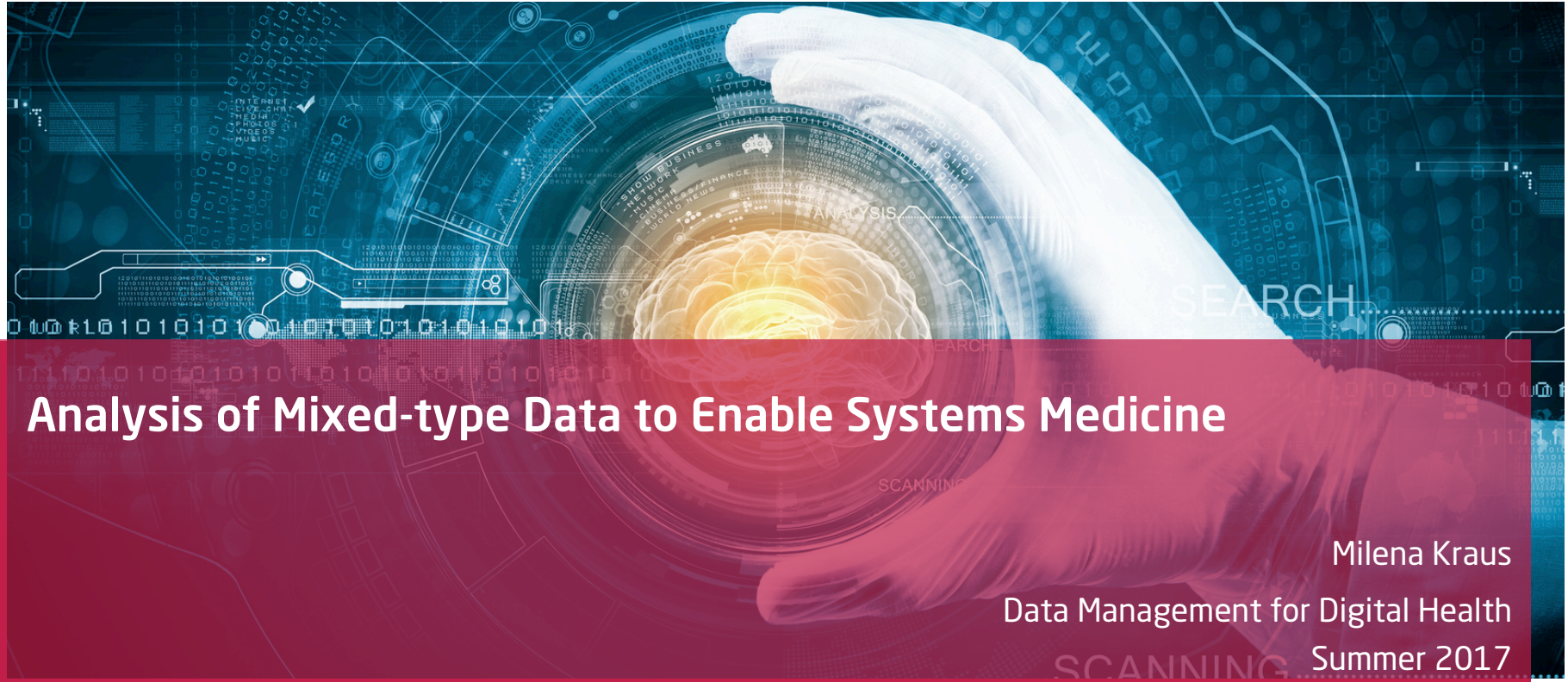
Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam



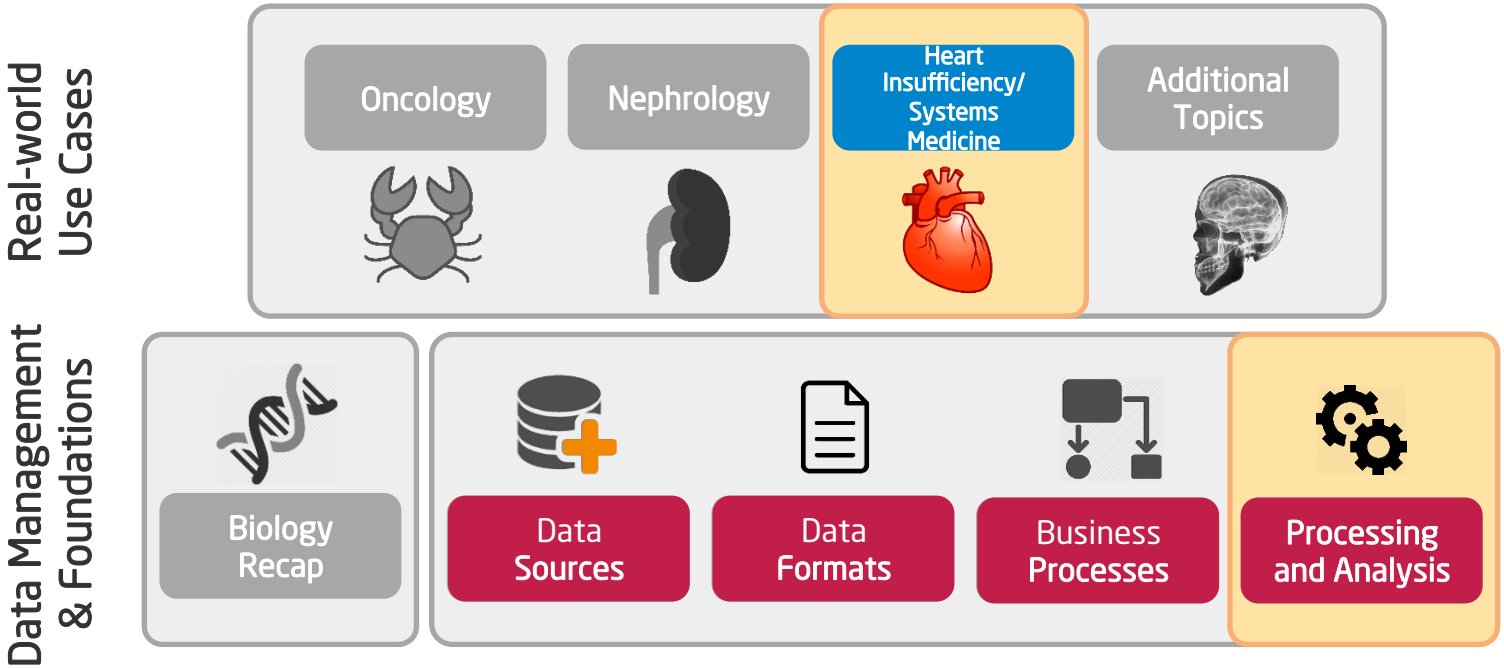
# Analysis of Mixed-type Data to Enable Systems Medicine

Milena Kraus

Data Management for Digital Health

Summer 2017

# Where are we?


Real-world Use Cases
Oncology
Nephrology
Heart Insufficiency/ Systems Medicine
Additional Topics
Data Management & Foundations
Biology Recap
Data Sources
Data Formats
Business Processes
Processing and Analysis

Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

# Recap Last Lecture

Milena Kraus

Data Management for Digital Health

Summer 2017

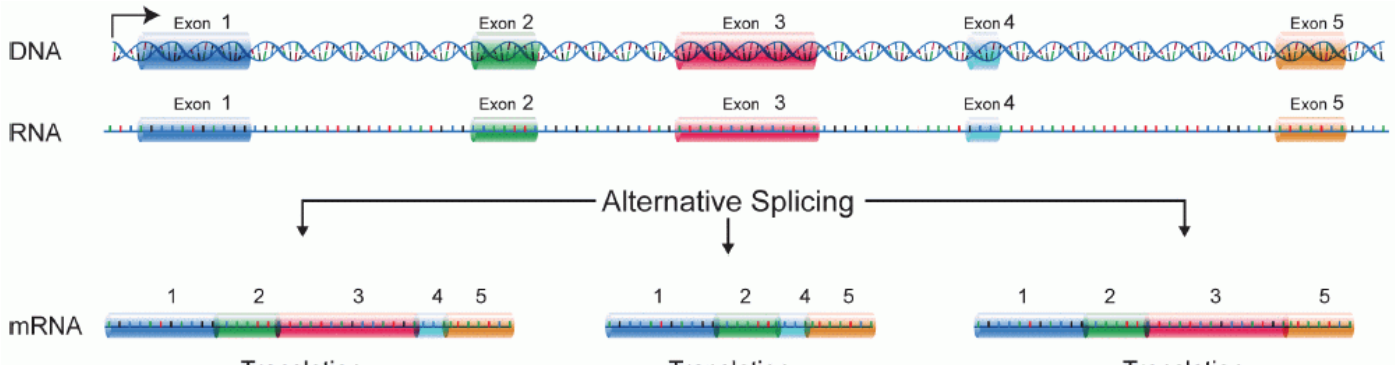
# Questions from last lecture:
# 1. What is the FM-Index?

