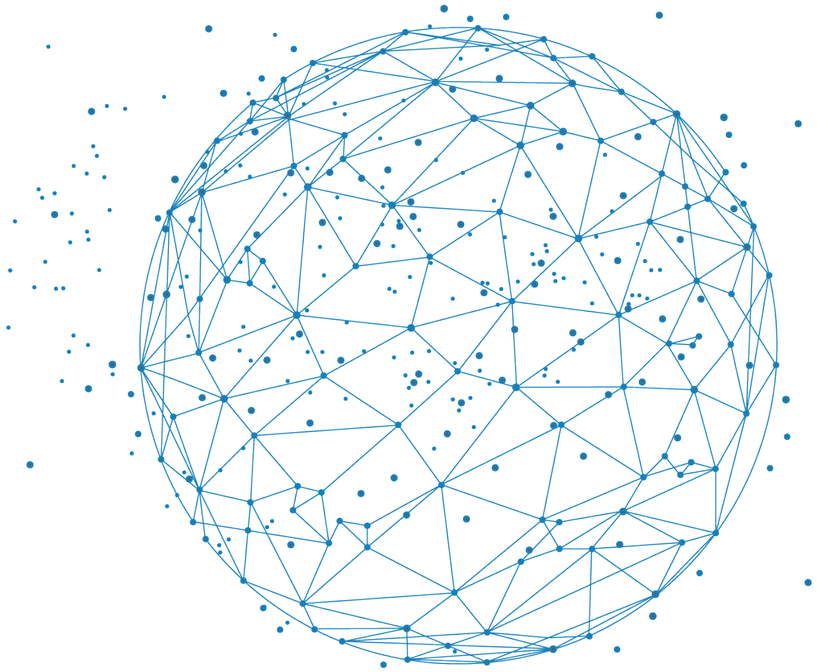
First edition 2023 in Fréjus

ÉCOLE THÉMATIQUE
ETBII
BIOINFORMATIQUE INTÉGRATIVE IFB



Theoretical part multivariate analyses

Carl Herrmann (Université Heidelberg)
Delphine Potier (CNRS Marseille)

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

elixir FRANCE

cnrs INRAE Inserm cea

INVESTIR L'AVENIR

Goals

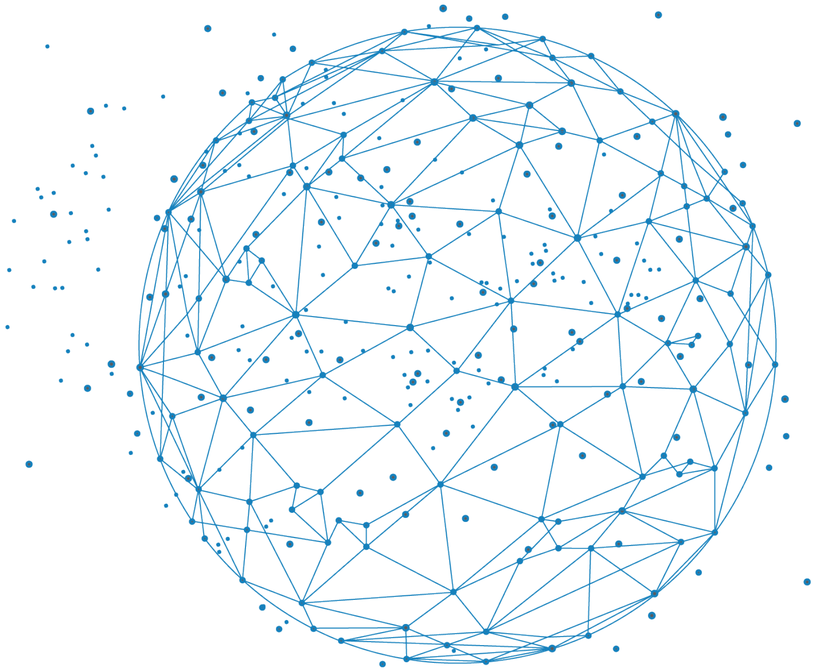
In this lecture, we plan to

- **Review** the fundamental concepts of multivariate statistics (variance, covariance,...)
- **Discuss** the required conditions (distribution, missing data,...)
- **Present** some statistical approaches in MVA and their implementations

At the end, you should be able to

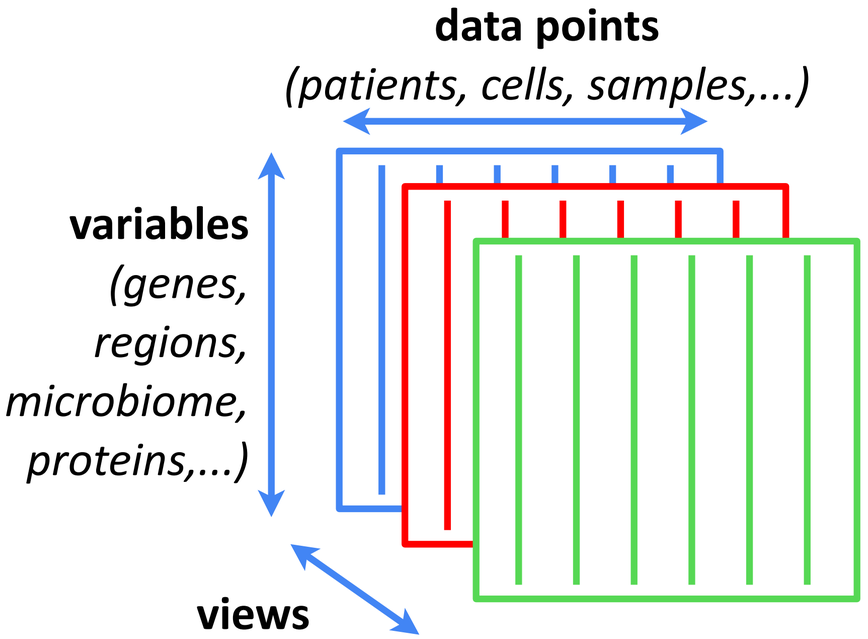
- **Distinguish** the categories of approaches
- **Understand** the vocabulary (factors, signatures, loadings,...)
- Have a better idea how to **select** appropriate tools for your setting.

General introduction to multivariate analyses



What is multivariate analysis?

- Multiple data points (= **observables**) described by multiple measurements (= **variables**)
- Multiple **views** (or modalities)
- Assumption: not all variables are independent
  - **which variables are related?**
  - **can we obtain a simpler description with less dimensions?**
  - **can we learn this description from multiple data types simultaneously?**

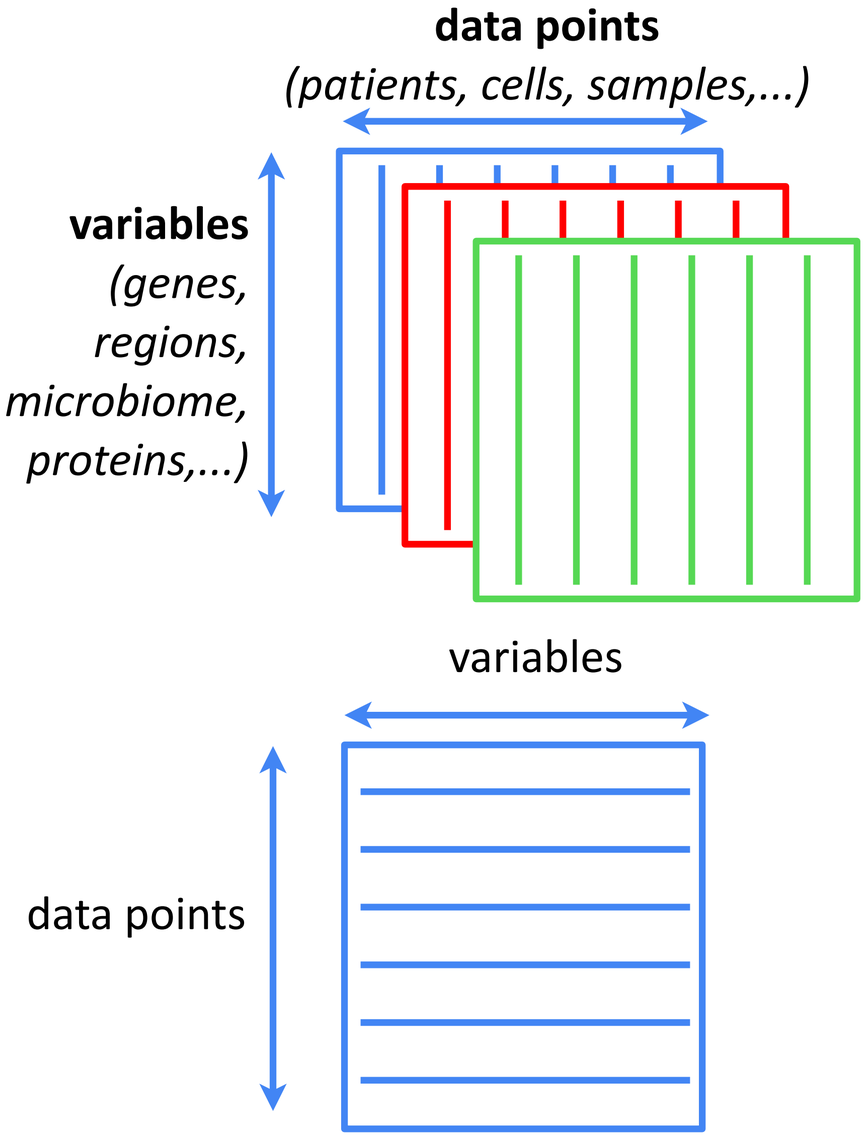


Data reduction

Data integration

What is multivariate analysis?

- Multiple data points (= **observables**) described by multiple measurements (= **variables**)
- Multiple **views** (or madalities)
- Assumption: not all variables are independent
  - **which variables are related?**
  - **can we obtain a simpler description with less dimensions?**
  - **can we learn this description from multiple data types simultaneously?**

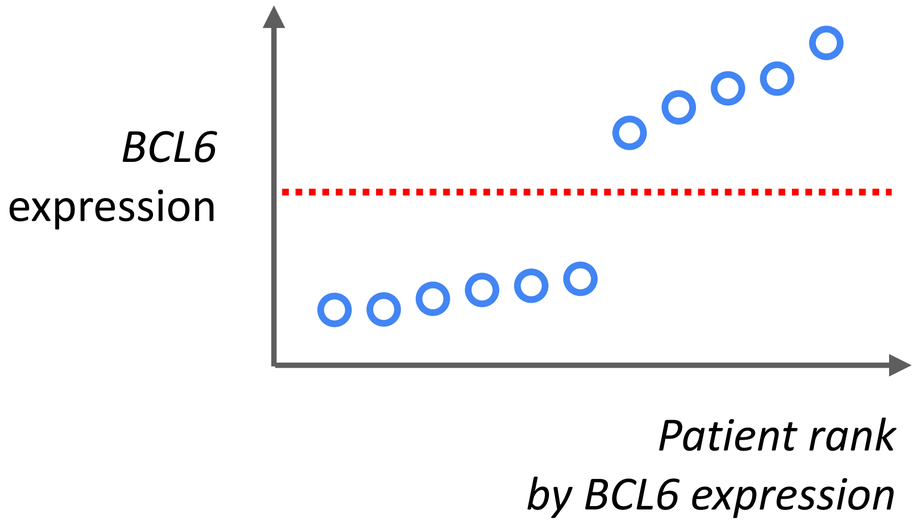


Beware how the matrix is oriented!!

Multivariate vs. Univariate

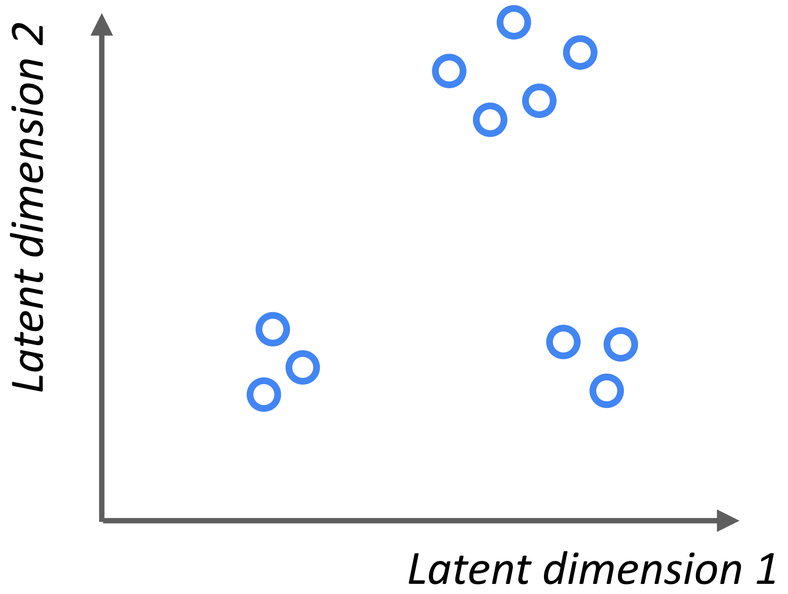
Univariate

- Does the expression of the gene *BCL6* define distinct groups of patients?



Multivariate

- Does the expression of all genes define distinct groups of patients?



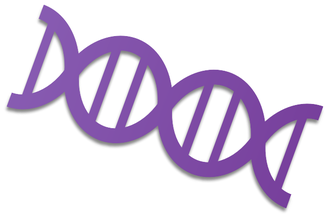
“Whole more than sum of the parts”

Epigenetic state (histone ChIP-seq)

Chromatin accessibility (ATAC-seq)

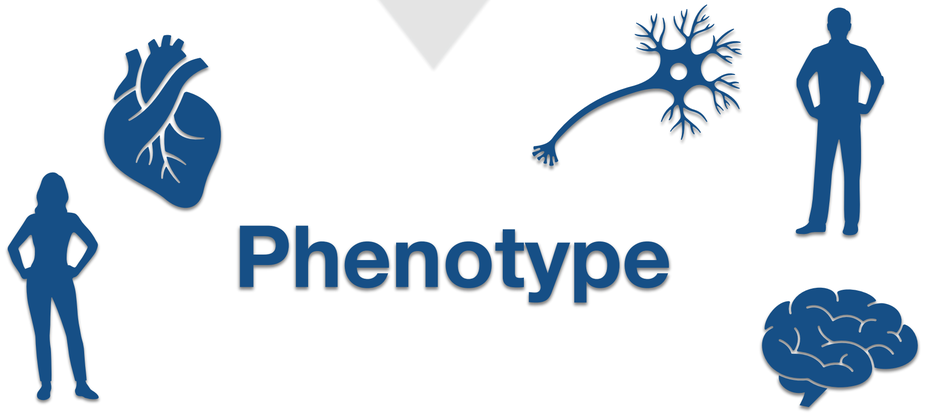
DNA methylation (WGB-seq)

Gene expression (RNAseq)



Genomic information (WGS)

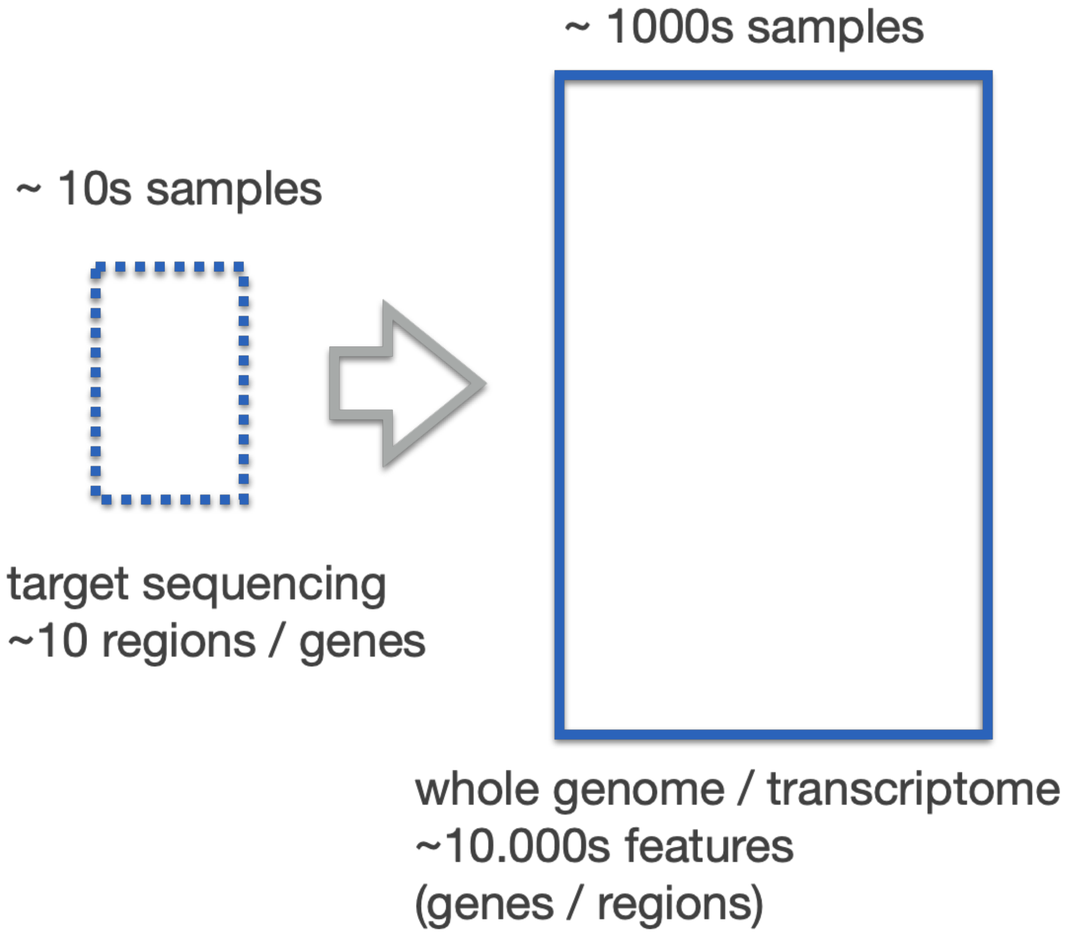
Clinical variables



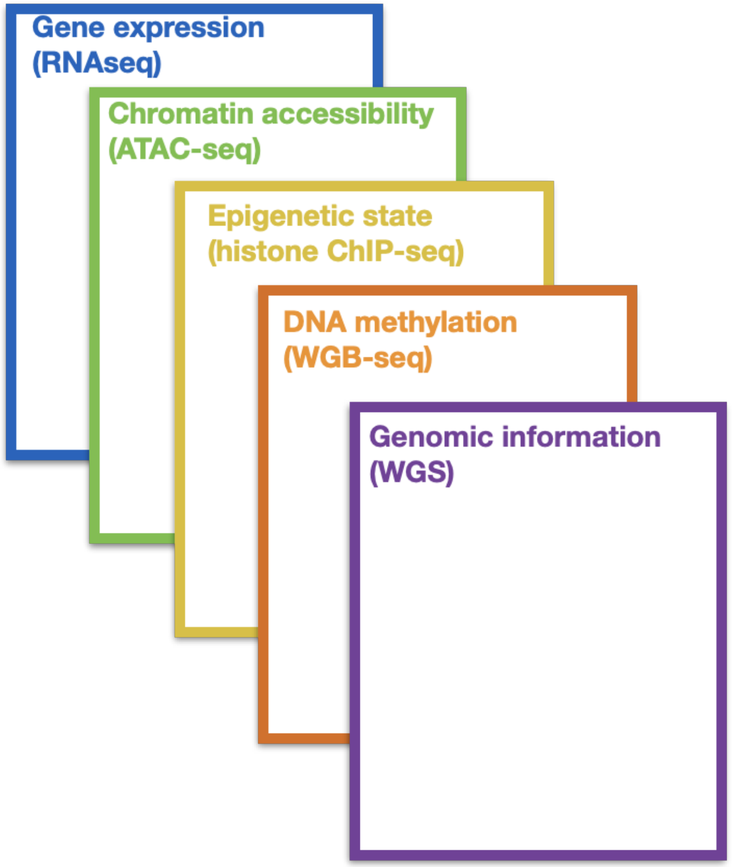
Phenotype

Available datasets

Large data matrices

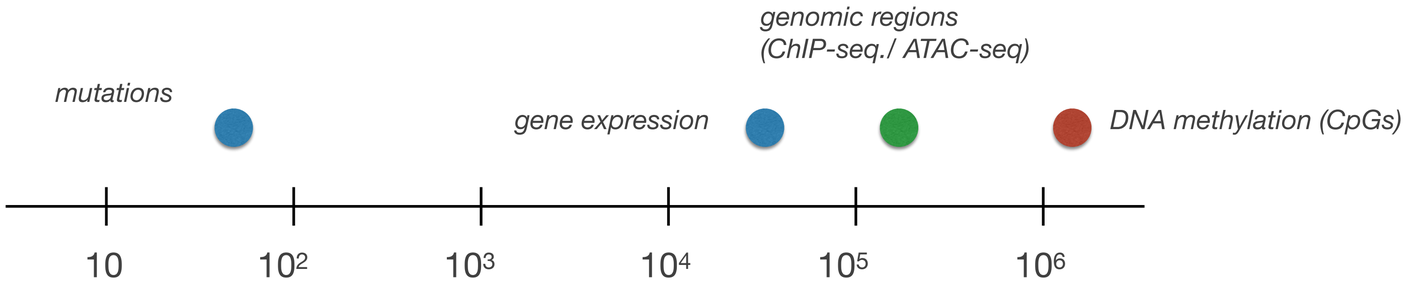


Many data types (“views")

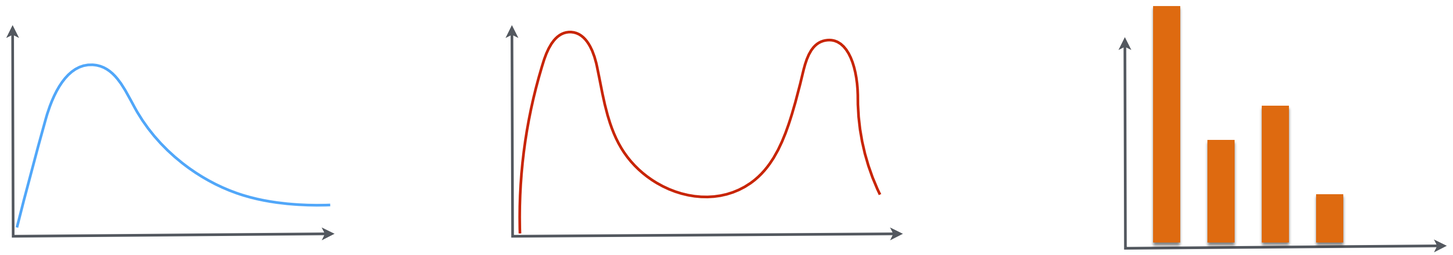


Challenges

- Different **dimensionalities** and **features**



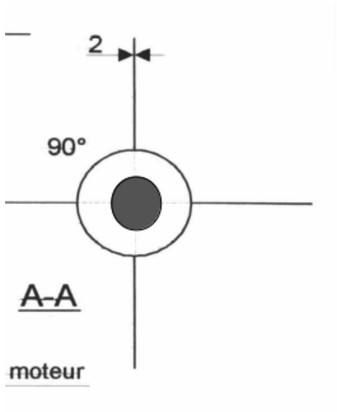
- Different types / **distributions** of data

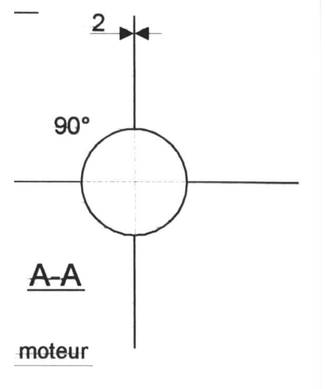


- **Missing data:** not all samples have measurements in all features and all views

Variance in the data

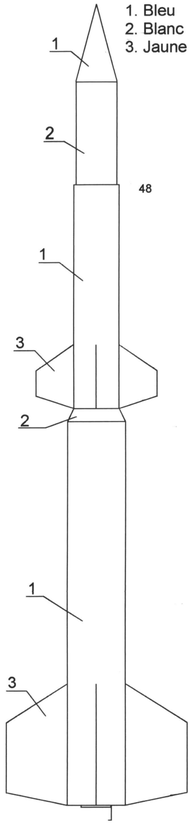
10% information



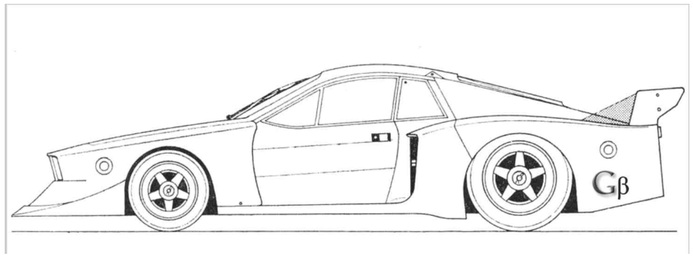


10% information

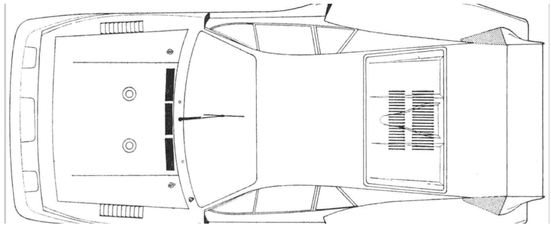
80% information



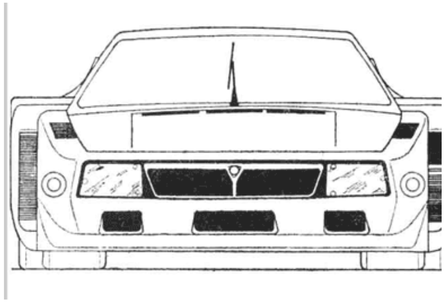
45% information



30% information



25% information



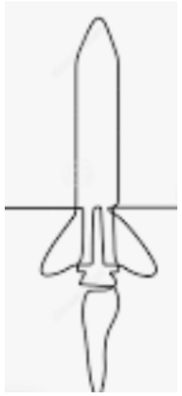
How can we determine the optimal viewing angle?

