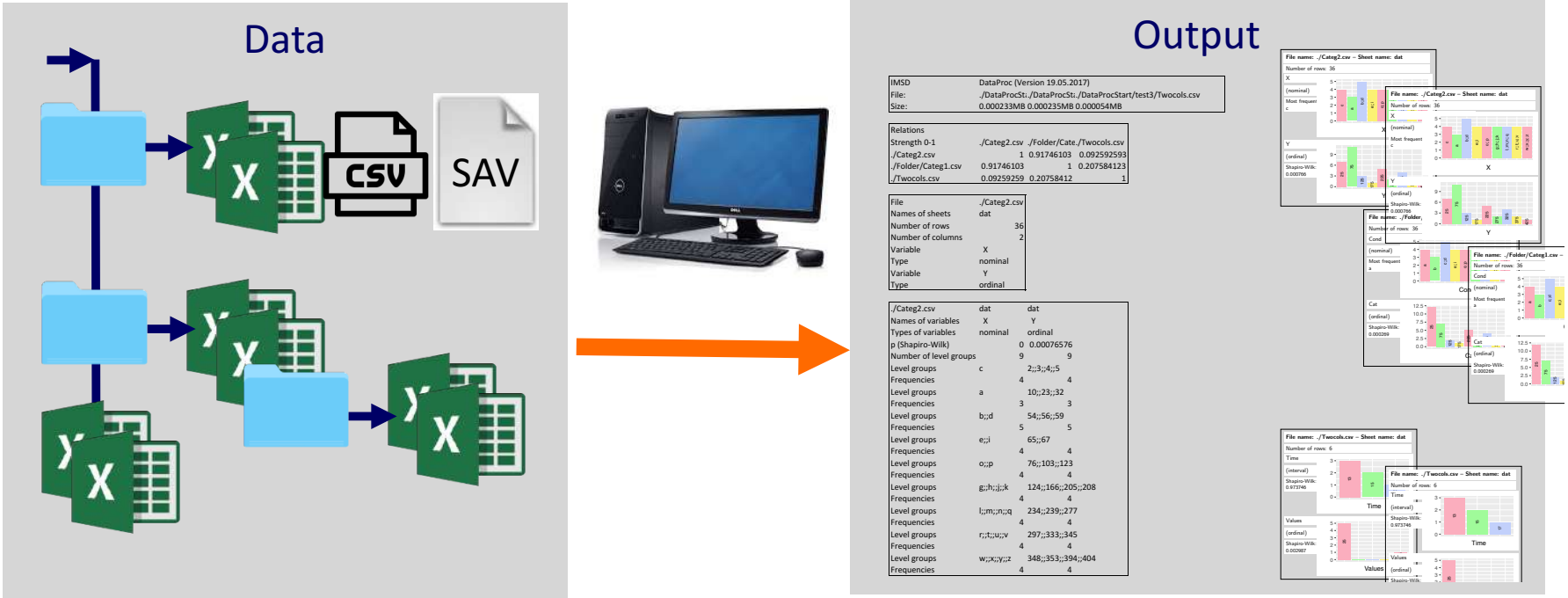
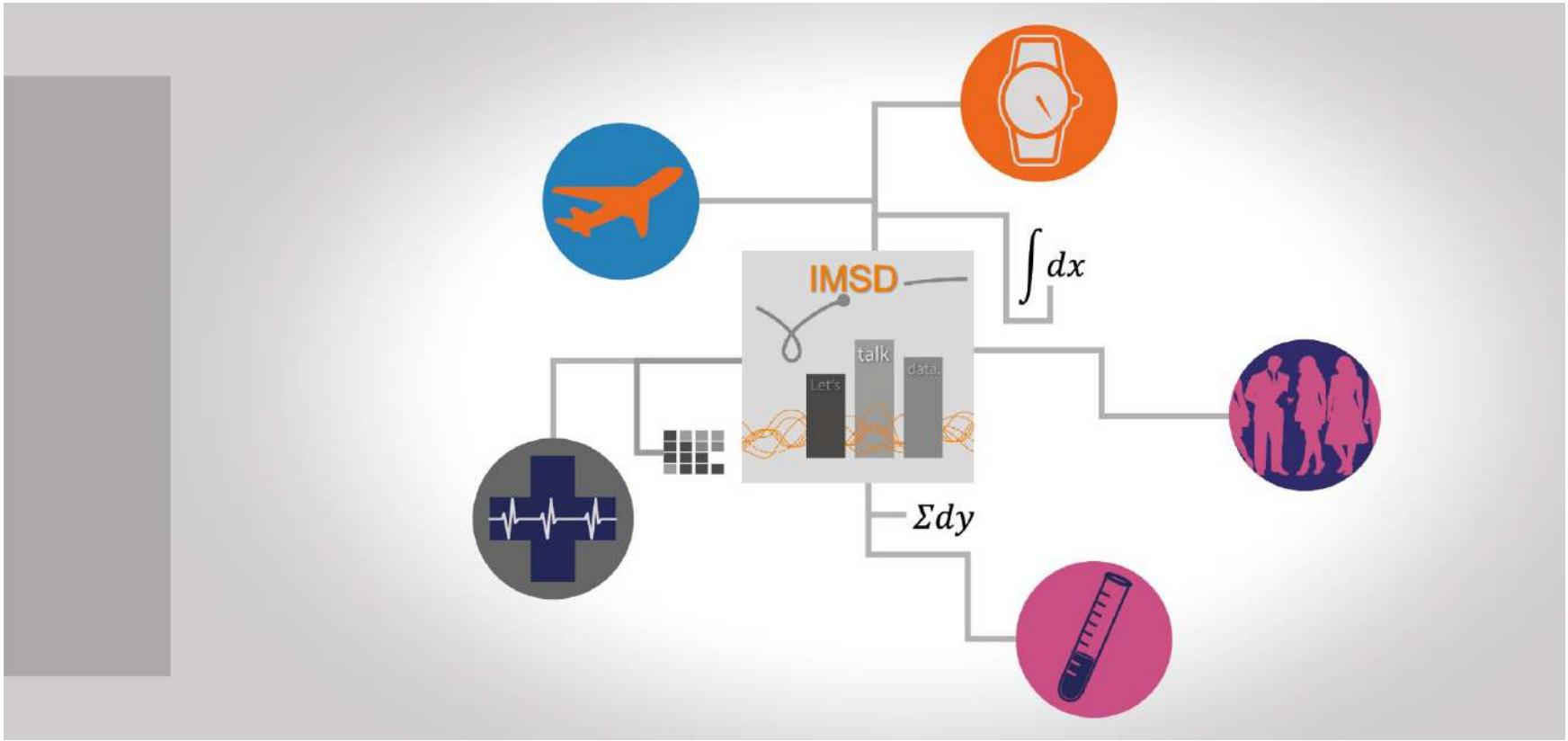
Automatic pre-processing of mass data collections of unknown structure

