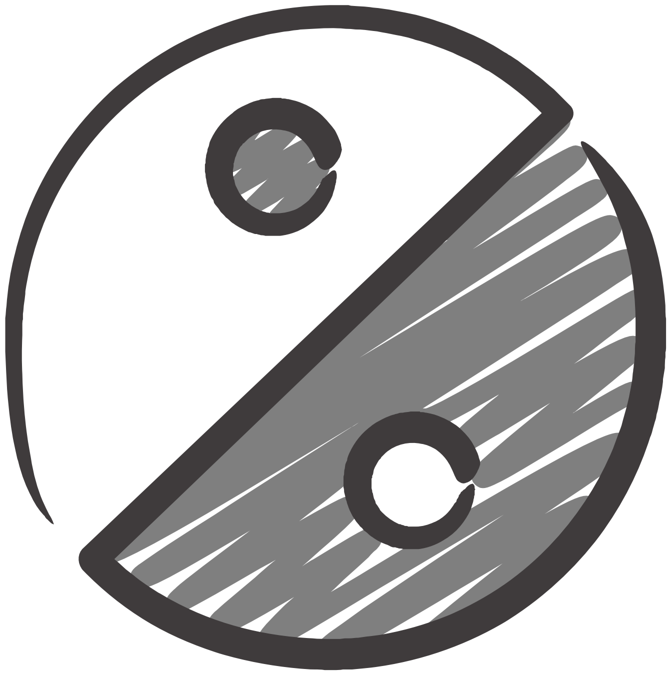
Computational methods for lexical semantic change

Nina Tahmasebi, PhD

University of Gothenburg

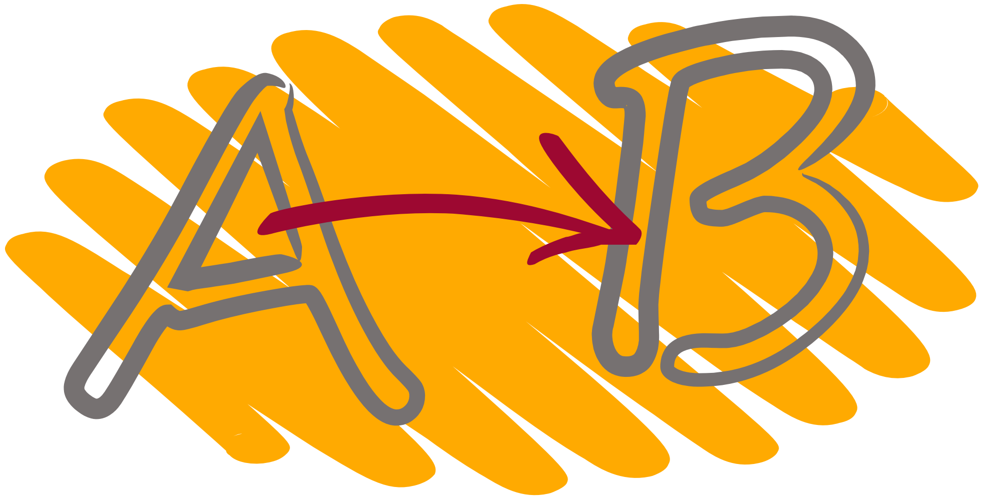
Helsinki, Finland, Feb. 18, 2019

Outline



- Computational methods for LSC
- Historical Linguistic perspective
- Evaluation

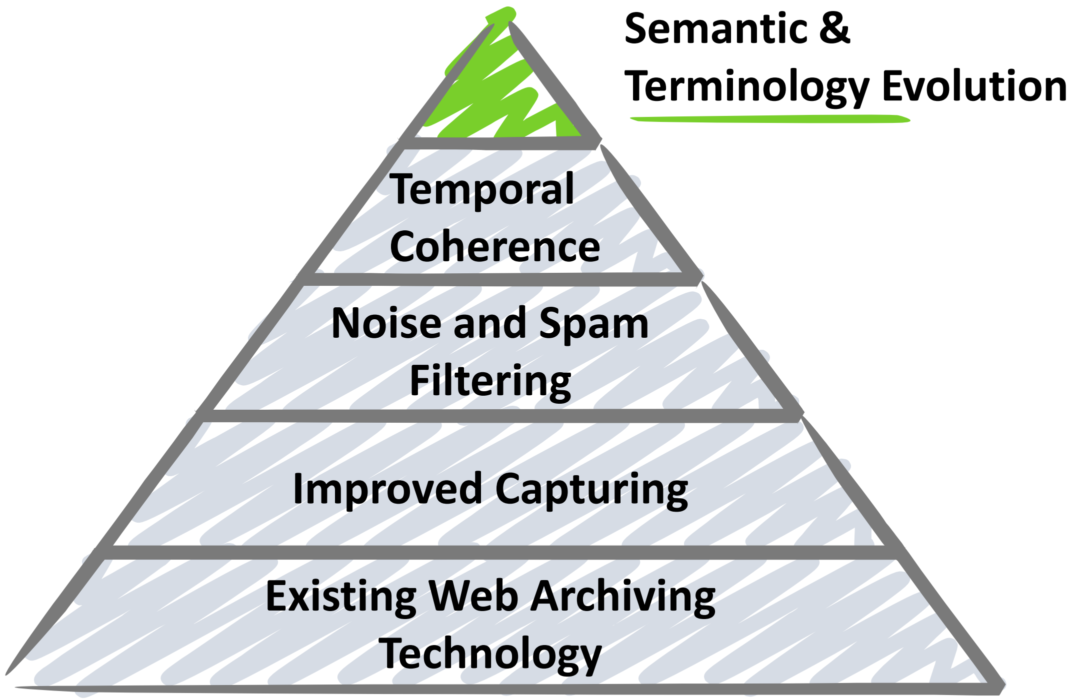
Computational lexical semantic change



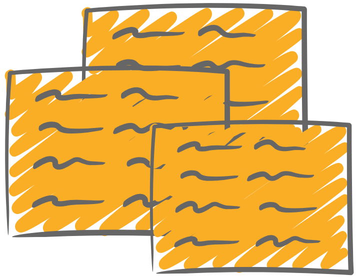
LiWA – Living Web Archives

dealing with terminology evolution

preparing for evolution aware access support



Increasing amount of historical texts in digital format

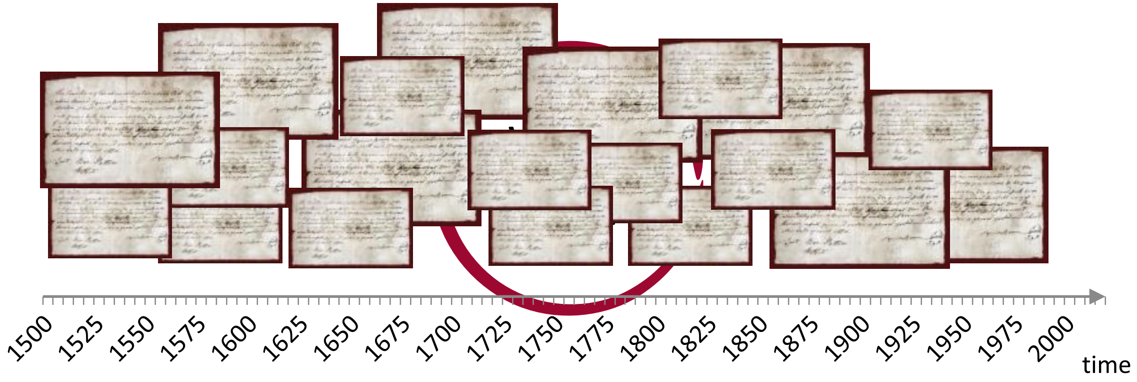


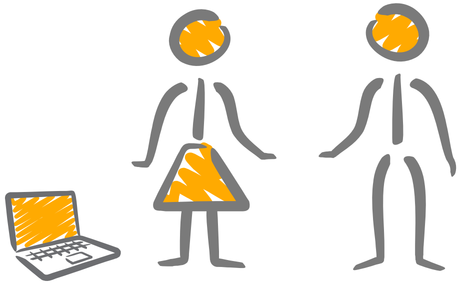
Easy digital access for anyone!
Not only scholars.

Possibility to **digitally analyze** historical documents at **large scale**.

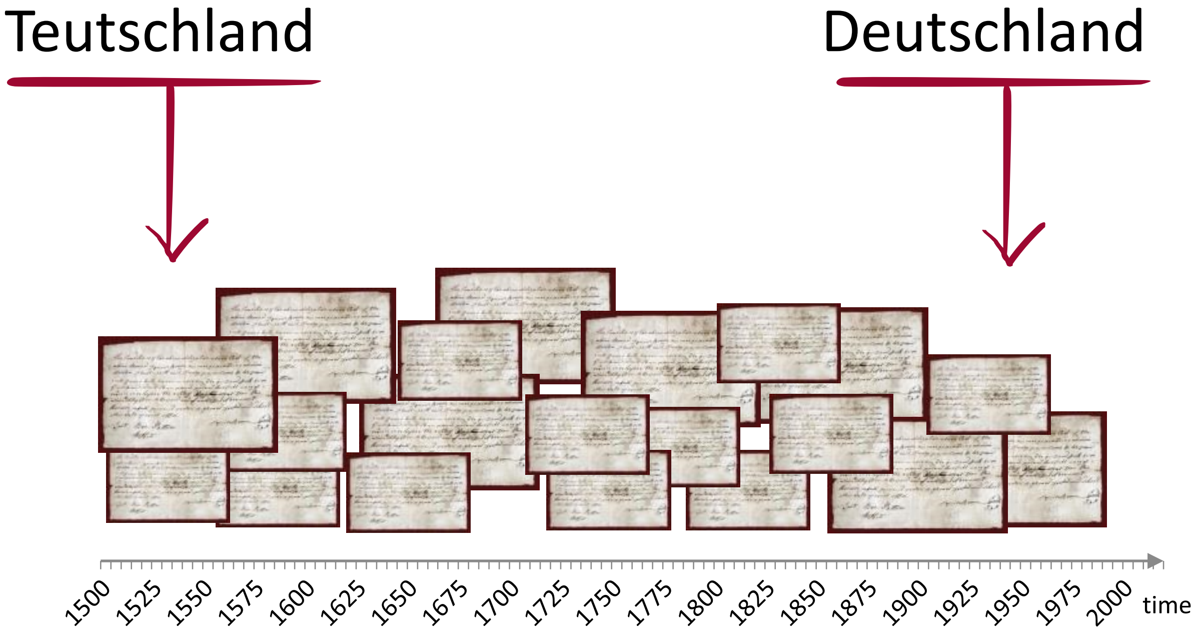
Information from primary sources
Not only modern interpretations.

