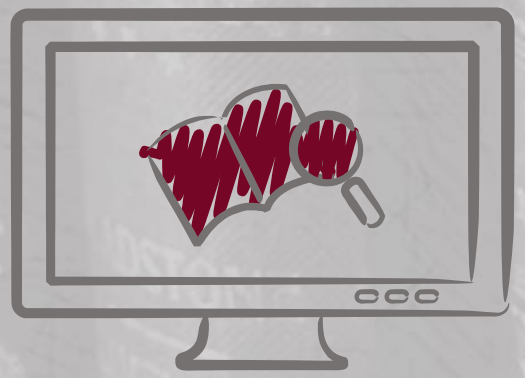
SPRÅKBANKENTEXT

RIKSBANKENS JUBILEUMSFOND
THE SWEDISH FOUNDATION FOR HUMANITIES AND SOCIAL SCIENCES

GÖTEBORGS UNIVERSITET



# Change is Key! An introduction to lexical semantic change

Nina Tahmasebi, Associate Professor & Simon Hengchen, Phd

University of Gothenburg

October 2022, KBR

**Digital Heritage Seminar Series: Lexical Semantic Change**

- 6 years
- 6 partner universities
- Members from 4 countries
- With advisors, 6 countries
- 13 people including PM and SE

KU LEUVEN

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung

Queen Mary
University of London

GÖTEBORGS UNIVERSITET

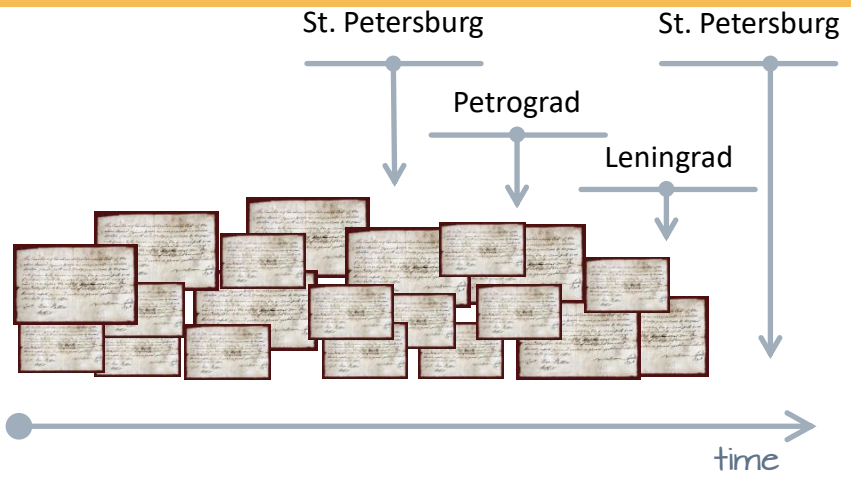
LUNDS UNIVERSITET

LINKÖPING UNIVERSITY


he asked, glancing at
at drop parritch.
was his own supper.
I can do fine wanting
ckens my cough.' He drank
an eye upon me as he drank
Let's see the letter,' said he
letter was for Mr Balfour; not for
think I am?' says he. 'Give me

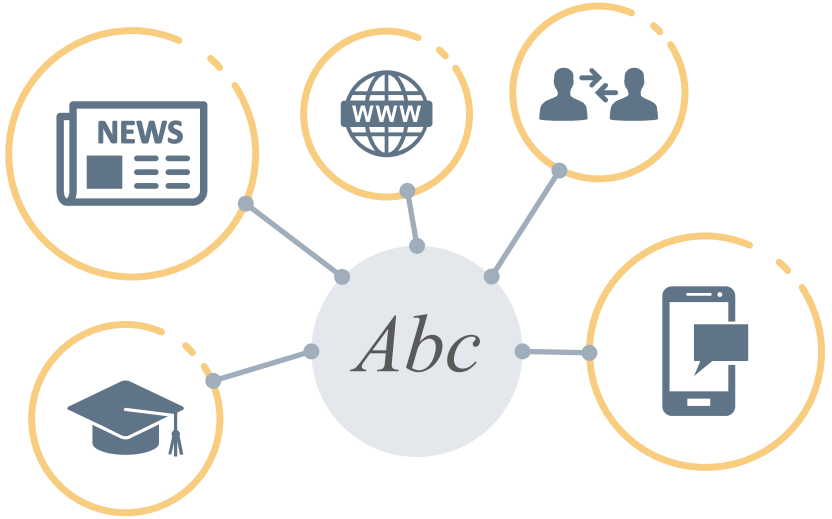
# Word meaning change

## Over time





## In different contexts (at the same time)



# main CHALLENGES for
## computational models of meaning and change

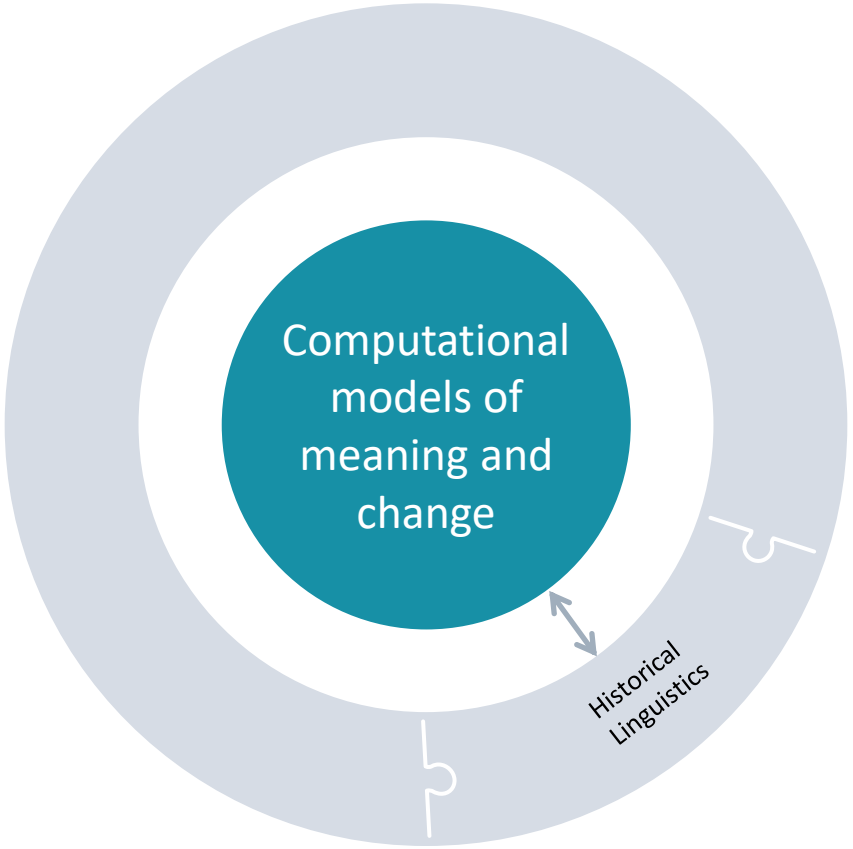
Computational models of meaning and change

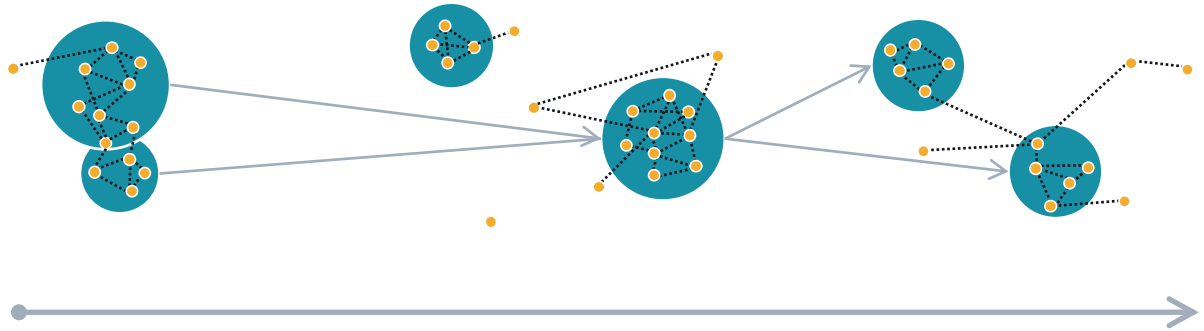
- Handle languages with smaller amounts of data
- Generalize to multiple languages
- Sense-aware models
- Find out WHAT changed, HOW and WHEN