Pierre A. Mandrin, Rodolphe Dewarrat
IMSD GmbH, Lavaterstrasse 103, 8002 Zürich, www.imsd.ch



Who we are

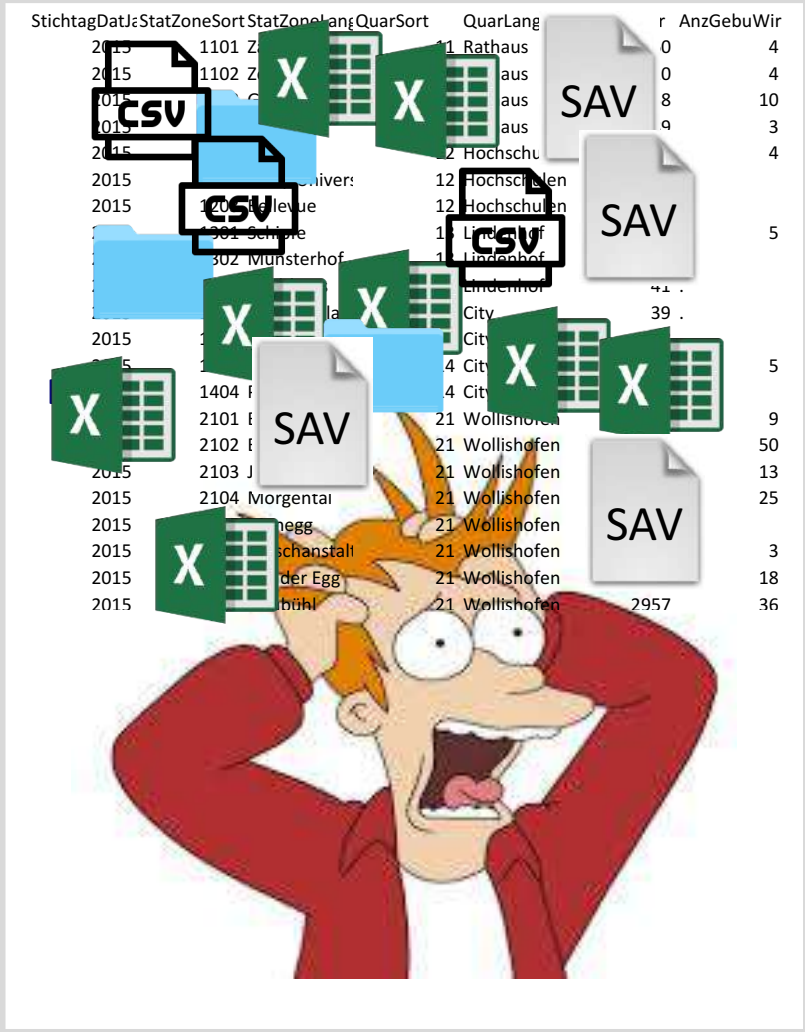
- Founded in 2012
- Project specific analyses of heterogenous data and optimizations
- 10-15 projects per year
- Customers in Switzerland and all over the world



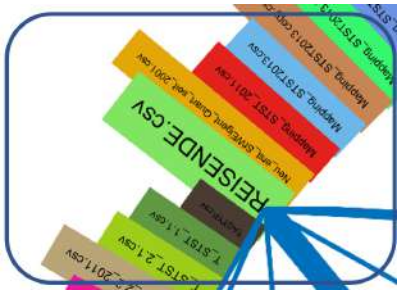
Automatic Pre-Processing – Unknown Structure