Variance explained by the model

data

models

explains 80% of the data variance



explains 20% of the data variance

Basic central concepts

- Variance

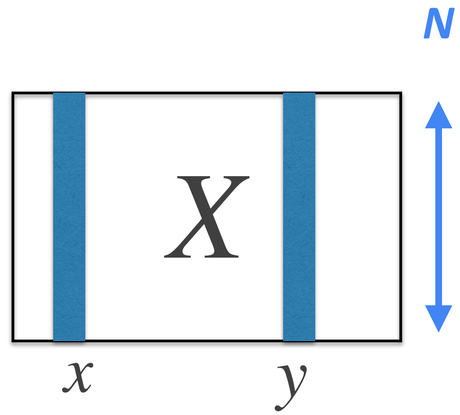
$$Var(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- Covariance

$$cov(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

- Correlation

$$cor(x, y) = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i - \bar{x})}{\sigma_x}\frac{(y_i - \bar{y})}{\sigma_y}$$



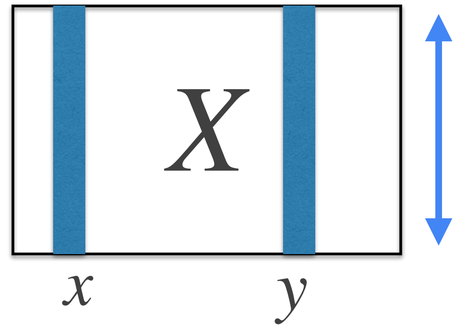
Basic central concepts

- Variance

$$Var(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \text{diag}(\frac{1}{N}X'_c \cdot X_c)$$

- Covariance

$$cov(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}X'_c \cdot X_c$$

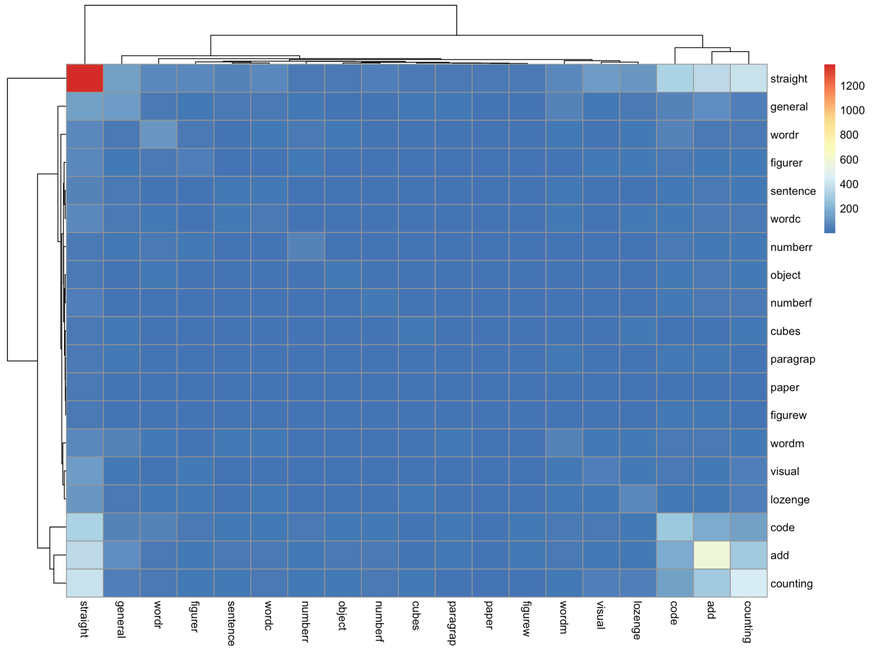


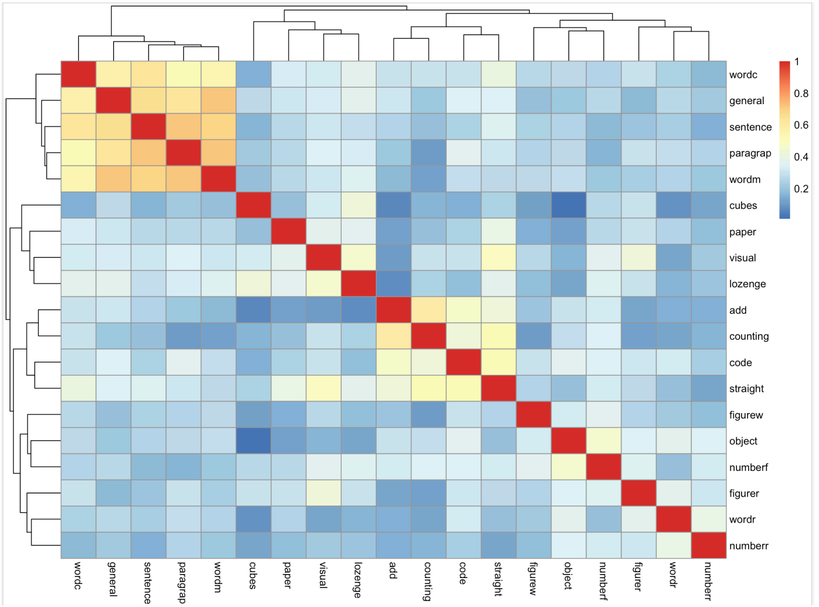
- Correlation

$$cor(x,y) = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i - \bar{x})}{\sigma_x}\frac{(y_i - \bar{y})}{\sigma_y} = \frac{1}{N}X'_{cs} \cdot X_{cs}$$

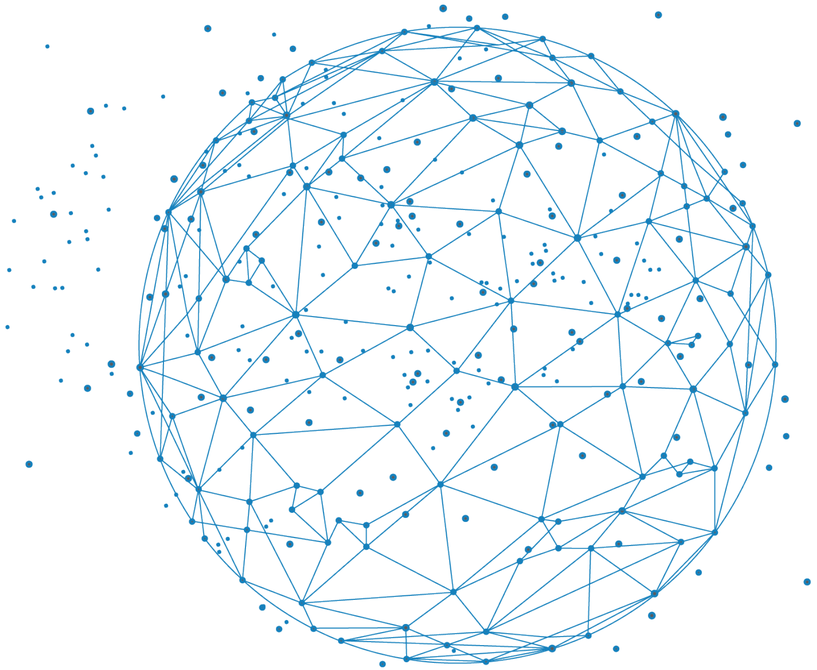
Variance-covariance and Correlation matrix

- **Variance/covariance matrix**
  - variance on the diagonal
  - covariance off-diagonal
  - symmetric matrix
- **Correlation matrix**
  - describes all pairwise correlation values
  - symmetric matrix
  - 1's in the diagonal





Multivariate analyses for multi-omics



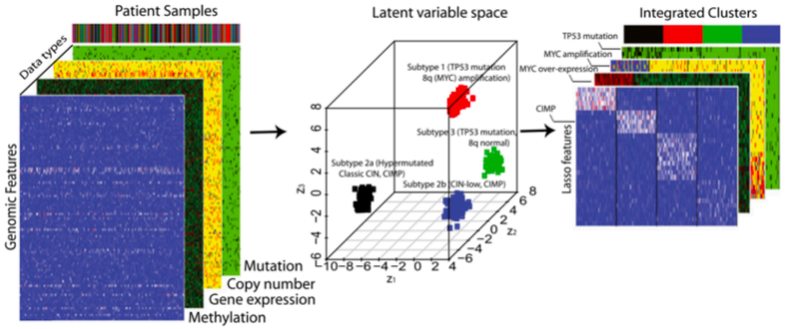
Various approaches for data reduction and integration

- **(Consensus) clustering** approaches
  - Clusters of Clusters (CoCA)
  - integrative clustering (iCluster)

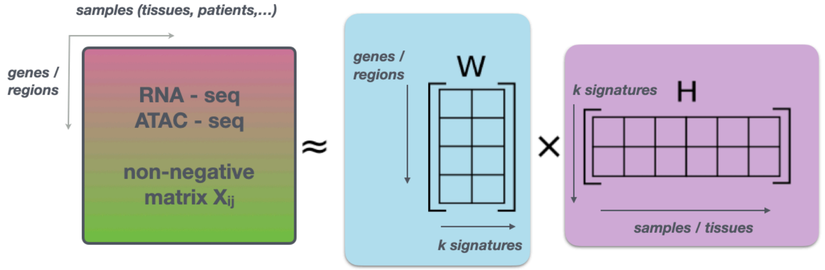
- **Linear approaches** approaches
  - Principal component analysis (PCA)
  - Non-negative matrix factorization (NMF)
  - Factor Analysis
  - Canonical correlation analysis

Matrix factorization approaches

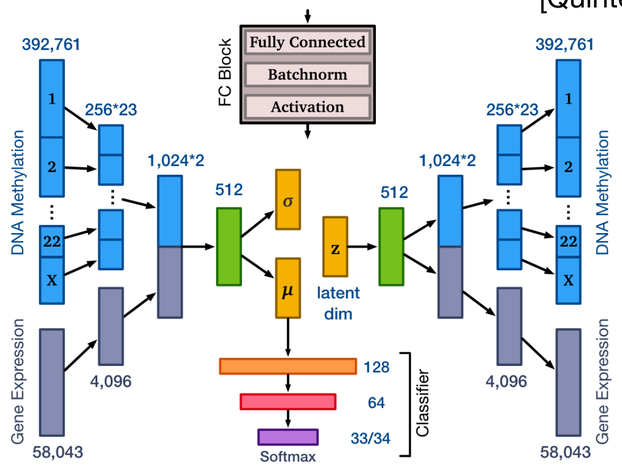
- **Neural network based** approaches
  - Autoencoders
  - Variational autoencoders



[Olshen et al., 2013]



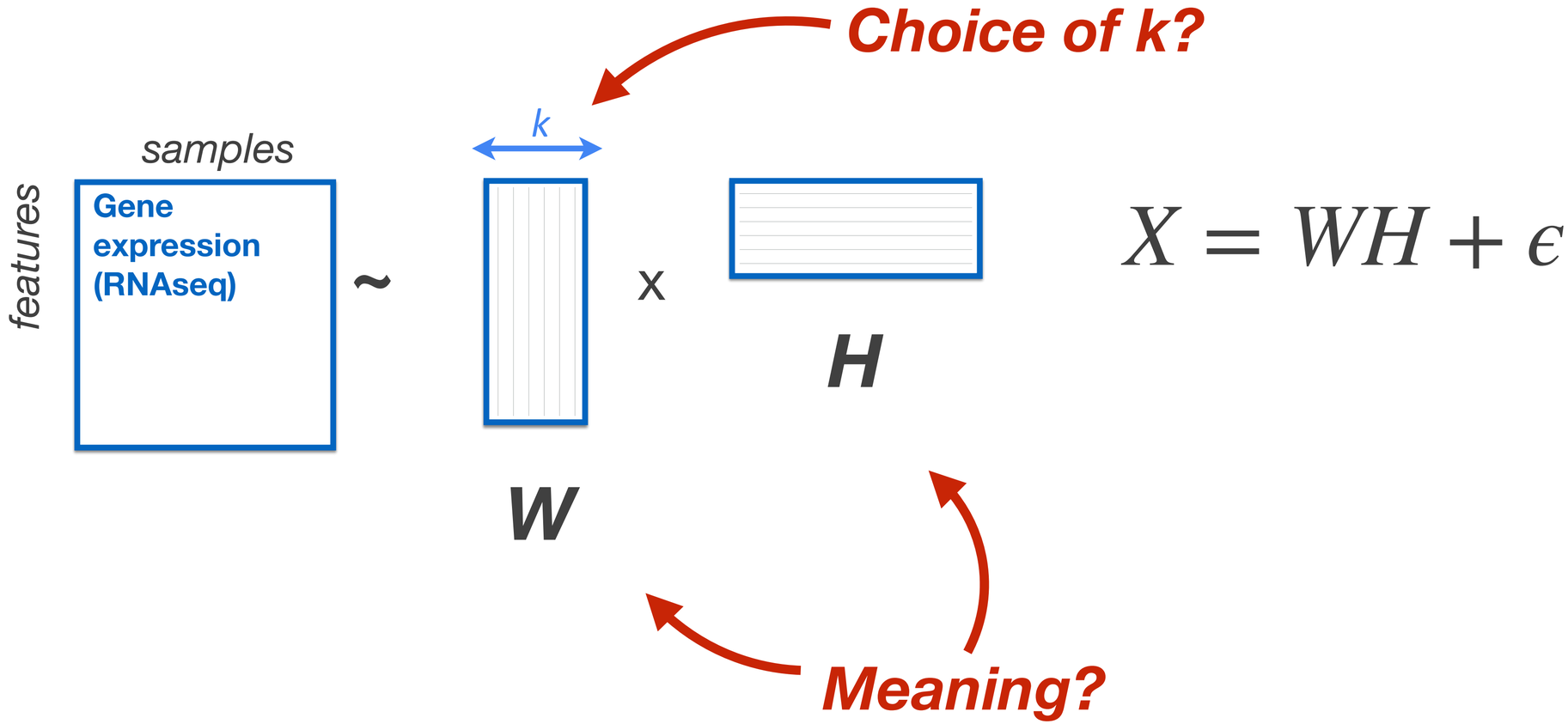
[Quintero et al., 2021]



[Zhang et al., 2019]

Matrix factorization

- approximate large data matrix using the product of 2 smaller matrices
- **columns of W = molecular signatures**

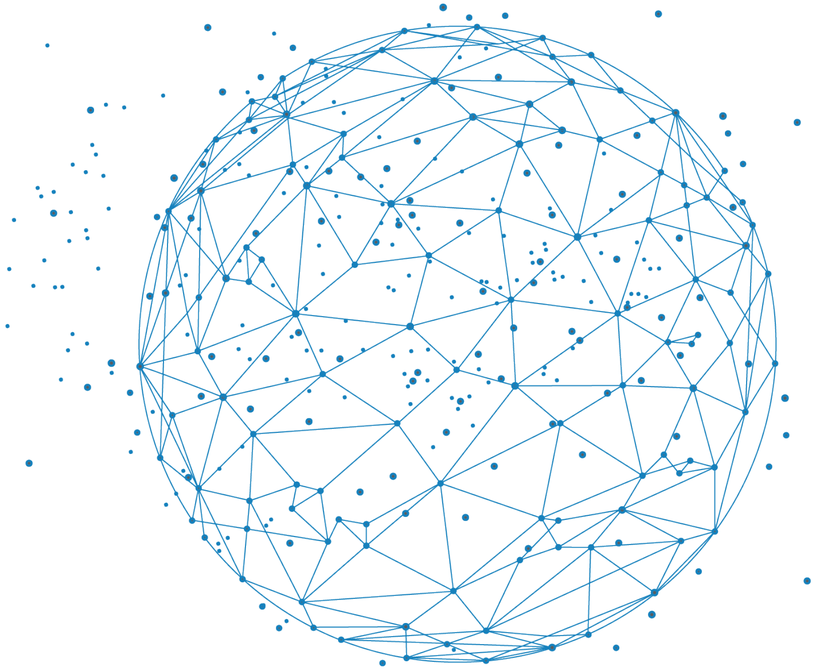


$$X = WH + \epsilon$$

Methods reviewed in this presentation

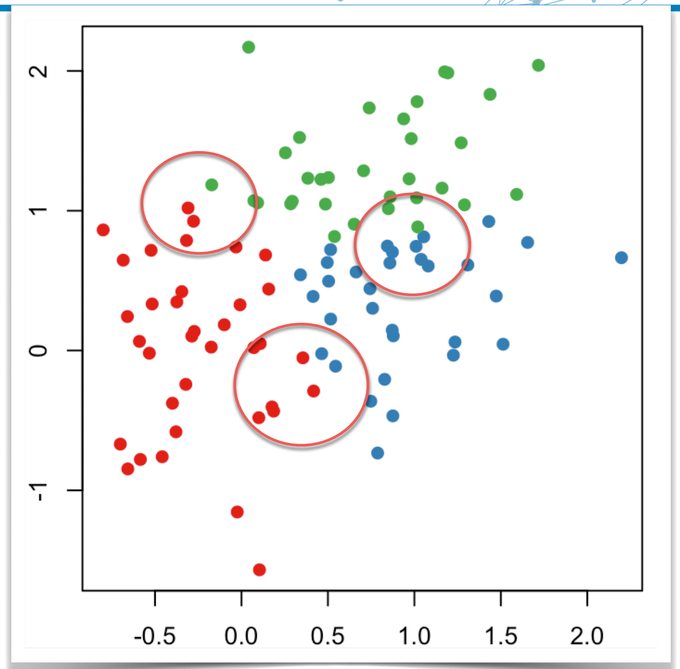
- Clustering approaches
- Principal component analysis (PCA)
- Exploratory factor analysis (EFA)
- Non-negative matrix factorization (NMF)

Clustering

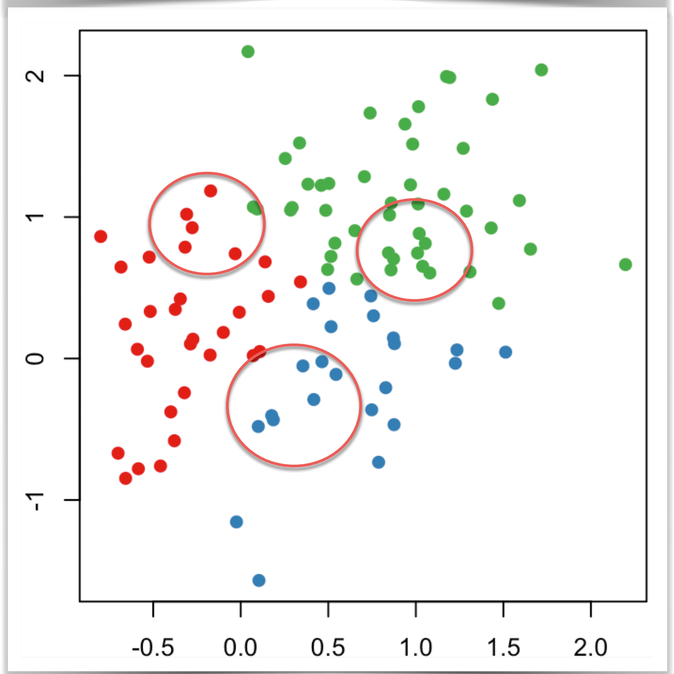


Clustering

- Clustering is the simplest **unsupervised** dimensional reduction method
  n data points → k << n clusters
- Many clustering methods:
  - k-means
  - k-medoids (PAM)
  - self-organizing maps (SOM)
  - …
- Sensitive to initialization of procedure, especially if the clusters not well separated!



k-means with k=3



k-means with k=3

Consensus clustering

- Idea of **consensus clustering**:
  if I cluster random subsamples of data points, how often will 2 points be found in the same cluster?

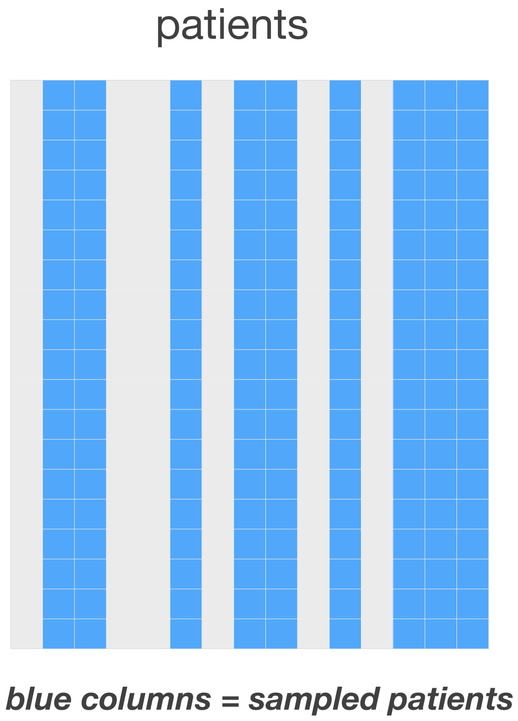
$D = \{e_1, \ldots, e_N\}$ expression profiles for N patients

$D^{(h)}$ subset of the patients (e.g. 80%)

$M^{(h)}$ result of clustering $D^{(h)}$

$M^{(h)}(i,j) = 1$ if (i,j) belong to the same cluster

$I^{(h)}(i,j) = 1$ if (i,j) both included in $D^{(h)}$



blue columns = sampled patients

$$m(i,j) = \frac{\sum_h M^{(h)}(i,j)}{\sum_h I^{(h)}(i,j)} \qquad d(i,j) = 1 - m(i,j)$$

→ Use the matrix d to perform (hierarchical) clustering

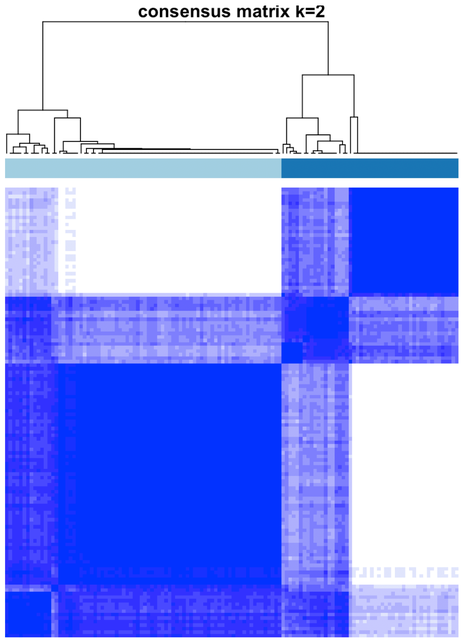
Consensus clustering

```
> results[[2]][["consensusMatrix"]][1:5,1:5]
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 1.0000000 1.0000000 0.9655172 1.0000000 1.0000000
[2,] 1.0000000 1.0000000 0.8857143 1.0000000 1.0000000
[3,] 0.9655172 0.8857143 1.0000000 0.9166667 0.8823529
[4,] 1.0000000 1.0000000 0.9166667 1.0000000 1.0000000
[5,] 1.0000000 1.0000000 0.8823529 1.0000000 1.0000000
> results[[3]][["consensusMatrix"]][1:5,1:5]
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 1.0000000 0.3548387 0.8620690 0.2413793 1.0000000
[2,] 0.3548387 1.0000000 0.1142857 1.0000000 0.4000000
[3,] 0.8620690 0.1142857 1.0000000 0.1388889 0.7941176
[4,] 0.2413793 1.0000000 0.1388889 1.0000000 0.3513514
[5,] 1.0000000 0.4000000 0.7941176 0.3513514 1.0000000
```

