



UNIVERSITÉ
DE GENÈVE



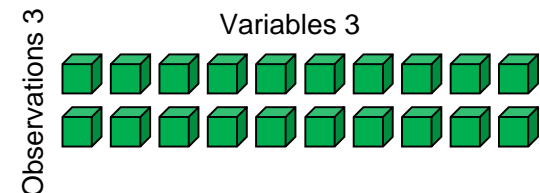
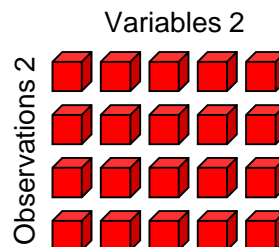
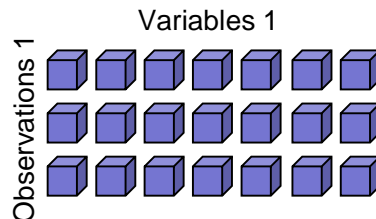
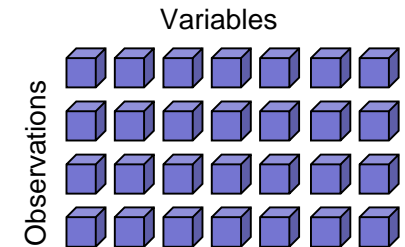
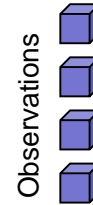
Swiss Institute of
Bioinformatics

PART II

MULTIOMICS DATA INTEGRATION

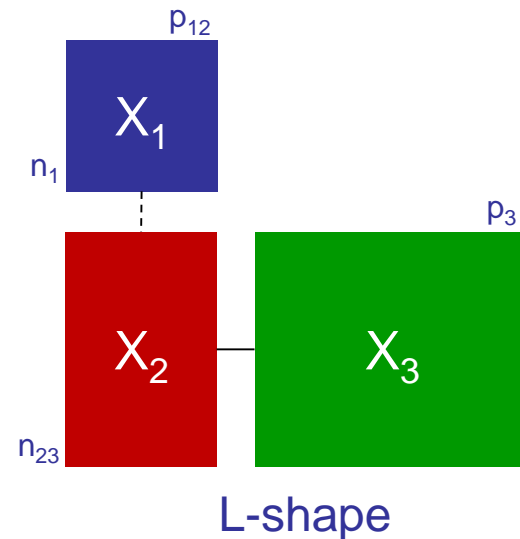
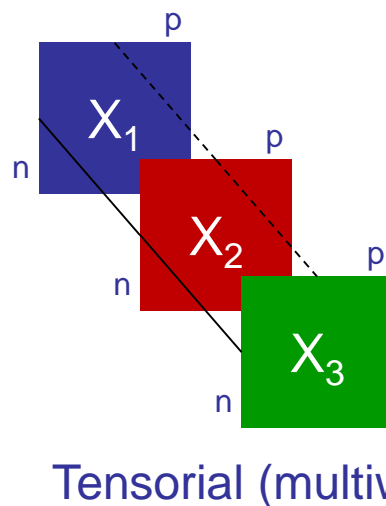
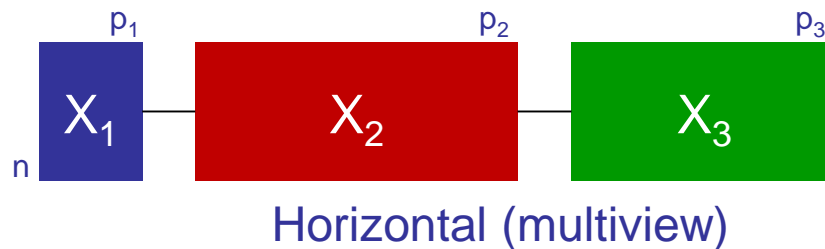
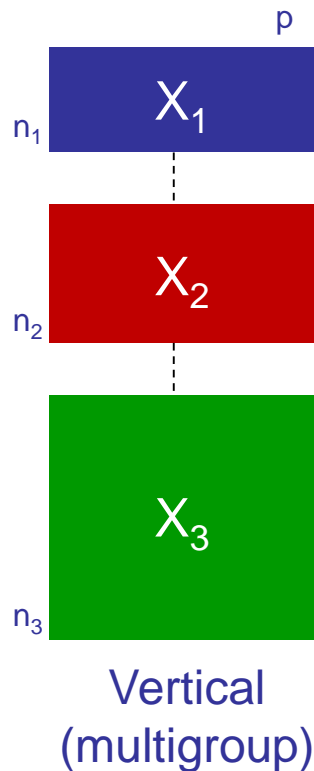
Data Structures

- 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)
- Multiblock data can be seen as a **collection of matrices**



Multiblock Data Structure

- Shared Observation mode ?
- Shared Variable mode ?
- Shared Observation or Variable mode ?
- Shared Observation and Variable mode ?



A Horizontal Multiblock Data Structure



The **Observation mode** (rows) is shared (n observations)
The **Variables mode** (columns) is specific (p_1, p_2, p_3 variables)
Many more variables than observations

Goals are the same as single-block data analysis

Find components (linear combinations of the initial variables) to
Describe – Discriminate – Classify – Predict

Objectives

- ✓ Specific applications: Collaboration or Competition



Combine data sources

- Gain an extended understanding of complex systems
 - coverage
 - mechanisms
- Improved prediction of new observations
 - more sensitive
 - more specific

Compare data sources

- Rank data sources (e.g. analytical methods)
- Block subsets selection

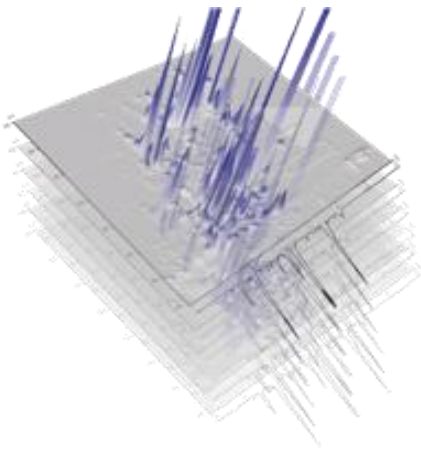


FAIRNESS BETWEEN BLOCKS ?

Data Complementarity

Relations between the input data sources:

- Complementary
 - The information represents different parts of the system
 - Obtain **more complete or new information**
(e.g. improved omics coverage or systems biology)
- Redundant
 - Two or more sources provide the same information
 - Increase the **confidence**
(e.g. multiple analytical platforms, biological layers)



COMMON AND/OR DISTINCT VARIATIONS ?

Data Pre-processing

- Centering
- Within-block scaling
 - usual scaling as performed for a single data block
e.g. unit variance or Pareto scaling
- Between-block scaling
 - each block of data is simultaneously scaled
 - different block weights = special properties

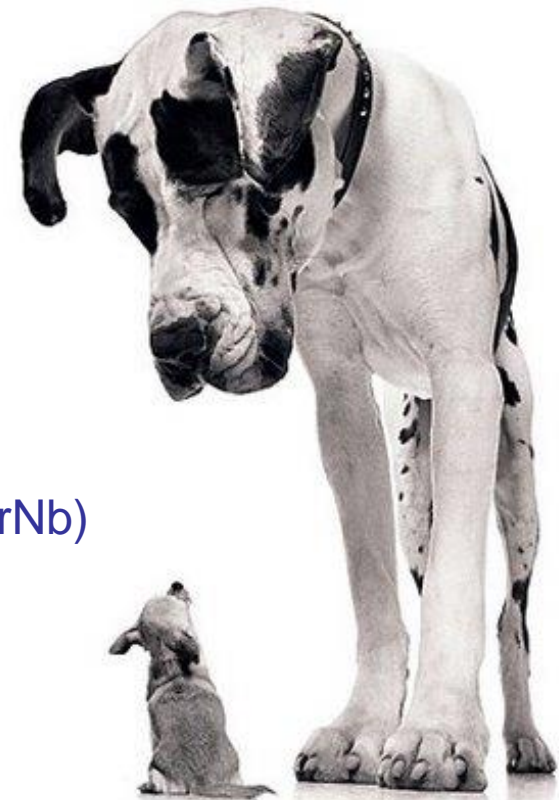
High-dimensional blocks will have more influence

→ Scaling according to the number of variables ($1/\text{VarNb}$)

Block with large range will have more influence

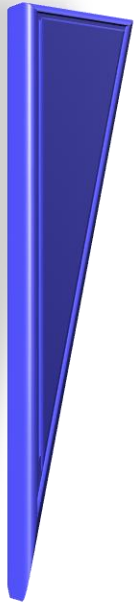
→ Scaling according to block inertia/norm

NB: Some methods are invariant to certain types of scaling



Data Fusion Terminology

Strategies for data integration based on abstraction levels



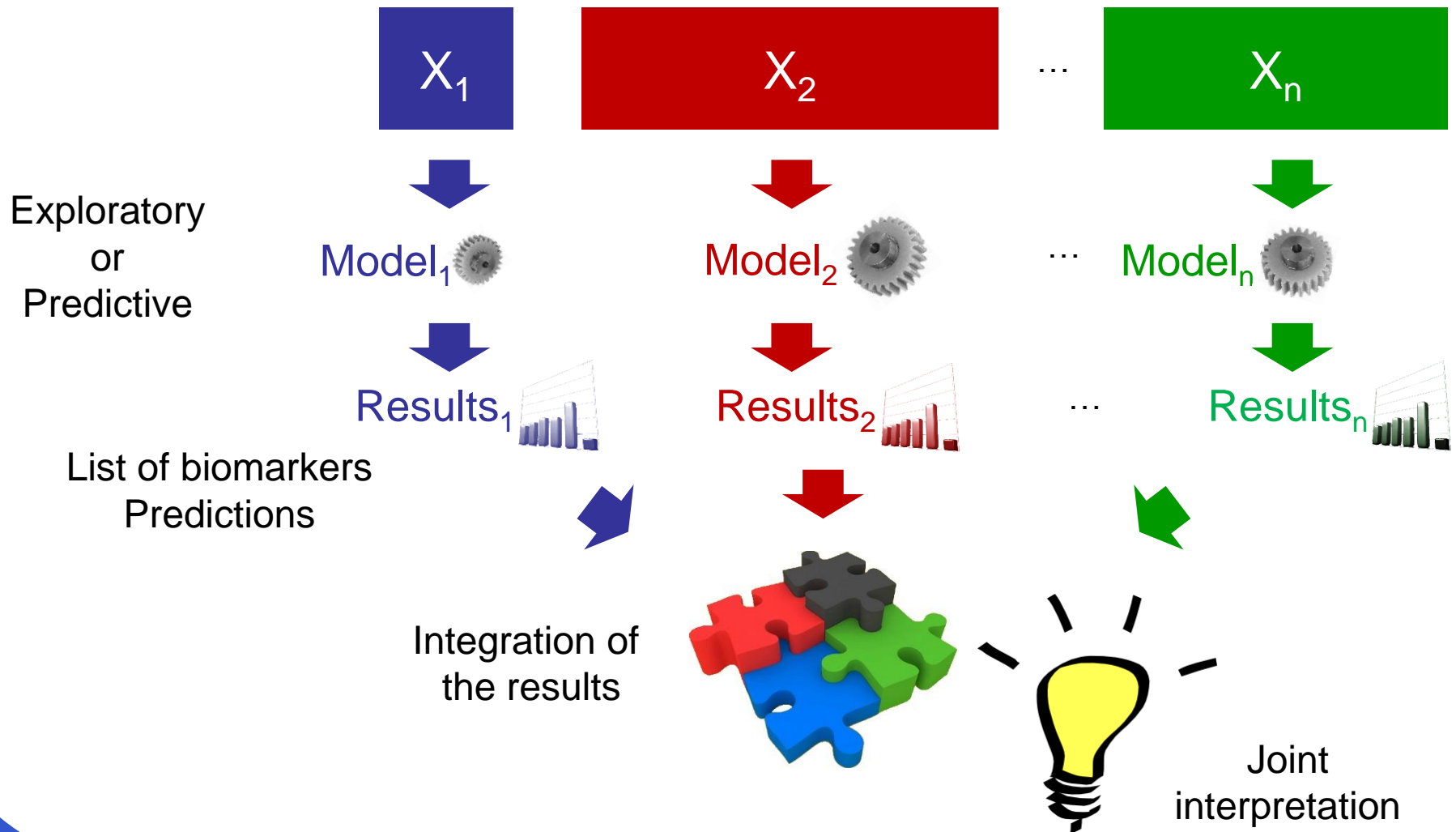
- High-level (symbolic representations or decisions)
→ information/decision fusion
- Mid-level (patterns or subsets extracted from the sources)
→ characteristics employed for other tasks
- Low-level (signals)
→ data fusion/aggregation/association

The terminology depends on the application domain



High-level Data Fusion

Integration of results from single blocks models



High-level Data Fusion

Separate evaluation of each data sources

The samples can be different between blocks → more flexibility



Fusion of outputs (unsupervised)

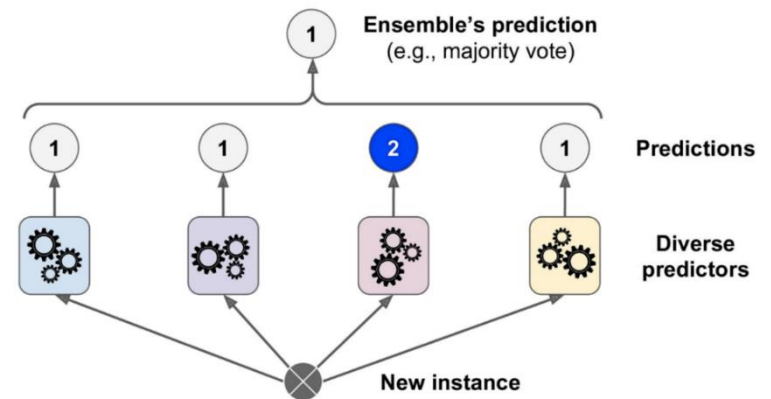
→ Joint interpretation of biomarker patterns or scores
e.g. **ontologies**, over-representation analysis

Decision fusion (supervised)

→ Joint prediction (ensemble learning)

Voting schemes

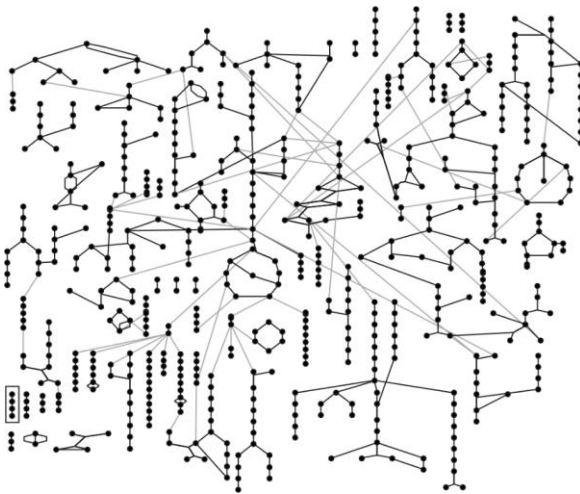
e.g. majority voting the class with the highest number of votes wins



