

# Data exploration: why and how

*Pietro Franceschi*

*Computational Biology - Research and Innovation Centre - Fondazione E. Mach*



# Data, knowledge & data analysis

Data



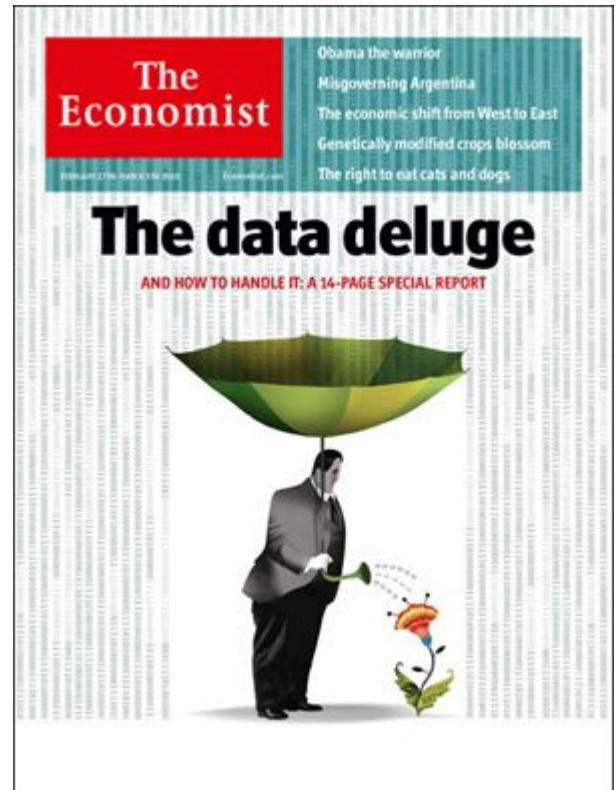
Knowledge

In *omic* sciences, bioinformatics, statistics, ... provide the tools to:

- Promote the **incremental** progress of science
- Guarantee the **validity** and the correctness of the results
- Facilitate the production of “**scientific**” results (of general validity ...)
- Be consistent
- Get the **maximum** from complex data

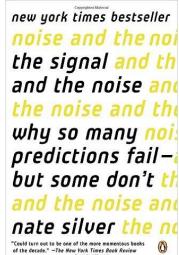
# Di che numeri parliamo ...

- Metabolomica Untargeted: ~10k variabili
- Metabolomica Targeted : ~ 300 molecole
- Proteomica: ~ 1k variabili
- Genomica e Metagenomica: 😱



# Signal, Noise & Data Driven Science

Measuring more does **not** necessarily mean **understanding** more



**Nate Silver**

*"The Signal and the Noise: Why So Many Predictions Fail, but Some Don't"*

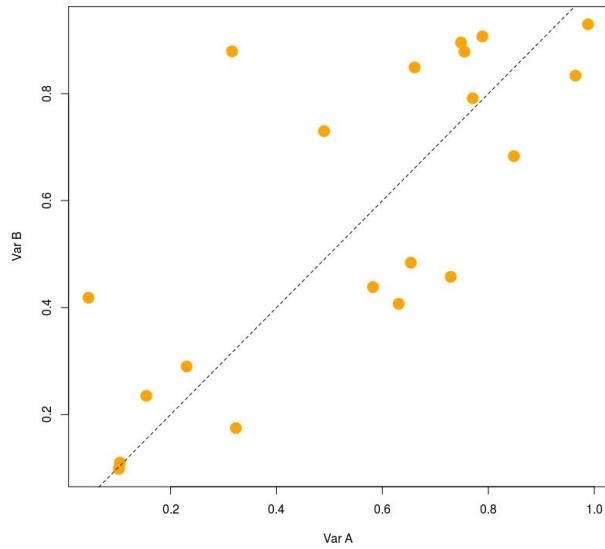
# “Noise” & false positives

# Data mining in a nutshell ...

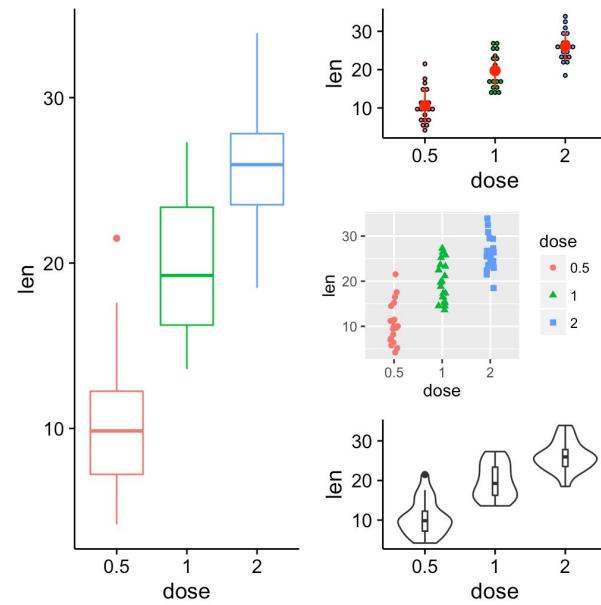
- Look for true Information (**signal**) inside a complex dataset ...
- Look for **organization** inside the data ...



# Two examples of organization ....



Two variables  
highly “correlated”



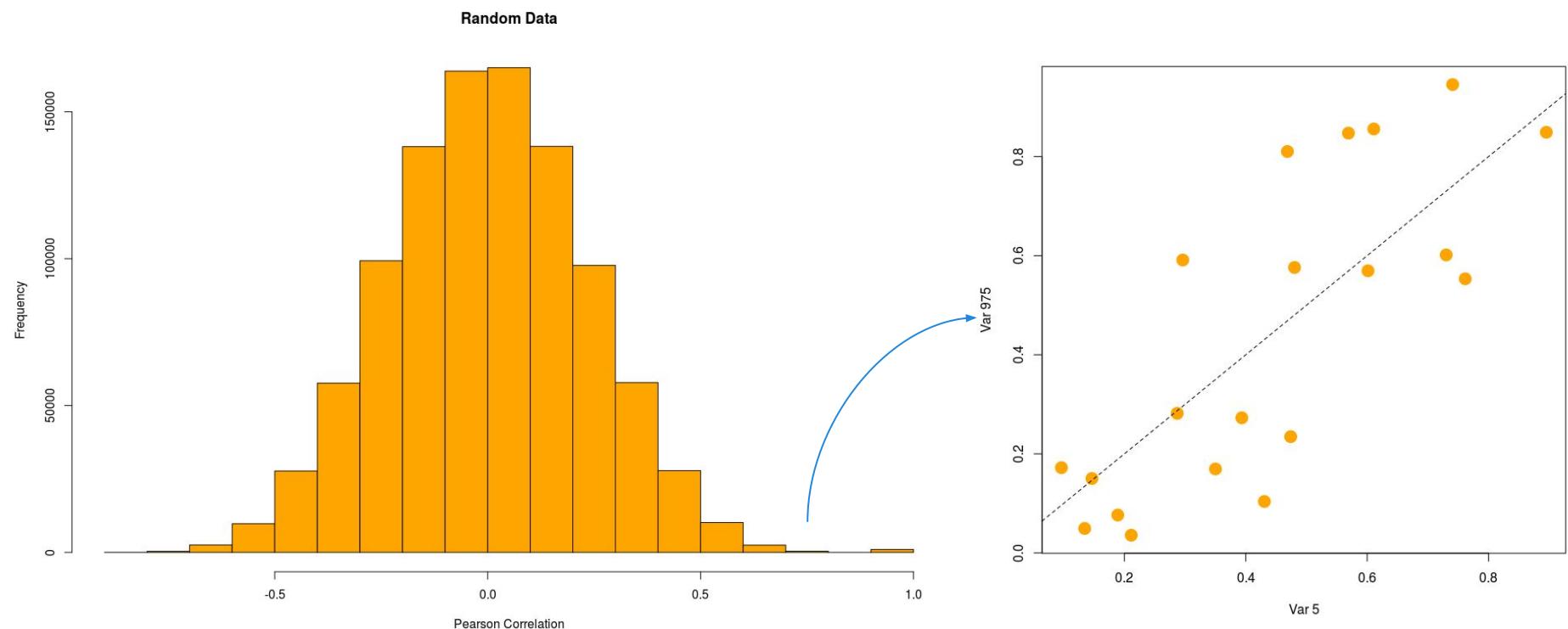
Biomarkers ...

# The challenge

- We construct a “fake” data matrix of 20 samples and 1000 variables
- We look correlated variables ([organized Information](#))
- There is not “true” information in the data



# An example of “self - organization” ?



# ALL FOR YOU

Questions for you ...

- Am I playing fair?
- Where is the trick?
- Why I call this “false positive”?
- Is there an error somewhere?

