

# Computational Toxicology and QSAR



by **Stefano Moro**

*Molecular Modeling Section (MMS)*

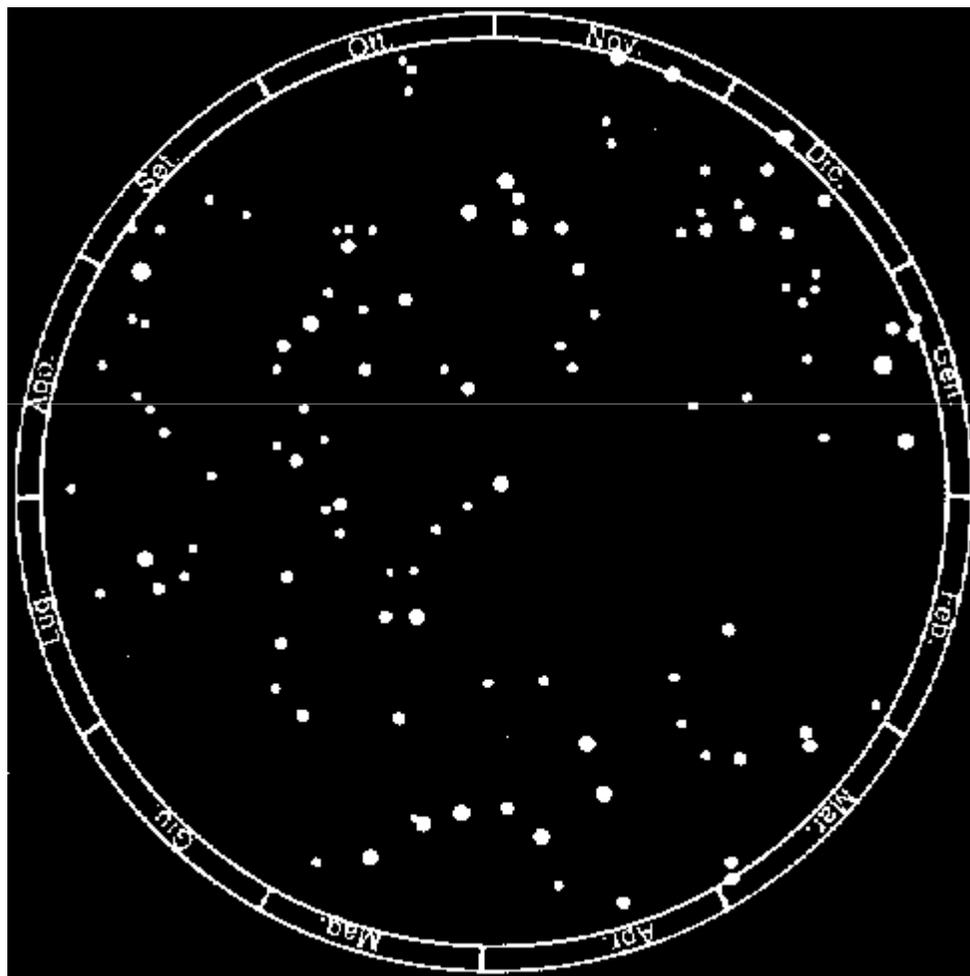
*Department of Pharmaceutical and Pharmacological Sciences*

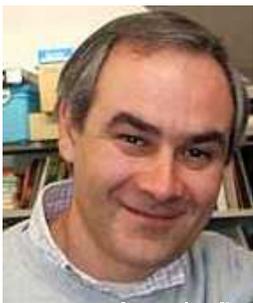
*University of Padova*

*©2016*



# (Q)SAR and surroundings...

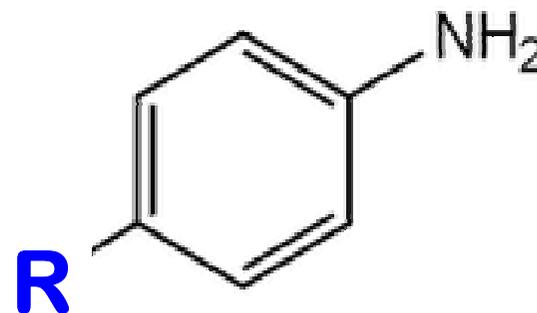




# The easiest way to describe the concept of (Q)SAR...



$f(x)$



Activities ( $LC_{50}$ ,  $\mu M$ )

$A_1$

$A_2$

$A_3$

...

$A_n$

Chemicals

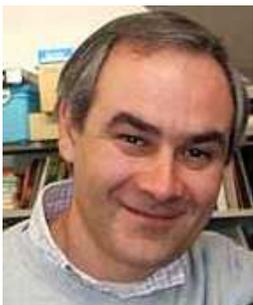
$R(1) = NH_2$

$R(2) = NHCH_3$

$R(3) = N(CH_3)_2$

...

$R(n) = NHC_6H_5$

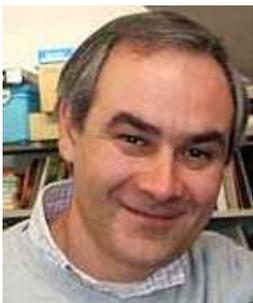


or alternatively...

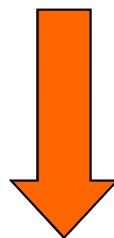
## Toxicological Space

	E.P. (A)	E.P. (B)	E. P. (C)	E.P. (D)	E.P. (E)
Chemical Space	Comp.1	LC <sub>1</sub> (A)			
	Comp.2	LC <sub>2</sub> (A)			
	Comp.3	LC <sub>3</sub> (A)			
	Comp.4	LC <sub>4</sub> (A)			
	Comp.n	LC <sub>n</sub> (A)			

Screening for the specific endpoint

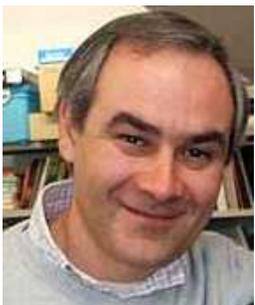


**(Q)SAR** it is not necessary to remember this definition:



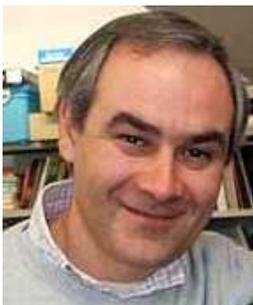
**Activities ( $IC_{50}$ ,  $\mu M$ )**

***Any in vivo or in vitro data is affected by both PHARMACODYNAMIC and PHARMACOKINETICS properties of the specific assay.***

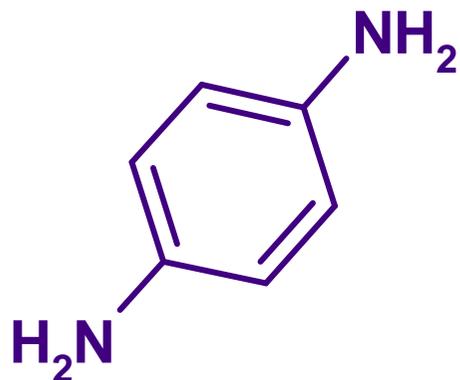


Please: don't start any quantitative structure-activity relationship if you're not confident of *reproducibility* of your data toxicological activity!!!

$LC_{50} =$	1.0	$\mu M \neq$	
	$1.0 \pm 1.0$	$\mu M \neq$	
	$1.0 \pm 0.5$	$\mu M \neq$	
	$1.0 \pm 0.1$	$\mu M \neq$	
	$1.0 \pm 0.05$	$\mu M$	



(Q)SAR: we are ready to this...



PM = 108,14

pKa = 6.2

PM = 108,14

Volume = 93,9

MP = 142

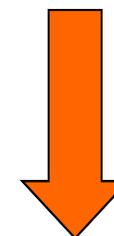
PSA = 52

nC = 6

logP = -0.3

...

Structure



Properties

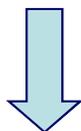


(Q)**S**AR: we are ready to this...

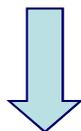
*Real world*

*Virtual world*

**Chemical Compound (CC)**

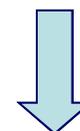


**Chemical Structure (CS)**



**Chemical Properties (CP)**

**Numerical  
representations of CS**



**Molecular Descriptors (MD)**



(Q)SAR: we are ready to this...

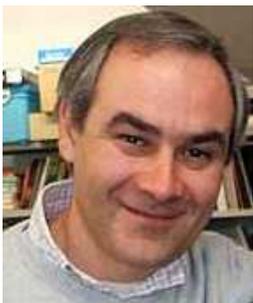
DRAGON 7.0 is able to calculate **5270** molecular descriptors.

The screenshot shows the DRAGON 7.0 software interface. At the top, there is a menu bar with 'File', 'Fingerprints', 'View', 'Analysis', and 'Settings'. Below the menu bar is a toolbar with various icons. The main window is divided into two panes: 'Viewer' and 'Info'. The 'Viewer' pane shows a 3D ball-and-stick model of a complex organic molecule. The 'Info' pane shows a table of molecular descriptors. The table has 13 columns: 'No.', 'NAME', 'CATS2D\_00\_DD', 'CATS2D\_01\_DD', 'CATS2D\_02\_DD', 'CATS2D\_03\_DD', 'CATS2D\_04\_DD', 'CATS2D\_05\_DD', 'CATS2D\_06\_DD', 'CATS2D\_07\_DD', 'CATS2D\_08\_DD', 'CATS2D\_09\_DD', and 'CATS2D\_00\_DA'. The table contains 24 rows of data, each representing a different descriptor.

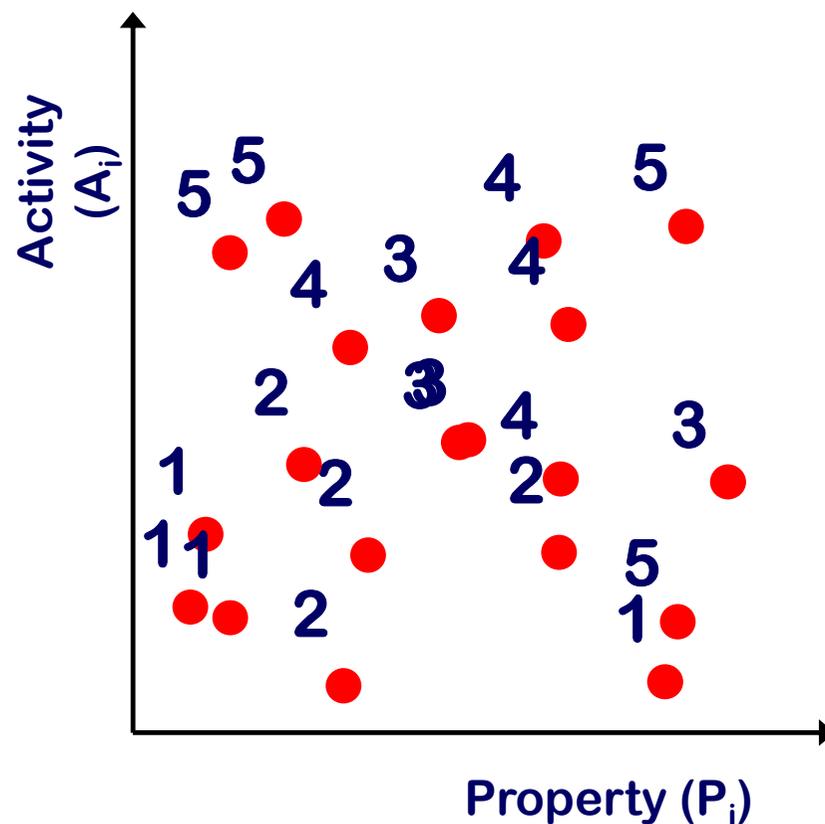
No.	NAME	CATS2D_00_DD	CATS2D_01_DD	CATS2D_02_DD	CATS2D_03_DD	CATS2D_04_DD	CATS2D_05_DD	CATS2D_06_DD	CATS2D_07_DD	CATS2D_08_DD	CATS2D_09_DD	CATS2D_00_DA
7	[N-]=[N+]=ClC=C(C=CCl	1	0	0	0	0	0	0	0	0	0	1
8	[N-]=[N+]=CC(=O)NCC(=O	1	0	0	0	0	0	0	0	0	0	0
9	[N-]=[N+]=CC(=O)OCC(N	2	0	0	0	1	0	0	0	0	0	1
10	[N-]=[N+]=NClCC(OCl)C	2	0	0	0	0	0	0	1	0	0	1
11	[N-]=[N+]=Nc1ccc(cc1)Nc	1	0	0	0	0	0	0	0	0	0	0
12	[N-]=[N+]=Nc4ccc3c(nlc	2	0	0	0	0	1	0	0	0	0	0
13	[N-]=[N+]=Nc4ccc3c(nlc	2	0	0	0	0	1	0	0	0	0	0
14	[N-]=[N+]=NCC(N)C(=O)C	2	0	0	0	1	0	0	0	0	0	1
15	[N-]=[N+]=NCC(O)CN=[N	1	0	0	0	0	0	0	0	0	0	1
16	[N-]=[N+]=Nc1ccc1	0	0	0	0	0	0	0	0	0	0	0
17	[N-]=[N+]=NCC1CCCC1	0	0	0	0	0	0	0	0	0	0	0
18	[O-][N+](=[O-])=ClCCCC1	0	0	0	0	0	0	0	0	0	0	0
19	[O-][N+](=Nc1ccc(c1)Cl)	0	0	0	0	0	0	0	0	0	0	0
20	[O-][N+](=Nc1ccc(OCC)cc	0	0	0	0	0	0	0	0	0	0	0
21	[O-][N+](=Nc1ccc1)c2cc	0	0	0	0	0	0	0	0	0	0	0
22	[O-][N+](=NCC)CC	0	0	0	0	0	0	0	0	0	0	0
23	[O-][N+](=O)N=C1NCCN	1	0	0	0	0	0	0	0	0	0	0
24	[O-][N+](=O)C(N)C=C/Cl=	0	0	0	0	0	0	0	0	0	0	0



R. Todeschini



**(Q)SAR**: follow me in this wonderful experience



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

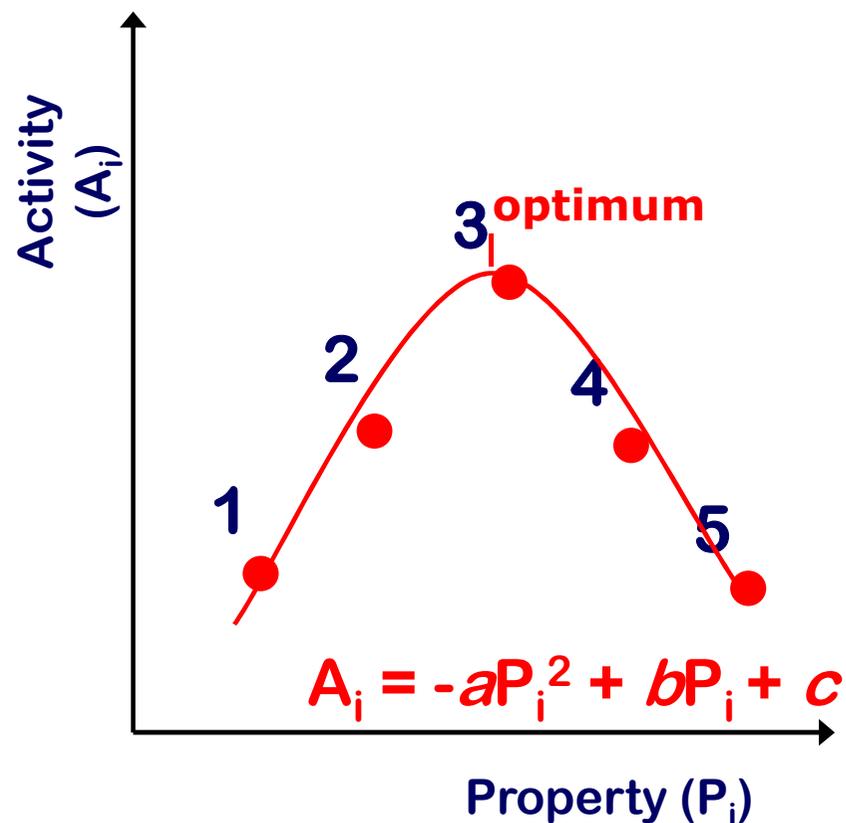
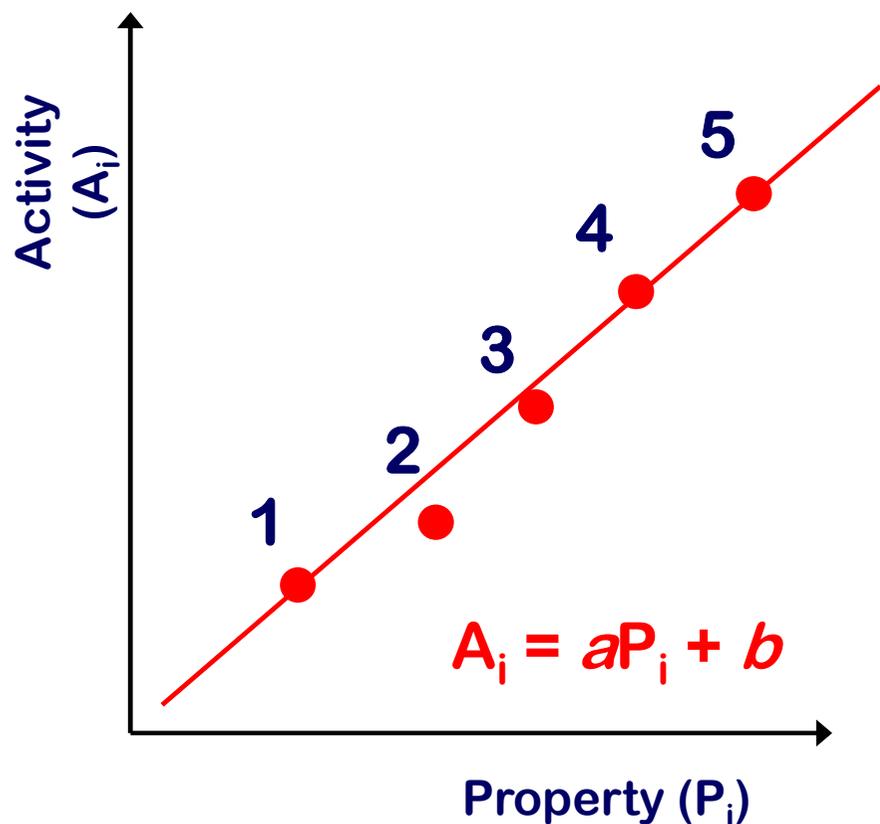


(Q)SAR: how can we select the good  
*“molecular descriptor(s)”*?

