Statistics…. introduction

Arnaud Gloaguen, Jimmy Vandel, Guillemette Marot

Swiss Institute of Bioinformatics

INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Statistics…. introduction

Arnaud Gloaguen, Jimmy Vandel, Guillemette Marot

inspired from Carl Herrmann (Heidelberg University), Delphine Potier (CIML, CNRS Marseille), Sébastien Déjean (IMT, Université de Toulouse) slides…

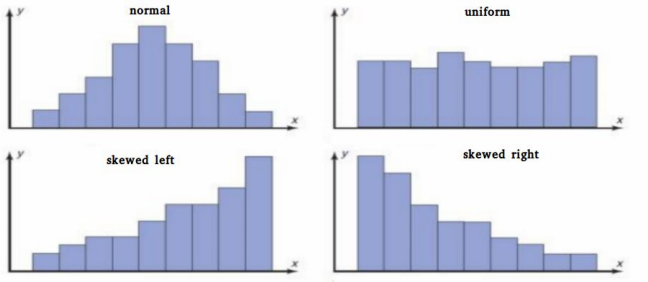
Swiss Institute of Bioinformatics

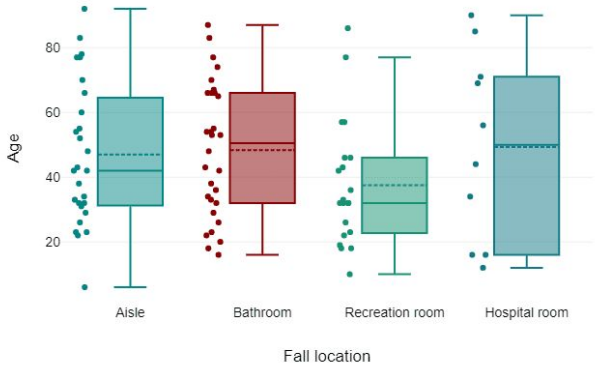
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

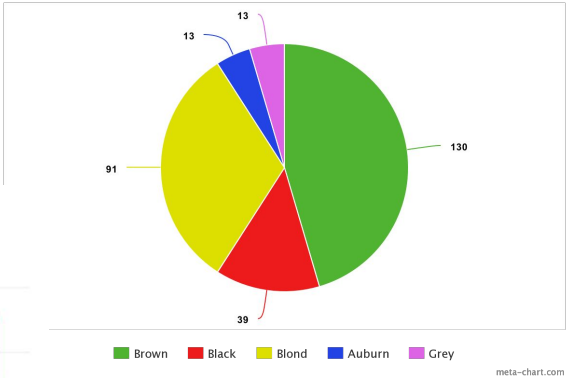
Statistics…. some vocabulary

→ Doing statistics… for what ?

- **descriptive statistics** : describe the characteristics or features of a dataset (sample/population)
  - distribution, skewness, outliers
  - mean/median/mode
  - variability (range/variance/standard deviation)





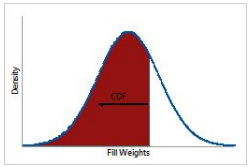


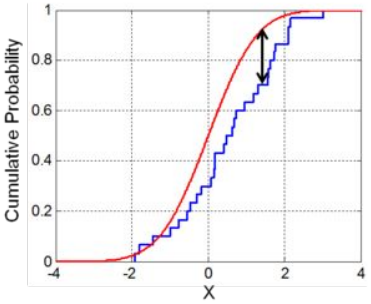
Statistics…. some vocabulary

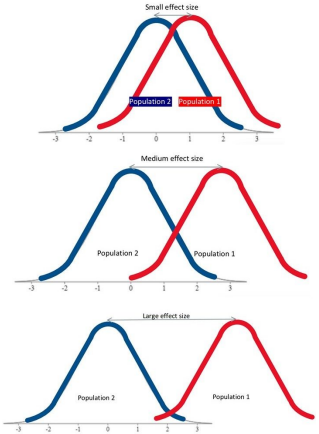
→ Doing statistics… for what ?

- **descriptive statistics** : describe the characteristics or features of a dataset (sample/population)
  - distribution, skewness, outliers
  - mean/median/mode
  - variability (range/variance/standard deviation)

- **inferential statistics** : draw meaningful conclusion about the dataset, and possibly generalize to a larger population
  - hypothesis testing





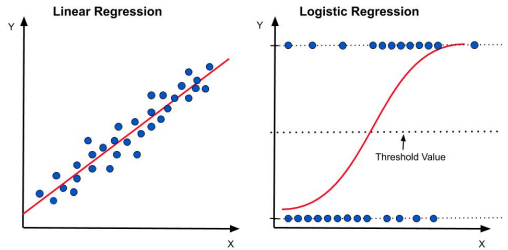


Statistics…. some vocabulary

→ Doing statistics… for what ?

- **descriptive statistics** : describe the characteristics or features of a dataset (sample/population)
  - distribution, skewness, outliers
  - mean/median/mode
  - variability (range/variance/standard deviation)

- **inferential statistics** : draw meaningful conclusion about the dataset, and possibly generalize to a larger population
  - hypothesis testing
  - modeling relationship (linear/logistic regression…)

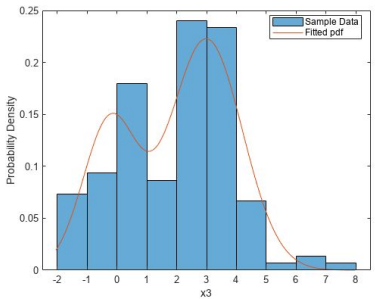


Statistics…. some vocabulary

→ Doing statistics… for what ?

- **descriptive statistics** : describe the characteristics or features of a dataset (sample/population)
  - distribution, skewness, outliers
  - mean/median/mode
  - variability (range/variance/standard deviation)

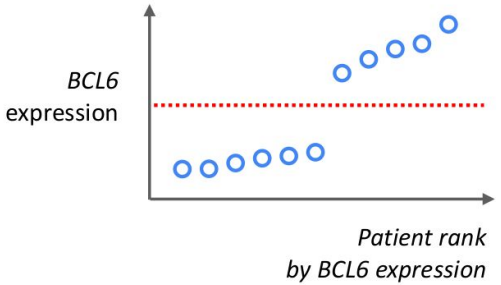
- **inferential statistics** : draw meaningful conclusion about the dataset, and possibly generalize to a larger population
  - hypothesis testing
  - modeling relationship (linear/logistic regression…)
  - probability estimation
  - confidence interval
  - …



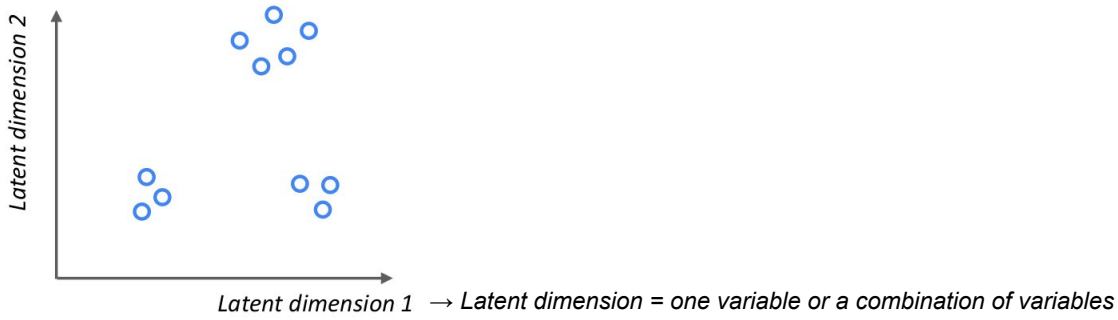
Statistics…. some vocabulary

→ Doing inferential statistics… considering what ?

- **univariate statistics** : analyze only one (‘uni’) variable at a time
  → for descriptive or inferential purposes



- **multivariate statistics** : analyze more than one (‘multi’) variables at a time



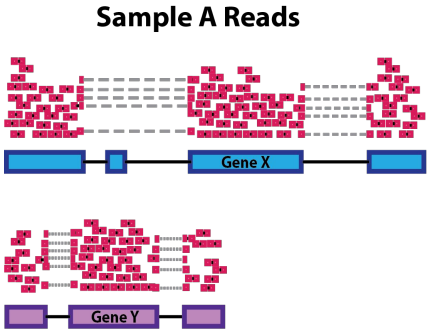
→ *Latent dimension = one variable or a combination of variables*

Statistics…. some vocabulary

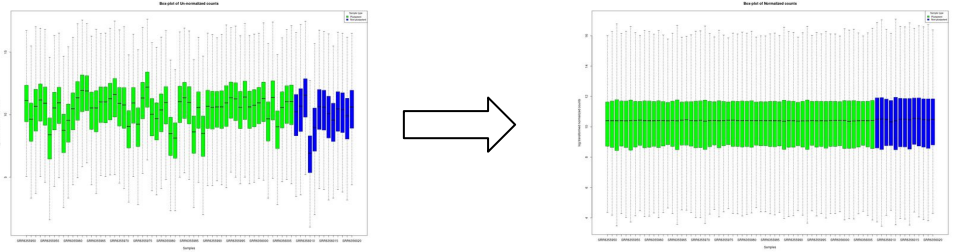
→ Doing multivariate inferential statistics… on what ? … on normalized data

- **Normalization** is a process designed to identify and correct “technical/experimental” biases without removing biological signal.
  Sources of bias: batch effect (lab condition, platform…), sequencing depth, sample quantity…

  - **within-sample normalization** :
    e.g. normalize expression of all genes within sample *A*



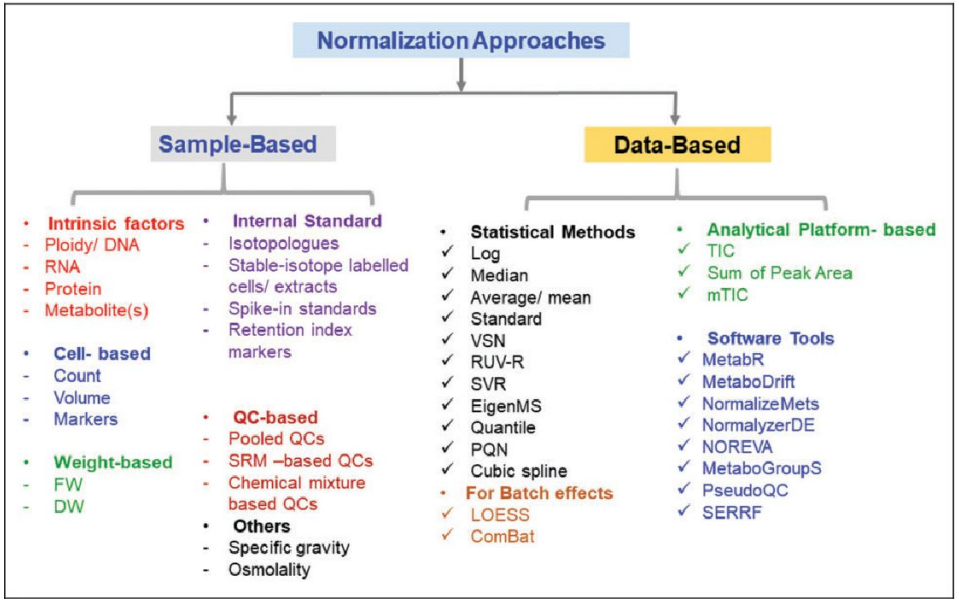
  - **between-sample normalization** :
    e.g. normalize expression of all genes between samples



Statistics…. some vocabulary

→ Doing multivariate inferential statistics… on what ? … on normalized data

- **Normalization strategies** : many exist, none of them is better than another, but guidelines/comparisons exist, some are omic dependant…

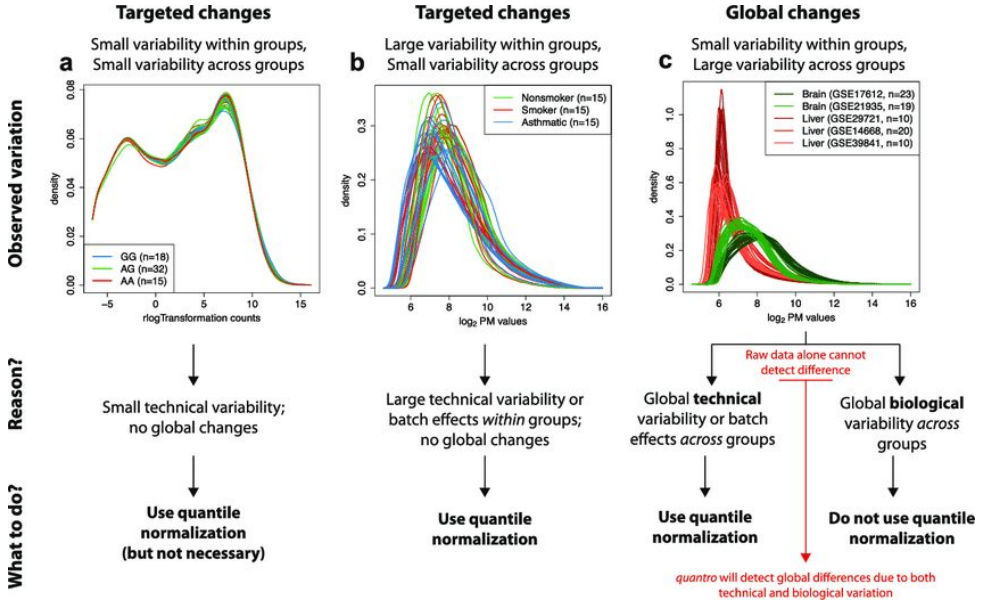


Misra BB. **Data normalization strategies in metabolomics**: Current challenges, approaches, and tools. *European Journal of Mass Spectrometry*. 2020;26(3):165-174. doi:10.1177/1469066720918446

Statistics…. some vocabulary

→ Doing multivariate inferential statistics… on what ? … on <u>normalized</u> data

- **Normalization strategies** : many exist, none of them is better than another, but guidelines/comparisons exist, some are omic dependant, and <u>should not be used automatically</u>!



Hicks, S.C., Irizarry, R.A. ***quantro: a data-driven approach to guide the choice of an appropriate normalization method.*** *Genome Biol 16, 117 (2015). doi:10.1186/s13059-015-0679-0*

Statistics…. some vocabulary

→ Doing multivariate inferential statistics on normalized data without missing values.

