Datenmodellierung 1: Modellierung

Vorlesung *Einführung in die Digital Humanities*
MSc Digital Humanities | Wintersemester 2020/21

Prof. Dr. Christof Schöch

Universität Trier

Sitzungsüberblick

1. Daten
2. Modell
3. Datenmodellierung
4. Wozu eigentlich Datenmodellierung?

1. Daten

Definition von ”Daten”

Data is a set of values of qualitative or quantitative variables.
("Data", Wikipedia)

Definition von ”Daten”

Data is "the absence of uniformity".
(Luciano Floridi 2010)

Definition von ”Daten”

Data in the humanities could be considered a digital, selectively constructed, machine-actionable abstraction representing some aspects of a given object of humanistic inquiry.
(Schöch 2013)

Digitale Daten

- diskret (statt kontinuierlich)
- binär repräsentiert (0 und 1)
- maschinenlesbar
- vervielfältigbar
- "Information wants to be free" (Stewart Brand, 1984)

Strukturierte Daten?

- strukturiert: bspw. Relationale Datenbanken
- semi-strukturiert: bspw. XML-Dateien, JSON-Dateien
- unstrukturiert: bspw. plain text oder Pixelgrafiken

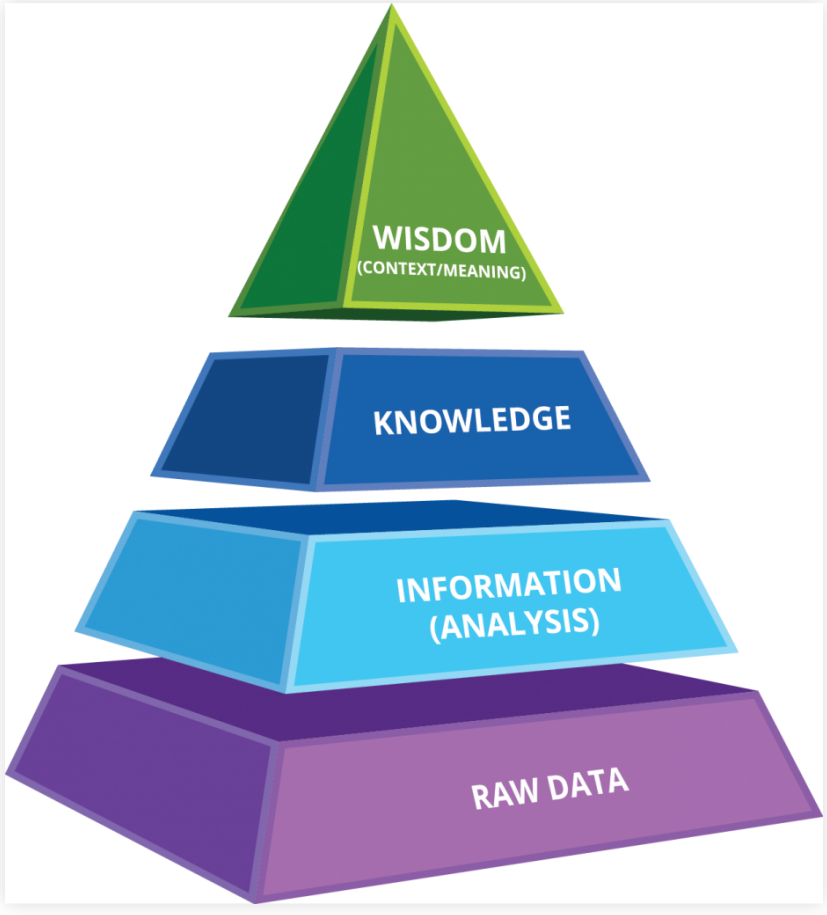
Datenstrukturen

- linear: bspw. Listen, Tabellen
- hierarchisch: bspw. XML oder JSON
- multi-relational: bspw. Netzwerke (Graphen)

Verwandte Begriffe

- Gegenstand
- Datensatz
- Datensammlung
- Datenbank
- Korpus
- Digitale Edition
- Information
- Wissen

DIKW-Pyramide



(Quelle: O.A, DIKW Pyramid, 2017, https://www.kisspng.com/png-dikw-pyramid-business-intelligence-knowledge-organ-2664729/)

DIKW: Beispiel Wahldaten 2020

Arizona (EV: 11)
Total Votes: Biden leads with 1,532,062 votes, Trump trails with 1,485,010 votes.

Timestamp	In The Lead	Vote Differential	Votes Remaining (est.)	Change	Block Breakdown	Block Trend	Hurdle
4 hours ago	Biden	47,052	309,034	5,713	Biden 53.4% / 46.6% Trump	Trump is averaging 54.1%	Trump needs 56.90% [0.198%]
6 hours ago	Biden	46,667	314,747	1,180	Biden 67.4% / 32.6% Trump	Trump is averaging 56.1%	Trump needs 56.71% [0.093%]
8 hours ago	Biden	46,257	315,927	75,314	Biden 43.0% / 57.0% Trump	Trump is averaging 57.0%	Trump needs 56.61% [0.053%]

Georgia (EV: 16)
Total Votes: Biden leads with 2,449,371 votes, Trump trails with 2,448,454 votes.

Timestamp	In The Lead	Vote Differential	Votes Remaining (est.)	Change	Block Breakdown	Block Trend	Hurdle
6 minutes ago	Biden	917	5,509	1,843	Biden 62.3% / 37.7% Trump	Trump is averaging 65.6%	Trump needs 57.70% [5.174%]
an hour ago	Trump	463	7,352	306	Trump 17.0% / 83.0% Biden	Biden is averaging 67.9%	Biden needs 52.52% [-1.193%]
2 hours ago	Trump	665	7,658	816	Trump 13.1% / 86.9% Biden	Biden is averaging 67.9%	Biden needs 53.71% [-3.134%]

North Carolina (EV: 15)
Total Votes: Trump leads with 2,732,120 votes, Biden trails with 2,655,383 votes.