→ Combining classification or prediction results for **improved accuracy**

Biological Networks

- Networks often represented as graphs
- Nodes represent **metabolites, proteins or genes**
- Edges represent the **functional links** between nodes (e.g. biochemical reaction, regulation or binding)
- **Changes in topology** can result in **novel properties** (comparisons of specific situations)



A metabolic network for Escherichia coli



Visualization of the results

Contribute to **mechanistic interpretation**

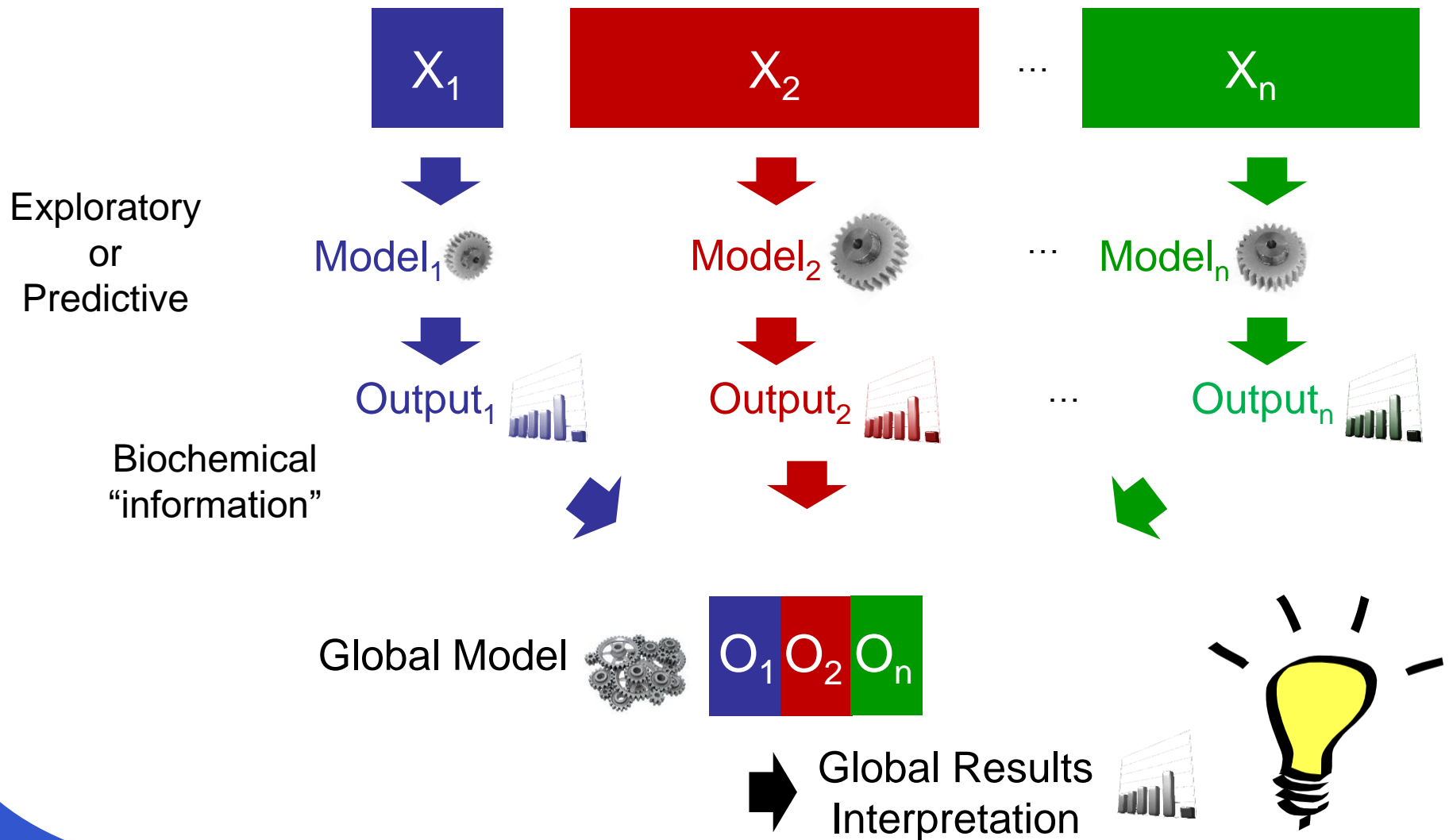
- most relevant biological processes
- regulatory relationships

Types of interactions:

- protein– metabolite (metabolic pathways)
- protein – protein (cell signaling, protein interactions)
- protein – gene (genetic networks)

Mid-level Data Fusion

Integration of results from single blocks models

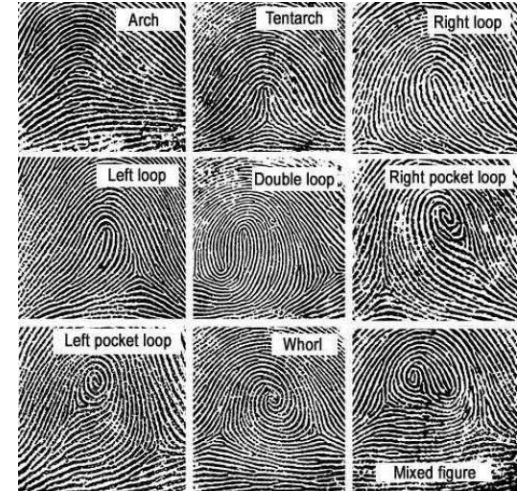


Feature Extraction vs. Selection

Several features can be combined together without loss or even with gain of information

→ **Feature extraction**

Combine the original variables into a smaller set of synthetic variables with **Projection Methods**



Some variables bear little or no useful information

→ **Feature selection**

Choose a subset of important features, ignoring the remaining unimportant variables with **Subset Selection Algorithms**

Mid-level Data Fusion

2-step procedure



- Middle-Up

- i. Multivariate model (PCA, CA, PLS) → dimensionality reduction
- ii. Second analysis based on the concatenated scores (PCA, PLS, clustering)



Handle heterogeneous datasets



What about model interpretation?



How many components to keep?

- Middle-Down

- i. Model or test (PLS, Fold change) → local variable selection
- ii. Second analysis based on the concatenated variables subsets



Easy interpretation of the final model



How to handle heterogeneous datasets?



How many variables to keep?

Limitations of High/Mid-level Data Fusion

No insight into the **links between** initial data blocks
Individual models may lead to a **substantial loss of biological information**

Results may be contradictory/heterogeneous

- Unsupervised: biological interpretation can be **tedious**
- Supervised: the result can be **inconclusive** in the case of ties

The combination of information may **not be relevant**

Different subsets of observations may lead to **conflicting conclusions** within multiple data sources

Interpretable patterns may be **impaired by combining** with other data sources



Low-level Multiblock Analysis

Consider the fact that **the same observations** are in the different blocks
→ assess the relationship between the variables and the data tables



Unsupervised analysis

- explorative analysis looking for structures and patterns
- links between variables in a single data block
 - links across data blocks

Supervised analysis

- predictive data analysis, emphasis is on a response block of data Y
- connections to one or more blocks of data
 - some blocks are **dependent** and others are **independent**

Multiblock Data Modeling

samples

X_1

X_2

X_3

X_n



- ✓ Think global by building a compromise accounting for all data with adequate weights
- ✓ Act local by maximizing the link between data blocks under a specific criterion, e.g. canonical correlation, co-inertia, partial least squares

Find the relevant information

- ✓ Role/importance of each data table
- ✓ Common/specific trends
- ✓ Links between variables of different nature



Low-level Data Fusion

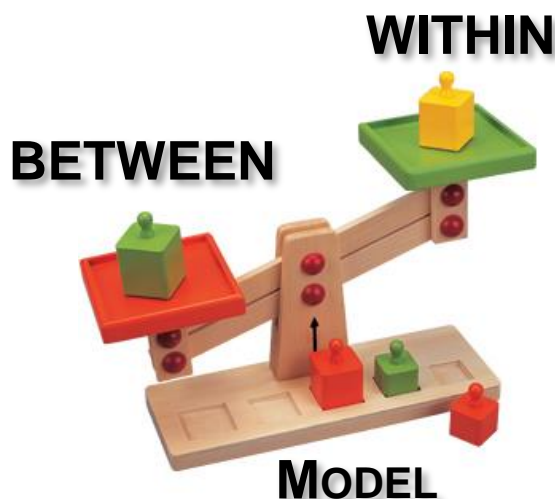
Factor analysis can be applied to blocks instead of initial variables



Each component related to one block is connected to all the components related to the other blocks and/or to a global component

Block components should verify two properties simultaneously

- Block components explain well their own block
- Block components are as correlated as possible with related blocks



The multiblock model build components as a compromise for explaining between-block and within-block variation

Different methods favor explaining more within- or between-block variation

Matrix Factorization

$$F_3 = X_3^1 w_3^1 + X_3^2 w_3^2 + \dots + X_3^n w_3^n$$

$$F_1 = X_1^1 w_1^1 + X_1^2 w_1^2 + \dots + X_1^n w_1^n$$

$$X_3 = \begin{bmatrix} | & \text{---} \end{bmatrix}$$

$$X_1 = \begin{bmatrix} | & \text{---} \end{bmatrix}$$

$$X_2 = \begin{bmatrix} | & \text{---} \end{bmatrix}$$

$$F_2 = X_2^1 w_2^1 + X_2^2 w_2^2 + \dots + X_2^n w_2^n$$

HOW TO BUILD COMPONENTS
HOW TO LINK COMPONENTS



CO-VARIANCE CRITERION



$$\max \sum_{j,k=1}^n \text{cov}(X_j w_j, X_k w_k)$$

Correlation, Variance and Covariance

Correlation is widely used to describe the relationship between variables

- linear relation between two variables (Pearson)
- non-metric relation based on the ranks of the variables (Spearman)

$$\text{corr}(x, y) = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)} \sqrt{\text{var}(y)}}$$

It can be extended to describe the relationship between data blocks



Extracting Structures From Data

Similar relations can be estimated **between matrices and components**

MATRICES

Similarity can be defined in very many different ways

The RV coefficient compares the **configurations of the samples**

- blocks with different numbers of variables
- modified version for high-dimensional data

COMPONENTS

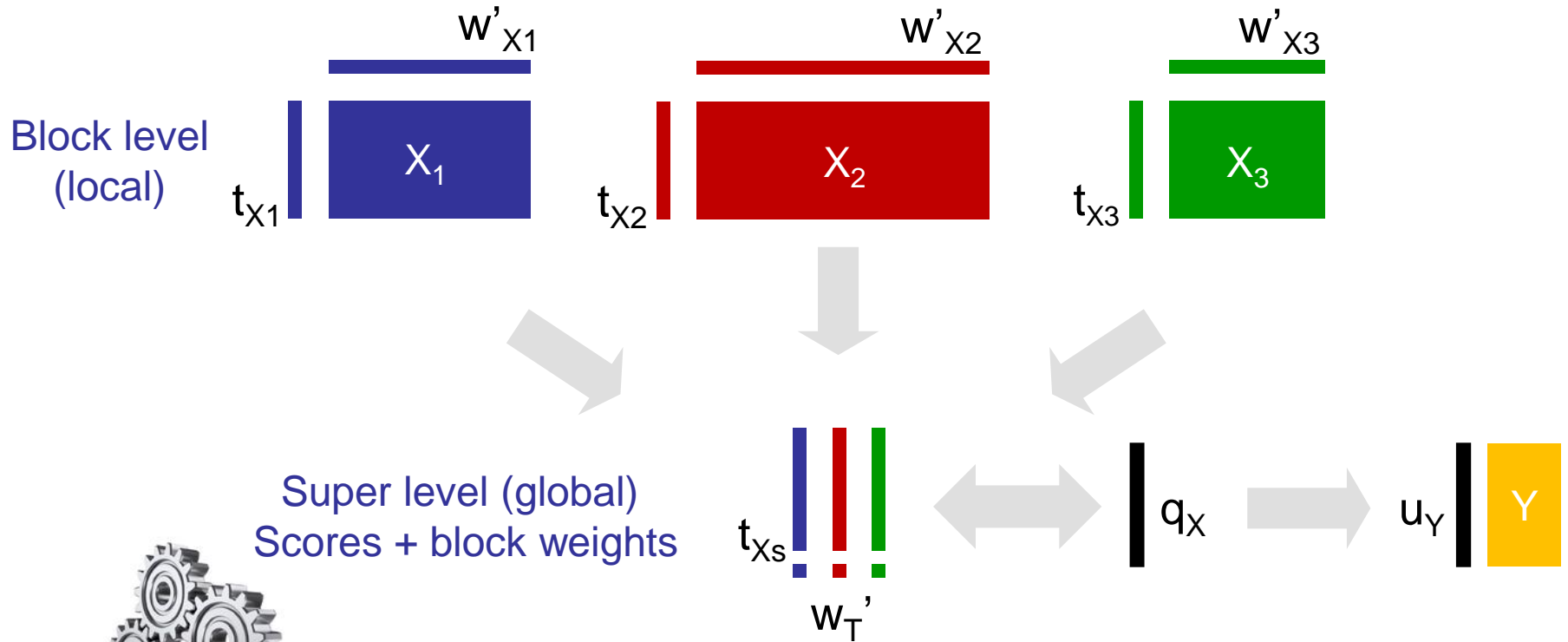
Components are linear combinations of the initial variables $X_j \mathbf{w}_j$

Links between components extracted from two blocks:

$$\text{cov}^2(X_j \mathbf{w}_j, X_k \mathbf{w}_k) = \text{var}(X_j \mathbf{w}_j) \text{cor}^2(X_j \mathbf{w}_j, X_k \mathbf{w}_k) \text{var}(X_k \mathbf{w}_k)$$