Text-based Digital Humanities

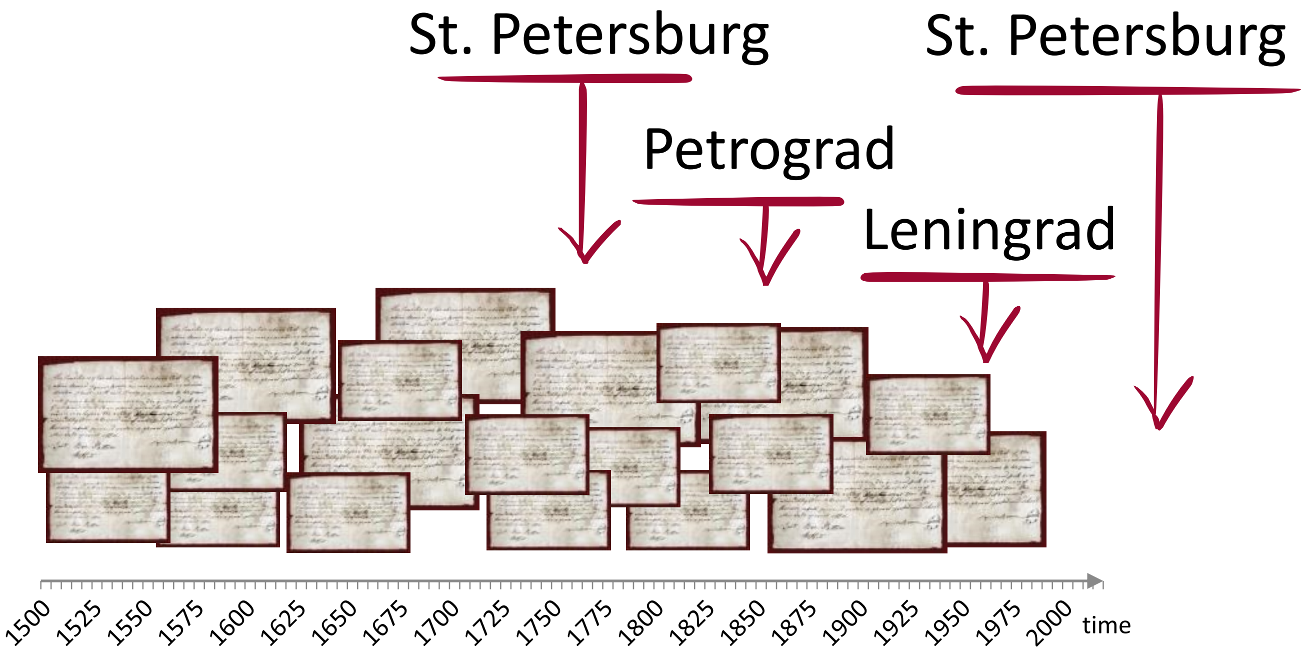


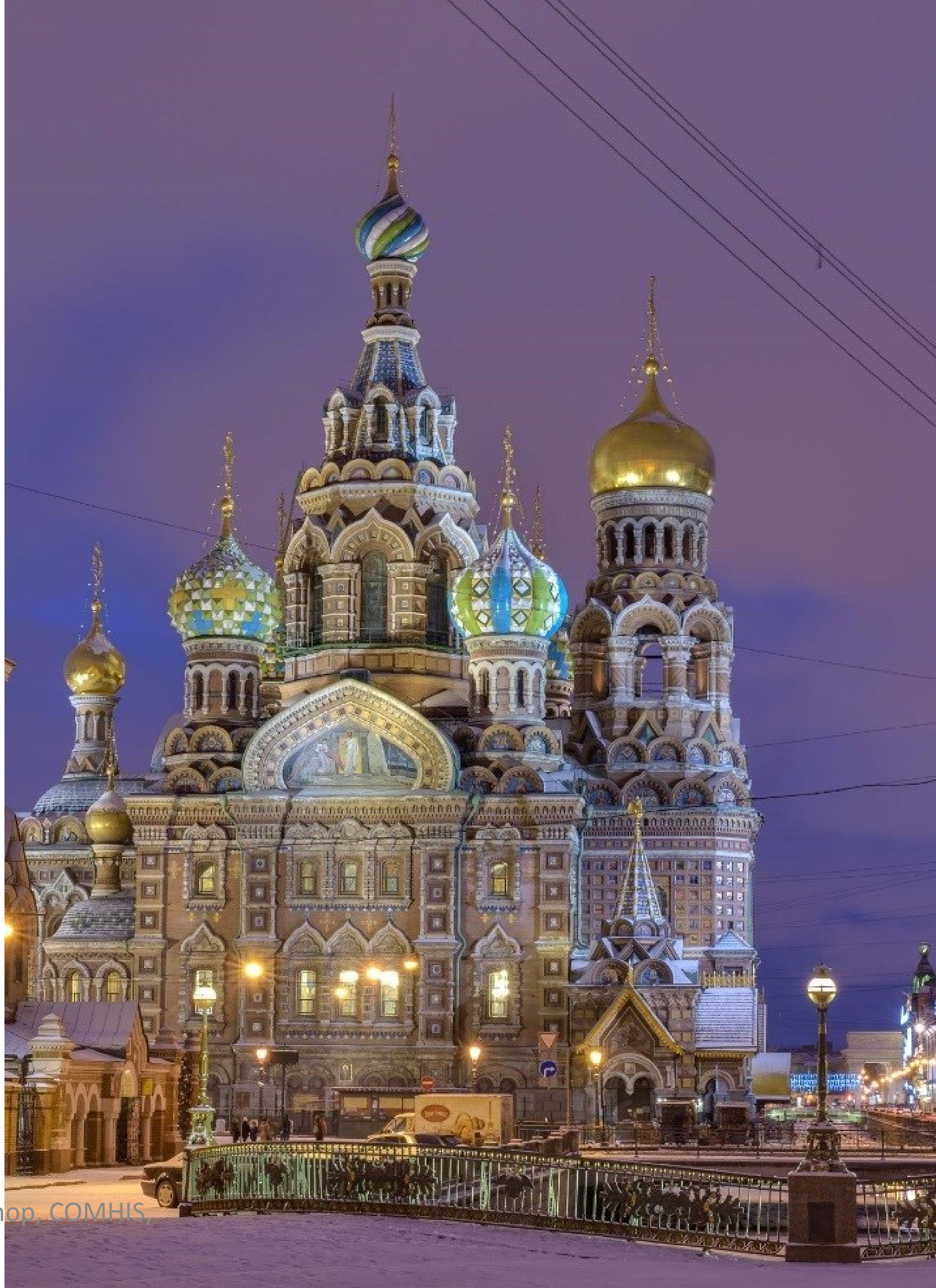


Spelling change

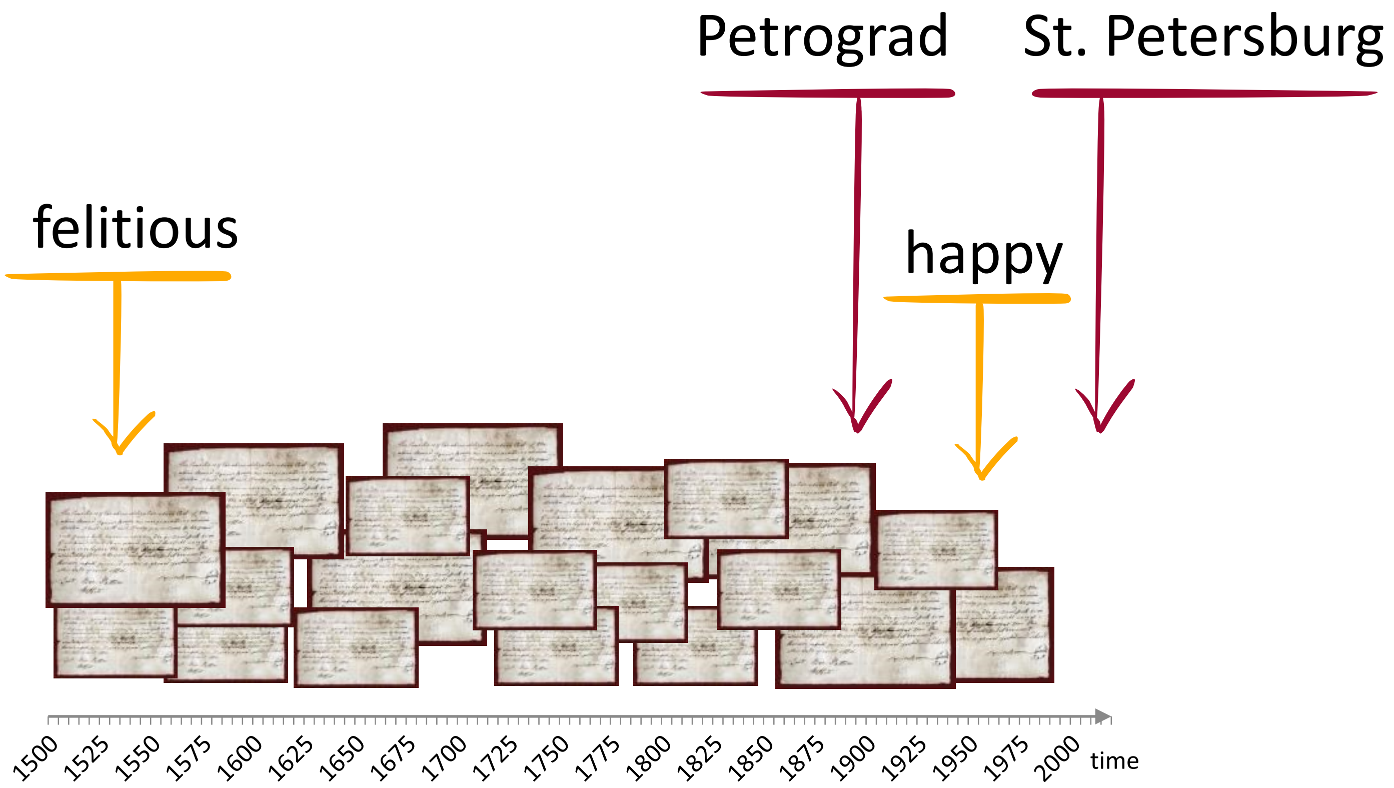


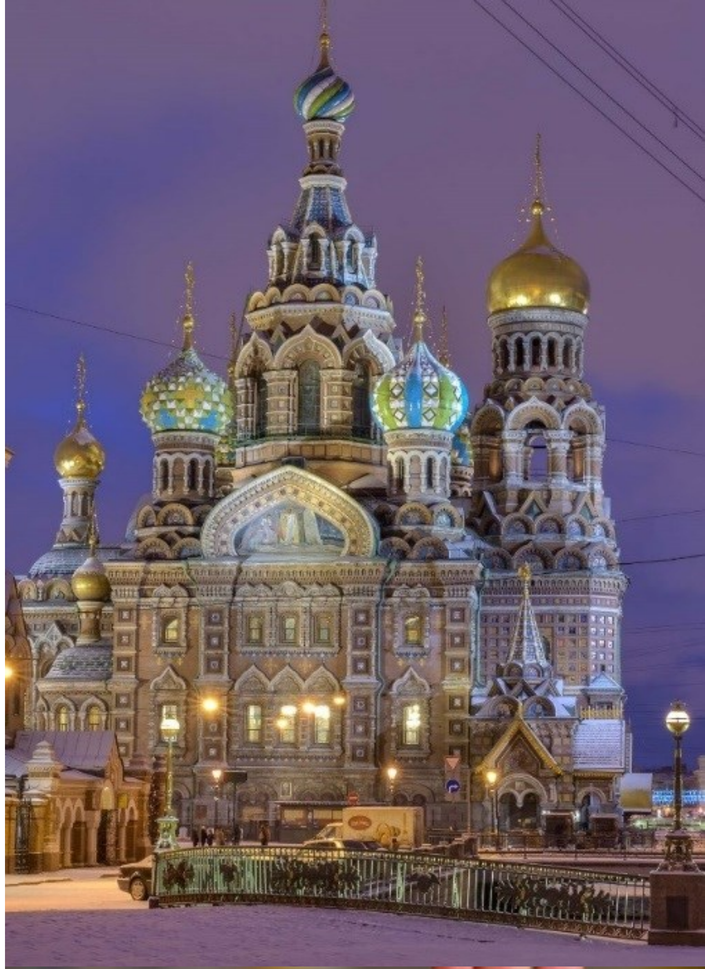
Lexical replacement:
Named entity change





Lexical replacement:





awesome

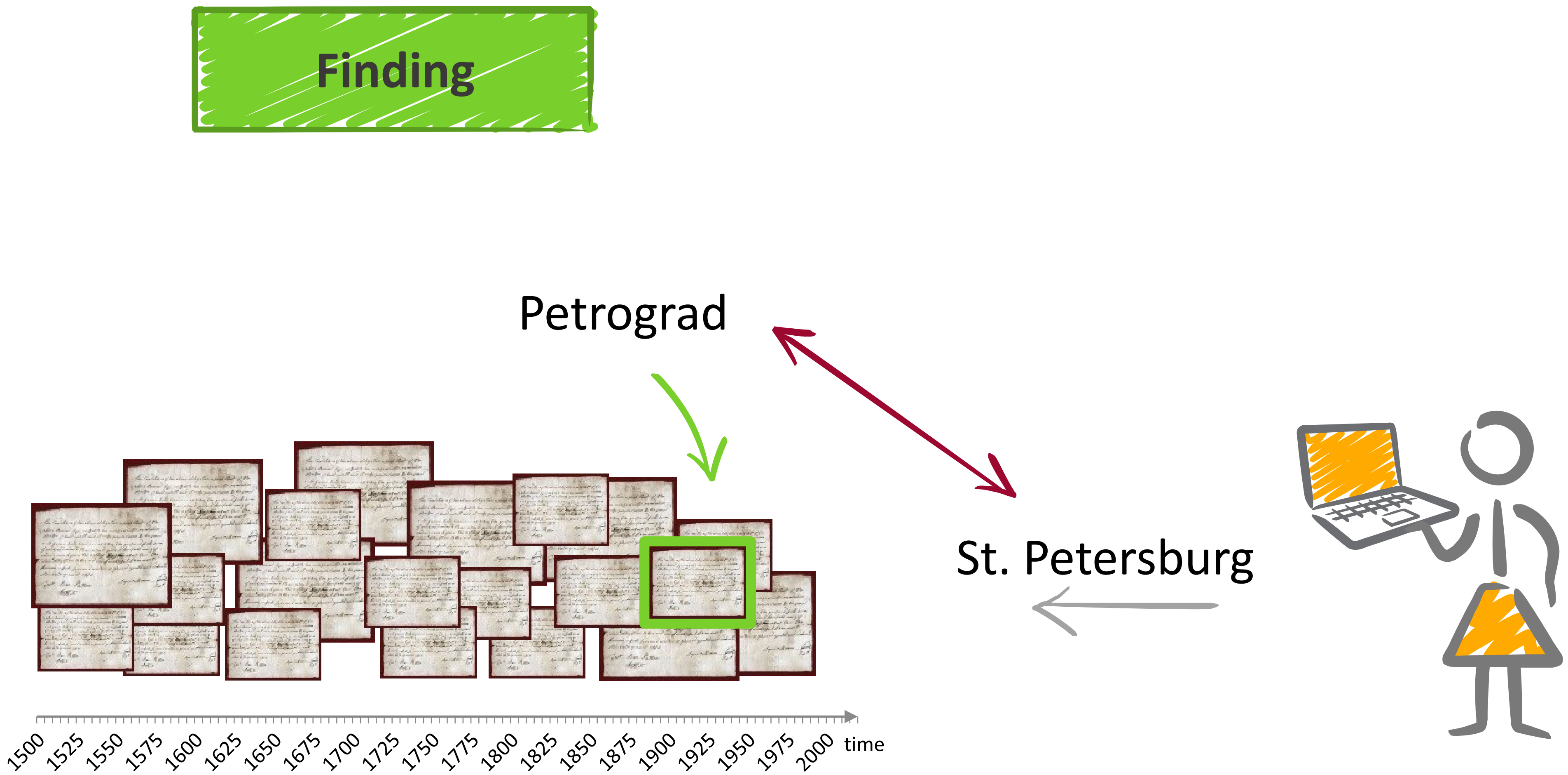
He was an **awesome** leader!

He was an **awesome** leader!

time

Kona → Qwinna → Qvinna → Kvinna

What is the problem?

