Webinar 1 - Data Analysis

Pietro Franceschi
pietro.franceschi@fmach.it

FEM - UBC

A clear evidence

Nowadays almost all biological/natural systems are characterized by quantitative (molecular) approaches

Biosystem Data Analysis

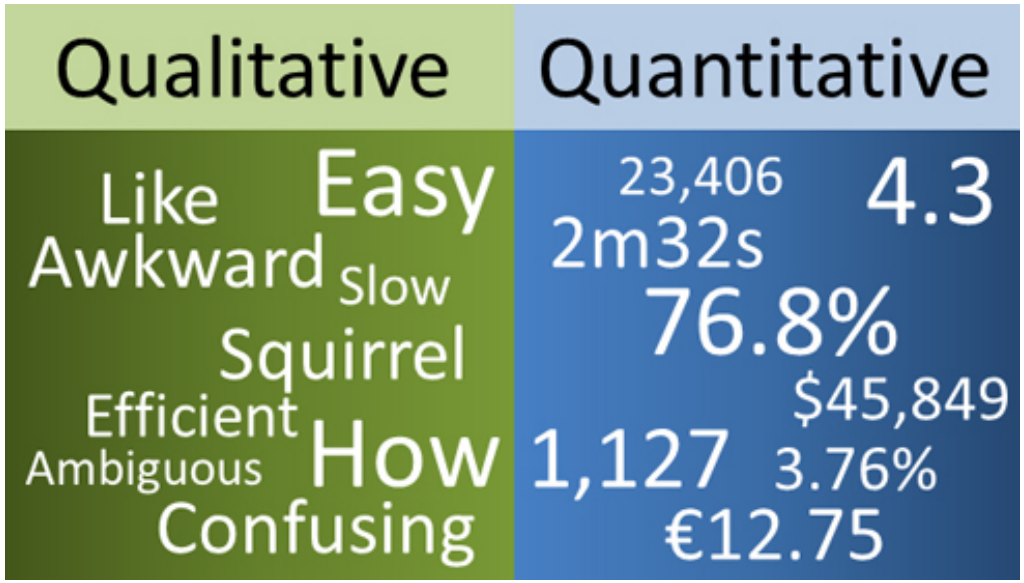
Develop and validate methods for organize, summarize and visualize complex biological data through the integration of ***bioinformatics*** and ***biostatistics***

Always more data

- Metabolomics
- Proteomics
- High throughput phenotyping
- Distributed network of sensors
- ...

Qualitative and quantitative

Technological advancement (experimental techniques, IT) is transforming the "soft sciences" in quantitative disciplines



Data, knowledge, data analysis

Bioinformatics, statistics, chemometrics, provide the tools to:

- Promote the incremental progress of science (*we stand on the shoulders of giants*)
- 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

Data Deluge

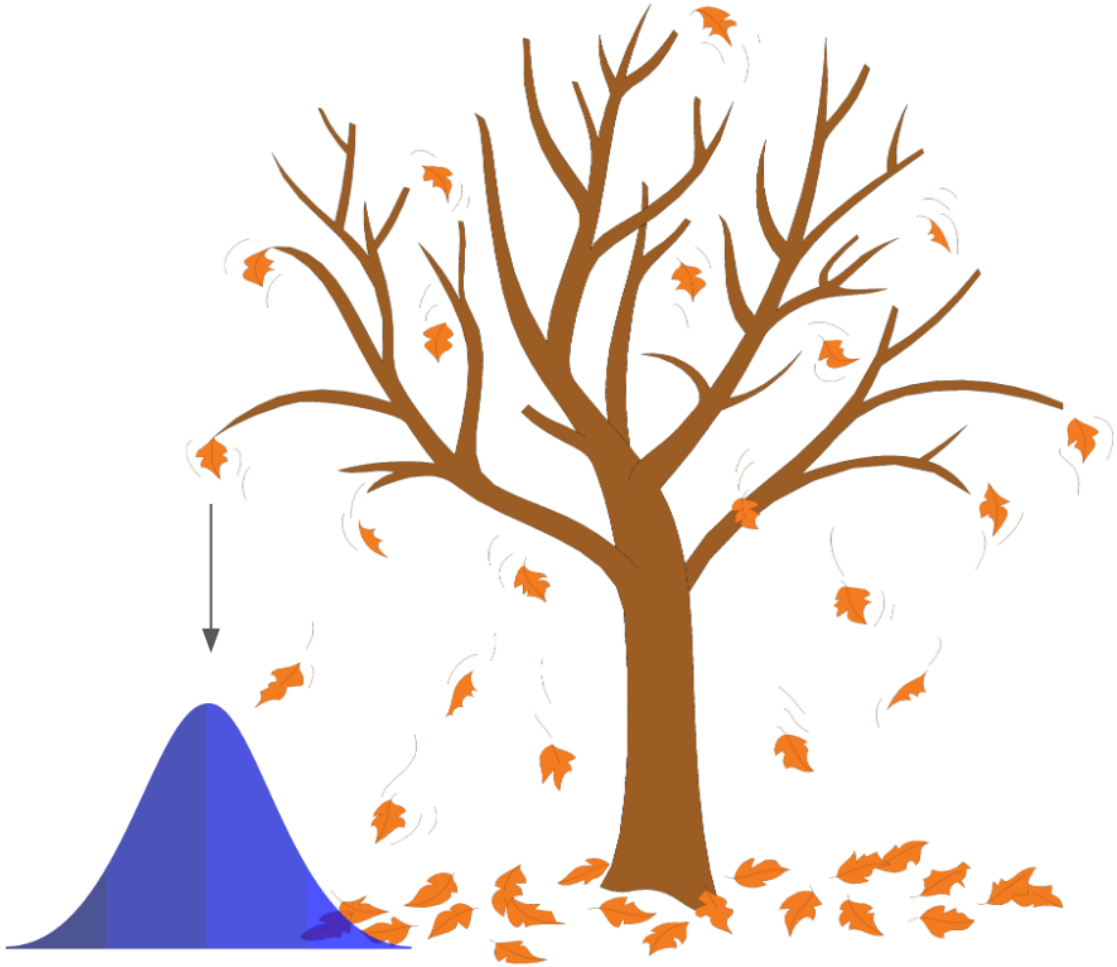
Measuring more does not necessarily mean understanding more