- missing values imputation is not mandatory, depends on downstream analysis and you can also remove corresponding samples/variables.
- if necessary, imputation strategy should be chosen carefully :

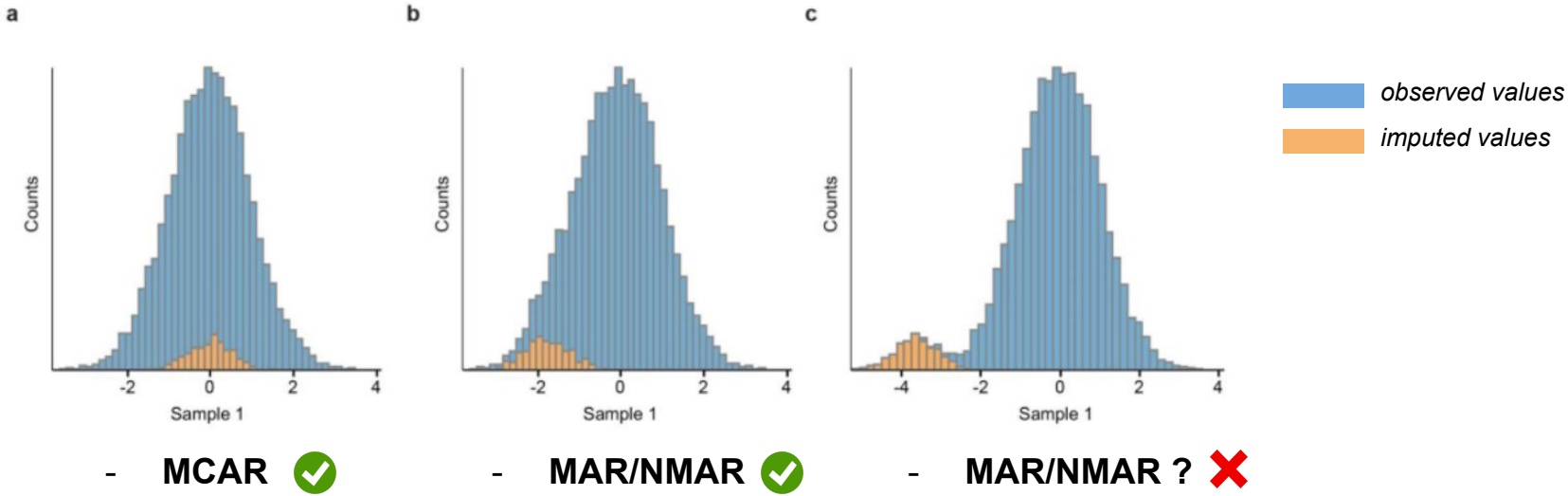
  - **missing completely at random (MCAR)** :
    → caused by external factor independent from observed data

  - **missing at random (MAR)**
    → caused by external fully known dependant factor, and so can be controlled

  - **not missing at random (NMAR)**
    → caused by external unknown dependant factor
    → due to the observed value (e.g. technical detection limits)

Statistics…. some vocabulary

→ Doing multivariate inferential statistics on normalized data without missing values.



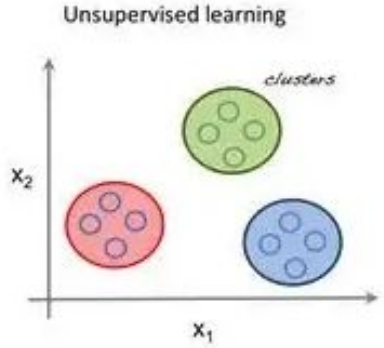
- **MCAR** ✓
- **MAR/NMAR** ✓
- **MAR/NMAR ?** ✗

Statistics…. some vocabulary

- **Unsupervised learning**

  → find hidden patterns, analyze and organize unlabelled samples

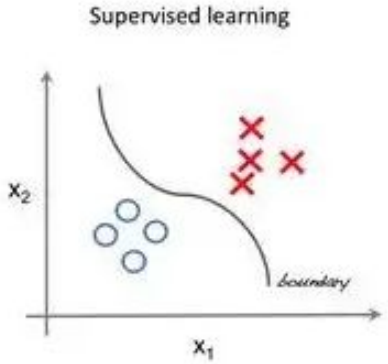
  e.g. clustering, dimension reduction, density estimation



- **Supervised learning**

  → use labelled samples and previous outputs to guess outcomes in advance (predictive model)

  e.g. classification task (categorical/numerical), regression (numerical)

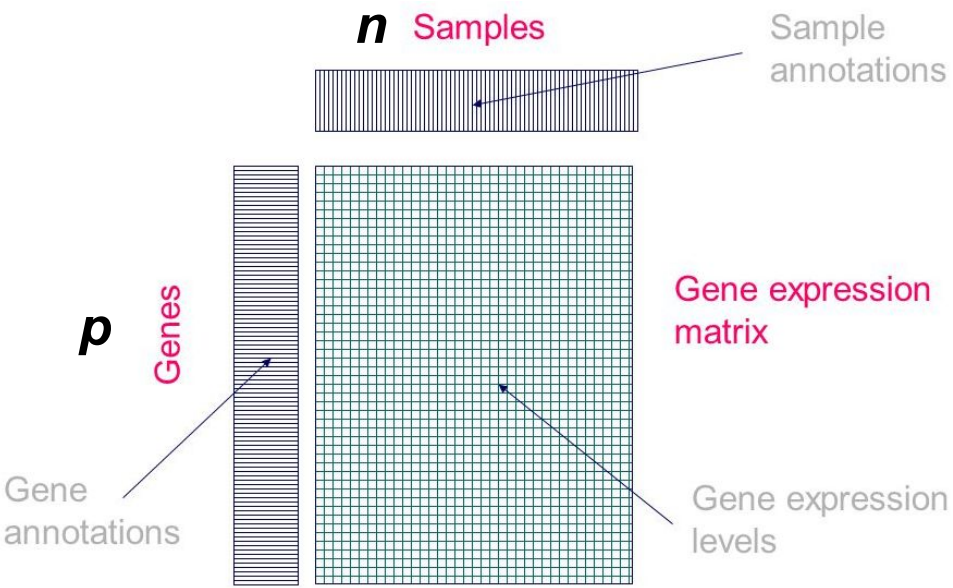


- **Semi-supervised learning**

  → only some labelled samples (not available, too expensive…)

Statistics…. some vocabulary

- **Matrix representation of data**



$$X = \overbrace{\begin{bmatrix} x_{11} & \cdots & x_{1n} \\ x_{21} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots \\ x_{p1} & \cdots & x_{pn} \end{bmatrix}}^{n} \Big\} p$$

x_{ij} : value of variable i for individual j.

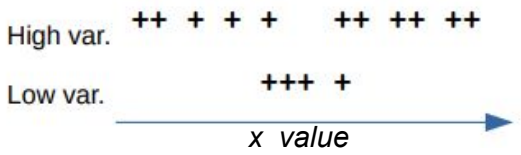
→ e.g. value of gene i for sample j

⚠ Or transposed, a n×p matrix instead of a p×n matrix !

Statistics…. some vocabulary

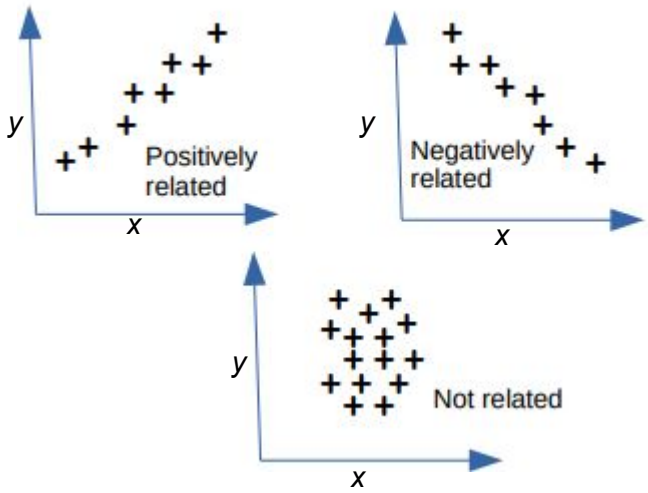
- **Variance:** indicator of spread for one variable x_i

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \text{with} \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



- **Covariance:** indicator of relationships for two variables *x* and *y*

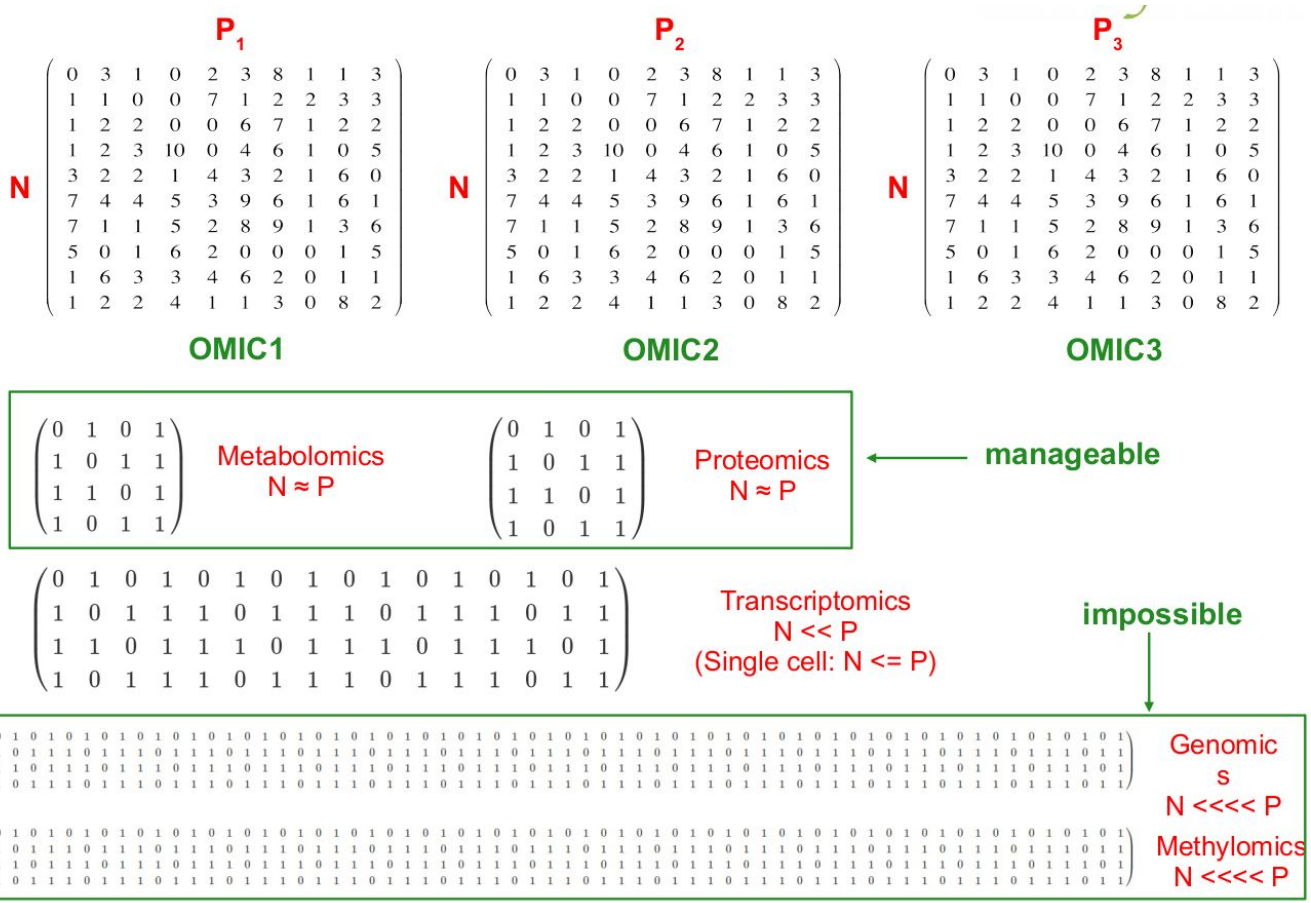
$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$



- **Correlation:** standardized covariance between -1 and 1

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \qquad \text{with} \qquad \sigma_X = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

Curse of dimensionality



P : number of features (genes, proteins, genetic variants…)

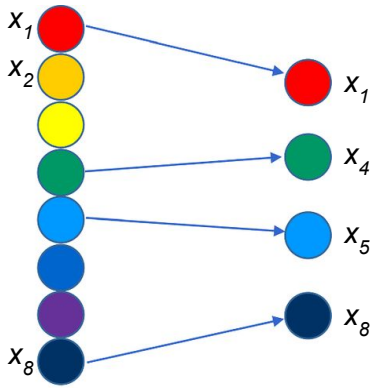
N : number of observations (samples, cells, nucleotides…)

Selection vs Extraction

- **Feature selection**

  → determine a smaller set of features minimizing (relevant) information loss

  e.g. filtering methods (correlation), recursive elimination, regularization…

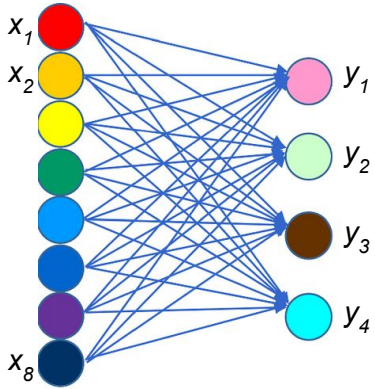


- **Feature extraction**

  → combine the input features into another set of variables in a linear or non-linear way: $y_1 = \alpha_1 * x_1 + \alpha_2 * x_2 + \alpha_3 * x_3 + \ldots.$

  e.g. **PCA**, PCoA, ICA…

  + regularization for sparse methods : sPCA, sNMF (i.e. some α_i forced to 0)



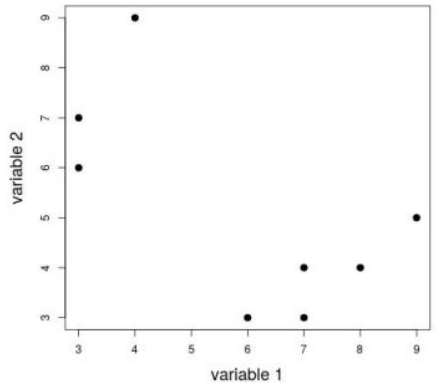
Dimensionality reduction : PCA

Problem: ***n*** samples, ***p*** quantitative variables (e.g. peptides, proteins, metabolites, mRNA, . . .)

Visualize pairwise relations by scatter plots

But when ***p*** is large ?