$f(x)$

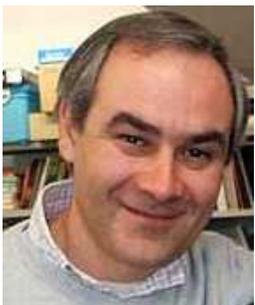


**(Q)SAR**: this is the base of a quantitative structure-activity relationship (QSAR): find *patterns!*

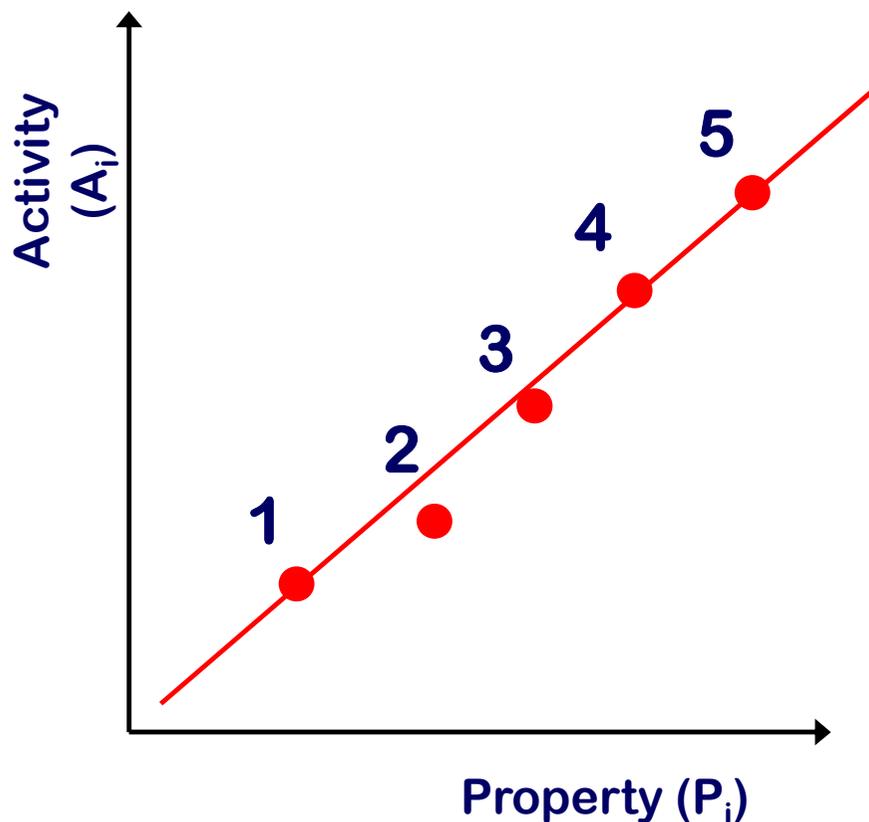




Yes, we can look for “*regularity*”  
(**pattern**) between the variability  
of molecular descriptors and the  
corresponding variability of  
experimental activities.

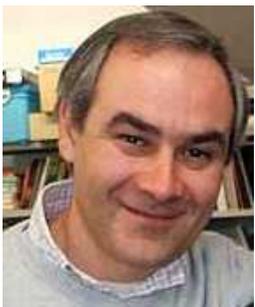


# The beauty of mathematics:



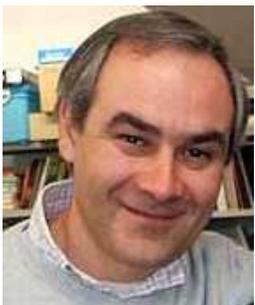
$$A_i = a P_i + b$$

**Discrete (few  $x$ - $y$  correspondences)**  
**Continuum ( $\infty$   $x$ - $y$  correspondences)**

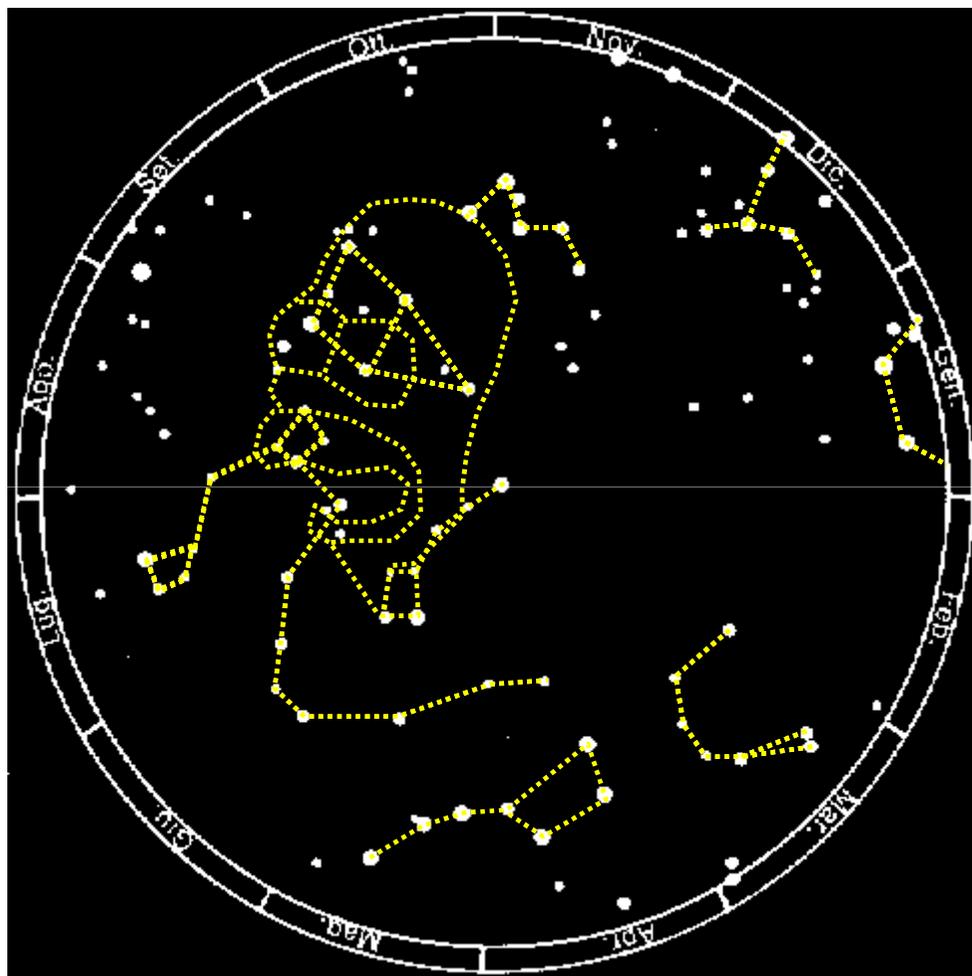


## Patterns are gorgeous:

- 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;

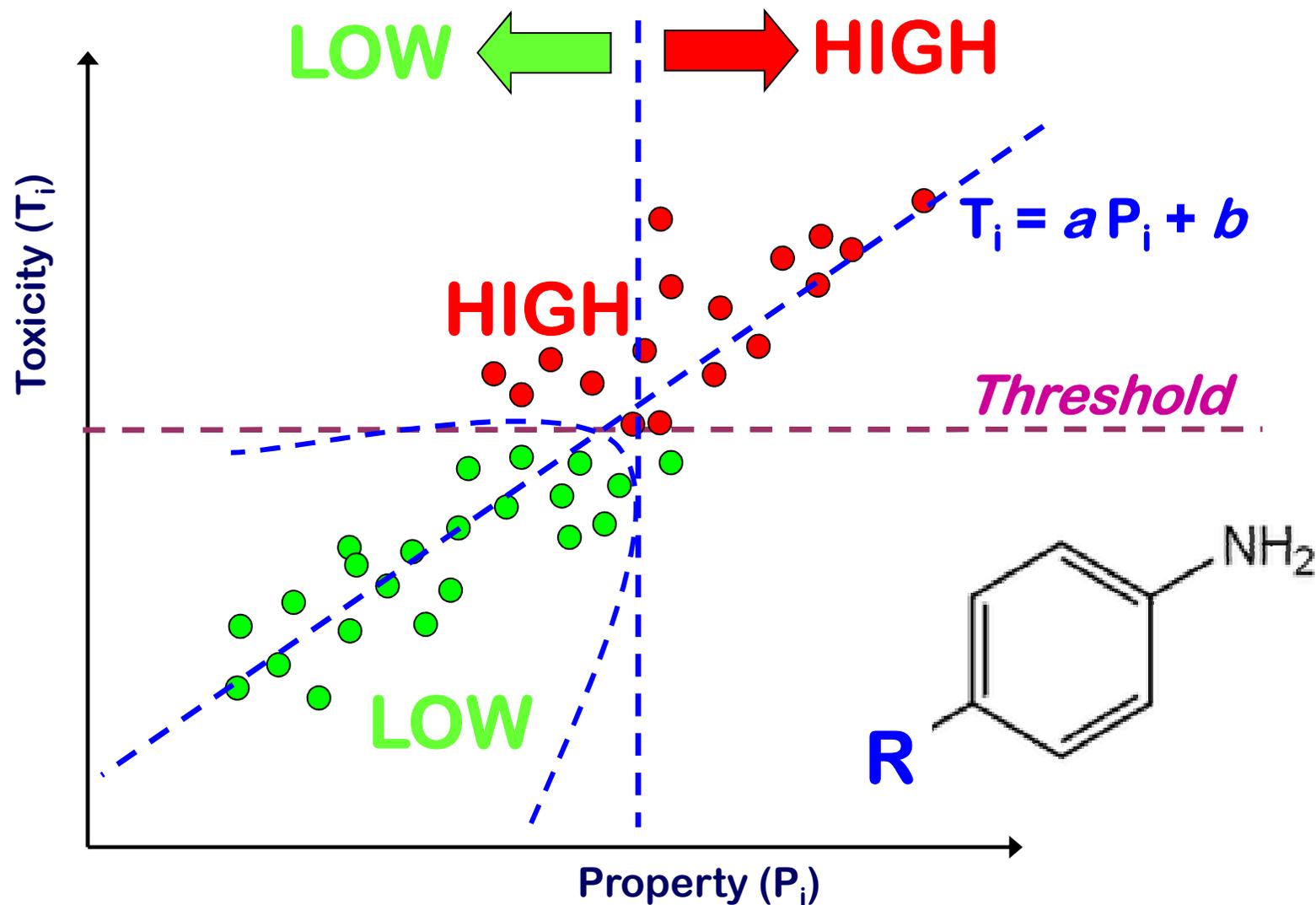


sometimes too gorgeous...



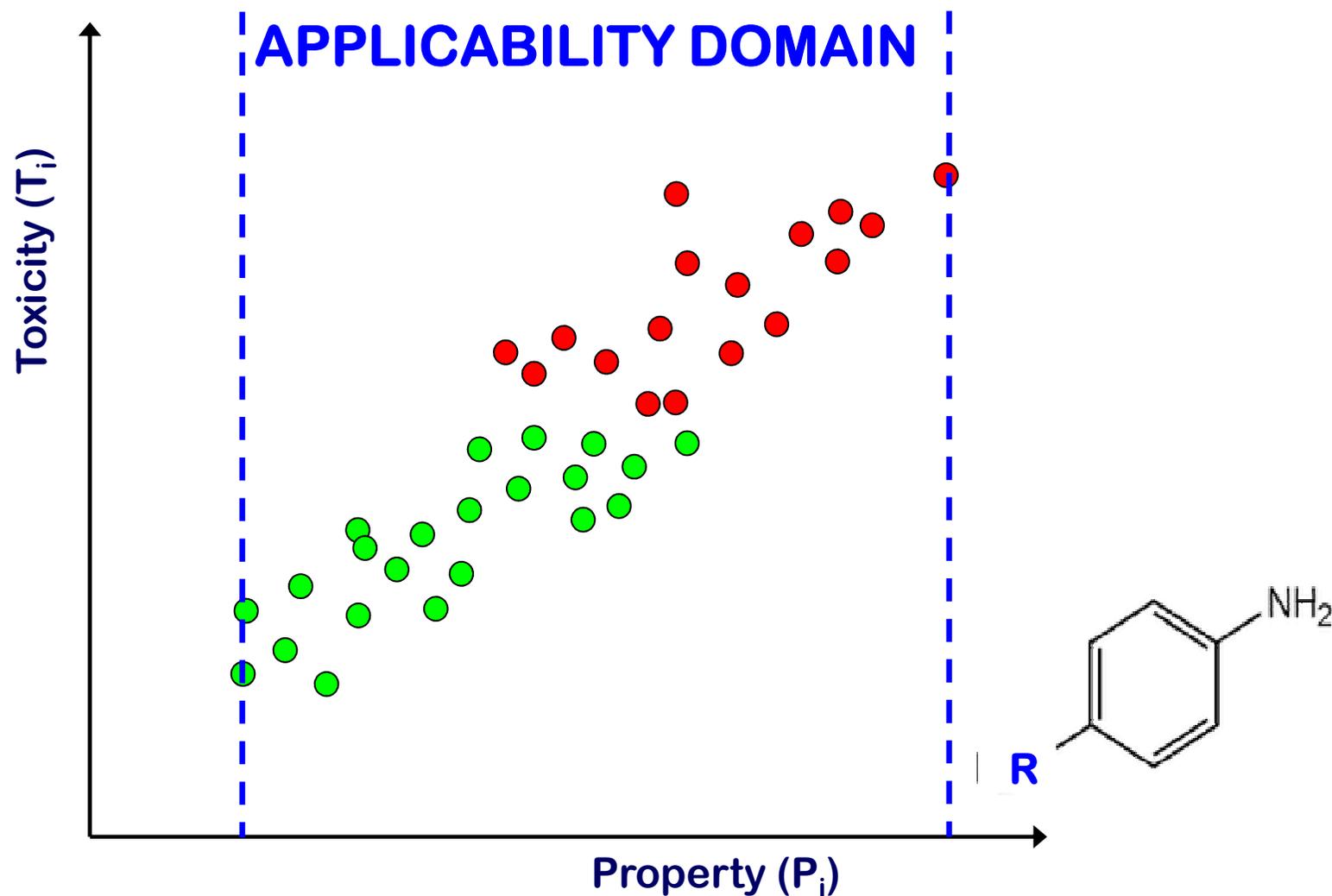


# The scatter plot: the best place where explore (Q)SAR.





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



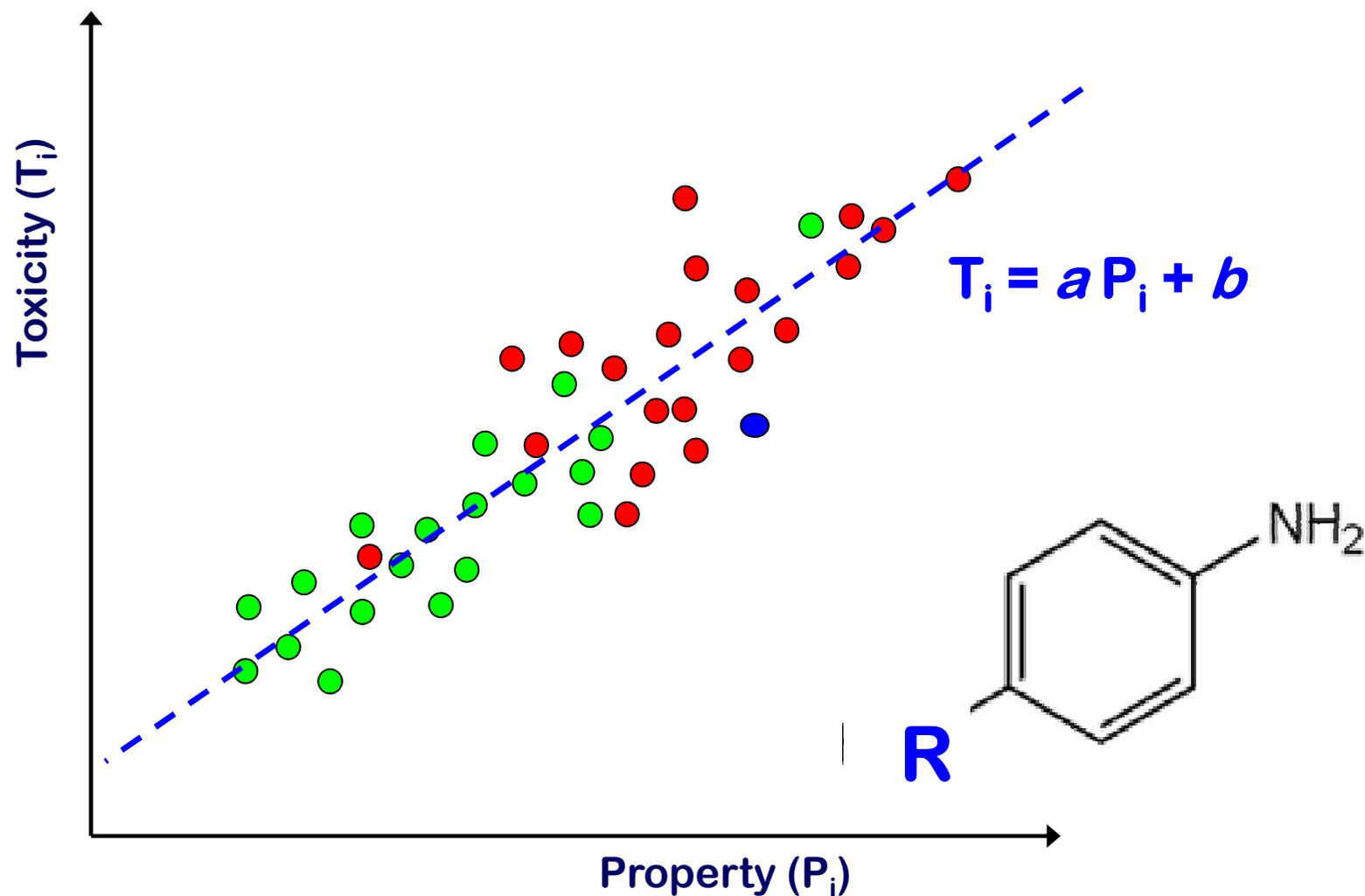


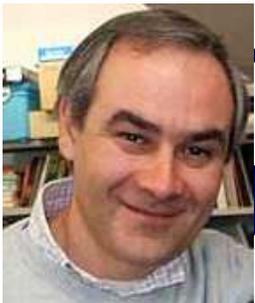
# sometimes too gorgeous...