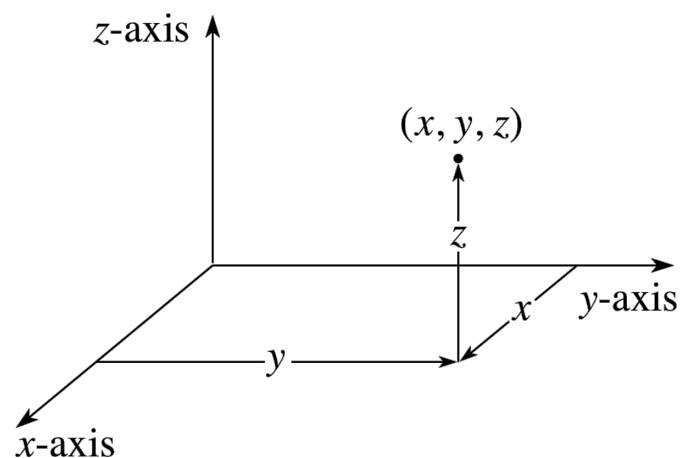
# Organized “by chance”

- We find **true** organization
- This is not the result of a true chemical/biological/physical process
- This organization is “true” **only for my dataset**
- There are no errors and no tricks
- Do you see the scientific consequences?

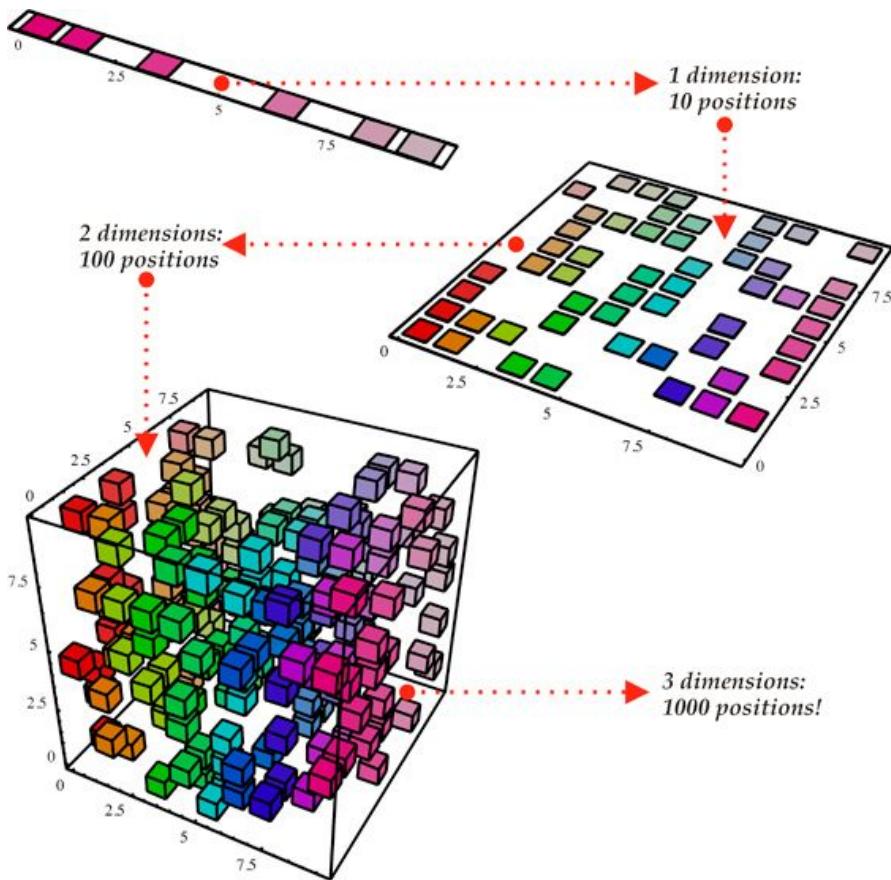


# n variables, n dimensions

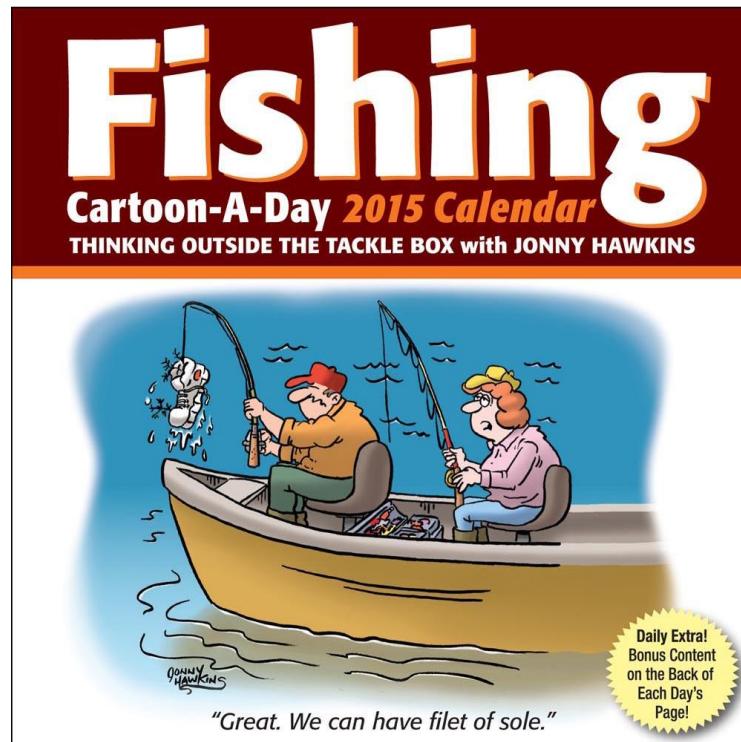
In a multivariate setting, each sample is a point in the multidimensional space



# Empty Space



Be careful, it could be a sole ...



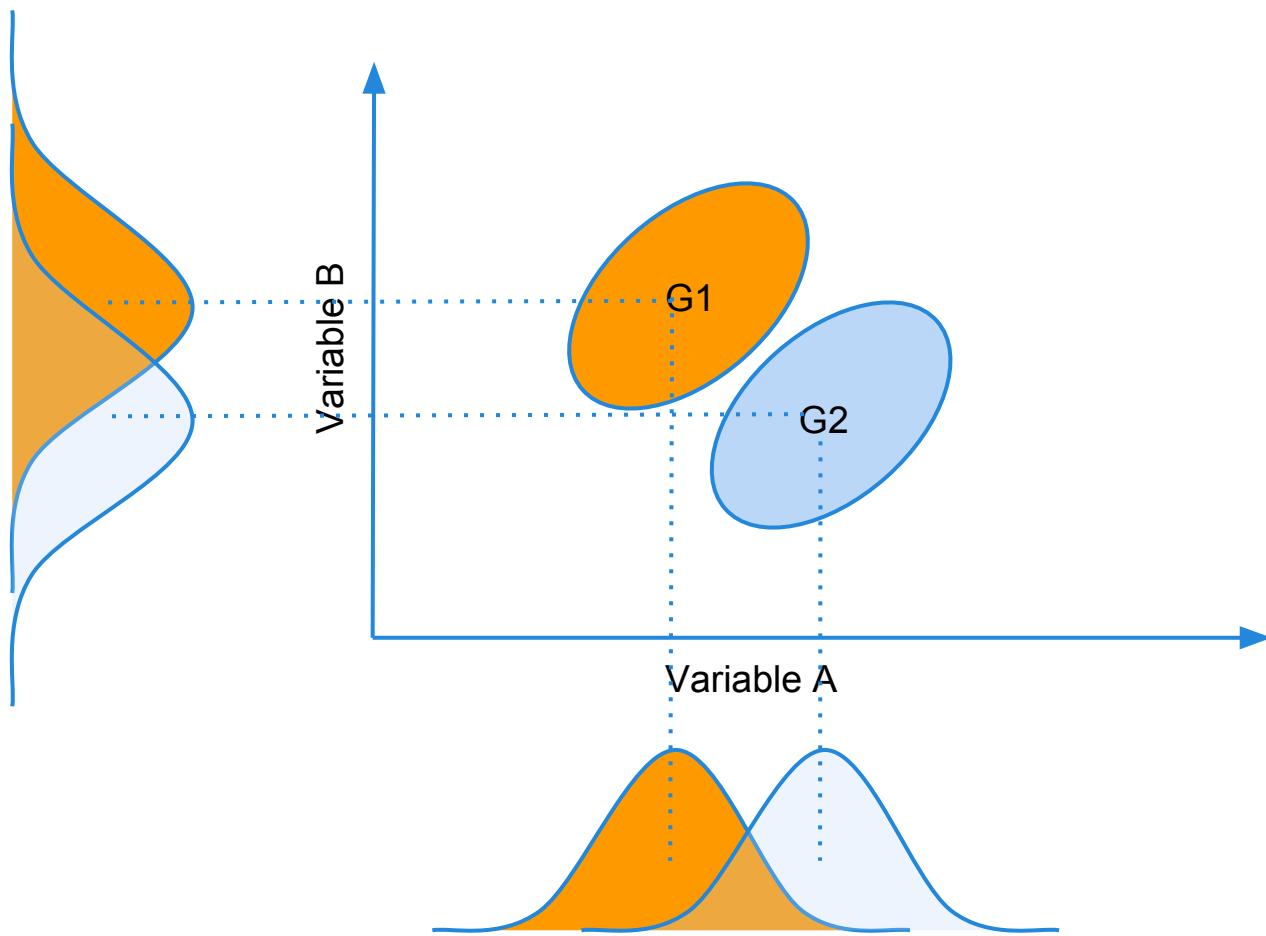
# Luckily enough ..