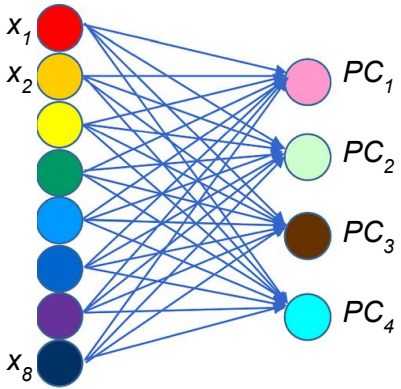
p=2

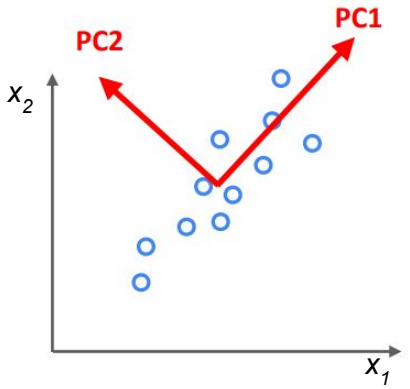


→ Need to reduce this large number of dimensions (***p***) to a smaller number of relevant variables, i.e variables which carry most of the information (or variance) of a dataset and without redundancy

PCA - Principle

Principle: *Find orthogonal axes (Principal Components) on which one can project sample to obtain a comprehensible space of reduced dimension.*





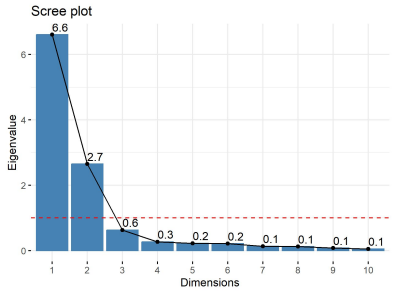
$$PC_1 = \alpha_1 * x_1 + \alpha_2 * x_2 + \alpha_3 * x_3 + \ldots.$$

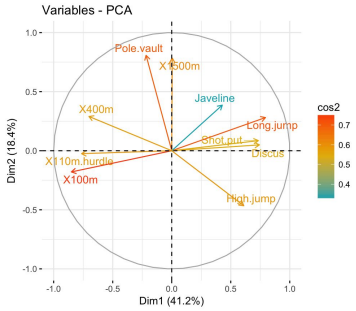
Projection is a distorting operation ⇒ we begin by looking for an axis on which the cloud of points is distorting the less possible during the projection.

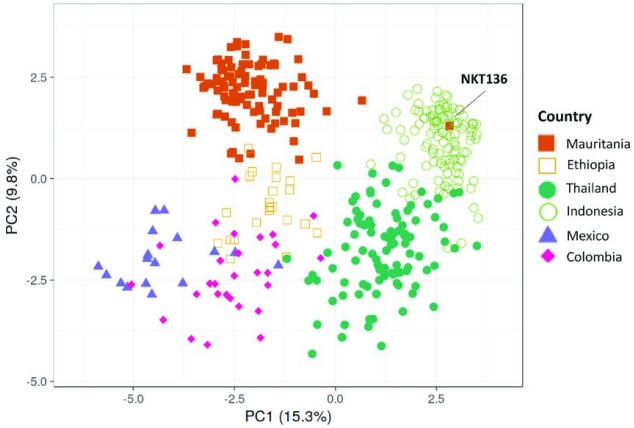
PCA - Goal

Main goal : explore the structure of the dataset to better understand the proximity between samples and detect possible bias → often used as a quality control step

- synthetize information and visualize points in a reduced dimension space
- describe links between variables and which ones explain most variability
- highlight homogeneous subgroups linked to biological effect
- detect aberrant samples







PCA - Computing

Computing PCA:

- Standardize the range of continuous initial variables
  → data homoscedasticity : the variance must be independent of the mean

$$z = \frac{x - \mu}{\sigma}$$

- Compute the covariance matrix **A**

$$\boldsymbol{A} \begin{bmatrix} var(x) & cov(x,y) & cov(x,z) \\ cov(x,y) & var(y) & cov(y,z) \\ cov(x,z) & cov(y,z) & var(z) \end{bmatrix}$$

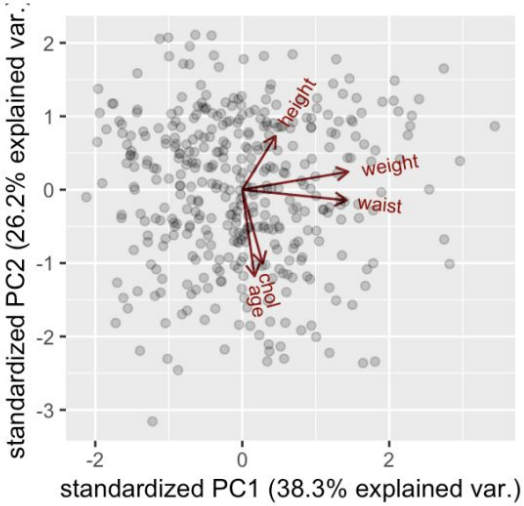
- Calculate the eigenvalues λ and eigenvectors for the covariance matrix → *solve* $|\mathbf{A}-\lambda\cdot\mathbf{I}| = 0$

- Sort eigenvalues λ and their corresponding eigenvectors

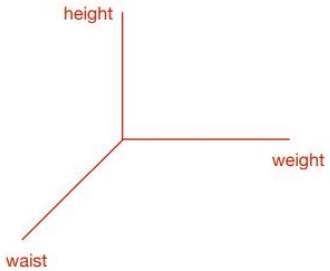
- Recast the data along the principal component axes

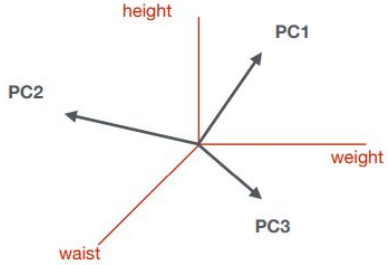
PCA - Plots

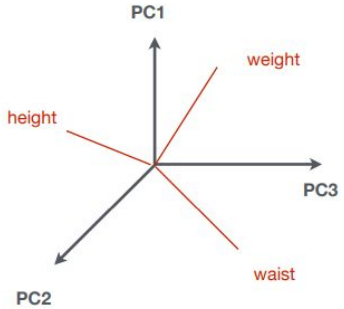
- each dot is a sample
- new coordinate system (PC_1, PC_2…)
- red arrows = contribution of each initial variable (old coordinate system)
- several 2D (2 PCs) plots : PC_1/PC_2
  PC_1/PC_3
  PC_2/PC_3
  …



<u>PCA biplot</u>
score plot + loading plot



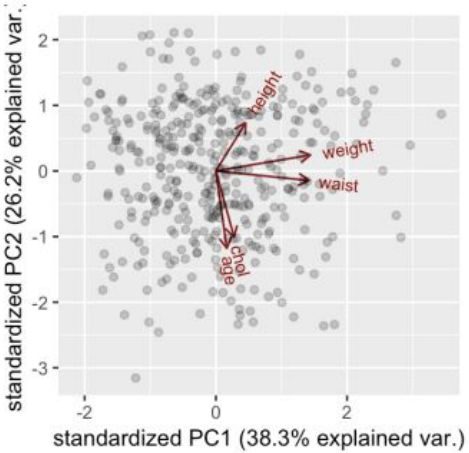


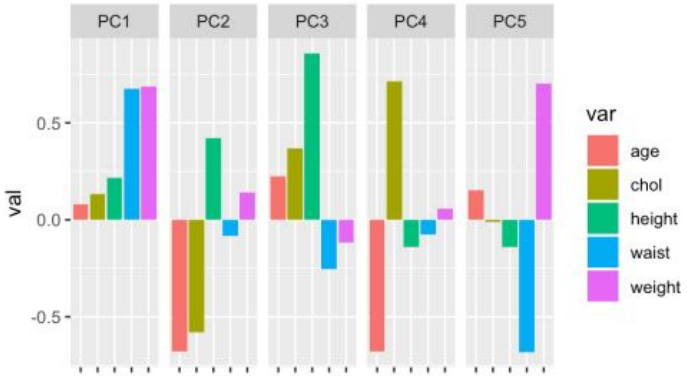


PCA - Components

- contribution of each initial variable to the PC_i : $\alpha, \beta, \gamma \ldots$are coefficients also called "loadings"
- some variables contribute in the same direction to some PCs (e.g. waist and height for PC_1), but opposite to others (PC_5)
- PC are orthogonal: no information redundancy between PC →reduce the “useful” representation space

$$PC_i = \alpha_i \cdot \text{age} + \beta_i \cdot \text{chol} + \gamma_i \cdot \text{height} + \delta_i \cdot \text{waist} + \epsilon_i \cdot \text{weight}$$

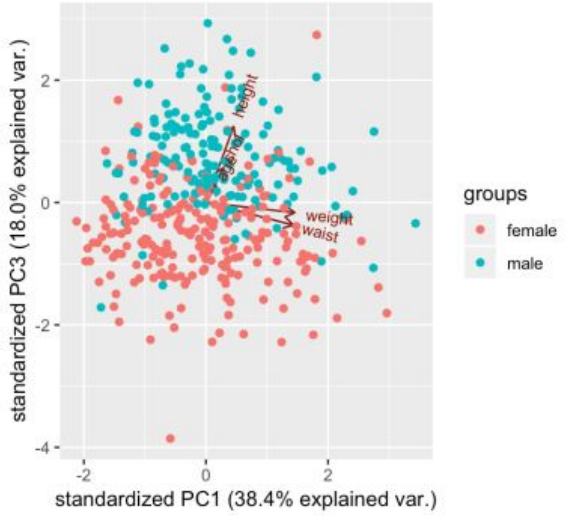


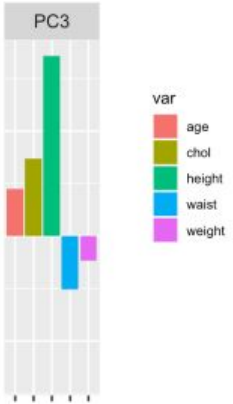


PCA - Biological interpretation

- PC plots can highlight new groups
- Example: PC_3 seems very associated to gender

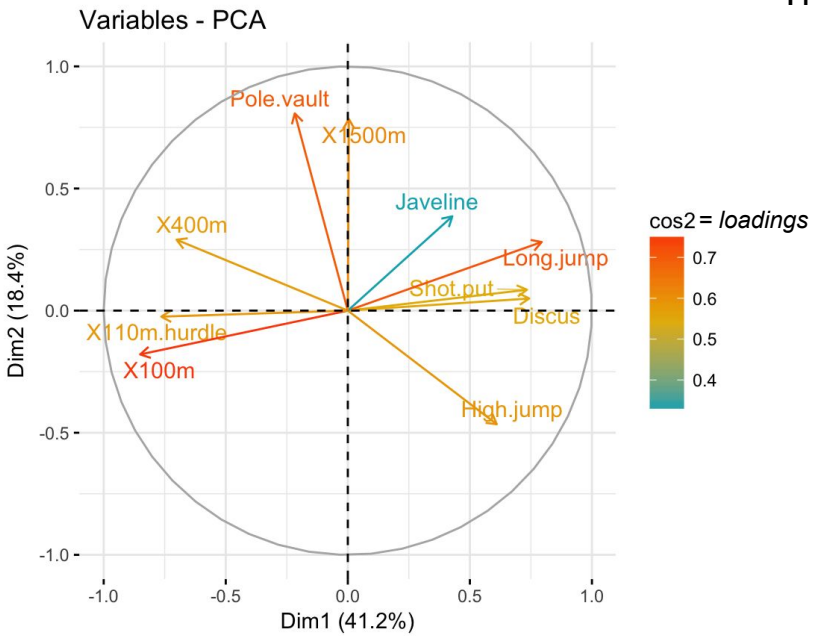
  → PC_3 loadings indicate that a combination of height and cholesterol separates men / women





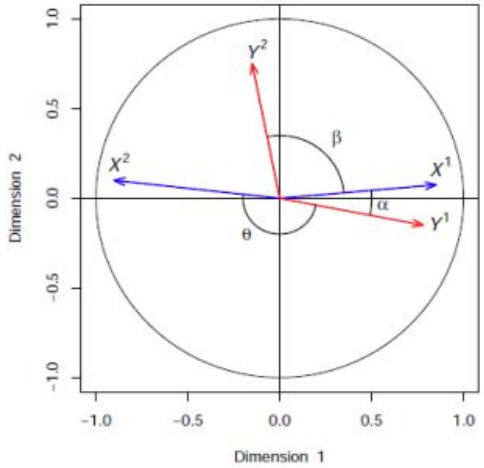
⚠ Be careful with visual proximity between 2 samples
→ depends on selected PC

PCA - Variable correlations in loading plots



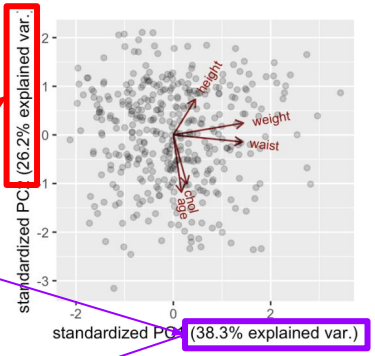
The correlation between two variables is represented as :

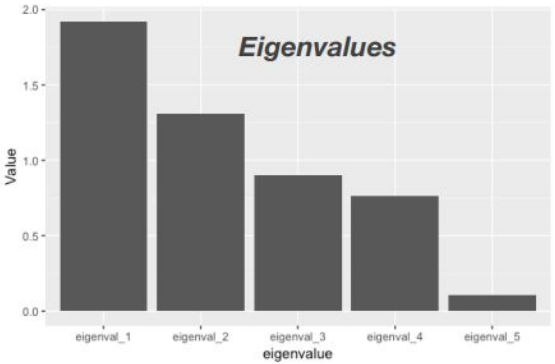
- an acute angle ($\cos(\alpha) > 0$) if it is positive
- an obtuse angle ($\cos(\theta) < 0$) if it is negative
- a right angle ($\cos(\beta) \approx 0$) if it is near zero

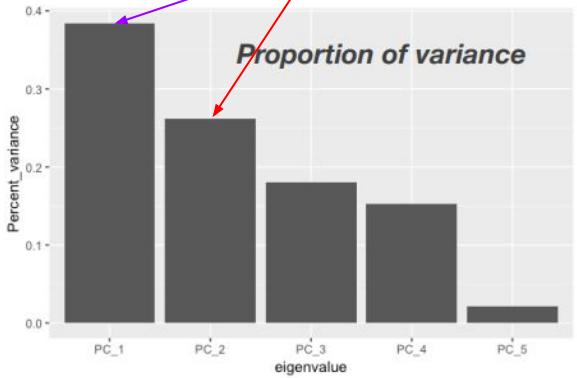


PCA - Scree plot

- Each PC explains some part of the total variance of the dataset
- This amount is proportional to the corresponding eigenvalue
- PC are ordered by decreasing eigenvalue (hence explained variance)



