GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Digital Humanities for Computer Scientists … or: How I became infected with the Indiana Jones virus

Marco Büchler
eTRAP Research Group
Göttingen Centre for Digital Humanities
Institute of Computer Science
Georg August University Göttingen, Germany

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Who am I?

- 2001/2 Head of Quality Assurance department in a software company
- 2006 Diploma in Computer Science on big scale co-occurrence analysis
- 2007- Consultant for several SME in IT sector
- 2008 Technical project management of eAQUA project
- 2011 PI and project manager eTRACES project
- 2013 PhD in „Digital Humanities“ on Text Reuse
- 2014- Head of Early Career Research Group eTRAP at Göttingen Centre for Digital Humanities

The Indiana Jones virus?

Overview

- Big Humanities Data
- Filling gaps in inscriptions
- Uncovering unexpected relations
- Text Reuse

Big (Humanities) Data

- **3 aspects** (by Ulrike Rieß, Big Data bestimmt die IT-Welt):
  - **Huge amount of data** that can‘t be processed and analyzed manually
  - **Less structured data**; e. g. in comparison to databases and data warehouse systems
  - Linked data between **heterogeneous and distributed resources**

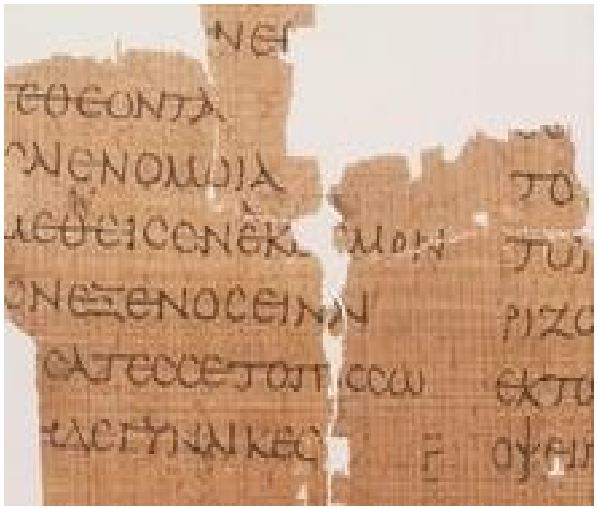
- The fastest growing sources of Big Data are text and images.

- Researchers easily get lost in the **information overload** (Big Data) and in the **information poverty** (Humanities Data).

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Filling (Guessing) gaps in inscriptions and Papyri