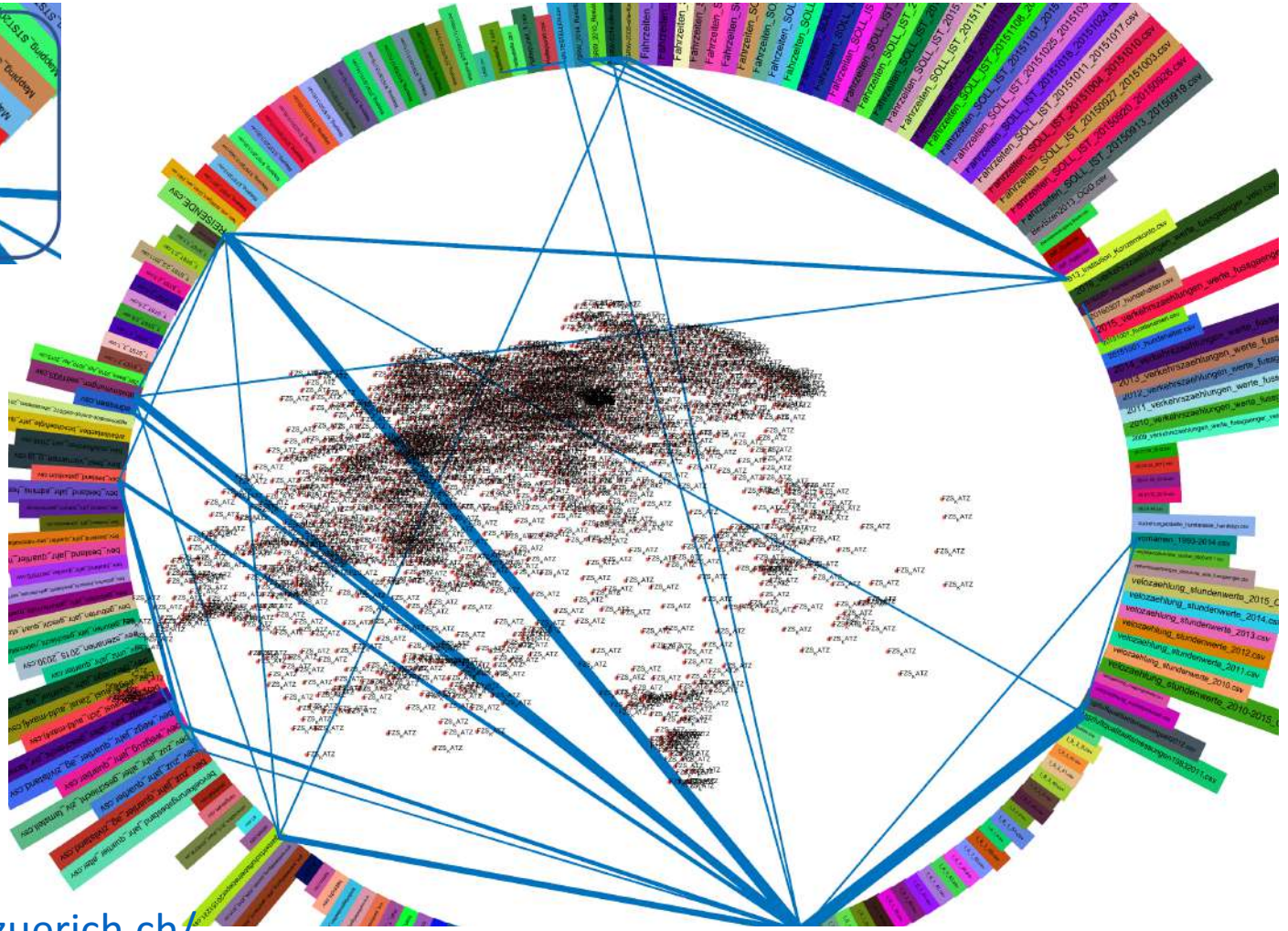
What are our Goals?

- Gain Overview over Large Data Sets
- Identify Data Structure
- Quick View Data Properties
- Avoid Manual Opening of Each File
- Pre-Visualization of All Data
- Aid to Prepare Detailed Analyses



A Related IMSD-Project – Open Data at a glance





Source:

Stadt Zürich
Open Data

https://data.stadt-zuerich.ch/

OUTLINE OF THIS TALK

- Guiding Principle of Pre-Processing: Using Data Categories
- Working Out Step by Step – Examples with small data sets
- Demonstration of the Tool for Data of an Experimental study
- Conclusions

OUTLINE OF THIS TALK

- Guiding Principle of Pre-Processing: Using Data Categories
- Working Out Step by Step – Examples with small data sets
- Demonstration of the Tool for Data of an Experimental study
- Conclusions

HOW TO EXTRACT INFORMATION FROM DATA OF UNKNOWN STRUCTURE

Specify Desired Features

Match Certain patterns to Given Categories

See e.g. D. W. Embley, C. Tao, S. W. Liddle - ER 2002, **2503** (2002) 322-337

No Input Specifications (Look for Data Characteristics)

Distinguish Data Features by a List of Categories

Compare Data Patterns

THIS PROJECT!

Prepare Data

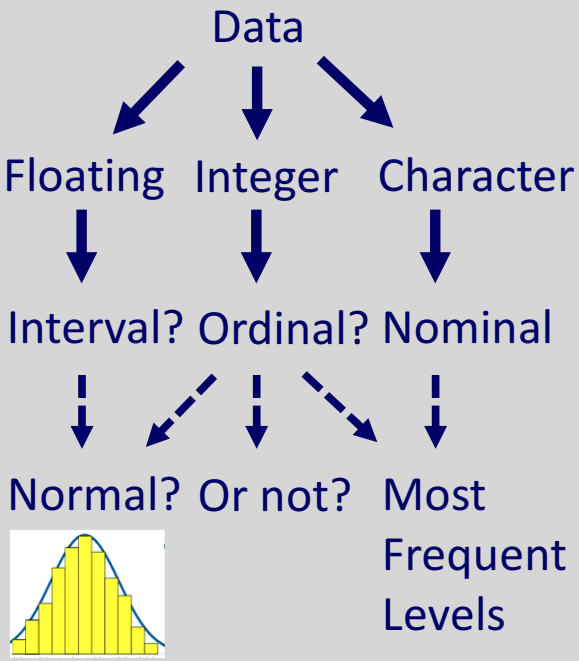
- Read files into tables / identify data types
- Identify header lines if possible

Sort Data by Categories

e.g.