# Our Research Questions

## Language level change

### Historical Linguistics

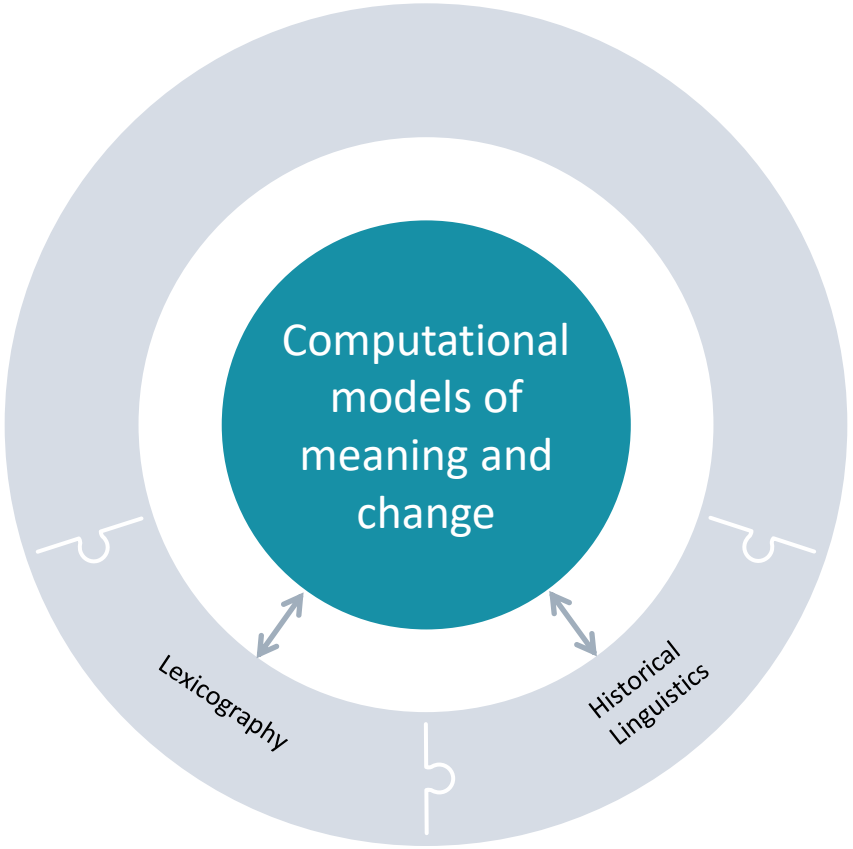


# Our Research Questions

## Language level change

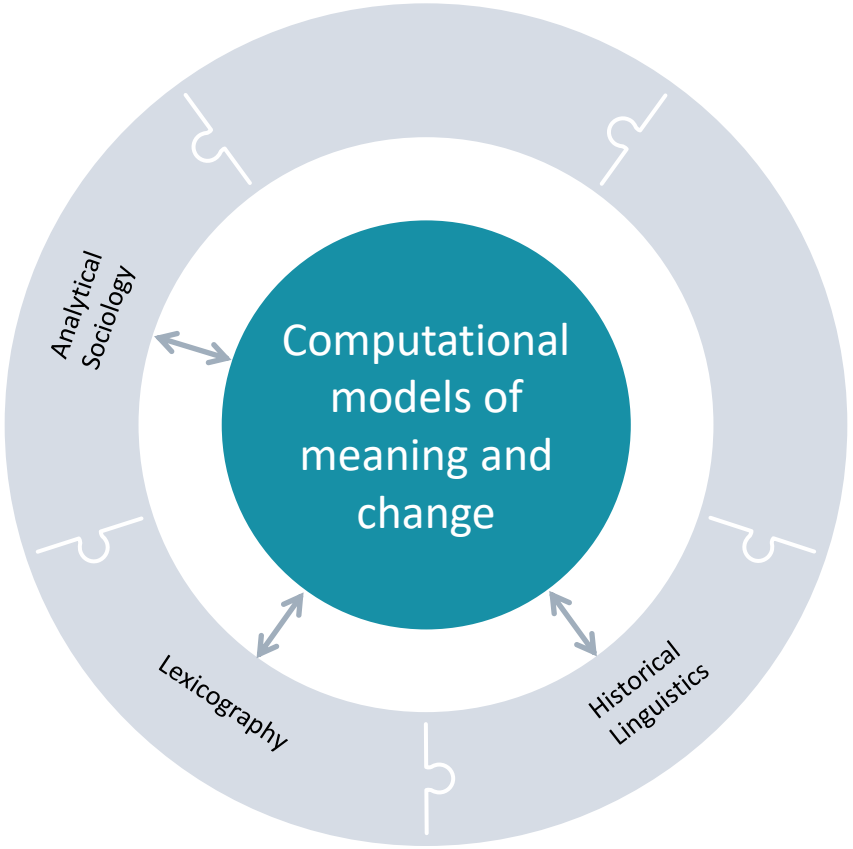
### Lexicography




# Our Research Questions

## Societal level change

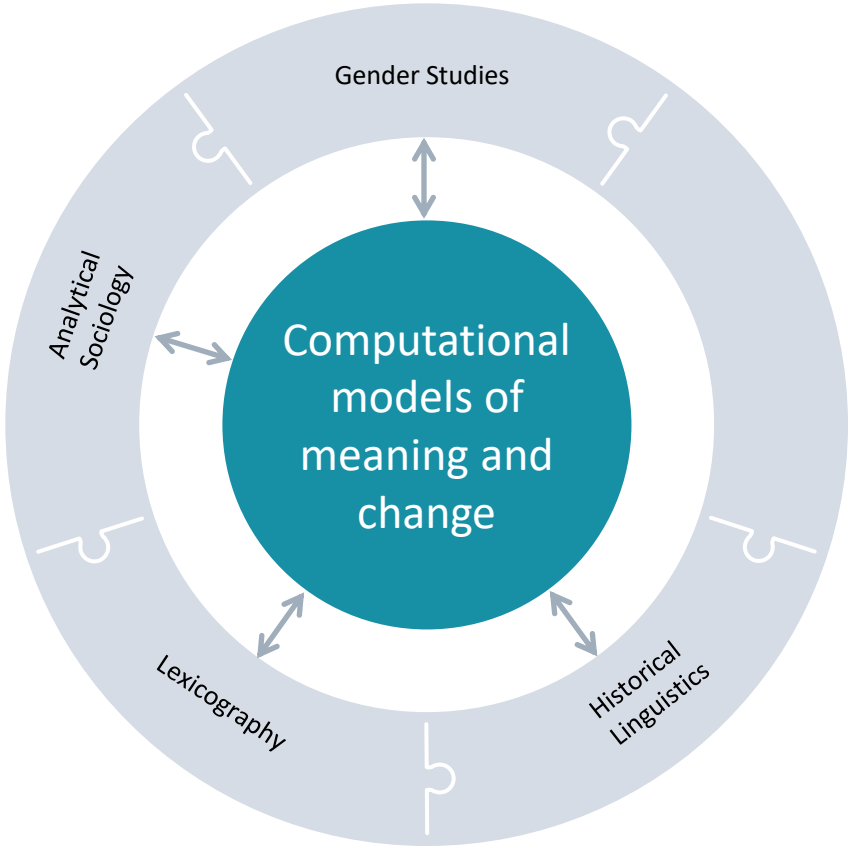
### Analytical Sociology

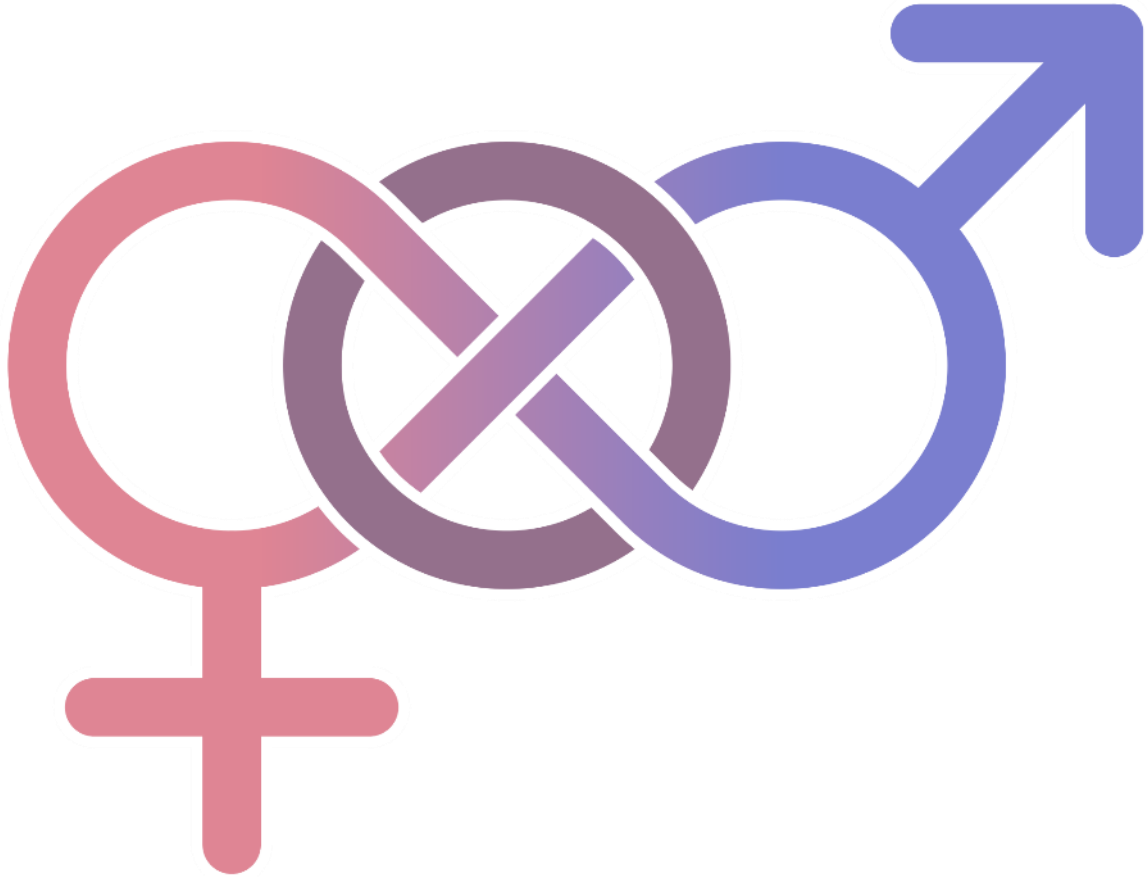



# Our Research Questions

## Societal level change

### Gender Studies

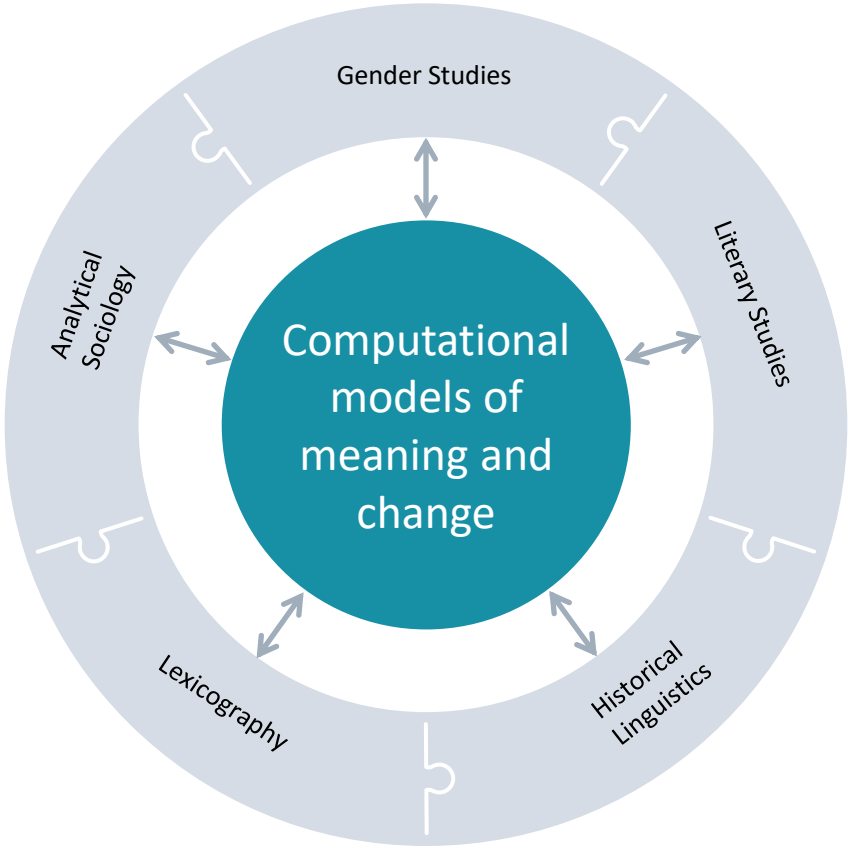


# Our Research Questions

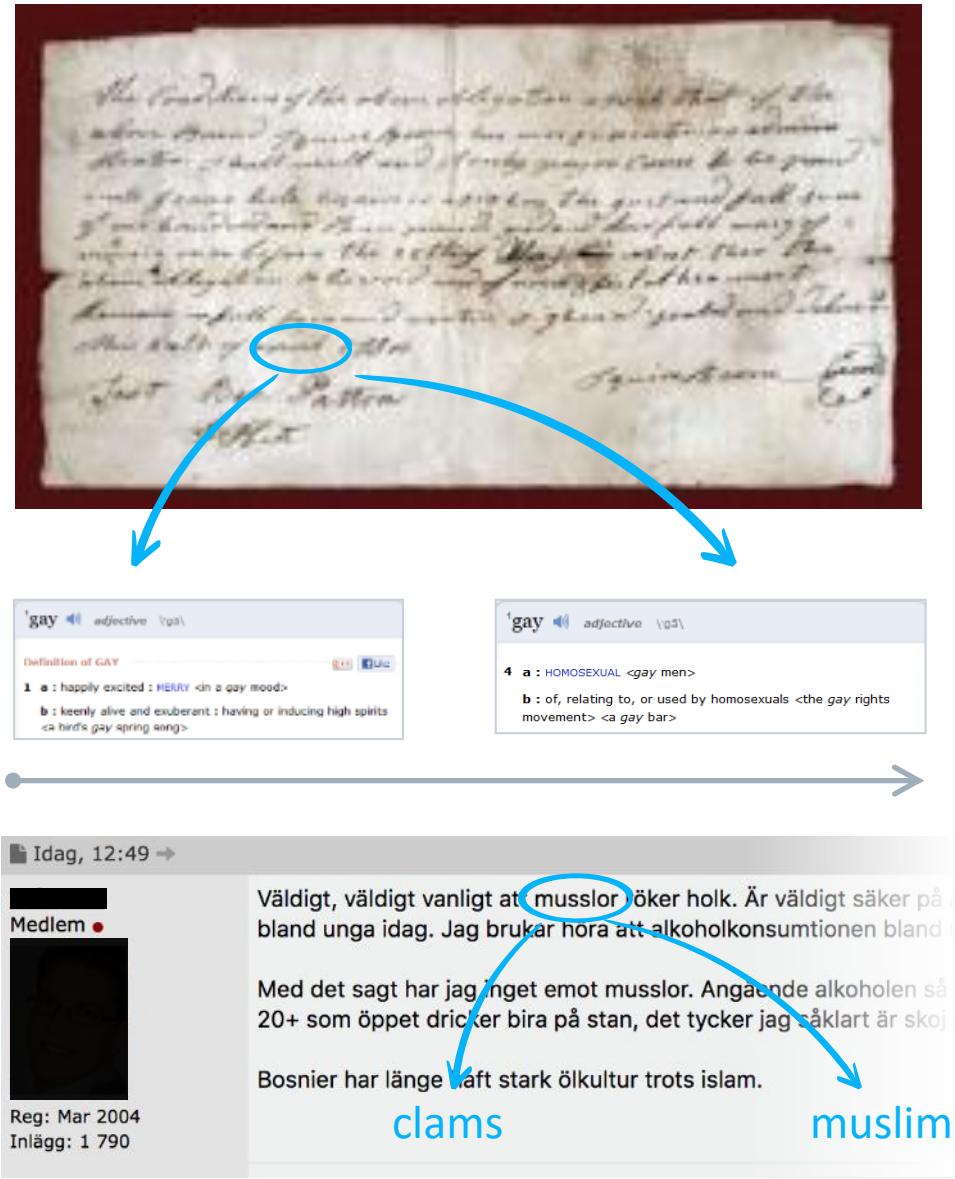
## Societal level change

### Literary Studies




# Our societal contribution

Meaning for everyone



https://www.changeiskey.org

Getting Started | Most Visited

**Change Is Key!** Sub-Projects People Events Publications About Funding Contact News Archive

# Change Is Key!

University of Gothenburg
University of Stuttgart
Queen Mary University of London
Lund University
Linköping University
KU Leuven

