



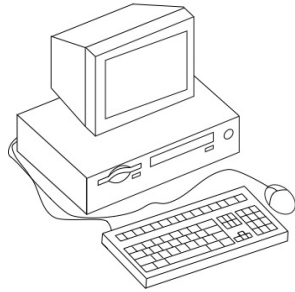
Finnish Institute of
Occupational Health

Interactive Visual Data Exploration with Subjective Feedback

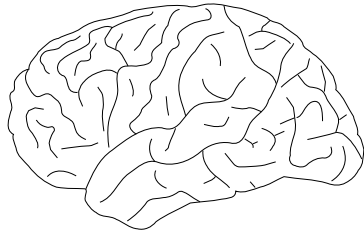
Kai Puolamäki, Bo Kang, Jefrey Lijffijt, Tijl De Bie

ECML PKDD 2016

21 September 2016

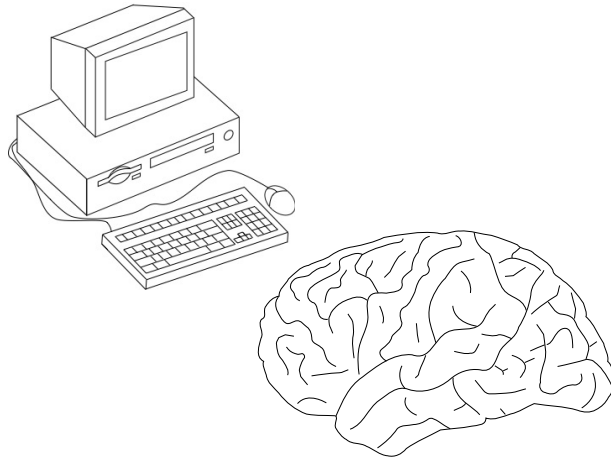


- + handling large data
- + handling high-dimensional data
- + making analytic comparisons
- identifying patterns truly relevant for the user
- black boxes, incomprehensible for the user



- + huge background knowledge
- + spotting patterns
- handling large, high-dimensional data
- making analytic comparisons

Computer figure by Christophe Dang Ngoc Chan, licensed under GFDL,
https://commons.wikimedia.org/wiki/File:Ordinateur_table_1990.svg



- + **handling large data**
- + **handling high-dimensional data**
- + **making analytic comparisons**
- ~~identifying patterns truly relevant for the user~~
- ~~black boxes, incomprehensible for the user~~

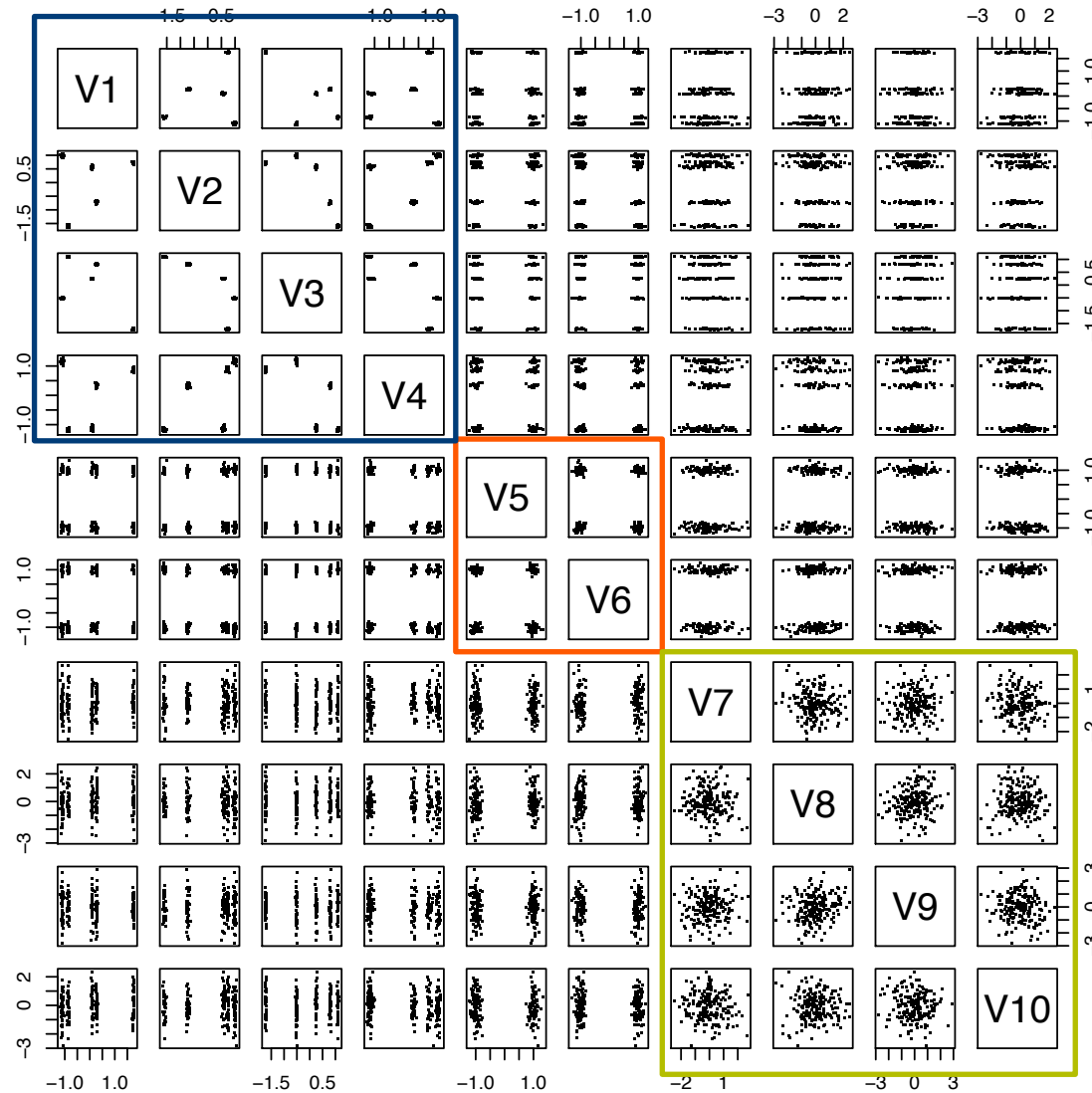
- + **huge background knowledge**
- + **spotting patterns**
- ~~handling large, high-dimensional data~~
- ~~making analytic comparisons~~

Computer figure by Christophe Dang Ngoc Chan, licensed under GFDL,
https://commons.wikimedia.org/wiki/File:Ordinateur_table_1990.svg

Toy example: 10-dimensional dataset

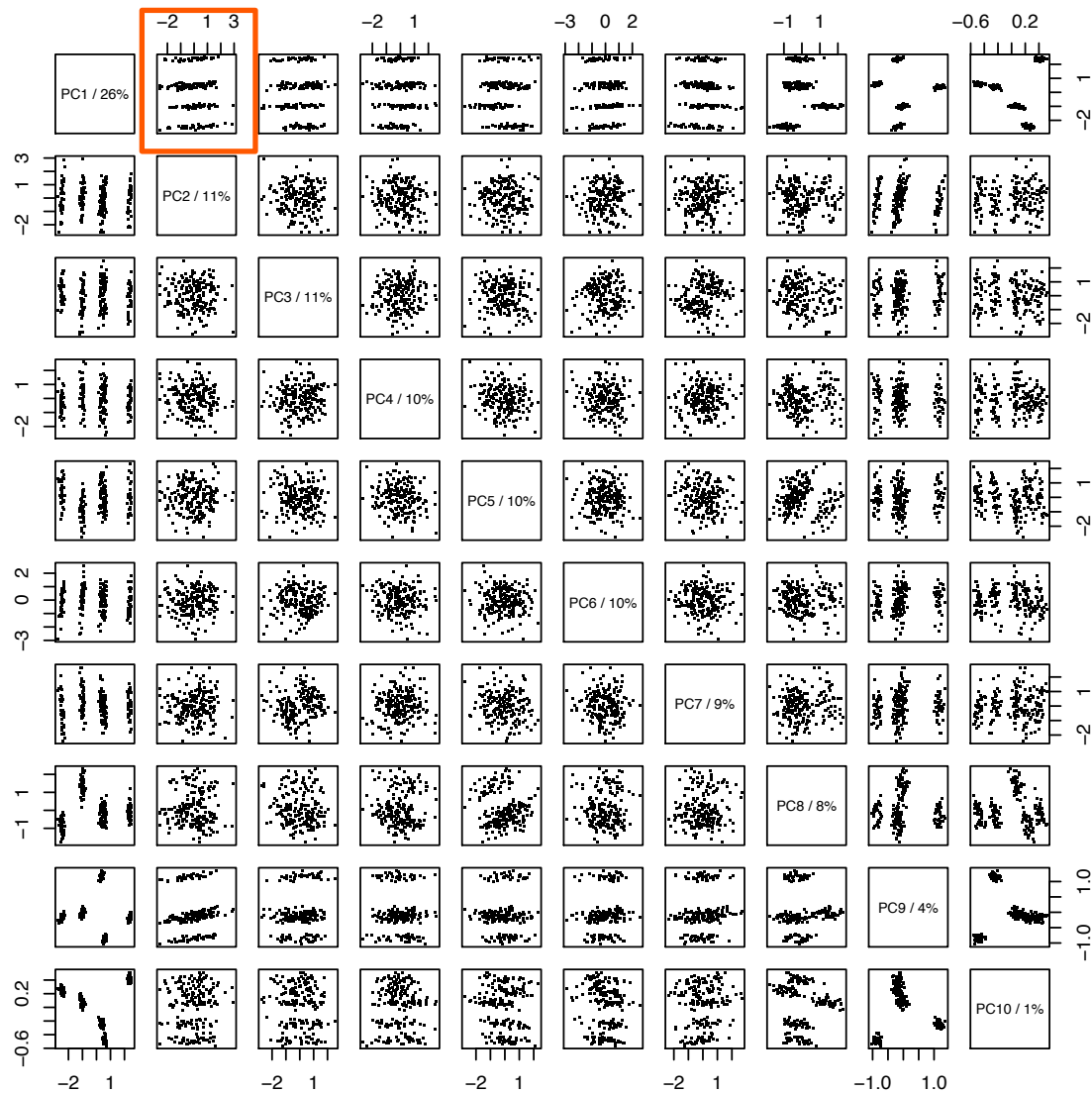
User already
knows this
clusters structure