Low-level Horizontal Multiblock Analysis

A collection of data blocks with shared observations



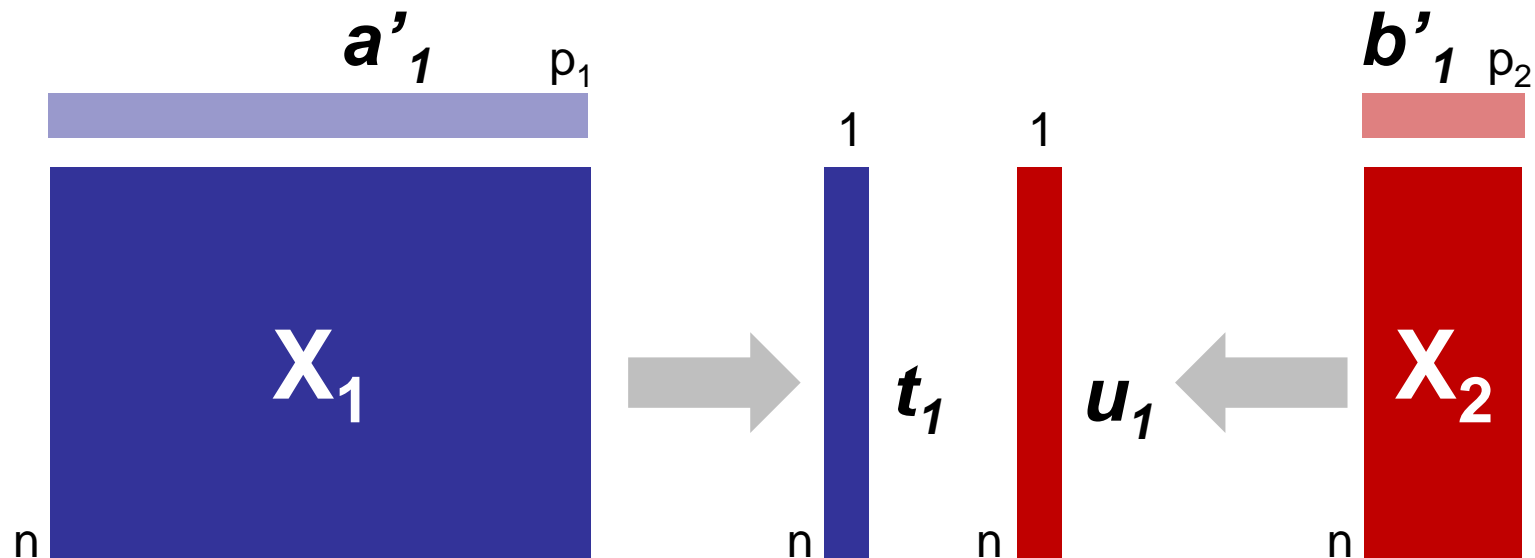
How to build the super level

It depends on the model



Relating Two Data Blocks – Multiblock

Linking function between latent variables
e.g. identity link, flexible link, partial identity link



$t_1 = Xa_1$ and $u_1 = Yb_1$
maximising the **link** between t_1 and u_1

→ Find **common trends** in data blocks

Methods and Criteria To Be Optimized

Many multiblock methods make implicitly or explicitly a **choice on which type of variation** is modelled

Covariance represents the amount of variation between the blocks but also **describes to some extent** the variation within the blocks



Method	Criterion	PLS path model	Mode	Scheme
(1) SUMCOR (Horst 1961)	$Max \sum_{j,k} Cor(F_j, F_k)$ or $Max \sum_j Cor(F_j, \sum_k F_k)$	Hierarchical	B	Centroid
(2) MAXVAR (Horst 1961) or GCCA (Carroll 1968)	$Max \{\lambda_{first}[Cor(F_j, F_k)]\} \text{ (a)}$ or $Max \sum_j Cor^2(F_j, F_{j+1})$	Hierarchical	B	Factorial
(3) SsqCor (Kettenring 1971)	$Max \sum_{j,k} Cor^2(F_j, F_k)$	Confirmatory	B	Factorial
(4) GenVar (Kettenring 1971)	$Min \{\det[Cor(F_j, F_k)]\}$			
(5) MINVAR (Kettenring 1971)	$Min \{\lambda_{last}[Cor(F_j, F_k)]\} \text{ (b)}$			
(6) Lafosse (1989)	$Max \sum_j Cor^2(F_j, \sum_k F_k)$			
(7) Mathes (1993) or Hanafi (2005)	$Max \sum_{j,k} Cor(F_j, F_k) $	Confirmatory	B	Centroid
(8) MAXDIFF (Van de Geer, 1984 & Ten Berge, 1988)	$Max_{all} \ w_j\ =1 \sum_{j \neq k} Cov(X_j w_j, X_k w_k)$			
(9) MAXBET (Van de Geer, 1984 & Ten Berge, 1988)	$Max_{all} \ w_j\ =1 \sum_{j,k} Cov(X_j w_j, X_k w_k)$			
(10) MAXDIFF B (Hanafi and Kiers 2006)	$Max_{all} \ w_j\ =1 \sum_{j \neq k} Cov^2(X_j w_j, X_k w_k)$			
(11) (Hanafi and Kiers 2006)	$Max_{all} \ w_j\ =1 \sum_{j \neq k} Cov(X_j w_j, X_k w_k) $			
(12) ACOM (Chessel and Hanafi 1996) or Split PCA (Lohmöller 1989)	$Max_{all} \ w_j\ =1 \sum_j Cov^2(X_j w_j, X_{j+1} w_{j+1})$ or $Min_{F,p_j} \sum_j \ X_j - F p_j^T\ ^2$	Hierarchical	A	Path-weighting
(13) CCSWA (Hanafi et al., 2006) or HPCA (Wold et al., 1996)	$Max_{all} \ w_j\ =1, Var(F)=1 \sum_j Cov^4(X_j w_j, F)$ or $Min_{\ F\ =1} \sum_j \ X_j X_j^T - \lambda_j F F^T\ ^2$			
(14) Generalized PCA (Casin 2001)	$Max \sum_j R^2(F, X_j) \sum_h Cor^2(x_{jh}, \hat{F}_j) \text{ (c)}$			
(15) MFA (Escofier and Pagès 1994)	$Min_{F,p_j} \sum_j \left\ \frac{1}{\sqrt{\lambda_{first}[Cor(x_{jh}, x_{jh})]}} X_j - F p_j^T \right\ ^2$	Hierarchical (applied to the reduced X_j)	A	Path-weighting
(16) Oblique maximum variance method (Horst 1965)	$Min_{F,p_j} \sum_j \left\ X_j \left(\frac{1}{n} X_j^T X_j \right)^{-1/2} - F p_j^T \right\ ^2$	Hierarchical (applied to the transformed X_j)	A	Path-weighting
		(d)		(e)

Different Weighting Strategy

How to **balance the influence** of the different blocks in a global analysis?

The block combination is based on **specific weighting schemes**:

- Data concatenation
→ each block as a **weight of one** (SUM-PCA, MBPCA)
- Unsupervised methods
→ weights depend on **block dispersion or agreement with a compromise** (Multiple Factor Analysis, STATIS, CCSWA)
- Supervised methods
→ block weights are **driven by the Y response** (MBPLS, block-PLS, consensus OPLS)



Multiblock Model Outputs

New common subspace

Common/distinct component(s)

Global/local observations scores

→ Pattern recognition

Blocks weights (contributions)

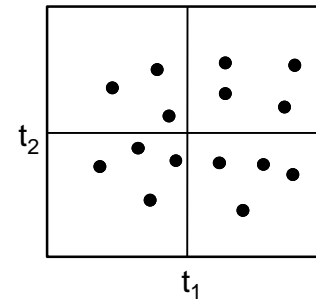
Loadings of the initial variables

Common/specific variation(s)

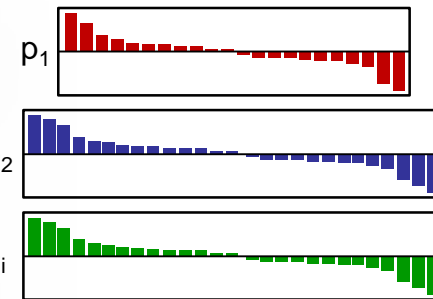
→ Balance between block weights

More complete interpretation:

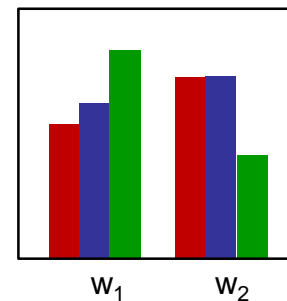
- Links between variables
- Links between blocks



Global/Local scores



Variables loadings



Blocks weights



Joint Matrix Factorization

Phylogeny of some multiblock methods and relations to basic data analysis methods

Green branch: Unsupervised multiblock

Yellow branch: Supervised multiblock

Red branch: Multiway

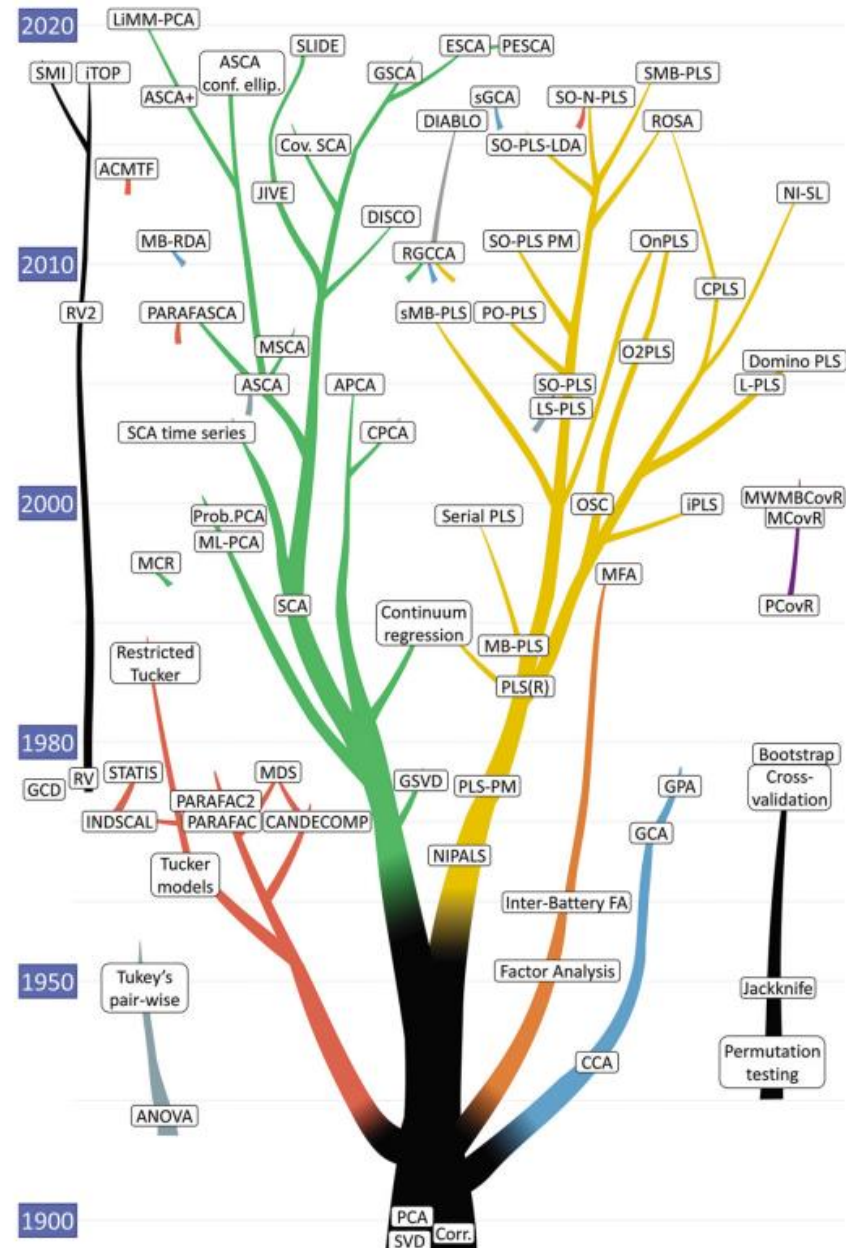
Blue branch: Correlation

Orange branch: Factor analysis

Black branch: Model validation

Multiblock Data Fusion in Statistics and Machine Learning: Applications in the Natural and Life Sciences
Wiley 2022

Age K. Smilde, Tormod Næs, Kristian Hovde Liland



Unsupervised Multiblock Analysis

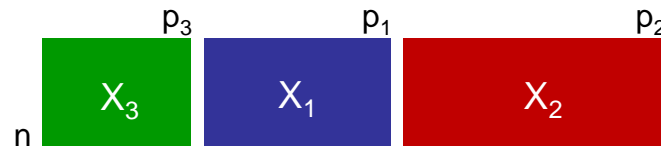
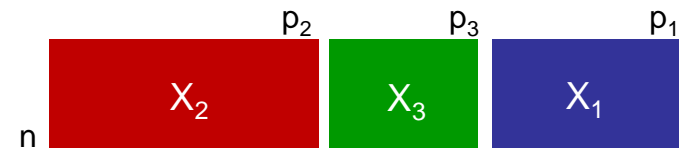
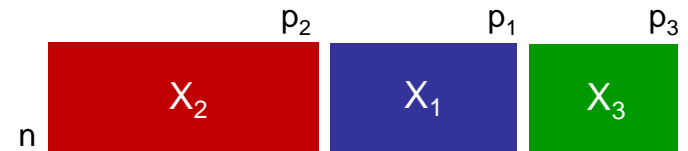
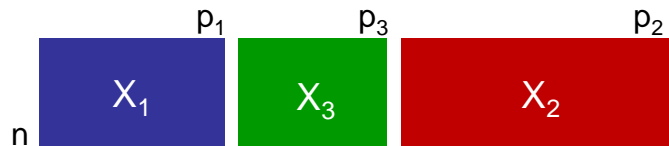
Generate hypotheses from the data blocks

Undirected links

All blocks are treated in the same way

→ the blocks are exchangeable

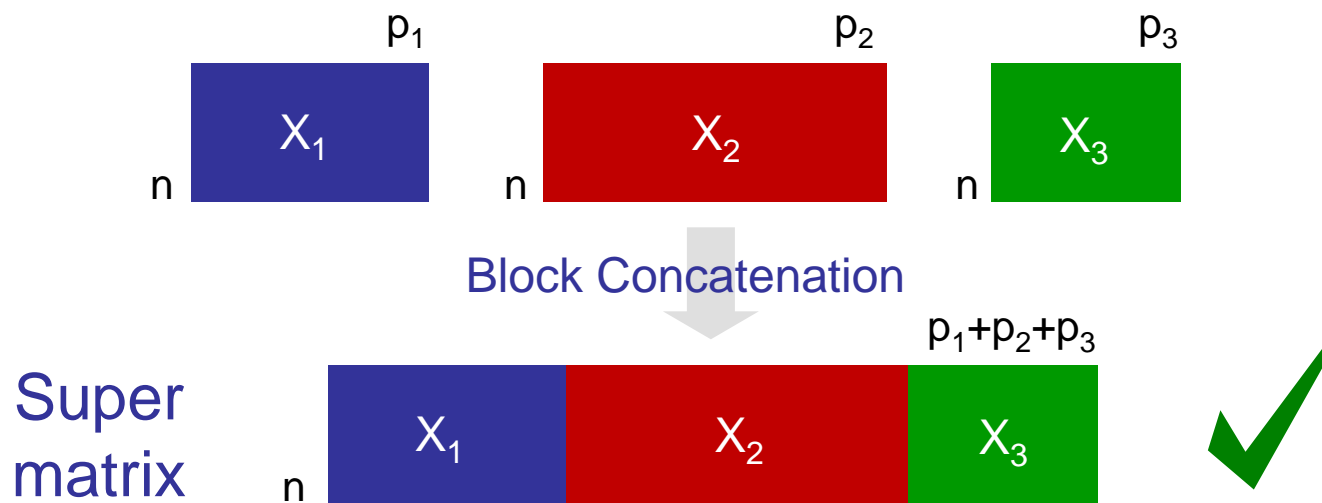
→ no block sequence



Data Concatenation

First simple idea (e.g. SUM-PCA) :

- 1) Block concatenation
- 2) Standard multivariate method



Focus on **common variation** only

Do not explain explicitly the **relationship** between the tables

Block weight is a function of the **number of variables** and **variance**

→ Some blocks may play a more important role **due to their size/scale**

Limitations of Data Concatenation

FAIRNESS BETWEEN BLOCKS

Weight of each block depends on its size/range



Do all blocks contribute equally to the model?