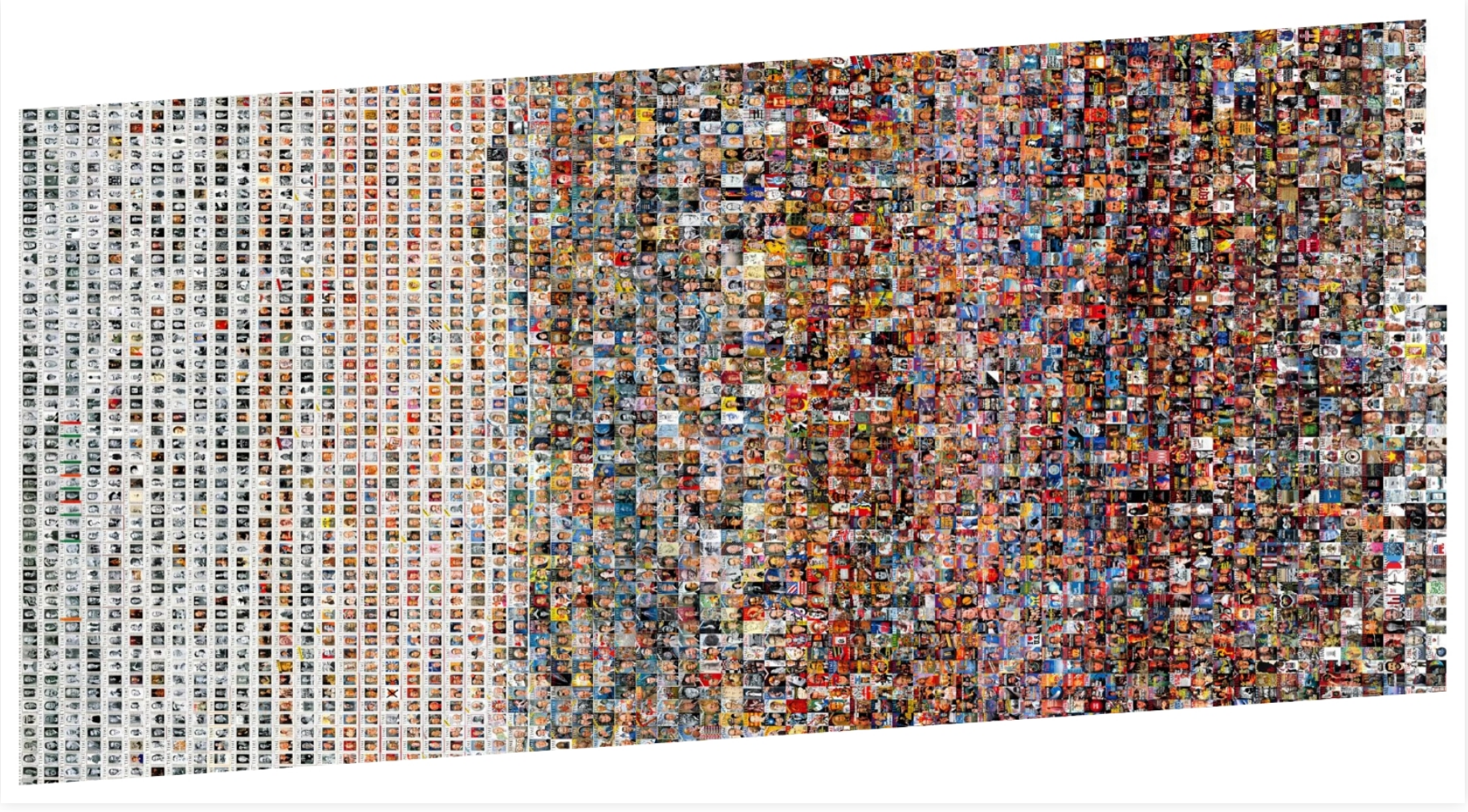
Timestamp	In The Lead	Vote Differential	Votes Remaining (est.)	Change	Block Breakdown	Block Trend	Hurdle
16 hours ago	Trump	76,737	190,621	2,233	Trump 50.0% / 50.0% Biden	Biden is averaging 50.0%	Biden needs 69.45% [0.213%]
17 hours ago	Trump	76,737	192,854	1,989	Trump 50.0% / 50.0% Biden	Biden is averaging 50.0%	Biden needs 69.24% [0.185%]
a day ago	Trump	76,737	194,843	0	N/A	Biden is averaging 50.0%	Biden needs 69.05% [-0.087%]

Nevada (EV: 6)
Total Votes: Biden leads with 604,251 votes, Trump trails with 592,813 votes.

Timestamp	In The Lead	Vote Differential	Votes Remaining (est.)	Change	Block Breakdown	Block Trend	Hurdle
16 hours ago	Biden	11,438	146,757	15	Biden 50.0% / 50.0% Trump	Trump is averaging 43.6%	Trump needs 52.86% [0.591%]
16 hours ago	Biden	11,438	173,045	913	Biden 30.9% / 69.1% Trump	Trump is averaging 43.6%	Trump needs 52.27% [-0.069%]
16 hours ago	Biden	11,787	173,958	12,189	Biden 51.4% / 48.6% Trump	Trump is averaging 42.7%	Trump needs 52.34% [0.284%]

(Quelle: Contributors, Visualisierung, 2020.