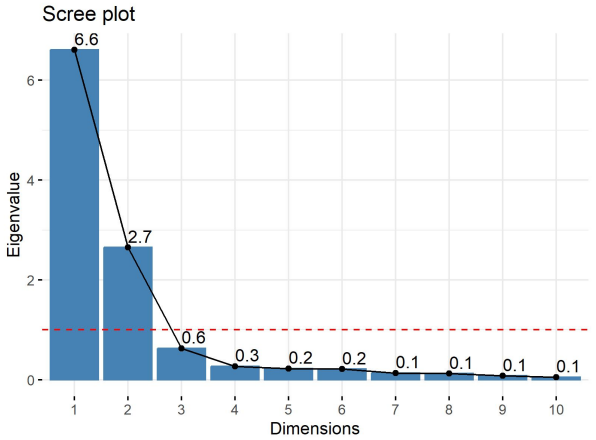




Considering PC1 & PC2 explains 63% of the total vairance

PCA - PCs number

- Several criteria to select the optimal subset of PC, without loosing too much information
- Proportion of total variance: keep PC such that the cumulative variance is above threshold
- Average eigenvalue criteria: keep PC which have eigenvalue larger than
  - mean eigenvalue (Kaiser rule) or
  - 70% of mean eigenvalue (Jottclife rule)

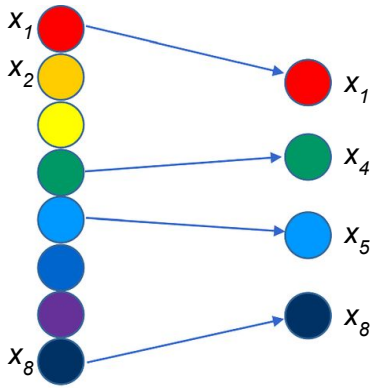


Extraction + Selection

- **Feature selection**

  → determine a smaller set of features minimizing (relevant) information loss

  e.g. filtering methods (correlation), recursive elimination, regularization

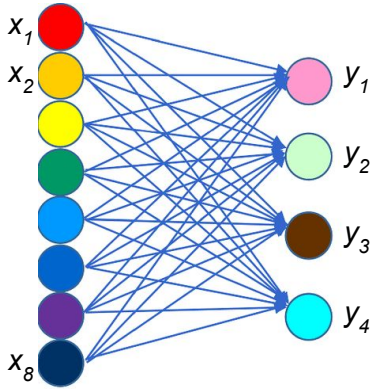


- **Feature extraction**

  → combine the input features into another set of variables in a linear or non-linear way : $y_1 = \alpha_1 * x_1 + \alpha_2 * x_2 + \alpha_3 * x_3 + \ldots$.

  e.g. **PCA**, PCoA, ICA…

  + regularization for sparse methods : **sPCA**, sNMF (i.e. some α_i forced to 0)



Sparse PCA : regularization

- To learn a more “simpler”/”comprehensive” model and avoid overfitting or inconsistency situations

- Linear combination of two functions f_1 and f_2 for a vector w : $f(\mathbf{w}) = f_1(\mathbf{w}) + \lambda\, f_2(\mathbf{w})$

  → adjusting the penalty λ (*regularization parameter*) give more/less weight to the regularizer f_2

- The simplest regularizer f_2 is the L0-norm $f(\mathbf{w}) = g(\mathbf{w}) + \lambda \|\mathbf{w}\|_0$

  with *g(w)* the objective/loss function to minimize and $\|\mathbf{w}\|_0 = \text{number of non-zero entries of } \mathbf{w}$

  → to minimize f(w) → minimize g(w) and limit the cost of the regularizer (ie limit w_0)

  $y = w_1{*}x_1 + w_2{*}x_2 + w_3{*}x_3 + w_4{*}x_4 + w_5{*}x_5$ ⇨ $y = 0{*}x_1 + w_2{*}x_2 + 0{*}x_3 + 0{*}x_4 + w_5{*}x_5$

- Alternative regularizer f_2 is the L1-norm $f(\mathbf{w}) = g(\mathbf{w}) + \lambda \|\mathbf{w}\|_1$ with $\|\mathbf{w}\|_1 = \sum_{n=0}^{N} |w_n|$

- Common regularization strategies : Lasso (L1), Ridge (L2) and Elastic Net (L1+L2)

Sparse PCA principle

- Objective to PCA: find linear combinations to maximize variability of projected data

$PC_1 : y_1 = w_1 * x_1 + w_2 * x_2 + w_3 * x_3 + \ldots.$

PCA

$$\underset{w_i:\,||w_i||_2^2=1}{\arg\max}\ \mathrm{Var}(Xw_i) \Rightarrow \underbrace{\underset{\mathbf{W},\mathbf{P}}{\operatorname{argmin}} \left\|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{P}^\top\right\|_F^2}_{\text{g(w) function to minimize}}$$

Sparse PCA

$$\Rightarrow \underset{\mathbf{W},\mathbf{P}}{\operatorname{argmin}} \left\|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{P}^\top\right\|_F^2 + \underbrace{\sum_{k=1}^{K} \lambda \|\mathbf{w}_k\|^2 + \sum_{k=1}^{K} \lambda_{1,k} \|\mathbf{w}_k\|_1}_{\text{regularizer (L2 + L1)}}$$

(Elastic-Net as proposed by Zou et al. (2006))

$PC_1 = 0 * x_1 + w_2 * x_2 + 0 * x_3 + 0 * x_4 + w_5 * x_5$

⚠ If PCA formulation are equivalents, sparse PCA formulations are not.

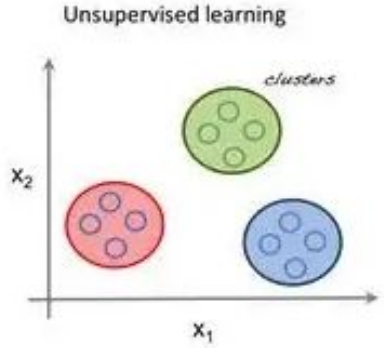
Guerra-Urzola, R. et al. A Guide for Sparse PCA: Model Comparison and Applications. Psychometrika (2021).

Statistics…. some vocabulary

- **Unsupervised learning**

  → find hidden patterns, analyze and organize unlabelled samples

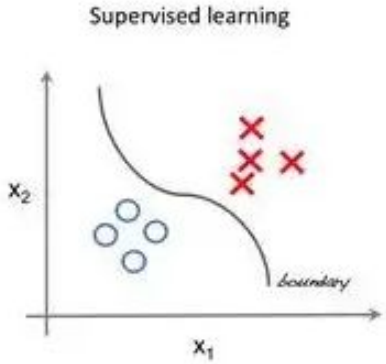
  e.g. clustering, dimension reduction, density estimation



- **Supervised learning**

  → use labelled samples and previous outputs to guess outcomes in advance (predictive model)

  e.g. classification task (categorical/numerical), regression (numerical)



- **Semi-supervised learning**

  → only some labelled samples (not available, too expensive…)

Differential analysis - Principle

Principle: Compare 2 or more sample groups (experimental conditions, treatment, time…)
e.g. healthy VS sick, old VS young…

Objective: detect differentially expressed (DE) genes/proteins/… between groups
→ analysis based on statistical tests (t-test…)
→ a gene/protein/… is “DE” if the difference is statistically significant between 2 groups, ie greater than any natural random variation

Specificities of omics:

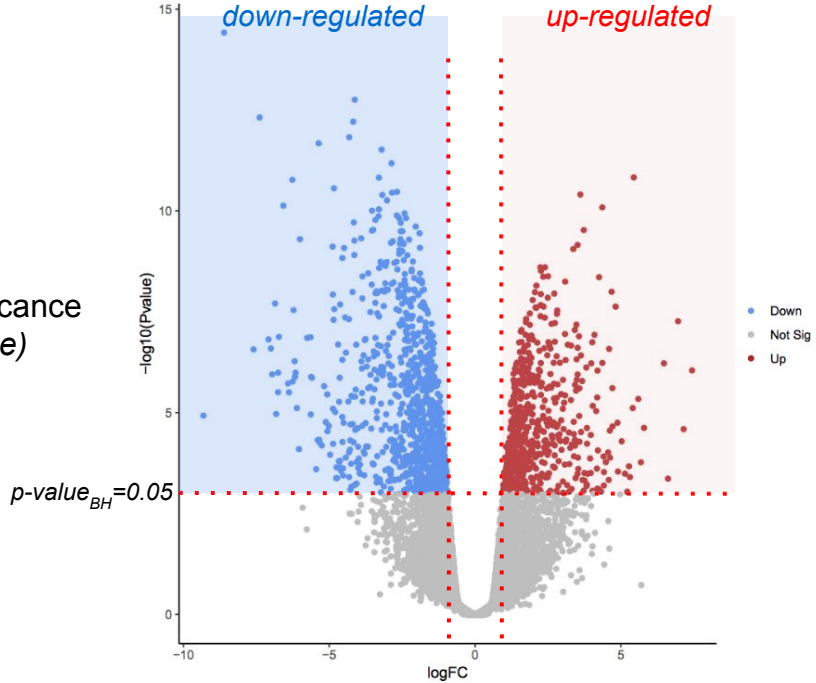
- few individuals
- many variables → many tests
- overdispersion problem (high variance)
- numerous possible bias
- omic specific data distribution

→ such analysis approaches exist for each omic

Differential analysis - Volcano Plot

”A gene/protein/… is declared differentially expressed if the observed difference between two conditions is statistically significant at 5% and the fold change is higher than 2”

Statistical significance
$-log_{10}(p\text{-}value)$



Biological significance
log_2FC

$log_2FC = 1 \rightarrow FC=2$
$log_2FC = -1 \rightarrow FC=0.5$

Overfitting, Cross-Validation & Regularization

Overfitting, Cross-Validation & Regularization

	x_1	x_2	x_3	x_4	y
	Intercept	Age	Nb_sisters	Neighbor'weight (kg)	Subject's Height (cm)
Subj1	1	5	1	1	90
Subj2	1	10	2	50	125
Subj3	1	15	1	80	160
Subj4	1	20	2	90	180

Overfitting, Cross-Validation & Regularization

	x_1	x_2	x_3	x_4	y
	Intercept	Age	Nb_sisters	Neighbor'weight (kg)	Subject's Height (cm)
TEST	Subj1	1	5	1	1	90
TRAIN	Subj2	1	10	2	50	125
	Subj3	1	15	1	80	160
	Subj4	1	20	2	90	180

Overfitting, Cross-Validation & Regularization

	x_1	x_2	x_3	x_4	y
	Intercept	Age	Nb_sisters	Neighbor'weight (kg)	Subject's Height (cm)
TEST	Subj1	1	5	1	1	90
TRAIN	Subj2	1	10	2	50	125
	Subj3	1	15	1	80	160
	Subj4	1	20	2	90	180

We are looking for $\beta_1, \beta_2, \beta_3$ and β_4 that minimizes $J_{TRAIN} = \sum_{i=2}^{4}(y_i - \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4})^2$.

Overfitting, Cross-Validation & Regularization

	x_1	x_2	x_3	x_4	y
	Intercept	Age	Nb_sisters	Neighbor'weight (kg)	Subject's Height (cm)
TEST	Subj1	1	5	1	1	90
TRAIN	Subj2	1	10	2	50	125
	Subj3	1	15	1	80	160
	Subj4	1	20	2	90	180

We are looking for $\beta_1, \beta_2, \beta_3$ and β_4 that minimizes $J_{TRAIN} = \sum_{i=2}^{4}(y_i - \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4})^2$.

Similarly, we can define $J_{TEST} = (y_1 - \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \beta_4 x_{14})^2$.

Overfitting, Cross-Validation & Regularization

	x_1	x_2	x_3	x_4	y
	Intercept	Age	Nb_sisters	Neighbor'weight (kg)	Subject's Height (cm)
TEST	Subj1	1	5	1	1	90
TRAIN	Subj2	1	10	2	50	125
	Subj3	1	15	1	80	160
	Subj4	1	20	2	90	180

We are looking for $\beta_1, \beta_2, \beta_3$ and β_4 that minimizes $J_{TRAIN} = \sum_{i=2}^{4}(y_i - \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4})^2$.

Similarly, we can define $J_{TEST} = (y_1 - \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \beta_4 x_{14})^2$.

Here, we are in "high-dimension" as $n < p$. The problem is ill-posed (more unknown parameters than equations).

Overfitting, Cross-Validation & Regularization

	x_1	x_2	x_3	x_4	y
	Intercept	Age	Nb_sisters	Neighbor'weight (kg)	Subject's Height (cm)
TEST	Subj1	1	5	1	1	90
TRAIN	Subj2	1	10	2	50	125
	Subj3	1	15	1	80	160
	Subj4	1	20	2	90	180

We are looking for $\beta_1, \beta_2, \beta_3$ and β_4 that minimizes $J_{TRAIN} = \sum_{i=2}^{4}(y_i - \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4})^2$.

Similarly, we can define $J_{TEST} = (y_1 - \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \beta_4 x_{14})^2$.

Here, we are in "high-dimension" as $n < p$. The problem is ill-posed (more unknown parameters than equations).

➔ It is possible to find an infinite number of solutions:

Overfitting, Cross-Validation & Regularization

	x_1	x_2	x_3	x_4	y
	Intercept	Age	Nb_sisters	Neighbor'weight (kg)	Subject's Height (cm)
TEST	Subj1	1	5	1	1	90
TRAIN	Subj2	1	10	2	50	125
	Subj3	1	15	1	80	160
	Subj4	1	20	2	90	180

We are looking for $\beta_1, \beta_2, \beta_3$ and β_4 that minimizes $J_{TRAIN} = \sum_{i=2}^{4}(y_i - \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4})^2$.

Similarly, we can define $J_{TEST} = (y_1 - \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \beta_4 x_{14})^2$.

Here, we are in "high-dimension" as $n < p$. The problem is ill-posed (more unknown parameters than equations).

➔ It is possible to find an infinite number of solutions:

	β_1	β_2	β_3	β_4	J_{train}	J_{test}
Solution 1	43.75	0	1.375	6.25	8.4e-22	1491.891
Solution 2	-7456.25	-1000	251.375	2506.25	1.1e-19	95817179
⋮						

Overfitting, Cross-Validation & Regularization

	x_1	x_2	x_3	x_4	y
	Intercept	Age	Nb_sisters	Neighbor'weight (kg)	Subject's Height (cm)
TEST	Subj1	1	5	1	1	90
TRAIN	Subj2	1	10	2	50	125
	Subj3	1	15	1	80	160
	Subj4	1	20	2	90	180

We are looking for $\beta_1, \beta_2, \beta_3$ and β_4 that minimizes $J_{TRAIN} = \sum_{i=2}^{4}(y_i - \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4})^2$.

Similarly, we can define $J_{TEST} = (y_1 - \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \beta_4 x_{14})^2$.

Here, we are in "high-dimension" as $n < p$. The problem is ill-posed (more unknown parameters than equations).

➔ It is possible to find an infinite number of solutions:

	β_1	β_2	β_3	β_4	J_{train}	J_{test}
Solution 1	43.75	0	1.375	6.25	8.4e-22	1491.891
Solution 2	-7456.25	-1000	251.375	2506.25	1.1e-19	95817179
⋮						

Go against the idea that age is the best explanatory variable.