Change is Key! is a research program in which we aim to create computational tools to turn text into a story of both our language, our societies and culture and how these have changed over time.

Firstly, we will develop corpus-based methods for detecting semantic change (over time) and variation (across social groups and media types). This will create general tools for the study and detection of language change at large-scale and directly benefit historical linguistics and lexicography. Secondly, in collaboration with researchers from each field, we aim to answer research questions in social sciences, gender studies, and literary studies.

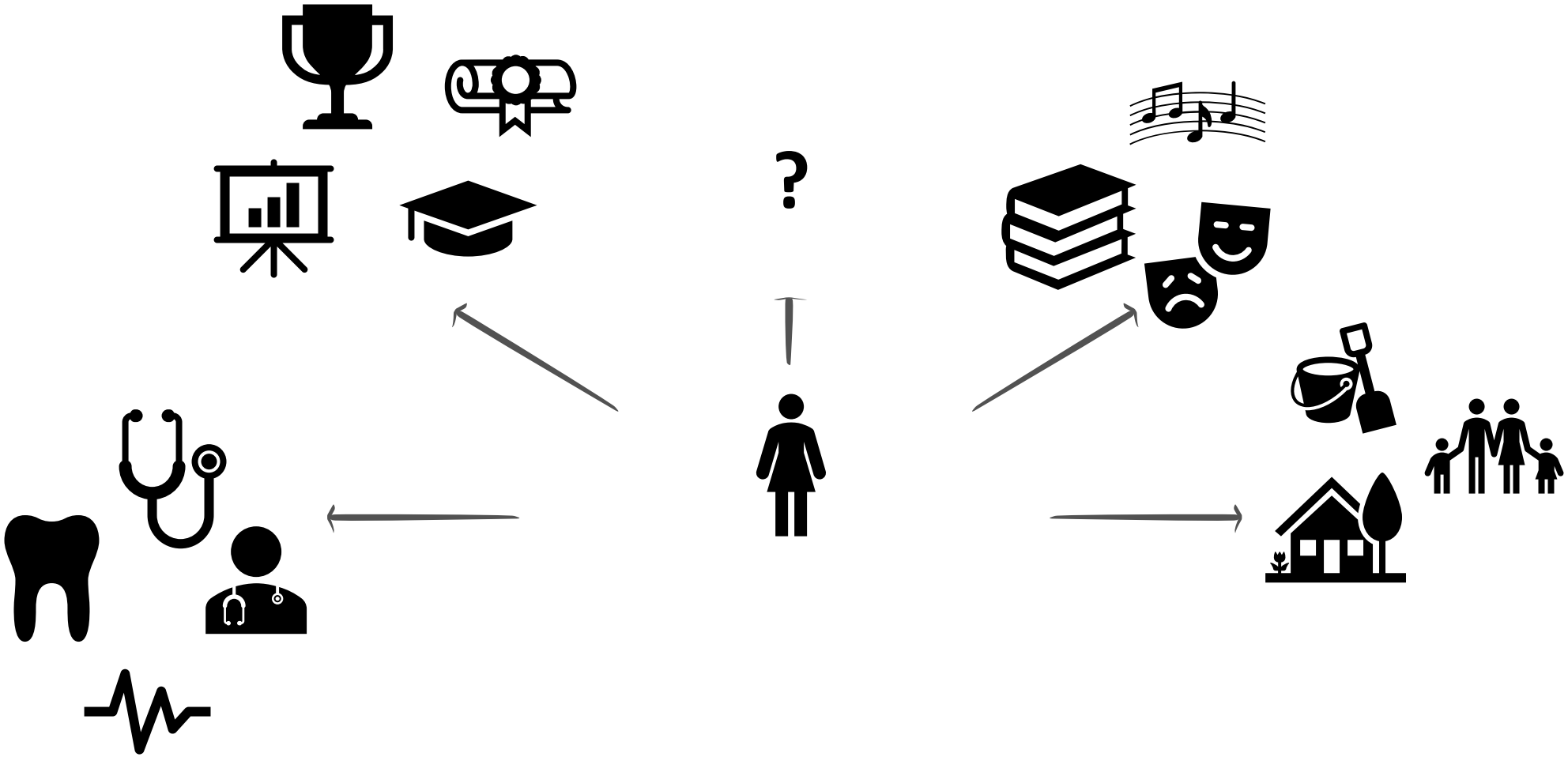
The program spans six years (2022 - 2027) with a total of 11 researchers, one research engineer and six partner universities.

We co-organized the third International Workshop on Computational Approaches to Historical Language Change 2022 (LChange'22) held end of May 2022 in Dublin.

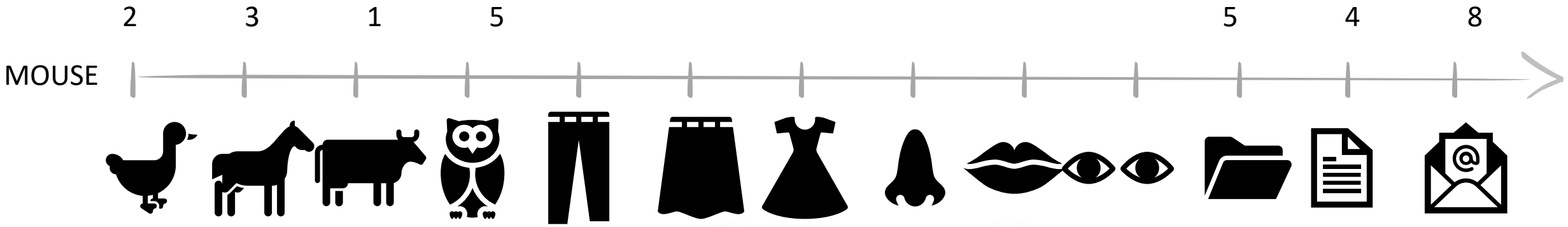
This research program is funded by the Riksbankens Jubileumsfond under reference number M21-0021 for a total of 33.5 Million SEK.

