UNIVERSITÉ
DE GENÈVE

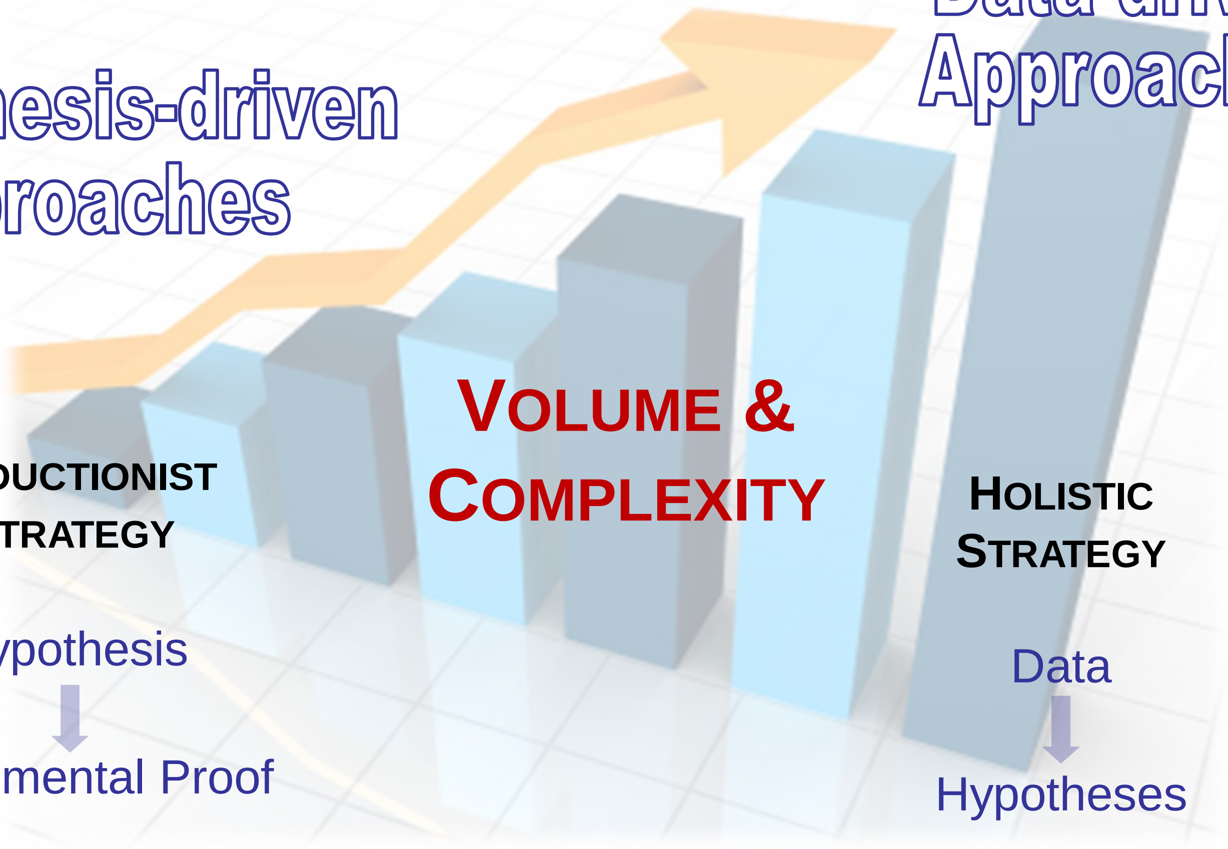
ISPSO
INSTITUT DES SCIENCES
PHARMACEUTIQUES
DE SUISSE OCCIDENTALE

SIB
Swiss Institute of
Bioinformatics

Multiomics
Data Analysis
and Integration

The Omics Data Explosion

Modern scientific technologies are able to generate massive datasets to describe specific phenotypes illustrating a biological phenomenon



