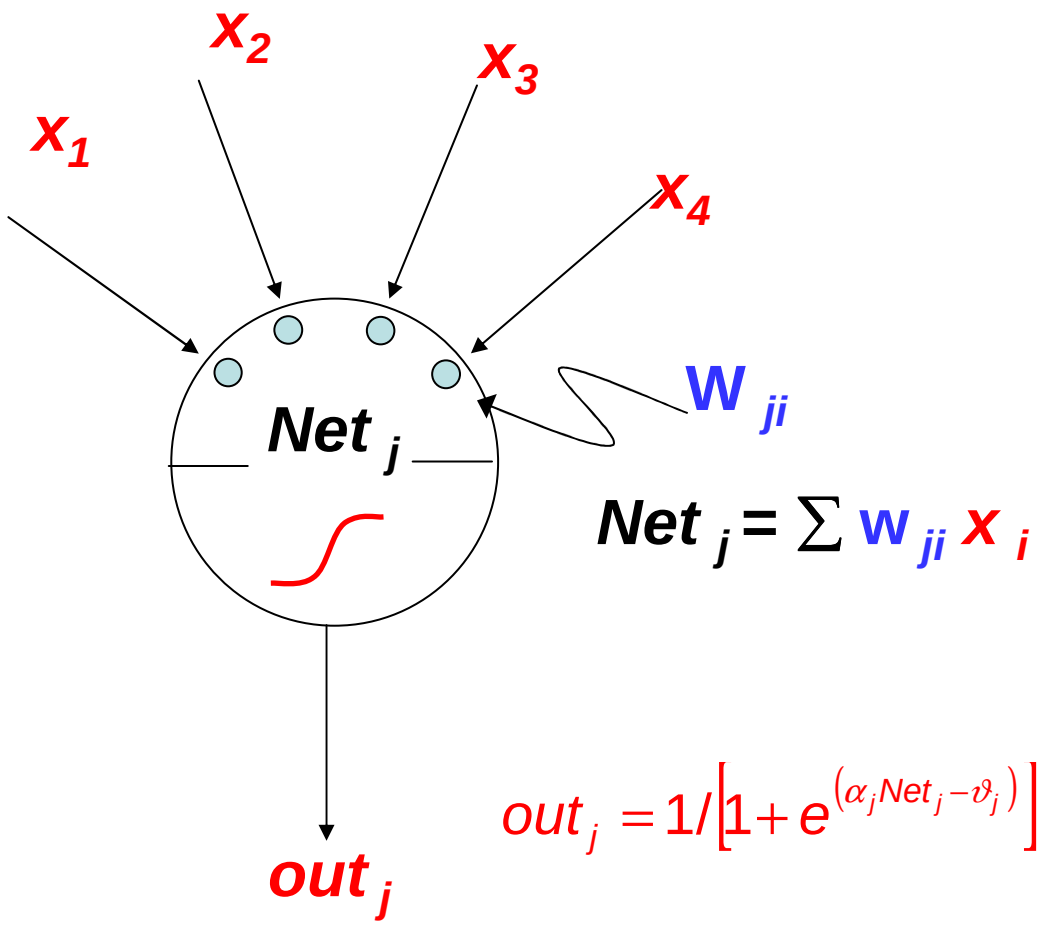
Phase 1 – Learning.



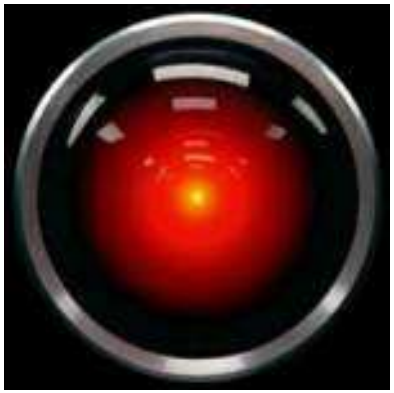
Artificial Neural Networks (ANN):

Phase 2 – Recognition.




x2
x3
x1
x4
W ji
Net j
Net j = ∑ w ji x i
out j = 1/[1+e^(α j Net j − ϑ j)]
out j

