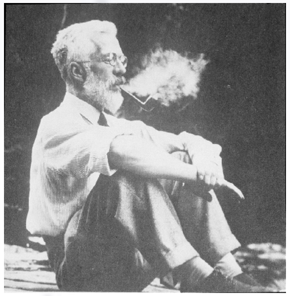
Applied Statistics

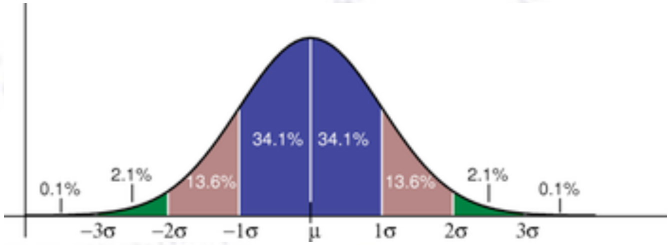
Correlations





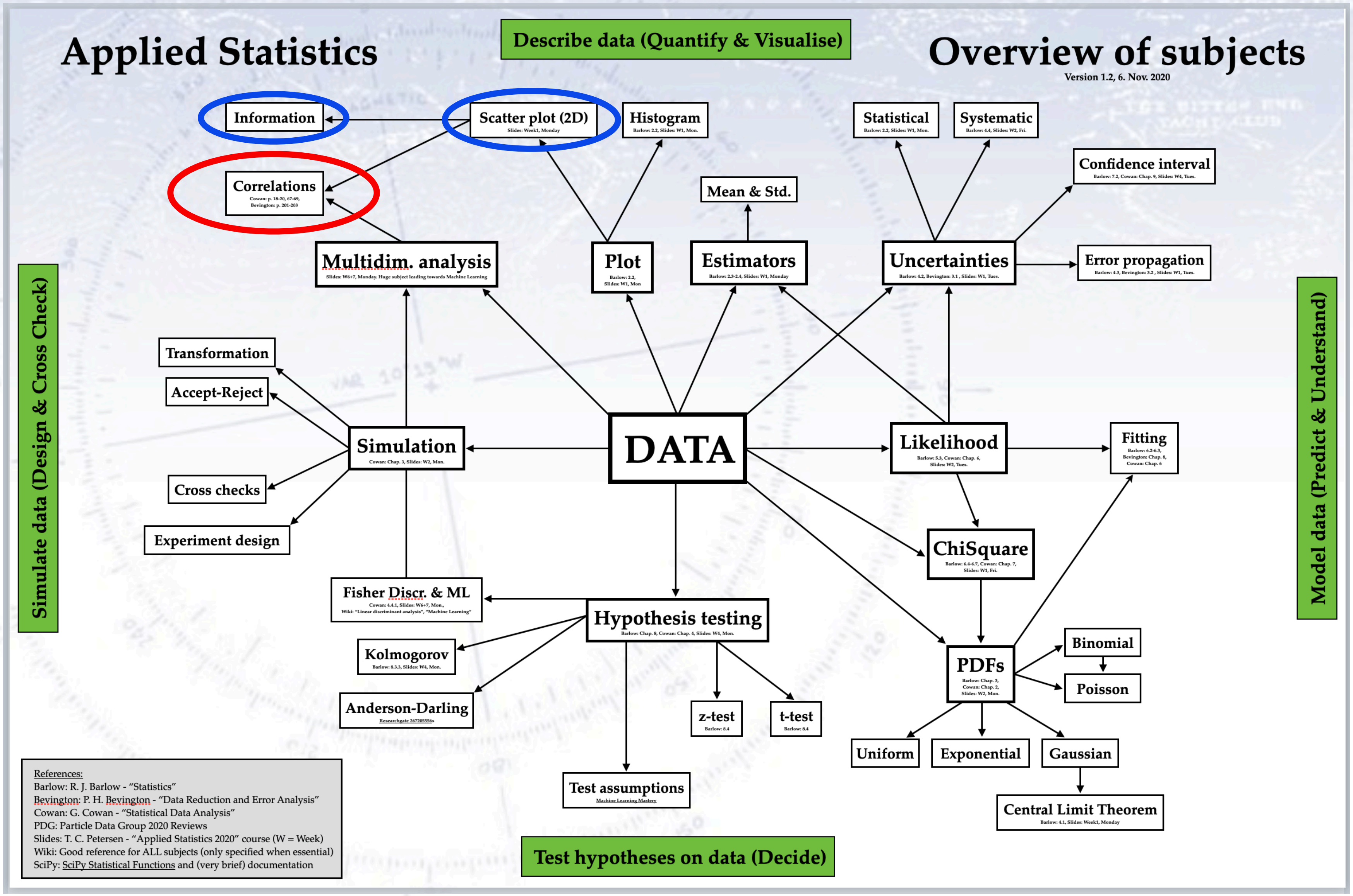



Troels C. Petersen (NBI)