These clusters
would be novel
and interesting
for the user

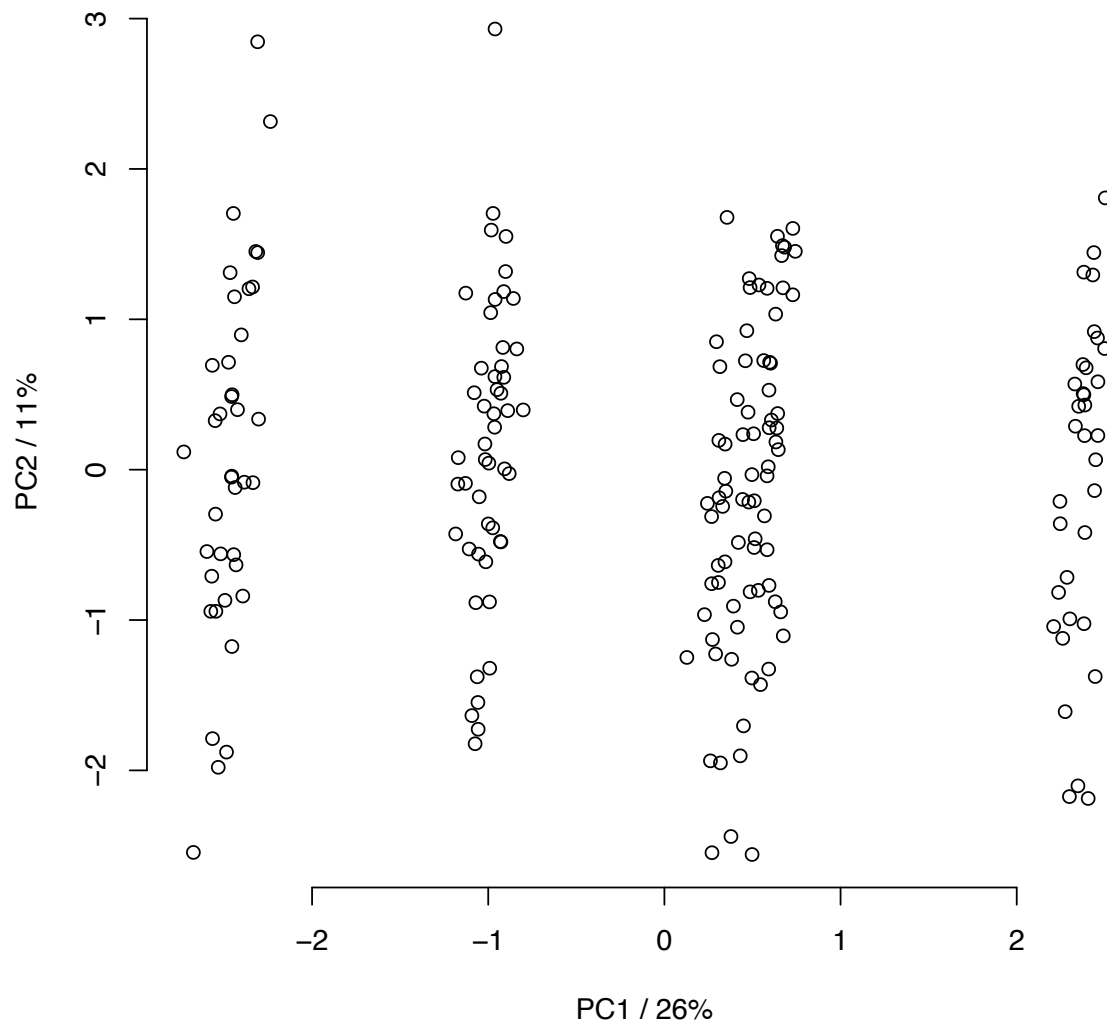


This is just
noise here

Principal components



Typical PC visualization: first 2 principal components





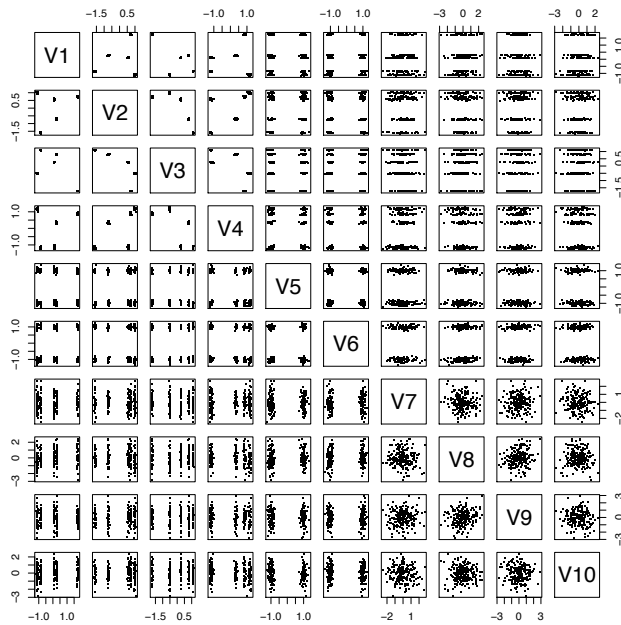
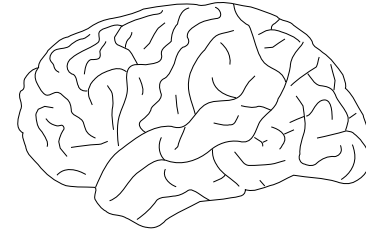
Finnish Institute of
Occupational Health

Our approach

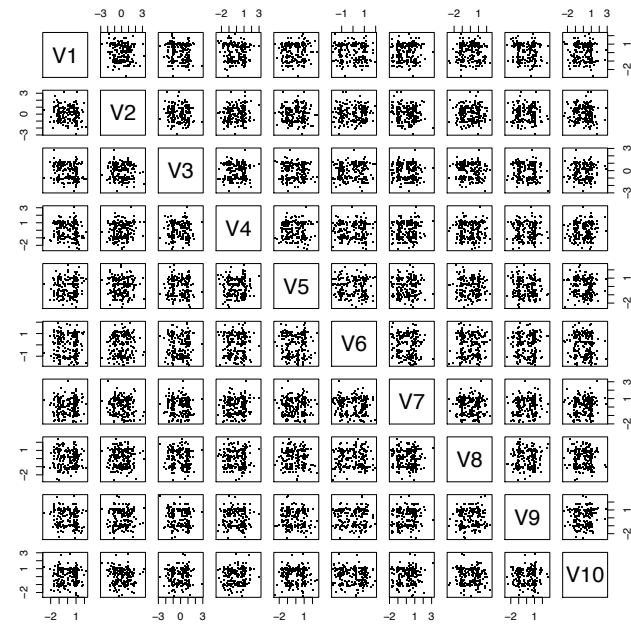
real world



user's background model
(= distribution over data sets)



real data

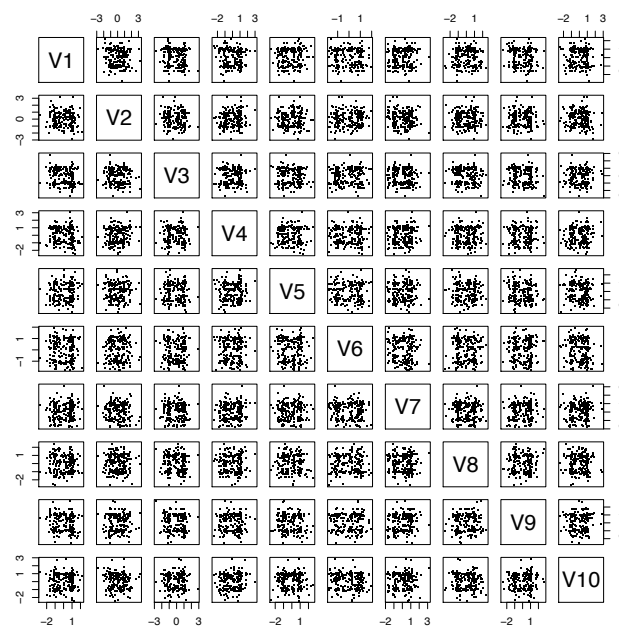
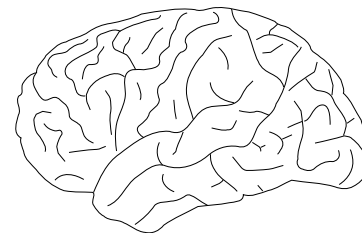


sample from background model


Background model is
a distribution over
possible data sets

Here: background
model is sampled by
permuting values of
real data (initially)


user's background model
(= distribution over data sets)



sample from background model



visualize difference
between real data and
background model



user tells what he or she has
absorbed from real data



update background model



iterate until done



visualize difference
between real data and
background model



user tells what he or she has
absorbed from real data



update background model



iterate until done

Task 1: visualize difference
between real data and
background distribution

Task 2: define visual
patterns by which user can
describe insights from data

Task 3: maintain description
of background model
(not discussed in this talk, see the paper)



visualize difference
between real data and
background model



user tells what he or she has
absorbed from real data



update background model



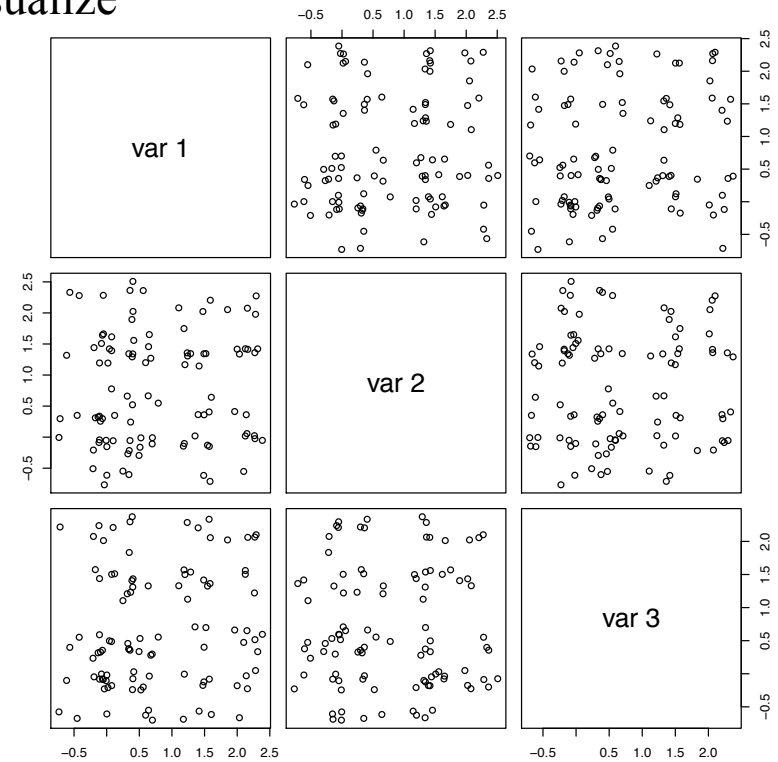
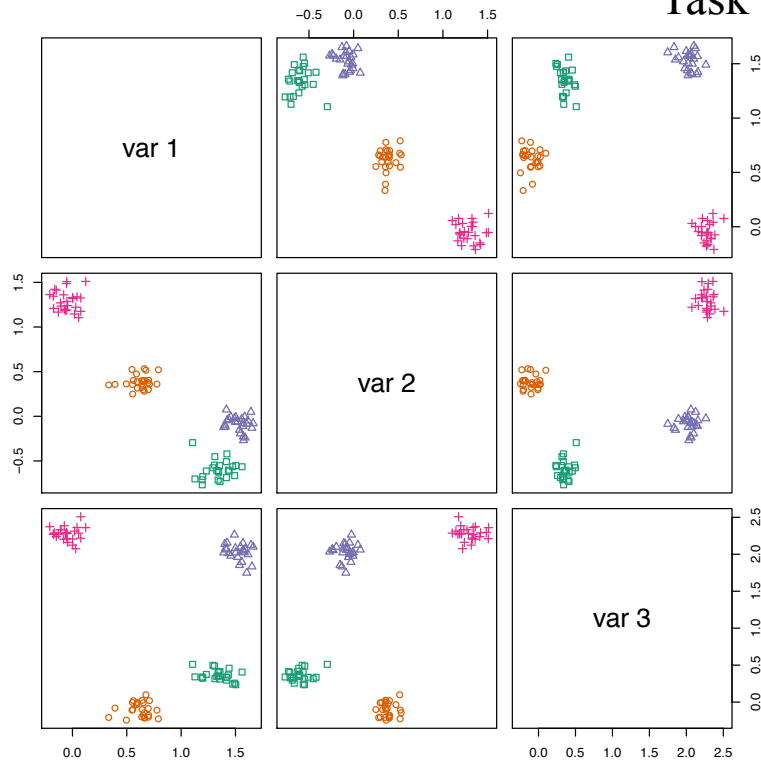
iterate until done

