QSAR:

Dead or Alive?

Arthur M. Doweyko

Princeton, NJ USA
Bristol-Myers Squibb
arthur.doweyko@bms.com

JCAMD *22*, 81-89 **(2008)**

Modern Day QSAR

Proactive QSAR invites computed descriptors:

LogP …….. estimated using dozens of different methods

Surface Area …….. descriptors map polarity

3D-QSAR ……. maps interactions and occupancy

QM ……. methods map charges, other atom-based properties

Connectivity Indices …describe molecular and electronic architecture

Molecular Shapes and fragments

Correlograms, EVA, GRID ……

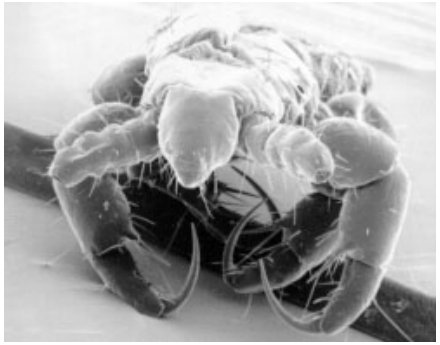
The plethora of descriptors is both a wonder and a bane !

The Correlation Problem: Spurious Correlations!

San Francisco Statistics:

The number of fire engines at each fire increases as do the damages at each fire.

Conclusion: Fire engines cause the damage!

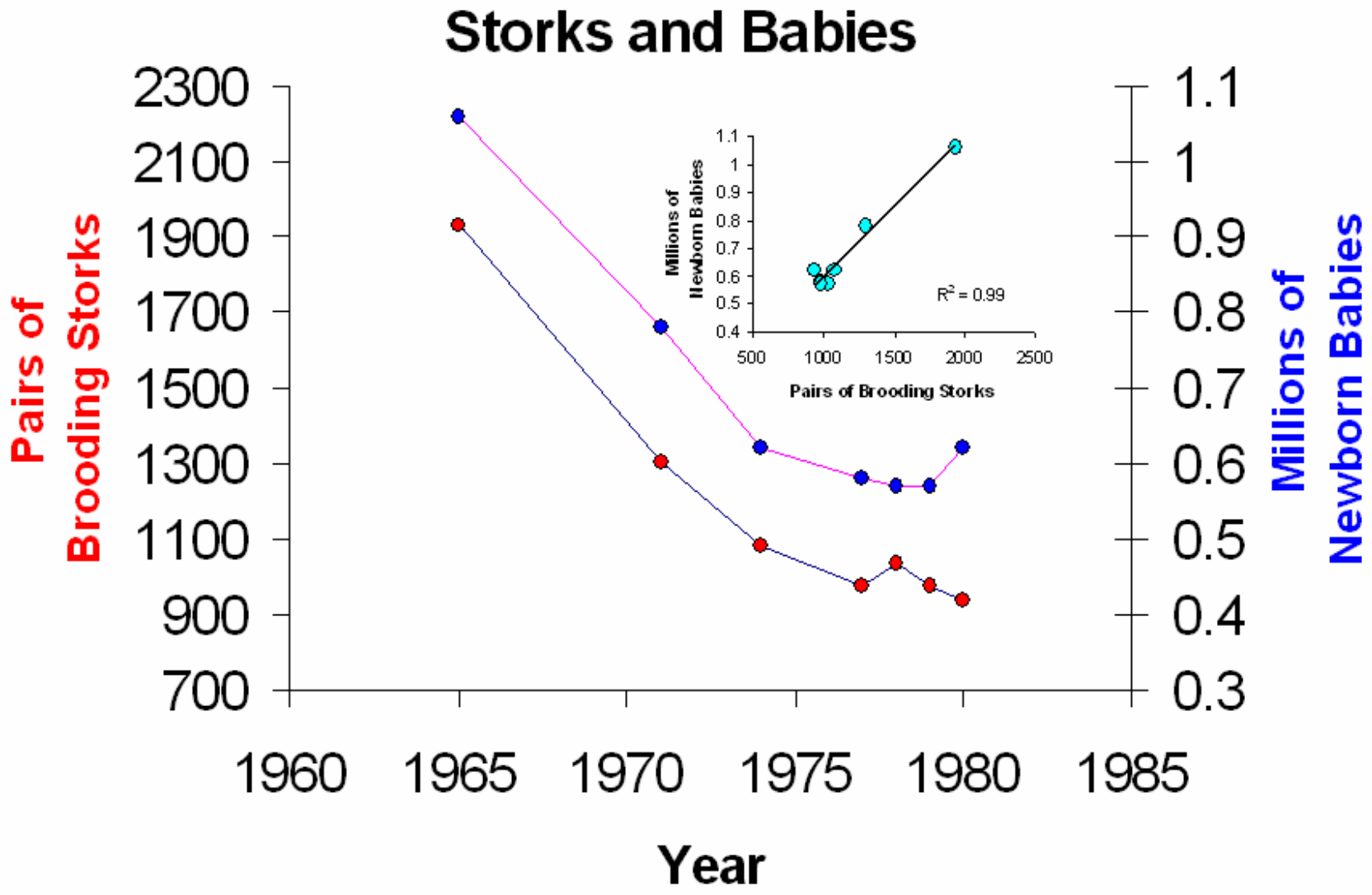


Several centuries of observations in the New Herbrides:

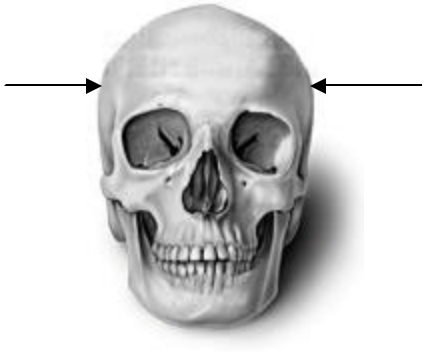
People who are healthy typically have lice.

People who are sick typically do not have lice.

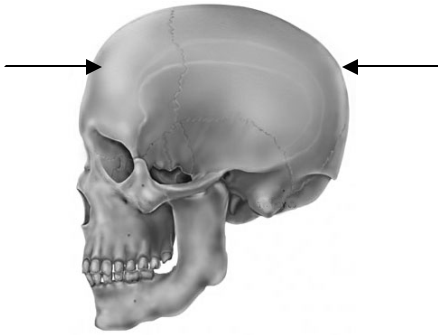
Conclusion: Lice make people healthy!