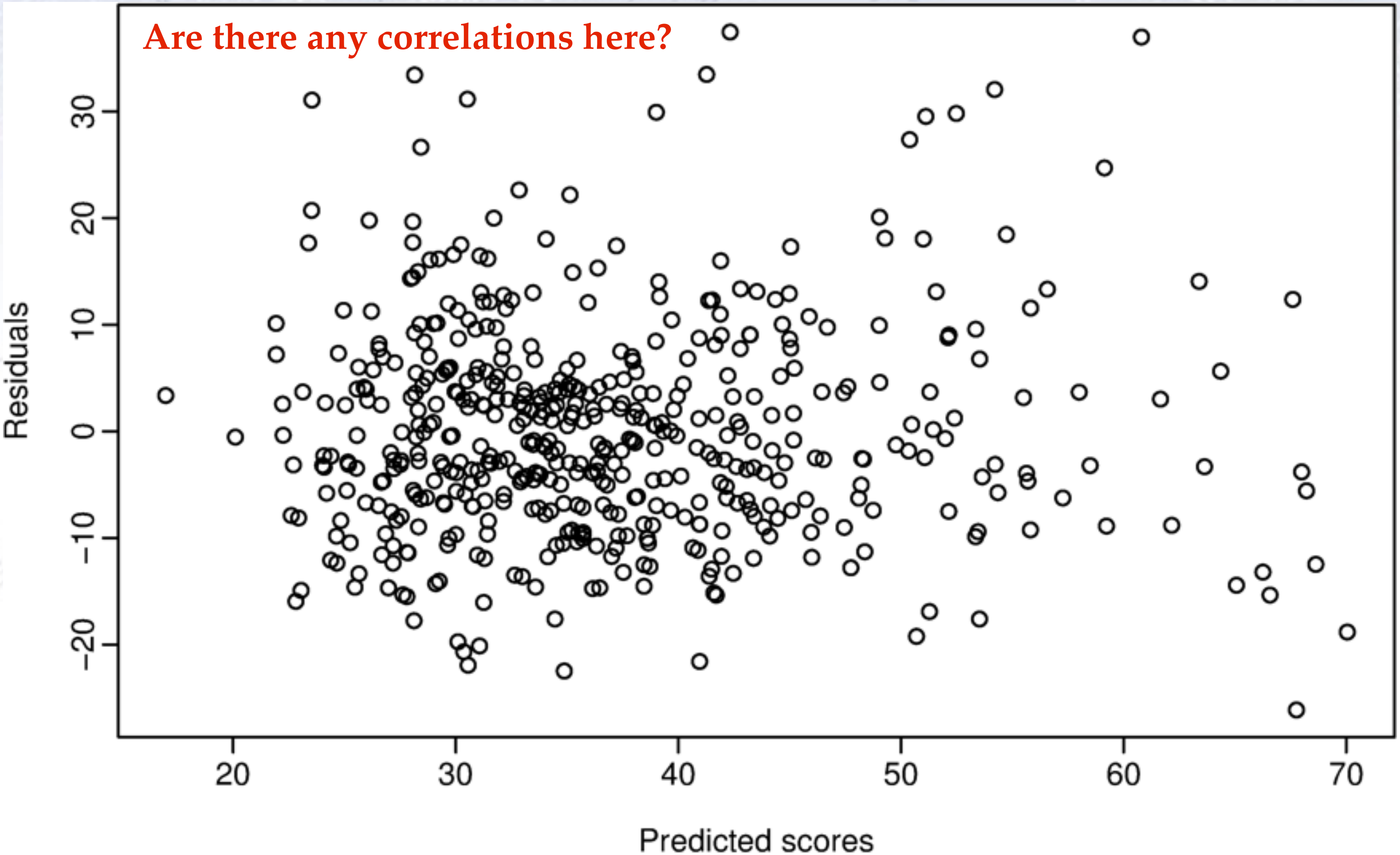


"Statistics is merely a quantisation of common sense"

