

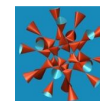


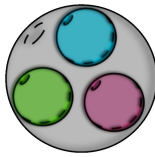
Qualitätsanalyse durch Clustern von Datenwerten

KONDA Hands-On Community-Workshop

Arno Kesper und Viola Wenz

02.02.2022, 10-16 Uhr





Zeitplan

10:00 Einführung (1 h)

11:00  Training (30 min)

11:30  Analyse Teil 1 (10 min + 1 h)

Pause (1 h)

 Analyse Teil 2 (1 h)

Individuelle
Einteilung

14:40  Fazit zur Analyse (15 min)

14:55  Auswertung Teil 1 (5 + 20 min)

15:20  Auswertung Teil 2 (20 min)

15:40 Abschlussbesprechung (20 min)









Download & Entpacken

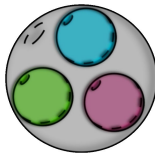
- Laden Sie sich das Clustering-Tool herunter:
`tinyurl.com/konda-dvctool`
- Entpacken Sie die Datei DataValueClustering.zip





Zeitplan

- **10:00 Einführung (1 h)**
- 11:00  Training (30 min)
- 11:30  Analyse Teil 1 (10 min + 1 h)
- Pause (1 h)
-  Analyse Teil 2 (1 h)
- 14:40  Fazit zur Analyse (15 min)
- 14:55  Auswertung Teil 1 (5 + 20 min)
- 15:20  Auswertung Teil 2 (20 min)
- 15:40 Abschlussbesprechung (20 min)



Ziele & Nutzen des Workshops

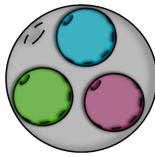
Für die Community

- Kennenlernen eines innovativen Tools zur explorativen Datenqualitätsanalyse
- Direkte Einflussnahme auf die zukünftige Verbesserung und Erweiterung des Tools

Für die Forschung & Entwicklung

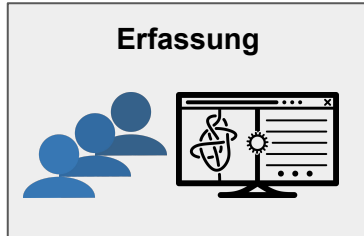
- Evaluation des Tools in praxisnahem Szenario
- Erkennen von Stärken, Schwächen und neuen Anforderungen

Datenqualitätsprobleme und ihre Ursachen



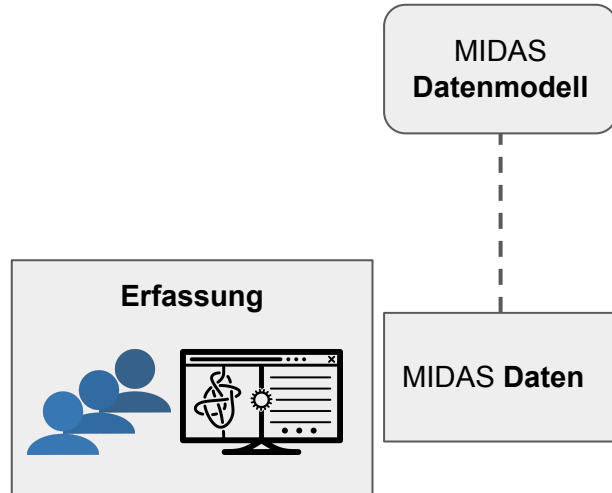


Datenqualitätsprobleme und ihre Ursachen



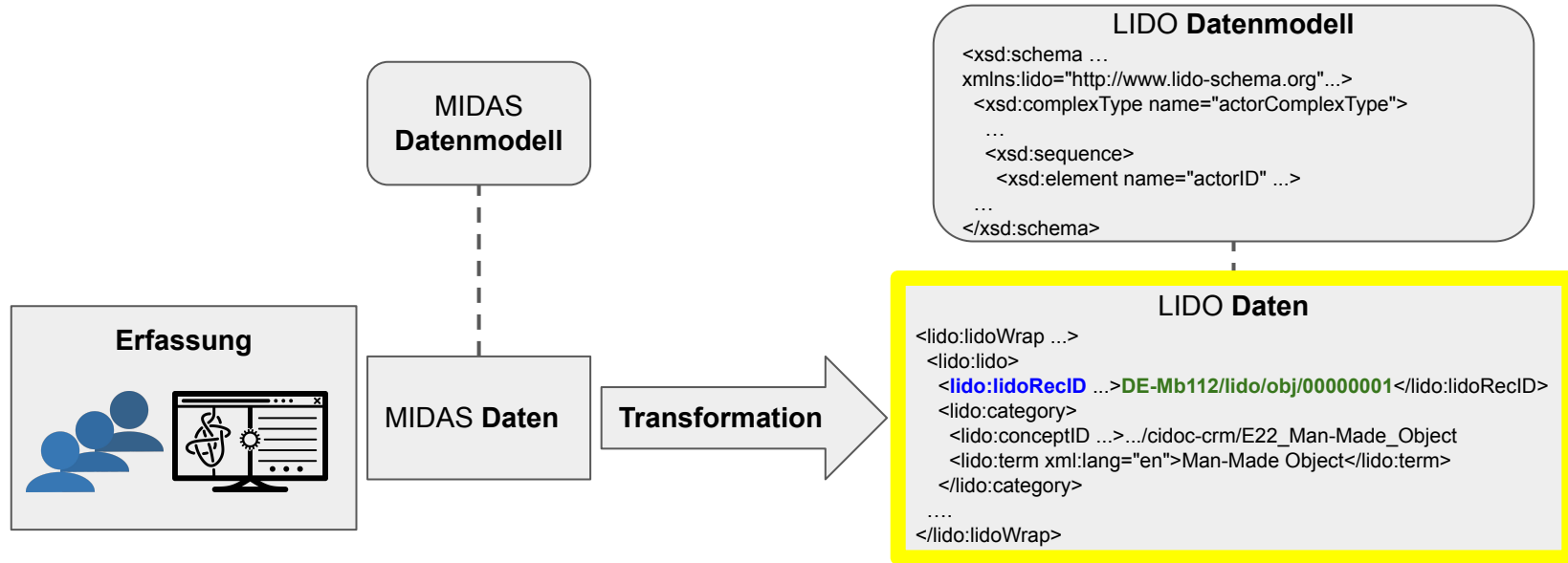


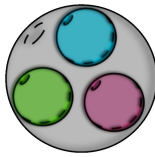
Datenqualitätsprobleme und ihre Ursachen