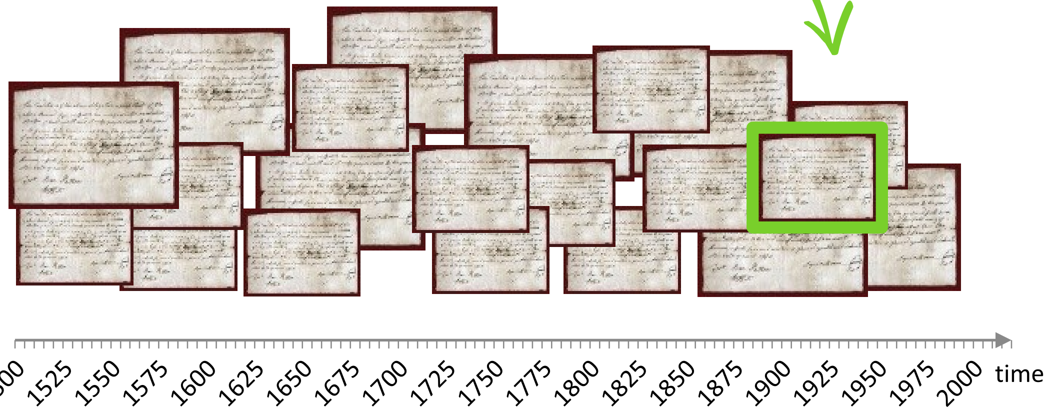


What is the problem?

Finding

Interpreting

Petrograd



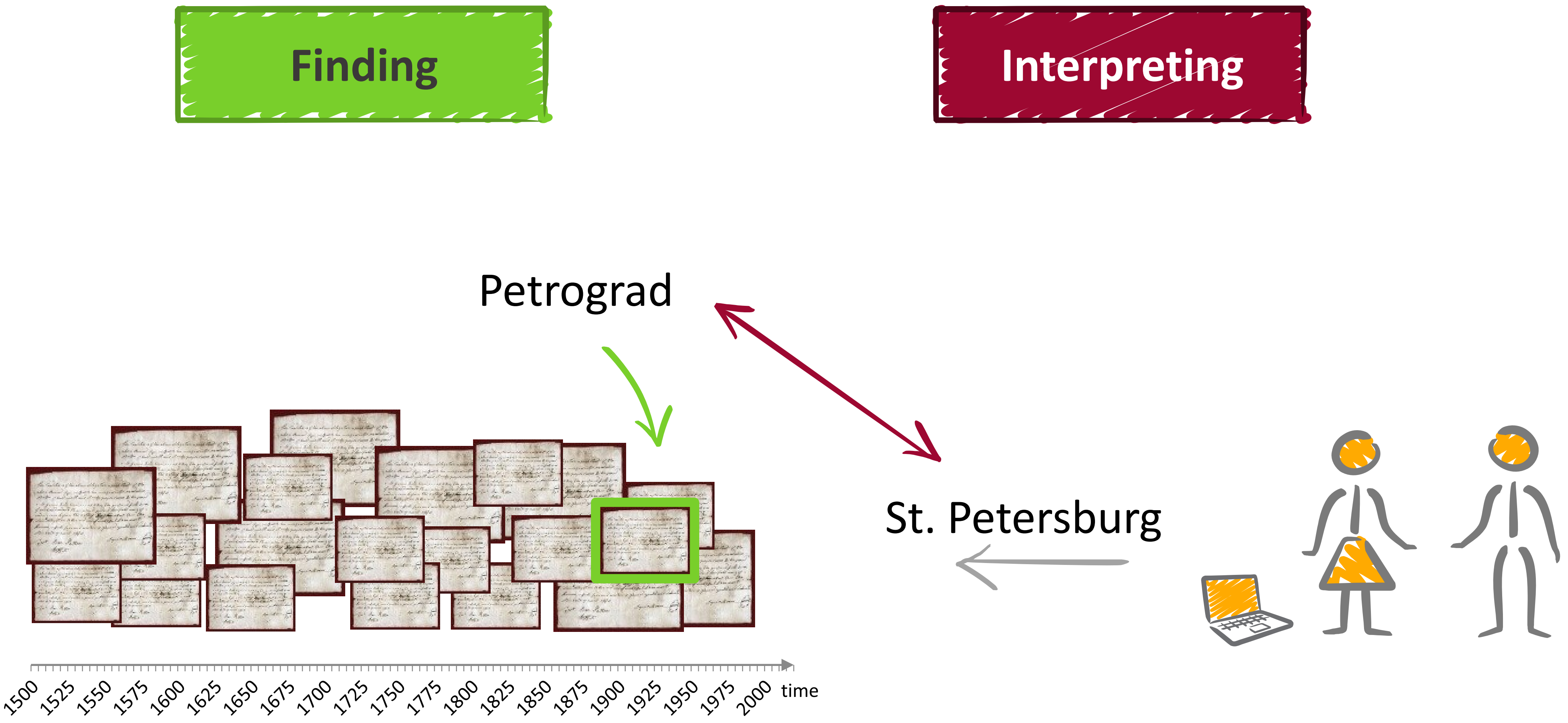
St. Petersburg

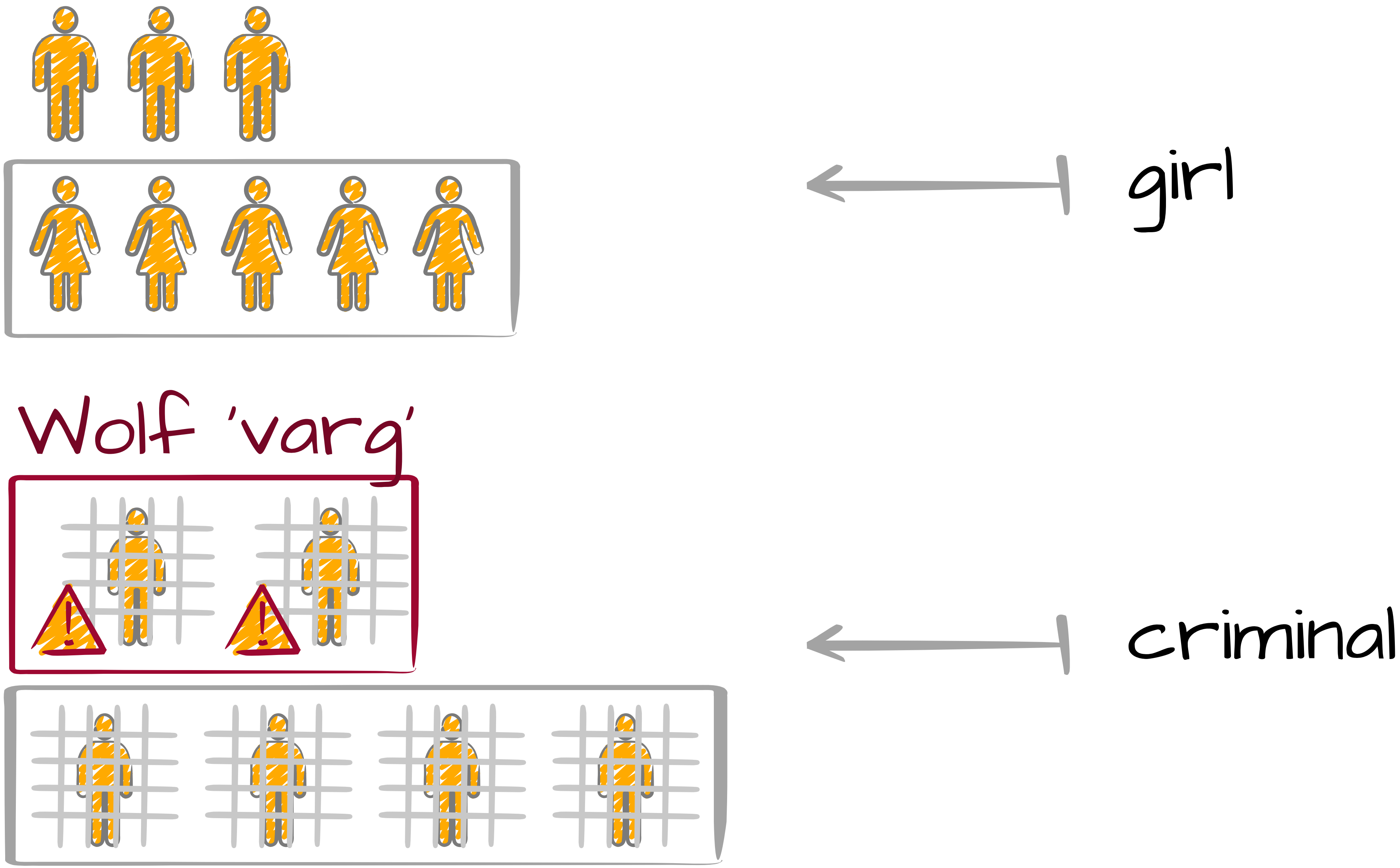
Sebastini's benefit last night at the Opera House was overflowing with the fashionable and **gay**

Sebastini's benefit last night at the Opera House was overflowing with the fashionable and **gay**

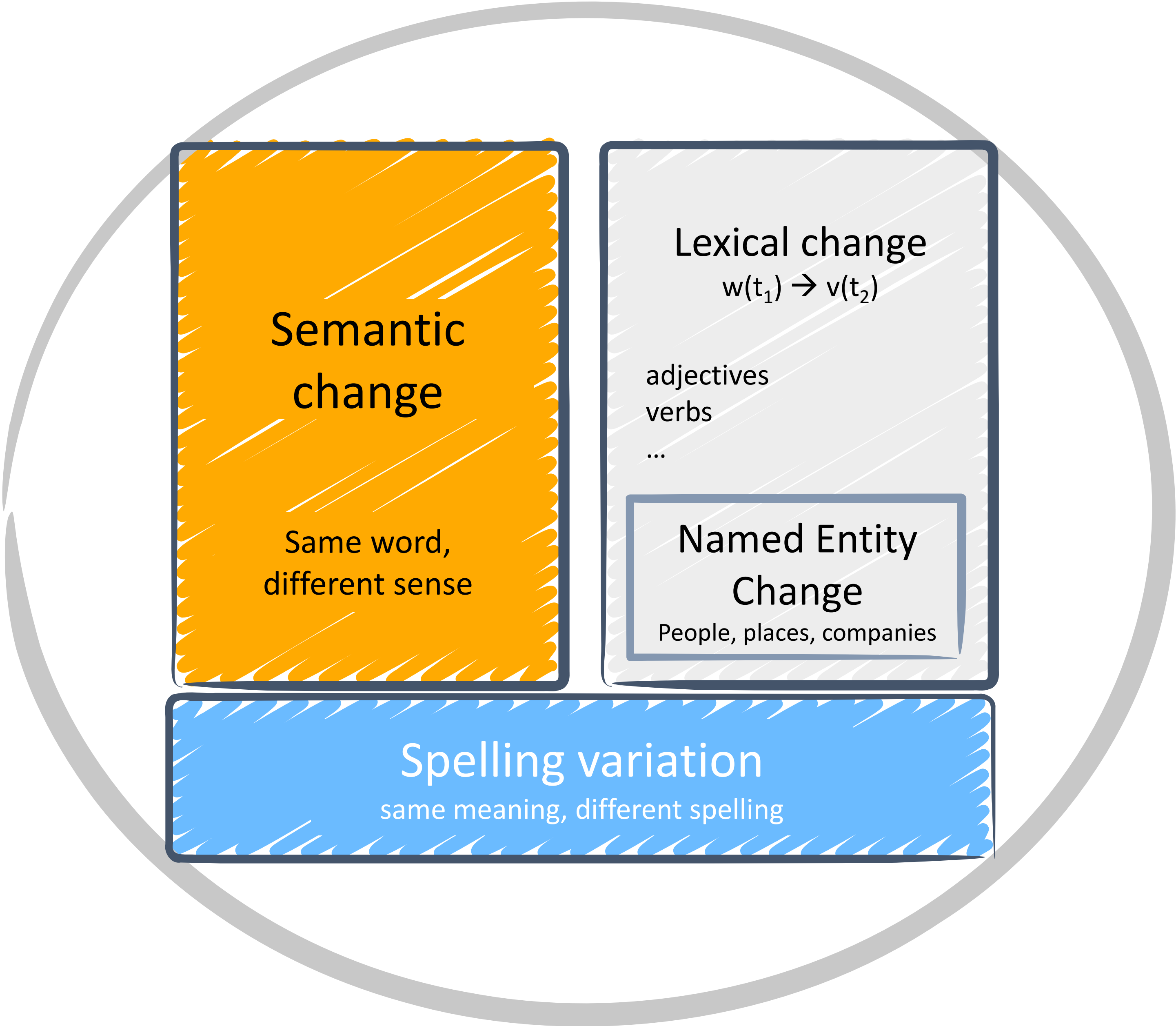
The Times, April 27th, 1787

What is the problem?