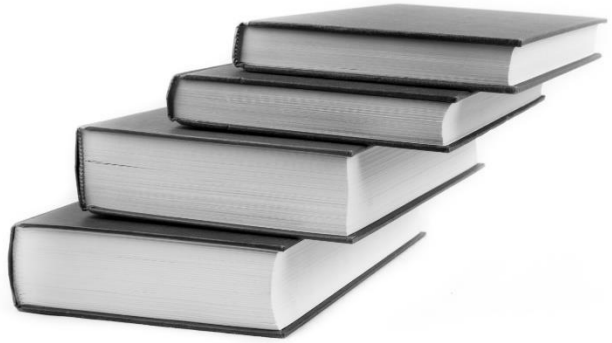
Changeiskey.org

# Methods for computational semantic change

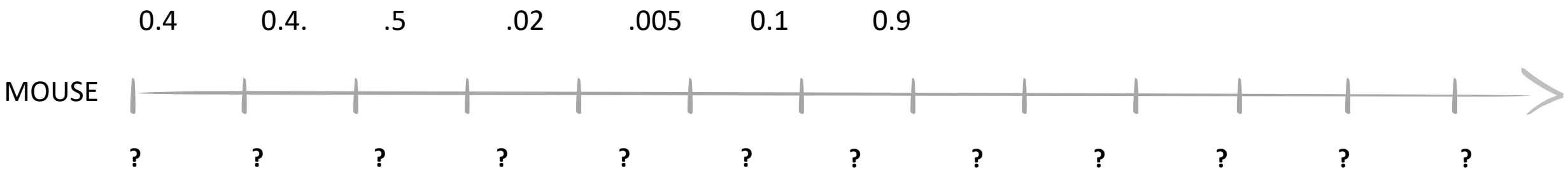


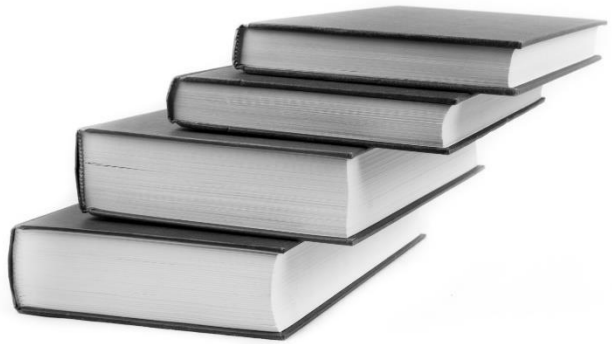
# Explicit, count-based vector representations

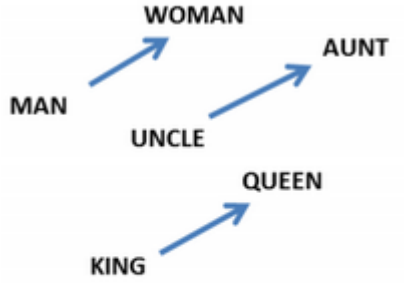


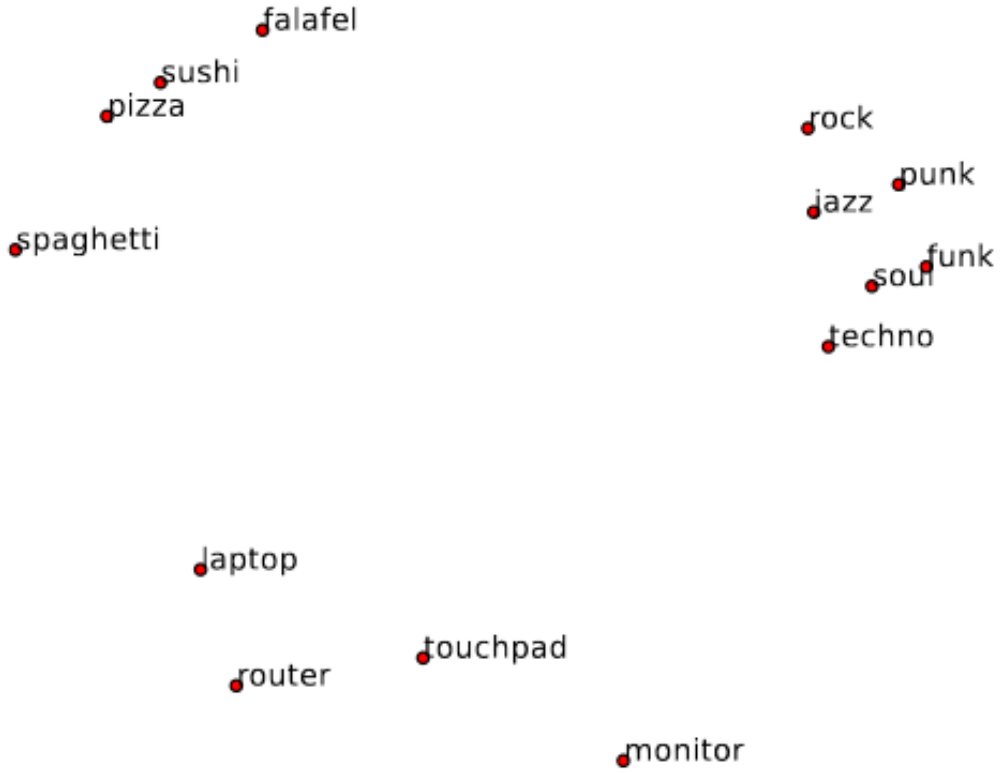
# Statistical, learned vector representations





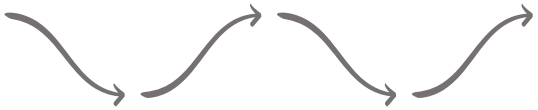

From Mikolov *et al.* (2013a)




Word embeddings shown in 2D instead of 50-100000

Image: Nieto Pina and Johansson, RANLP'15

# Pipeline



signal change

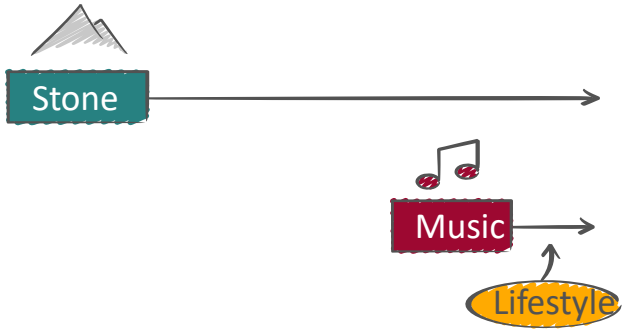
signal
topic, cluster, vector…

collective text

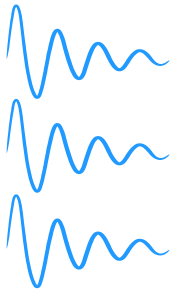
individual text

individual

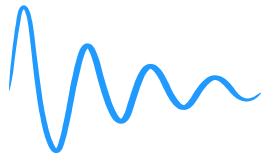
# Rock




Sense-differentiated



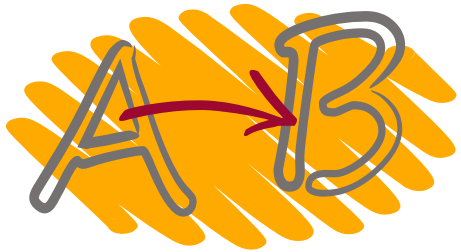
word



word

Single-sense

# Change type




- Novel word
- Novel word sense
- Novel related ws
- Novel unrelated ws
- Death
- Broadening
- Narrowing
- Join
- Split