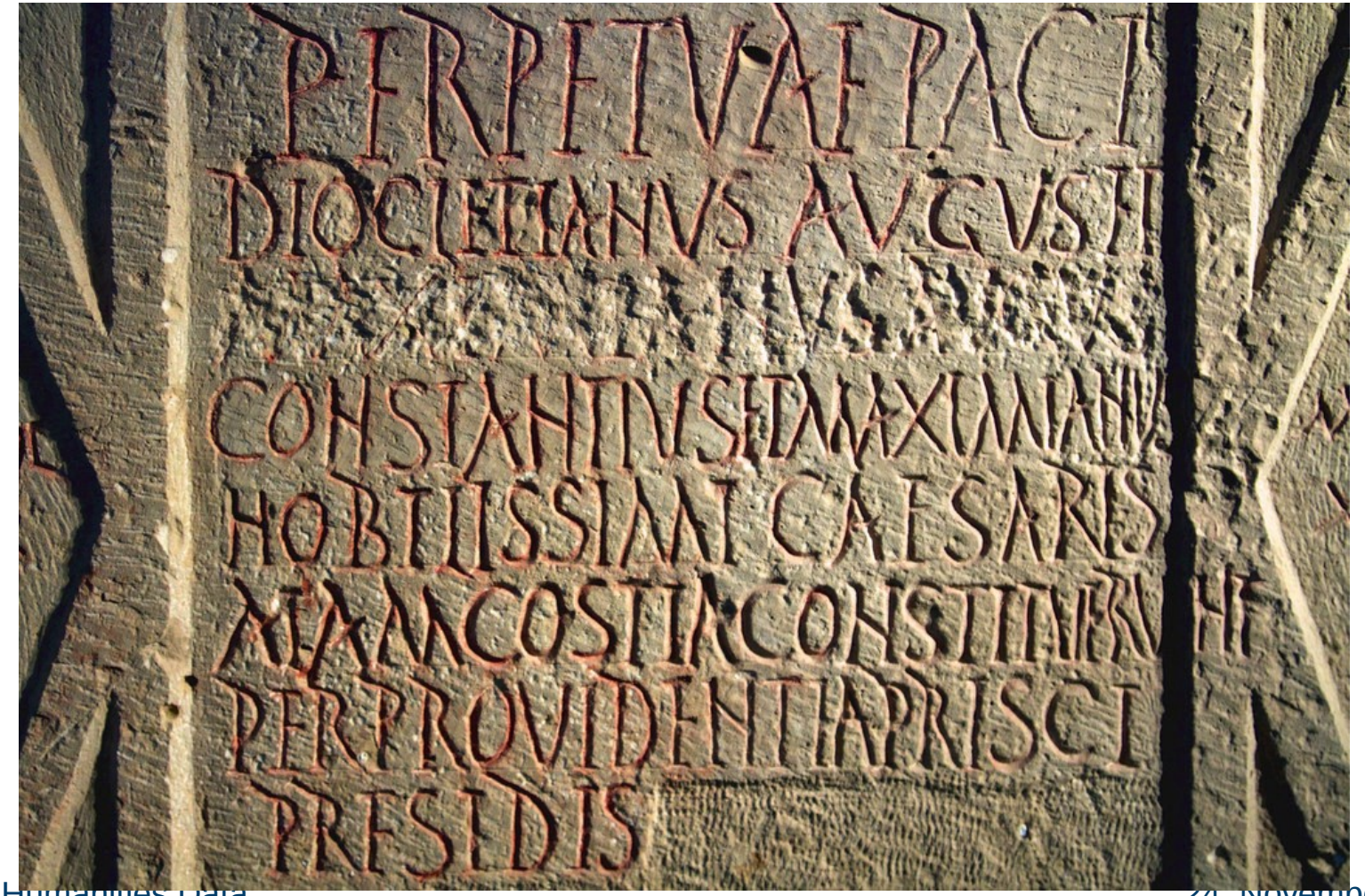
The data



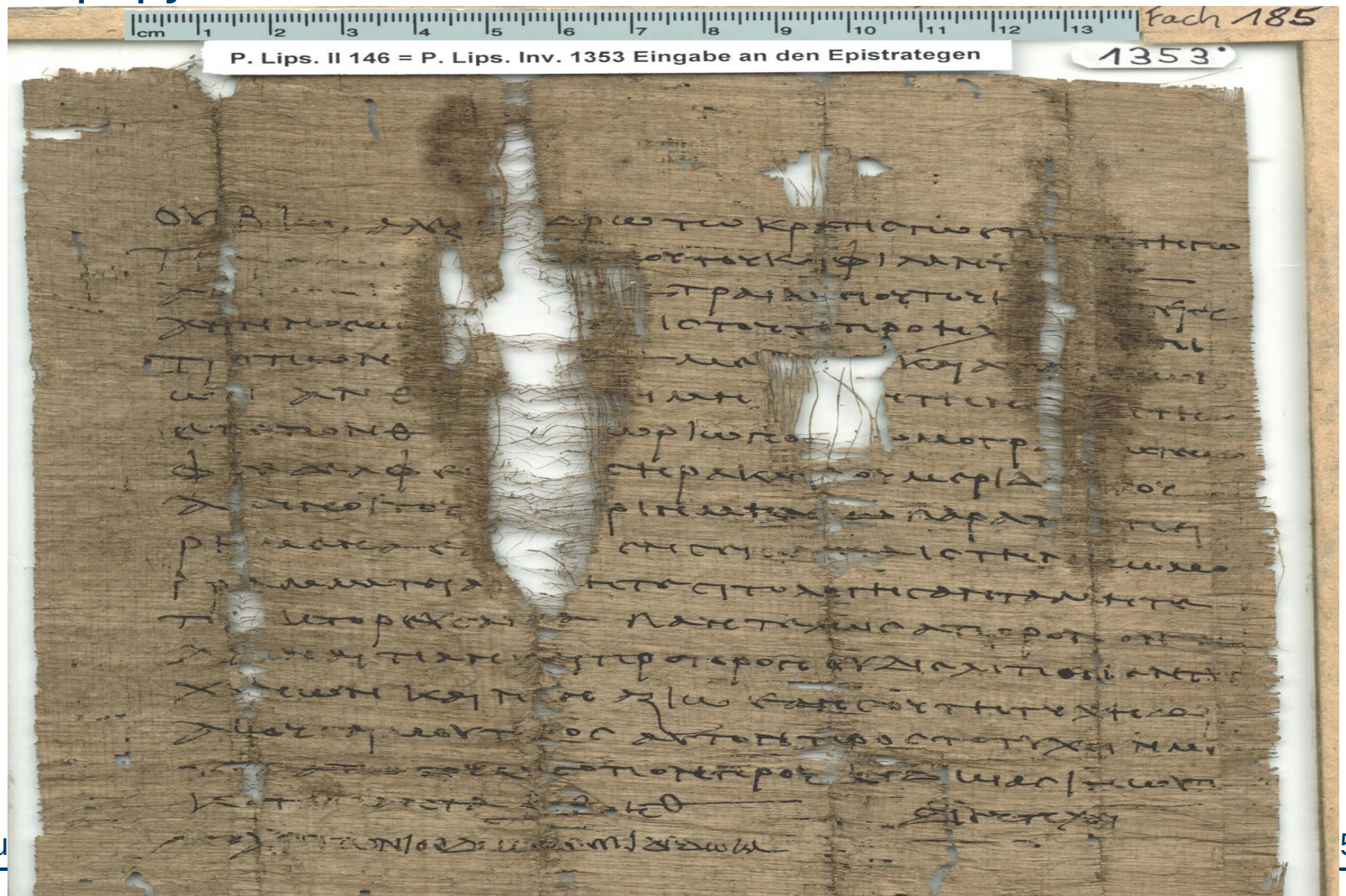
Textcorrection

Possible passage in the text:
Platon Timaios, 38c7 bis 38d4 (from TLG-Online):
σώματα δὲ αὐτῶν ἑκάστων ποιήσας ὁ θεὸς ἔθηκεν εἰς τὰς @1 περιφορὰς ἃς ἡ θατέρου περίοδος ᾔειν, ἑπτὰ οὔσας ὄντα (d.) ἑπτά, σελήνην μὲν εἰς τὸν περὶ γῆν πρῶτον, ἥλιον δὲ εἰς τὸν δεύτερον ὑπὲρ γῆς, ἑωσφόρον δὲ καὶ τὸν ἱερὸν Ἑρμοῦ λεγόμενον εἰς [τὸν] τάχει μὲν ἰσόδρομον ἡλίῳ κύκλον ἰόντας,τὴν δὲ ἐναντίαν εἰληχότας αὐτῷ δύναμιν·

Damnatio Memoriae



The papyrus


Fach 185
1353
P. Lips. II 146 = P. Lips. Inv. 1353 Eingabe an den Epistrategen

Transcribed data

Οὐιβίῳ **Ἀλεξά[ν]δρῳ** τῷ κρατίστῳ ἐπιστρατήγῳ
παρὰ Ἀντ[ωνίου Δ]όμνου τοῦ καὶ Φιλαντι[νό]ου
Ἀντωνίο[υ Ῥωμανο]ῦ Τραιανείου τοῦ κα[ὶ Στρα]τείου
Ἀντινοέως. [οὐκ ἂν] εἰς τοῦτο προήχθ[η]ν, ἐπι-
τρόπων [μέγιστ]ε, μέ[τριος] καὶ ἀπράγμων
ὢν ἄνθρ[ωπος,] εἰ μὴ [ὕβρι]ν τὴν μ[εγ]ίστην
ἐπεπόνθ[ειν ὑπὸ] Ὡρίωνο[ς κ]ωμογρα[μ]ματέως
Φ[ι]λαδελφεί[ας τῆ]ς Ἡρακλείδου μερίδο[ς] τοῦ
Ἀρσινοίτου. [οὐ χά]ριν μην[ύ]ω παρὰ τ[ὰ ἀ]πει-
ρημένα ἐα[υτὸ]ν ἐνσείσαντα εἰς τὴν κωμο-
γραμματείαν [μ]ήτε σιτολογήσαντα μήτε
πρ[α]κτορεύσαντα παντελῶς ἄπορον ὄν[τ]α.
δι᾽ ἣν αἰτίαν καὶ πρότερον οὐ διέλιπον ἐντυγ-
χάνων καὶ νῦν ἀξιῶ, ἐάν σου τῇ τύχῃ δόξ[ῃ],
ἀκοῦσαί μου π[ρ]ὸς αὐτὸν πρὸς τὸ τυχεῖν με
τῆς ἀπὸ σοῦ [μι]σοπονήρου ἐγδ[ι]κίας, ἵν᾽ ὦ ὑπὸ [σ]οῦ
κατὰ πάντα βεβοηθ(ημένος). διευτύχει
Ἀντώνιος Δόμνος ἐπιδέδωκα.

Input form

Text

Οὐιβίῳ [_]λεξά[____] τῷ κρατίστῳ ἐπιστρατήγῳ
παρὰ Ἀντ[ωνίου Δ]όμνου τοῦ καὶ Φιλαντι[νό]ου.
Ἀντωνίο[υ Ῥωμανο]ῦ Τραιανείου τοῦ κα[ὶ Στρα]τείου
Ἀντινοέως. [οὐκ ἂν] εἰς τοῦτο προήχθ[η]ν, ἐπι-
τρόπων [μέγιστ]ε, μέ[τριος] καὶ ἀπράγμων
ὢν ἄνθρ[ωπος,] εἰ μὴ [ὕβρι]ν τὴν μ[εγ]ίστην
ἐπεπόνθ[ειν ὑπὸ] Ὡρίωνο[ς κ]ωμογρα[μ]ματέως
Φ[ι]λαδελφεί[ας τῆ]ς Ἡρακλείδου μερίδο[ς] τοῦ
Ἀρσινοίτου. [οὐ χά]ριν μην[ύ]ω παρὰ τ[ὰ ἀ]πει-
ρημένα ἑα[υτὸ]ν ἐνσείσαντα εἰς τὴν κωμο-

TLG PHI7 Epiduke Send

#	Sentence

Parsed input (parsed for Leiden conventions)

Text

```
Οὐιβίῳ [_]λεξά[____] τῷ κρατίστῳ ἐπιστρατήγῳ
παρὰ Ἀντ[ωνίου Δ]όμνου τοῦ καὶ Φιλαντι[νό]ου.
Ἀντωνίο[υ Ῥωμανο]ῦ Τραιανείου τοῦ κα[ὶ Στρα]τείου
Ἀντινοέως. [οὐκ ἂν] εἰς τοῦτο προήχθ[η]ν, ἐπι-
τρόπων [μέγιστ]ε, μέ[τριος] καὶ ἀπράγμων
ὢν ἄνθρ[ωπος,] εἰ μὴ [ὕβρι]ν τὴν μ[εγ]ίστην
ἐπεπόνθ[ειν ὑπὸ] Ὡρίωνο[ς κ]ωμογρα[μ]ματέως
Φ[ι]λαδελφεί[ας τῆ]ς Ἡρακλείδου μερίδο[ς] τοῦ
Ἀρσινοίτου. [οὐ χά]ριν μην[ύ]ω παρὰ τ[ὰ ἀ]πει-
ρημένα ἑα[υτὸ]ν ἐνσείσαντα εἰς τὴν κωμο-
```

TLG
PHI7
Epiduke
Send

#	Sentence
0	Οὐιβίῳ [_]λεξά[____] τῷ κρατίστῳ ἐπιστρατήγῳ παρὰ Ἀντ[ωνίου] [Δ]όμνου τοῦ καὶ Φιλαντι[νό]ου Ἀντωνίο[υ] Ῥωμανο]ῦ Τραιανείου τοῦ κα[ὶ] [Στρα]τείου Ἀντινοέως
1	[οὐκ] [ἂν] εἰς τοῦτο προήχθ[η]ν ἐπι- τρόπων [μέγιστ]ε μέ[τριος] καὶ ἀπράγμων ὢν ἄνθρ[ωπος] εἰ μὴ [ὕβρι]ν τὴν μ[εγ]ίστην ἐπεπόνθ[ειν] [ὑπὸ] Ὡρίωνο[ς] [κ]ωμογρα[μ]ματέως Φ[ι]λαδελφεί[ας] [τῆ]ς Ἡρακλείδου μερίδο[ς] τοῦ Ἀρσινοίτου
2	[οὐ] [χά]ριν μην[ύ]ω παρὰ τ[ὰ] [ἀ]πει- ρημένα ἑα[υτὸ]ν ἐνσείσαντα εἰς τὴν κωμο- γραμματείαν [μ]ήτε σπολογήσαντα μήτε πρ[α]κτορεύσαντα παντελῶς ἄπορον ὄν[τ]α
3	δι ἣν αἰτίαν καὶ πρότερον οὐ διέλιπον ἐντυγ- χάνων καὶ νῦν ἀξιῶ ἐάν σου τῇ τύχῃ δόξ[ῃ] ἀκοῦσαί μου π[ρ]ὸς αὐτὸν πρὸς τὸ τυχεῖν με τῆς ἀπὸ σοῦ [μι]σοπονήρου ἐγδ[ι]κίας ἵν ὦ ὑπὸ [σ]οῦ κατὰ πάντα βεβοηθ(ημένος)
4	διευτύχει Ἀντώνιος Δόμνος ἐπιδέδωκα

