



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG



# Die explorative Visualisierung von Texten

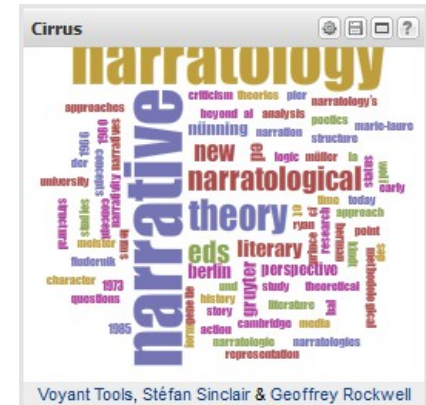
Evelyn Gius und Marco Petris

25.02.2015 DHd2015, Graz



# Probleme

- häufig reine Textdaten
- Blackbox
- langer Weg zur Visualisierung





# Hermeneutische Ebenen

The screenshot shows a text analysis interface with two main panels. The left panel displays a document titled "1862 Journal" and "Der Tod" with several lines of text. Each line is overlaid with multiple horizontal bars of different colors (blue, red, green, yellow, purple), representing different levels of annotation. The text includes:

Nun ist der Herbst da, und der Sommer wird nicht zurückkehren; niemals werde ich ihn wiedersehen ...

Das Meer ist grau und still, und ein feiner, trauriger Regen geht hernieder. Als ich das heute morgen sah, habe ich vom Sommer Abschied genommen und den Herbst begrüßt, meinen vierzigsten Herbst, der nun wirklich unerbittlich heraufgezogen ist. Und unerbittlich wird er jenen Tag bringen, dessen Datum ich manchmal leise vor mich hin spreche, mit einem Gefühl von Andacht und stillem Grauen ...

The right panel shows the "Active Tagsets" and "Active Markup Collections" for the document. It lists various markup collections and their tag colors:

- narrative\_levels (blue)
- narrative\_level\_transgre (purple)
- narrative\_level (light blue)
- narrative\_level\_function (cyan)
- additional\_categories (dark blue)
- perspective (blue)

Below this, the "Tag Instance" panel shows a detailed view of a specific tag instance:

- anaphoric\_time-point (pink)
- reference\_direction (yellow)
  - forwards
- anterior\_narrating (cyan)
  - anteriority\_standpoint (yellow)
    - figural\_anteriority
- future (green)
  - semantic\_meaning (not set)
  - verb\_form (not set)

At the bottom of the interface, there are navigation controls (back, forward, search) and a zoom slider set to 14%.