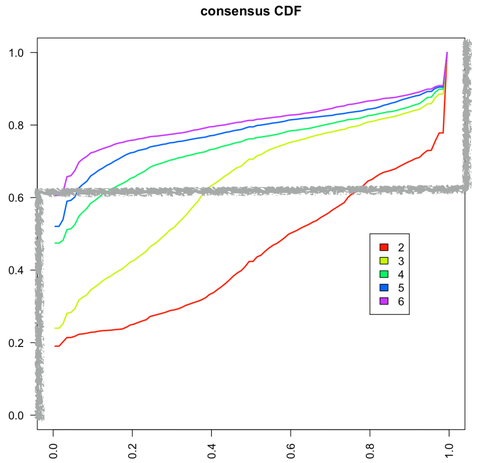
similarity matrix for $k = 2$

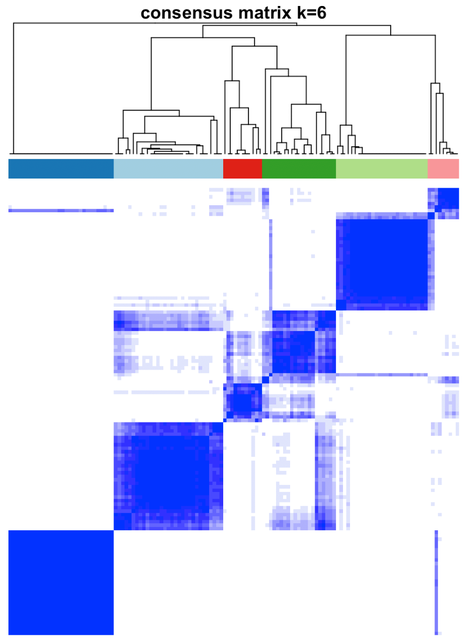
similarity matrix for $k = 3$

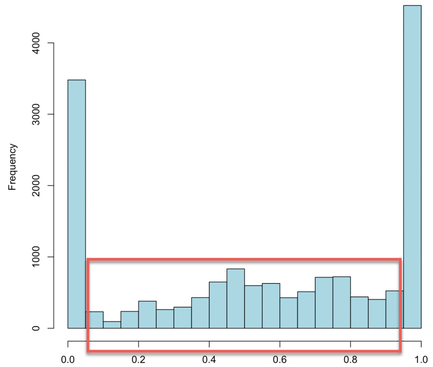
Consensus Clustering



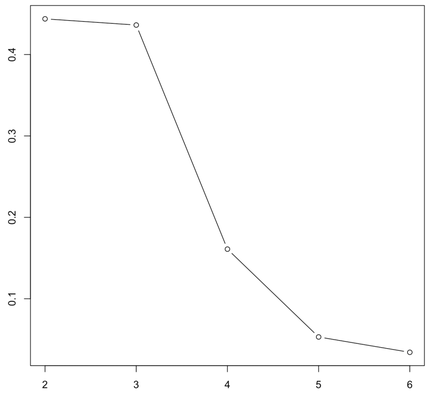
Ideal shape would
be a step function

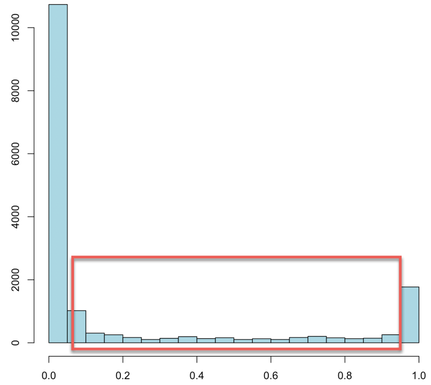






Differential AUC





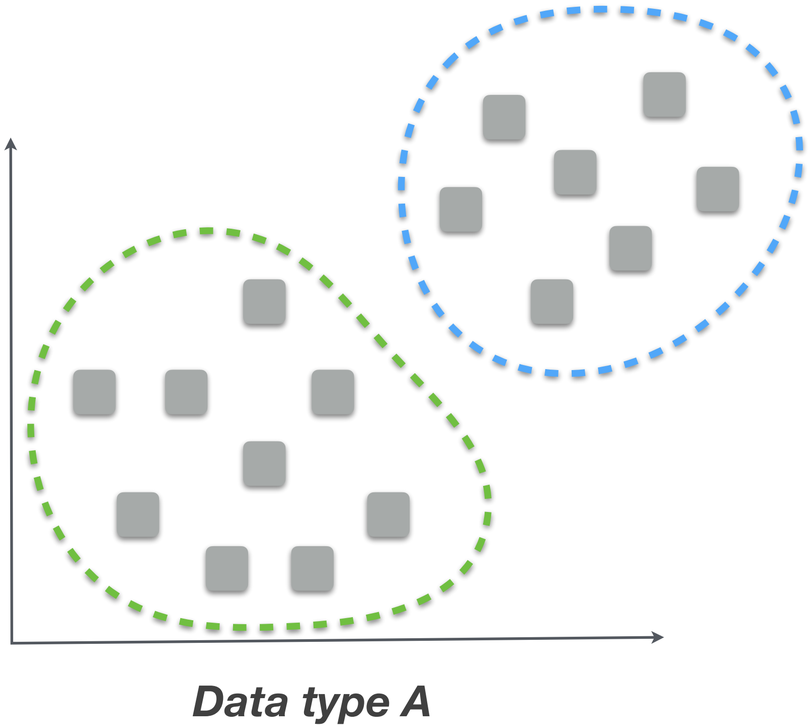
badly assigned samples

Optimal K when AUC no longer increases

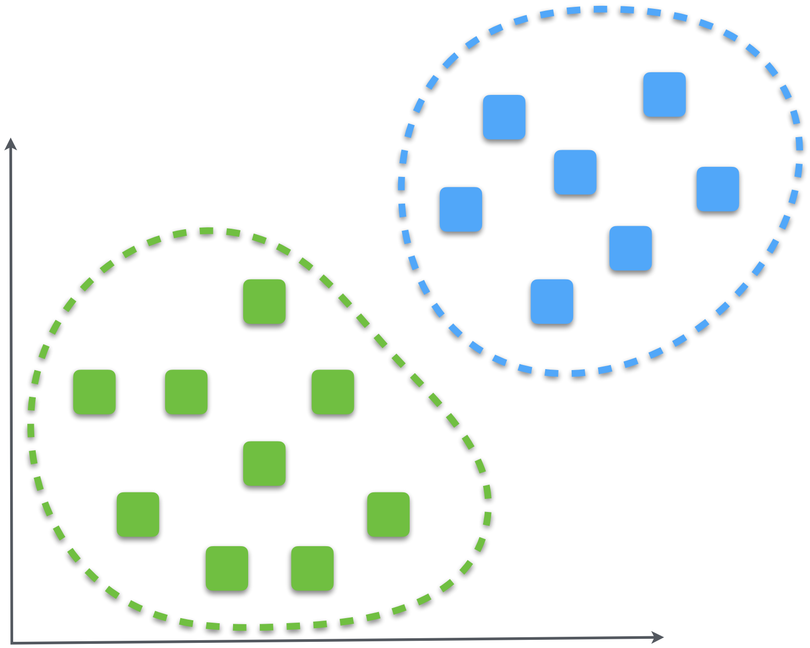
badly assigned samples

[Monti et al., 2003]

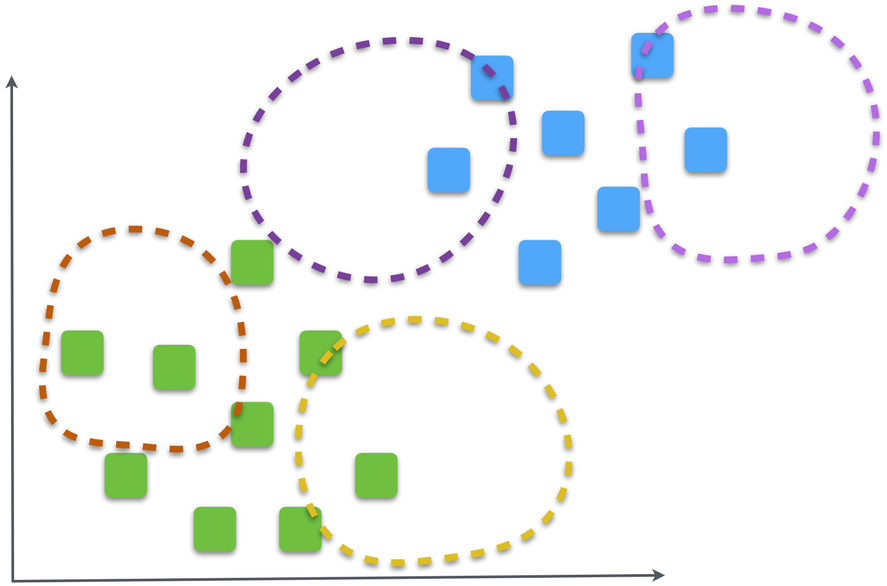
Clustering over multiple data?



Clustering over multiple data?



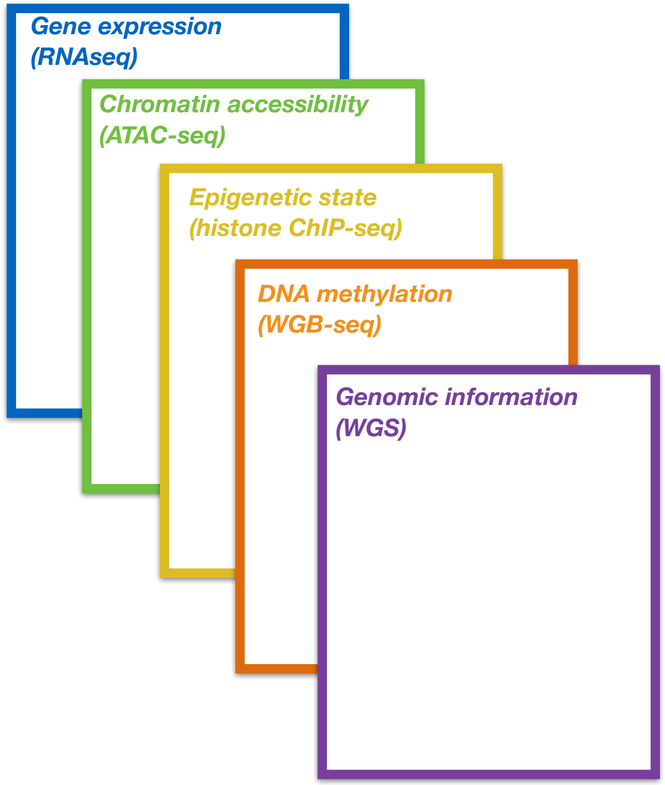
Data type A



Data type B

Cluster of Cluster Analysis (CoCA)

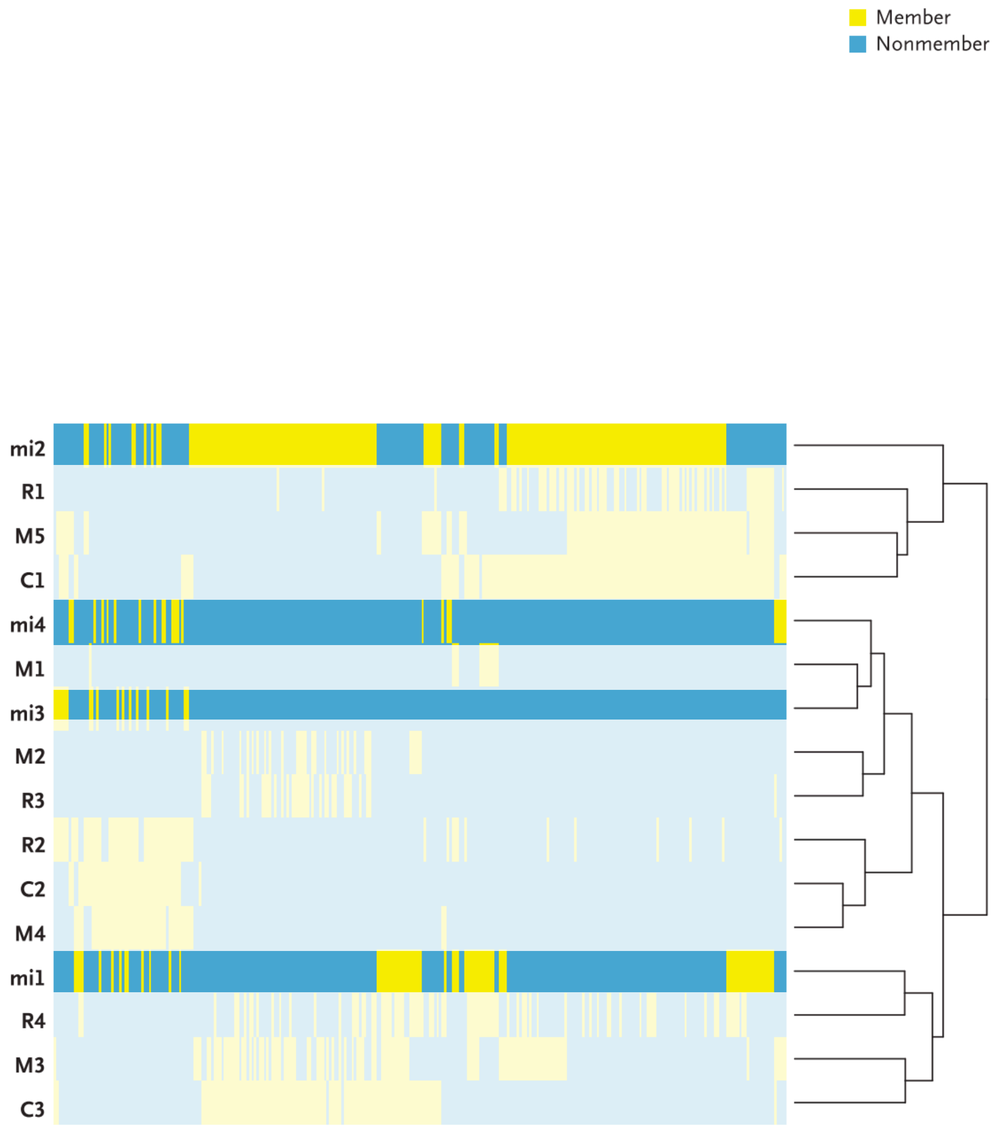
- Cluster each omics data separately
  - each clustering can use a different clustering algorithm (k-means, PAM,…)
  - each omics datatype can lead to distinct number of clusters
- Represent each sample by an **indicator vector** showing to which cluster it belongs in each omic
  s_3 = (**1** , **3**, **2** , **3** , **1**)
- Cluster the samples based on this indicator vector using **consensus clustering**



→ ***late integration***

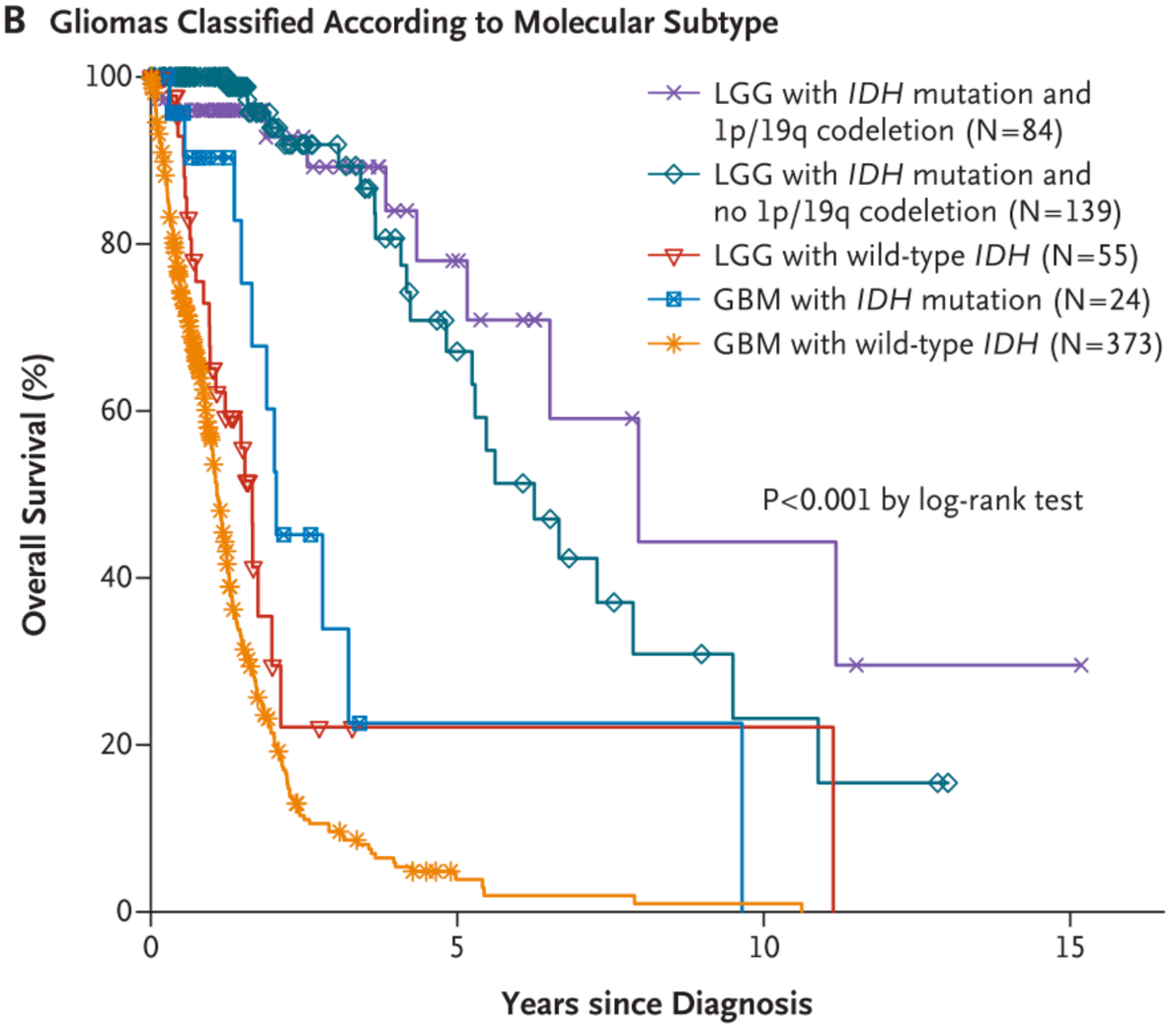
Application: low-grade glioma

- TCGA: integrative clustering of low-grade glioma (brain tumor)
- Available data *(n=293)*:
  - mRNA expression (R)
  - micro-RNA expression (mi)
  - Copy-number variation (C)
  - DNA-methylation (M)
- Result: 3 robust subtypes which disagree with histological subtypes!



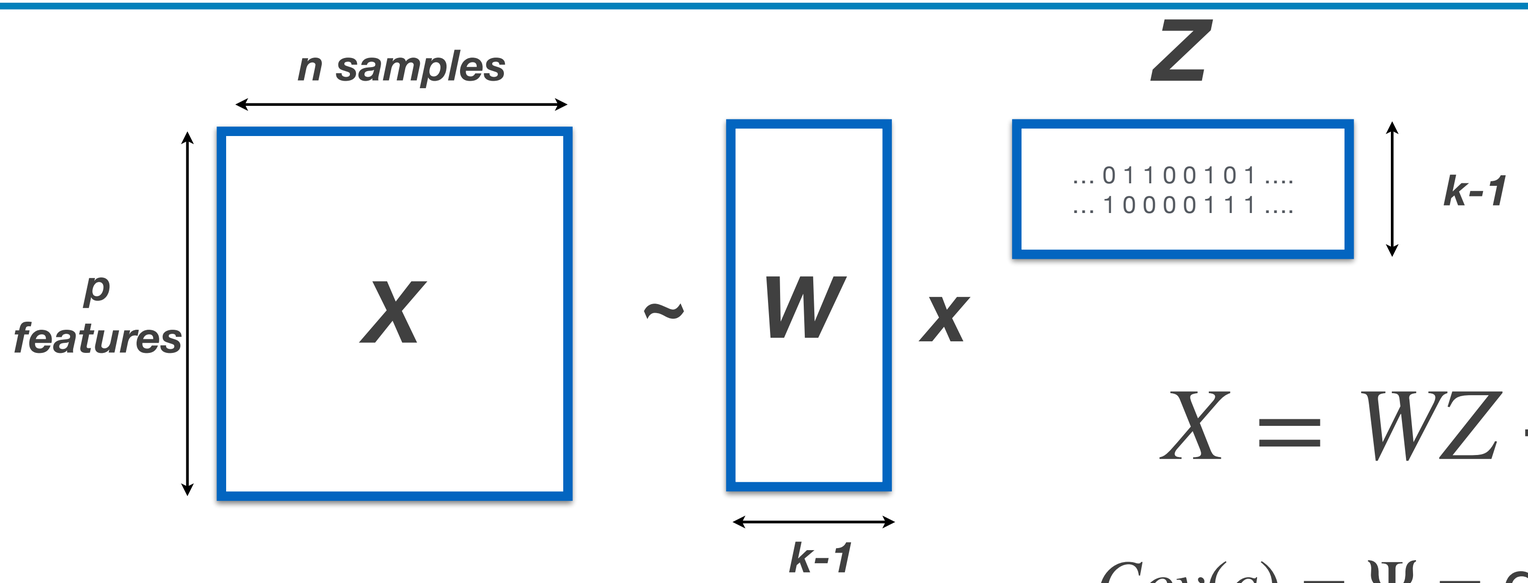
[Brat et al., NJEM, 2015]

Application: low-grade glioma



[Brat et al., NJEM, 2015]

iCluster



$$X = WZ + \epsilon$$

$$Cov(\epsilon) = \Psi = \text{diag}(\psi_1, \psi_2, \ldots, \psi_p)$$

- Goal: identify *k* clusters of samples in the dataset (i.e. Z) such that the inter-cluster distance is maximized
- *Z* is the **indicator function**
  - z_{ij} = 1 : sample j belongs to cluster i
  - z_{ij} = 0 : sample j does not belong to cluster i

[Shen et al., Bioinformatics 2010]

iCluster

$$X = WZ + \epsilon$$

$$Cov(\epsilon) = \Psi = \text{diag}(\psi_1, \psi_2, \ldots, \psi_p)$$

- ***X*** is **observed**
- ***W*** and Ψ are unknown parameters (these are numbers!)
- ***Z*** is the unknown **latent variable** (this is a random variable!)
- Bayesian formulation: binary $Z \rightarrow$ continuous Z^*
- Prior distribution : $Z^* \sim \mathcal{N}(0, I)$
- Goal: maximize posterior probability $E[Z^* \mid X]$

iCluster

$$X = WZ + \epsilon \qquad \text{Cov}(\epsilon) = \Psi = \text{diag}(\psi_1, \psi_2, \ldots, \psi_p)$$

- Find optimal solution using **Expectation-Maximization**

Initial random values for $(W^{(0)}, \Psi^{(0)})$

(Expectation Step)

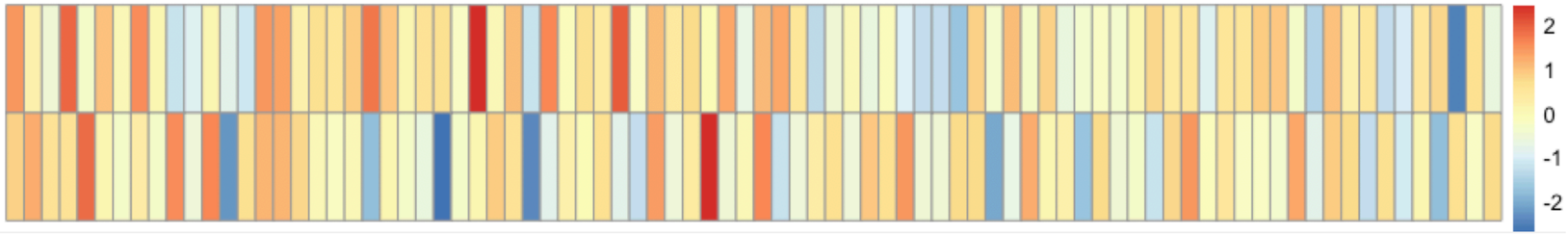
Estimate $Z^{(t)}$
using $(W^{(t-1)}, \Psi^{(t-1)}, X)$

(Maximization Step)

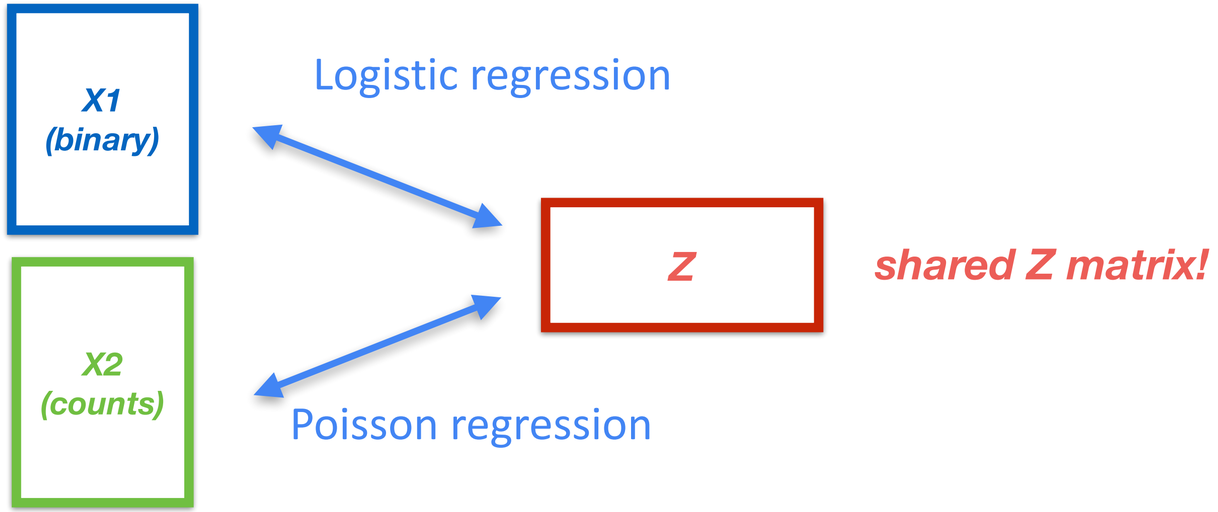
Estimate $(W^{(t+1)}, \Psi^{(t+1)})$
using $(Z^{(t)}, X)$

iCluster

Infered posterior probability E[Z|X] (for k = 2)*



Cluster indicator for k = 2 clusters



- Different types of data (binary, count data, continuous data,…) can be taken into account using different conditional probabilities
  - X_i is binary: **logistic** regression

$$\log \frac{P\left(x_{ijt}=1 \middle| \mathbf{z}_i\right)}{1-P\left(x_{ijt}=1 \middle| \mathbf{z}_i\right)} = \alpha_{jt} + \boldsymbol{\beta}_{jt}\mathbf{z}_i$$

i = sample, j = feature, t = view)