Strategy 1: use only information of the word

Word length + „survived“ pattern

[_]λεξά[___]

Interpreted word : _λεξά____

Length : 9

Candidate	Score	☐ Word length	☐ Neighboured letter bigrams	☐ Word similarity (letters)	☐ Named Entity	☐ Word bigram	☐ Semantic context	☐ Classification	Show
Ἀλεξάνδρᾳ	2	1.0		1.0					
Ἀλεξάνρου	2	1.0		1.0					
Ἀλεξάνδρα	2	1.0		1.0					
Ἀλεξάνδρῳ	2	1.0		1.0					
Ἀλεξάρχου	2	1.0		1.0					
Ἀλεξάνδου	2	1.0		1.0					
ἐνεχάραξα	1	1.0							
ὁμολογεῖ	1	1.0							
ὁλοκλήρον	1	1.0							

Strategy 2: use only of context information

Word bigrams + co-occurrences + classification

[_]λεξά[___]

Interpreted word : _λεξά____

Length : 9

Candidate	Score	☐ Word length	☐ Neighboured letter bigrams	☐ Word similarity (letters)	☐ Named Entity	☐ Word bigram	☐ Semantic context	☐ Classification	Show
νομοῦ	3					0.5	0.4	0.000	
Ἀλεξάνδρῳ	3					1.0	0.8	0.003	
ἀπόδος	2					0.5		0.001	
Δόμνου	2						0.8	0.040	
ἀνέτεινα	2						0.2	0.025	
Αὐρηλίου	2						0.4	0.000	
Ἀχιλλεῖ	2					0.5	0.2		
Ἑπτὰ	2						0.2	0.010	
διαδεχομένῳ	2						0.2	0.250	
Σεουηριανῷ	2					0.5	0.2		
Αἰγύπτου	2						0.4	0.001	
ἡγεμόνι	2						0.2	0.001	
Λικιννιανῷ	2						0.2	0.200	

The „real“ Strategy 2: use only of context information and removing any information about the damaged word

Reparsing

Text

```
Οὐιβίῳ [_] τῷ κρατίστῳ ἐπιστρατήγῳ
παρὰ Ἀντ[ωνίου Δ]όμνου τοῦ καὶ Φιλαντι[νό]ου.
Ἀντωνίο[υ Ῥωμανο]ῦ Τραιανείου τοῦ κα[ὶ Στρα]τείου
Ἀντινοέως. [οὐκ ἂν] εἰς τοῦτο προήχθ[η]ν, ἐπι-
τρόπων [μέγιστ]ε, μέ[τριος] καὶ ἀπράγμων
ὢν ἄνθρ[ωπος,] εἰ μὴ [ὕβρι]ν τὴν μ[εγ]ίστην
ἐπεπόνθ[ειν ὑπὸ] Ὡρίωνο[ς κ]ωμογρα[μ]ματέως
Φ[ι]λαδελφεί[ας τῆ]ς Ἡρακλείδου μερίδο[ς] τοῦ
Ἀρσινοίτου. [οὐ χά]ριν μην[ύ]ω παρὰ τ[ὰ ἀ]πει-
ρημένα ἑα[υτὸ]ν ἐνσείσαντα εἰς τὴν κωμο-
```

TLG
PHI7
Epiduke
Send

#	Sentence
0	Οὐιβίῳ [_] τῷ κρατίστῳ ἐπιστρατήγῳ παρὰ Ἀντ[ωνίου] [Δ]όμνου τοῦ καὶ Φιλαντι[νό]ου Ἀντωνίο[υ] Ῥωμανο]ῦ Τραιανείου τοῦ κα[ὶ] [Στρα]τείου Ἀντινοέως
1	[οὐκ] [ἂν] εἰς τοῦτο προήχθ[η]ν ἐπι- τρόπων [μέγιστ]ε μέ[τριος] καὶ ἀπράγμων ὢν ἄνθρ[ωπος] εἰ μὴ [ὕβρι]ν τὴν μ[εγ]ίστην ἐπεπόνθ[ειν] [ὑπὸ] Ὡρίωνο[ς] [κ]ωμογρα[μ]ματέως Φ[ι]λαδελφεί[ας] [τῆ]ς Ἡρακλείδου μερίδο[ς] τοῦ Ἀρσινοίτου
2	[οὐ] [χά]ριν μην[ύ]ω παρὰ τ[ὰ] [ἀ]πει- ρημένα ἑα[υτὸ]ν ἐνσείσαντα εἰς τὴν κωμο- γραμματείαν [μ]ήτε σπολογήσαντα μήτε πρ[α]κτορεύσαντα παντελῶς ἄπορον ὄν[τ]α
3	δι ἣν αἰτίαν καὶ πρότερον οὐ διέλιπον ἐντυγ- χάνων καὶ νῦν ἀξιῶ ἐάν σου τῇ τύχῃ δόξ[η] ἀκοῦσαί μου π[ρ]ὸς αὐτὸν πρὸς τὸ τυχεῖν με τῆς ἀπὸ σοῦ [μι]σοπονήρου ἐγδ[ι]κίας ἵν ὦ ὑπὸ [σ]οῦ κατὰ πάντα βεβοηθ(ημένος)
4	διευτύχει Ἀντώνιος Δόμνος ἐπιδέδωκα

Word bigrams + co-occurrences + classification

[_]

Interpreted word : _

Length : 1

Candidate	Score	☐ Word length	☐ Neighboured letter bigrams	☐ Word similarity (letters)	☐ Named Entity	☐ Word bigram	☐ Semantic context	☐ Classification	Show
νομοῦ	3					0.5	0.4	0.000	
Ἀλεξάνδρῳ	3					1.0	0.8	0.003	
ἀπόδος	2					0.5		0.001	
Δόμνου	2						0.8	0.040	
ἀνέτεινα	2						0.2	0.025	
Αὐρηλίου	2						0.4	0.000	
Ἀχιλλεῖ	2					0.5	0.2		
Ἑπτὰ	2						0.2	0.010	
διαδεχομένῳ	2						0.2	0.250	
Σεουηριανῷ	2					0.5	0.2		
Αἰγύπτου	2						0.4	0.001	
ἡγεμόνι	2						0.2	0.001	
Λικιννιανῷ	2						0.2	0.200	

Strategy 3: Complete strategy

Multiple options

[_]λεξά[___]