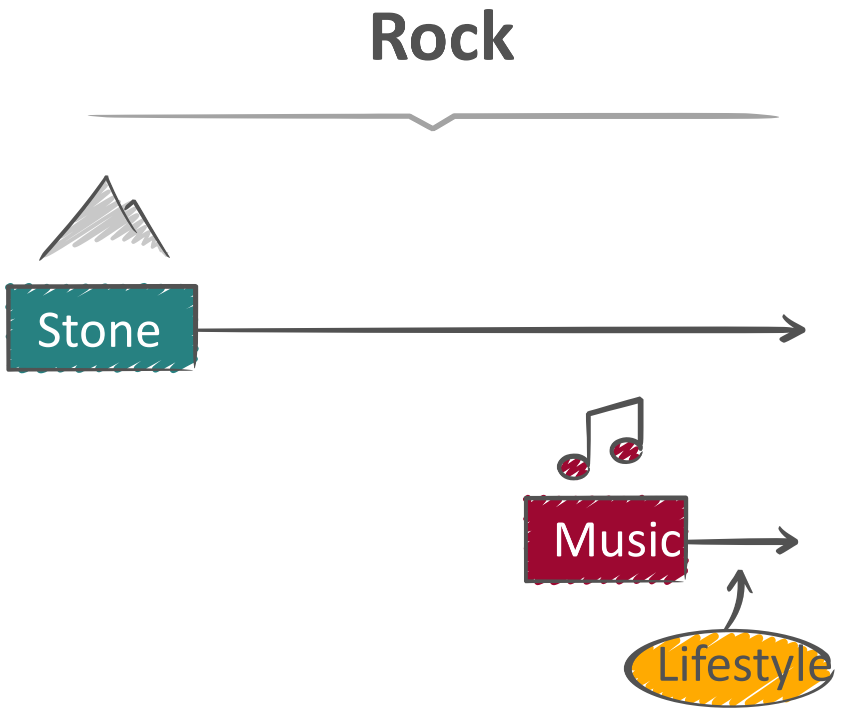
girl
Wolf 'varg'
criminal


Semantic change
Same word, different sense
Lexical change
$w(t_1) \rightarrow v(t_2)$
adjectives
verbs
…
Named Entity Change
People, places, companies
Spelling variation
same meaning, different spelling

Aims

To find word sense changes **automatically** by

1. Modeling word senses
2. Comparing these over time



To find **what** changes, **how** it changed and **when** it changed

Vision

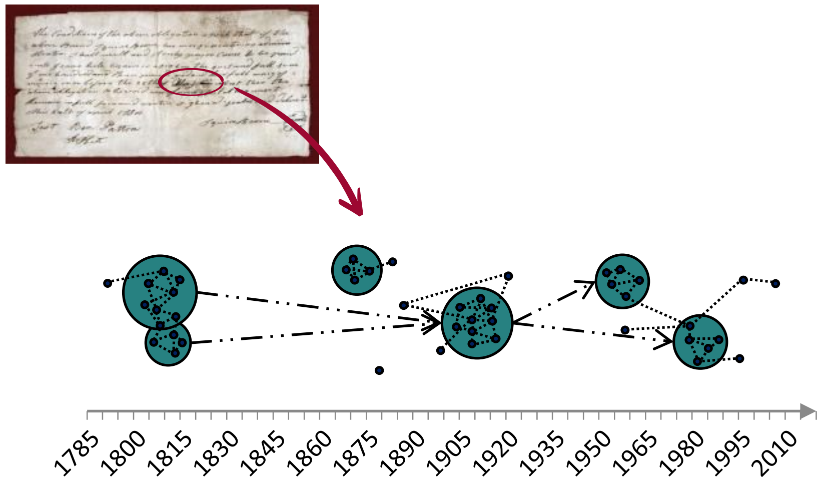
Given a word in a document at time *t*

1

Mark words that are likely to have a changed meaning

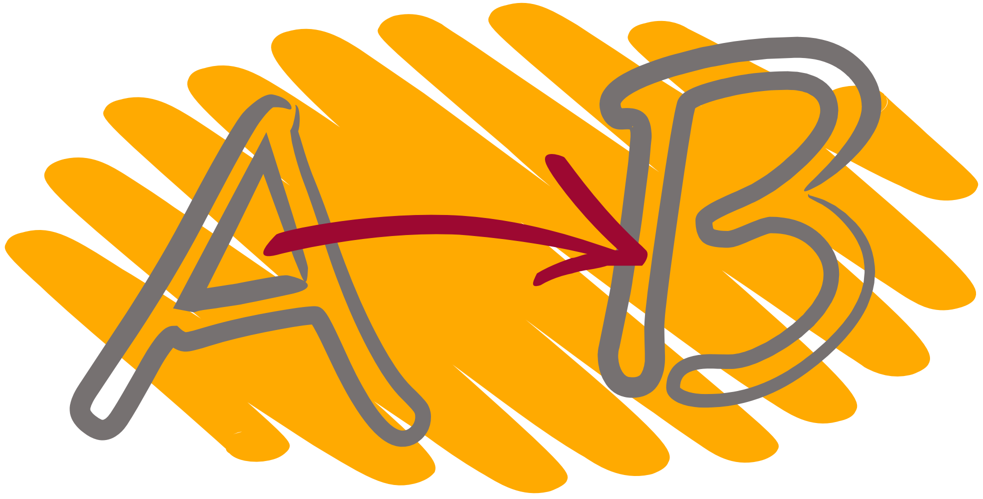
2

Find all changes to the word

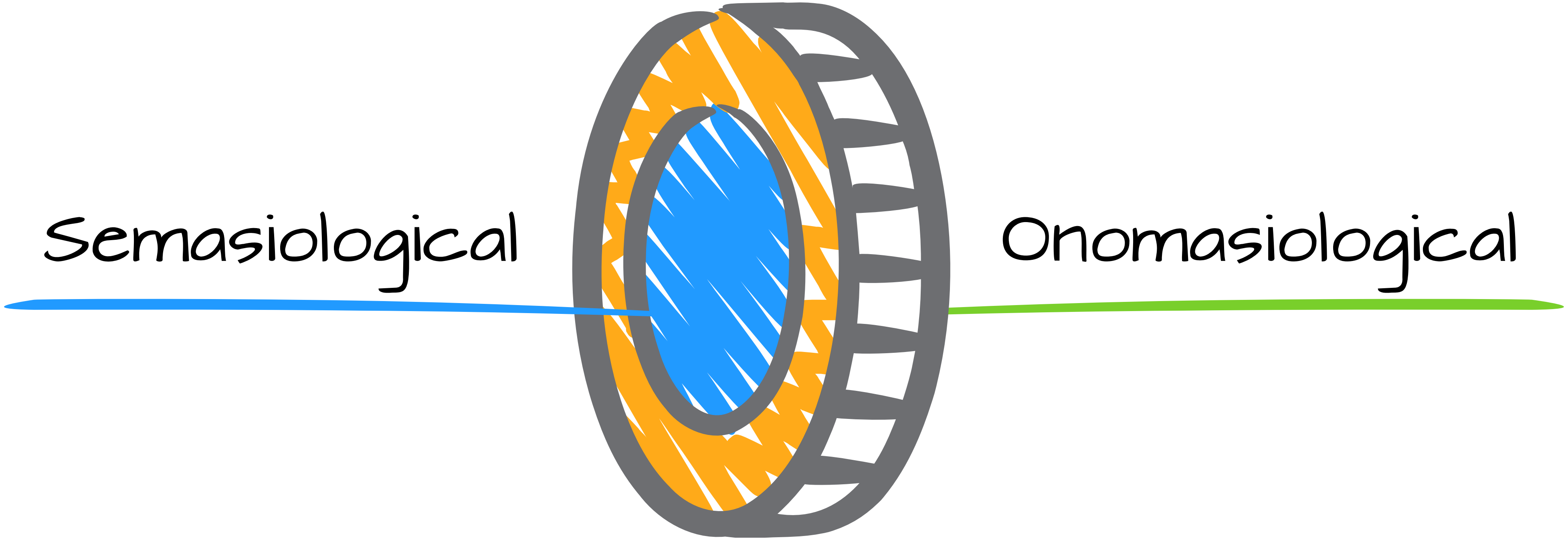


Lexical semantic change

The (historical) linguistic perspective




LSC


Semasiological
Onomasiological

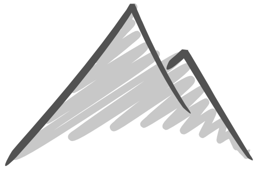
Semasiological perspective

word

concept/meaning

Rock

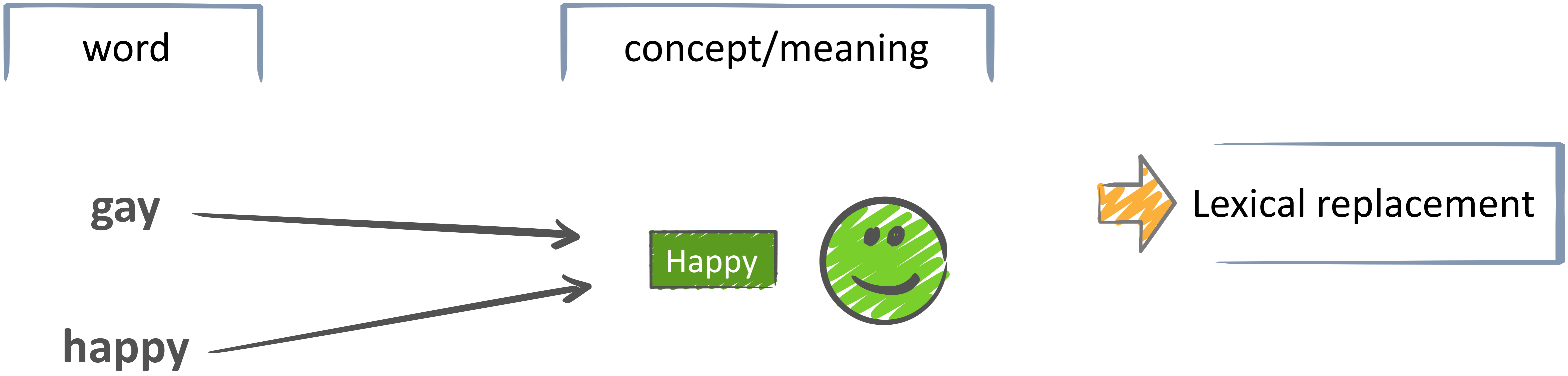
Stone



Music

Semantic change

Onomasiological perspective



Ono- and Semasiological are interlinked!