  - X_i is count data: **Poisson** regression

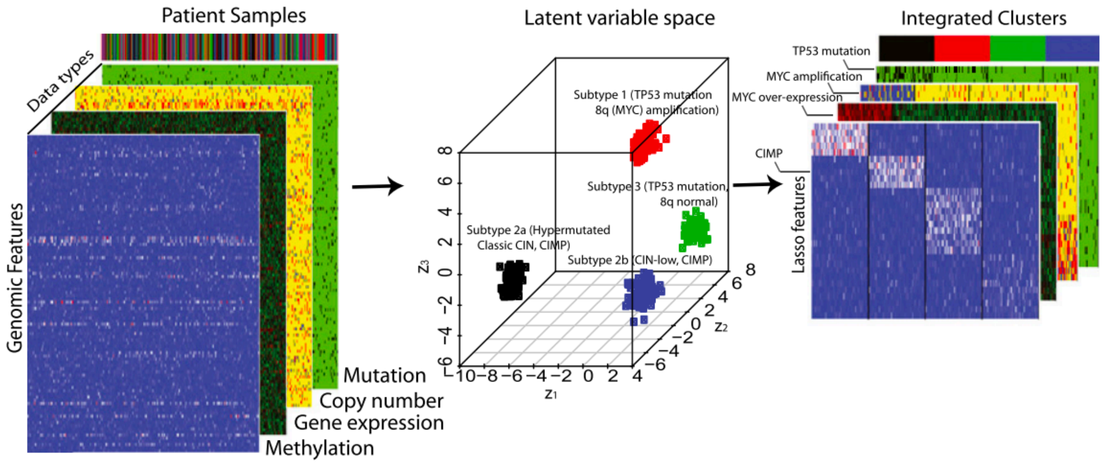
$$\log\left(\lambda\left(x_{ijt} \middle| \mathbf{z}_i\right)\right) = \alpha_{jt} + \boldsymbol{\beta}_{jt}\mathbf{z}_i$$

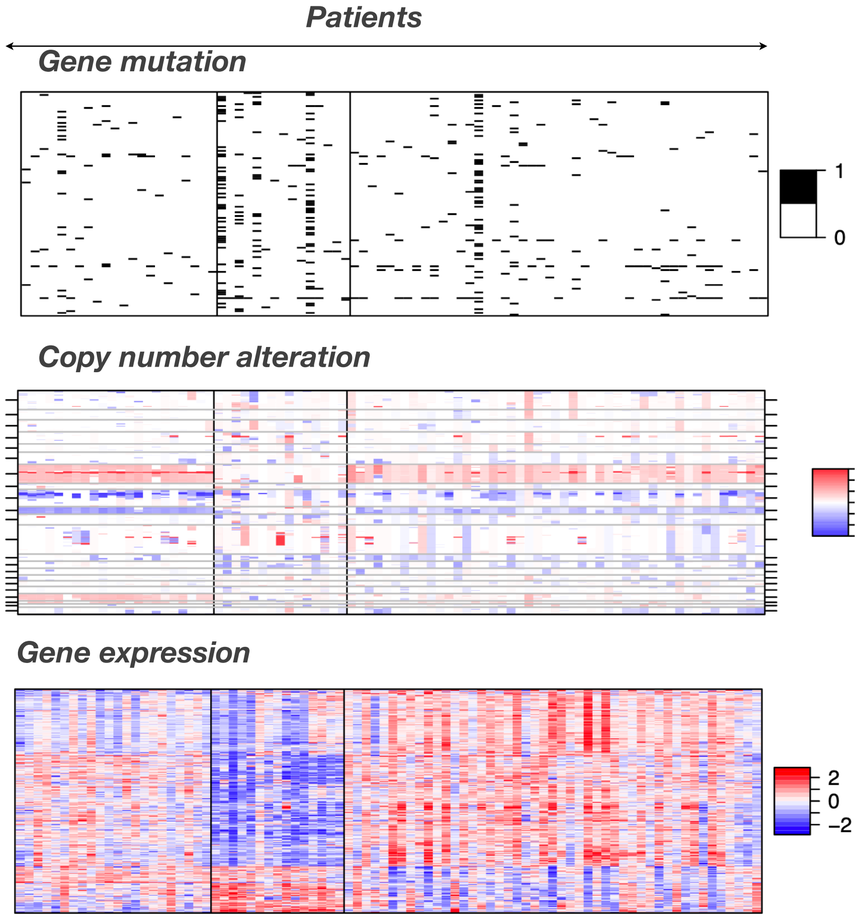
  - X_i is continuous: **linear** regression

$$x_{ijt} = \alpha_{jt} + \boldsymbol{\beta}_{jt}\mathbf{z}_i + \varepsilon_{ijt}$$

[Mo et al., PNAS 2013]

iCluster+

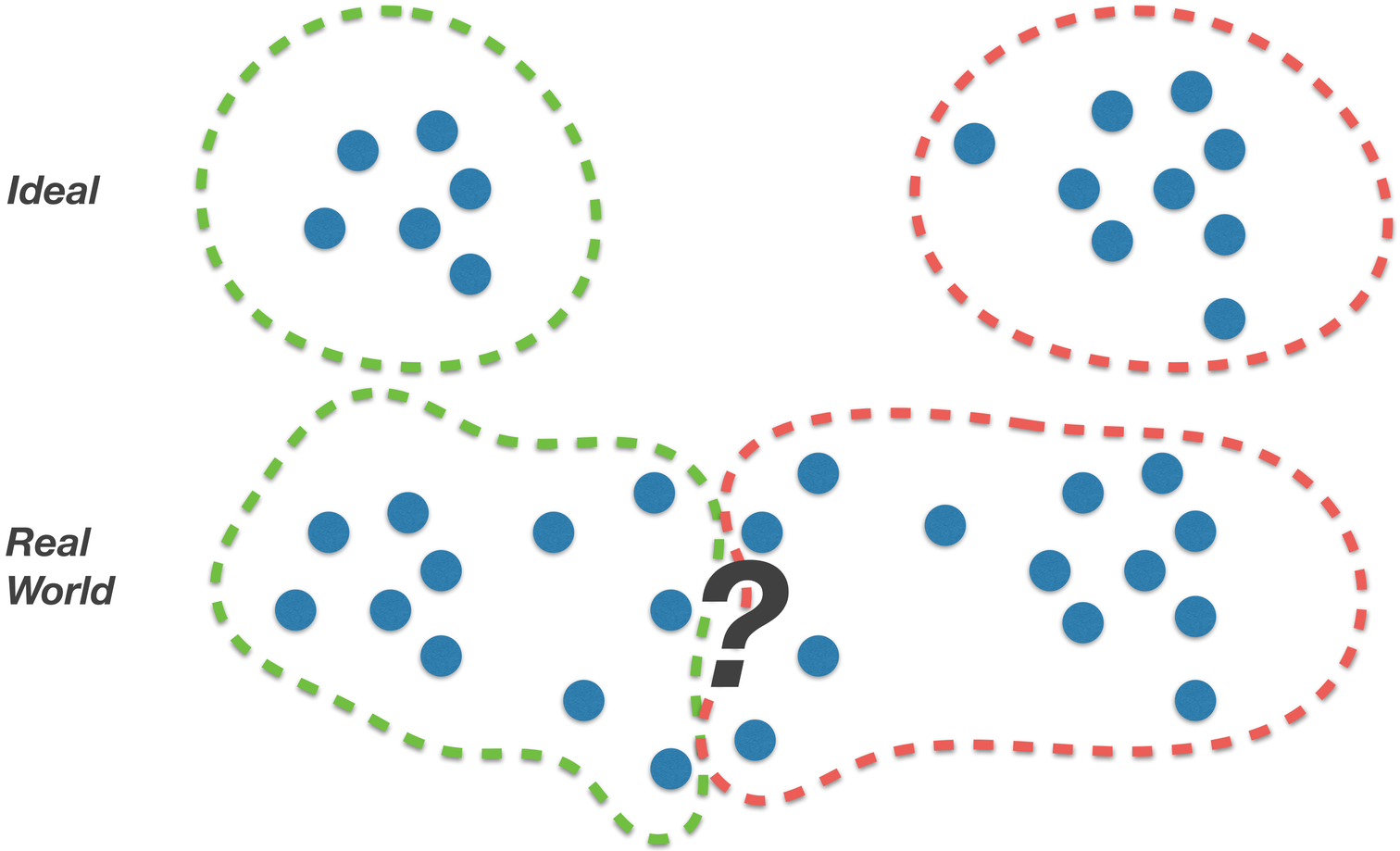




- Application: **TCGA glioblastoma** datasets
  - **gene mutations**
    (120 genes x 84 patients)
  - **copy-number alterations**
    (5512 regions x 84 patients)
  - **gene expression**
    (1740 top variable genes x 84 patients)

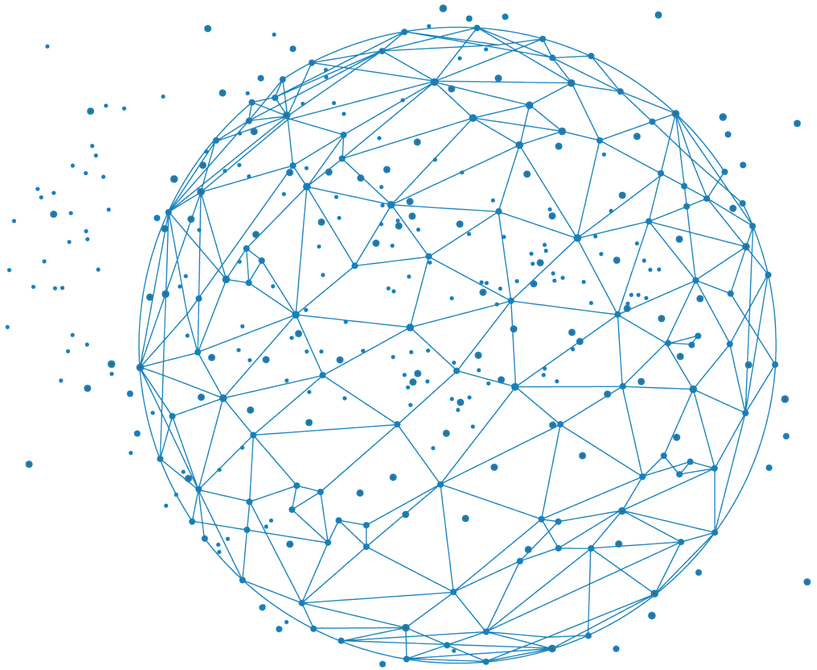
[Mo et al., PNAS 2013]

Limitations of Clustering



We needs methods allowing a "fuzzy"assignment of samples clusters → signatures

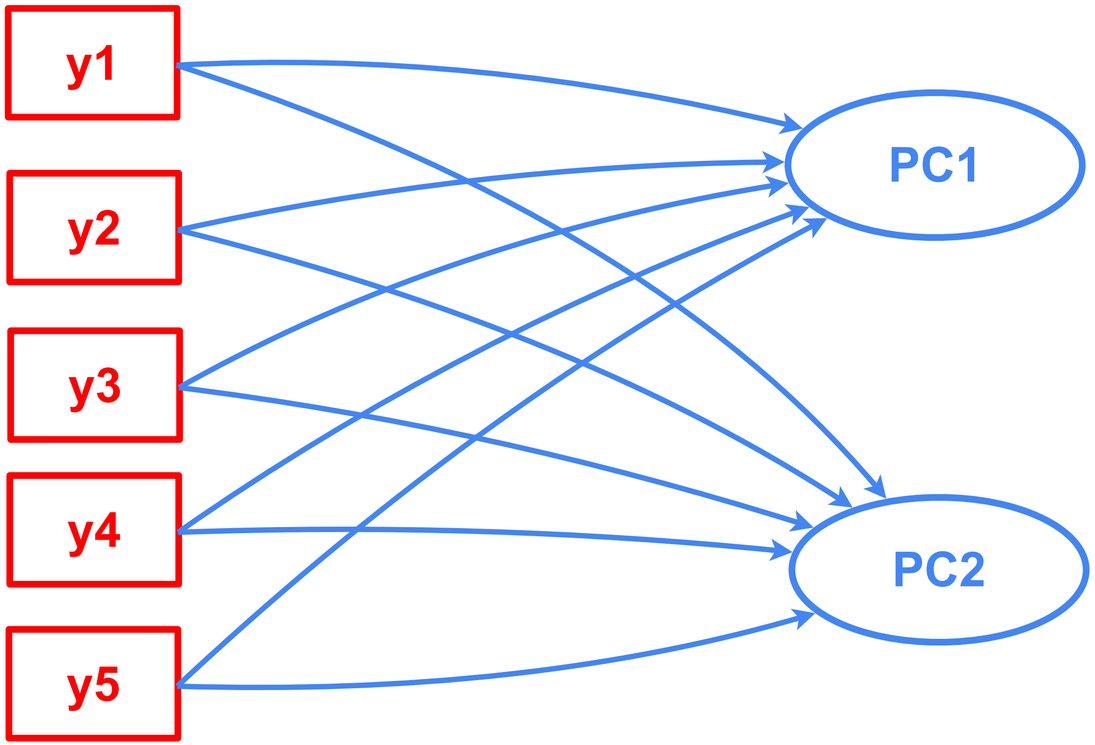
Principal Component Analysis (PCA)

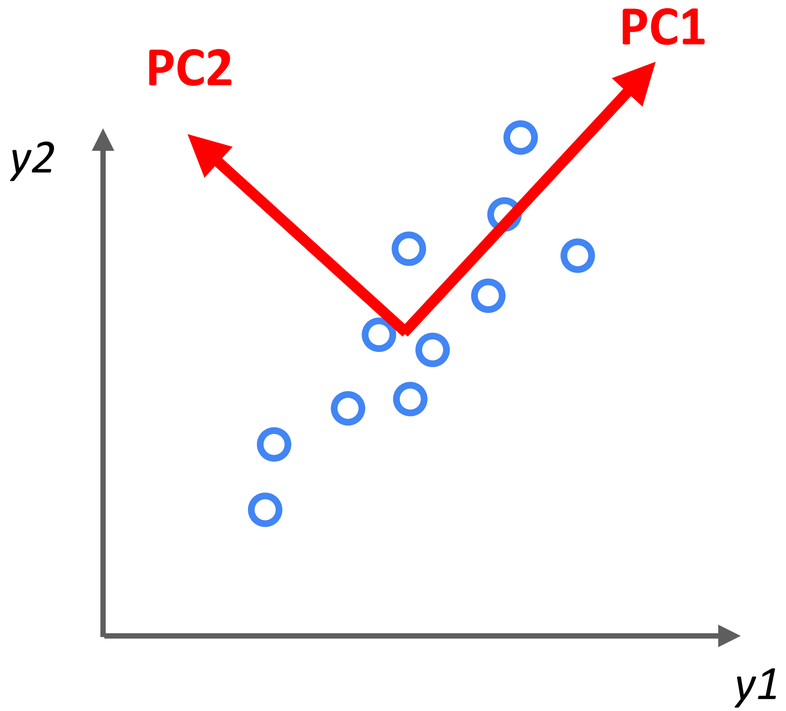


Graphical abstract

Observed variables
(e.g. genes)

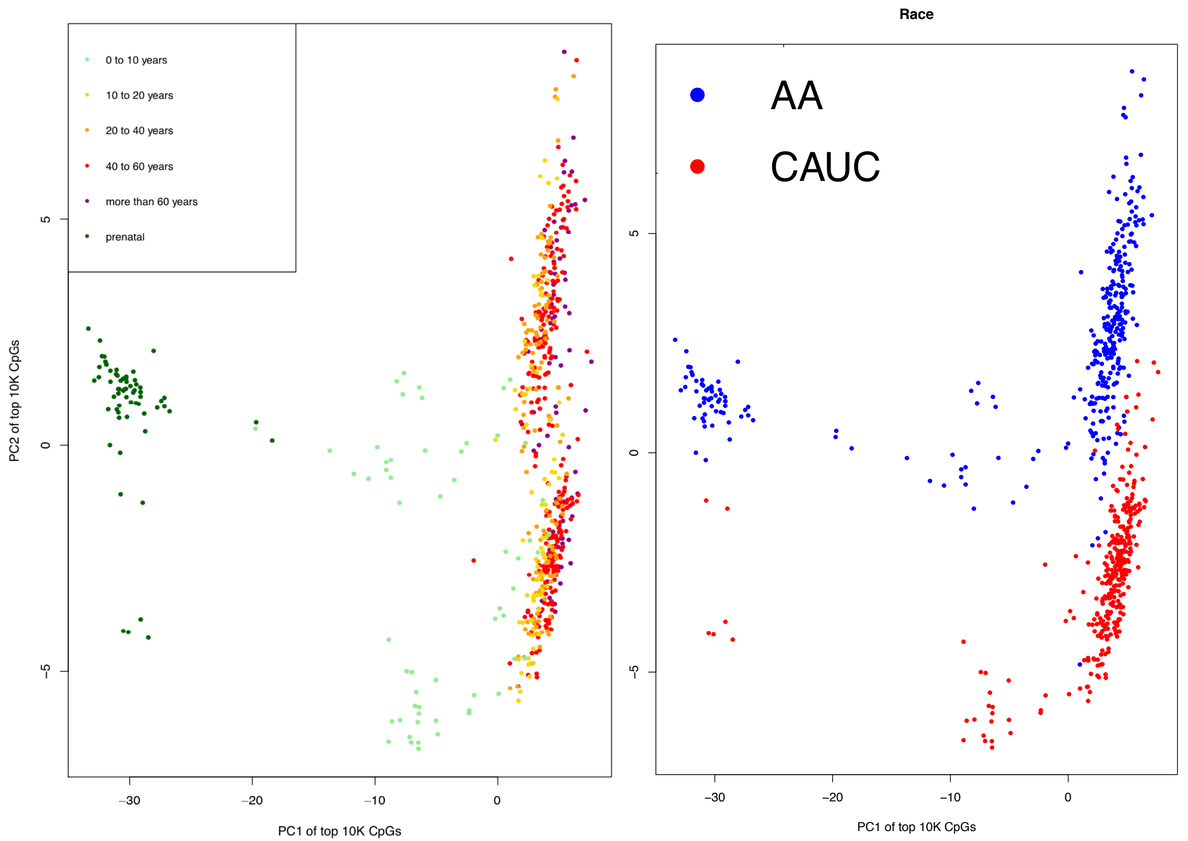
Principal components
("metagenes")





Principal component analysis

- Dataset have a very **high dimensionality** (e.g. number of genes)
- Need to reduce this large number of dimensions to a **smaller number of relevant variables**
- Relevant variables = variables which carry **most of the information** (or variance) of a dataset
- These new variables are **orthogonal**
- Goal: identify **directions** in the data corresponding to **biological effects**



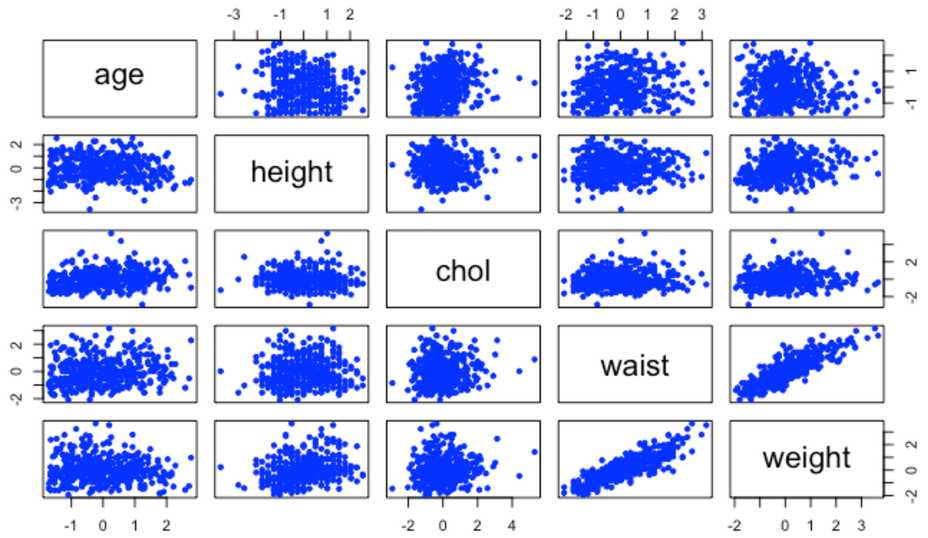
Example of DNA methylation of blood samples in patient cohort (Jana Dalhoff)
data matrix : 400.000 CpG positions / 250 patients

Correlation structure

- if two variables are **strongly correlated**, they are partly redundant: knowing the variation of one, you have information about how the second variables changes

- if two variables have **little correlation**, each variable carries information not contained in the other

- **The more diagonal a correlation matrix is, the more information is revealed by the variables**

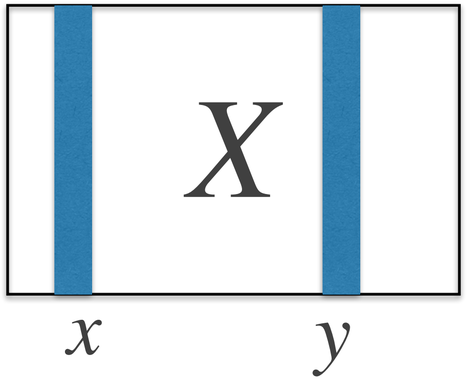


Correlation and covariance

$$cov(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}X'_c \cdot X_c$$

$$cor(x,y) = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i - \bar{x})}{\sigma_x}\frac{(y_i - \bar{y})}{\sigma_y} = \frac{1}{N}X'_{cs} \cdot X_{cs}$$

Z-transformation



Computing PCA

1. Consider the **correlation matrix** *A*

	age	height	chol	waist	weight
age	1.00000000	-0.09479919	0.23990232	0.15255761	-0.06269027
height	-0.09479919	1.00000000	-0.05853973	0.05661532	0.25298143
chol	0.23990232	-0.05853973	1.00000000	0.11245805	0.05932074
waist	0.15255761	0.05661532	0.11245805	1.00000000	0.84955930
weight	-0.06269027	0.25298143	0.05932074	0.84955930	1.00000000

2. Determine its *n* eigenvalues and *n* eigenvectors and build the n x n matrix *V* from all the n eigenvectors as columns

```
$values
[1] 1.9201374 1.3081302 0.9011191 0.7635241 0.1070892

$vectors
           [,1]       [,2]       [,3]       [,4]       [,5]
[1,] -0.0782536  0.66340112 -0.1957637  0.70124582 -0.15396822
[2,] -0.2139712 -0.41884235 -0.8557896  0.16410053  0.13957979
[3,] -0.1338086  0.59835882 -0.3893703 -0.68726253  0.01108988
[4,] -0.6768556  0.08069999  0.2542954  0.06898591  0.68258976
[5,] -0.6870622 -0.14115337  0.1141285 -0.06508820 -0.70054229
```

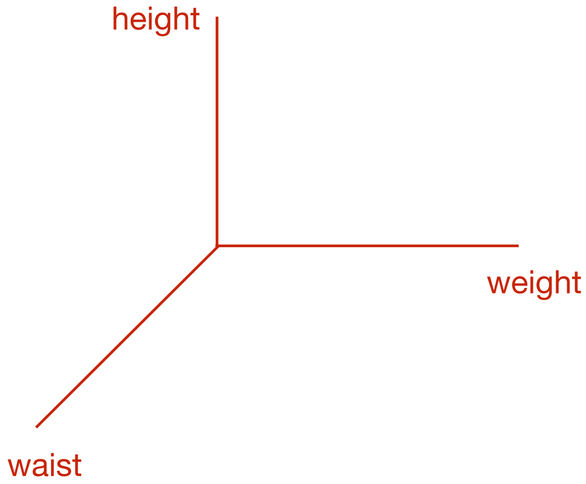
3. Compute $S = V' \cdot A \cdot V$

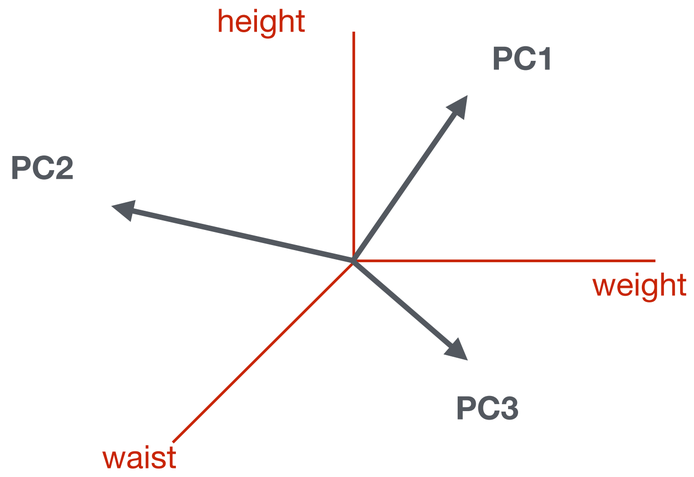
diagonal matrix! — transposed *V* matrix

