Omics data integration and variable selection

ANR MetaboHUB 2.0 project, WP1.4

Bougel Céline

21 novembre 2024

Job title: Research engineer in data integration

Project: MetaboHUB (ANR)

- WP1: Scaling Up: towards large cohort studies
- task 4: Data fusion and integration for large-scale investigations

Contract period: Recruitment 14 months (until August + extension 2 months)

2 missions → ConsensusOPLS

↘ Variable selection

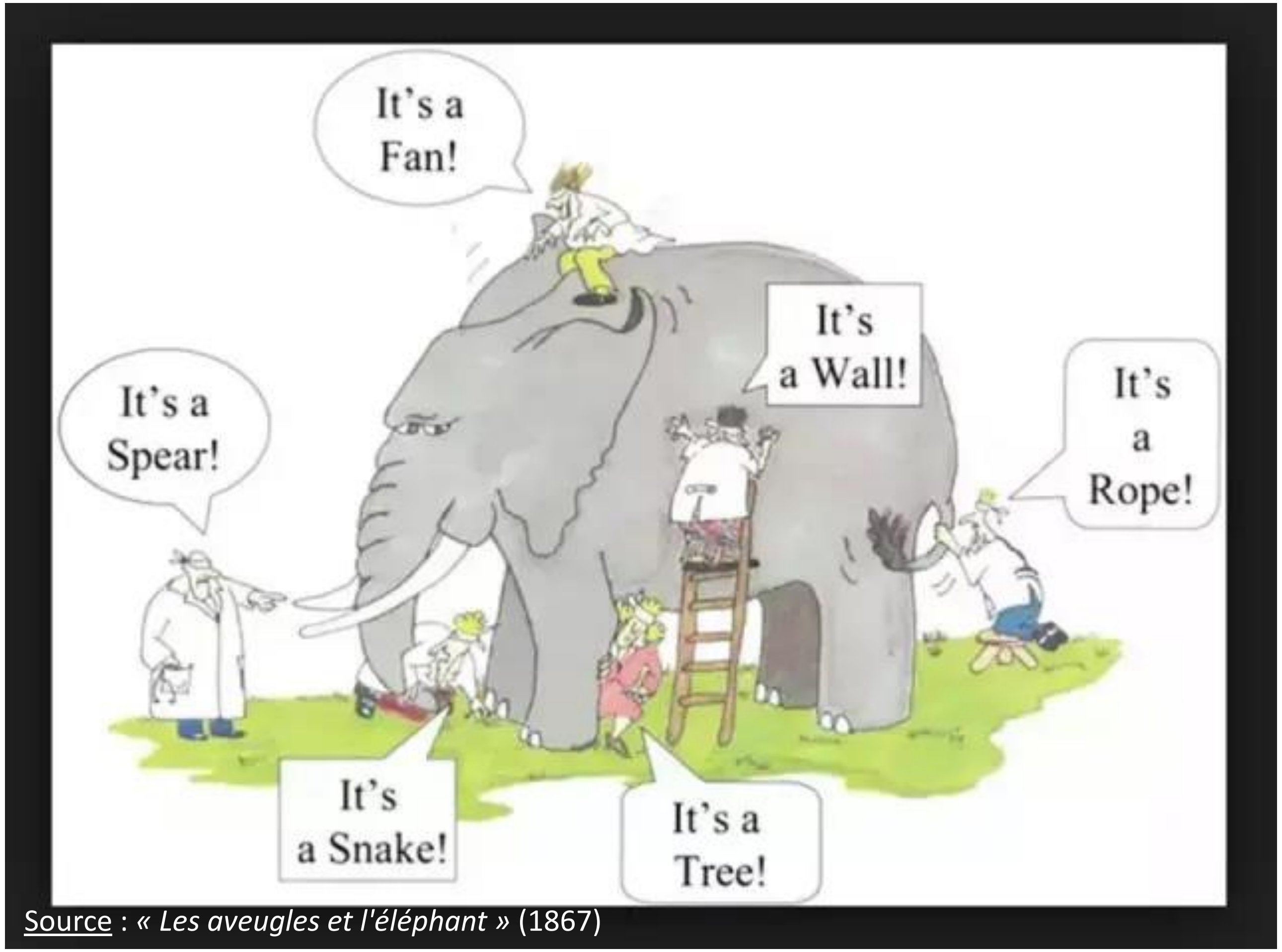
Where does the need for data integration come from?

To understand and model the complexity of biological systems.

Where does the need for data integration come from?

To understand and model the complexity of biological systems.

Multi-omics
integration enables us to

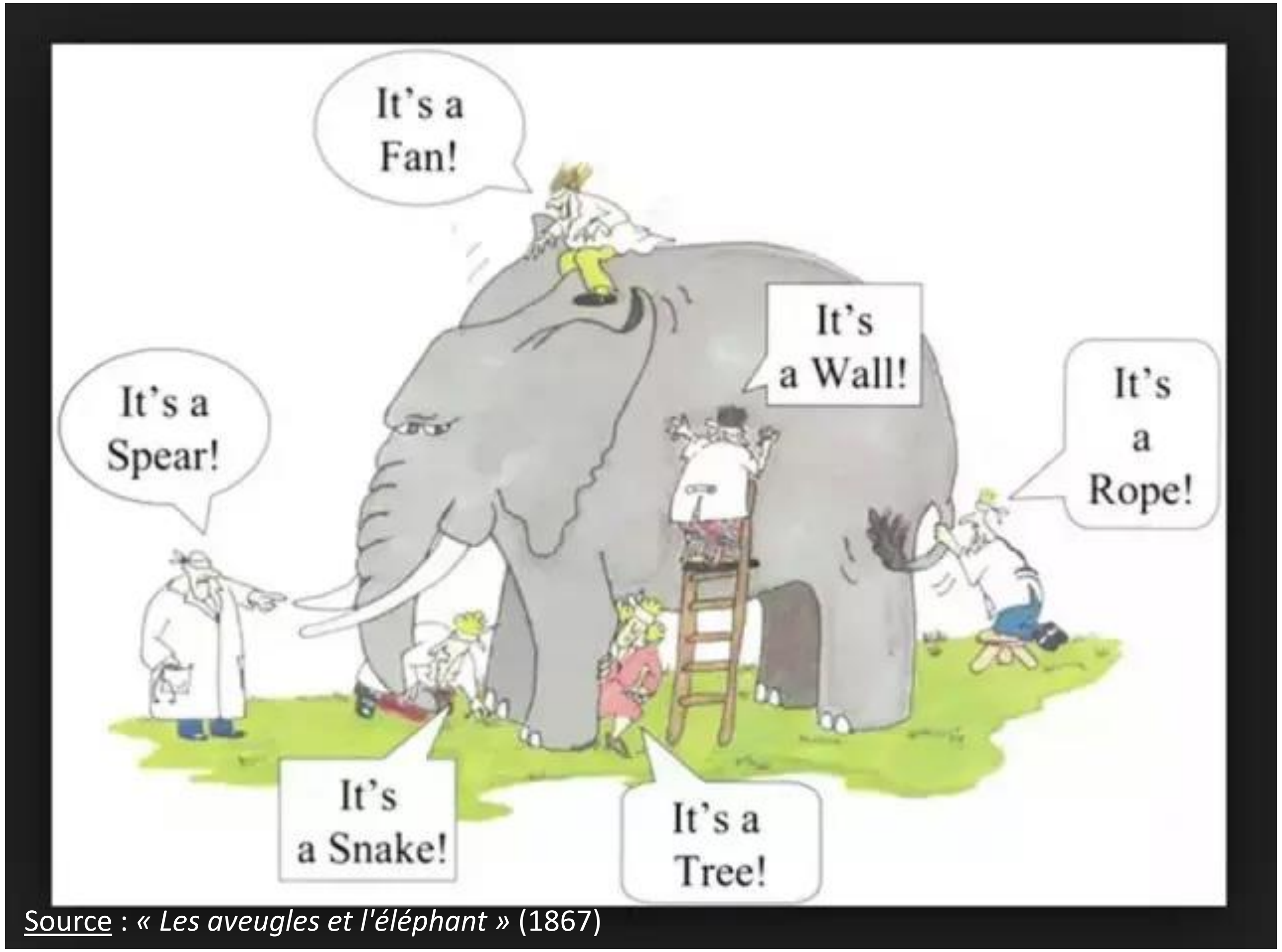


Source : *« Les aveugles et l'éléphant »* (1867)

Where does the need for data integration come from?

To understand and model the complexity of biological systems.

Multi-omics and multi-techniques integration enables us to

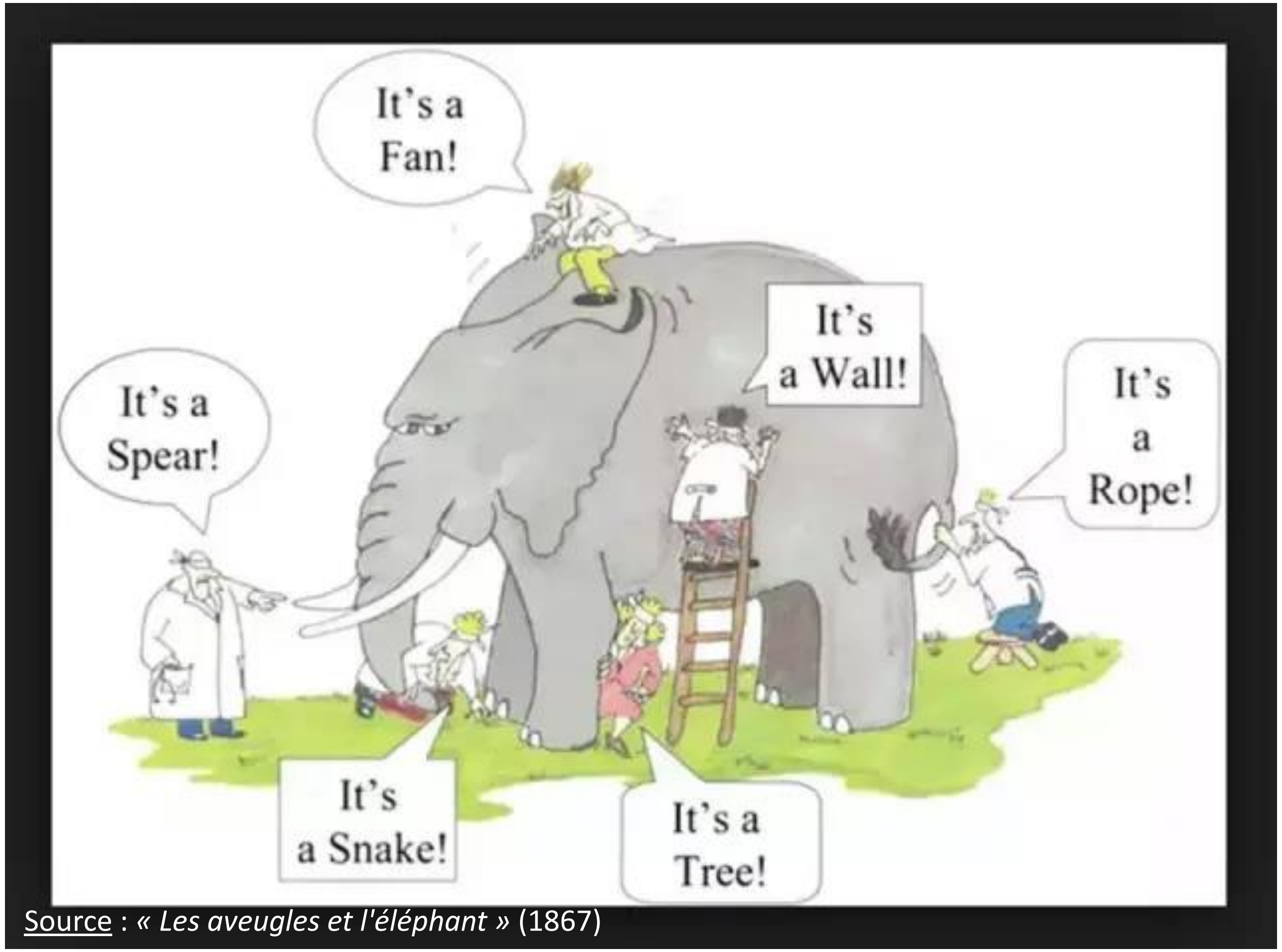


Source : *« Les aveugles et l'éléphant »* (1867)

Where does the need for data integration come from?

To understand and model the complexity of biological systems.

Multi-omics and multi-techniques integration enables us to build a **global and interconnected vision** of biological responses, impossible to obtain with a single data source.



Source : *« Les aveugles et l'éléphant »* (1867)

What is data integration?

Several sources/ Blocks

Comprehensive modelling

What is data integration?

Several sources/ Blocks

Raw data

Comprehensive modelling

Usable information

What is data integration?

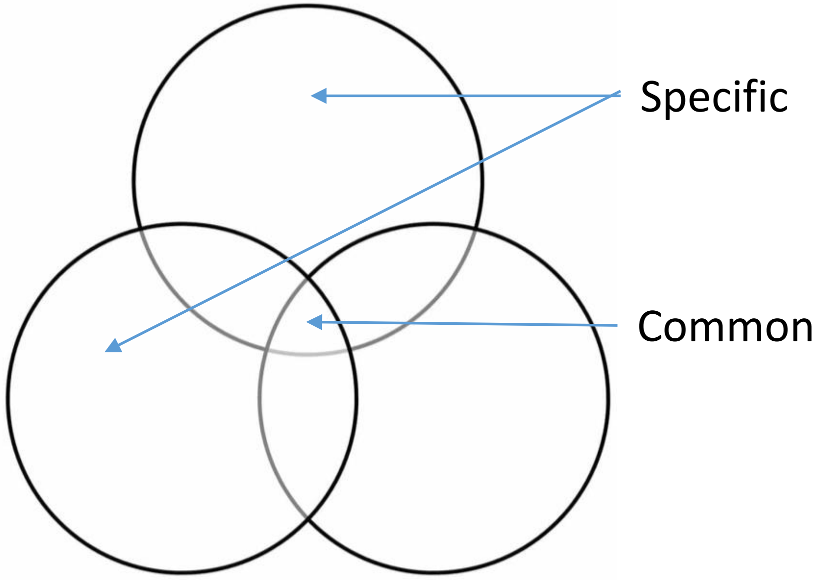
Several sources/ Blocks

Raw data

Comprehensive modelling

Usable information

Why?