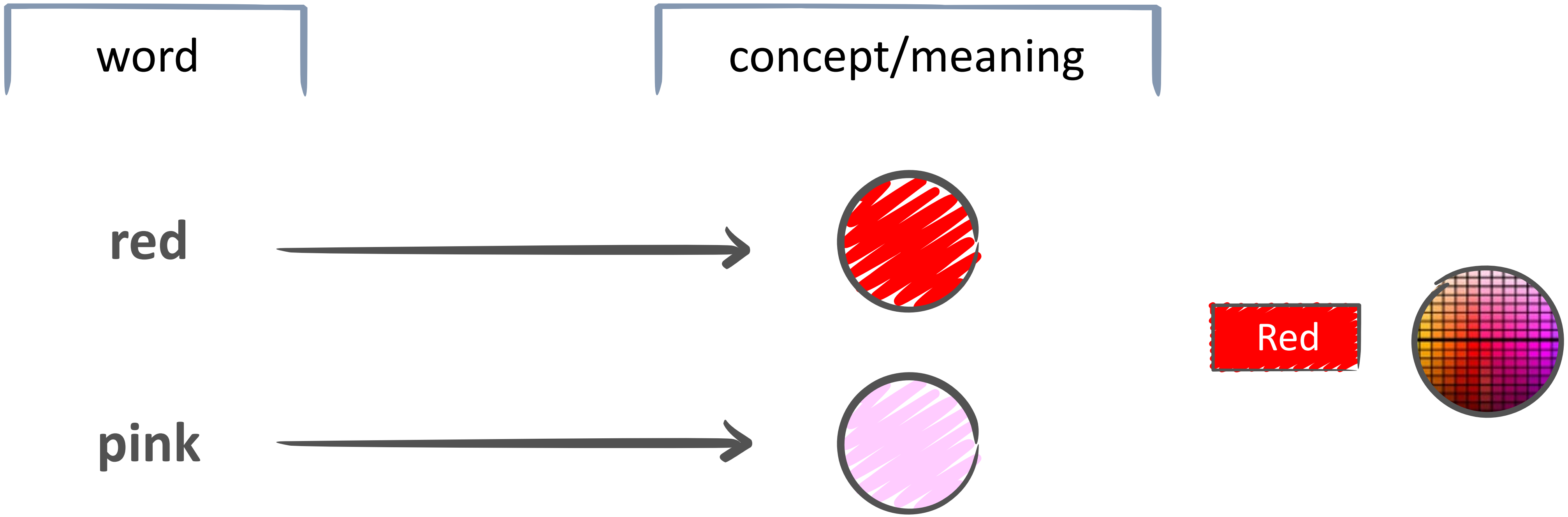
word

concept/meaning

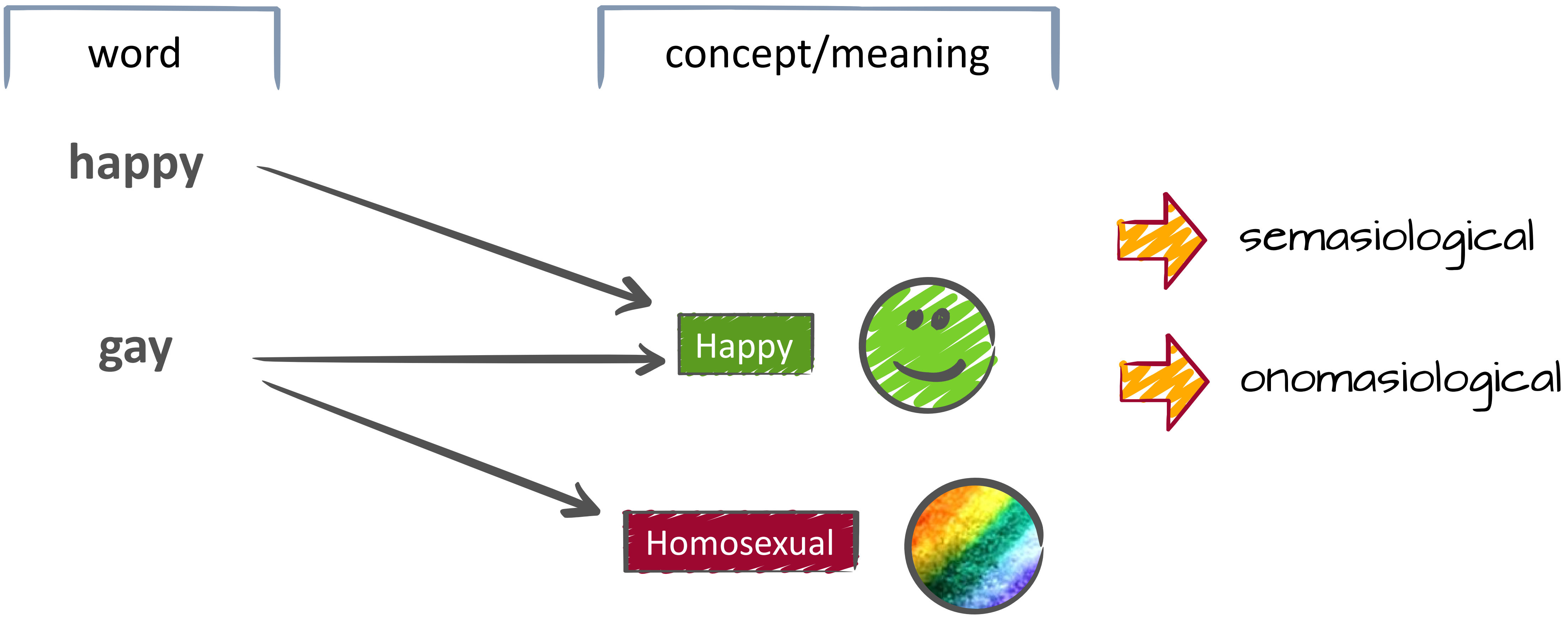
red

Red

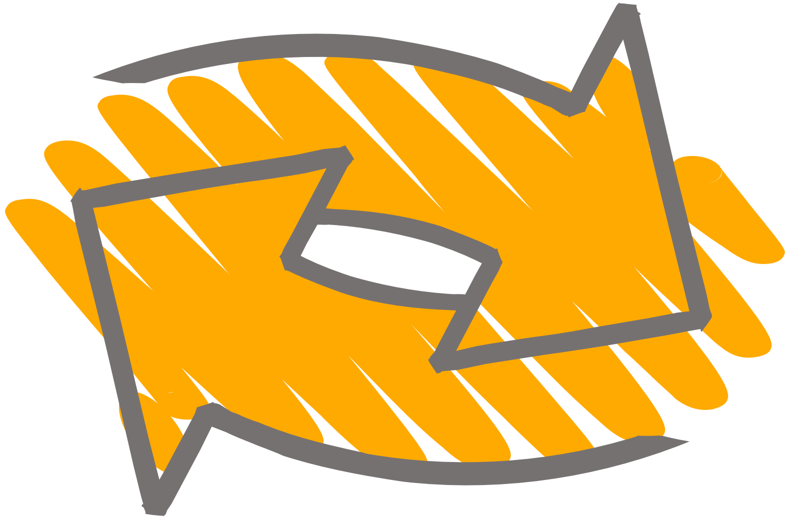
Ono- and Semasiological are interlinked!



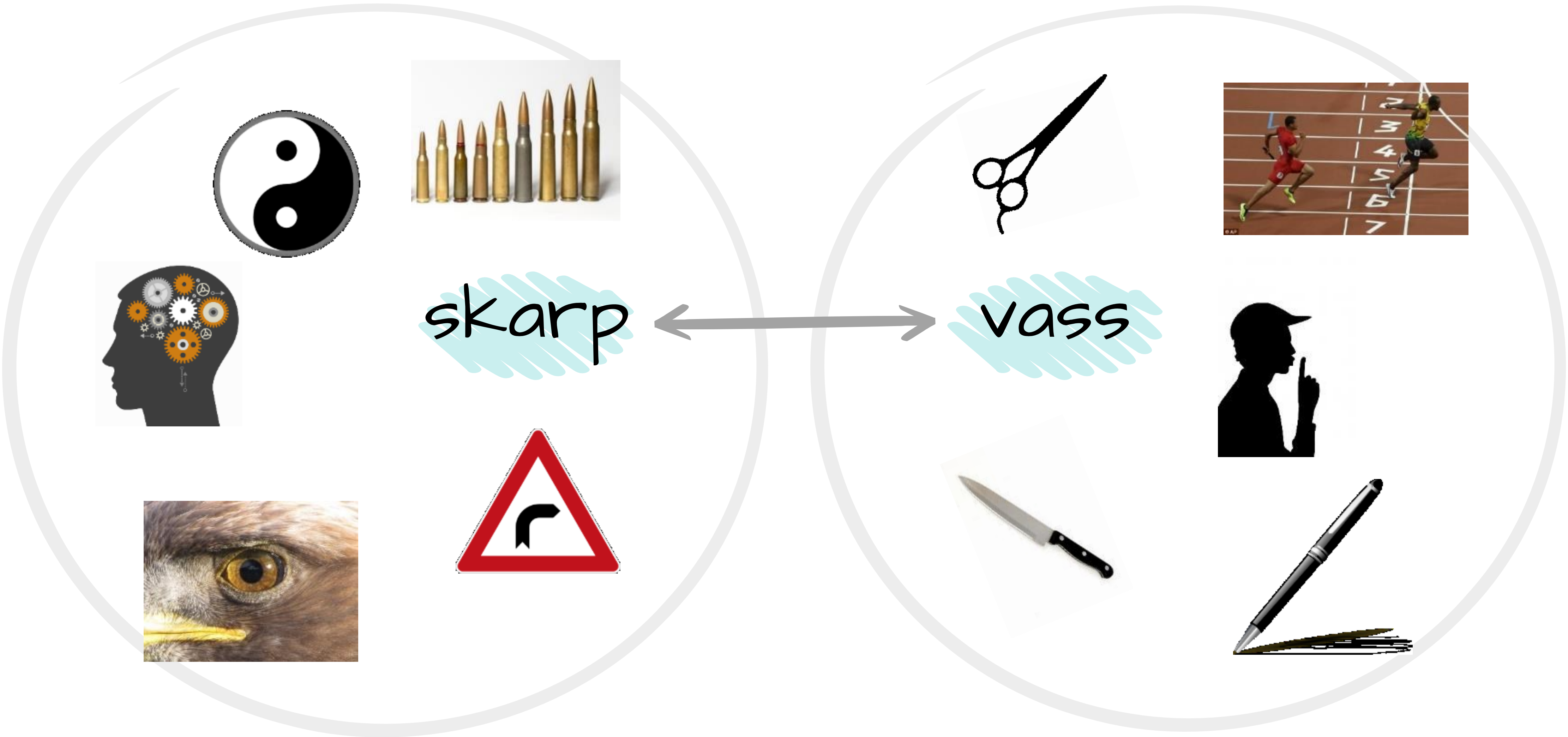
One more example



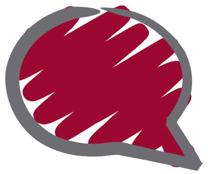
Why?



A division of the semantic field 'sharp'



skarp ⟷ vass



skarp (adjektiv)

Skarp Attribut		Adverbial skarp	
1. kritik	1008	1. lika	120
2. kontrast	822	2. mycket	227
3. ammunition	358	3. ganska	82
4. version	299	4. så	270
5. gräns	246	5. alltför	24
6. blick	213	6. liten	34
7. kritiker	141	7. föga	32
8. varning	151	8. riktig	75
9. kurva	125	9. osedvanlig	9
10. kant	77	10. tillräcklig	18
11. analys	89	11. i går	7
12. sväng	73	12. oväntad	10
13. bild	169	13. samtidig	8
14. protest	82	14. oerhörd	17
15. tillsägelse	34	15. uppknappt	4

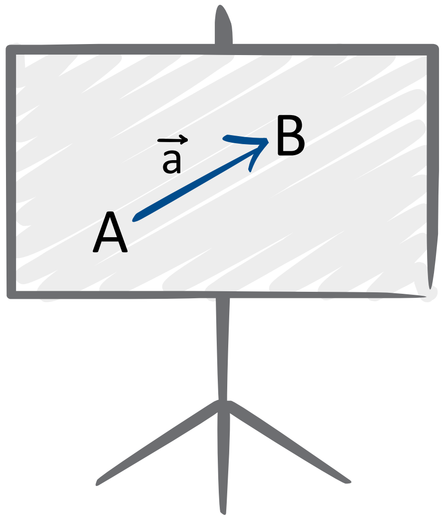
vass (adjektiv)

Vass Attribut		Adverbial vass	
1. kniv	803	1. riktig	529
2. spurt	200	2. lika	230
3. penna	114	3. tillräcklig	31
4. avslutning	118	4. jävligt	29
5. tunga²	79	5. jävlig	29
6. tunga	79	6. ruskig	14
7. avslutare	54	7. jäklig	14
8. kant	72	8. grön	5
9. egg	45	9. oerhörd	16
10. speed	36	10. ovanlig	11
11. slutspeed	26	11. ganska	41
12. sax	30	12. ruggig	5
13. satir	34	13. speciell	8
14. målskytt	44	14. onekligen	5
15. sten²	36	15. invändig	2

A corpus comparison: https://spraakbanken.gu.se/korp/

Methods for computational semantic change

Some terminology

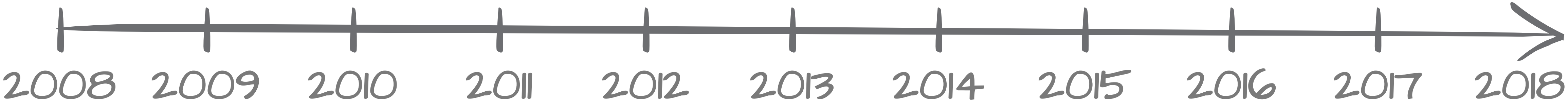


Vector (1, 4, 3) (=3 dimensions)

Topic modeling

- embeddings
- neural embeddings
- dynamic embeddings

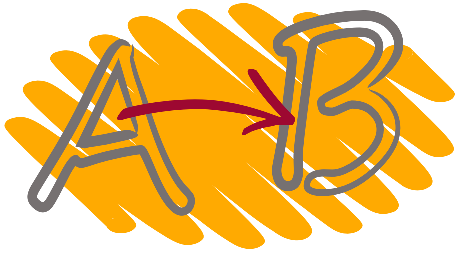
Single-sense



- topic models
- word sense induction

Sense-differentiated

Change type



Novel word

Novel word sense

Novel related ws

Novel unrelated ws

Death

Broadening

Narrowing

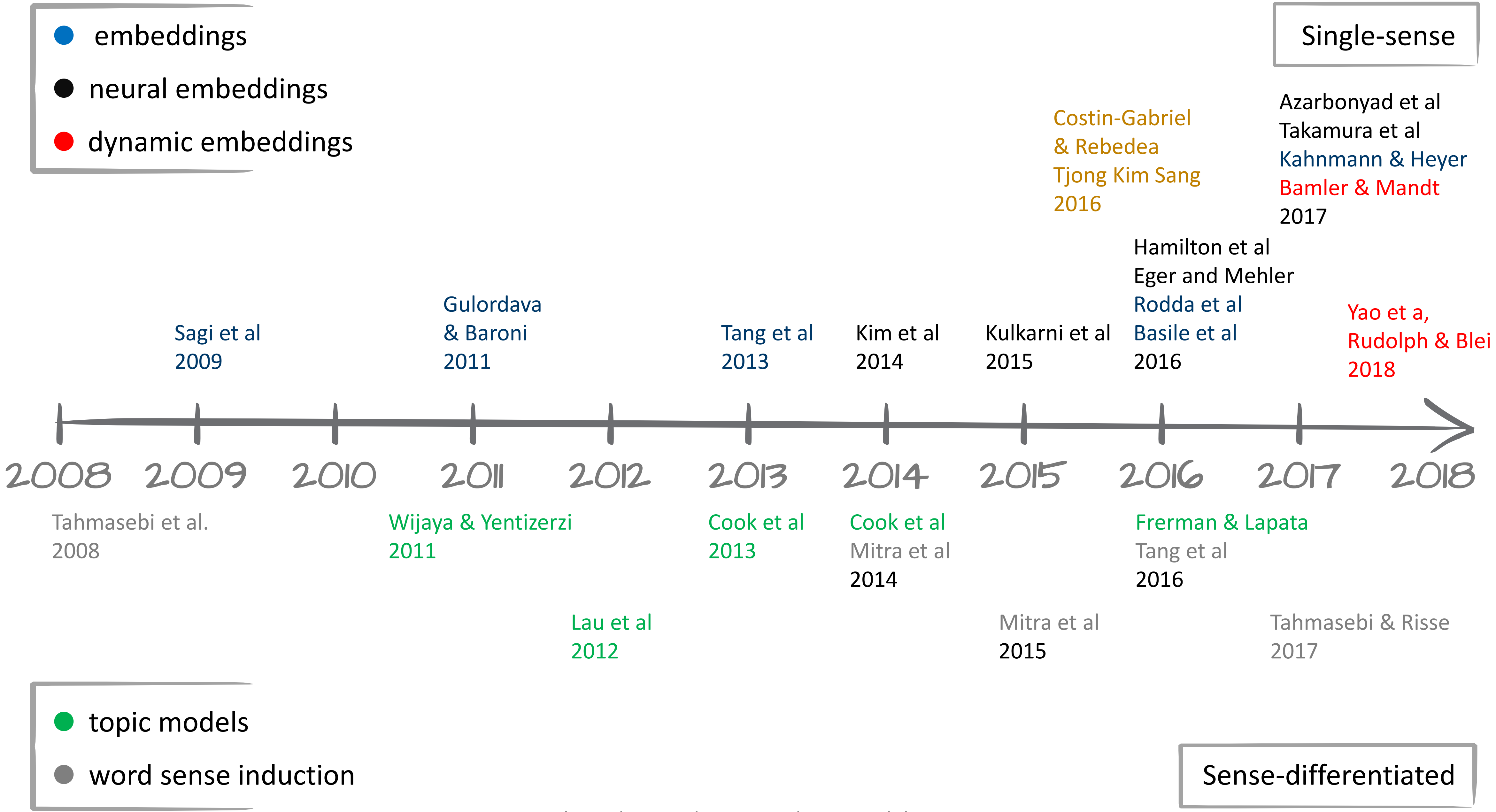
Join

Split

Change

Sense-differentiated

Single-sense


embeddings
neural embeddings
dynamic embeddings
Single-sense
Costin-Gabriel & Rebedea
Tjong Kim Sang
2016
Azarbonyad et al
Takamura et al
Kahnmann & Heyer
Bamler & Mandt
2017
Hamilton et al
Eger and Mehler
Rodda et al
Basile et al
2016
Sagi et al
2009
Gulordava & Baroni
2011
Tang et al
2013
Kim et al
2014
Kulkarni et al
2015
Yao et a,
Rudolph & Blei
2018
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
Tahmasebi et al.
2008
Wijaya & Yentizerzi
2011
Cook et al
2013
Cook et al
Mitra et al
2014
Frerman & Lapata
Tang et al
2016
Lau et al
2012
Mitra et al
2015
Tahmasebi & Risse
2017
topic models
word sense induction
Sense-differentiated

