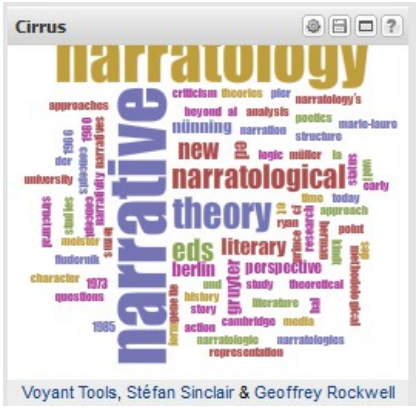
Die explorative Visualisierung von Texten

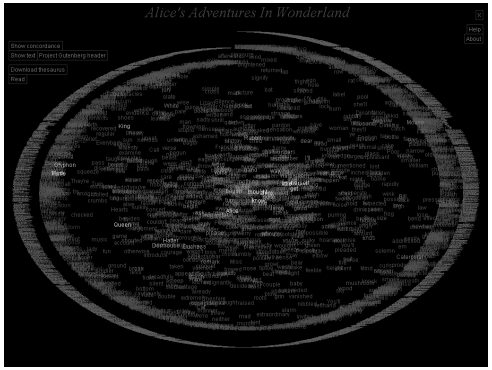
Evelyn Gius und Marco Petris

25.02.2015 DHd2015, Graz

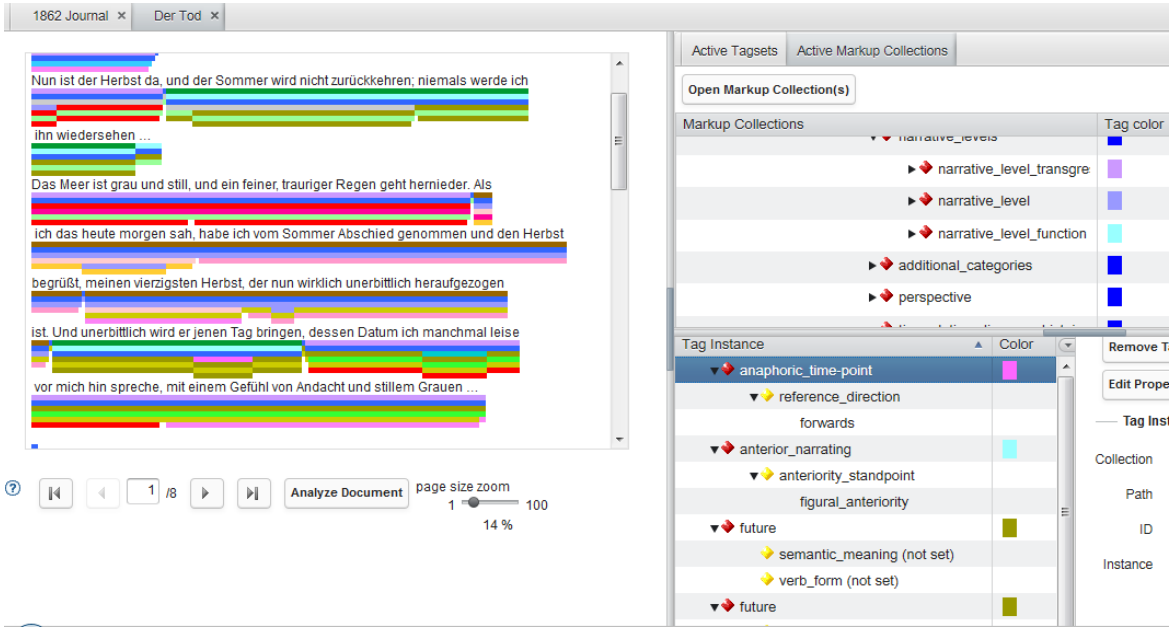
Probleme

- häufig reine Textdaten
- Blackbox
- langer Weg zur Visualisierung





Hermeneutische Ebenen



Ziele

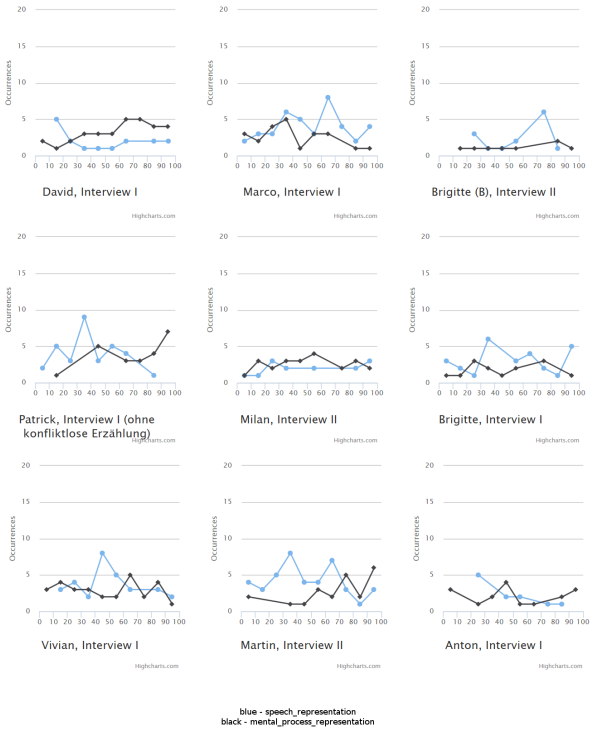
- We need to conceive of every metric ‘as a factor of X’, where X is a point of view, agenda, assumption, presumption, or simply a convention. By qualifying any metric as a factor of some condition, the character of the ‘information’ shifts from self-evident ‘fact’ to constructed interpretation motivated by a human agenda. (Drucker 2014:131)
- kein Blackbox-System
- hypothesengetrieben
- heuristisches Werkzeug