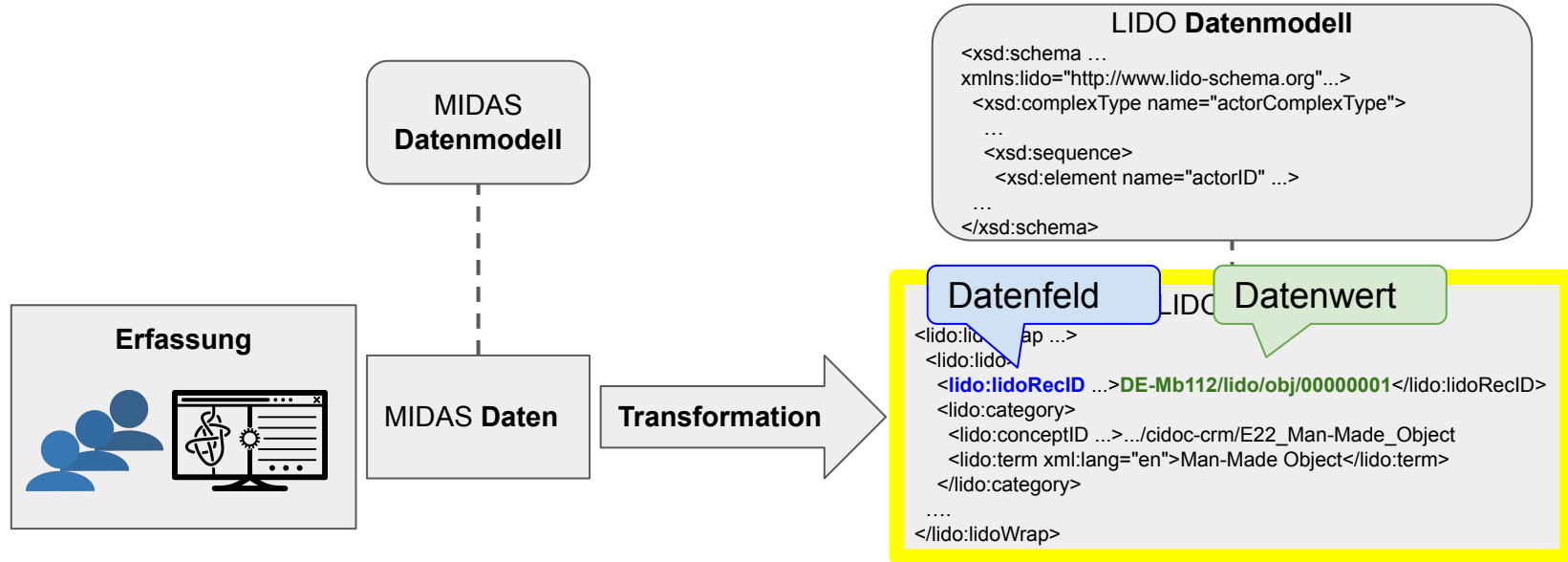


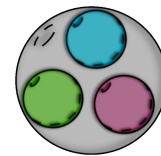
Datenqualitätsprobleme und ihre Ursachen



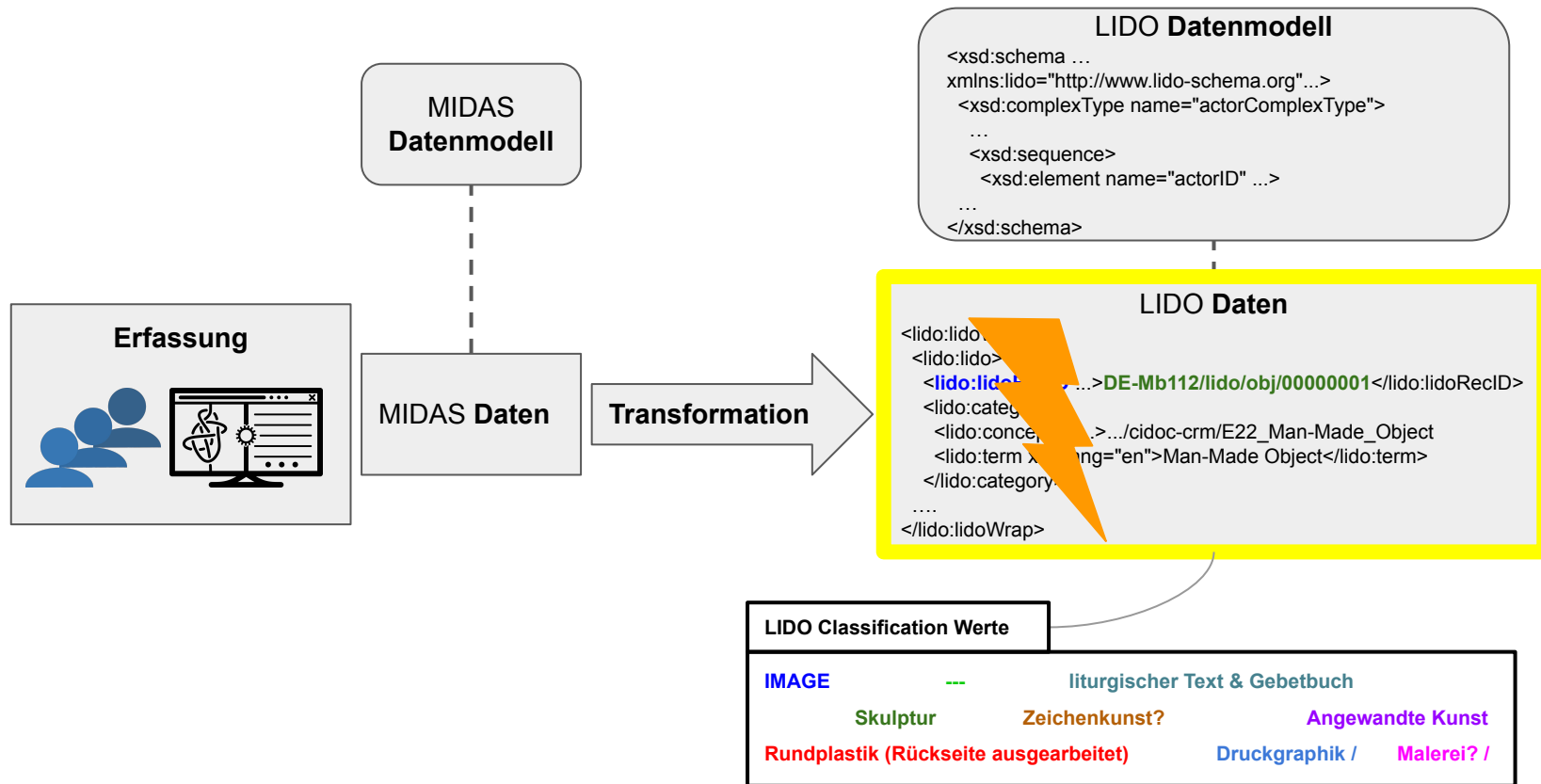


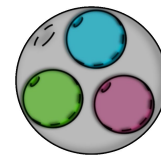
Datenqualitätsprobleme und ihre Ursachen