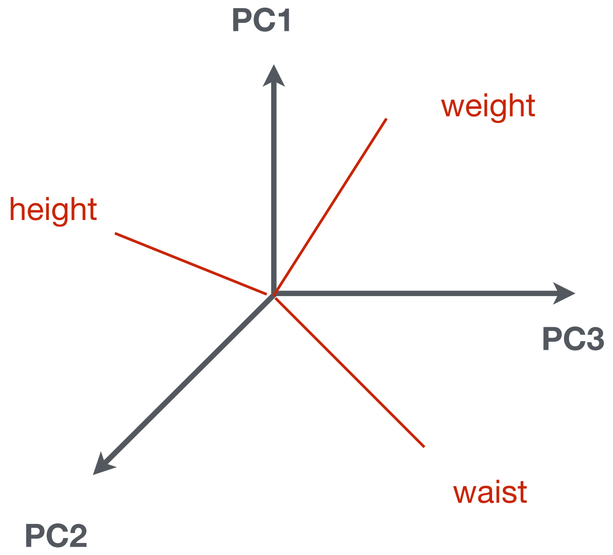
```
     [,1]  [,2]  [,3]  [,4]  [,5]
[1,] 1.92 0.000 0.000 0.000 0.000
[2,] 0.00 1.308 0.000 0.000 0.000
[3,] 0.00 0.000 0.901 0.000 0.000
[4,] 0.00 0.000 0.000 0.764 0.000
[5,] 0.00 0.000 0.000 0.000 0.107
```

New coordinate system

- *V* is the **rotation matrix** transforming the initial variables into new variables called **principal components**





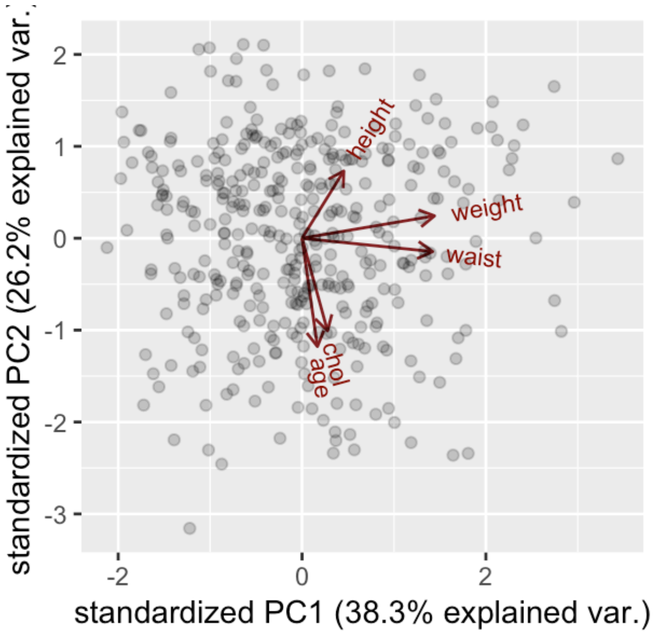


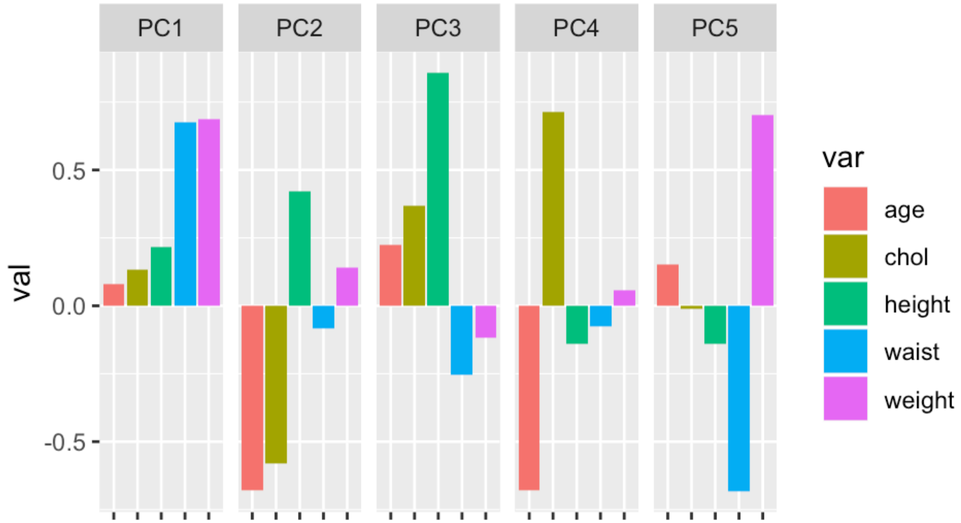
PCA biplot

- each **dot** is a sample / patient
- new coordinate system is (PC1,PC2)
- **Red arrows** indicate the contribution of each “old" coordinate to the PCs

Principal components

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

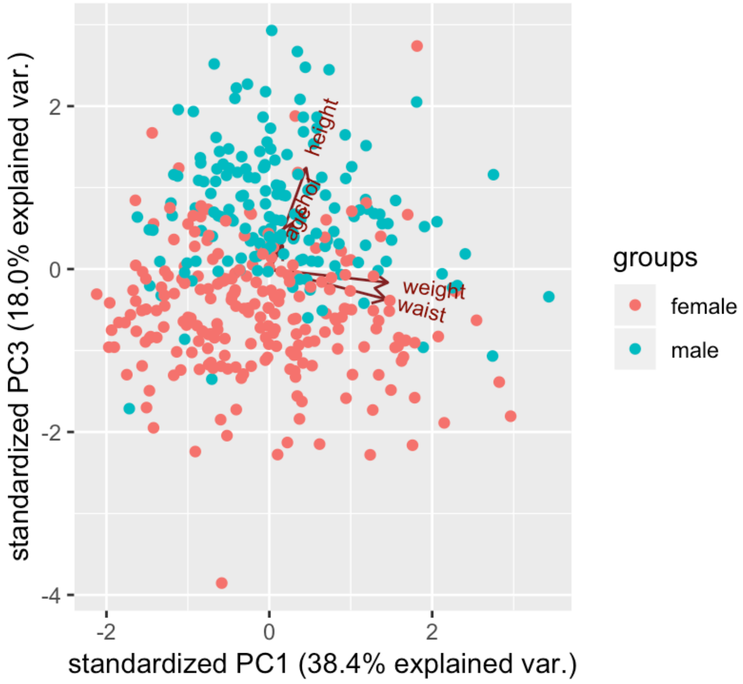




- contribution of each variable to the principal components (coefficients are called *"**loadings**"*)
- some variables contribute in the same direction to some PCs (e.g. waist and height for PC1), but opposite to others (PC5)

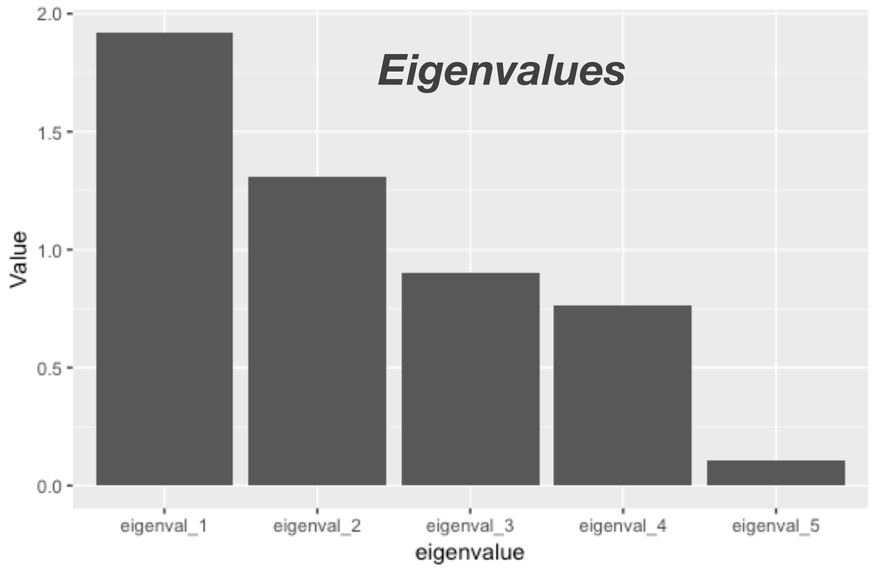
Identifying interesting PCs

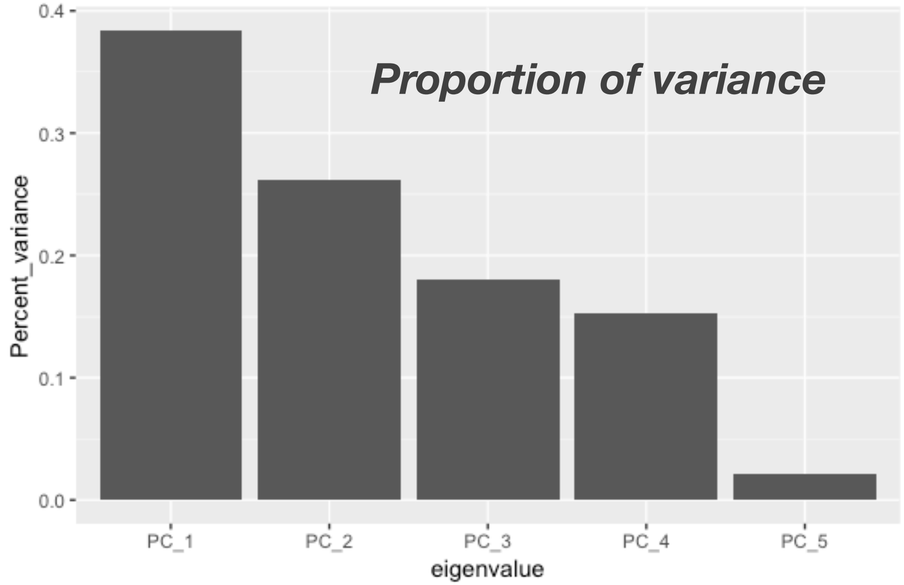
- PC plots can highlight a new group structure
- Example: **PC3** seems very associated to gender
- indicates that a combination of height and cholesterol does separate men /women



Number of PCs?

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





Considering PC1 & PC2 explains 63% of the total vairance

Choosing the number of PCs

- several criteria to select the optimal subset of PCs, without loosing too much information

- **Proportion of variance:**
  keep PCs such that the cumulative variance is above threshold

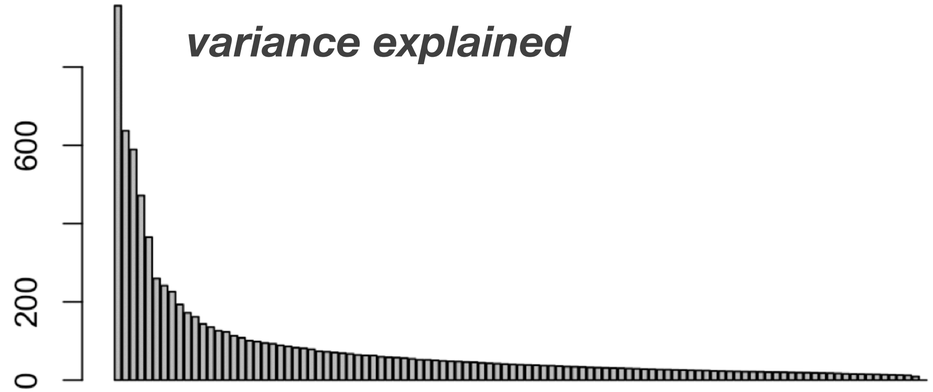
$$\sum_{i=1}^{k} \frac{\lambda_i}{\sum \lambda_i} \geq \text{var}_{min}$$

- **Average eigenvalue criteria:**
  keep PCs which have eigenvalue larger than
  - mean eigenvalue (Kaiser rule) or
  - 70% of mean eigenvalue (Jottclife rule)

$$\lambda_i \geq \bar{\lambda}$$

Application to gene expression

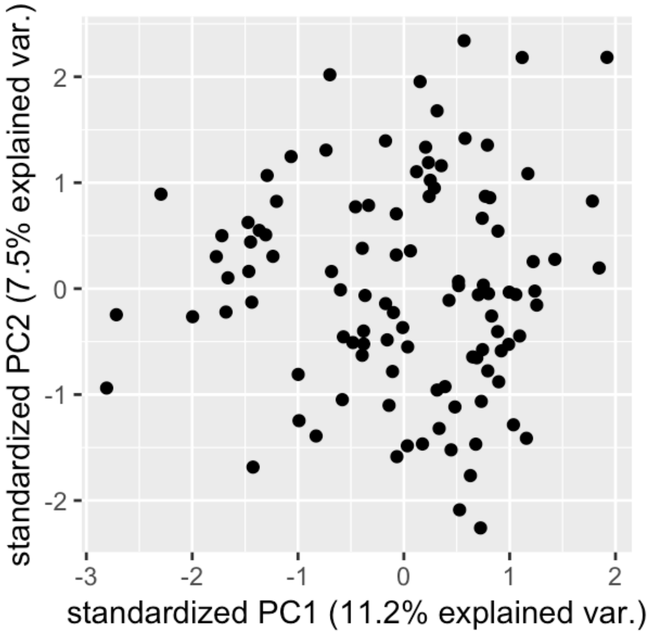
- Gene expression dataset of **breast cancer patients**
- 2 groups: ER+ and ER- patients
- Dimension: k = 105 patients / n = 8534 genes
  (here: n >> k)
- pre-processing:
  - **scale** the gene expression across patients
  - **center** the gene expression across patients
- How many principal components do we get?
  → **k** (this has to do with the rank of the data matrix)

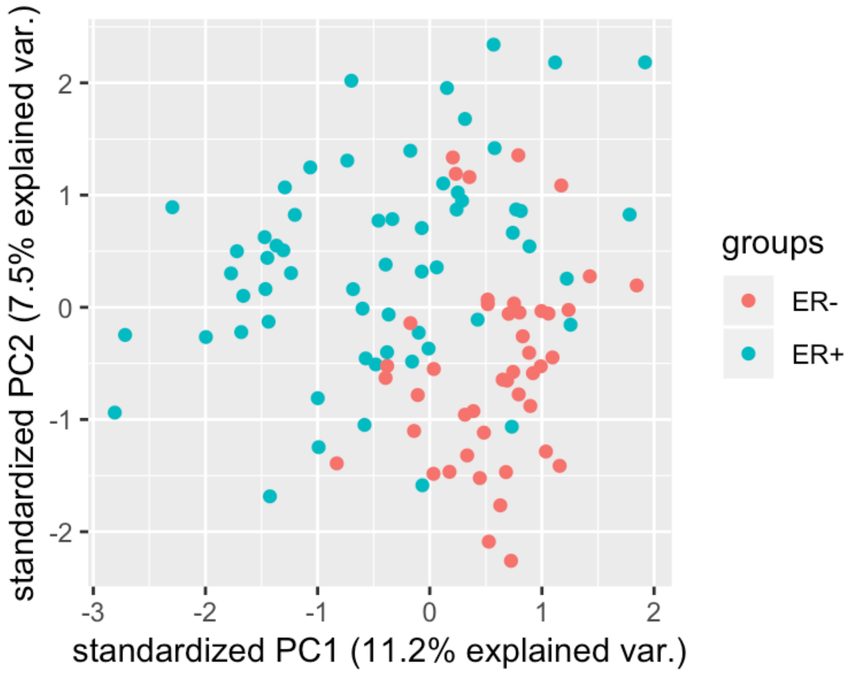


105 principal components

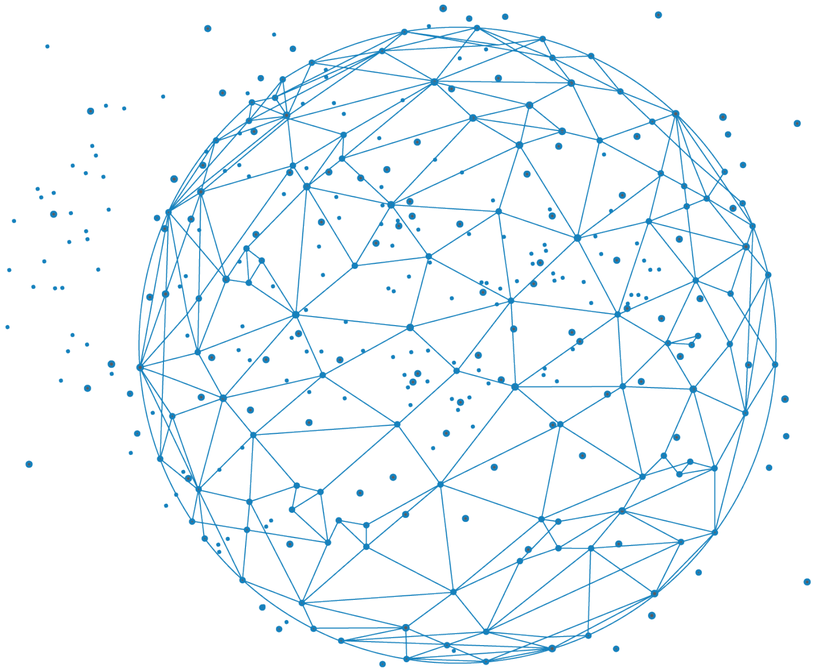
Application to gene expression

- PC1 separates ER+ from ER- patients

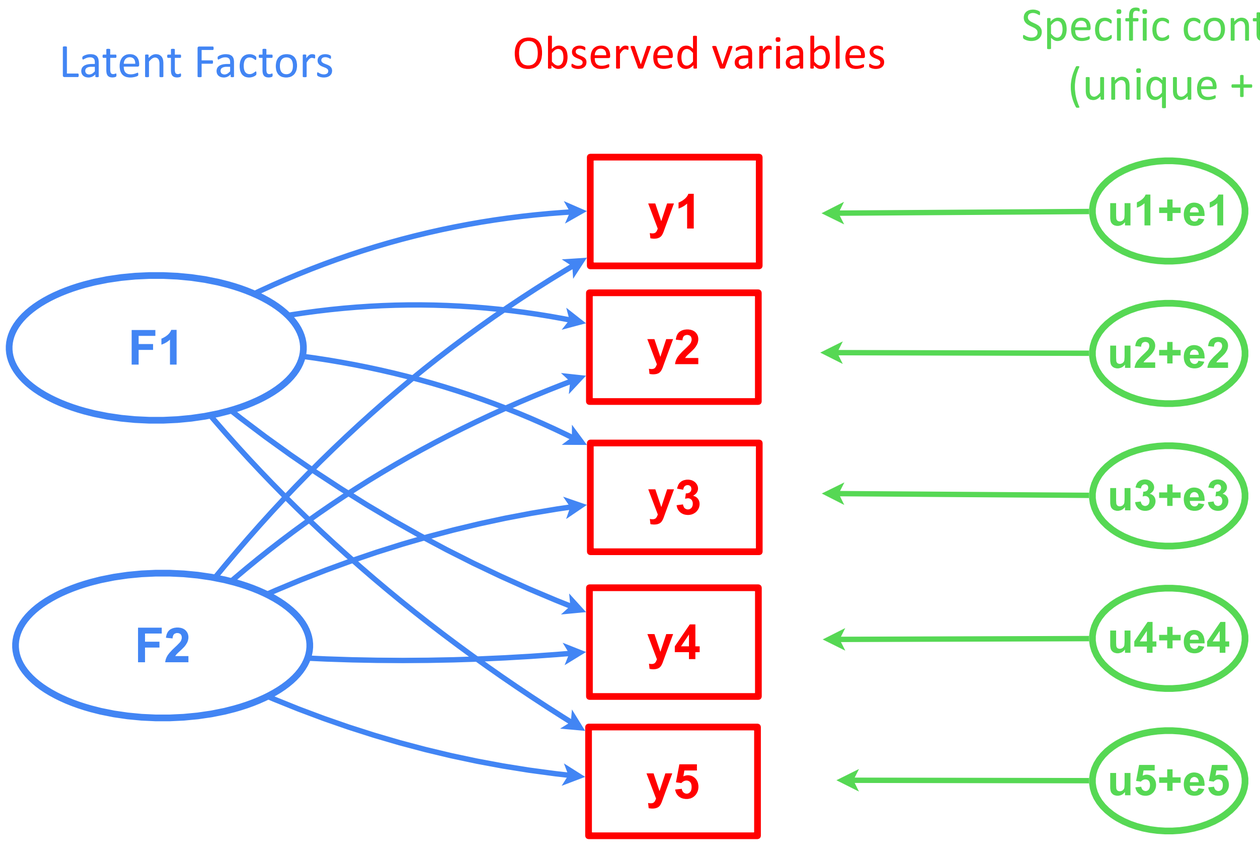




Exploratory Factor Analysis (EFA)

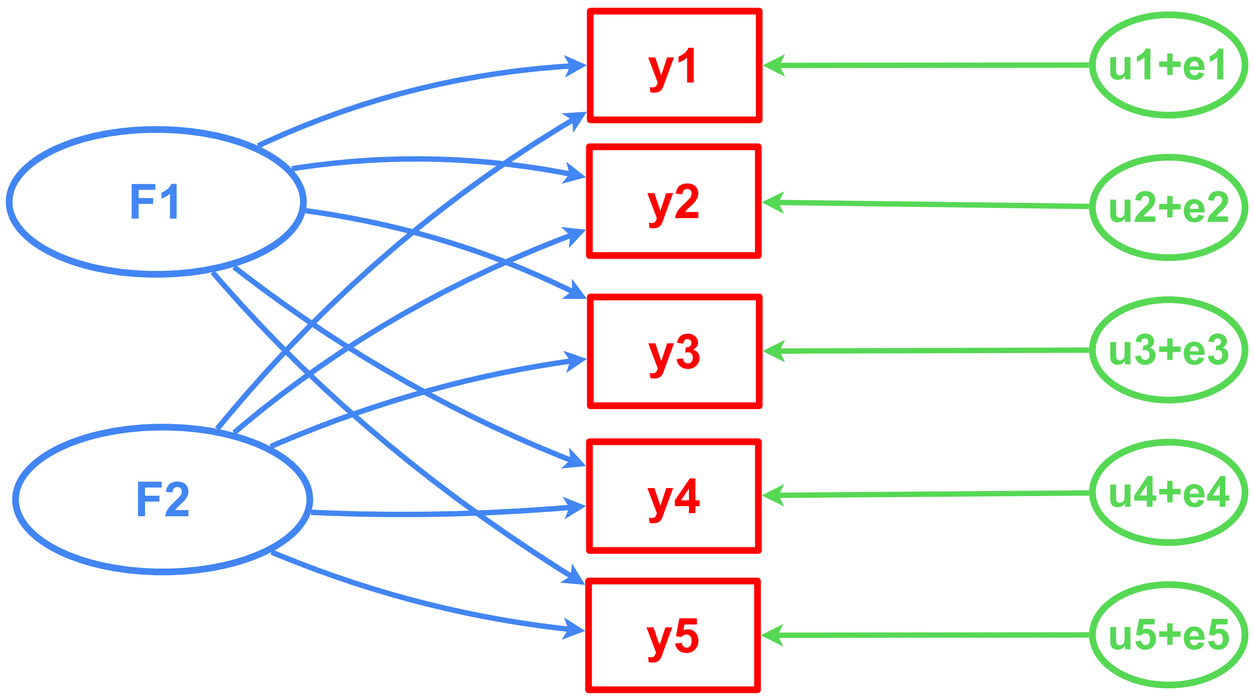


Exploratory Factor Analysis



- Observed variables are assumed to be the manifestation of underlying **latent factors**
- These factors are **orthogonal** (non-correlated)
- Each variable has also a **specific contribution** (u) and a **measurement error** (e)