## 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



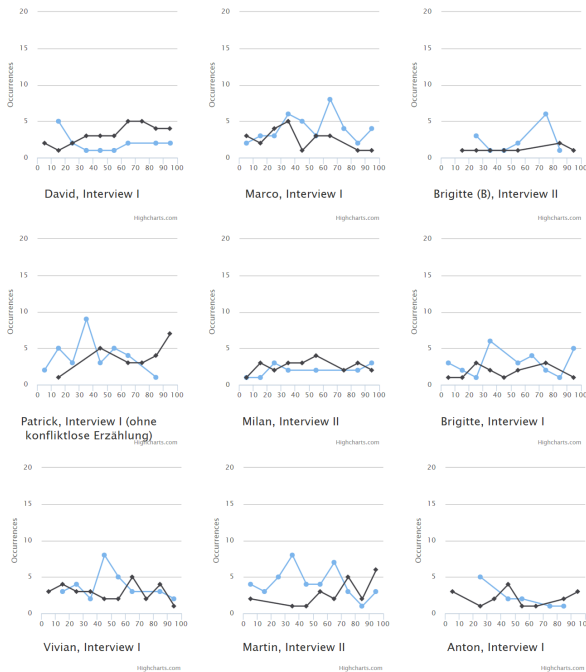
# Visualisierung multivariater Daten

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

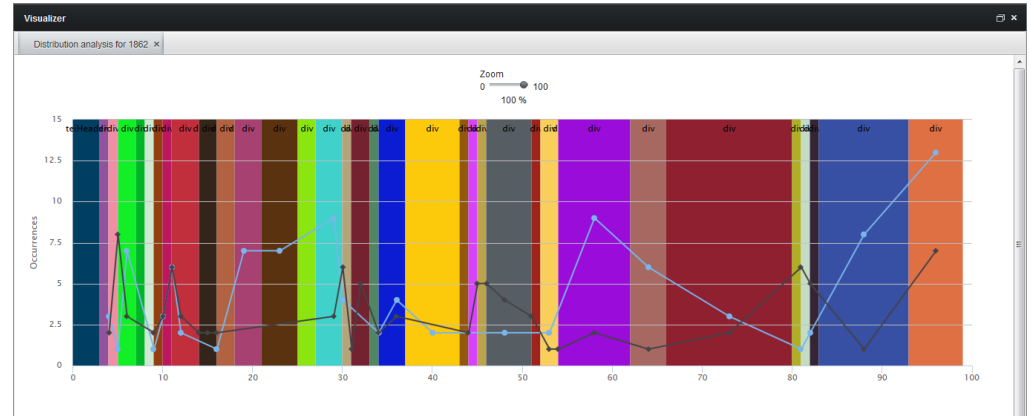
Ward et al. (2010)



# Multivariate Textdaten



blue - speech\_representation  
black - mental\_process\_representation



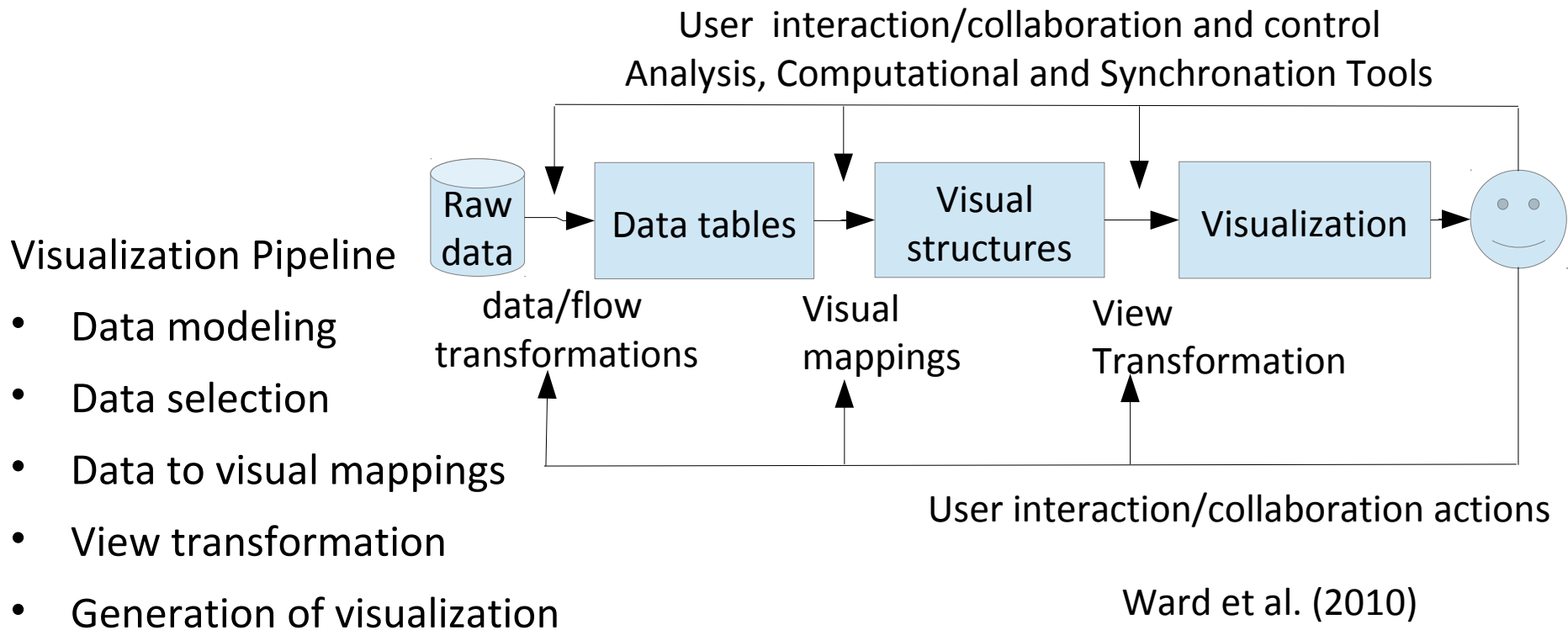
1862 Journal von Eliza Baylies Chapin Wheaton,  
TAPAS Project