Hasso Plattner Institut

- FM Index: an index combining the BWT with a few small auxiliary data structures
- "FM" supposedly stands for "Full-text Minute-space."
- Core of index consists of F and L from Burrows-Wheeler-Matrix:
  - F can be represented very simply (1 integer per alphabet character)
  - And L is compressible
- Potentially very space-economical!
- For more information: https://www.youtube.com/watch?v=kvVGj5V65io

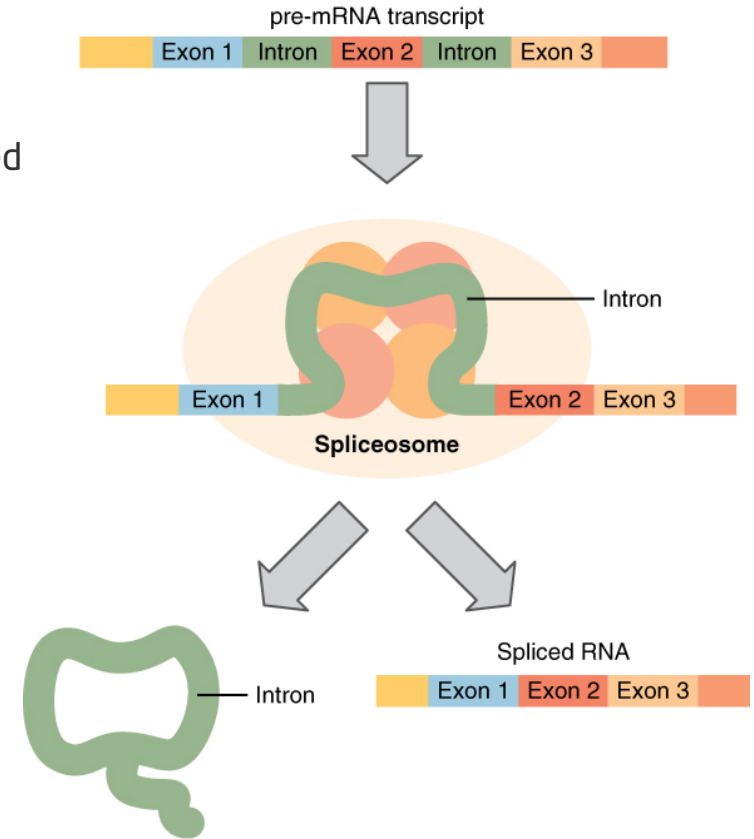
# Questions from last lecture:
# 2. Is exon order preserved in alternative splicing?

- Very few literature on that topic but phenomenon has been observed and is called "exon scrambling"
- Evidence on the scrambling happening in the alternative splicing process does not exist
- A very new species of circular mRNAs may be prone to exon scrambling



# Questions from last lecture:
# 2. Is exon order preserved in alternative splicing?

- Researchers hypothesize that this phenomenon is very rare because spliced introns are circular and are usually degraded right away
- Very few literature on that topic but phenomenon has been observed and is called “exon scrambling”
- Evidence on the scrambling happening in the alternative splicing process does not exist
- A very new species of circular mRNAs may be prone to exon scrambling




http://ascaffa.github.io/research.html

# Questions from last lecture:
# 3. What is the difference of RNAseq and DGE analysis results?

Hasso Plattner Institut

- RNAseq analysis results in a count table (Samples x Genes) and is analyzed by, e.g.,
  - Clustering
  - PCA

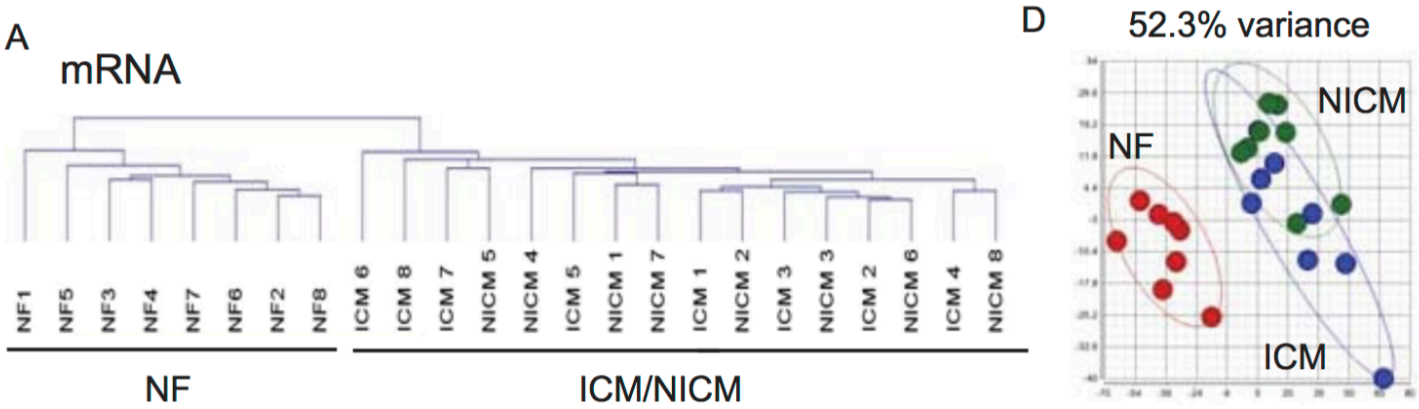
- Differential Gene Expression analysis results in a list of genes, which
  - Differ significantly in groups of samples specified by the researcher (e.g. diseased vs. healthy)
  - Need further exploration via annotation databases (e.g. gene X is involved in lipid metabolism)

- A combined analysis is given in a clustered heatmap

# Questions from last lecture:
# 3. What is the difference of RNAseq and DGE analysis results?

- RNAseq analysis results in a count table (Samples x Genes) and is analyzed by, e.g.,
  - Clustering (e.g. via a distance matrix)
  - PCA

| Gene | Sample 1 | Sample 2 |
|---|---|---|
| A | 1 | 1 |
| B | 1,5 | 1,5 |
| C | 5 | 5 |
| D | 3 | 4 |
| E | 4 | 4 |




Yang, Kai-Chien, et al. "Deep RNA sequencing reveals dynamic regulation of myocardial noncoding RNA in failing human heart and remodeling with mechanical circulatory support." Circulation (2014): CIRCULATIONAHA-113.

# Questions from last lecture:
# 3. What is the difference of RNAseq and DGE analysis results?

Hasso Plattner Institut

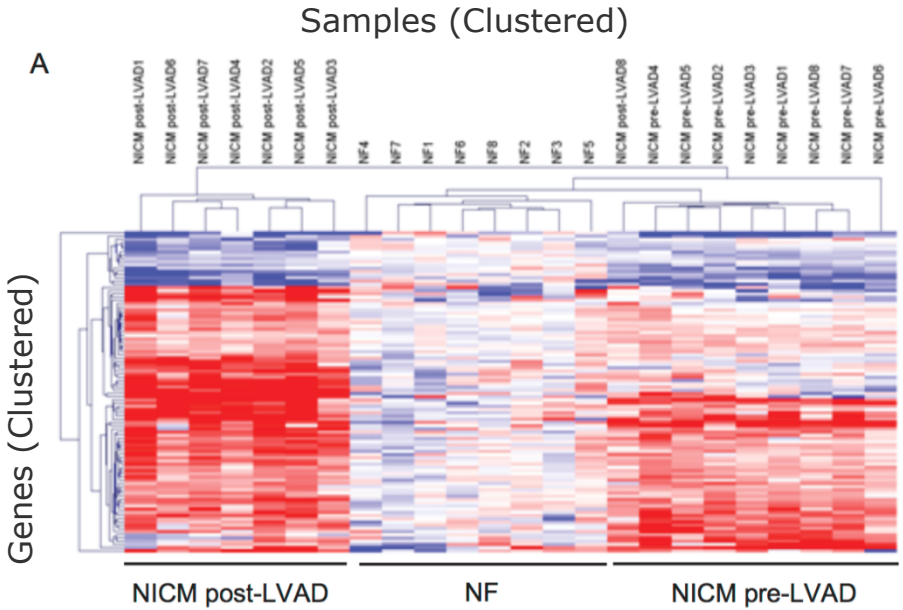
- Differential Gene Expression is a comparison of sample groups, e.g., healthy vs. diseased, as specified by the researcher
- Analysis results in a list of genes (target_id) that show differences between groups with a measure of
  - Statistical significance (p value)
  - Fold change (b) - How much more or less of the gene is expressed in diseased (condition) when compared to healthy (control)?

| target_id | pval | qval | b | se_b | mean | var |
|---|---|---|---|---|---|---|
| ENST00000263923 | 8.945596e-21 | 2.995125e-19 | -6.068921 | 0.6492325 | 2.678976 | 11.233403 |
| ENST00000510861 | 1.085725e-20 | 3.620863e-19 | -5.585684 | 0.5988514 | 2.806024 | 9.376466 |
| ENST00000005178 | 1.060794e- | 2.877073e- | -5.943046 | 0.7162216 | 2.278376 | 10.630427 |

# Questions from last lecture:
# 3. What is the difference of RNAseq and DGE analysis results?

Combined analysis in clustered heatmaps:

- Reduce set of genes to be clustered to the ones that show significant changes (DGE)
- Cluster normalized counts
  - Gene-wise and
  - Sample-wise
- Color of heatmap = normalized count value




Yang, Kai-Chien, et al. "Deep RNA sequencing reveals dynamic regulation of myocardial noncoding RNA in failing human heart and remodeling with mechanical circulatory support." Circulation (2014): CIRCULATIONAHA-113.