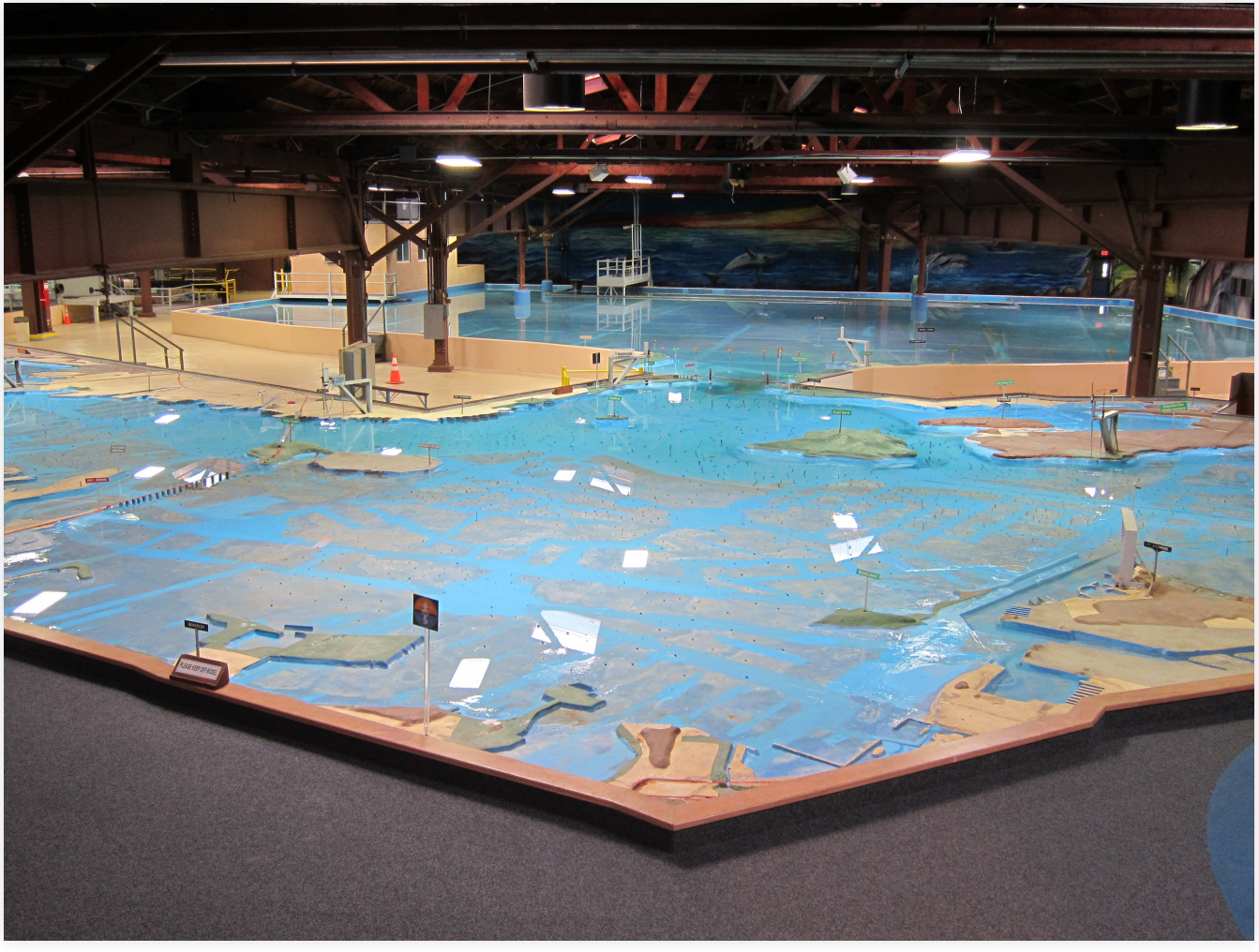
DIKW: Beispiel Time Magazine



(Quelle: Lev Manovich, Timeline Project, URL, 2010.)

2. Modell

San Francisco Bay: Simulation



(Bildquelle: Wikipedia, public domain.)

Royal Navy: Planspiel

(Bildquelle: Wikimedia Commons.)