**Task 1: visualize difference
between real data and
background distribution**

**Task 2: define visual
patterns by which user can
describe insights from data**

**Task 3: maintain description
of background model**
(not discussed in this talk, see the paper)

Task 1: Visualize



Task 1: Visualize projection of the largest difference between real data and background distribution

Task 1: Visualize



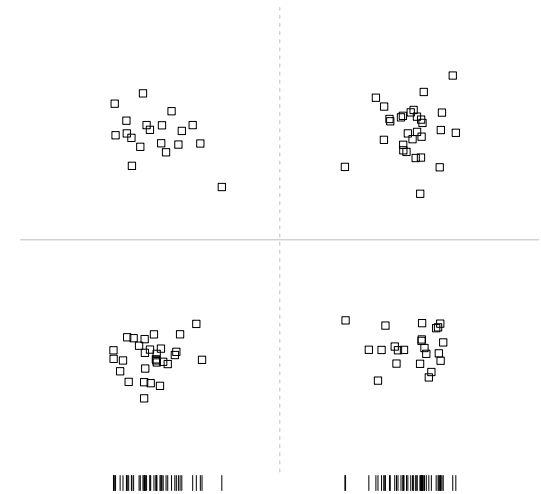
Easier task:
find 1D projection to which
difference is maximized