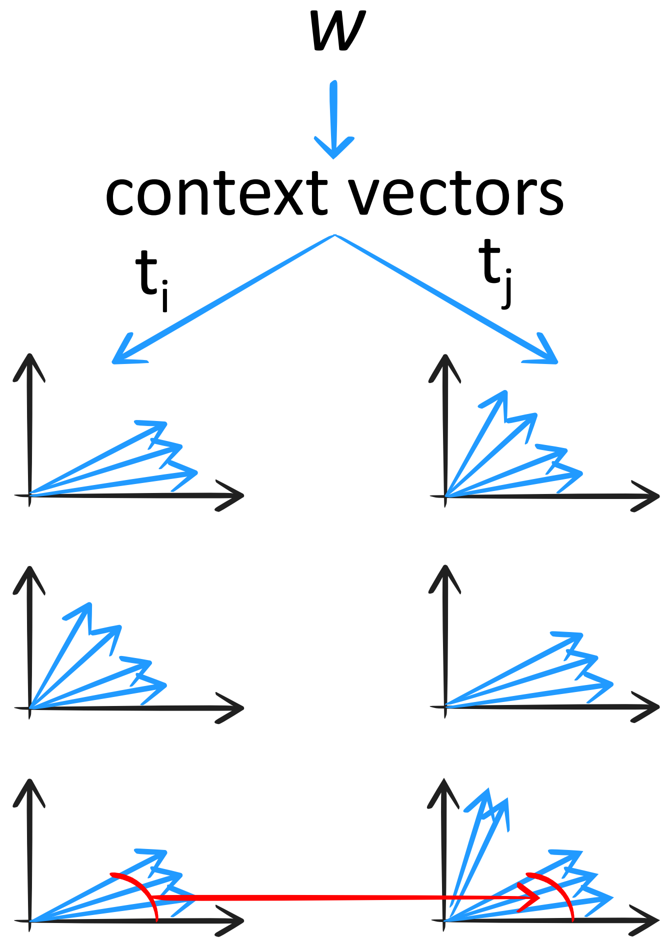
Outline

- embeddings / context-based methods
- neural embeddings
- dynamic embeddings

- topic models
- word sense induction

Context-based method

Sagi et al.
GEMS 2009



Data set split in approp. sets

Broadening of sense

Narrowing of sense

With grouping:
Added/removed sense

BUT: 1. No discrimination between senses **2.** No alignment of senses over time!

Word embedding-based models

Kulkarni et al. WWW'15

- Project a word onto a vector/point (POS, frequency and embeddings)
- Track vectors over time

Kim et al. LACSS 2014

Basile et al. CLiC-it 2016

Hamilton et al. ACL 2016

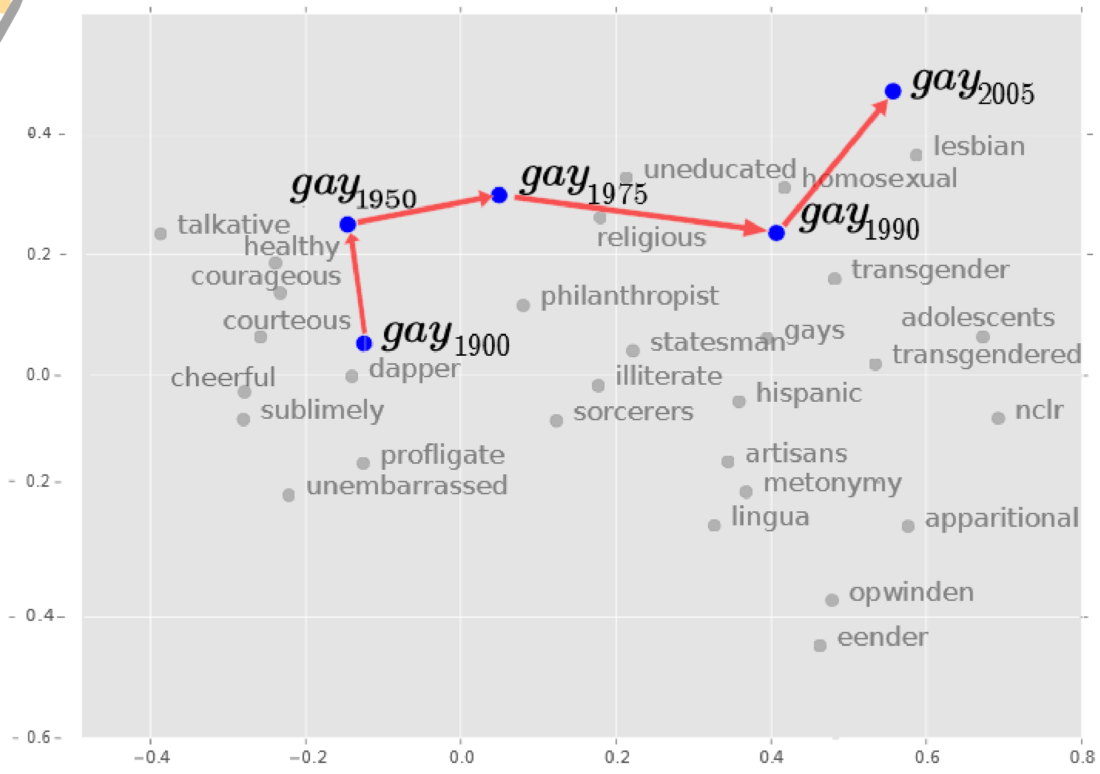


Image: Kulkarni et al. WWW'15

Dynamic Embeddings

Share data across all time points

Avoids aligning

Bamler & Mandt:

- Bayesian Skip-gram

Yao et al:

- PPMI embeddings

Rudolph & Blei:

- Exponential family embeddings (Beronoulli embeddings)

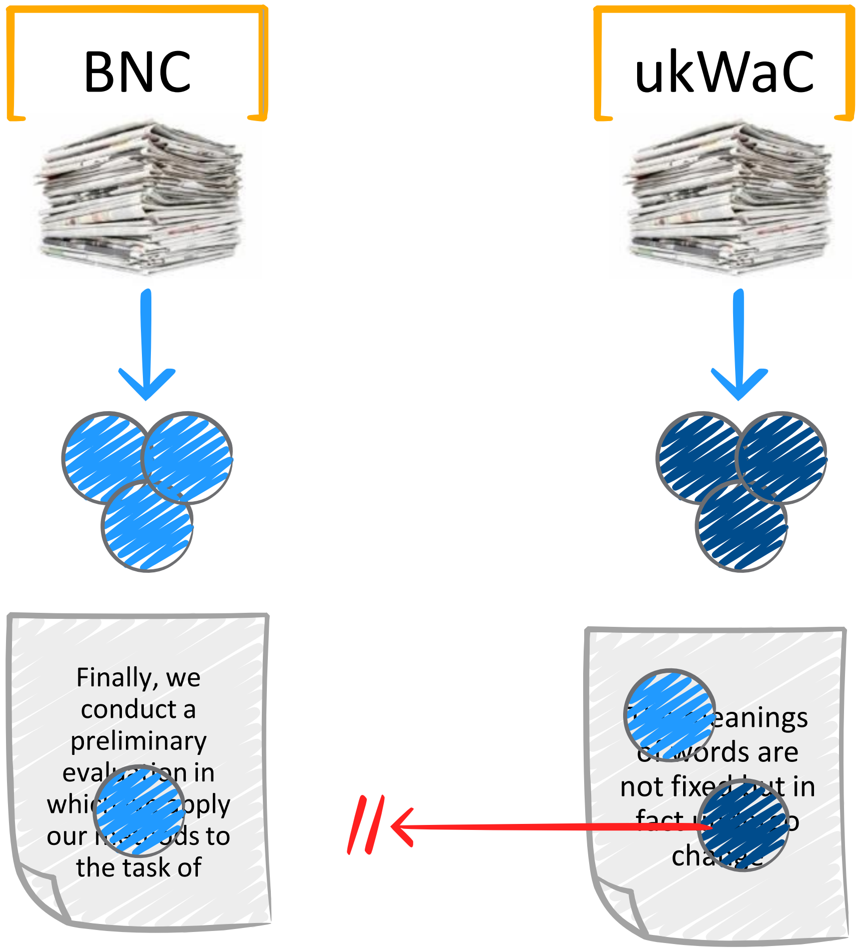
Sharing data is **highly beneficial!**

Topic-based methods

1. Topic model (HDP)
2. Assign topics to all instances of a word.
3. If a word sense WS_i is assigned to collection 2 but not 1 then WS_i is a **novel** word sense.

BUT:

A. Only two time points (typically there is much noise!)
B. No alignment of senses over time!



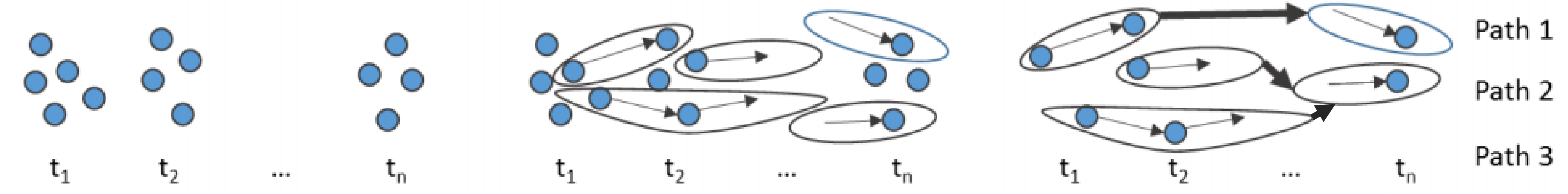
Lau et al.
EACL 2014

Wijaya & Yeniterzi
DETECT '11

Cook et al.
Coling 2014

Fermann & Lapata
TACL 2016

Word sense induction



Step 1:

Word sense discr.