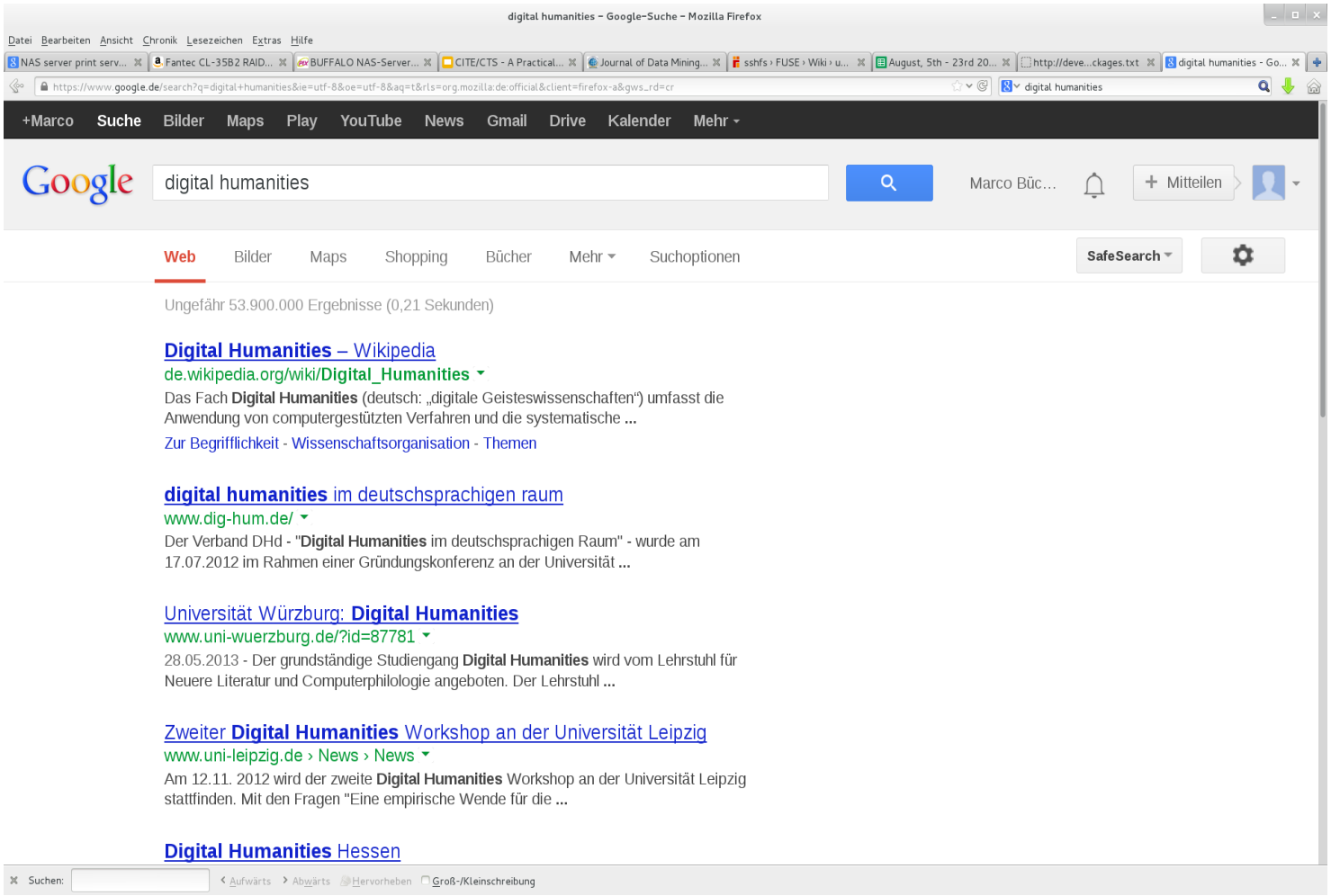
Interpreted word : _λεξά____

Length : 9

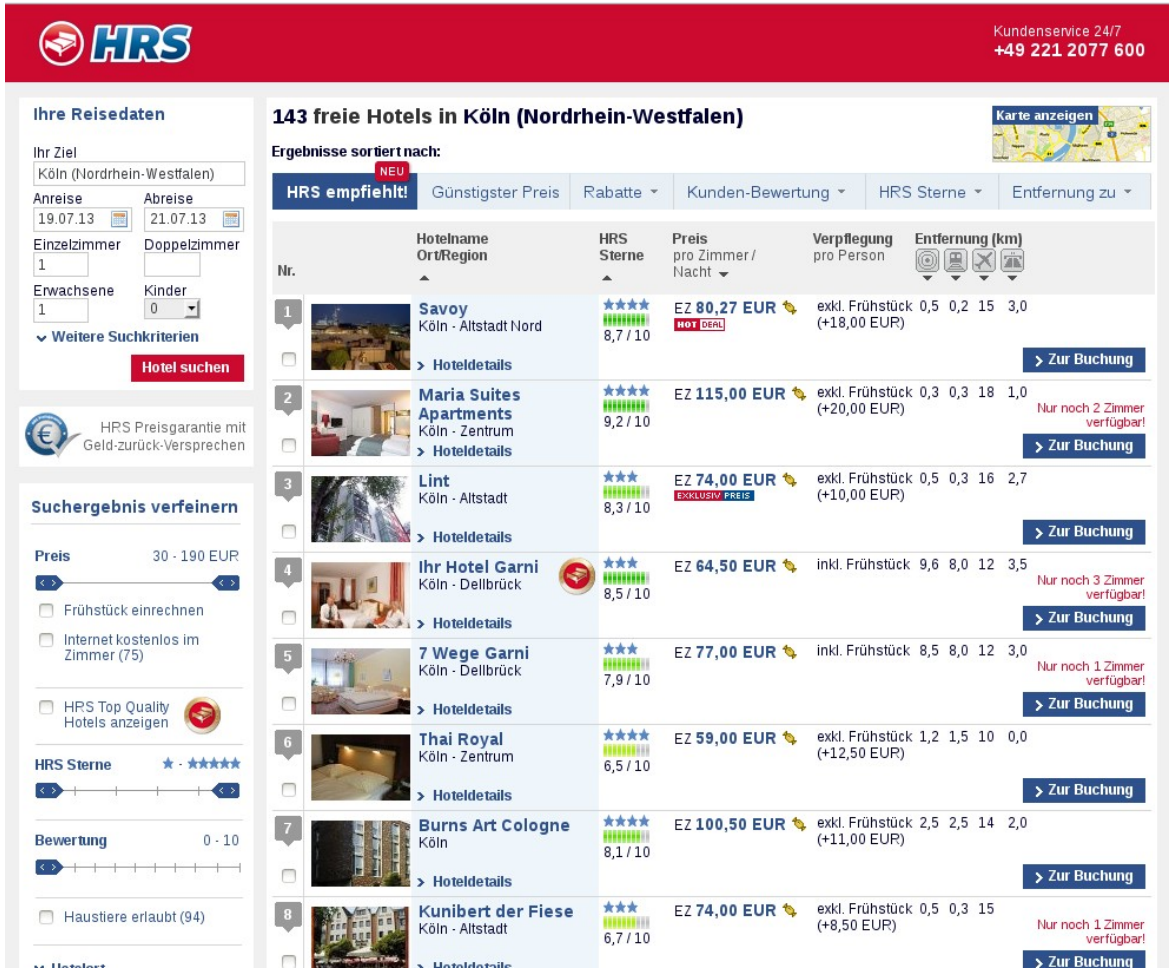
Candidate	Score	☐ Word length	☐ Neighboured letter bigrams	☐ Word similarity (letters)	☐ Named Entity	☐ Word bigram	☐ Semantic context	☐ Classification	Show
Ἀλεξάνδρῳ	5	1.0		1.0		1.0	0.8	0.003	
γενομένην	3	1.0				0.5	0.2		
διοίκησιν	3	1.0					0.2	0.008	
νομοῦ	3					0.5	0.4	0.000	
Ἀντινοέως	3	1.0					0.8	0.011	
βιβλιδίου	3	1.0					0.4	0.003	
ἐπιπρόπων	3	1.0					0.2	0.005	
Ἀντινοέων	3	1.0					0.4	0.002	
Δημητρίωι	3	1.0					0.2	0.002	
βιβλιδίων	3	1.0					0.2	0.002	
Στρατείου	3	1.0					0.6	0.214	
στρατηγὸς	3	1.0					0.2	0.001	
Ἀλεξάρχου	2	1.0		1.0					

Search & Find

How to find relevant information in massive data?



How to find relevant information in massive data?



Definition of co-occurrence

Definition of co-occurrences:

- Common occurrence of at least two objects/events within a dedicated window
- Possible windows in Humanities: line, sentence, paragraph, document, author, century

Motivation:

- Psycholinguistic experiments: Given a word: What is the first word that comes to mind for test subjects?