Exploratory Factor Analysis



unique contribution

$$y_i = a_{i1} F_1 + a_{i2} F_2 + u_i + e_i$$

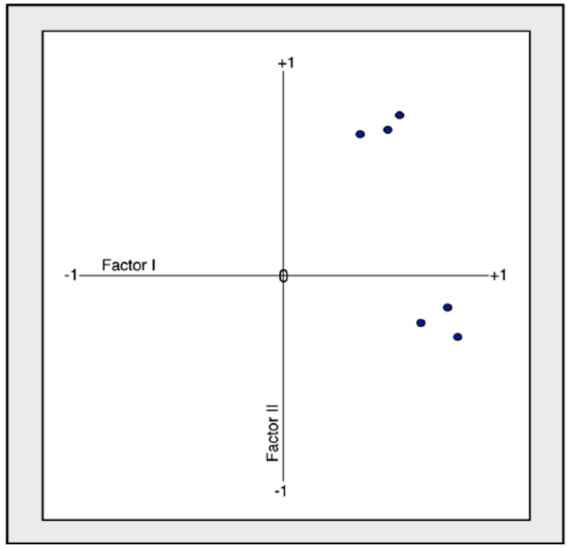
factor loadings

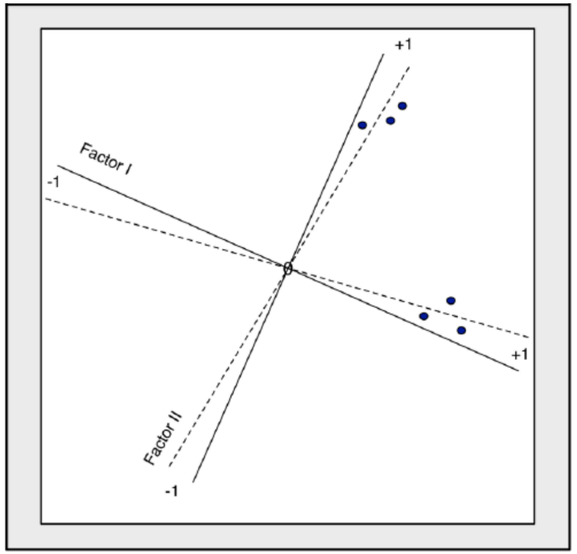
Measurement error

$$Var(y_i) = \underbrace{a_{i1}^2 \, Var(F_1) + a_{i2}^2 \, Var(F_2)}_{\text{communality } h^2} + \underbrace{Var(u_i) + Var(e_i)}_{\text{specificity } u^2}$$

Factor rotation

- Factors are defined up to a **rotation**
- The rotation can be
  - orthogonal: rotated factors remain uncorrelated
  - oblique: rotated factors become correlated





plain = orthogonal
dahed = oblique

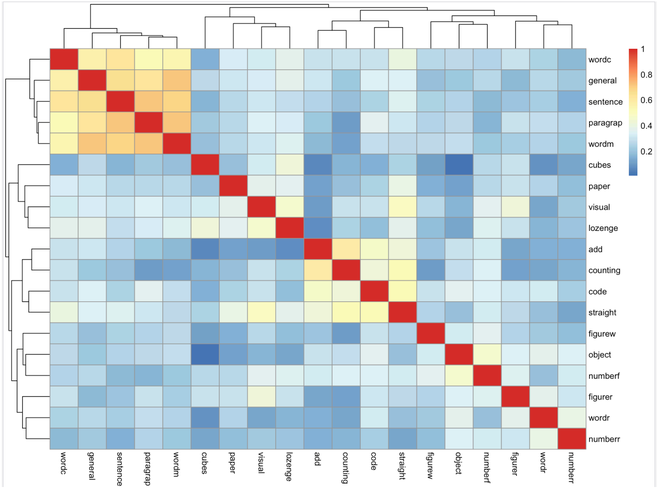
Example of EFA

original data: cognitive test results on n=145 persons

```
> fadata
   visual cubes paper lozenge general paragrap sentence wordc wordm add code counting straight wor
1      23    19    13       4      46       10       17    22    10  69   65       82      156   1
2      33    22    12      17      43        8       17    30    10  65   60       98      195   1
3      34    24    14      22      36       11       19    27    19  50   49       86      228   1
4      29    23    12       9      38        9       19    25    11 114   59      103      144   1
5      16    25    11      10      51        8       25    28    24 112   54      122      160   1
6      30    25    12      20      42       10       23    28    18  94   84      113      201   1
7      36    33    19      36      69       17       25    42    41 129   96      139      333   1
8      28    25    10       9      35       10       18    29    11  96   83       95      174   1
9      30    25    15      11      32       11       21    35     8 103   67      114      197   1
10     20    25    13       6      39        9       21    27    16  89   49      101      178   1
11     27    26    13       6      27       10       16    25    13  88   35      107      137   1
12     32    21    16       8      27        1        7    29    11 103   62      136      154   1
```

visual	scores on visual perception test, test 1
cubes	scores on cubes test, test 2
paper	scores on paper form board test, test 3
lozenge	scores on lozenges test, test 4
general	scores on general information test, test 5
paragrap	scores on paragraph comprehension test, test 6
sentence	scores on sentence completion test, test 7
wordc	scores on word classification test, test 8
wordm	scores on word meaning test, test 9
add	scores on add test, test 10
code	scores on code test, test 11
counting	scores on counting groups of dots test, test 12
straight	scores on straight and curved capitals test, test 13
wordr	scores on word recognition test, test 14
numberr	scores on number recognition test, test 15
figurer	scores on figure recognition test, test 16
object	scores on object-number test, test 17
numberf	scores on number-figure test, test 18
figurew	scores on figure-word test, test 19

correlation structure



Factor analysis (k=4)

	Factor1	Factor2	Factor3	Factor4
visual	0.536	0.176	0.392	-0.249
cubes	0.330		0.302	-0.228
paper	0.440	0.110	0.247	-0.147
lozenge	0.505		0.358	-0.253
general	0.762	-0.238	-0.113	
paragrap	0.759	-0.338		
sentence	0.762	-0.322	-0.166	
wordc	0.701			
wordm	0.762	-0.381		
add	0.455	0.475	-0.451	
code	0.545	0.367		0.103
counting	0.434	0.593	-0.238	-0.162
straight	0.592	0.393		-0.289
wordr	0.394		0.149	0.362
numberr	0.352	0.139	0.219	0.315
figurer	0.435	0.183	0.425	0.192
object	0.445	0.241		0.522
numberf	0.454	0.383	0.221	0.157
figurew	0.389	0.115	0.133	0.202

after rotation

	Factor1	Factor2	Factor3	Factor4
visual		0.747		
cubes		0.571		
paper		0.485		
lozenge		0.683		
general	0.760			
paragrap	0.806			-0.103
sentence	0.862			
wordc	0.555	0.147		0.141
wordm	0.856			-0.117
add		-0.245	0.117	0.806
code			0.290	0.420
counting	-0.150	0.165		0.773
straight		0.489	-0.124	0.484
wordr			0.567	
numberr			0.544	
figurer		0.376	0.501	-0.164
object		-0.244	0.766	0.114
numberf	-0.168	0.271	0.446	0.166
figurew			0.381	

scores of original observations

	Factor1	Factor2	Factor3	Factor4
[1,]	-0.439011	-1.78968	-0.74163	-1.16528
[2,]	-0.640320	0.33018	0.17654	-0.65755
[3,]	-0.057138	0.81855	-1.35900	-1.40696
[4,]	-0.554279	-1.07389	-0.70366	0.26768
[5,]	0.681781	-1.70449	0.18772	0.53174
[6,]	0.219437	0.32968	1.04977	0.20680
[7,]	2.611266	3.18222	2.64516	2.36497
[8,]	-0.479476	-0.53915	0.28261	-0.31006
[9,]	-0.232114	-0.22414	-0.95923	0.47125
[10,]	-0.069679	-1.38991	-0.83592	-0.24703

Summary: Exploratory Factor Analysis

Assumptions

- **Sampling adequacy**
  enough observations per variable
  → *Kaiser-Meyer-Olkin (KMO) test*
- No **multicolinearity** (singular correlation matrix!)
- Covariance matrix should not be the identity matrix!
  → *Bartlett test*
- More observations than variables

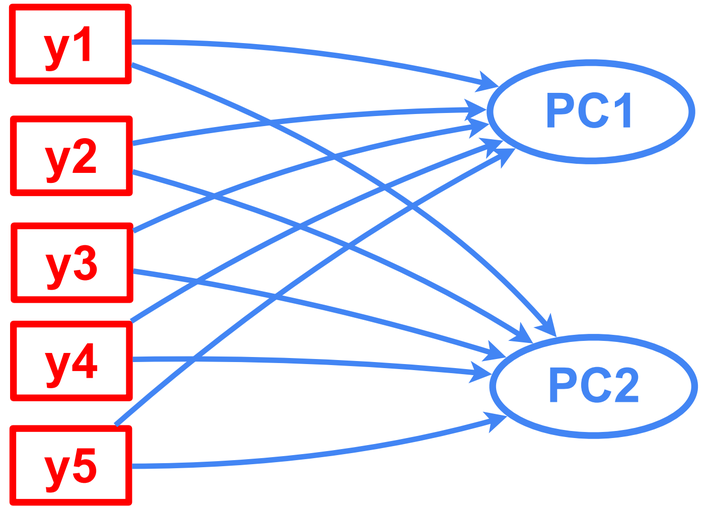
Questions/Challenges

- Factors are determined **up to a rotation**
- Rotation can be
  - **orthogonal** (rotated factors still uncorrelated) or
  - **oblique** (rotated factors are correlated)
- Proper number of factors remains to be determined
  → heuristic (Kaiser rule, knee-plot,...)

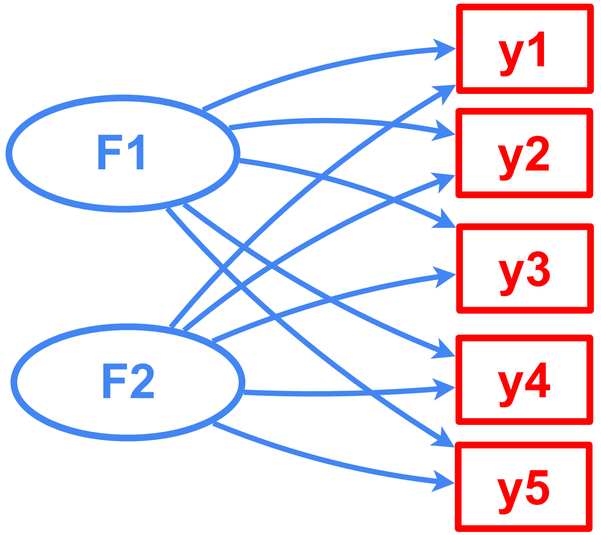
Difference PCA / EFA

"Basically, researchers tend to:

- use PCA if they are on a **fishing expedition** trying to find patterns in their data and have no theory to base the analysis on, or

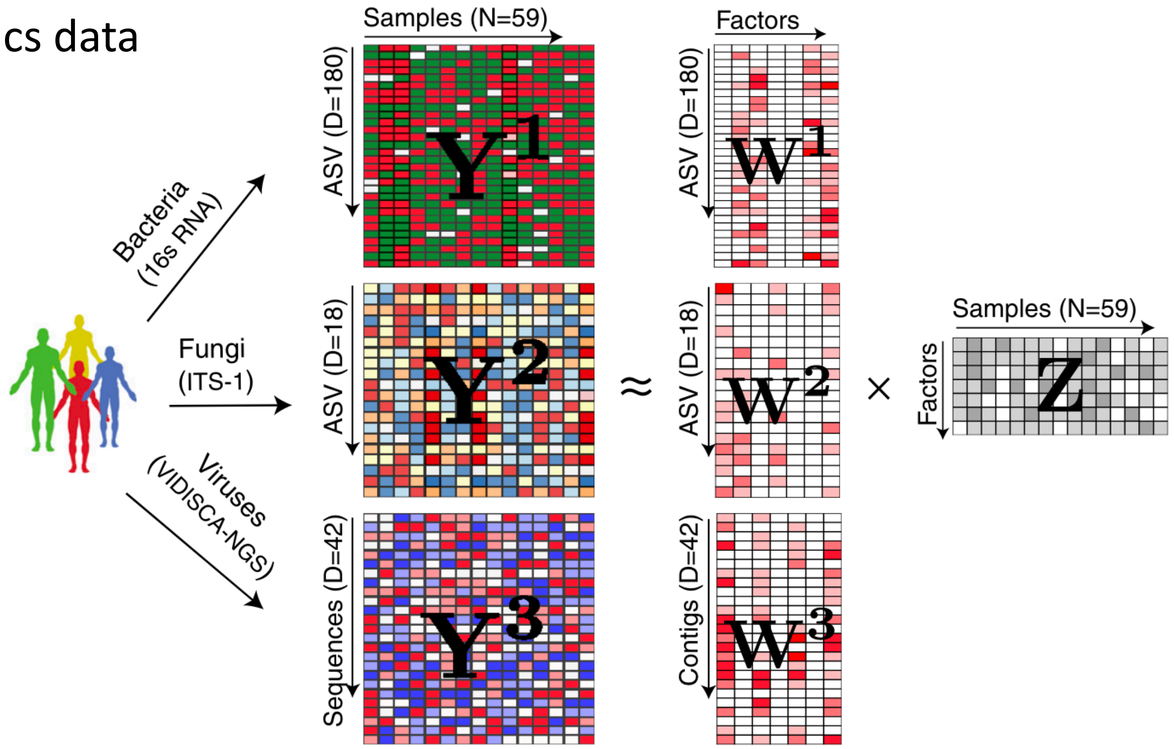


- use EFA if they have a **well-grounded theory** to base their analysis on. Generally, the second strategy is considered to be the stronger form of analysis."



Multi-Omics Factor Analysis (MOFA)

Omics data



Metadata

```
> metadata
     sample Age    Sexs                  Diagnosis Category Penicillins Cephalosporins Carbapenems Macrolides Aminoglycosides Quinolones
 1:  TKI_F1  89 Female          Sepsis, pulmonary   Sepsis       FALSE          FALSE       FALSE      FALSE            TRUE       TRUE
 2:  TKI_F2  74 Female          Sepsis, pulmonary   Sepsis        TRUE          FALSE       FALSE      FALSE           FALSE      FALSE
 3:  TKI_F3  61   Male          Sepsis, pulmonary   Sepsis       FALSE           TRUE       FALSE       TRUE           FALSE      FALSE
 4:  TKI_F4  67 Female          Sepsis, pulmonary   Sepsis       FALSE           TRUE       FALSE      FALSE           FALSE      FALSE
 5:  TKI_F5  72   Male          Sepsis, pulmonary   Sepsis       FALSE           TRUE       FALSE      FALSE           FALSE      FALSE
 6:  TKI_F6  57 Female          Sepsis, pulmonary   Sepsis        TRUE           TRUE       FALSE      FALSE            TRUE      FALSE
 7:  TKI_F7  68   Male          Sepsis, pulmonary   Sepsis       FALSE           TRUE       FALSE      FALSE           FALSE      FALSE
 8:  TKI_F8  62 Female          Sepsis, pulmonary   Sepsis       FALSE           TRUE       FALSE       TRUE           FALSE      FALSE
 9:  TKI_F9  71   Male          Sepsis, pulmonary   Sepsis        TRUE           TRUE       FALSE      FALSE           FALSE      FALSE
10: TKI_F10  74   Male          Sepsis, pulmonary   Sepsis        TRUE           TRUE       FALSE       TRUE           FALSE      FALSE
11: TKI_F11  66 Female          Sepsis, abdominal   Sepsis        TRUE           TRUE        TRUE      FALSE            TRUE      FALSE
```

$$Y^m = W^m \cdot Z + \epsilon^m$$

- Matrices W^m and Z are learned through **bayesian inference**
- Implementation favors **sparsity**
  - sparsity of the *W* matrices
  - sparsity of the *Z* matrix
- Different models for Y^m, ϵ^m
  - Poisson model (count)
  - Bernouilli model (binary)
  - Gaussian model (continuous)

[Argelaguet, MSB 2018]

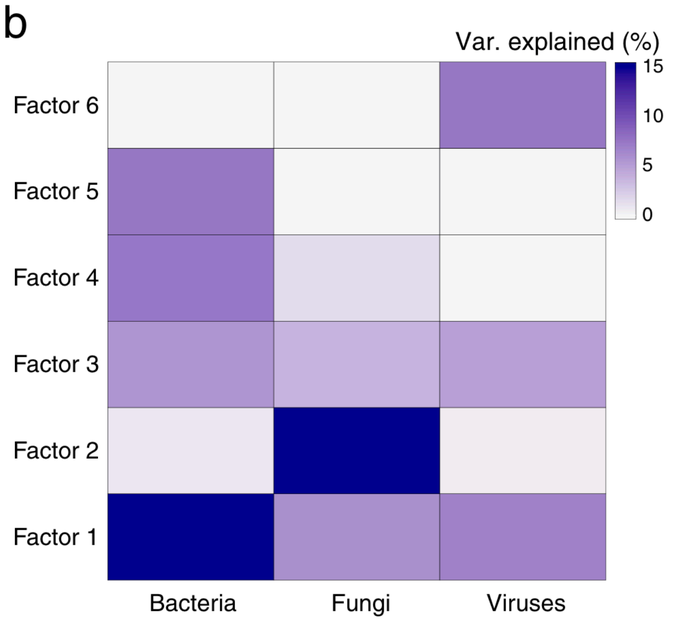
Multi-Omics Factor Analysis (MOFA): variance explained

Total variance explained in each view and each factor

$$R^2_{m,k} = 1 - \left(\sum_{n,d} y^m_{nd} - z_{nk} w^m_{kd} - \mu^m_d\right)^2 / \left(\sum_{n,d} y^m_{nd} - \mu^m_d\right)^2$$

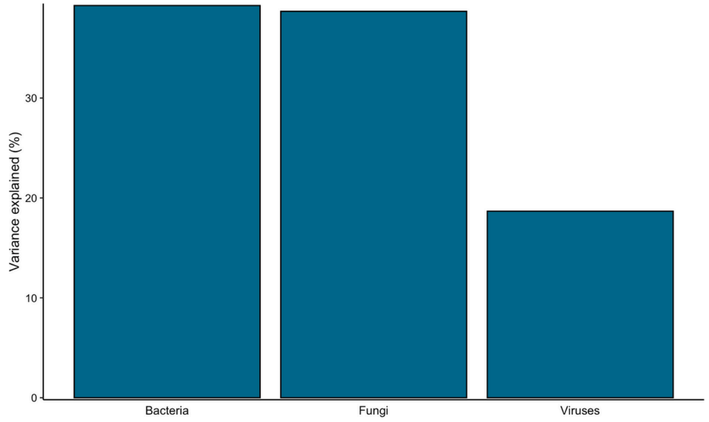
Residual variance in view *m* and factor *k*

Total variance in view *m*



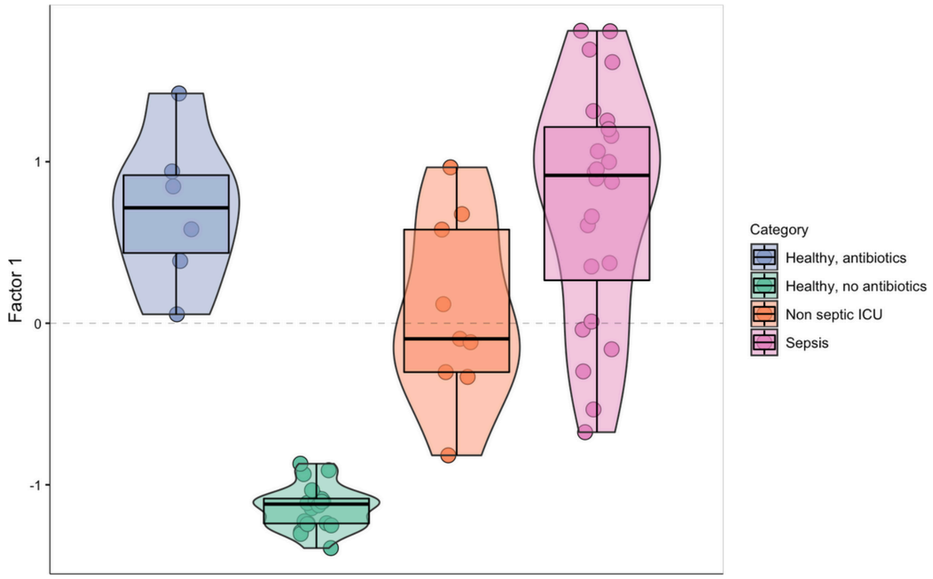
Total variance explained in each view

$$R^2_m = 1 - \left(\sum_{n,d} y^m_{nd} - \sum_k z_{nk} w^m_{kd} - \mu^m_d\right)^2 / \left(\sum_{n,d} y^m_{nd} - \mu^m_d\right)^2$$

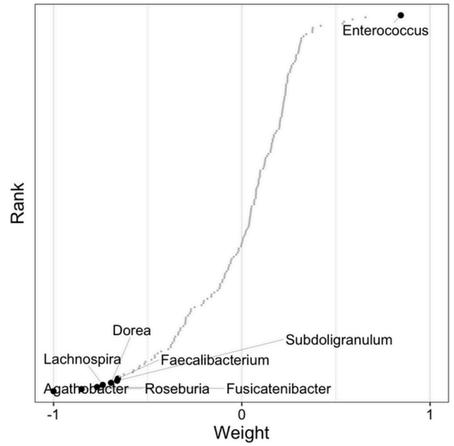


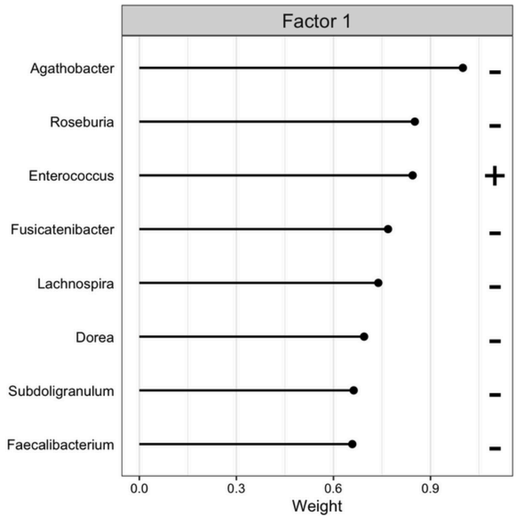
MOFA: post-hoc interpretation of factors

Analysis 1
Association of factors with groups
(Z matrix)

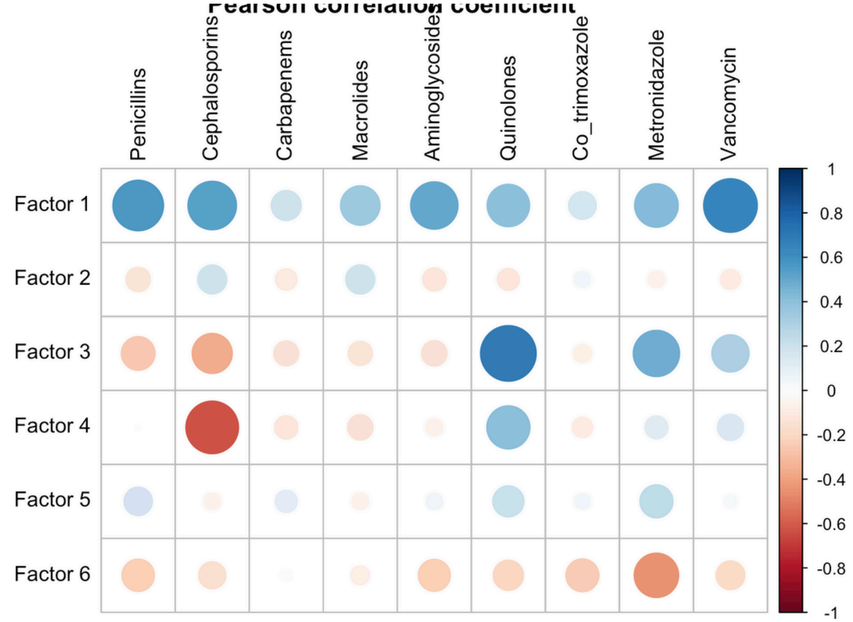


Analysis 2
Weights in factors for each view
(W matrix)

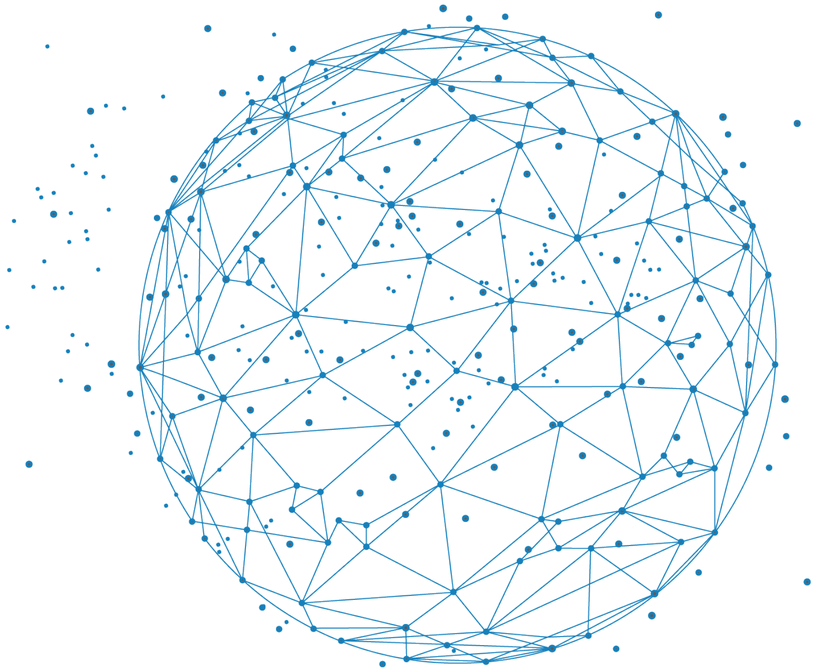




Analysis 3
Correlation of factors with covariates
(Z matrix)