Hypothesis-driven Approaches

Data-driven Approaches

Reductionist Strategy

Hypothesis

↓

Experimental Proof

Volume & Complexity

Holistic Strategy

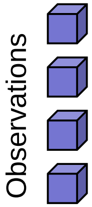
Data

↓

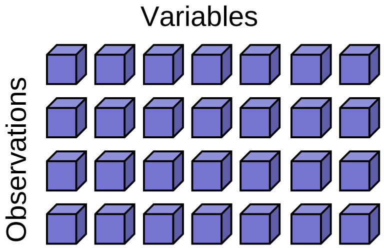
Hypotheses

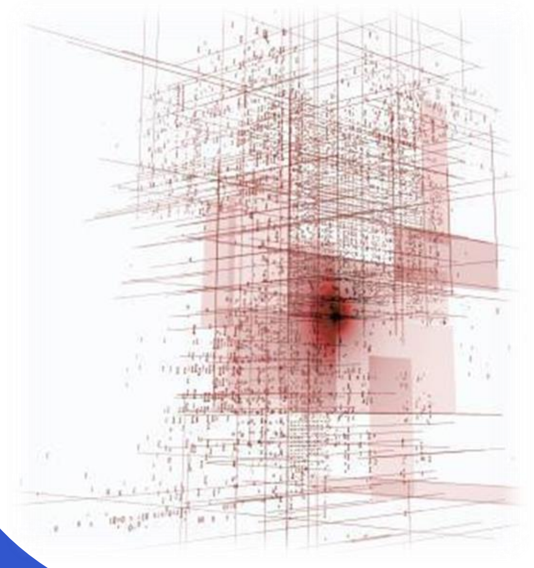
Data Structures

- One-way data is a vector, with a single data value for each element of the single dimension (n)



- Two-way data is a matrix, with a single data value for each element of two separate dimensions (n,p)