# Questions from last lecture:
# 4. Which elements of the count table are used for PCA?

- $G_i$ vectors span the space of the gene transcriptional responses (counts)
- $A_j$ vectors span the space of the assay (patient) expression profiles

| Gene | Sample 1 | Sample 2 |
|---|---|---|
| A | 1 | 1 |
| B | 1,5 | 1,5 |
| C | 5 | 5 |
| D | 3 | 4 |
| E | 4 | 4 |

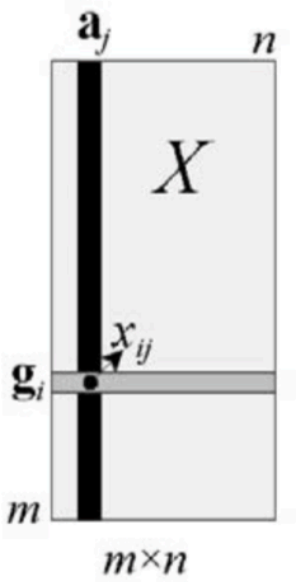



Image from: Wall, Michael E., Andreas Rechtsteiner, and Luis M. Rocha. "Singular value decomposition and principal component analysis." *A practical approach to microarray data analysis*. Springer US, 2003. 91-109.

# Singular Value Decomposition
# Basis for PCA

$$X = USV^{\mathrm{T}}$$

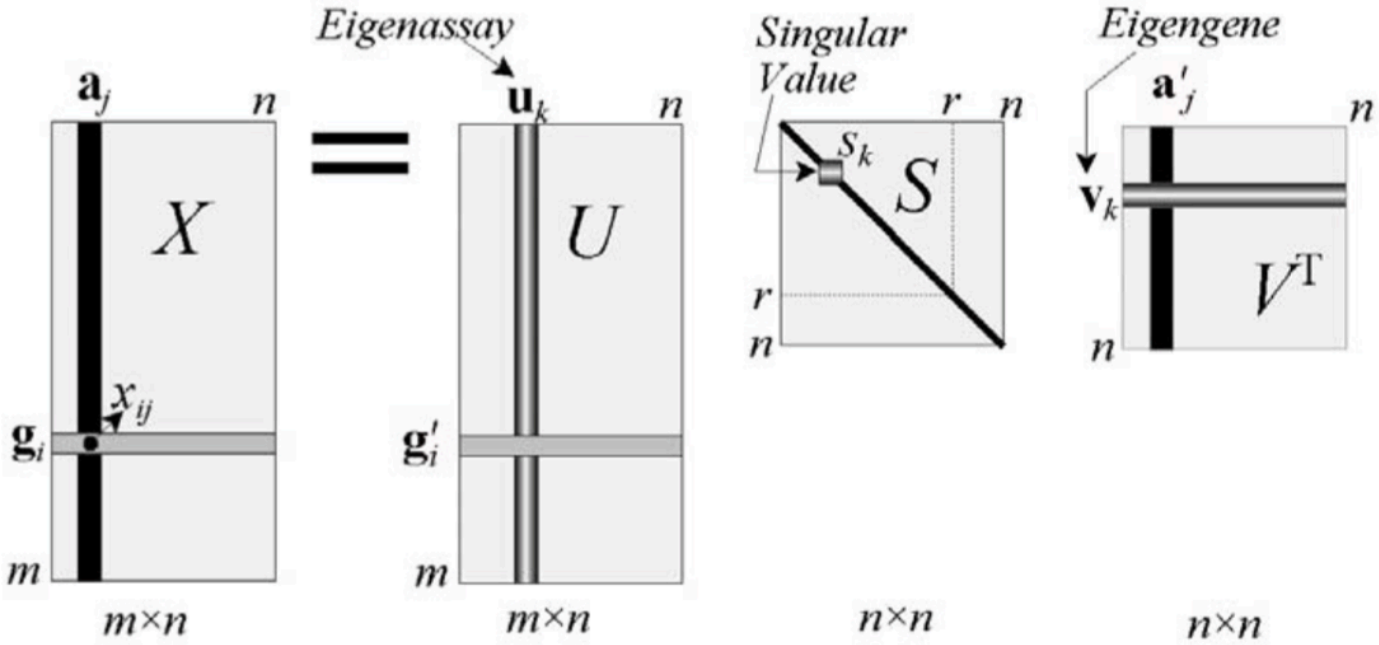



Image from: Wall, Michael E., Andreas Rechtsteiner, and Luis M. Rocha. "Singular value decomposition and principal component analysis." *A practical approach to microarray data analysis*. Springer US, 2003. 91-109.

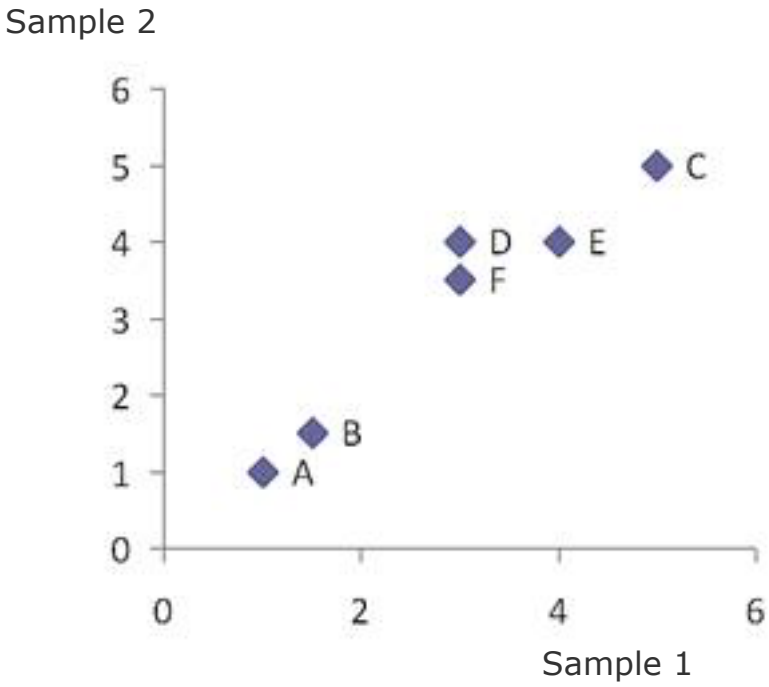
# Interpretation of SVD/PCA

- Number of first important components is sometimes associated with the number of underlying biological processes that give rise to the patterns in the data
- The last components mostly resemble noise within the data
- Loadings (S) can be used to find those genes contributing most to found variance
- Infer biological meaning/processes to the significant
  - eigenassays (in the case of diagnostic applications)
    - E.g. patients can be separated into two groups (healthy vs diseased)
  - eigengenes (in the case of systems biology applications)
    - E.g. genes that can be separated into two groups (expressed via fat metabolism vs glucose metabolism)

# Detailed Clustering Example

- Suppose a table of RNAseq count data

| Gene | Sample 1 | Sample 2 |
|---|---|---|
| A | 1 | 1 |
| B | 1,5 | 1,5 |
| C | 5 | 5 |
| D | 3 | 4 |
| E | 4 | 4 |






Altered example from http://people.revoledu.com/kardi/tutorial/Clustering/Distance%20Matrix.htm

# Calculating the Distance Matrix for Genes
# Euclidean Distance

- Euclidean distance

$$d_{ij} = \left( \sum_k \left( x_{ik} - x_{jk} \right)^2 \right)^{\frac{1}{2}}$$

- Calculate distance between gene A and B

$$d_{AB} = \left( (1-1.5)^2 + (1-1.5)^2 \right)^{\frac{1}{2}} = \sqrt{\tfrac{1}{2}} = 0.7071$$

- Another example on distance between D and F

$$d_{DF} = \left( (3-3)^2 + (4-3.5)^2 \right)^{\frac{1}{2}} = 0.5$$



Altered example from http://people.revoledu.com/kardi/tutorial/Clustering/Distance%20Matrix.htm

# Distance Matrix for all Genes
# Euclidean Distance

- Distance matrix of dimension 6x6
- Distances and therefore also the matrix are symmetric
- Matrix diagonal is 0 as the distances to each other are 0

| Dist | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | 0.00 | 0.71 | 5.66 | 3.61 | 4.24 | 3.20 |
| B | 0.71 | 0.00 | 4.95 | 2.92 | 3.54 | 2.50 |
| C | 5.66 | 4.95 | 0.00 | 2.24 | 1.41 | 2.50 |
| D | 3.61 | 2.92 | 2.24 | 0.00 | 1.00 | 0.50 |
| E | 4.24 | 3.54 | 1.41 | 1.00 | 0.00 | 1.12 |
| F | 3.20 | 2.50 | 2.50 | 0.50 | 1.12 | 0.00 |

Altered example from http://people.revoledu.com/kardi/tutorial/Clustering/Distance%20Matrix.htm

# Agglomerative Hierarchical Clustering
# Merge Genes to Become a cluster

- Merge Gene F and D to become one cluster as they have the smallest distance to each other

| Dist | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | 0.00 | 0.71 | 5.66 | 3.61 | 4.24 | 3.20 |
| B | 0.71 | 0.00 | 4.95 | 2.92 | 3.54 | 2.50 |
| C | 5.66 | 4.95 | 0.00 | 2.24 | 1.41 | 2.50 |
| D | 3.61 | 2.92 | 2.24 | 0.00 | 1.00 | 0.50 |
| E | 4.24 | 3.54 | 1.41 | 1.00 | 0.00 | 1.12 |
| F | 3.20 | 2.50 | 2.50 | 0.50 | 1.12 | 0.00 |