Guess Type of Variable

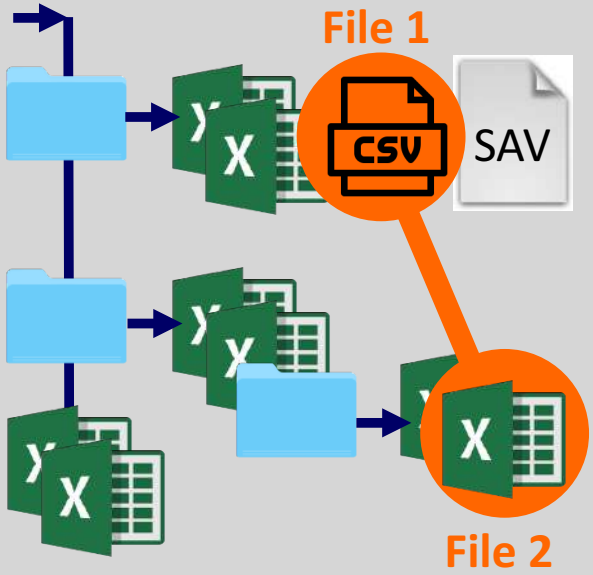
Normally Distributed? (Shapiro-Wilk-Test)



Dog	Days
Chihuahua	348
Akita	123
Dackel	345
Dackel	124
Chihuahua	234
Akita	59
Akita	3
Boxer	56
Beagle	394
Chihuahua	404
Beagle	67
Mops	65
chihuahua	54
Boxer	76

Compare Data Across Files

File Collectivities:



Compare Pair-Wise:
Similarity of
- Data Levels
- Frequencies

("Normalized Scalar Product" *R*)