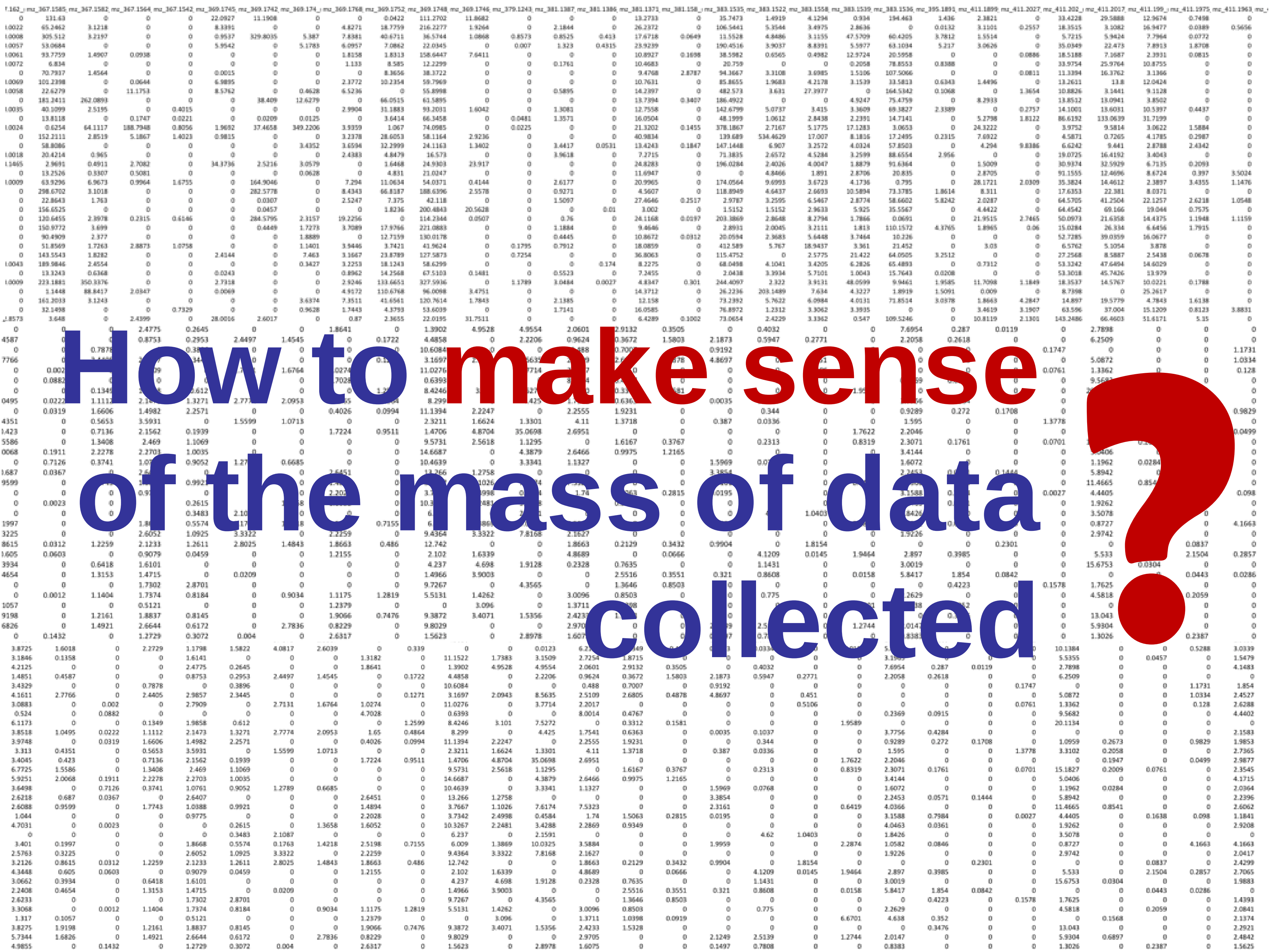


High dimensionality (n << p)
Multicollinearity between variables
Missing values
Biological/analytical variability

Adding extra dimensions leads to an exponential increase of the hypothesis space size
→ Relevant hypotheses become harder to find

!

How to make sense of the mass of data collected ?



Knowledge Discovery In Omics

Analytics

Data Production
- ✓ Sample preparation
- ✓ Data acquisition

Data Processing
- ✓ Signal extraction
- ✓ Filtering
- ✓ Normalisation
- ✓ Annotation

Signal

Data Mining

Knowledge

Information Content

Biological Interpretation
- ✓ Extract relevant information
- ✓ Link to existing knowledge
- ✓ Biological validation

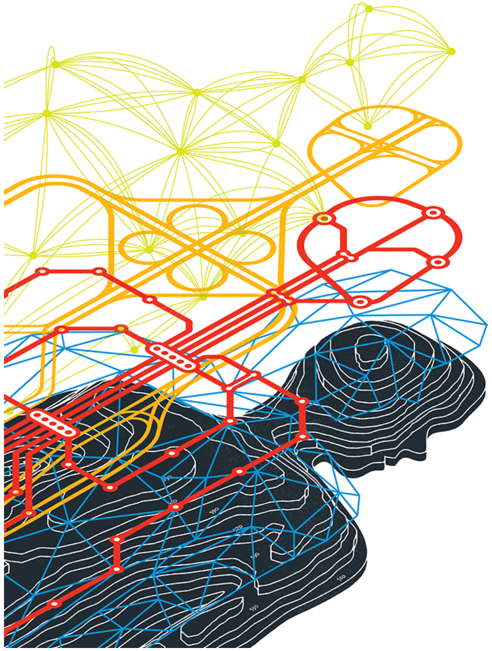
Chemometrics

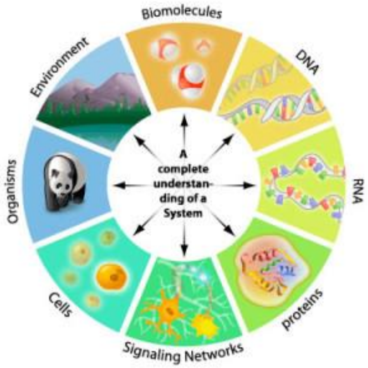
Multivariate Analysis
- ✓Exploration
- ✓ Classification
- ✓ Pattern Recognition
- ✓ Variables contribution