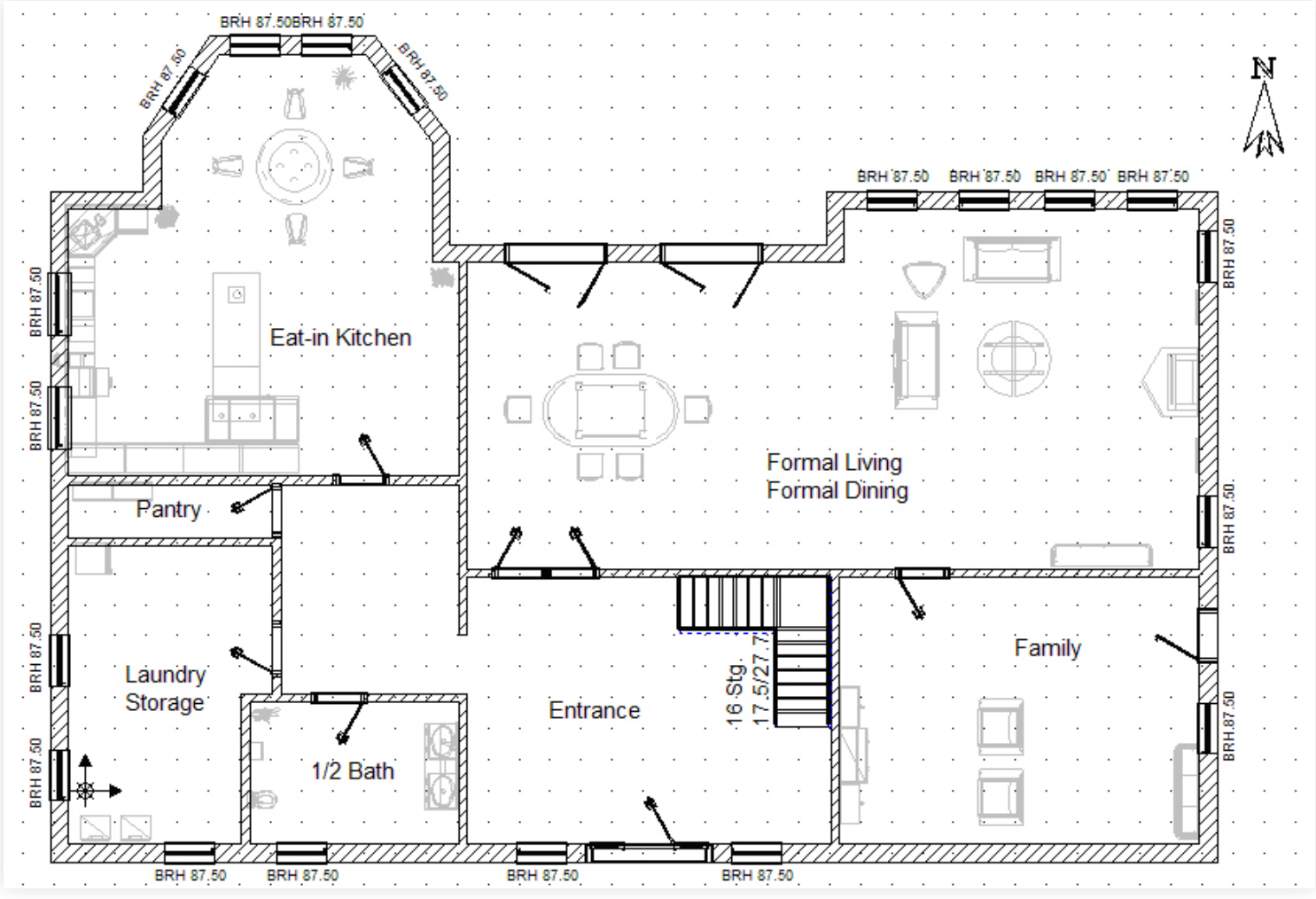
Modellflugzeug

(Quelle: KPWM Spotter at the English language Wikipedia, CC BY-SA 3.0, Link.)

Modellflugzeug

(Quelle: User Carl @FellowCreative, https://www.flickr.com/photos/fellowcreative/8056510218/, CC-BY NC 2.0.)

Bauplan



(Quelle: User Boereck, "A sample floor plan for a single-family home", https://en.wikipedia.org/wiki/Floor_plan#/media/File:Sample_Floorplan.jpg, public domain.)