The variables are not independent



- Biological dependence
- Chemical dependence
- ....

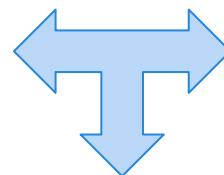
# Multivariate vs Univariate



There ain't no such  
thing as a free lunch.

... IN STATISTICS

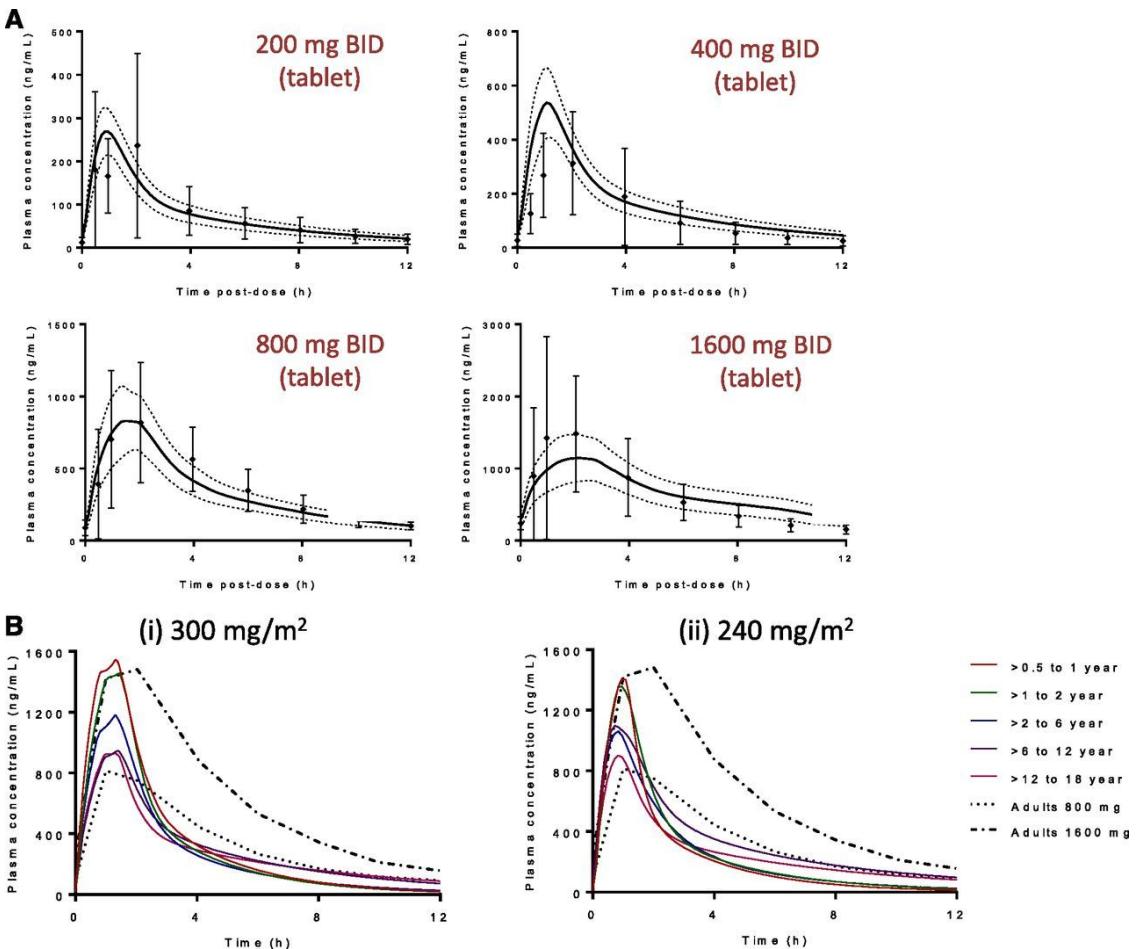
SAMPLES



MODEL  
COMPLEXITY

KNOWLEDGE

# Pharmacokinetics



# Understanding (Modelling) vs Classifying

To classify something it is not necessary to understand why ...



*To recognize a tree your brain does not need to know the details of photosynthesis*



# The NEW ENGLAND JOURNAL of MEDICINE

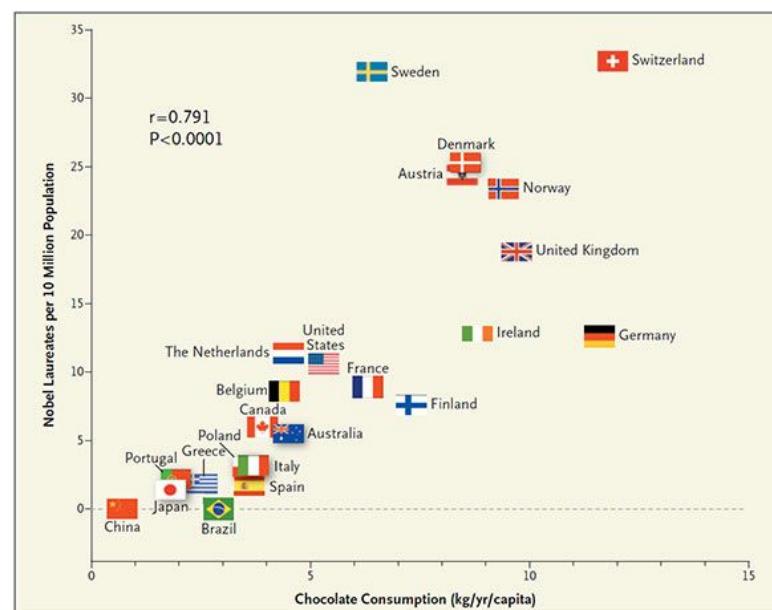
## ***Chocolate Consumption, Cognitive Function, and Nobel Laureates***

Franz H. Messerli, M.D.

*N Engl J Med 2012; 367:1562-1564 October 18, 2012 DOI: 10.1056/NEJMMon1211064*

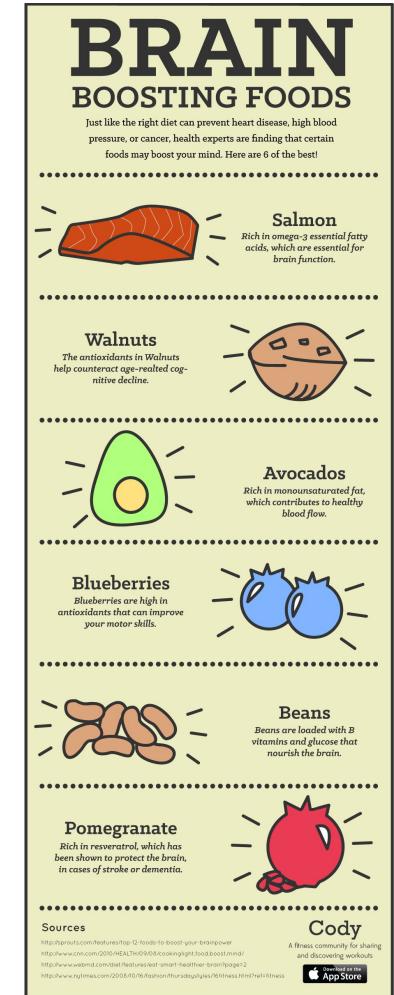
*Chocolate consumption could hypothetically improve cognitive function not only in individuals but in whole populations. Could there be a correlation between a country's level of chocolate consumption and its total number of Nobel laureates per capita?*

There was a close, significant linear correlation ( $r=0.791$ ,  $P<0.0001$ ) between chocolate consumption per capita and the number of Nobel laureates per 10 million persons in a total of 23 countries (Fig. 1)