Overfitting, Cross-Validation & Regularization

	x_1	x_2	x_3	x_4	y
	Intercept	Age	Nb_sisters	Neighbor'weight (kg)	Subject's Height (cm)
TEST	Subj1	1	5	1	1	90
TRAIN	Subj2	1	10	2	50	125
	Subj3	1	15	1	80	160
	Subj4	1	20	2	90	180

We are looking for $\beta_1, \beta_2, \beta_3$ and β_4 that minimizes $J_{TRAIN} = \sum_{i=2}^{4}(y_i - \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4})^2$.

Similarly, we can define $J_{TEST} = (y_1 - \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \beta_4 x_{14})^2$.

OVERFITTING

Here, we are in "high-dimension" as $n < p$. The problem is ill-posed (more unknown parameters than equations).

➔ It is possible to find an infinite number of solutions:

	β_1	β_2	β_3	β_4	J_{train}	J_{test}
Solution 1	43.75	0	1.375	6.25	8.4e-22	1491.891
Solution 2	-7456.25	-1000	251.375	2506.25	1.1e-19	95817179
⋮						

Go against the idea that age is the best explanatory variable.

Overfitting, Cross-Validation & Regularization

Figure extracted from https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b

Overfitting, Cross-Validation & Regularization

Cross-Validation allows to evaluate the generalization power of a model and realize if the model overfits or not.

Figure extracted from https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b

Overfitting, Cross-Validation & Regularization

Cross-Validation allows to evaluate the generalization power of a model and realize if the model overfits or not.

A lot of sampling possibilities are available to perform Cross-Validation (CV). The most well-known is K-fold CV:

Figure extracted from https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b

Overfitting, Cross-Validation & Regularization

Cross-Validation allows to evaluate the generalization power of a model and realize if the model overfits or not.

A lot of sampling possibilities are available to perform Cross-Validation (CV). The most well-known is K-fold CV:

Figure extracted from https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b

Overfitting, Cross-Validation & Regularization

Cross-Validation allows to evaluate the generalization power of a model and realize if the model overfits or not.

A lot of sampling possibilities are available to perform Cross-Validation (CV). The most well-known is K-fold CV:

One way to avoid overfitting is by preforming regularization.

Figure extracted from https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b

Overfitting, Cross-Validation & Regularization

Cross-Validation allows to evaluate the generalization power of a model and realize if the model overfits or not.

A lot of sampling possibilities are available to perform Cross-Validation (CV). The most well-known is K-fold CV:

One way to avoid overfitting is by preforming regularization.

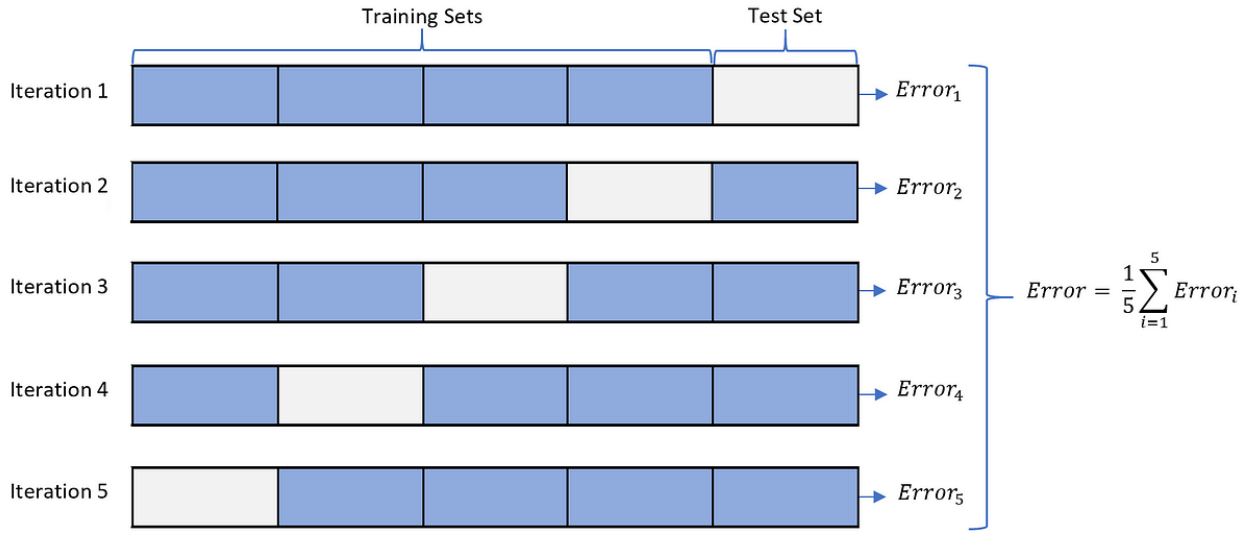
Regularization consists in adding more constraints to the model in order to reduce the space of solutions.

Figure extracted from https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b

Overfitting, Cross-Validation & Regularization

Cross-Validation allows to evaluate the generalization power of a model and realize if the model overfits or not.

A lot of sampling possibilities are available to perform Cross-Validation (CV). The most well-known is K-fold CV:



One way to avoid overfitting is by preforming regularization.

Regularization consists in adding more constraints to the model in order to reduce the space of solutions.

Multiple regularizations are available such as Ridge or LASSO regularizations.

Figure extracted from https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b

Overfitting, Cross-Validation & Regularization

Cross-Validation allows to evaluate the generalization power of a model and realize if the model overfits or not.

A lot of sampling possibilities are available to perform Cross-Validation (CV). The most well-known is K-fold CV:

One way to avoid overfitting is by preforming regularization.

Regularization consists in adding more constraints to the model in order to reduce the space of solutions.

Multiple regularizations are available such as Ridge or LASSO regularizations.

Here, we choose to regularize the model by forcing it to have a low number of variables.

Figure extracted from https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b

Overfitting, Cross-Validation & Regularization

Overfitting, Cross-Validation & Regularization

So let us consider all models with either 2 or 3 variables (with at least the intercept each time).

Overfitting, Cross-Validation & Regularization

So let us consider all models with either 2 or 3 variables (with at least the intercept each time).

By doing so, we add respectively 2 (ex: $\beta_2 = 0$ and $\beta_4 = 0$) or 1 constraint (idem).

Overfitting, Cross-Validation & Regularization

So let us consider all models with either 2 or 3 variables (with at least the intercept each time).

By doing so, we add respectively 2 (ex: $\beta_2 = 0$ and $\beta_4 = 0$) or 1 constraint (idem).

For all these possible models, let us compute J_{TRAIN} and J_{TEST}:

Overfitting, Cross-Validation & Regularization

So let us consider all models with either 2 or 3 variables (with at least the intercept each time).

By doing so, we add respectively 2 (ex: $\beta_2 = 0$ and $\beta_4 = 0$) or 1 constraint (idem).

For all these possible models, let us compute J_{TRAIN} and J_{TEST}:

Variables considered	J_{TRAIN}	J_{TEST}
(x_1, x_2)	3.750000e+01	100
(x_1, x_3)	2.403846e+01	959.8081
(x_1, x_4)	1.512500e+03	4900
(x_1, x_2, x_3)	1.831567e-22	203.0625
(x_1, x_2, x_4)	6.464166e-24	225
(x_1, x_3, x_4)	8.664767e-22	1491.8906

Overfitting, Cross-Validation & Regularization

So let us consider all models with either 2 or 3 variables (with at least the intercept each time).

By doing so, we add respectively 2 (ex: $\beta_2 = 0$ and $\beta_4 = 0$) or 1 constraint (idem).

For all these possible models, let us compute J_{TRAIN} and J_{TEST}:

Variables considered	J_{TRAIN}	J_{TEST}
(x_1, x_2)	3.750000e+01	100
(x_1, x_3)	2.403846e+01	959.8081
(x_1, x_4)	1.512500e+03	4900
(x_1, x_2, x_3)	1.831567e-22	203.0625
(x_1, x_2, x_4)	6.464166e-24	225
(x_1, x_3, x_4)	8.664767e-22	1491.8906

OVERFITTING

Overfitting, Cross-Validation & Regularization

So let us consider all models with either 2 or 3 variables (with at least the intercept each time).

By doing so, we add respectively 2 (ex: $\beta_2 = 0$ and $\beta_4 = 0$) or 1 constraint (idem).

For all these possible models, let us compute J_{TRAIN} and J_{TEST}:

Variables considered	J_{TRAIN}	J_{TEST}
(x_1, x_2)	3.750000e+01	100
(x_1, x_3)	2.403846e+01	959.8081
(x_1, x_4)	1.512500e+03	4900
(x_1, x_2, x_3)	1.831567e-22	203.0625
(x_1, x_2, x_4)	6.464166e-24	225
(x_1, x_3, x_4)	8.664767e-22	1491.8906

Best model

OVERFITTING

Overfitting, Cross-Validation & Regularization

So let us consider all models with either 2 or 3 variables (with at least the intercept each time).

By doing so, we add respectively 2 (ex: $\beta_2 = 0$ and $\beta_4 = 0$) or 1 constraint (idem).

For all these possible models, let us compute J_{TRAIN} and J_{TEST}:

Variables considered	J_{TRAIN}	J_{TEST}
(x_1, x_2)	3.750000e+01	100
(x_1, x_3)	2.403846e+01	959.8081
(x_1, x_4)	1.512500e+03	4900
(x_1, x_2, x_3)	1.831567e-22	203.0625
(x_1, x_2, x_4)	6.464166e-24	225
(x_1, x_3, x_4)	8.664767e-22	1491.8906

Best model

OVERFITTING

CV was also used here so set an hyper-parameter: «the number of variables to keep in the model».

Overfitting, Cross-Validation & Regularization

So let us consider all models with either 2 or 3 variables (with at least the intercept each time).

By doing so, we add respectively 2 (ex: $\beta_2 = 0$ and $\beta_4 = 0$) or 1 constraint (idem).

For all these possible models, let us compute J_{TRAIN} and J_{TEST}:

Variables considered	J_{TRAIN}	J_{TEST}
(x_1, x_2)	3.750000e+01	100
(x_1, x_3)	2.403846e+01	959.8081
(x_1, x_4)	1.512500e+03	4900
(x_1, x_2, x_3)	1.831567e-22	203.0625
(x_1, x_2, x_4)	6.464166e-24	225
(x_1, x_3, x_4)	8.664767e-22	1491.8906

Best model

OVERFITTING

CV was also used here so set an hyper-parameter: «the number of variables to keep in the model».

Here apparently, keeping only 2 variables leads to the best model with the variable «Age», which was expected.

Overfitting, Cross-Validation & Regularization

Figure taken from https://typeset.io/resources/top-reasons-for-research-paper-rejection/

Overfitting, Cross-Validation & Regularization

Overfitting can be handled with regularization.

Figure taken from https://typeset.io/resources/top-reasons-for-research-paper-rejection/

Overfitting, Cross-Validation & Regularization

Overfitting can be handled with regularization.

Cross-Validation can both help to:

Figure taken from https://typeset.io/resources/top-reasons-for-research-paper-rejection/

Overfitting, Cross-Validation & Regularization

Overfitting can be handled with regularization.

Cross-Validation can both help to:

1. **realize if the model overfits or not**

Figure taken from https://typeset.io/resources/top-reasons-for-research-paper-rejection/

Overfitting, Cross-Validation & Regularization

Overfitting can be handled with regularization.

Cross-Validation can both help to:

1. **realize if the model overfits or not**
2. **tune the hyper-parameters (associated with the regularization).**

Figure taken from https://typeset.io/resources/top-reasons-for-research-paper-rejection/

Overfitting, Cross-Validation & Regularization

Overfitting can be handled with regularization.

Cross-Validation can both help to:

1. **realize if the model overfits or not**
2. **tune the hyper-parameters (associated with the regularization).**

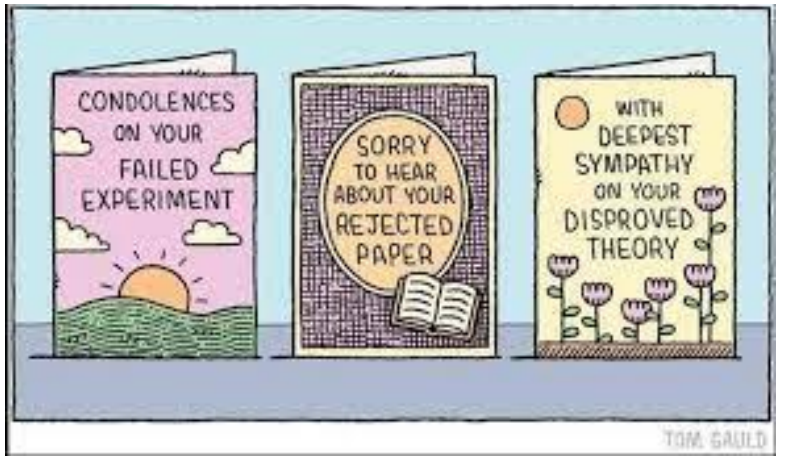


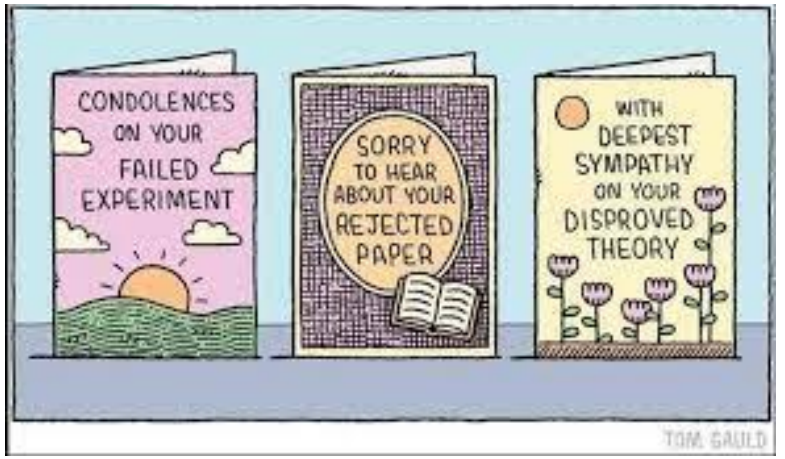
Figure taken from https://typeset.io/resources/top-reasons-for-research-paper-rejection/

Overfitting, Cross-Validation & Regularization

Overfitting can be handled with regularization.

Cross-Validation can both help to:

1. **realize if the model overfits or not**
2. **tune the hyper-parameters (associated with the regularization).**



Classical mistake to avoid with Cross-Validation: «**Double Dipping**».