What is data integration?

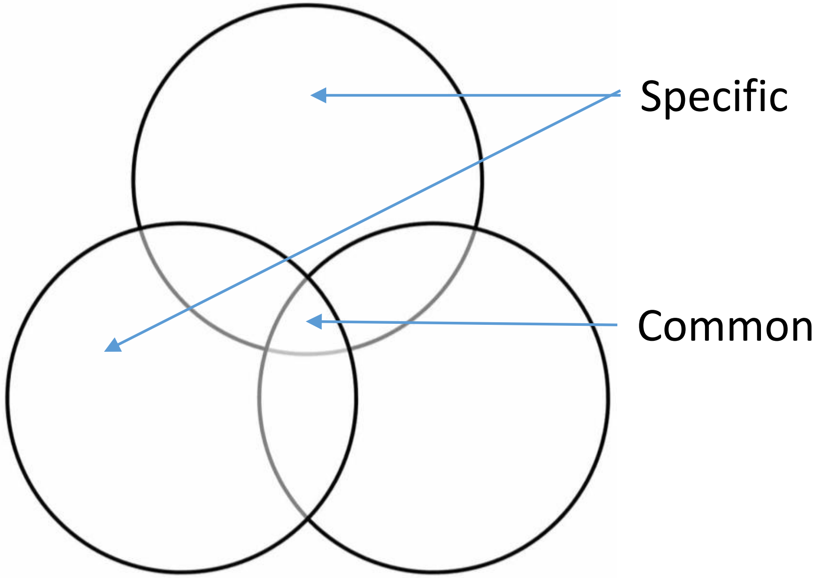
Several sources/ Blocks

Raw data

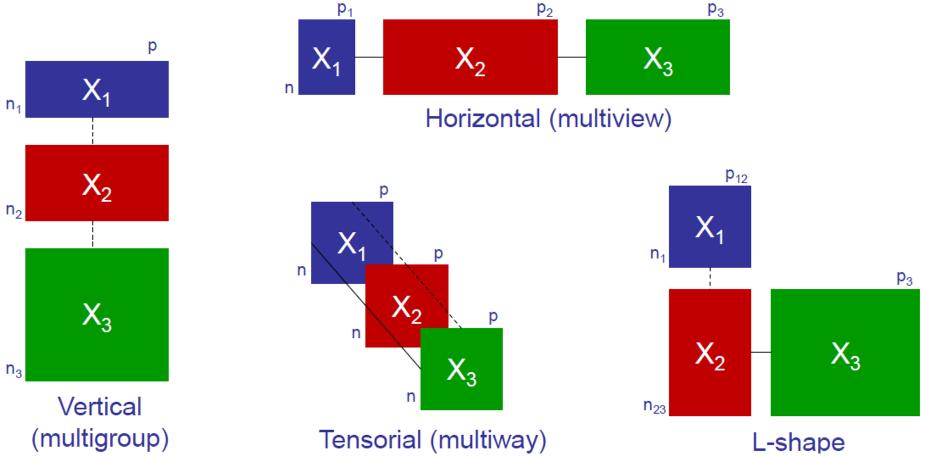
Comprehensive modelling

Usable information

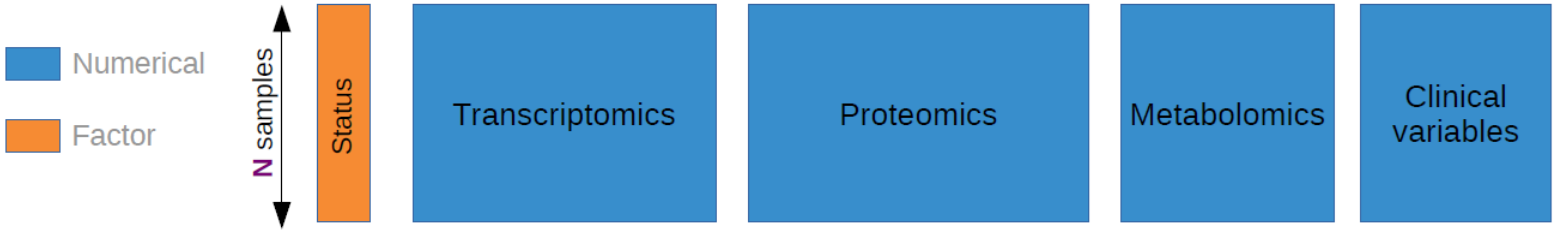
Why?



How?



How?



🟦 Numerical

🟧 Factor

N samples

Status

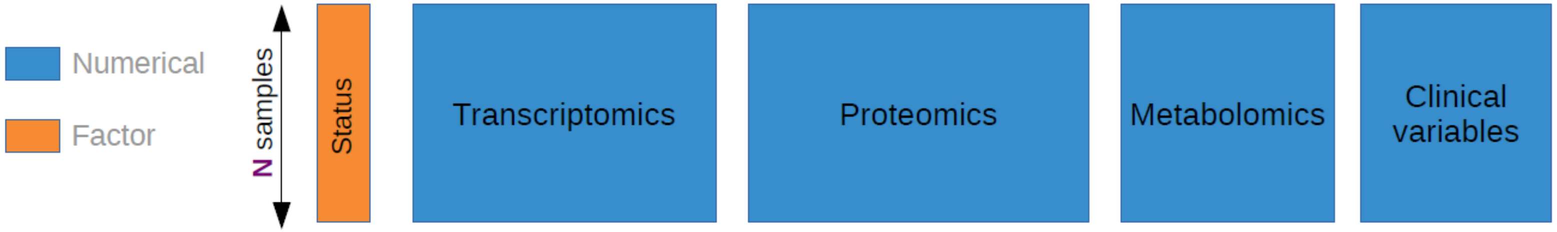
Transcriptomics

Proteomics

Metabolomics

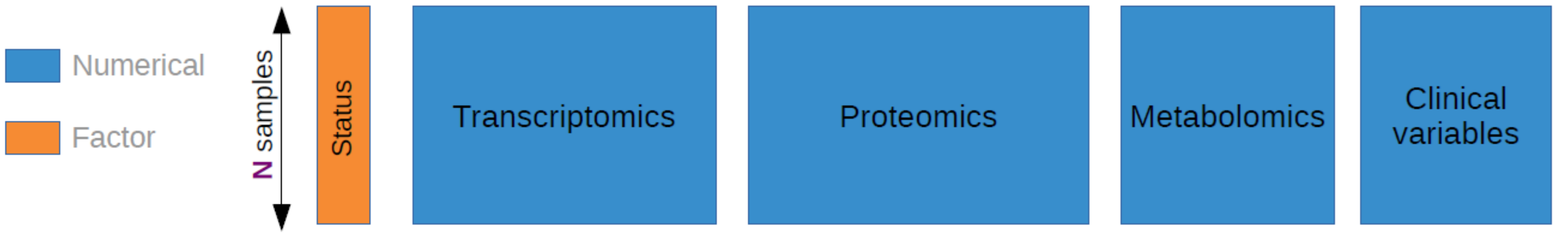
Clinical variables

How?



- ○ Variables >> subjects
- ○ Colinearity
- ○ Heterogeneity
  - Numbers of var
  - Magnitude

How?



🟦 Numerical

🟧 Factor

N samples

Status

Transcriptomics

Proteomics

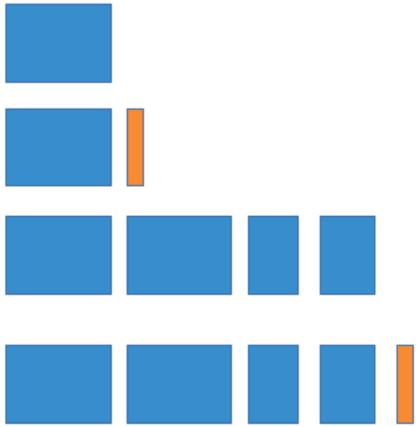
Metabolomics

Clinical variables

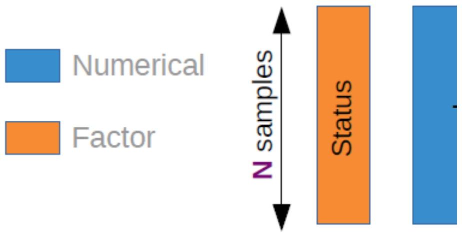
- Variables >> subjects
- Colinearity
- Heterogeneity
  - Numbers of var
  - Magnitude

Which? Multivariate analysis and dimensionality reduction

- Multivariate unsupervised
  Principal Components Analysis (PCA)
- Multivariate supervised
  Projection to Latent Structure Discriminat Analysis (PLS-DA)
- Multi-block unsupervised
  Canonical Correlation Analysis (CCA) or PLS (2 blocks), Generalized CCA (>2 blocks)
- Multi-block supervised
  Generalized Canonical Correlation Discriminant Analysis (GCC-DA)



How?



Which? Multivariate ar

- Multivariate unsupervised
  Principal Components Analysis (PCA)
- Multivariate supervised
  Projection to Latent Structure Discriminat Analysis (PLS-DA)
- Multi-block unsupervised
  Canonical Correlation Analysis (CCA) (
  PLS (2 blocks), Generalized CCA (>2 l
- Multi-block supervised
  Generalized Canonical Correlation Discriminant Analysis (GCC-DA)

Metabolomic data analysis with chemometrics Boccard and Rudaz (2013)

CHEMOMETRICS

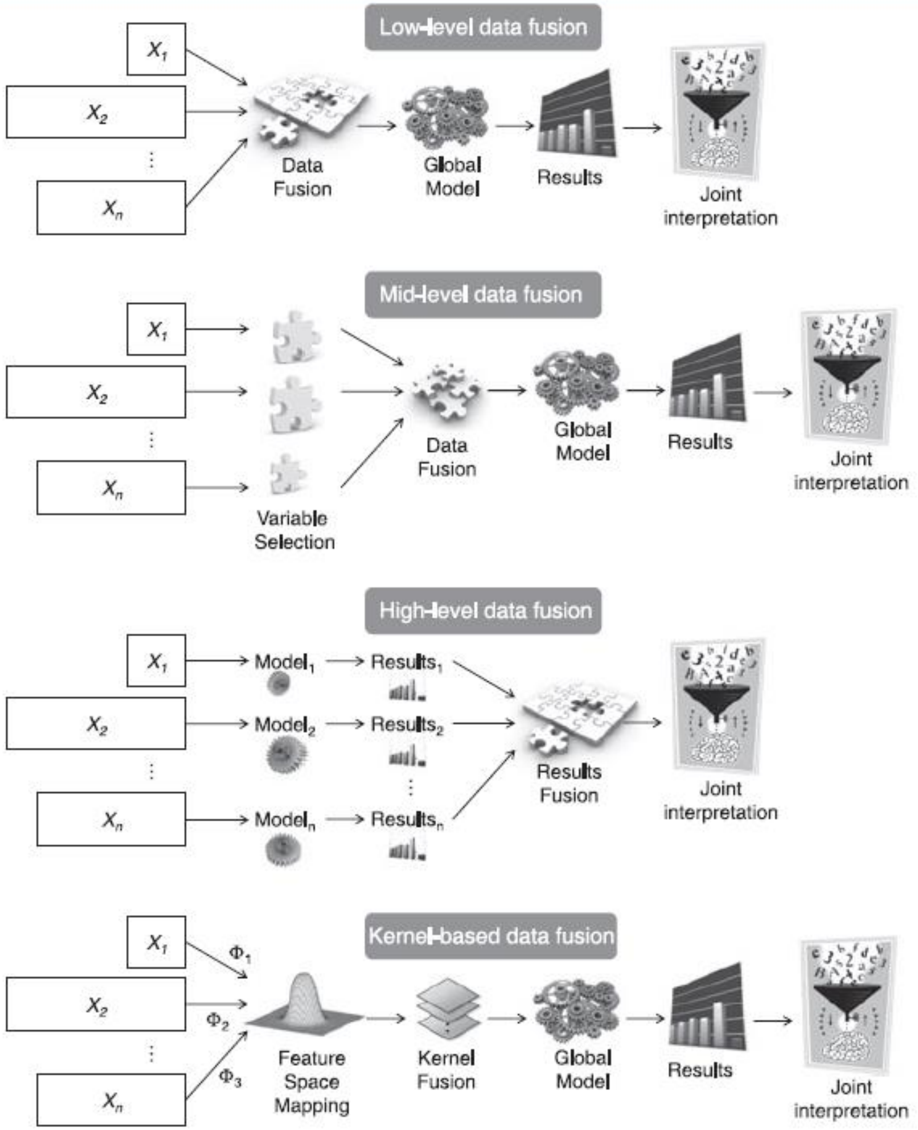
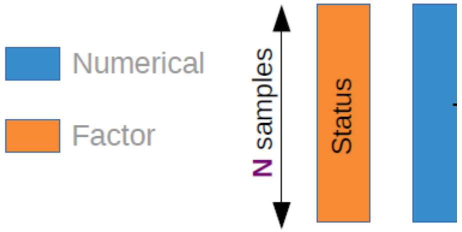


Figure 3. The four main approaches to data fusion.

- ○ Variables >> subjects
- ○ Colinearity
- ○ Heterogeneity
  - Numbers of var
  - Magnitude

How?