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

An unavoidable chain

$$\text{Quantification} \Rightarrow \text{Measure} \Rightarrow \text{Variability}$$

Why we need statistics



Some Ideas

- The leaf always fall (i.e there is a general rule!)

Some Ideas

- The leaf always fall (i.e there is a general rule!)
- . . . but it is not at all easy to say where

Some Ideas

- The leaf always fall (i.e there is a general rule!)
- . . . but it is not at all easy to say where
- Every leaf is different . . . so measuring once is not sufficient

Some Ideas

- The leaf always fall (i.e there is a general rule!)
- . . . but it is not at all easy to say where
- Every leaf is different . . . so measuring once is not sufficient
- Probability!

Some Ideas

- The leaf always fall (i.e there is a general rule!)
- . . . but it is not at all easy to say where
- Every leaf is different . . . so measuring once is not sufficient
- Probability!
- If it would be an apple the prediction would be easier

Some Ideas

- The leaf always fall (i.e there is a general rule!)
- . . . but it is not at all easy to say where
- Every leaf is different . . . so measuring once is not sufficient
- Probability!
- If it would be an apple the prediction would be easier
- The spread depends on the relative importance of the **interaction** between the leaf and the environment

Sources of Variability

- Biological variability
- Technical variability
  - Sample collection
  - Sample handling
  - Sample analysis
  - **Data analysis**

Technical variability

Technical variability is also a problem of **management**

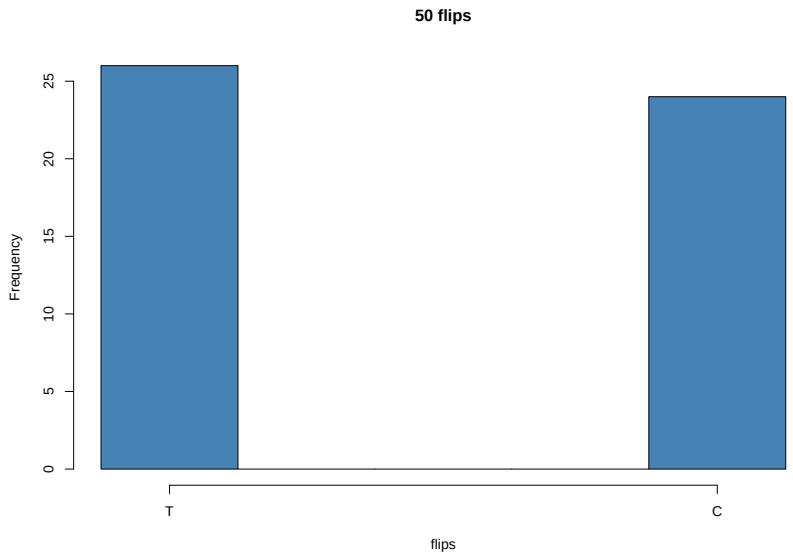
- Collecting the sample
- Handling the sample
- Storing the sample
- . . . the better the instrument, the more you see