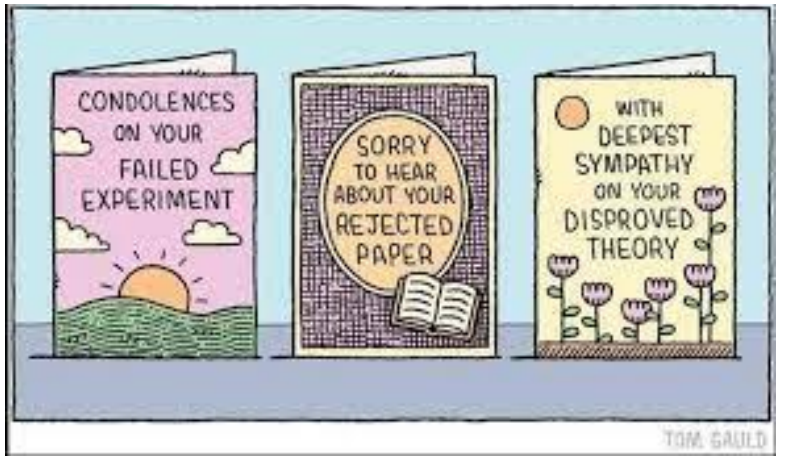
Figure taken from https://typeset.io/resources/top-reasons-for-research-paper-rejection/

Overfitting, Cross-Validation & Regularization

Overfitting can be handled with regularization.

Cross-Validation can both help to:

1. **realize if the model overfits or not**
2. **tune the hyper-parameters (associated with the regularization).**



Classical mistake to avoid with Cross-Validation: «**Double Dipping**».

➔ The whole point of Cross-Validation is to keep the train and the test sets **independant** from each other.

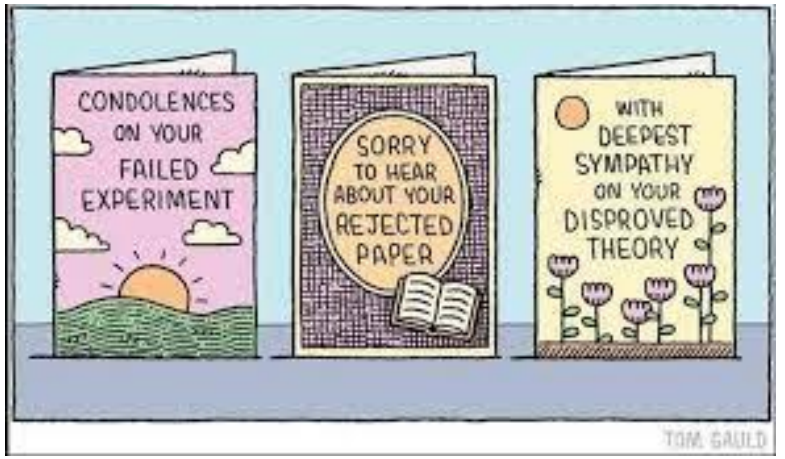
Figure taken from https://typeset.io/resources/top-reasons-for-research-paper-rejection/

Overfitting, Cross-Validation & Regularization

Overfitting can be handled with regularization.

Cross-Validation can both help to:

1. **realize if the model overfits or not**
2. **tune the hyper-parameters (associated with the regularization).**



Classical mistake to avoid with Cross-Validation: «**Double Dipping**».

➔ The whole point of Cross-Validation is to keep the train and the test sets **independant** from each other.

This is no longer the case when for example:

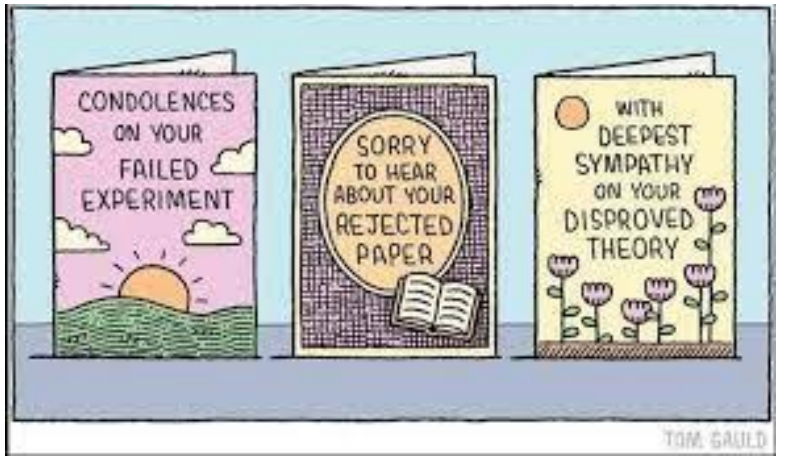
Figure taken from https://typeset.io/resources/top-reasons-for-research-paper-rejection/

Overfitting, Cross-Validation & Regularization

Overfitting can be handled with regularization.

Cross-Validation can both help to:

1. **realize if the model overfits or not**
2. **tune the hyper-parameters (associated with the regularization).**



Classical mistake to avoid with Cross-Validation: «**Double Dipping**».

➔ The whole point of Cross-Validation is to keep the train and the test sets **independant** from each other.

This is no longer the case when for example:

1. Normalization accross subjects is performed on the whole data-set.

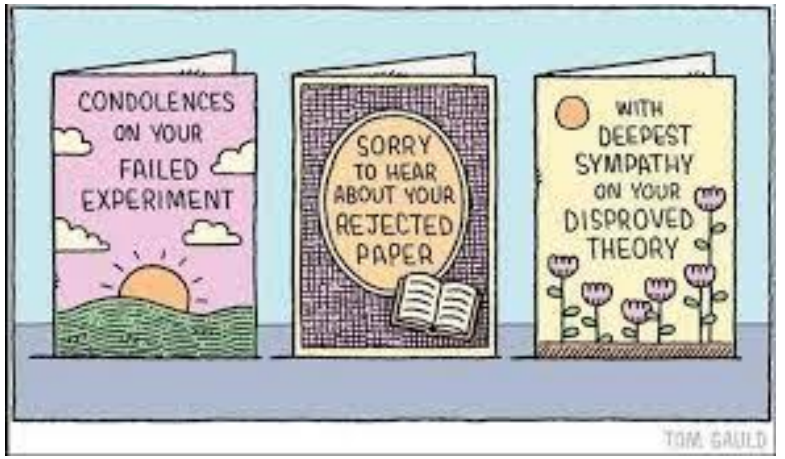
Figure taken from https://typeset.io/resources/top-reasons-for-research-paper-rejection/

Overfitting, Cross-Validation & Regularization

Overfitting can be handled with regularization.

Cross-Validation can both help to:

1. **realize if the model overfits or not**
2. **tune the hyper-parameters (associated with the regularization).**



Classical mistake to avoid with Cross-Validation: «**Double Dipping**».

➔ The whole point of Cross-Validation is to keep the train and the test sets **independant** from each other.

This is no longer the case when for example:

1. Normalization accross subjects is performed on the whole data-set.
2. Variable selection is performed on the whole data-set (ex: differentially expressed genes)

Figure taken from https://typeset.io/resources/top-reasons-for-research-paper-rejection/

Swiss Institute of Bioinformatics

ÉCOLE THÉMATIQUE ETBII BIOINFORMATIQUE INTÉGRATIVE IFB

Omics integration

General aspects

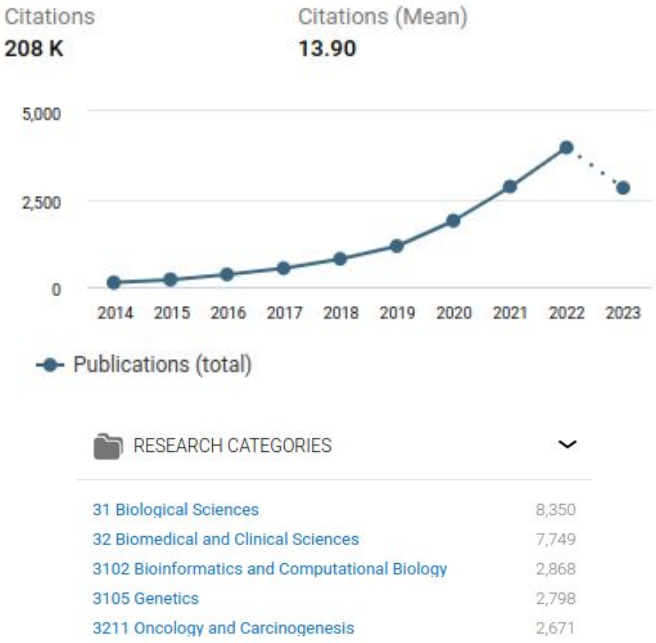
INSTITUT FRANÇAIS DE BIOINFORMATIQUE IFB

elixir FRANCE

cnrs INRAE Inserm cea

Rise in popularity

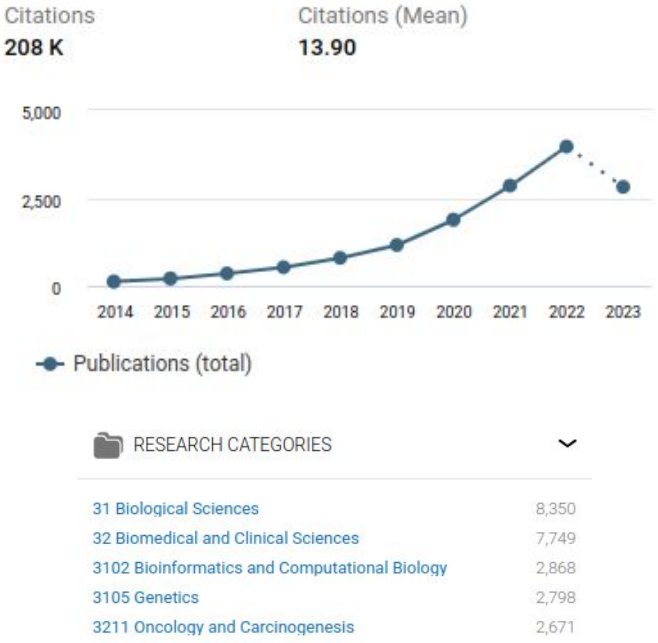
"Multi-omics" citations



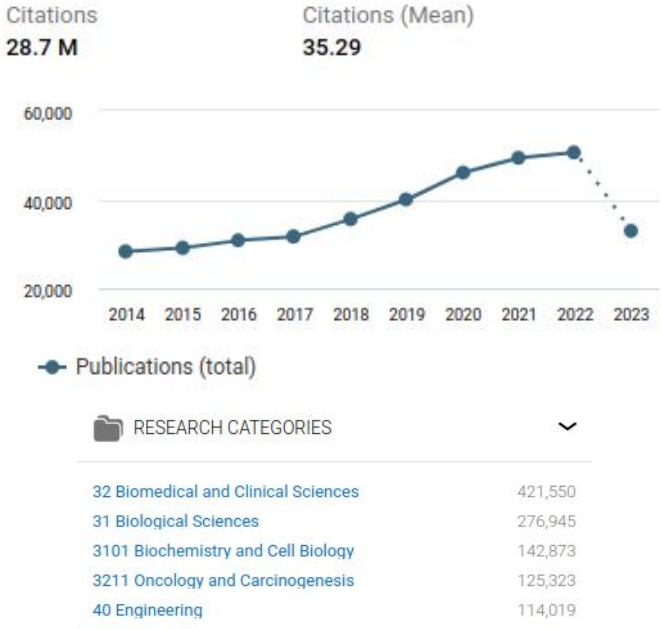
https://app.dimensions.ai/discover/publication *(8th Jan. 2023 : 132,863,611 referenced publications)*

Rise in popularity

“Multi-omics” citations

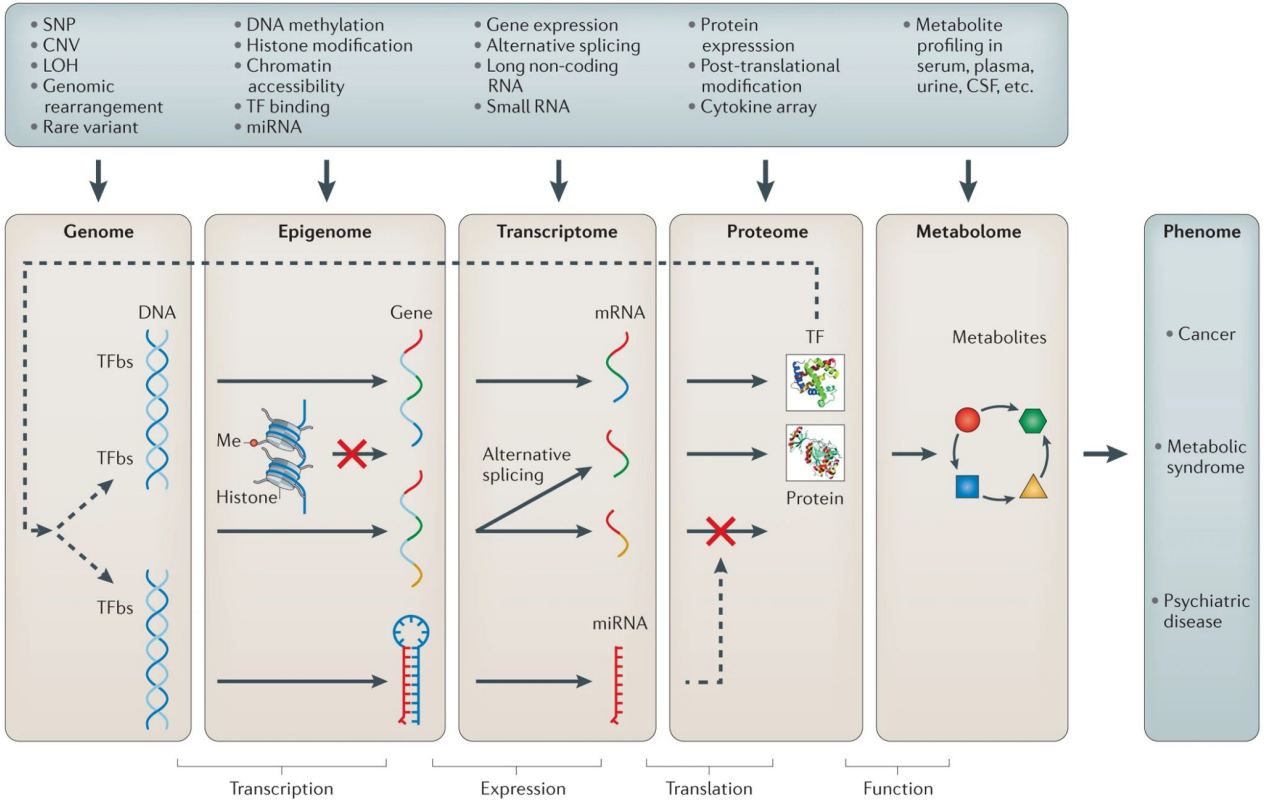


“Single-cell” citations



https://app.dimensions.ai/discover/publication (23th Aug. 2023 : 138,395,868 referenced publications)

Omics… which ones ?