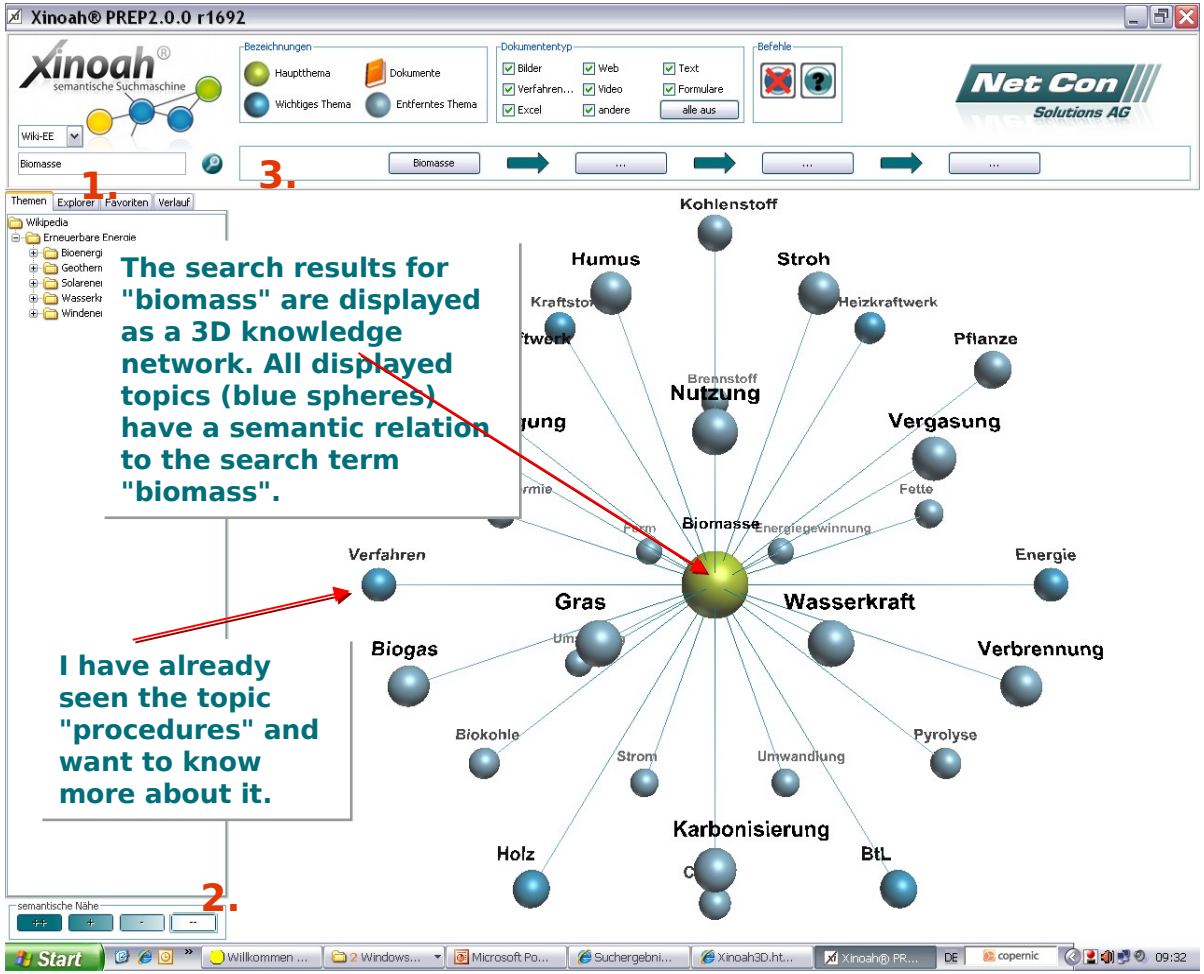
Definition of co-occurrence

Motivation:

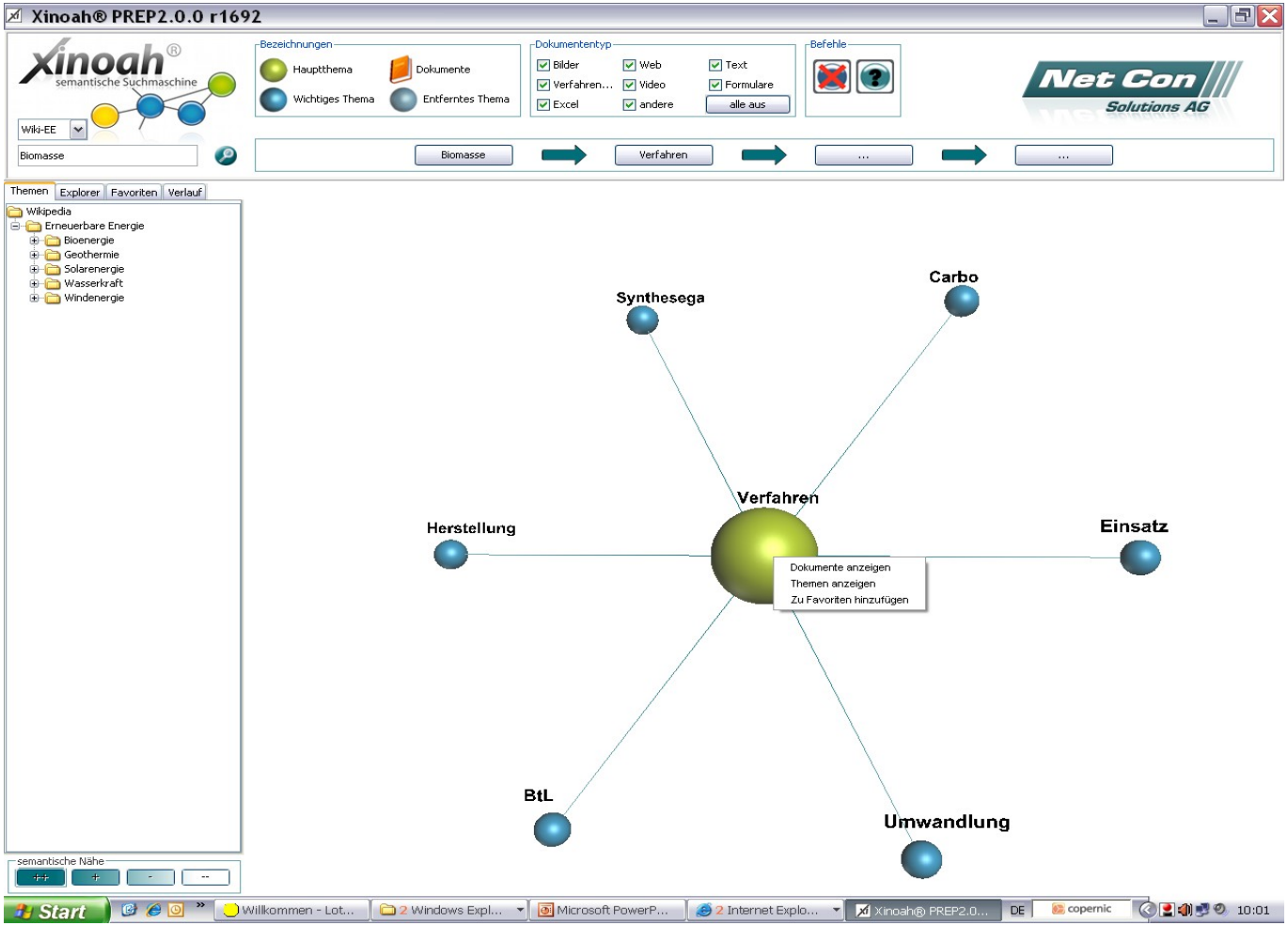
- Psycholinguistic experiments: Given a word: What is the first word that comes to mind for test subjects?

Stimulus	Response Prob.	# of Prob.'s	Co-occurrence	Significance
butter	Bread	60	Bread	51
	soft	40	Cheese	49
	Milk	32	Sugar	29
	Margarine	27	Milk	23
	Cheese	20	Margarine	22
	Fat	16	Farina	18
	yellow	14	Eggs	16
	Bread and butter	8	Pound	14
	Box / can	6	Meat	13
	eat	6		

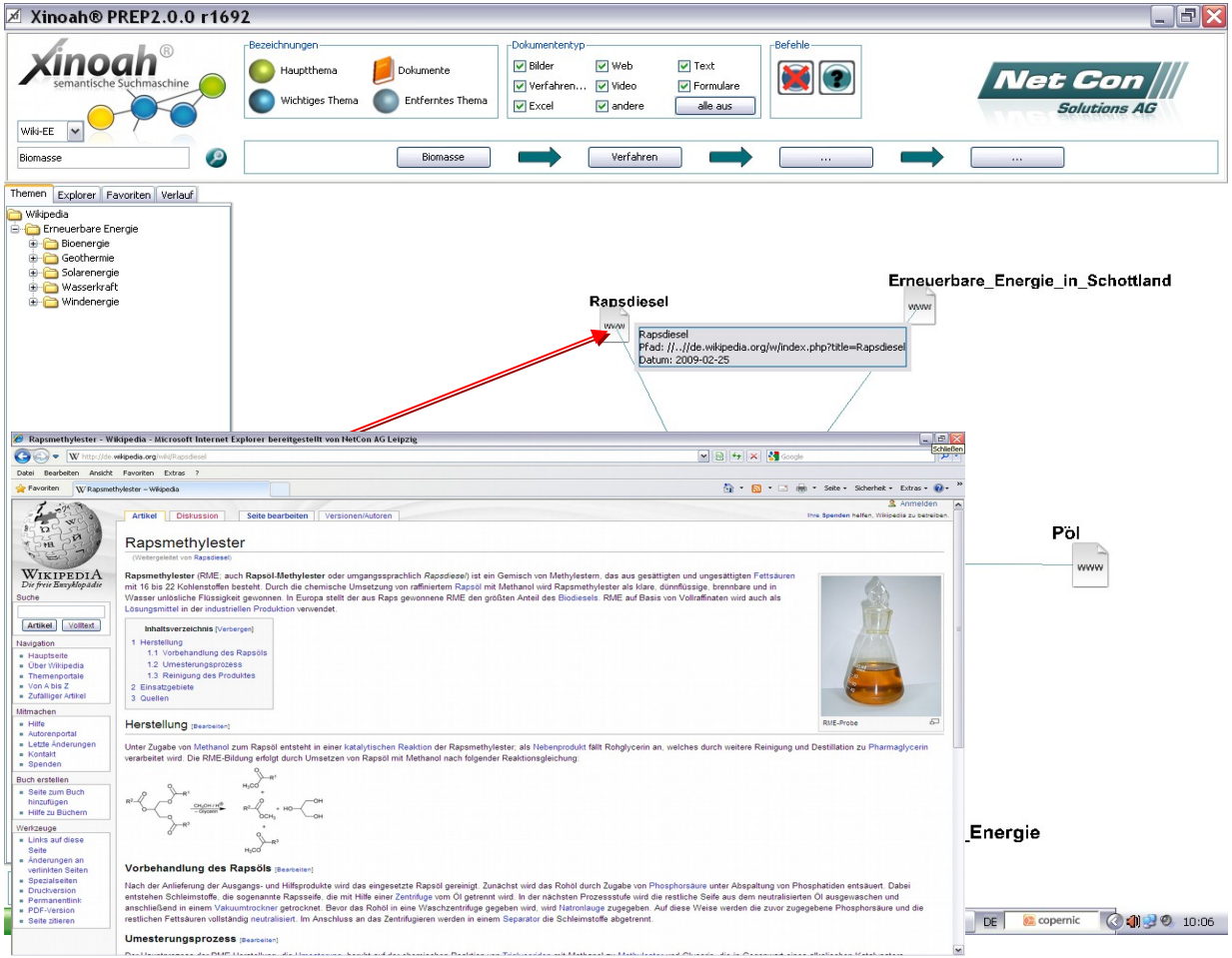
How to find relevant information in massive data?