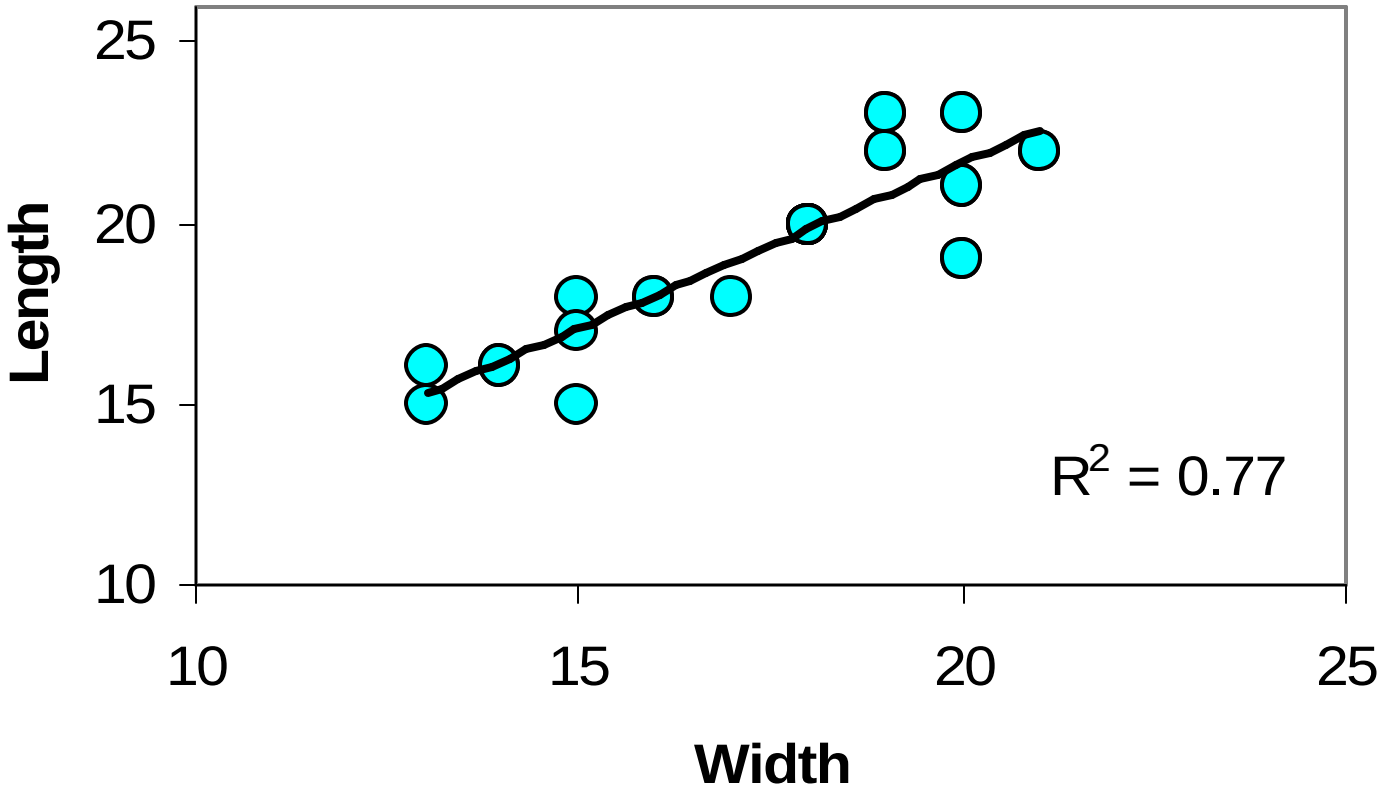
Sies, H. *Nature*, **332**, 495 (1988)



Watch the data!

