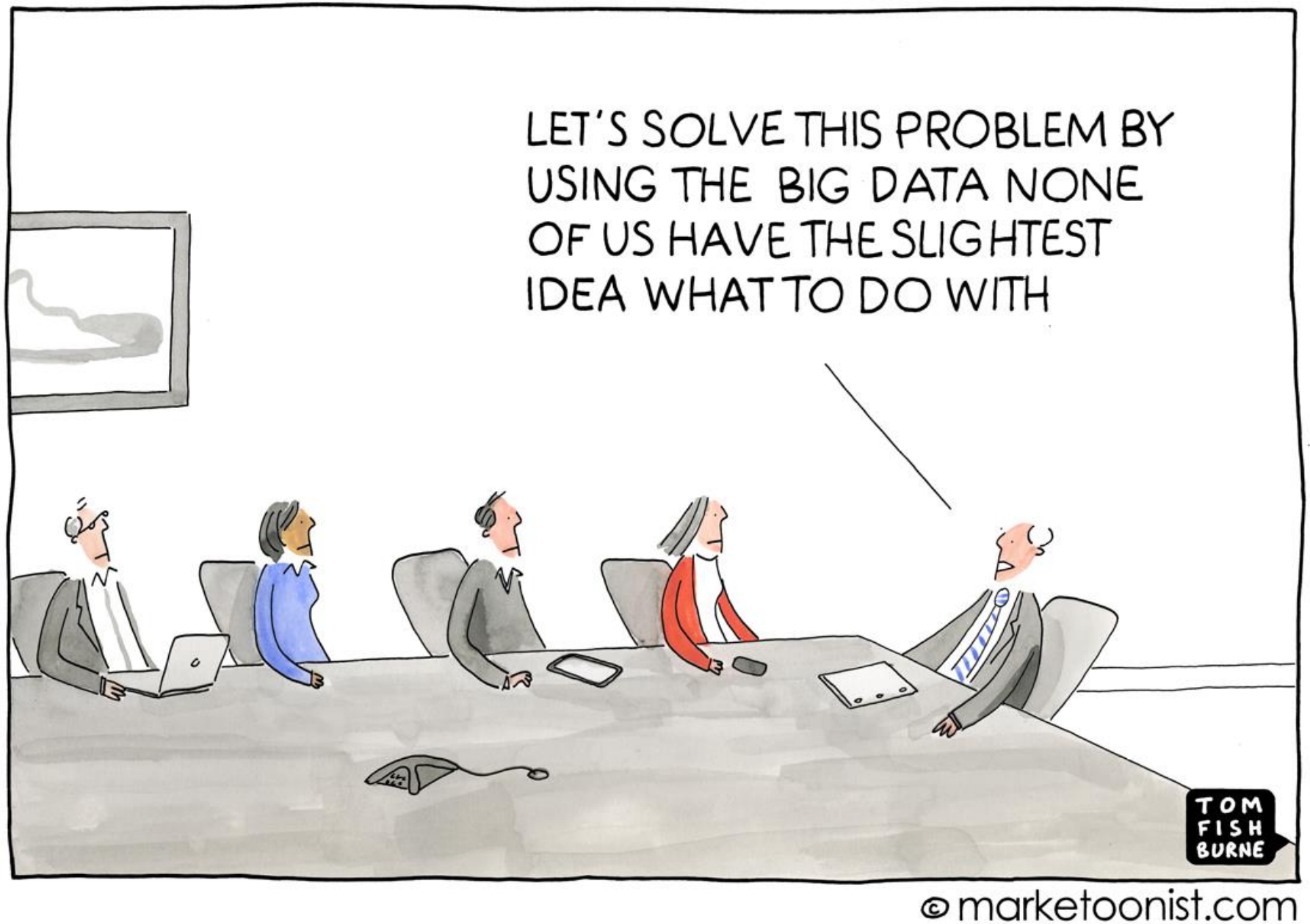
Big Data in Psychology:

Statistical methods for linked high-dimensional with traditional data

Katrijn Van Deun, Tilburg University


LET'S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH
TOM FISH BURNE
© marketoonist.com

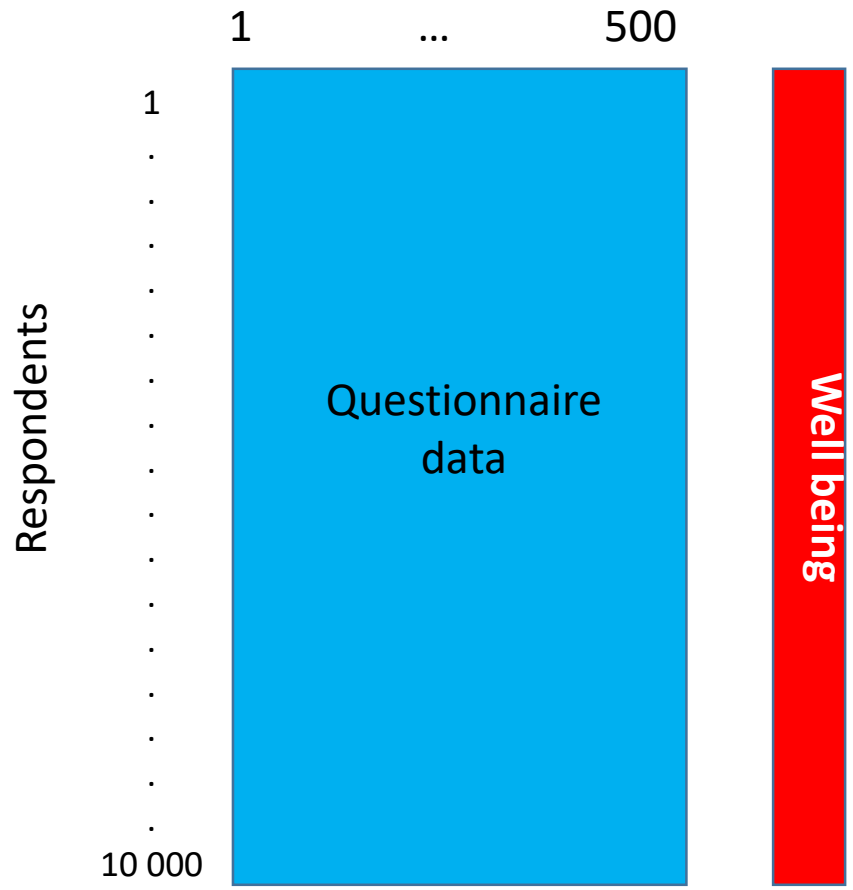
Outline

- Background: Psychological research in a time of Big Data
- Challenges for data analysis
- Part 1: Exploration
- Part 2: Prediction
- Illustrations
  - Exploration: 500 families
  - Prediction: Systems vaccinology
- Playtime: RegularizedSCA

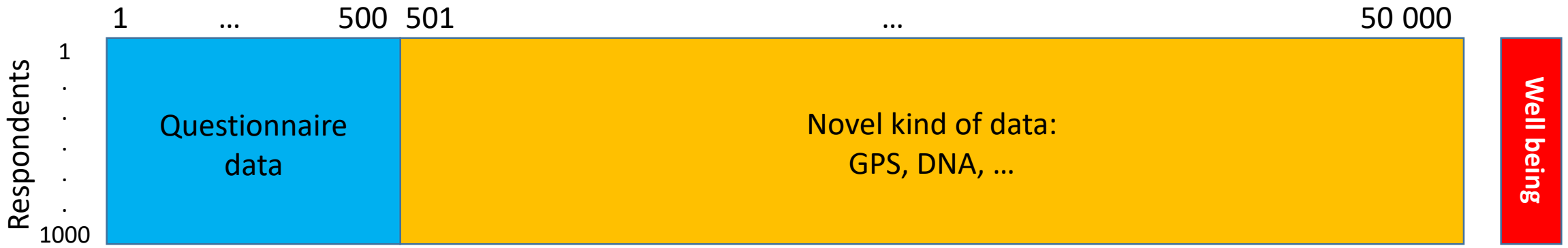
Background:
Psychological research in a time of Big Data

Big Data in the (Behavioral) Sciences

- Everything is measured
  - What we think and do: Social media, Web browsing behavior
  - Where we are and with whom: GPS tracking, cameras
  - At a very detailed level: Experience sampling, neuron, DNA
- Data are shared
  - Open data: in science, governments (open government data)
- Data are linked
  - Government
  - Science: multi-disciplinary
  - Linked Data web-architecture