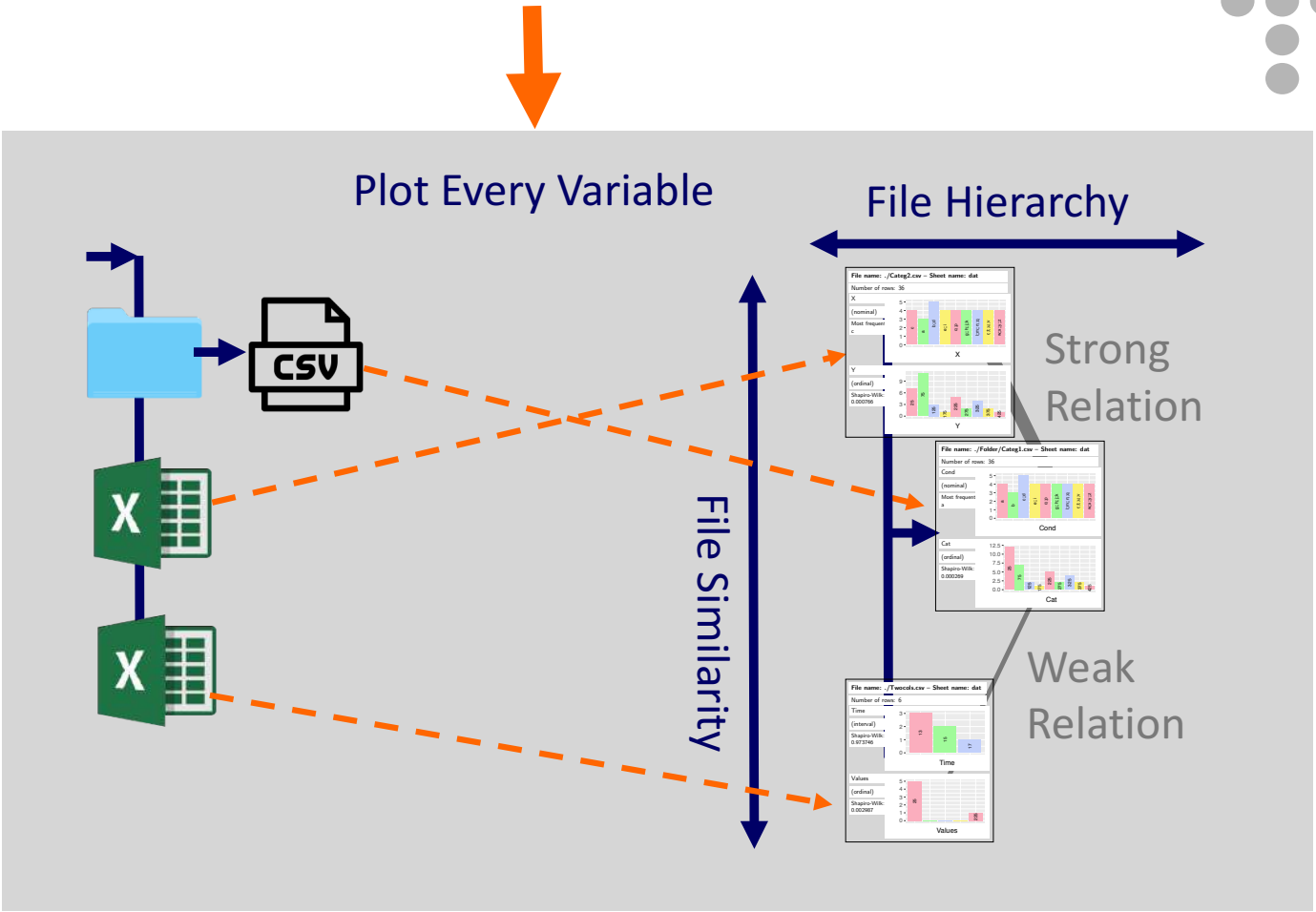
File 1

Dog	Days
Chihuahua	348
Akita	123
Dackel	345
Dackel	124
Chihuahua	234
Akita	59
Akita	3
Boxer	56

File 2

Breed	Age (d)
Akita	348
Akita	123
Dackel	345
Dackel	24
Chihuahua	234
Akita	59
Akita	3
Boxer	56

Visualize
All Data
Information



OUTLINE OF THIS TALK

- Guiding Principle of Pre-Processing: Using Data Categories
- Working Out Step by Step – Examples with simple data sets
- Demonstration of the Tool for Data of an Experimental study
- Conclusions

Data Preparation and Graph Presentation

Variable X:
Character (nominal)

Variable Y:
Integer (ordinal?)

Identify header lines (pure characters)

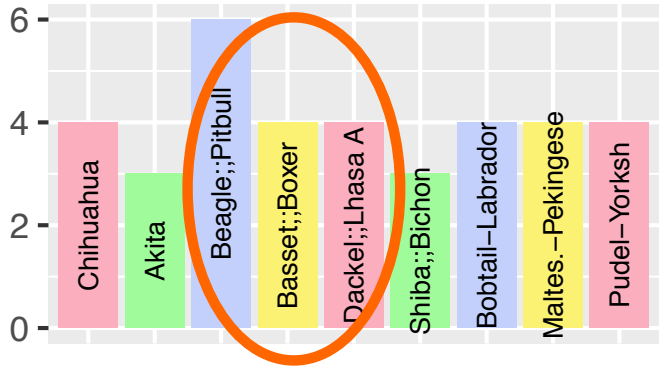
Identify body (data columns)

Dog	Days
Chihuahua	348
Akita	123
Dackel	345
Dackel	124
Chihuahua	234
Akita	59
Akita	3

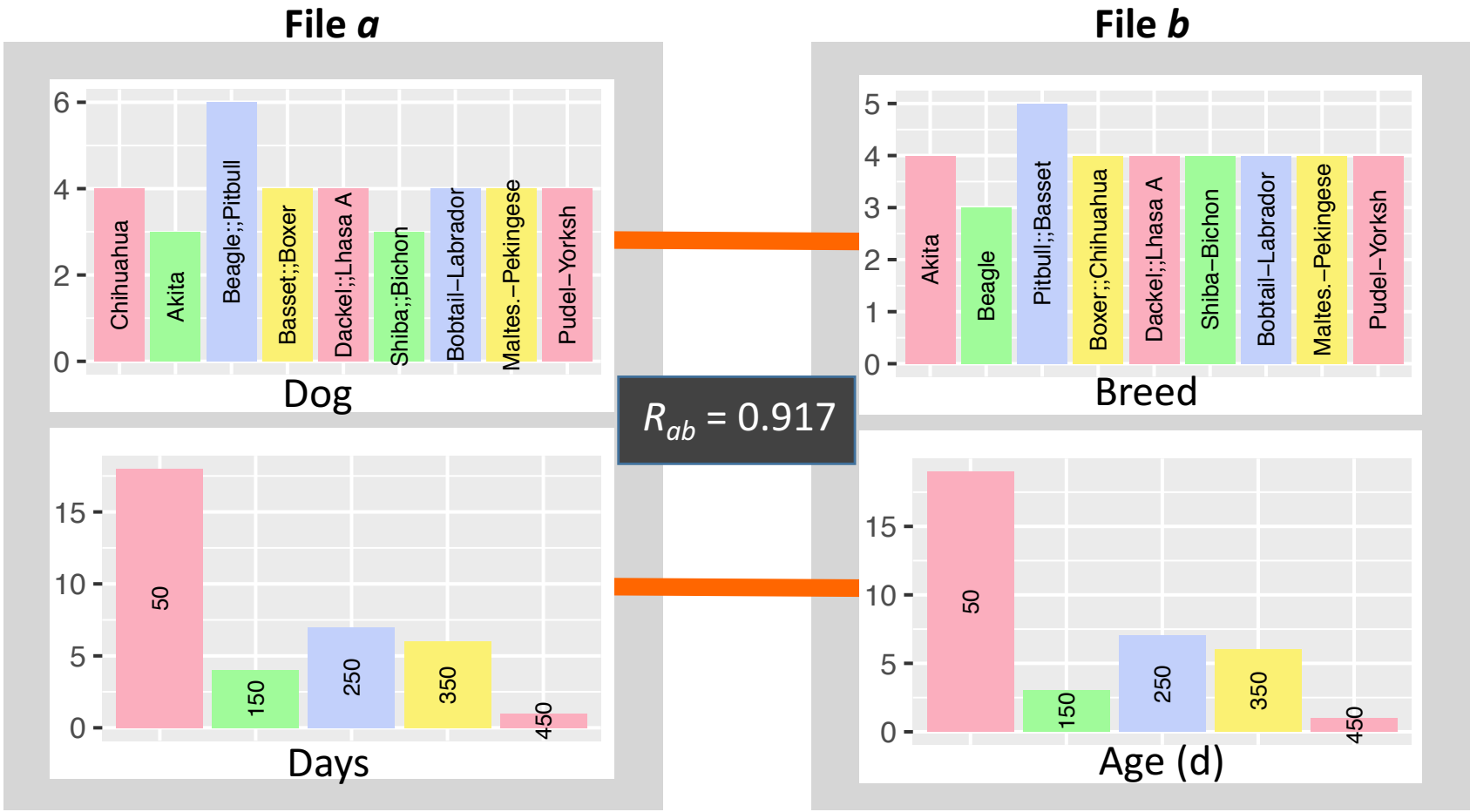
Levels *l*	Frequencies *f*
Akita	3
Chihuahua	2
Dackel	2