How to find relevant information in massive data?



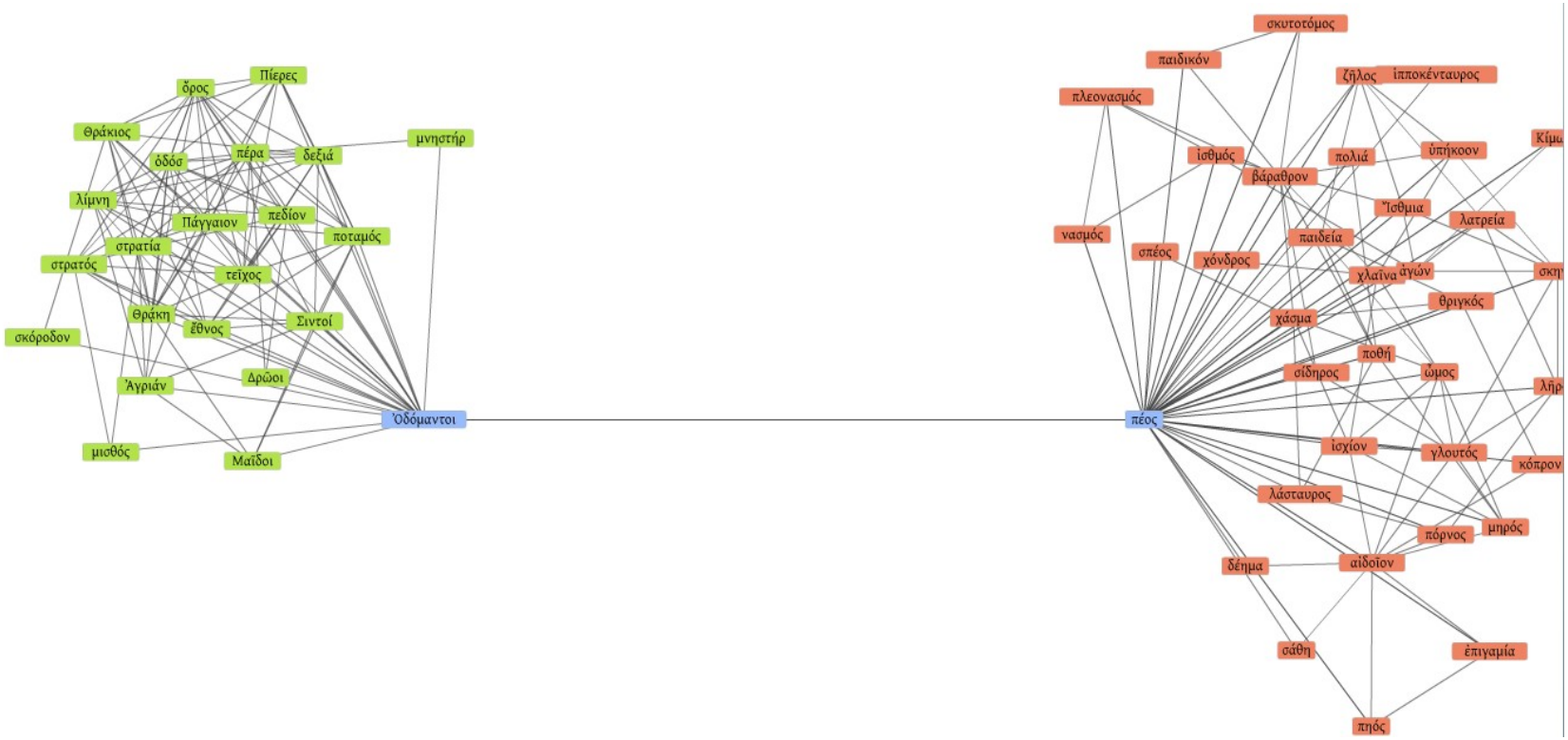
How to find relevant information in massive data?



GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Contrastive Semantics

Visualisation of Contrastive semantics



Main properties

- **Contrast:**

$$contrast(w_i, w_j) = \begin{cases} 1 - sim_{dice}(w_i, w_j) & \text{if } sim_{dice}(w_i, w_j) \leq eps \\ 0 & \text{if } sim_{dice}(w_i, w_j) > eps \end{cases}$$

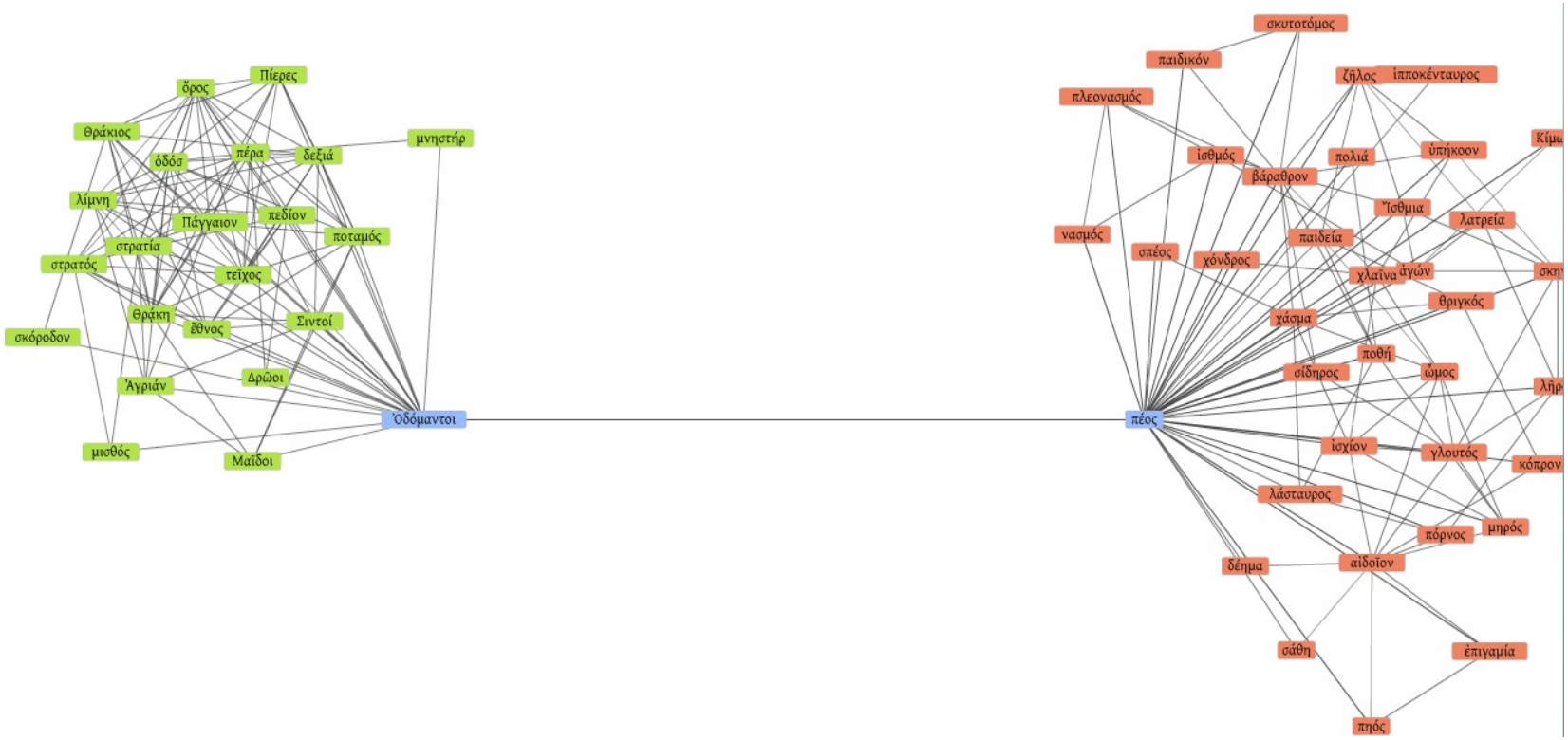
$$sim_{dice}(w_i, w_j) = 2 * \frac{|K_{w_i} \cap K_{w_j}|}{|K_{w_i}| + |K_{w_j}|}$$

- **Locality**:

$$dist(w_i, w_j) \leq eps_{dist} \text{ aus } (w_i, w_j) \in C$$

- **Frequency range of contrastive semantic relations:**
  - Generally less than 10 times of common occurrences

Connectivity?