- ⊙ regression models (quantitative response)
- ⊙ classification models (qualitative response)
- ⊙ ranking models (ordered response)



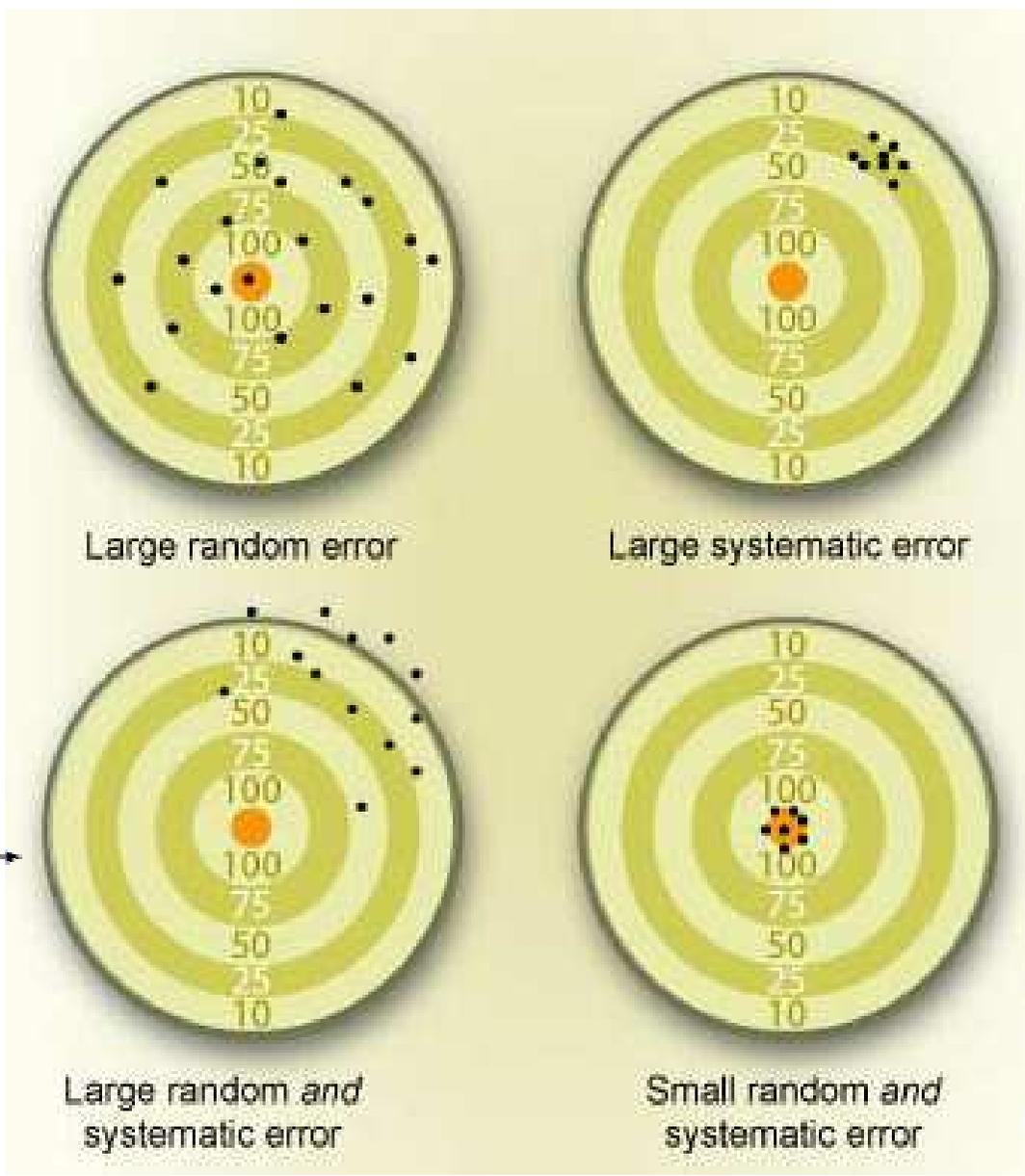
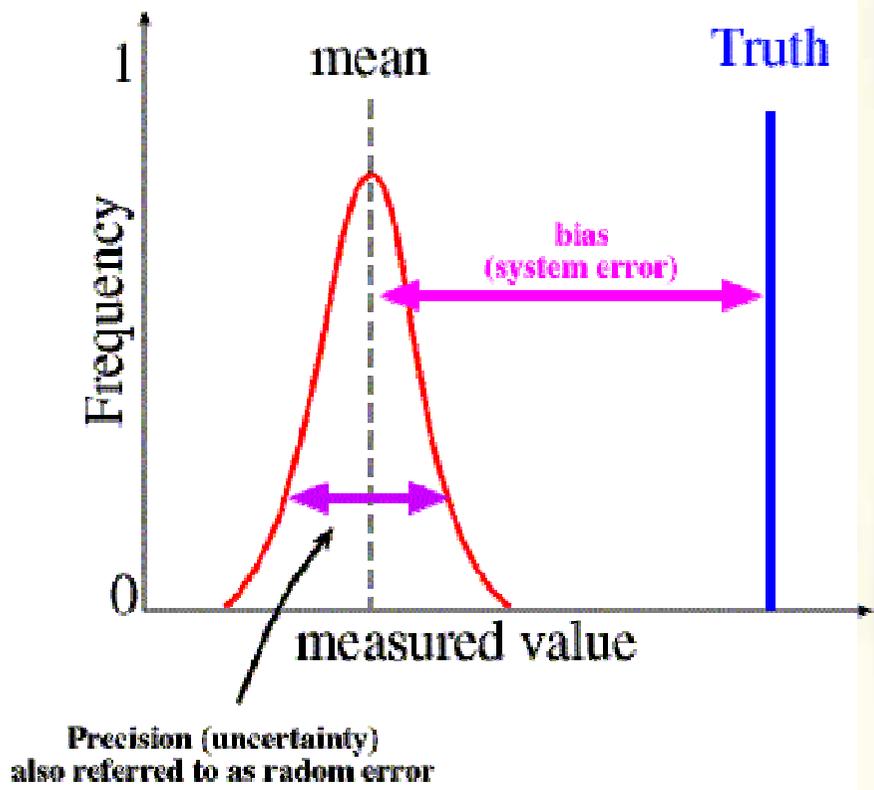
# The scatter plot: the best place where explore (Q)SAR.

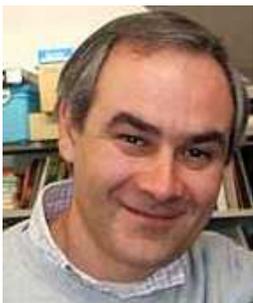




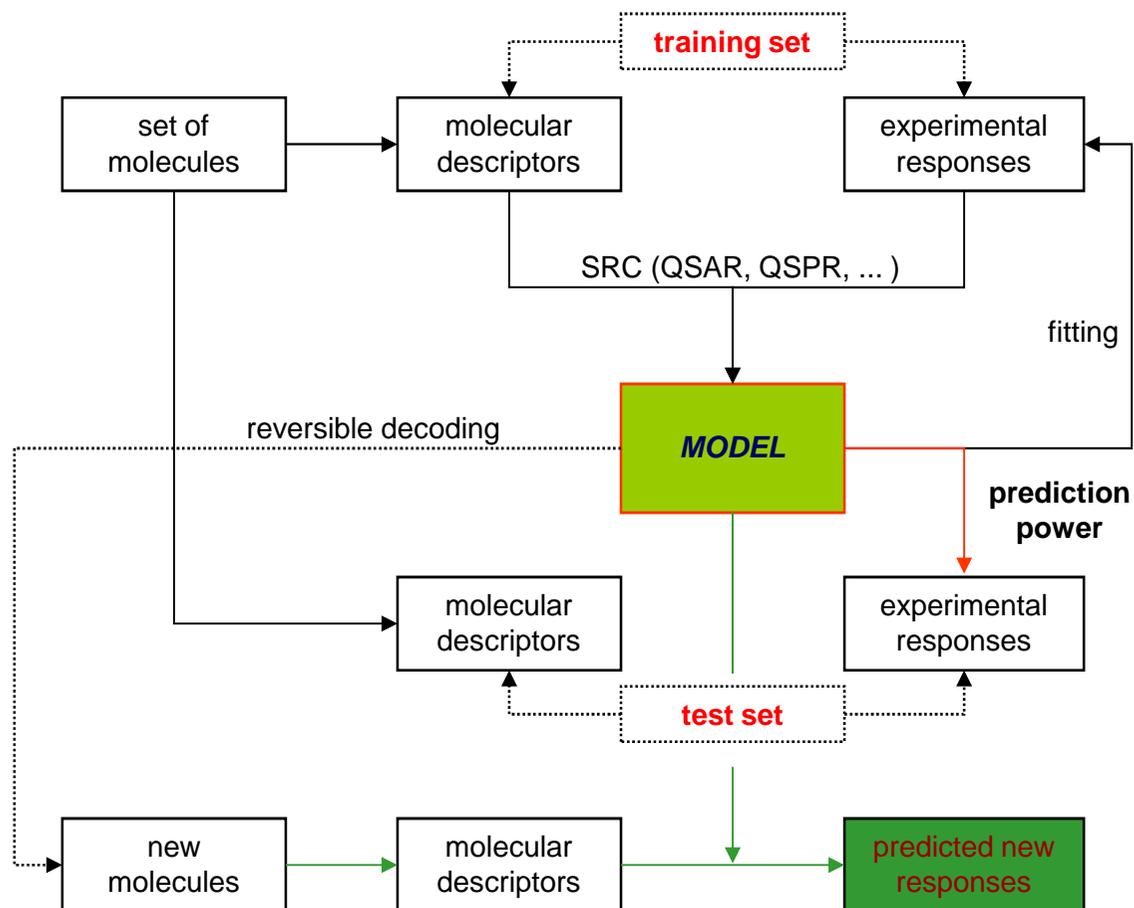
## The two statistical **gold** rules do build up linear models:

- For each independent variable (*molecular descriptor*) you need at least five (5) dependent variable values (*activities*).
- The dependent variable values (*activities*) must be accurate and precise.



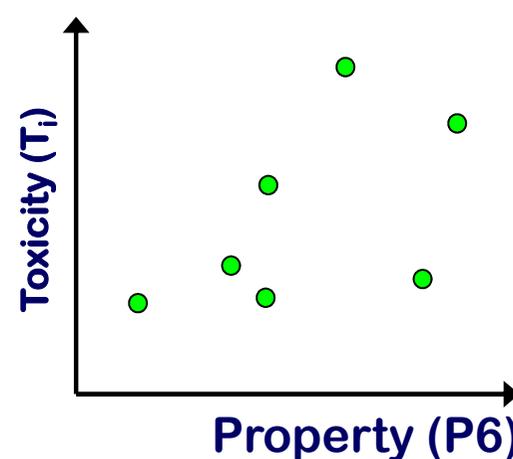
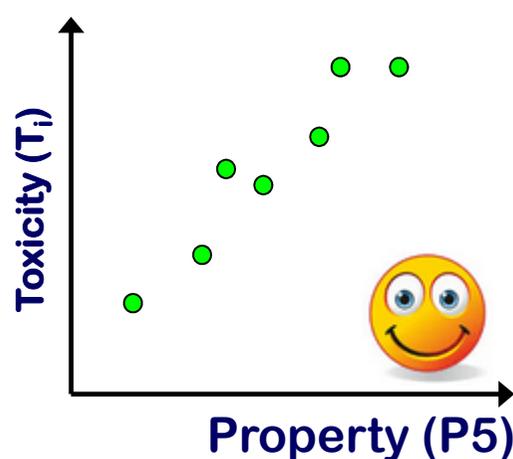
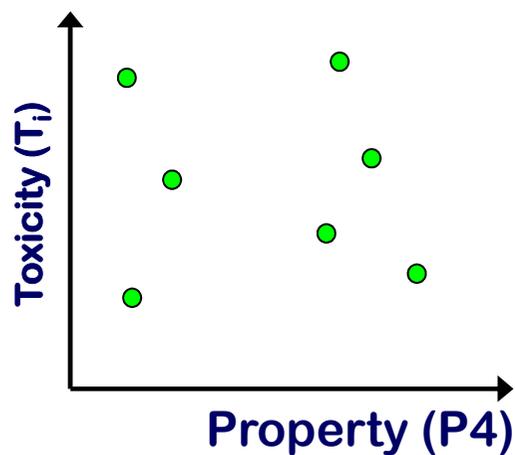
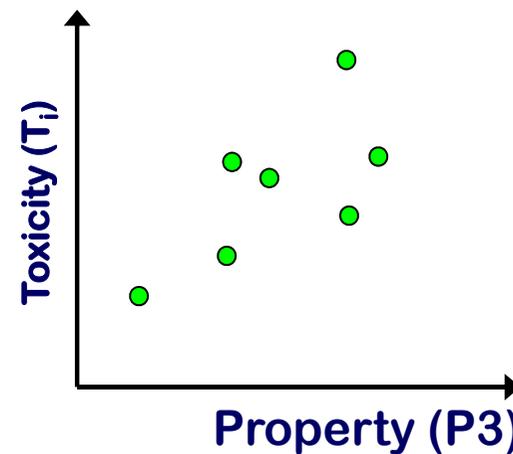
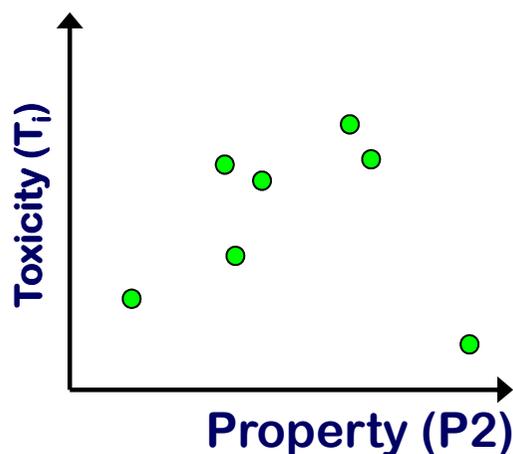
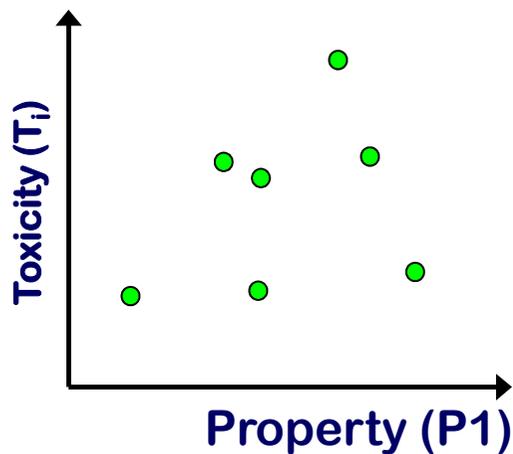


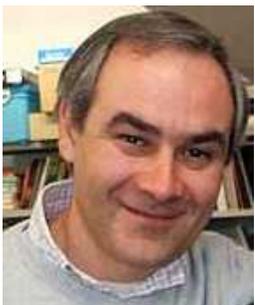
# Here a possible work-flow:



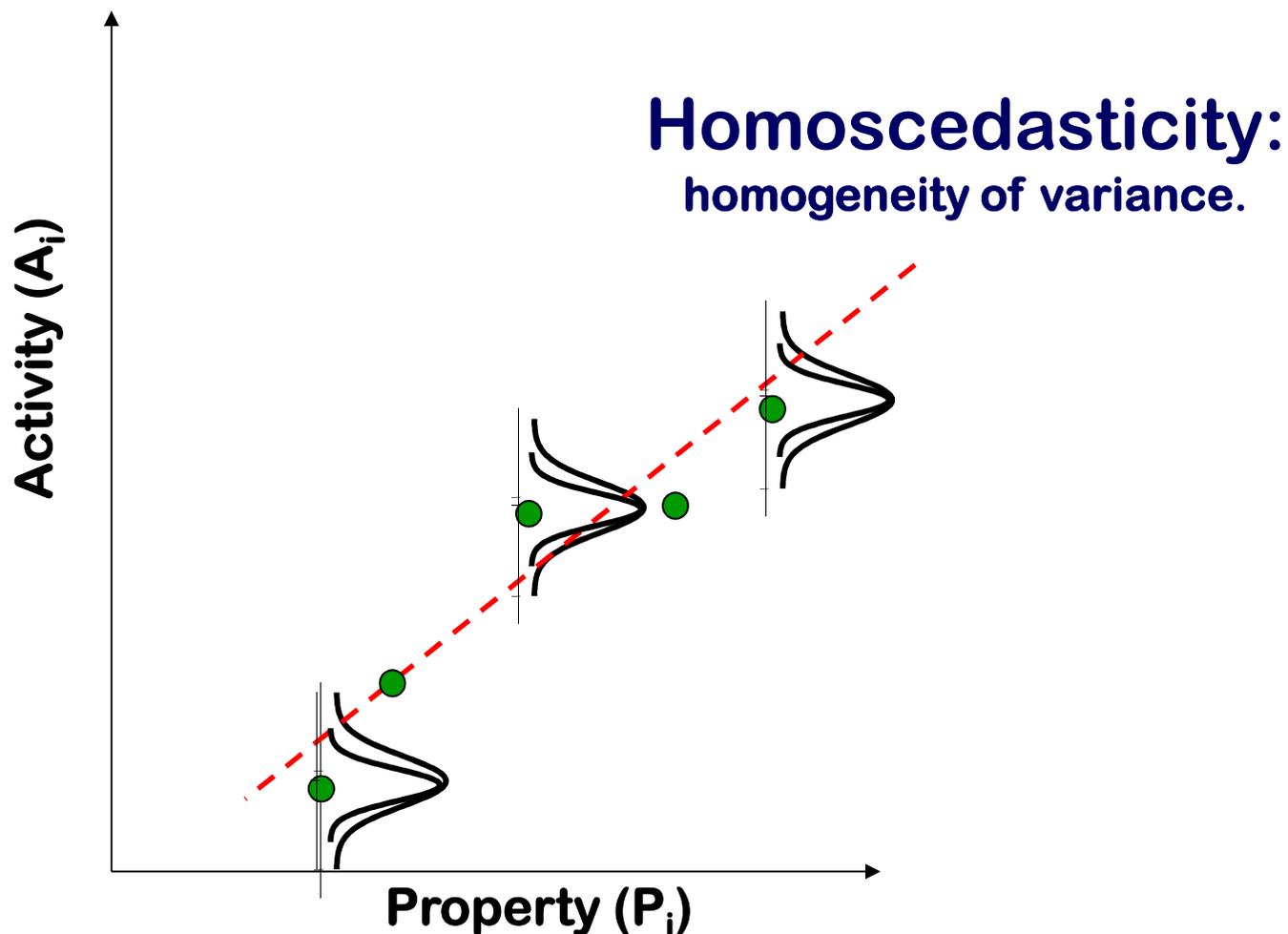


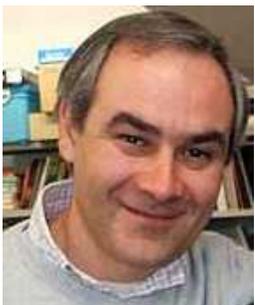
# How we can select the *good* descriptors?