A view of HAL 9000's Main Terminal



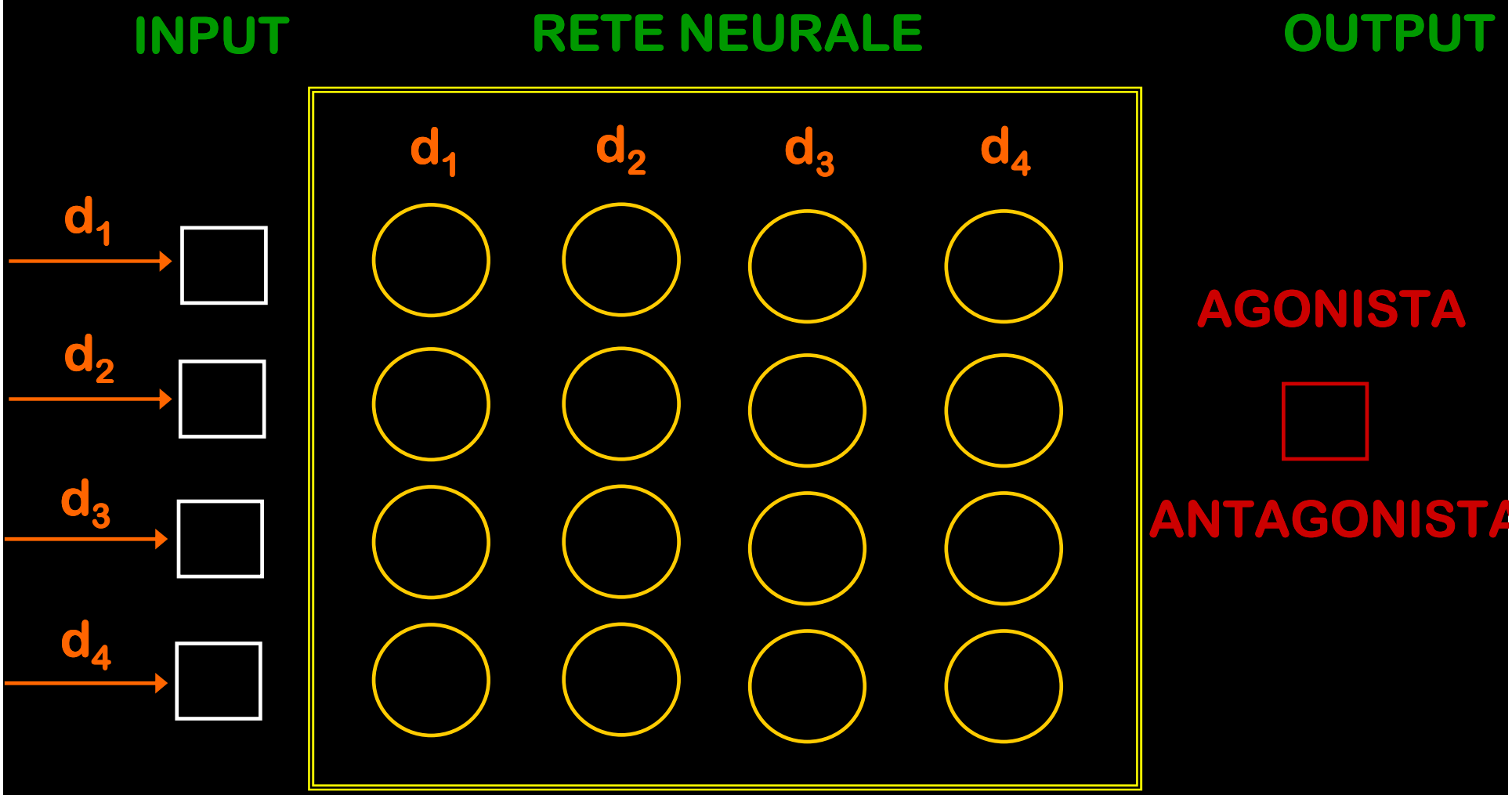
HAL's iconic camera eye.

A view of HAL 9000's Central Core in the *Discovery*.

Let me put it this way, Mr. Amer. The 9000 series is the most reliable computer ever made. No 9000 computer has ever made a mistake or distorted information. We are all, by any practical definition of the words, foolproof and incapable of error.

Artificial Neural Networks (ANN):

Applicazione in chimica farmaceutica.



Now, it is time to back at the nature of molecular descriptors.


TITLE
"Progettazione e Sviluppo di un Farmaco"
DIRECTOR
Stefano Moro
CAMERA
Chimica e Tecnologia Farmaceutiche
DATE
SCENE
TAKE