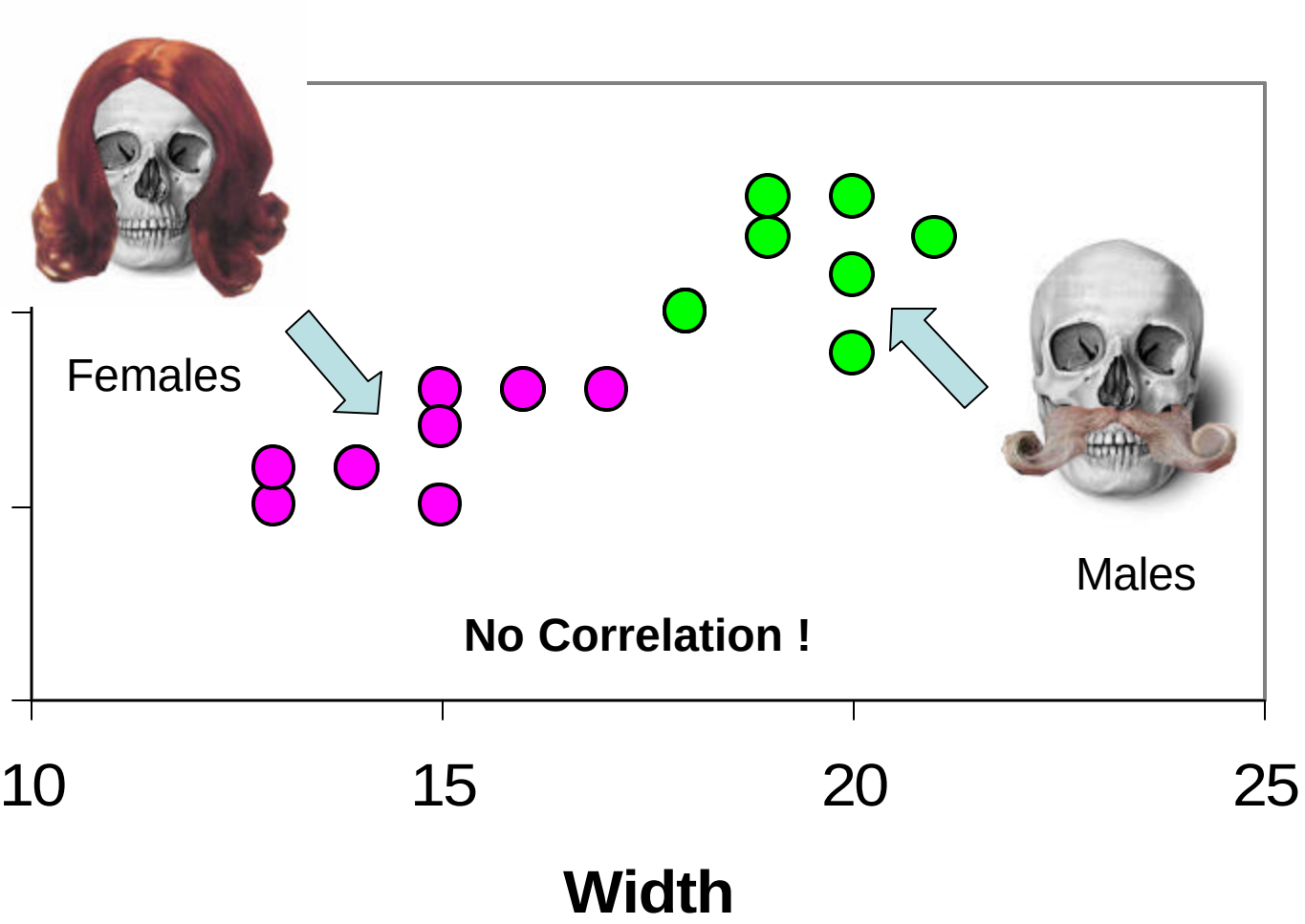


Skull Sizes

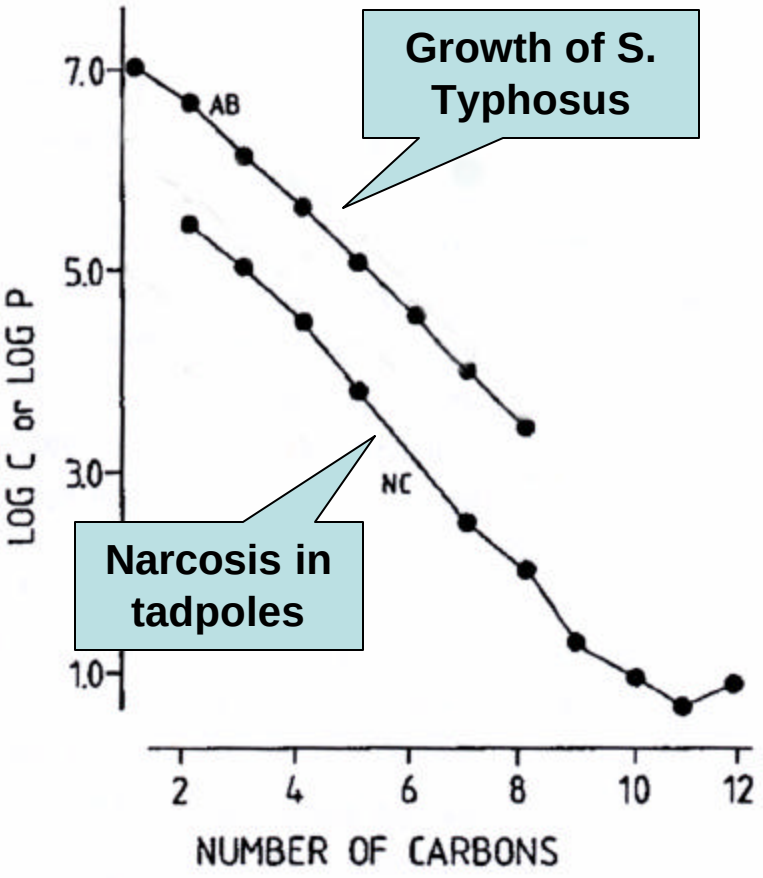


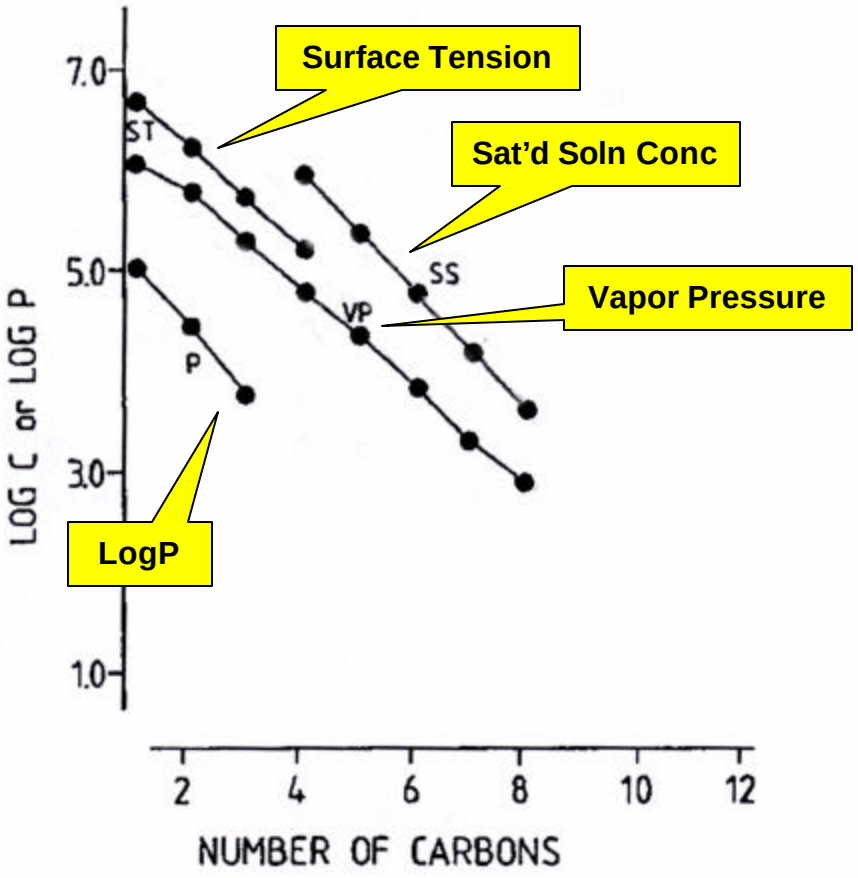
Stigler, S. *Perspectives in Biology and Medicine*, **48**, S88 (2005)

Skull Sizes


Lengthh
25
20
15
10
10
15
20
25
Width
Females
Males
No Correlation !