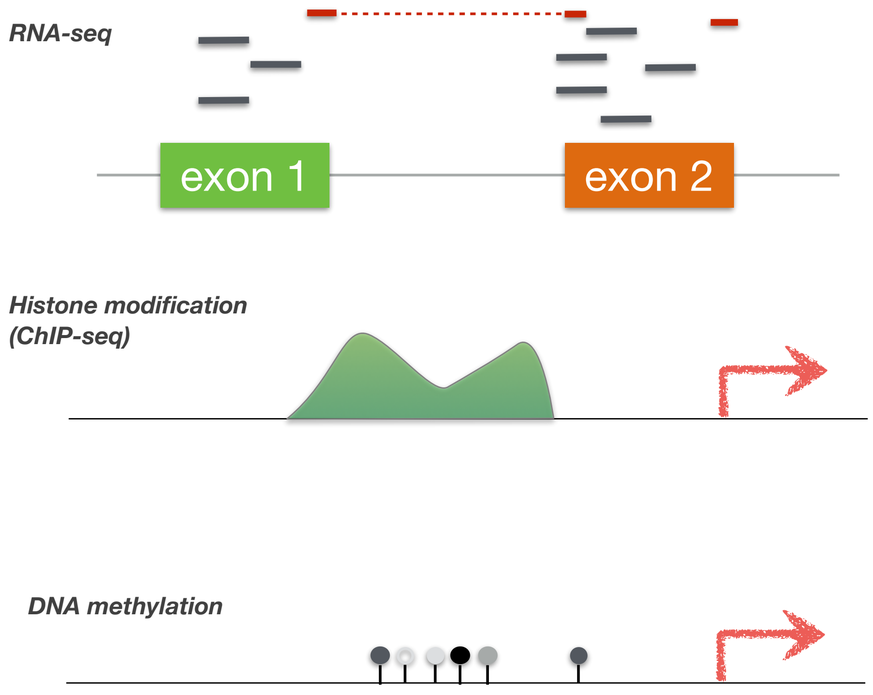


Non-negative matrix factorization



Non-negative matrix factorization

- Most datasets in modern genomics are by essence non-negative
- Read counts in RNA-seq
- Methylation *b*-values in DNA methylation arrays
- Integrated signal aver genomic regions



we can apply parts-base decomposition of the data

Non-negative matrix factorization

$$X \sim WH \qquad \text{with} \quad X \geq 0, W \geq 0, H \geq 0$$

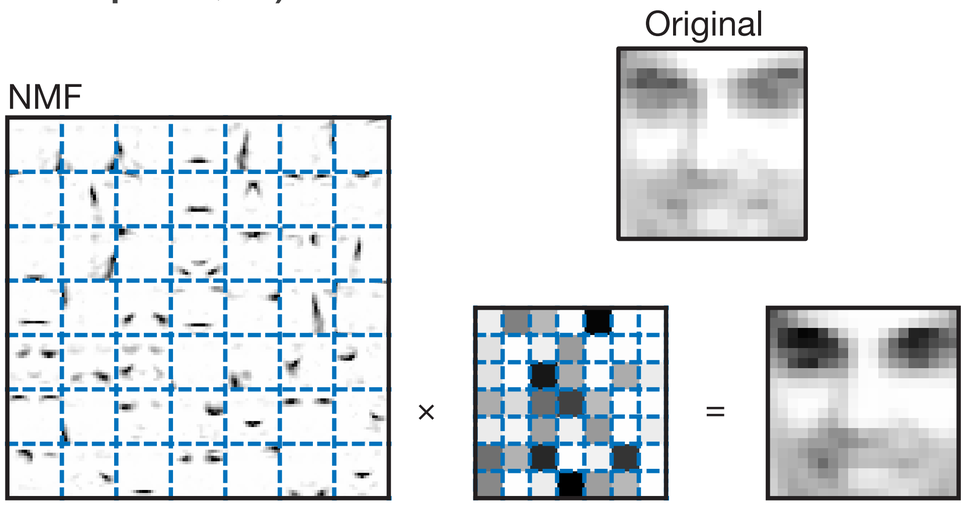
$X : N \times M$ matrix

$N =$ number of features (genes, regions,...)

$M =$ number observations (patients,samples,...)

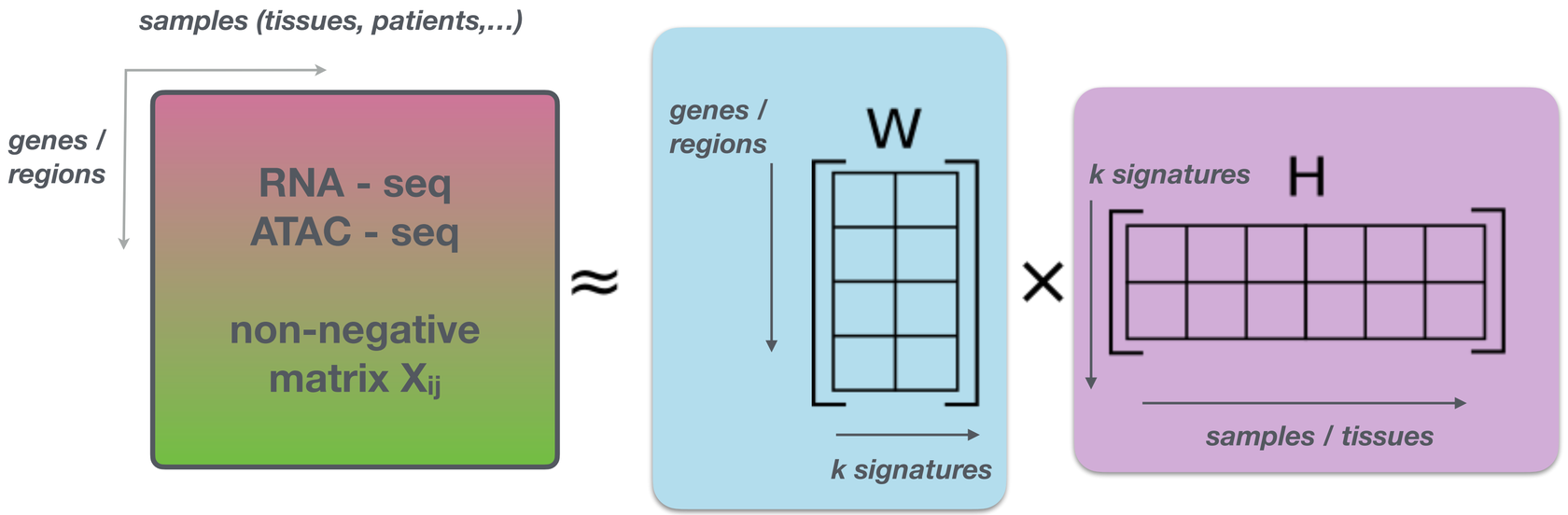
NMF in essence similar to PCA,
but non-negativity implies

- a **better interpretability** of the signatures
- a **natural sparseness** of the decomposition



[Lee, Seung 1999]

NMF



X : original data matrix

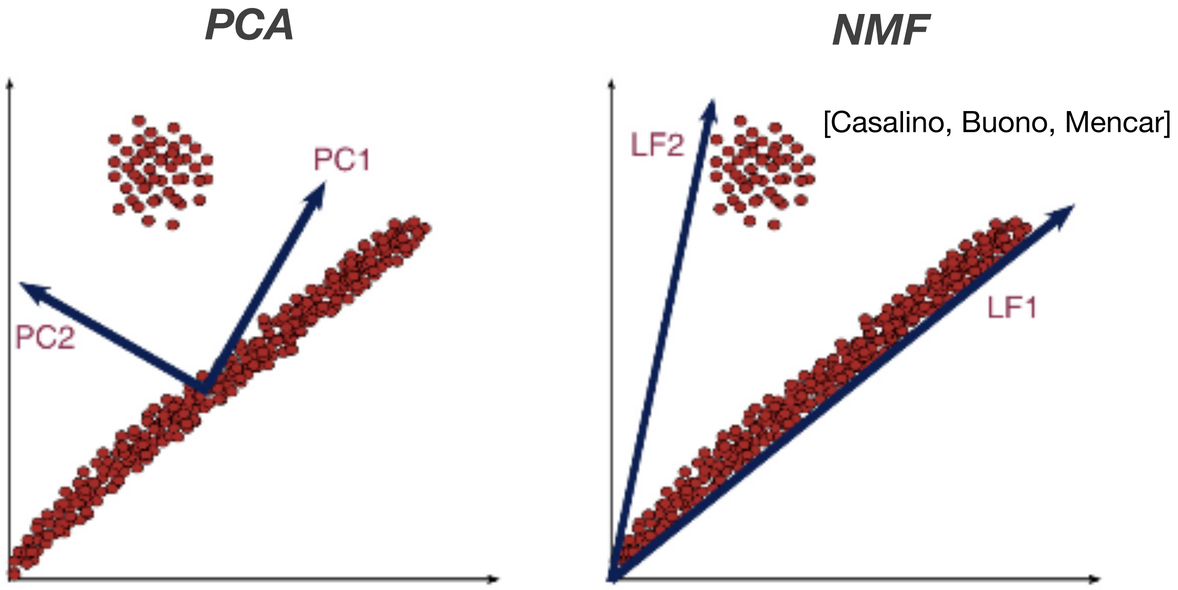
columns of W : k signatures (genes, regions,…)

columns of H : exposures to the k signatures

→ Genomic signatures + features of the signature

NMF vs. PCA

- PCA defines orthogonal directions explaining most variance
- NMF signatures (or *latent factors LF*) define the hypercone containing all data points
- There is **no natural ranking of the NMF-signatures** (unlike PCs); choice of the number of signatures is crucial!



because of the non-negative constraint, only point inside the cone can be reconstructed using the basis vectors

NMF vs. PCA

(a) (b) (c) (d) (e) (f)

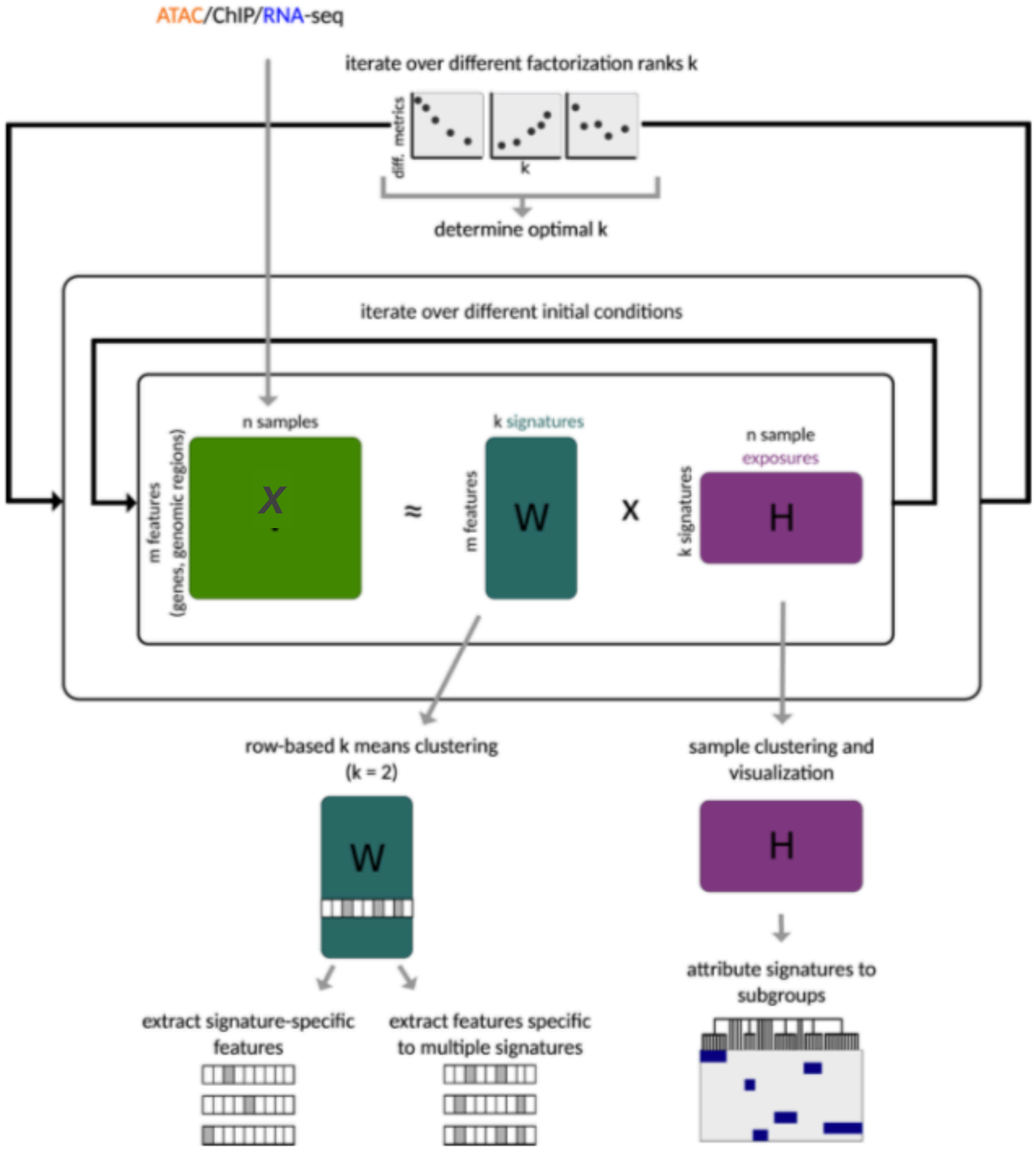
Figure 4.5: Base images of dataset $\mathbf{D_{face}}$ after applying the PCA

(a) (b) (c) (d) (e) (f)

Part are more easily interpretable in NMF

[Nikolaus]

Implementation



- Iteration over **update equations** (~ 10.000s, inner iteration)
- Iterate of set of **initial conditions** (~ 10s, outer iteration)
- Iterate over different **number of signatures** to be extracted

How to choose k?

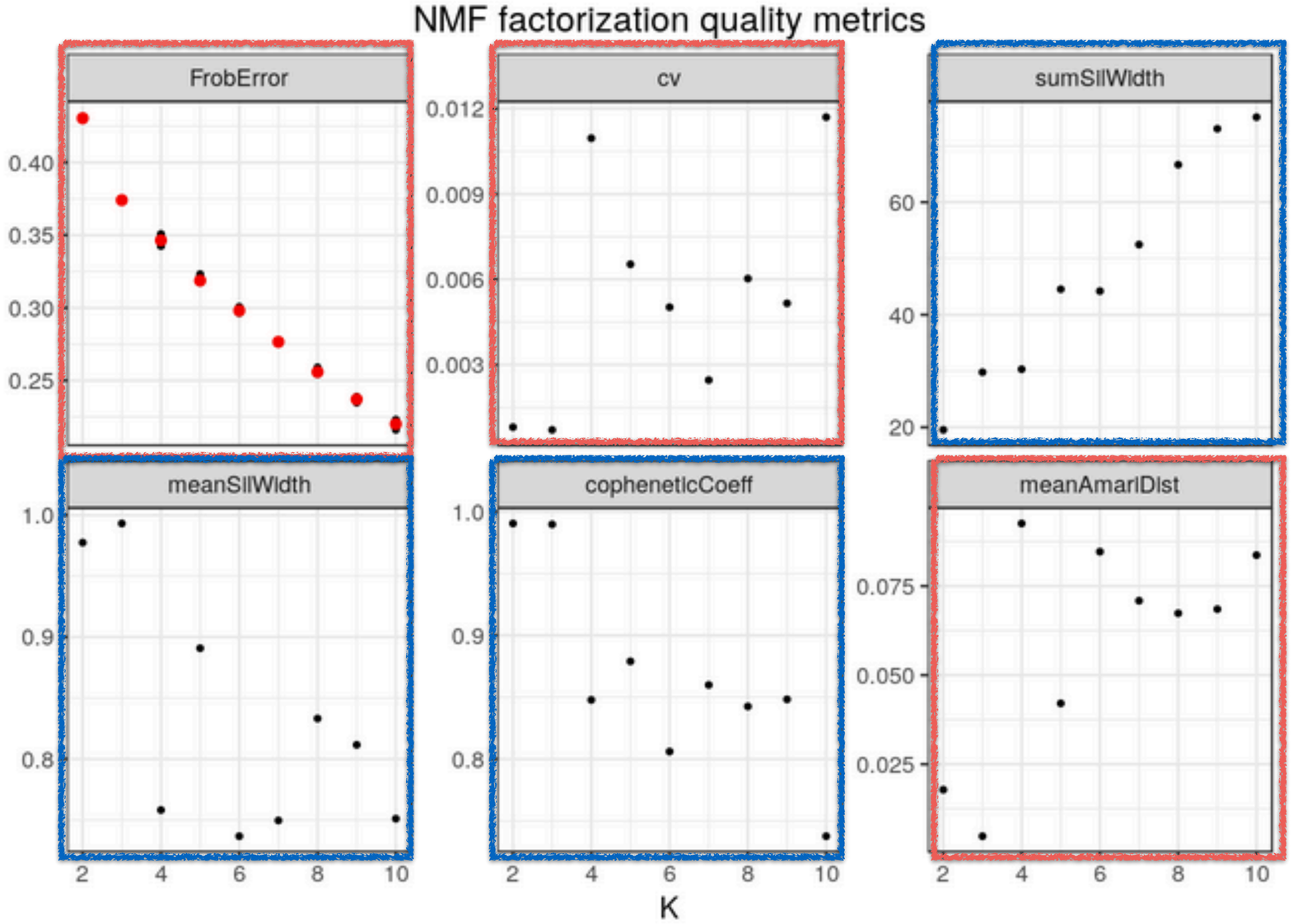
- Accuracy of matrix decomposition: how well does *WH* represent *V*?
  - **Froebenius error** should be small
  - **Amari distance** should be small

- Stability of solutions: how variable are the solutions using different random initializations?
  - **Coefficient of variation** should be small

- Groups of samples should be homogeneous: how well does each sample belong to its group?
  - **Silhouette coefficient** should be large

- Clustering should well represent the original data
  - **Cophenetic coefficient** should be large

How to choose k?



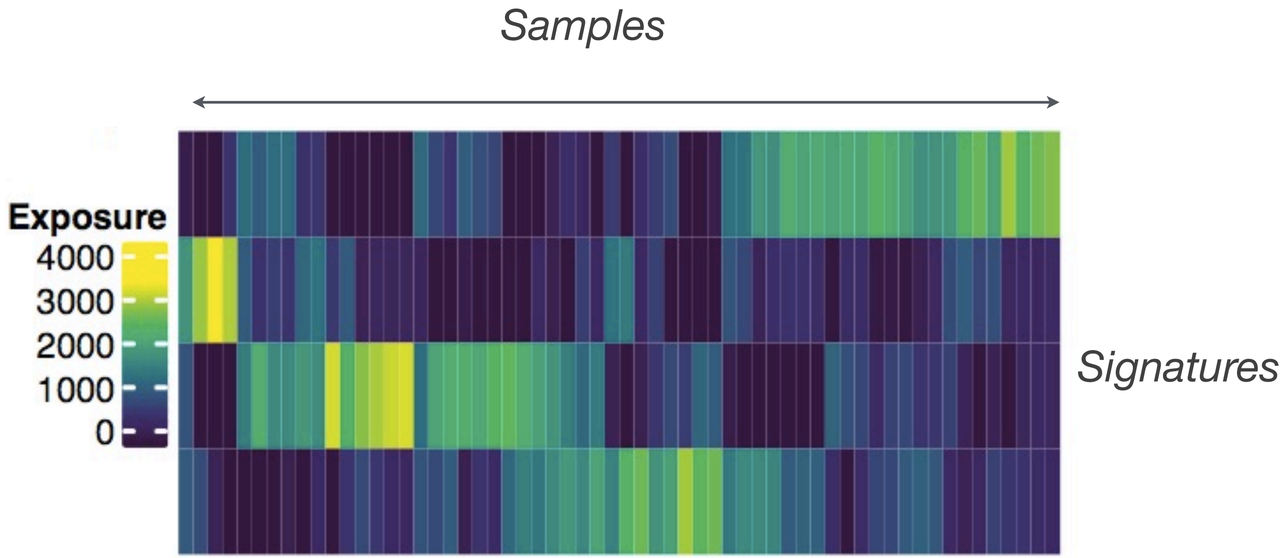
large is better

small is better

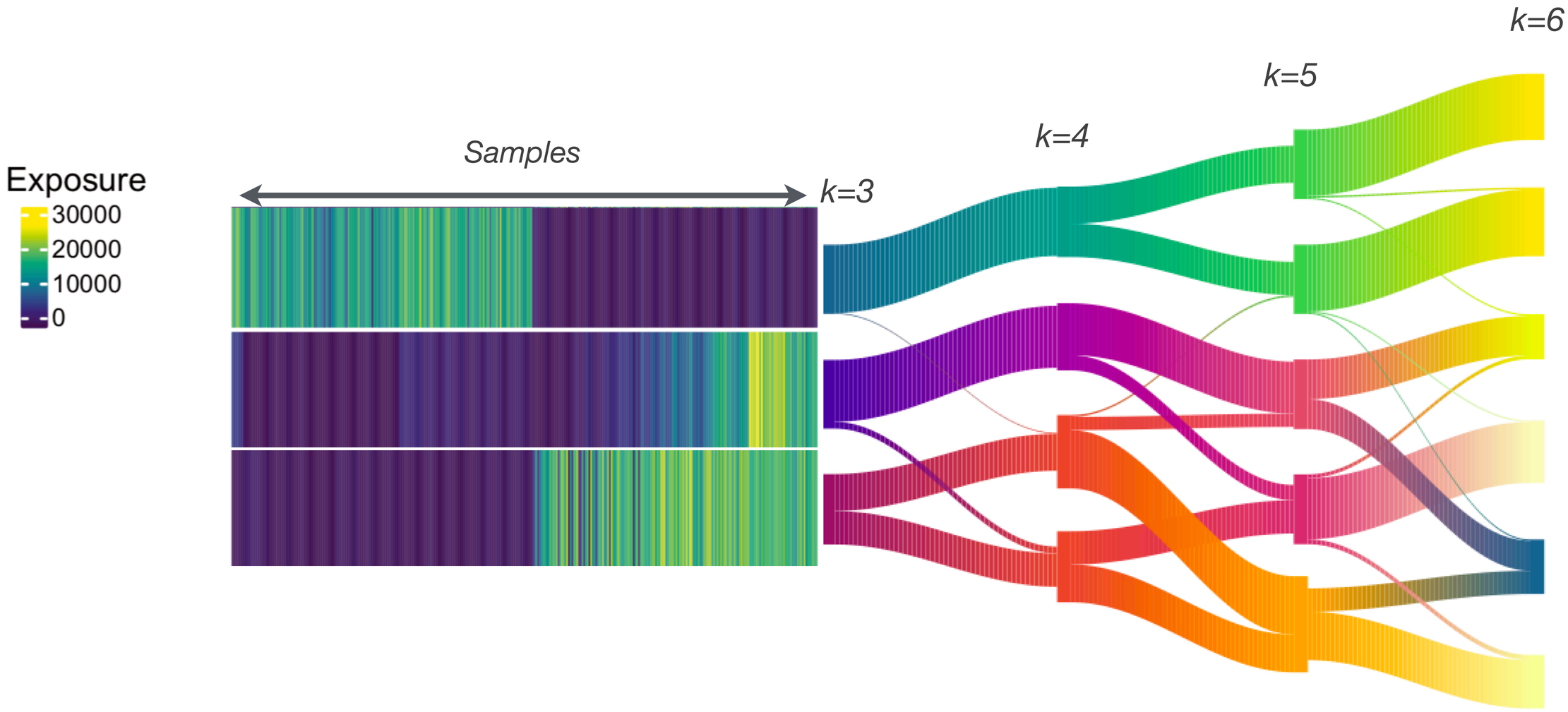
k = 5 appears to be a good choice

Exposure matrix H

- A sample can have “exposure” to multiple signatures
- Gradient of exposures (unlike hard clustering)
- sparseness: many coefficients are (almost) 0 in W and H matrix