Smallest distance between two Genes/Clusters

Altered example from http://people.revoledu.com/kardi/tutorial/Clustering/Distance%20Matrix.htm

# Recalculate Distance Matrix

- Single linkage: The distance between two clusters is the distance between the nearest neighbors in those clusters
- Both distances can be found in the previous distance matrix, the minimum is taken as new distance for cluster $d_{(D,F)}$

$$d_{(D,F)\mapsto A} = \min\left(d_{DA}, d_{FA}\right) = \min\left(3.61, 3.20\right) = 3.20$$

**Min Distance (Single Linkage)**

| Dist | A | B | C | D, F | E |
|---|---|---|---|---|---|
| A | 0.00 | 0.71 | 5.66 | ? | 4.24 |
| B | 0.71 | 0.00 | 4.95 | ? | 3.54 |
| C | 5.66 | 4.95 | 0.00 | ? | 1.41 |
| D, F | ? | ? | ? | 0.00 | ? |
| E | 4.24 | 3.54 | 1.41 | ? | 0.00 |

Single linkage

Altered example from http://people.revoledu.com/kardi/tutorial/Clustering/Distance%20Matrix.htm

# Recalculate Distance Matrix

- Continue to recalculate as shown on the previous slide
- And start over by merging genes/clusters as in the previous steps

**Min Distance (Single Linkage)**

| Dist | A | B | C | D, F | E |
|---|---|---|---|---|---|
| A | 0.00 | 0.71 | 5.66 | 3.20 | 4.24 |
| B | 0.71 | 0.00 | 4.95 | 2.50 | 3.54 |
| C | 5.66 | 4.95 | 0.00 | 2.24 | 1.41 |
| D, F | **3.20** | **2.50** | **2.24** | 0.00 | 1.00 |
| E | 4.24 | 3.54 | 1.41 | **1.00** | 0.00 |

Altered example from http://people.revoledu.com/kardi/tutorial/Clustering/Distance%20Matrix.htm

# Series of all Distance Matrices

| Dist | A | B | C | D, F | E |
|---|---|---|---|---|---|
| A | 0.00 | 0.71 | 5.66 | 3.20 | 4.24 |
| B | 0.71 | 0.00 | 4.95 | 2.50 | 3.54 |
| C | 5.66 | 4.95 | 0.00 | 2.24 | 1.41 |
| D, F | **3.20** | **2.50** | **2.24** | 0.00 | 1.00 |
| E | 4.24 | 3.54 | 1.41 | **1.00** | 0.00 |

| Dist | (A,B) | C | (D, F), E |
|---|---|---|---|
| (A,B) | 0.00 | 4.95 | 2.50 |
| C | 4.95 | 0.00 | 1.41 |
| (D, F), E | **2.50** | **1.41** | 0.00 |

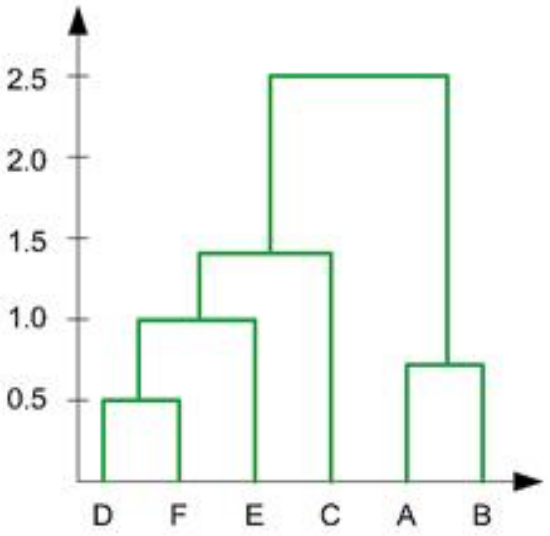
| Dist | A,B | C | (D, F) | E |
|---|---|---|---|---|
| A,B | 0 | 4.95 | 2.50 | 3.54 |
| C | **4.95** | 0 | 2.24 | 1.41 |
| (D, F) | **2.50** | 2.24 | 0 | 1.00 |
| E | **3.54** | 1.41 | 1.00 | 0 |

| Dist | (A,B) | (D, F), E),C |
|---|---|---|
| (A,B) | 0.00 | 2.50 |
| ((D, F), E),C | **2.50** | 0.00 |

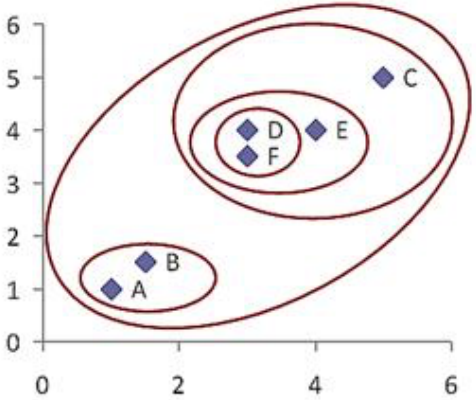


Altered example from http://people.revoledu.com/kardi/tutorial/Clustering/Distance%20Matrix.htm

# Resulting Dendrogram and (Concentric) Clusters



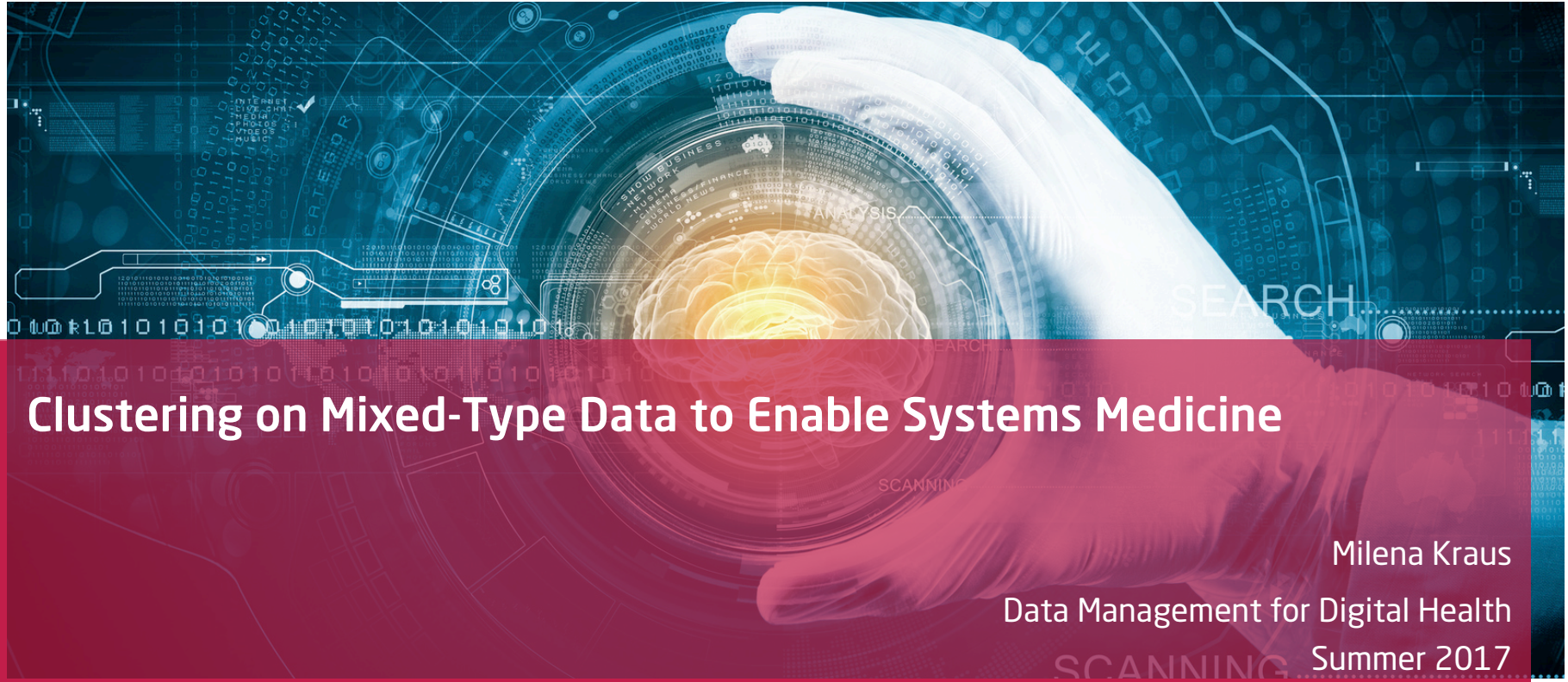

- Cluster(D,F) = 0.5
- Cluster(A,B) = 0.71
- Cluster((D,F,)E) = 1.00
- Cluster(((D,F), E), C) = 1.41
- Cluster((((D,F), E), C), (A,B)) = 2.5




Altered example from http://people.revoledu.com/kardi/tutorial/Clustering/Distance%20Matrix.htm

Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam



# Clustering on Mixed-Type Data to Enable Systems Medicine

Milena Kraus

Data Management for Digital Health

Summer 2017

# Motivation

- Many “simple” clustering strategies are based on a similarity measure, that is calculated through numerical values (Euclidean or Manhattan distance)
- Systems medicine datasets are heterogeneous and contain numerical and categorical data, e.g., count data from RNAseq experiments or ethnicity from patient characteristics

**How to enable a strategy of a combined, holistic analysis of the complete mixed-type systems medicine data set?**

# Multivariate Distance Matrix Through Gower's Similarity Coefficient

- Applied on datasets containing continuous, binary and categorical variables at the same time.
- Gower's General Similarity Coefficient $S_{ij}$ compares two cases i and j and is defined as follows

$$S_{ij} = \frac{\sum_{k}^{n} w_{ijk} S_{ijk}}{\sum_{k}^{n} w_{ijk}}$$

- Where:
  $S_{ijk}$ denotes the contribution provided by the k-th variable, and
  $w_{ijk}$ is usually 1 or 0 depending if the comparison is valid for the k-th variable.