- Traditional data: Health & Retirement Study



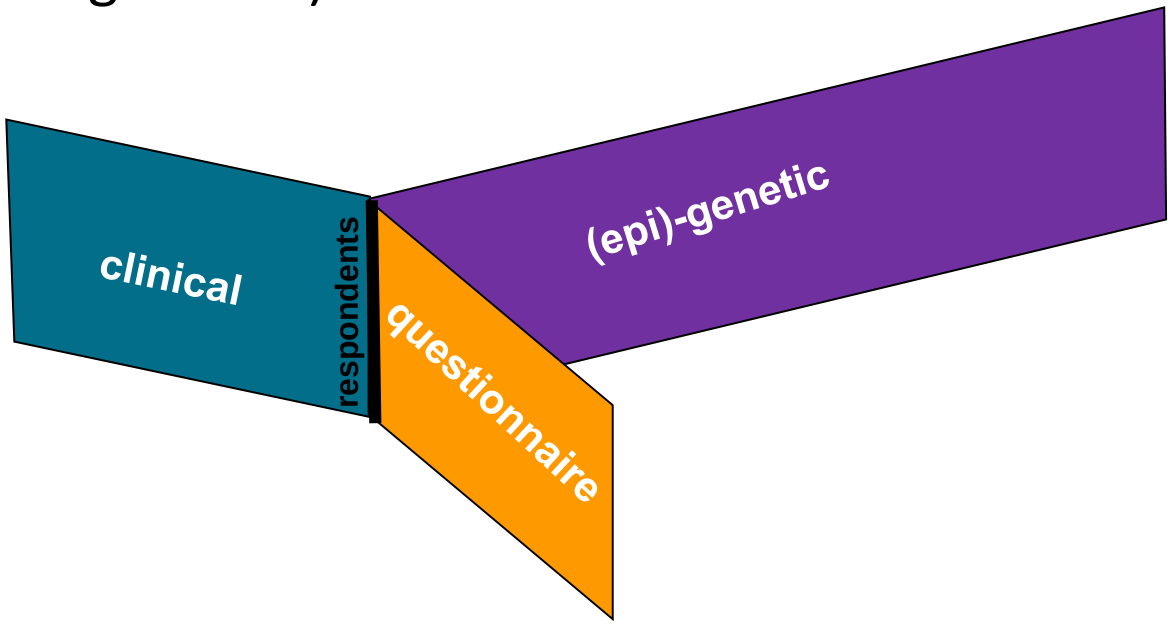
- Health & Retirement Study: traditional + novel type of data

  - Multiple sources, **heterogeneous** in nature
  - **High-dimensional** (*p*>>*n*)

  } Big Data

- Illustration II: ALSPAC household panel data

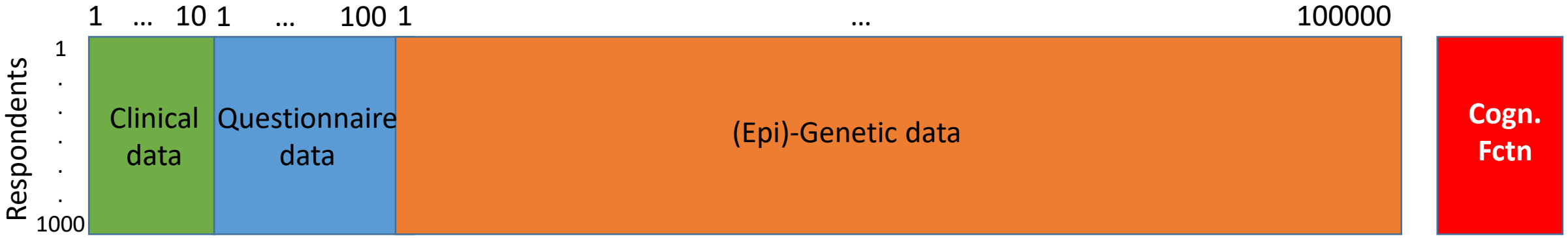
  - Multiple sources / multi-block (heterogeneous)
  - High-dimensional
  - *(no outcome)*



- Illustration III: ADNI data

  - Multi-block data

  - *(Outcome)*



The Promises of linked Traditional with novel data

⇒ Extremely information-rich data

- 1) Gives insight in the interplay between multiple factors / system or multi-disciplinary point of view (Exploratory; hypothesis generating)

Eg. gene-environment interactions: **Link** susceptibility genes with protective/risk-provoking environmental conditions

- 2) Adds context, detail => deeper understanding + more accurate prediction

Eg. Similar income, social network, health but difference in well-being?

However, statistical tools fall short … how to find the **linked variables**?

Non-theory driven nature of Big Data

- Novel / Big data:
  - NOD: Naturally Occurring Data
  - High-throughput screening (eg, GWAS)

    => untargeted

  Furthermore, little theory available on these novel sorts of data.

- Relevant variables? Introduces a **variable selection problem**.

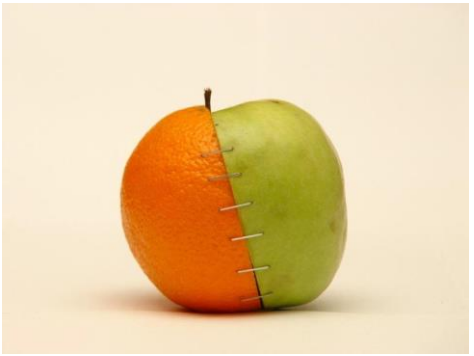
The hidden link

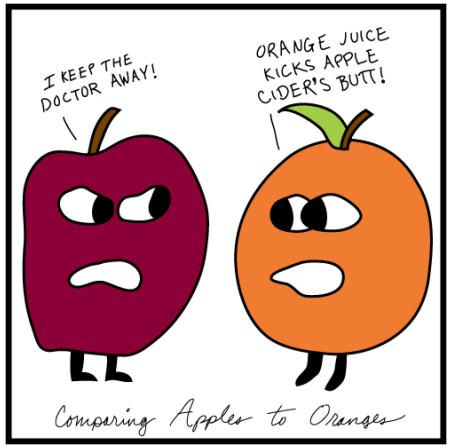
- Traditional and novel data are very different:
  - Different sizes of data blocks, different noise/measurement levels
  - Stronger correlations of variables within blocks than between blocks
    - Traditional: Eg, response tendencies, general psychological processes
    - Novel: Eg, (general) biological processes

    => Link variables / shared mechanisms are hidden while block-specific mechanisms dominate