Visualisierung multivariater Daten

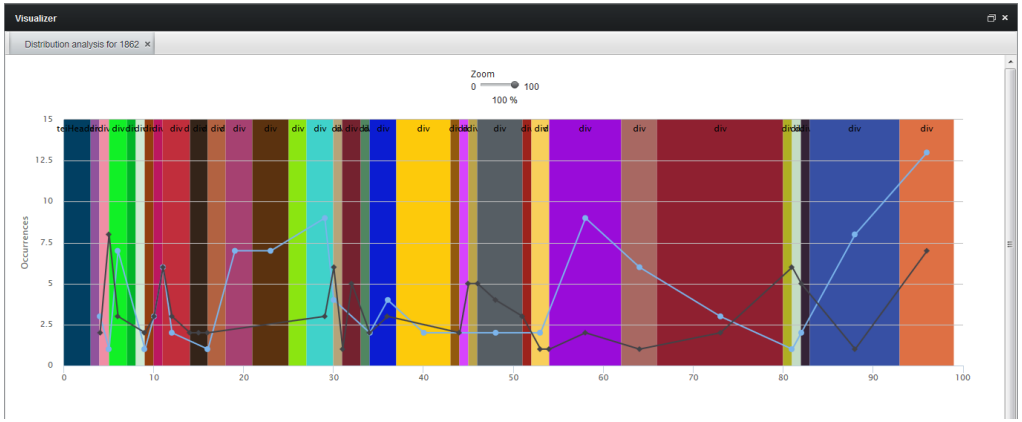
- dimension subsetting
- dimension embedding
- multiple displays
- dimension reduction

Ward et al. (2010)

Multivariate Textdaten



Erzählen über Konflikte, Gius (2013), multiple displays



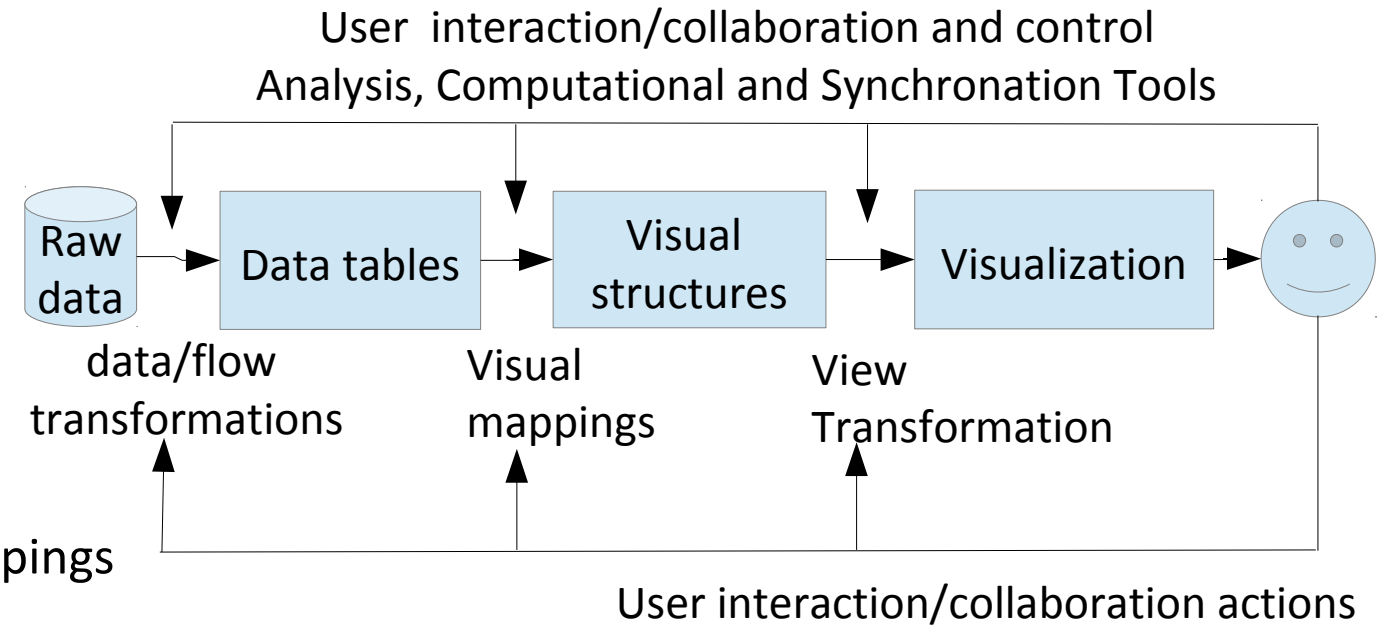
1862 Journal von Eliza Baylies Chapin Wheaton, TAPAS Project