Task 1: Visualize

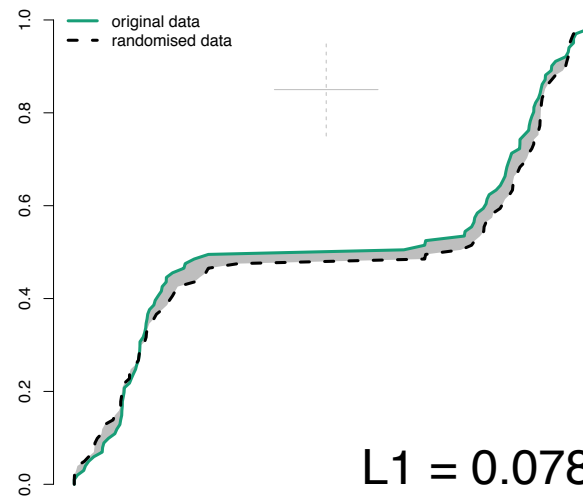
original data, rotation = 0



randomised data, rotation = 0



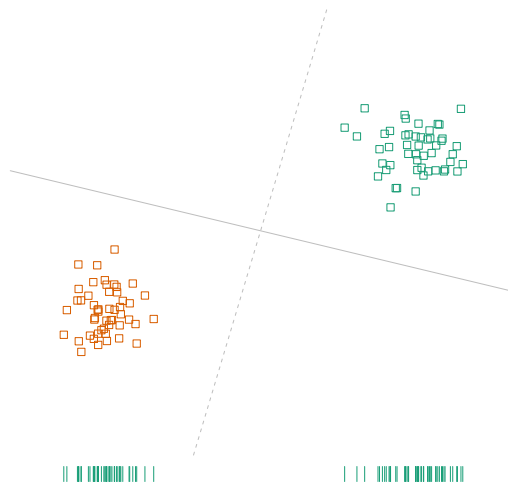
rotation = 0



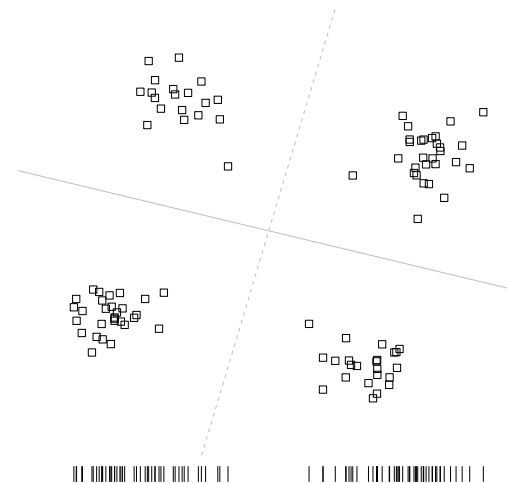
$L1 = 0.078$

Task 1: Visualize

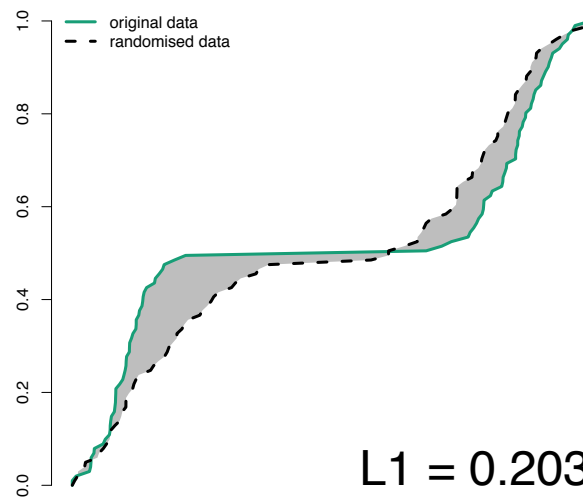
original data, rotation = 15



randomised data, rotation = 15



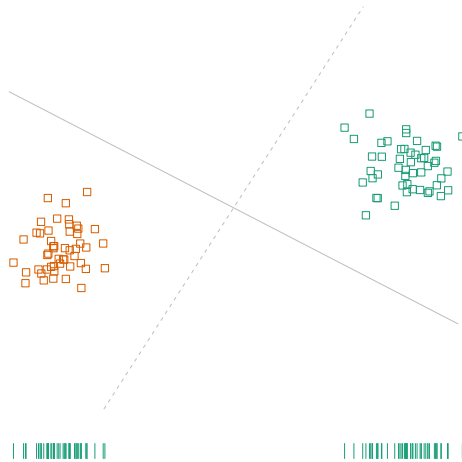
rotation = 15



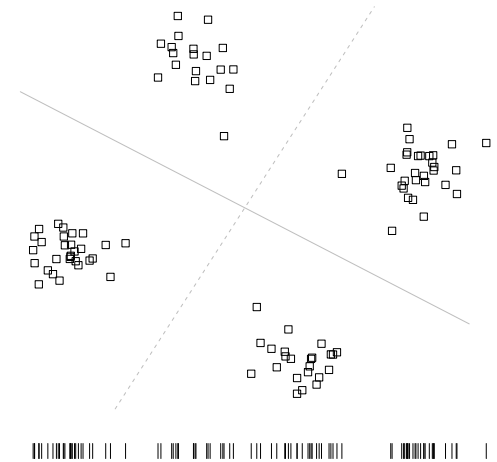
L1 = 0.203

Task 1: Visualize

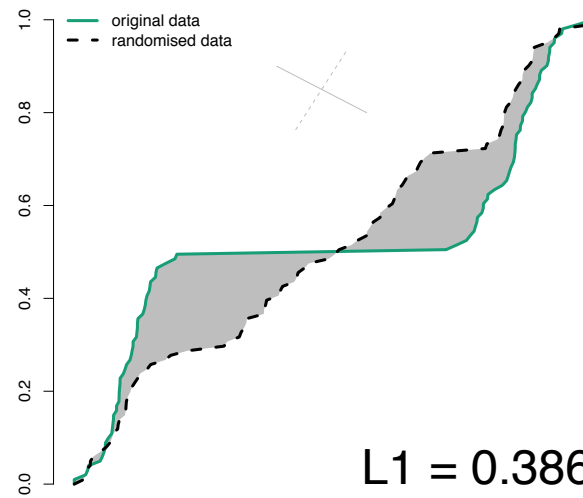
original data, rotation = 30



randomised data, rotation = 30

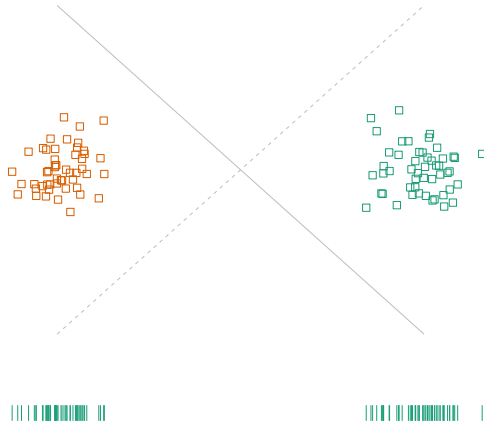


rotation = 30

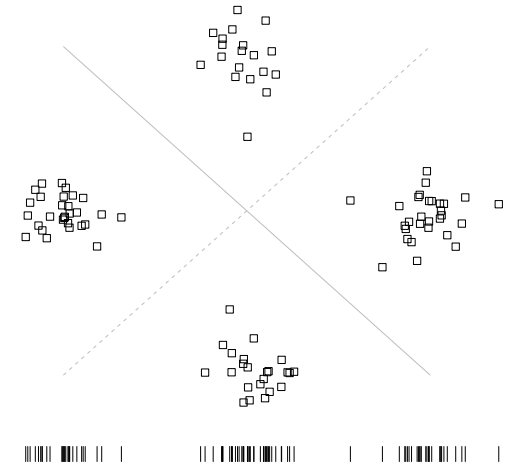


Task 1: Visualize

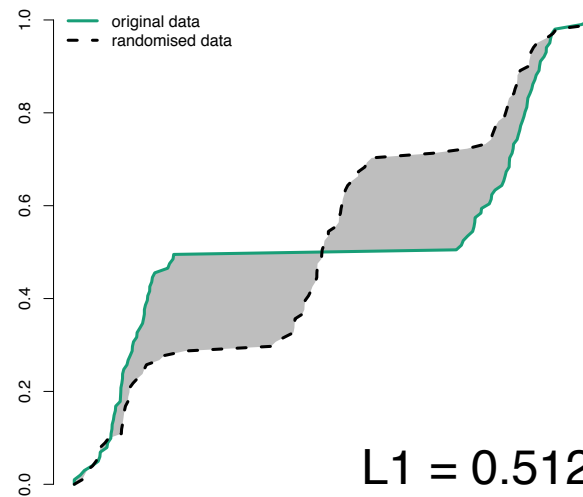
original data, rotation = 45



randomised data, rotation = 45



rotation = 45

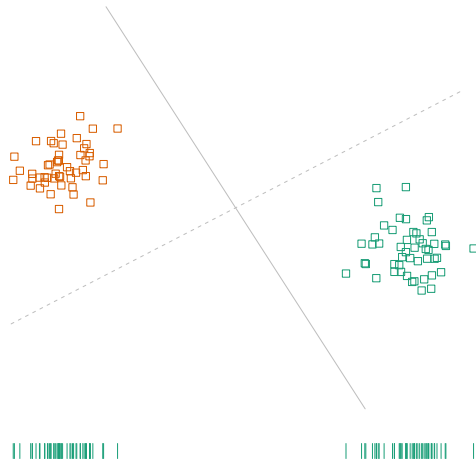


L1 = 0.512

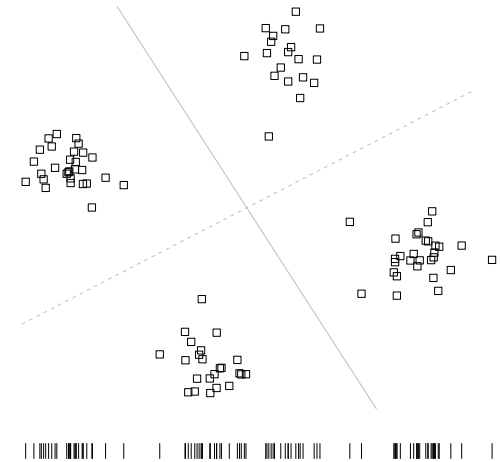
MAX

Task 1: Visualize

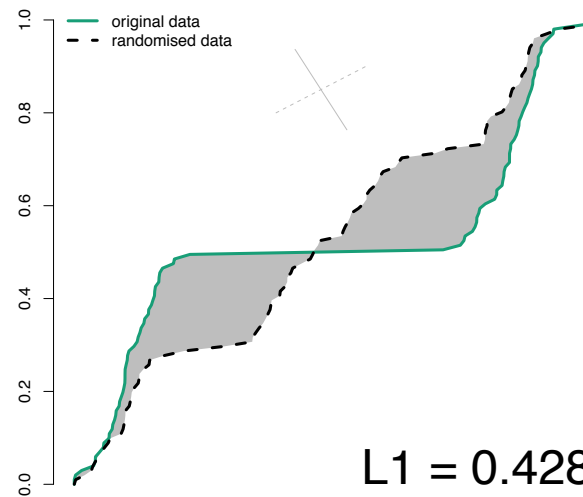
original data, rotation = 60



randomised data, rotation = 60

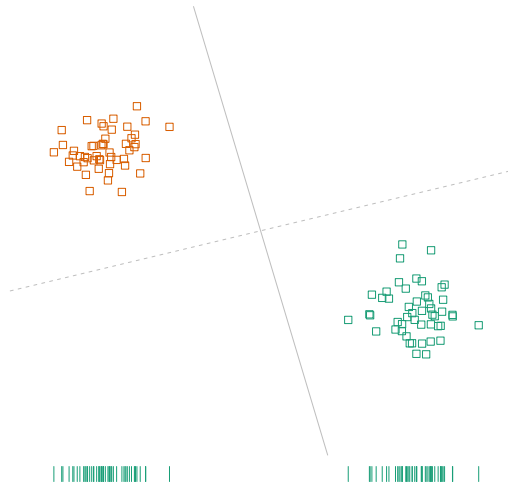


rotation = 60

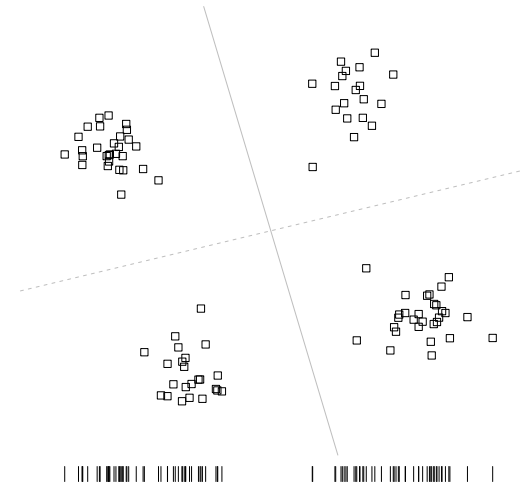


Task 1: Visualize

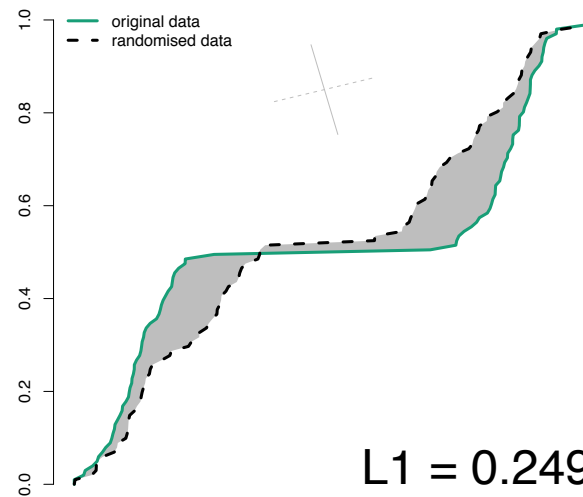
original data, rotation = 75



randomised data, rotation = 75

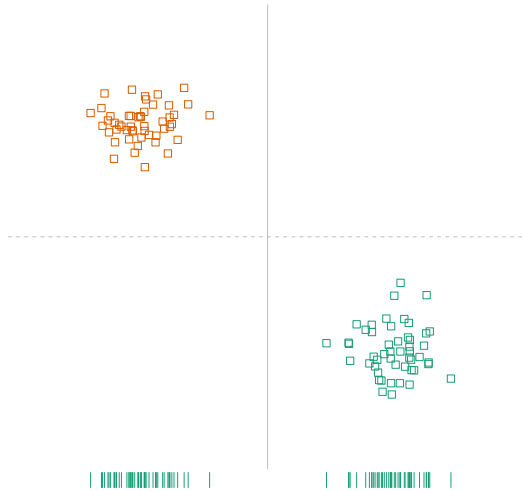


rotation = 75

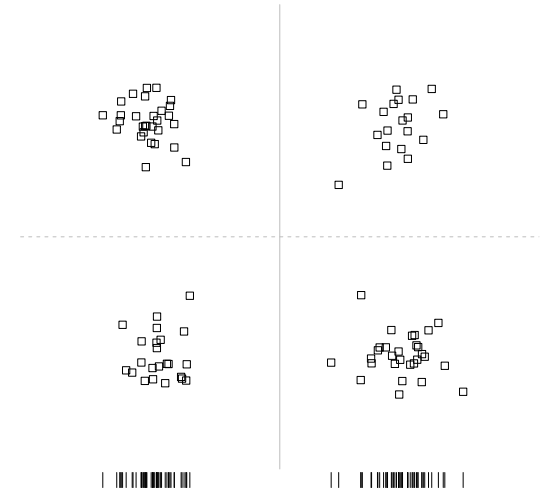