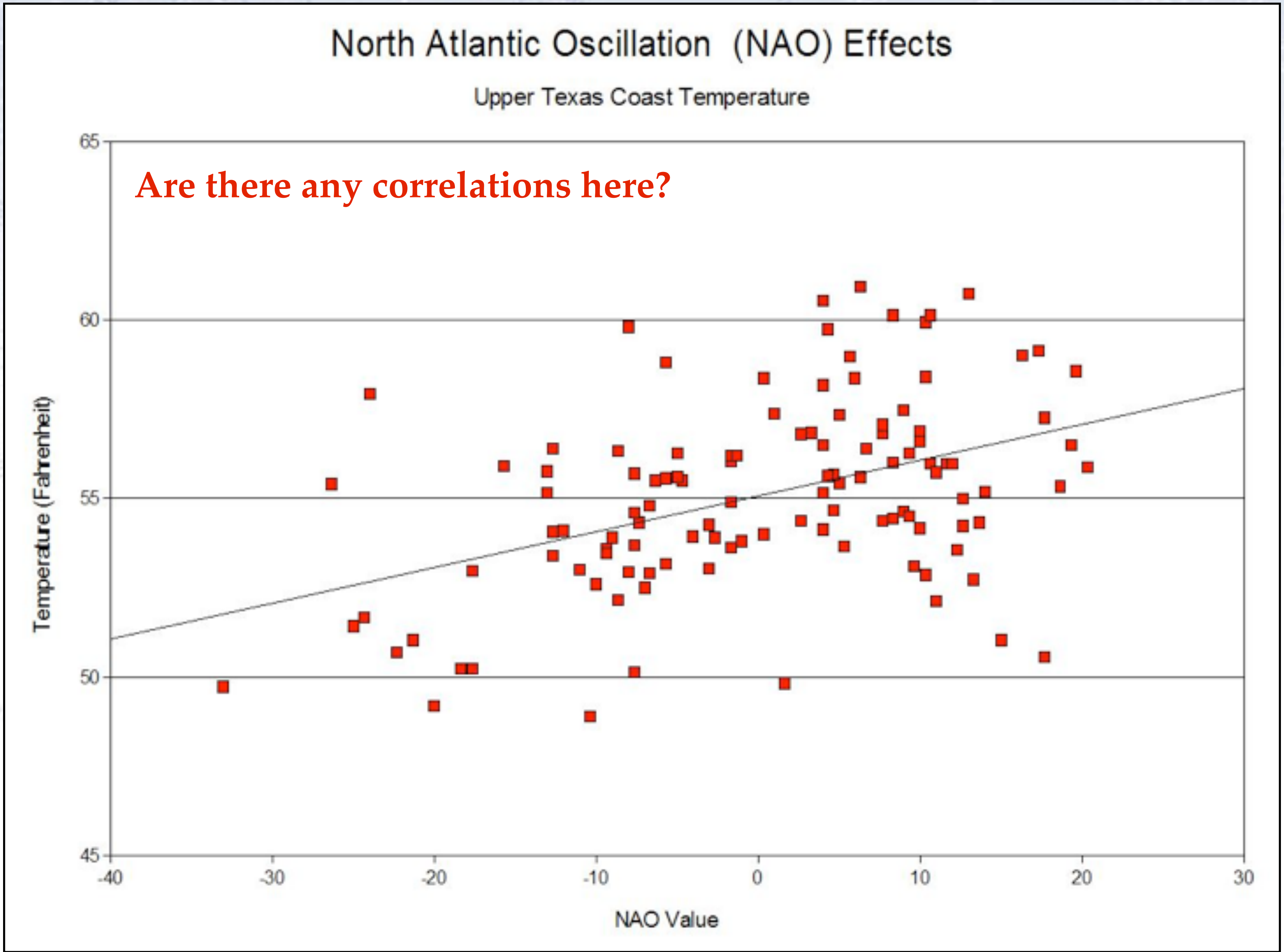
Correlations



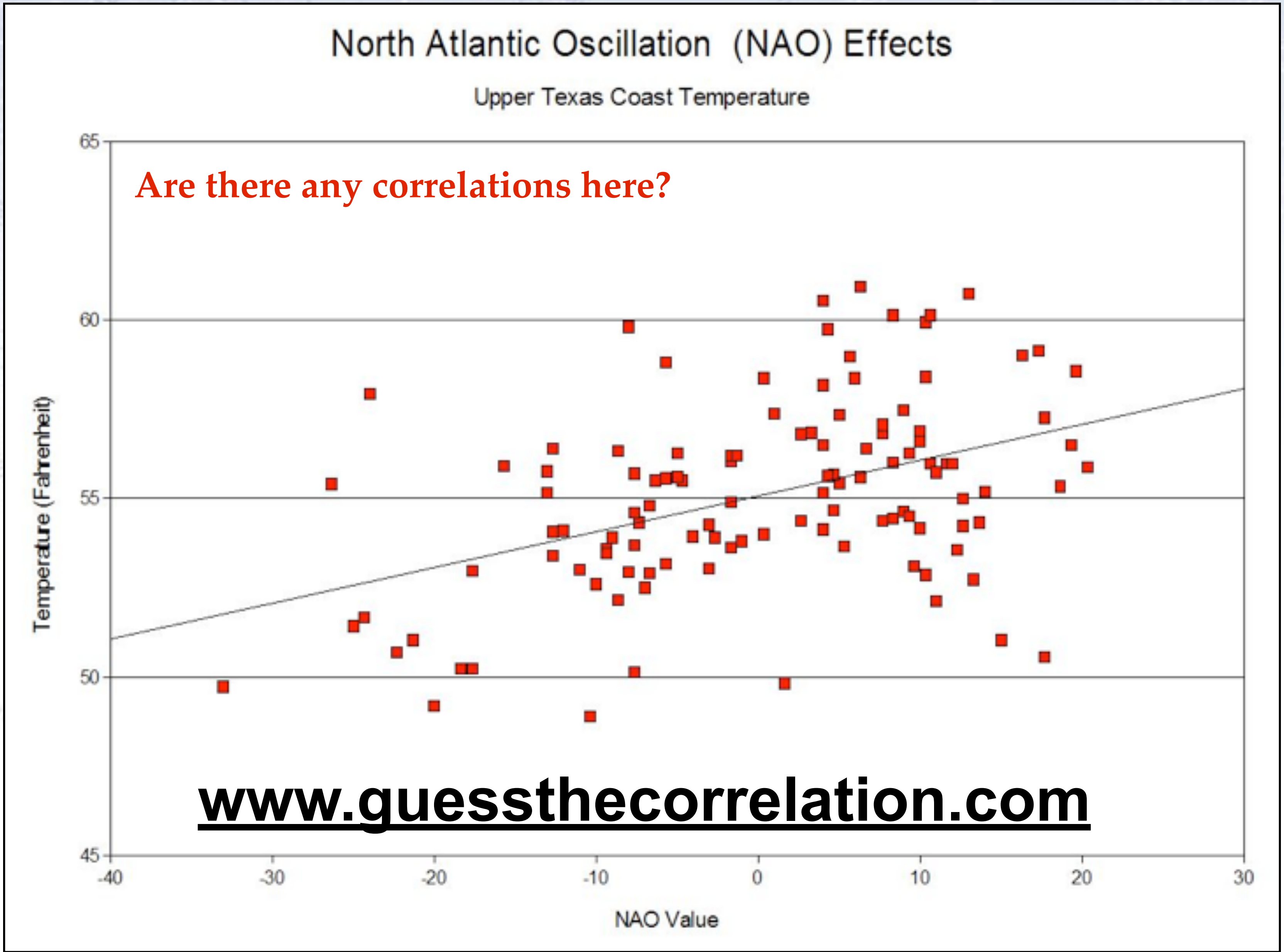
Correlation



Correlation



Correlation



www.guessthecorrelation.com

Correlation

Recall the definition of the Variance, V:

$$V = \sigma^2 = \frac{1}{N}\sum_{i}^{n}(x_i - \mu)^2 = E[(x-\mu)^2] = E[x^2] - \mu^2$$

Correlation

Recall the definition of the Variance, V:

$$V = \sigma^2 = \frac{1}{N}\sum_{i}^{n}(x_i - \mu)^2 = E[(x-\mu)^2] = E[x^2] - \mu^2$$

Likewise, one defines the **Covariance, $\mathbf{V_{xy}}$:**

$$V_{xy} = \frac{1}{N}\sum_{i}^{n}(x_i - \mu_x)(y_i - \mu_y) = E[(x_i - \mu_x)(y_i - \mu_y)]$$

Correlation