The hidden link

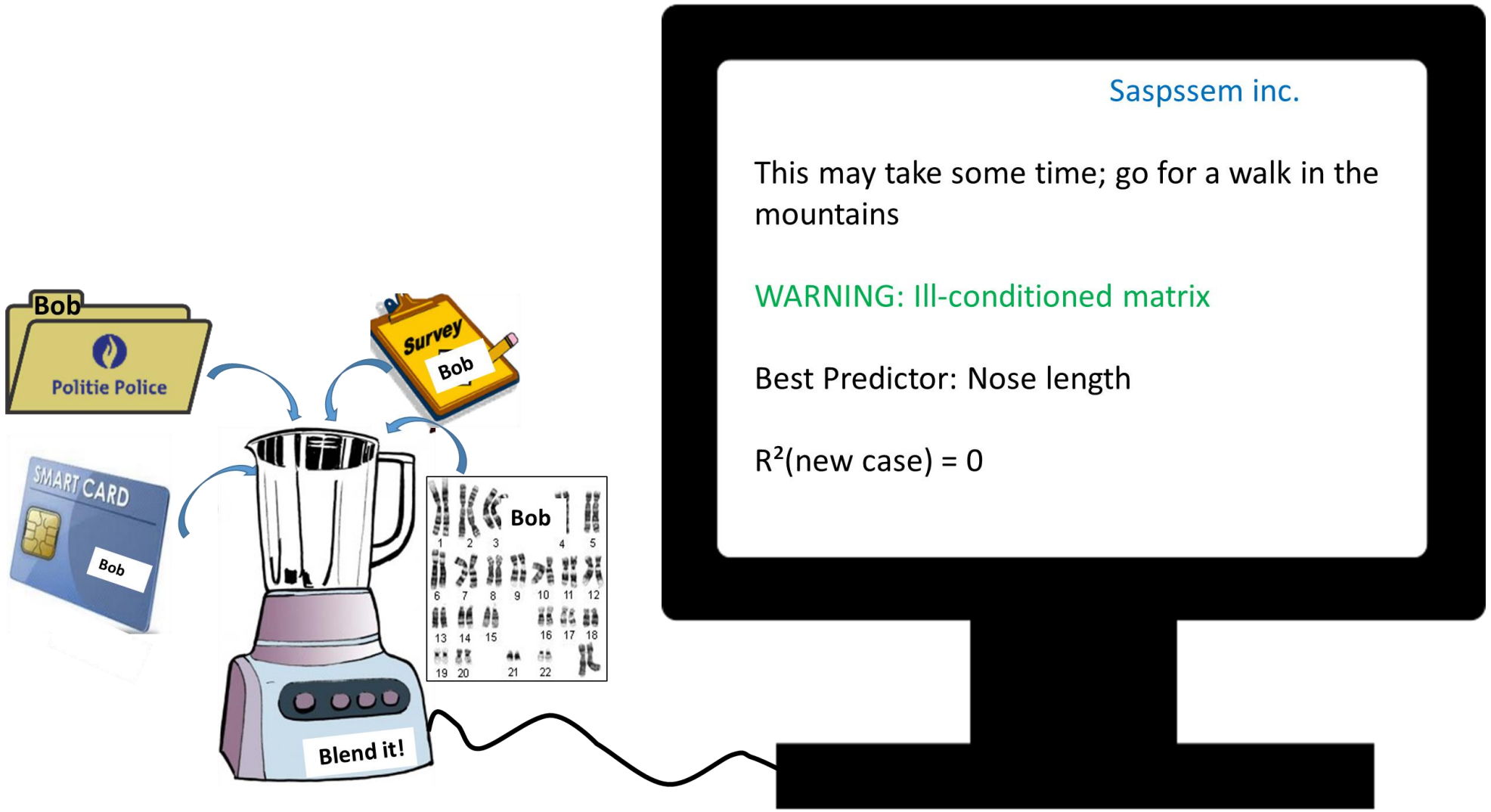
- Popular but bad ideas for analysing linked (traditional with) big data:
  - 1) glueing the data together and analysing them as one block





  - 2) analysing each block separately and comparing the results

Furthermore …

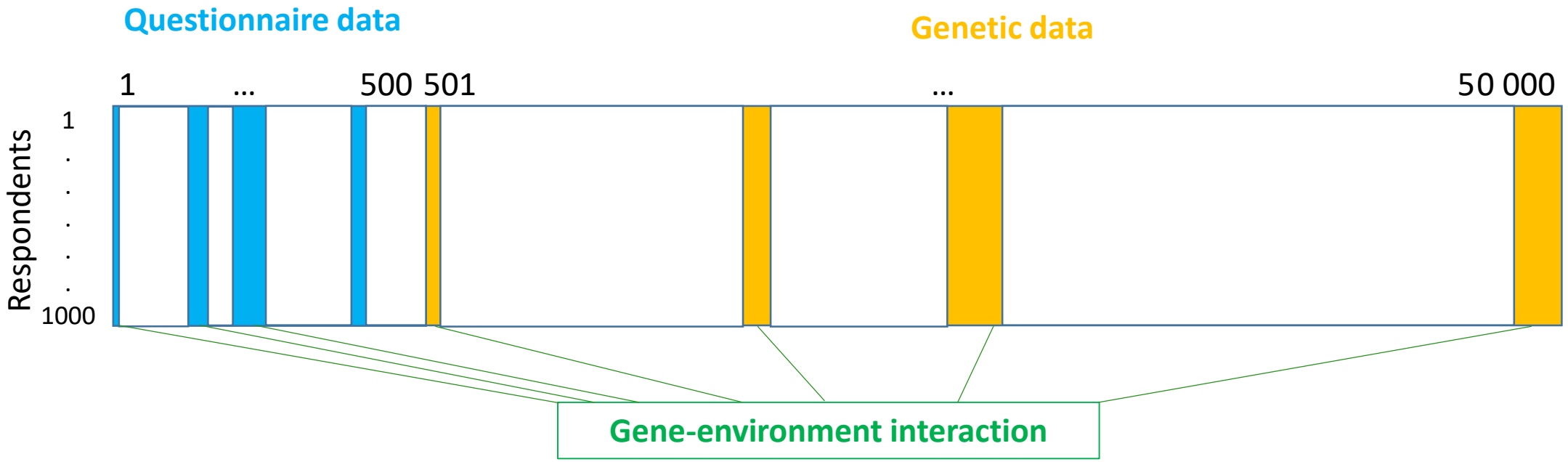

Bob
Politie Police
Survey
Bob
SMART CARD
Bob
Bob
Blend it!
Saspssem inc.
This may take some time; go for a walk in the mountains
WARNING: Ill-conditioned matrix
Best Predictor: Nose length
R²(new case) = 0

In sum: The challenges

- Automated selection of variables needed, in the (ultra-) high-dimensional setting (=many more variables than observations)
- Integrative approach: find shared mechanisms, even if hidden / dominated by source-specific sources of variation
- Computational efficiency (time + memory)
- Avoid capitalization on chance

PART I:
Sparse Common Components

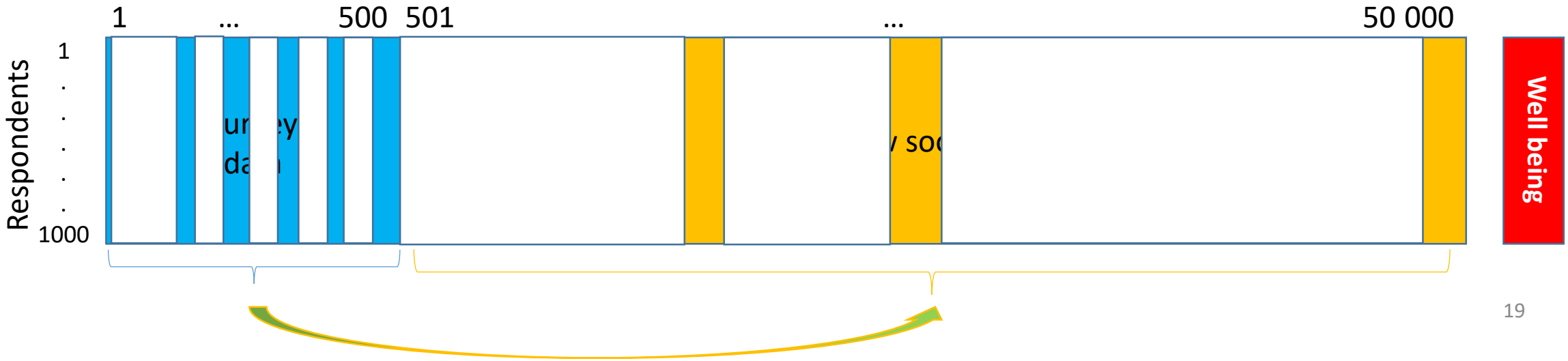
Revealing the linked variables



Sparse CoCo: Basic principles

- Detection of relevant variables
- Find shared / linked mechanisms throughout the data blocks

=> Selection of **linked** variables (between blocks)



Solving the challenges

- Exploratory approach needed

- Find structural sources of variation / shared mechanisms

  $\Rightarrow$Usual suspect: Principal component analysis (Exploratory Factor Analysis)

  $\Rightarrow$Sparse CoCo is an extension of PCA

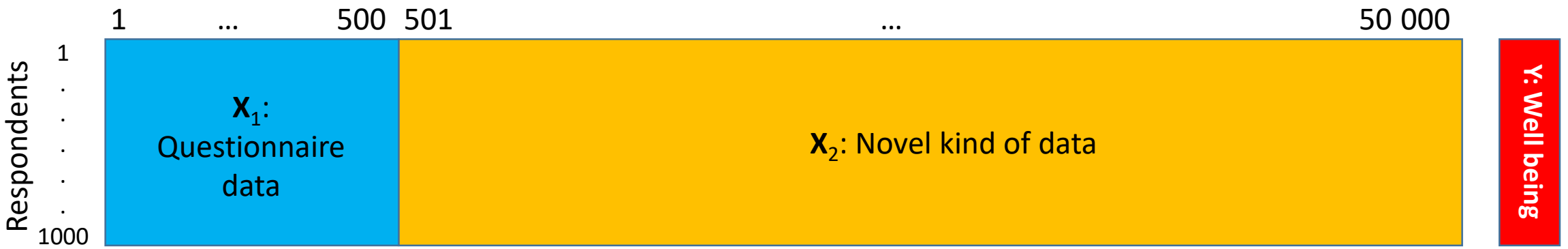
The first suggestion I would have for the authors would therefore be to present what they are doing in terms that will be more recognizable to readers. In the discussion they imply that their method could be applied to factor analysis as well as PCA. If this is truly the case, then I'd suggest that they rework their presentation and present this as an extension of factor analysis instead of PCA. This would dramatically increase reader interest in the paper.