(curvature clustering) individual time slices

Step 2:

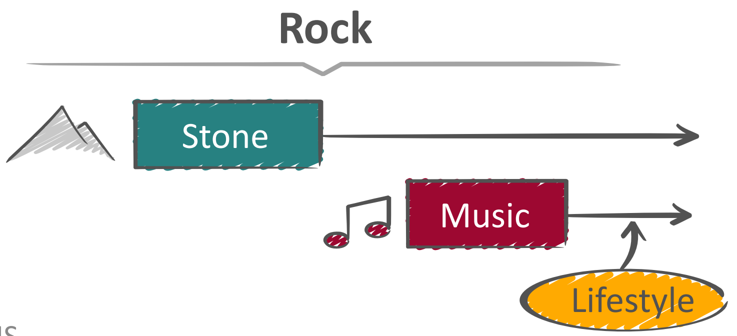
Detecting stable senses

→ units

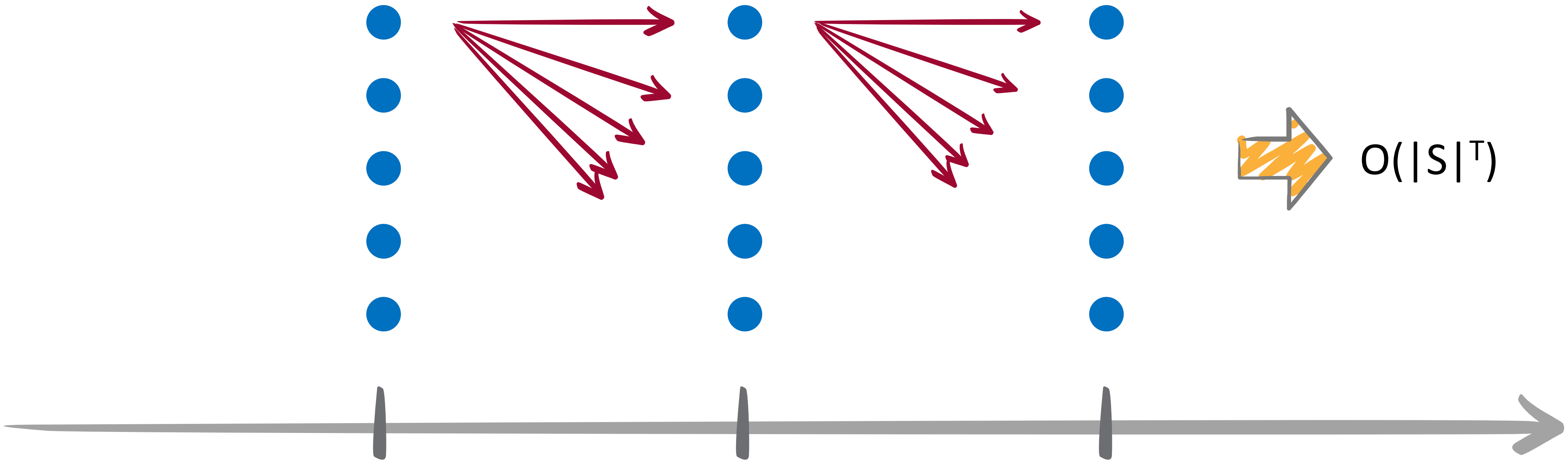
Step 3:

Relating units

→ Paths



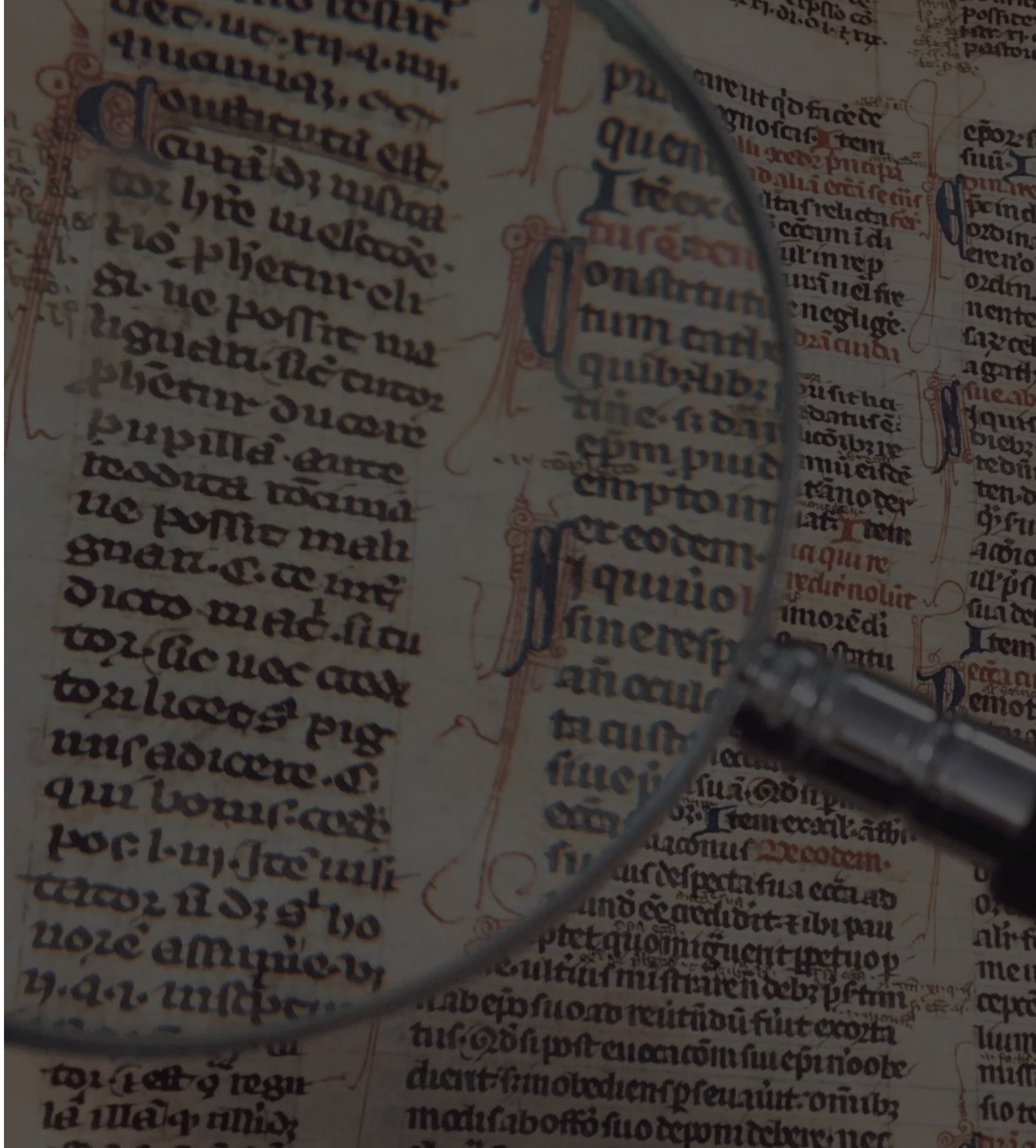
Complexity



$O(|S|^T)$

How?

Evaluation



NLP pipeline: From text to result

Text-mining method

Dimensions

Filtering: Function words

Filtering: Stopwords

Part-of-speech tagging

Lemmatization

Tokenization

like (only verbs)

room (frequency filtering)

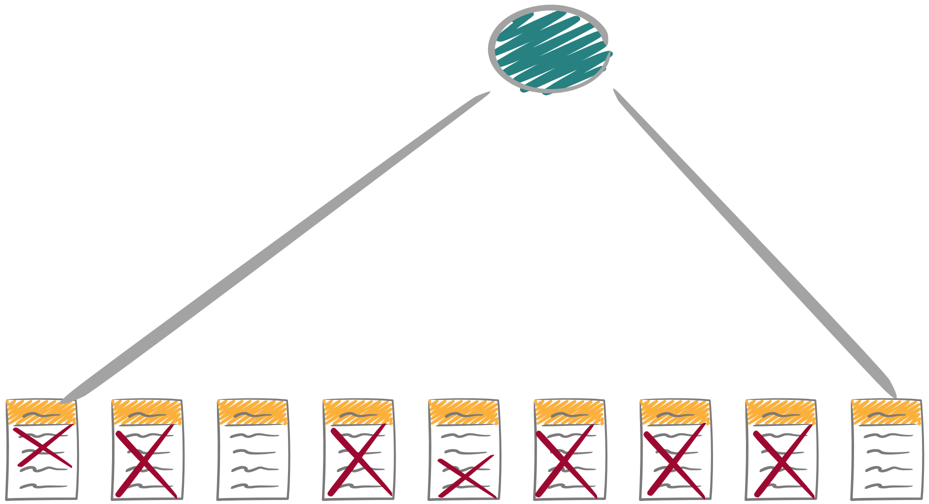
room sheet. (only nouns)

I like room sheet. (after lemmatization)

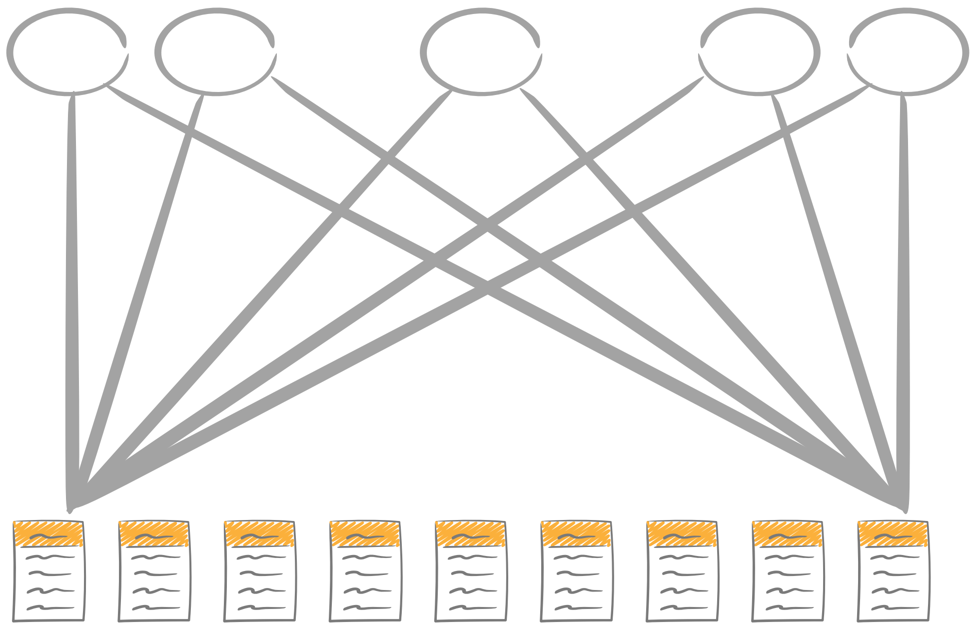
I like room sheets. (after stop word filtering)

I like the room but not the sheets.

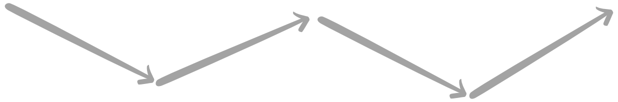
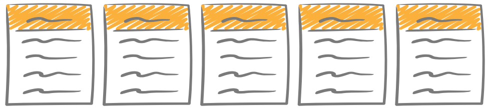
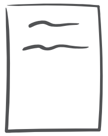
Viewpoint on the data



Viewpoint on the data (cont'd)



Evaluation

		minimum	optimum	medium
	signal change			
	signal topic, cluster, vector…			
	collective text			
	individual text			
	individual			

Evaluation

	minimum	optimum	medium
signal change			
collective text			

Evaluation

	minimum	optimum	medium
signal change			
individual text			
individual			?

Evaluation

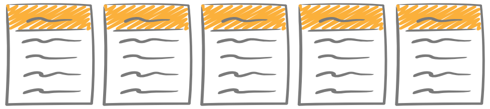
minimum

optimum

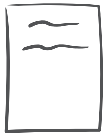
medium

signal

topic, cluster, vector…

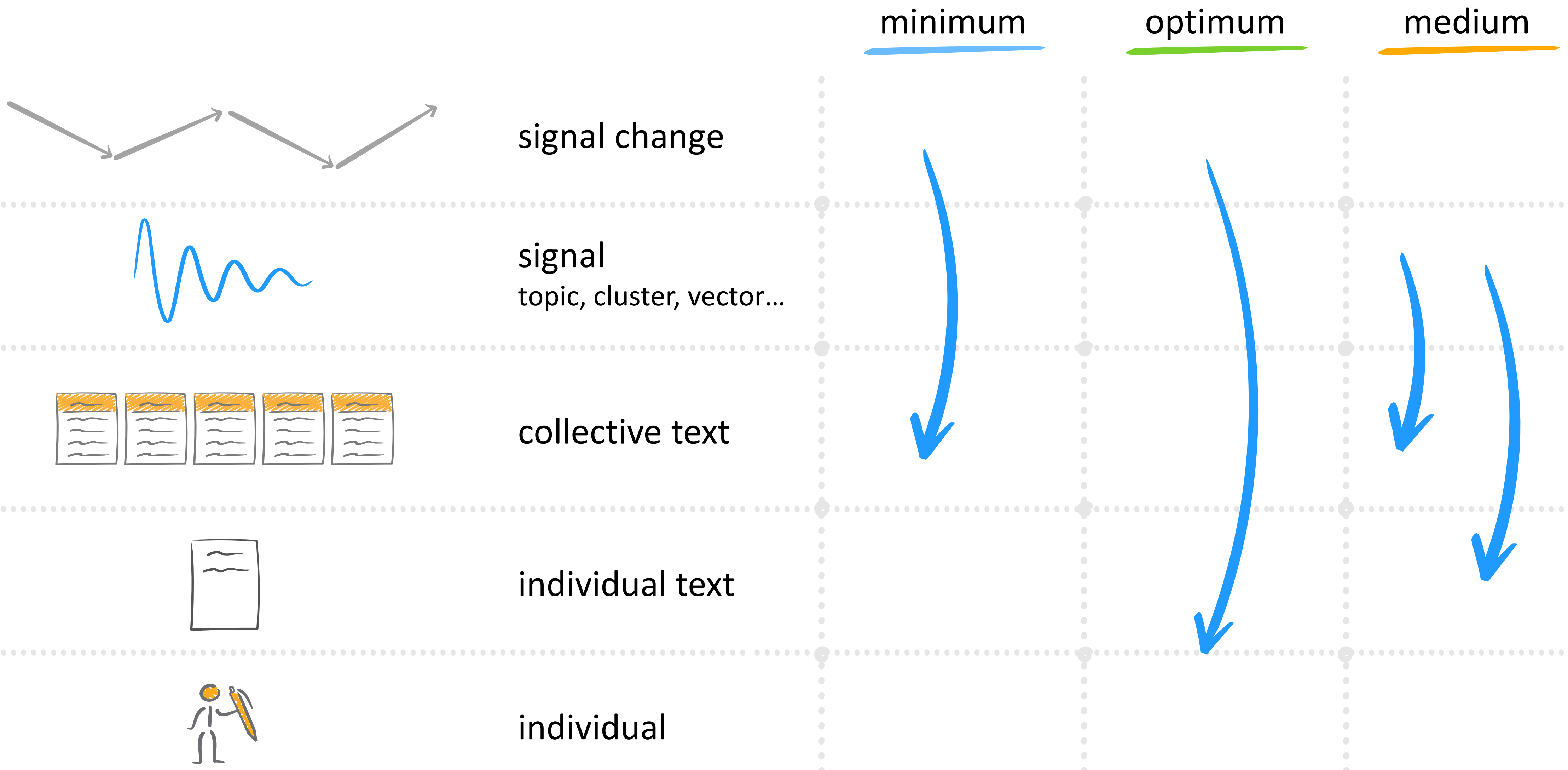


collective text



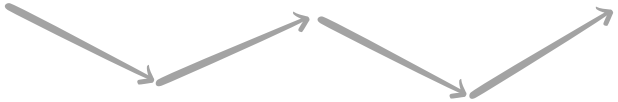
individual text

Evaluation



minimum
optimum
medium
signal change
signal
topic, cluster, vector…
collective text
individual text
individual