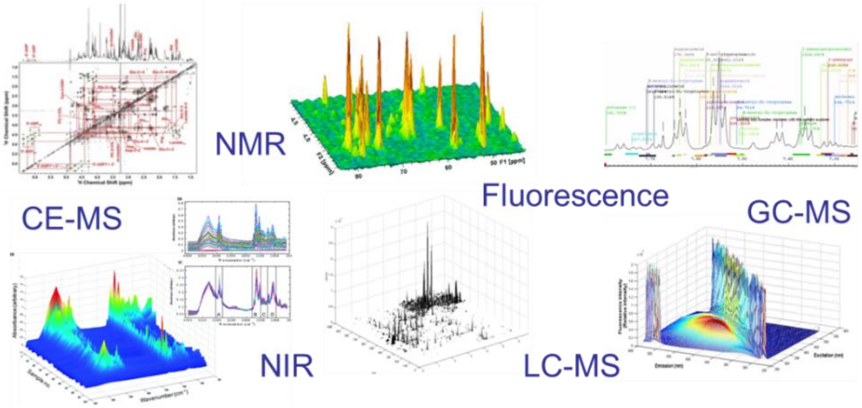
Bioinformatics

Multiple Data Sources Omics

- ✓ Different biological scales
  - ✓ Cell/tissue/organism
  - ✓ Systems biology
- ✓ Different stages of a process
  - ✓ Dose
  - ✓ Toxicity
  - ✓ Disease progression
- ✓ Different analytical techniques
  - ✓ Heterogeneous data
  - ✓ Separation or spectral methods