An anonymous reviewer

are equivalent in (14.81) for any $q \times q$ orthogonal $\mathbf{R}$. This leaves a certain subjectivity in the use of factor analysis, since the user can search for rotated versions of the factors that are more easily interpretable. This aspect has left many analysts skeptical of factor analysis, and may account for its lack of popularity in contemporary statistics. Although we will not go into

The Elements of Statistical Learning (2009)

- Notation and naming conventions
  - *data block*: denotes the different data sources forming the *multiblock* data
  - $\mathbf{X}_k$: data block k (with k=1,…,K); the outcome(s) - if present - is denoted by **Y** (**y** if univariate)
  - Each of the data blocks: same set of observation units (respondents)



The many faces of PCA

- 1. Linear combination with maximal variance

  $\max_{\mathbf{w}} \mathbf{w}\mathbf{X}'\mathbf{X}\mathbf{w}$ with $\mathbf{w}'\mathbf{w} = 1$ and $\boldsymbol{w}_r{}'\boldsymbol{w}_q = \mathbf{0}$ for $r \neq q$

- 2. Data reconstruction/dimensionality reduction

  $\min_{\mathbf{T},\mathbf{P}} \left\| \boldsymbol{X} - \mathbf{T}\mathbf{P}' \right\|^2$ s.t. **T'T=I**

- 3. Data reconstruction/dimensionality reduction

  $\min_{\mathbf{W},\boldsymbol{P}} \left\| \boldsymbol{X} - \mathbf{X}\mathbf{W}\mathbf{P}' \right\|^2$ s.t. **P'P=I**

- 4. And the SVD:

  $\mathbf{X} = \mathbf{U}\mathbf{S}\boldsymbol{V}' + \boldsymbol{E}$ with $\mathbf{U}'\mathbf{U} = \mathbf{I}$, $\mathbf{V}'\mathbf{V} = \mathbf{I}$, and $\mathbf{S}$ diagonal

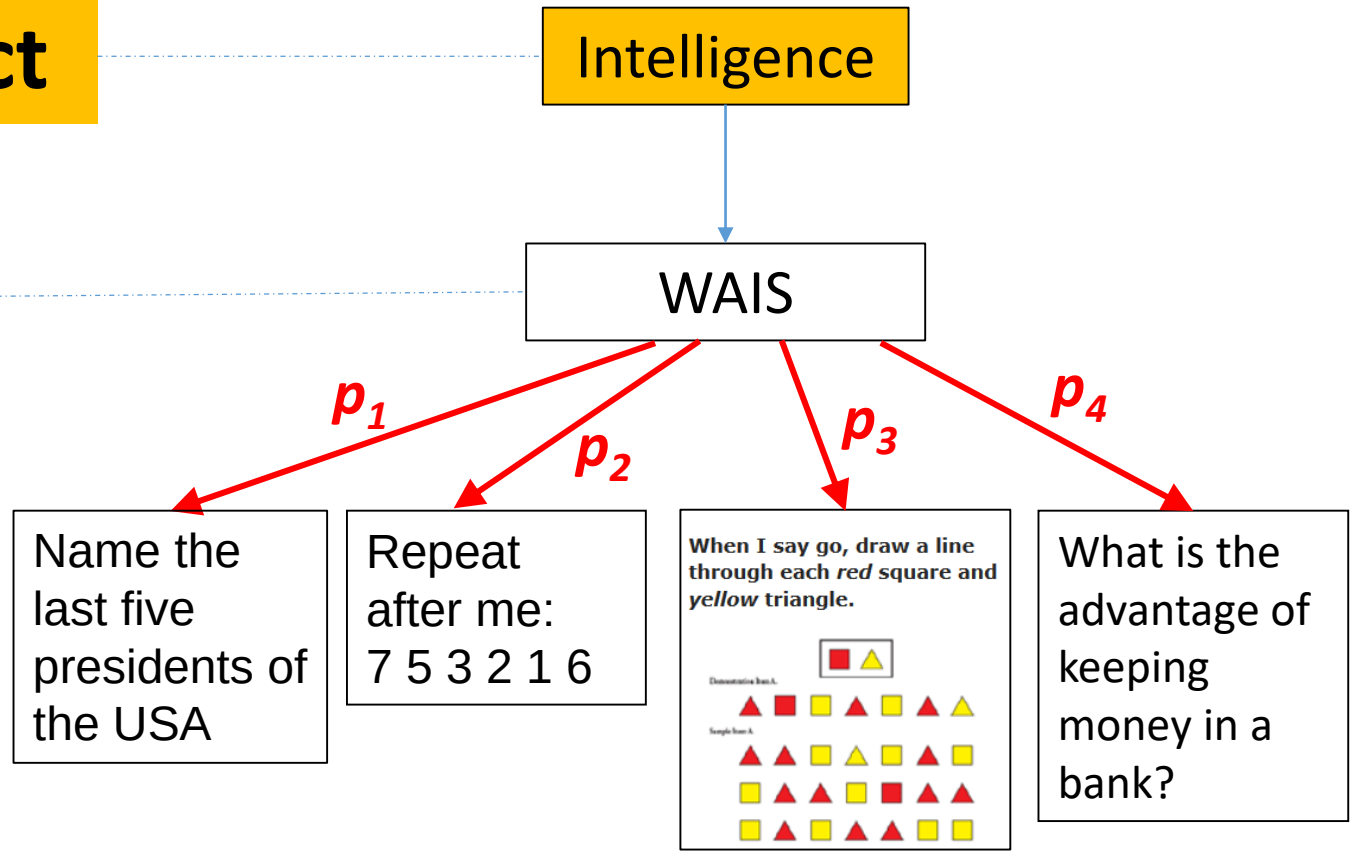
Do the distinctions matter? Aren't they all the same/equivalent?

Focus on loadings **P**: Indicators *reflect* the concept

Concept = construct

Measure: scale

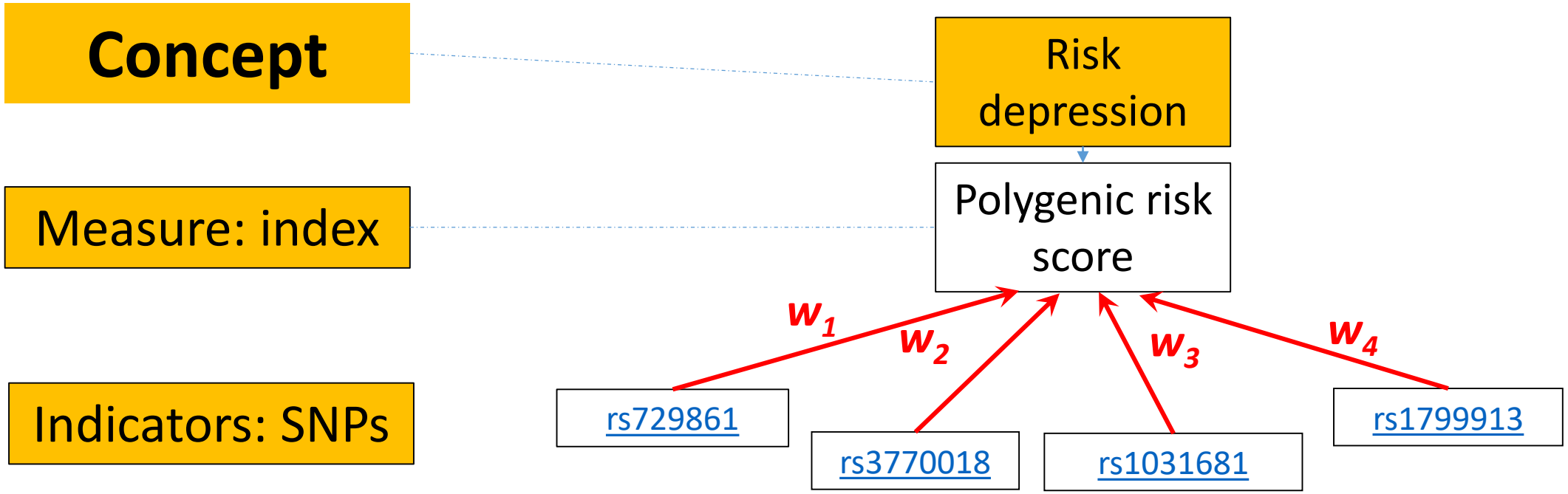
indicators



Scale: measures a concept that is **latent,** non-observable

Indicators reflect the “effects” of the concept

Focus on weights **W**: Indicators *form* the concept



Index: measures a concept that depends on **observable** characterisics
In this case indicators “make up” the concept; reification