The common shape of variability

- What we observe is the result of a "chain" of processes (e.g. gene $\rightarrow$ protein $\rightarrow$ metabolite)
- We never observe only one chain (e.g we consider *many* people with similar metabolism)
- The fact that we have noise on the "chain" produces variability in the output
- This variability has a bell shaped profile, which is more often than not **Gaussian**

Coin toss

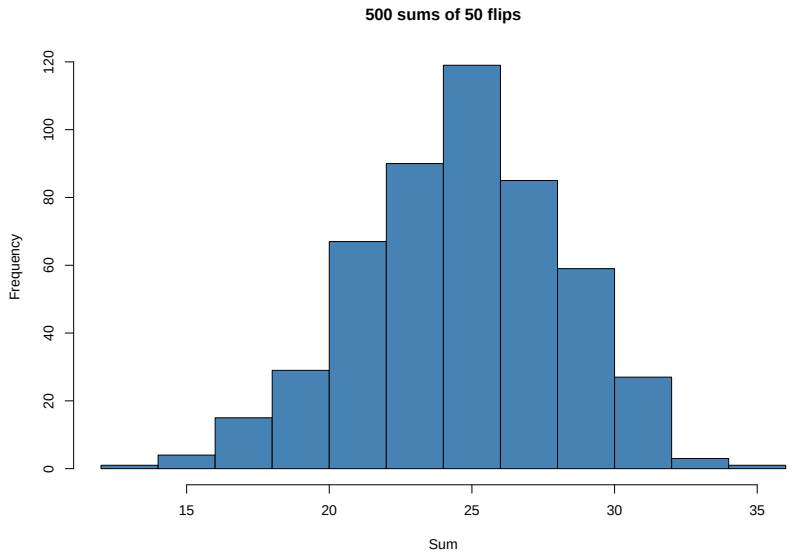
The distribution of the outcomes of 50 tosses of the same coin



This “biological” process is clearly non normally distributed, but has variability

Sum of 50 coin tosses

Suppose that now my "biological" process is the result of the sum of 50 coin tosses where T counts as 1 and C as 0. What is the distribution of the results of 500 sums?



Here the "biological" process yields normally distributed data!

The empirical low of chance

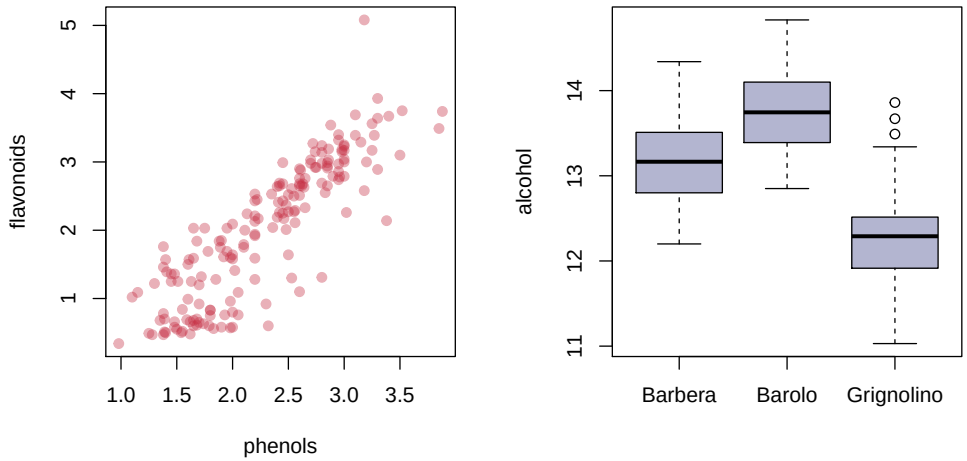
The Law

In a sequence of experiments performed on the same conditions the relative frequency of a phenomenon gets closer to the probability of the phenomenon itself ... and the goodness of this approximation improves as the number of experiments increases.