Recall the definition of the Variance, V:

$$V = \sigma^2 = \frac{1}{N}\sum_i^n (x_i - \mu)^2 = E[(x-\mu)^2] = E[x^2] - \mu^2$$

Likewise, one defines the **Covariance, V_{xy}:**

$$V_{xy} = \frac{1}{N}\sum_i^n (x_i - \mu_x)(y_i - \mu_y) = E[(x_i - \mu_x)(y_i - \mu_y)]$$

"Normalising" by the widths, gives Pearson's (linear) correlation coefficient:

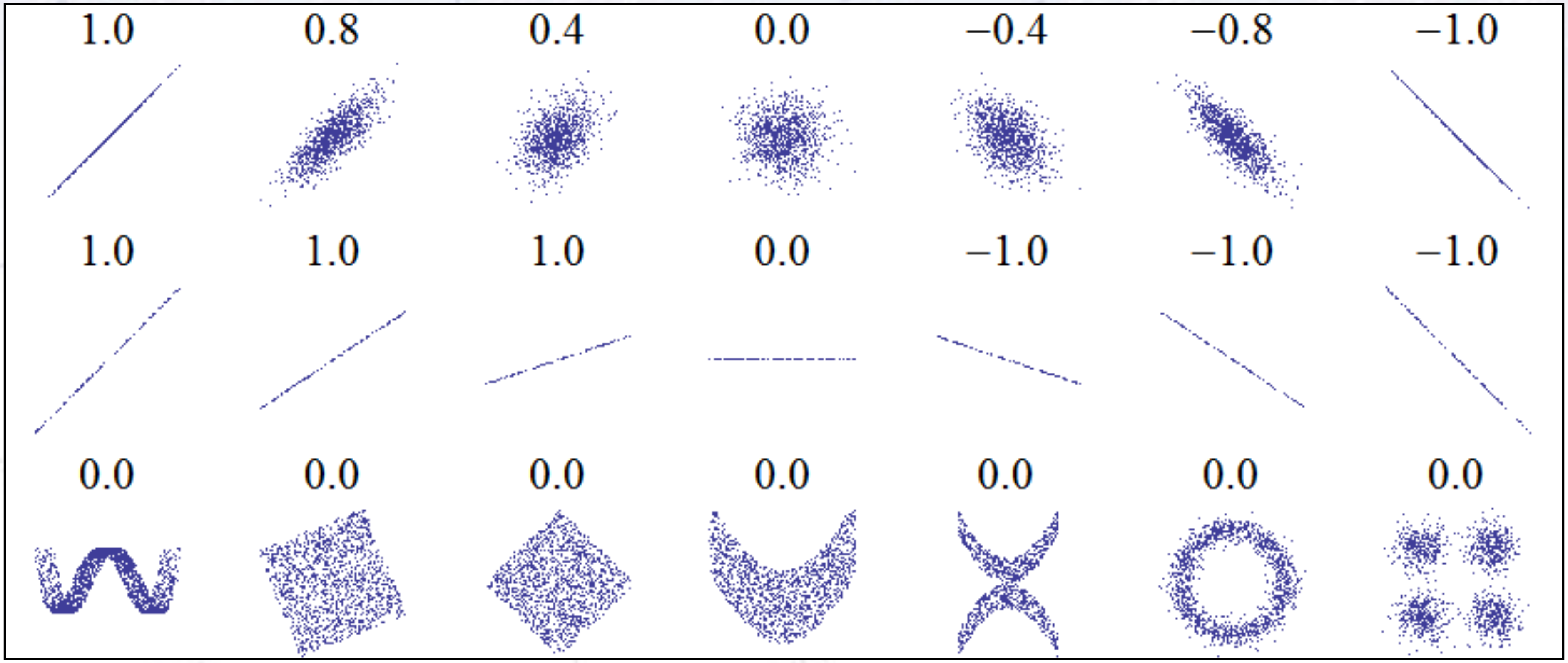
$$\rho_{xy} = \frac{V_{xy}}{\sigma_x \sigma_y}$$

$$-1 < \rho_{xy} < 1$$

$$\sigma(\rho) \simeq \sqrt{\frac{1}{n}(1-\rho^2)^2 + O(n^{-2})}$$

Correlation

Correlations in 2D are in the Gaussian case the "degree of ovalness"!