Here: focus on data reconstruction, considering both the weight (formative) and loading (reflective) PCA model

- **Principal** component models
  - Weight based variant:

$$\mathbf{X}_k = \mathbf{X}_k \mathbf{W}_k \mathbf{P}_k^T + \mathbf{E}_k \quad \text{s.t. } \mathbf{W}_k^T \mathbf{W}_k = \mathbf{I}, \qquad \textbf{(1)}$$

$$= \mathbf{T}_k \mathbf{P}_k^T + \mathbf{E}_k$$

with $\mathbf{W}_k$ ($J_k \times R$) the component weights, $\mathbf{T}_k$ ($I \times R$) the component scores, and $\mathbf{P}_k$ ($J_k \times R$) the component loadings

Interpretation of component $\mathbf{t}_{rk}$ based on J_k (!) regression weights:

$t_{irk} = \sum_j w_{jrk} x_{ijk}$ ($T \sim \mathbf{X}_k$ -> *high-dimensional regression*)

- Principal component models
  - Loading based variant

$\mathbf{X}_k = \mathbf{T}_k\mathbf{P}_k^T + \mathbf{E}_k$ s.t. $\mathbf{T}_k^T\mathbf{T}_k = \mathbf{I}$, **(2)**

Interpretation of component $\mathbf{t}_{rk}$ based on J_k (!) correlations:

$r(\mathbf{x}_{jk},\mathbf{t}_{rk})=p_{jrk}$ ($X_{jk} \sim T_{rk}$ -> *low-dim. regression, indep. pred.*)

- Note: In a least squares approach subject to $\mathbf{P}_k^T\mathbf{P}_k=\mathbf{I}$, we have $\mathbf{W}_k=\mathbf{P}_k$

- **Simultaneous** component analysis

For all *k:*

$$\mathbf{X}_k = \mathbf{T}\mathbf{P}_k^T + \mathbf{E}_k \text{ s.t. } \mathbf{T}^T\mathbf{T}=\mathbf{I} \qquad (3)$$

-> same component scores for all data blocks!

1. *Weight based* variant:

$$[\mathbf{X}_1 \dots \mathbf{X}_K] = \mathbf{T} [\mathbf{P}_1^T \dots \mathbf{P}_K^T] + [\mathbf{E}_1 \dots \mathbf{E}_K^T]$$
$$= [\mathbf{X}_1 \dots \mathbf{X}_K] [\mathbf{W}_1^T \dots \mathbf{W}_K^T]^T [\mathbf{P}_1^T \dots \mathbf{P}_K^T] + [\mathbf{E}_1 \dots \mathbf{E}_K^T]$$

2. *Loading based* variant:

$$\mathbf{X}_{conc} = \mathbf{T}\ \mathbf{P}_{conc}^T + \mathbf{E}_{conc} \qquad (4)$$

- Simultaneous components
  - Model the largest source of variation in the concatenated data, often this is source-specific variation or a mix of common and specific variation
  - Guarantee same source of variation for the different data blocks
  - But … do not guarantee that the components are common!
    - See also Schouteden et al. (BRM, 2013): Rotation to common & specific components

- **Common component analysis (CoCo)**
  - Account for dominating source specific variation
  - Common component model:

$\mathbf{X}_{conc} = \mathbf{TP}_{conc}'$ such that $\mathbf{T'T=I}$ and $\mathbf{P}_{conc}$ has **common / specific structure**

	P1			P2			
Common	x	x	x	x	x	**x**	x
Spec1	x	x	x	**0**	**0**	**0**	**0**
Spec1	**0**	**0**	**0**	x	x	x	x

  - Structure can be imposed (*constrained analysis, if K small / structure known*)
  - Or: $\mathbf{X}_k = \mathbf{X}_{conc}\mathbf{W}_{conc}\mathbf{P}_k'$ with the $\mathbf{W}_{conc}$ having **common / specific structure**

- So far, so good …
- Yet:
  - Interest in relevant / important variables (social factors, genetic markers)?
  - Interpretation of the components based on 1000s of loadings/weights is infeasible
  - PCA estimates are not consistent in case of high-dimensional data (Johnstone & Lu)

$\Rightarrow$Need to **automatically** select the "relevant" variables!

$\Rightarrow$**Sparse common components** are needed.

- Structured **sparsity**:
  - Sparse common components: few non-zero loadings in each data block $\mathbf{X}_k$
  - (Sparse) specific components: (few) non-zero loadings only in one/few data blocks $\mathbf{X}_k$

	X1			X2			
Common	x	**0**	**0**	x	x	**0**	x
Dist1	x	x	x	**0**	**0**	**0**	**0**
Dist1	**0**	**0**	**0**	x	**0**	x	x

Variable selection: How to?

- Sparse analysis
  - Impose restriction on the loadings/weights: many should become zero
  - S-O-A: add penalties (e.g., lasso) to the objective function

- Sparse SCA: Objective function

Add penalty known to have variable selection properties to SCA objective function:

Minimize over **T** and $\mathbf{P}_{Conc}$ and such that $\mathbf{T}'\mathbf{T} = \mathbf{I}$ and $\mathbf{P}_{conc}$ constrained

$$\underbrace{\|\boldsymbol{X}_{Conc} - \mathbf{T}\mathbf{P}'_{Conc}\|^2}_{\text{Fit / SCA}} \quad + \quad \underbrace{\sum_{r,k} \lambda_{r,k} \left|\mathbf{p}_{r,k}\right|_1}_{\text{Penalty}}$$

with $\left|\mathbf{p}_{r,k}\right|_1 = \sum_j \left|p_{j_k r}\right|$ the L_1 penalty or lasso tuned by $\lambda_{r,k} \geq 0$

-> shrinks and selects variables
-> penalty can also be applied to the weights:

$$\|\boldsymbol{X}_{Conc} - \boldsymbol{X}_{Conc}\boldsymbol{W}_{Conc}\mathbf{P}'_{Conc}\|^2 \quad + \quad \sum_{r,k} \lambda_{r,k} \left|\mathbf{w}_{r,k}\right|_1$$

Why not rotation to simple structure?

- Must read : Cadima & Jolliffe (1995)
- Ordinary simultaneous component analysis, common interpretation practice (VARIMAX + thresholding)
  - %VAF: .31 (if calculated with neglected loadings set equal to zero)
- Sparse SCA, lasso in action
  - %VAF: .35 (>.31, VARIMAX!!!!)