Evaluation

		minimum	optimum	medium
	signal change			

Evaluation

Top/bottom results

3 ways

Pre-determined list of
- Positive examples
- Negative examples

Controlled data

	prechosen # pos	prechosen # neg	top	entity (S)ingle/ (P)airs	eval. method (M)anual/ (A)utomatic	time span	time # points	# classes	classes	modes time aware	modes sense diff
Sagi, Kaufmann, and Clark (2009a)	4	0		S	M	569y	4	2	broad./narrow.	no	no
Gulordava and Baroni (2011)	0	0	100[54]	S	M	40y	2	1	change	no	no
Tang, Qu, and Chen (2013)	33	12		S	M	59	59	3	B/N/novel/change[55]	no	no
Kim et al. (2014)	0	0	10/10[56]	S/P[57]	M	110	110	1	change	yes[58]	no
Kulkarni et al. (2015)	20	0	20[59]	S	M/A	105y/12y/2y	21/13/24	1	change	yes	no
Hamilton, Leskovec, and Jurafsky (2016b)	28	0	10[60]	S/P	M	200/190	20	1	change	no	no
Rodda, Senaldi, and Lenci (2016)	0	0	50	S	M	1200y	2	1	change	no	no
Eger and Mehler (2016)	0	0	21[61]	S/P	M	200/190	20/19	1	change	no	no
Basile et al. (2016)	40	0		S	M	170	17	1	change	yes	no
Azarbonyad et al. (2017)	24	0	5/5[62]	S	M	20/11	2/2	1	change	no	no
Takamura, Nagata, and Kawasaki (2017)	10	0	100/20[63]	S/P	M	-[64]	2	1	change	no	no
Kahmann, Niekler, and Heyer (2017)	4	0		S	M	≤ 1[65]	48	1[66]	change	no	no
Bamler and Mandt (2017)	6	0	10	S/P	M[67]	209/230/7	209/230/21	1	change	no	no
Yao et al. (2018)	4/1888[68]	0		S	M/A	27	27	1	change	no	no
Wijaya and Yeniterzi (2011)	4	2		S	M	500[69]	500	2[70]	change novel	yes	yes[71]
Lau et al. (2012)	5	5		S	M	43 y	2	1	novel	no	yes
Cook et al. (2013)	0	0	30	S	M	14	2	1	novel	no	yes
Cook et al. (2014)	7/13	50/164		S	M	43y/17y	2/2	1	novel	no	yes
Mitra et al. (2015)[72]	0	0	69/50	S	M/A	488/2	8/2	3	split/join/novel[73]	no	yes
Frermann and Lapata (2016)	4	0	200	S	M/A	311	16	2	change/novel	no	yes
Tang, Qu, and Chen (2016)[74]	197	0		S	M	59	59	6	B/N/novel/change[75]	no	yes
Tahmasebi and Risse (2017a)	35	25		S	M	222y	221	4	novel,B/N,stable	yes	yes

https://languagechange.org/publication/2018-surveypaper/

Data sets

Table 3
Datasets used for diachronic conceptual change detection. Non-English ·

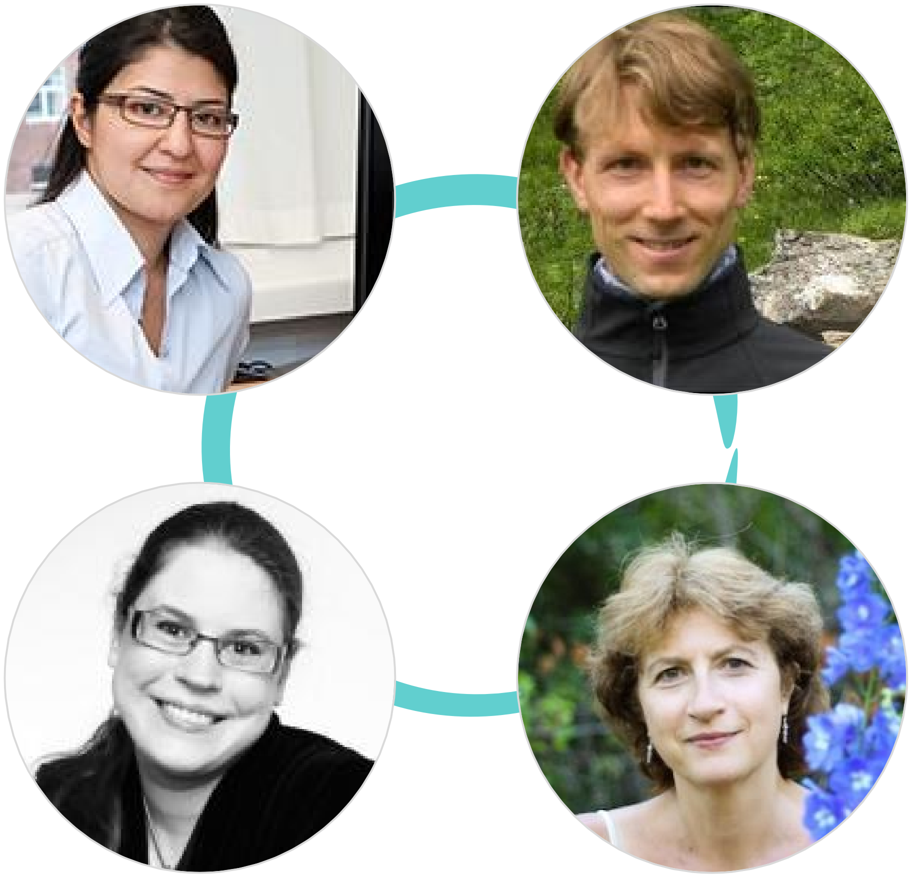
Sagi, Kaufmann, and Clark (2009a)	Helsinki corpus
Gulordava and Baroni (2011)	Google Ngram
Wijaya and Yeniterzi (2011)	Google Ngram
Lau et al. (2012)	British National Corpus (BNC), ukWaC
Cook et al. (2013)	Gigawords corpus
Cook et al. (2014)	BNC, ukWaC, Sibol/Port
Mihalcea and Nastase (2012)	Google books
· Basile et al. (2016)	Google Ngram (Italian)
· Tang, Qu, and Chen (2013, 2016)	Chinese People's Daily
Kim et al. (2014)	Google Ngram
Kulkarni et al. (2015)	Google Ngram, Twitter, Amazon movie reviews
Mitra et al. (2015)	Google Ngram, Twitter
Hamilton, Leskovec, and Jurafsky (2016b)	COHA, Google Ngram
· Eger and Mehler (2016)	COHA, Süddeutsche Zeitung, PL[76]
Azarbonyad et al. (2017)	New York Times Annotated Corpus, Hansard
· Rodda, Senaldi, and Lenci (2016)	Thesaurus Linguae Graecae
Frermann and Lapata (2016)	DATE corpus
Takamura, Nagata, and Kawasaki (2017)	Wikipedia (English and Japanese)
Kahmann, Niekler, and Heyer (2017)	Guardian (non-public)
Tahmasebi and Risse (2017a)	Times Archive, New York Times Annotated Corpus
Bamler and Mandt (2017)	Google Ngram, State of the Union addresses, Twitter
Yao et al. (2018)	New York Times (non-public)
Rudolph and Blei (2018)	ACM abstracts, ML papers ArXiv, U.S. Senate speech

Towards automatic language change detection

VR funded

6 million sek (+ cofunding Språkbanken ~700k sek)
2019 – 2022

4 year project: https://languagechange.org/



Overall goal is to bridge the gap between the four of us and all that can benefit from the results.

Main goals

Wp1: Swedish word sense induction

- Using sense-differentiated dynamic embeddings

Wp2: Semantic change

- On the basis of Wp1

Wp3: Lexical replacements

- On the basis of Wp1
- Or using other textual clues

Wp4: Applications

- Applied sociology, historical linguistics, history of concepts, …

WP*: Evaluation

- Integrated in all work packages

Planned activities

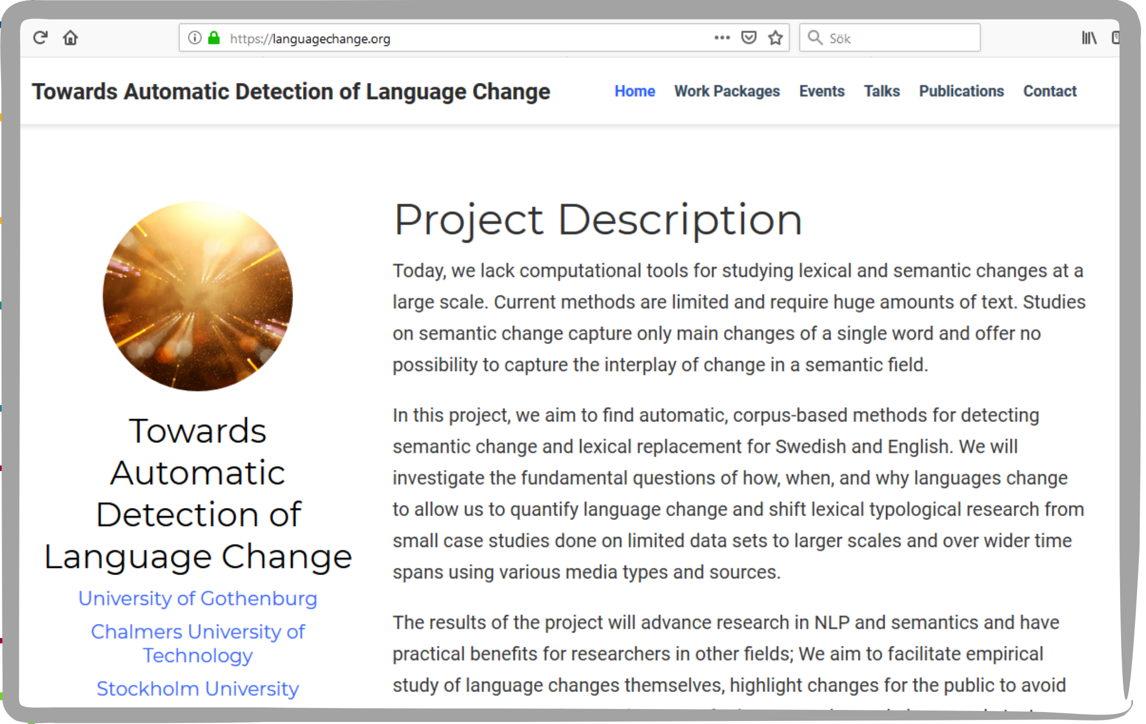
News-list (news@languagechange.org)

Introductory videos to LS change

Workshops (next at ACL2019)

Work on evaluation
(possibly in a SemEval task)

Talks (Stuttgart / Frankfuhrt spring 2019)



Project timeline

SC = Semantic Change
LR = Lexical Replacement
DH = Digital Humanities
SS = Social Science

2019

Feb – Workshop Helsinki
August – ACL workshop
September – Project Meeting

Word sense induction
Evaluation of SC
Hypotheses from historical linguistics

2020

SLTC workshop?
DHSS conf?

Manual study of SC
Annotation for SC
Lexical Replacement

2021

*ACL workshop?

Manual LR-study
Collaboration with DH and SS

2022

Final project event

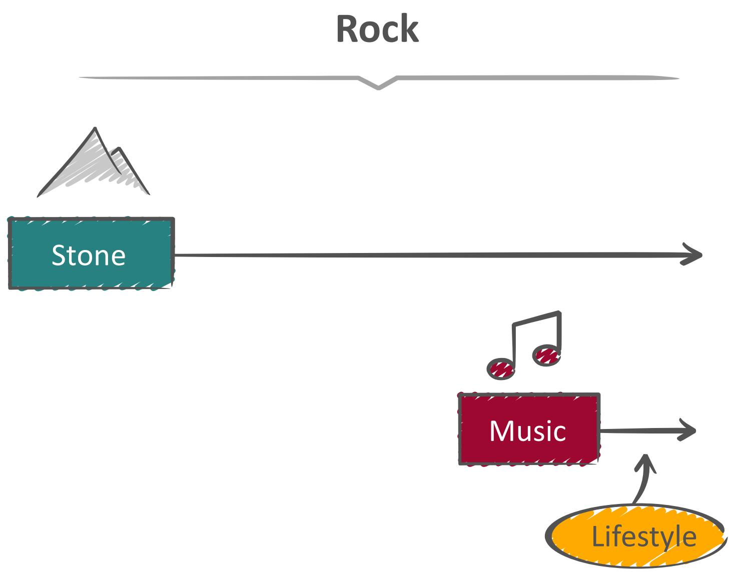
Semantic fields
LR-SC interchange

Conclusions

Complexity in
- Multiple senses
- Many time points

Not all data are big data!

Evaluation
→ Common datasets and methods!
→What is the result valid for?



Thank you for listening!

Nina.tahmasebi@gu.se

nina@tahmasebi.se