Some observations

- **Identified clusters:**
  - As shown in examples comedy
  - Sarcasm
  - Cynicism
  - Artificial ambiguity like „Michael Schumacher the red king“ (translated from a German corpus)
  - Scope to gnomology

- **Is there a relation between contrastive semantics and textual reuse?**
  - Clearly, yes.
  - „Evaluation results“: More than 90% of the contrastive semantics have a relation to text reuse

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Historical Text Reuse

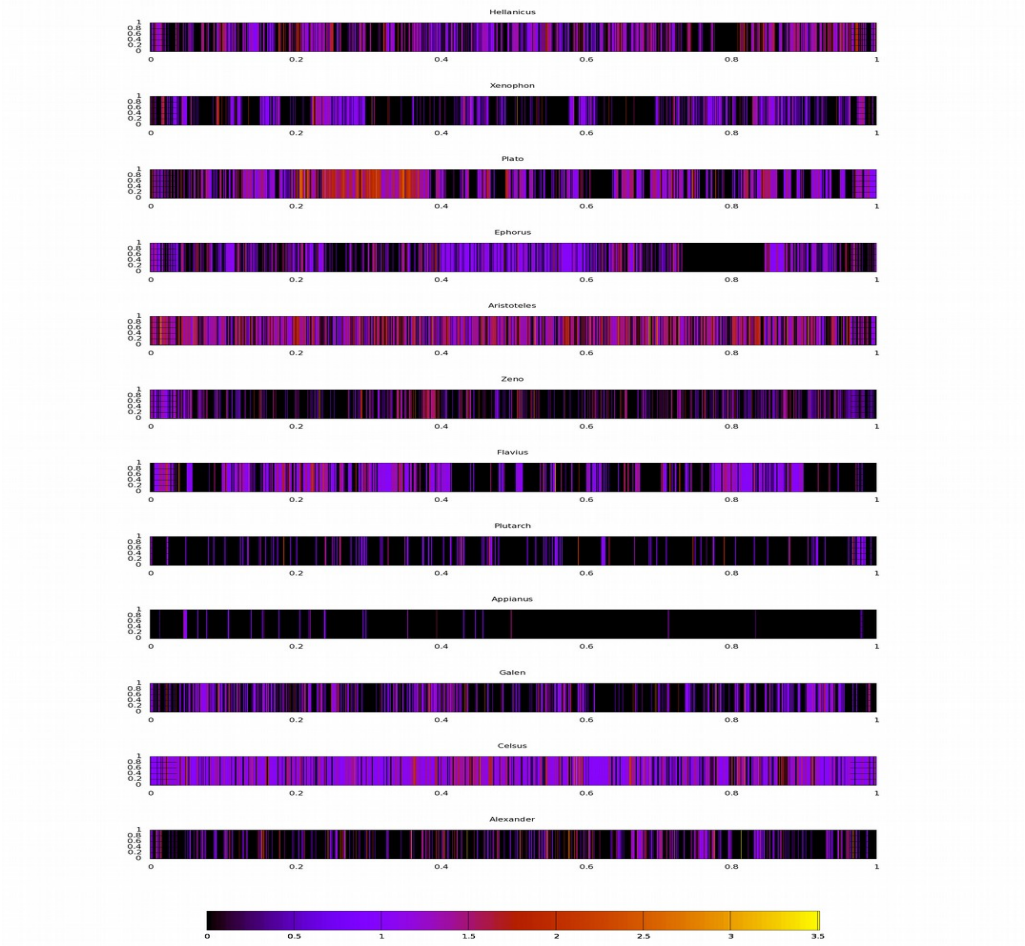
Text Reuse for Humanities and Computer Science

- **Question**: Why is Text Reuse so relevant for Humanities and Computer Science?
- **Premise**: The amount of digitally available data is growing exponentially (Big Data)

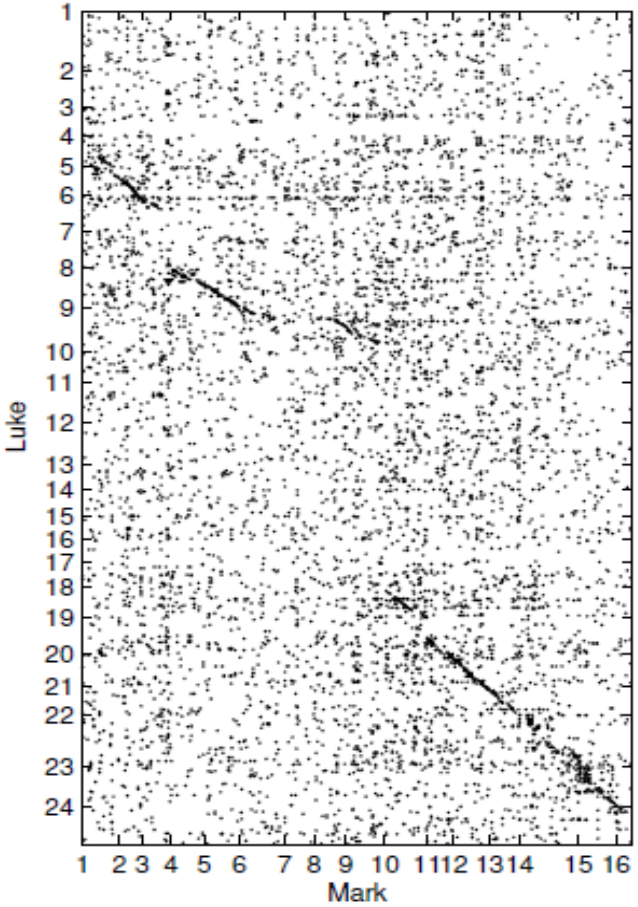
- Humanities:
  - Lines of transmission and textual criticism
  - Transmissions of ideas/thoughts under different circumstances and conditions

- Computer Science:
  - Text Decontamination for stylometry and authorship attribution, dating of texts
  - gen. Text Mining, Corpus Linguistics