		PC1	PC2	PC3
Discrete Emotions	ANGRY	0	0	0
	ANGRY_1	0	0	0
	DEPRE	**0.24**	0	0
	DEPRE_1	**0.23**	0	0
	SAD	**0.24**	0	0
	SAD _1	**0.24**	0	0
	ANXIOUS	**0.21**	0	0
	ANXIOUS _1	**0.21**	0	0
	RELAXED	0	**0.23**	0
	RELAXED_1	0	**0.20**	0
	HAPPY	0	0	0
	HAPPY_1	**-0.20**	0	0
Emotion Regulation	RUMINATION	**0.22**	0	0
	RUMINATION_1	**0.20**	0	0
	REFLECTION	0	**0.26**	0
	REFLECTION_1	0	**0.28**	0
	RE-APPRAISAL	0	**0.23**	0
	RE-APPRAISAL_1	0	**0.25**	0
	SUPPRESSION	**0.20**	0	0
	SUPPRESSION_1	0	0	0
	SOCIAL SHARING	0	**0.31**	0
	SOCIAL SHARING_1	0	**0.32**	0
	DISTRACTION#	0	0	**-0.22**
	DISTRACTION_1	0	0	**-0.20**
Appraisals	IMPORTANCE_1	0	**0.21**	0
	IMPORTANCE_1	0	**0.21**	0
	(DIS)ADVANTAGEOUS	0	**0.26**	0
	(DIS)ADVANTAGEOUS_1	0	**0.22**	0
	OTHER RESPONSIB	0	0	**0.41**
	OTHER RESPONSIB_1	0	0	**0.40**
	SELF-RESPONSIB	0	0	**-0.39**
	SELF-RESPONSIB_1	0	0	**-0.40**
	CONTROLLABILI	0	0	**-0.20**
	CONTROLLABILI_1	0	0	**-0.22**
	EMOT COPING	**-0.21**	0	0
	EMOT COPING_1	**-0.22**	0	0

- **In-house Algorithm: Alternating procedure**

Given fixed tuning parameters and number of common and distinctive components, do

0. Initialize $\mathbf{P}_{conc}$

1. Update **T** conditional upon $\mathbf{P}_{conc}$

   Closed form: **T**=**UV**' with **U** and **V** from the SVD of $\mathbf{X}_C\mathbf{P}$ (*I*×*R* -> *small for H-D data*)

2. Update $\mathbf{P}_{conc}$ conditional upon **T**

   Coordinate descent (see next)

3. Check stop criteria (convergence of the loss, maximum number of iterations) and return to step 1 or terminate

Coordinate descent

$$L = \|\boldsymbol{X}_{Conc} - \mathbf{T}\mathbf{P}'_{Conc}\|^2 + \sum_{r,k} \lambda_{r,k} |\mathbf{p}_{r,k}|_1$$

$$= k + \sum_{j_k,r} \left\{ \left(\sum_i \left(x_{ij} - t_{ir} p_{j_k r} \right)^2 \right) + \lambda_{r,k} |p_{j_k r}| \right\}$$

For which the root can be found using subgradient techniques (this is a standard lasso regression problem)

Note that the problem is separable over *j* and that **T** is orthogonal

=> all $p_{j_k r}$ can be updated simultaneously (fast!!!)

=> fixing of loadings / constrained analysis is 

- Hence, the following soft thresholding update of the loadings:

$$p_{j_k r} = \begin{cases} \dfrac{\sum_i x_{i j_k} t_{ir} - \lambda_r/2}{1 + d_{j_k r}} & \text{if} \quad p_{j_k r} > 0 \\ \dfrac{\sum_i x_{i j_k} t_{ir} + \lambda_r/2}{1 + d_{j_k r}} & \text{if} \quad p_{j_k r} < 0 \\ 0 & \text{else} \end{cases}$$

which can be calculated for all loadings of component *r* simultaneously using simple vector and matrix operations!!!

=> *Highly efficient (time+memory), scalable to large data*

- Algorithm: Weight based variant (sparseness on weights)

  - Similar type of algorithm can be constructed

  - Expression to estimate $w_{j_k^* r^*}$ using coordinate descent (cycle over all $R \sum_k J_k$ coefficients); sparse group lasso case

$$w_{j_k^* r^*} = \begin{cases} \dfrac{\sum_{i,k,j_k} x_{ij_k^*} r_{ij_k} p_{j_k r^*} - \lambda_{1,r}/2}{1 + \lambda_{2,r}} & \text{if } w_{j_k^* r^*} > 0 \\ \dfrac{\sum_{i,k,j_k} x_{ij_k^*} r_{ij_k} p_{j_k r^*} + \lambda_{1,r}/2}{1 + \lambda_{2,r}} & \text{if } w_{j_k^* r^*} < 0 \\ 0 & \text{else} \end{cases}$$

Ridge penalty has to be included

The expression for an individual coefficient is not very expensive but has to be calculated many times. Also here adding elementwise constraints is 

- Algorithm

  - Coordinate-wise approach allows to fix coefficients

  - This can be used to define the specific components by fixing to zero those coefficients corresponding to the block(s) that is not accounted for by the component

  - Often no prior knowledge on number of specific components; for which block(s) they are specific => we also included a group lasso penalty (performs selection at the block level)

  - Is run with input of the nr of components; their status (common/specific) and/or values for the (group) lasso tuning parameter => **MODEL SELECTION**

Some illustrations

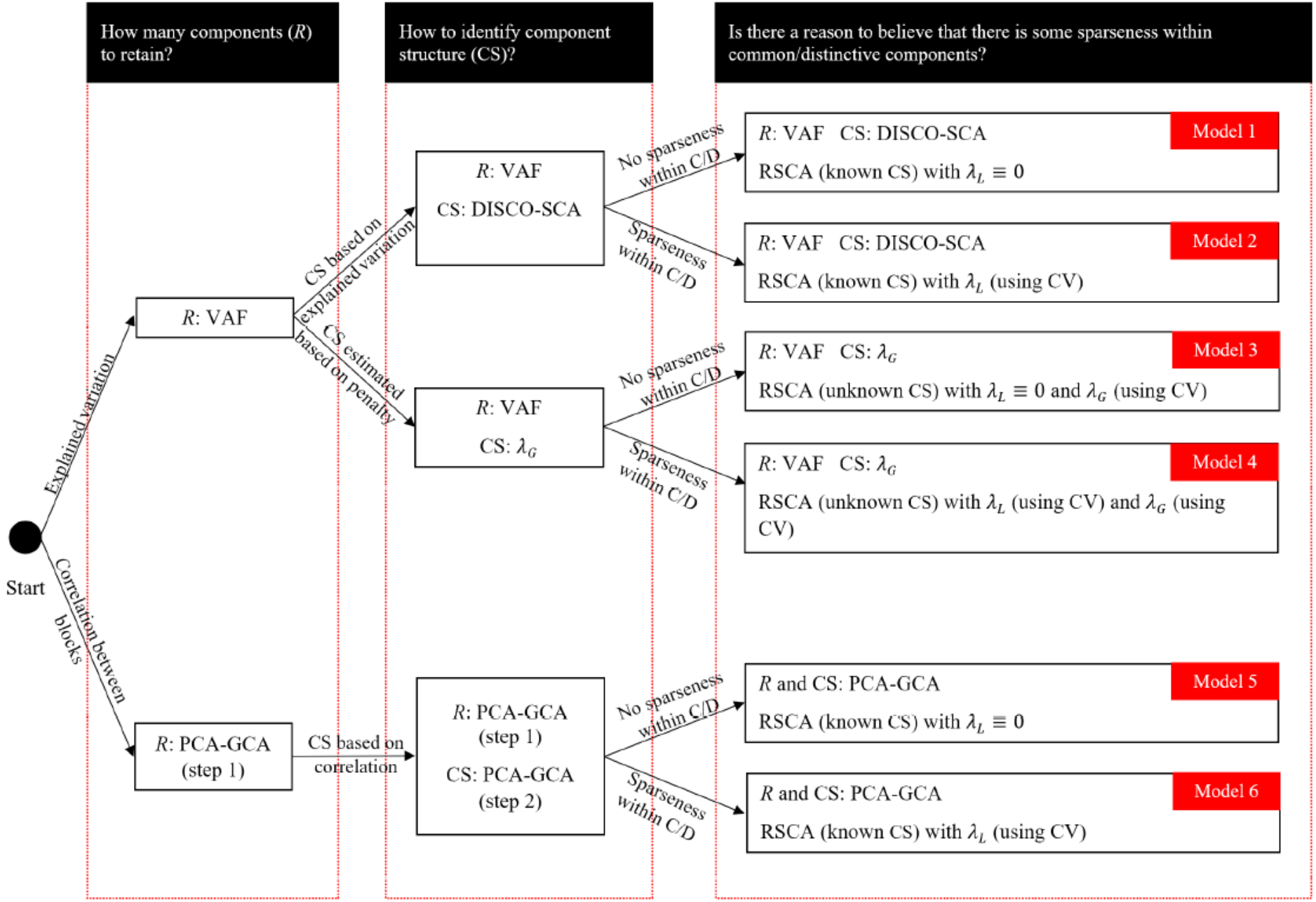
500 Family Study (Schneider & Linda, 2008)