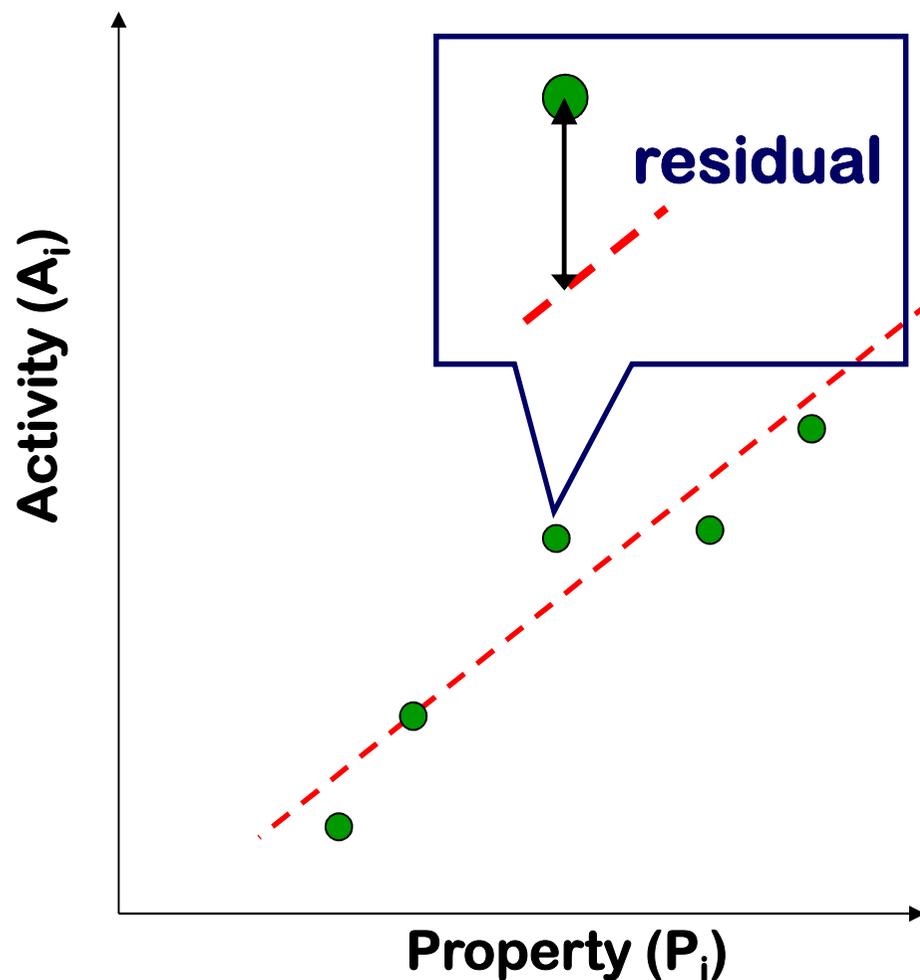
Now, how can we select the “*good*” linear model:



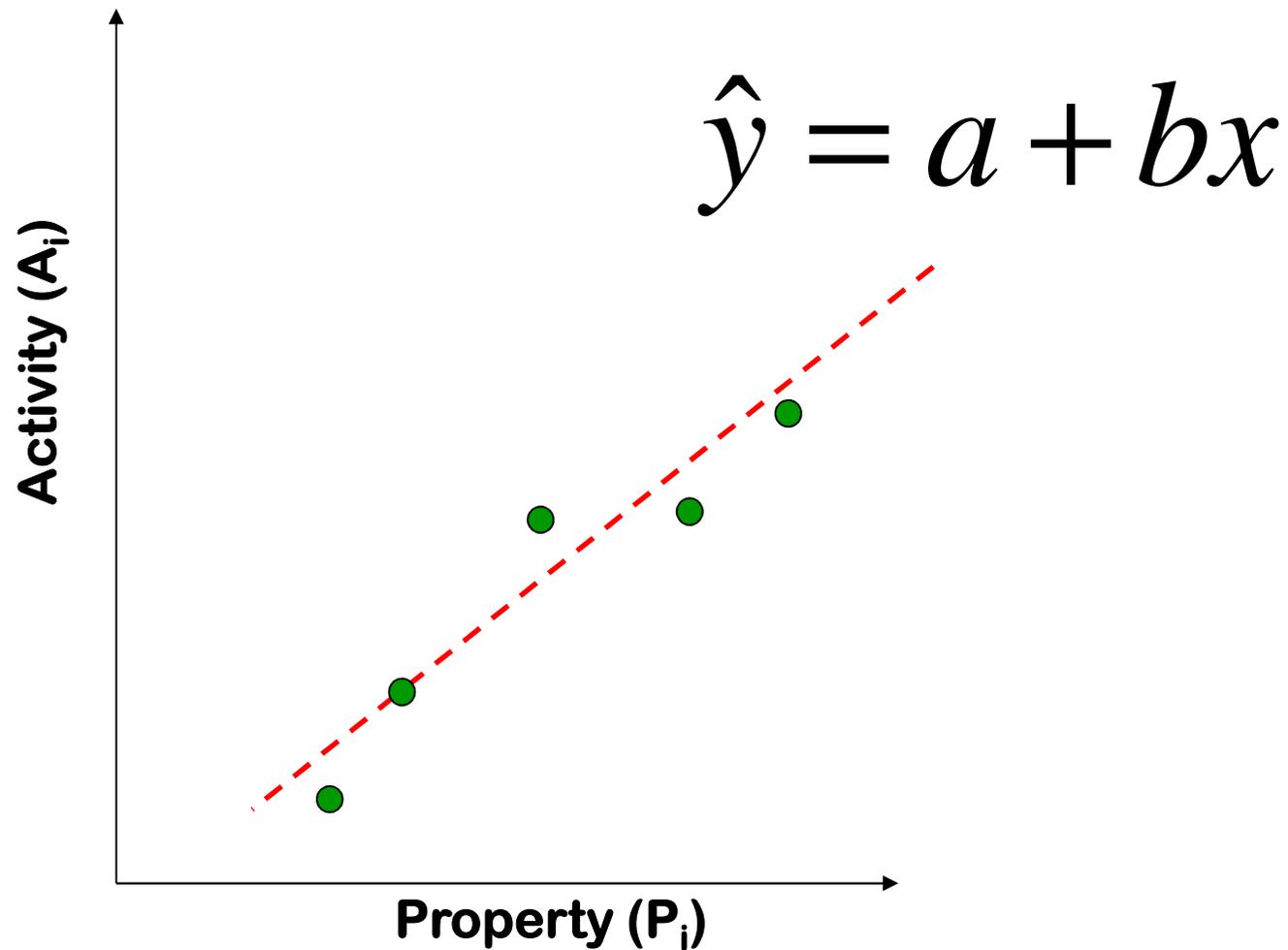


# do you remember... Least Squares Analysis?

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 ( $\hat{y}$ ) and observed values ( $y$ ) squared.



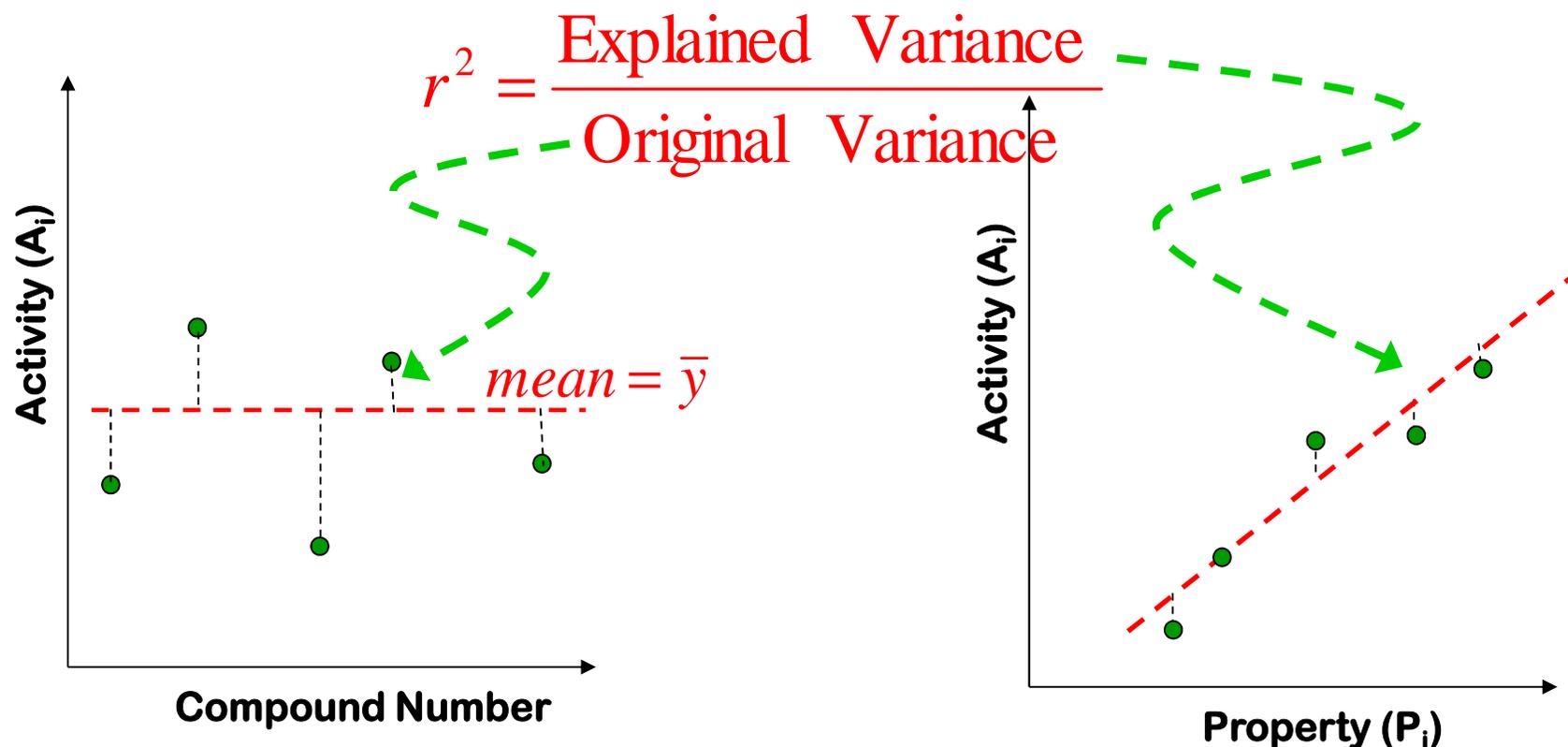
$$e = \hat{y} - y$$



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

$$a = \frac{\sum y}{n} - b \frac{\sum x}{n}$$

**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:



**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:**

$$ESS = \sum_{i=1}^N (y_{i,calc} - \bar{y})^2$$

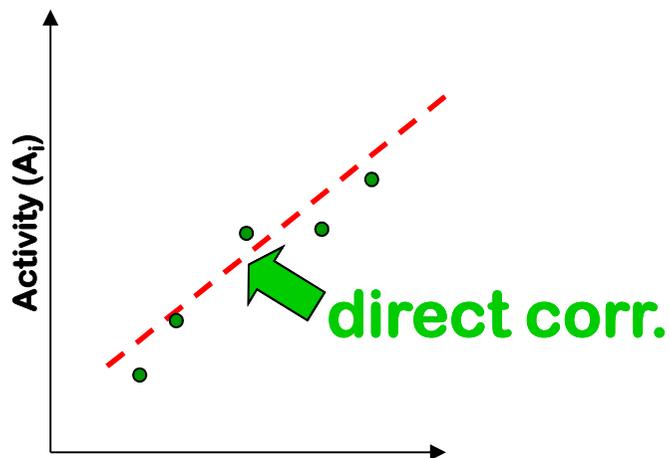
- **Variance around regression line:**

$$RSS = \sum_{i=1}^N (y_i - y_{calc,i})^2$$

$$r^2 = \frac{ESS}{TSS} \equiv \frac{TSS - RSS}{TSS} \equiv 1 - \frac{RSS}{TSS} \quad 0 < r^2 < 1$$

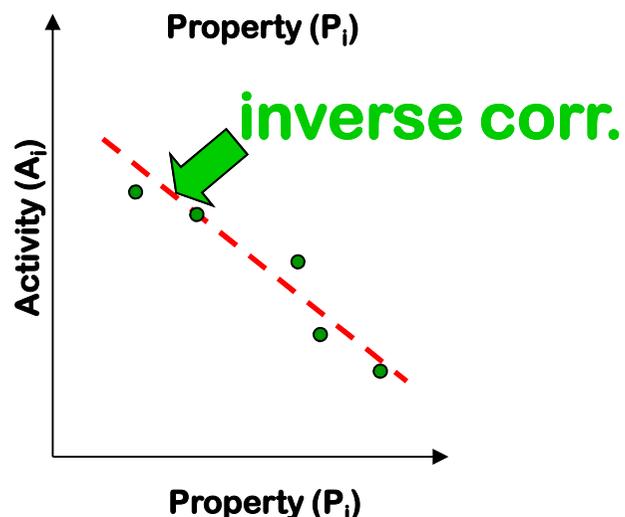
Possible values reported for  $r^2$  fall between 0 and 1. For example: with  $r^2$  of 0.83, you can say that 83% of the variability in activity can be explained by the different value of the selected molecular property. The remaining 17% of variability is due to other unexplained factors.

**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:



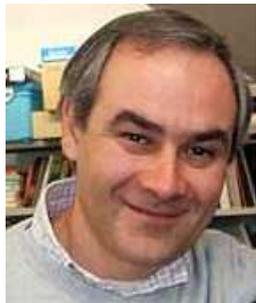
$$r^2 = \frac{ESS}{TSS} \quad r = \sqrt{\frac{ESS}{TSS}}$$

$$0 < r < 1$$

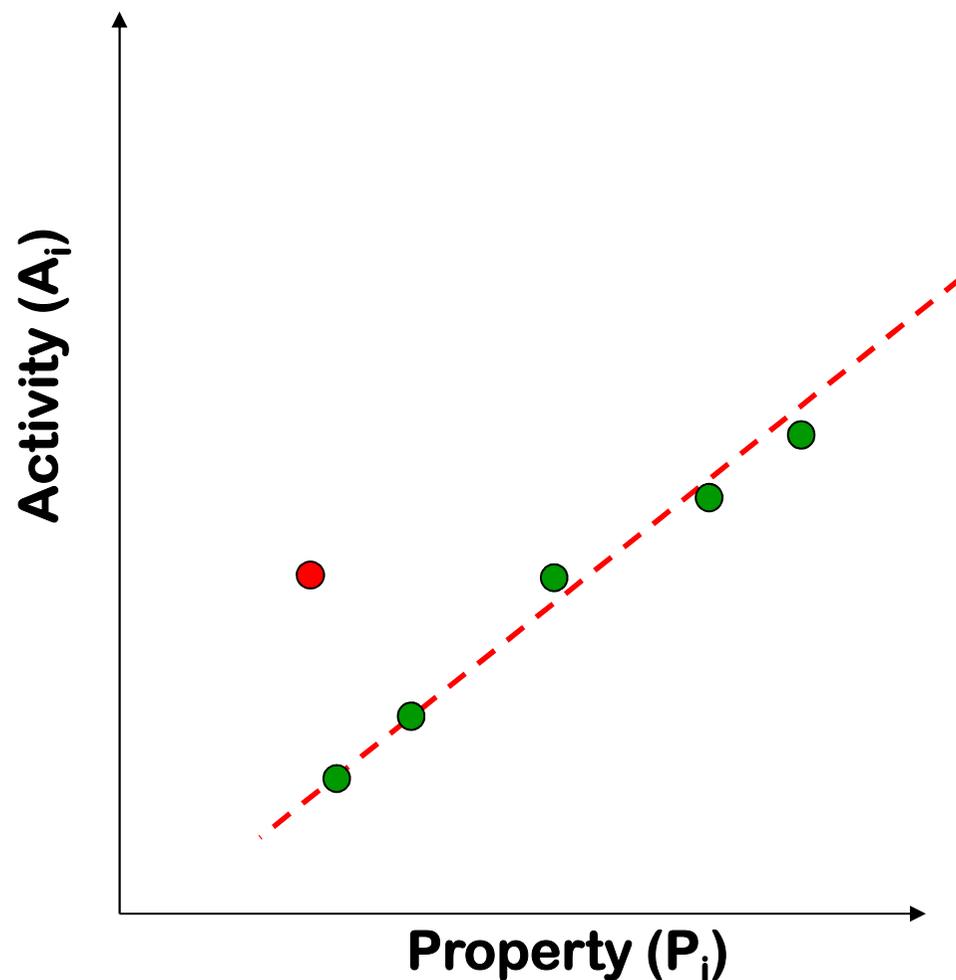


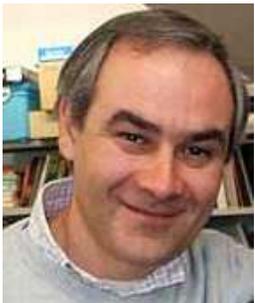
A perfect correlation of  $\pm 1$  occurs only when the data points all lie exactly on a straight line. A correlation greater than **0.8** would be described as strong, whereas a correlation less than **0.5** would be described as weak.