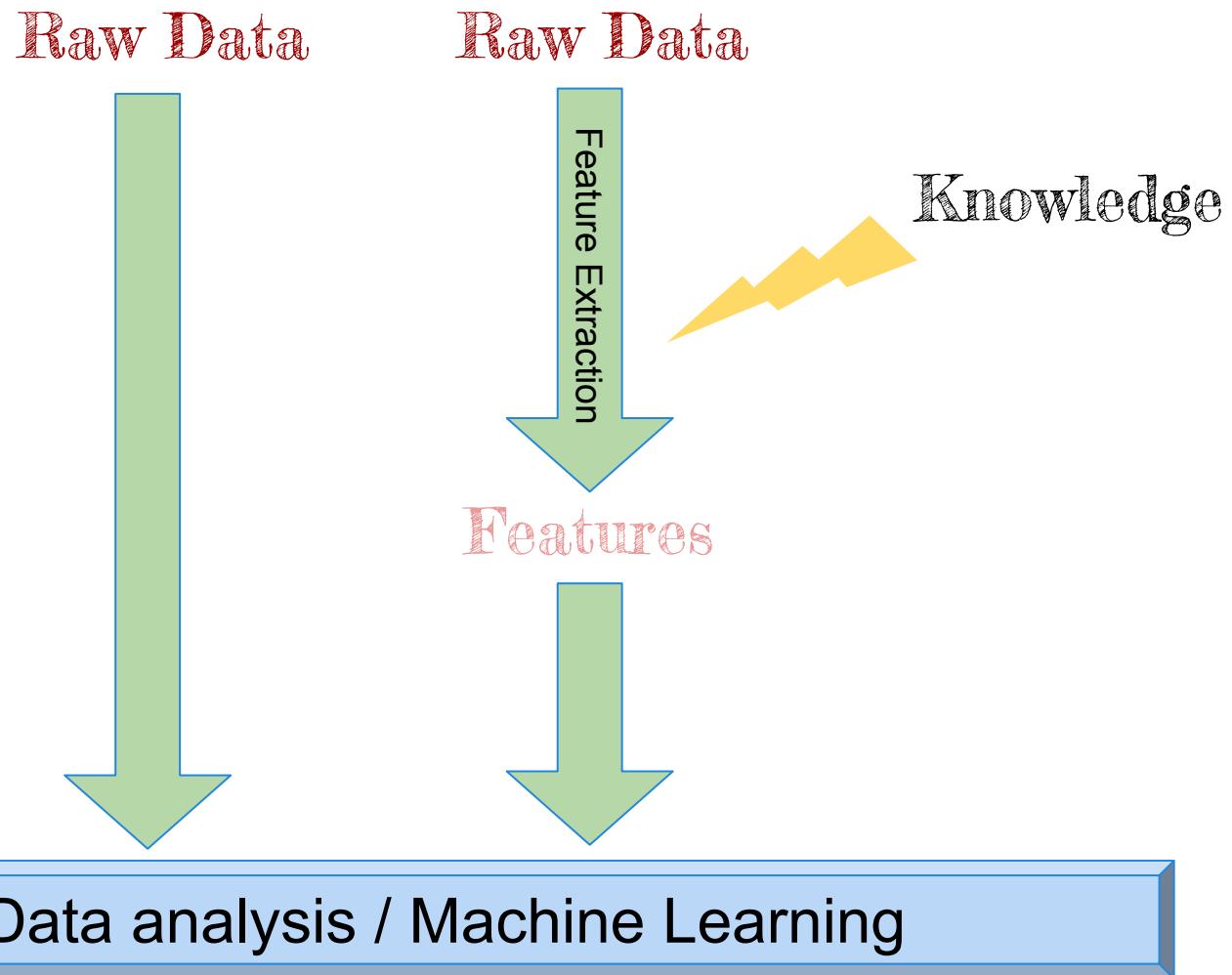
# ... food for brain

- It could be not easy to spot the predicting variables if you use a complex “predictor” (classifier,...)
- Sometimes is not needed! (Think to google)
- Modeling and predicting are not synonyms
- It depends on what is your objective ...



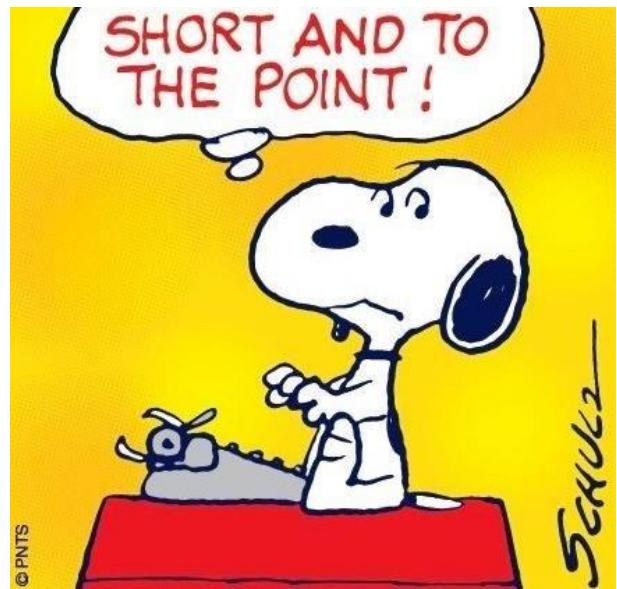
**Things to look at ...**

# What variables should I use?

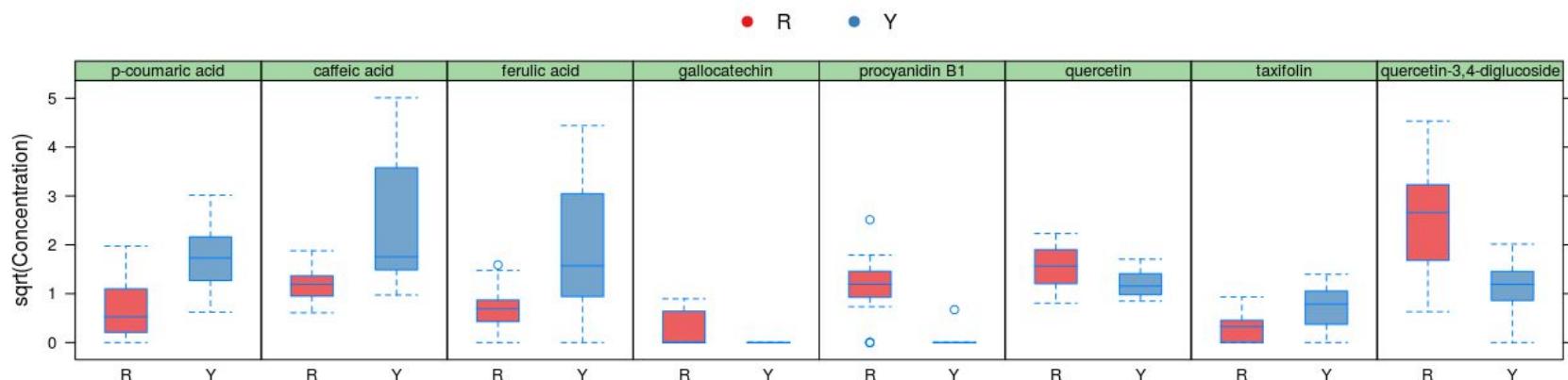


# ... food for brain

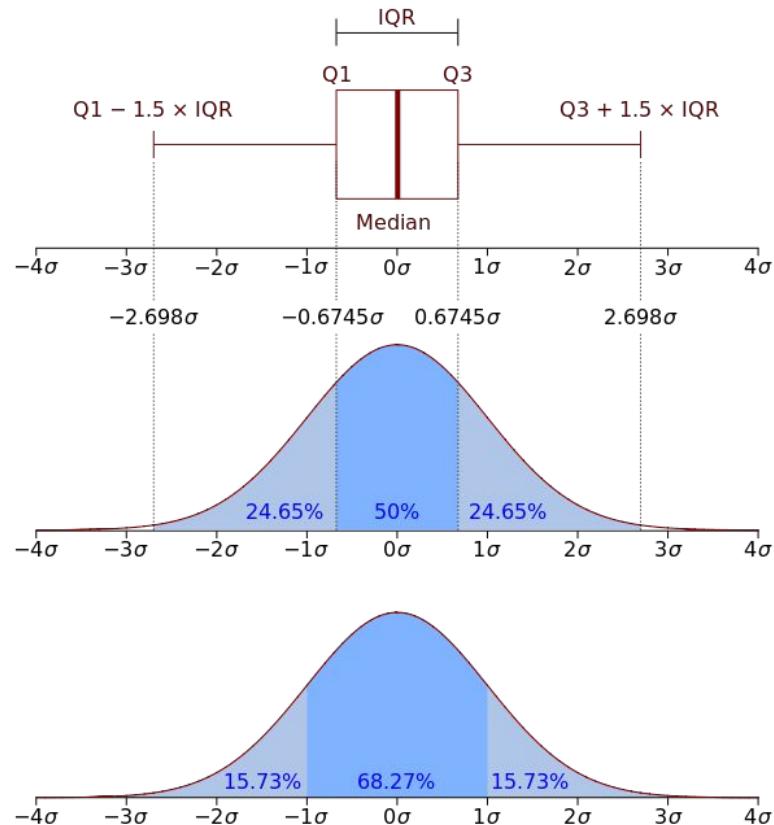
- The “raw” content of information in the raw data is bigger ... but also the noise is bigger
- The process of feature extraction distill useful knowledge from noise
- Something can get lost
- It really depends on the problem ...
- What would you do in image analysis, metabolomics, metagenomics?