Should all blocks play a role in each component?

Would it be useful to give more weight to informative blocks?

TYPE OF VARIATION EXPLAINED

Average trends of variations between observations



May not be relevant to explain block patterns

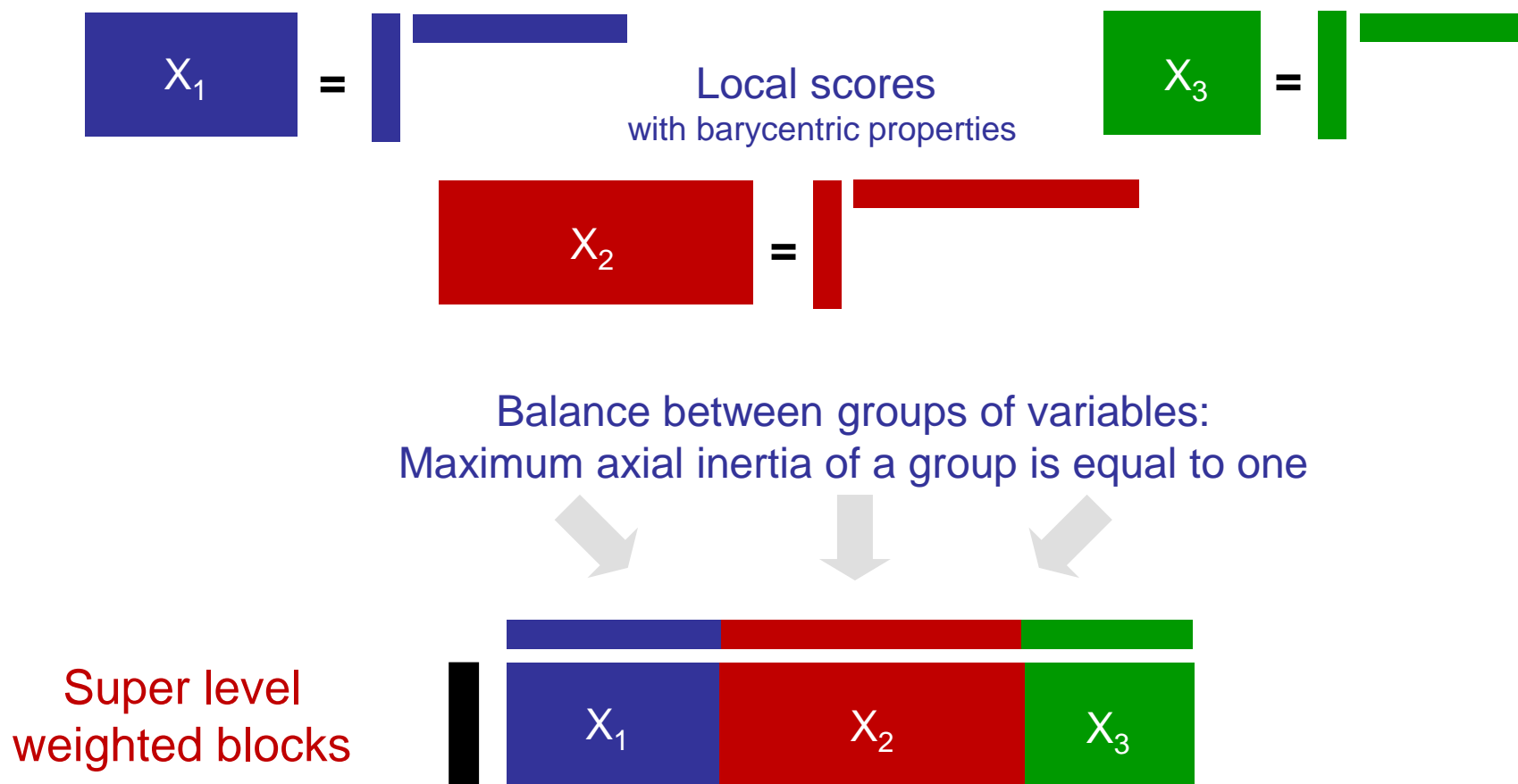


Do we only want to explain common variation between blocks or also within blocks?



Multiple Factor Analysis

Data concatenation with **block weighting** based on the first latent variable



Multiple Factor Analysis

Focus on **common variation** only

Balancing **maximum axial inertia** rather than the total inertia
(= the number of variables in PCA with UV scaling)

→ Maximum axial inertia of each group equals one

MFA assigns to each variable of group a weight equal to the **inverse of the first eigenvalue** of the individual analysis

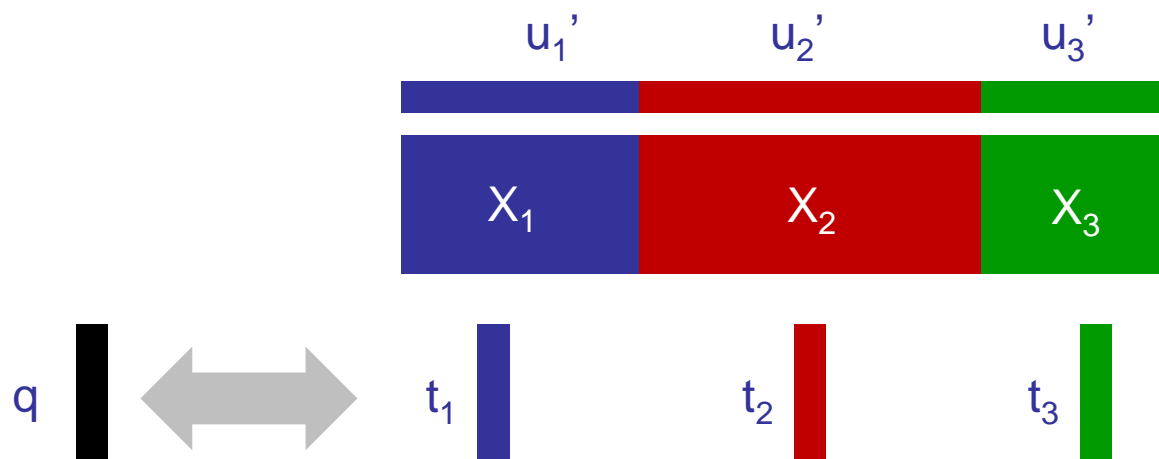
MFA weighting scheme takes into account that a multidimensional group influences naturally more axes than a one-dimensional group

MFA is based on **weighted concatenated PCA**



Multiple Co-inertia Analysis

Maximization of the link between the canonical variables



q : first standardised principal component of the merged data matrix

$$\max \sum_{k=1}^n \text{cov}^2(X_k w_k, q) \quad \text{to maximize with} \quad \begin{cases} q : \text{super level} \\ t_k = X_k w_k \end{cases}$$

→ Emphasis is put on common information and **structural similarities**

Multiple Co-inertia Analysis Model

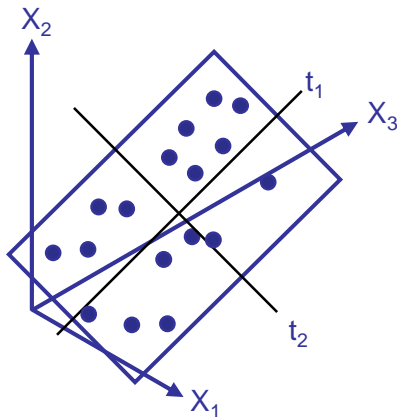
Multiple Co-inertia Analysis (MCoA) components are extracted according to their **explained variance**

→ Similar to PCA

MCoA leads to components

- well explaining their own block
- as **positively correlated** as possible to the **first principal component** of the concatenated data table

Each block is deflated with respect to its associated vectors of loadings



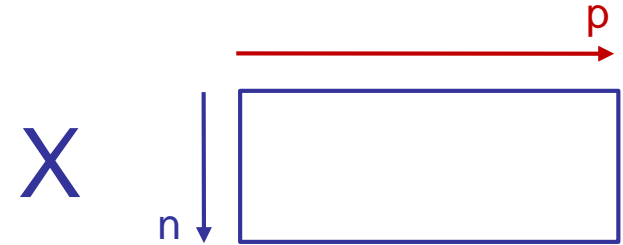
→ the **vectors of loadings** within each block are constrained to be **orthogonal**

→ global **vectors of scores** are also **orthogonal**

Different Types of Data Matrix

- **The data matrix**

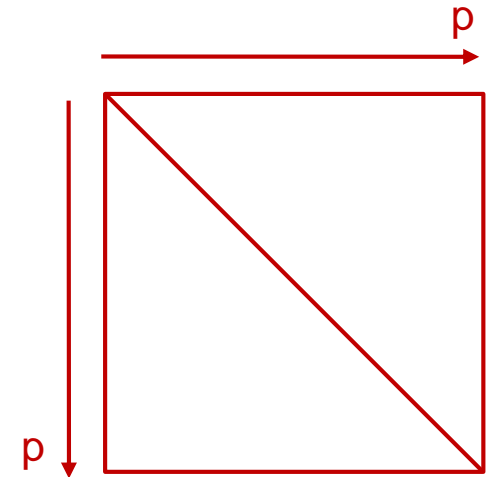
p variables for each of n samples
presented in a rectangular matrix
 n rows and p columns



How to extract the
new axes of the subspace ?

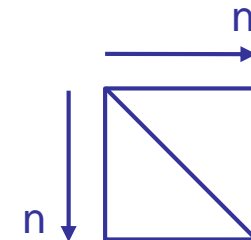
- **The covariance/correlation matrix**
Similarity between every pair of
✓ variables

$$X'X$$



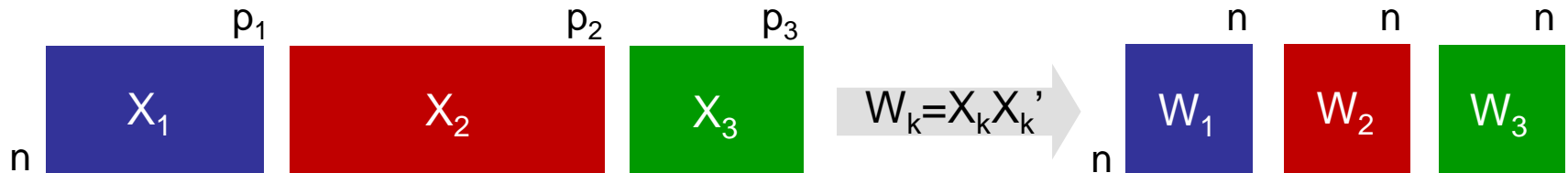
- **The association matrix**
Similarity between every pair of
✓ samples

$$XX'$$



STATIS Method

The scalar product defines an **association matrix** for each data block
→ Similarity between observations within a block (covariances)

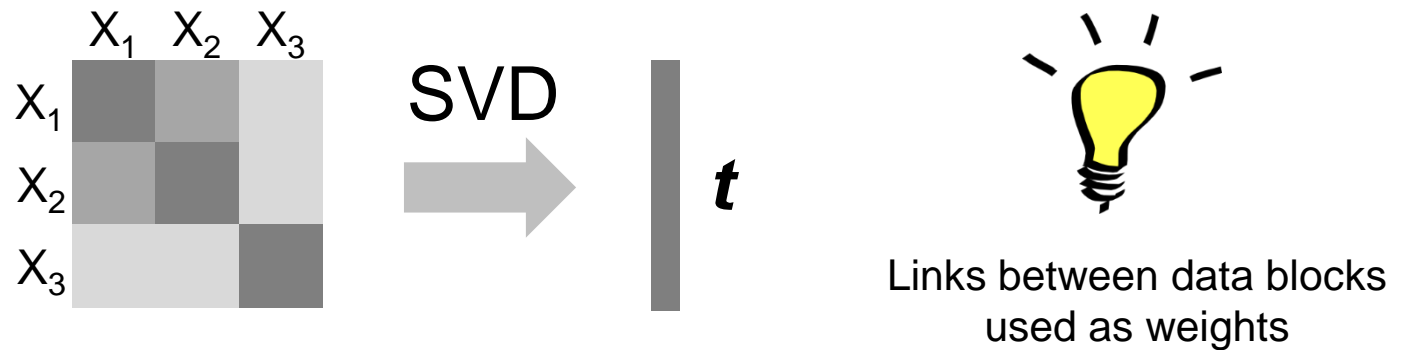


The weighting scheme is based on the RV coefficient between W_i
Cosine based on the Frobenius norm

$$R_V = \frac{\text{trace}\{\mathbf{XX}^T \mathbf{YY}^T\}}{\sqrt{(\text{trace}\{\mathbf{XX}^T \mathbf{XX}^T\}) \times (\text{trace}\{\mathbf{YY}^T \mathbf{YY}^T\})}}$$

Decomposition of the matrix of similarities between all W_i
→ Identity link between data blocks

STATIS Method



STATIS is based on **weighted concatenated PCA**

Variables are **weighted similarly** whether they contribute to the links between blocks and the compromise or not

The strength of the link between data blocks is evaluated **without accounting for their dimension**

→ Emphasis is put on **global similarity** (*i.e.* inter-structure analysis)