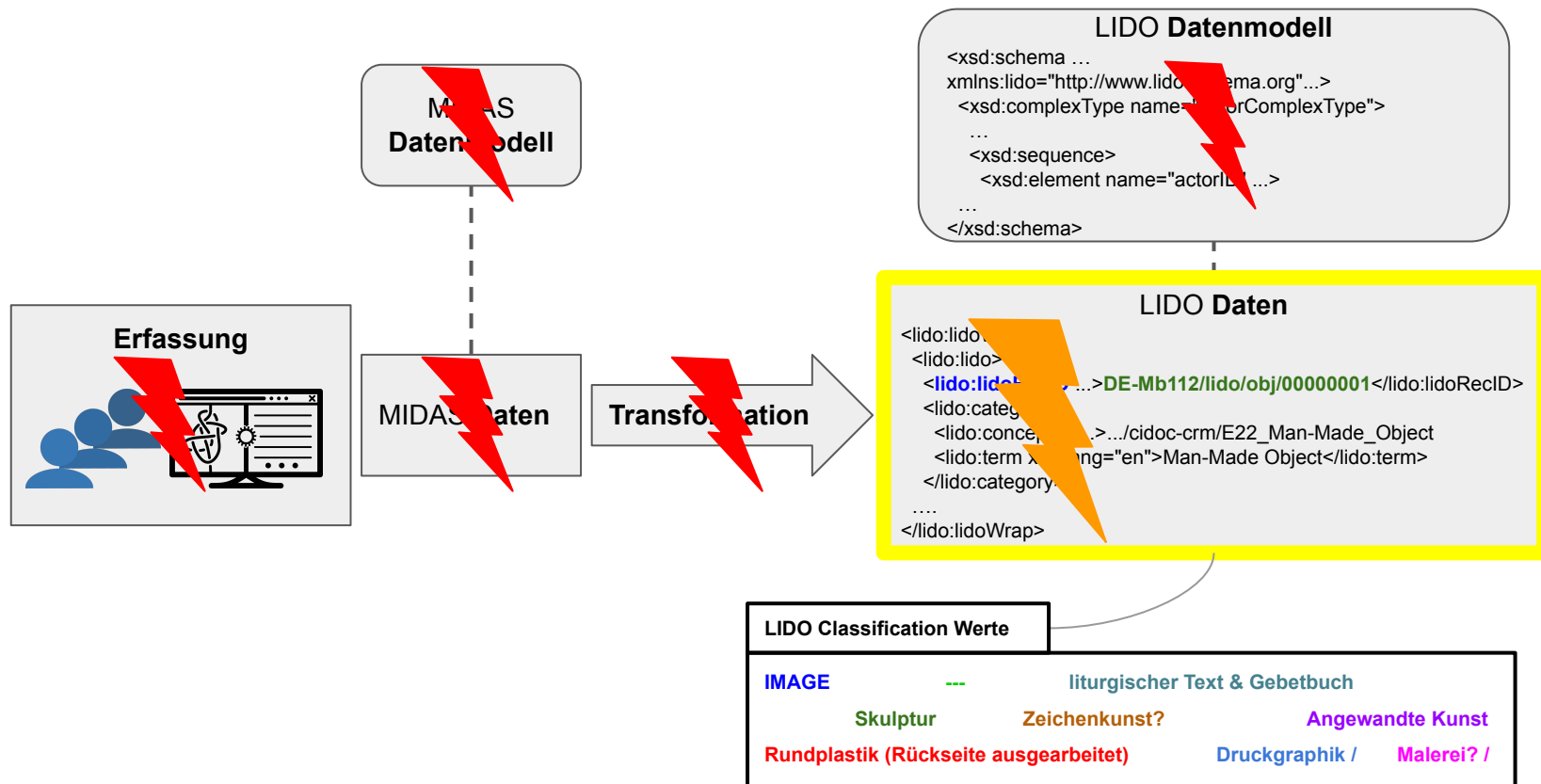


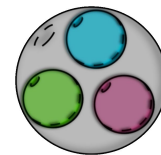
Datenqualitätsprobleme und ihre Ursachen



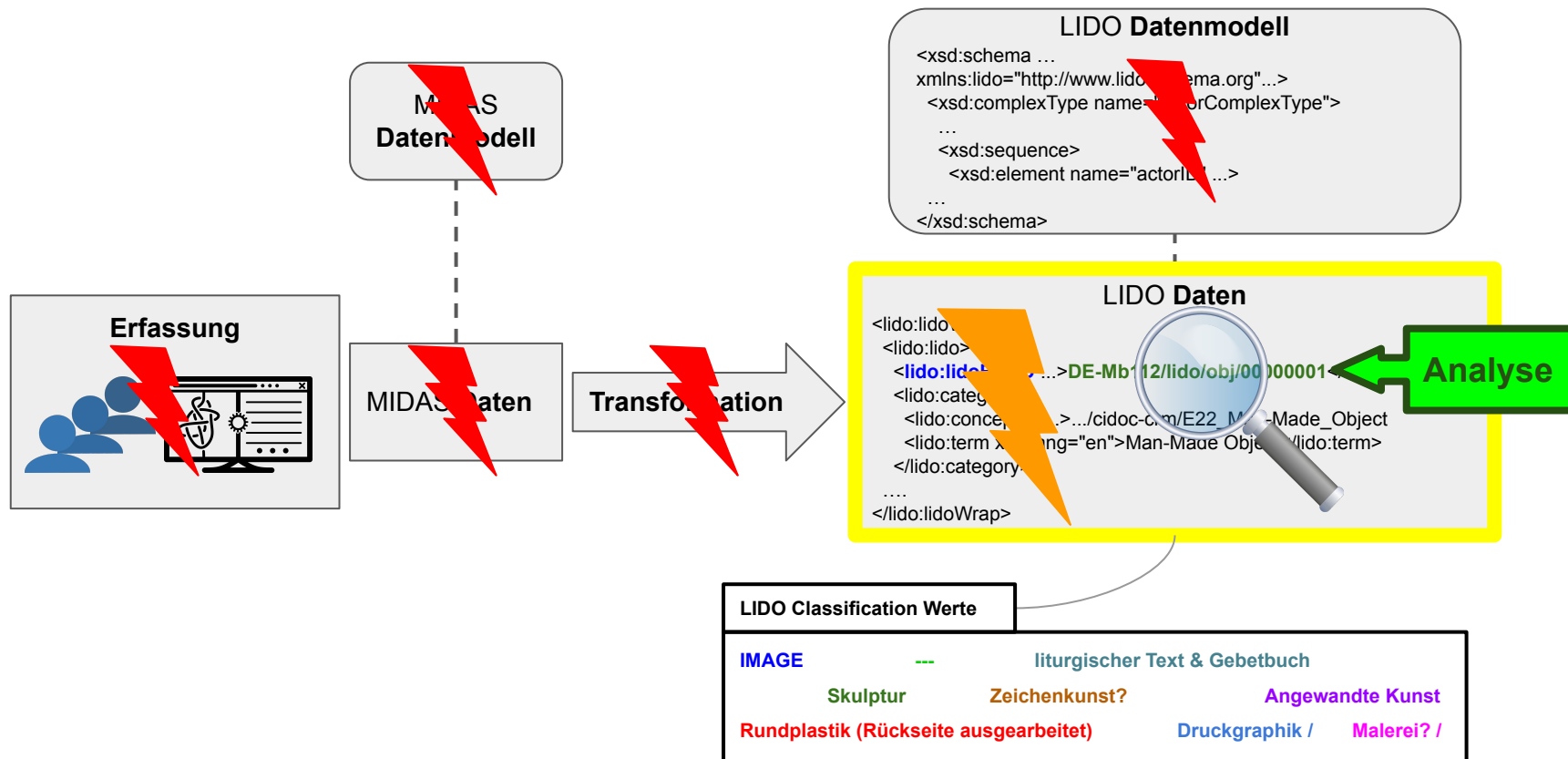


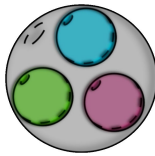
Datenqualitätsprobleme und ihre Ursachen