Which? Multivariate ar

- Multivariate unsupervised
  Principal Components Analysis (PCA)
- Multivariate supervised
  Projection to Latent Structure Discriminat Analysis (PLS-DA)
- Multi-block unsupervised
  Canonical Correlation Analysis (CCA) (
  PLS (2 blocks), Generalized CCA (>2 l
- Multi-block supervised
  Generalized Canonical Correlation Discriminant Analysis (GCC-DA)

Metabolomic data analysis with chemometrics Boccard and Rudaz (2013) CHEMOMETRICS

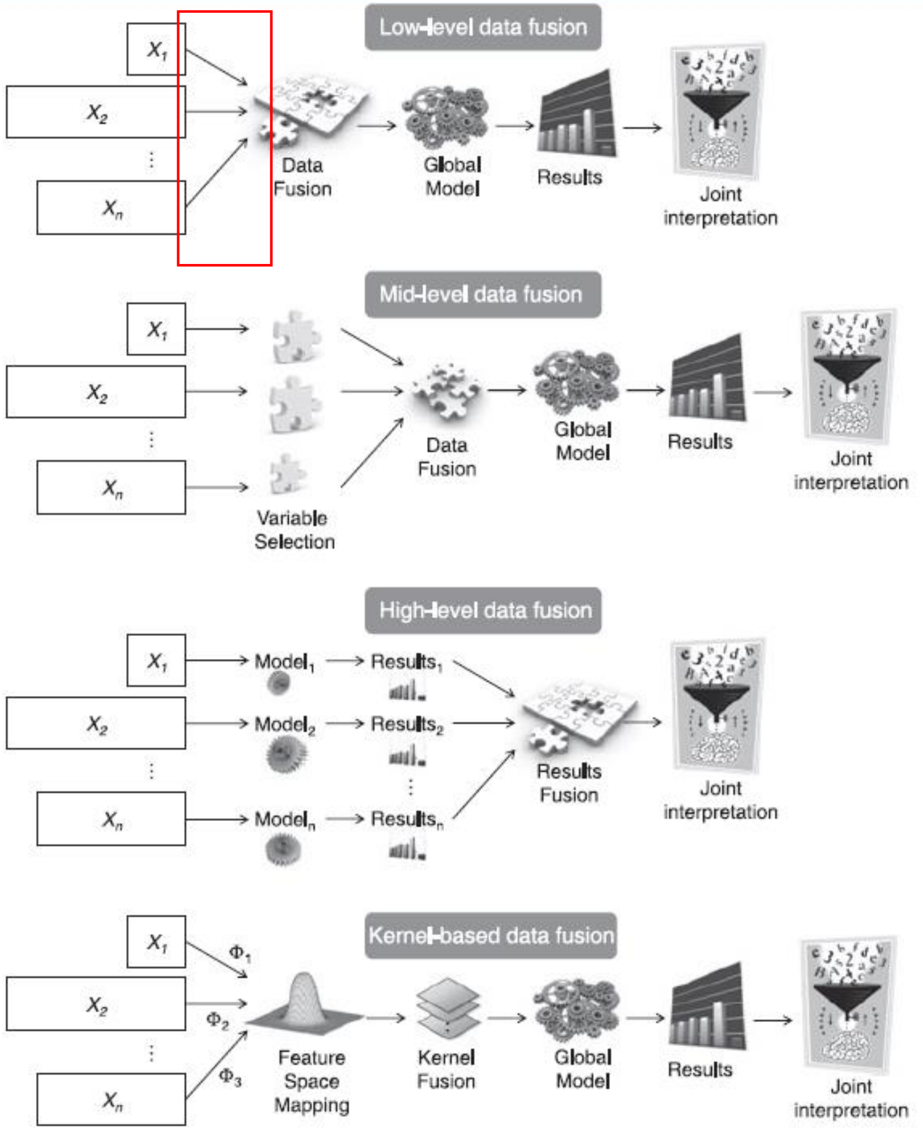
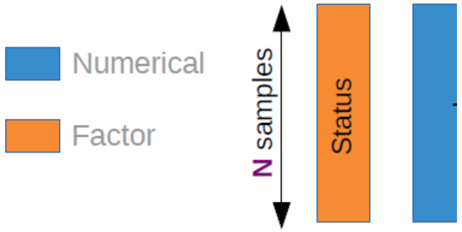


Figure 3. The four main approaches to data fusion.

- ○ Variables >> subjects
- ○ Colinearity
- ○ Heterogeneity
  - Numbers of var
  - Magnitude

How?



Which? Multivariate ar

- Multivariate unsupervised
  Principal Components Analysis (PCA)
- Multivariate supervised
  Projection to Latent Structure Discriminat Analysis (PLS-DA)
- Multi-block unsupervised
  Canonical Correlation Analysis (CCA)
  PLS (2 blocks), Generalized CCA (>2
- Multi-block supervised
  Generalized Canonical Correlation Discriminant Analysis (GCC-DA)

Metabolomic data analysis with chemometrics Boccard and Rudaz (2013)

CHEMOMETRICS

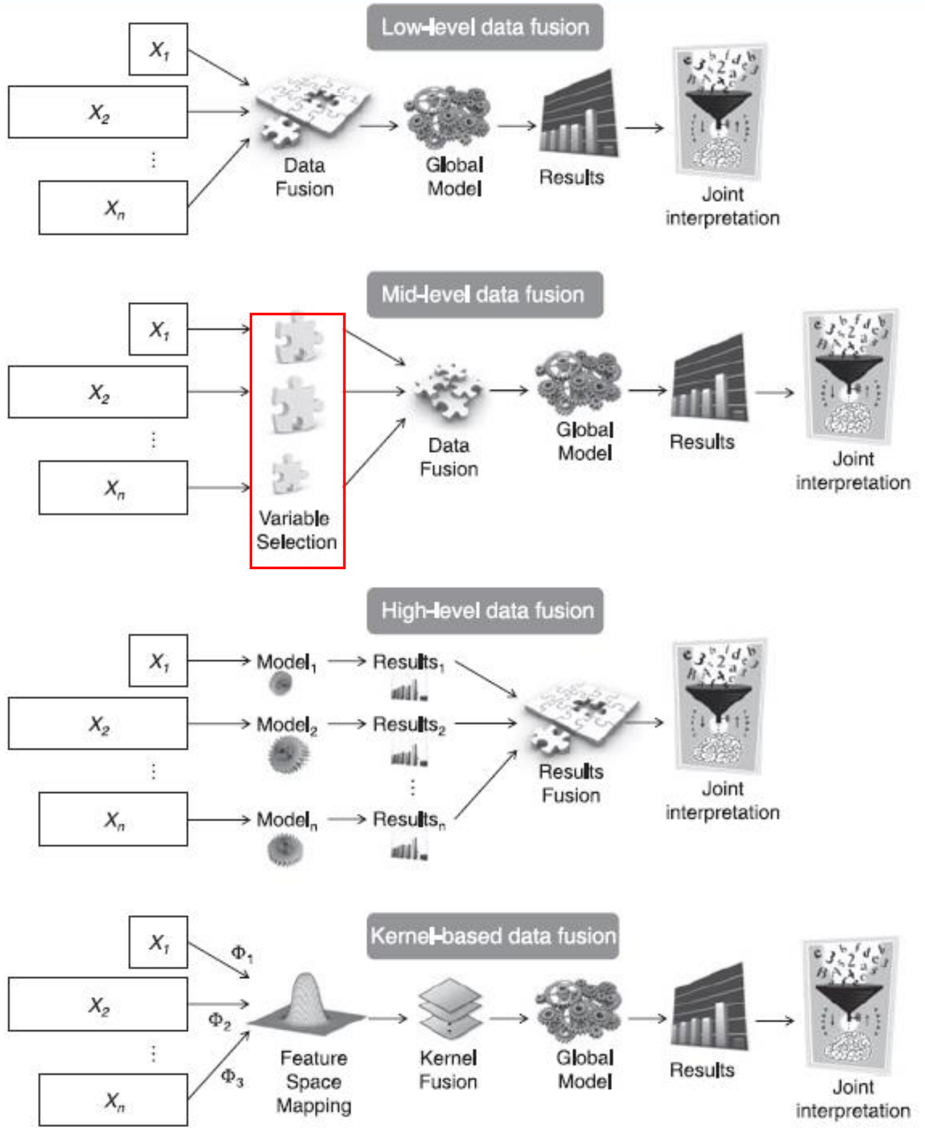


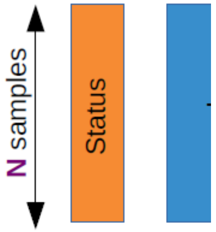
Figure 3. The four main approaches to data fusion.

- Variables >> subjects
- Colinearity
- Heterogeneity
  - Numbers of var
  - Magnitude

How?

🟦 Numerical

🟧 Factor



Which? Multivariate ar

- Multivariate unsupervised
  Principal Components Analysis (PCA)
- Multivariate supervised
  Projection to Latent Structure Discriminat Analysis (PLS-DA)
- Multi-block unsupervised
  Canonical Correlation Analysis (CCA) (
  PLS (2 blocks), Generalized CCA (>2 l
- Multi-block supervised
  Generalized Canonical Correlation Discriminant Analysis (GCC-DA)

Metabolomic data analysis with chemometrics Boccard and Rudaz (2013) CHEMOMETRICS

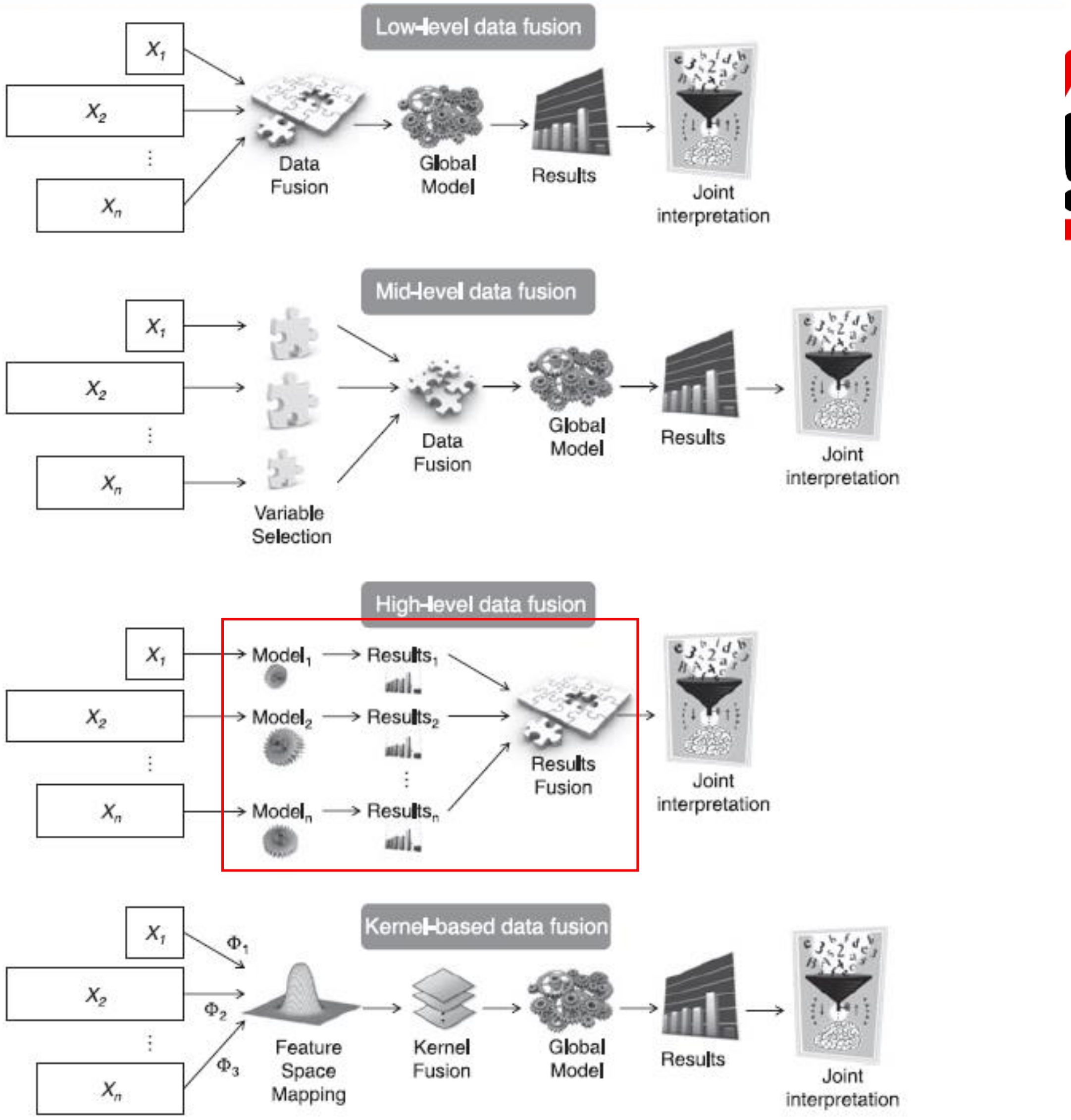
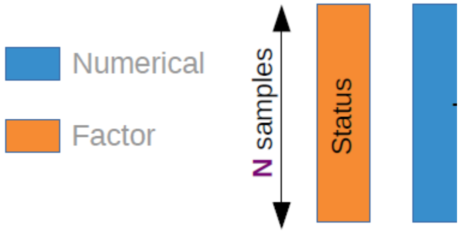


Figure 3. The four main approaches to data fusion.

- ○ Variables >> subjects
- ○ Colinearity
- ○ Heterogeneity
  - Numbers of var
  - Magnitude

How?