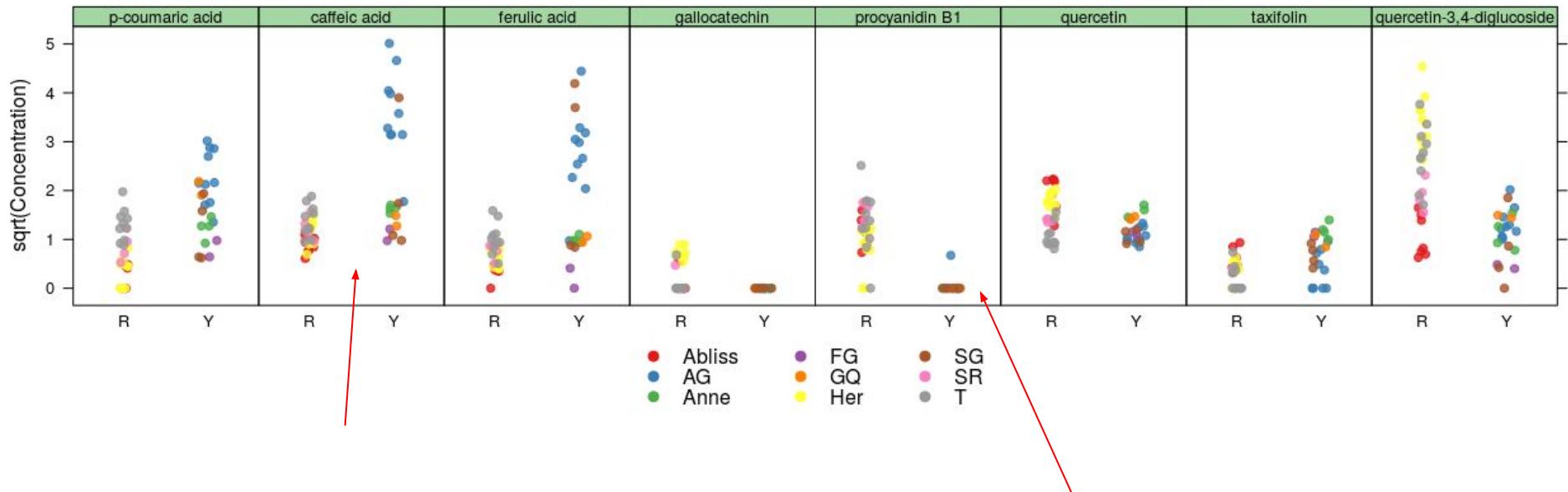
# Confounding Factors



# Box plots .... ;-)



# Confounding Factors

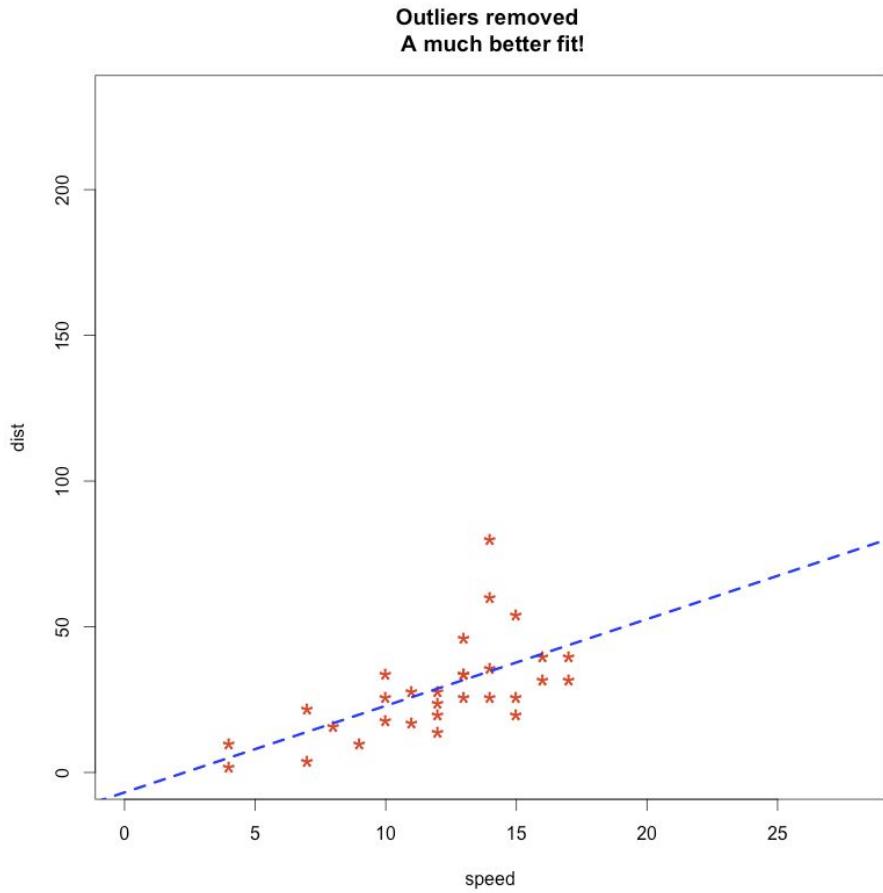
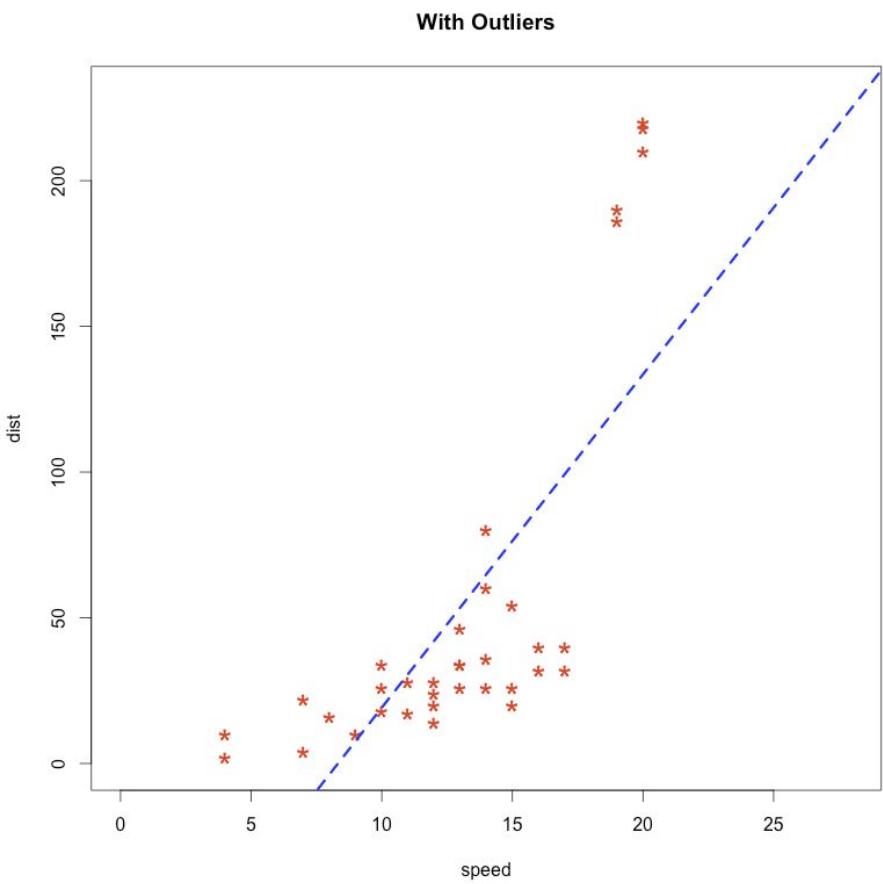


This seems to be not related to color ....