dimension embedding

Erzählen über Konflikte, Gius (2013),  
multiple displays



# Von den Daten zur Visualisierung

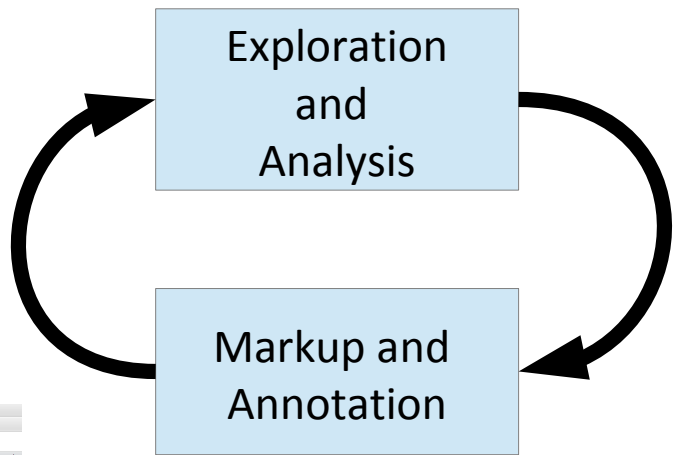




# Von den Daten zur Visualisierung und zurück

The screenshot shows the 'Analysator' interface with a search query 'tag="narrative"'. The results table is as follows:

Tag Definition	Frequency	Visible in View	DocumentCollection	Left Context	Keyword	Right Con
!Time Taggestimmteinstellung_discours-InformationenAnschonAnstapfe	17	<input type="checkbox"/>	Hiba Der Tod	verloren, sind aufen	Wie was ist	Wie es si
!Time Taggestimmteinstellung_discours-InformationenAnschonAnstapfeAnstapfeAnstapfe	6	<input type="checkbox"/>	Hiba Der Tod	ist in einem Angebots	Zwilling, die, die in ein	
!Time Taggestimmteinstellung_discours-InformationenAnschonAnstapfeAnstapfeAnstapfeAnstapfe	11	<input type="checkbox"/>	Hiba Der Tod	meinen einzigen verlost,	der das wert	Und unet
!Time Taggestimmteinstellung_discours-InformationenAnschonAnstapfeAnstapfeAnstapfeAnstapfe	3	<input checked="" type="checkbox"/>				
!Time Taggestimmteinstellung_discours-InformationenAnschonAnstapfeAnstapfeAnstapfeAnstapfe	6	<input type="checkbox"/>				
!Time Taggestimmteinstellung_discours-InformationenAnschonAnstapfeAnstapfeAnstapfeAnstapfe	6	<input type="checkbox"/>				



The screenshot shows a document viewer with text and colored annotations. The sidebar on the right lists 'Active Tags' and 'Active Markup Collections'. The 'Active Markup Collections' section includes:

- !narrative\_level
- !narrative\_level\_function
- !narrative\_level
- !additional\_categories
- !perspective

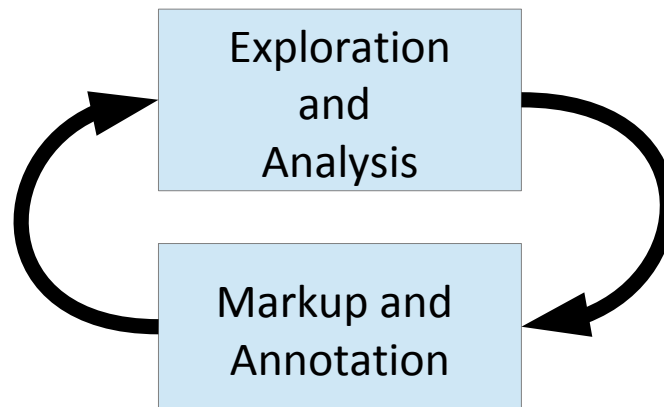
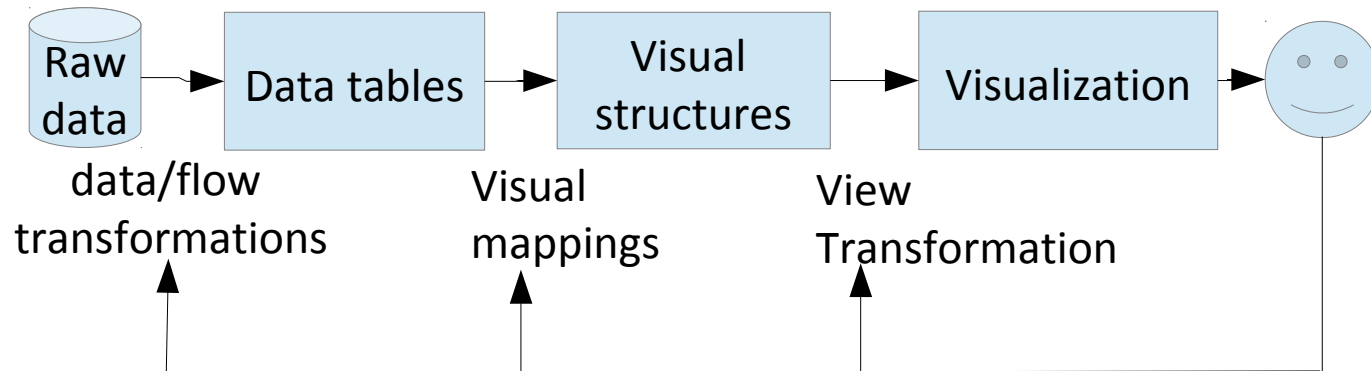
The 'Tag Instance' section shows:

- !narrative\_level
- !narrative\_level\_function
- !narrative\_level
- !perspective



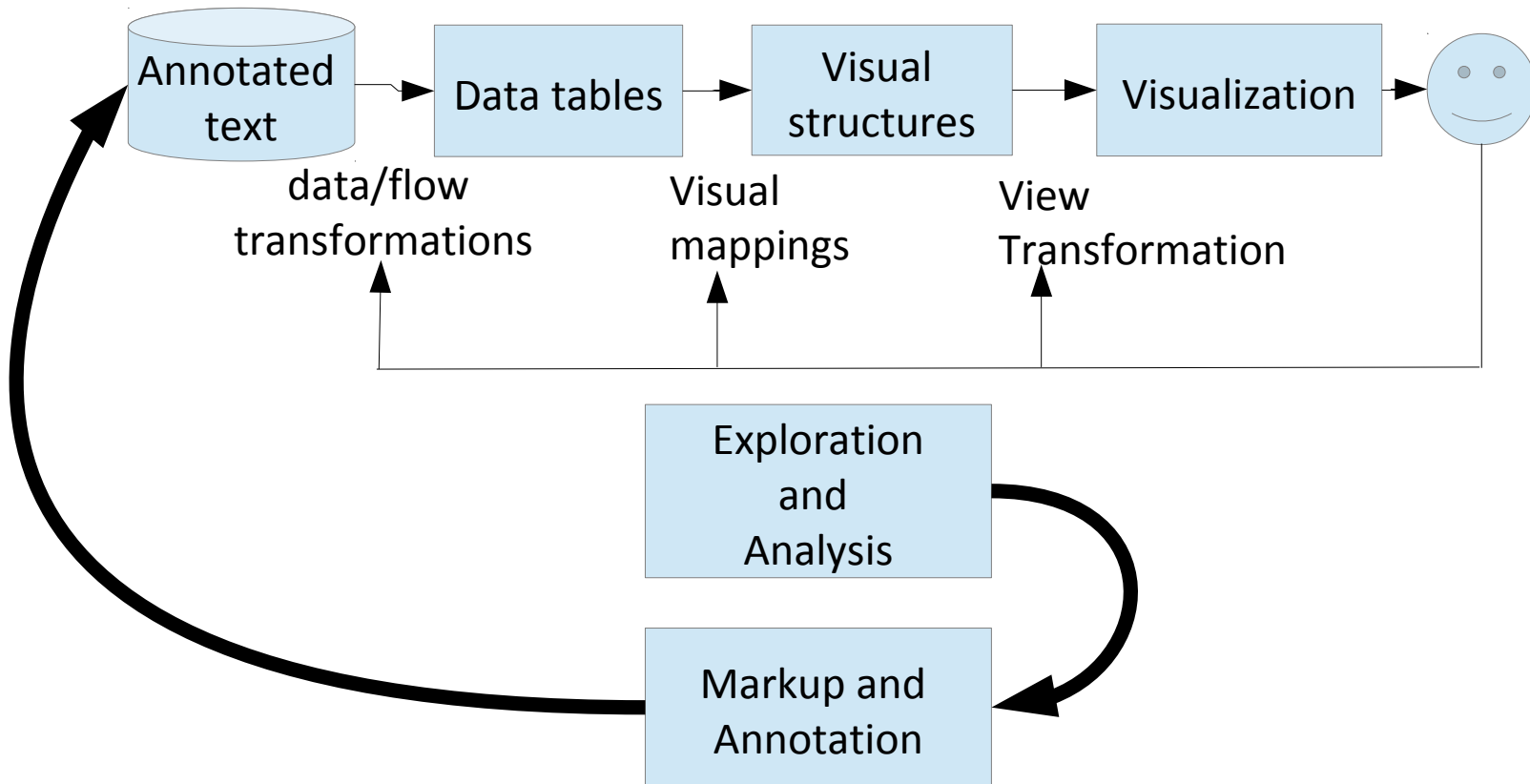


# Von den Daten zur Visualisierung und zurück



