

A generic model for data fusion

Iven Van Mechelen¹ and Age K. Smilde²

¹Research Group on Quantitative Psychology &
Center for Computational Systems Biology, K.U. Leuven, Belgium
Iven.VanMechelen@psy.kuleuven.be

²Biosystems Data Analysis, Swammerdam Institute for Life Sciences,
University of Amsterdam, The Netherlands

Acknowledgement

This talk has benefited from contributions of the following colleagues and collaborators:

Eva Ceulemans

Henk Kiers

Jan Schepers

Martijn Schouteden

Robert Van den Berg

Mariët van der Werf

Katrijn Van Deun

Tom Wilderjans

“data fusion”

“analysis of coupled/linked data”

“multiset data analysis”

“integrative data analysis”

generic model

Overview of this talk:

1. data and problem
2. model
3. research challenges

Overview of this talk:

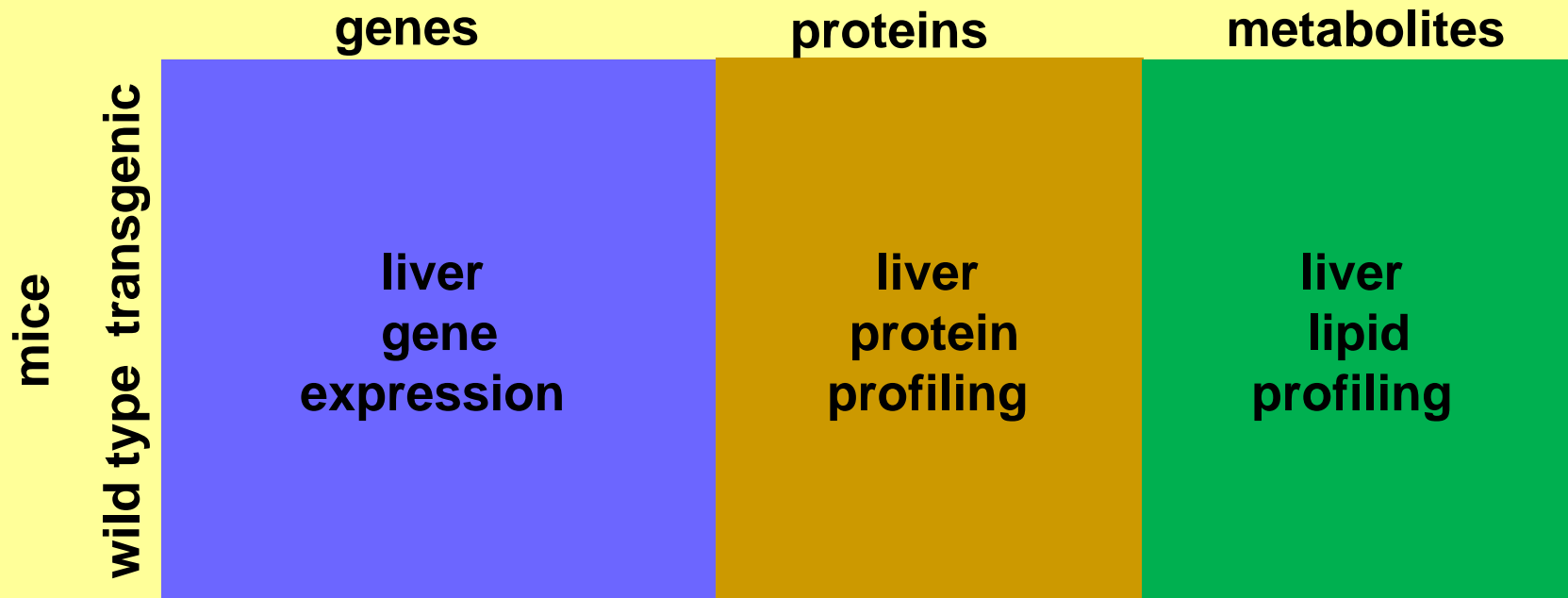
1. data and problem
2. model
3. research challenges