dimension embedding

Von den Daten zur Visualisierung

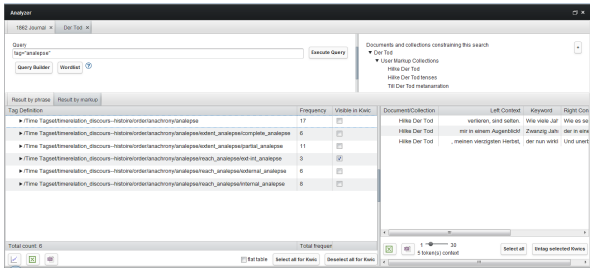
Visualization Pipeline

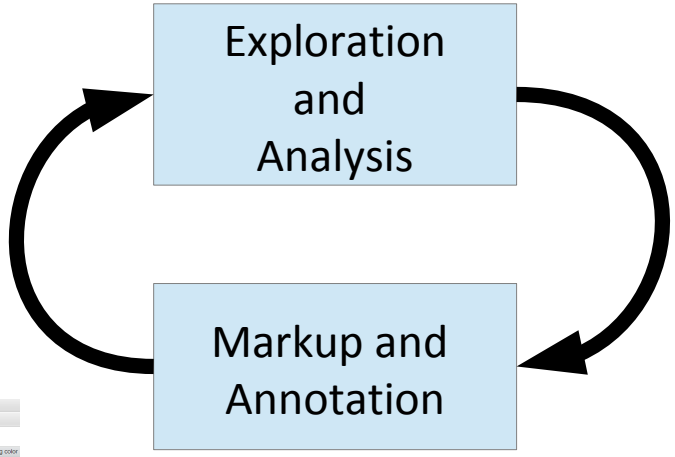
- Data modeling
- Data selection
- Data to visual mappings
- View transformation
- Generation of visualization

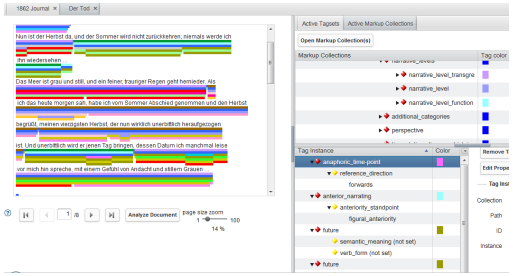


Ward et al. (2010)

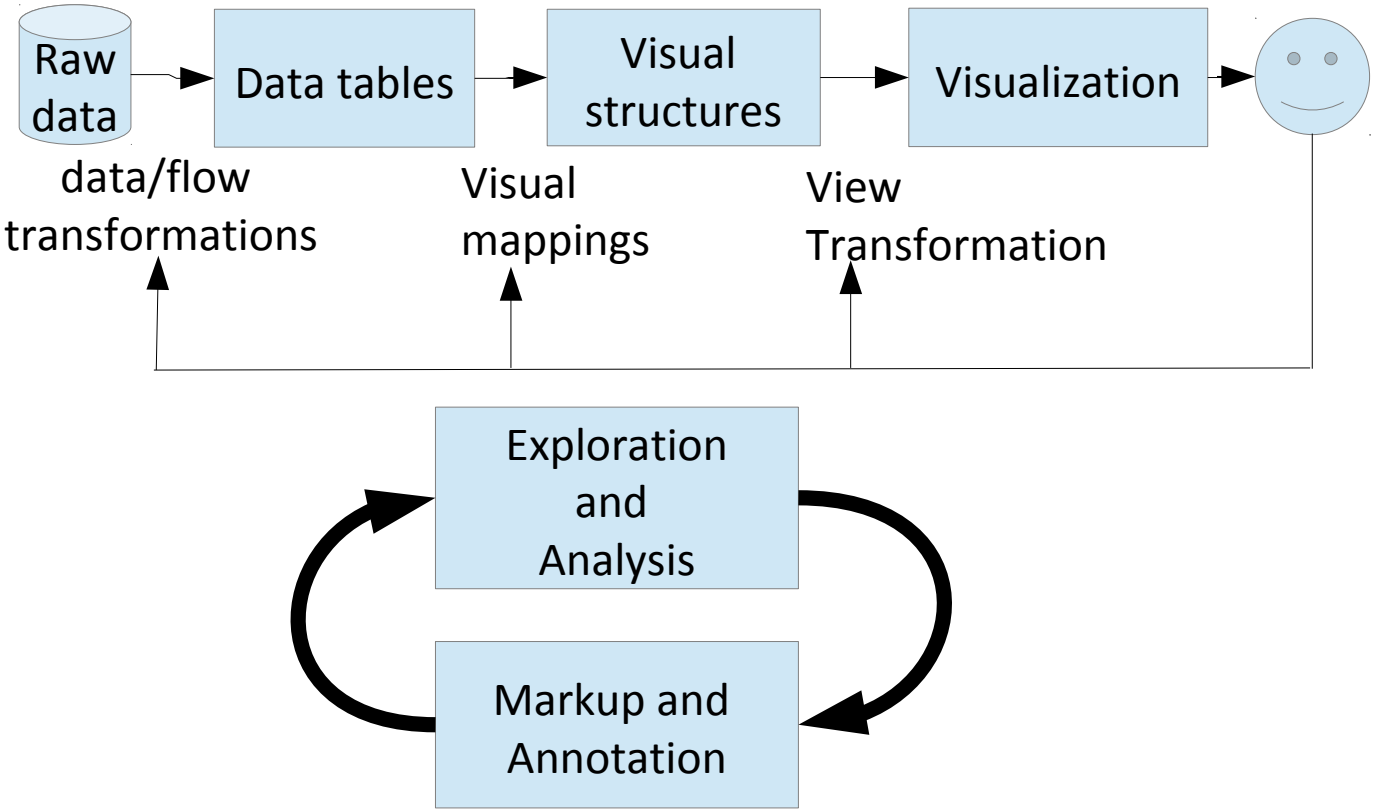
Von den Daten zur Visualisierung und zurück