NMR

Fluorescence

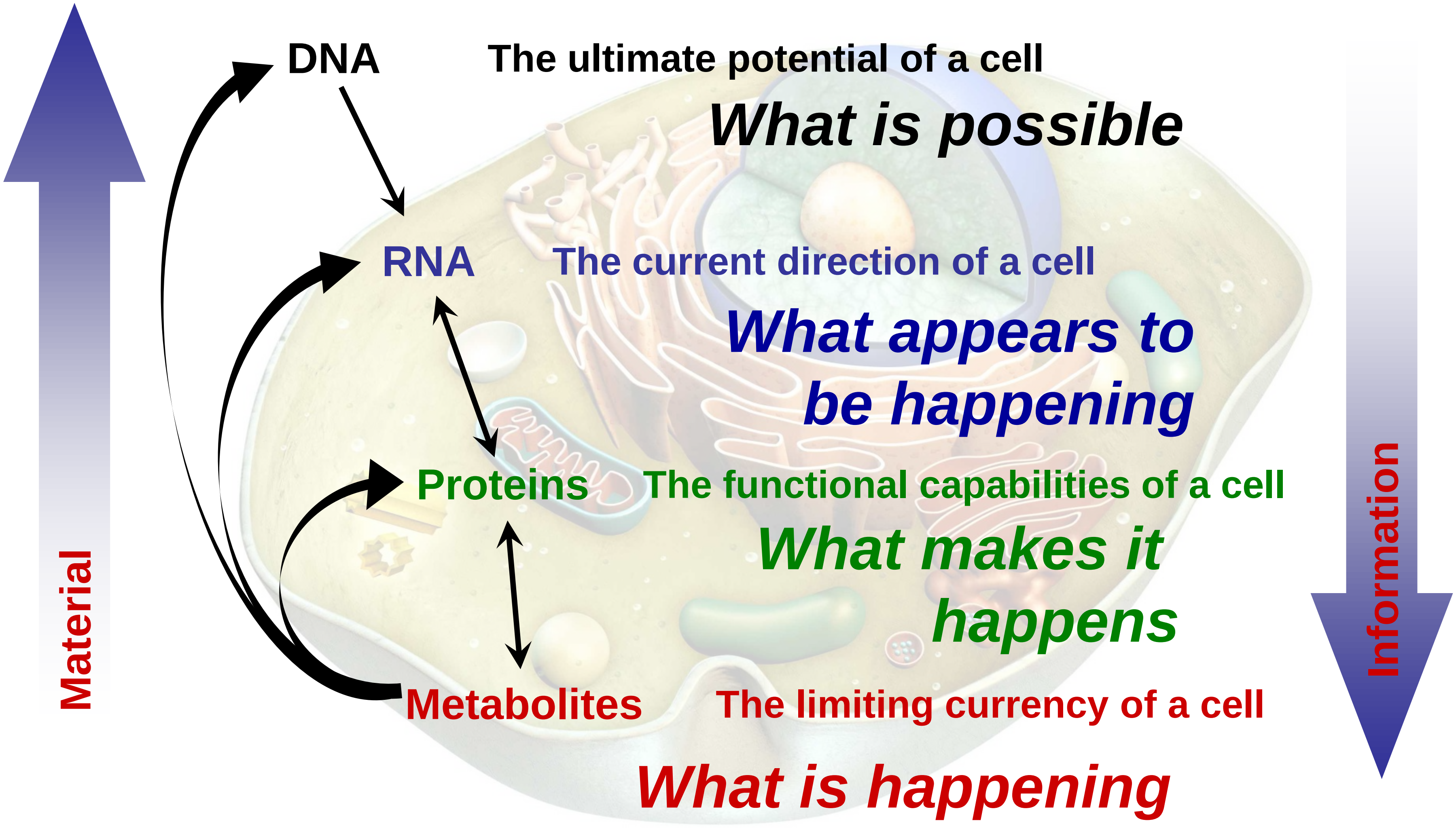
CE-MS

GC-MS

NIR

LC-MS

MultiOmics & Systems Roles



DNA **The ultimate potential of a cell**

What is possible

RNA **The current direction of a cell**

What appears to be happening

Proteins **The functional capabilities of a cell**

What makes it happens

Metabolites **The limiting currency of a cell**

What is happening

Embracing Complexity

How does a complex system work?

✗ Examine separately springs, gears, shafts, etc. how they fit together

or

✓ Consider all the elements at once and how they fit and interact together

Data Integration

Multigroup Analysis

Multiview Analysis

Data Fusion

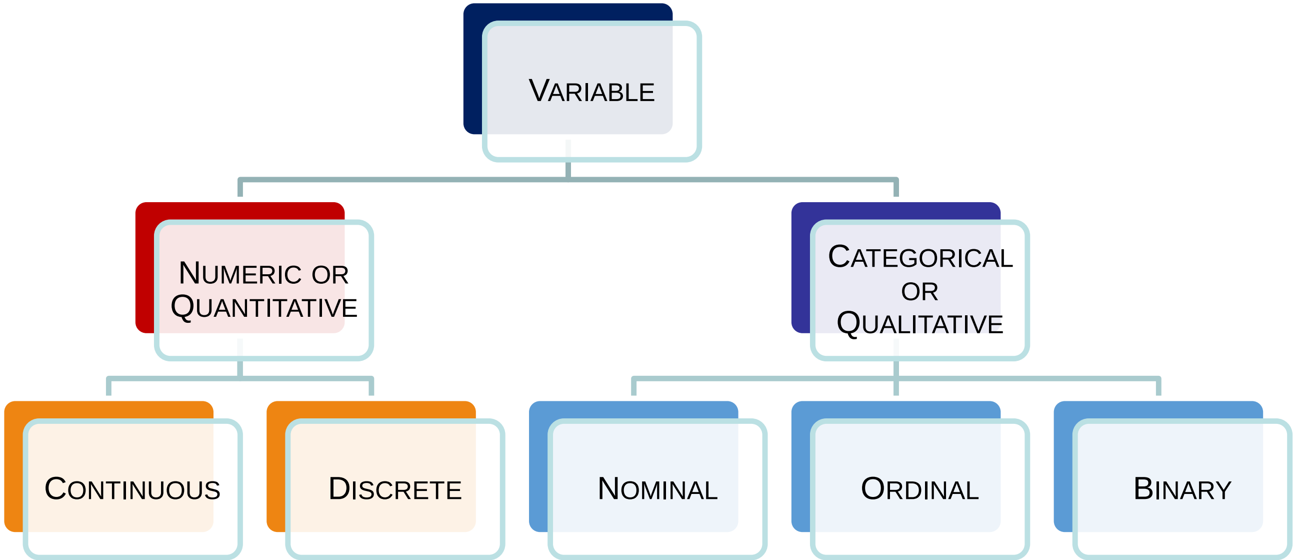
?