As a second step, the intra-structure can be investigated

→ Assess similarities/differences between individuals/variables

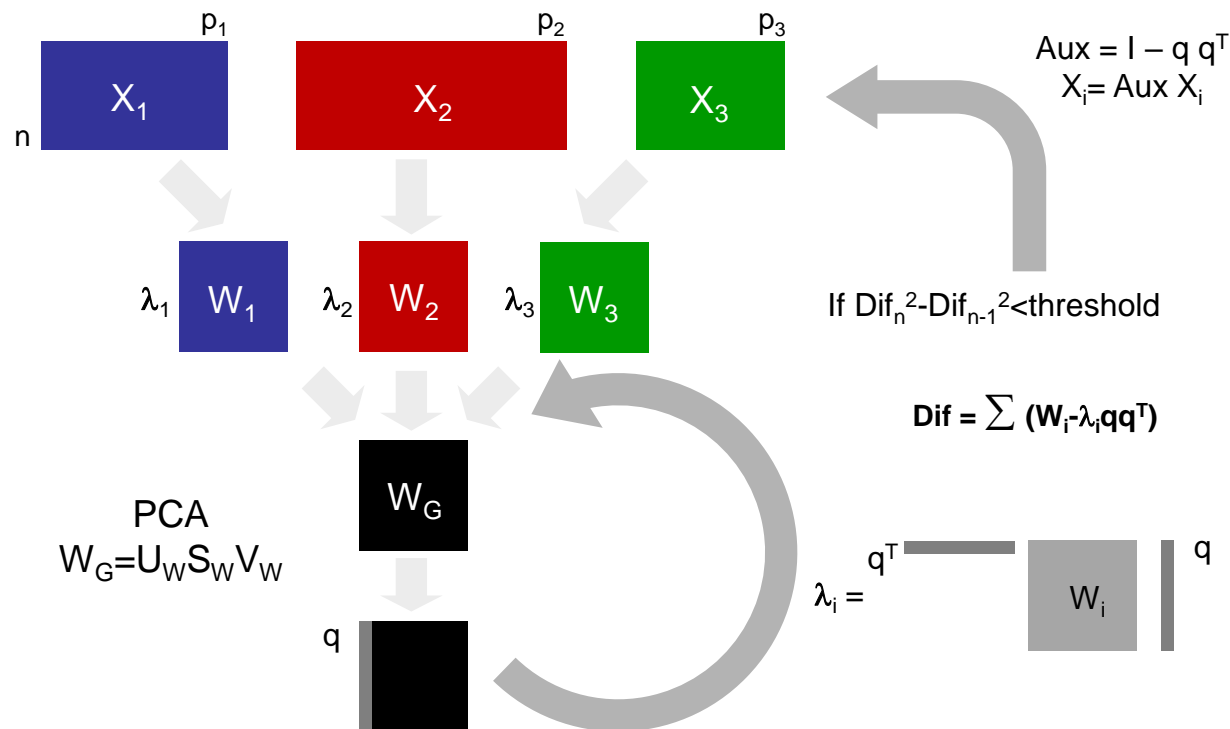
ComDim Method

Common Dimensions

Common Components and Specific Weights Analysis (CCSWA)

Scalar product + iterative block weighting

→ Block weighting is different from a component to another



Link between blocks:
covariance⁴

$$\max \sum_{k=1}^n cov^4(X_k w_k, q)$$

CCSWA Model

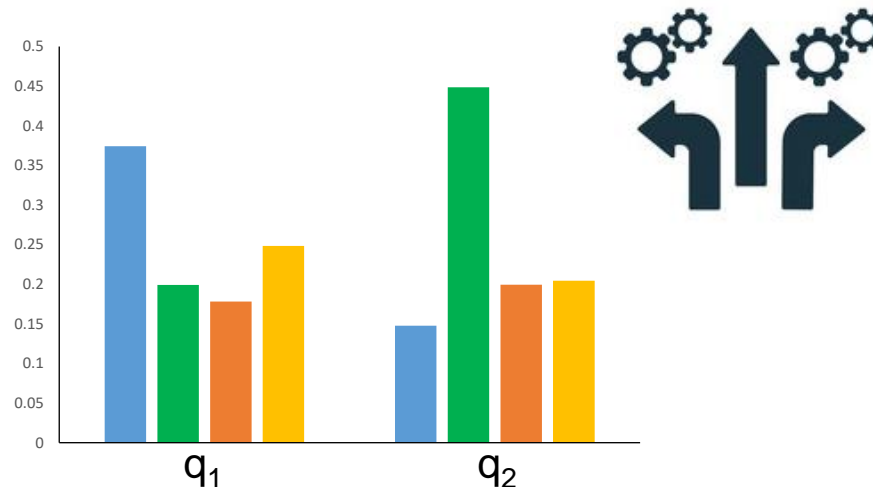
CCSWA components are extracted according to their **explained variance**
→ Similar to PCA

But **more flexibility is included** !

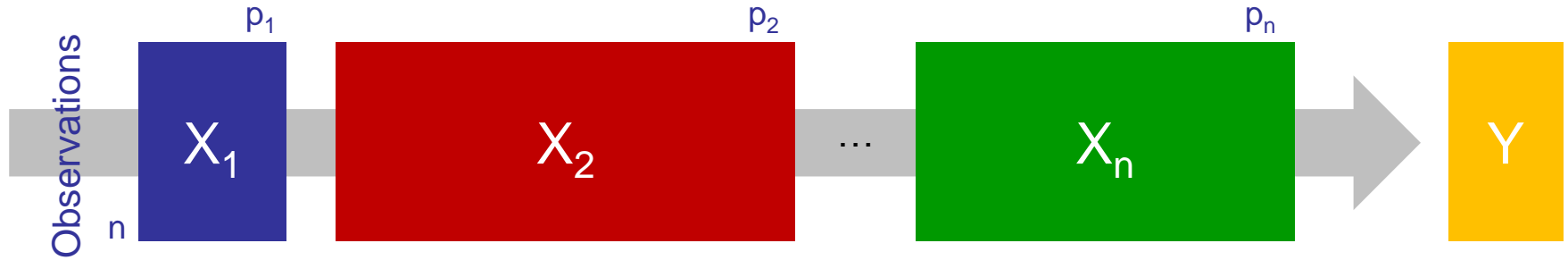
Data blocks can contribute or not to a component

- First components will tend to aggregate several data blocks
→ large variance
- Higher components may grasp more specific trends
→ lower variance

Block
saliences



Multiblock Supervised Analysis



Add another block, but with a different role in the system
→ the blocks are **no longer exchangeable**

Predictive relationship
→ regression approach (usually PLS-type)

BOTH PREDICTION ABILITY AND INTERPRETATION ARE IMPORTANT



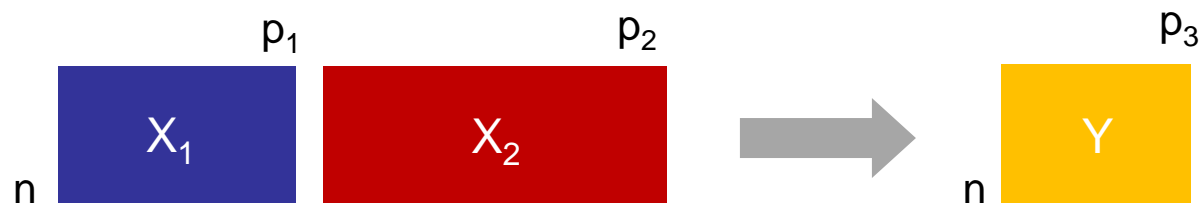
Some methods account for **the sequence of blocks (hierarchy)**
Choosing the linking structure is an a priori decision (domain-specific)

Multiblock Prediction Using PLS

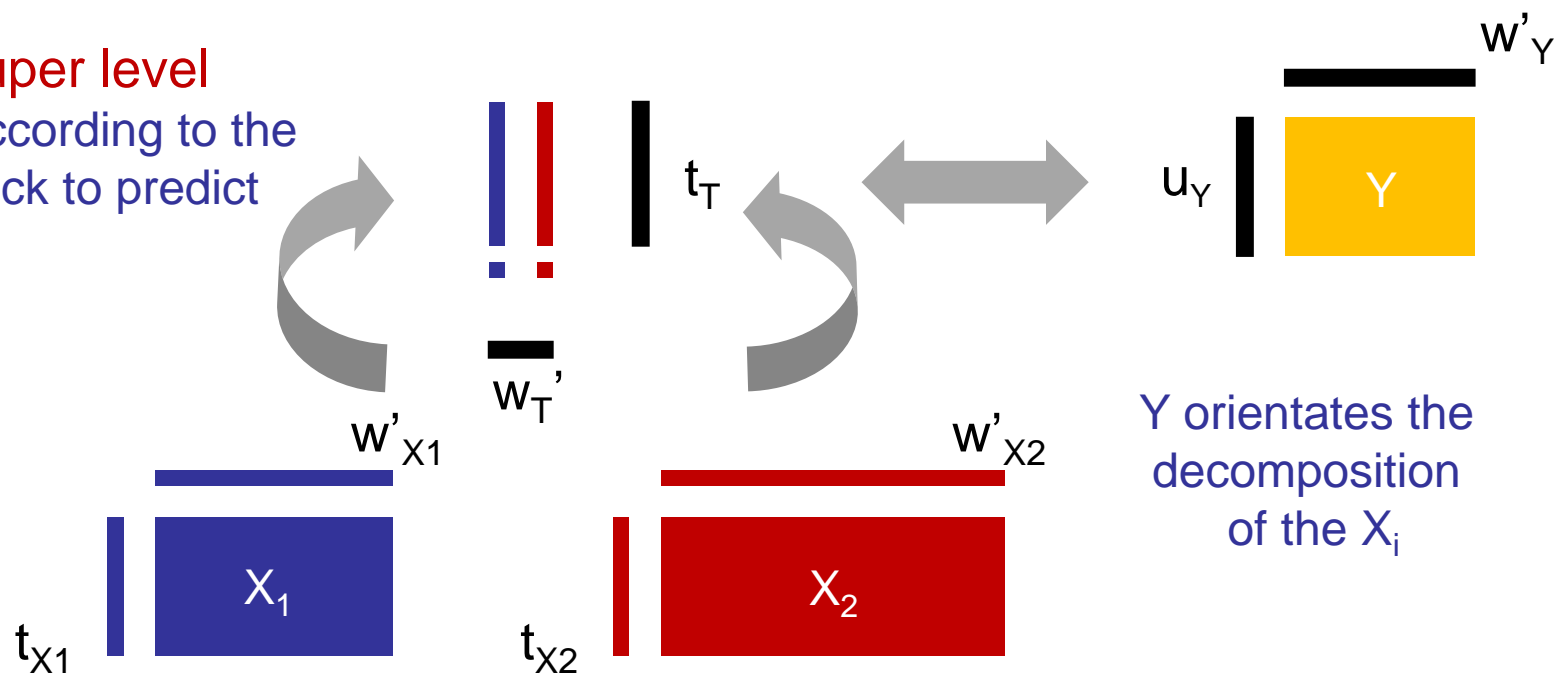
A collection of X_i blocks and one Y response block

→ Multiblock PLS (MBPLS)

→ Scaling with the square root of the number of variables for each block



Super level
built according to the
 Y block to predict



Deflation In Multiblock PLS

Several multiblock PLS algorithms have been presented in the literature
→ the deflation strategy is different

Deflation of X

- using block scores leads to **inferior prediction** of Y
- using super scores gives the **same predictions** as concatenated PLS

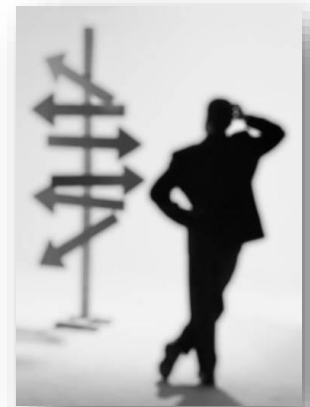
BUT

the information of the blocks gets **mixed up**

If Y is deflated using the super score, these problems disappear

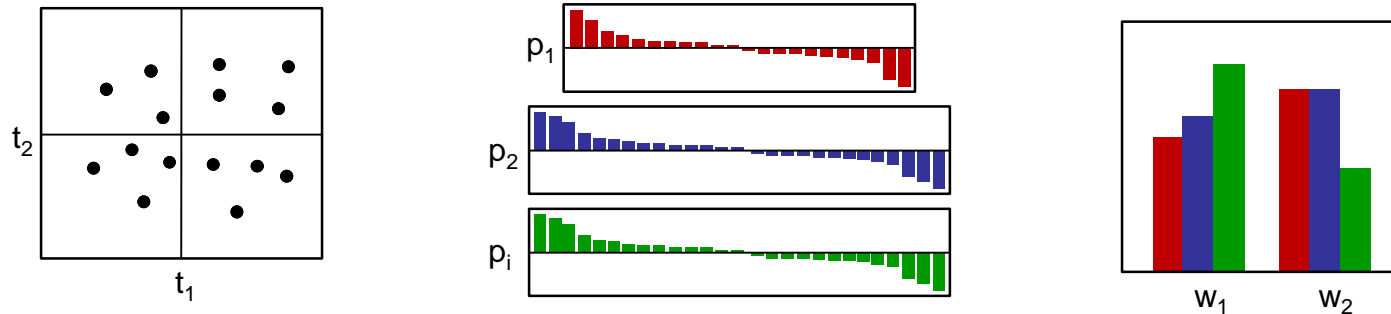
Conclusion (somewhat disappointing) :

- 1) calculate **a PLS model using concatenated data**
- 2) estimate the multiblock parameters from this model



Multiblock PLS

Interpretation is typically based on **score** and **loading** plots

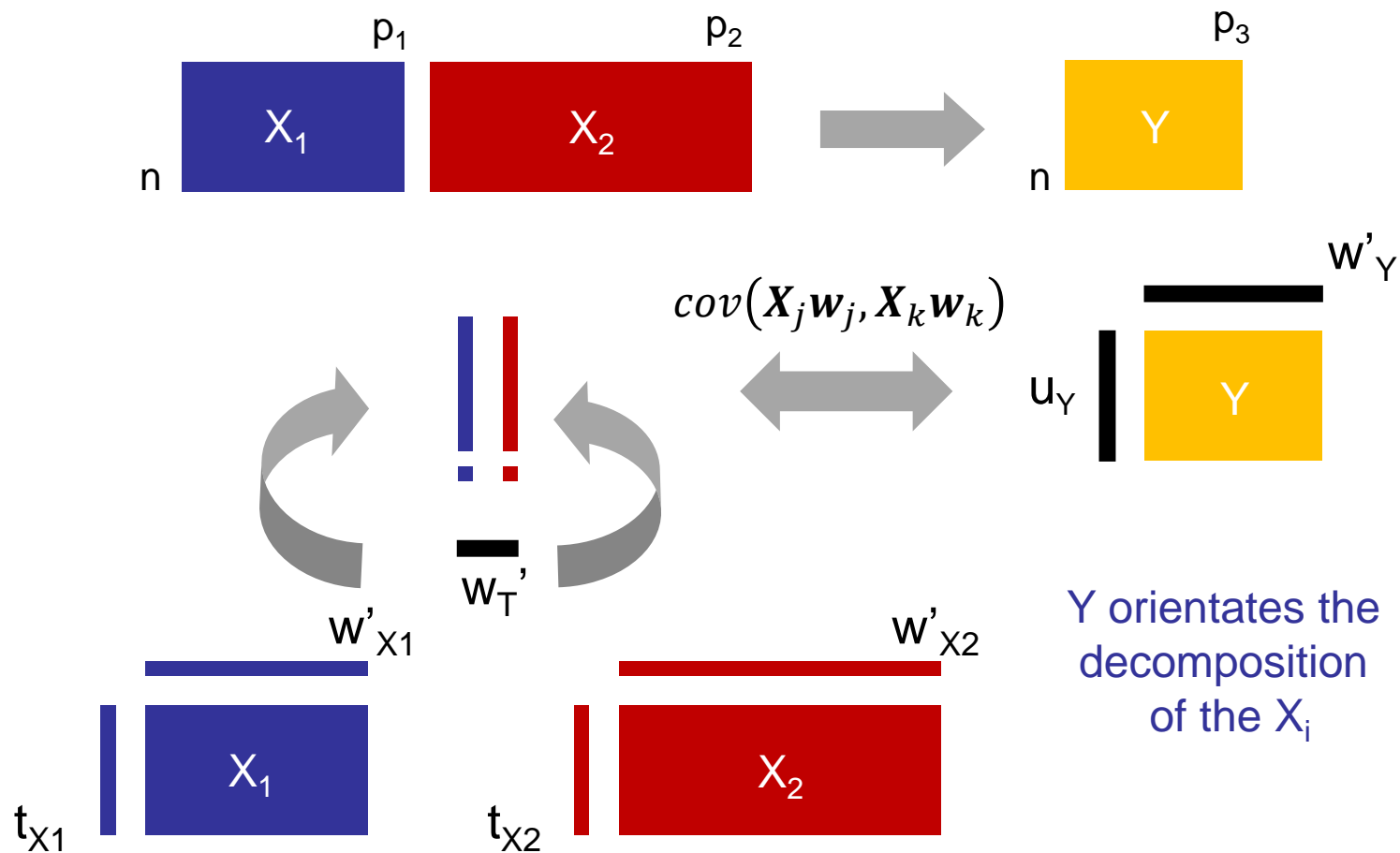


The **same number of components** is extracted for all blocks
→ This may complicate interpretation in cases with very different underlying dimensionality