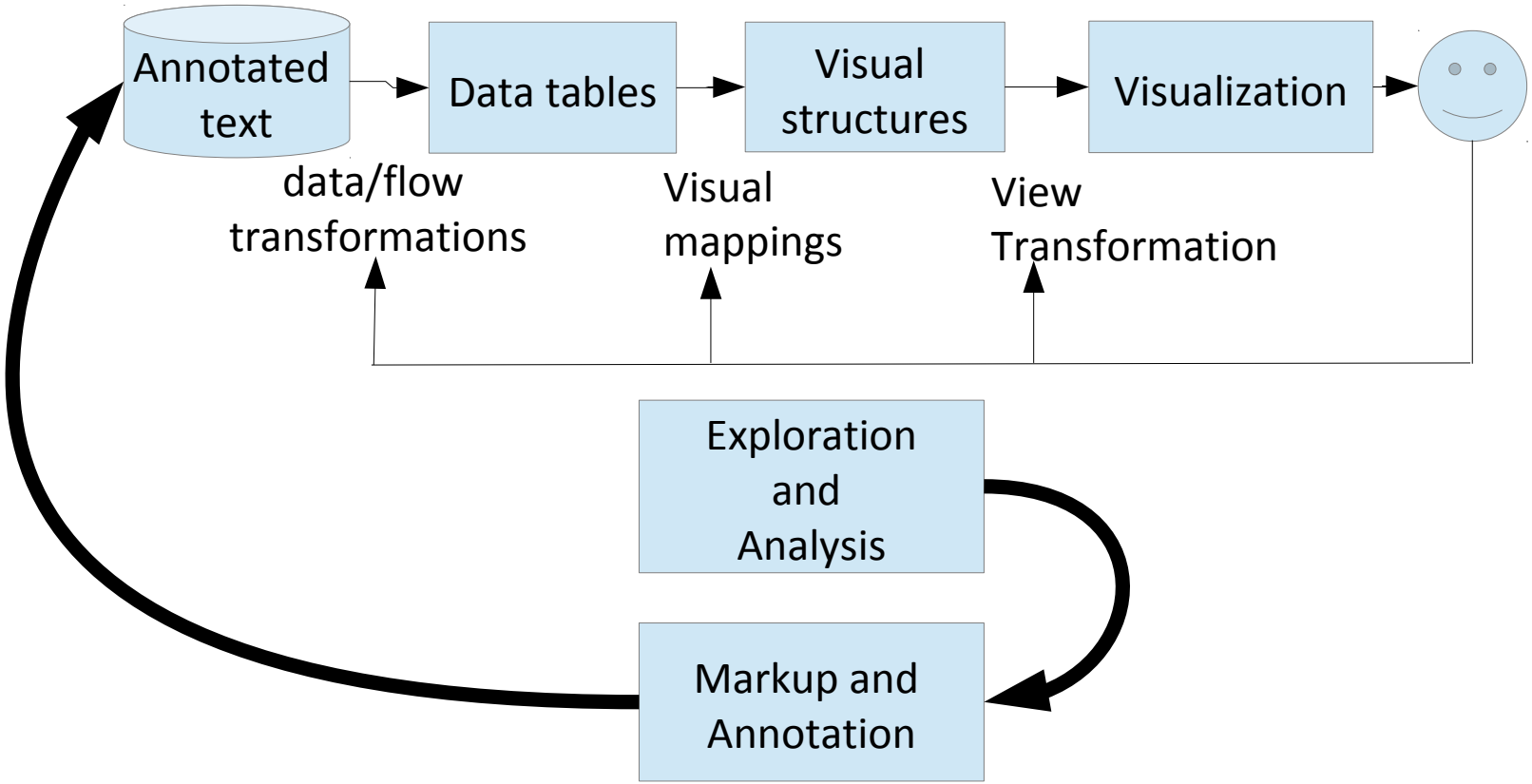




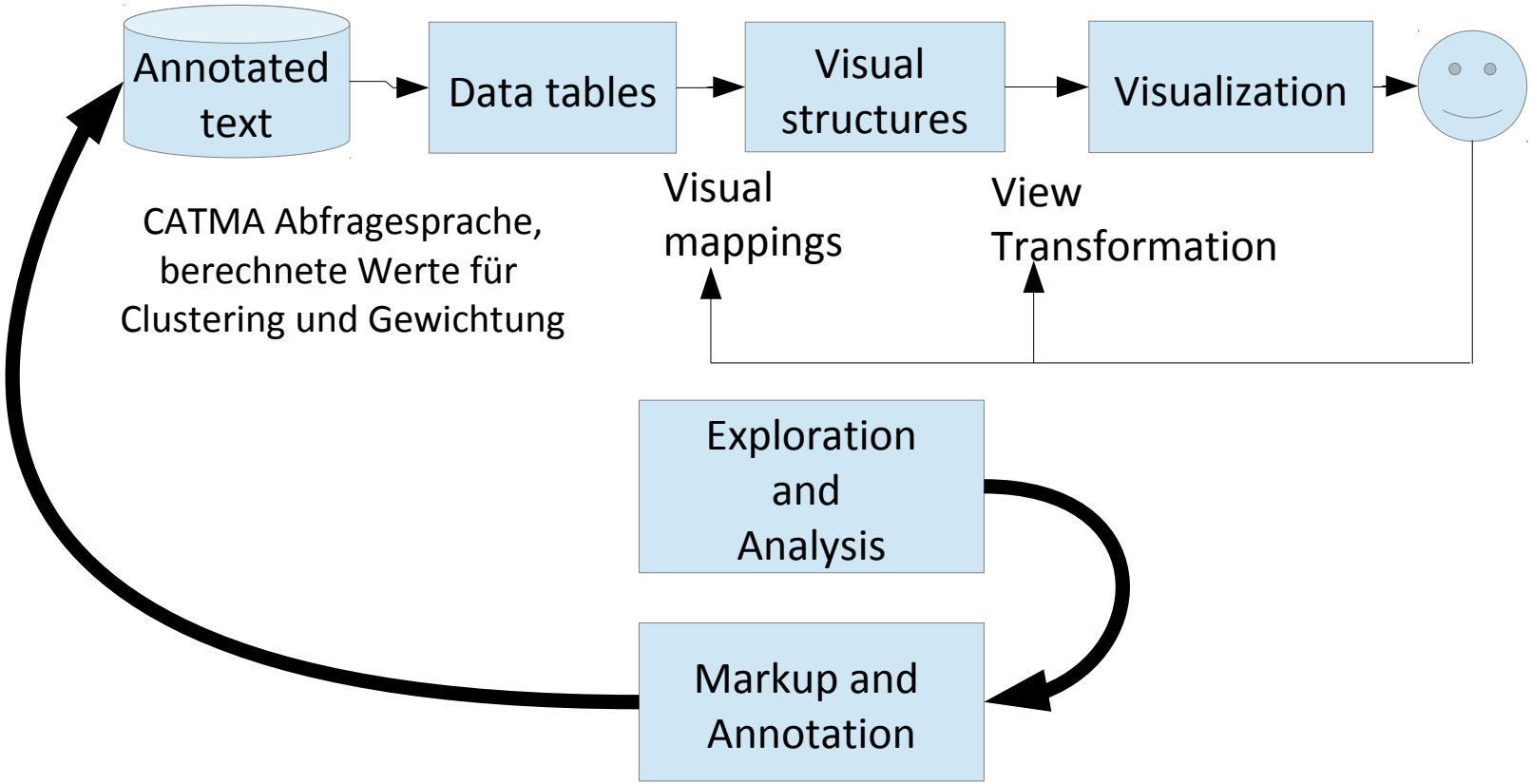
Von den Daten zur Visualisierung und zurück