Task 1: Visualize

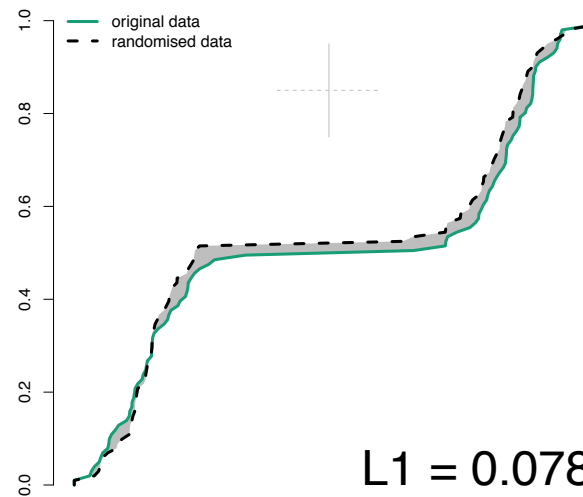
original data, rotation = 90



randomised data, rotation = 90



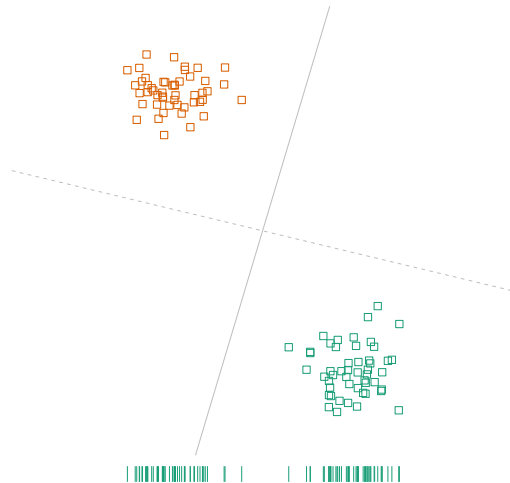
rotation = 90



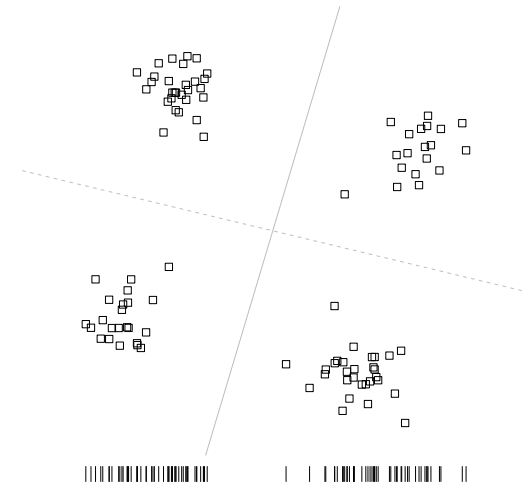
L1 = 0.078

Task 1: Visualize

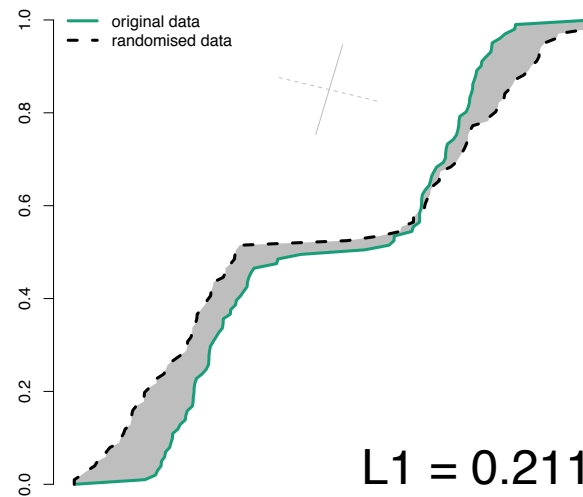
original data, rotation = 105



randomised data, rotation = 105

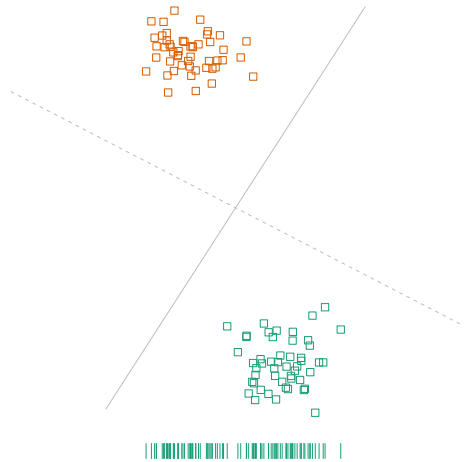


rotation = 105

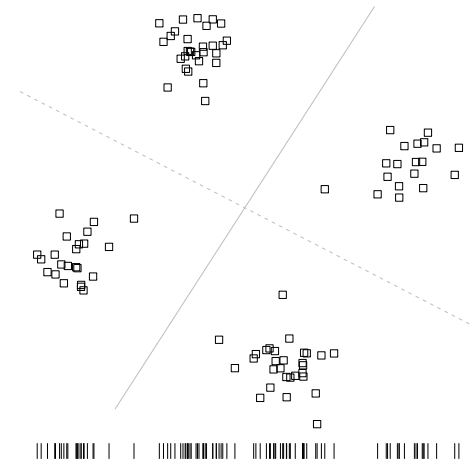


Task 1: Visualize

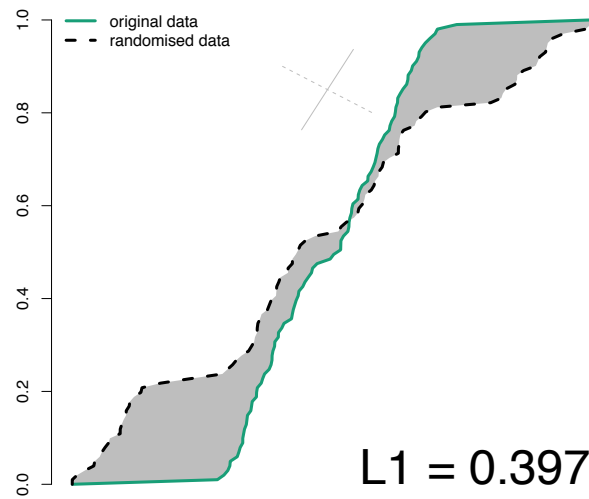
original data, rotation = 120



randomised data, rotation = 120

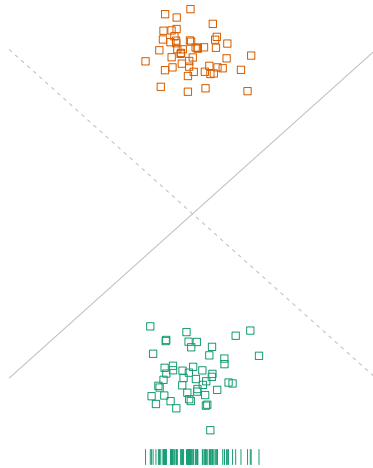


rotation = 120

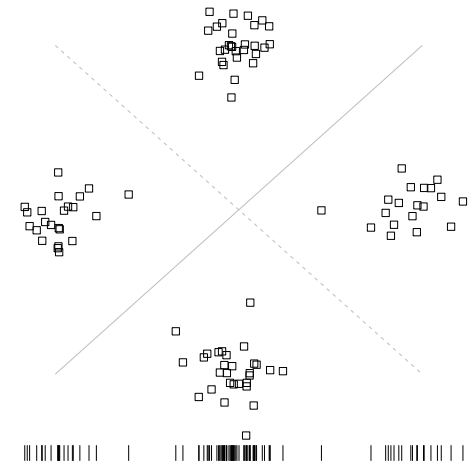


Task 1: Visualize

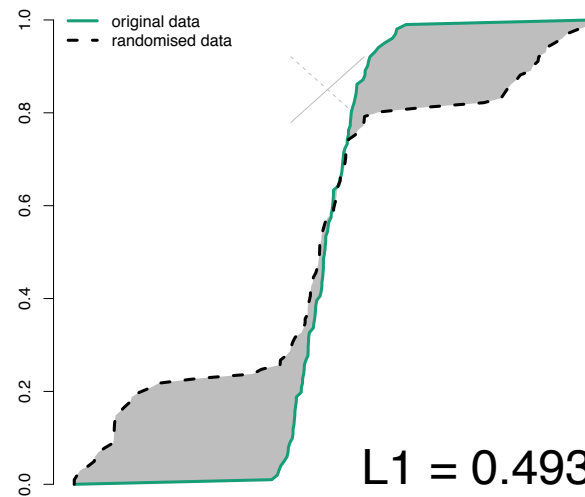
original data, rotation = 135



randomised data, rotation = 135

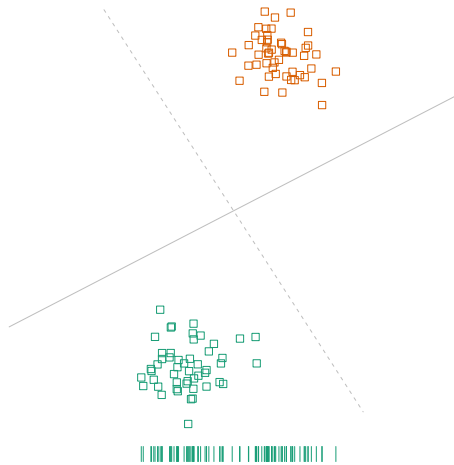


rotation = 135

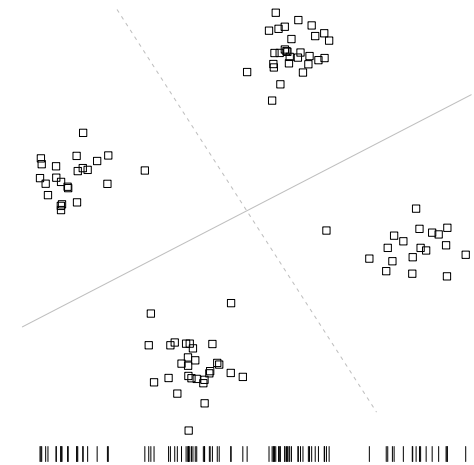


Task 1: Visualize

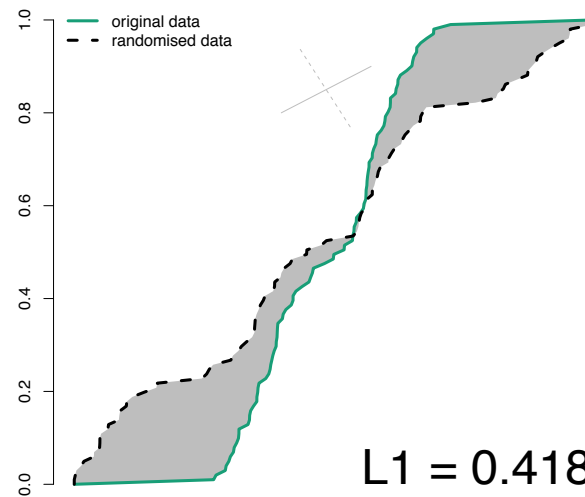
original data, rotation = 150



randomised data, rotation = 150

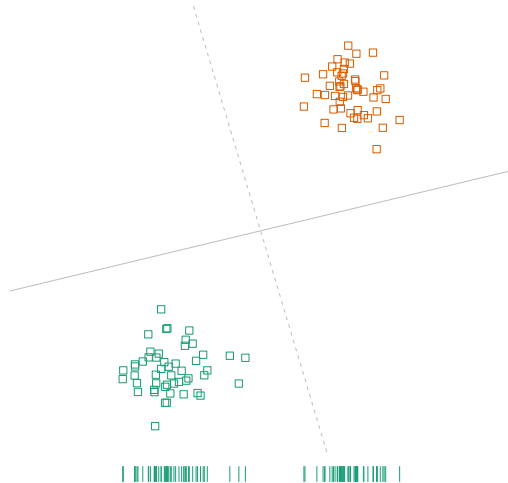


rotation = 150

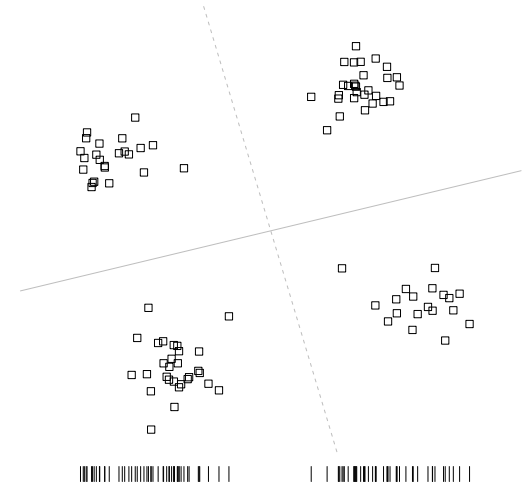


Task 1: Visualize

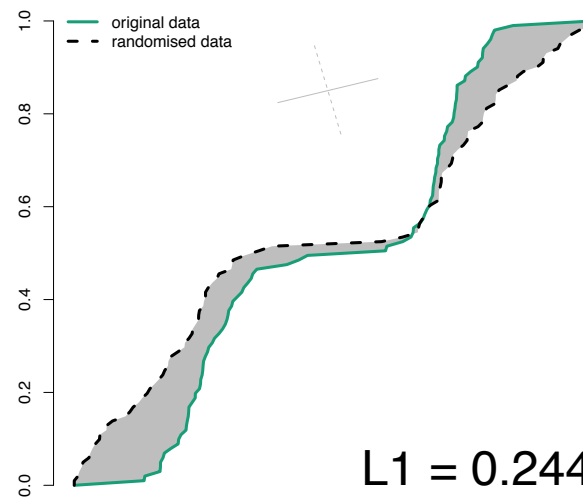
original data, rotation = 165



randomised data, rotation = 165



rotation = 165

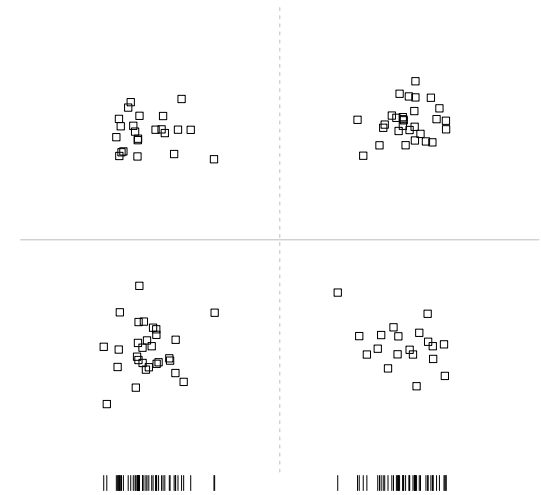


Task 1: Visualize

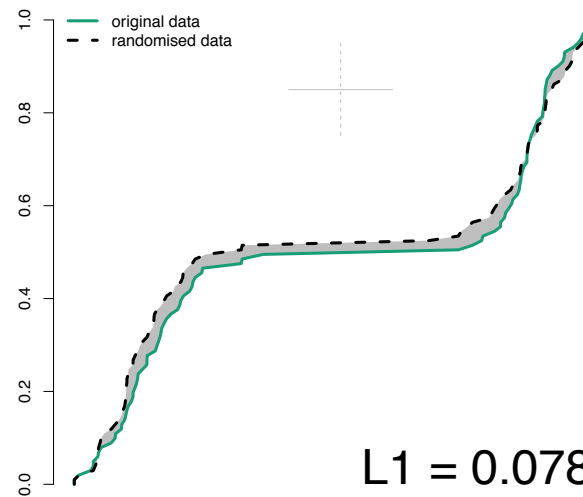