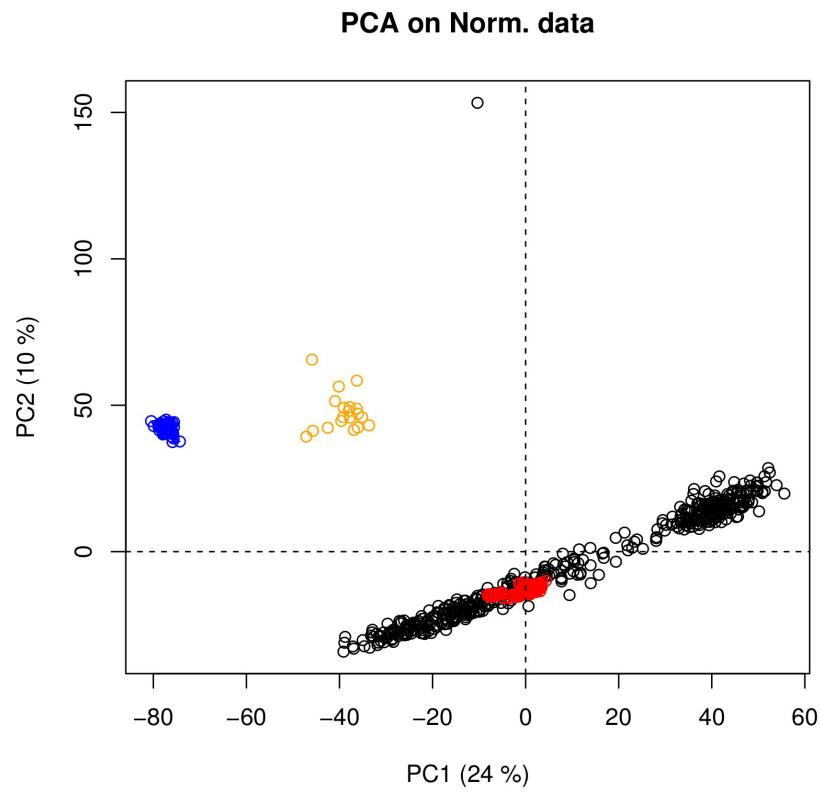
These look like zeroes ... what does **zero** mean?

In machine learning , hidden confounding factors could bias validation schemes

# Outliers



# Outliers



# ... let's discuss ...

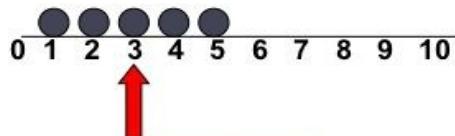
- What is an outlier?
- How should I treat it?
- When is fair to get rid of a sample?

ALL FOR  
YOU

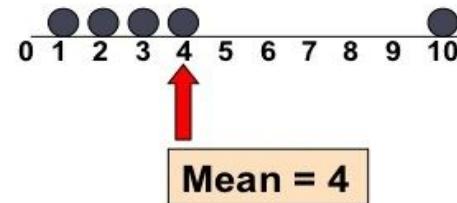
# Rely on robust statistics and methods ...

## The Mean: Sensitive To Outliers

- The most common measure of the central tendency
- Affected by extreme values (outliers)



$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

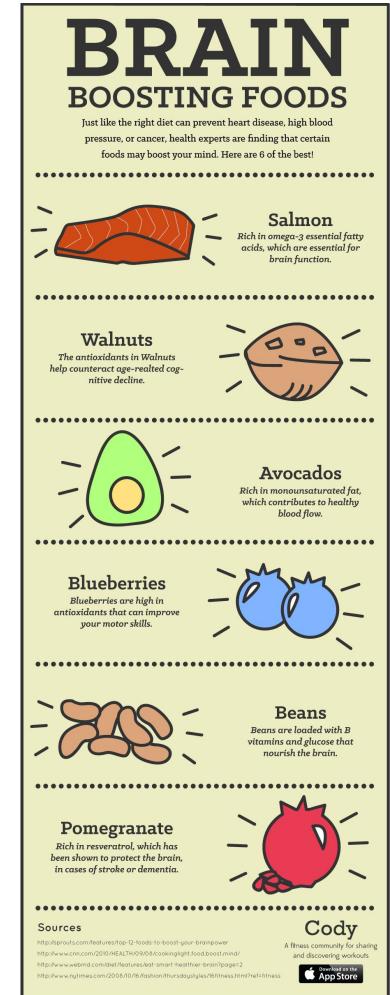


$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

The mean is not **robust** the median is better

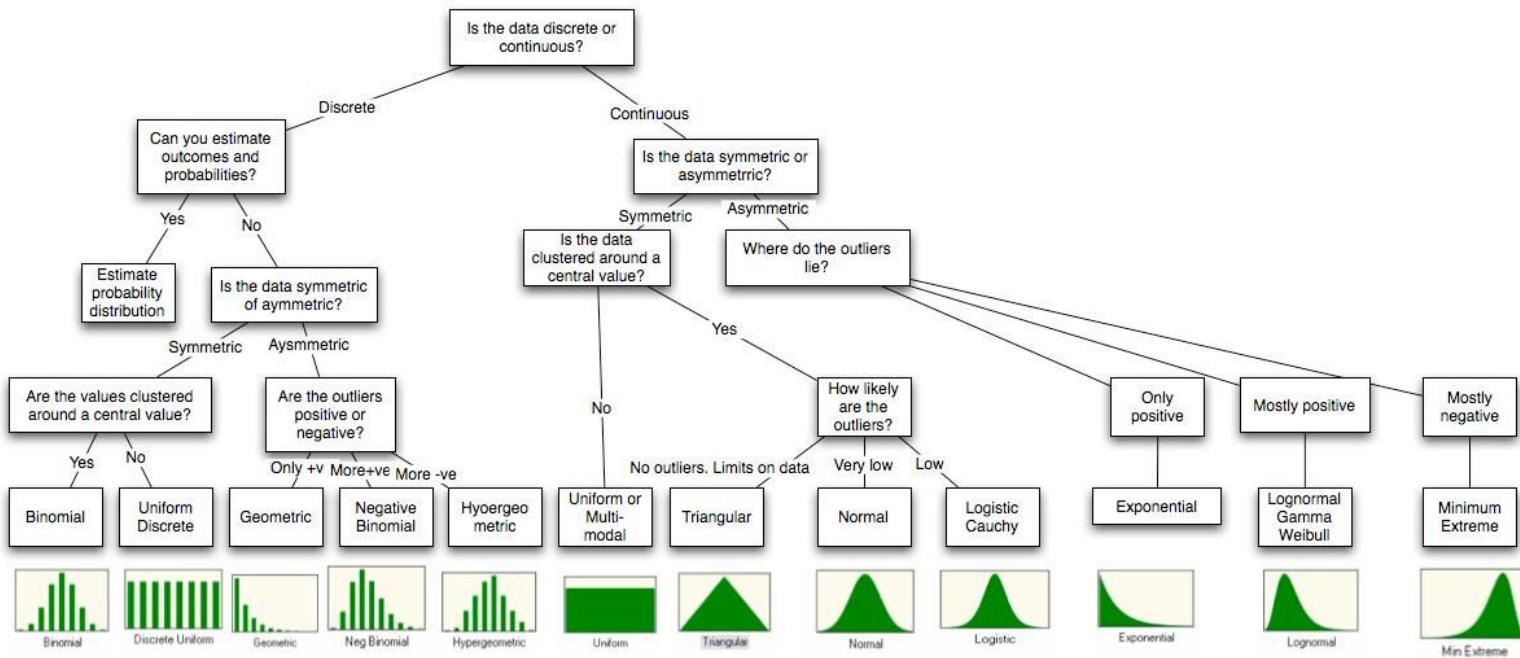
# ... food for brain

- Use boxplots wisely and not to hide the fact that you have few samples.
- If a sample looks as an outlier go back to the data and try to discover there are good reasons to discard it
- If not try to rely on robust statistics



# Which is the distribution of your data?

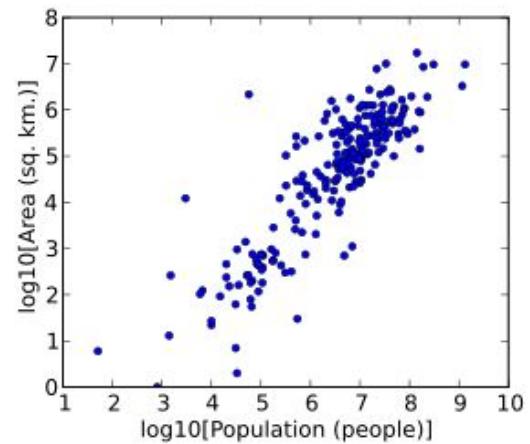
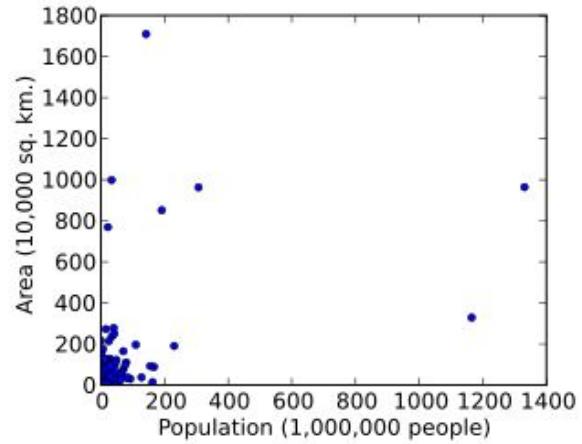
Many statistical methods contains hidden assumptions on the “acceptable” distribution of the data. As a rule, everything works better if data are normally distributed



