# The Mushroom Database

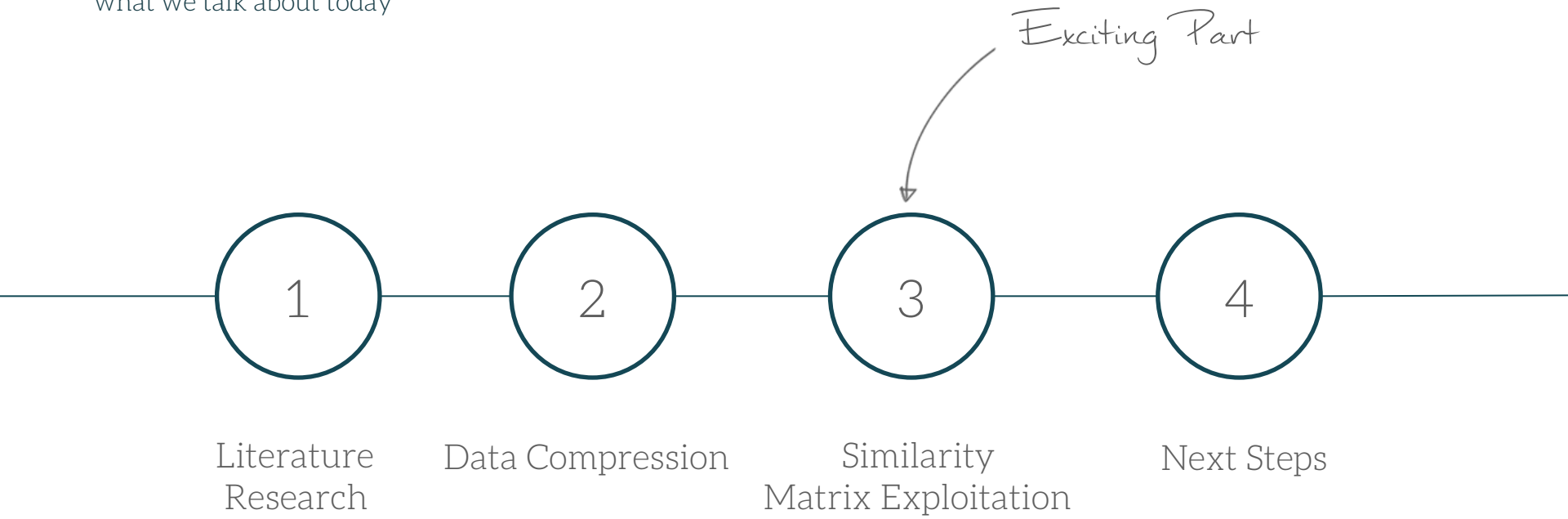
## Descriptive Mining II

5. November. 2014

ROSTLAB.

# Agenda

what we talk about today



# Literature Research

what we found out

# Data Compression

Deleting of data

## Delete Redundant Features

*Threshold >95%*

- Gill Attachment (97.42%)
- Veil Type (100%)
- Veil Color (97.54%)

## Delete Redundant Characteristics

*Threshold >95%*

- Gill Spacing: characteristic 'distant'
- Stalk Root: characteristics 'cup' and 'rhizomorphs'.
- Ring Type: characteristics 'cobwebby', 'sheathing' and 'zone'

## Delete Strong Correlated Features

*Threshold >95%*

- Stalk Surface Above Ring + Stalk Surface Below Ring: 77.01%
- Stalk Color Above Ring + Stalk Color Below Ring: 62.38%