Prediction ability (Q^2Y) is usually measured by **cross-validation** and/or **test set** prediction

Block-PLS

The covariance between pairs of components is maximized
→ special case of Generalized Canonical Correlation Analysis

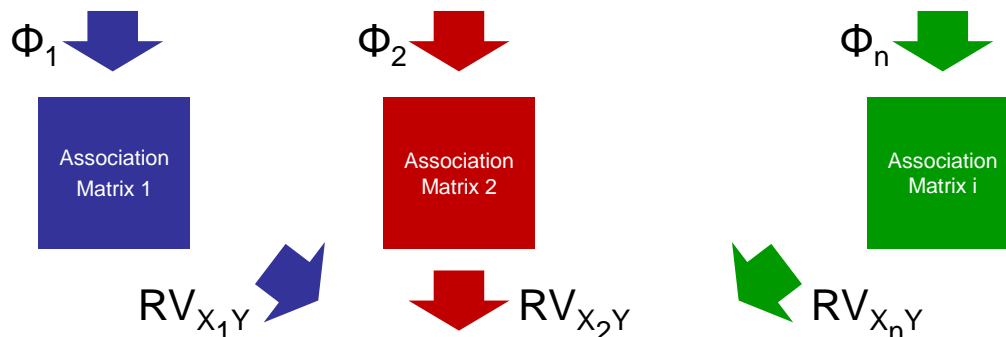


Consensus OPLS Workflow

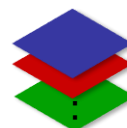
Heterogeneous data
($n \times k_i$)



Association matrices XX^T
(linear kernel, $n \times n$)



Weighted sum matrix
(linear combination, RV coefficients)

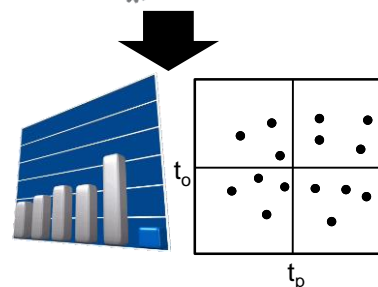


Kernel OPLS algorithm
(dual form)



$$X = t_p p_p^T + t_o p_o^T + E$$
$$Y = t_p q_p^T + F$$

Global results & joint interpretation



Consensus OPLS Model

Multiblock predictive model

New common subspace

Predictive/orthogonal component(s)

Consensus observations scores

Blocks weights (contributions)

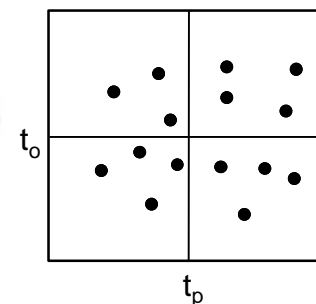
Loadings of the initial variables

Common/specific variation(s)

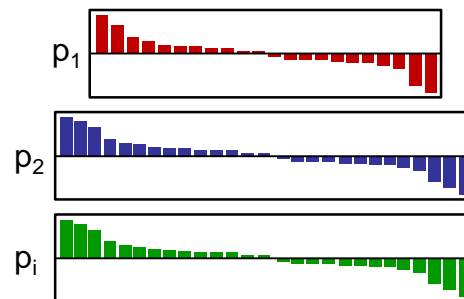
→ Balance between block weights

Easy interpretation

Compromise between prediction ability and data description

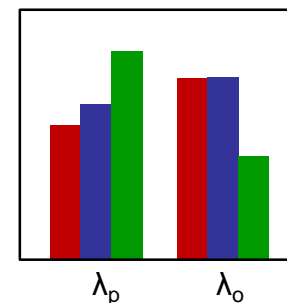


Consensus scores



Variables loadings

$$p_i = \frac{X_i^T * t}{(t^T * t)}$$



Blocks weights

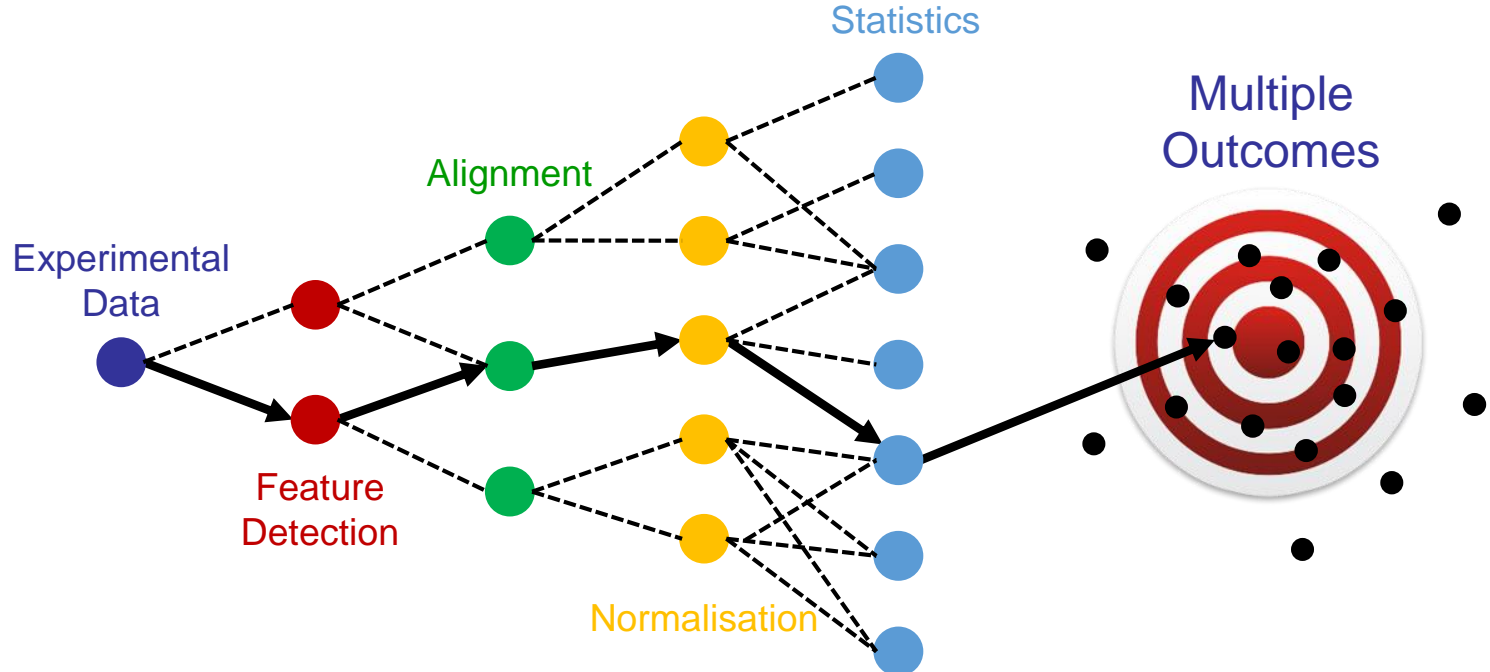
$$\lambda_i = t^T * K_i * t$$

Data Processing Parameter Space

Data can lead to different results depending on **processing parameters**
→ The parameter space in data processing is huge

Each model obtained is just **one of many representations** of the actual data

Evaluation of the sensitivity of crucial parameter to find proper values
→ better understand the underlying data structure



Model Validation

Validation needs to be done at different levels!

- Theoretical appropriateness:
Does the model used fit the **goal** and the **data structure**?
- Computational correctness
Is the solution a local or a **global optimum**?
- Statistical reliability
Are the **assumptions** appropriate?
Are the **solutions stable** under resampling?
- Explanatory validity
Can **new knowledge** be gained from the model?



Statistical Model Evaluation

How well does the model **describe** the data?

- too little structure (**underfitting**): poor prediction/description
- too much noise (**overfitting**): poor generalization ability

The ideal model captures

- ✓ all of the **replicable structure** in the data
- ✓ none of the **noise**

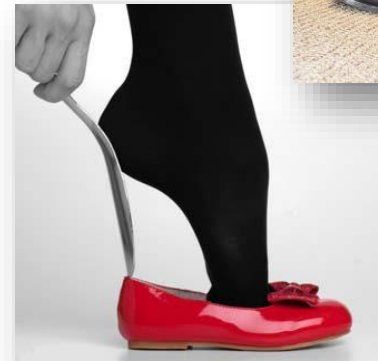
CRITERIA TO INVESTIGATE

Unsupervised

- stability of found patterns

Supervised

- predictive model performance (generalization ability)
- significance of contributions to a model (factors, blocks or variables)



Model Evaluation

Indexes of **model quality**:

R^2X : how well the model describes the X blocks

R^2Y : how accurate are the model predictions

BUT with a limited number of observations:

✗ not sufficiently informative

✗ completely optimistic



How to evaluate the **generalization ability** of a model?



Data-driven methods based on **resampling**

→ No underlying distribution assumptions

(i) Cross-validation (Leave-One-Out, K-fold)

(ii) Permutation testing

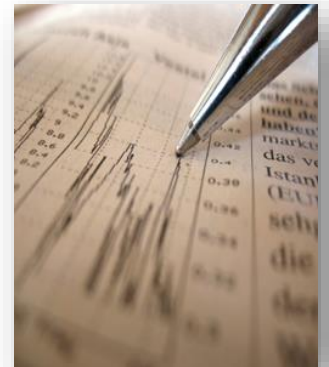
(iii) Bootstrap

Cross-Validation

Cross-validation is a statistical method for validating a predictive model

Main idea: predictive virtues can only be assessed for **unseen data**

- a) Divide the data into **a training set** and **a validation set**
- b) Fit a model to the **training set**
- c) **Predict the validation set** with the model
- d) Repeat the procedure with other training and test sets



Average the quality of the predictions across the validation sets

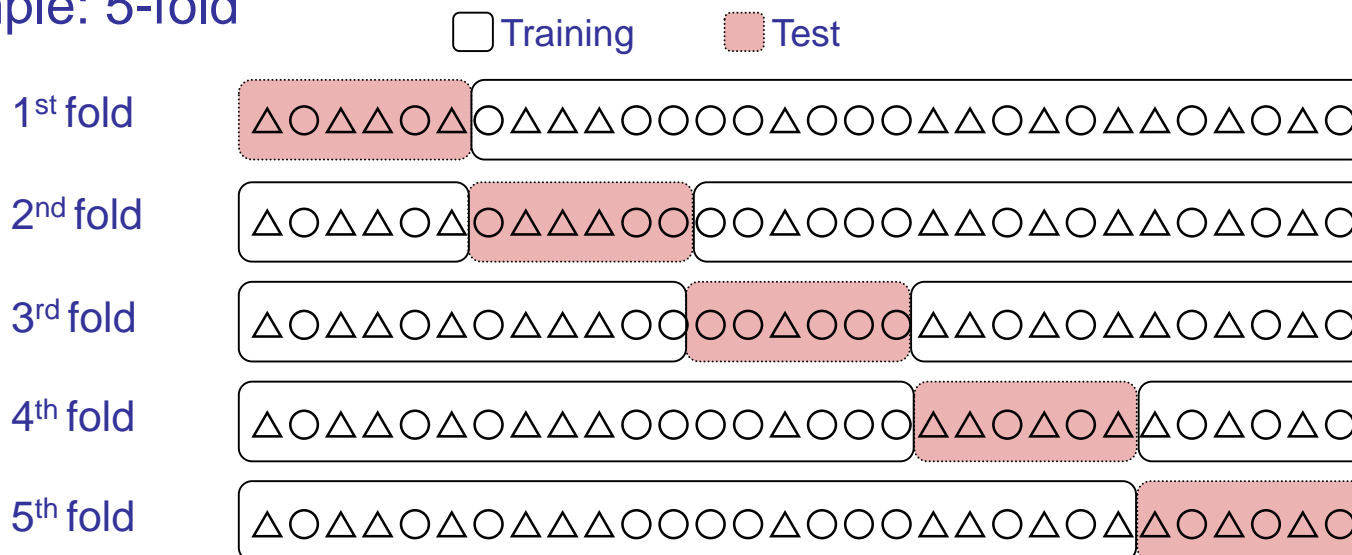
→ overall measure of the **prediction accuracy**

→ quality index: **goodness of prediction Q^2Y**

K-fold Cross-Validation

- ✓ The dataset is randomly **partitioned** into **K** subsets
- ✓ A **single subset** is retained as the **validation data** for testing the model
- ✓ The **remaining K-1** subsets are used as **training data**
- ✓ The cross-validation process is then **repeated K times** (the folds)
- ✓ Each of the **K** subsets is used exactly once as the validation data
- ✓ The **K** results from the folds are then combined

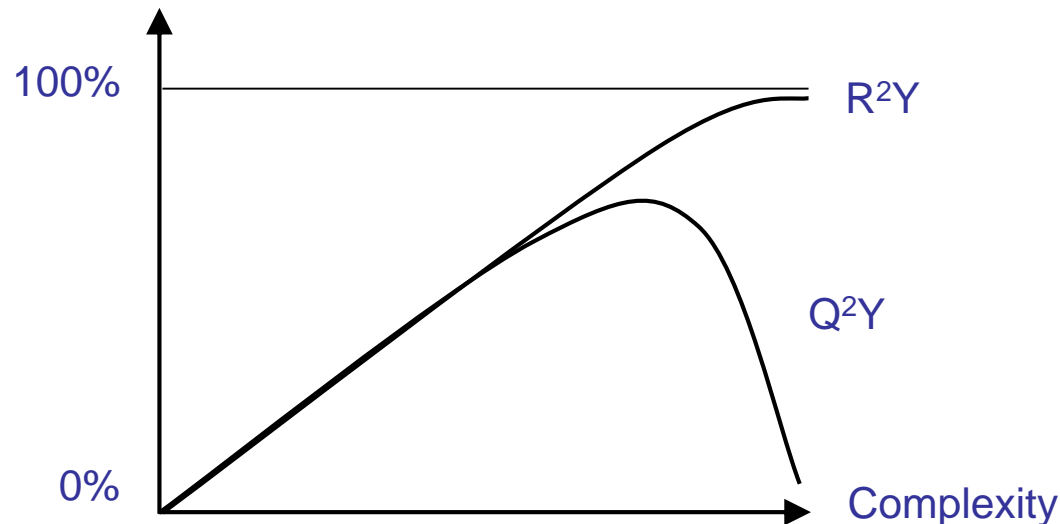
Example: 5-fold



R^2Y vs. Q^2Y

R^2Y : goodness of fit (the portion of data fitted by the model)
→ converge to 100% when adding successive parameters
(increasing the complexity of the model)