1. Data and problem

- examples of data sets

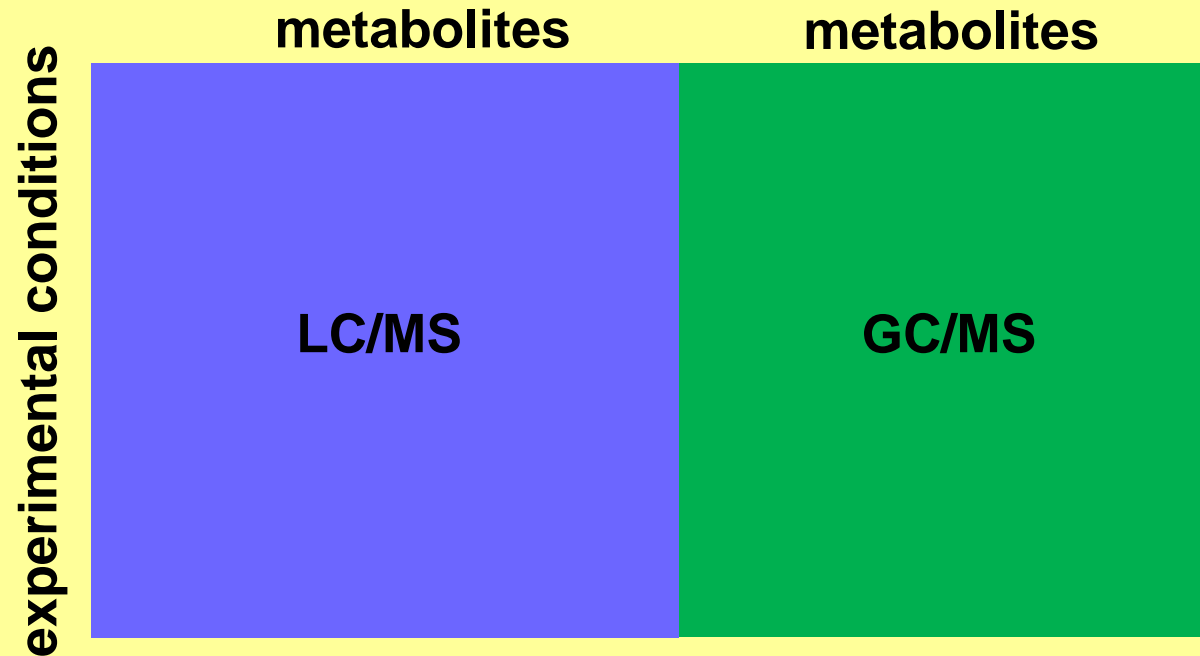
example 1

Clish et al. (2004): artherosclerotic disease model



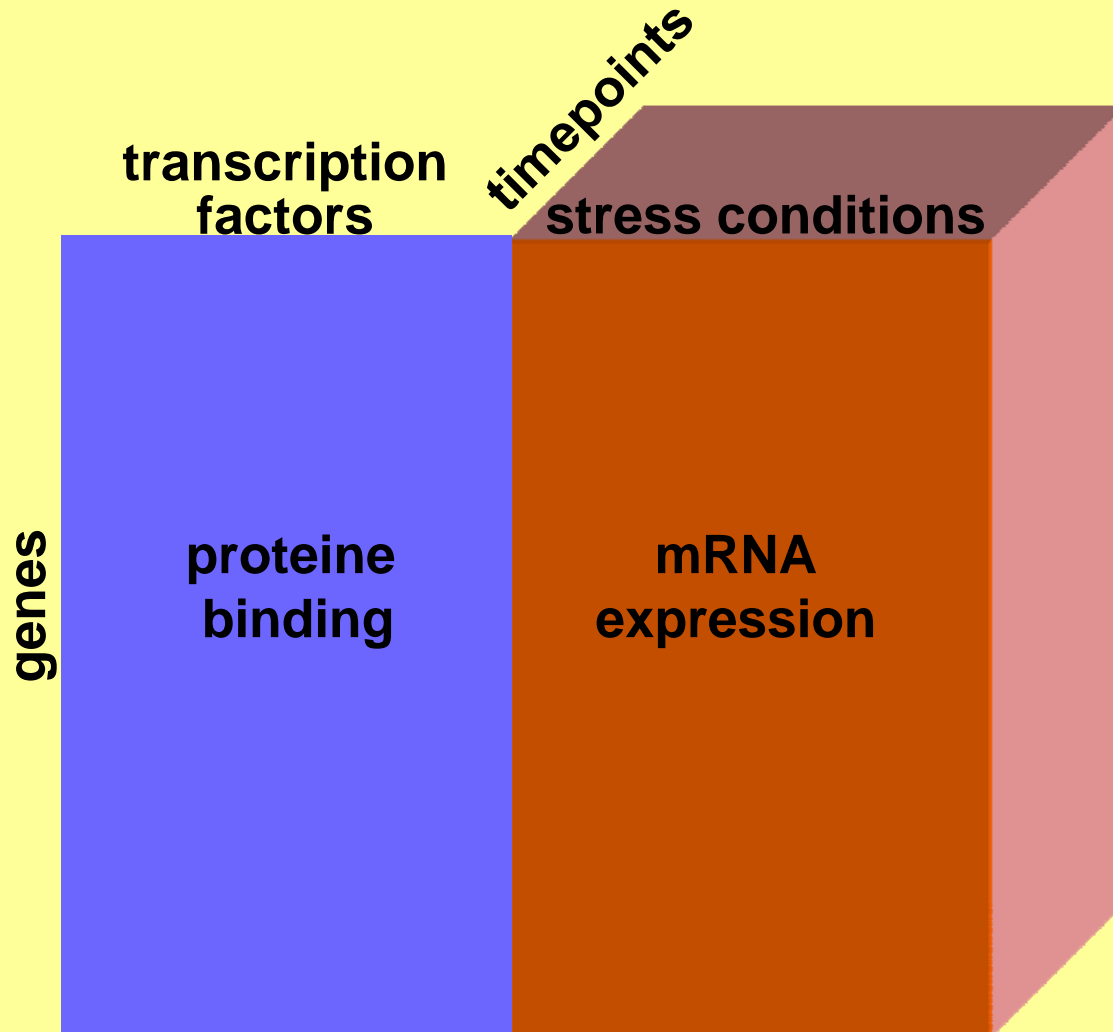
example 2

Smilde et al. (2005): comprehensive view on microbial metabolome



example 3

Omberg et al. (2007): yeast cell cycle time course



1. Data and problem

- examples of data sets
- generic structure of the data

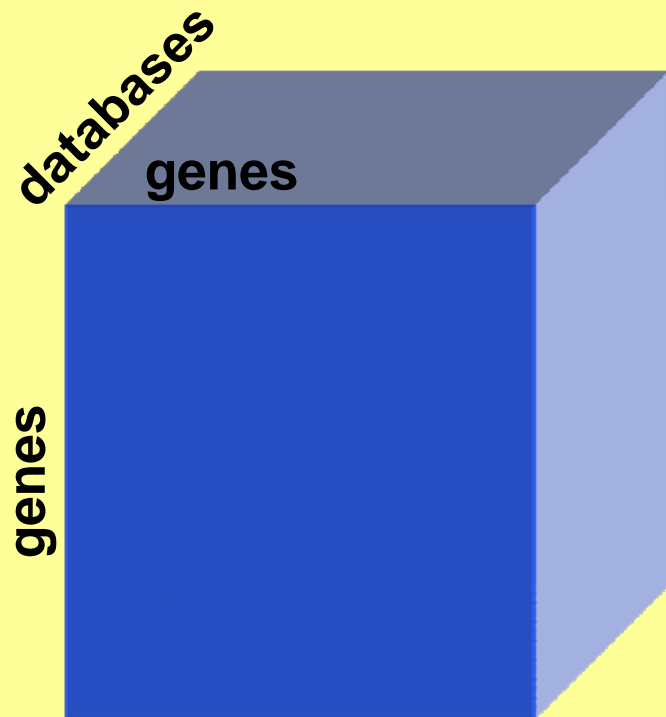
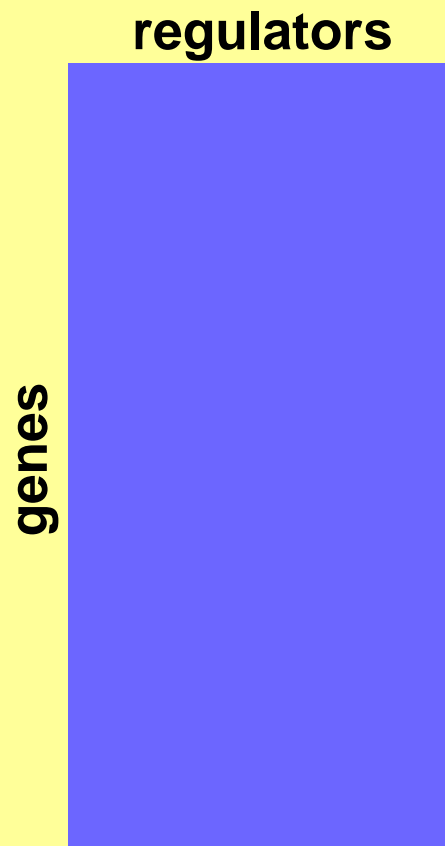
generic structure of the data

data block: mapping $\mathbf{B} : \mathcal{S}_1 \times \mathcal{S}_2 \times \dots \times \mathcal{S}_n \rightarrow Y$