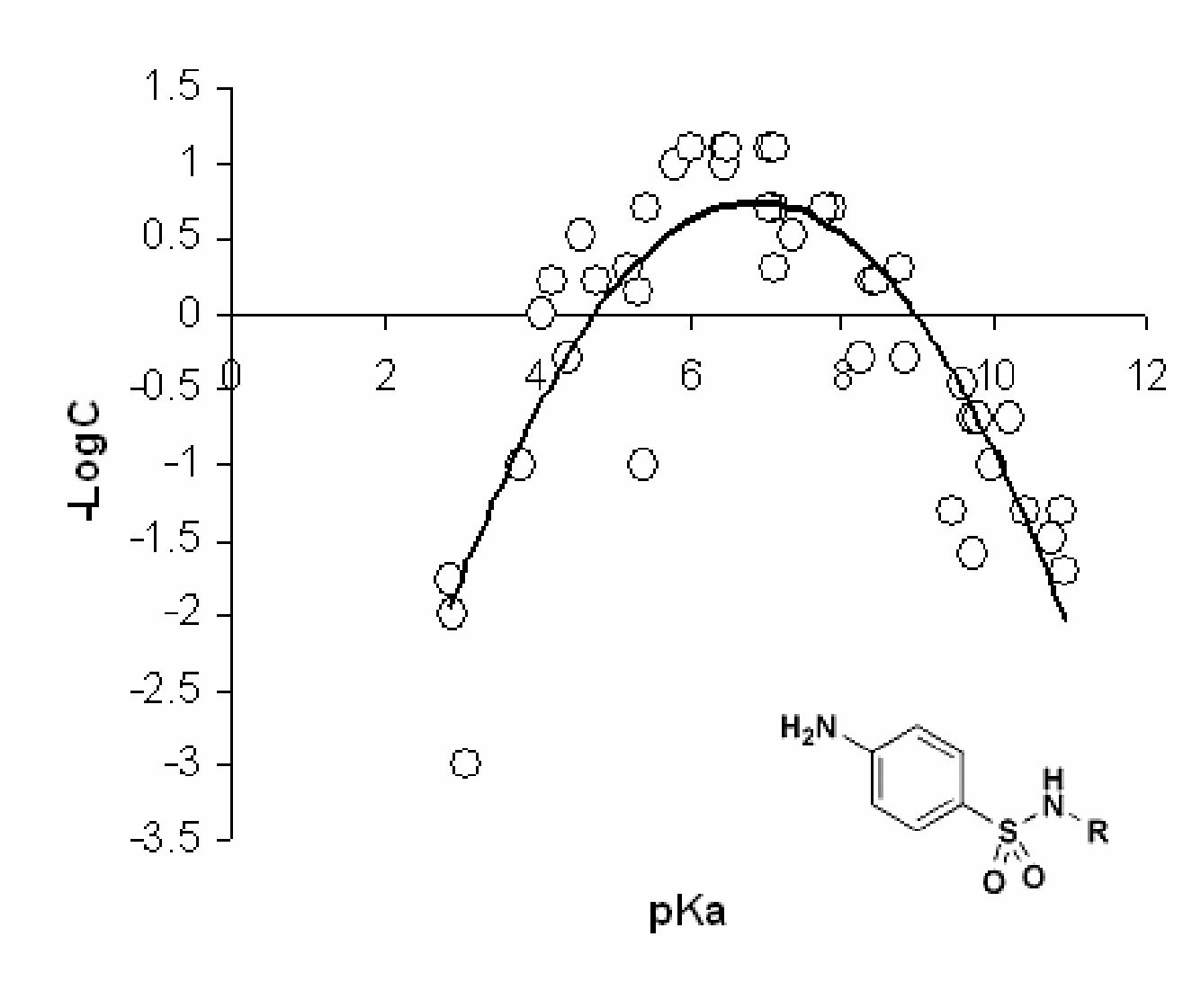
Effects of Alkyl Alcohols





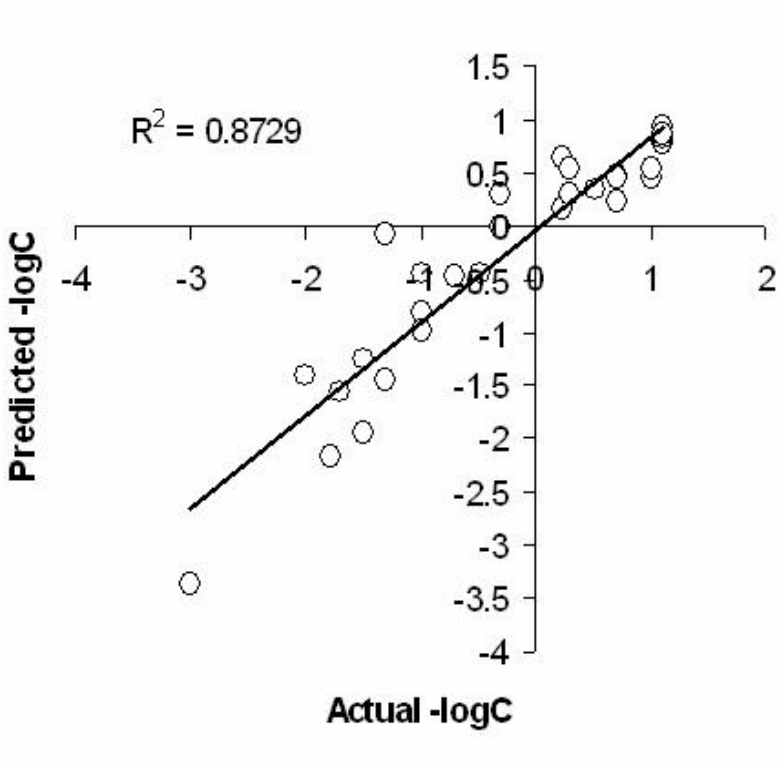
Barlow, R. B. *Trends in Pharmacological Sciences*, **1**, 124 (1979)

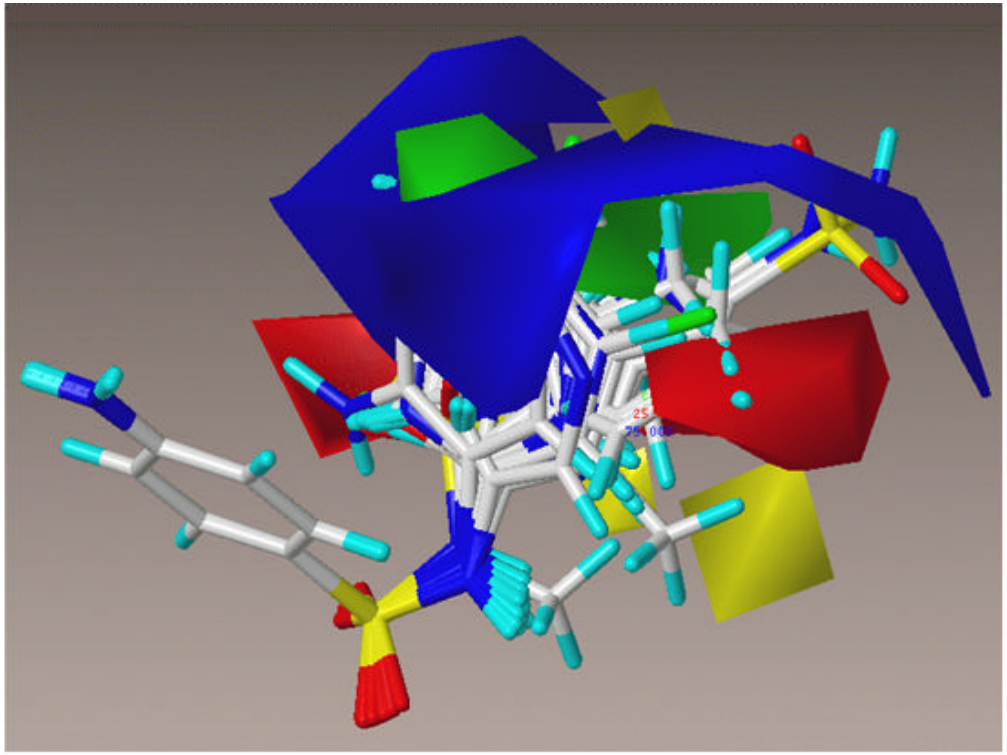
Sulfanilamides and pKa



Bell, P. H. and Roblin, Jr., R. O. *JACS*, **64**, 2905 (1942)