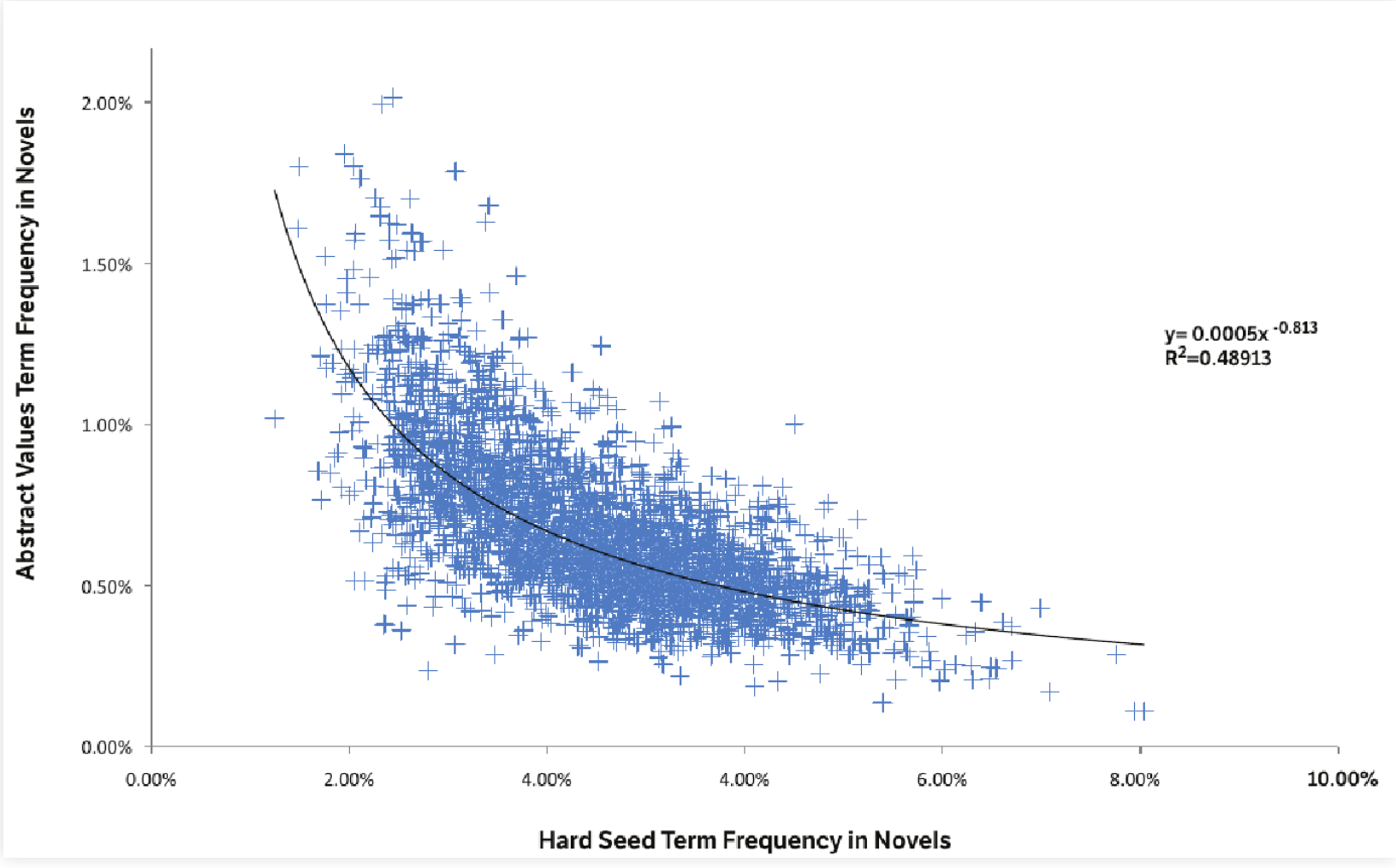
Modell: verwandte Begriffe

- Beschreibung (deskriptiv)
- Theorie (erklärend)
- Begriff (benennend)
- Abstraktion (vereinfachend)
- Auswahl (selektiv)
- Klassifikation (ordnend)
- Formalisierung (explizit)
- Simulation (dynamisch)

Modellbegriffe

- Perspektive:
  - Modell von (repräsentiert einen Gegenstand)
  - Modell für (zweckorientiert, handlungsleitend)
- Gegenstandsmodell vs. Prozessmodell
- Statistisches Modell vs. Datenmodell

Statistisches Modell



Quelle: Ryan Heuser, Long Le-Khac: „A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method." in: Literary Lab Pamphlet, 4, 2012, S. 31.)

Datenmodell: Beispiel

```
default namespace = "htttp://www.tei-c.org/ns/1.0"
start =
  element TEI {
    element text {
      element body {
        element div {
          element head { text },
          element p { text }+
          }+
        }
      }
    }
  }
s
```

Eine Formulierung eines Datenmodells
(Schema in Relax NG compact syntax)

3. Datenmodellierung

Datenmodellierung: Definition

The term "data modeling" in computer science is most typically used in a fairly restrictive sense for the modeling of relational databases, while the digital humanities has a more general understanding of the term: data modeling is the [representation] of some segment of the world in such a way to make some aspects computable.
(Flanders/Jannidis 2016)

"A data model [...] is an abstract, self-contained, logical definition of the data structures, data operators, and so forth, that together make up the abstract machine with which users interact." (Data, in Flanders/Jannidis 2016)

Drei Ebenen des Datenmodells

1. Konzeptuelles Datenmodell
   - Abstrakte Modellierung: Entitäten, Eigenschaften, Relationen
2. Logisches Datenmodell
   - Umsetzung des konzeptuellen Datenmodells (Datenbankstruktur, XML-Schema)
3. Physisches Datenmodell
   - Hardwarenahe Implementierung des logischen Datenmodells (bspw. in einem DBMS)

Drei Abstraktionslevels (Beispiel)

1. Bestimmte digitale Kodierung eines bestimmten Briefs
   - konkrete XML-Datei = (modellierte) Instanz
2. Menge von Elementen mit Eigenschaften und Relationen bei Briefen
   - konkretes Schema = (logisches) Datenmodell
3. Eine formale Sprache, mit der die Elemente und Relationen definiert werden können
   - bspw. Relax NG oder XML = (allgemeine) Metasprache / Metamodell

4. Wozu eigentlich Datenmodellieren?

”The residue of uniqueness”

Willard McCarty

- Autor des Buchs *Humanities Computing*, 2005
- Moderiert seit 1987 die Humanist-List
- Hat 2016 den Busa-Award der ADHO erhalten

(Bildquelle: http://www.mccarty.org.uk/IMG_8042.jpg)

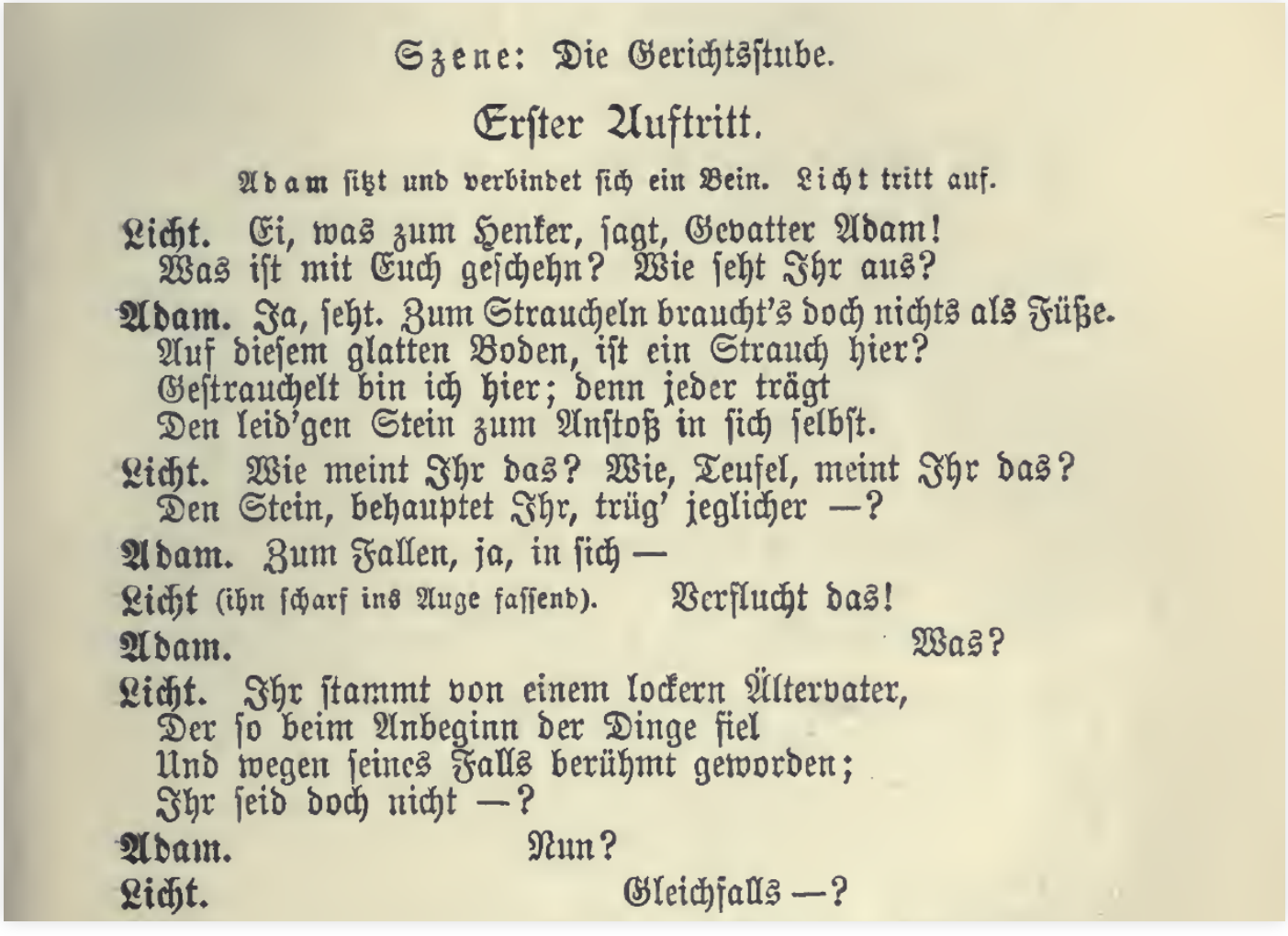
Datenmodellierung nach McCarty...