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

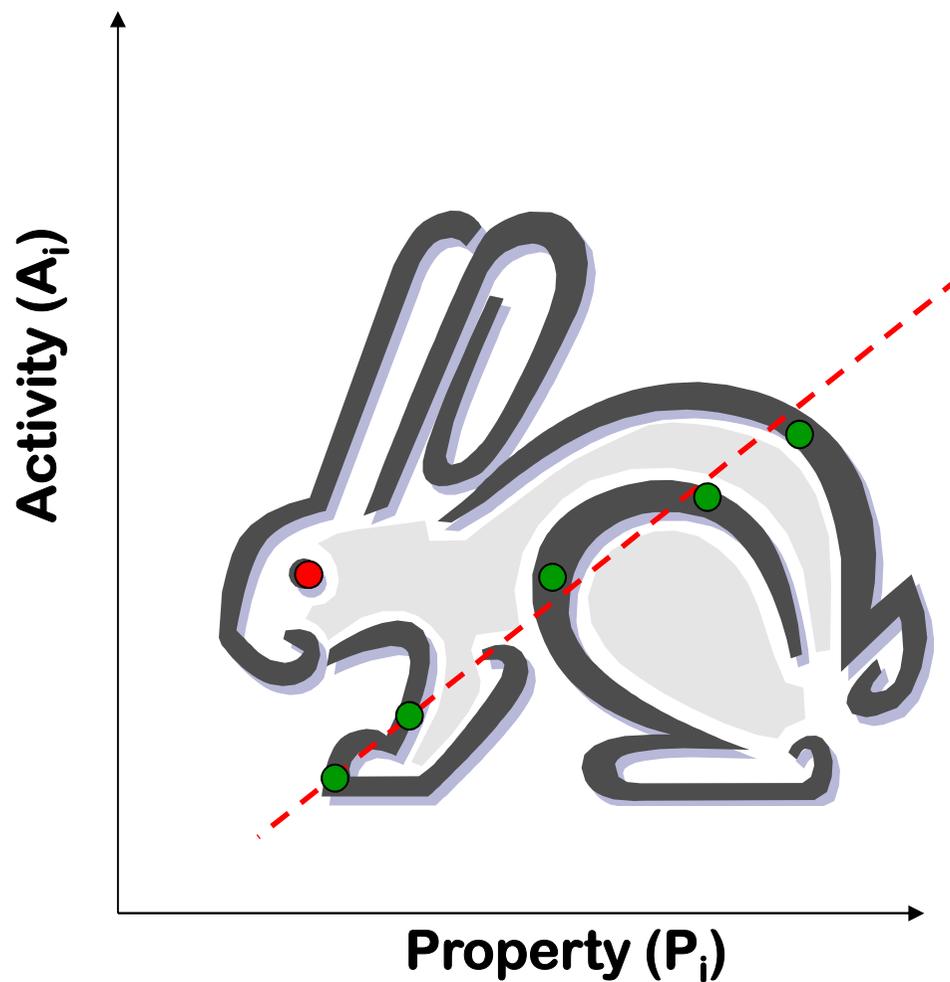


**How do we deal with them, usually?**

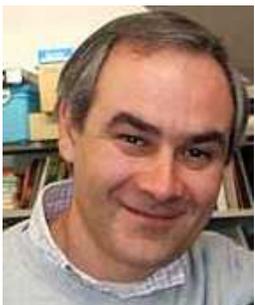




**Be carefull...**



**... the rabbit is out there!!!**



## (Q)SAR: a concrete example...

# Aquatic toxicity: non-polar narcotics and neutral organic compounds



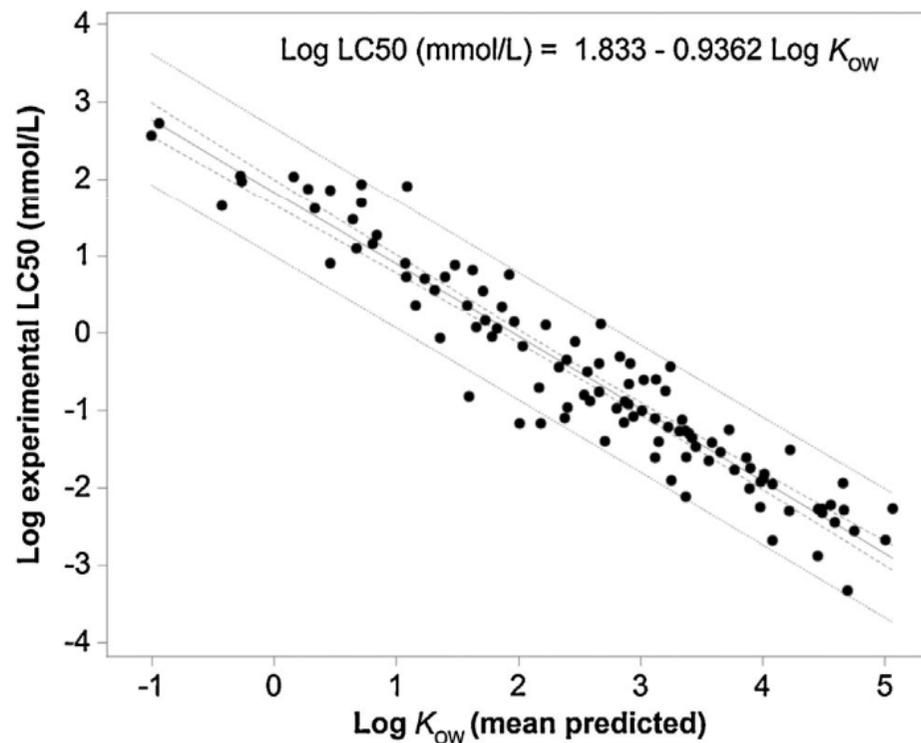
*Poecilia reticulata*

*Oncorhynchus mykiss*

*Cyprinus carpio*

**Training set:**

**$n = 161$   $r^2 = 0.912$   $sd = 0.413$**

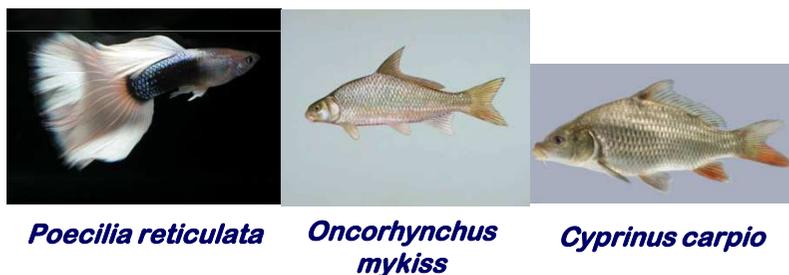


Handbook of Pesticide Toxicology: Principles. Chapter 29. Robert Irving Krieger (2001, 2 edition)

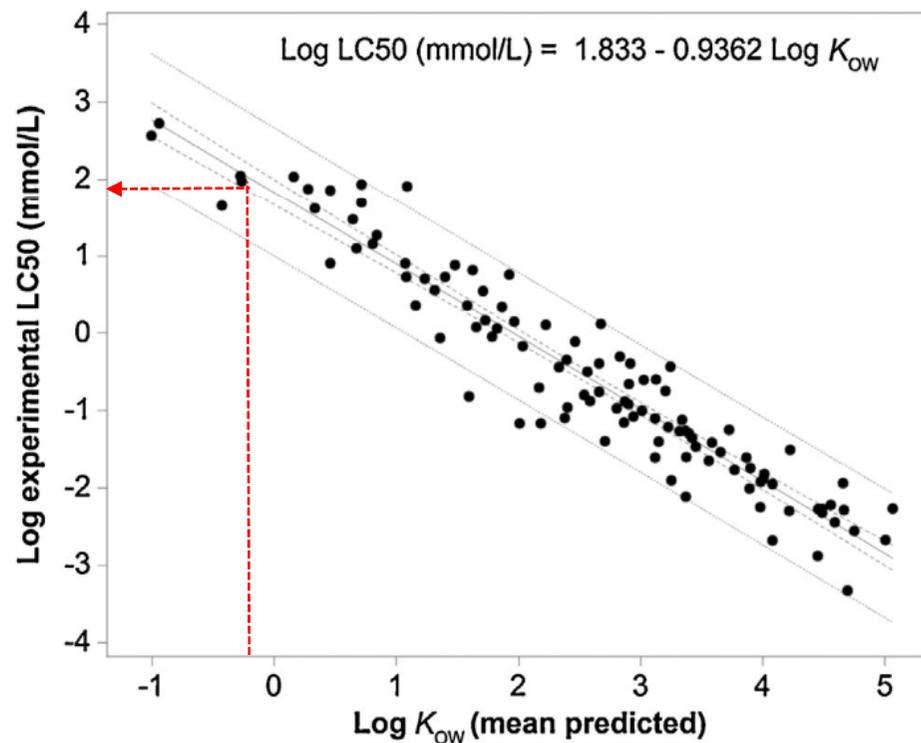
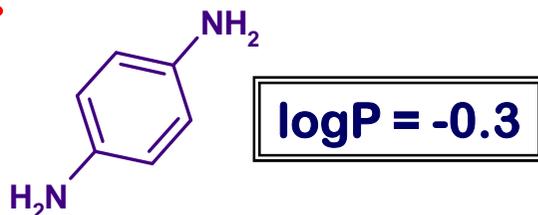


## (Q)SAR: a concrete example...

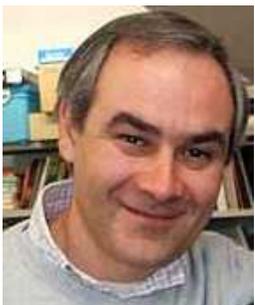
# Aquatic toxicity: non-polar narcotics and neutral organic compounds



### Test set:

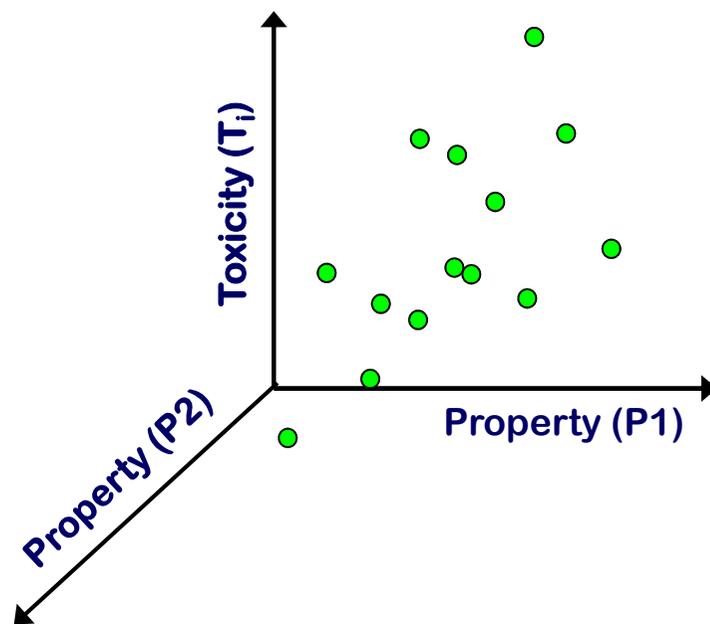


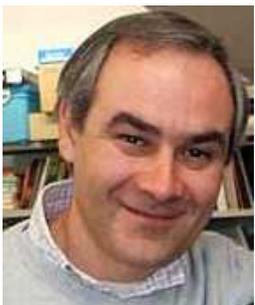
Handbook of Pesticide Toxicology: Principles. Chapter 29. Robert Irving Krieger (2001, 2 edition)



## (Q)SAR: a different scenario...

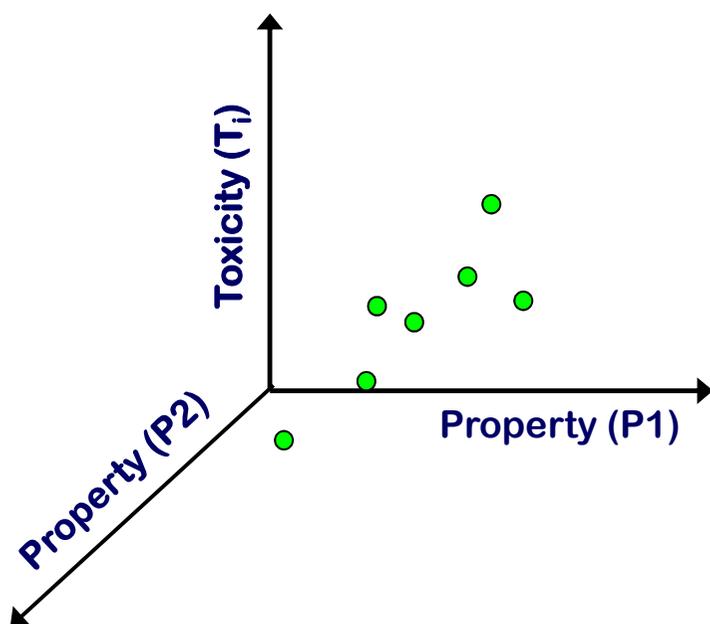
and if the experimental activity depended on more than one descriptor?





## (Q)SAR: multiple regression analysis (MRA)

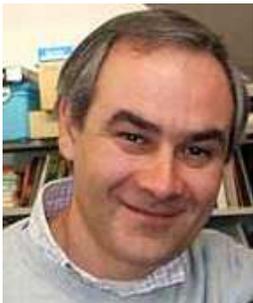
and if the experimental activity depended on more than one descriptor?



$$y_1 = a_1x_1 + b_1x_2 + z_1$$

...