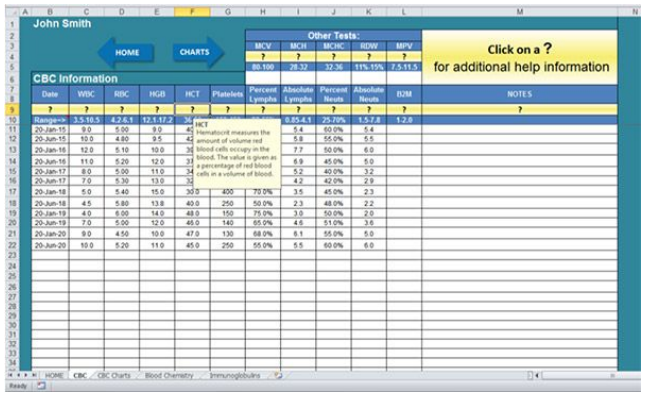
Ritchie, M., Holzinger, E., Li, R. et al. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet 16, 85–97 (2015).

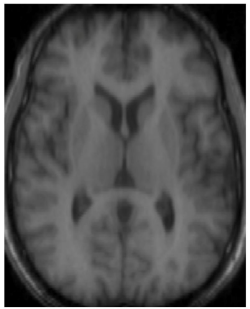
But also ?

Other data ?

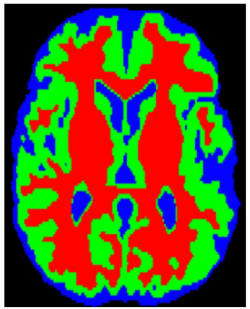
- clinical data
- imaging data (full data or extracted characteristics)
- new omics fields : fluxomics, ionomics, microbiomics, glycomics…
- biological knowledge : DNA/protein, protein/protein interactions

→ a priori in model definition/construction





(a) Axial slice

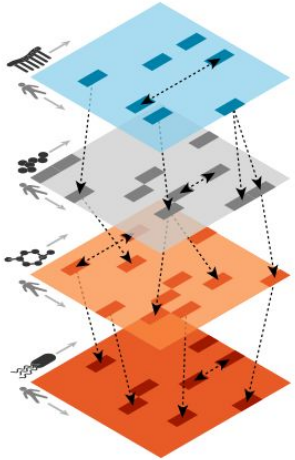


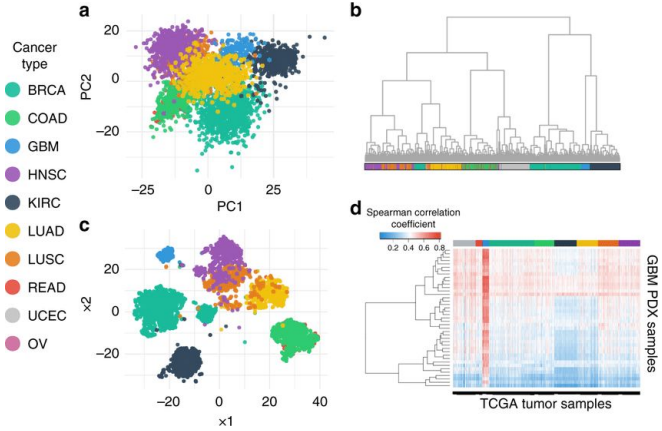
(b) Tissue segmentation

https://www.statcan.gc.ca/en/data-science/network/image-segmentation

Integration: why ?

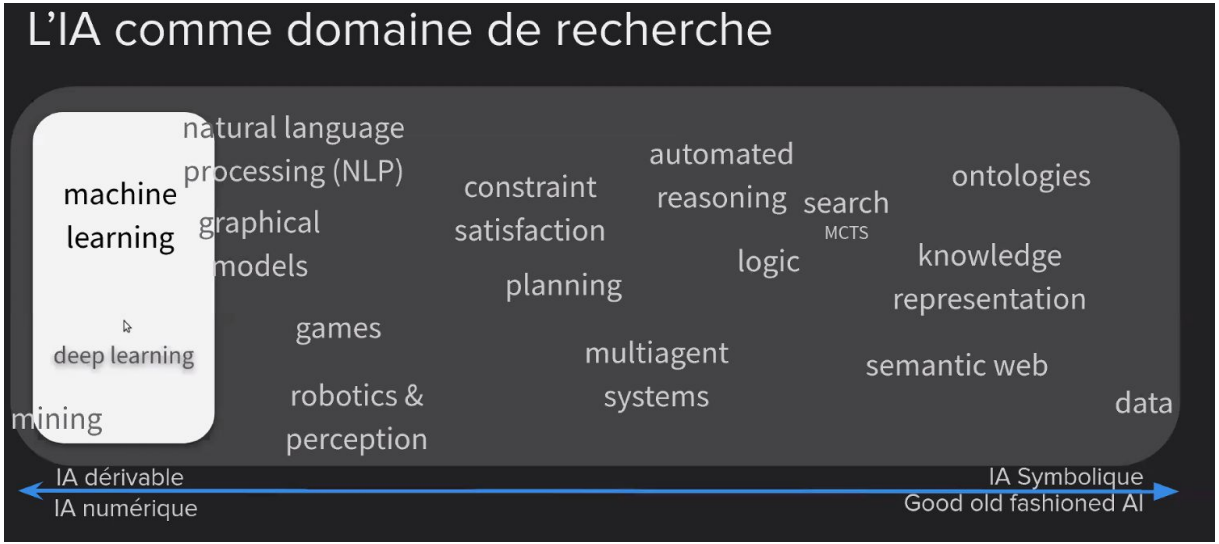
- Disease subtyping and classification
- Biomarkers prediction : diagnostic, disease drivers
- Deep insights into disease biology

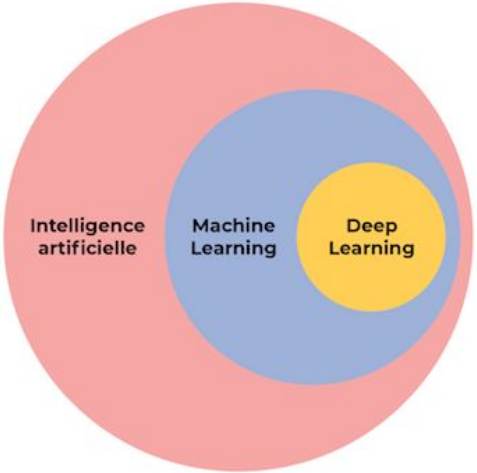




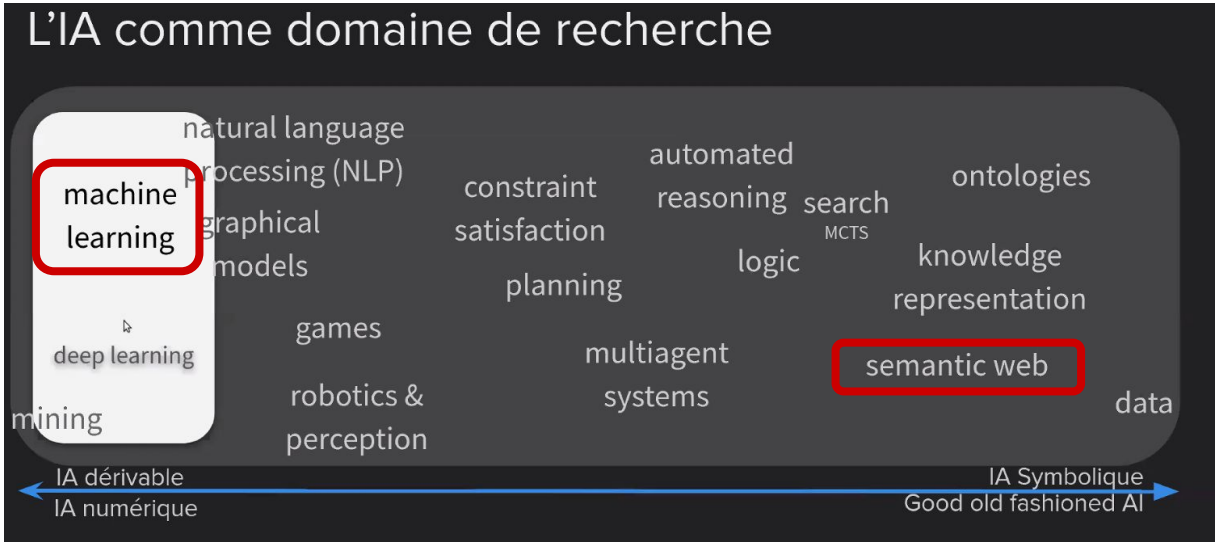
Vasileios et al (2018). Drug and disease signature integration identifies synergistic combinations in glioblastoma. Nature Communications. 9.

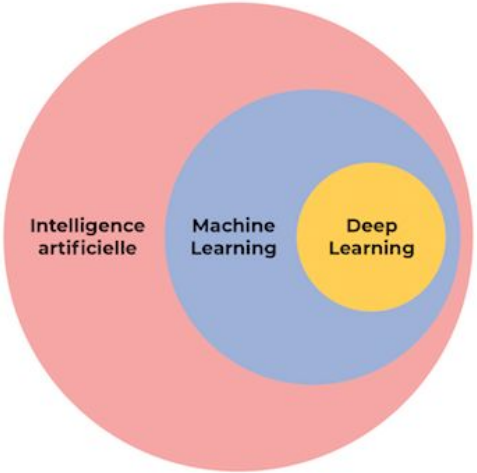
Integration: how ?



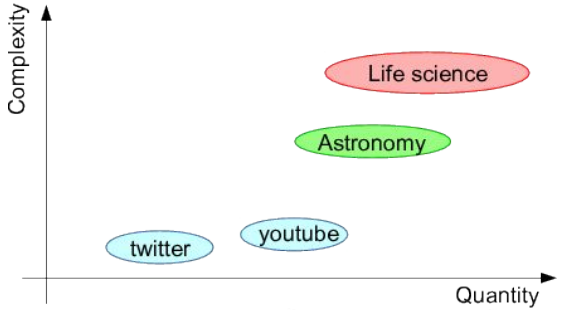


Integration: how ?





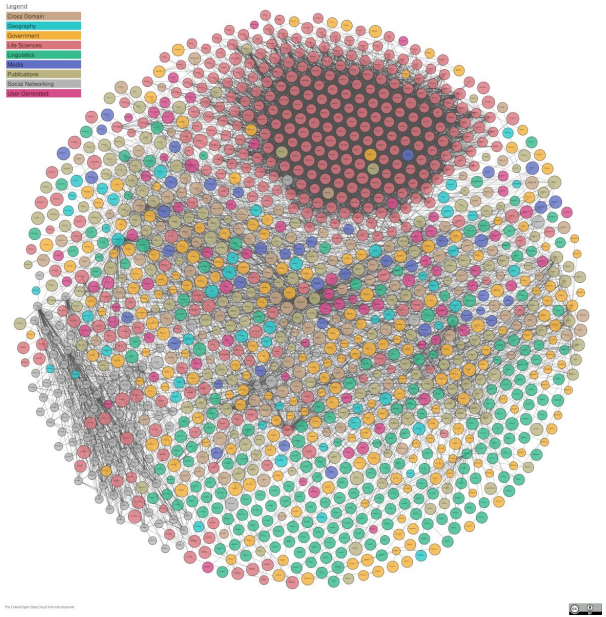
Integration with semantic web



Big Data: Astronomical or Genomical?

Zachary D. Stephens[1], Skylar Y. Lee[1], Faraz Faghri[2], Roy H. Campbell[2], Chengxiang Zhai[3], Miles J. Efron[4], Ravishankar Iyer[1], Michael C. Schatz[5]*, Saurabh Sinha[3]*, Gene E. Robinson[6]*

PLOS Biology | DOI:10.1371/journal.pbio.1002195 July 7, 2015



Life science: 1600+ reference databases

→ integrating heterogeneous data and knowledge is (badly) needed!

Editorial > Nucleic Acids Res. 2022 Jan 7;50(D1):D1-D10. doi: 10.1093/nar/gkab1195.

The 2022 Nucleic Acids Research database issue and the online molecular biology database collection

Daniel J Rigden [1], Xosé M Fernández [2]