- ... als zentrale Tätigkeit der Digital Humanities
- ... als Prozess, nicht als Ergebnis
- ... als iterativer Prozess
- ... als erkenntnisfördernder Prozess
- ... als leitende Grundlage für folgende Analysen

Beipiel: Theaterstücke

- als Bildscan eines Buchs (mit Transkription): PDF
- als TEI-kodierte Textdatei: XML
- als Term-Dokument-Matrix: CSV
- Als "Zwischenformat": XML
- als Adjazenz-Matrix: CSV
- ... und davon abgeleitete Analysen

Modell: Bildscan



Szene: Die Gerichtsstube.

Erster Auftritt.

Adam sitzt und verbindet sich ein Bein. Licht tritt auf.

Licht. Ei, was zum Henker, sagt, Gevatter Adam!
Was ist mit Euch geschehn? Wie seht Ihr aus?

Adam. Ja, seht. Zum Straucheln braucht's doch nichts als Füße.
Auf diesem glatten Boden, ist ein Strauch hier?
Gestrauchelt bin ich hier; denn jeder trägt
Den leid'gen Stein zum Anstoß in sich selbst.

Licht. Wie meint Ihr das? Wie, Teufel, meint Ihr das?
Den Stein, behauptet Ihr, trüg' jeglicher —?

Adam. Zum Fallen, ja, in sich —

Licht (ihn scharf ins Auge fassend). Verflucht das!

Adam. Was?

Licht. Ihr stammt von einem lockern Ältervater,
Der so beim Anbeginn der Dinge fiel
Und wegen seines Falls berühmt geworden;
Ihr seid doch nicht —?

Adam. Nun?

Licht. Gleichfalls —?

Modell: XML-Datei (nach TEI)

```
   </div>
</div>
<div subtype="work:no" xml:id="tg134" n="/Literatur/M/Kleist, Heinrich von/Dramen/Der zerbrochene Krug/1. Auftritt">
   <div>
      <desc>
         <title>1. Auftritt</title>
      </desc>
   </div>
   <div type="text" xml:id="tg134.2">
      <div type="h4">
         <head type="h4" xml:id="tg134.2.1">Erster Auftritt</head>
         <stage rend="zenoPC" xml:id="tg134.2.4">
            <hi rend="italic" xml:id="tg134.2.4.1">Adam sitzt und verbindet sich ein Bein. Licht tritt auf.</hi>
         </stage>
         <lb xml:id="tg134.2.5"/>
         <sp>
            <speaker xml:id="tg134.2.6">LICHT.</speaker>
            <lg>
               <l rend="zenoPLm4n4" xml:id="tg134.2.7">Ei, was zum Henker, sagt, Gevatter Adam!</l>
               <l rend="zenoPLm4n4" xml:id="tg134.2.8">Was ist mit Euch geschehn? Wie seht Ihr aus?</l>
            </lg>
         </sp>
         <sp>
            <speaker xml:id="tg134.2.9">ADAM.</speaker>
            <lg>
               <l rend="zenoPLm4n4" xml:id="tg134.2.10">Ja, seht. Zum Straucheln braucht's doch nichts, als Füße.</l>
               <l rend="zenoPLm4n4" xml:id="tg134.2.11">Auf diesem glatten Boden, ist ein Strauch hier?</l>
               <l rend="zenoPLm4n4" xml:id="tg134.2.12">Gestrauchelt bin ich hier; denn jeder trägt</l>
               <l rend="zenoPLm4n4" xml:id="tg134.2.13">Den leid'gen Stein zum Anstoß in sich selbst.</l>
            </lg>
         </sp>
         <sp>
            <speaker xml:id="tg134.2.14">LICHT.</speaker>
            <l rend="zenoPLm4n4" xml:id="tg134.2.15">Nein, sagt mir, Freund! Den Stein trüg jeglicher –?</l>
         </sp>
         <sp>
            <speaker xml:id="tg134.2.16">ADAM.</speaker>
            <l rend="zenoPLm4n4" xml:id="tg134.2.17">Ja, in sich selbst!</l>
         </sp>
```