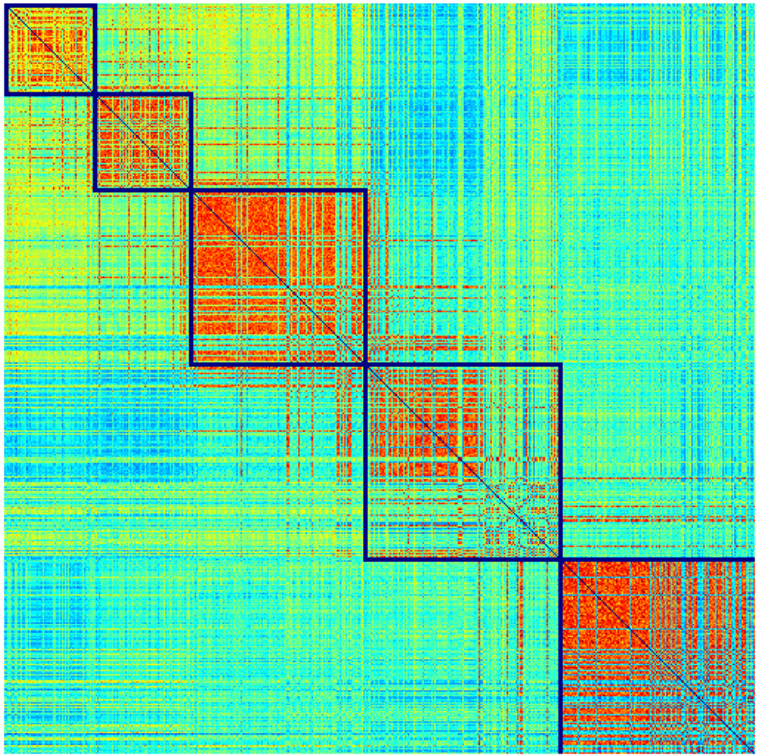
# Data Compression
grouping of data

## Group Characteristics
*Threshold <1%*

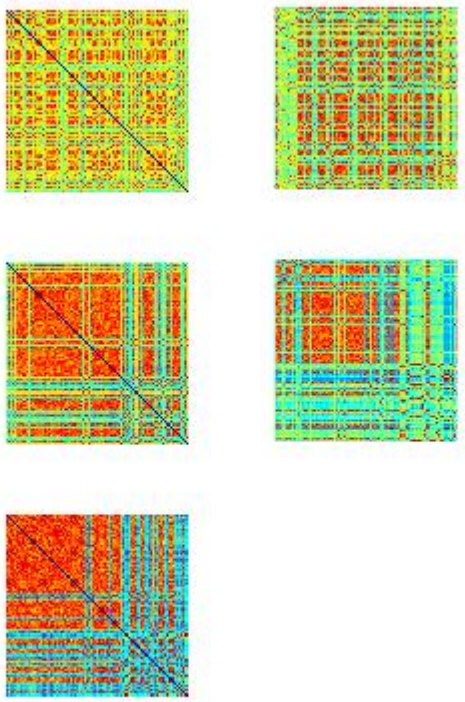
- Cap Color: 'brown' + 'cinnamon' = 'brown'. 'green' + 'pink' = 'misc'
- Ring Type: 'flaring' + 'none' = 'misc‘
- Odor: 'musty' + 'foul' = 'foul'