Sulfanilamides and Structure

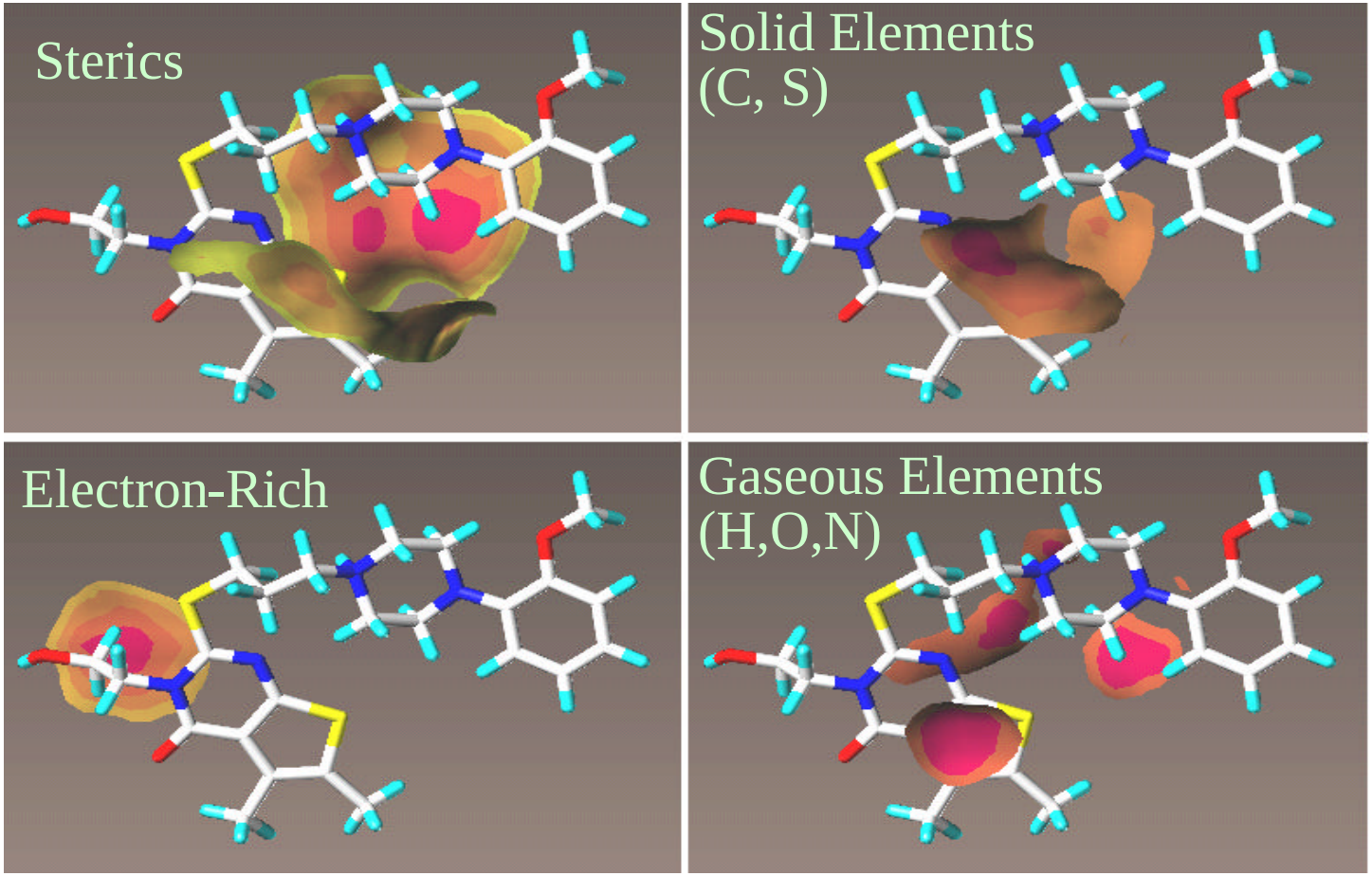




3D-QSARs Developed for 14 5HT1A Ligands

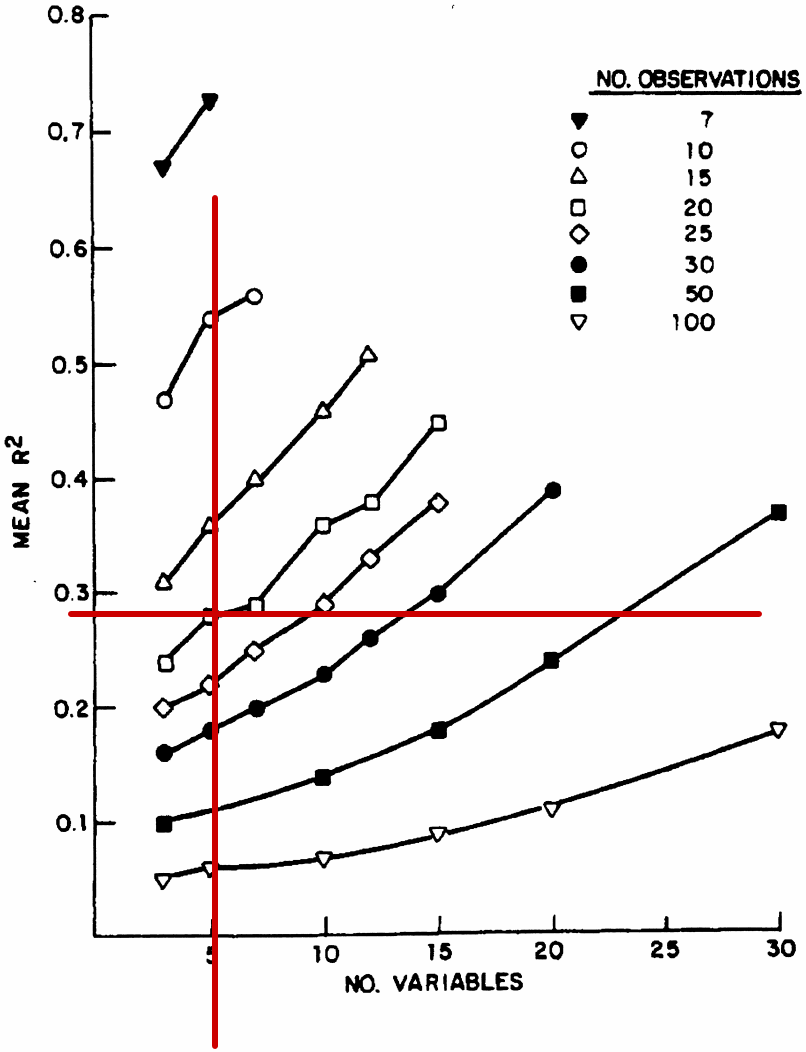
Normal Atom Descriptors
$q^2 = 0.83$

SLG Descriptors
$q^2 = 0.77$



Doweyko, A. *JCAMD*, **18**, 587 (2004)

Chance Correlations



What does this mean?

Starting with 5 possible independent variables…

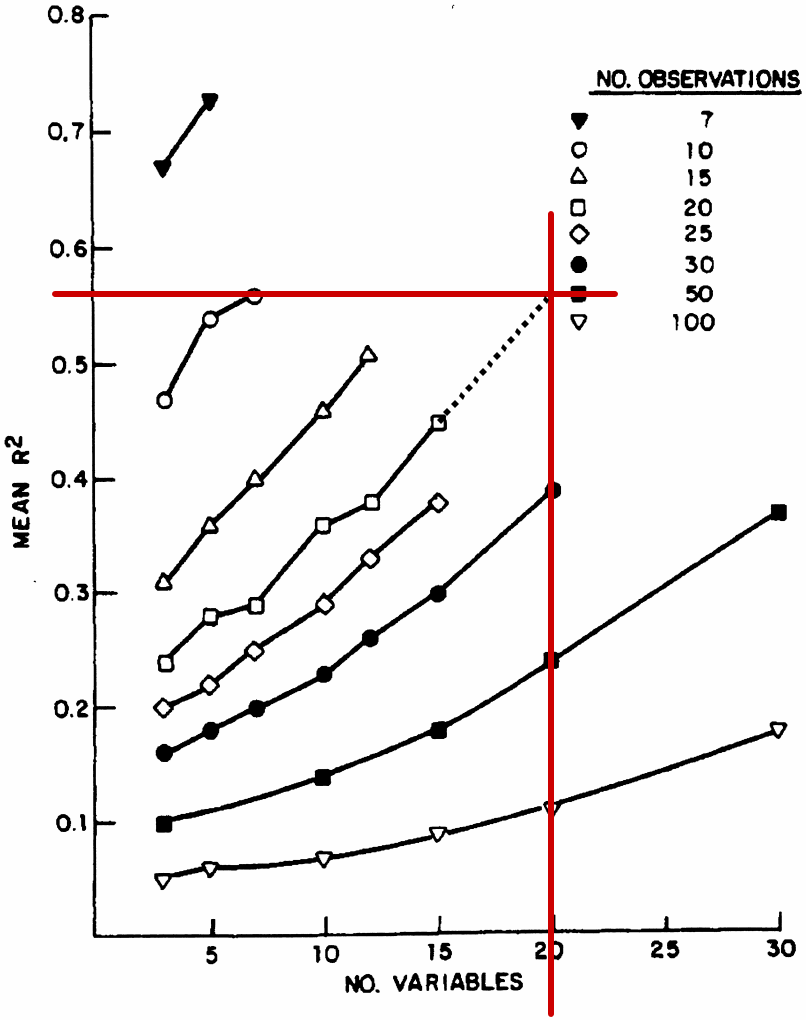
and 20 observations…

The average equation obtained by chance would exhibit a mean r^2 of 0.28 and contain 1.30 variables.

Topliss, J. and Edwards, R. *J. Med. Chem*. **22**, 1238 (1979)

See also Topliss, J and Costello, R. J. *J. Med. Chem*. **15**, 1066 (1972)

Chance Correlations



What does this mean?