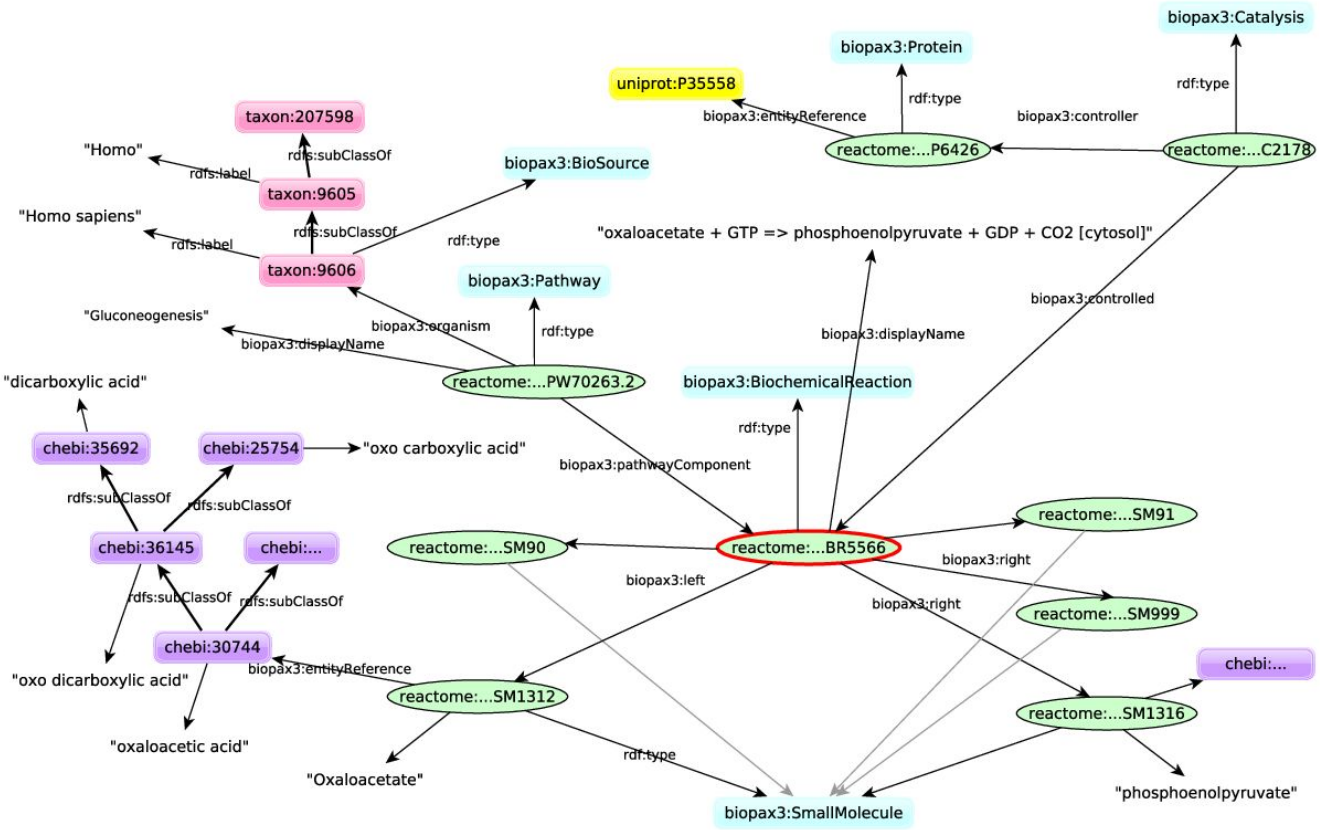
Affiliations + expand

PMID: 34986604 PMCID: PMC8728296 DOI: 10.1093/nar/gkab1195

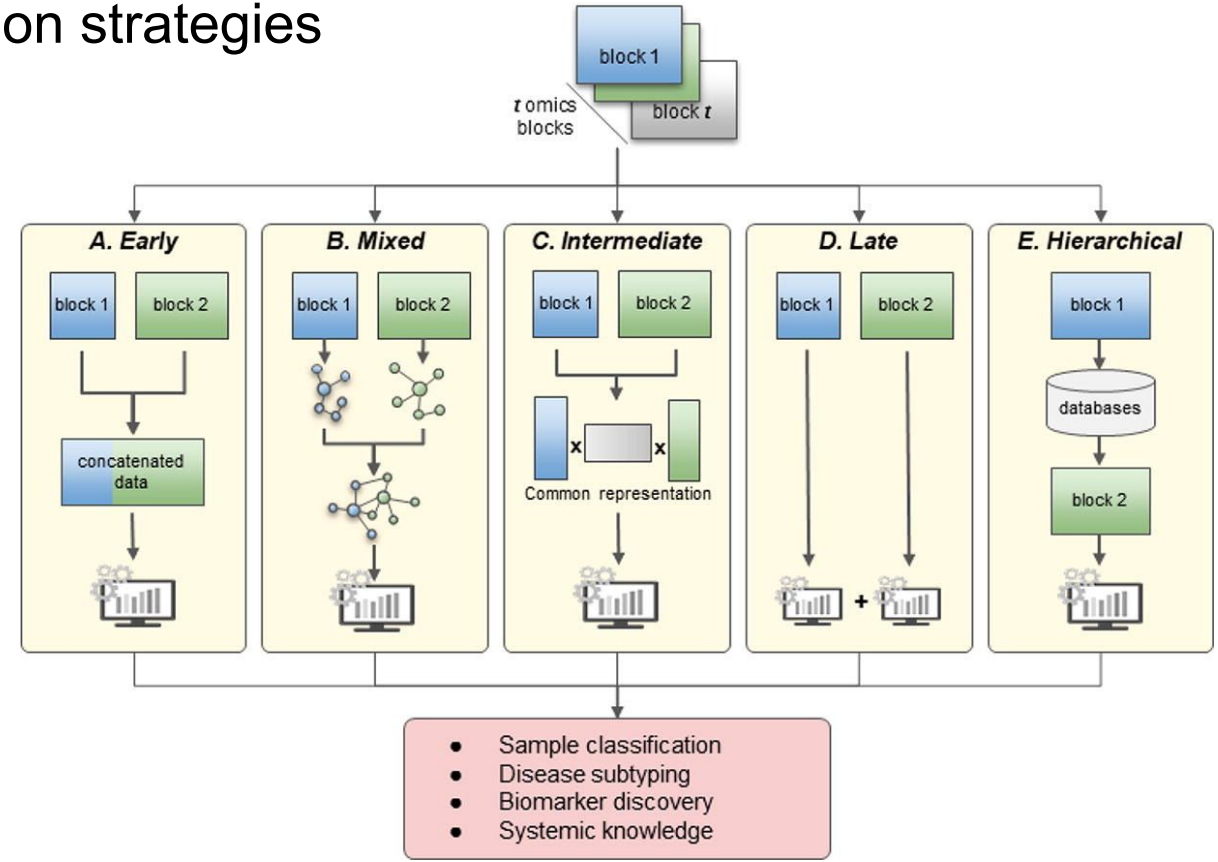
Semantic Web = framework for:

- **integrating** data and knowledge
- **querying**
- **reasoning**

Integration with semantic web

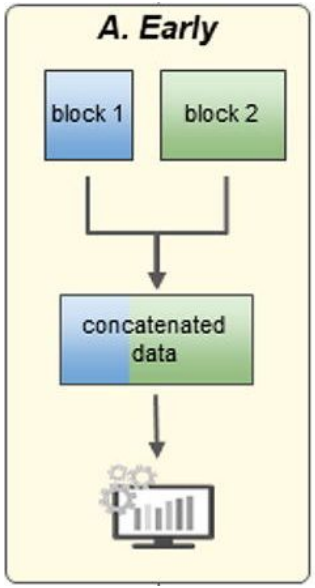


Integration strategies



Picard M. et al. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021.

Integration strategies



Concatenate every omics datasets into a single large matrix.

Pros :

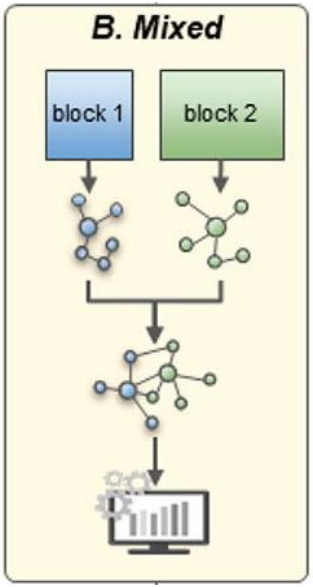
- conceptually simple
- easy implementation
- directly uncovers interactions between omics

Cons :

- technically complicated (noisy and high dimensional concatenated matrix)
- requires to have omics on the same samples or same variables
- imbalanced omics datasets
- ignores the specific data distribution of each omics

Picard M. et al. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021.

Integration strategies



Transform independently each omics dataset into a simpler representation before integration.

Pros :

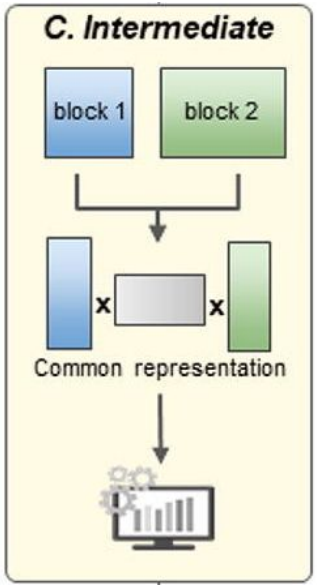
- new representation is less dimensional and less noisy
- less heterogeneity between omics
- classical approaches can be used on combined representation

Cons :

- choice of the transformation method is not trivial
- requires correspondence between variables in the new representation
- information loss during transformation

Picard M. et al. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021.

Integration strategies



Jointly integrate the multi-omics datasets without prior transformation.

Pros :

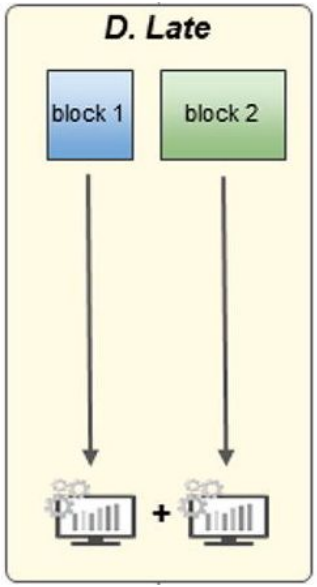
- reduce information loss
- discover the joint inter-omics structure
- highlight the complementary information in each omics

Cons :

- could require robust pre-processing step to reduce heterogeneity
- common latent space assumption

Picard M. et al. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021.

Integration strategies



Apply machine learning models separately on each omics dataset and then combine results.

Pros :

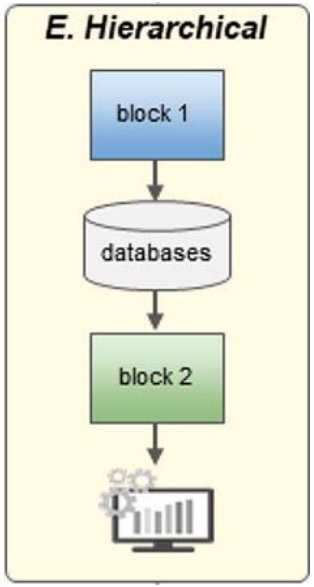
- avoid (numerous) challenges of direct omics integration
- use tools designed specifically for each omics
- classical approaches can be used to combine results

Cons :

- cannot capture inter-omics interactions
- complementarity information between omics is not exploited

Picard M. et al. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021.

Integration strategies



Include prior knowledge of omics relationships.

Pros :

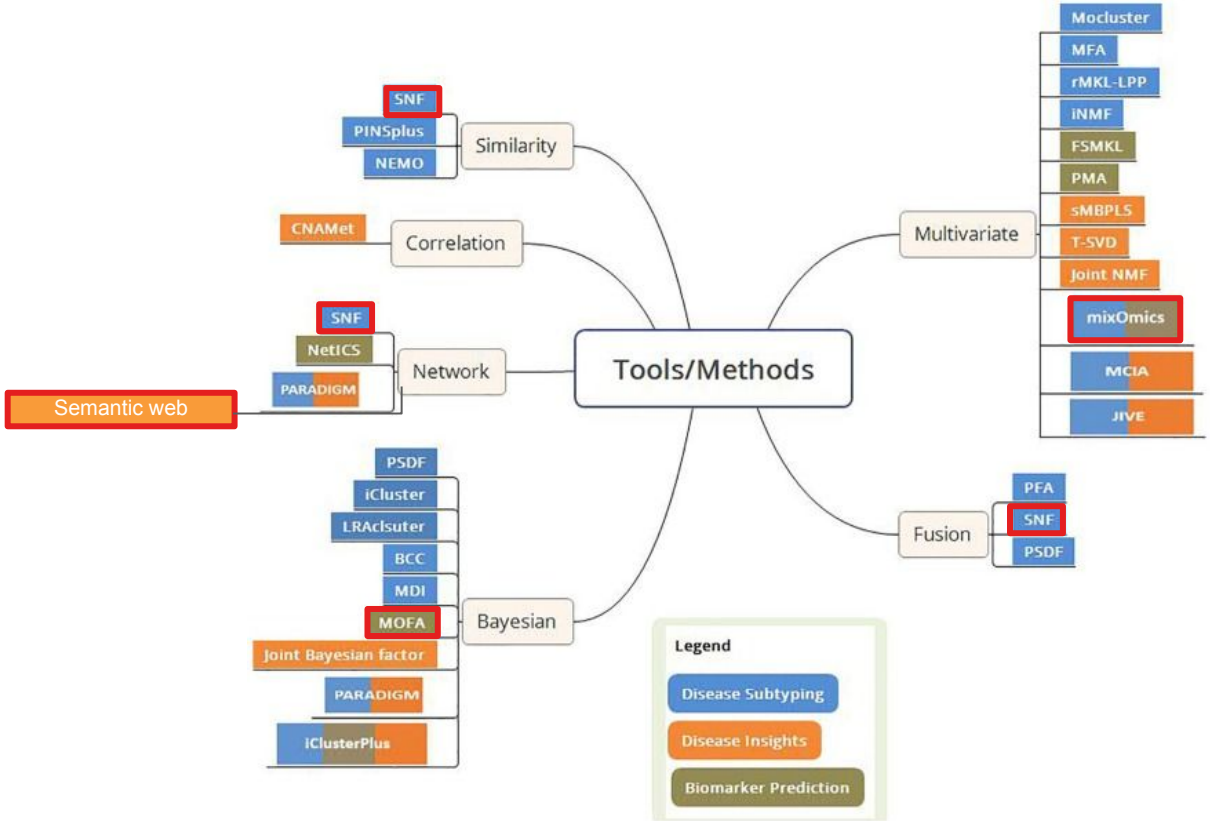
- reduced complexity (sequential integration)
- integrate external knowledge

Cons :

- less generic than previous strategies

Picard M. et al. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021.

Integration approaches



Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. Bioinform Biol Insights. 2020

Limits of integration approaches

Integration approaches are not magic!

You will still need to:

- carefully check design and confounding factors
- perform specific data pre-processing for each omic
- impute missing values* (different meaning → different strategy)
- choose your integration strategy based on your objective and your data (ex. matching between omics) → still no standard pipelines
- some omics bring more noise than answers

Web-applications

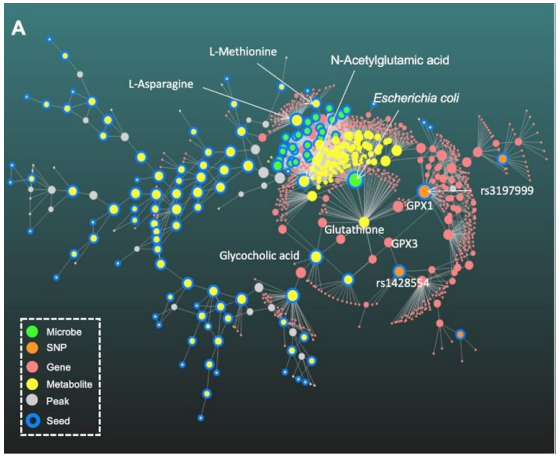
PaintOmics *(T. Liu et al. PaintOmics 4: new tools for the integrative analysis of multi-omics datasets supported by multiple pathway databases, Nucleic Acids Research, Volume 50, Issue W1, 2022.)*

3Omics *(K. Tien-Chueh et al. 3Omics: A web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. BMC systems biology. 7. 64, 2013)*

XCMSOnline *(EM. Forsberg et al. Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online. Nat Protoc. 13(4):633-651, 2018)*

Galaxy-P project *(Galaxy-P Project. galaxyp.org.)*

OmicsNet *(G. Zhou et al., OmicsNet 2.0: a web-based platform for multi-omics integration and network visual analytics, Nucleic Acids Research, Volume 50, Issue W1, 5 , 2022.)*



...

References

Misra BB, Langefeld CD, Olivier M, Cox LA. Integrated Omics: Tools, Advances, and Future Approaches. J Mol Endocrinol, 2018.

Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. Bioinform Biol Insights, 2020.

Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J., 2021.

Benfeitas R, Viklund J, Ash706, Robinson J, Manoharan L, Fasterius E, Oskolkov N, Francis R, Anton M. (2020). NBISweden/workshop_omics_integration: Lund, 2020/10/05 (Version course2010). Zenodo. https://doi.org/10.5281/zenodo.4084627

Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, Milanesi L. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics, 17 Suppl 2(Suppl 2):15, 2016.

Ritchie, M., Holzinger, E., Li, R. et al. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet 16, 85–97, 2015.