Modell: Term-Dokument-Matrix

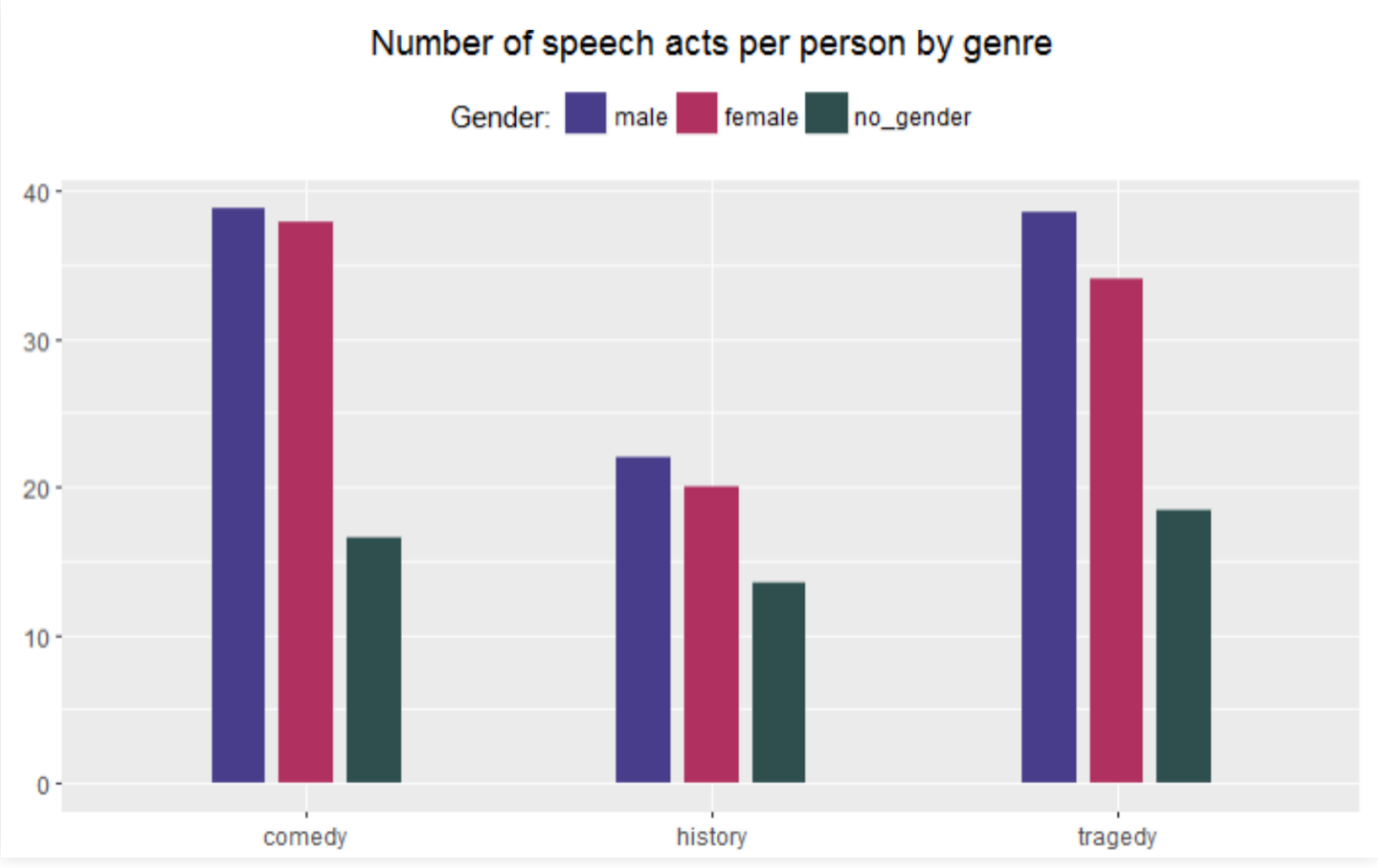
	A	B	C	D	E	F	G
1		CorneilleP_tc0189	CorneilleP_tc0196	CorneilleT_tc0222	CorneilleT_tc0226	RacineJ_tc0656	RacineJ_tc0661
2	de	3.567	3.779	3.414	2.970	3.858	3.528
3	et	2.733	2.764	1.969	2.530	1.834	2.050
4	vous	2.822	2.191	1.969	1.345	2.872	1.982
5	le	1.878	2.029	2.461	2.561	2.006	2.159
6	à	2.139	2.484	2.159	2.230	1.834	1.907
7	l	1.889	1.583	2.055	2.680	1.738	1.519
8	que	1.894	1.835	1.956	1.594	1.795	1.396
9	je	1.961	1.380	1.532	1.397	2.006	2.159
10	il	1.494	1.371	1.643	1.713	1.560	0.981
11	un	1.389	1.438	1.249	1.625	0.974	1.539
12	la	1.161	1.511	1.403	1.340	1.222	1.417
13	en	1.528	1.425	1.390	1.511	1.101	0.913
14	qu	1.422	1.362	1.538	1.506	1.108	0.817
15	les	1.261	1.632	1.120	0.771	1.305	1.757
16	d	1.339	1.366	1.181	1.459	1.019	1.287
17	est	1.139	0.960	1.489	1.247	0.949	0.960
18	ce	1.222	1.141	1.070	0.988	0.891	0.715
19	pour	1.067	0.978	1.347	1.097	0.681	0.572
20	ne	0.917	0.834	0.849	0.890	1.101	0.926
21	qui	0.772	0.960	0.904	0.797	0.751	0.647
22	n	0.950	0.739	0.818	0.719	0.643	0.708
23	si	0.806	0.852	0.806	0.787	0.579	0.531
24	m	0.750	0.613	0.892	0.792	0.694	0.647