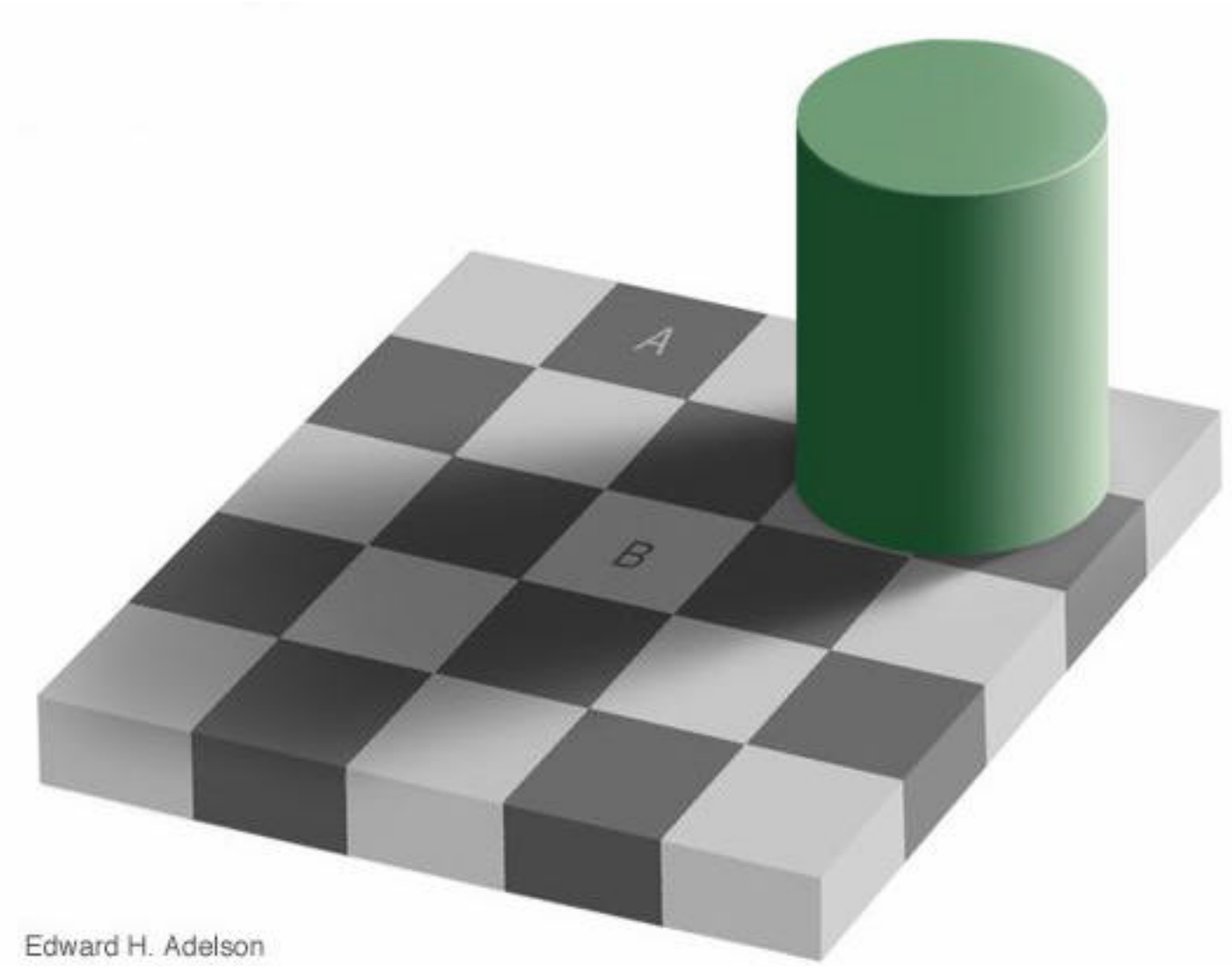
Starting with 20 possible independent variables…

and 20 observations…

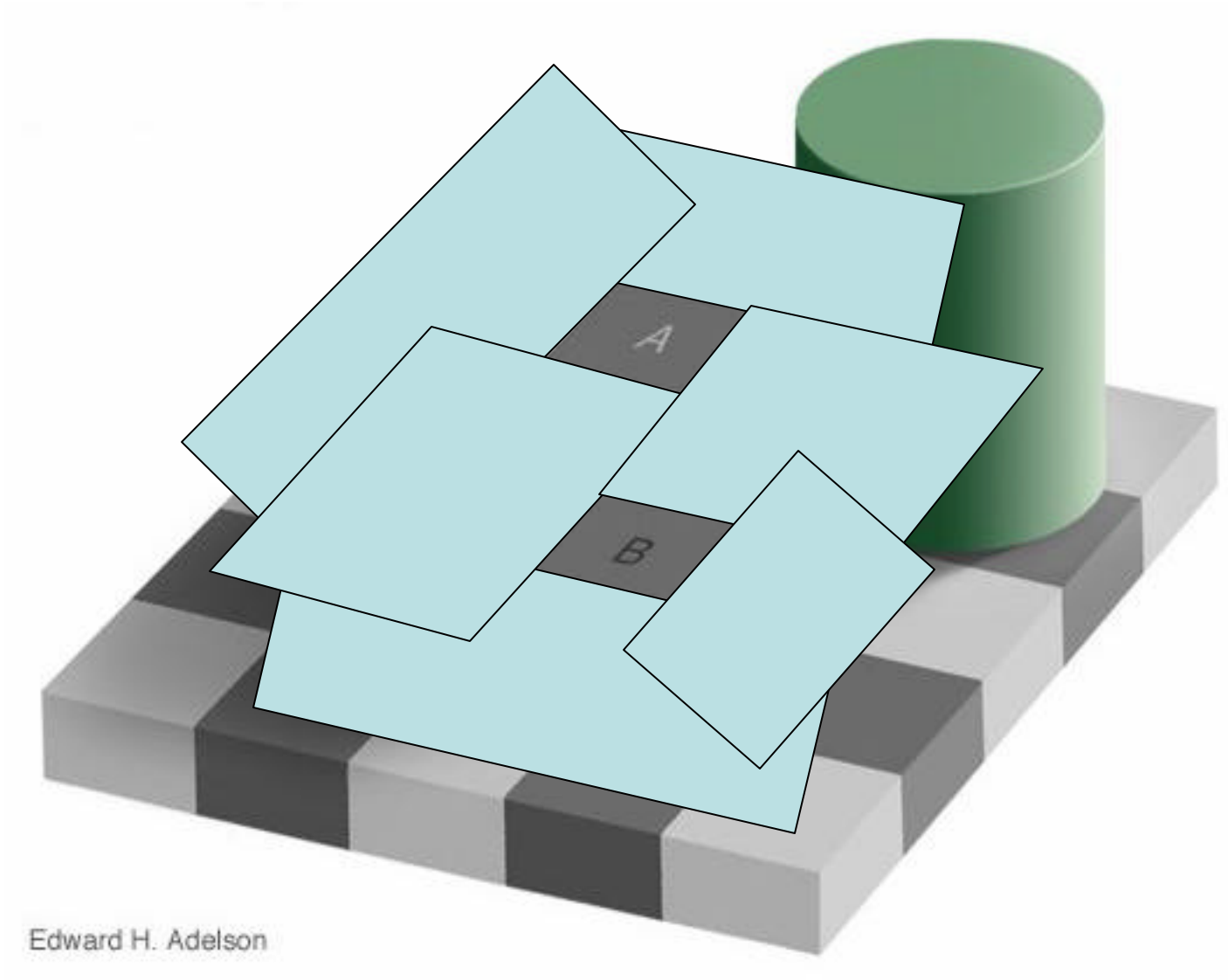
The average equation obtained by chance would exhibit a mean r^2 of 0.56 and contain 3.53 variables.

Topliss, J. and Edwards, R. *J. Med. Chem*. **22**, 1238 (1979)

See also Topliss, J and Costello, R. J. *J. Med. Chem*. **15**, 1066 (1972)



Edward H. Adelson

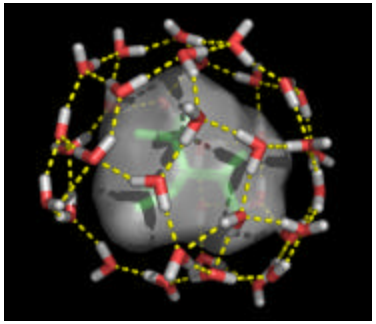


Edward H. Adelson

Illusory Correlations

The belief that hydrophobic molecules associate with one another in solution due to van der Waals forces (London dispersion forces).