Datenqualitätsprobleme und ihre Ursachen





Motivation

Beobachtung: Werte im selben Feld sind oft heterogen

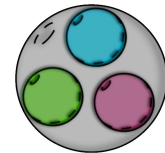
LIDO Classification

IMAGE	--	liturgischer Text & Gebetbuch
Skulptur	Zeichenkunst?	Angewandte Kunst
Rundplastik (Rückseite ausgearbeitet)	Druckgraphik /	Malerei? /

Wie verschafft man sich hier einen Überblick?

Was verrät die Heterogenität bzgl. Datenqualität?

Wo liegt die Ursache der Heterogenität?



Running Example: Classification

/lidoWrap/lido/descriptiveMetadata/objectClassificationWrap/classificationWrap/classification/term

classificationWrap (element)

Definition: A wrapper for classification information.

Sequence:

classification (0-unbounded)

Definition: Concepts used to categorize an object / work by grouping it together with others on the basis of similar characteristics.
How to record: The category belongs to a systematic scheme (classification) which groups objects of similar characteristics according to uniform aspects. This grouping / classification may be done according to material, form, shape, function, region of origin, cultural context, or historical or stylistic period. In addition to this systematic grouping it may also be done according to organizational divisions within a museum (e.g., according to the collection structure of a museum). If the object / work is assigned to multiple classifications, repeat this element. Preferably taken from a published controlled vocabulary.

Extension (base [lido:conceptComplexType](#))

Attribute [lido:type](#)

Attribute [lido:sortorder](#)

LIDO Classification		
IMAGE	--	liturgischer Text & Gebetbuch
Skulptur	Zeichenkunst?	Angewandte Kunst
Rundplastik (Rückseite ausgearbeitet)	Druckgraphik /	Malerei? /



Das Clustering-Tool

Heterogene Datenwerte

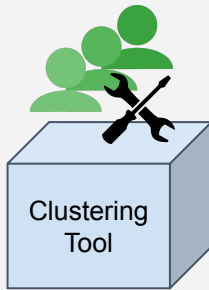
IMAGE
--
Angewandte Kunst
liturgischer Text & Gebetbuch
Zeichenkunst?
...

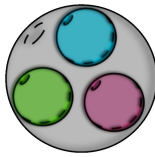


Das Clustering-Tool

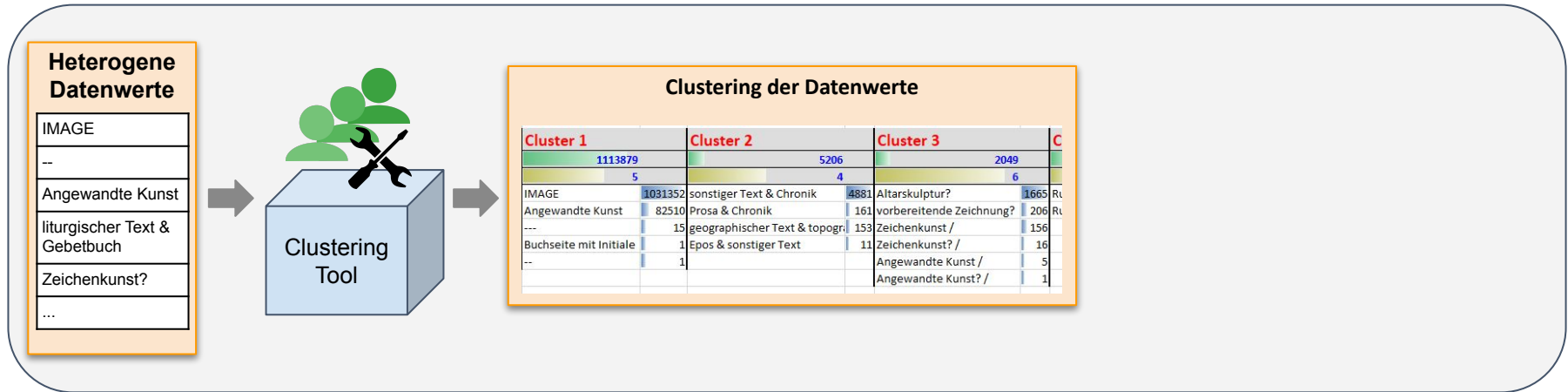
Heterogene Datenwerte

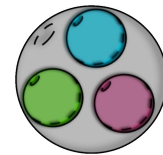
IMAGE
--
Angewandte Kunst
liturgischer Text & Gebetbuch
Zeichenkunst?
...