# Similarity Matrix Exploitation

The Similarity Matrix

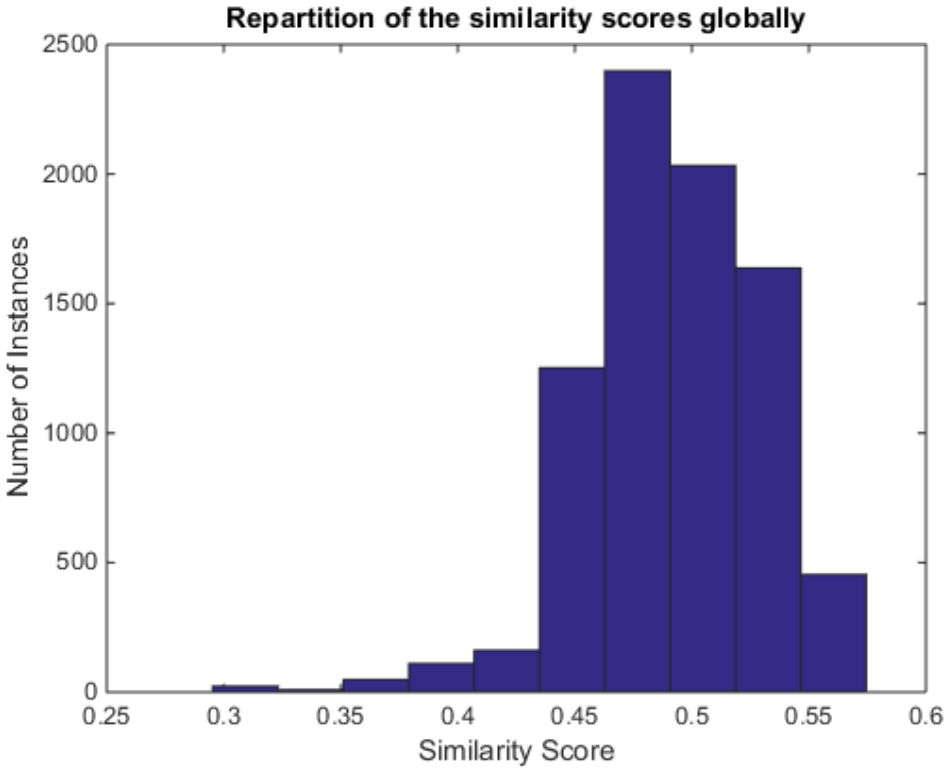


The Initial Subgroups



# Average Similarity Distribution

Whole Similarity Matrix

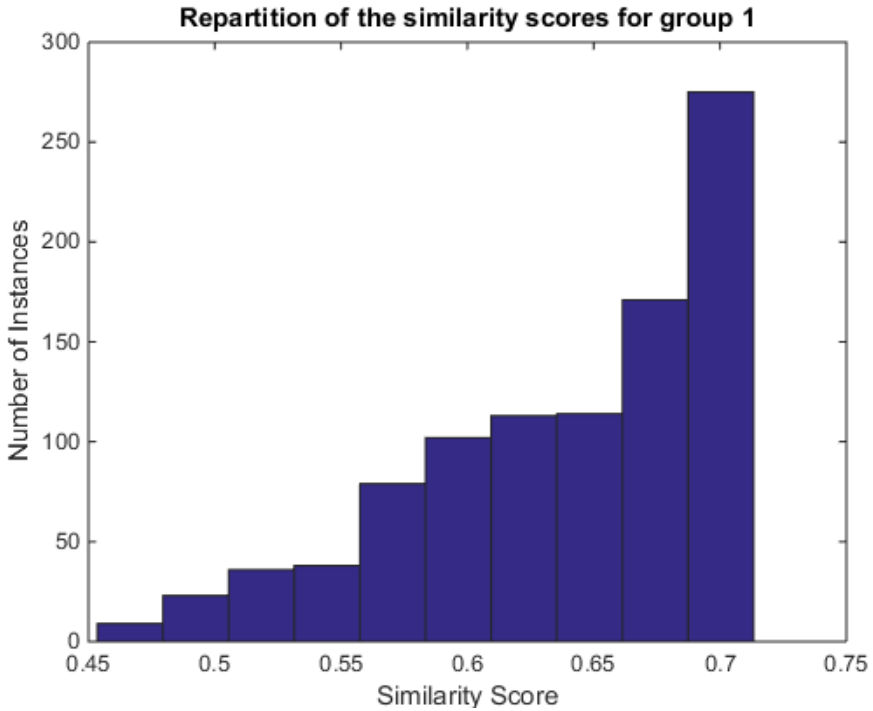



- 0 no attribute in common
- 1 all attributes in common
- For *each* instance with *all other* instances

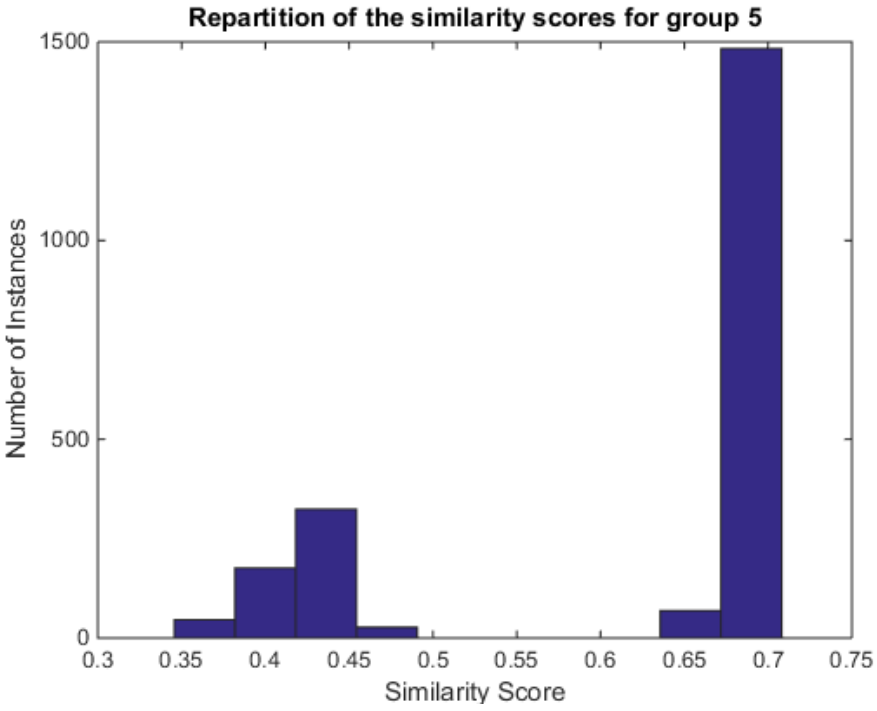
→ Helps to identify outliers

# Average Similarity Distribution

Group




Group 5



# Next Steps

what our future tasks will be

# Thank You

for your undivided attention