While vdW interactions can account for 0.1-1 kcal/mole, the Hydrophobic Effect (-T?S) can account for much more:
e.g., buried –CH2- 0.8-1.6 kcal/mole



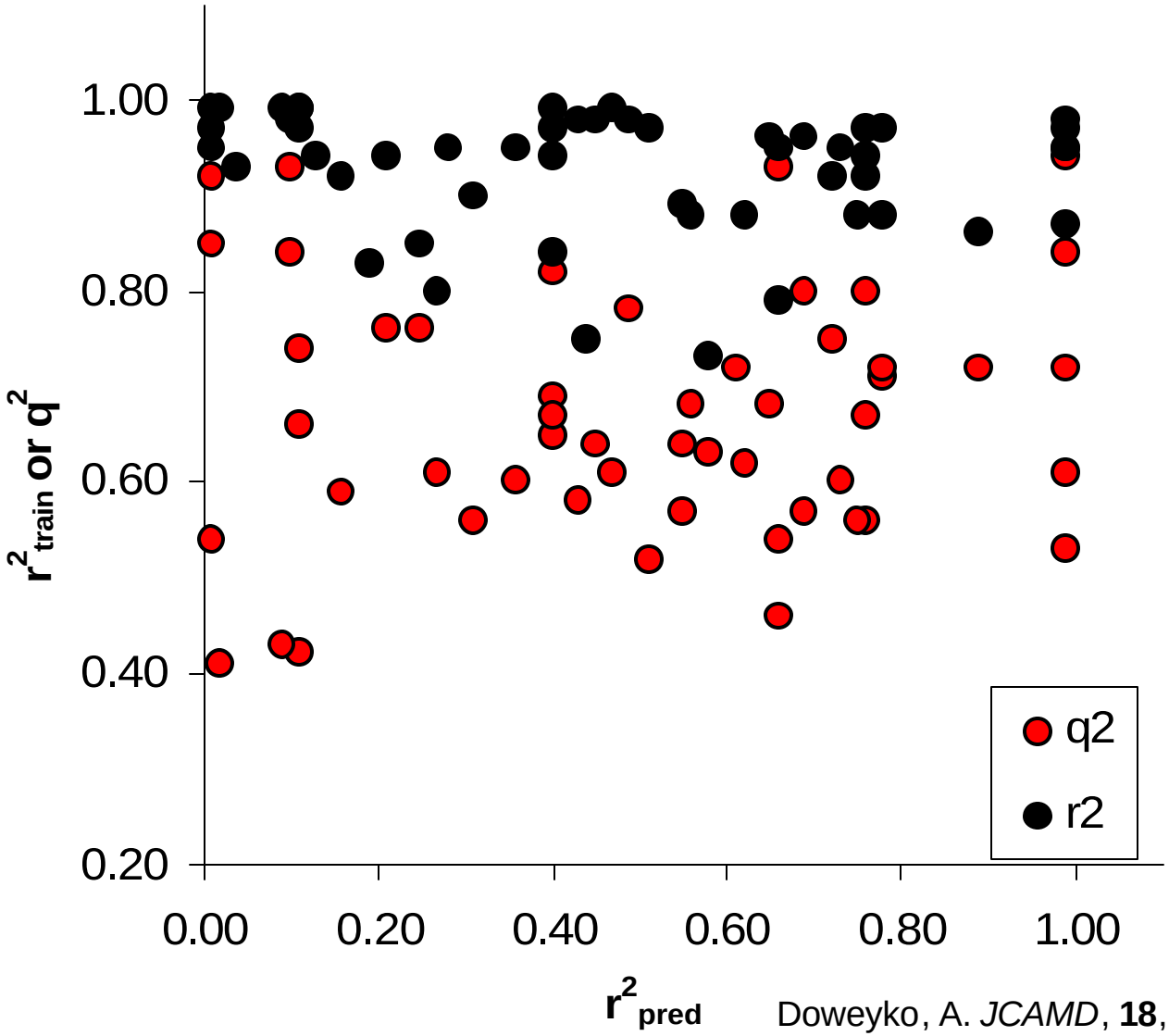
The belief that significant hydrogen bonding occurs between C-F and C-H.

Repeatedly shown to be an artifact of x-ray crystallography … simply reflecting crystal packing.

The belief that q^2 reflects the predictive ability of a QSAR model.

This one is very troubling!

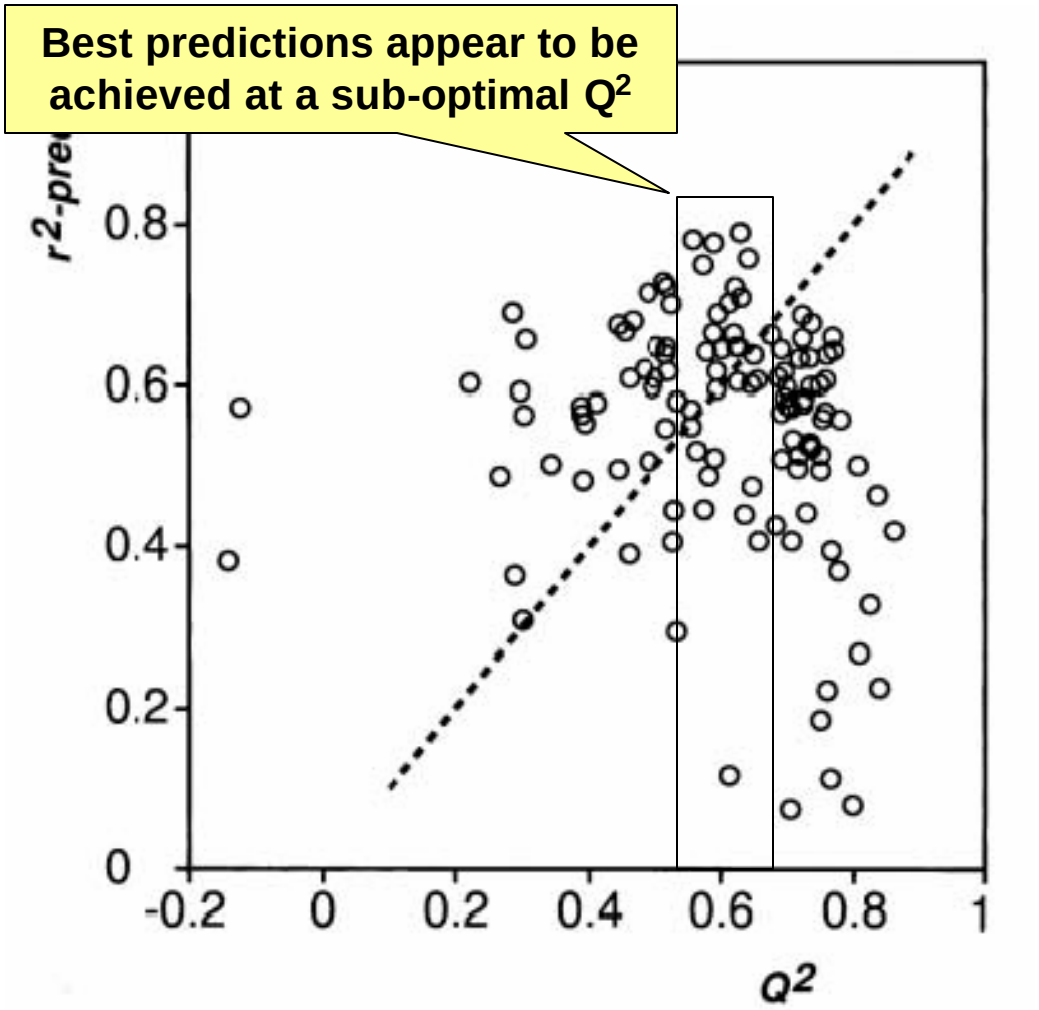
What's up with q^2?



q^2 and test set r^2_{pred} obtained from 37 3D-QSAR studies published in the past decade (60 models).

Doweyko, A. *JCAMD*, **18**, 587 (2004)
See also Tropsha, A. *J. Mol. Graph. Model*, **20**, 269 (2002)

The Kubinyi Paradox



Results of a similarity-based QSAR study aimed at developing models using the Cramer steroid set (1-21) for training, predicting steroids 22-31.

Six different alignment paradigms and four different property sets were used to generate the models.