Which? Multivariate ar

- Multivariate unsupervised
  Principal Components Analysis (PCA)
- Multivariate supervised
  Projection to Latent Structure Discriminat Analysis (PLS-DA)
- Multi-block unsupervised
  Canonical Correlation Analysis (CCA) (PLS (2 blocks), Generalized CCA (>2 l
- Multi-block supervised
  Generalized Canonical Correlation Discriminant Analysis (GCC-DA)

Metabolomic data analysis with chemometrics Boccard and Rudaz (2013) CHEMOMETRICS

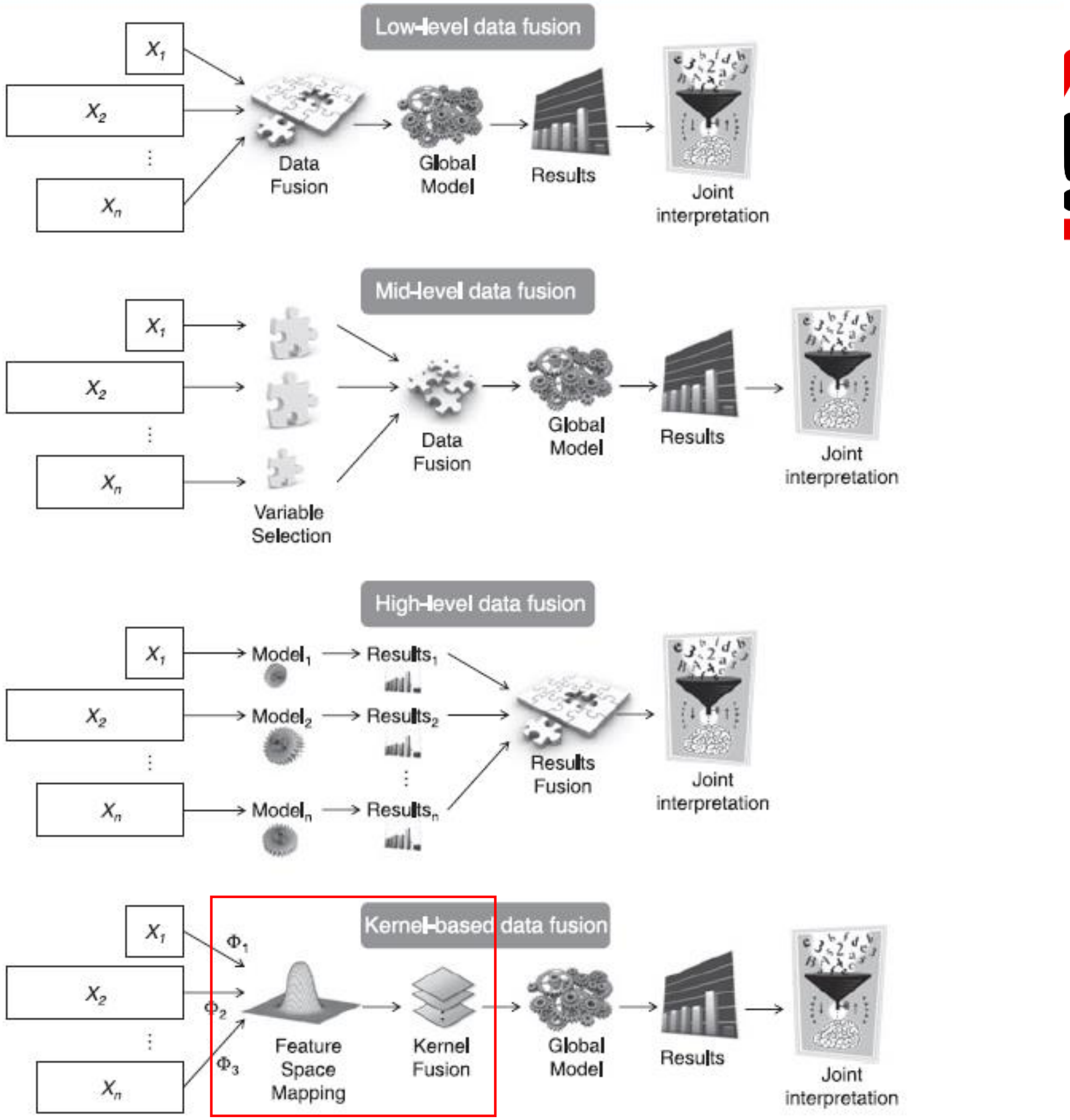
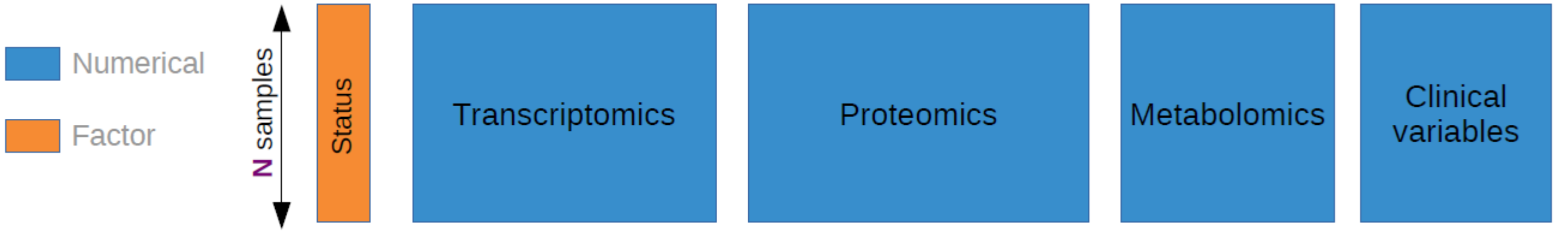


Figure 3. The four main approaches to data fusion.

- ○ Variables >> subjects
- ○ Colinearity
- ○ Heterogeneity
  - Numbers of var
  - Magnitude

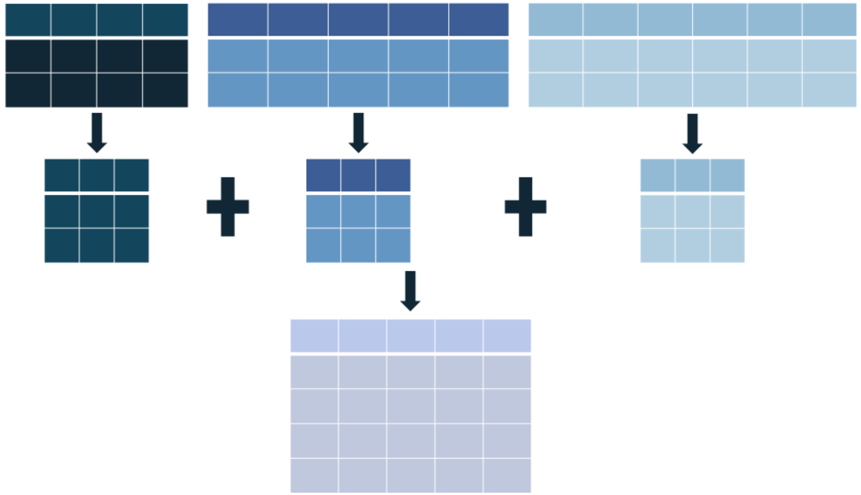
How?



- Variable >> subjects
- Colinearity
- Heterogeneity
  - Numbers of var
  - Magnitude

Which? Multivariate analysis and dimensionality reduction

ConsensusOPLS



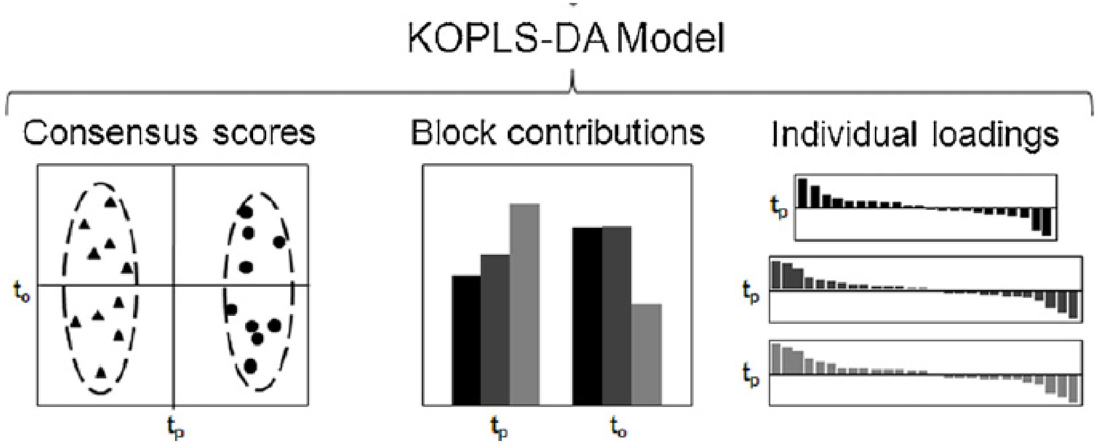
Data matrix

Kernel = transforming data into another space
to improve data representation capability
Metric = Similarity
Types = linear

Consensus kernel = weighted sum of kernels
by modified RV coefficient
with the response matrix

kOPLS(-DA) model
on the
meta-kernel
+ cross validation
# opt comp

And what do we do with it?



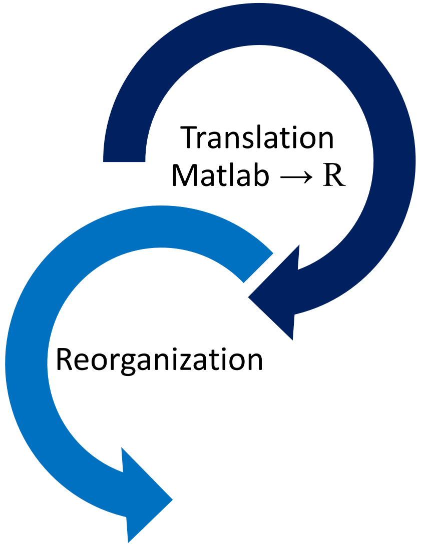
Boccard et Rutledge (2013)

Key indicators :

- ✓ R^2 = the R-squared coefficient
- ✓ Q^2 = the Stone-Geisser Q^2 coefficient
- ✓ DQ^2 = the discriminant Q^2 index
- ✓ Permutations tests

ConsensusOPLS: An R package for Multi-Block Data Fusion

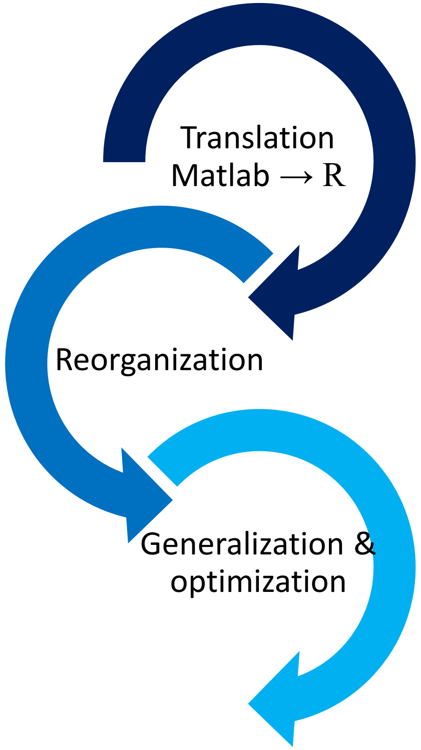
ConsensusOPLS: An R package for Multi-Block Data Fusion



Script-to-script conversion ...with all that this implies (format of objects to be adapted, calling/using functions, ...)

From script to package

ConsensusOPLS: An R package for Multi-Block Data Fusion



Translation Matlab → R

Reorganization

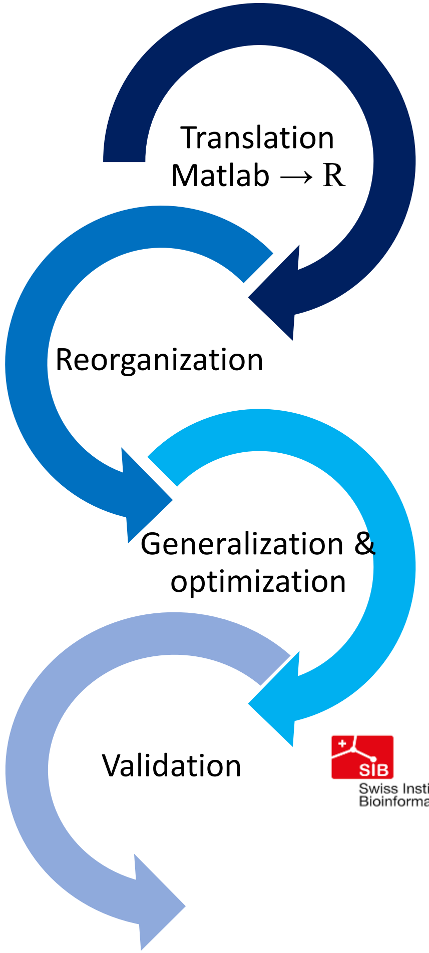
Generalization & optimization

Script-to-script conversion ...with all that this implies (format of objects to be adapted, calling/using functions, ...)

From script to package

	MATLAB .	R .
Response & Comp pred	Only 2 class 1	≥ 2 classes ≥ 1
Kernel	Linear (= polynomial order 1)	Polynomial order ≥ 1 (non linear) + Gaussian
Permutations	Sequential	Parallelized
Outputs		Add Variable importance in Projection (VIP) + synthetics indicators
User friendly		Add main results with print(model)

ConsensusOPLS: An R package for Multi-Block Data Fusion



Script-to-script conversion ...with all that this implies (format of objects to be adapted, calling/using functions, ...)

From script to package

	MATLAB .	R .
Response & Comp pred	Only 2 class 1	≥ 2 classes ≥ 1
Kernel	Linear (= polynomial order 1)	Polynomial order ≥ 1 (non linear) + Gaussian
Permutations	Sequential	Parallelized
Outputs		Add Variable importance in Projection (VIP) + synthetics indicators
User friendly		Add main results with print(model)

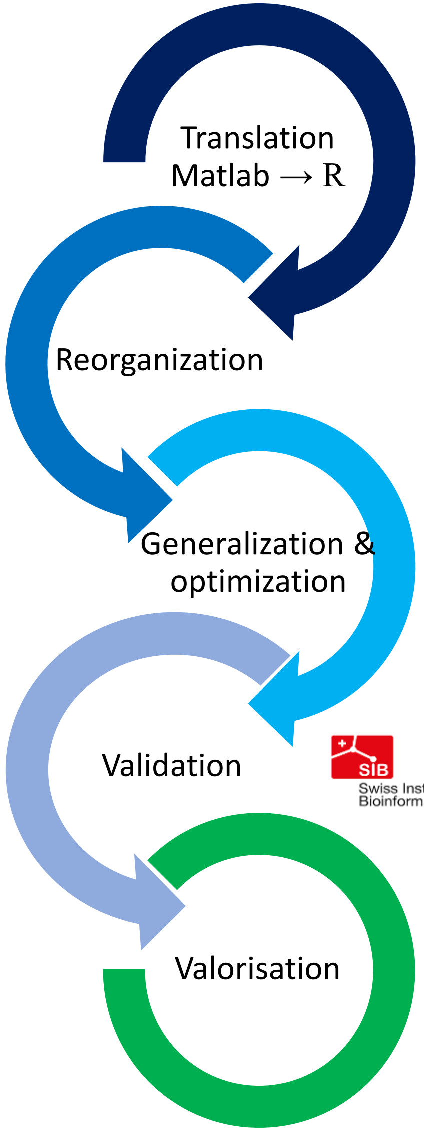
SIB Swiss Institute of Bioinformatics

On demo data (14 subjects, 3 blocs (metabolomic = 150, microarray = 200 and proteomics = 100 variables): completed

On real data:

→ internal project data set (OCTOPUS): completed
→ ProMetIs metabolomics data: completed
→ Similarity Network Fusion (SNF) data: in progress

ConsensusOPLS: An R package for Multi-Block Data Fusion



Script-to-script conversion ...with all that this implies (format of objects to be adapted, calling/using functions, ...)