Adapted from http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/MetQ/Talk6.pdf

# Gower's Similarity Coefficient
# Continuous Variables

- Gower similarity defines the value of $S_{ijk}$ for continuous variables as follows:

$$S_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{r_k}$$

- Where $r_k$ is the **range** of values for the k−th variable.

Adapted from http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/MetQ/Talk6.pdf

# Gower's Similarity Coefficient
# Nominal Variables

- The value of $S_{ijk}$ for nominal variables is 1 if $x_{ik} = x_{jk}$ or 0 if $x_{ik}$ != $xj_k$.
- Thus $S_{ijk}$= 1 if cases i and j have the same state for attribute k, or 0 if they have different states, and
- $w_{ijk}$ = 1 if both cases have observed states for attribute k.

$$S_{ij} = \frac{\sum_{k}^{n} w_{ijk} S_{ijk}}{\sum_{k}^{n} w_{ijk}}$$

Adapted from http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/MetQ/Talk6.pdf

# Gower's Similarity Coefficient
# Binary Variables

- For a binary variable (or dichotomous character), the Gower similarity defines the components of similarity and the weight according to the table

| Value of attribute k | | | | |
|---|---|---|---|---|
| Case i | + | + | - | - |
| Case j | + | - | + | - |
| $S_{ijk}$ | 1 | 0 | 0 | 0 |
| $w_{ijk}$ | 1 | 1 | 1 | 0 |

$$S_{ij} = \frac{\sum_{k}^{n} w_{ijk} S_{ijk}}{\sum_{k}^{n} w_{ijk}}$$

- where + denotes that attribute k is present and - denotes that attribute k is absent.

Adapted from http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/MetQ/Talk6.pdf

# Summary
# Gower's Similarity Coefficient

- Converts numerical, binary and nominal data types into numerical data
- The split of variable into S and w enables an automatic mechanism to ignore missing data
- No coverage of ordinal data
- Resulting distance matrix can then be used in distance-based algorithms, e.g., hierarchical clustering

# Hamming Distance and Ranking

- Hamming distance can be applied to categorical data
- Categorical data is converted to an array of binary data (present/absent feature)
- d(x, y) between two vectors x, y ∈ F (n) is the number of coefficients in which they differ, e.g.
  - in F(5) d(00111, 11001) = 4
  - in F(4) d(HAUS, MAUT) = 2
- Ordinal data are categorical values following an order, e.g., good, stable, bad
- Ordinal data can be converted via their rank

→ For examples on how to use Hamming distance and ranking, go through the following example

# Before we start with the example…

- Please consider the difference between distance and similarity
  - Gower's similarity coefficient should be high for high similarity between observations
  - The following examples uses distances and thus is low for similar observations
- Nomenclature:
  - Nominal = Categorical
  - The expression "numerical" is mostly used for continuous data
  - Ordinal and binary are usually numbers that are not continuous

# Example: Create a Multivariate Distance Matrix
# Data Description

- Time on Heart-Lung-Support during ventricle valve replacement in minutes – numerical data
- Implanted valve model is the model of valve the surgeon chose during surgery and can be exclusively 1 out of 4 models (1, 2, 3, 4) – exclusive categorical data
- Pre-surgery drugs are medications that the patient took before he/she underwent surgery. The patient may have taken a combination of different drugs, e.g., beta-blockers, ACE inhibitors, pain medication, diuretics, statins, other medications) – categorical values (not exclusive)
- Quality of Life (QoL) is a score cumulated through an interview with a patient. It is a value on an ordinal scale with five values: -2 = very bad, -1 = bad, 0 = ok, 1 = good, 2 = very good – ordinal data
- Gender is encoded as binary data

# Example: Create a Multivariate Distance Matrix Data Set

| Patient | Time | Valve | Drugs | QoL | Gender |
|---|---|---|---|---|---|
| A | 30 | 1 | 1, 2, 3 | 2 | M |
| B | 30 | 3 | 4, 6 | 1 | F |
| C | 60 | 2 | 1, 2 | 2 | M |
| D | 45 | 1 | 5 | -1 | M |

The given example is adapted from

http://people.revoledu.com/kardi/tutorial/Similarity/MutivariateDistance.html

# Example: Create a Multivariate Distance Matrix
# Algorithm Overview

Algorithm:

1. Convert data into coordinates based on measurement scale
2. Determine distance matrix for each feature variable based on coordinate
3. Normalize the distance matrix into range of [0, 1]
4. Aggregate the distance matrix

# Example: Create a Multivariate Distance Matrix
# 1. Step: Convert into Coordinates

- Numerical data stays as it is
  - Time on Heart-Lung-Support

# Example: Create a Multivariate Distance Matrix
# 1. Step: Convert into Coordinates

- Numerical Data stays as it is
- Binary data is converted to 0 and 1
  - Gender (M/F) is converted to 0 and 1

# Example: Create a Multivariate Distance Matrix
# 1. Step: Convert into Coordinates

- Numerical Data stays as it is
- Binary data is converted to 0 and 1
- Ordinal Data is converted to values between 0 and 1 via their rank
  - Quality of Life

| | | | | | | |
|---|---|---|---|---|---|---|
| Original index | -2 | -1 | 0 | 1 | 2 | = i |
| Converted to rank | 1 | 2 | 3 | 4 | 5 | = r |
| | | | | | | R = max (r) = 5 |
| Normalized rank | 0 | 1/4 | 1/2 | 3/4 | 1 | $x = \frac{r-1}{R-1}$ |

# Example: Create a Multivariate Distance Matrix
# 1. Step: Convert into Coordinates

- Numerical Data stays as it is
- Binary data is converted to 0 and 1
- Ordinal Data is converted to values between 0 and 1 via its rank
- Nominal, exclusive data is represented as a combination of multiple binary dummy variables, needed dummy variables are smaller than the number of values
  - Implanted valve model (1, 2, 3, 4) can be represented via two binary dummy variables (DV1, DV2)

| Valve model | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| DV1 | 0 | 1 | 0 | 1 |
| DV2 | 0 | 0 | 1 | 1 |

# Example: Create a Multivariate Distance Matrix
# 1. Step: Convert into Coordinates

- Numerical Data stays as it is
- Binary data is converted to 0 and 1
- Ordinal Data is converted to values between 0 and 1 via its rank
- Nominal, exclusive data is represented as a combination of multiple binary dummy variables, needed dummy variables are smaller than the number of values
- Nominal, non-exclusive data is represented by one binary dummy variable per value
  - Pre-surgery drugs will be converted, e.g., from “ACE Inhibitor, Beta-Blocker, pain medication” to (1, 1, 1, 0, 0 ,0)

# Example: Create a Multivariate Distance Matrix Converted Data Set

| Patient | Time | Valve | Drugs | QoL | Gender |
|---|---|---|---|---|---|
| A | 30 | (0,0) | (1,1,1,0,0,0) | 1 | 1 |
| B | 30 | (0,1) | (0,0,0,1,0,1) | 3/4 | 0 |
| C | 60 | (1,0) | (1,1,0,0,0,0) | 1 | 1 |
| D | 45 | (0,0) | (0,0,0,0,1,0) | 1/4 | 1 |

# Example: Create a Multivariate Distance Matrix Algorithm Overview

Algorithm:

1. Convert data into coordinates based on measurement scale
2. Determine distance matrix for each feature variable based on coordinate
3. Normalize the distance matrices into range of [0, 1]
4. Aggregate the distance matrix

# Example: Create a Multivariate Distance Matrix
# 2. Step: Calculate all Distances Separately

- Numerical data
- In this example we use Manhattan distance for numerical data
- Normalization by dividing through maximum distance = 30 min

Distance matrix

| Time | A | B | C | D |
|---|---|---|---|---|
| A | 0 | 0 | 30 | 15 |
| B | 0 | 0 | 30 | 15 |
| C | 30 | 30 | 0 | 15 |
| D | 15 | 15 | 15 | 0 |
| max | 30 | | min | 0 |

| Patient | Time |
|---|---|
| A | 30 |
| B | 30 |
| C | 60 |
| D | 45 |

Normalized distance matrix

| Time | A | B | C | D |
|---|---|---|---|---|
| A | 0 | 0 | 1 | 0,5 |
| B | 0 | 0 | 1 | 0,5 |
| C | 1 | 1 | 0 | 0,5 |
| D | 0,5 | 0,5 | 0,5 | 0 |

# Example: Create a Multivariate Distance Matrix
# 2. Step: Calculate all Distances Separately

- Binary data in dummy variables
- Hamming distance for arrays of binary data
- Normalization by dividing through maximum distance = 2

Distance matrix

| **Mode** | **A** | **B** | **C** | **D** |
|---|---|---|---|---|
| **A** | 0 | 1 | 1 | 0 |
| **B** | 1 | 0 | 2 | 1 |
| **C** | 1 | 2 | 0 | 1 |
| **D** | 0 | 1 | 1 | 0 |
| max | 2 | | min | 0 |

| Patient | Valve |
|---|---|
| A | (0,0) |
| B | (0,1) |
| C | (1,0) |
| D | (0,0) |