- Three data blocks: 195 families
  - Father -> 8 variables
  - Mother -> 8 variables
  - Child -> 7 variables

  Here: part of questionnaires

- Analyzed with the RegularizedSCA package (available from CRAN)
  - This is the sparse loading variant
  - Several options for model selection; includes the DISCO rotation

- Exploratory analysis

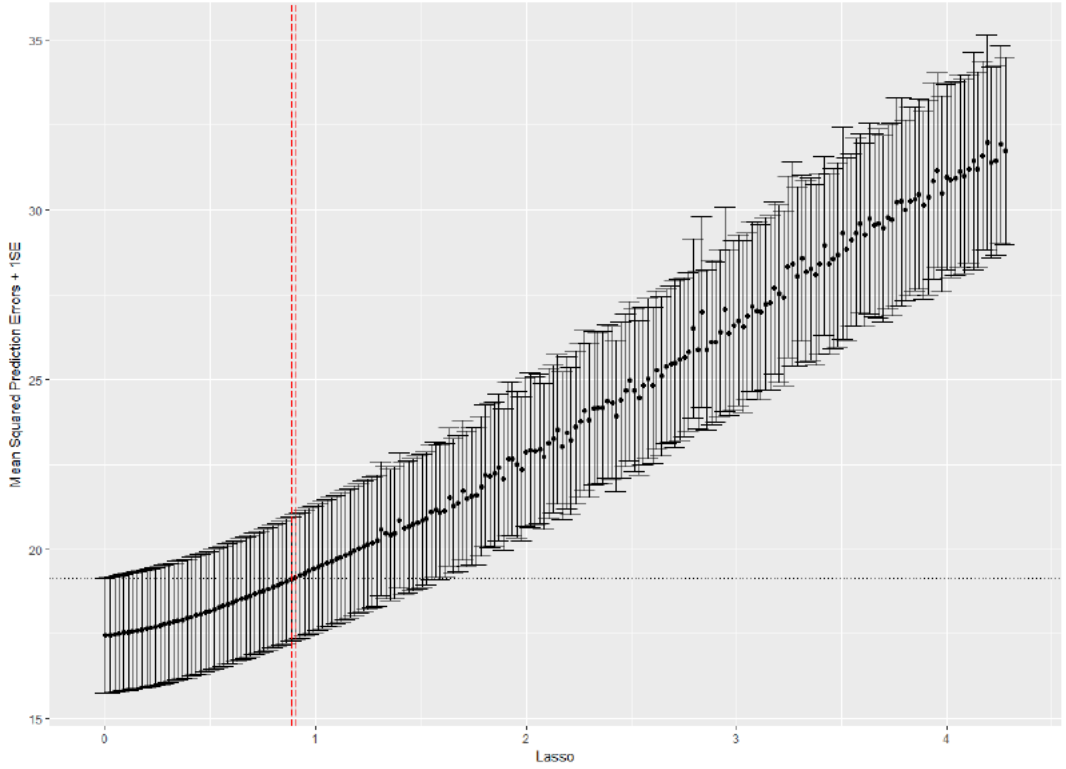

How many components (R) to retain?
How to identify component structure (CS)?
Is there a reason to believe that there is some sparseness within common/distinctive components?
Start
Explained variation
Correlation between blocks
R: VAF
R: PCA-GCA (step 1)
CS based on explained variation
CS estimated based on penalty
CS based on correlation
R: VAF
CS: DISCO-SCA
R: VAF
CS: λG
R: PCA-GCA (step 1)
CS: PCA-GCA (step 2)
No sparseness within C/D
Sparseness within C/D
No sparseness within C/D
Sparseness within C/D
No sparseness within C/D
Sparseness within C/D
R: VAF CS: DISCO-SCA
RSCA (known CS) with λL ≡ 0
Model 1
R: VAF CS: DISCO-SCA
RSCA (known CS) with λL (using CV)
Model 2
R: VAF CS: λG
RSCA (unknown CS) with λL ≡ 0 and λG (using CV)
Model 3
R: VAF CS: λG
RSCA (unknown CS) with λL (using CV) and λG (using CV)
Model 4
R and CS: PCA-GCA
RSCA (known CS) with λL ≡ 0
Model 5
R and CS: PCA-GCA
RSCA (known CS) with λL (using CV)
Model 6

- Model Selection Step 1: Number of components?
  - VAF method => 5 components

```
Proportion of VAF for each component of each block:
          [,1]       [,2]       [,3]       [,4]       [,5]       [,6]       [,7]       [,8]       [,9]      [,10]
[1,] 0.1902580 0.08505923 0.08672370 0.07153828 0.06447496 0.05526066 0.02852091 0.03936437 0.05352701 0.05906965
[2,] 0.1366034 0.09455771 0.06367872 0.08605901 0.10354762 0.08630231 0.03933645 0.04906133 0.04728068 0.02230673
[3,] 0.1940918 0.10392886 0.11339839 0.06370639 0.02089805 0.03280644 0.09052555 0.03910193 0.01721554 0.02715787
```

- Model selection Step 2: Status components and degree of sparseness?
  - Cross-validation for both group lasso and lasso tuning parameters

```
Recommended tuning parameter values are:
       Lasso Group Lasso
[1,] 2.82068     1.28369
```