Change

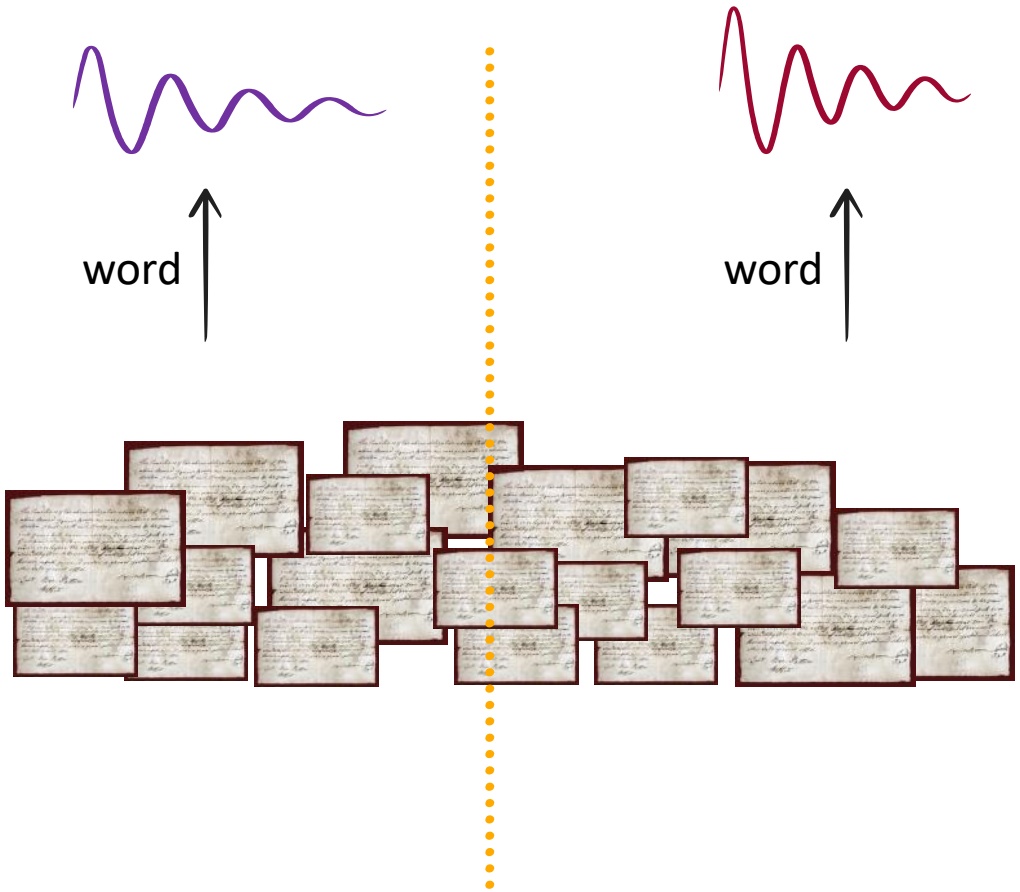
Sense-differentiated

Single-sense

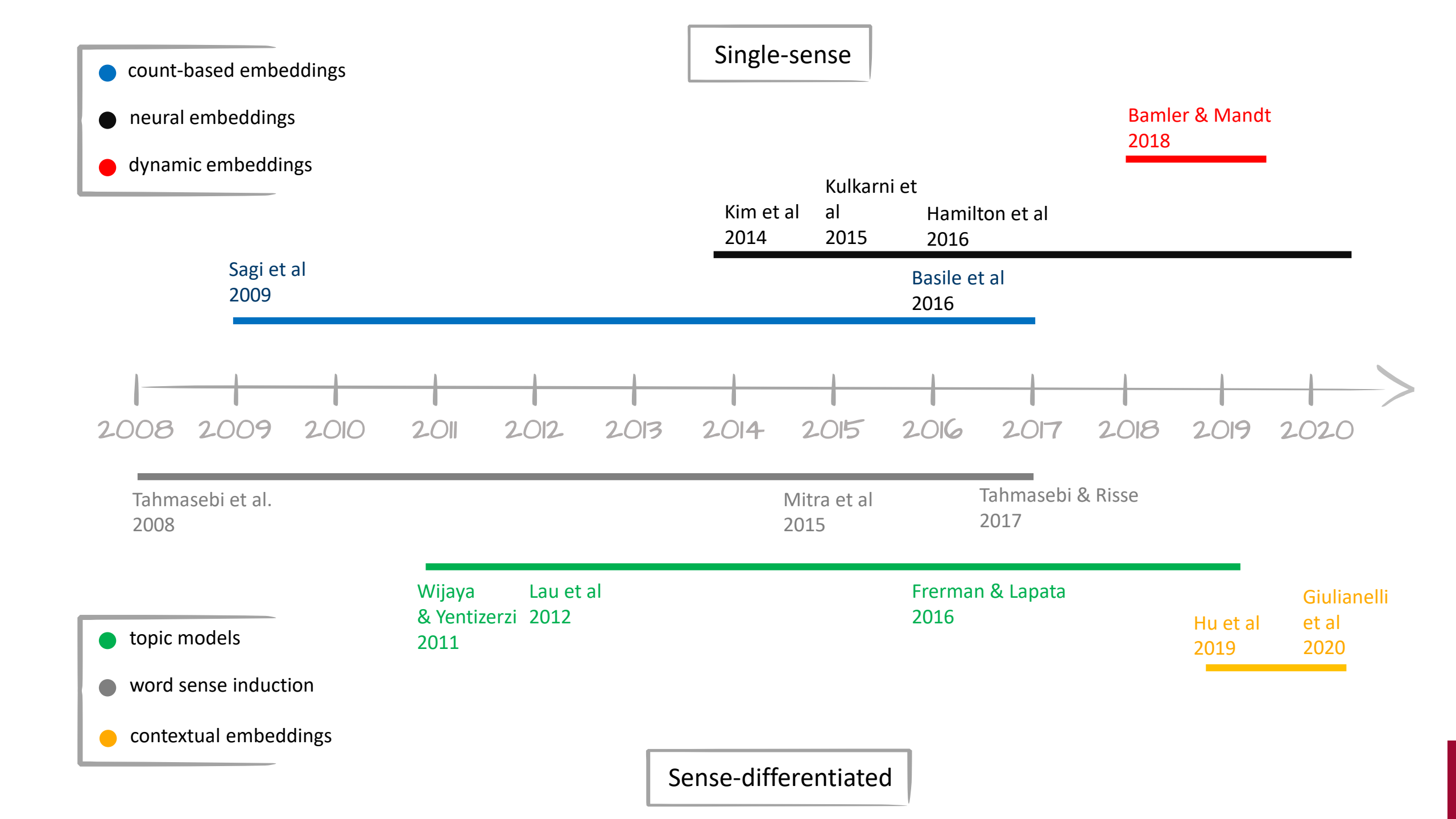
# Difficulty:

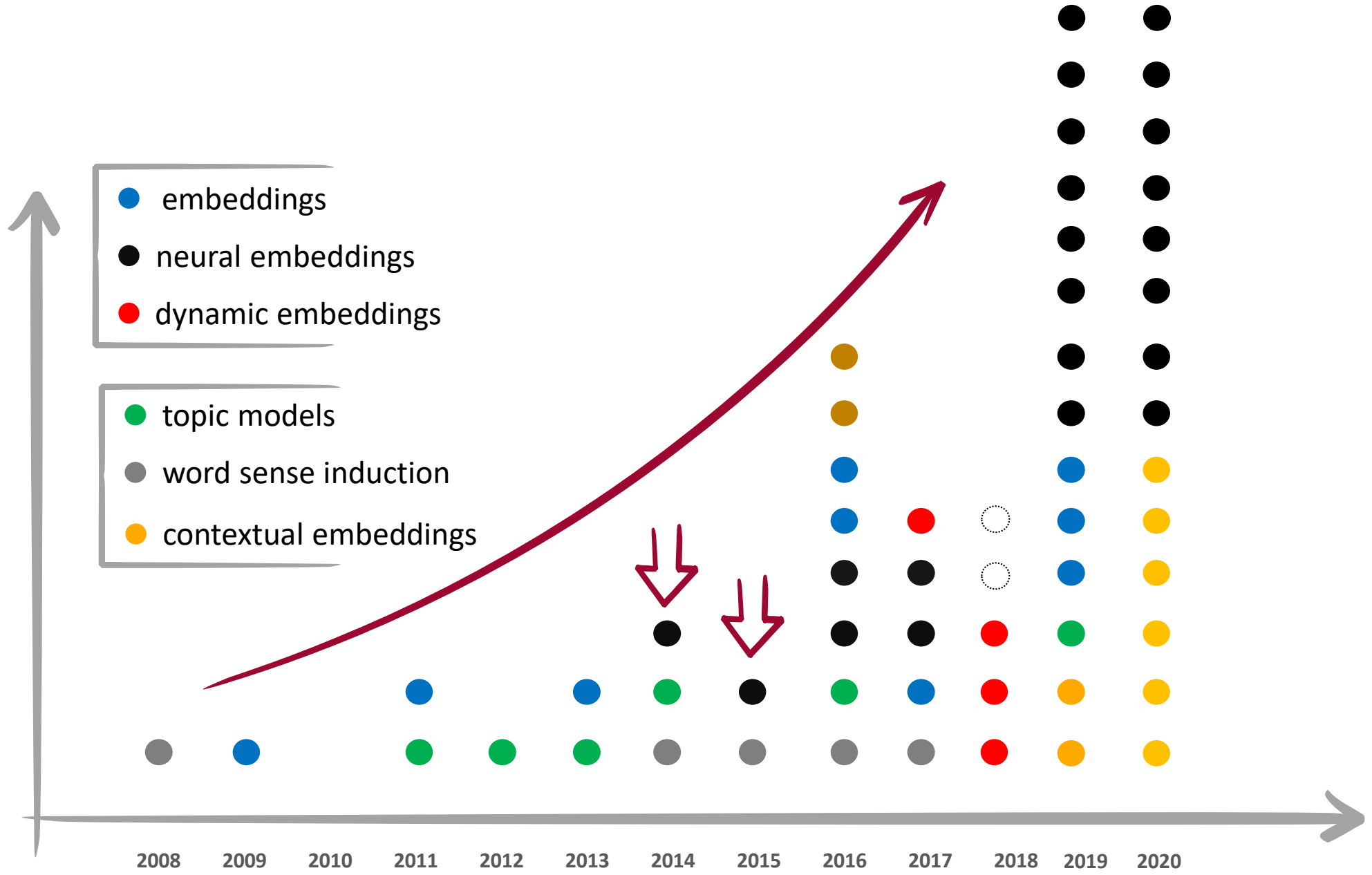
What does a word mean?

When are two meanings the same?




Single-sense


count-based embeddings
neural embeddings
dynamic embeddings
topic models
word sense induction
contextual embeddings
Single-sense
Sense-differentiated
Bamler & Mandt 2018
Kim et al 2014
Kulkarni et al 2015
Hamilton et al 2016
Sagi et al 2009
Basile et al 2016
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Tahmasebi et al. 2008
Mitra et al 2015
Tahmasebi & Risse 2017
Wijaya & Yentizerzi 2011
Lau et al 2012
Frerman & Lapata 2016
Hu et al 2019
Giulianelli et al 2020

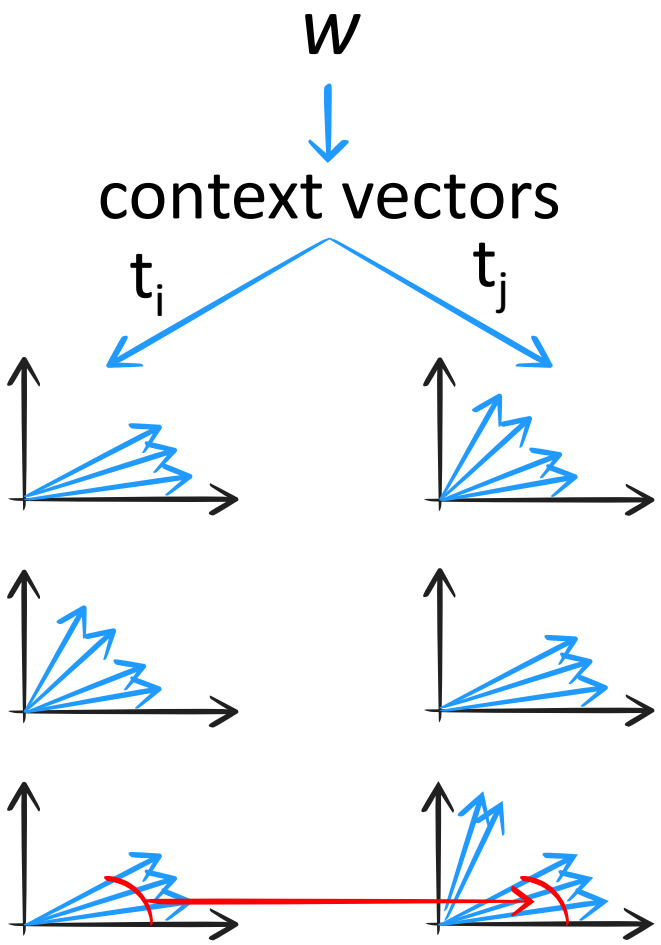

embeddings
neural embeddings
dynamic embeddings
topic models
word sense induction
contextual embeddings
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020