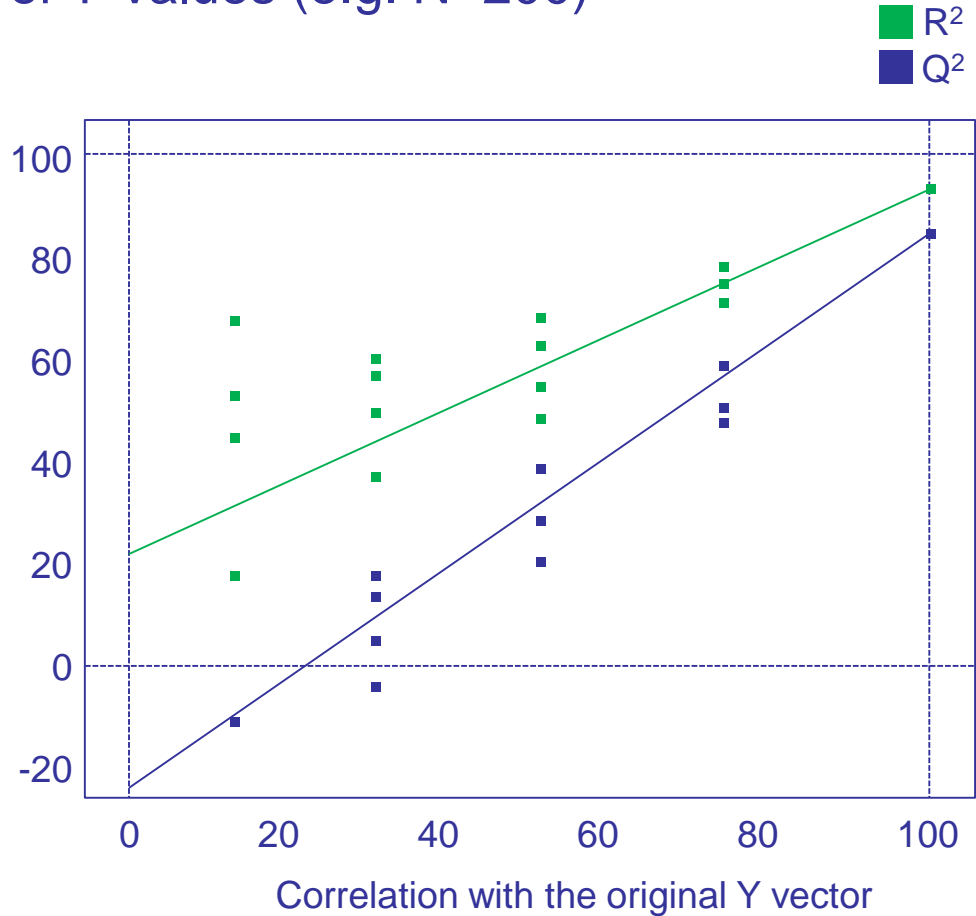
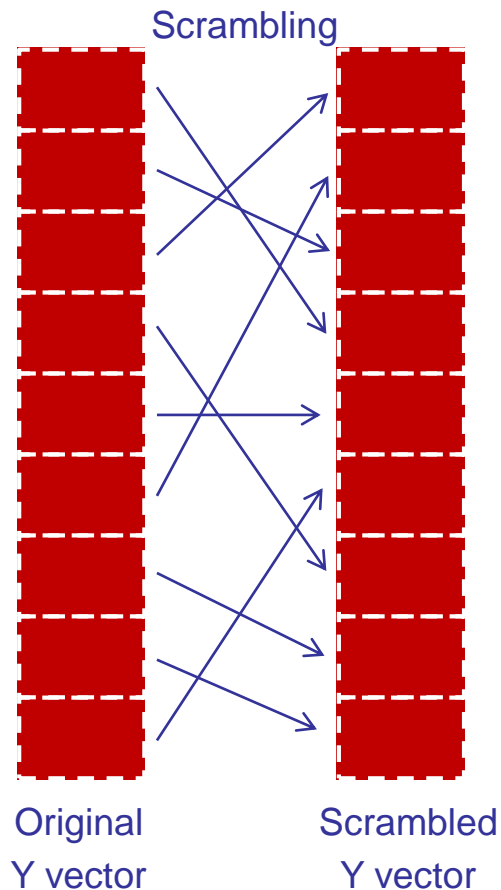
Q^2Y : cross-validated goodness of prediction
✓ increase if valuable predictors are added
✗ decrease if worthless predictors are added



Permutation Tests

Y-scrambling or Y-shuffling

→ Random Permutations of Y Values (e.g. N=200)

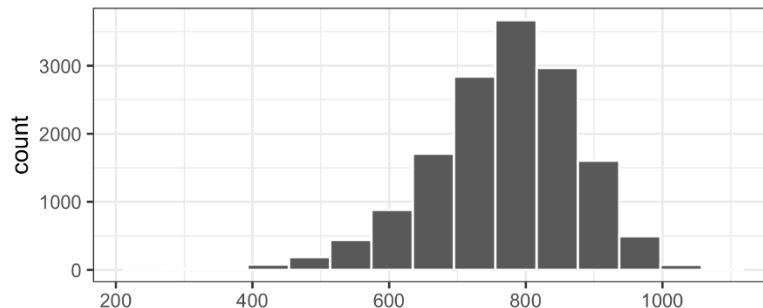


→ Diagnostic by examining the intercepts

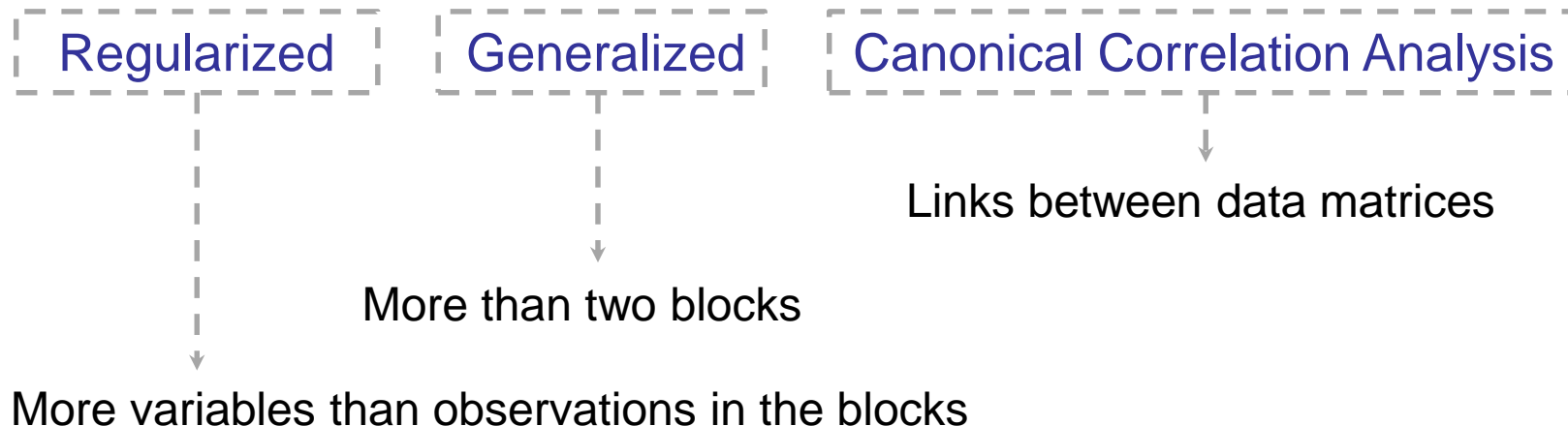
Bootstrap

- The bootstrap is a flexible and powerful statistical tool to **quantify uncertainty** associated with model parameters
- The basic idea is to randomly draw datasets **with replacement** from the training data (same size as the original training set)
- This is done n times, producing n bootstrap datasets and n corresponding sets of model parameters
- The distribution of the model parameters can then be estimated from the bootstrap sampling (e.g. variance)

Simulation-based
distribution



RGCCA



RGCA is a very versatile method

It covers many **multiblock data methods as special cases**

→ Unifying framework for sequential multiblock methods

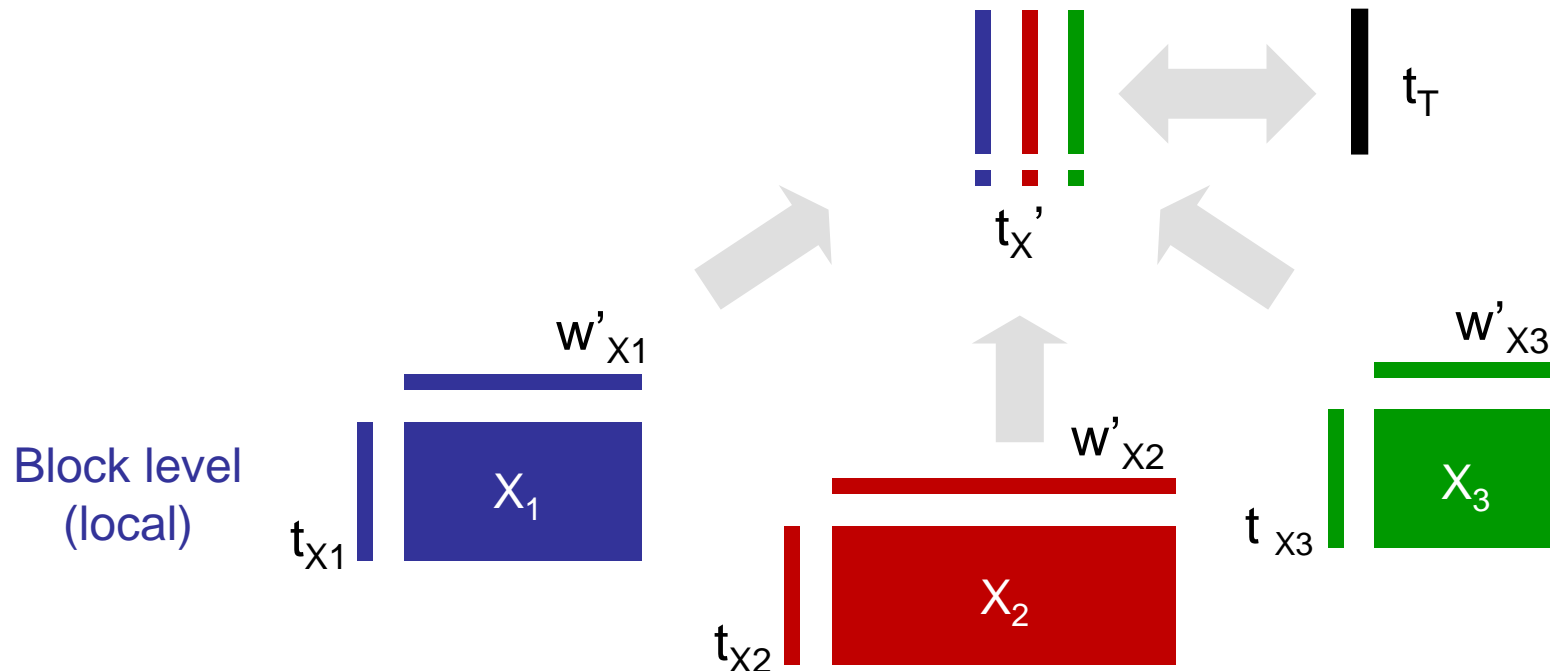
Fairness between blocks is related to the **dispersion of the correlations**

Small variance indicates fairness

Large variance indicates a block selection behavior

RGCCA

Maximize the sum of covariances between linear combinations of the blocks under specific constraints (shrinkage)



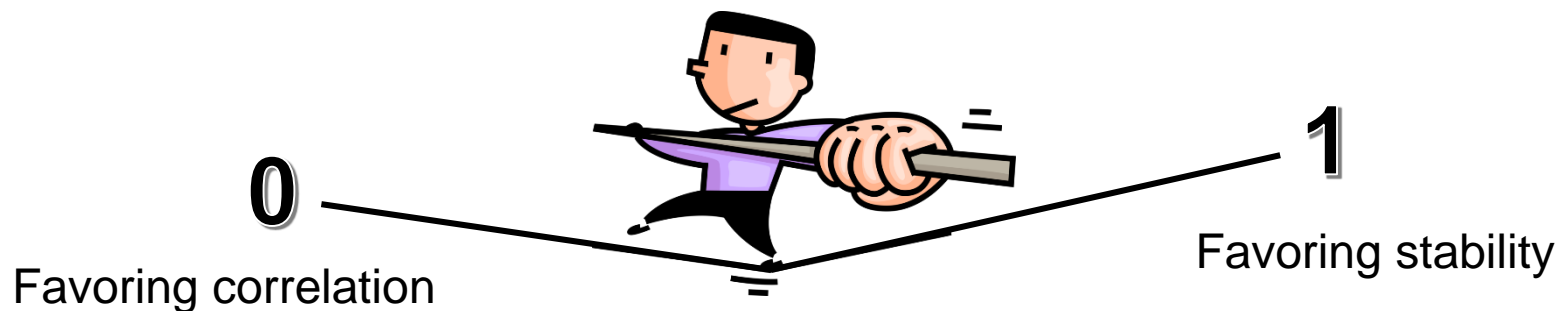
$$\max \sum_{j,k=1}^J c_{jk} g \left(\text{cov} \left((X_j \mathbf{w}_j, X_k \mathbf{w}_k) \right) \right) \quad \text{s.t.} \quad \mathbf{w}_j^t \mathbf{M}_j \mathbf{w}_j = 1, \forall j$$

$$\mathbf{M}_j = \tau_j \mathbf{I} + (1 - \tau_j)(1/n) \mathbf{X}_j^t \mathbf{X}_j$$

Choice of the shrinkage constant τ_j

$$\mathbf{M}_j = \tau_j \mathbf{I} + (1 - \tau_j)(1/n)\mathbf{X}_j^t \mathbf{X}_j$$

τ_j is a tunable parameter (shrinkage constant)



Optimal shrinkage constant values can be found automatically
e.g. using Schäfer & Strimmer formula

$$\tau_j^* = \underset{\tau_j}{\operatorname{argmin}} \mathbb{E} \left[\left\| \hat{\Sigma}_j(\tau_j) - \Sigma_j \right\|_F^2 \right]$$