For Large Numbers of Levels – Group the Lower Frequency Levels!

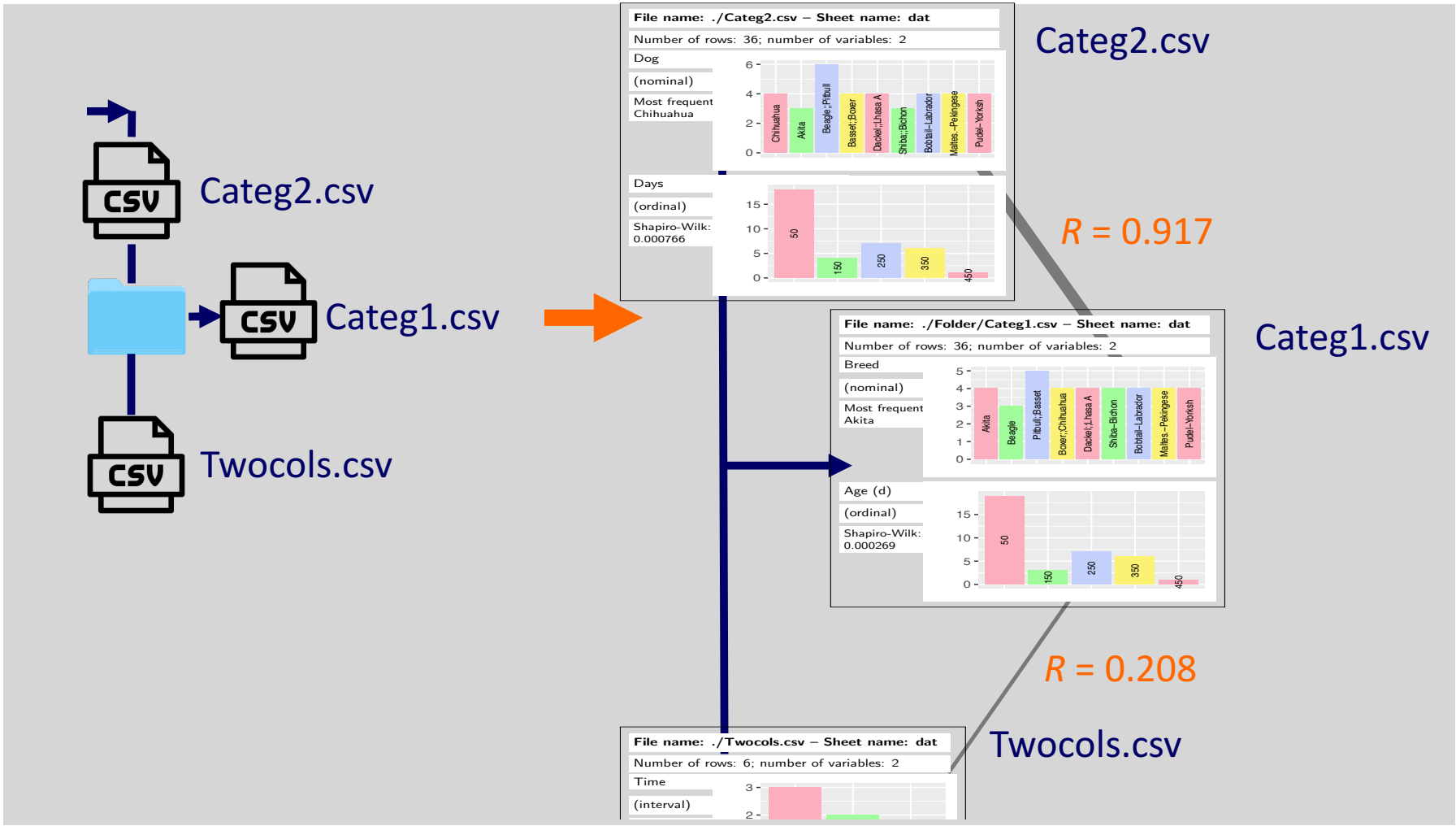
Level groups	Chihuahua
Frequencies	4
Level groups	Akita
Frequencies	3
Level groups	Beagle;;Pitbull
Frequencies	6
Level groups	Boxer;;Dackel
Frequencies	4
Level groups	Lhasa A;;Shiba
Frequencies	4
Level groups	Basset;;Bullt



“Normalized Scalar Product” $R_{ab} = \frac{\sum_l f_a(l)\, f_b(l)}{[\sum_l f_a^2(l) \sum_l f_b^2(l)]^{1/2}}$