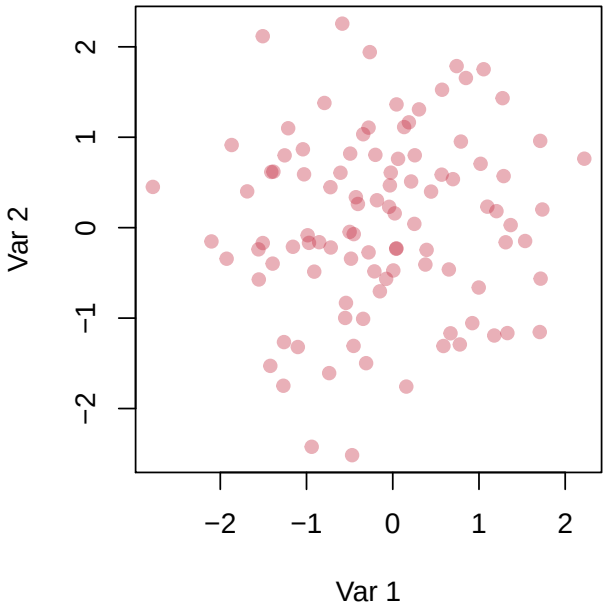
Data Analysis in a Nutshell

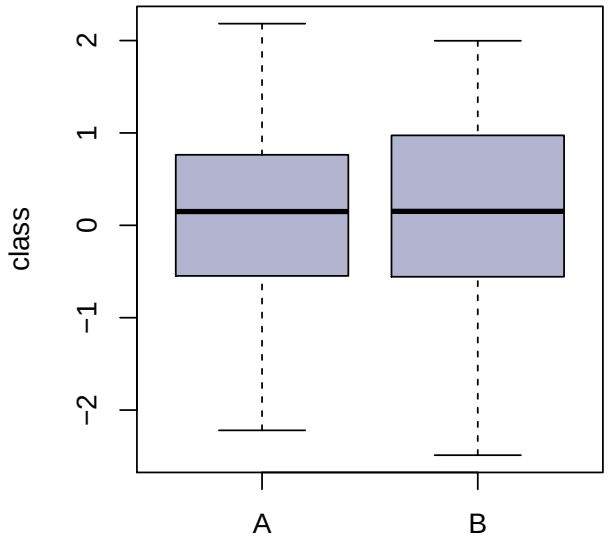
- Highlight the presence of **organization** inside complex datasets
- . . . trying to measure with which ***confidence*** one can say that this organization is true at the population level (inference!)

Is what I'm observing ***true beyond my sample***?

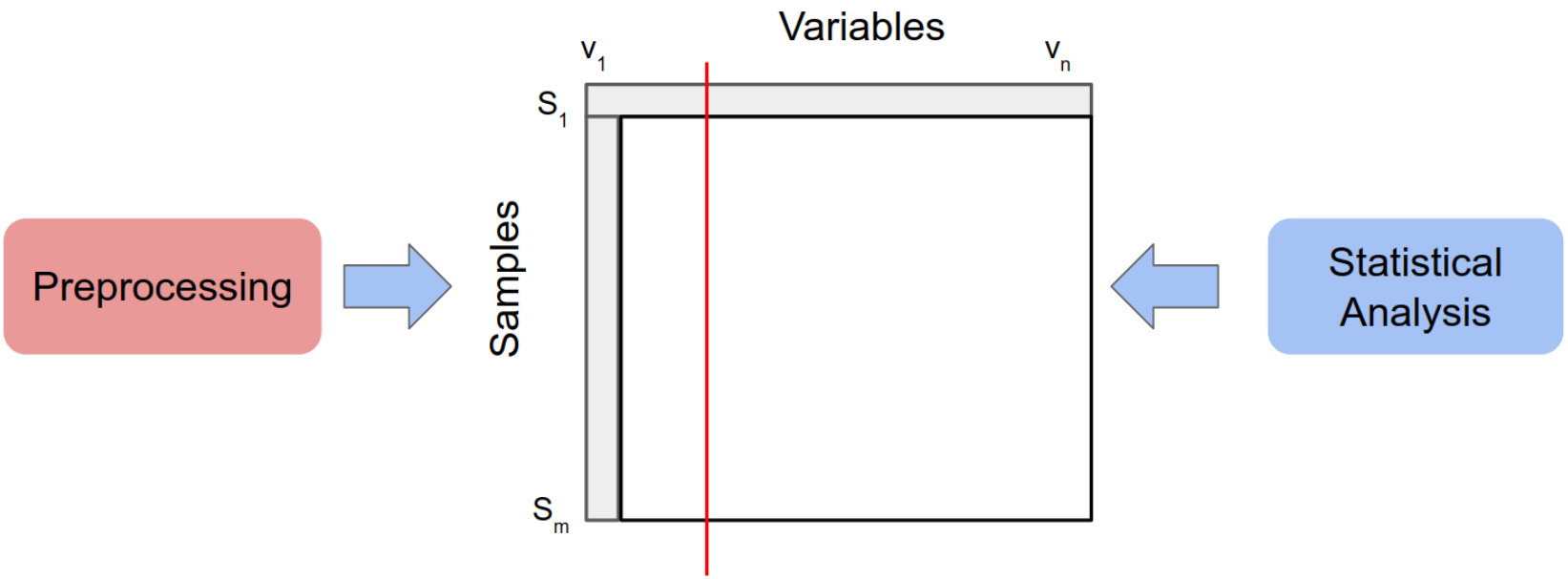


No organization . . .