Normalized distance matrix

| **Mode** | A | B | C | D |
|---|---|---|---|---|
| A | 0 | 0,5 | 0,5 | 0 |
| B | 0,5 | 0 | 1 | 0,5 |
| C | 0,5 | 1 | 0 | 0,5 |
| D | 0 | 0,5 | 0,5 | 0 |

# Example: Create a Multivariate Distance Matrix
# 2. Step: Calculate all Distances Separately

- Binary data
- Hamming distance for arrays of binary data
- Normalization by dividing through maximum distance = 6

Distance matrix

| Drugs | A | B | C | D |
|---|---|---|---|---|
| A | 0 | 5 | 1 | 4 |
| B | 5 | 0 | 4 | 3 |
| C | 1 | 4 | 0 | 3 |
| D | 4 | 3 | 3 | 0 |
| max | 6 | | min | 0 |

| Patient | Drugs |
|---|---|
| A | (1,1,1, 0,0,0) |
| B | (0,0,0, 1,0,1) |
| C | (1,1,0, 0,0,0) |
| D | (0,0,0, 0,1,0) |

Normalized distance matrix

| Drugs | A | B | C | D |
|---|---|---|---|---|
| A | 0,00 | 0,83 | 0,17 | 0,67 |
| B | 0,83 | 0,00 | 0,67 | 0,50 |
| C | 0,17 | 0,67 | 0,00 | 0,50 |
| D | 0,67 | 0,50 | 0,50 | 0,00 |

# Example: Create a Multivariate Distance Matrix
# 2. Step: Calculate all Distances Separately

- Binary data
- Hamming distance for arrays of binary data
- Normalization was performed via ranking procedure

Distance matrix

| QoL | A | B | C | D |
|---|---|---|---|---|
| A | 0 | 0,25 | 0 | 0,75 |
| B | 0,25 | 0 | 0,25 | 0,5 |
| C | 0 | 0,25 | 0 | 0,75 |
| D | 0,75 | 0,5 | 0,75 | 0 |
| max | 1 | | min | 0 |

| Patient | QoL |
|---|---|
| A | 1 |
| B | 3/4 |
| C | 1 |
| D | 1/4 |

# Example: Create a Multivariate Distance Matrix
# 2. Step: Calculate all Distances Separately

- Binary data
- Hamming distance
- Normalization not needed for true binary data

Distance matrix

| **Gender** | **A** | **B** | **C** | **D** |
|---|---|---|---|---|
| **A** | 0 | 1 | 0 | 0 |
| **B** | 1 | 0 | 1 | 1 |
| **C** | 0 | 1 | 0 | 0 |
| **D** | 0 | 1 | 0 | 0 |
| max | 1 | | min | 0 |

| Patient | Gender |
|---|---|
| A | 1 |
| B | 0 |
| C | 1 |
| D | 1 |

# Example: Create a Multivariate Distance Matrix
# 3. Step: Aggregate all Distances into one Matrix

- Aggregation is the weighted average of the distance
- We assume that all variables have the same weight
- Average is achieved by dividing through 5 (variables)

Sum of all distance matrices

| Sum | A | B | C | D |
|---|---|---|---|---|
| A | 0,00 | 2,58 | 1,67 | 1,92 |
| B | 2,58 | 0,00 | 3,92 | 3,00 |
| C | 1,67 | 3,92 | 0,00 | 2,25 |
| D | 1,92 | 3,00 | 2,25 | 0,00 |

Averaged distance matrix

| Average | A | B | C | D |
|---|---|---|---|---|
| A | 0,00 | 0,52 | 0,33 | 0,38 |
| B | 0,52 | 0,00 | 0,78 | 0,60 |
| C | 0,33 | 0,78 | 0,00 | 0,45 |
| D | 0,38 | 0,60 | 0,45 | 0,00 |

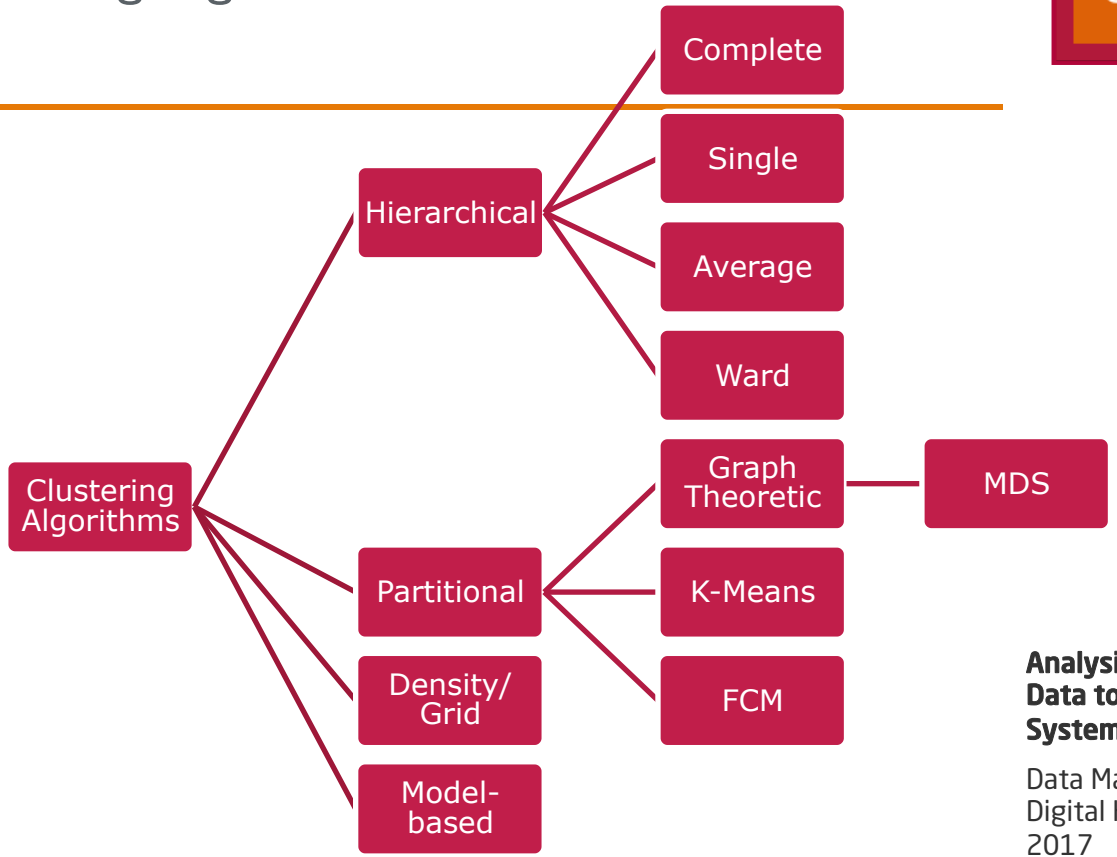
→ Proceed with distance based clustering Algorithms

# Summary
# Distance-based Clustering on Mixed-type Data

- All variables are converted to numerical data, e.g.,
  - through Gower's Similarity Measure
  - Or other methods like Hamming distance or ranking
- Normalization (mostly) to values between [0,1]
- Distance matrix can then be used in clustering
- Weights of variables need to be known or adjusted by the user

# Classification of Clustering Algorithms

- Very common approaches are hierarchical and partitional clustering techniques
- Model-based approaches gain more and more attention and have been extended to also account for mixed-type data



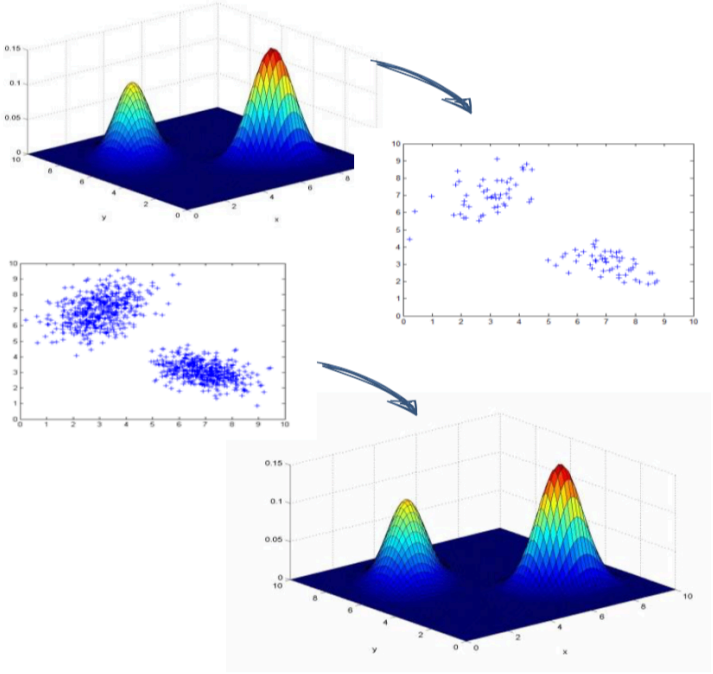

Adapted from http://slideplayer.com/slide/9351875/

# Model-based Clustering Methods

- Attempt to optimize the fit between the data and some mathematical model or density
- Assumption: Data are generated by a mixture of underlying probability distributions
- Techniques:
  - Expectation-Maximization (EM)
  - Conceptual Clustering
  - Neural Networks Approach

https://de.slideshare.net/Krish_ver2/35-model-based-clustering

# Expectation Maximization (EM)



- Statistical approach for finding maximum likelihood estimates of parameters in probabilistic models
- EM as clustering algorithm underlying assumption:
  - Observations are drawn from one of several components of a mixture distribution.
- Main idea:
  - Define clusters as probability distributions