Overall Visualisation – Simple Data Set



OUTLINE OF THIS TALK

- Guiding Principle of Pre-Processing: Using Data Categories
- Working Out Step by Step – Examples with small data sets
- Demonstration of the Tool for Data of an Experimental study
- Conclusions

Example: Experimental Study on Similarity Based Learning**

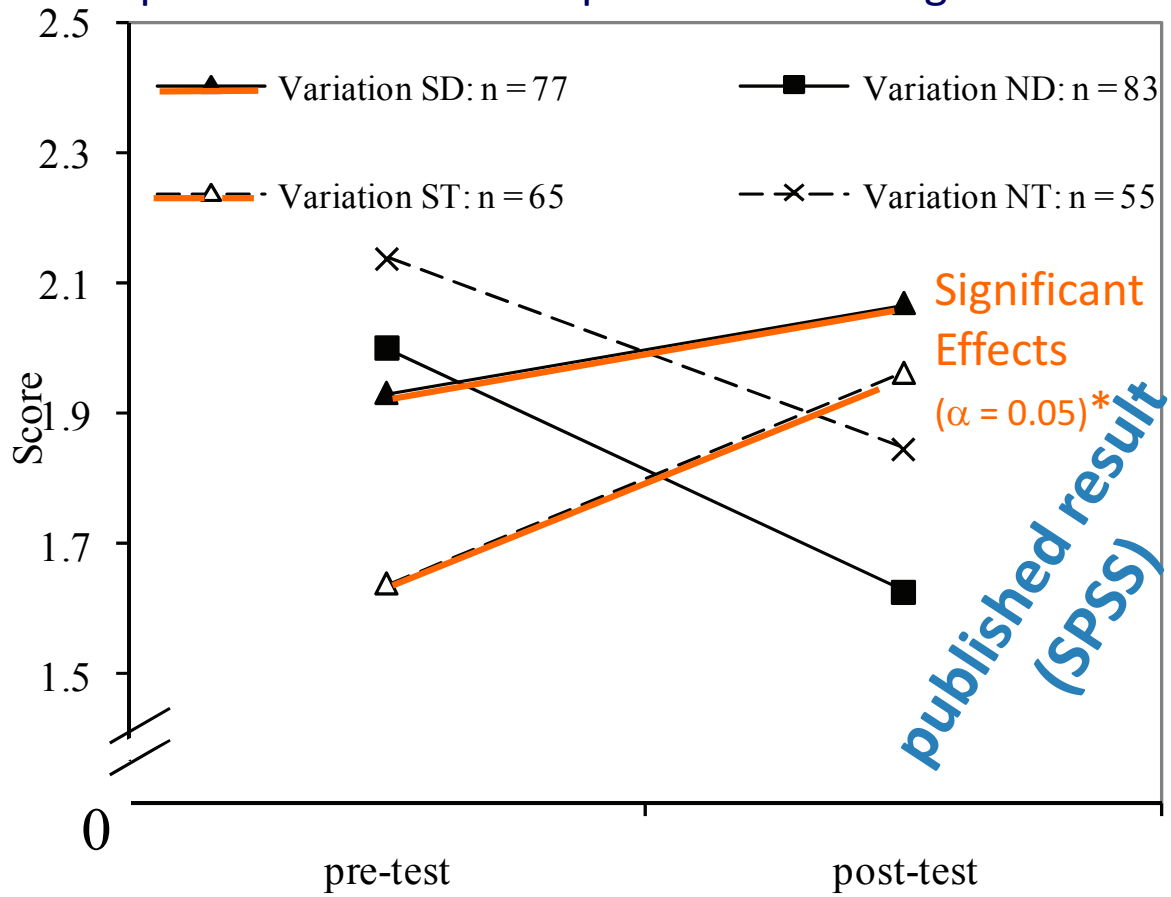
What Type of Instruction is Good for Concept Learning?

Can we combine analogies with discovery learning units and keep benefits of both techniques?

** P. Mandrin, D. Preckel, Effect of similarity-based guided discovery learning on conceptual performance, School Science and Mathematics, **109** (2009), 133–145

Example: Experimental Study on Similarity Based Learning

Conceptual Performance Improvement during interventions



Experimental Groups:

SD: Similarity based Discovery Learning

ST: Similarity based Text Learning

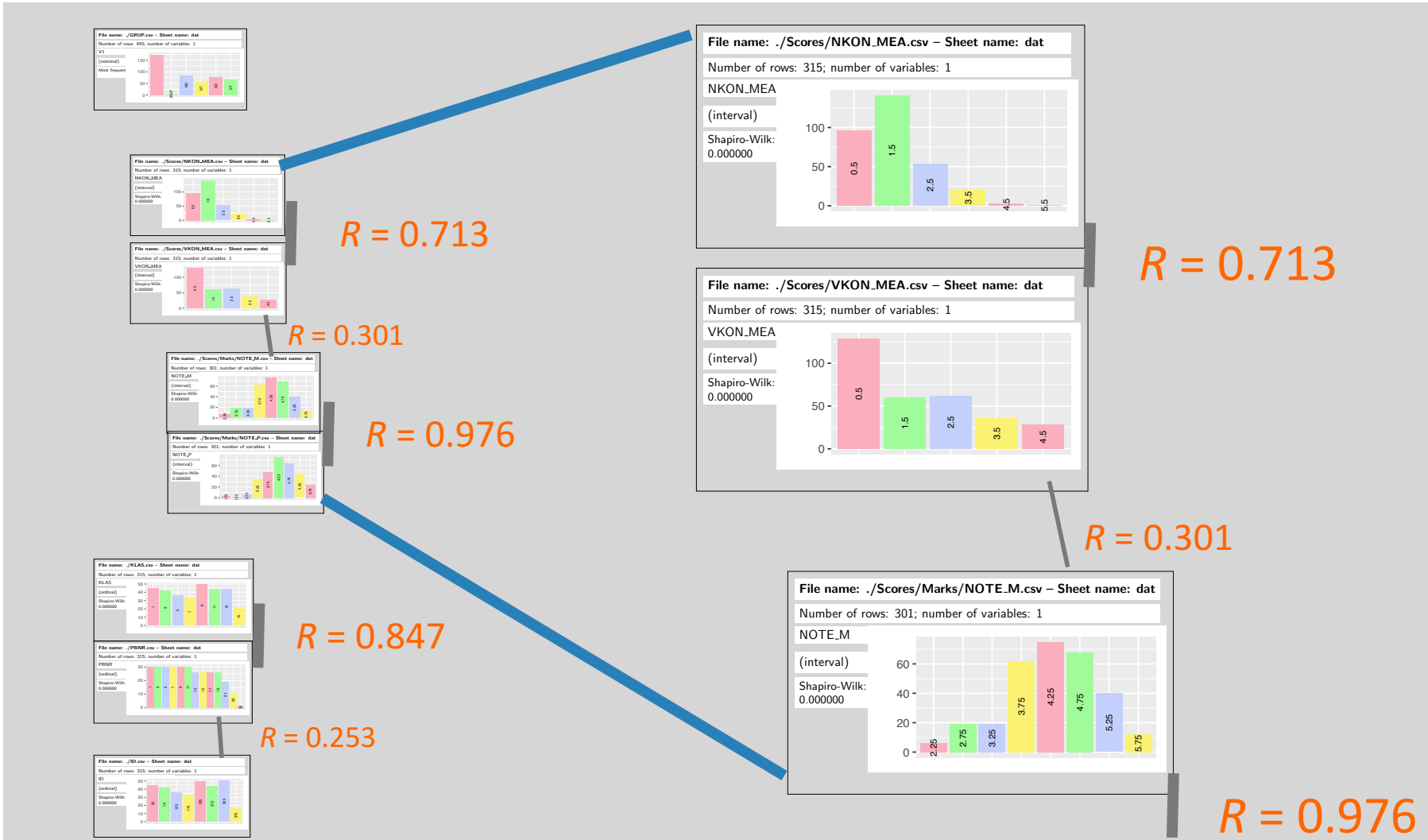
ND: Normal Discovery Learning

Control Group:

NT: Normal Text Learning

* LSD- and Bonferoni-test

Preprocessor: Overview of data (8 files / 2 subfolders)



Can we Refine the Pre-Analysis?
Yes: Dependence between Variables (Score vs Group)

Preprocessor Results:

Variable: Concept Score after intervention	Shapiro-Wilk-test (poss. normally distributed if >0.05†)
SD	0.9428
ND	0.9371
ST	0.9256
NT	0.9214
Group Effect	$p < 0.05$ *

SPSS Results (as published**):

Variable: concept score after intervention	Multiple Comparisons
SD	$p < 0.01$ *
ND	n.s.
ST	$p < 0.05$ *
NT	n.s.
Group Effect	$p < 0.05$ *

* Significant ($\alpha = 0.05$) † to be supplemented by a qq-plot

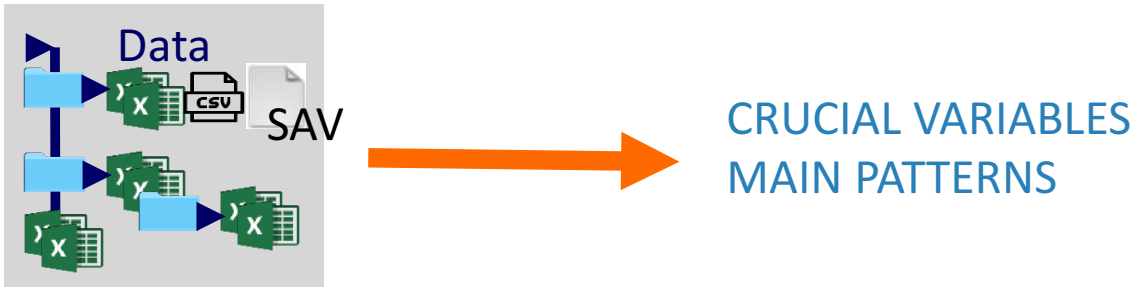
** P. Mandrin, D. Preckel, Effect of similarity-based guided discovery learning on conceptual performance, School Science and Mathematics, **109** (2009), 133–145

OUTLINE OF THIS TALK

- Guiding Principle of Pre-Processing: Using Data Categories
- Working Out Step by Step – Examples with small data sets
- Demonstration of the Tool for Data of an Experimental study
- Conclusions

Conclusions

- Preprocessing with Unknown Data Structure
- Arbitrary Data are Preprocessed Fully Automatically (Robustness)
- Requires Extensive Categorization of Data
- Beware: Every Model of Categorization has its Limitations
- Graphs must be simple (information reduction)
- Similar Data Patterns are Effectively Identified Across Files
- In-Depth Statistical Analyses also Possible but Non-Exhaustive (Test Reliabilities)
- Applications: Efficient Support for Quick Overview and Preparation of Data Mining



IMSD

THANK YOU!