Modell: Adjazenzmatrix

	I.1	I.2	I.3	I.4	II.1	II.2	III.1	III.2	III.3	VI.1	VI.2	VI.3	VI.4	VI.5	V.1	V.2	V.3	V.4	V.5	V.6
Iphigenie	1	1	1	1	0	1	1	0	1	1	1	1	1	1	0	0	1	1	1	1
Thoas	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
Orest	0	0	0	0	1	0	1	1	1	0	0	0	0	0	0	0	0	1	1	1
Pylades	0	0	0	0	1	1	0	0	1	0	0	0	1	0	0	0	0	0	1	0
Arkas	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0

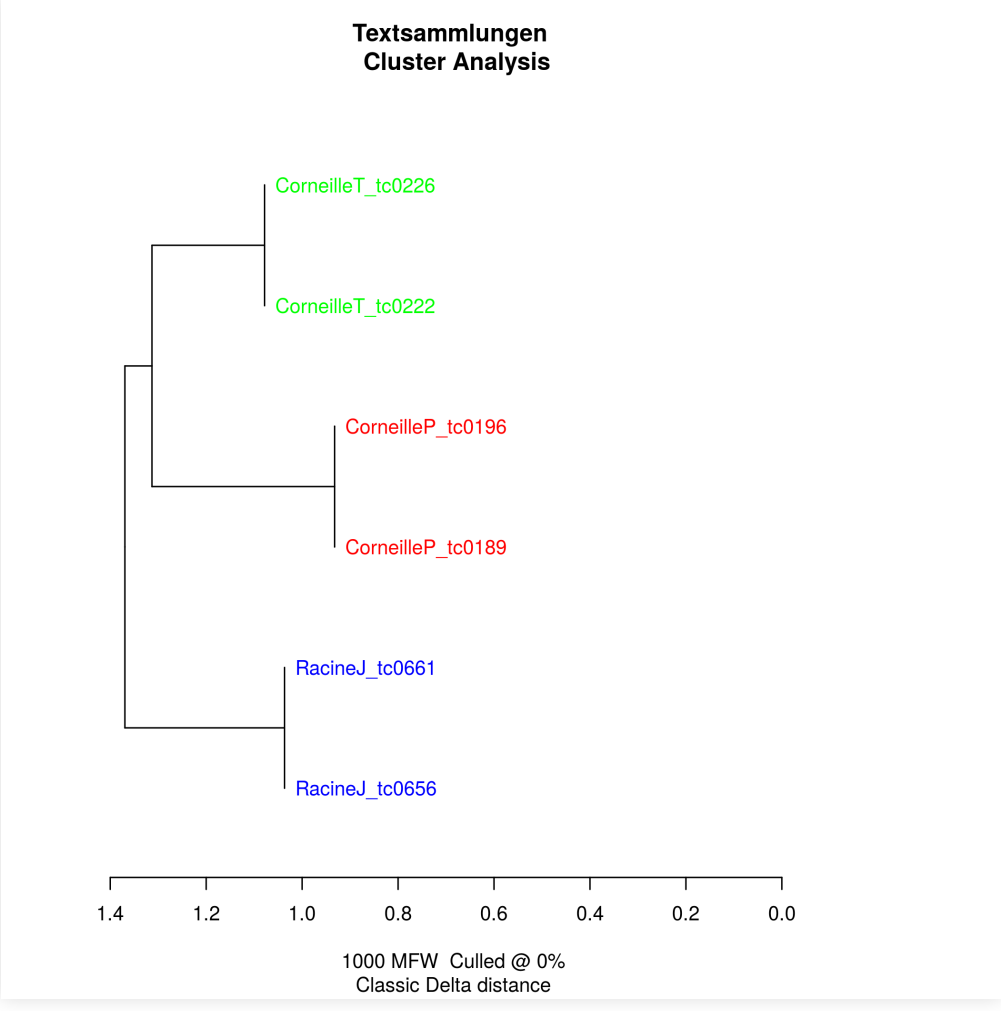
(Quelle: Trilcke, Peer, „Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft“. In: Philip Ajouri, Katja Mellmann u. Christoph Rauen (Hg.): Empirie in der Literaturwissenschaft, Münster 2013, S. 201–247, 226.)

Analyse: Replikenverteilung



Grundlage: TEI-Kodierung

Analyse: Textähnlichkeit



Grundlage: Term-Dokument-Matrix

Abschluss

Fragen?

Lektürehinweise

- Fotis Jannidis, "Grundlagen der Datenmodellierung", in: *Digital Humanities: Eine Einführung*, hg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein. Stuttgart: Metzler, 2017, S. 99-108.

Weitere Empfehlungen

- Julia Flanders und Fotis Jannidis. "Data modeling", in: *The New Companion to Digital Humanities*, ed. by Susan Schreibman, Ray Siemens and John Unsworth. Oxford: Blackwell, 2016.

Darüber hinaus

- Willard McCarty. *Humanities Computing*. 2005.
- Richard Jean So. "All Models Are Wrong". In: *PMLA* 132.3, 2017, 668-673.
- Michael Weisberg. *Simulation and Similarity. Using Models to UNderstand the World*. Oxford UP, 2013.

Nächster Termin

- 26.11.: Thema: Datenmodellierung 2: Datenbanken
  Vorbereitung: Kapitel "Datenbanken" in der Dateiablage

Christof Schöch, 2017
http://www.christof-schoech.de