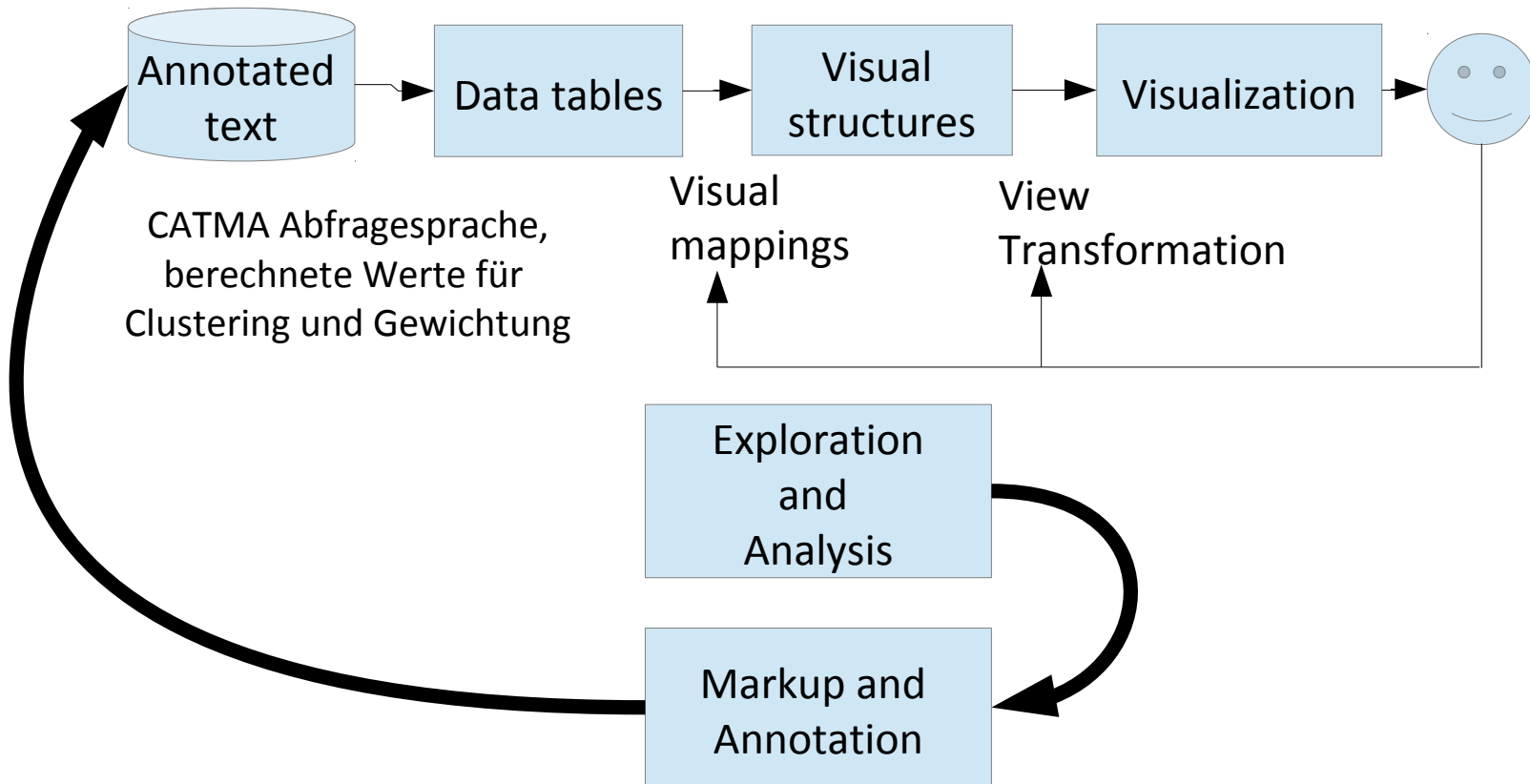


# Von den Daten zur Visualisierung und zurück



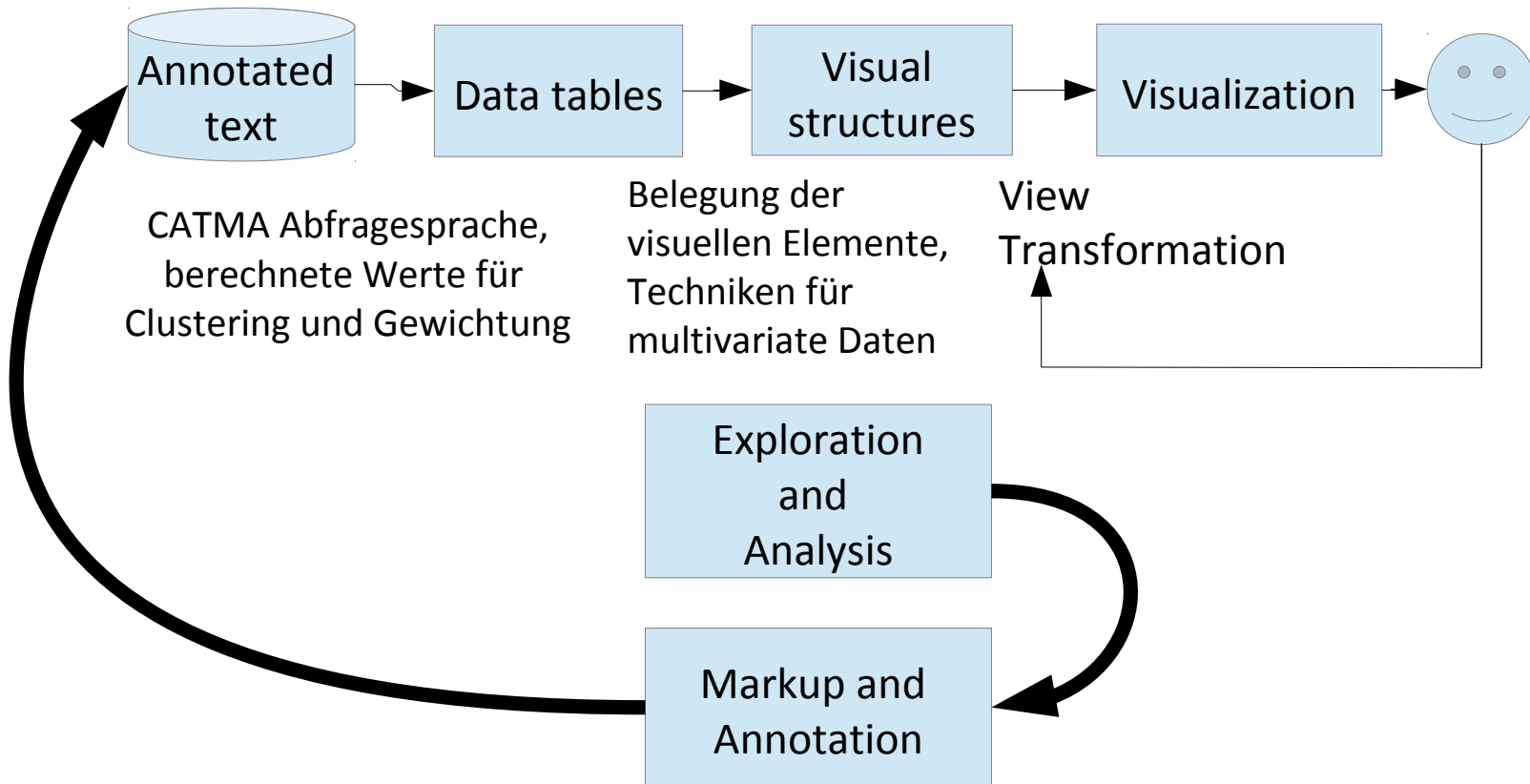


# Von den Daten zur Visualisierung und zurück