original data, rotation = 180



randomised data, rotation = 180



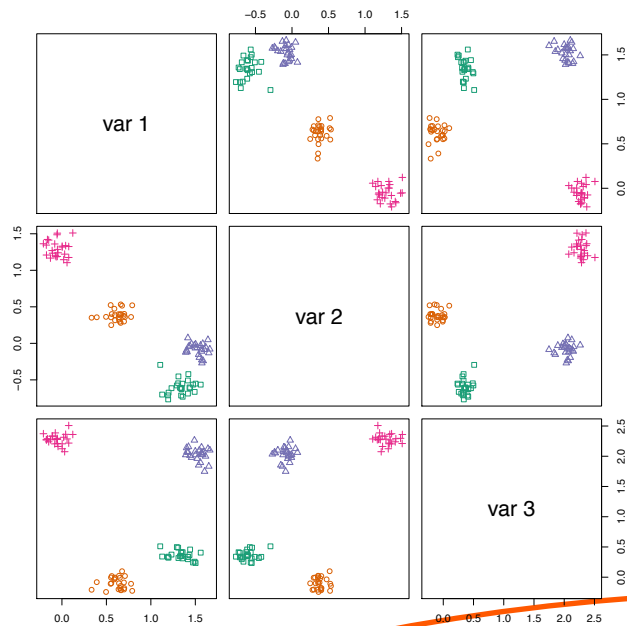
rotation = 180



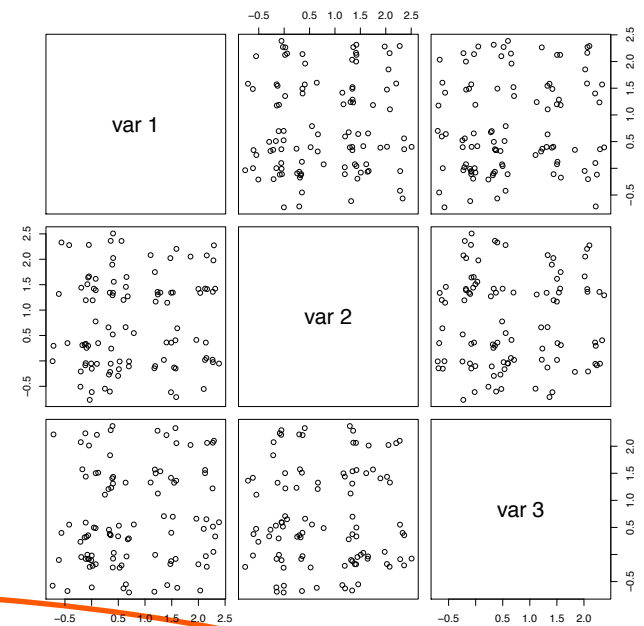
Task 1: Show projection of the largest difference between real data and background distribution

Solution: Use *projection pursuit* to find 2D to which the real data and background distributions differ most

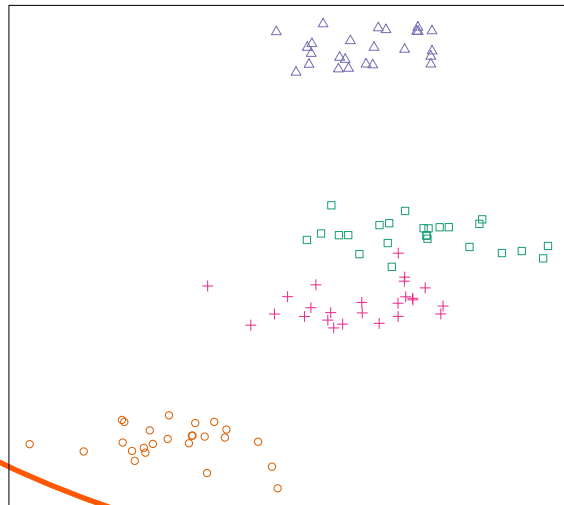
Task 1: Visualize



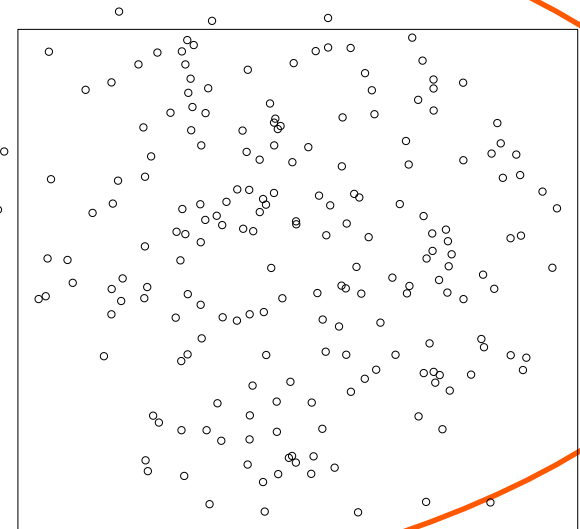
real data



sample from background model



2D projection
showing
maximal
difference
between real
data and
background
distribution





visualize difference
between real data and
background model



user tells what he or she has
absorbed from real data



update background model



iterate until done

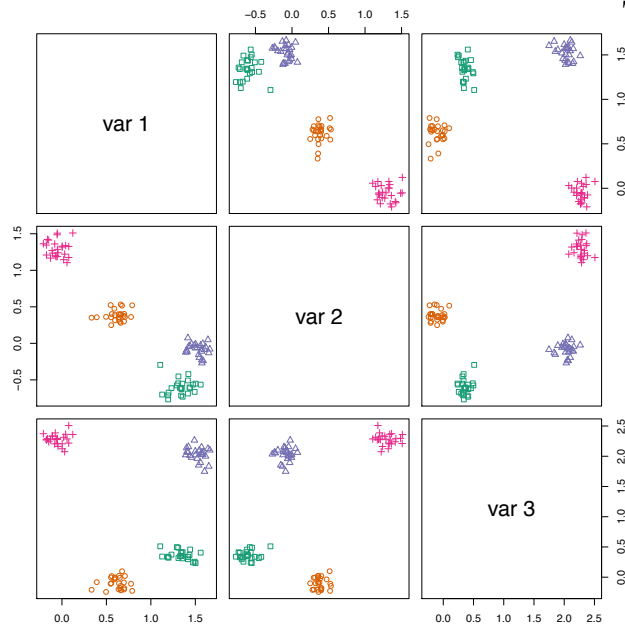
Task 1: visualize difference
between real data and
background distribution

**Task 2: define visual
patterns by which user can
describe insights from data**

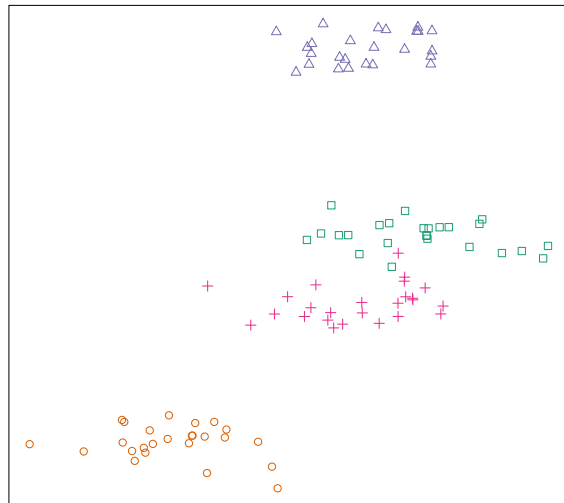
Task 3: maintain description
of background model
(not discussed in this talk, see the paper)

Task 2: Modifying background distribution with constraints

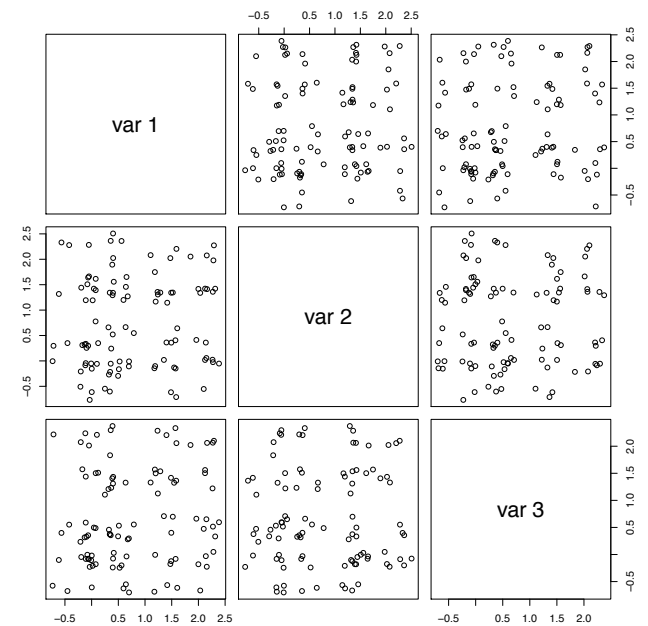
Task 2: Constraints



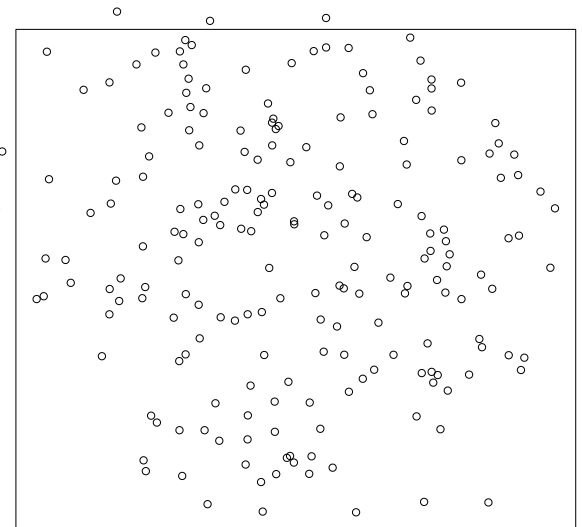
real data



2D projection
showing
maximal
difference
between real
data and
background
distribution