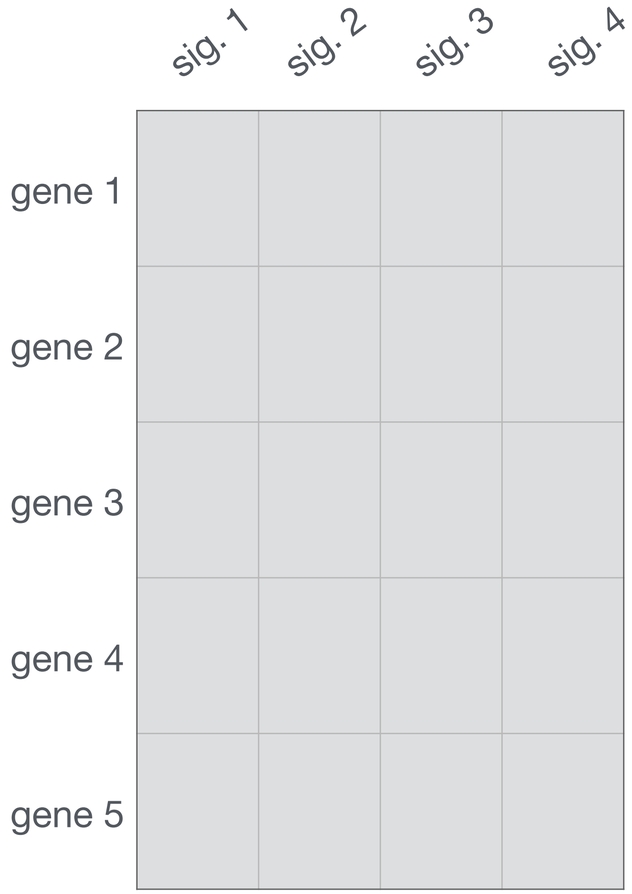


Stability of signatures



Signature matrix W

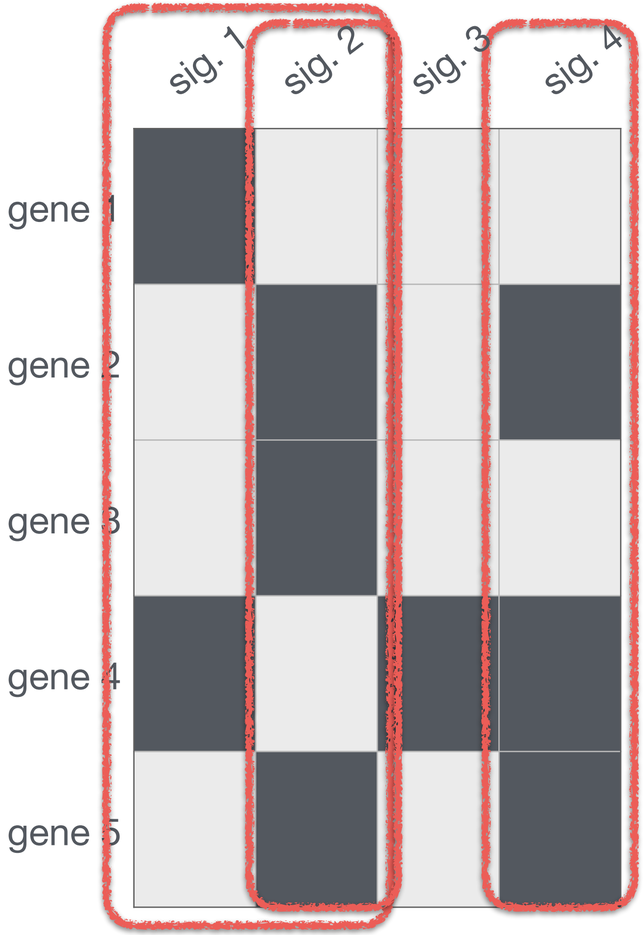
- the W matrix gives the “definition” of the signatures in terms of features contributing
- applying k-means (k=2) to each row of the W matrix



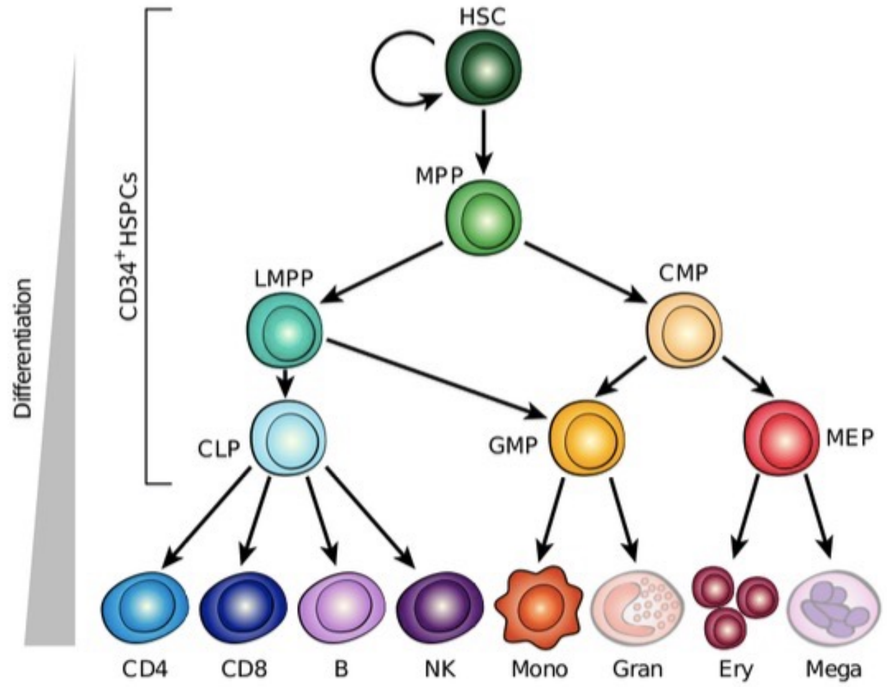
Signature matrix W

- the W matrix gives the “definition” of the signatures in terms of features contributing
- applying k-means (k=2) to each row of the W matrix

  - **single-signature features**:
    → gene 1 / 3
  - **multi-signature features**:
    → gene 2 / 4 / 5
  - signatures 1 and 2 share no feature
  - signatures 2 and 4 share 2 features

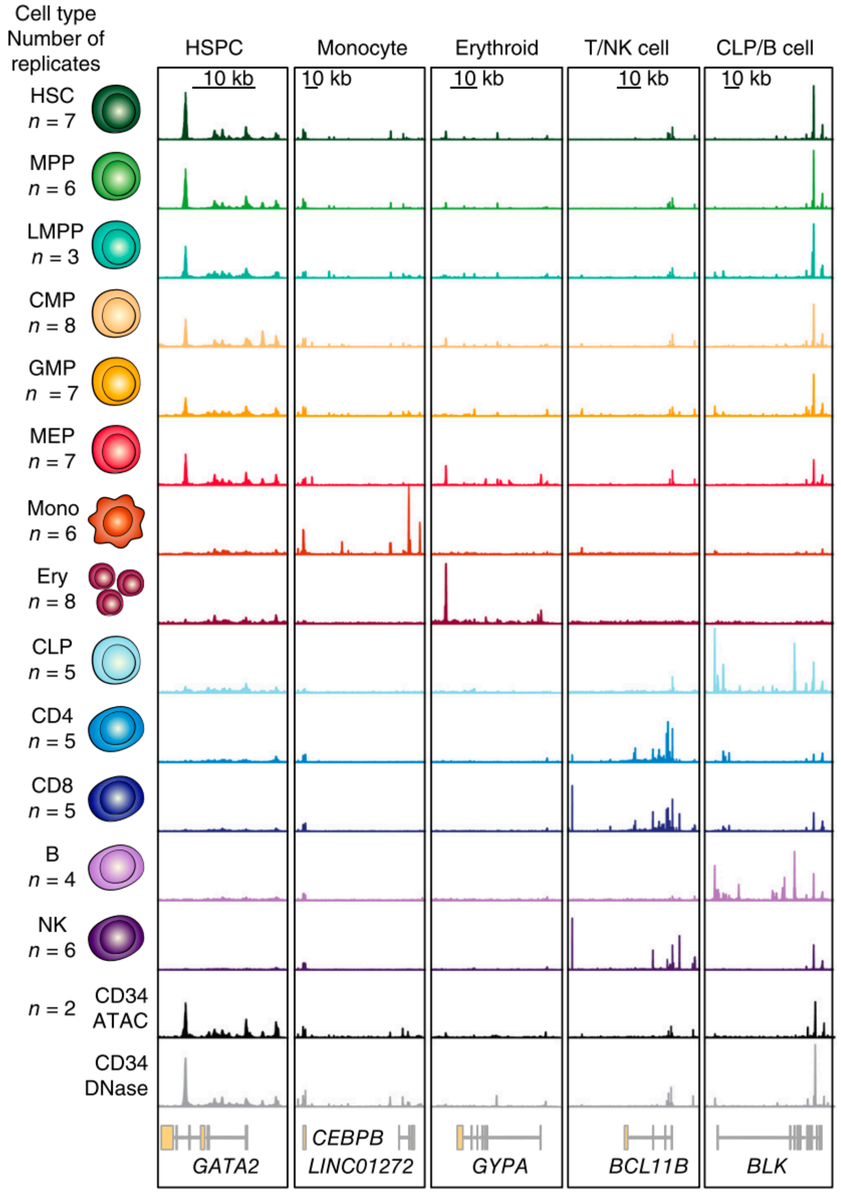


Example of use case



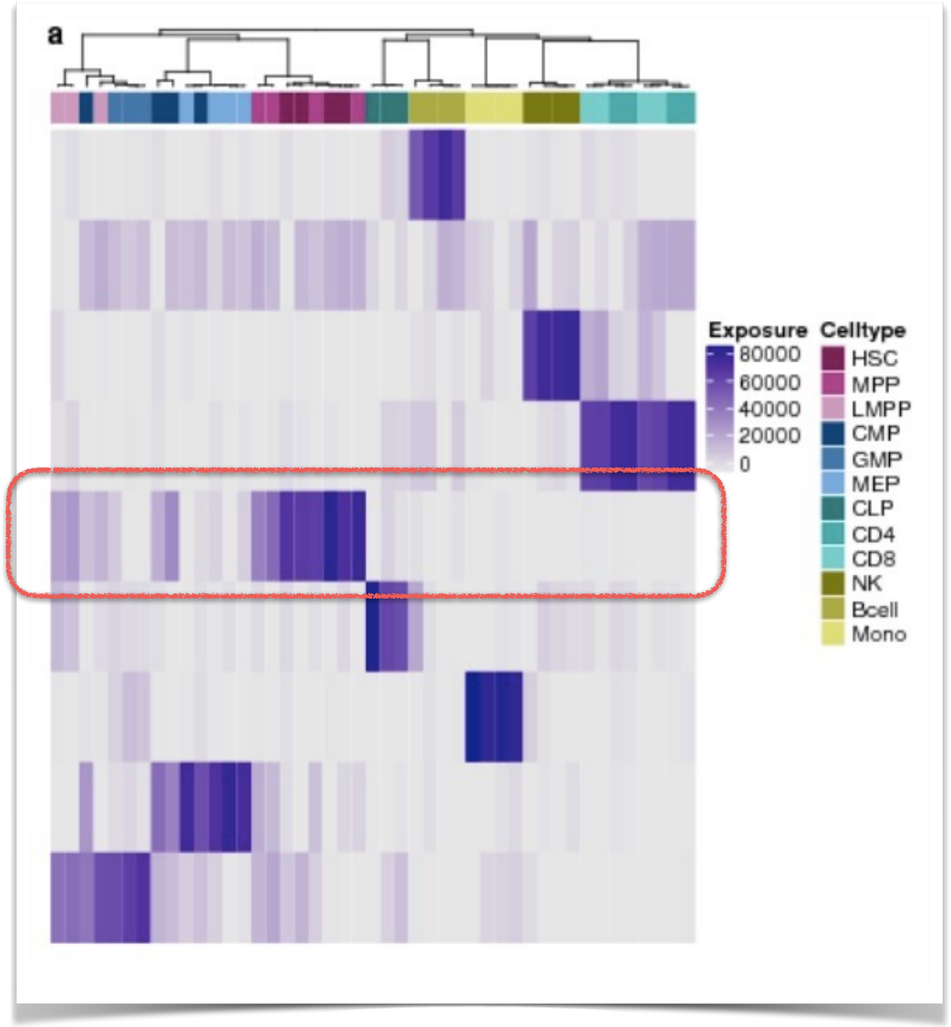
Combined RNA-seq (gene expression) and chromatin acessibility (ATAC-seq) from purified blood populations

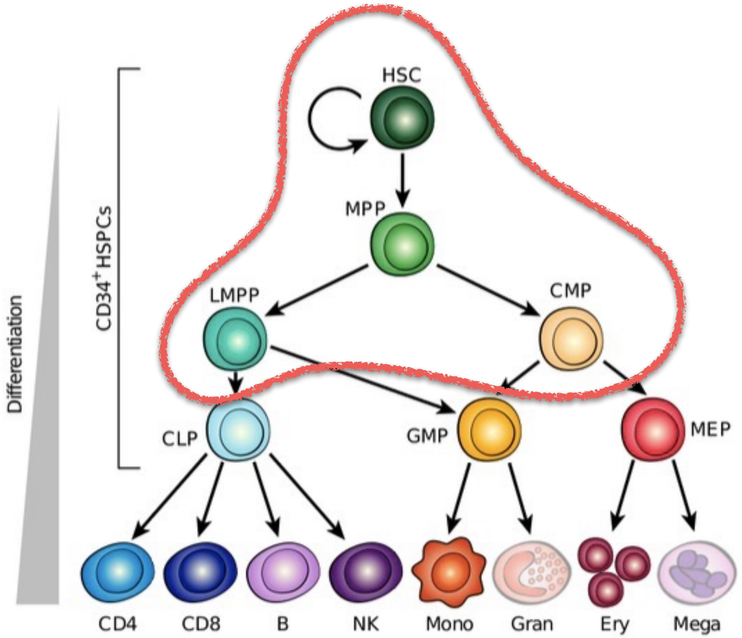
[Corces et al. Nat. Gen (2016)]



Interpreting signatures

RNA-seq



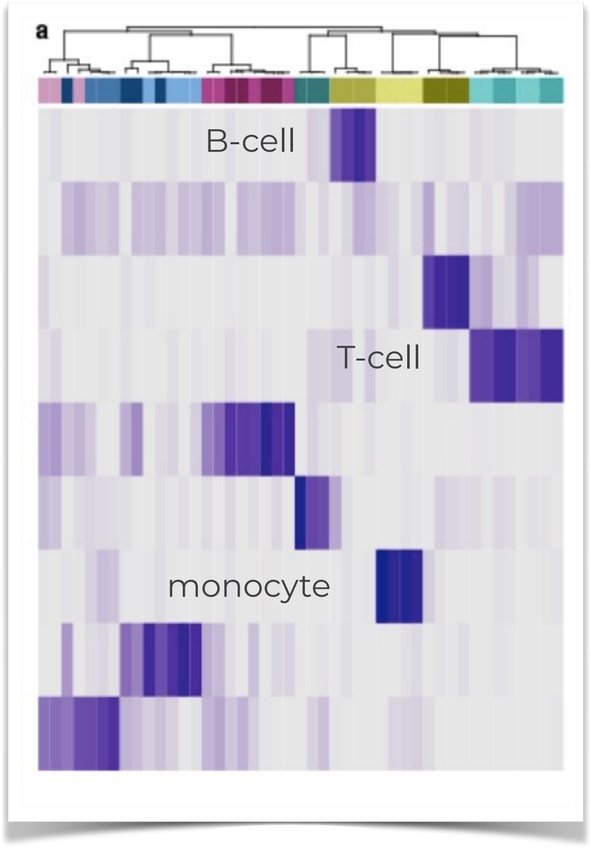


Stemness-signature fades away, as differentiation progresses

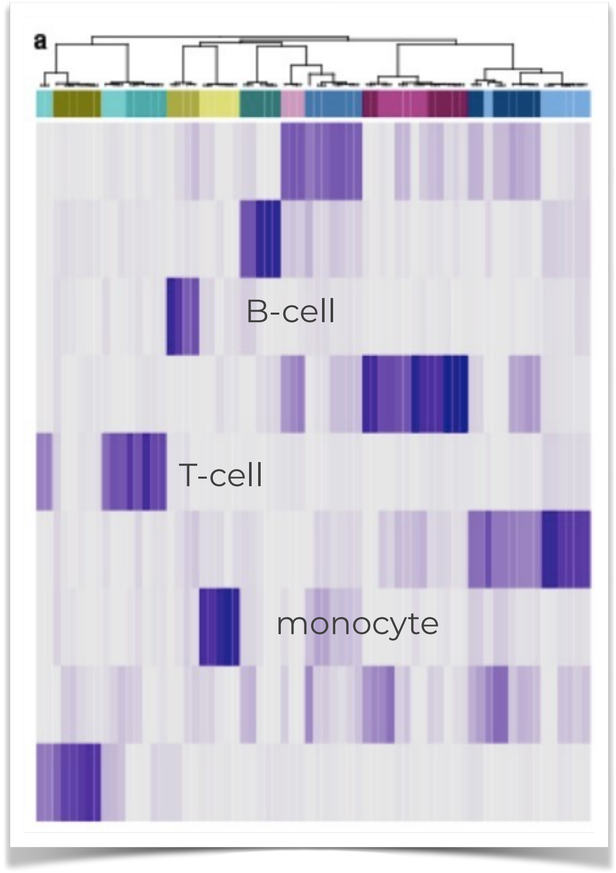
[Corces et al. Nat. Gen (2016)]

Associating signatures

RNA-seq

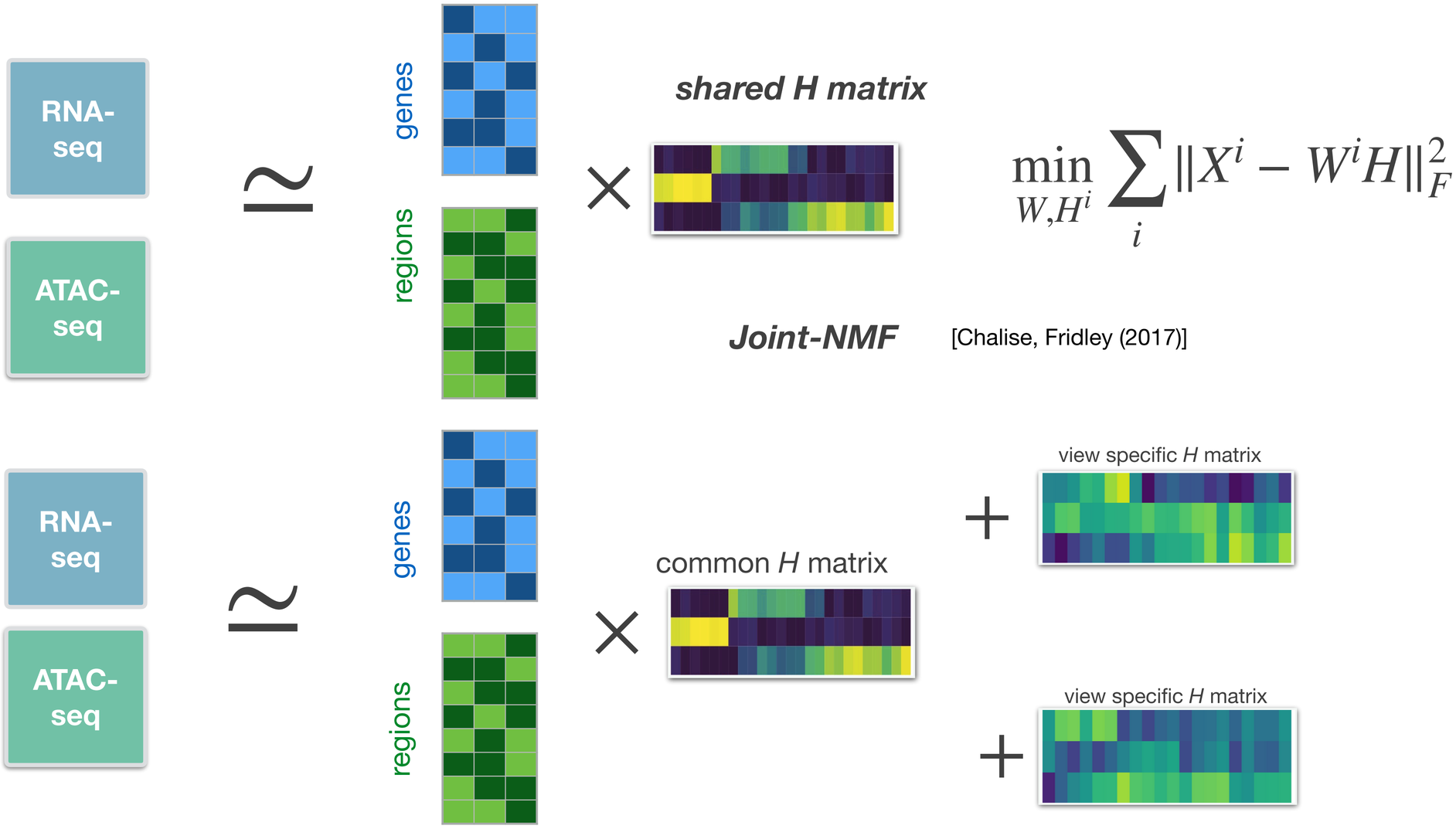


ATAC-seq



[Corces et al. Nat. Gen (2016)]

Integrating multiple datasets using NMF



$$\min_{W,H^i} \sum_i \|X^i - W^i H\|_F^2$$

Joint-NMF [Chalise, Fridley (2017)]

integrative-NMF [Yang, Michailidis (2015)]

[Andres Quintero]

integrative NMF

general reconstruction error

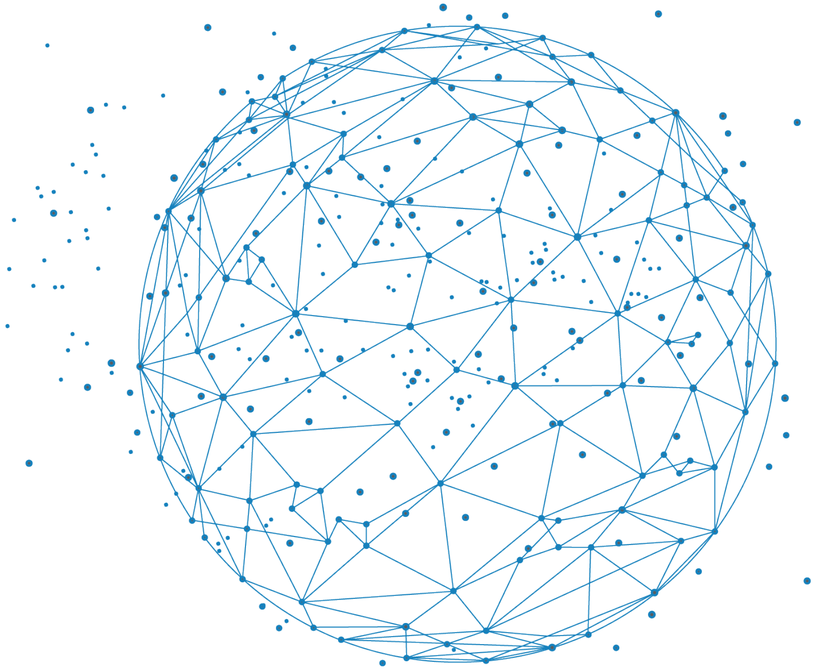
heterogeneous part

$$\min_{W^i,H,H^i} \left(\boxed{\sum_i \|X^i - W^i(H + H^i)\|_F^2} + \boxed{\lambda \sum_i \|W^i H^i\|_F^2} \right)$$

- integrative NMF identifies both **homogeneous** effects between datasets (H) as well as **heterogeneous** (H^i)
- λ is a homogeneity parameters
  - large values will promote the **homogeneous** effects
  - small values will promote the **heterogeneous** effects

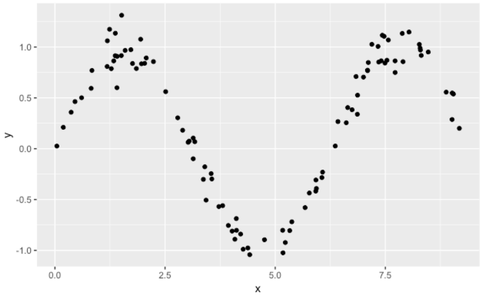
[Yang, Michailidis (2015)]

Keep in mind



Key concepts

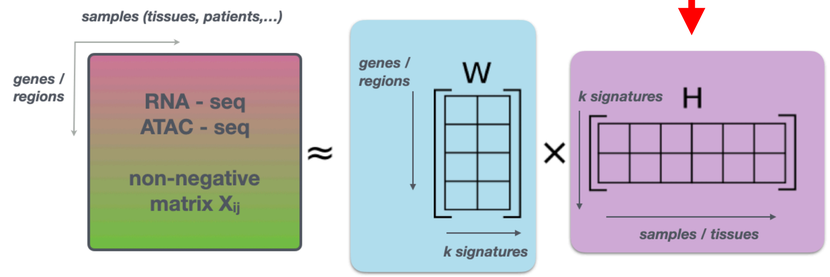
- These methods are **linear methods**, which makes assumptions about **linear co-variation** of the variables (correlation is a linear measure!)



- Some consider the **total variance** (of a variable or a data set), some determine the **shared/specific** part (e.g. PCA vs. EFA)

- We have described **unsupervised** multivariate approaches; can be enriched with prior knowledge (e.g. graph-NMF)

can be initialized with prior information



Vocabulary cheat sheet

- **Views / modalities**
  → different types of data

- **Latent factor / signature / Principal component**
  → lower dimensional representation

- **Variance / covariance**
  → data spread, joint variation

- **Homogeneous**
  (= communality, shared)
  → amount of shared variance

- **Heterogeneous**
  (= uniqueness, specific)
  → amount of specific variance