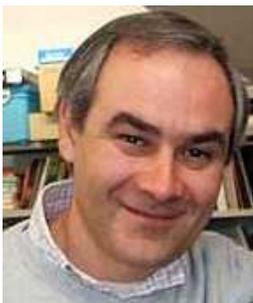
$$y_n = a_nx_n + b_nx_n + z_n$$



# Multiple Regression Analysis (MRA)

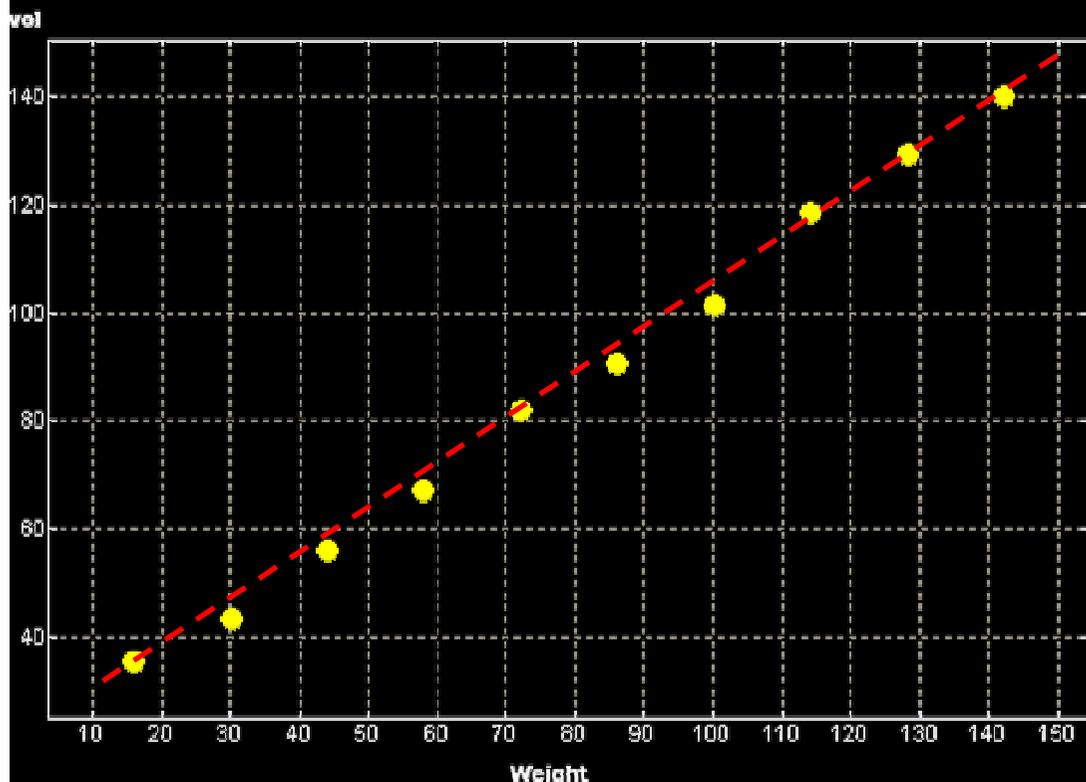
## *Requirements:*

- 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

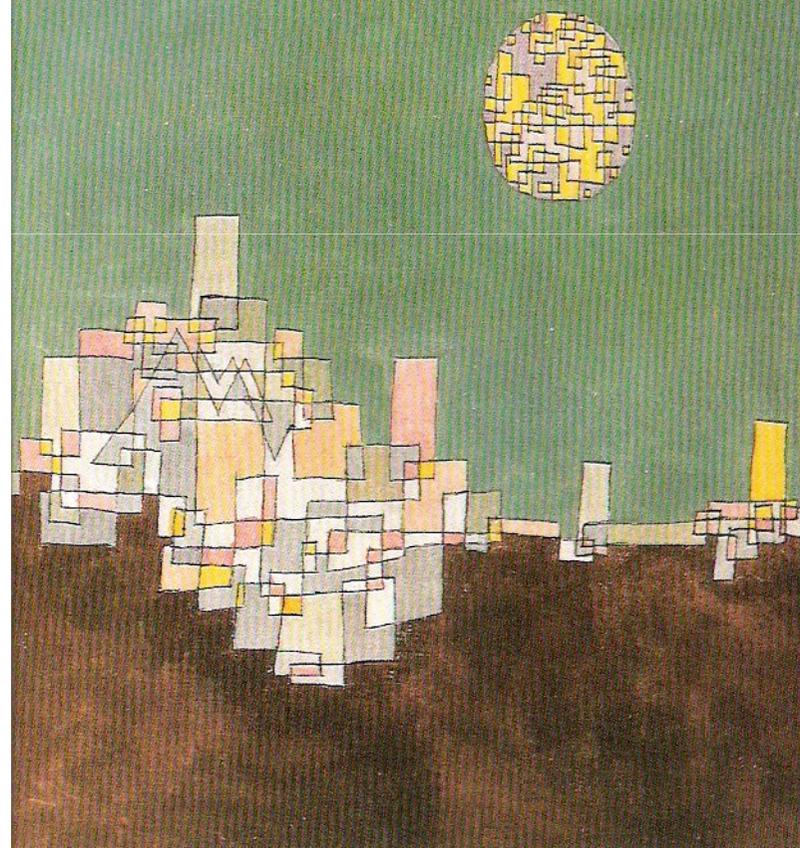
$$r^2 = 0.9973$$

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

Signet Classic

451-CE2290 \* (CANADA \$5.99) \* U.S. \$4.95

EDWIN  
A. ABBOTT  
**FLATLAND**  
A ROMANCE OF  
MANY DIMENSIONS



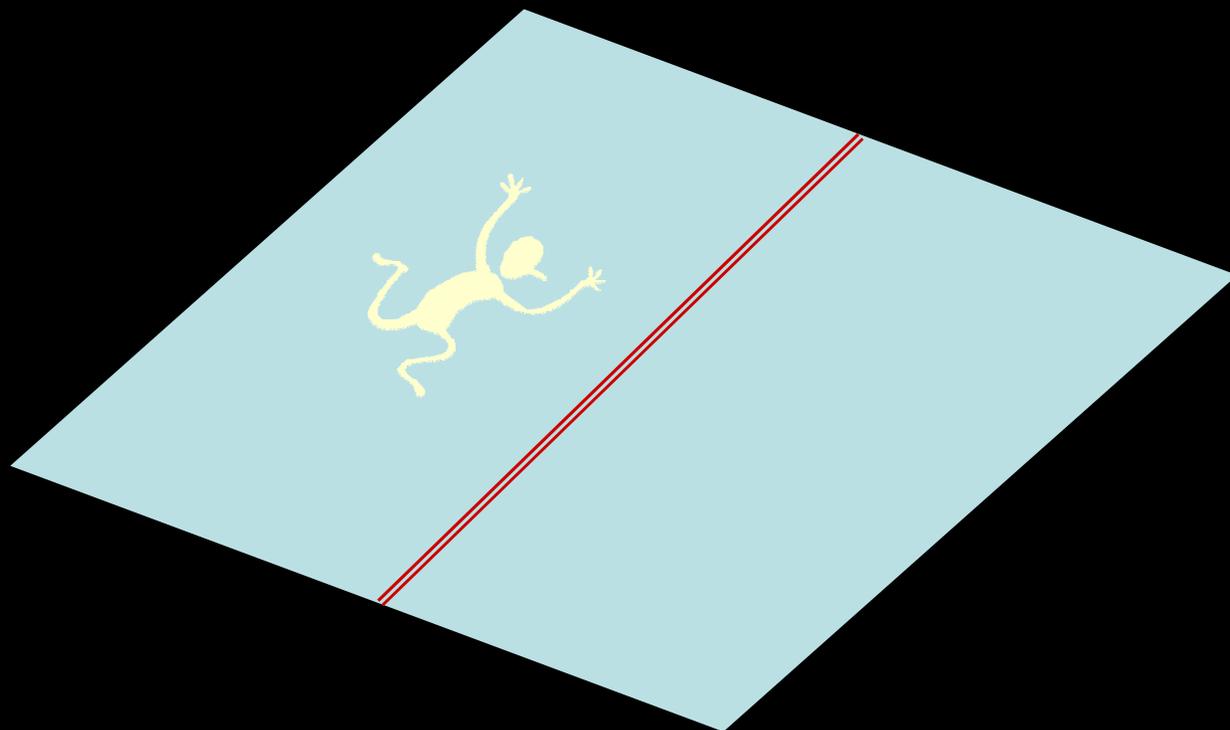
MS  
M

Confidential and Property of ©2012 Molecular Modeling Section  
Dept. Pharmaceutical and Pharmacological Sciences – University of Padova - Italy

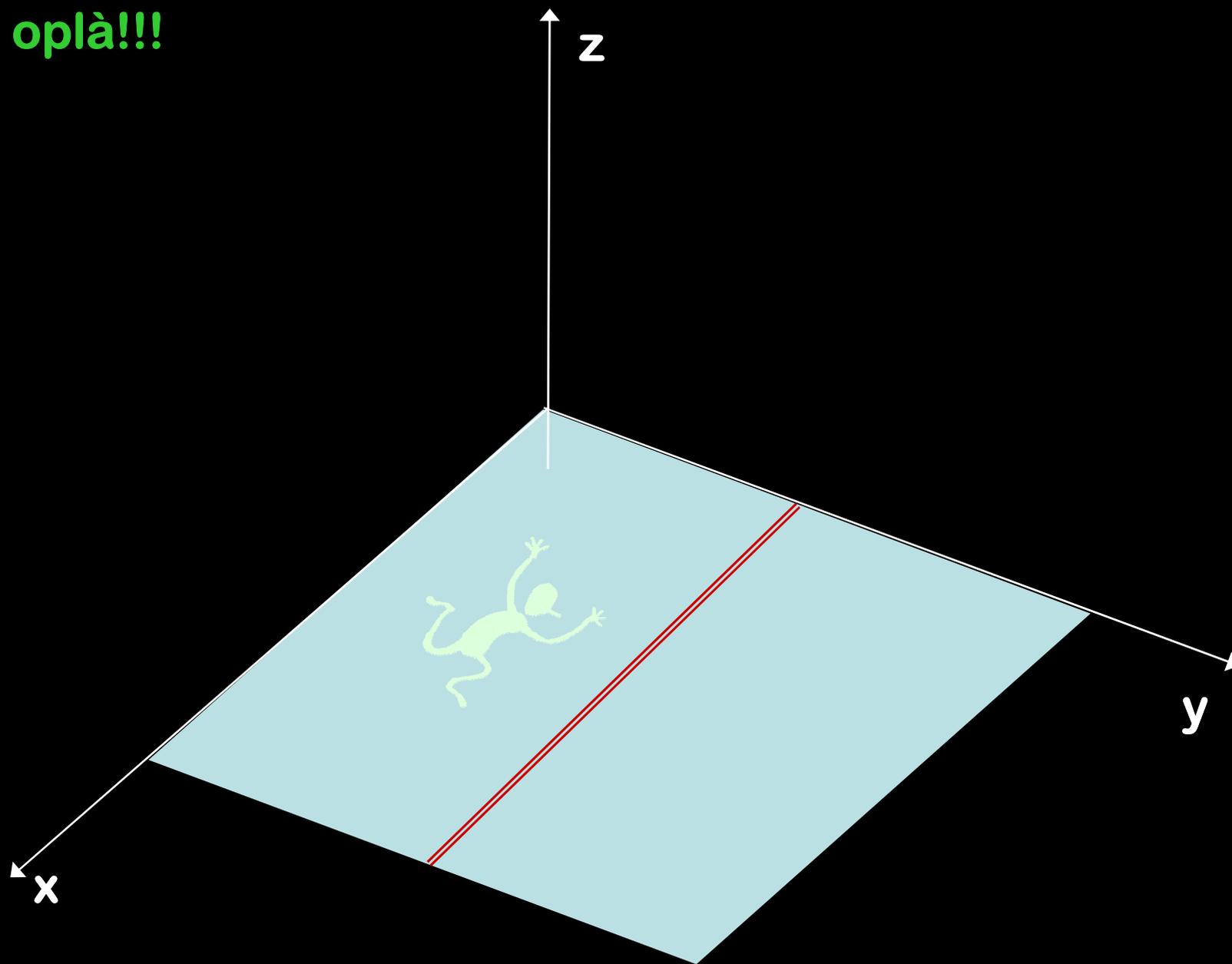
S. MORO – SSVGRC



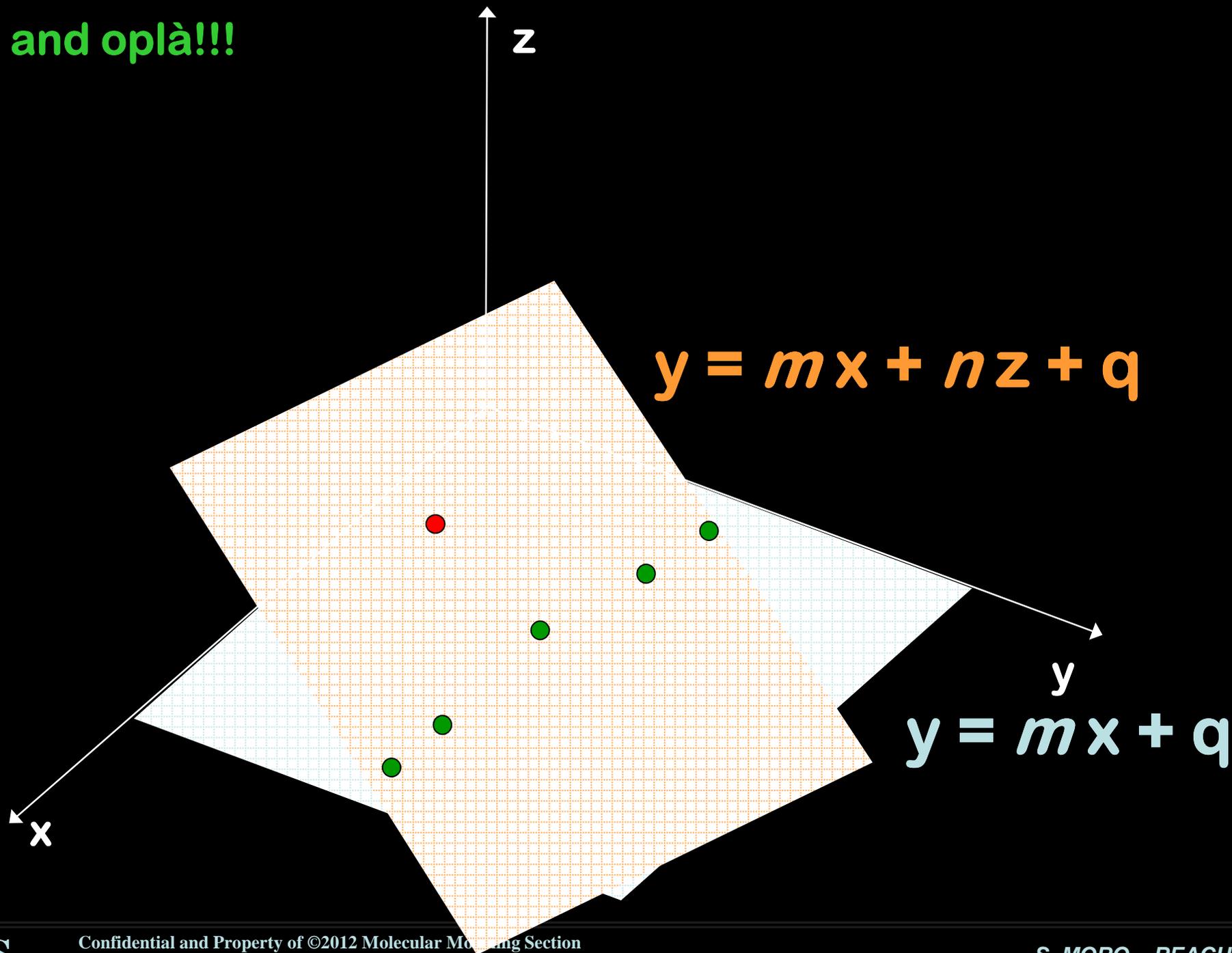
# MRA approaches can transform the life in Flatland!

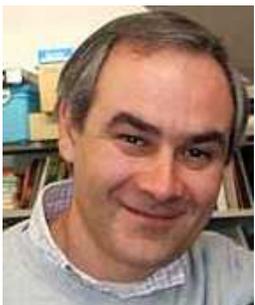


... opla!!!



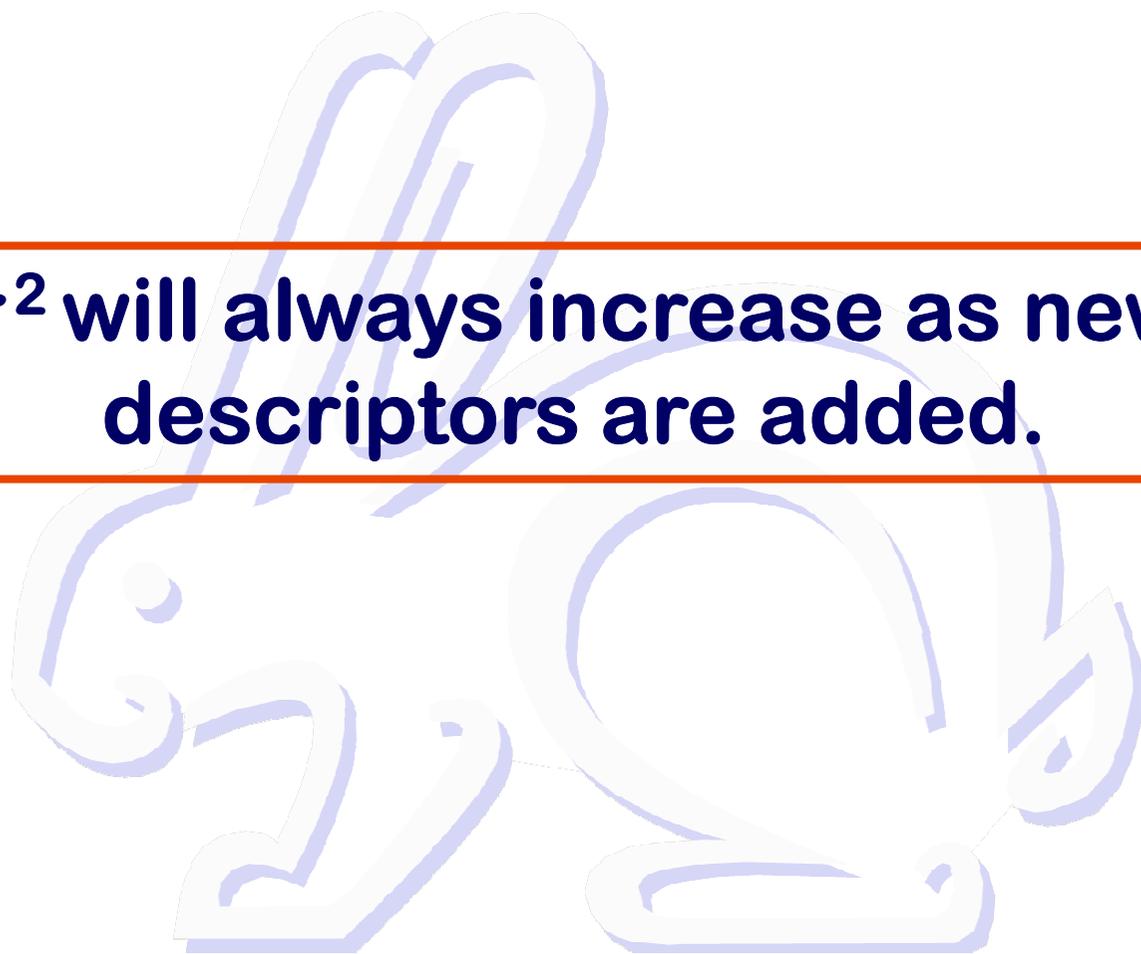
... and oplà!!!