From script to package

On demo data (14 subjects, 3 blocs (metabolomic = 150, microarray = 200 and proteomics = 100 variables): completed

On real data:

→ internal project data set (OCTOPUS): completed

→ ProMetIs metabolomics data: completed

→ Similarity Network Fusion (SNF) data: in progress

CRAN publication

Poster RFMF (june 2024, Saint-Malo)

Poster JOBIM **(june 2024, Toulouse)**

Application Note (in writing)

	MATLAB .	R .
Response & Comp pred	Only 2 class 1	≥ 2 classes ≥ 1
Kernel	Linear (= polynomial order 1)	Polynomial order ≥ 1 (non linear) + Gaussian
Permutations	Sequential	Parallelized
Outputs		Add Variable importance in Projection (VIP) + synthetics indicators
User friendly		Add main results with print(model)

ConsensusOPLS: An R package for Multi-Block Data Fusion

Which data?

ProMetIS: proteomics and metabolomics data integration [1]

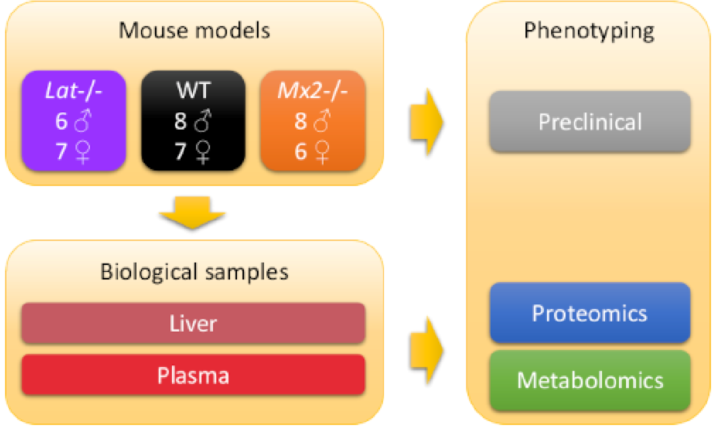
Post-processed data – mass spectrometry datasets

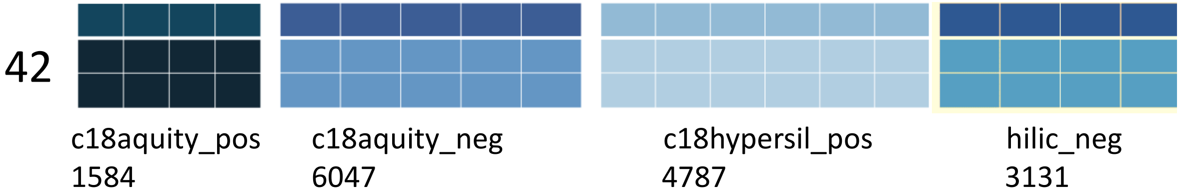
N = 42 mice

Y = Mutant (Lat or Mx2)
vs Control (WT)

Biological samples = plasma

Phenotyping = metabolomic





[1] Imbert, A., Rompais, M., Selloum, M. *et al.* ProMetIS, deep phenotyping of mouse models by combined proteomics and metabolomics analysis. *Sci Data* **8**, 311 (2021).

ConsensusOPLS: An R package for Multi-Block Data Fusion

Which data?

ProMetIS: proteomics and metabolomics data integration [1]

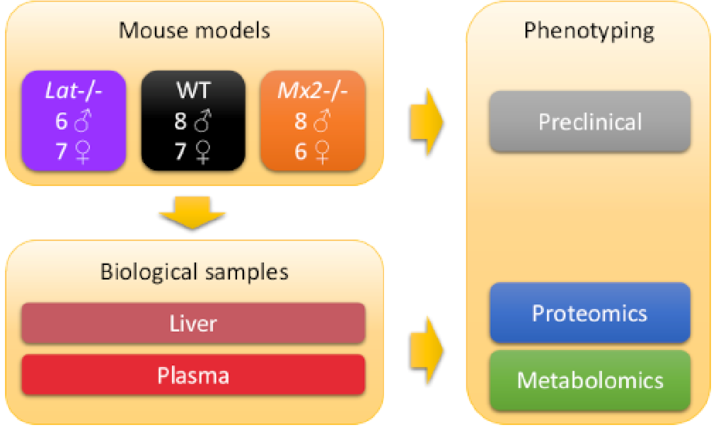
Post-processed data – mass spectrometry datasets

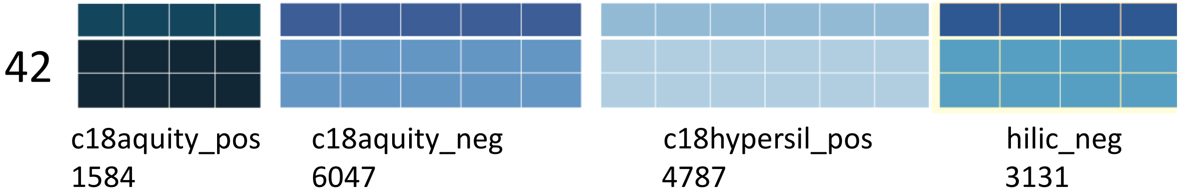
N = 42 mice

Y = Mutant (Lat or Mx2)
vs Control (WT)

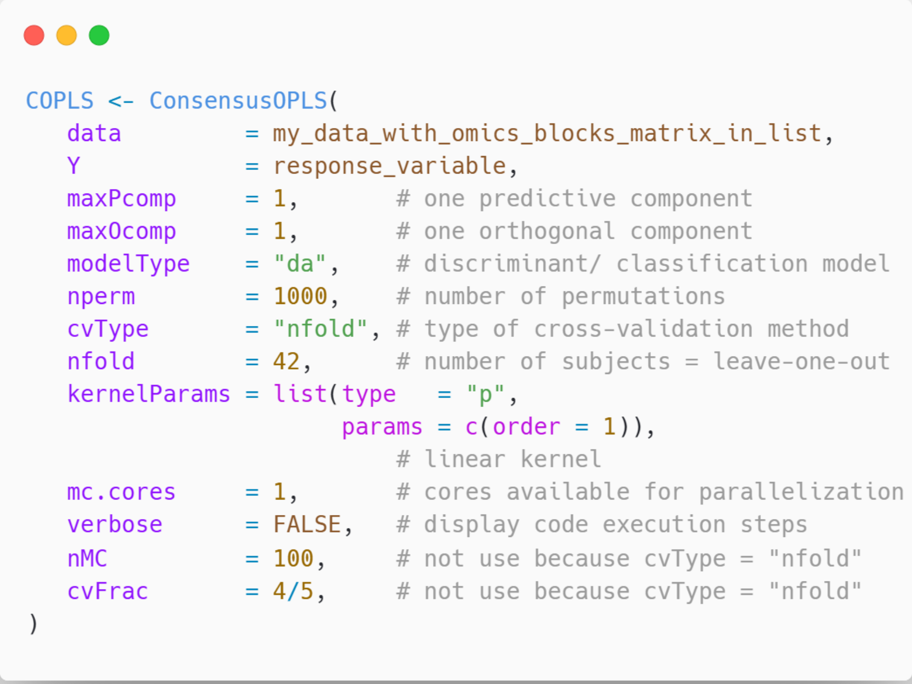
Biological samples = plasma

Phenotyping = metabolomic





How to use?

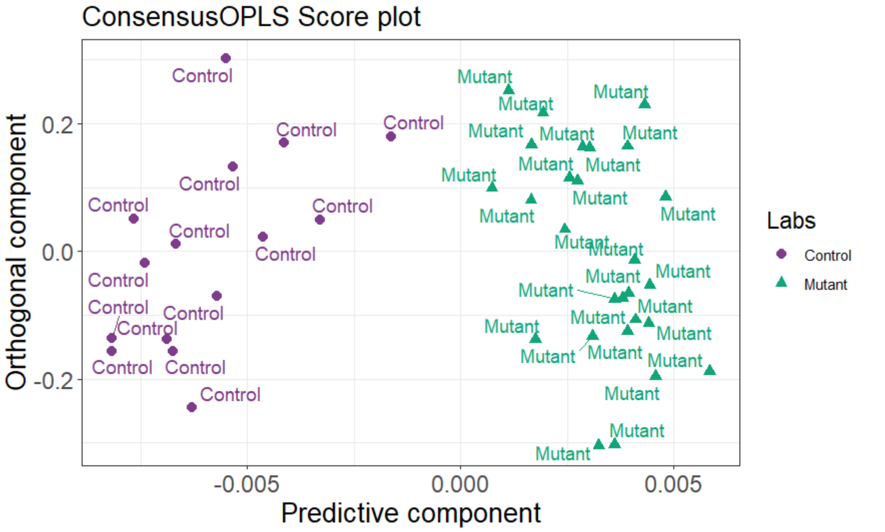


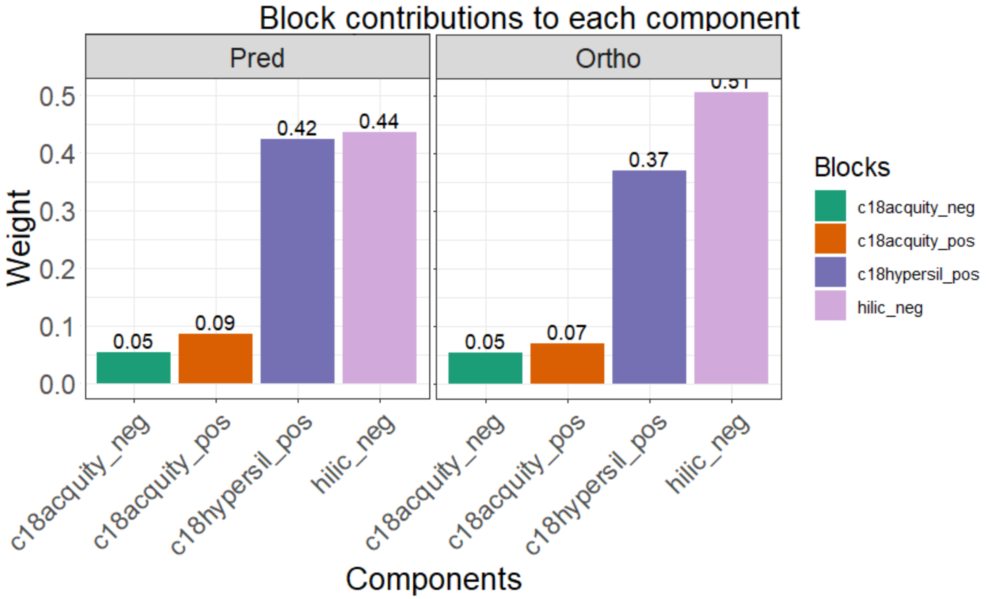
```
COPLS <- ConsensusOPLS(
   data         = my_data_with_omics_blocks_matrix_in_list,
   Y            = response_variable,
   maxPcomp     = 1,       # one predictive component
   maxOcomp     = 1,       # one orthogonal component
   modelType    = "da",    # discriminant/ classification model
   nperm        = 1000,    # number of permutations
   cvType       = "nfold", # type of cross-validation method
   nfold        = 42,      # number of subjects = leave-one-out
   kernelParams = list(type   = "p",
                       params = c(order = 1)),
                           # linear kernel
   mc.cores     = 1,       # cores available for parallelization
   verbose      = FALSE,   # display code execution steps
   nMC          = 100,     # not use because cvType = "nfold"
   cvFrac       = 4/5,     # not use because cvType = "nfold"
)
```

[1] Imbert, A., Rompais, M., Selloum, M. *et al.* ProMetIS, deep phenotyping of mouse models by combined proteomics and metabolomics analysis. *Sci Data* **8**, 311 (2021).

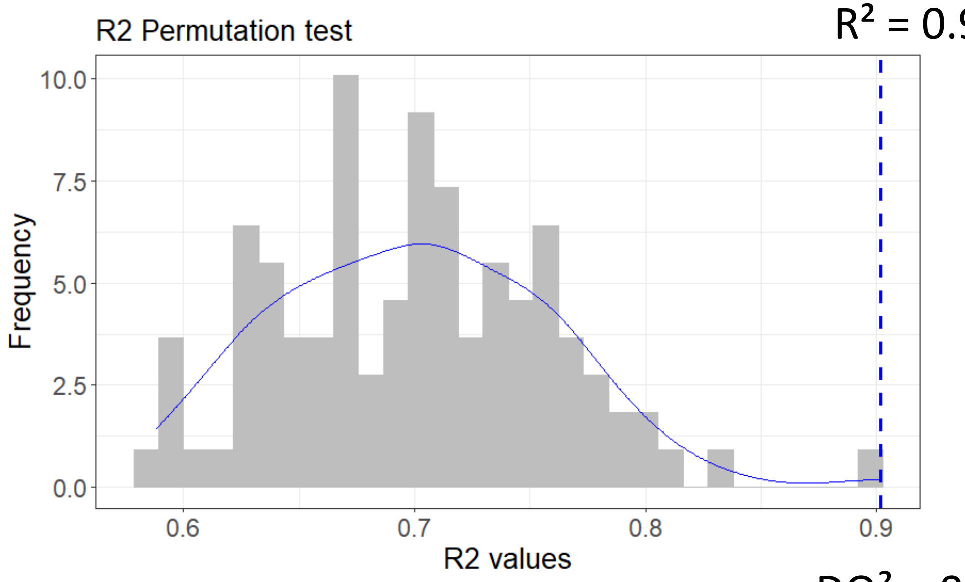
ConsensusOPLS: An R package for Multi-Block Data Fusion

Which results?