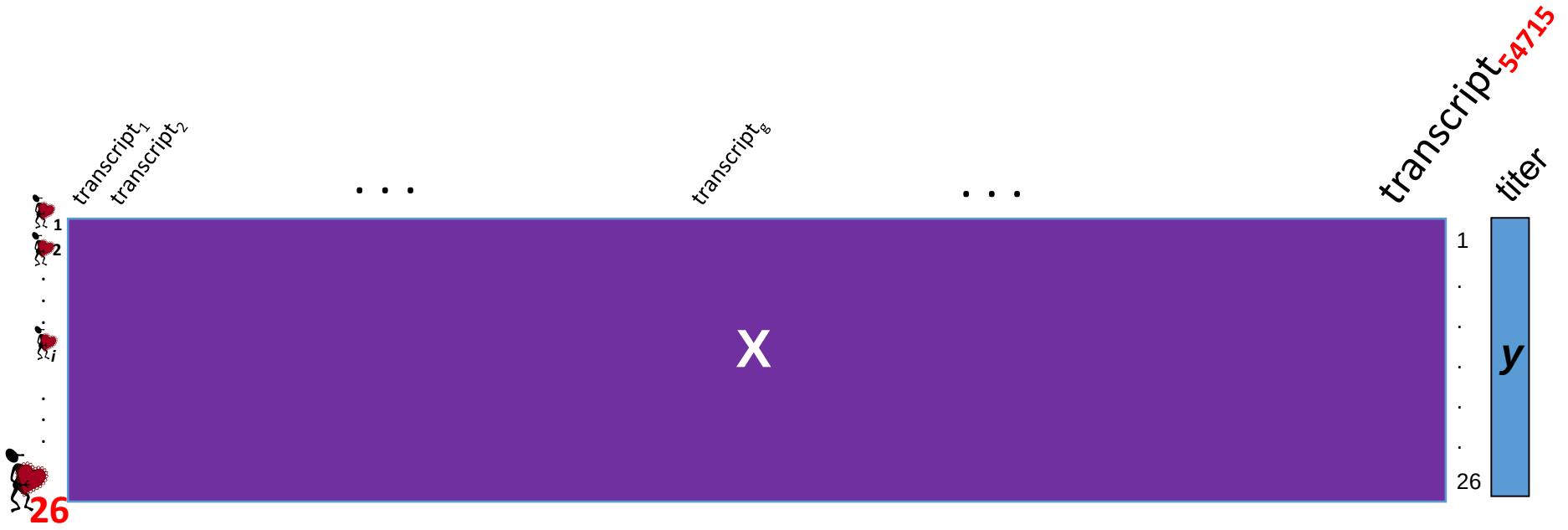


	[,1]	[,2]	[,3]	[,4]	[,5]
M: Relationship with partners	0.000000	12.063380	0.000000	0.000000	0.000000
M: Argue with partners	-5.656729	5.606819	0.000000	0.000000	0.000000
M: Childs bright future	-8.583745	0.000000	0.000000	0.000000	0.000000
M: Activities with children	-4.462827	0.000000	0.000000	-9.015597	0.000000
M: Feeling about parenting	-8.757033	2.750746	0.000000	0.000000	0.000000
M: Communation with children	-9.077420	0.000000	0.000000	-3.305209	0.000000
M: Argue with children	-9.268780	0.000000	0.000000	3.193377	0.000000
M: Confidence about oneself	-6.680594	7.242076	0.000000	0.000000	0.000000
D: Relationship with partners	0.000000	11.903001	0.000000	0.000000	0.000000
D: Argue with partners	0.000000	5.104832	-9.236275	0.000000	0.000000
D: Childs bright future	-3.171864	0.000000	-5.817937	0.000000	0.000000
D: Activities with children	0.000000	0.000000	0.000000	-11.016351	0.000000
D: Feeling about parenting	-3.940123	0.000000	-6.861383	-2.153304	0.000000
D: Communation with children	0.000000	0.000000	0.000000	-8.373867	0.000000
D: Argue with children	-5.098655	0.000000	-9.855232	0.000000	0.000000
D: Confidence about oneself	0.000000	5.561253	-8.181819	0.000000	0.000000
Self confidence/esteem	-5.963305	0.000000	0.000000	0.000000	-8.467937
Academic performance	0.000000	0.000000	0.000000	0.000000	-7.091520
Social life and extracurricular	0.000000	0.000000	0.000000	0.000000	-4.260376
Importance of friendship	0.000000	0.000000	0.000000	0.000000	-9.715712
Self Image	-2.560587	0.000000	0.000000	0.000000	-10.153244
Happiness	0.000000	0.000000	0.000000	0.000000	-9.424702
Confidence about the future	0.000000	0.000000	0.000000	0.000000	-7.516765
	Common	Parents	Father	Parents	Child

PART 2:
Prediction in a high-dimensional context, the psychologist's way

Illustration: An insightful linear model for high-dimensional prediction

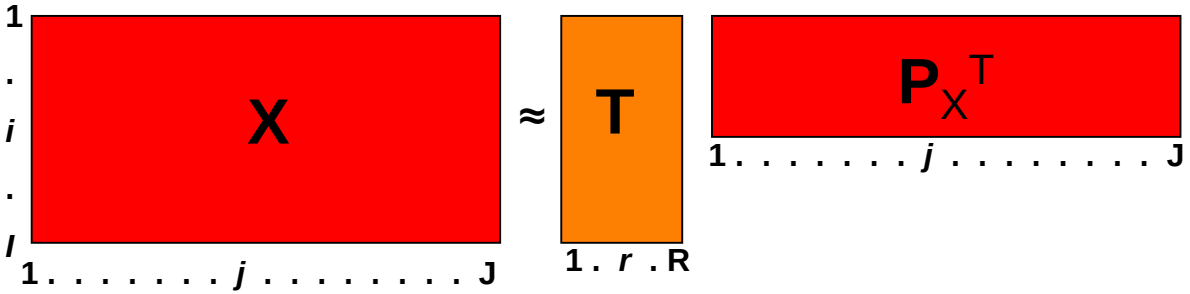
- Motivating example: Systems vaccinology (Nakaya et al., 2011)



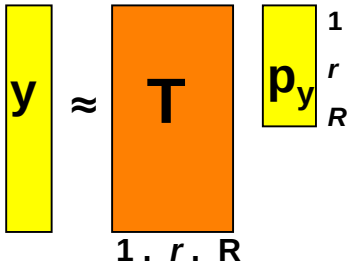
Sparse Principal Covariates Regression

- Extension of PCovR (de Jong & Kiers, 1992)
- *Simultaneous* sparse PCA and regression
- Stability selection (Meinshausen & Buhlmann, 2010)

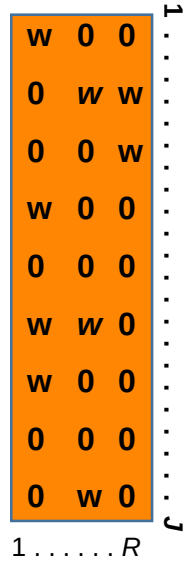
$$\mathbf{X} = \mathbf{XWP}_x^T + \mathbf{E}_x = \mathbf{TP}_x^T + \mathbf{E}_x \text{ s.t. } \mathbf{T}^T\mathbf{T} = \mathbf{I}$$



$$\boldsymbol{y} = \mathbf{XW}\boldsymbol{p}_y + \mathbf{E}_y = \mathbf{T}\boldsymbol{p}_y + \mathbf{E}_y \text{ s.t. } \mathbf{T}^T\mathbf{T} = \mathbf{I}$$



W sparse



Systems vaccinology example: Comparison of results with sparse PLS

Table 1 Fit of modeled to observed data for three methods: SPCovR, spls, and SGCCA. Displayed are the variance accounted for by the components in the block of covariates and the squared correlation between the modeled and observed outcome for the 2008 and 2007 season. The model was constructed using the 2008 data.

Method	VAF	$r(\hat{y}, y)^2$	$r(\hat{y}_{2007}, y_{2007})^2$
SPCovR	0.19	0.42	0.79
spls		0.99	0.55
SGCCA	0.11	1	0.53

$r(\hat{y}, y)^2$ for elastic net = 0,06 (200 non-zero regression weights, tuned with CV using glmnet)

- SPCovR: Biological content of selected transcripts

Significantly enriched gene ontology classes

	Biological process	Nr of genes found	Nr of genes expected	+/−	*P*-value
PC1	rRNA methylation	5	.21	+	2.03*E*−02
	Cellular macromolecule metabolic process	89	58.65	+	1.68*E*−02
	Nucleic acid metabolic process	60	34.14	+	2.84*E*−02
	Cellular component organization or biogenesis	75	47.15	+	3.86*E*−02
	Gene expression	57	31.88	+	3.30*E*−02
PC2	Leukocyte activation	18	4.59	+	6.11*E*−03
	Cell activation	20	5.36	+	2.88*E*−04
	Immune system process	31	13.13	+	2.79*E*−02
	Immune effector process	19	5.25	+	9.41*E*−03
	Negative regulation of metabolic process	32	14.16	+	4.59*E*−02

- Sparse PLS: no (significant) terms found

DISCUSSION

- We presented a generalization of sparse PCA to the multiblock case
  - Sparsity can be imposed accounting for the block structure
  - Sparsity can be imposed either on the weights or the loadings
  - Several (combinations of) penalties possible

=> Flexible modeling framework that includes several known methods as special cases: E.g., SPCA (Zou et al., 2006); sPCA-rSVD (Shen & Huang, 2007); SCA (Escofier & Pagès, 1983); and PCA

- We also presented an extension of PCovR to the high-dimensional setting

- This method balances prediction with interpretation and –in a sense – presents the best of two worlds (machine learning/data mining and psychometrics)

- Gave highly promising results on worst-case definion of Big Data for the statistician (this is, (ultra-)high-dimensional)

- Stability selection is a key ingredient

- Planned developments / in progress
  - Merger of the two projects: Prediction based on sparse common covariates
  - Extension to long data and the presence of heterogeneity (Presentation of Shuai Yuan)
  - (Further) develop model selection methods
  - Causal relations: network approach (Presentation Pia Tio) and beyond
  - Improve efficiency R package, add weight based approach, add visualization

  - Proof of the pudding: Joint analysis questionnaire + genetic (and other) types of data

Much to learn, we still have

- Software:
  - RegularizedSCA package available from CRAN
  - Sparse Principal Covariates regression available from GitHub:
    https://github.com/katrijnvandeun/SPCovR

- References:

  - Gu, Z., & Van Deun, K. (2018). RegularizedSCA: Regularized simultaneous component based data integration. *Under review*.

  - Van Deun, K., Crompvoets, E.A.V., & Ceulemans, E. (2018). Obtaining insights from high-dimensional data: Sparse principal covariates regression. *BMC Bioinformatics, 19:104*.

Thx to the cool people!

Mr. Regularized SCA

Mrs. SNAC

Mr. Sparse W

Mr. Clusterwise sparse CoCo

Mr. Compare

Thank you!

&