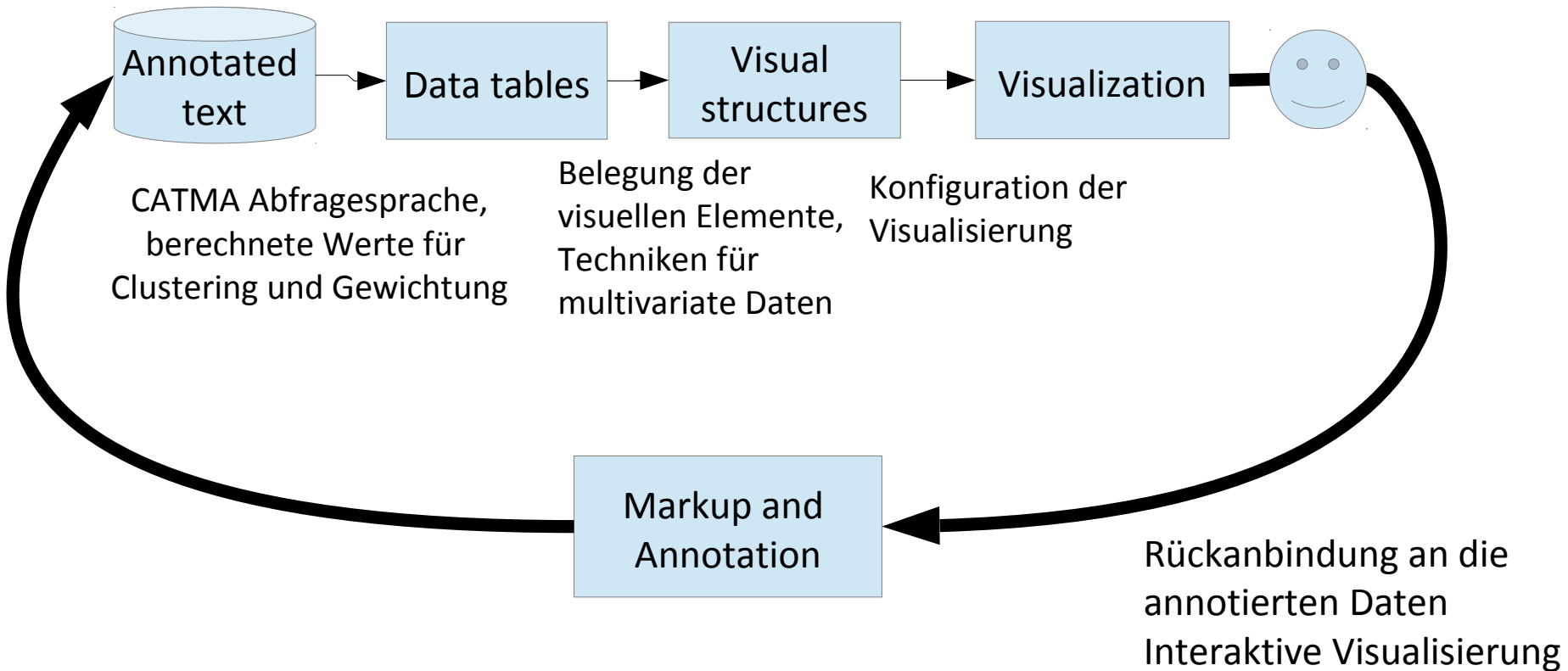


# Von den Daten zur Visualisierung und zurück





## Von den Daten zur Visualisierung und zurück





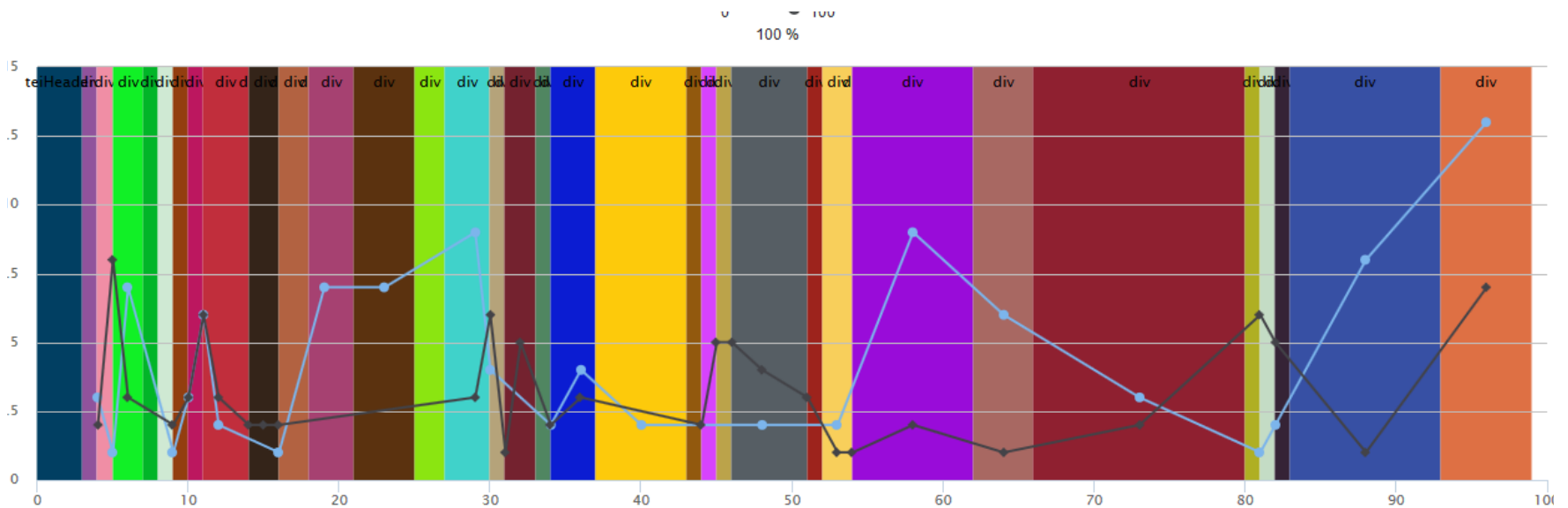