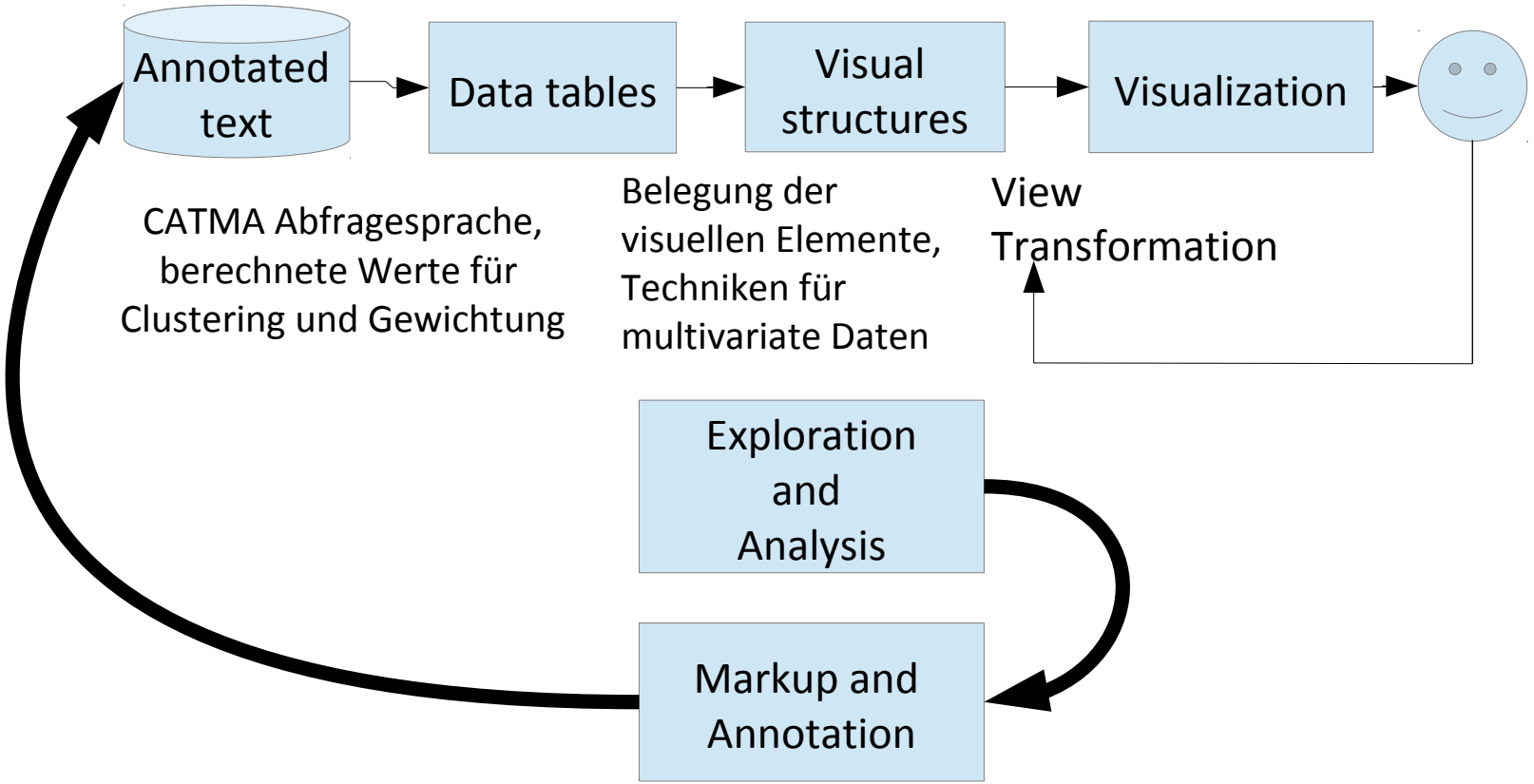
Von den Daten zur Visualisierung und zurück