Note how ALL of the bottom distributions have $\varrho = 0$, despite obvious correlations!

Correlation Matrix

The correlation matrix V_{xy} explicitly looks as:

$$V_{xy} = \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 & \cdots & \sigma_{1N}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2N}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_N^2 & \sigma_{N2}^2 & \cdots & \sigma_{NN}^2 \end{bmatrix}$$

Very specifically, the calculations behind are:

$$V = \begin{bmatrix} \mathrm{E}[(X_1-\mu_1)(X_1-\mu_1)] & \mathrm{E}[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_1-\mu_1)(X_n-\mu_n)] \\ \mathrm{E}[(X_2-\mu_2)(X_1-\mu_1)] & \mathrm{E}[(X_2-\mu_2)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_2-\mu_2)(X_n-\mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n-\mu_n)(X_1-\mu_1)] & \mathrm{E}[(X_n-\mu_n)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_n-\mu_n)(X_n-\mu_n)] \end{bmatrix}.$$

Correlation and Information

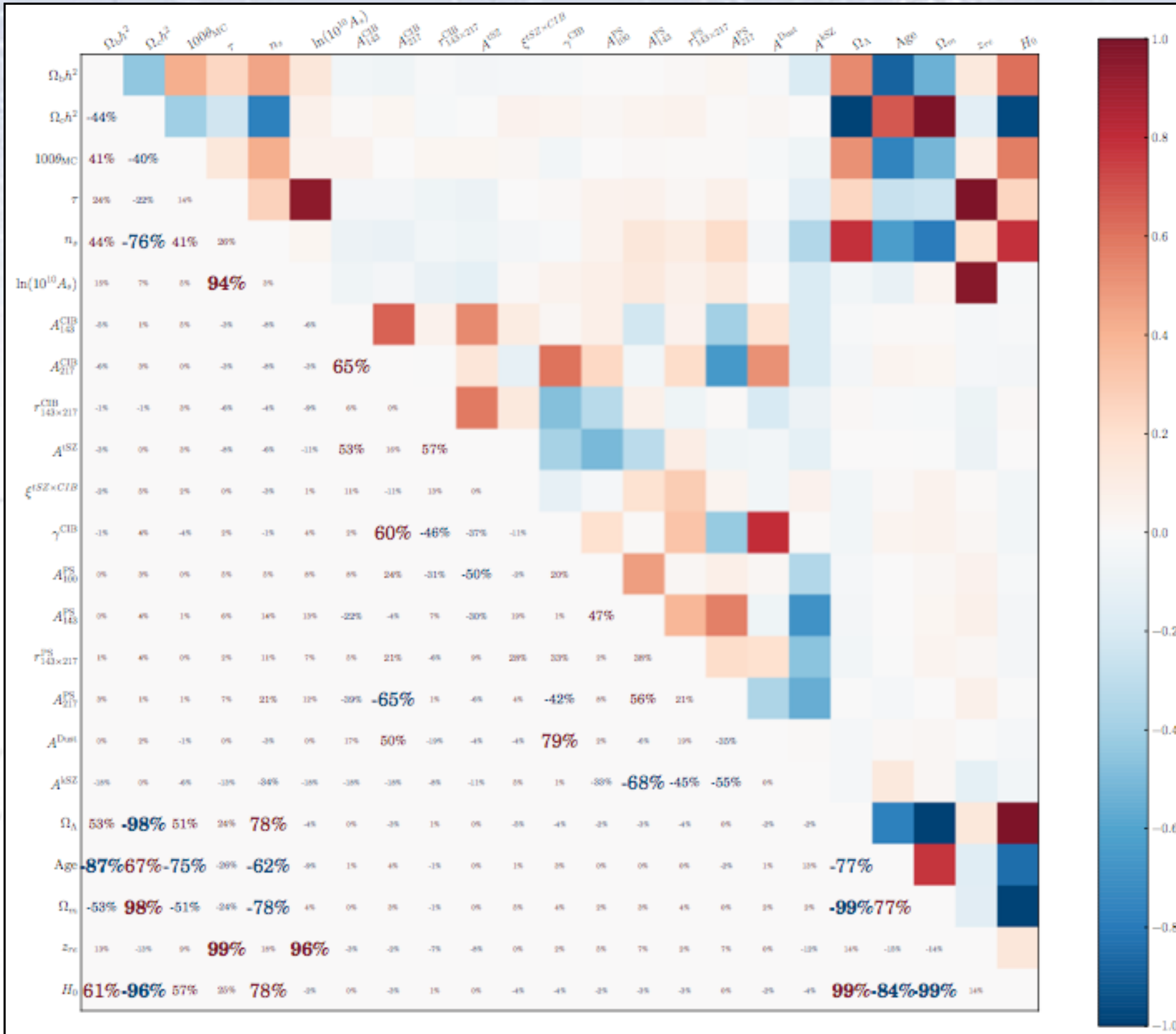
Correlations influence results in complex ways!