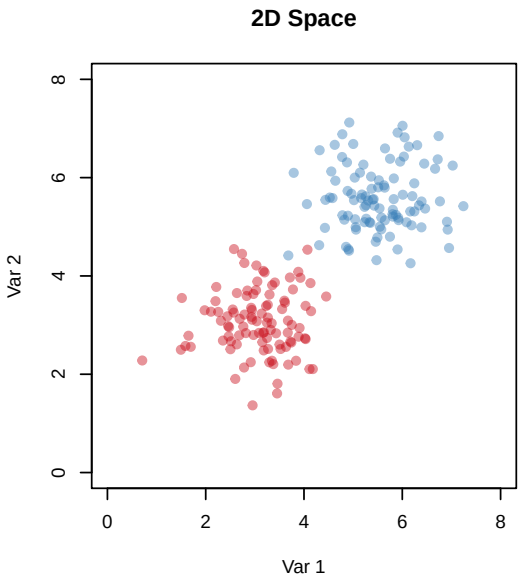
Data Matrix

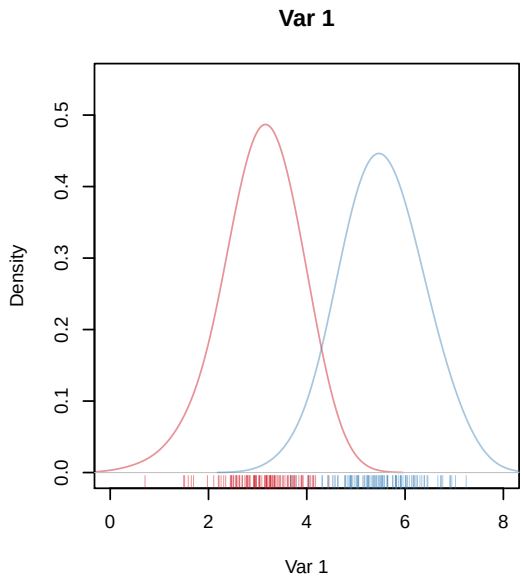

Variables
v_1
v_n
S_1
Samples
S_m
Preprocessing
Statistical Analysis

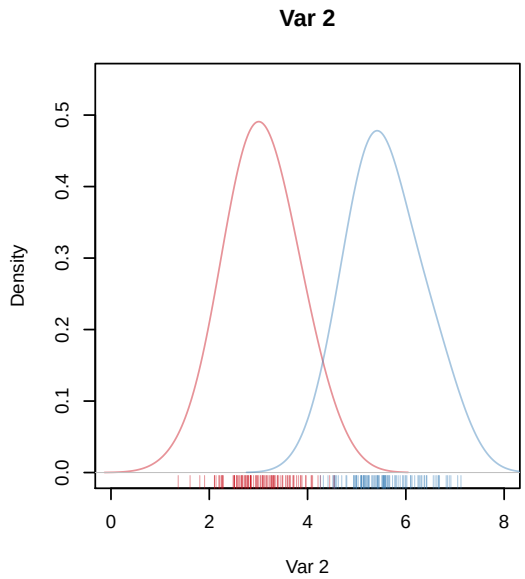
Multivariate vs Univariate

- **Univariate approach**: each experimental variable is analyzed/visualized autonomously
- **Multivariate approach**: Each sample (observation) is a point in the n dimensional space of the variables. E.g If we measure three properties the space is three dimensional