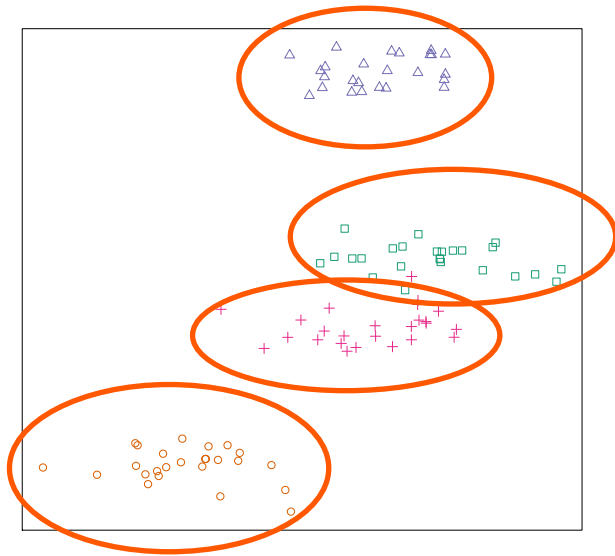
sample from background model



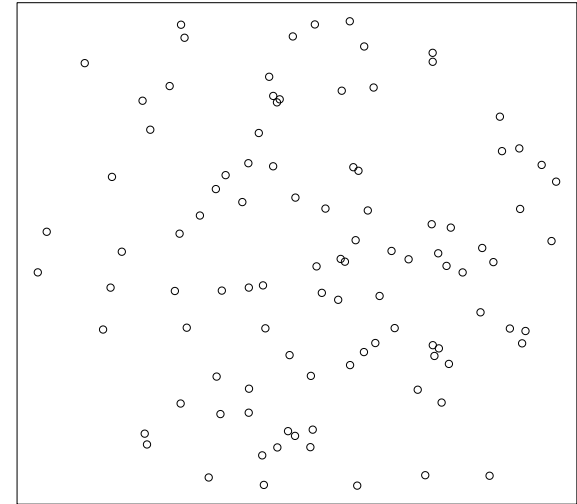
Here: 2 types of visual constraints

- **2D constraints:** *“I know the positions of a set of points in the shown 2D projection”*
- **clustering constraints:** *“The set of points are nearby maybe in other directions as well.”* (allows inputting insight not obvious from visualization!)

Task 2: Constrains



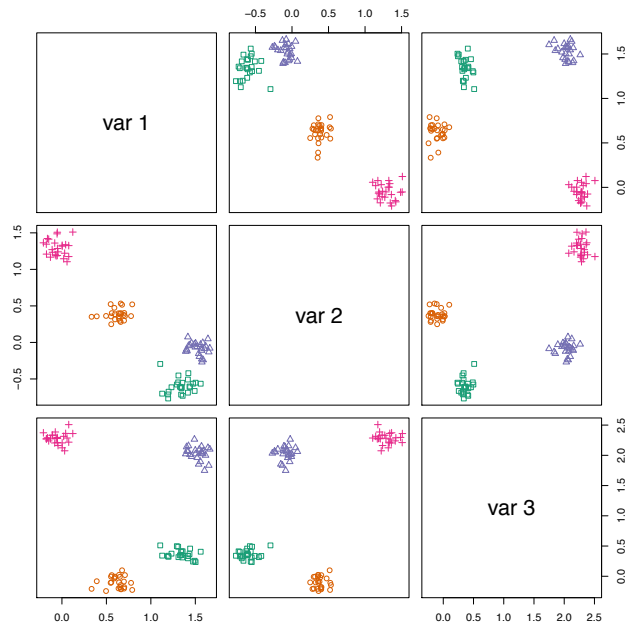
real data



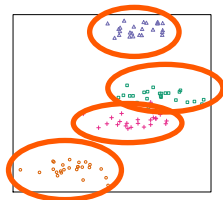
sample from background model

User marks 4 sets of points in real data which are different from background distribution

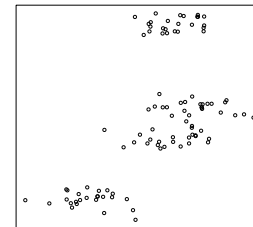
Task 2: Constrains



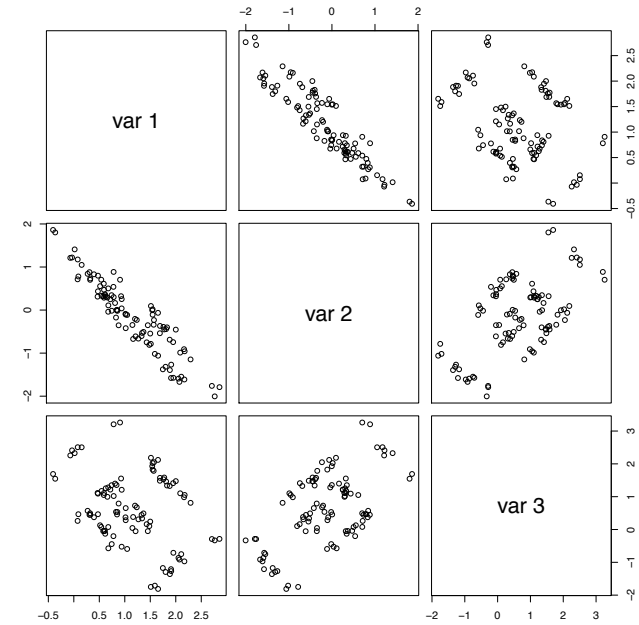
real data



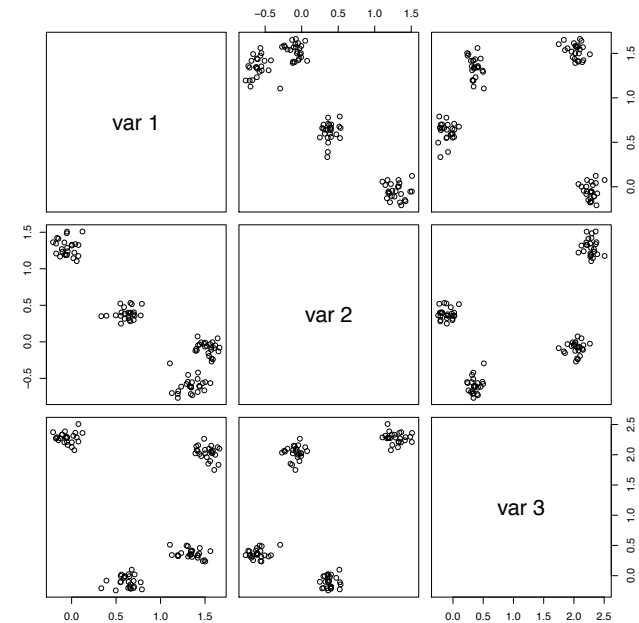
**2D
constraint**



**clustering
constraint**



sample from background model





visualize difference
between real data and
background model



user tells what he or she has
absorbed from real data



update background model



iterate until done

Task 1: visualize difference
between real data and
background distribution

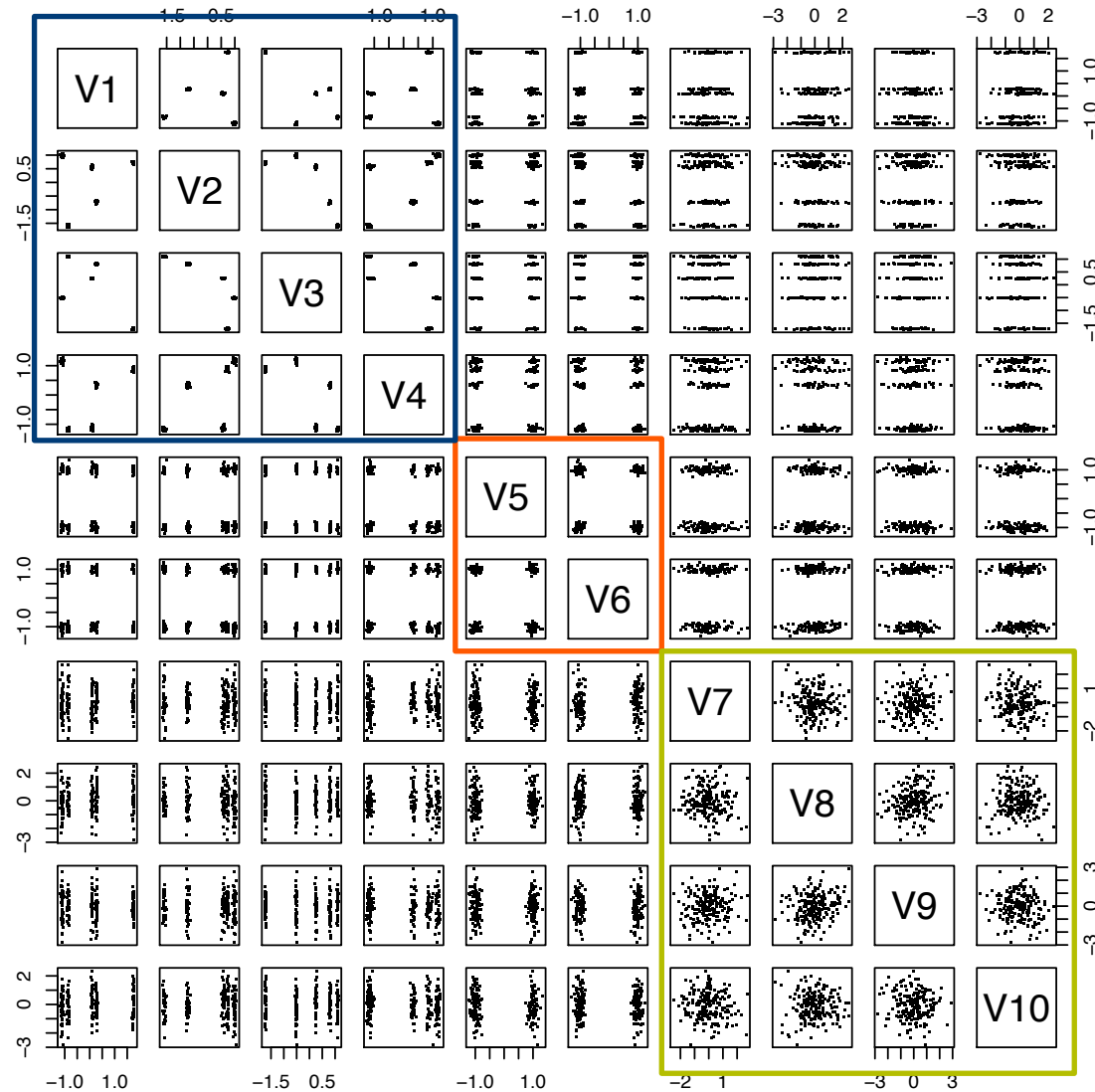
Task 2: define visual
patterns by which user can
describe insights from data