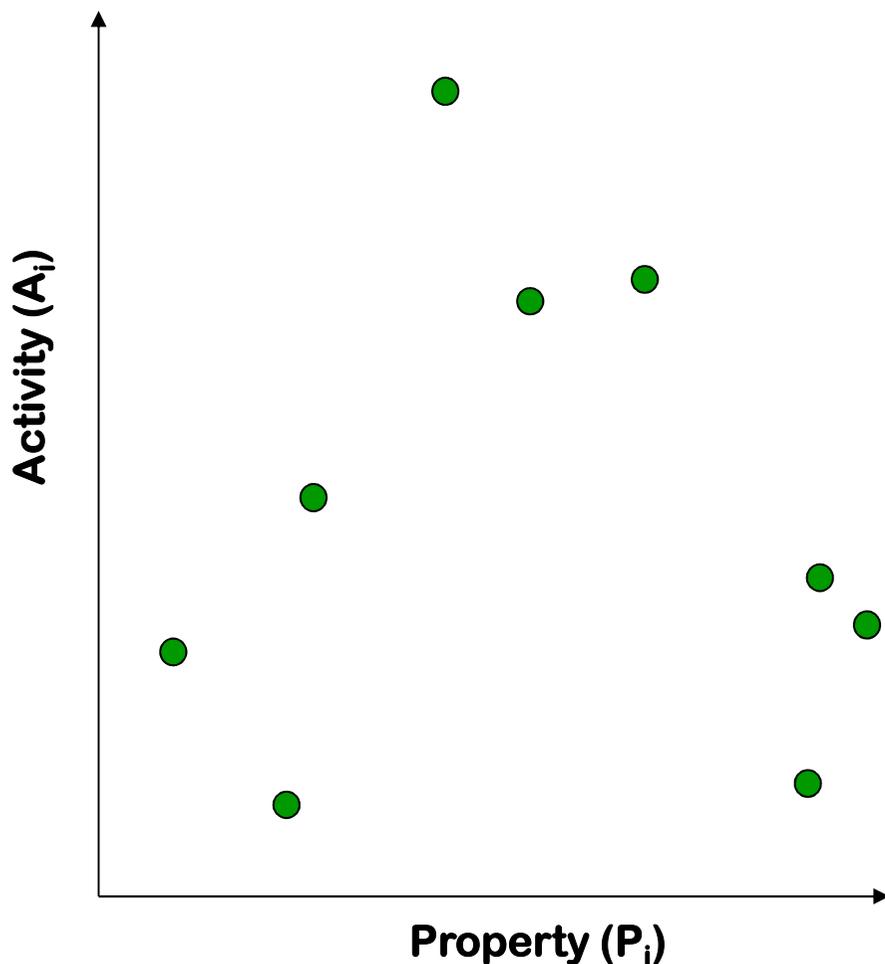
## Here is the MRA nightmare:

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





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

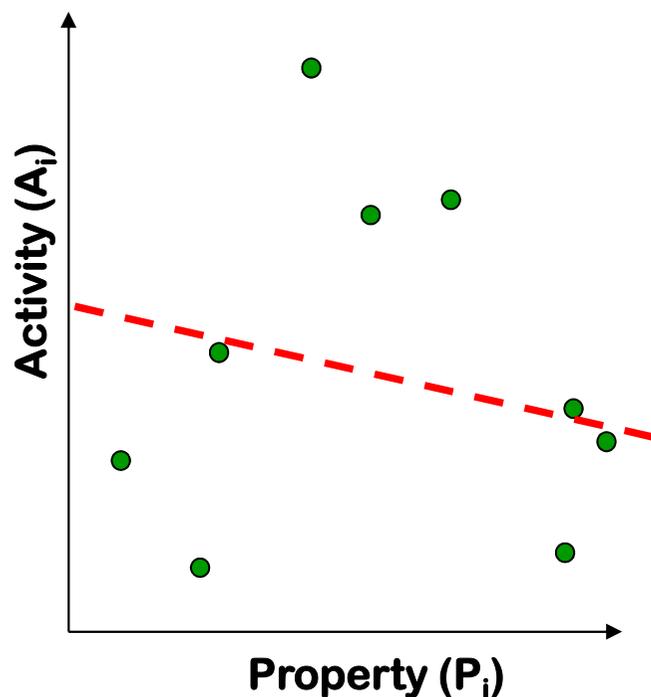


$$y = f(x) + \text{noise}$$

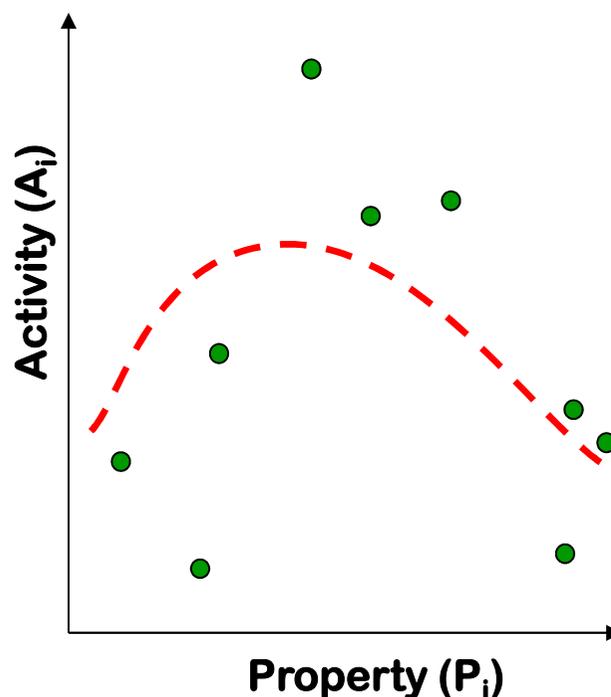
How we can deal with these data?

Let's consider three different methods...

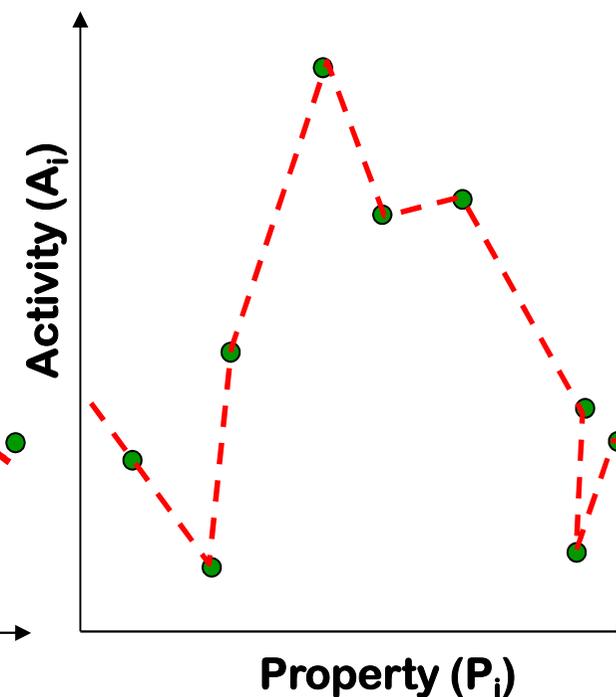
## Linear



## Quadratic



## Joint-the-dots

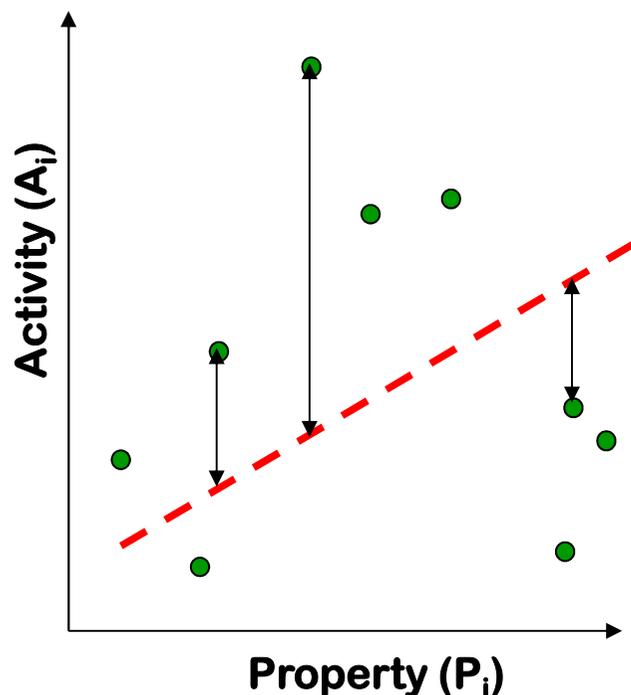


**“How well are you going to predict future data drawn from the same distribution?”**

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



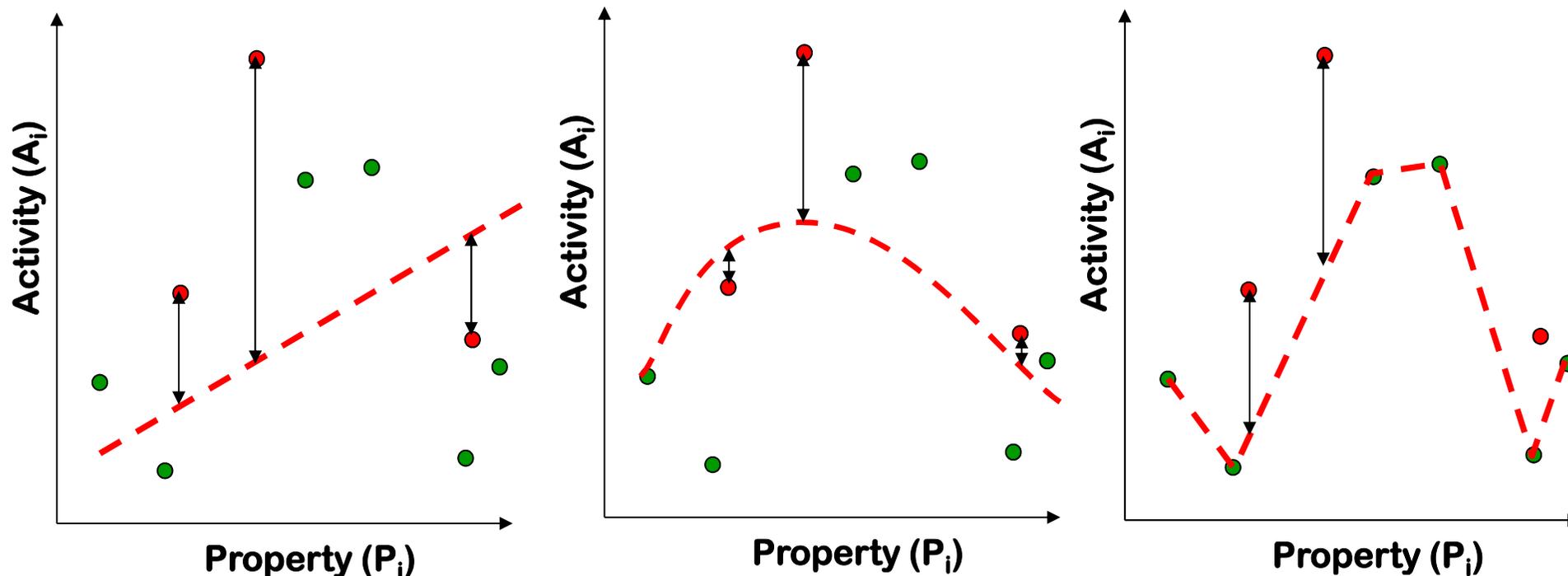
# The “*test set*” method...looks this:



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:*

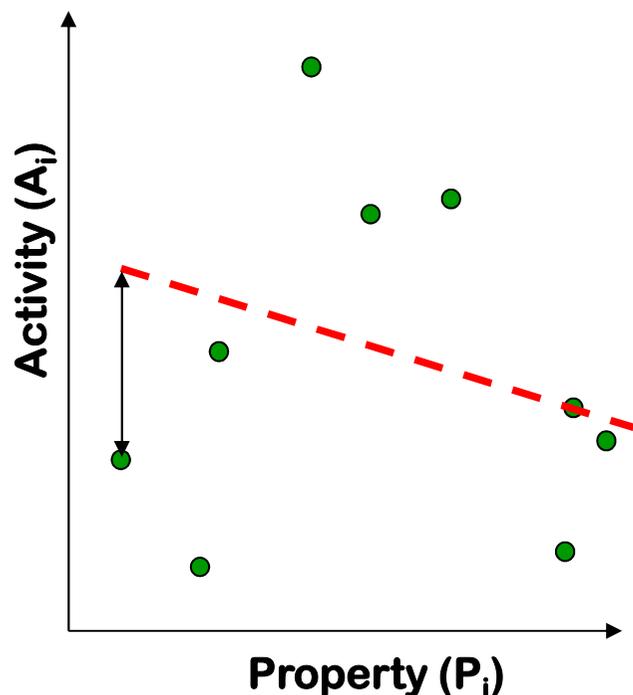
- Wastes data: we get an estimate of the best method to apply to 30% less data;
- If we don't have much data, our test-set might just be lucky or unlucky.



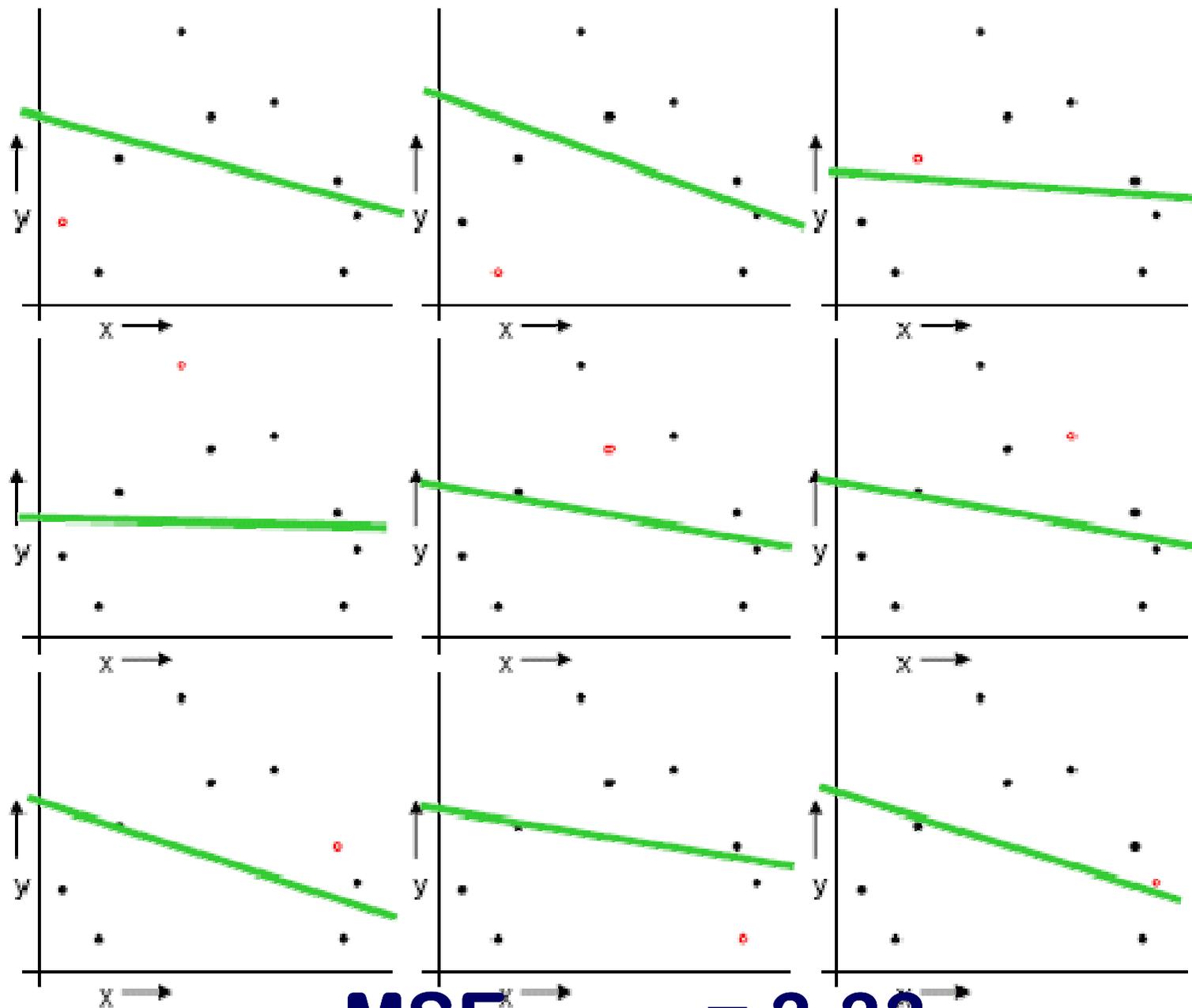
# or “*LOOCV*” (Leave-One-Out Cross Validation) method... looks this:

For each data consider this loop:

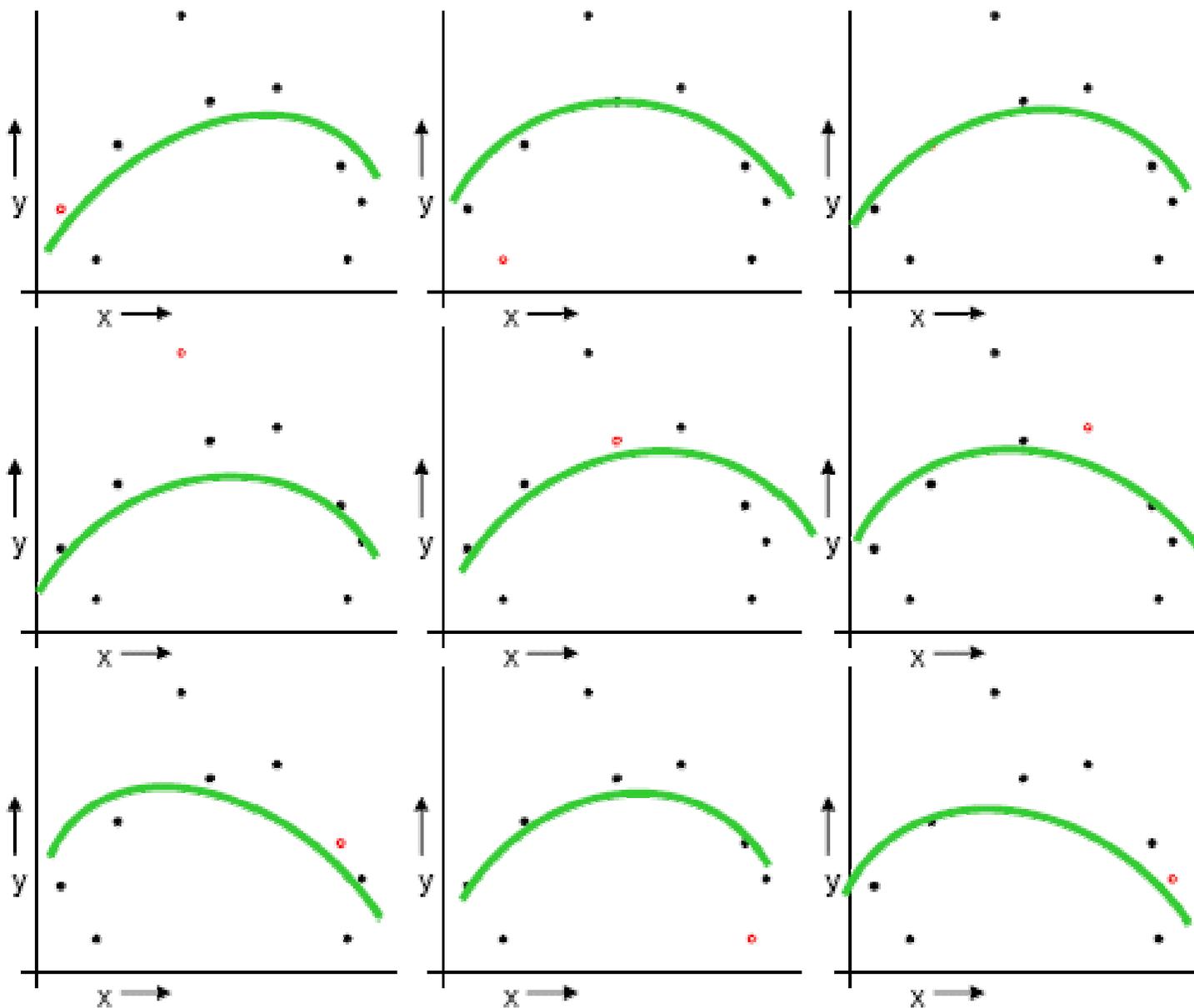
1. Select the **first**  $(x_i, y_i)$  data;
2. Temporary remove  $(x_i, y_i)$  from the data set;
3. Train on the remaining  $n-1$  datapoints;
4. Note your error  $(x_i, y_i)$ ;