# Data transformation

In many cases “normality” can be restored by data transformation

Common transformations are square root or logarithm



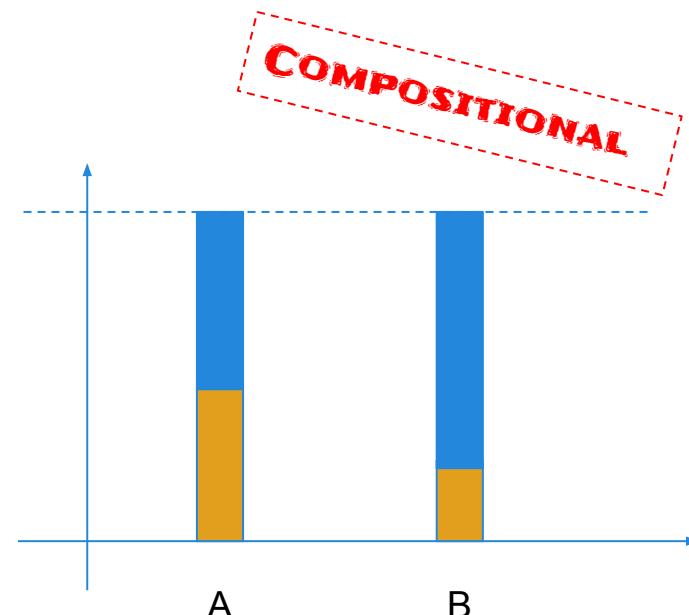
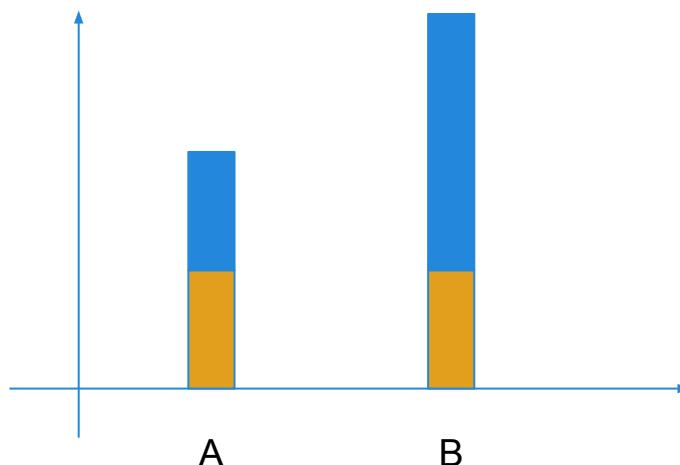


# Compositional data

From Wikipedia, the free encyclopedia

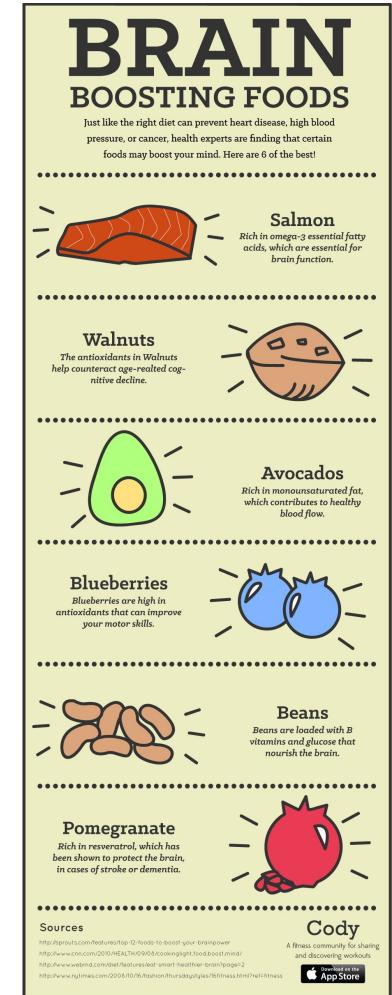
In statistics, **compositional data** are quantitative descriptions of the parts of some whole, conveying exclusively relative information. Measurements involving probabilities, proportions, percentages or ppm can all be thought of as compositional data.

## Relative Abundances/ Relative concentrations



# ... food for brain

- The fact that in conditions A and B the sum of the variables should be the same introduces correlation
- In the first case the variable “blue” is a marker and “yellow” is not.
- In the second case both are markers!



# ... let's discuss ...

- Can omic- data be compositional?
- Why?
- Is this a problem?

ALL FOR  
YOU

# Variable range and magnitude

Almost invariably, we are dealing at the same time with variables of low and high intensity (concentrations, abundances, ...)

The “structure” of our dataset could then be dominated by high intensity variables.



*The exact Representation of those two very Remarkable Persons, Mr. Blamfield the Staffordshire Giant, & Mr. Coss the Norfolk Dwarf, Drawn by Mr. Hackmann Feryngton, as he has Moulded them from the Life, & are to be seen there as of late. The one measuring full 7 feet 4 inches high, the other not exceeding 3 feet high. & each of these extraordinary Persons Departed this Life at the Age of thirty four. At the hundredth and last part of his life in a small Lane near Temple-Bar: the Dwarf near Oldgate - Published according to Act of Parliament May 4, 1771. sold by Stans, Robert Sigravur & M. Hawsworth Seller near the Sun Inn in Gray's Inn Lane, & R. Dibdin, 36 Holborn. —*

# Scaling

The process of making variable comparable is called **scaling**.

As usual this can be done in many different way - unit variance scaling is popular - and the scaling will affect the outcomes of the analysis

Mean **centering** is also a commonly applied strategy to make variable comparison easier



# ... let's discuss ...

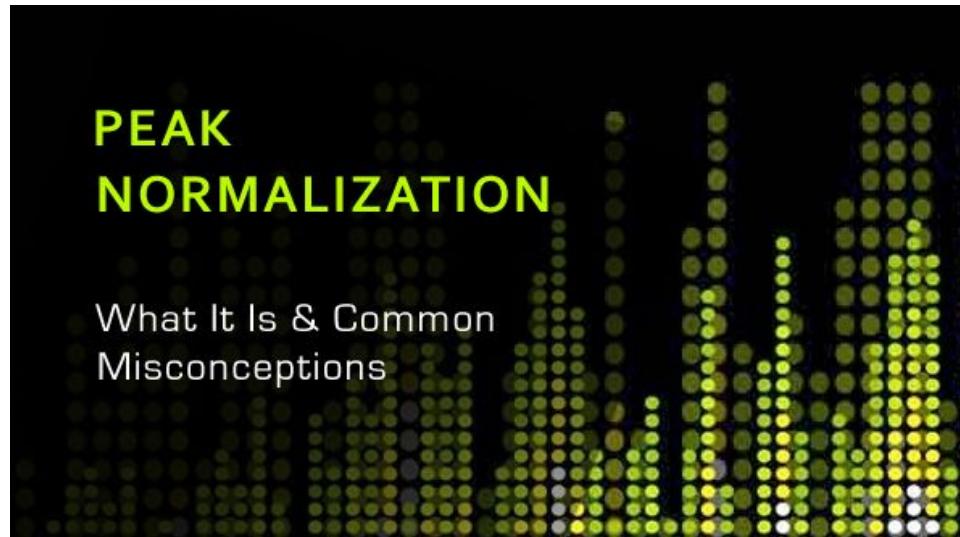
- Is scaling always fair?
- Do you see its limitations?
- How can I decide the “best” scaling approach?

ALL FOR  
YOU

# Making samples comparable ...

In many cases (metabolomics, metagenomics, ...) the overall response of the data acquisition pipeline is not constant across the samples.