number of sets in Cartesian product of domain = # ways

number of different sets in Cartesian product of domain =
modes

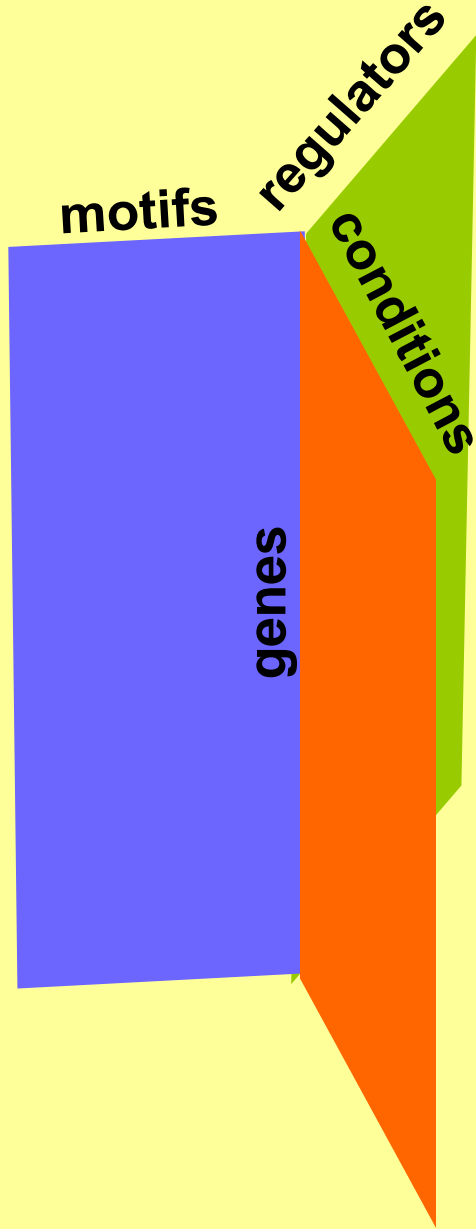
examples

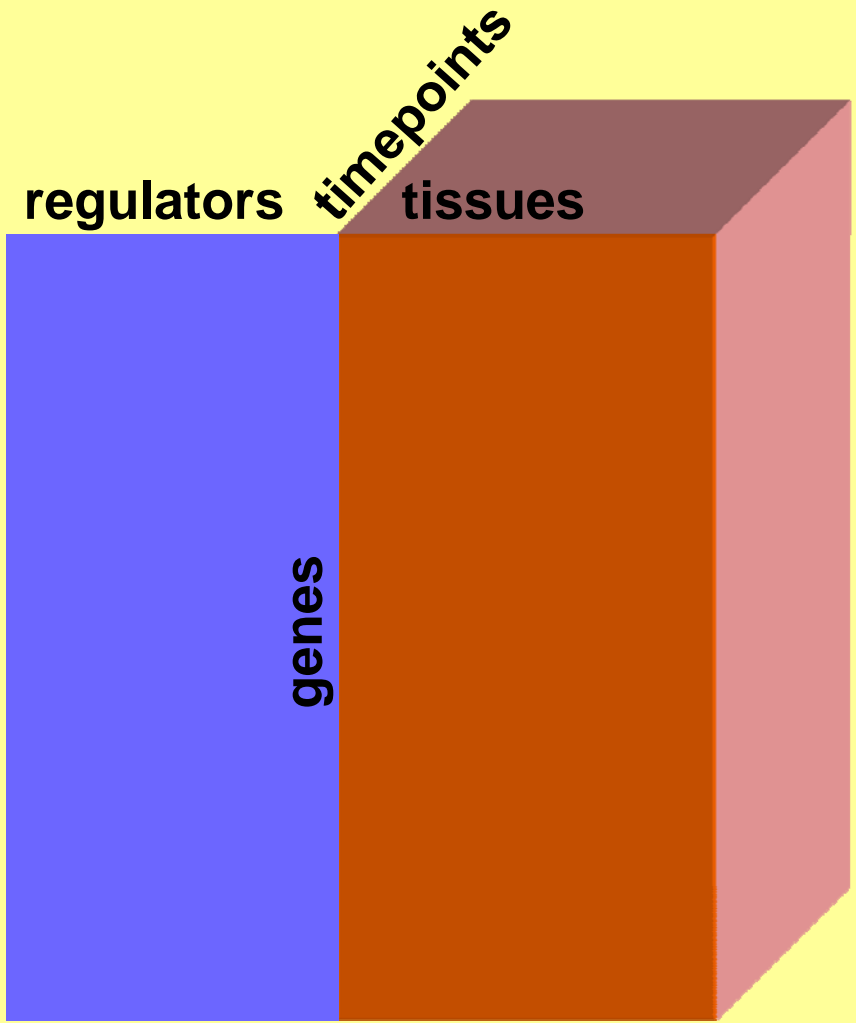


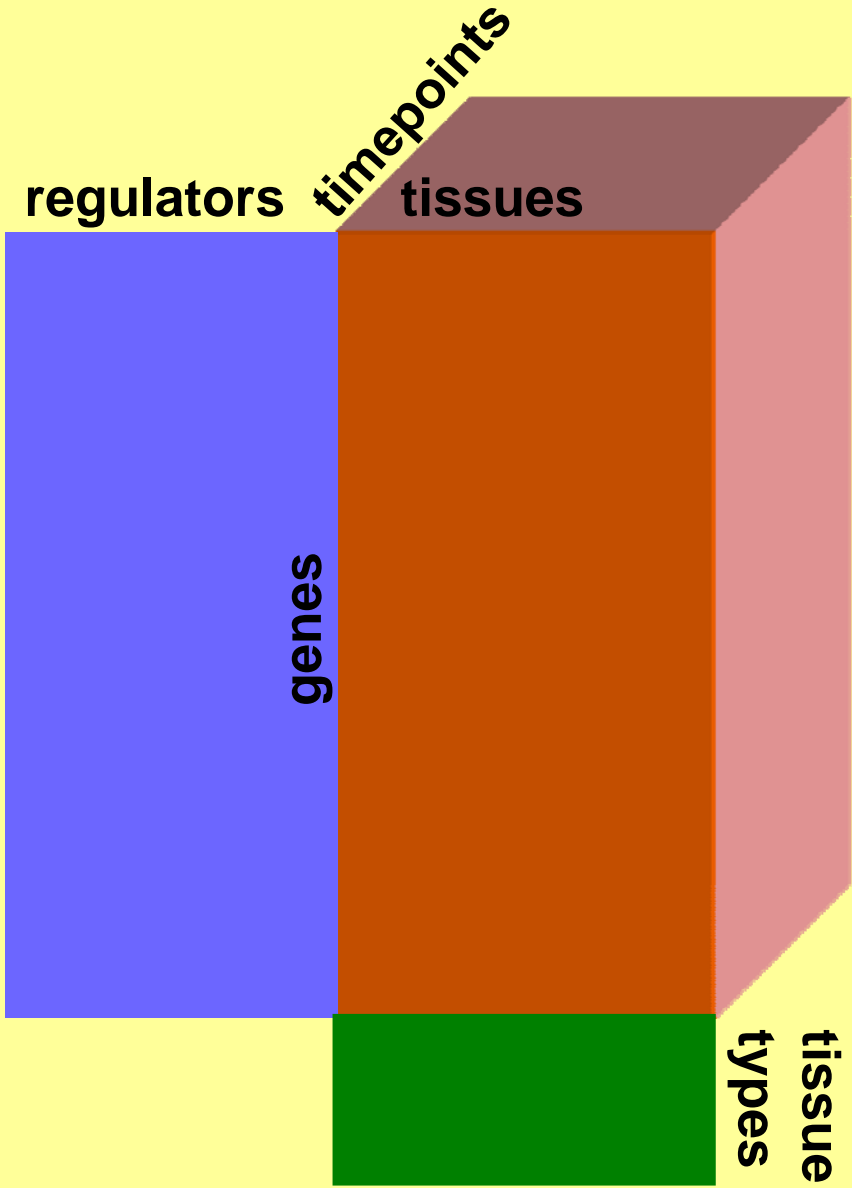
generic structure of the data (ctd)

coupled data: set of data blocks with shared modes

examples





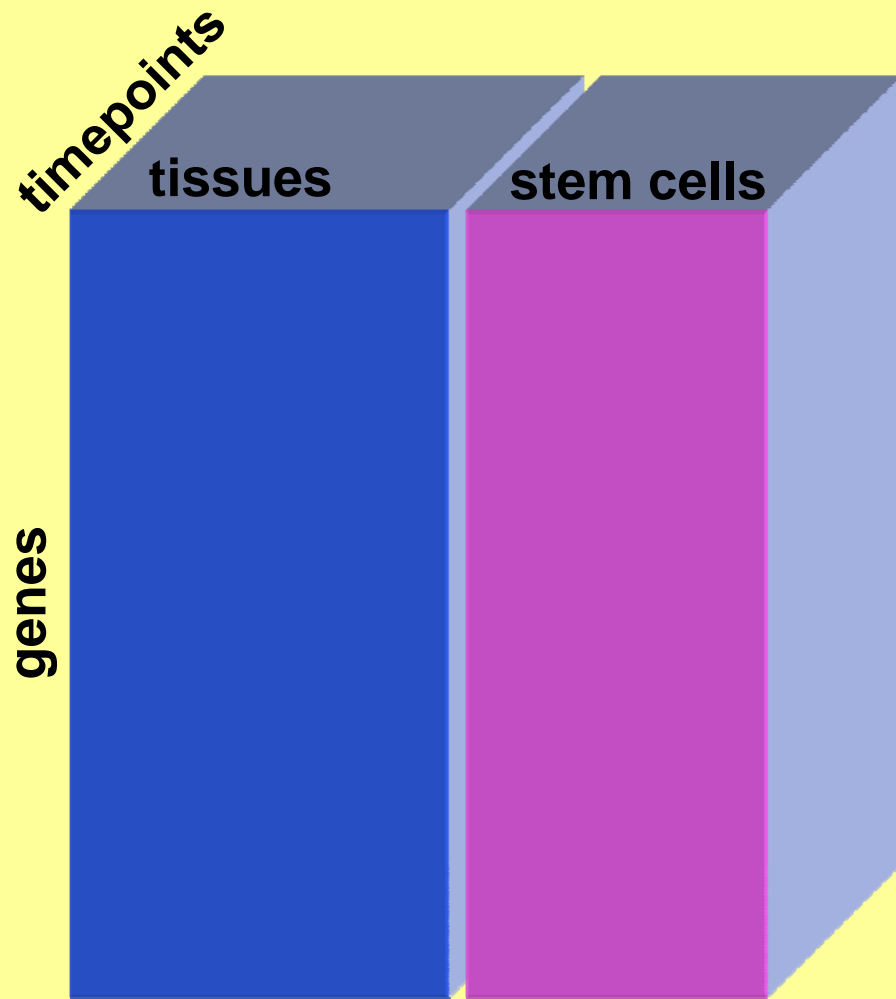


generic structure of the data (ctd)

coupled data: set of data blocks with joint modes

examples

note 1: data blocks may share multiple modes



generic structure of the data (ctd)

coupled data: set of data blocks with joint modes

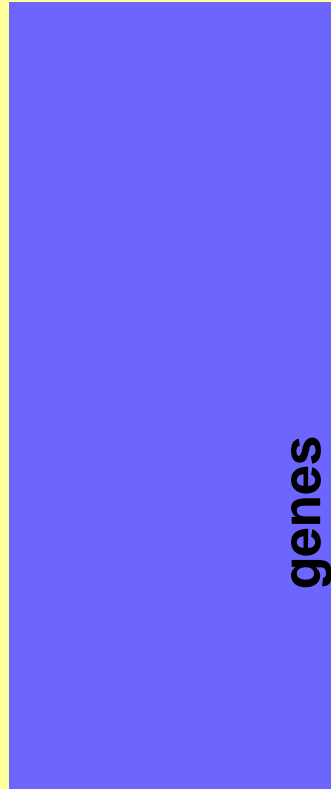
examples

note 1: data blocks may share multiple modes

note 2: possibility of partially shared modes

gene expression data stemming from two different organisms

conditions



conditions



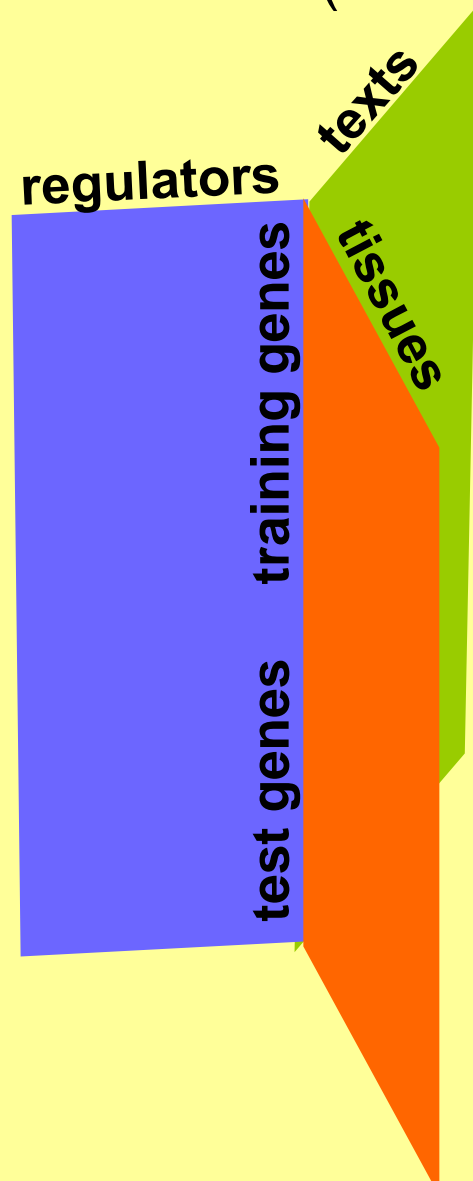
genes

1. Data and problem

- examples of data sets
- generic structure of the data
- examples of problems

example 1

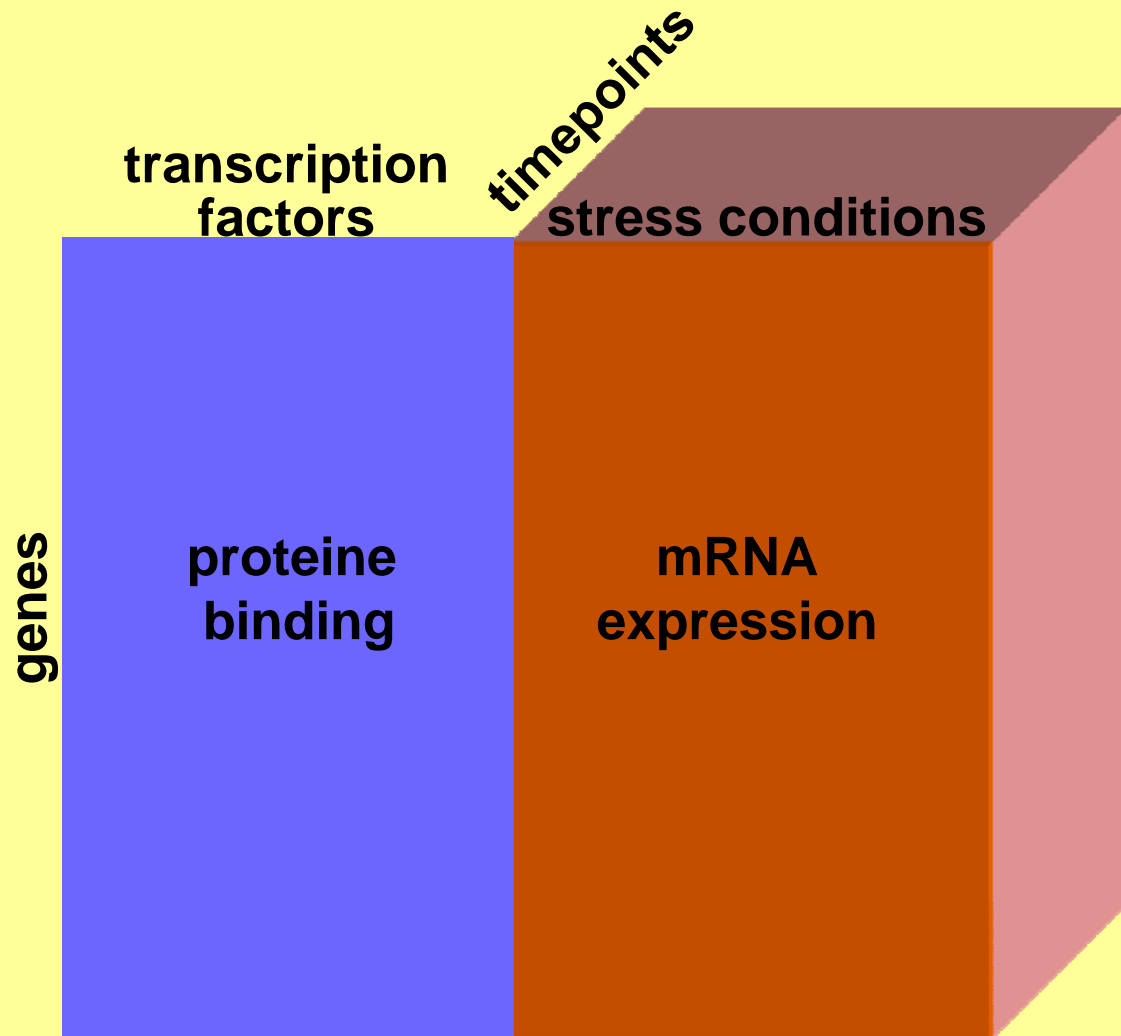
Aerts et al. (2004): gene prioritization