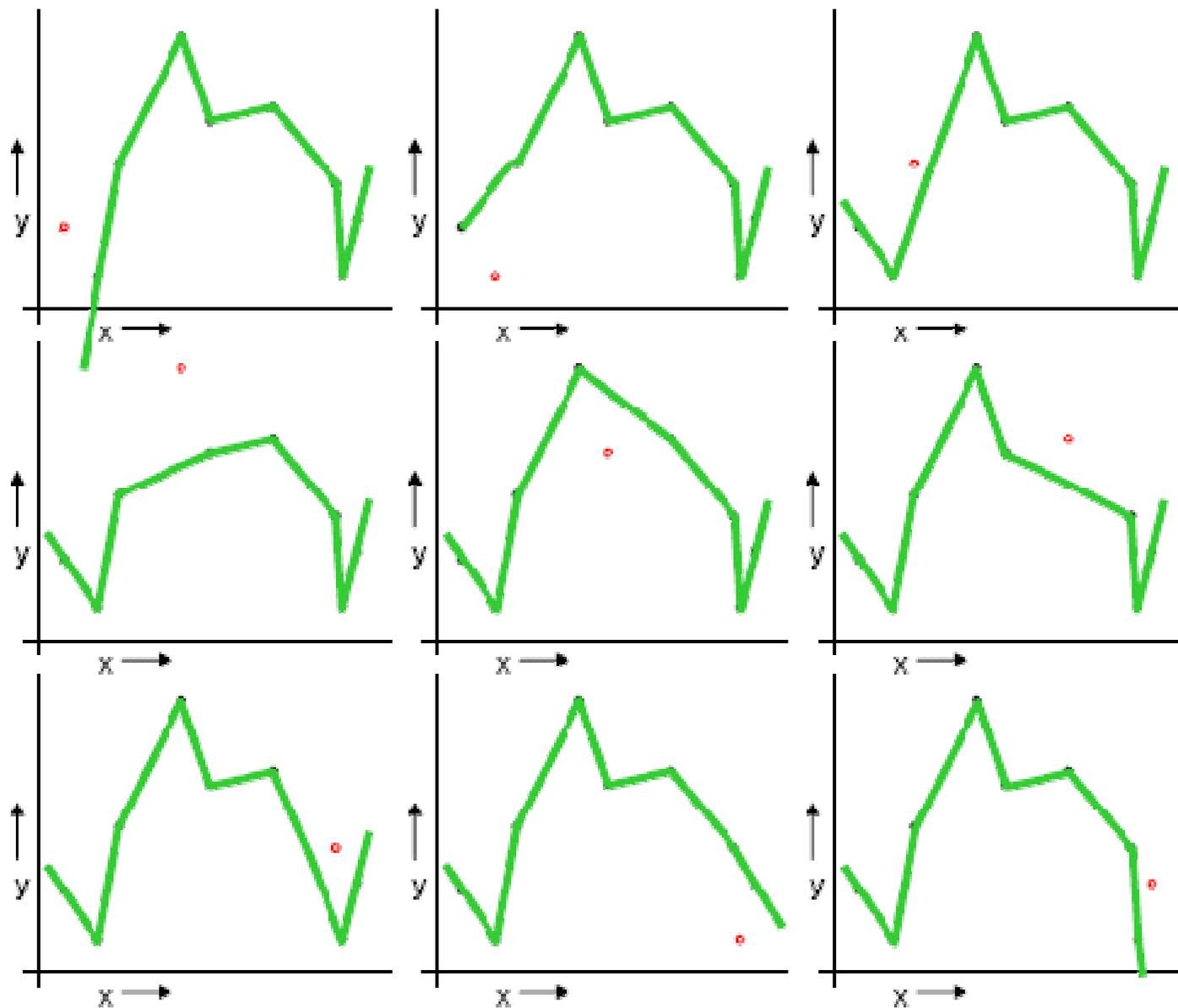
When you've done all points, report the mean squared errors (MSE).



$$\text{MSE}_{\text{Loccv}} = 3.33$$



$$\text{MSE}_{\text{Loocv}} = 0.96$$



$$\text{MSE}_{\text{Loccv}} = 2.12$$

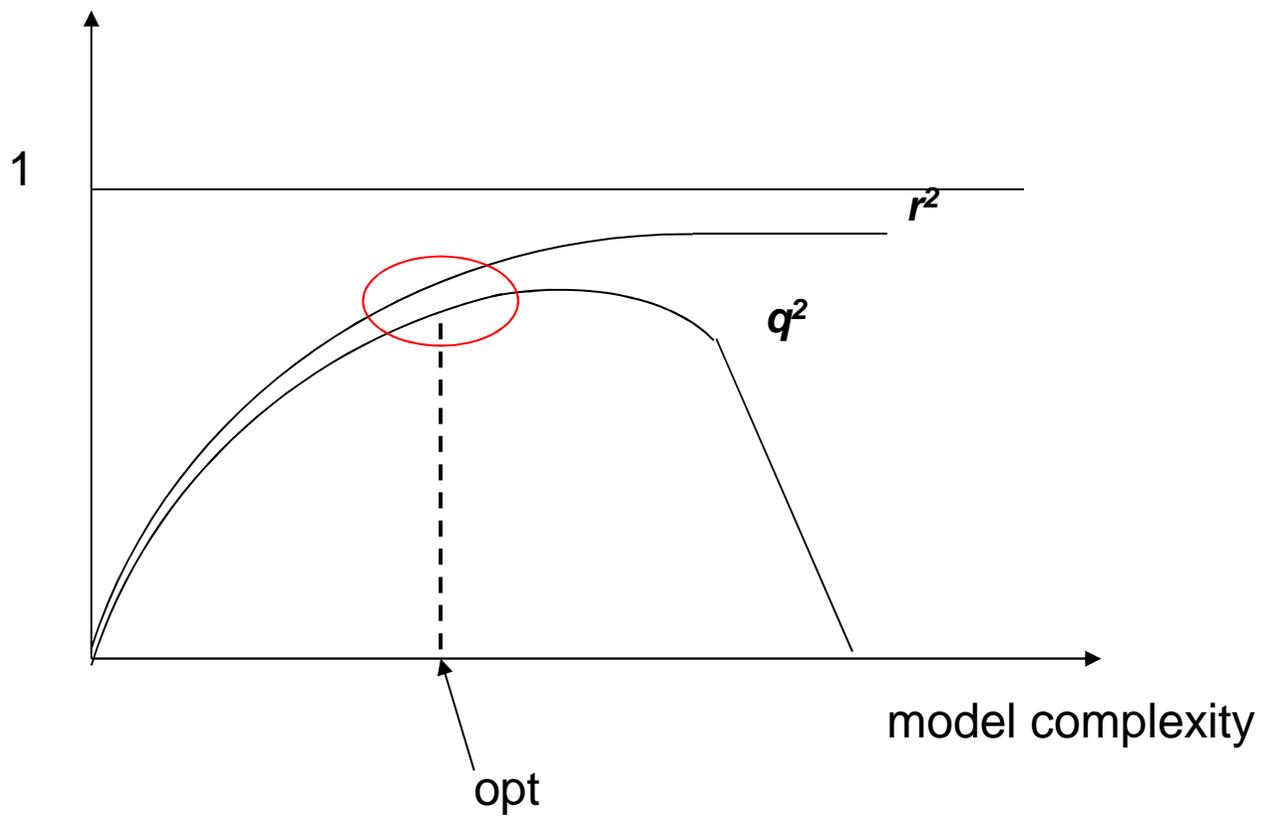


# Cross-validation coefficient ( $Q^2$ )

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

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

**$Q^2$  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.**





# So... which kind of validation?

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



## Another important consideration using MRA technique:

#	MR	logP	Volume	PM	Surface	density	n. X atoms
CH <sub>2</sub> Cl <sub>2</sub>	1,62959	1,30436	67,1359	84,933	176,1169	1,63677	3
CHCl <sub>3</sub>	2,01731	1,73808	75,7514	119,378	186,8237	2,03576	4
CCl <sub>4</sub>	2,35508	2,42116	83,2702	153,823	194,0565	2,3224	5
CF <sub>3</sub> CHBrCl	2,35642	2,36112	88,098	197,381	206,6438	2,85284	7
CHCl <sub>2</sub> CHCl	2,92829	2,49472	94,2376	167,85	215,3294	2,26061	6
Cl <sub>2</sub> C=CHCl	2,46705	2,28836	115,831	131,389	241,6985	1,42863	5
CCl <sub>2</sub> =CCl <sub>2</sub>	2,82835	3,37472	132,106	165,834	257,2367	1,46129	6

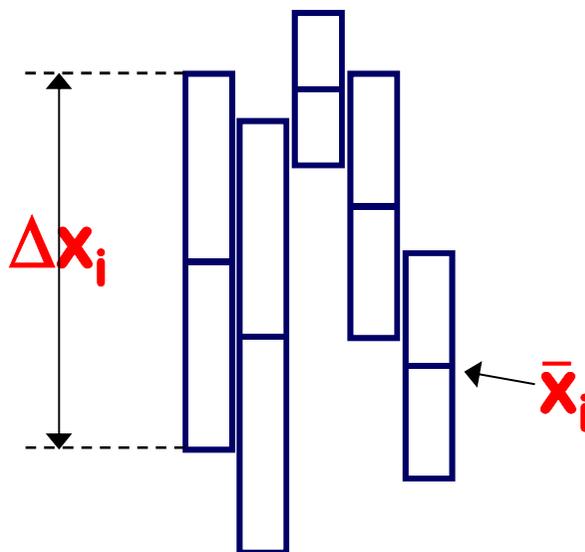
$$\bar{x}_j = \quad 2.37 \quad 2.28 \quad 93.77 \quad 145.80 \quad 211.13 \quad 1.99 \quad 5.10$$

$$\Delta x_j = \quad 1.30 \quad 2.07 \quad 44.00 \quad 112.45 \quad 81.12 \quad 1.42 \quad 4.00$$

## data scaling and data centering

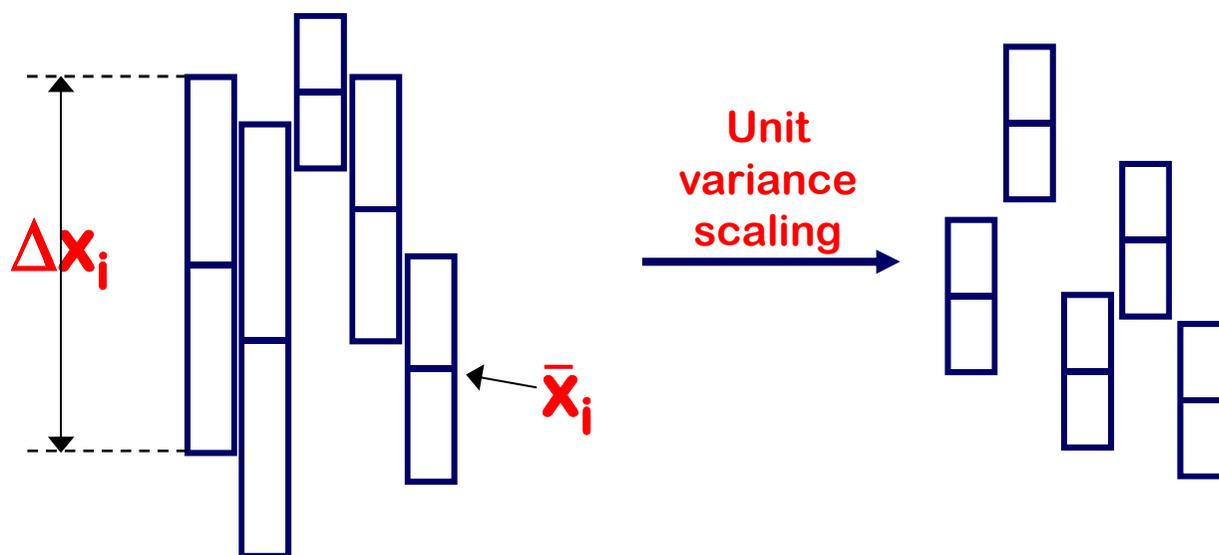
# data scaling and data centering

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

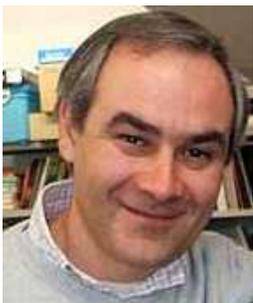


# data scaling and data centering

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



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



## Back to the real case:

#	MR	logP	Volume	PM	Surface	density	n. X atoms
CH <sub>2</sub> Cl <sub>2</sub>	1,62959	1,30436	67,1359	84,933	176,1169	1,63677	3
CHCl <sub>3</sub>	2,01731	1,73808	75,7514	119,378	186,8237	2,03576	4
CCl <sub>4</sub>	2,35508	2,42116	83,2702	153,823	194,0565	2,3224	5
CF <sub>3</sub> CHBrCl	2,35642	2,36112	88,098	197,381	206,6438	2,85284	7
CHCl <sub>2</sub> CHCl	2,92829	2,49472	94,2376	167,85	215,3294	2,26061	6
Cl <sub>2</sub> C=CHCl	2,46705	2,28836	115,831	131,389	241,6985	1,42863	5
CCl <sub>2</sub> =CCl <sub>2</sub>	2,82835	3,37472	132,106	165,834	257,2367	1,46129	6

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

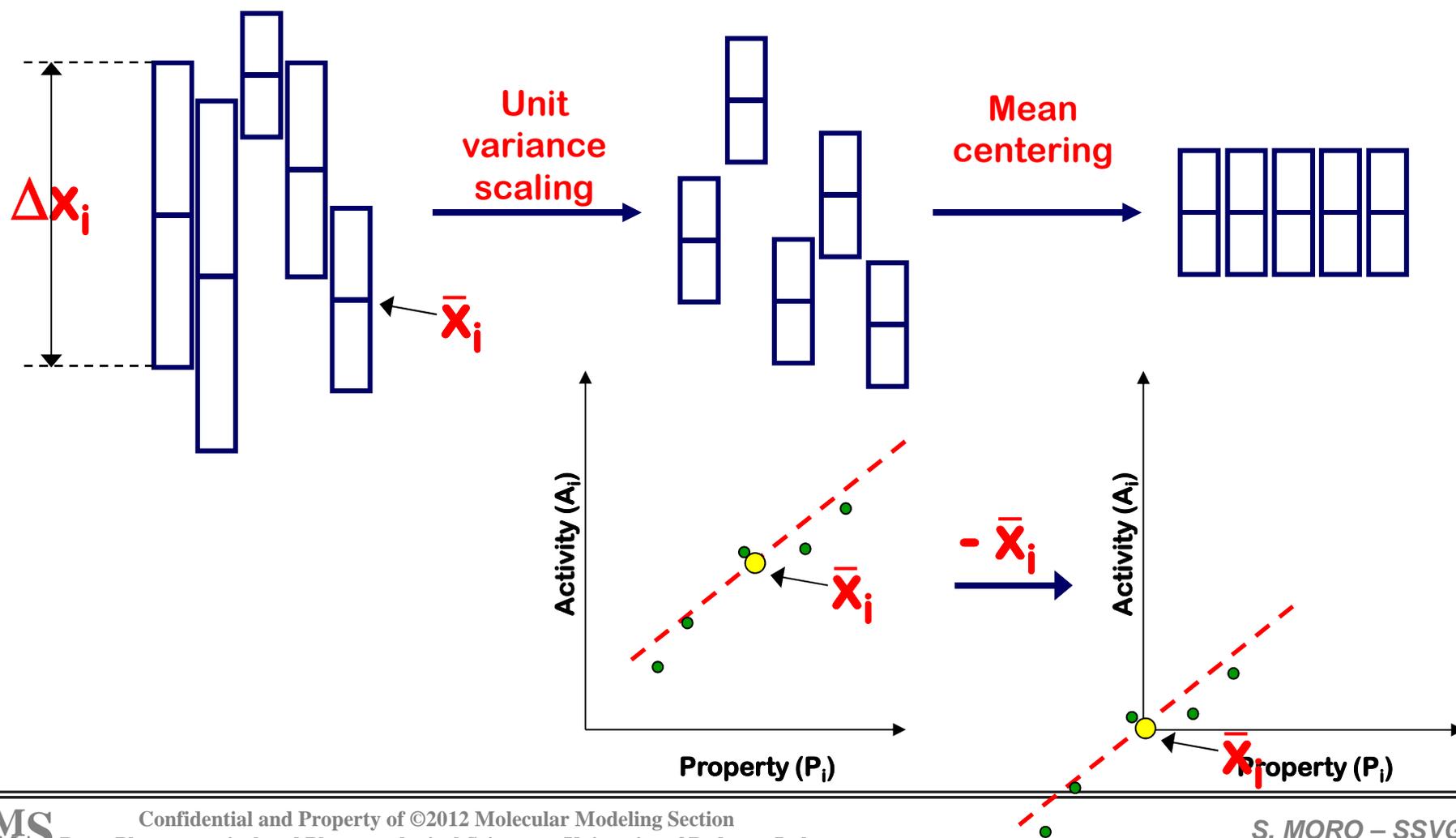
$$\Delta x_i = \quad \mathbf{1.30} \quad \mathbf{2.07} \quad \mathbf{44.00} \quad \mathbf{112.45} \quad \mathbf{81.12} \quad \mathbf{1.42} \quad \mathbf{4.00}$$

$$\sigma_i = \quad \mathbf{0.45} \quad \mathbf{0.65} \quad \mathbf{22.85} \quad \mathbf{37.02} \quad \mathbf{29.46} \quad \mathbf{0.52} \quad \mathbf{1.34}$$

$$\Delta x_i / \sigma_i = \quad \mathbf{2.89} \quad \mathbf{3.18} \quad \mathbf{1.92} \quad \mathbf{3.04} \quad \mathbf{2.75} \quad \mathbf{2.73} \quad \mathbf{2.98}$$

# 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.