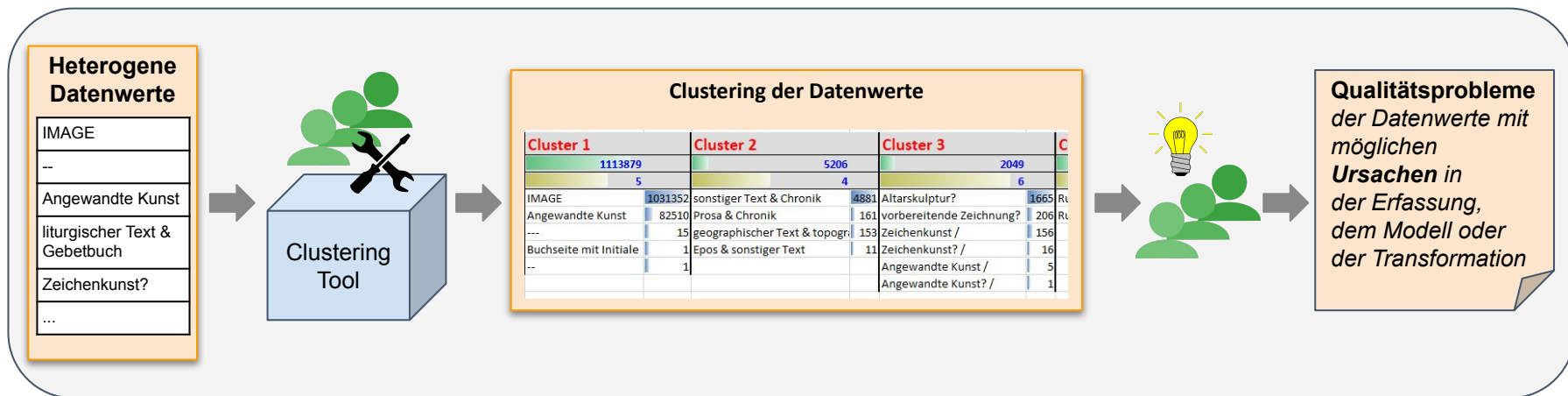


Das Clustering-Tool



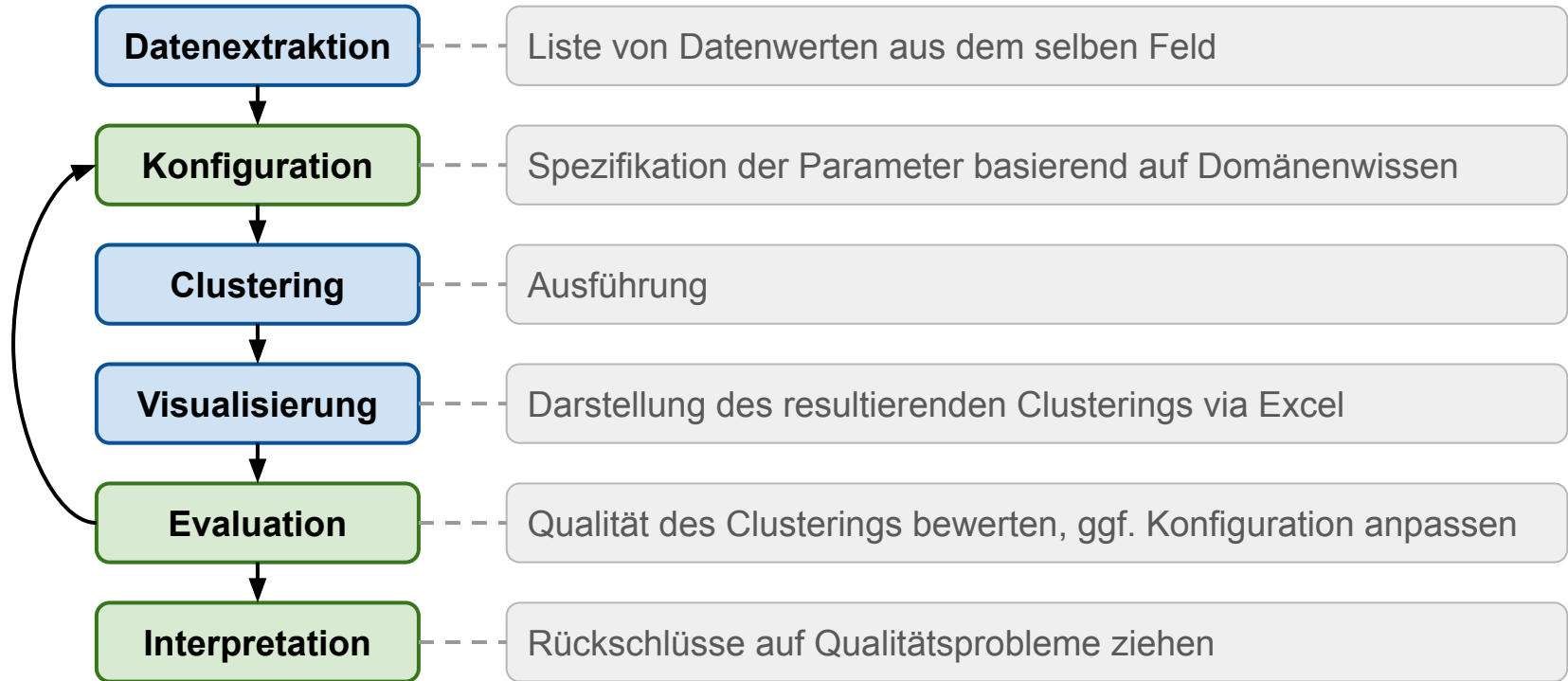


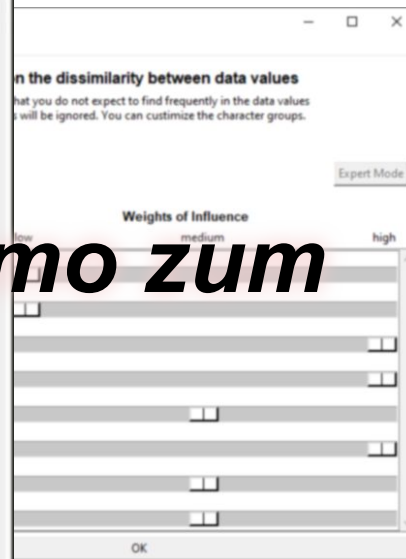
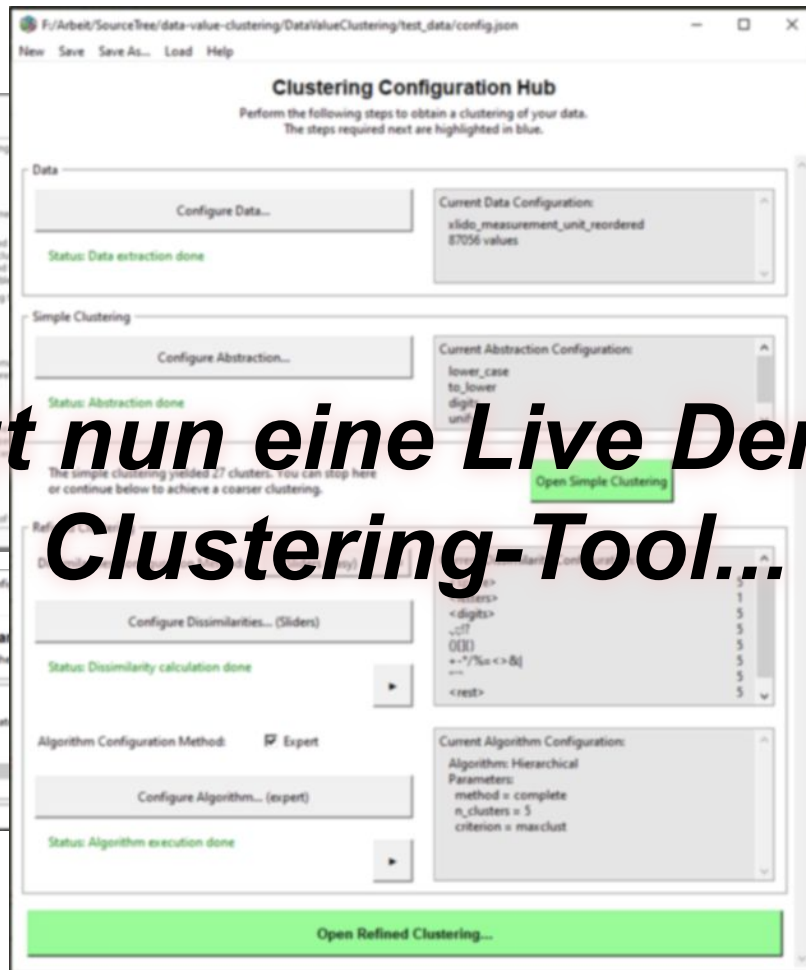
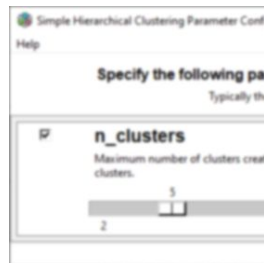
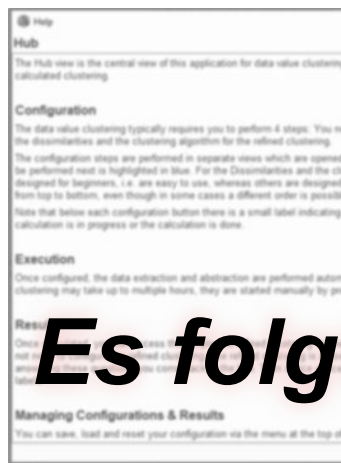
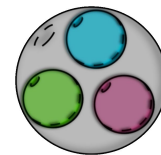
Das Clustering-Tool