which test genes are most similar to the training genes that are known to be associated with the disease under study?

example 2

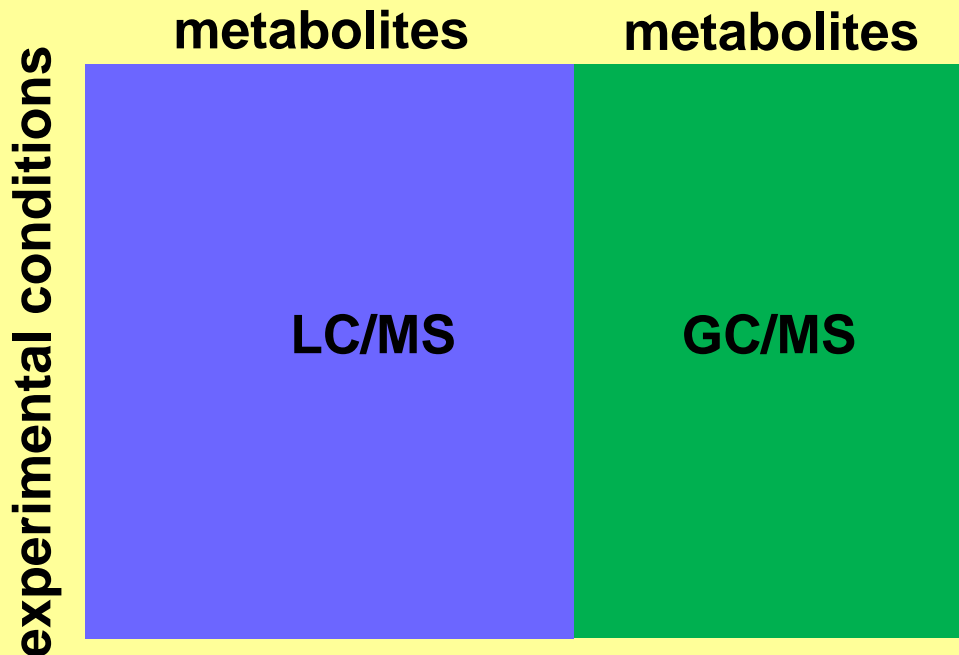
Omberg et al. (2007): yeast cell cycle time course



does transcription binding predict gene expression profiles over time?

example 3

Smilde et al. (2005): comprehensive view on microbial metabolome



which are the common and distinctive aspects of the metabolome that are captured by the two measurement platforms?

1. Data and problem

- examples of data sets
- generic structure of the data
- examples of problems
- **generic aspects of questions**

generic aspects of questions

which information is to be derived from the data?

- which information is to be derived from each data block?
 - full information (data reconstruction)?
 - partial information (e.g., derived similarity or discrimination information; information on dependence/interaction between different data modes)?

- which information is to be derived from the whole of the data blocks?
 - consensus/vote/average?
 - commonalities and differences between blocks?
 - information on linking relation(s) between blocks)?

generic aspects of questions (ctd)

are different data blocks exchangeable in terms of role or not? (e.g., predictor block vs. criterion block)

do different data blocks have different levels of priority or importance?

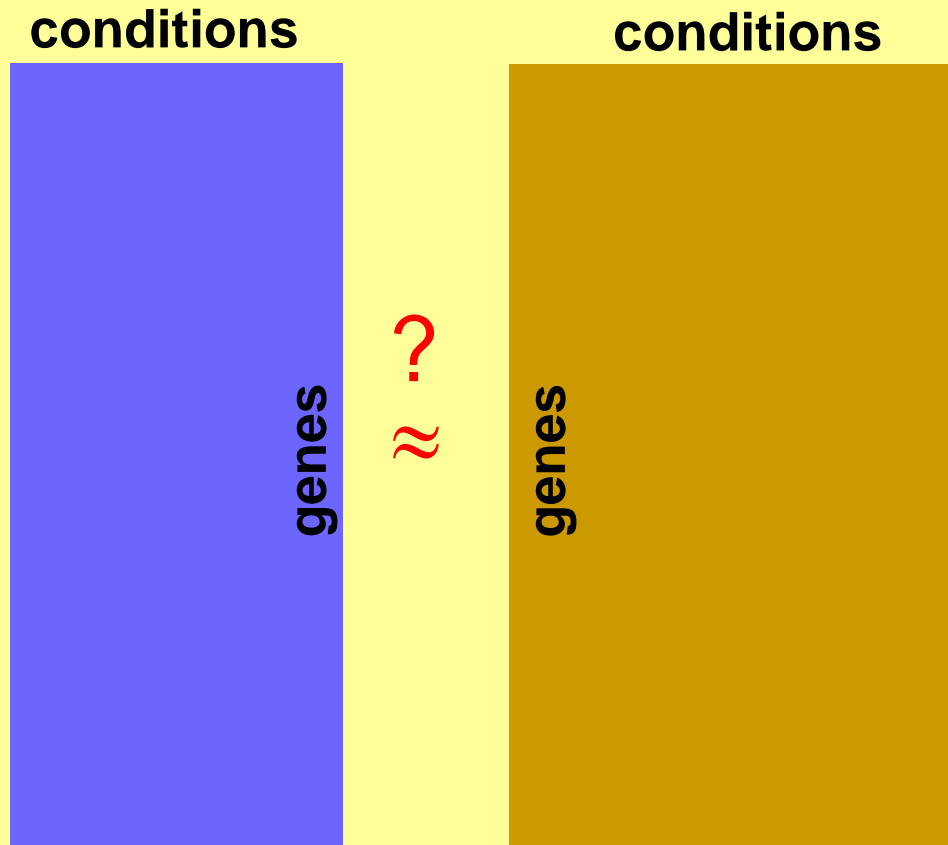
1. Data and problem

- examples of data sets
- generic structure of the data
- examples of problems
- generic aspects of questions
- complicating factors

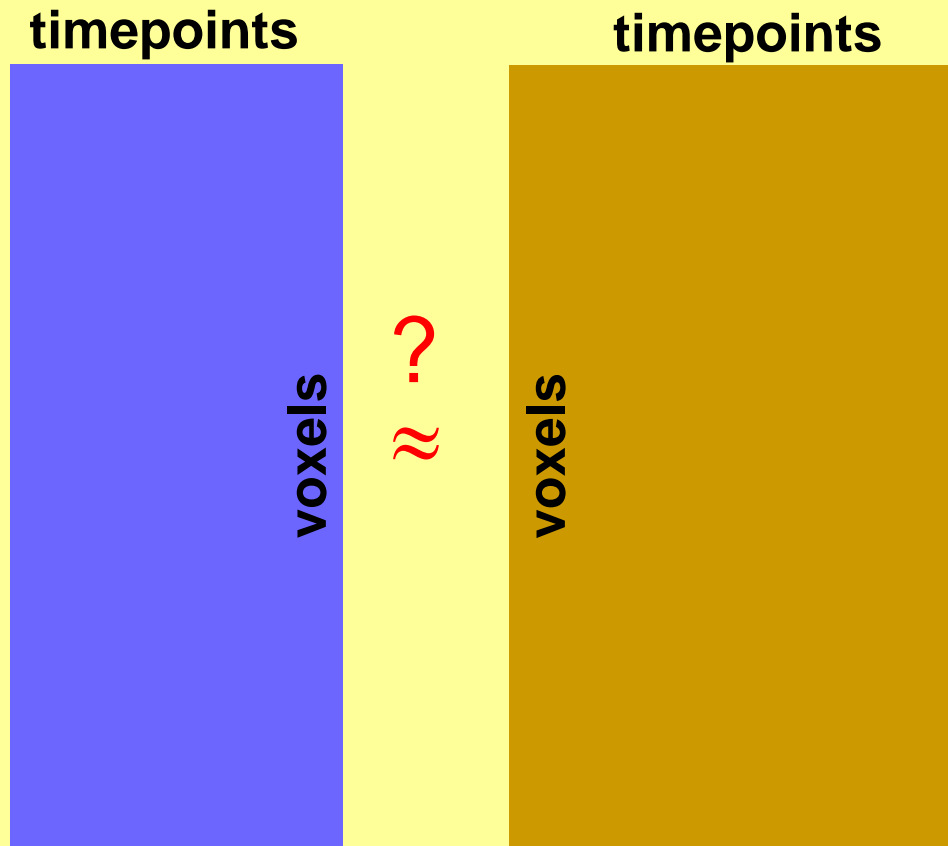
complicating factors

alignment/mapping problem

gene expression data stemming from two different organisms



fMRI data stemming from two different persons



complicating factors

alignment/mapping problem

problem of comparability

comparability/commensurability/conditionality

possibilities: block-conditionality, mode-conditionality
(e.g., gene-conditionality), other

relation to preprocessing

complicating factors

alignment/mapping problem

problem of comparability

problem of heterogeneity in size

complicating factors

alignment/mapping problem

problem of comparability

problem of heterogeneity in size

problem of heterogeneity in noise level

1. Data and problem

- examples of data sets
- generic structure of the data
- examples of problems
- generic aspects of questions
- complicating factors
- **our aims**

our aims

data reconstruction (full information to be derived from each data block)

interest both in commonalities and differences between blocks and in linking relation(s) between blocks)

data blocks exchangeable in terms of role and of priority/ importance

our aims (ctd)

no:

- exclusive focus on mapping/alignment
- meta analysis (high level data fusion)
- regression-type of models
- canonical correlation type of models
- models of derived similarities

yes:

- low level data fusion

Overview of this talk:

1. data

2. model

3. research challenges

2. Model

- preliminary note
- submodel per data block
- linking structure between different submodels
- examples of instantiations

2. Model

- preliminary note
- submodel per data block
- linking structure between different submodels
- examples of instantiations

preliminary note

deterministic heart

optional stochastic periphery

(see, e.g., Leenen et al., 2008, *Psychometrika*)

2. Model

- preliminary note
- **submodel per data block**
- linking structure between different submodels
- examples of instantiations

Submodel per data block

(Van Mechelen & Schepers, 2007, *CSDA*)

assume $(I_1 \times \dots \times I_n \times \dots \times I_N)$ N -way N -mode data block **B**

reduction / quantification of mode n by means of an $I_n \times P_n$ matrix: A^n

gene ₁	3.2	5.2	gene ₁	0	1	0
gene ₂	4.1	-6.7	gene ₂	1	1	0
gene ₃	5.8	3.9	gene ₃	1	0	0
gene ₄	1.0	-2.1	gene ₄	1	0	1
gene ₅	-2.3	8.0	gene ₅	0	0	0
...			...			

Submodel per data block (ctd)

quantification of a mode: special cases

- lowdimensional representation
- partitioning
- nested clustering
- identity matrix

gene ₁	1	0
gene ₂	0	1
gene ₃	1	0
gene ₄	0	1
gene ₅	0	1
...		

gene ₁	1	1	0
gene ₂	1	1	0
gene ₃	1	0	1
gene ₄	1	0	1
gene ₅	1	0	1
...			

Submodel per data block (ctd)

association between quantifications of N data modes:

- $(P_1 \times \dots \times P_n \times \dots \times P_N)$ core array \mathbf{W}
- decomposition rule:

$$\mathbf{B} = f(A^1, \dots, A^N, \mathbf{W}) + \mathbf{E}$$

with within any mode $n = 1, \dots, N$:

$f(A^1, \dots, A^N, \mathbf{W})_{i_1 \dots i_n \dots i_N}$ depends only on the i_n^{th} row of A^n

Submodel per data block (ctd)

special cases of decomposition function f :

1. generalized Cartesian product (Carroll & Chaturvedi, 1995):

$$f(A^1, \dots, A^N, \mathbf{W})_{i_1 \dots i_n \dots i_N} = \sum_{p_1=1}^{P_1} \dots \sum_{p_n=1}^{P_n} \dots \sum_{p_N=1}^{P_N} a_{i_1 p_1}^1 \dots a_{i_n p_n}^n \dots a_{i_N p_N}^N w_{p_1 \dots p_n \dots p_N}$$

special case: Tucker_N models

$$f(A^1, A^2, A^3, \mathbf{W})_{i_1 i_2 i_3} = \sum_{p_1=1}^{P_1} \sum_{p_2=1}^{P_2} \sum_{p_3=1}^{P_3} a_{i_1 p_1}^1 a_{i_2 p_2}^2 a_{i_3 p_3}^3 w_{p_1 p_2 p_3}$$

Submodel per data block (ctd)

special cases of decomposition function f :

1. generalized Cartesian product:

$$f(A^1, \dots, A^N, \mathbf{W})_{i_1 \dots i_n \dots i_N} = \sum_{p_1=1}^{P_1} \dots \sum_{p_n=1}^{P_n} \dots \sum_{p_N=1}^{P_N} a_{i_1 p_1}^1 \dots a_{i_n p_n}^n \dots a_{i_N p_N}^N w_{p_1 \dots p_n \dots p_N}$$

special case: several forms of biclustering

(see also Van Mechelen et al., 2004, *SMMR*)

$$f(A^1, A^2, \mathbf{W})_{i_1 i_2} = \sum_{p_1=1}^{P_1} \sum_{p_2=1}^{P_2} a_{i_1 p_1}^1 a_{i_2 p_2}^2 w_{p_1 p_2}$$

$$f(A^1, A^2, \mathbf{W})_{i_1 i_2} = \sum_{p_1=1}^{P_1} \sum_{p_2=1}^{P_2} a_{i_1 p_1}^1 a_{i_2 p_1}^2 w_{p_1 p_2}$$

		$A_{\bullet 1}^1$	$A_{\bullet 2}^1$								V_1	V_2	V_3	V_4	V_5	V_6	V_7		
A	O_1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	O_1	
	O_2	1	0	0	2	2	2	0	0	0	0	0	0	0	0	0	0	O_2	
	O_3	1	0	0	2	2	2	0	0	0	0	0	0	0	0	0	0	O_3	
	O_4	1	1	0	2	2	5	3	3	0	0	0	0	0	0	0	0	O_4	
	O_5	0	1	0	0	0	0	3	3	3	0	0	0	0	0	0	0	O_5	
	O_6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	O_6	
W	$A_{\bullet 1}^2$	2	0	0	1	1	1	0	0	0								$A_{\bullet 1}^2$	B
	$A_{\bullet 2}^2$	0	3	0	0	0	1	1	1	0								$A_{\bullet 2}^2$	
	$A_{\bullet 1}^1$	$A_{\bullet 2}^1$	V_1	V_2	V_3	V_4	V_5	V_6	V_7										

Submodel per data block (ctd)

special cases of decomposition function f :

2. generalized Cartesian product in (Max, \times) structure:

$$f(A^1, \dots, A^N, \mathbf{W})_{i_1 \dots i_n \dots i_N} = \text{Max}_{p_1=1}^{P_1} \dots \text{Max}_{p_n=1}^{P_n} \dots \text{Max}_{p_N=1}^{P_N} a_{i_1 p_1}^1 \dots a_{i_n p_n}^n \dots a_{i_N p_N}^N w_{p_1 \dots p_n \dots p_N}$$

(see, e.g., De Schutter et al., 2006; Van Mechelen et al., 2007, *Psychometrika*)

Submodel per data block (ctd)

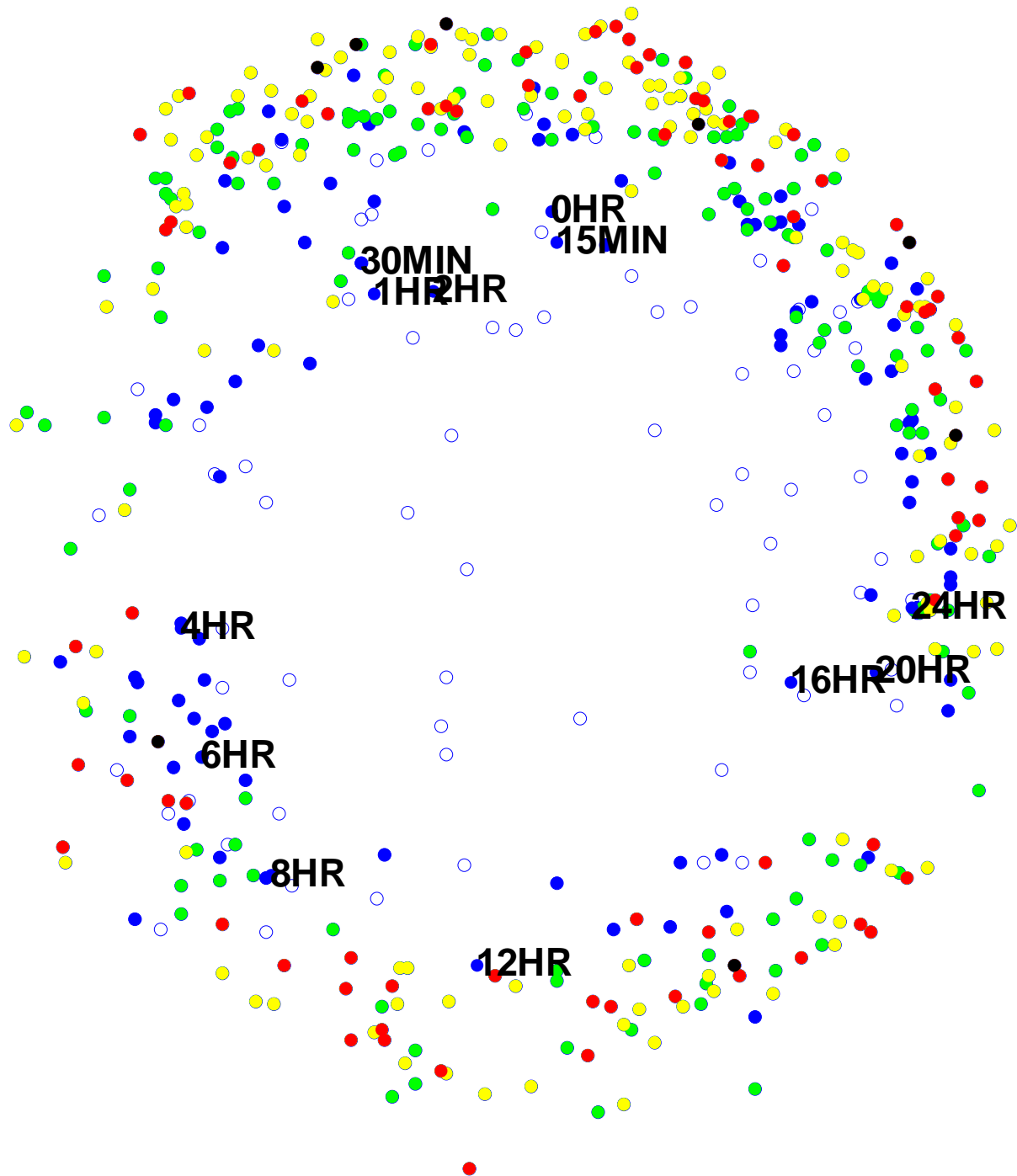
special cases of decomposition function f :