# Word-level
# semantic change

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

# 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

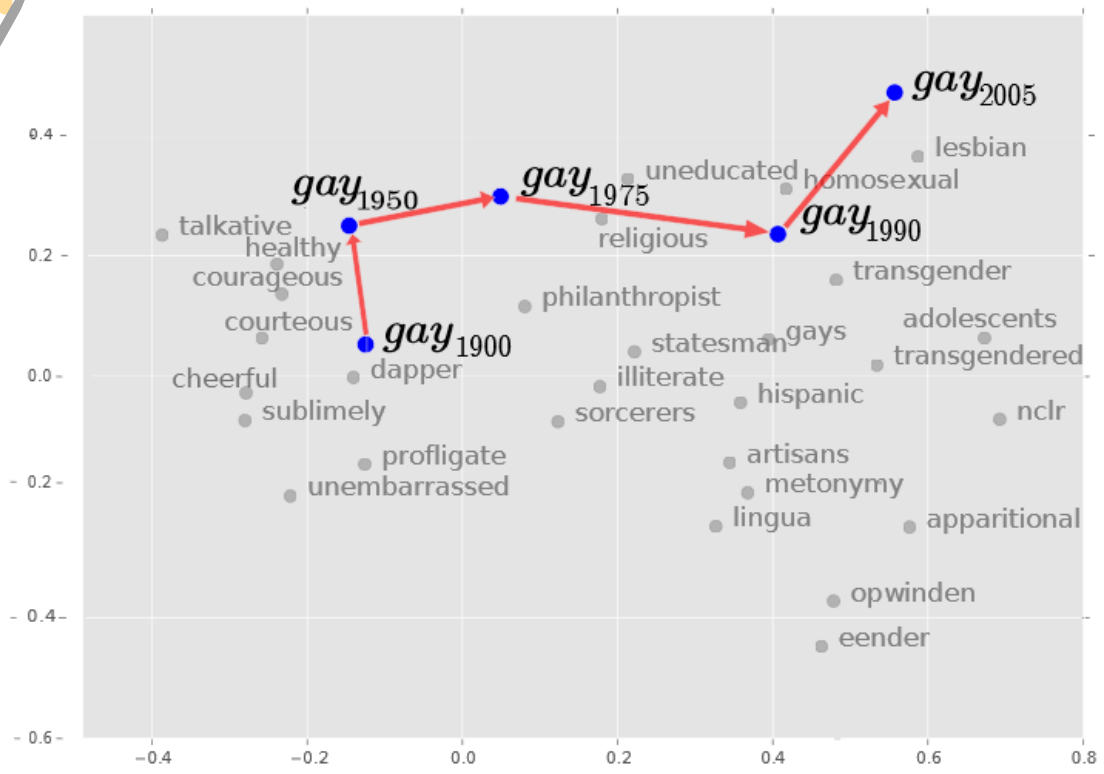
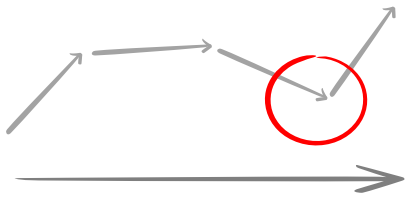
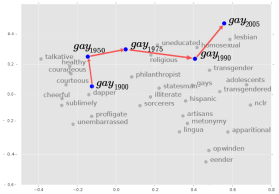



Image: Kulkarni et al. WWW'15

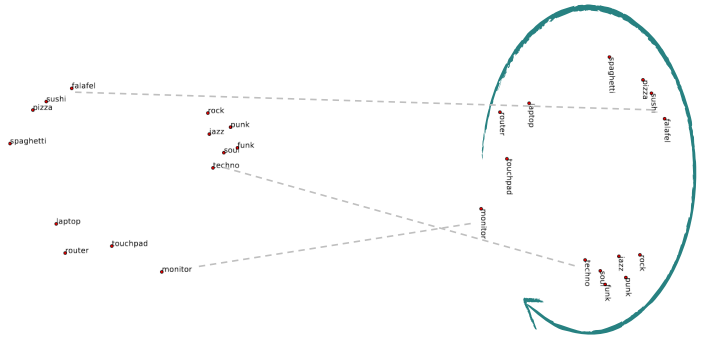
# LSC – individually trained embedding spaces



Track an individual word w over time

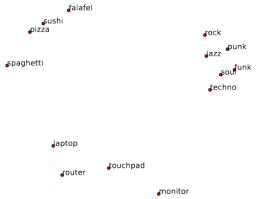



Change point/degree detection




multiple time points

**align**




Single-point embedding space $t_i$

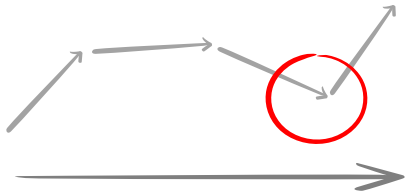
3 Change degree/ point

2 Alignment

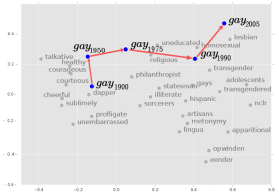
1 Embedding space

Vector space image: Nieto Pina and Johansson, RANLP'15

# LSC – dynamic embedding spaces



Track an individual word w over time




Change point/degree detection


Align while training

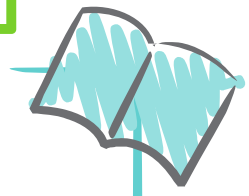
# Dynamic Embeddings

Share data across all time points

Avoids aligning

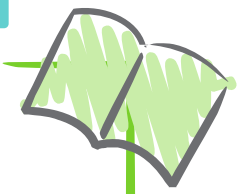
Bamler & Mandt:

- Bayesian Skip-gram



Yao et al:

- PPMI embeddings



Rudolph & Blei:

- Exponential family embeddings (Beronoulli embeddings)

Sharing data is **highly beneficial!**

# Temporal Referencing

Share contexts across all time points
Indivudal vectors for words for each bin
Avoids aligning

Dubossarsky et al

- SGNS
- PPMI embeddings


Sharing data is **highly beneficial!**

# Sense-differentiated semantic change

- topic models
- word sense induction
- contextual embeddings

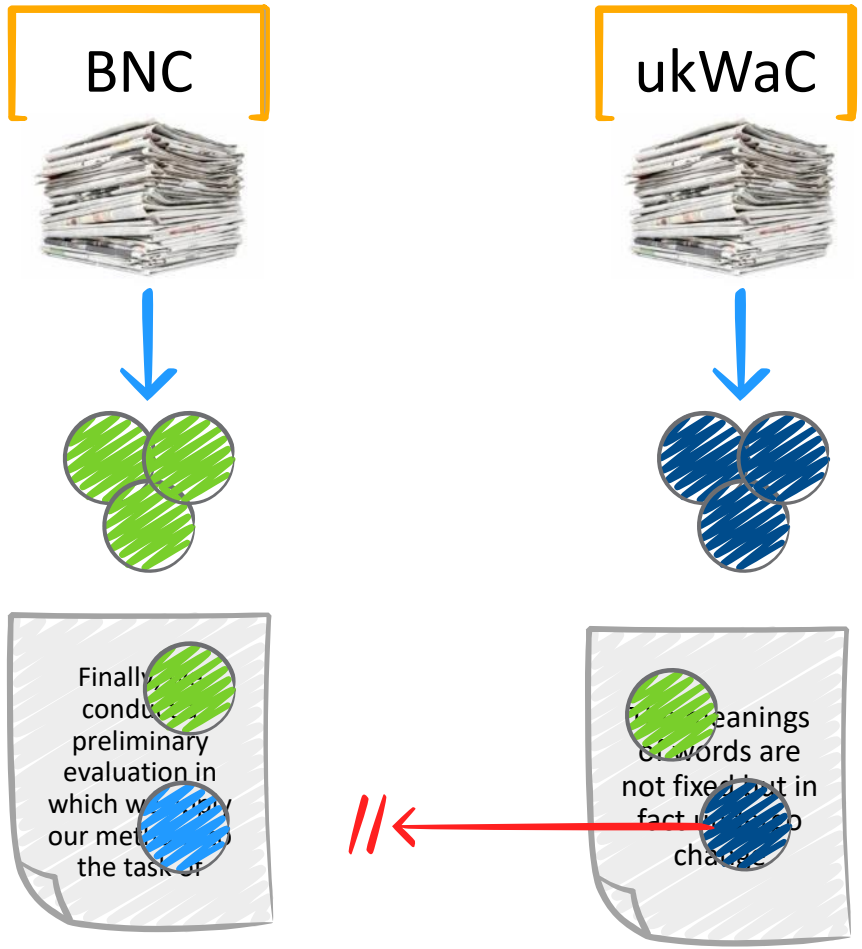
# Topic-based methods

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!