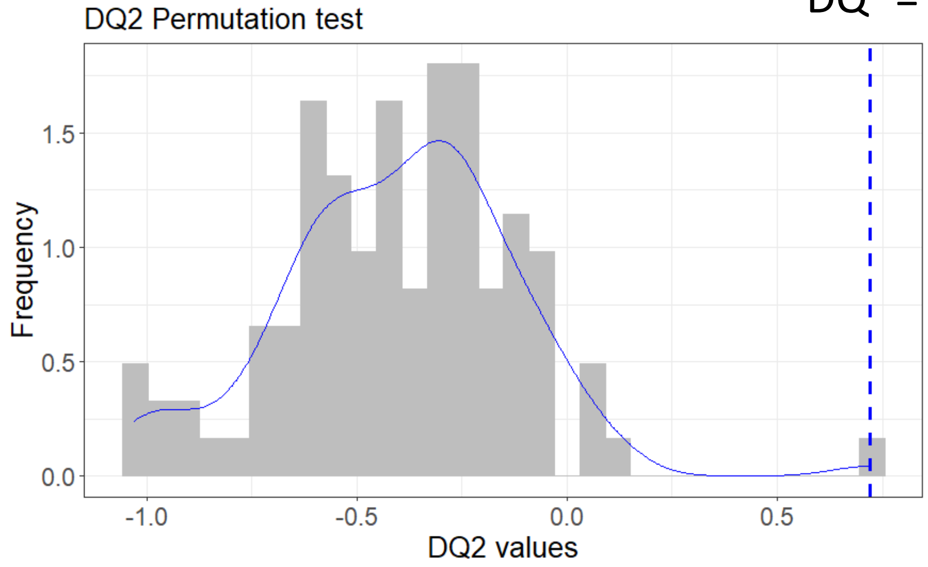




R² = 0.9018



DQ² = 0.7262



Why?

- Variables >> subjects
- Colinearity
- Heterogeneity
  - Numbers of var

What does variable selection mean in our project?

Use **kernel-based** variable selection methods
to eliminate continuous information **redundancy and noise**
in the data

Why?

- Variables >> subjects
- Colinearity
- Heterogeneity
  - Numbers of var

What does variable selection mean in our project?

Use **kernel-based** variable selection methods
to eliminate continuous information **redundancy and noise**
in the data

Why?

- ○ Variables >> subjects
- ○ Colinearity
- ○ Heterogeneity
    - Numbers of var

What are the decision criteria?

❑Improved performance of the ConsensusOPLS model on key indicators (R^2 and Q^2/ DQ^2)

❑Reduction in the number of highly correlated variables

Authors and ref.

Taguchi et al. 2022

Novel feature selection method via kernel tensor decomposition for improved multi-omics data analysis

DOI [10.1186/s12920-022-01181-4](10.1186/s12920-022-01181-4).

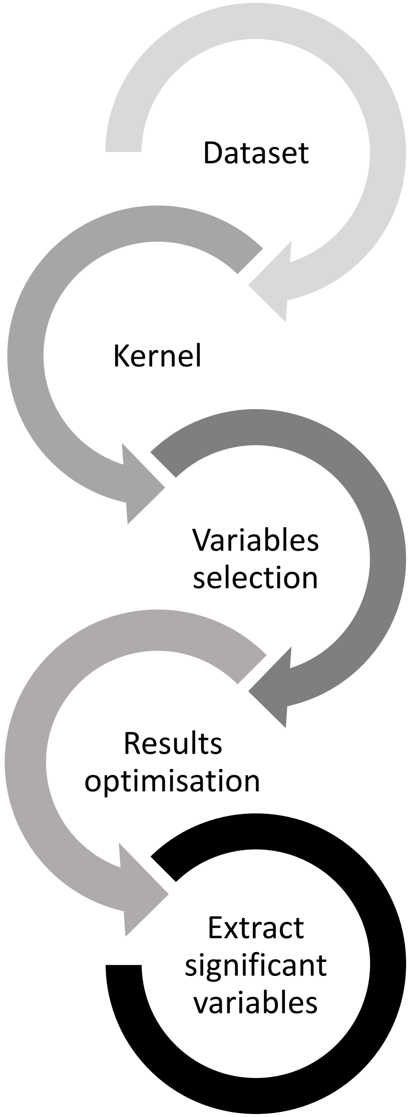
Unsupervised

Code already [available in R](), which I've reformatted into generalizable, automatic functions for better reproducibility.

Benefits:

- Captures complex, multi-dimensional and non-linear interactions
- Selects important features

Extended Kernel Tensor Decomposition (KTD)-based method



Raw data X in list form

Kernel transformation: the first features have a stronger dependency on the others (the most representative).

Tensor formation (n-dimensional cube) followed by Higher-order singular-value decomposition (HOSVD)

H0: the derivative of the singular values of the tensor decomposition (u) follows a normal distribution.
Thus p represents the probability that the importance of one component of the tensor (i.e. variable) is due to chance compared to the other variables (unsupervised).

Reduced dataset

Authors and ref.

Brouard et al. 2022

Feature selection for kernel methods in systems biology

DOI 10.1093/nargab/lqac014.

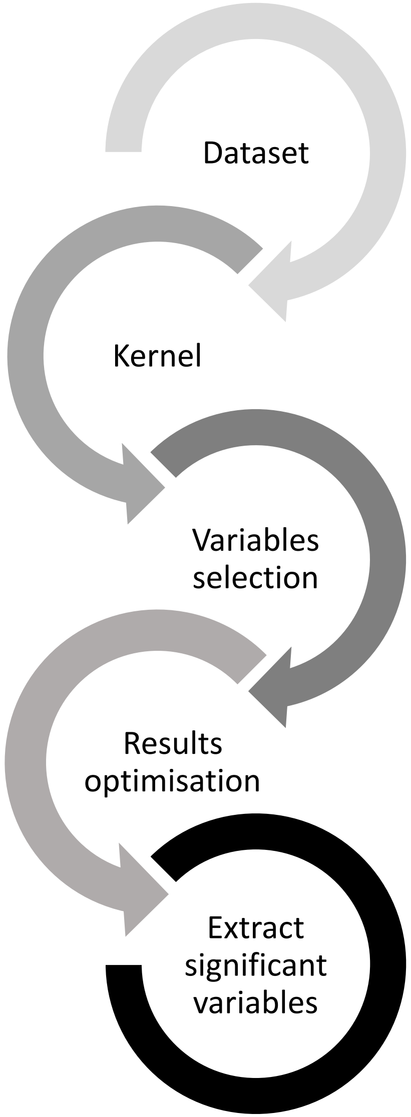
Unsupervised

Code already available in R, requires the use of the reticulate package. I use it in generalizable and automatic functions.

Benefits:

- Complex data interactions
- Reduces data redundancy
- Suitable for large-scale data
- Preserves kernel structure
- Can be used in a supervised environment with data a priori

Unsupervised Kernel Feature Selection (UKFS)



Dataset

Raw data X in list form

Kernel

Kernel transformation: the first features have a stronger dependency on the others (the most representative).

Variables selection

Method similar to Lasso (L1 penalty):
Definition of a penalized distortion criterion

$$\mathbf{w}^* := \underset{\mathbf{w}\in(\mathbb{R}+)^p}{\operatorname{argmin}} \|\mathbf{K}_x^{\mathbf{w}} - \mathbf{K}_x\|_F^2 + \lambda\|\mathbf{w}\|_1, \qquad (2)$$

Results optimisation

Optimize the selected feature subset by reformulating (2) as follows:

$$\underset{\mathbf{w}\in(\mathbb{R}+)^p}{\operatorname{argmin}} \quad f(\mathbf{w}) + \lambda g(\mathbf{w}) \qquad (3)$$

And using the proximal descending gradient, in particular with Forward-backward Splitting (FBS)

Extract significant variables

Reduced dataset

Authors and ref.

Yamada et al. 2019

High-Dimensional Feature Selection by Feature-Wise Kernelized Lasso

DOI 10.1162/NECO_a_00537.

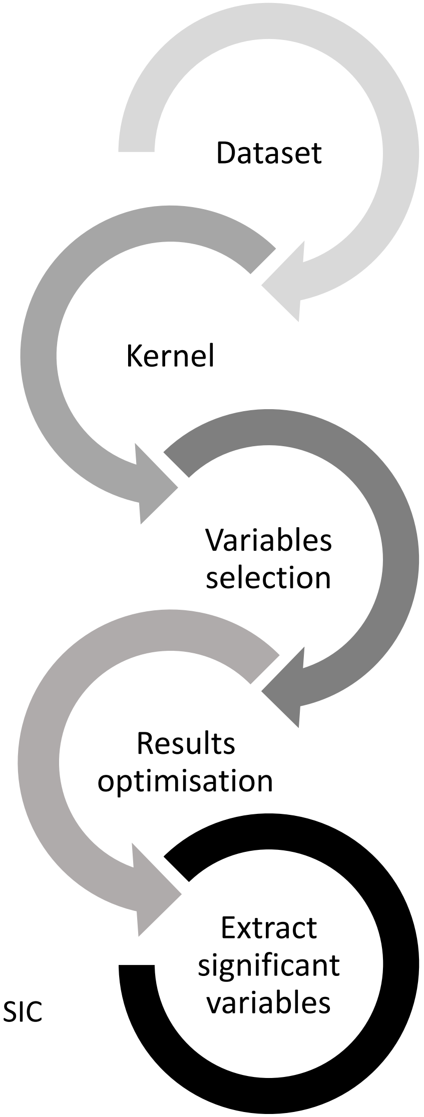
Supervised

Code already available in Python, requires the use of the reticulate package. I use it in generalizable and automatic functions..

Benefits:

- Applicable to large datasets (features >> samples)
- Captures non-linear relationships between inputs and outputs
- Takes into account information redundancy between features and outcomes (through HSIC calculation).
- Different kernel types for inputs and outputs (e.g.: regression, Gaussian + Gaussian; classification, Delta + Gaussian).

Hilbert-Schmidt independance criterion (HSIC) Lasso



Raw data X in list form and vector Y

Calculation of dependency measures between each variable and the response Y using the Hilbert-Schmidt independence criterion (HSIC) = kernel based independance measure

HSIC Lasso (with L1 penalisation):

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^d} \quad \frac{1}{2}\|\bar{L} - \sum_{k=1}^{d}\alpha_k \bar{K}^{(k)}\|_{\text{Frob}}^2 + \lambda\|\boldsymbol{\alpha}\|_1, \qquad \text{s.t.} \quad \alpha_1, \ldots, \alpha_d \geq 0, \tag{3}$$

Optimization of selected feature subset in HSIC-weighted feature space

Reduced dataset

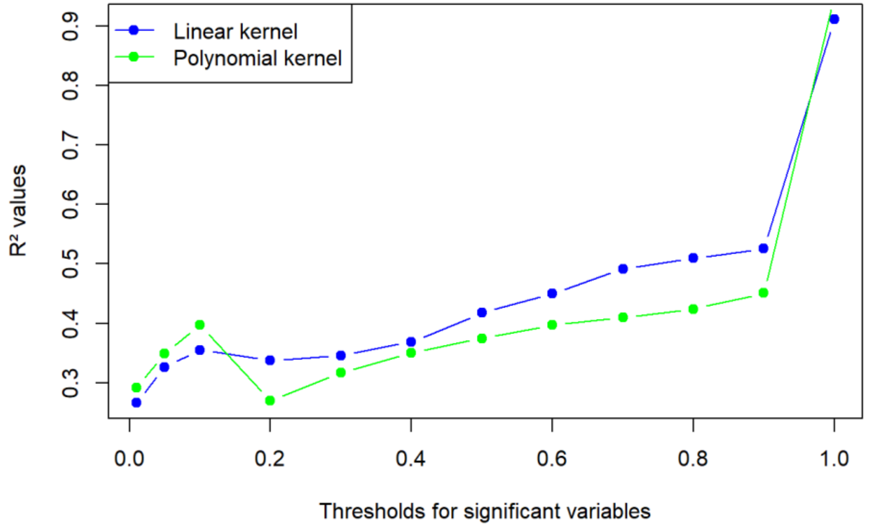
Extended Kernel Tensor Decomposition (KTD)-based method

Taguchi 2022 :

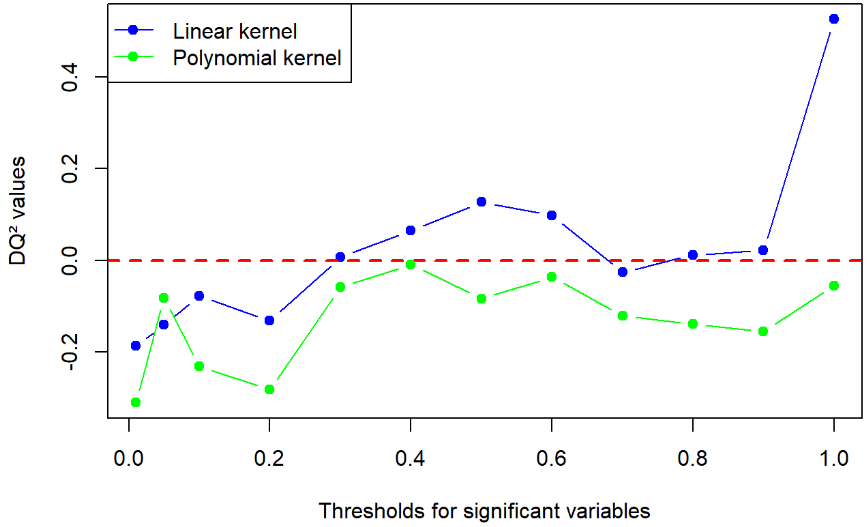
ProMetIs

Pareto scaled

ConsensusOPLS R² evolution according to p-values of significant results variations



ConsensusOPLS DQ² evolution according to p-values of significant results variations



Comments:

✓ Before selection :
$R^2 = 0.9497$ et $DQ^2 = -0.0546$.

✓ After selection :
Positive values only between 0.04 ($DQ^2 = 0.0645$) and 0.6 ($DQ^2 = 0.0985$), with a maximum at 0.5 ($DQ^2 = 0.1271$).