Task 3: maintain description
of background model
(not discussed in this talk, see the paper)

Return to the original 10-dimensional dataset

User already knows this clusters structure

These clusters would be novel and interesting for the user



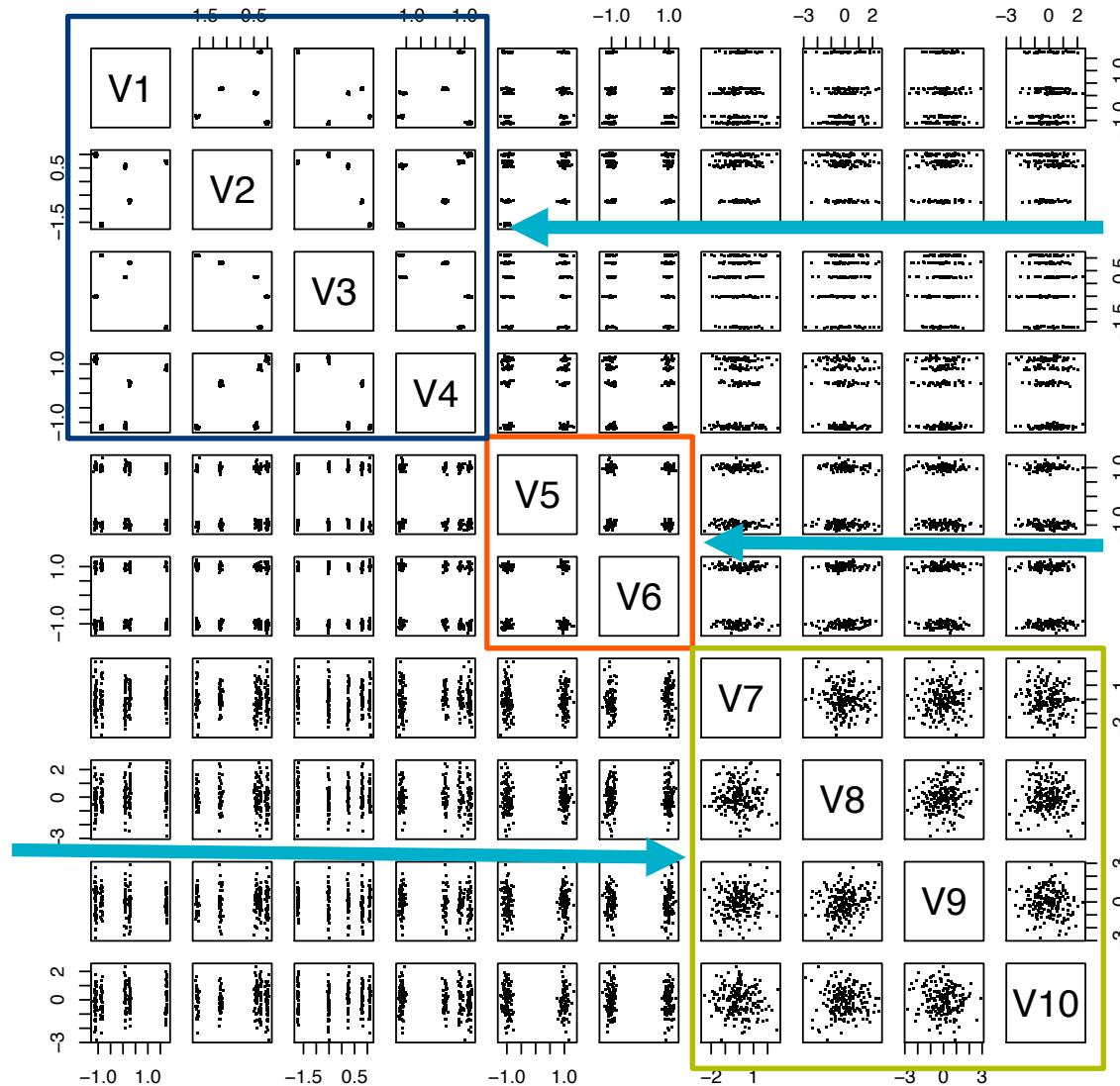
This is just noise here

Return to the original 10-dimensional dataset

User already knows this clusters structure

These clusters would be novel and interesting for the user

System does not show noise because it is similar to the background model



System shows this first and user can “skip” it with clustering constraint

Next system shows this structure

This is just noise here

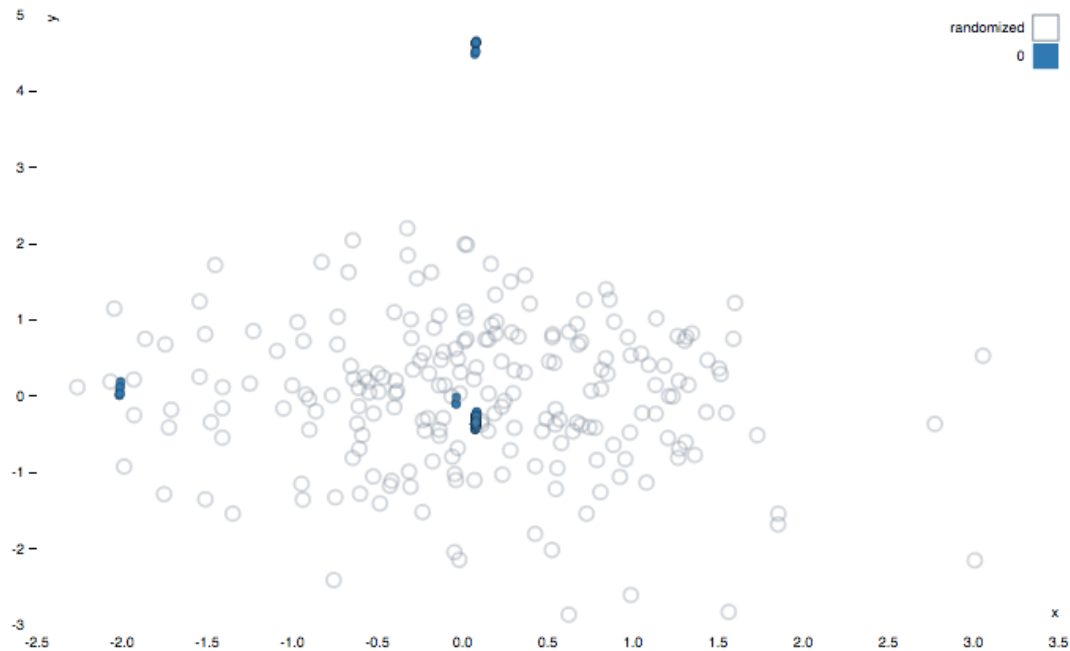
UCI Adult Dataset Case Study

Update Background Model

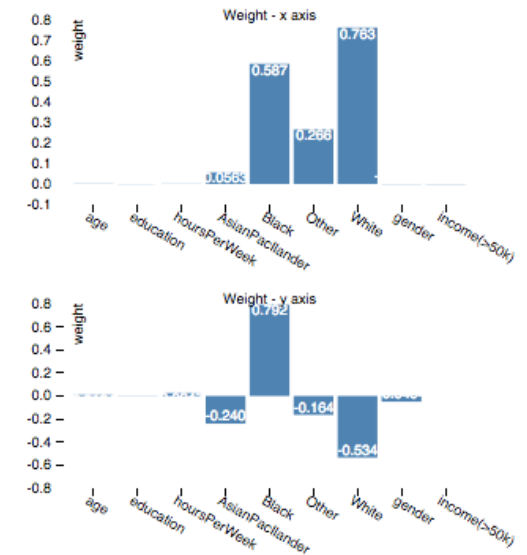
Feedback (Cluster Constraint)

Feedback (2D Constraint)

Projection:

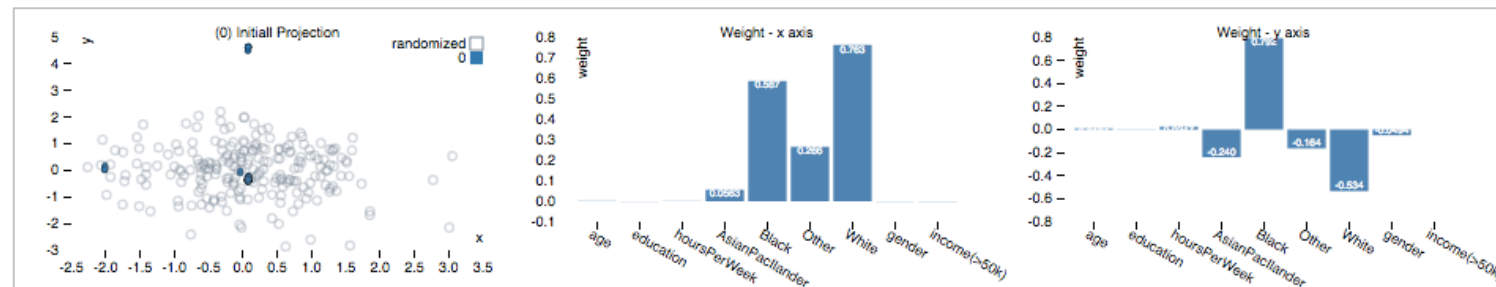


Weight Vectors:



Marked Cluster Center:

Snapshots:



Not discussed

- Runtime (works quite fast)
- Detailed mathematical derivation of the problem
- Example data sets
- See the conference and demo track papers for more details!

Concluding remarks

- This is a generic framework for explorative data mining:
 - User has a background model that is assumed to be a distribution over data sets
 - Computer can model the background distribution and show the user interesting differences between it and the real data
 - User can modify his/her background distribution and inform the computer about this
 - Constrained randomization (CORAND) / max.entropy (FORSIED)
- Open questions / future work:
 - Real user tests still to do
 - Does this make any sense cognitively?
 - Other data types / interactions



Finnish Institute of
Occupational Health



UNIVERSITEIT
GENT

Please come to see our demo on Thursday or
try it online (link at the paper)!



ACADEMY
OF FINLAND

Tekes



erc

European Research Council

Established by the European Commission