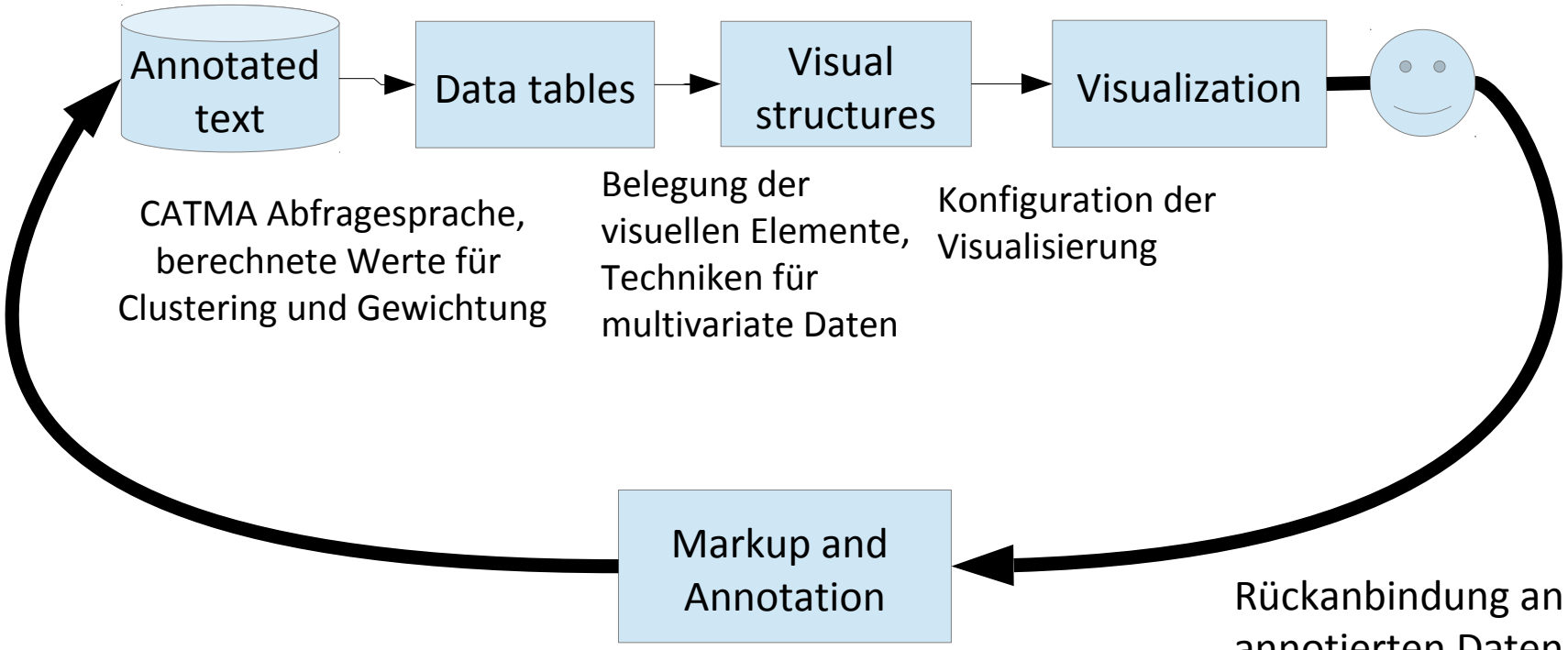
Von den Daten zur Visualisierung und zurück



Von den Daten zur Visualisierung und zurück



Von den Daten zur Visualisierung und zurück

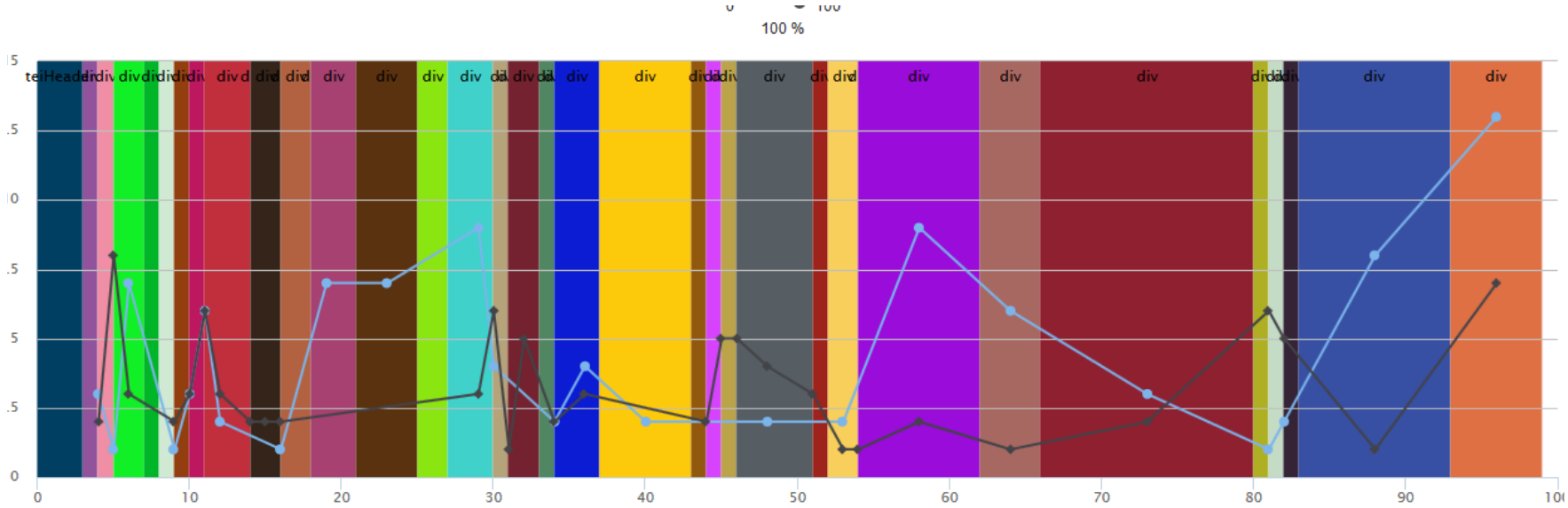


Datenauswahl

zur Verfügung gestellte Variablen

- Metadaten der Dokumente (Titel, Autor, Jahr, Dokumentgröße …)
- Textsequenzen
- Tag bzw. Typ einer Annotation,
- Properties der Annotation und die für den annotierten Text vergebenen Werte,
- Textsequenz einer annotierten Fundstelle,
- relative/absolute Größe des annotierten Textes,
- Vorkommenshäufigkeit der Annotation,
- Vorkommenshäufigkeit bestimmter Propertywerte,
- Annotationskontext der Annotation
- Position im Text (via Zeichen oder TokenOffset),
- Textkontext des annotierten Textes (variable Anzahl von Token),
- Vorkommenshäufigkeit des annotierten Textes,
- weitere berechnete Kategorien, wie der zFaktor oder der TF-IDF
- Vorkommenshäufigkeit pro Klasseneinteilung (chunks, Annotationen)