To look for biomarkers it is necessary to correct for that



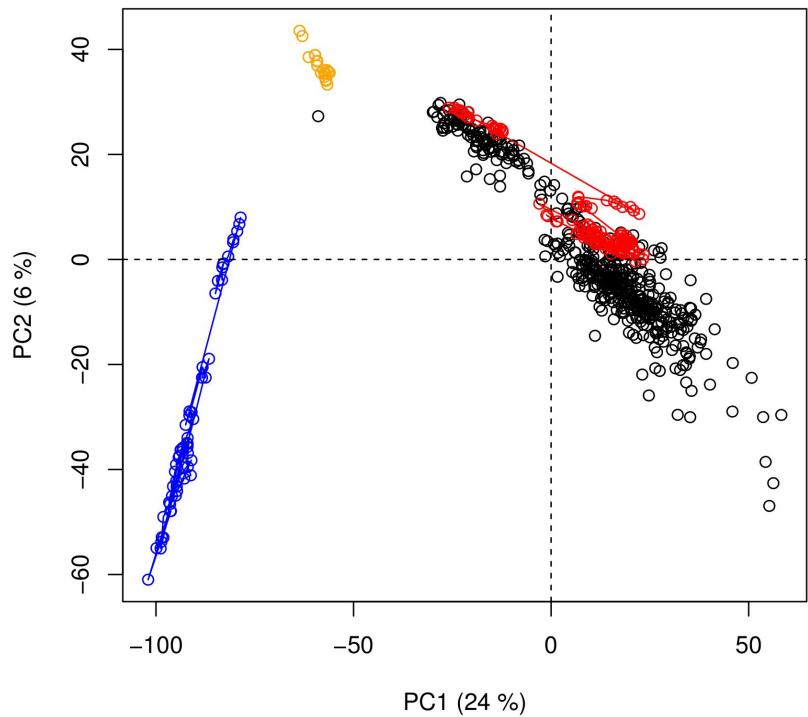
## Normalization

# ... food for brain

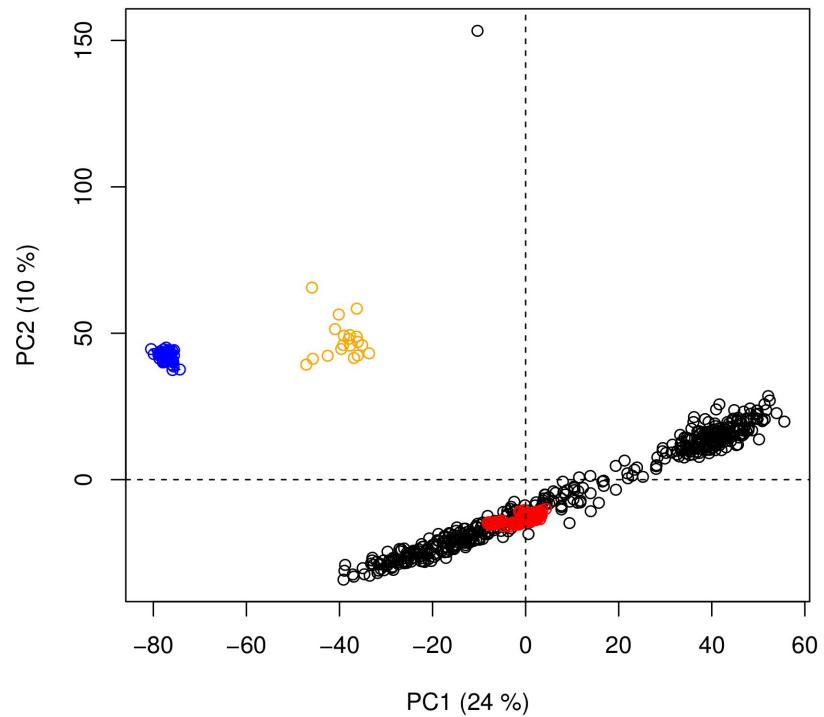
- It would be better to avoid normalization
- If normalization is done relying on relative abundances the data are becoming compositional!
- A wise organization of the “sample list” taking into account the experimental design is of great help
- Normalization strategies are domain specific!



**PCA**



**PCA on Norm. data**





**WIKIPEDIA**  
The Free Encyclopedia

[Main page](#)

[Contents](#)

[Featured content](#)

Not logged in [Talk](#) [Contributions](#) [Create account](#) [Log in](#)

[Article](#) [Talk](#)

[Read](#) [Edit](#) [View history](#)

Search Wikipedia



# Missing data

From Wikipedia, the free encyclopedia

In statistics, **missing data**, or **missing values**, occur when no **data value** is stored for the variable in an **observation**. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.

- Preprocessing errors
- Low intensity
- Low abundance
- ...



## I could ...

- Forget them and live in peace
- Fill them with a value ... which one?
- Take out the features with too many nans ...
- ... but then what I can do with the rest?

Peace   
Love   
Happiness 

The process of substituting missing values with meaningful number is called imputation

... let's discuss ...

ALL FOR  
YOU

Which is the best way to perform imputation?

#1

LETS GO  
LIVE!!

# Towards Statistical Significance

# Data, knowledge & data analysis

Data

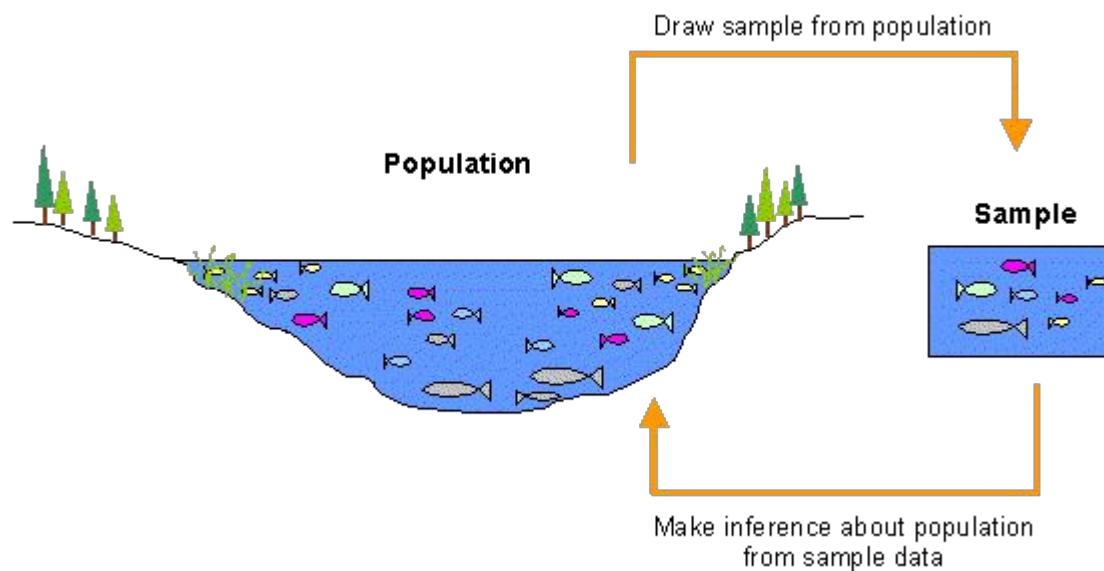


Knowledge

In *omic* sciences, bioinformatics, statistics, ... provide the tools to:

- Promote the **incremental** progress of science
- Guarantee the **validity** and the correctness of the results
- Facilitate the production of “scientific” results (of general validity ...)
- Be consistent
- Get the **maximum** from complex data

# Population, sampling and inference



# The challenge

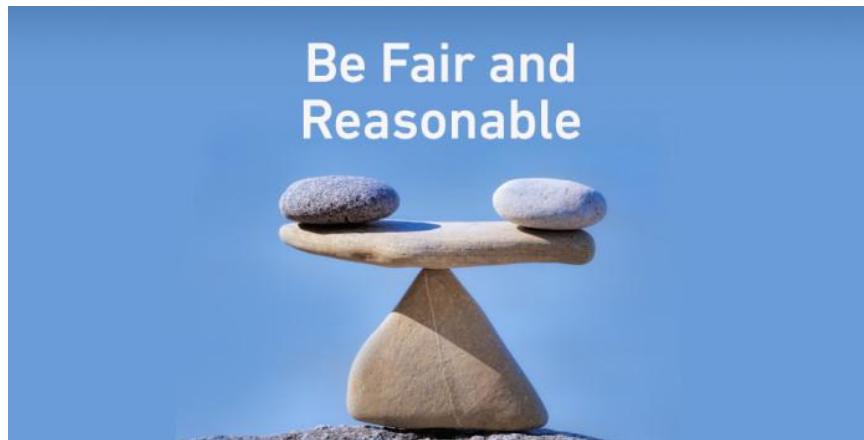
- A sampling can be unlucky ...  
but the bigger the better ;-)
- Organization can be true only for  
the sample and not for the  
population
- We cannot eliminate this, so we  
are never sure that what we see  
has general validity



The best we can do is to try to quantify our uncertainty

# A reasonable approach ... in words

- Suppose that the organization is absent at the population level ( $H_0$ )
- Choose a “statistic” to quantify organization
- Calculate the probability of obtaining “at least” the observed level of organization only by chance (p-value!)
- Set a threshold of acceptable incertitude (1%, 5%)



If the probability is below the threshold, one can reasonably expect that the observed organization is true at the population level

# ... food for brain

- In all this approach I'm only testing the absence of organization ... not the presence of a specific level of organization ( $H_0$ , not  $H_1$ )
- In any case I'm never sure!
- The best I can do is quantify the incertitude ...