→ each object has a certain probability of belonging to each cluster

  - Iteratively improve the parameters of each distribution (e.g. center, "width" and "height" of a Gaussian distribution) until some quality threshold is reached

Prof. Emmanuel Müller – Big Data Analytics lecture (WS 2016)

# Metabolic Syndrome (MetS)

- Metabolic syndrome is associated with the risk of developing cardiovascular disease and type 2 diabetes.
- In the USA, about a quarter of the adult population have metabolic syndrome
- prevalence increases with age, with racial ethnic minorities being particularly affected.
- Insulin resistance, metabolic syndrome, and prediabetes are closely related to one another and have overlapping aspects.
- Cause might be an underlying disorder of energy utilization and storage
- The cause of the syndrome is an area of ongoing medical research.

https://en.wikipedia.org/wiki/Metabolic_syndrome

# Metabolic Syndrome (MetS)
# Diagnosis Criteria

Table 1: A person with 3 or more of the abnormalities listed below is diagnosed as having the MetS.

| | |
|---|---|
| Fasting glucose concentration | $\geq$5.5 mmol $l^{-1}$ or treatment of previously diagnosed diabetes. |
| Serum TAG concentration | $\geq$ 1.5 mmol $l^{-1}$ or treatment of previously diagnosed lipidemia. |
| Serum HDL-c concentration | $<$ 1.04 mmol $l^{-1}$ (Men)<br>$<$ 1.29 mmol $l^{-1}$ (Women) |
| Blood pressure | Systolic BP $\geq$ 130 mm Hg, Diastolic BP $\geq$ 85 mm Hg or treatment of previously diagnosed hypertension. |
| Waist Circumference | $>$ 94 cm (Men), $>$ 80 cm (Women) |

McParland, Damien, et al. "Clustering high dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data." *arXiv preprint arXiv:1606.05107* (2016).

# Clustering on Mixed-type Data
# McParland et al. 2016

Hasso Plattner Institut

## Clustering high dimensional mixed data to uncover sub-phenotypes:joint analysis of phenotypic and genotypic data

Damien McParland, Catherine M. Phillips, Lorraine Brennan, Helen M. Roche, Isobel Claire Gormley

*(Submitted on 16 Jun 2016)*

The LIPGENE-SU.VI.MAX study, like many others, recorded high dimensional continuous phenotypic data and categorical genotypic data. LIPGENE-SU.VI.MAX focuses on the need to account for both phenotypic and genetic factors when studying the metabolic syndrome (MetS), a complex disorder that can lead to higher risk of type 2 diabetes and cardiovascular disease. Interest lies in clustering the LIPGENE-SU.VI.MAX participants into homogeneous groups or sub-phenotypes, by jointly considering their phenotypic and genotypic data, and in determining which variables are discriminatory.

A novel latent variable model which elegantly accommodates high dimensional, mixed data is developed to cluster LIPGENE-SU.VI.MAX participants using a Bayesian finite mixture model. A computationally efficient variable selection algorithm is incorporated, estimation is via a Gibbs sampling algorithm and an approximate BIC-MCMC criterion is developed to select the optimal model.

Two clusters or sub-phenotypes (`healthy' and `at risk') are uncovered. A small subset of variables is deemed discriminatory which notably includes phenotypic and genotypic variables, highlighting the need to jointly consider both factors. Further, seven years after the LIPGENE-SU.VI.MAX data were collected, participants underwent further analysis to diagnose presence or absence of the MetS. The two uncovered sub-phenotypes strongly correspond to the seven year follow up disease classification, highlighting the role of phenotypic and genotypic factors in the MetS, and emphasising the potential utility of the clustering approach in early screening. Additionally, the ability of the proposed approach to define the uncertainty in sub-phenotype membership at the participant level is synonymous with the concepts of precision medicine and nutrition.

# McParland et al. 2016
# Data Set – LIPGENE-SU.VI.MAX - Original

- 1754 participants
- 827 variables in total
- 26 continuous variables, e.g., fasting glucose concentration, waist circumference and plasma fatty acids
- 801 categorical SNP variables, e.g.., rs512535 of the APOB gene which is represented by three genotypes: AA, GG or AG in the data



McParland, Damien, et al. "Clustering high dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data." *arXiv preprint arXiv:1606.05107* (2016).

# McParland et al. 2016
# Data Set – LIPGENE-SU.VI.MAX - Cleaning

- Data cleansing included
  - Dismissal of of sparsely represented variables or participants (missing data)
  - Merging of recessive homozygous with the corresponding heterozygous genotypes, as they represent the same function
    - Two genotypes → binary variable [0,1]
    - Three genotypes → two dummy variables that define the category when combined (similar to chart 42)
  - Standardization/Normalization of continuous data

McParland, Damien, et al. "Clustering high dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data." *arXiv preprint arXiv:1606.05107* (2016).

# McParland et al. 2016
# Data Set – LIPGENE-SU.VI.MAX – Final Set

- Final set of data:
  - 505 participants x 738 variables
  - 225 “healthy” participants and 280 MetS patients
  - 26 continuous clinical measurements
  - 371 binary SNPs and 341 nominal SNPs



McParland, Damien, et al. "Clustering high dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data." *arXiv preprint arXiv:1606.05107* (2016).

# Data Modeling and Clustering through a Probabilistic Approach

Hasso Plattner Institut

Data is modeled via:

- Factor Analysis Model for numerical (continuous) data
- Item Response Theory (IRT) model for binary data
- Multinomial probit type model for nominal data

McParland, Damien, et al. "Clustering high dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data." *arXiv preprint arXiv:1606.05107* (2016).

# Mathematical Models

- Continuous data

$$\underline{z}_i = \underline{\mu} + \Lambda \underline{\theta}_i + \underline{\epsilon}_i$$

- Binary data

$$z_{ij} | \underline{\theta}_i \sim N(\mu_j + \underline{\lambda}_j^T \underline{\theta}_i, 1).$$

- Categorical data

$$\underline{z}_{ij} | \underline{\theta}_i \sim \mathrm{MVN}_2(\underline{\mu}_j + \Lambda_j \underline{\theta}_i, \mathbf{I})$$

All follow a common structure and can be combined in one model

# Data Modeling and Clustering through a Probabilistic Approach

Hasso Plattner Institut

Data is modeled via:

- Factor Analysis Model for numerical (continuous) data
- Item Response Theory (IRT) model for binary data
- Multinomial probit type model for nominal data

All follow a common structure and can be combined in one model

Clustering is performed via:

- Finite mixture model:
  - Maximum Likelihood Estimation (ME) via Bayesian Inference and Markov Chain Monte Carlo (MCMC) sampling model balances goodness of fit of clustering strategy with need for simplicity and calculates the optimal number of needed clusters to explain the data

McParland, Damien, et al. "Clustering high dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data." *arXiv preprint arXiv:1606.05107* (2016).

# Online Variable Selection

- Dataset still contains a large amount of variables of which a lot only show very little variability across all clusters
- Information on these variables to not contribute to the overall clustering result
- Online variable selection is performed while the clustering algorithm runs
- VR = Variance Ratio = Variance of all participants currently in the cluster/Variance of the variable across all participants
- VR is specified by a threshold
  - Low VR → Variable is discriminates between clusters
  - Large VR → Variable takes similar values across all clusters

McParland, Damien, et al. "Clustering high dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data." *arXiv preprint arXiv:1606.05107* (2016).