Multiset Analysis

Multitable Analysis

Multiblock Analysis

Nature Of The Data



QUANTITATIVE

- Continuous: numeric variables that can take any value between a certain set of real numbers
- Discrete: numeric variables that only consist of integers

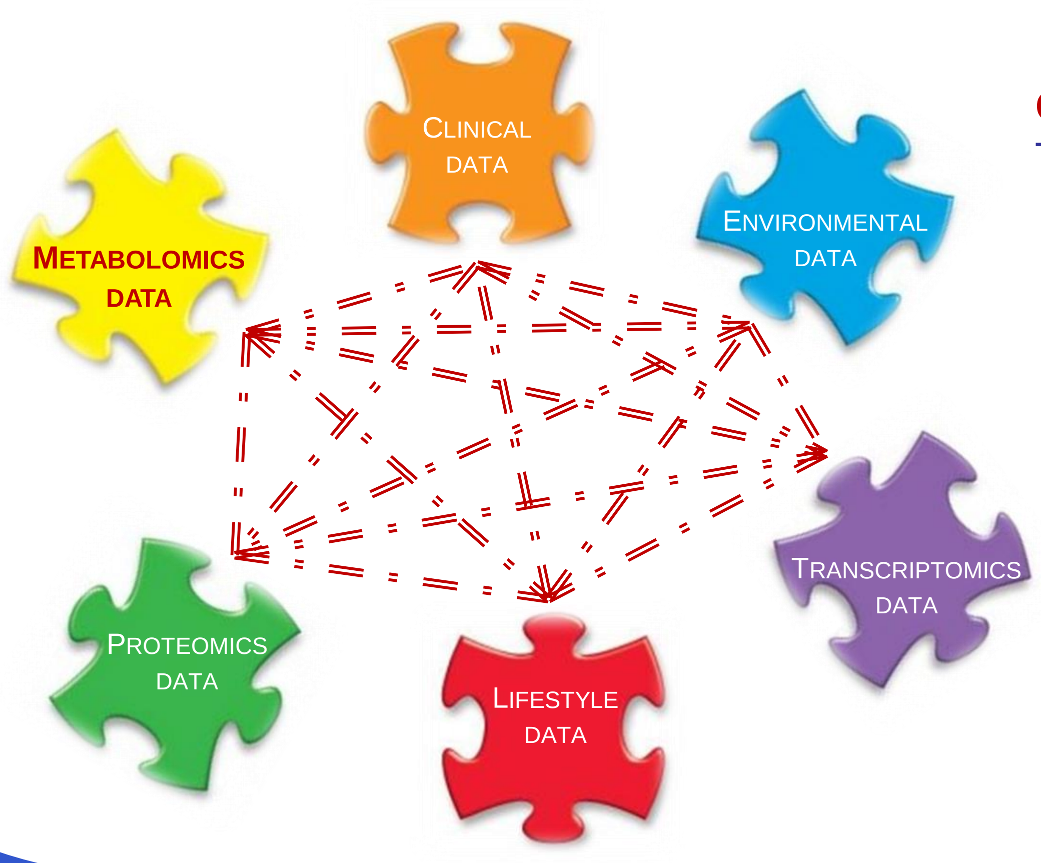
QUALITATIVE

- Nominal: categorical variable that cannot be ranked
- Ordinal: categorical variable that can be ranked
- Binary: categorical variable that is either true or false

Data Homo/Heterogeneity

Homogeneous data: data blocks all measured on the same scale
e.g. quantitative data

Heterogeneous data: data blocks measured on different scales
e.g. quantitative, ordinal, qualitative, binary