Why Multivariate







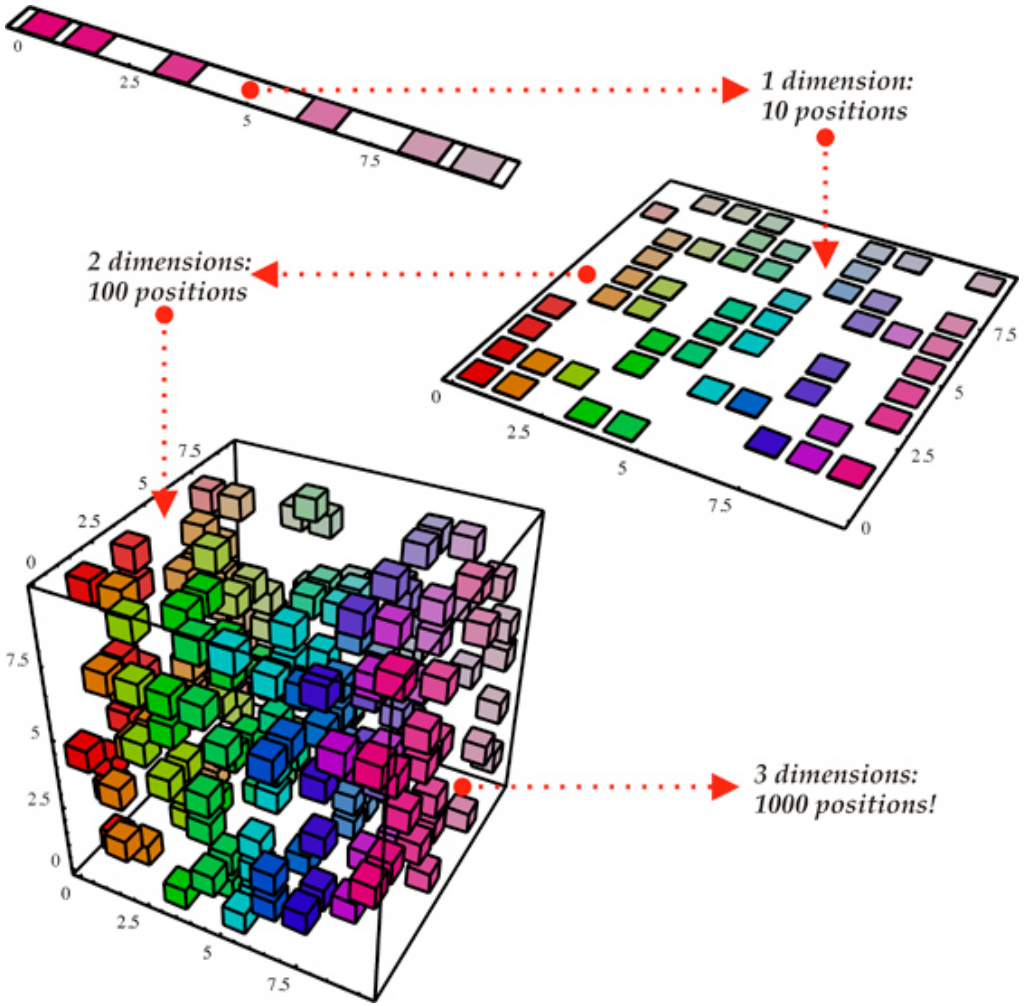
How Big is the space ?

- Untargeted metabolomics $\sim$ 1000/10000 variables/dimensions
- Targeted proteomics $\sim$ 1000/3000 variables/dimensions
- Targeted metabolomics $\sim$ 100/200 variables/dim
- NGS, Metagenomics, spectroscopy ...

How Many Samples?

Most experiments are performed on 10-100 samples. This is the number of points in the multidimensional space