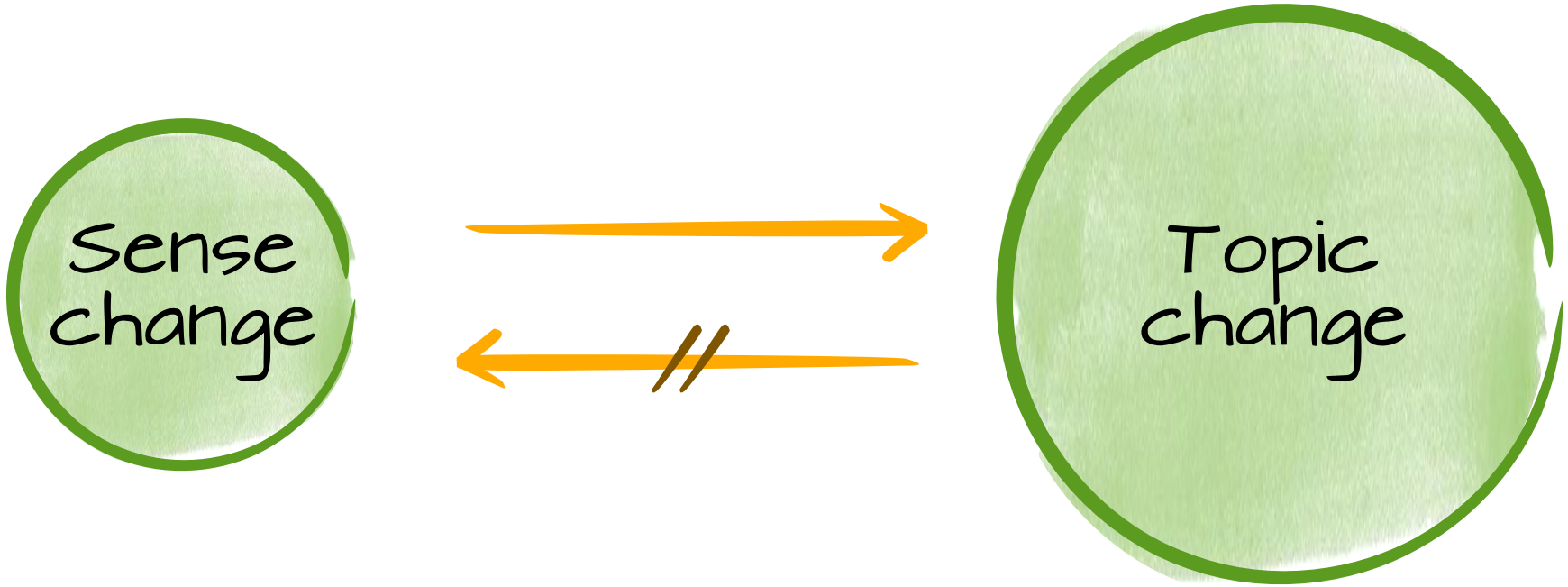

Lau et al. EACL 2014

Cook et al. Coling 2014

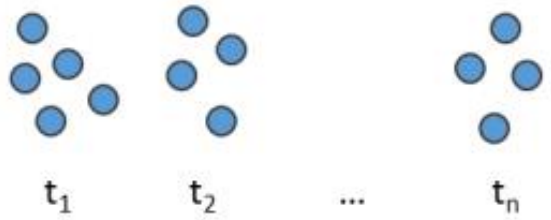
Wijaya & Yeniterzi DETECT '11

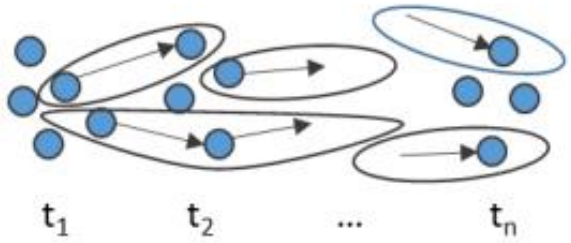
Frermann & Lapata TACL 2016

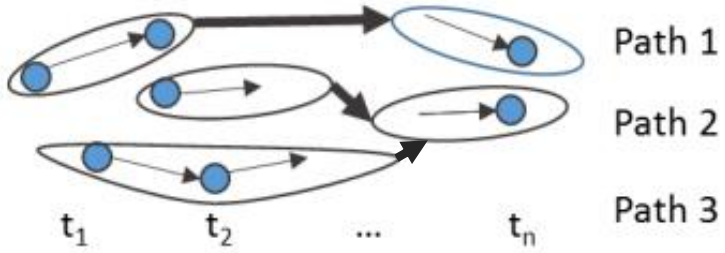
# Downsides topic models



# Word sense induction




**Step 1:**

Word sense induction

(curvature clustering) individual time slices

**Step 2:**

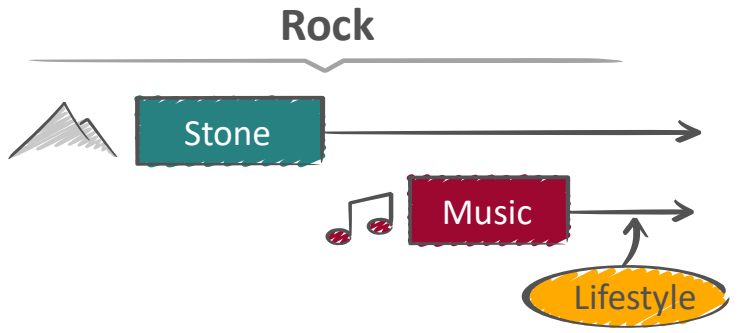
Detecting stable senses

→ units

**Step 3:**

Relating units

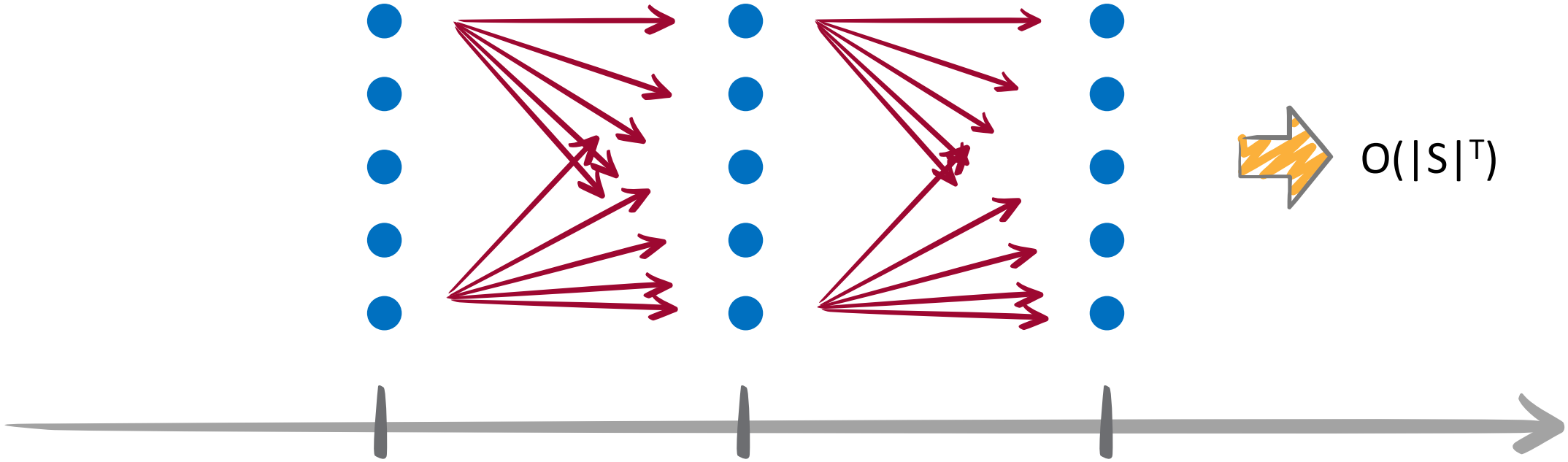
→ Paths



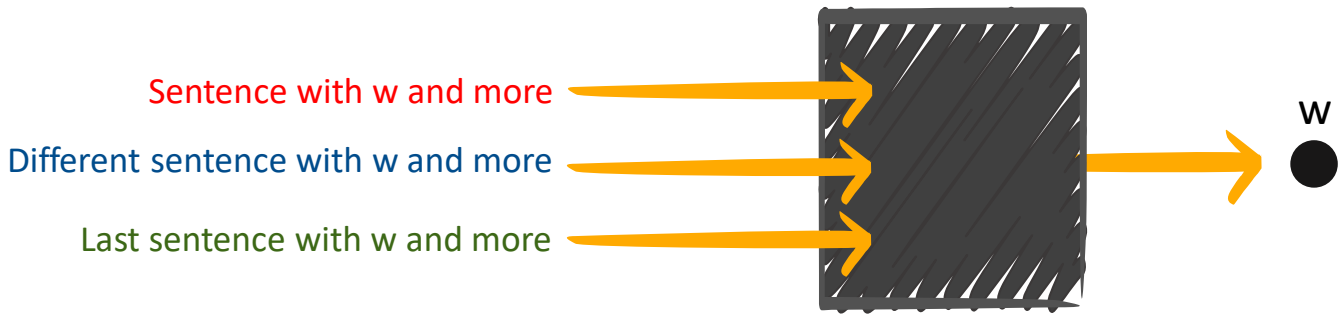



Tahmasebi & Risse, RANLP2017

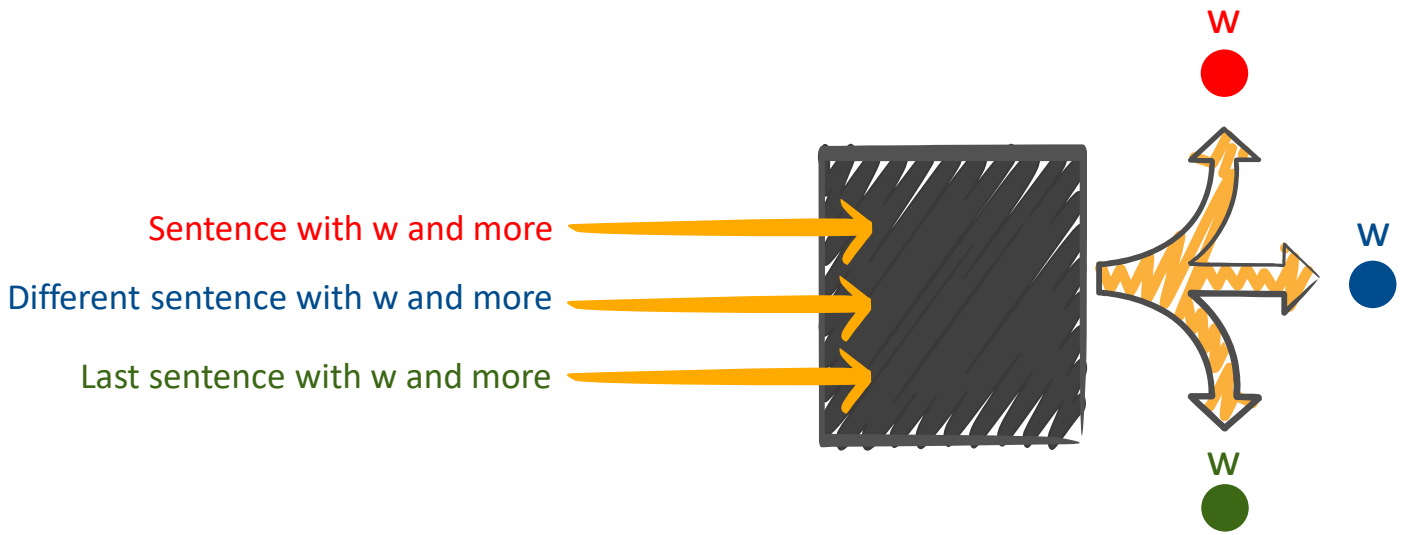
# Complexity



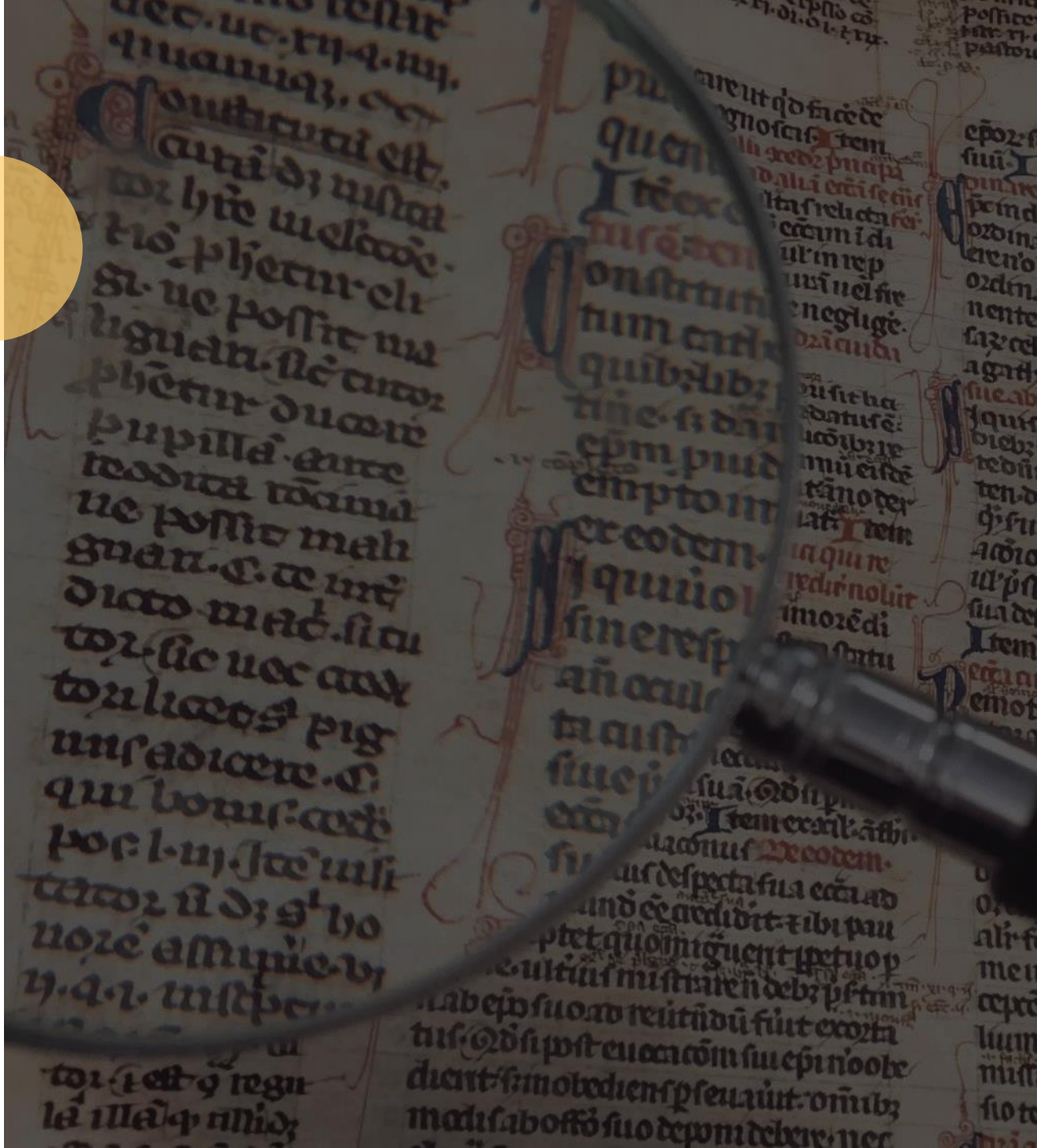
# Type-based embedding methods



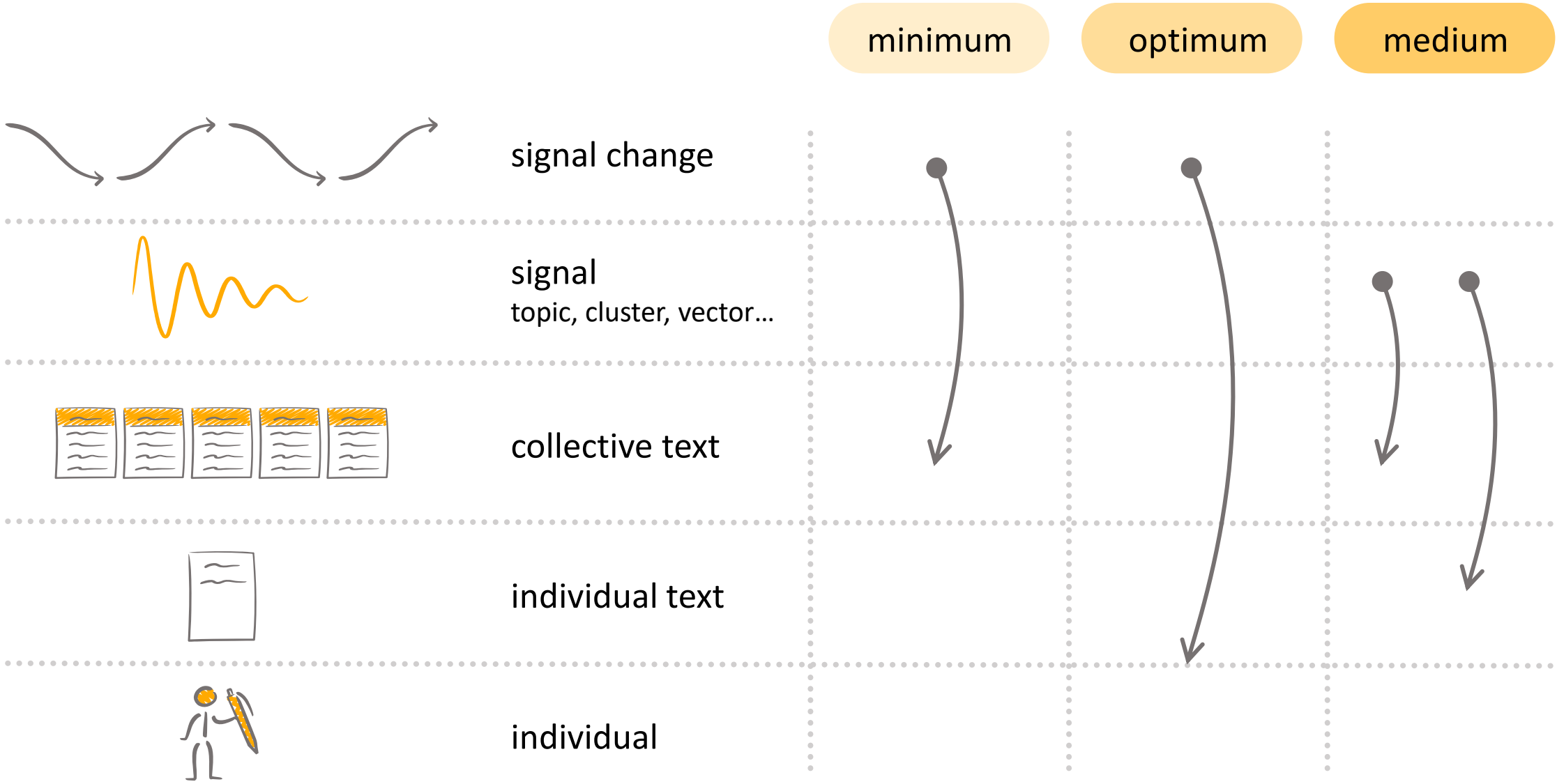
# Token-based embedding methods



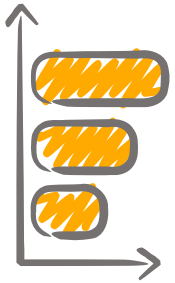
# Evaluation

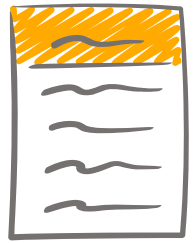


# Evaluation


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

# Evaluation





Pre-determined list of:

Top/bottom results

Controlled data

- Positive examples
- Negative examples




**3 ways**

# Summary of methods

- Most co-occurrence methods
  - are outperformed by type-embeddings
- Type-embeddings
  - average embeddings
  - need alignment across corpora