# Datenauswahl

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

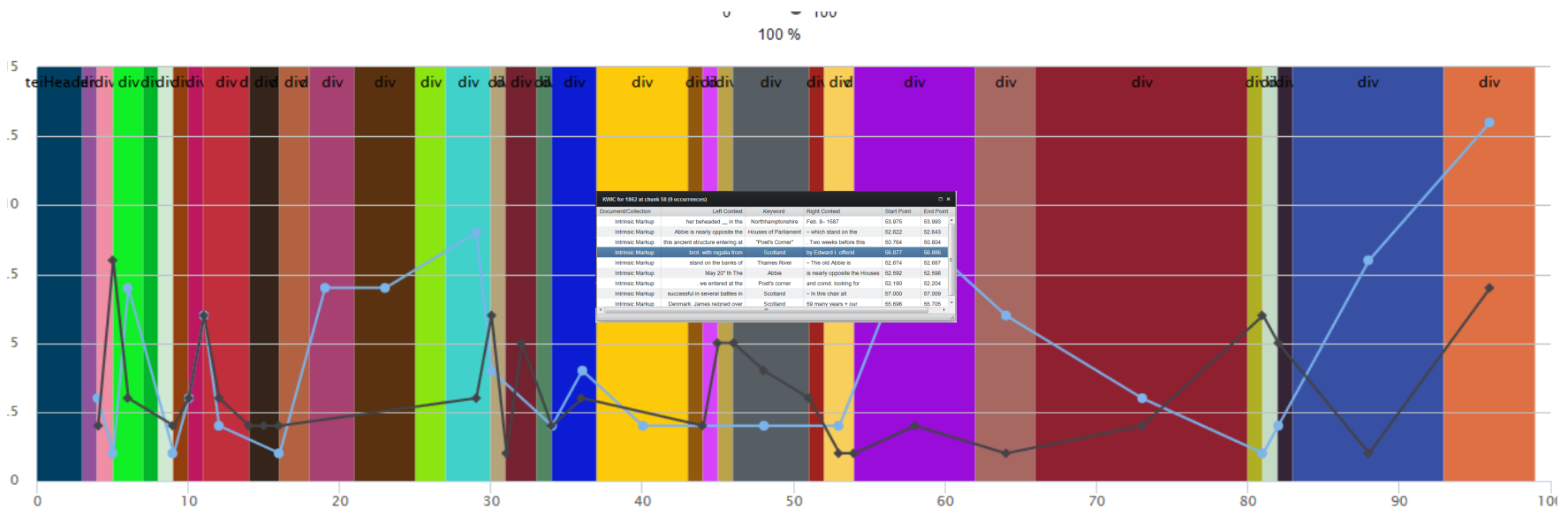


1862 Journal von Eliza Baylies Chapin Wheaton, TAPAS Project

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



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG



Vielen Dank für Ihre Aufmerksamkeit!

<http://www.catma.de>

<http://www.heureclea.de>