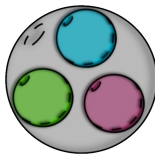


Workflow des Clustering-Ansatzes





Es folgt nun eine Live Demo zum Clustering-Tool...



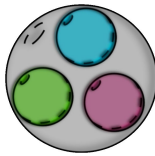
Alternative Methode: Werteliste(n)

	A	B	C
1	IMAGE		
2	Tafelmalerei		
3	Malerei		
4	IMAGE		
5	Tafelmalerei		
6	Malerei		
Raw Data Occurrences Alphabetical			

	B	C	D	E
2	Data	#Occurrences		
3	IMAGE	431887		
4	Skulptur	122769		
5	Architektur	100617		
6	Malerei	70312		
Raw Data Occurrences Alphabetical				

	B	C	D	E
2	Data	#Occurrences		
3	--	1		
4	---	15		
5	Altarskulptur	5667		
6	Altarskulptur /	3		
Raw Data Occurrences Alphabetical				

Zeitplan



10:00 Einführung (1 h)

Fragebogen zum Alias

11:00  Training (30 min)

11:30  Analyse Teil 1 (10 min + 1 h)

Fragebogen zur Werteliste

Pause (1 h)

Fragebogen zum Clustering

 Analyse Teil 2 (1 h)

14:40  Fazit zur Analyse (15 min)

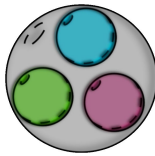
Fragebogen zum Fazit zur Analyse

14:55  Auswertung Teil 1 (5 + 20 min)

15:20  Auswertung Teil 2 (20 min)

Fragebogen zur Auswertung

15:40 Abschlussbesprechung (20 min)



Welche Daten betrachten wir?

Format: LIDO v1.0 (www.lido-schema.org/schema/v1.0/lido-v1.0-schema-listing.html)

Thema: Objekte der materiellen Kultur

Entstehung:

- Manuelle Erfassung in MIDAS
- Transformation zu LIDO

Felder:

- Training: “ShapeMeasurements”
- Analyse: “ActorName” und “RepositoryName”

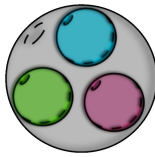


Aktuelle Aufgabe: Alias ausdenken

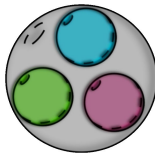
- Denken Sie sich ein Alias aus und notieren Sie es lokal
- Tragen Sie das Alias hier ein:

`tinyurl.com/konda`





5 Minuten Pause im Gange...









Aktuelle Aufgabe: Alias Ordner checken

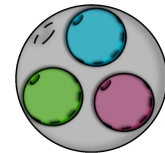
- Bei Google abmelden oder Privates Fenster nutzen
- Unterordner in Drive checken:
`tinyurl.com/konda-dvcworkshop`
 - Anleitung (PDF)
 - Zuteilung zu einer Gruppe
 - Anweisungen wann welche Daten mit welcher Methode analysiert werden sollen usw.
 - Links zu Fragebögen und LIDO Dokumentation
 - Werteliste (XLSX)
- ↻ Später kommt hinzu:
 - Ihre beiden ausgefüllten Fragebögen (PDF)
 - Zwei zu bewertende ausgefüllte Fragebögen einer anderen Person (PDF)





Zeitplan

- 10:00 Einführung (1 h)
- 11:00  **Training (30 min)**
- 11:30  Analyse Teil 1 (10 min + 1 h)
- Pause (1 h)
-  Analyse Teil 2 (1 h)
- 14:40  Fazit zur Analyse (15 min)
- 14:55  Auswertung Teil 1 (5 + 20 min)
- 15:20  Auswertung Teil 2 (20 min)
- 15:40 Abschlussbesprechung (20 min)



ShapeMeasurements

/lidoWrap/lido/descriptiveMetadata/objectIdentificationWrap/objectMeasurementsWrap/objectMeasurementsSet/objectMeasurements/shapeMeasurements

shapeMeasurements (0-unbounded)

Definition: The shape of an object / work. Used for unusual shapes (e.g., an oval painting).
How to record: Example values: oval, round, square, rectangular, and irregular.

Simple content

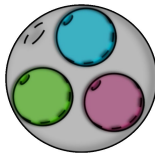
Extension (base [lido:textComplexType](#))

Attribute [lido:sortorder](#)

LIDO Shape Measurements			
oval	fünfeckig (unregelmäßig)	Saalkirche? /	Kopfbahnhof
	Eckhaus mit Turm	24-zeilig	Doppelstatue
gerahmt	5/ 8-Schluss	gerahmt (Architekturrahmen)	

Aktuelle Aufgabe: Training

30 min









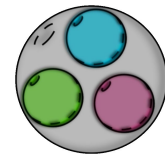
- Entpacken Sie die Datei DataValueClustering.zip
 - Führen Sie die Anwendung DataValueClustering.exe aus
 - Machen Sie sich mit dem Clustering-Tool vertraut
 - Nutzen Sie die Feldwerte ShapeMeasurements
 - Durchlaufen Sie alle Konfigurationsphasen
 - Betrachten Sie das resultierende Simple und Refined Clustering
- Stellen Sie jegliche Fragen im Plenum





Zeitplan

10:00	Einführung (1 h)	
11:00	 Training (30 min)	
➤ 11:30	 Analyse Teil 1 (10 min + 1 h)	 Individuelle Einteilung
	Pause (1 h)	
	 Analyse Teil 2 (1 h)	
14:40	 Fazit zur Analyse (15 min)	
14:55	 Auswertung Teil 1 (5 + 20 min)	
15:20	 Auswertung Teil 2 (20 min)	
15:40	Abschlussbesprechung (20 min)	



ActorName

/lidoWrap/lido/descriptiveMetadata/**eventWrap**/eventSet/event/**eventActor**/actorInRole/actor/**nameActorSet**/appellationValue

actorComplexType (complex type)

Definition: Contains identifying and indexing actor information.

How to record: Data values of the type attribute: person, corporation, family, group.

Sequence:

actorID ([lido:identifierComplexType](#) 0-unbounded)

nameActorSet ([lido:appellationComplexType](#) 1-unbounded)

Definition: A wrapper for name elements.