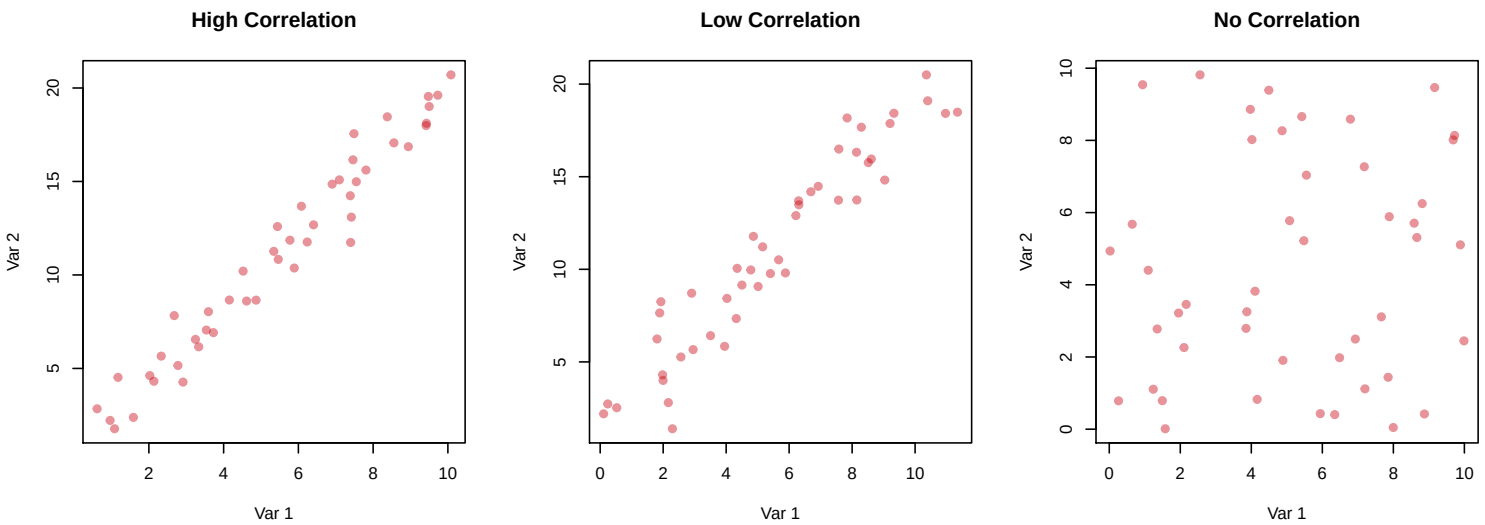
The course of dimensionality



Variable Dependence

- The variables we measure are **not independent** (e.g. network of genes, associated proteins, ecc . . .)
- The effective size of the space occupied by the samples is smaller than the number of variables
- This is equivalent to say that the cloud of samples only populates a limited part of the available multivariate space

What I mean with not independent



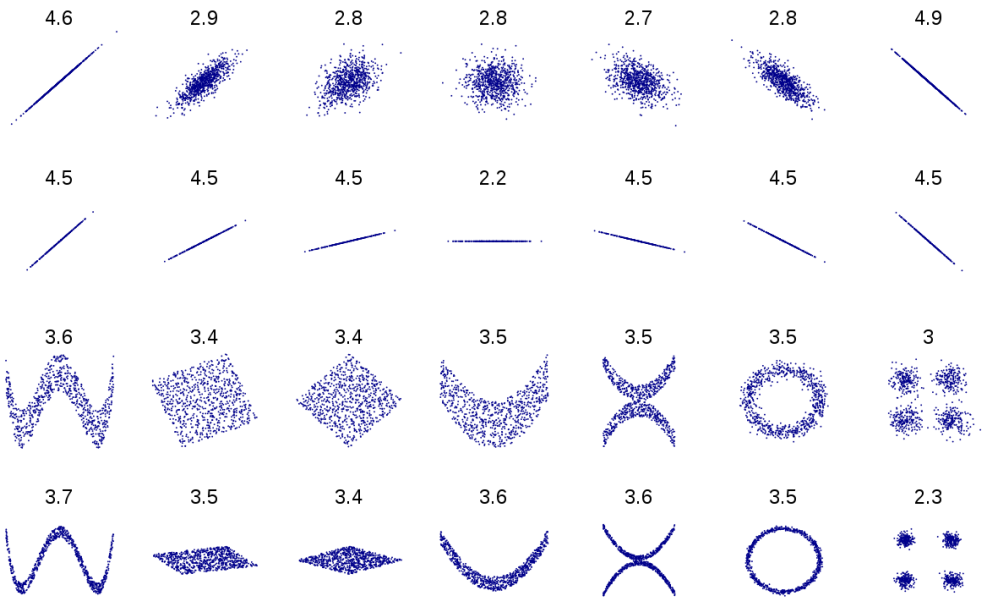
In presence of dependence between the variables the samples are occupying a smaller part of the available space

Type of variable dependence

- "Analytical"/chemical
- Biological
- ...

Measuring variable dependence

- **Pearson correlation**: measure the strength of a *linear relationship*
- **Rank based measures**: measure the strength of a monotone association
- **Mutual Information**: measures, ideally, the strength of any form of association



Conservation of misery

The data analyst dilemma

$$\text{Data} \longleftrightarrow \text{Complexity} \longleftrightarrow \text{Knowledge}$$

- parametric vs non parametric tests
- variable association measure
- powerful machine learning approach
- ...

The positive implication of variable dependence

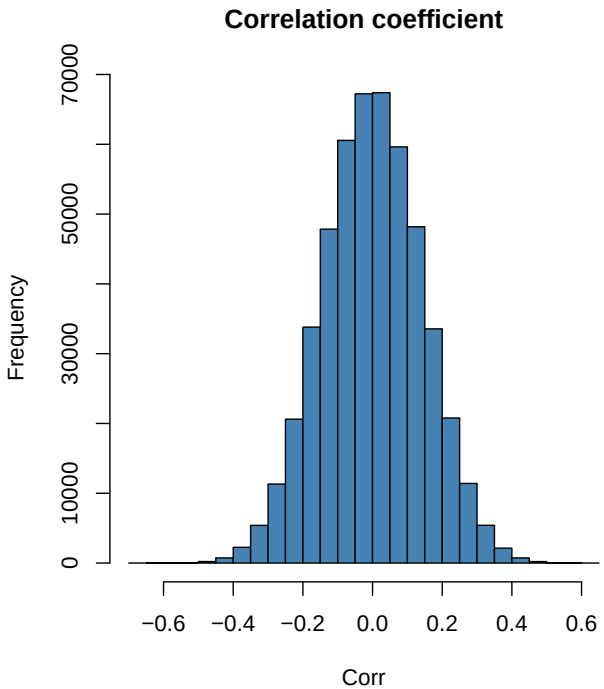
- Multivariate approaches are potentially more "effective" since they use explicitly variable association
- We can make reasonably **good science** even from experiments with **relatively small number of samples**
- Buying a more expensive instrument can be useful . . . :-)