They need to be taken into account, for example in **Error Propagation!**

Correlations may contain a significant amount of information.

We will consider this more when we play with multivariate analysis.

Planck example



Rank correlations

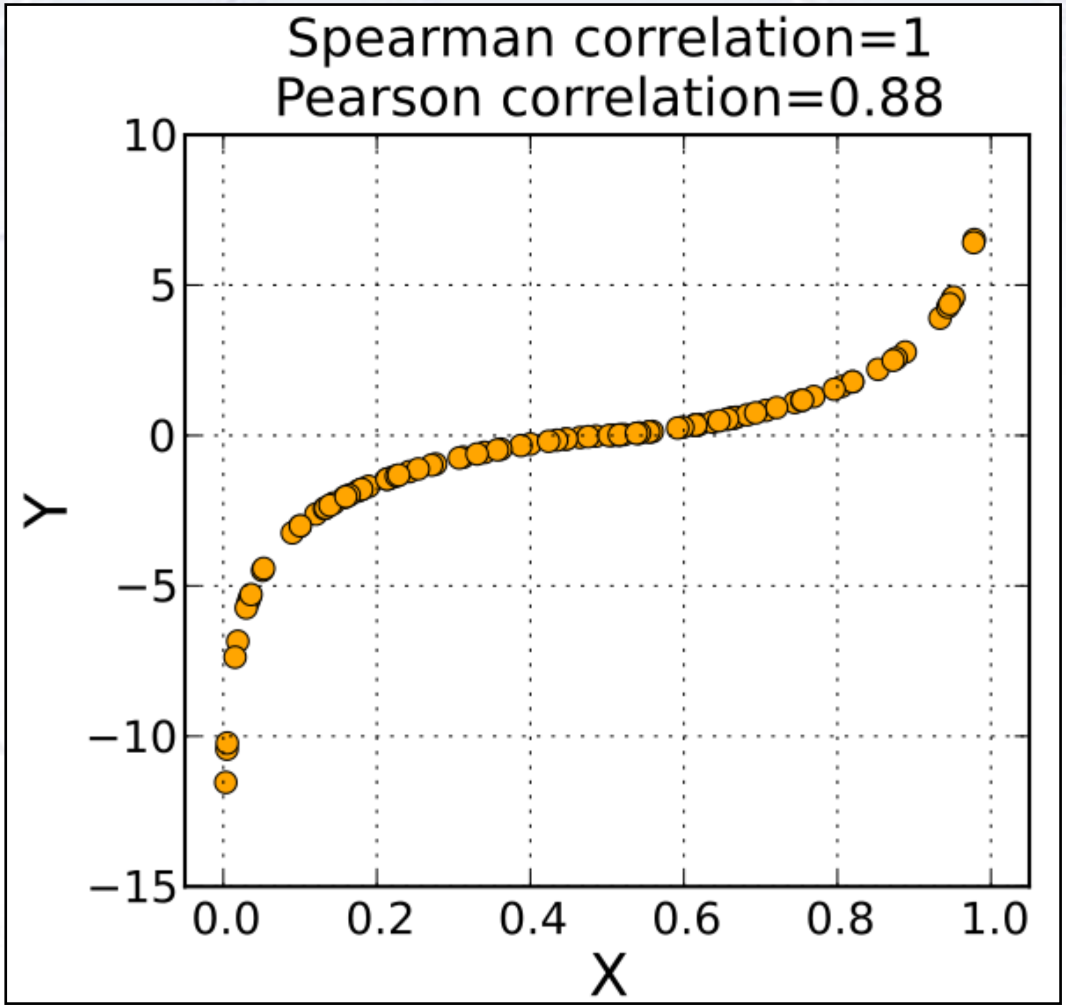
Sometimes, variables are perfectly correlated, just not linearly:

In this case the Pearson correlation is not the best measure.

Rank correlation compares the ranking between the two sets, and therefore gets a good measure of the correlation (see figure).