MOFA & DIABLO

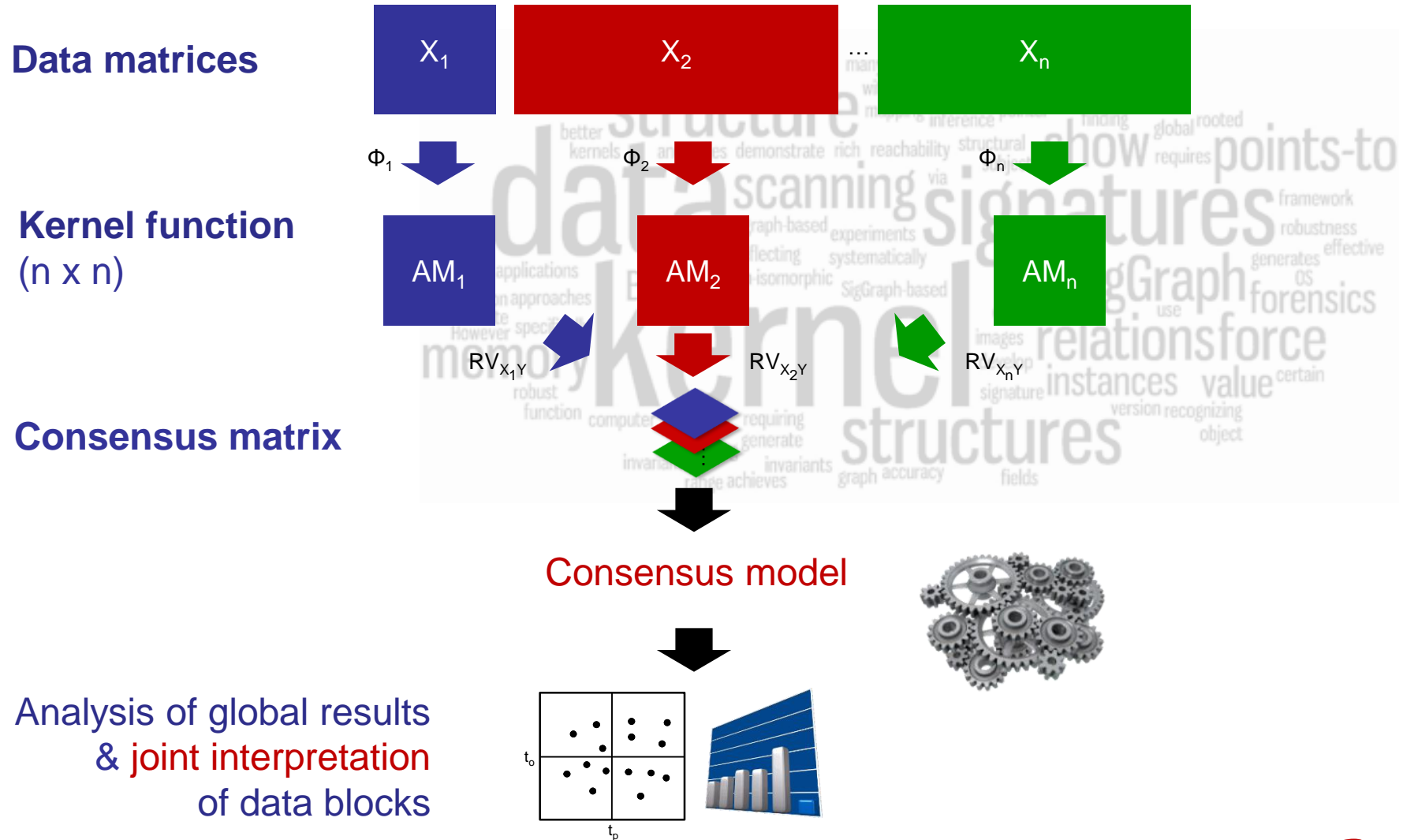
Models based on factor analysis with penalties

Experience is required for **selecting adequate parameters**



- Multi-omics factor analysis (MOFA)
 - Model parameters estimated within a **Bayesian approach with priors**
 - Different penalties can be imposed on the weights
 - **Probabilistic estimation** (different distributions can be used for heterogeneous data)
 - Properties of the estimated scores and loadings **are not always clear**
- Data Integration Analysis for Biomarker discovery using a Latent component method for Omics studies (DIABLO)
 - Sparse implementation of RGCCA
 - Lasso-type penalties on the weights

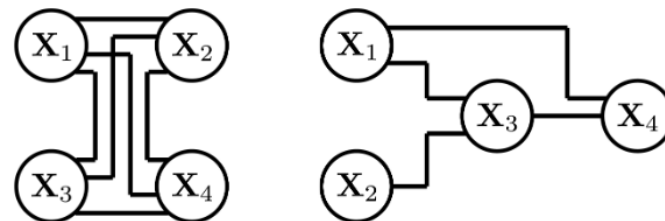
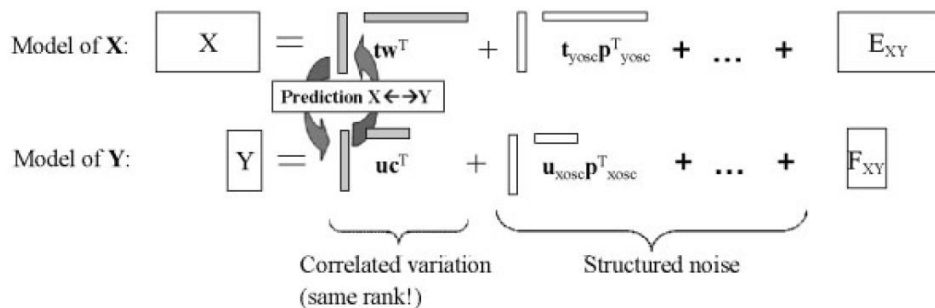
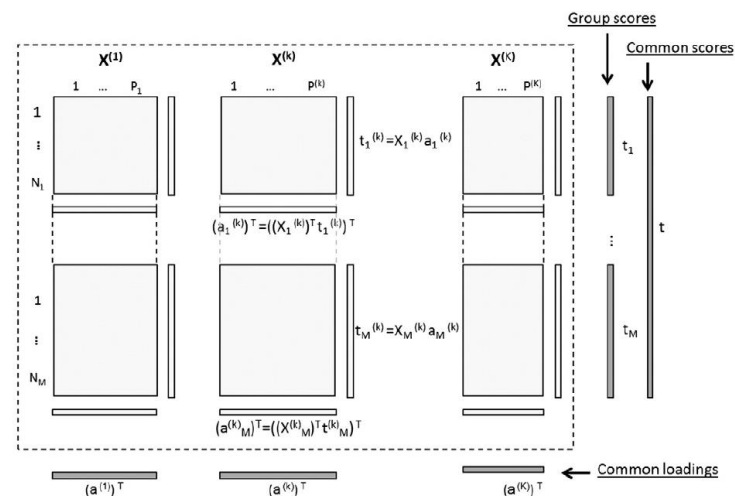
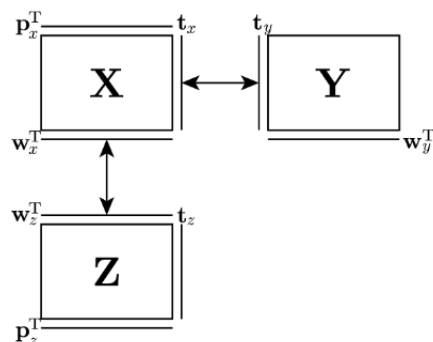
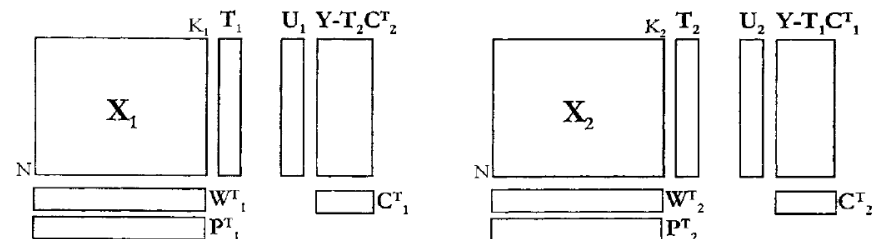
Kernel-based Data Integration



WHAT ABOUT INTERPRETATION ?

Other Multiblock Methods

- Serial PLS
- L-PLS
- GOMCIA
- PLS-Path Modelling
- OnPLS
- DISCO
- JIVE
- NetPCA
- Sparse methods
- and more...



Data Storytelling



Probably the most important question is
which method to use in what situation?

This depends on the goal of the analysis, on the knowledge domain and on the properties of the methods and the data

- Data can have very different nature/structure
- There is no ready made recipe
- Each dataset has its own specificities



- Explore different approaches
- Be creative

Some Take Home Messages



Different data sources can be combined for a **more complete description** of complex data using two-level of interpretation (global and local)



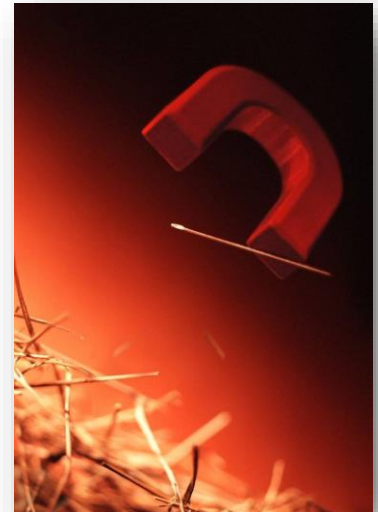
Dedicated chemometric methods allow **common and/or specific directions of variations** to be extracted from the data blocks



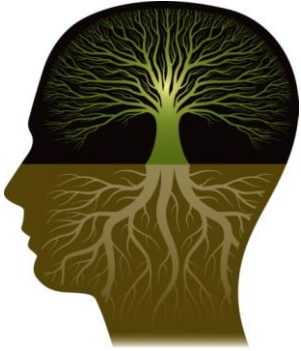
Unsupervised multiblock analysis takes the **relationship between X_i blocks** into account



Supervised multiblock analysis takes **the relationship with the Y response(s)** into account



Toward Biological Insights



Understand links between multiblock data sets
Combine different biological layers
Compare data sources
Choose between analytical methods (or combinations)

Often makes **interpretation easier**
Relate data sources to **common or specific patterns**
Understand **mechanisms** leading to phenotypes
Highlight **subsets of variables** in different blocks

BUT

First **define the question(s)** to be answered !



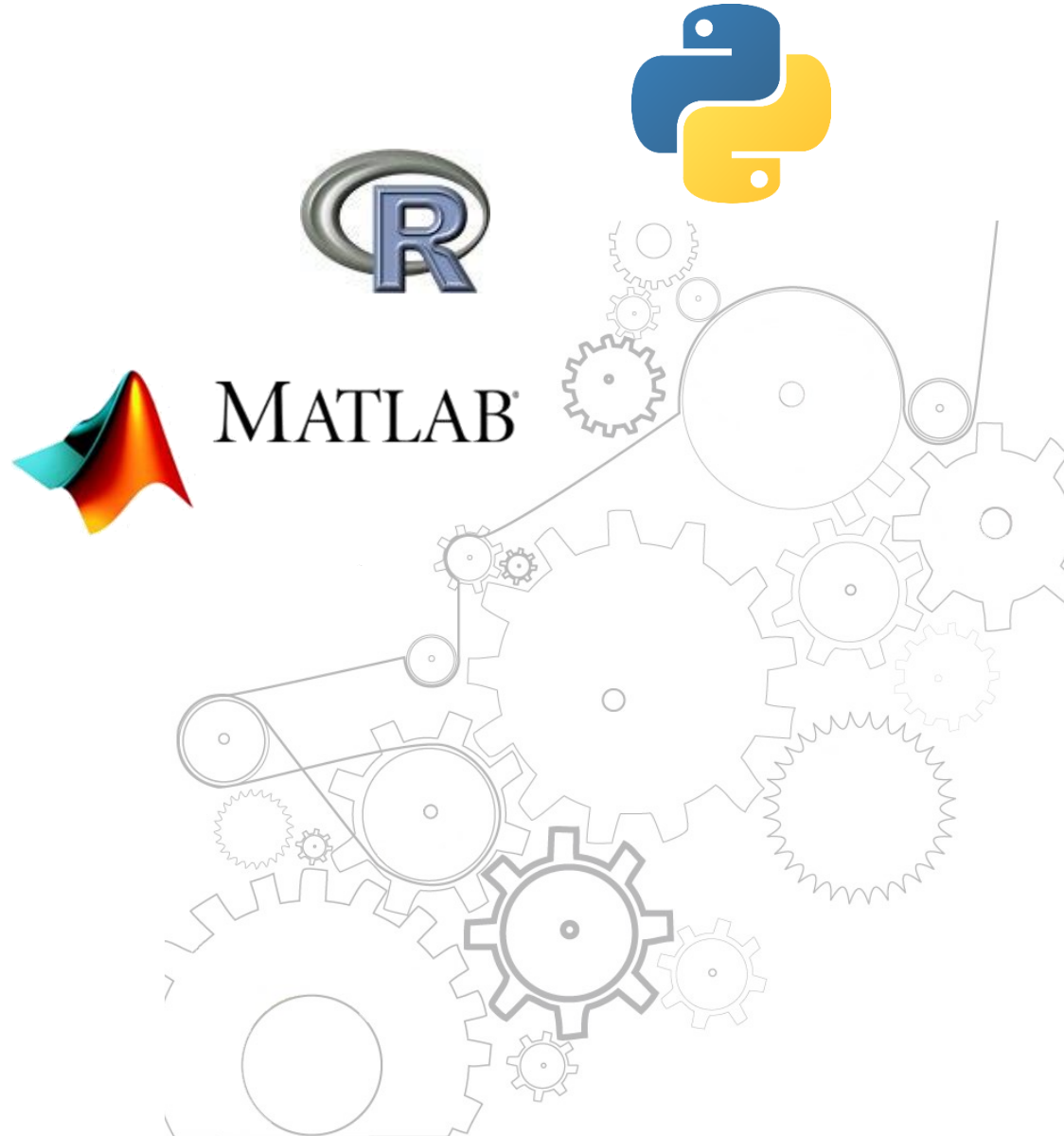
Texas-sharpshooter fallacy

Toolboxes

Packages:

- MBAnalysis
- CCA
- mixOMics
- ade4
- omicade4
- multiblock
- SAISIR
- MBToolbox

...many others



An Excellent Book For Further Reading

Multiblock Data Fusion in Statistics and Machine Learning: Applications in the Natural and Life Sciences

Wiley 2022

ISBN: 978-1-119-60097-8

Age K. Smilde

Tormod Næs

Kristian Hovde Liland

- I. Concepts & Theory
- II. Unsupervised/supervised methods
- III. Complex structures
- IV. Alternative methods
- V. Software (multiblock R package)