How to record: if there exists more than one name for a single actor, repeat Name Actor Set.

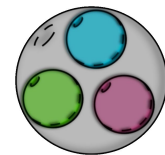
Notes: Indicates names, appellations, or other identifiers assigned to an individual, group of people, firm or other corporate body, or other entity.

nationalityActor (0-unbounded)

vitalDatesActor ([lido:dateComplexType](#) 0-1)

genderActor ([lido:textComplexType](#) 0-unbounded)

LIDO ActorName		
unbekannt	Rembrandt Harmensz, van Rijn	Rembrandt
Jean (Berry, Herzog, 1, le Magnifique)	?	Miro, Joan
Palant, Werner (Graf, 2) & Palant, Elverad von & Bergerhausen, Margarete von		



RepositoryName (Teil 1)

/lidoWrap/lido/descriptiveMetadata/objectIdentificationWrap/repositoryWrap/repositorySet/repositoryName/legalBodyName/appellationValue

repositorySetComplexType (complex type)

Definition: Wrapper for designation and identification of the institution of custody and, possibly, indication of the exact location of the object.

How to record: If there are several designations known, e.g., a current one and former ones (see: type attribute), repeat the element. Data values of the type attribute: current, former.

Sequence:

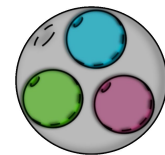
repositoryName ([lido:legalBodyRefComplexType](#) 0-1)

Definition: Unambiguous identification, designation and weblink of the institution of custody.

workID (0-unbounded)

repositoryLocation ([lido:placeComplexType](#) 0-1)

LIDO RepositoryName		
? (Paris, Prinz) (private Sammlung) (Paris)	3. Register	
Ägyptisches Museum (Kairo)		
Apotheke	zwischen Ruhrortsbrücke und Eisenbahnhofen	
Universitätsmuseum für Kunst und Kulturgeschichte (Marburg)		
Zur Krone	Sankt Lorenz ob Katsch	x



RepositoryName (Teil 2)

/lidoWrap/lido/descriptiveMetadata/objectIdentificationWrap/repositoryWrap/repositorySet/repositoryName/legalBodyName/appellationValue

legalBodyRefComplexType (complex type)

Definition: Reference information to a legal body.

Sequence:

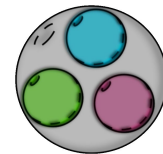
legalBodyID ([lido:identifierComplexType](#) 0-unbounded)

legalBodyName ([lido:appellationComplexType](#) 0-unbounded)

Definition: Appellation of the institution or person.

legalBodyWeblink ([lido:webResourceComplexType](#) 0-unbounded)

LIDO RepositoryName		
? (Paris, Prinz) (private Sammlung) (Paris)	3. Register	
Ägyptisches Museum (Kairo)		
Apotheke	zwischen Ruhrortsbrücke und Eisenbahnhofen	
Universitätsmuseum für Kunst und Kulturgeschichte (Marburg)		
Zur Krone	Sankt Lorenz ob Katsch	x



Es folgt nun eine Live Demo zu den Fragebögen...

Fragebogen zum Clustering

Bitte beschreiben Sie im Folgenden jedes identifizierte Problem der Datensätze und seine möglichen Ursachen.

Bitte beschreiben Sie im Folgenden jedes identifizierte Problem der Datensätze und seine möglichen Ursachen.

Single vs. Refined Clustering

Fragebogen zur Werteliste

Bitte beschreiben Sie im Folgenden jedes identifizierte Problem der Datensätze und seine möglichen Ursachen.

Bitte beschreiben Sie im Folgenden jedes identifizierte Problem der Datensätze und seine möglichen Ursachen.

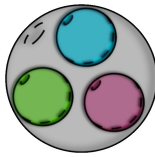


Zeitplan

10:00	Einführung (1 h)	
11:00	 Training (30 min)	
➤ 11:30	 Analyse Teil 1 (10 min + 1 h)	 Individuelle Einteilung
	Pause (1 h)	
	 Analyse Teil 2 (1 h)	
14:40	 Fazit zur Analyse (15 min)	
14:55	 Auswertung Teil 1 (5 + 20 min)	
15:20	 Auswertung Teil 2 (20 min)	
15:40	Abschlussbesprechung (20 min)	

Aktuelle Aufgabe: Analyse und Pause

1+1+1 h



- Befolgen Sie Ihre persönliche Anleitung
 - Zuerst Analyse Teil 1
 - Danach Analyse Teil 2
- Bearbeitungszeit jeweils max. 1h
 - Früher aufhören wenn Sie keine weiteren Probleme finden
- Einstündige Mittagspause nach Belieben

[tinyurl.com/
konda-dvcworkshop](https://tinyurl.com/konda-dvcworkshop)



Vorschlag

11:40 - 12:40 Analyse Teil 1
12:40 - 13:40 Pause
13:40 - 14:40 Analyse Teil 2

Aktuelle Aufgabe: Analyse und Pause

1+1+1 h



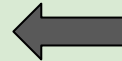
- Befolgen Sie Ihre persönliche Anleitung
 - Zuerst Analyse Teil 1
 - Danach Analyse Teil 2
- Bearbeitungszeit jeweils max. 1h
 - Früher aufhören wenn Sie keine weiteren Probleme finden
- Einstündige Mittagspause nach Belieben

[tinyurl.com/
konda-dvcworkshop](https://tinyurl.com/konda-dvcworkshop)



Vorschlag

11:40 - 12:40 Analyse Teil 1

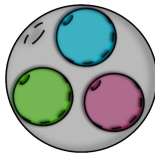


12:40 - 13:40 Pause

13:40 - 14:40 Analyse Teil 2

Aktuelle Aufgabe: Analyse und Pause

1+1+1 h



- Befolgen Sie Ihre persönliche Anleitung
 - Zuerst Analyse Teil 1
 - Danach Analyse Teil 2
- Bearbeitungszeit jeweils max. 1h
 - Früher aufhören wenn Sie keine weiteren Probleme finden
- Einstündige Mittagspause nach Belieben

[tinyurl.com/
konda-dvcworkshop](https://tinyurl.com/konda-dvcworkshop)



Vorschlag

11:40 - 12:40 Analyse Teil 1

12:40 - 13:40 Pause

13:40 - 14:40 Analyse Teil 2



Aktuelle Aufgabe: Analyse und Pause

1+1+1 h



- Befolgen Sie Ihre persönliche Anleitung
 - Zuerst Analyse Teil 1
 - Danach Analyse Teil 2
- Bearbeitungszeit jeweils max. 1h
 - Früher aufhören wenn Sie keine weiteren Probleme finden
- Einstündige Mittagspause nach Belieben

[tinyurl.com/
konda-dvcworkshop](https://tinyurl.com/konda-dvcworkshop)



Vorschlag

11:40 - 12:40 Analyse Teil 1

12:40 - 13:40 Pause

13:40 - 14:40 Analyse Teil 2





Zeitplan

10:00 Einführung (1 h)