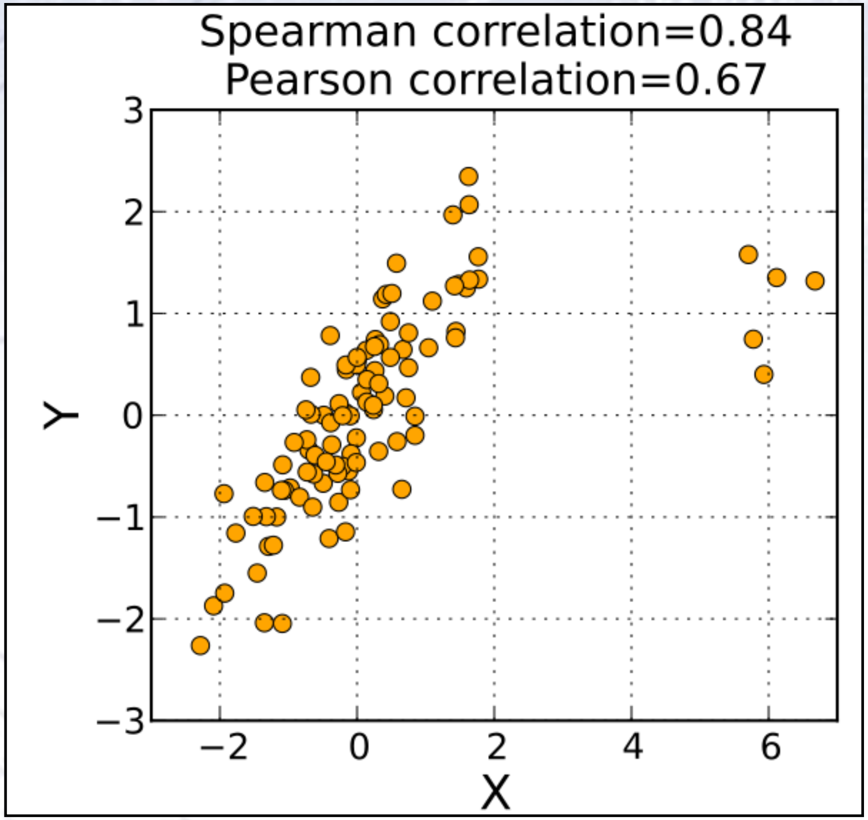
The two main cases of rank correlations are:

- Spearman's rho
- Kendall's tau



Rank correlations

An additional advantage is, that the rank correlation is less sensitive to outliers:



The two rank correlations are special cases of a more general rank correlation.

Typically, Spearman's rank correlation is used.

The definition is:

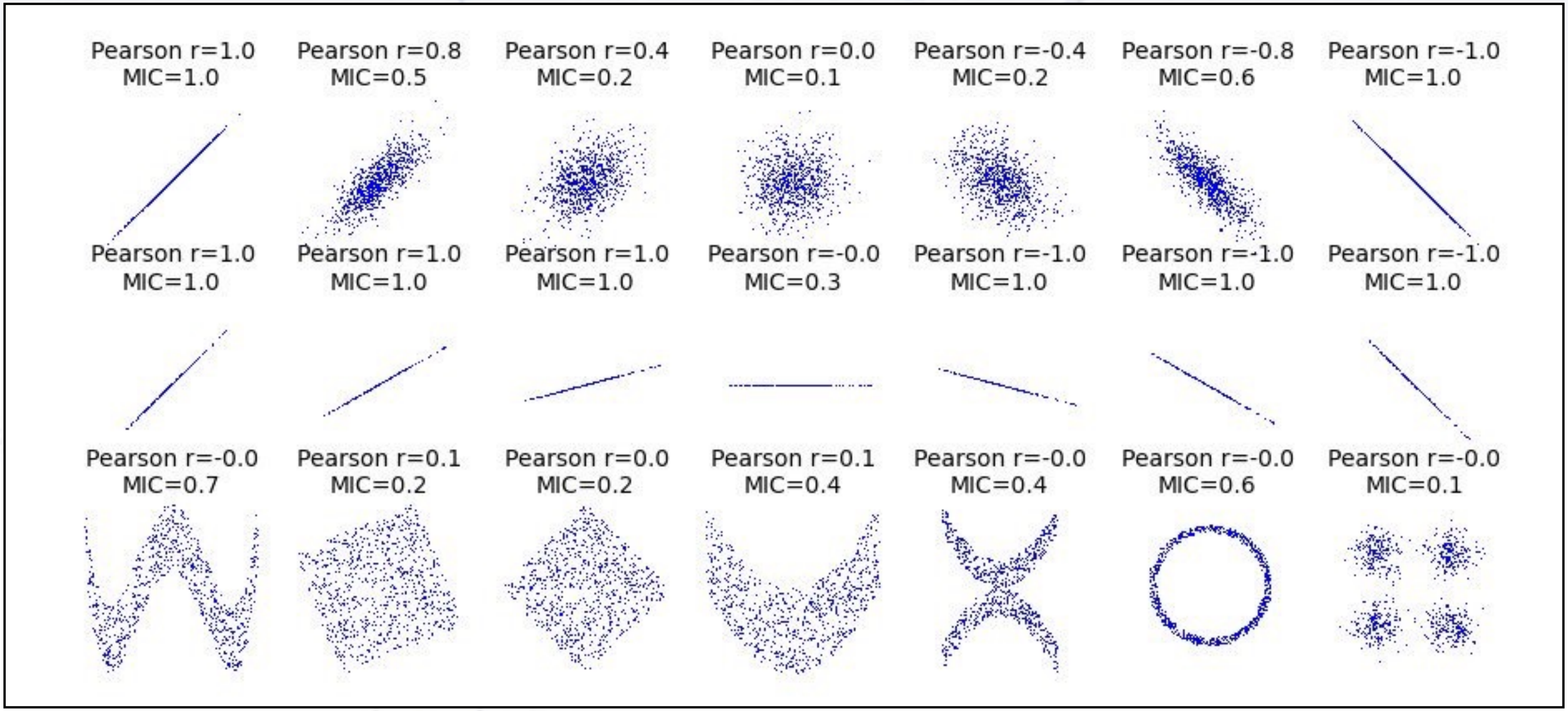
$$\rho = 1 - 6\sum_i (r_i - s_i)^2/(n^3 - n)$$

where r_i and s_i is the rank of the i'th element.