Kubinyi, H., Hamprecht, F. A., Mietzner, T. *J. Med. Chem*. **41**, 2553 (1998)

Cramer, R. D. et al. *J. Am. Chem. Soc*., **110**, 5959 (1988)

The Overfitting Issue

Embedding measurement error into the correlation…

…adding noise to the resulting QSAR.

The Overfitting Issue

Embedding measurement error into the correlation…

…adding noise to the resulting QSAR.

Alas…An unavoidable problem.

The Overfitting Issue

Embedding measurement error into the correlation…

…adding noise to the resulting QSAR.

Alas…An unavoidable problem.

Exceeding the bounds dictated by the data…

…once again, adding noise to the resulting QSAR.

The Overfitting Issue

Embedding measurement error into the correlation…

…adding noise to the resulting QSAR.

Alas…An unavoidable problem.

Exceeding the bounds dictated by the data…

…once again, adding noise to the resulting QSAR.

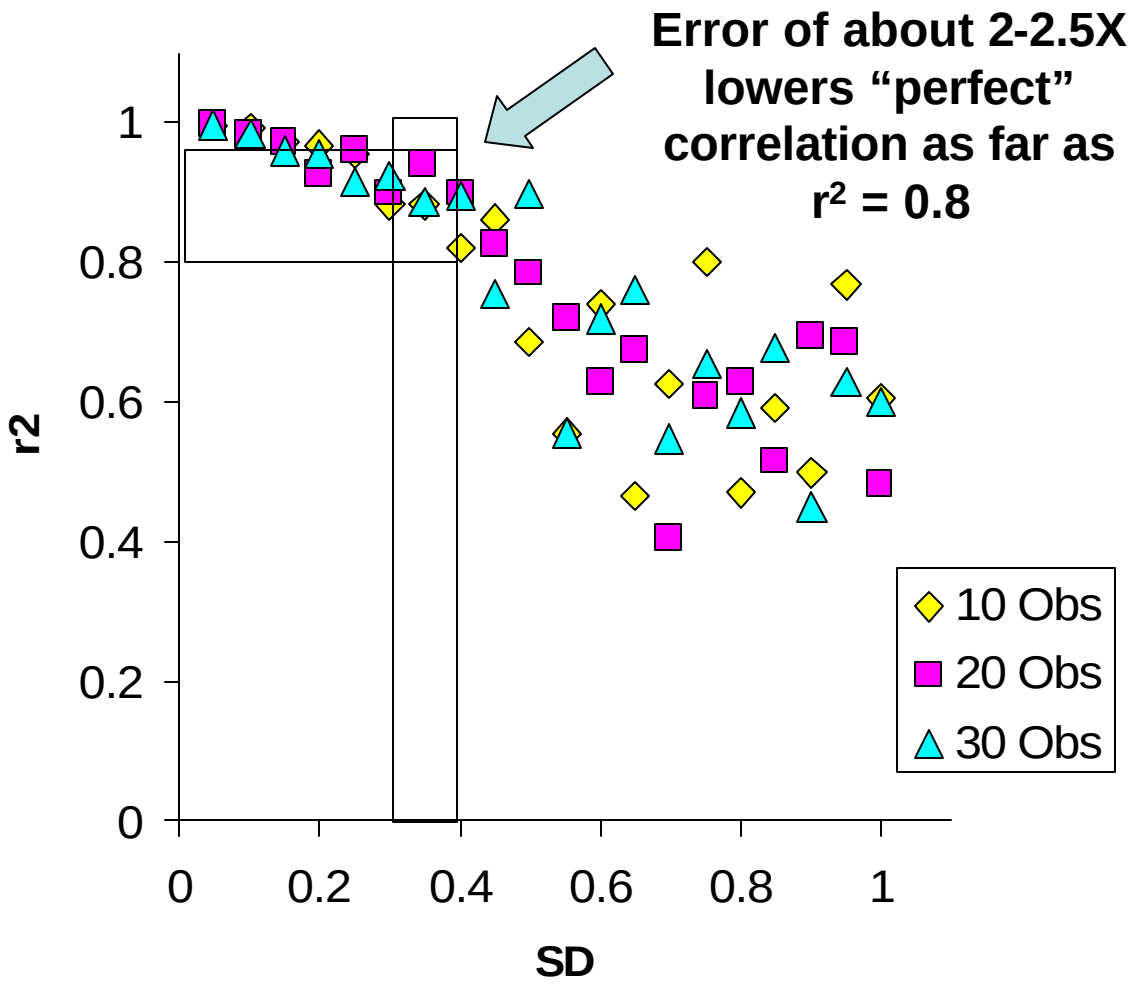
An avoidable problem!

The Effect of Observation Error

Example

Using 10 observations and SD = 0.20

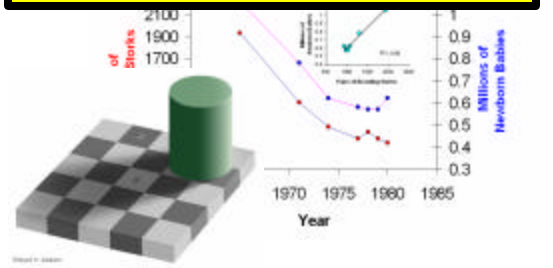
1	**1.00**	**0.73**
2	**1.50**	**1.66**
3	**2.00**	**2.17**
4	**2.50**	**2.41**
5	**2.75**	**2.66**
6	**3.00**	**3.33**
7	**3.20**	**3.02**
8	**3.50**	**3.27**
9	**3.80**	**3.58**
10	**4.00**	**4.23**

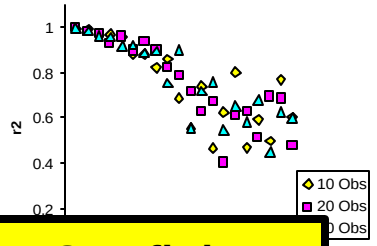


See also Doweyko, A. *et al*., *J. Med. Chem*., **26**, 475 (1983) and Doweyko, A. *JCAMD*, **18**, 587 (2004)

The QSAR Obstacle Course

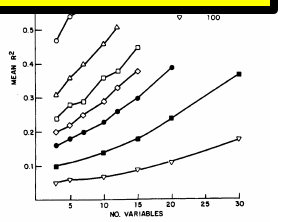
Illusory Spurious Correlations

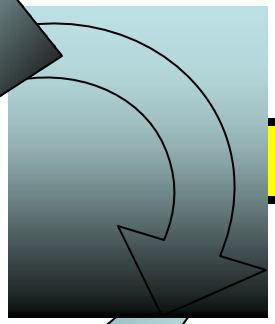


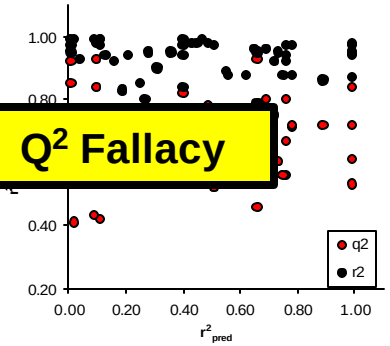


Overfitting

Chance Correlations







Q^2 Fallacy

Predictive QSAR based on possible causation

Meaningless and Uninterpretable Descriptors

QSAR remains valuable:

At the least, it provides a retrospective explanation for SAR

At the most, it provides synthetic guidance leading to testable hypotheses

---

QSPR is used extensively to predict molecular properties (solubility, vapor pressure, LogP…)

It plays a central role

In Experimental Design (effectively utilized to optimize process yields, formulations, product quality…)

As a database tool (similarity/diversity metrics)

In compound library design

In the parameterization of docking/scoring paradigms

As an estimator of binding affinity in LRM (Linear Response Method)

It’s Alive !