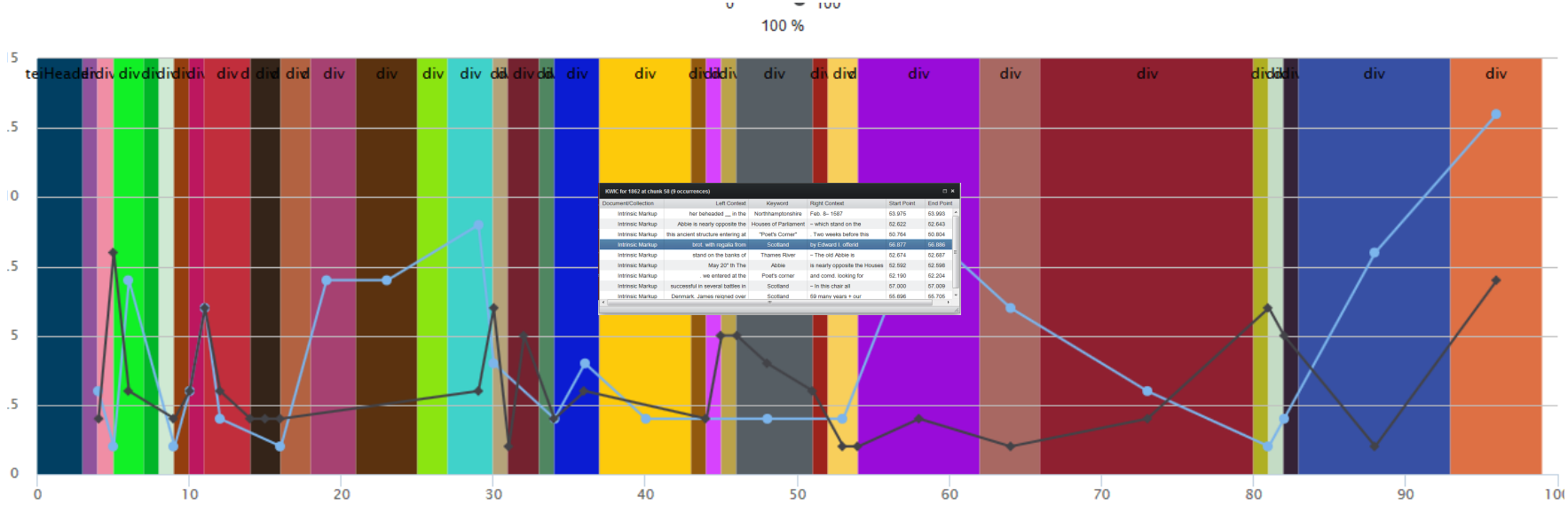
Beispiel Datenauswahl und -transformation



1862 Journal von Eliza Baylies Chapin Wheaton, TAPAS Project

Ziel: Verteilung von Orts- und Personennamen pro Tagebucheintrag mit Darstellung der Eintraggröße in Bezug auf die Tagebuchlänge

Beispiel Datenauswahl und -transformation



1862 Journal von Eliza Baylies Chapin Wheaton, TAPAS Project

Ziel: Verteilung von Orts- und Personennamen pro Tagebucheintrag mit Darstellung der Eintraggröße in Bezug auf die Tagebuchlänge

Beispiel Datenauswahl und -transformation

- Abfrage 1 - Tagebucheinträge
- tag = "div" property = "type" value="entry", tag = "teiHeader"
- tagname, range, documentsize
- range mit documentsize auf Prozent skalieren
- Plotband: size: Prozent range, name: tagname

Beispiel Datenauswahl und -transformation

- Abfrage 2 - Ortsnamen
- tag = "name" property="type" value="place" where tag="text" boundary
- tagname, range, documentsize
- Clusterbildung via range-Inklusion: Zählen pro Tagebucheintrag-range
- Positionierung mittig zur Tagebucheintrag-range
- Tagebucheintrag-range/2 mit documentsize auf Prozent skalieren
- line: y-Werte: geclusterte Vorkommnisse x-Werte: Prozentwert
- gleiches Verfahren für Personennamen
- Interface?

Auswahl der Visualisierung

- Visualisierung beeinflusst Datentransformation

- Heuristisches Werkzeug?
- Best Practices

Vielen Dank für Ihre Aufmerksamkeit!

http://www.catma.de

http://www.heureclea.de