Non-linear correlations

Non-linear correlations (associations) are harder to measure, but possible:

- Maximal Information Coefficient (MIC), see reference and Wikipedia on MIC.
- Mutual Information (MI), linked to entropy, see Wikipedia on MI and SKLearn.
- Distance Correlation (DC) between paired vectors, see Wikipedia on DC.

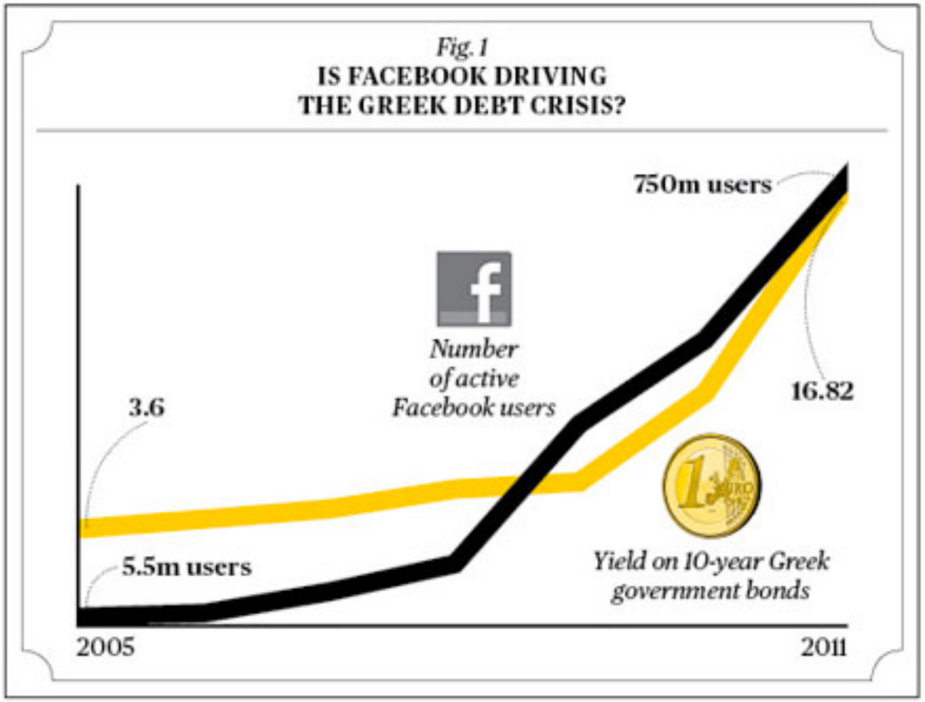


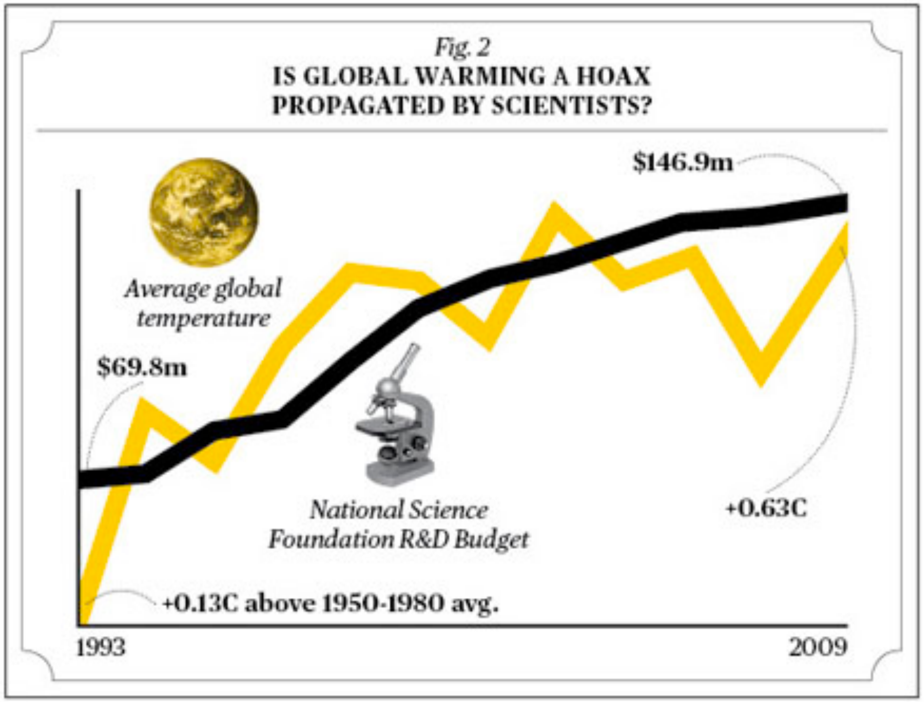
Original paper: "Detecting Novel Associations in Large Data Sets" (2011). Science 334 (6062): 1518–1524.

Correlation Vs. Causation

"Com hoc ergo propter hoc"

(with this, therefore because of this)





It is a common mistake to think that correlation proves causation…

Correlation Vs. Causation

"Com hoc ergo propter hoc"

(with this, therefore because of this)