=> invalid model

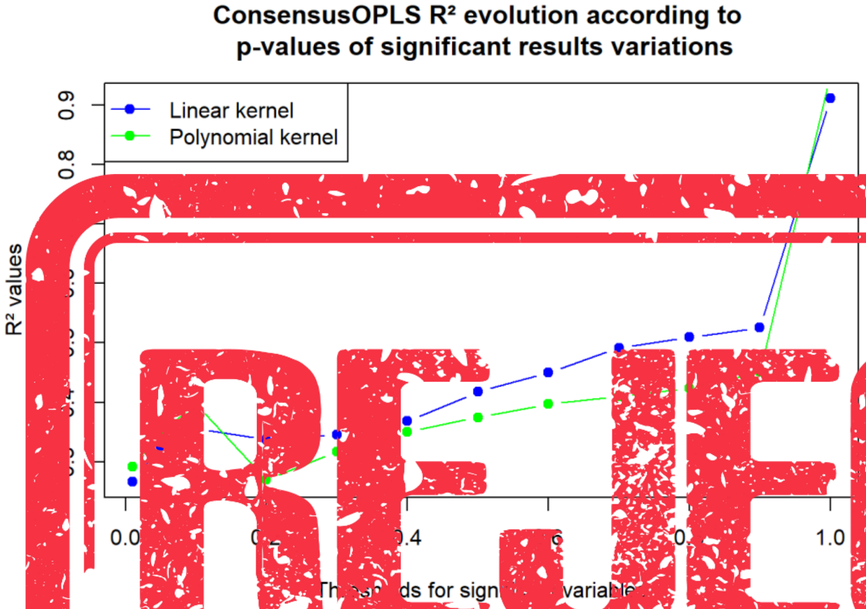
Extended Kernel Tensor Decomposition (KTD)-based method

Taguchi 2022 :

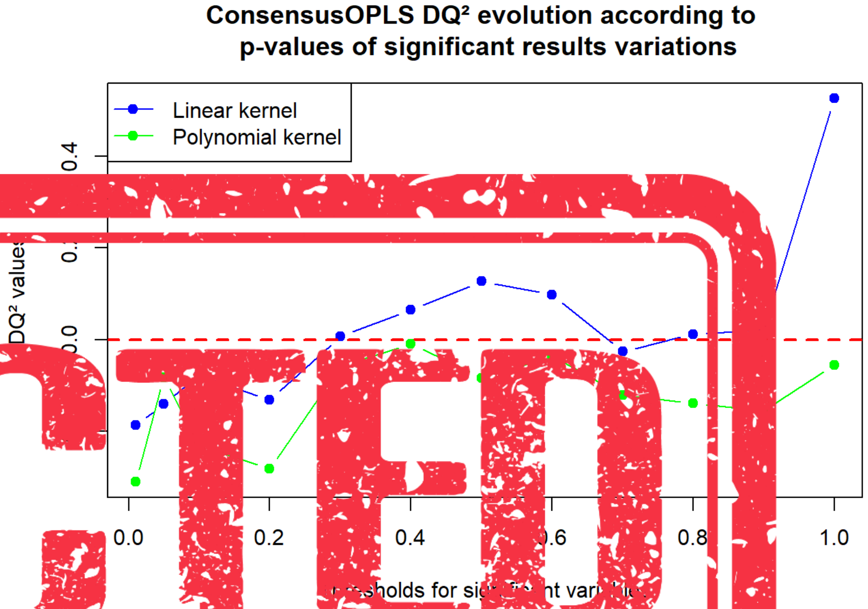
ProMetIs

Pareto scaled

ConsensusOPLS R² evolution according to p-values of significant results variations



ConsensusOPLS DQ² evolution according to p-values of significant results variations



REJECTED

Comments:

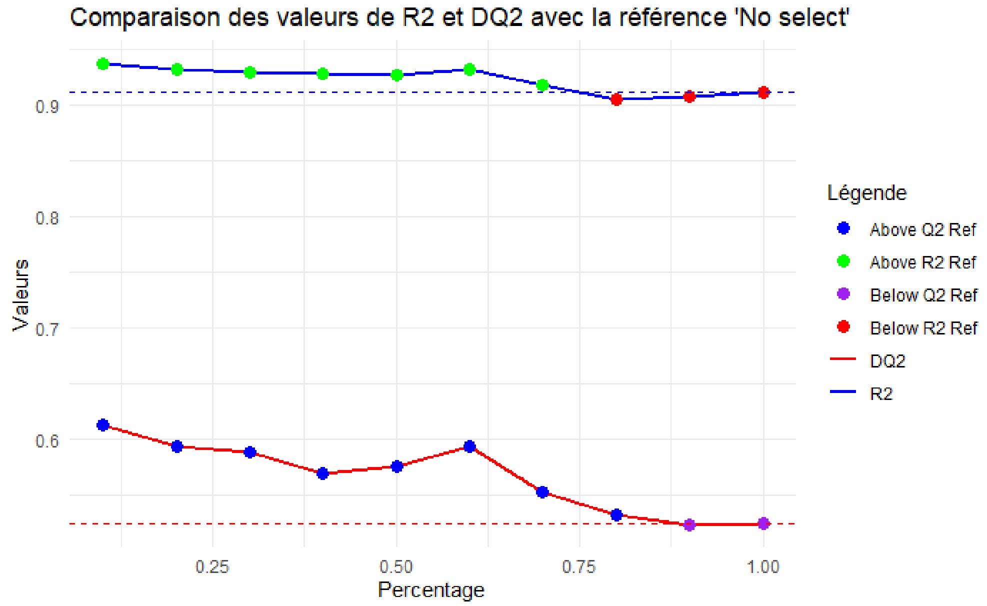
- ✓ Before selection :
  $R^2 = 0.9497$ et $DQ^2 = -0.0546$
- ✓ After selection :
  Positive values only between 0.04 ($DQ^2 = 0.0645$) and 0.6 ($DQ^2 = 0.0985$), with a maximum at 0.5 ($DQ^2 = 0.1271$).

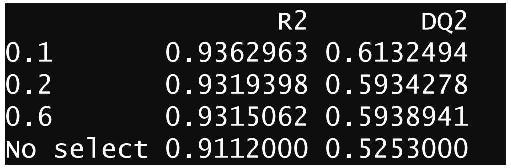
  => invalid model

Unsupervised Kernel Feature Selection (UKFS)

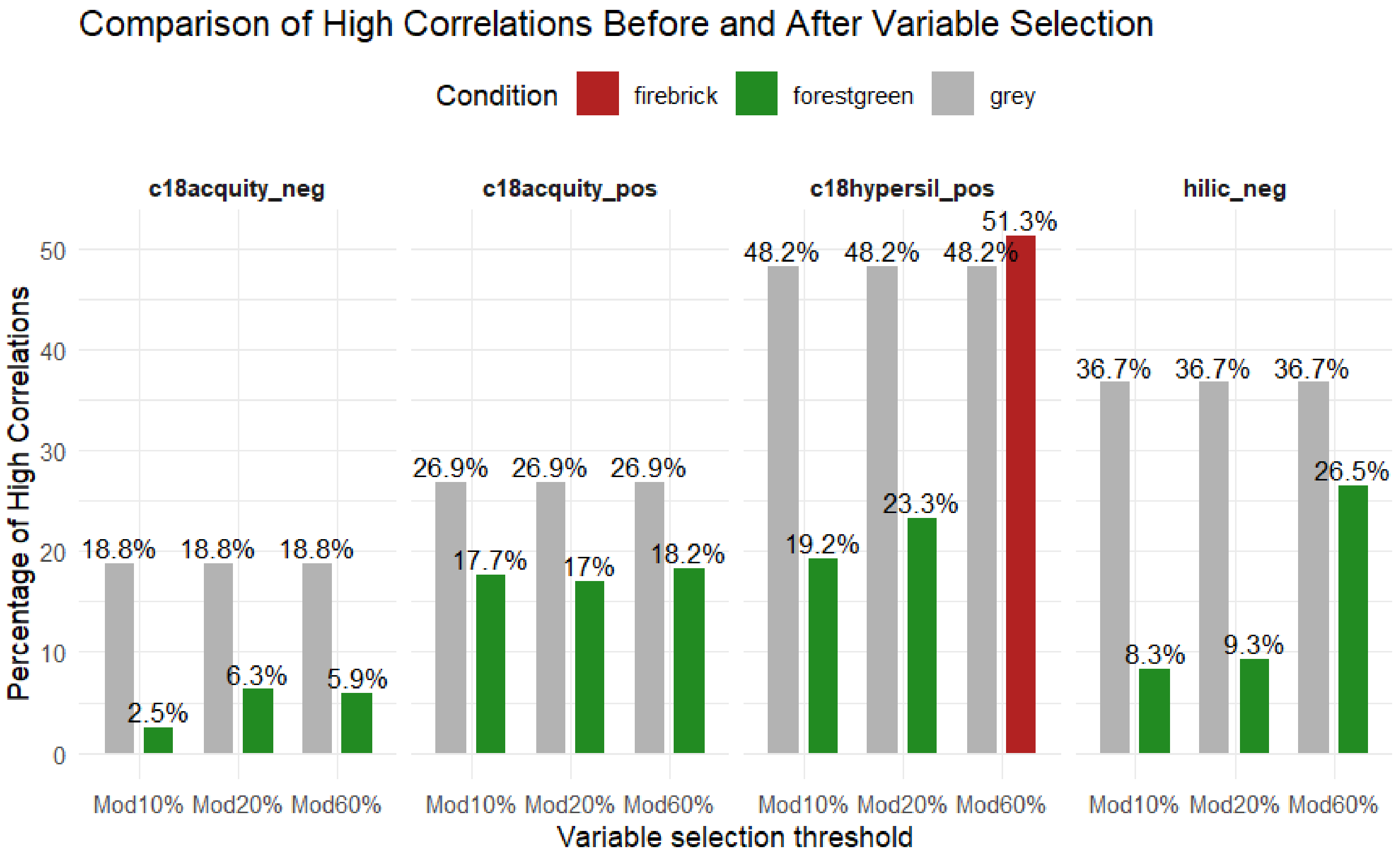
Brouard 2022:

ProMetIs - **Pareto scaled**





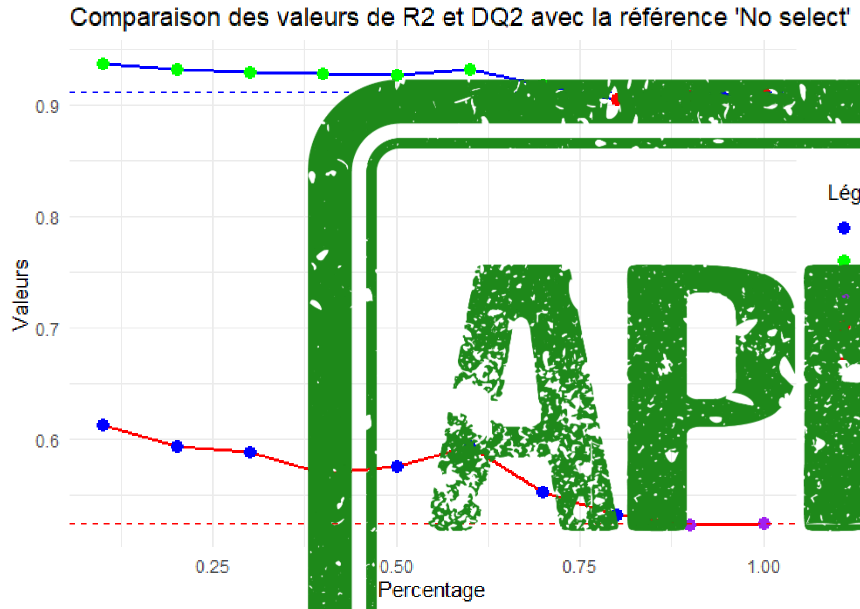
```
                R2       DQ2
0.1        0.9362963 0.6132494
0.2        0.9319398 0.5934278
0.6        0.9315062 0.5938941
No select 0.9112000 0.5253000
```

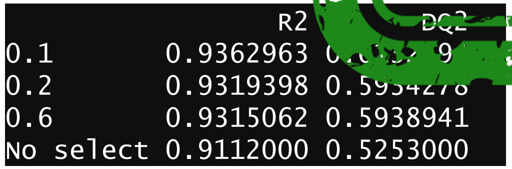


Unsupervised Kernel Feature Selection (UKFS)

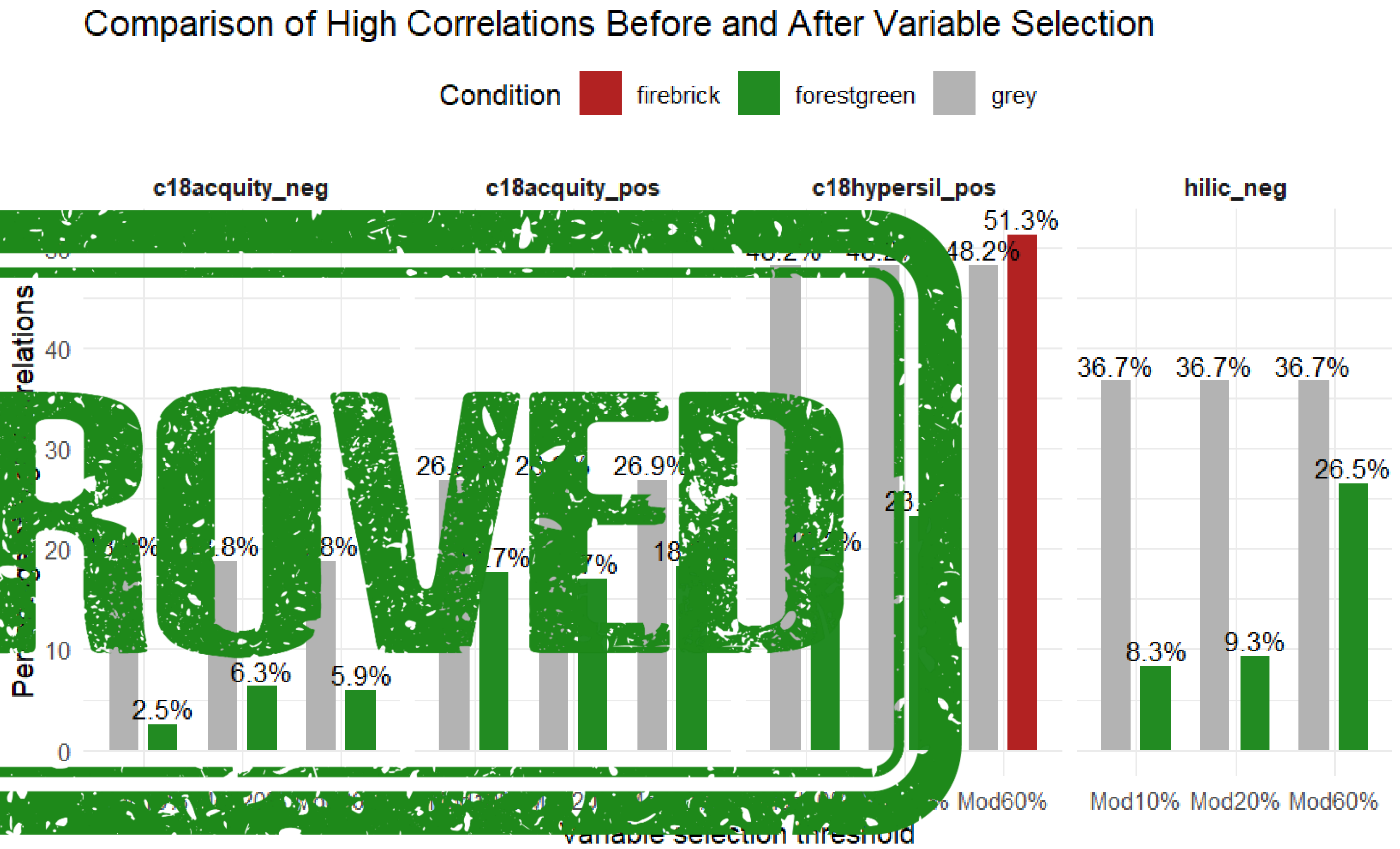
Brouard 2022:

ProMetIs - **Pareto scaled**





```
                  R2       DQ2
0.1        0.9362963 [illegible]
0.2        0.9319398 0.5934278
0.6        0.9315062 0.5938941
No select  0.9112000 0.5253000
```



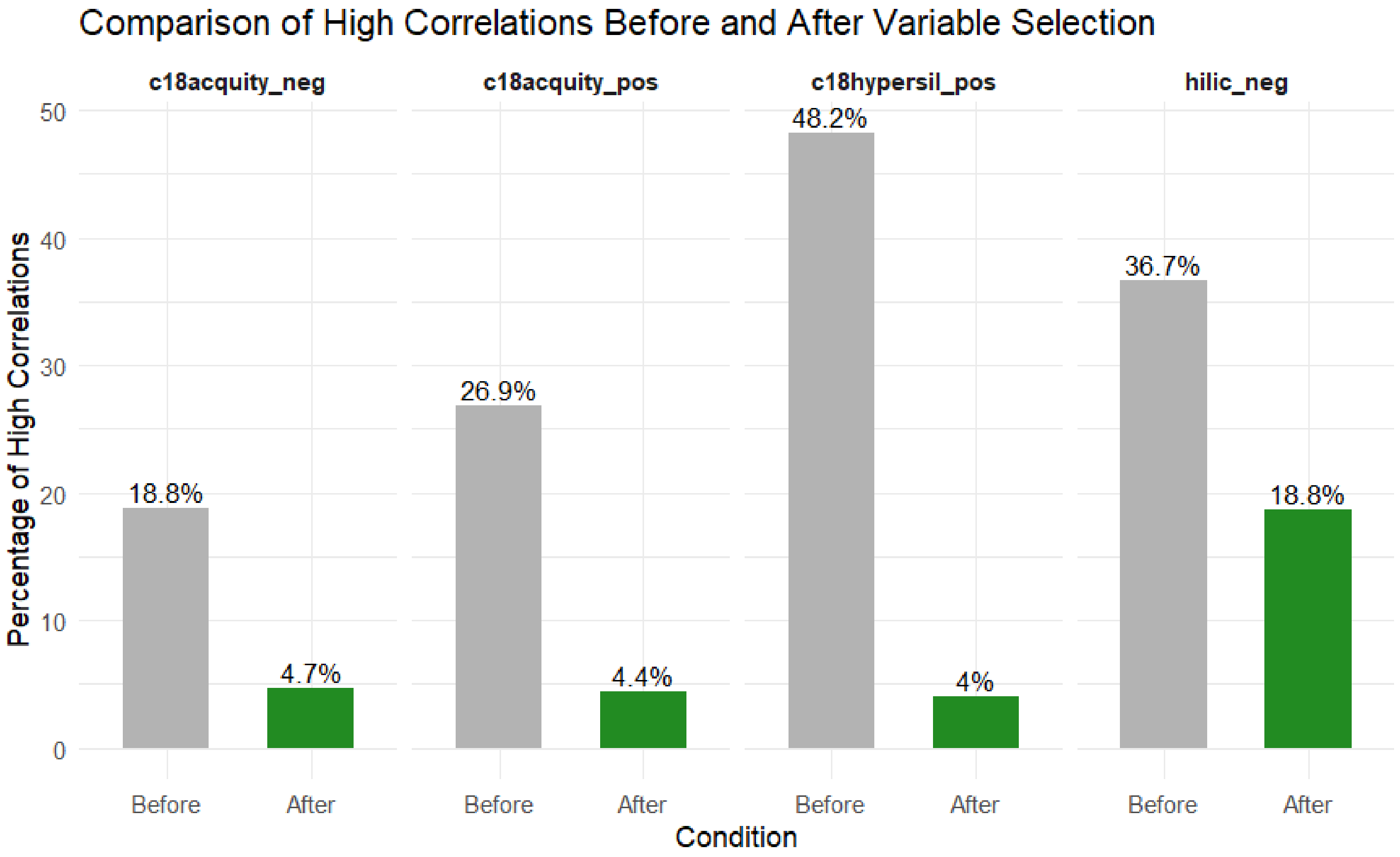
Hilbert-Schmidt independance criterion (HSIC) Lasso

Yamada 2019 :

ProMetIS

```
##                  R2       DQ2
## 0.1        0.969762 0.9088585
## 0.2        0.969762 0.9088585
## 0.3        0.969762 0.9088585
## 0.4        0.969762 0.9088585
## 0.5        0.969762 0.9088585
## 0.6        0.969762 0.9088585
## 0.7        0.969762 0.9088585
## 0.8        0.969762 0.9088585
## 0.9        0.969762 0.9088585
## No select 0.911200 0.5253000
```

```
                0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
c18acquity_neg   43  43  43  43  43  43  43  43  43
c18acquity_pos   45  45  45  45  45  45  45  45  45
c18hypersil_pos  50  50  50  50  50  50  50  50  50
hilic_neg        32  32  32  32  32  32  32  32  32
```



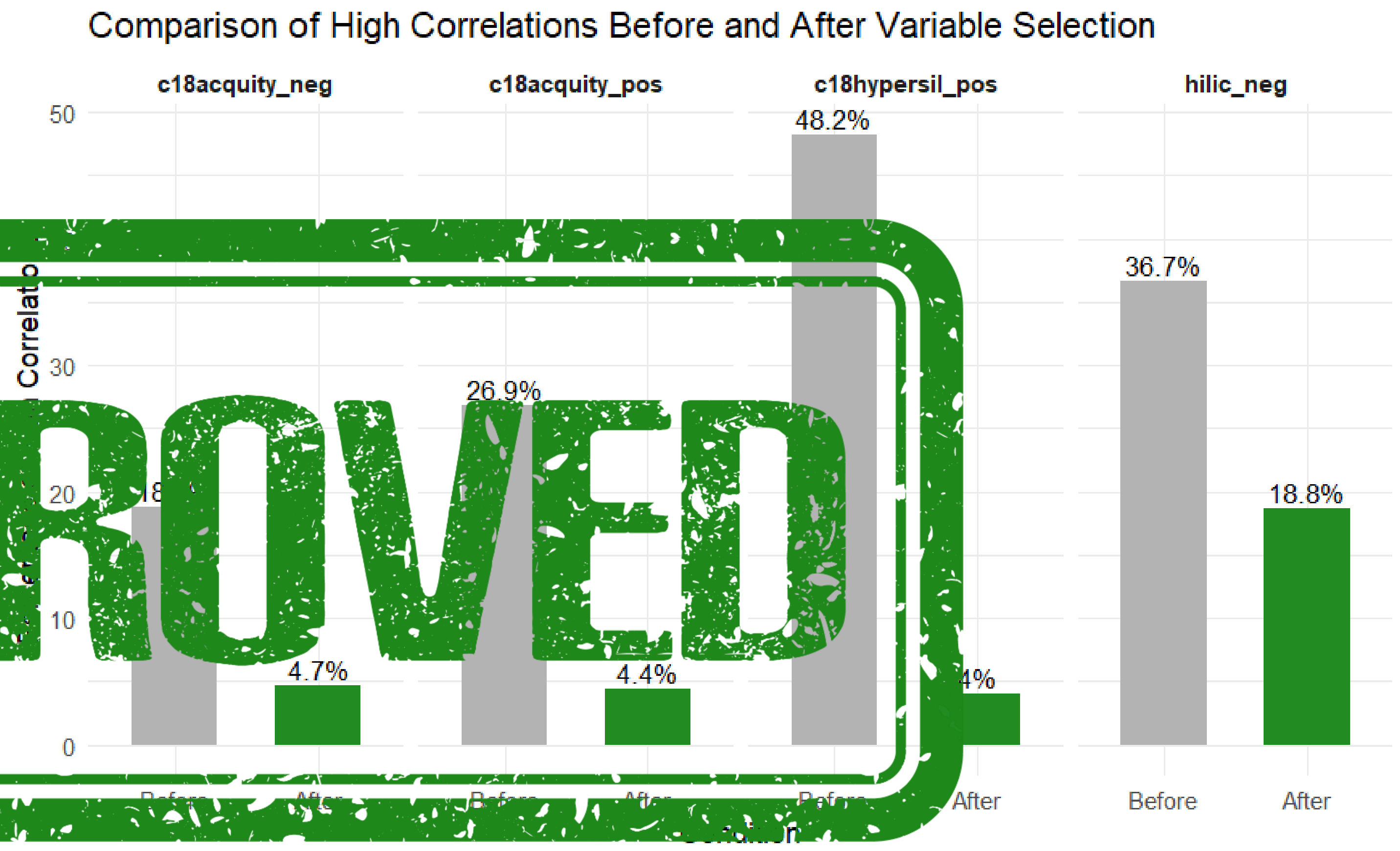
Hilbert-Schmidt independance criterion (HSIC) Lasso

Yamada 2019 :

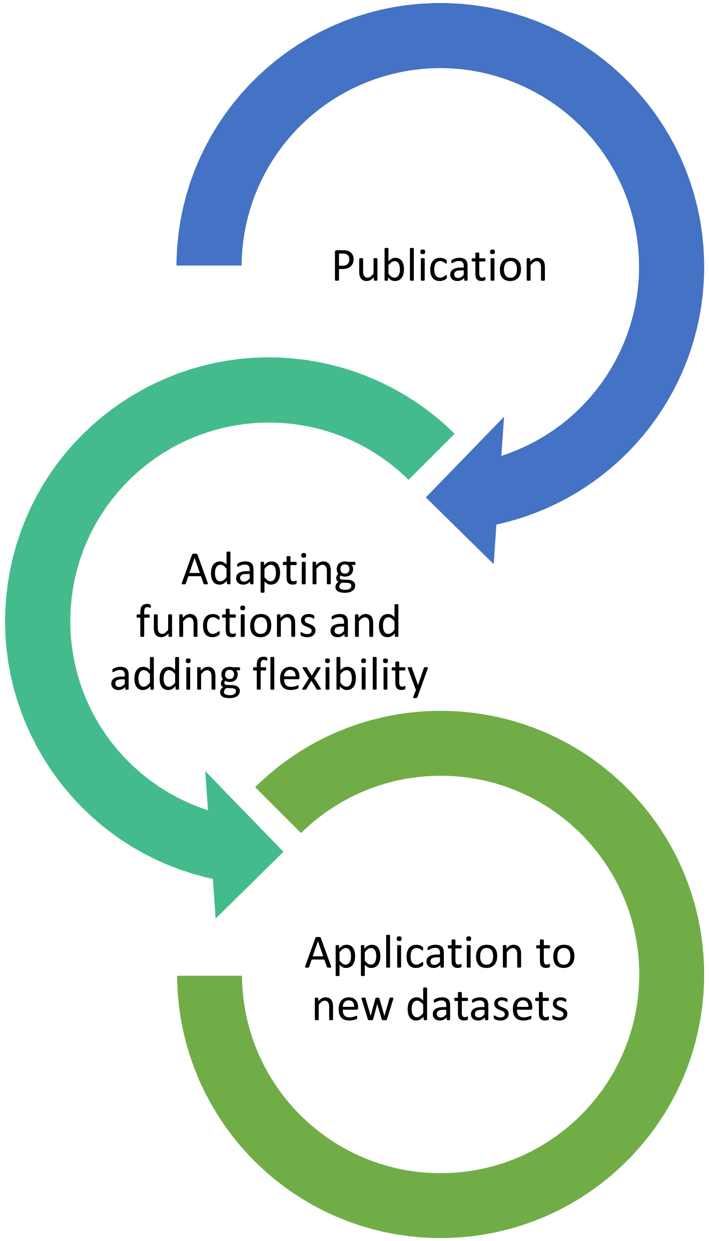
ProMetIS

```
##                   R2       DQ2
## 0.1         0.969762 0.9088585
## 0.2         0.969762 0.9088585
## 0.3         0.969762 0.9088585
## 0.4         0.969762 0.9088585
## 0.5         0.969762 0.9088585
## 0.6         0.969762 0.9088585
## 0.7         0.969762 0.9088585
## 0.8         0.969762 0.9088585
## 0.9         0.969762 0.9088585
## No select 0.911200 0.[illegible]53000
```

```
                  0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
c18acquity_neg     43  43  [illegible]  43  43  43  43  [illegible]  43
c18acquity_pos     45  45  [illegible]  45  45  45  45  [illegible]  45
c18hypersil_pos    50  50  [illegible]  50  50  50  50  50  50
hilic_neg          32  32  [illegible]  32  32  32  32  32  32
```



APPROVED


Publication
Adapting functions and adding flexibility
Application to new datasets

Merci pour votre attention

```
 /\__/\
(=^•^=)
  (")(")_/
```