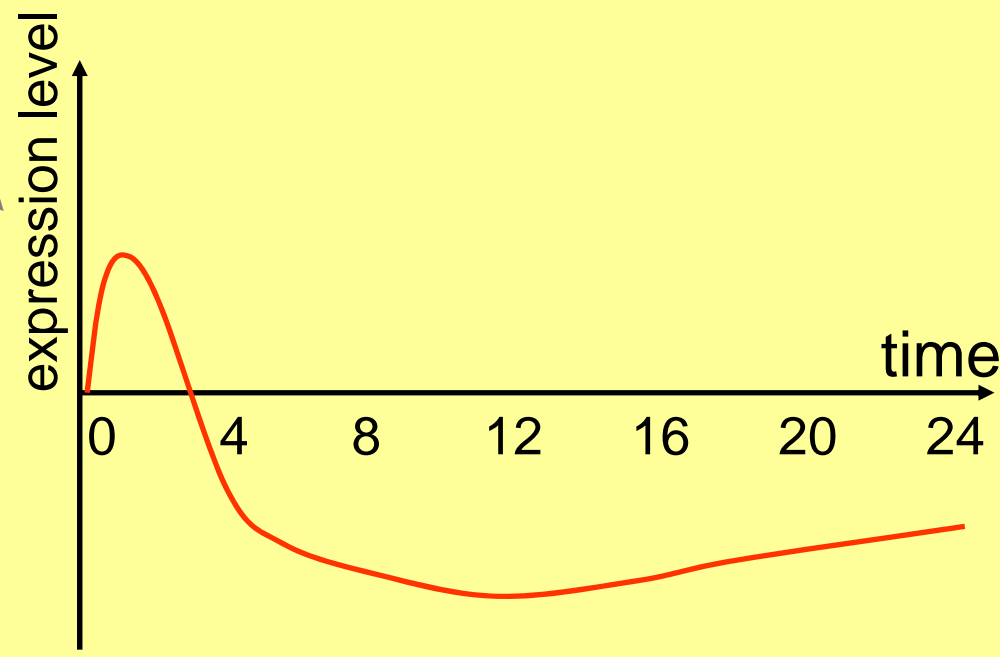
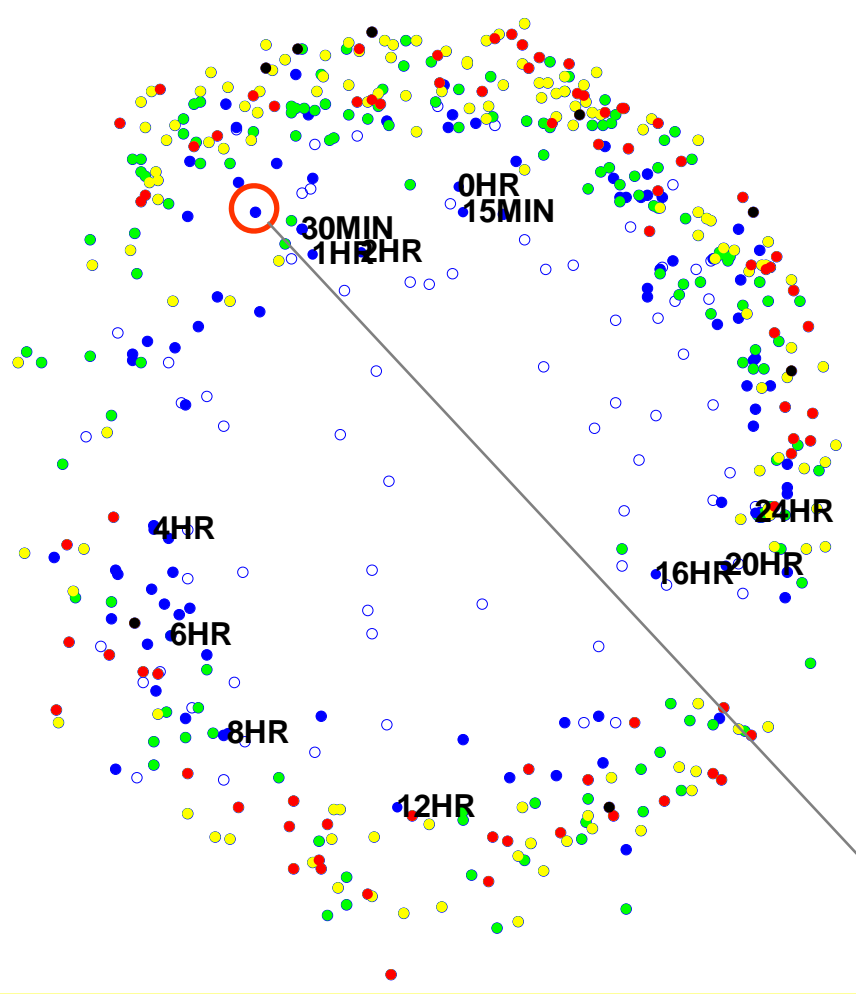
3. weighted Minkowski distance; e.g.,

$$f(A^1, A^2, A^3, \mathbf{W})_{i_1 i_2 i_3} = \left[\sum_{\rho_1=1}^{P_1} \sum_{\rho_2=1}^{P_2} \sum_{\rho_3=1}^{P_3} a_{i_3 \rho_3}^3 w_{\rho_1 \rho_2 \rho_3} \left| a_{i_1 \rho_1}^1 - a_{i_2 \rho_2}^2 \right|^k \right]^{\frac{1}{k}}$$

$$f(A^1, A^2)_{i_1 i_2} = \left[\sum_{\rho_1=1}^{P_1} \left(a_{i_1 \rho_1}^1 - a_{i_2 \rho_1}^2 \right)^2 \right]^{\frac{1}{2}}$$

(see also Van Deun et al., 2007, *BMC Bioi*)





2. Model

- preliminary note
- submodel per data block
- **linking structure between different submodels**
- examples of instantiations

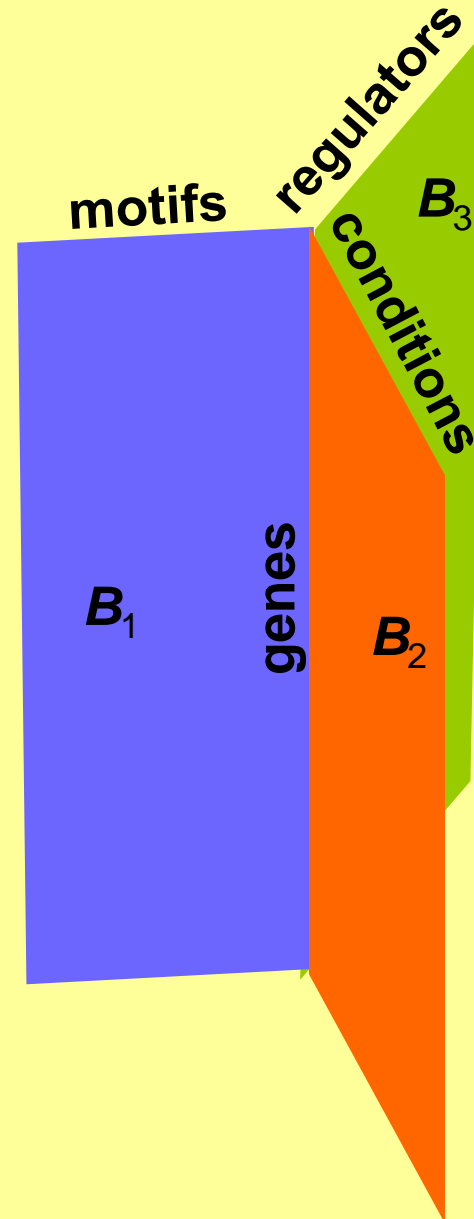
linking structure between different submodels