11:00  Training (30 min)

11:30  Analyse Teil 1 (10 min + 1 h)

Pause (1 h)

 Analyse Teil 2 (1 h)

➤ 14:40  **Fazit zur Analyse (15 min)**

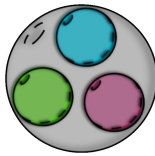
14:55  Auswertung Teil 1 (5 + 20 min)

15:20  Auswertung Teil 2 (20 min)

15:40 Abschlussbesprechung (20 min)

Aktuelle Aufgabe: Fazit zur Analyse

15 min



Füllen Sie den Fragebogen zum Fazit zur Analyse aus



Details: siehe Anleitung im eigenen Ordner



Zeitplan

10:00 Einführung (1 h)

11:00  Training (30 min)

11:30  Analyse Teil 1 (10 min + 1 h)

Pause (1 h)

 Analyse Teil 2 (1 h)

14:40  Fazit zur Analyse (15 min)

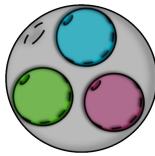
➤ 14:55  **Auswertung Teil 1 (5 + 20 min)**

15:20  Auswertung Teil 2 (20 min)

15:40 Abschlussbesprechung (20 min)

Aktuelle Aufgabe: Auswertung Teil 1

20 min

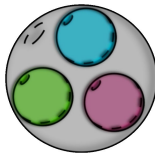


- Auswertung eines ausgefüllten Fragebogens einer anderen Person (PDF in Ordner)
 - Bewertung von beschriebenen Problemen und Ursachen
 - Diese Person hat das Feld mit der anderen Methode analysiert



- Befolgen Sie dazu Ihre persönliche Anleitung → Auswertung Teil 1

`tinyurl.com/
konda-dvcworkshop`



Zeitplan

10:00 Einführung (1 h)

11:00  Training (30 min)

11:30  Analyse Teil 1 (10 min + 1 h)

Pause (1 h)

 Analyse Teil 2 (1 h)

14:40  Fazit zur Analyse (15 min)

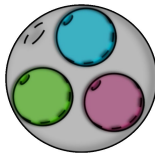
14:55  Auswertung Teil 1 (5 + 20 min)

➤ 15:20  **Auswertung Teil 2 (20 min)**

15:40 Abschlussbesprechung (20 min)

Aktuelle Aufgabe: Auswertung Teil 2

20 min



- Auswertung eines ausgefüllten Fragebogens einer anderen Person (PDF in Ordner)
 - Bewertung von beschriebenen Problemen und Ursachen
 - Diese Person hat das Feld mit der anderen Methode analysiert



- Befolgen Sie dazu Ihre persönliche Anleitung → Auswertung Teil 2

`tinyurl.com/
konda-dvcworkshop`



Zeitplan

10:00 Einführung (1 h)

11:00  Training (30 min)

11:30  Analyse Teil 1 (10 min + 1 h)

Pause (1 h)

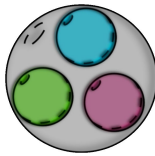
 Analyse Teil 2 (1 h)

14:40  Fazit zur Analyse (15 min)

14:55  Auswertung Teil 1 (5 + 20 min)

15:20  Auswertung Teil 2 (20 min)

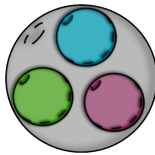
➤ **15:40 Abschlussbesprechung (20 min)**



Was passiert als nächstes?

Anhand der heute gesammelten Daten wollen wir u.A. folgende Forschungsfragen beantworten:

- Mit welcher Methode können *mehr* Datenprobleme und mehr mögliche Ursachen in der Erfassung, dem Datenmodell und der Datentransformation identifiziert werden?
 - Mit welcher Methode können Probleme *schneller* identifiziert werden?
 - Die *Nützlichkeit* welcher Methode wird als Höher wahrgenommen?
 - Welche *Arten von Problemen* werden mit den Methoden häufig identifiziert? Gibt es Unterschiede?
 - Wo liegen Stärken und Schwächen des Clustering-Tools?
 - Was können wir am Tool verbessern?
- Tool besser an Anforderungen der Community anpassen
 - Forschungspapier schreiben



Bei Interesse...

Unsere Publikationen sind auf Zenodo verfügbar:

`zenodo.org/communities/konda-project/`

Source Code des Clustering-Tools:

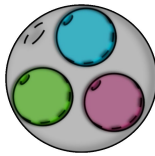
`github.com/Project-KONDA/data-value-clustering`

Interesse an der Auswertung, Weiterentwicklungen usw. gern per E-Mail äußern:

`arno.kesper@uni-marburg.de`

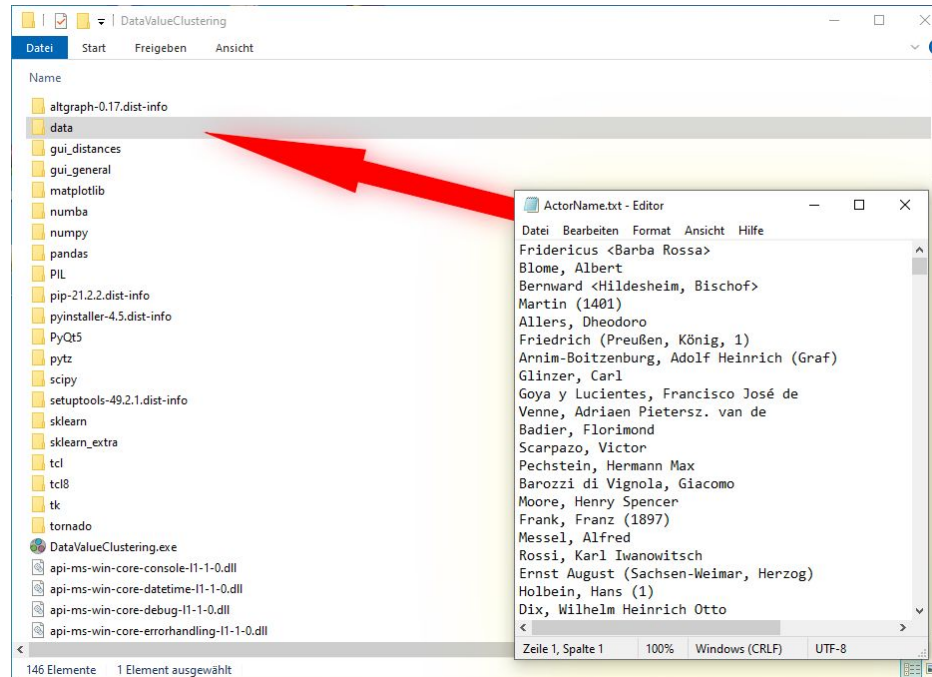
`viola.wenz@uni-marburg.de`

`konda@uni-marburg.de`



Import eigener Daten

- als TXT-Datei in **DataValueClustering/data/**





Fragen, Anmerkungen, Kritik,
Verbesserungsvorschläge, ...?



Vielen Dank für die Teilnahme!