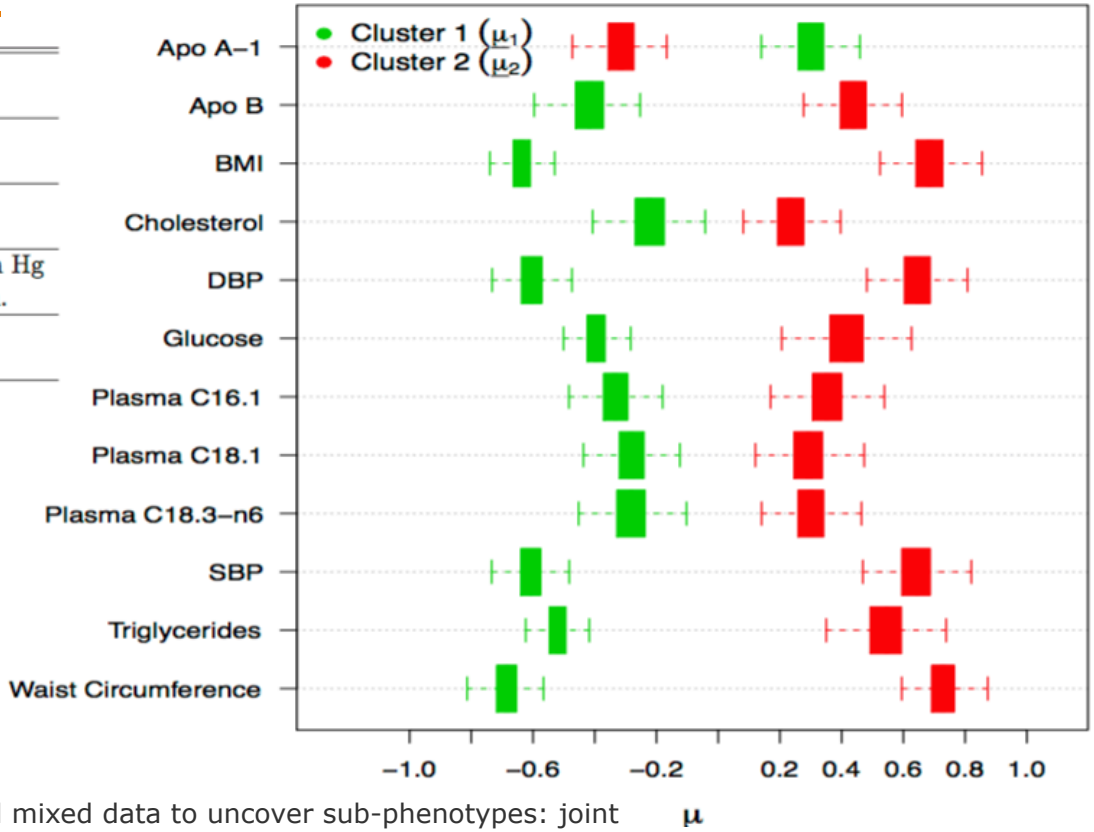
# McParland et al. 2016
# Results

- Reduction of considered variables from 738 to 25
- Clustering reveals a “healthy” and an “at-risk” sub-phenotype
- All data types were represented highlighting the importance of considering all available data
  - 12 continuous clinical measurements
  - 2 binary SNPs
  - 11 nominal SNPs
- Some of the identified 25 variables were already known and described by the literature, some were unknown and thus need further investigation (e.g. novel SNPs)

McParland, Damien, et al. "Clustering high dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data." *arXiv preprint arXiv:1606.05107* (2016).

# McParland et al. 2016
# Results

| | |
|---|---|
| Fasting glucose concentration | $\geq$5.5 mmol $l^{-1}$ or treatment of previously diagnosed diabetes. |
| Serum TAG concentration | $\geq$ 1.5 mmol $l^{-1}$ or treatment of previously diagnosed lipidemia. |
| Serum HDL-c concentration | $<$ 1.04 mmol $l^{-1}$ (Men) $<$ 1.29 mmol $l^{-1}$ (Women) |
| Blood pressure | Systolic BP $\geq$ 130 mm Hg, Diastolic BP $\geq$ 85 mm Hg or treatment of previously diagnosed hypertension. |
| Waist Circumference | $>$ 94 cm (Men), $>$ 80 cm (Women) |




McParland, Damien, et al. "Clustering high dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data." *arXiv preprint arXiv:1606.05107* (2016).

Table 3: Characteristics of the set of 13 binary and nominal SNP variables deemed to be discriminatory.(Source: NCBI SNP data base http://www.ncbi.nlm.nih.gov/SNP/)

| Gene | SNP | SNP type | Chromosome | Associated biological pathway |
|---|---|---|---|---|
| *ADD1* | rs17777371 | Adducin 1 | Flanking_3UTR chromosome 4 | Blood pressure regulation |
| *APOB* | rs512535 | Apolipoprotein B | Intronic chromosome 2 | Lipid metabolism |
| *APOL1* | rs136147 | Apolipoprotein L1 | Intronic chromosome 22 | Lipid metabolism |
| *CETP* | rs4784744 | Cholesterol ester transfer protein | Intronic chromosome 16 | Lipid metabolism |
| *FABP1* | rs2970901 | Fatty acid binding protein 1, liver | Flanking_5UTR chromosome 2 | Lipid metabolism |
| *GYS1* | rs2270938 | Glycogen synthase 1 | Intronic chromosome 19 | Glucose homeostasis |
| *INSIG1* | rs9770068 | Insulin Induced Gene 1 | Intronic chromosome 7 | Lipid metabolism, innate immunity. |
| *LRP2* | rs2544377 | LDL receptor related protein 2 | Intronic chromosome 2 | Lipid metabolism |
| *OLR1* | rs1050289 | Oxidized low density lipoprotein (lectin-like) receptor 1 | 3UTR chromosome 12 | Lipid metabolism |
| *SLC25A14* | rs2235800 | Solute Carrier Family 25 (Mitochondrial Carrier, Brain), Member 14 or UCP5 | Intronic x chromosome | Oxidative phosphorylation |
| *SLC27A6* | rs185411 | Solute Carrier Family 27 (Fatty acid transported), member 6 | Intronic chromosome 5 | Lipid metabolism |
| *SLC6A14* | rs2071877 | Solute carrier family 6 (amino acid transporter), member 14 | Intronic x chromosome | Amino acid transporter |
| *THYN1* | rs570113 | Thymocyte nuclear protein 1 | Intronic chromosome 11 | Amino acid metabolism |

McParland, Damien, et al. "Clustering high dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data." *arXiv preprint arXiv:1606.05107* (2016).

# McParland et al. 2016
# Results

Table 2: Cross tabulation of sub-phenotype membership (based on fitting the MFA-MD model to the initial phenotypic and genotypic data) and MetS diagnosis (based on the diagnosis criterion in Table 1 on seven year follow up phenotypic data only). The Rand index is 0.73 (adjusted Rand index = 0.46).

| | | Follow up data | |
|---|---|---|---|
| | | Healthy | MetS |
| Initial data | Cluster 1 ('Healthy') | 220 | 42 |
| | Cluster 2 ('At risk') | 39 | 204 |

Table 4: Cross tabulation of MetS diagnoses from initial and follow up data. The Rand index is 0.69 (adjusted Rand index is 0.38)

| | | Follow up data | |
|---|---|---|---|
| | | Healthy | MetS |
| Initial data | Healthy | 194 | 31 |
| | MetS | 65 | 215 |

McParland, Damien, et al. "Clustering high dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data." *arXiv preprint arXiv:1606.05107* (2016).

# McParland et al. 2016
# Results

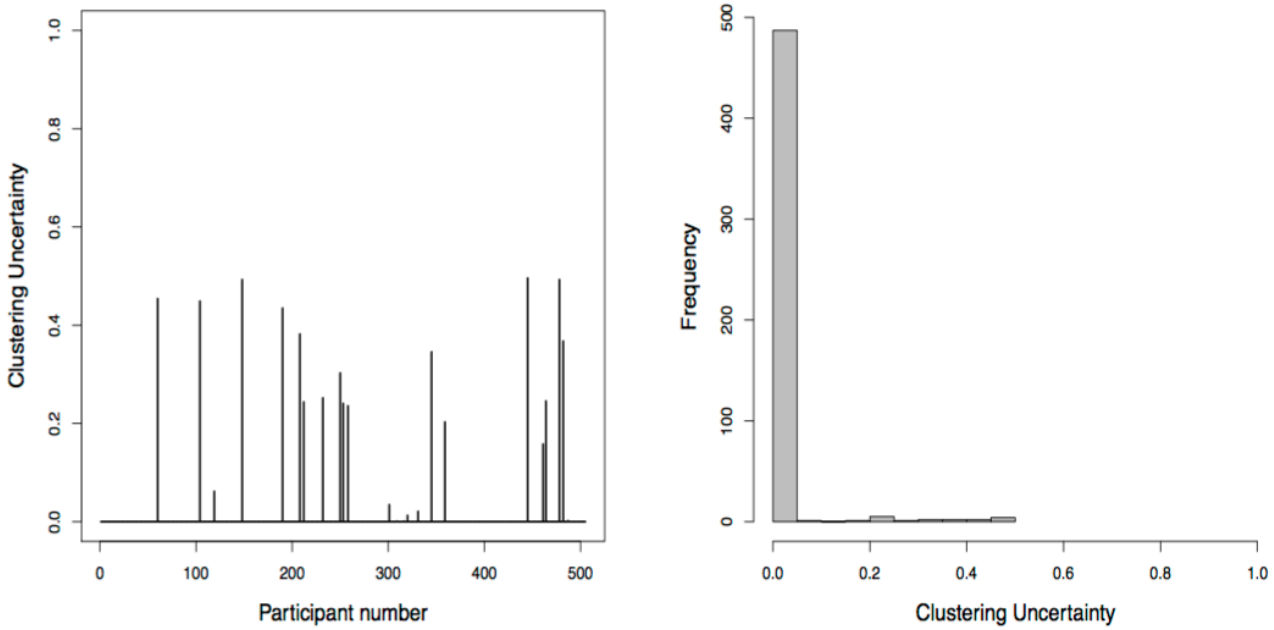



Figure 5: (a) The participant specific clustering uncertainties and (b) the histogram of the clustering uncertainties across all participants, under the optimal MFA-MD model.

# Summary
# Model-based Clustering

- Model-based clustering approaches were published more frequently in the last years
- Models are based on complex mathematical and statistical methods
- McParland et al. show that
  - It is beneficial to include all available information into the clustering process to infer all informative variables
  - Propose a valuable model for Metabolic Syndrome to explore, e.g., risk factors, genetic components and the value of clinical measurements
  - Patients that are not clearly represented by one cluster may need special attention to enable personalized care
- Model-based clusters can be studied for multiple purposes and enable a very deep exploration of the disease, patients and their characteristics

# Thank you!
# Any further Questions?

Hasso Plattner Institut