Continuous
Transcripts/Proteins/Metabolites

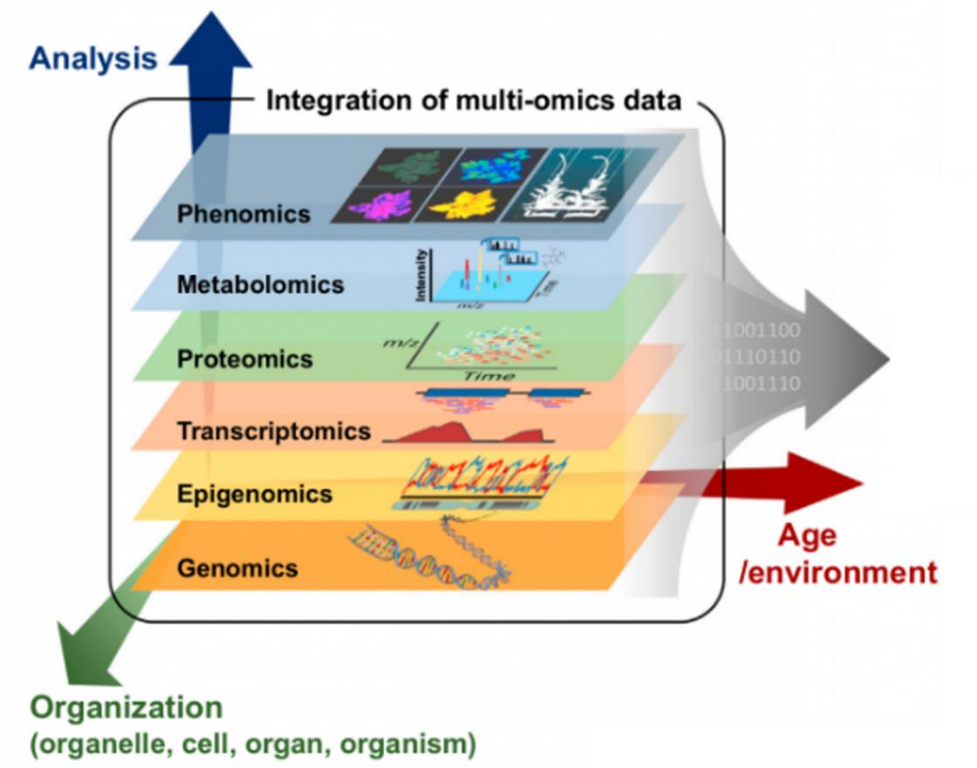
Discrete
Clinical scores

Ordinal
Lifestyle

Nominal
SNPs
Clinical data

Binary
Methylation
Clinical conditions

MultiOmics Data Integration

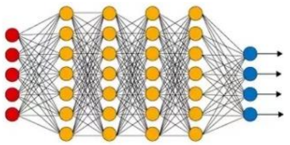


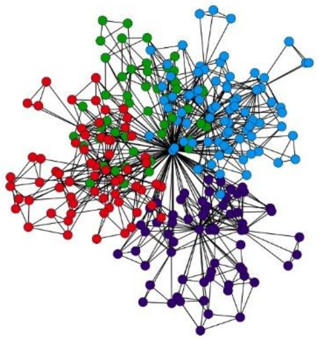
Aims

- Molecular signatures
- Biological processes
- Mechanistic insights
- Interplay between layers
- Holistic view

Methods

- Matrix factorization
- Network-based approaches (multiplex, multilayer)
- Bayesian approaches
- Machine learning (embeddings)





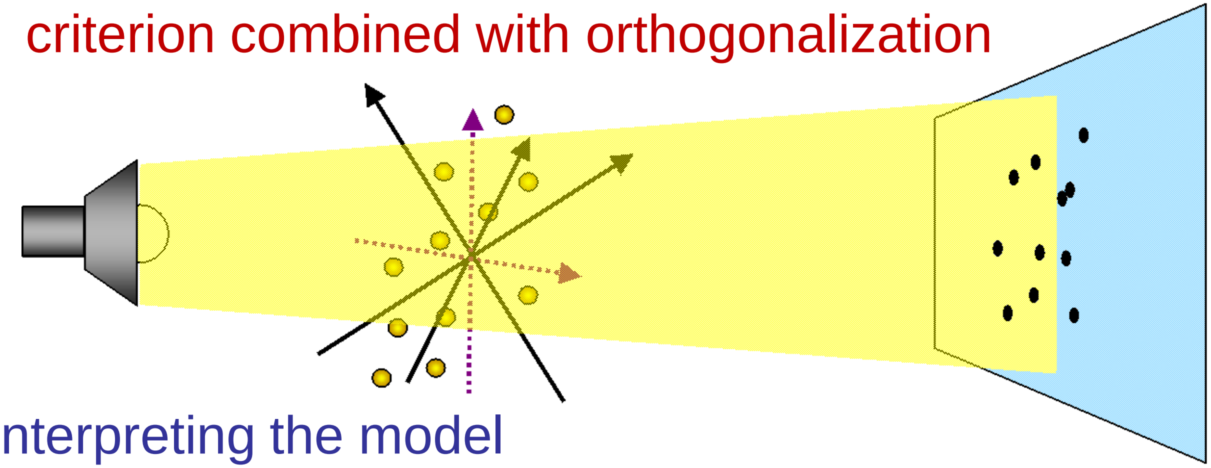
Methods Based On Components

Data High Dimensionality
Variables Correlations
Biological Variability
Experimental Noise