identity link

$$B_1 = f_1(A^1, A_1^2, W_1) + E_1$$

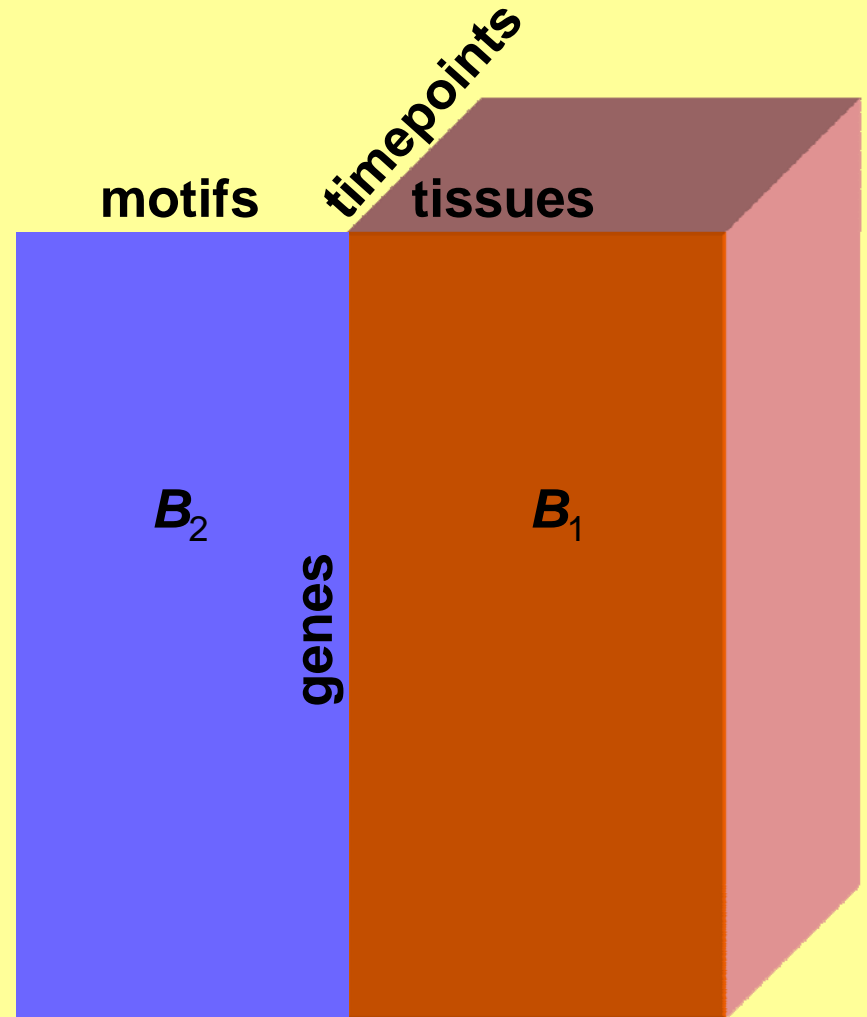
$$B_2 = f_2(A^1, A_2^2, W_2) + E_2$$

$$B_3 = f_3(A^1, A_3^2, W_3) + E_3$$



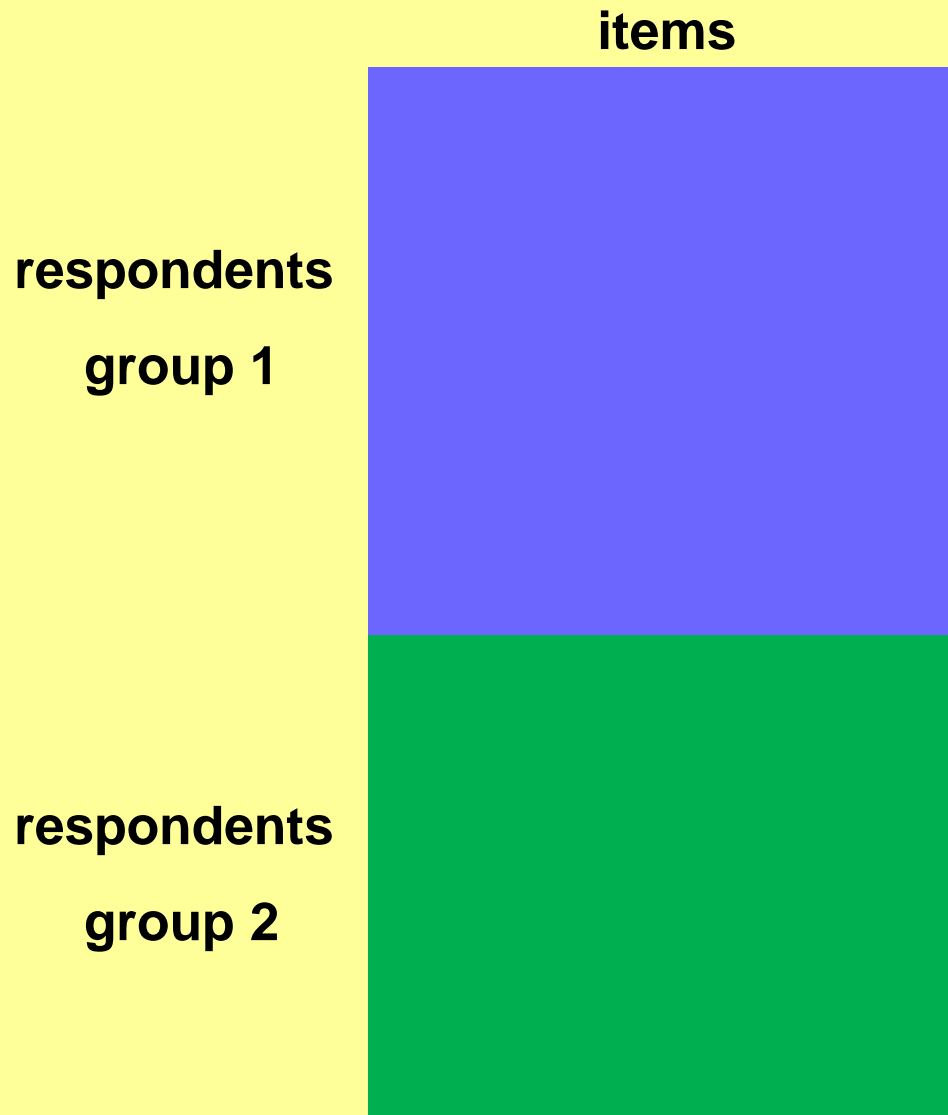
$$B_1 = f_1(A^1, A_1^2, A_1^3, W_1) + E_1$$

$$B_2 = f_2(A^1, A_2^2, W_2) + E_2$$



linking structure between different submodels

identity link (possibly with exceptions ...)



problem of (configural) measurement invariance (Meredith, 1993)
(psychometrics: bias, differential item functioning)

linking structure between different submodels

identity link (possibly with exceptions ...)

in case of binary quantifications:

nestedness (in particular in case of partitions: one partition is refinement of the other – asymmetric relation!)

logical relation (partition classes implied by first quantification are nested within partition classes implied by second quantification)

in case of real-valued quantifications: spaces induced by two quantification are in a space-subspace relation

2. Model

- preliminary note
- submodel per data block
- linking structure between different submodels
- **examples of instantiations**

examples of instantiations

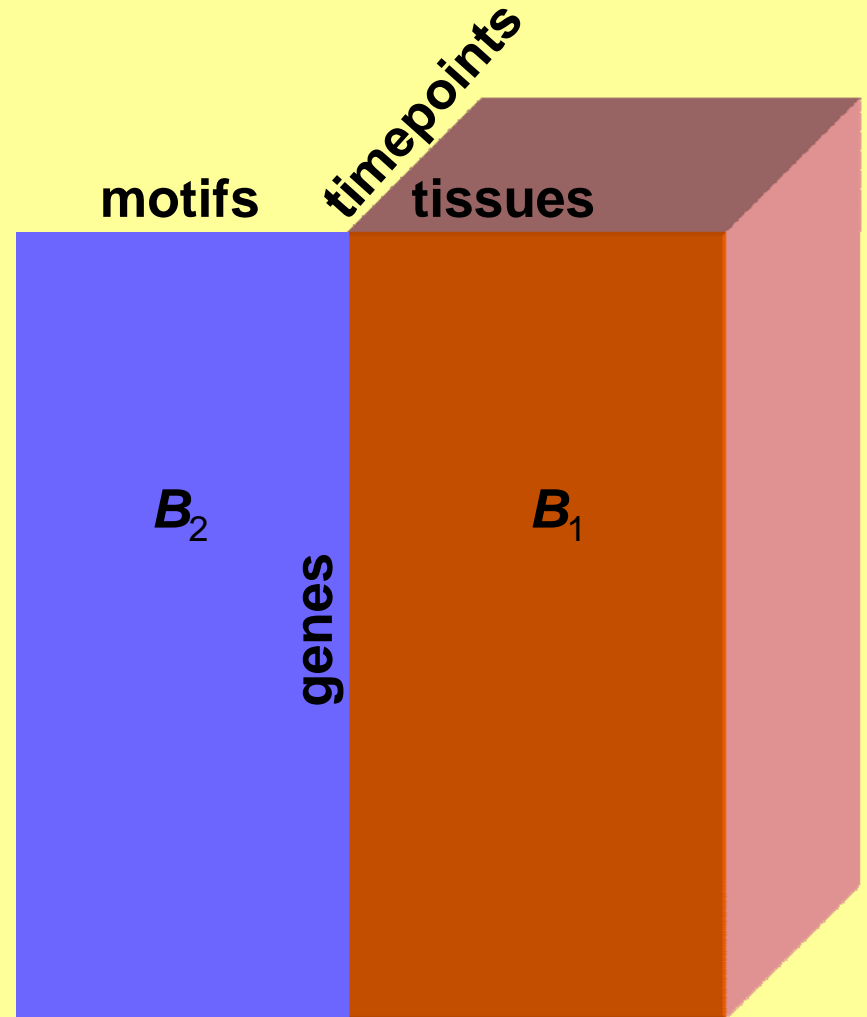
simultaneous component models (see talk Van Deun)

linked-mode PARAFAC-PCA (Wilderjans et al., in press, *BRM*)

CHIC (Wilderjans et al., 2008, *Psychometrika*)

$$B_1 = A^1(A_1^2 \odot A_1^3)^T + E_1$$

$$B_2 = A^1(A_2^2)^T + E_2$$



Overview of this talk:

1. data

2. model

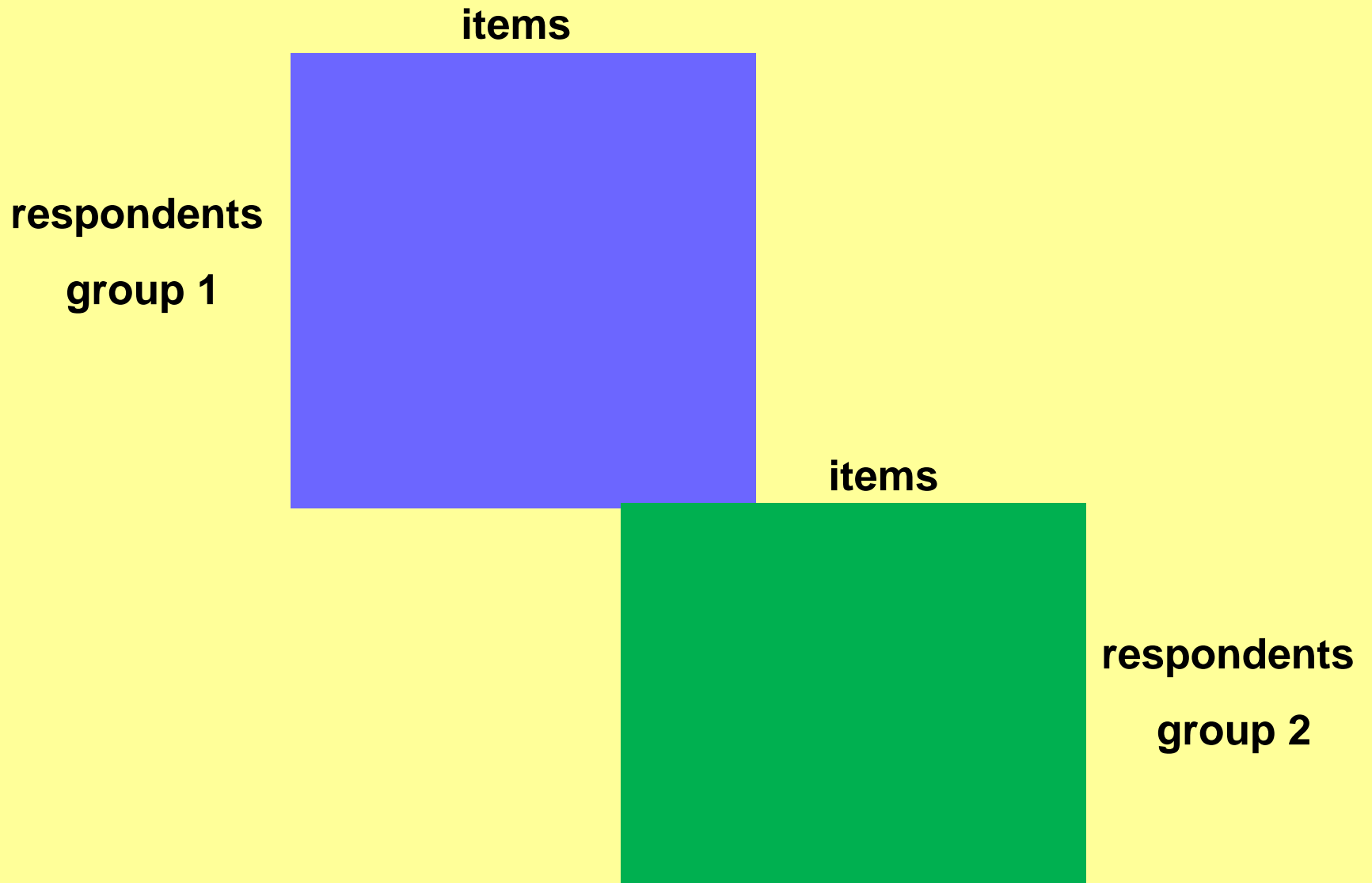
3. research challenges

3. Research challenges

- design data collection
- model
- objective function
- algorithmics
- data-analytic issues

3. Research challenges

- design data collection
- model
- objective function
- algorithmics
- data-analytic issues



psychometrics: problem of test equating (choice of anchoring items)

3. Research challenges

- design data collection
- **model**
- objective function
- algorithmics
- data-analytic issues

model

specify how specific existing models are subsumed by generic model (naming, mathematical characteristics, model interrelations, ...)

development novel models that meet substantive (psychological, biological, etc.) needs

3. Research challenges

- design data collection
- model
- **objective function**
- algorithmics
- data-analytic issues

objective function

determine optimal weights for different data blocks / elements / data entries to improve quality of inferences (e.g., Wilderjans et al., 2009, *CSDA*)

development of novel objective functions that put higher emphasis on commonality

investigation of quality of objective functions through simulation studies

3. Research challenges

- design data collection
- model
- objective function
- **algorithmics**
- data-analytic issues

algorithmics

construction suitable algorithms (?ALS, ?sequential procedures)

challenges: avoiding local optima; computational efficiency

evaluation algorithmic performance in simulation studies
(see, e.g., Schepers et al., 2006, *CSDA*)

3. Research challenges

- design data collection
- model
- objective function
- algorithmics
- data-analytic issues

data-analytic issues

model selection

- splitting/decoupling modes
- nature of submodels (which modes are to be reduced, nature of reduction, extent of reduction; decomposition function; choice of possible constraints; stochastic assumptions)
- linking structure (including measurement invariance issue)

uniqueness/identifiability

representation of uncertainty (see, e.g., Ceulemans & Kiers, in press, *BJMSP*; Timmerman et al., 2009, *BJMSP*)

interpretation of shared quantifications (including decomposition into common and distinctive parts)

References

see also <http://ppw.kuleuven.be/okp/>

- Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.-C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P., & Moreau, Y. (2006). Gene prioritization through genomic data fusion. *Nature Biotechnology*, *24*, 537-544.
- Carroll, J.D., & Chaturvedi, A. (1995). A general approach to clustering and multidimensional scaling of two-way, three-way, or higher-way data. In R.D. Luce, M. D'Zmura, D.D. Hoffman, G.J. Iverson, & A.K. Romney (Eds.), *Geometric representations of perceptual phenomena* (pp. 295–318). Mahwah: Erlbaum.
- Ceulemans, E., & Kiers, H.A.L. (in press). Discriminating between strong and weak structures in three-mode principal component analysis. *British Journal of Mathematical & Statistical Psychology*.

References (ctd)

- Clish, C., Davidov, E., Oresic, M., Plasterer, T.N., Lavine, G., Londo, T., Meys, M., Snell, P., Stochaj, W., Adourian, A., Zhang, X., Morel, N., Neumann, E., Verheij, E., Vogels, J.T.W.E., Havekes, L.M., Afeyan, N., Regnier, F., Van Der Greef, J., & Naylor, S. (2004). Integrative biological analysis of the APOE*3-Leiden transgenic mouse. *Omics: A Journal of Integrative Biology*, 8,3-13.
- Curran, P.J., & Hussong, A.M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, 14, 81-100.
- De Schutter, B., Schepers, J., & Van Mechelen, I. (2006). On algorithms for a binary-real (max,x) matrix approximation problem. In *Proceedings of the 45th IEEE Conference on Decision & Control* (pp. 5168-5173). San Diego: IEEE Control System Society.

References (ctd)

- Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, D. F., & Meulders, M. (2005). Multiple imputation for model checking: Completed-data plots with missing and latent data. *Biometrics*, *61*, 74-85.
- Leenen, I., Van Mechelen, I., Gelman, A., & De Knop, S. (2008). Bayesian hierarchical classes analysis. *Psychometrika*, *73*, 39-64.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525-543.
- Schepers, J., Ceulemans, E., & Van Mechelen, I. (2008). Selecting among multi-mode partitioning models of different complexities: A comparison of four model selection criteria. *Journal of Classification*, *25*, 67-85.
- Schepers, J., Van Mechelen, I., & Ceulemans, E. (2006). Three-mode partitioning. *Computational Statistics & Data Analysis*, *51*, 1623-1642.

References (ctd)

- Omberg, L., Golub, G.H., & Alter, O. (2007). A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. *Proceedings of the National Academy of Sciences USA*, 104, 18371-18376.
- Smilde, A.K. van der Werf, M.J., Bijlsma, S., van der Werff-van der Vat, B.J.C., & Jellema, R.H. (2005). Fusion of mass spectrometry-based metabolomics data. *Analytical Chemistry*, 77, 6729-6736.
- Timmerman, M.E., Kiers, H.A.L., Smilde, A.K., Ceulemans, E., & Stouten, J. (2009). Bootstrap confidence intervals in multilevel simultaneous component analysis. *British Journal of Mathematical & Statistical Psychology*, 62, 299-318.
- Van Deun, K., Marchal, K., Heiser, W., Engelen, K., & Van Mechelen, I. (2007). Joint mapping of genes and conditions via multidimensional unfolding analysis. *BMC Bioinformatics*, 8, 181.
- Van Mechelen, I., Bock, H.-H., & De Boeck, P. (2004). Two-mode clustering methods: A structural overview. *Statistical Methods in Medical Research*, 13, 363-394.

References (ctd)

- Van Mechelen, I., Lombardi, L., & Ceulemans, E. (2007). Hierarchical classes modeling of rating data. *Psychometrika*, *72*, 475-488.
- Van Mechelen, I., & Scheepers, J. (2007). A unifying model involving a categorical and/or dimensional reduction for multimode data. *Computational Statistics & Data Analysis*, *52*, 537-549.
- Wilderjans, T.F., Ceulemans, E., Kiers, H.A.L., & Meers, K. (in press). The LMPCA program: A graphical user interface for fitting the Linked-Mode PARAFAC-PCA model to coupled real-valued data. *Behavior Research Methods*.
- Wilderjans, T.F., Ceulemans, E., & Van Mechelen, I. (2008). The CHIC model: A global model for coupled binary data. *Psychometrika*, *73*, 729-751.
- Wilderjans, T.F., Ceulemans, E., & Van Mechelen, I. (2009). Simultaneous analysis of coupled data blocks differing in size: A comparison of two weighting schemes. *Computational Statistics and Data Analysis*, *53*, 1086-1098.