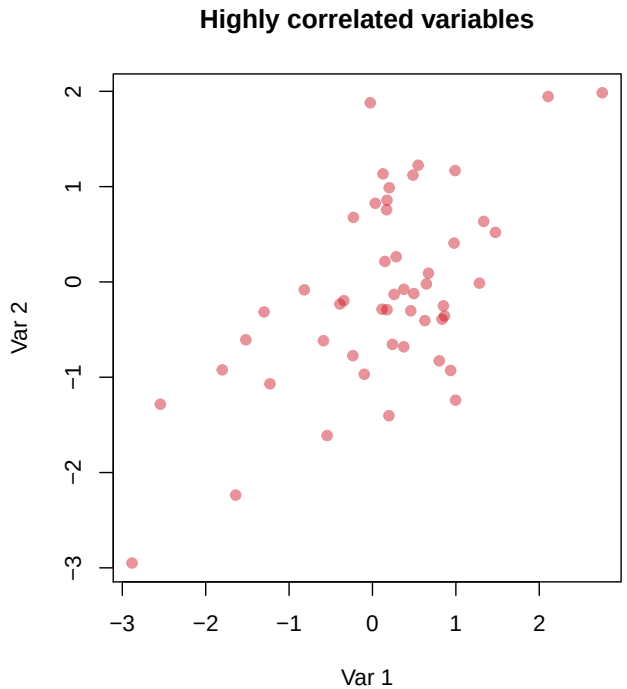
There are also negative implications . . .

- With small number of samples we'll always find **spurious organization**
- This will always lead to FALSE DISCOVERIES

Null dataset

- 50 samples
- 1000 variables (analytes, metabolites, wavelengths, . . .)
- Filled with only random numbers





False Discoveries

False Discovery

- Any form of organization which is visible in my dataset, but cannot be generalized at the population level
- It is **not an error**, but an inherent result of chance during **sampling** ... we can see it a sort of "bad luck"
- In presence of high variability sampling issues can be determinant!
- I need to **validate** my outcomes (new data, new experiments)

Association and Causation

The NEW ENGLAND JOURNAL of MEDICINE

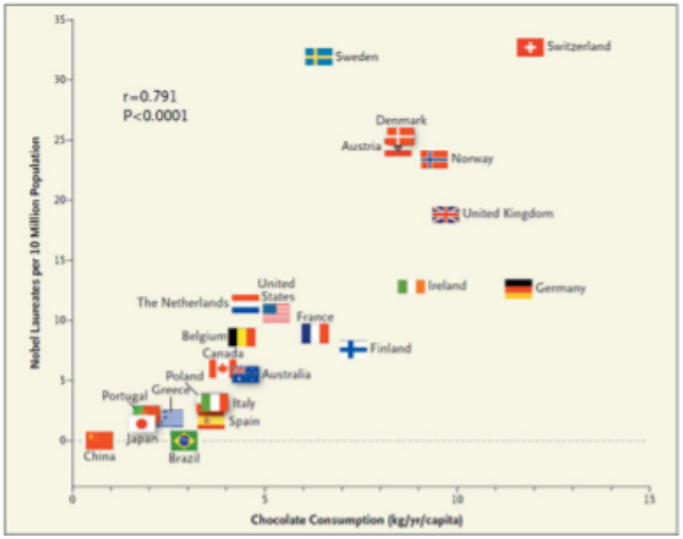
Chocolate Consumption, Cognitive Function, and Nobel Laureates

Franz H. Messerli, M.D.

N Engl J Med 2012; 367:1562-1564 October 18, 2012 DOI: 10.1056/NEJMon1211064

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)



Causality, association, observational studies

- A clear cut answer requires an **experiment**
- In a health context we often have to rely on **observational studies**
- It is important to **validate** (new samples, model systems, . . .)
- Causation it is not necessary for prediction