Projection methods

- ✓ analyze datasets of high dimensionality
- ✓ provide knowledge about systems
- ✓ find unsuspected relationships
- ✓ summarize the data with a small number of factors

Linear combination of the initial variables
→ maximization/minimization some criterion combined with orthogonalization



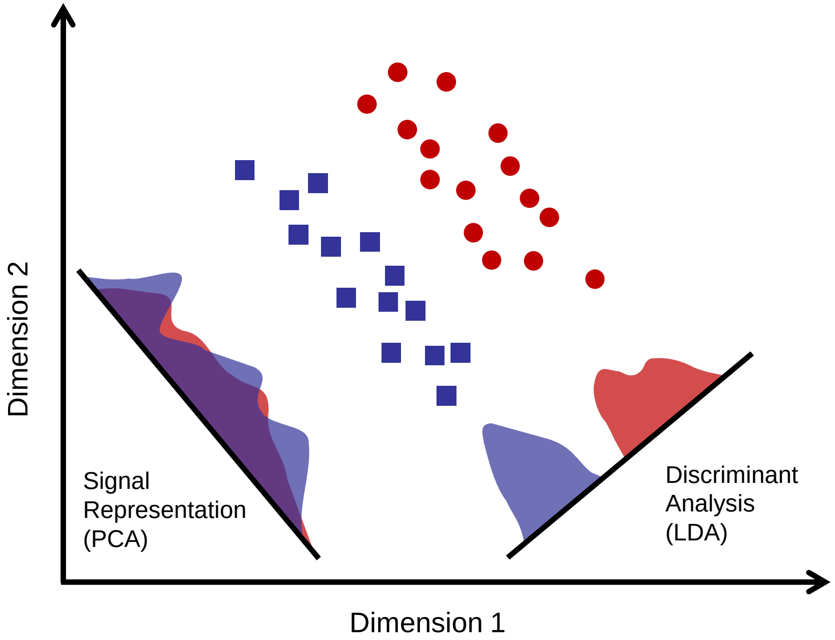
Interpreting the model

- Visualize the samples' distribution
- Visualize correlations between variables

Model Objectives

Search for a subspace providing an effective representation of the data
Build a multivariate model (PCA, PLS, OPLS)
Analyse the model

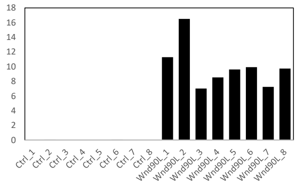
- ✓ Search for patterns/groupings
- ✓ Prediction performance



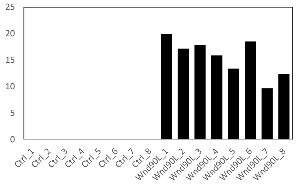
Evaluate the variables' contributions
Rank the variables

Find the most relevant biomarkers

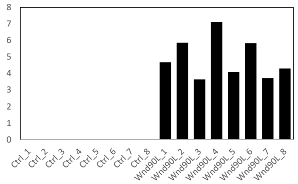
m/z 225.1125 @ 1.3 min

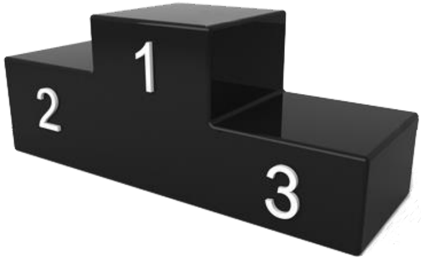


m/z 209.1167 @ 2.4 min



m/z 322.2011 @ 3.0 min





→ MOLECULAR SIGNATURES