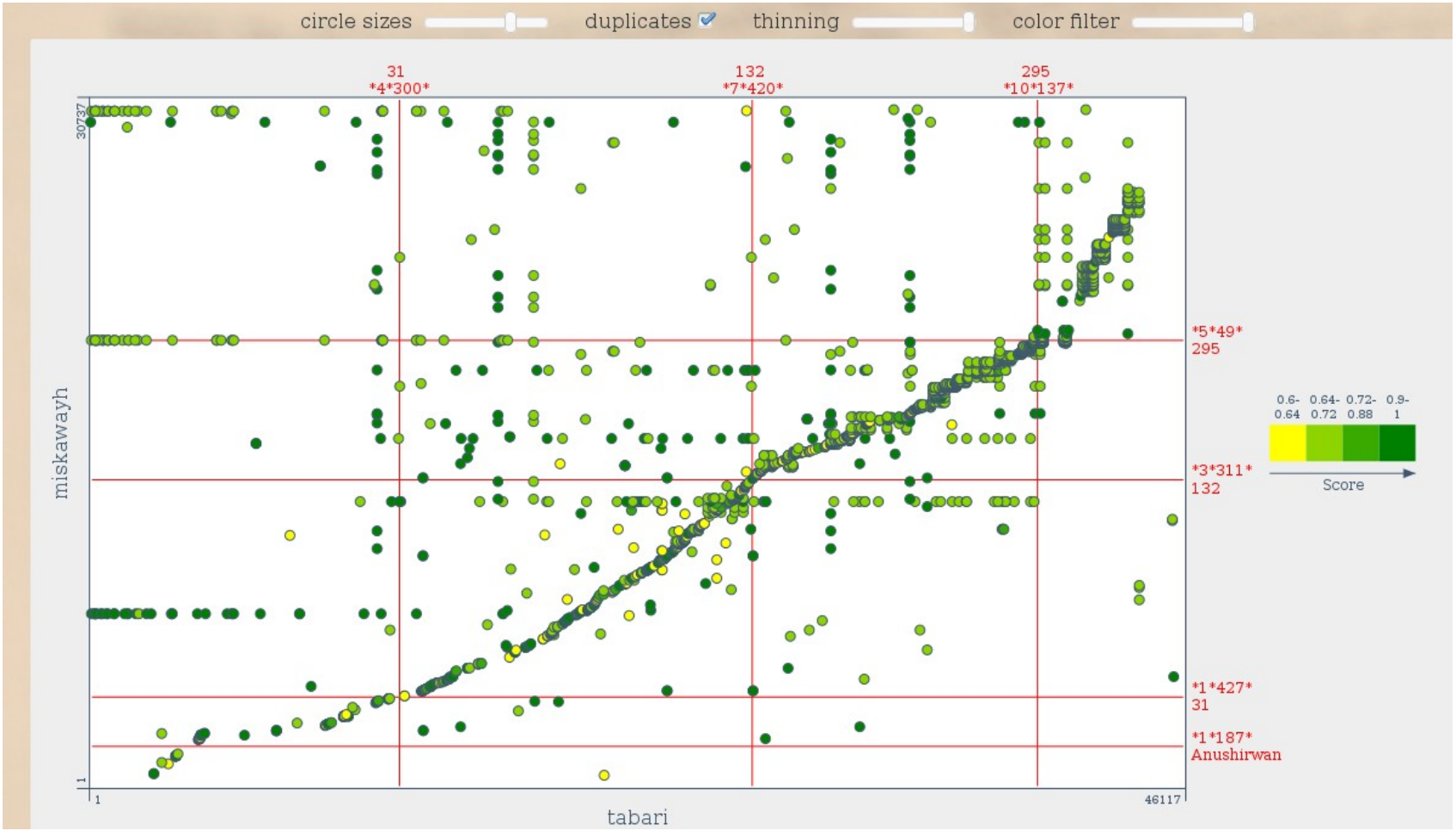
Temperature Map



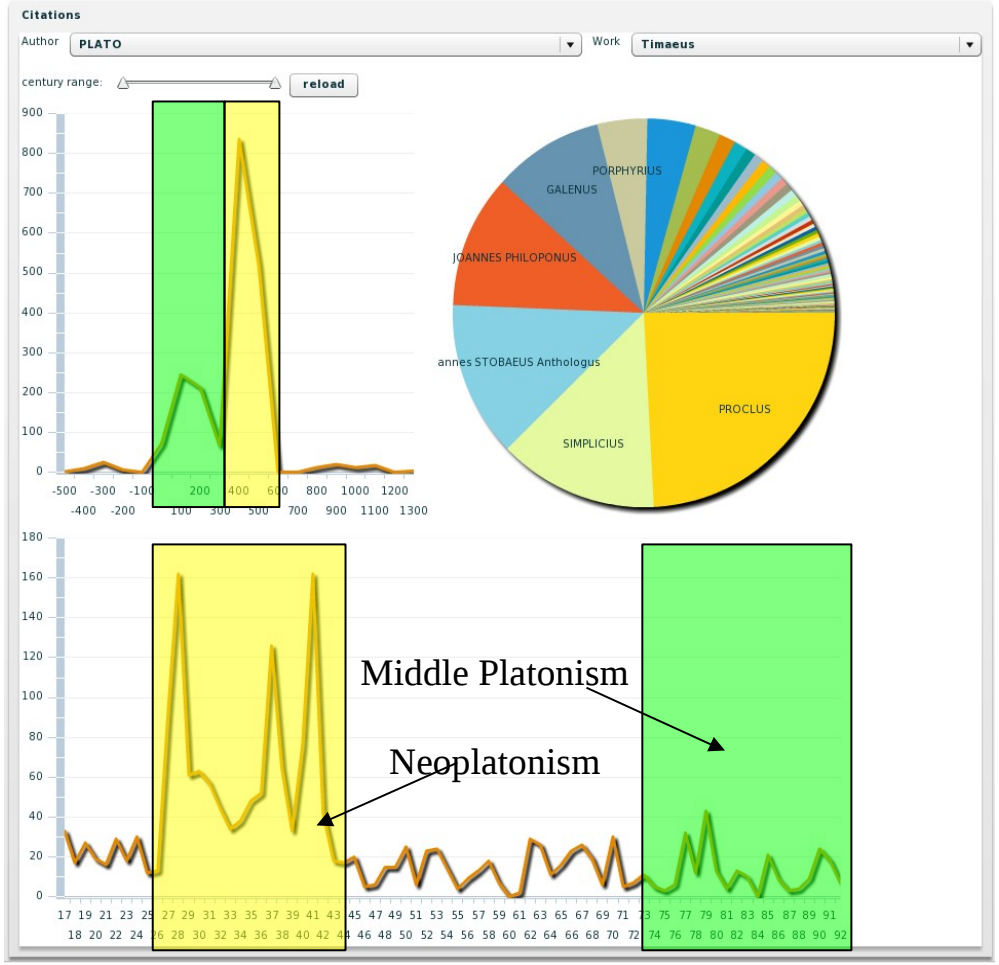
Visualisation problem: Dotplot view



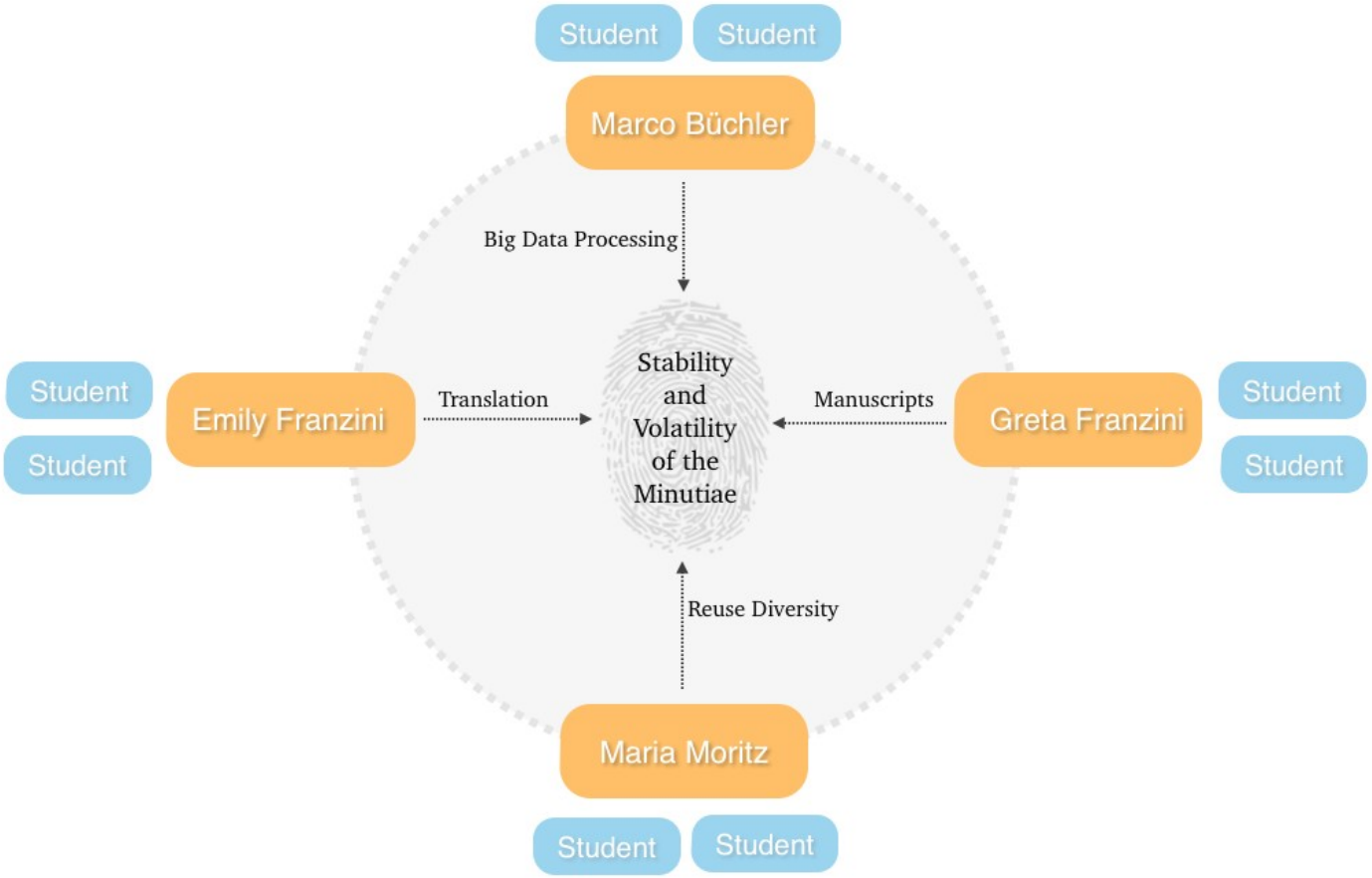
Dotplot view



Visualisation problem: Macro view



eTRAP – Electronic Text Reuse Acquisition Project



GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Thank you!

"*Stealing from one is plagiarism, stealing from many is research*" (Wilson Mitzner, 1876-1933)

eTRAP
Electronic Text Reuse Acquisition Project
INSTITUTE OF COMPUTER SCIENCE
GÖTTINGEN CENTRE FOR DIGITAL HUMANITIES

SPONSORED BY THE
Federal Ministry
of Education
and Research

Visit us at http://etrap.gcdh.de