  - need very much data
- Dynamic embeddings
  - ‘remember’ too much historical
- Topic-based method
  - have little correspondence to senses
  - (and run badly on too large datasets)
- WSI-based method
  - have typically too low coverage
- Contextual embeddings
  - need to be clustered into senses

SPRÅKBANKENTEXT

RIKSBANKENS JUBILEUMSFOND
THE SWEDISH FOUNDATION FOR HUMANITIES AND SOCIAL SCIENCES

GÖTEBORGS UNIVERSITET

# Thank you!

Nina.tahmasebi@gu.se

nina@tahmasebi.se

# Detecting semantic change in historical texts

## A case study on Flemish socialist newspapers using LDA

Simon Hengchen, PhD
KBR Digital Heritage Seminar – October 18th 2022

# Foreword: affiliation imbroglio


- Currently working at my company
- Steering Committee member of *Change is Key!*
- … also formally a guest researcher *(gästforskare)* at Språkbanken Text within the University of Gothenburg
- … also a lecturer at the Université de Genève

… but the work presented here was done during my PhD at the Université libre de Bruxelles, as part of the BELSPO BRAIN-BE project TIC-Belgium (coordinated by UGent), with data provided by AMSAB-ISG and computational power by the FNRS.

# Context

- Words change meaning
- This makes historical interpretation more complicated
- Historians in the project discussed the need for a way to detect those words

# Problem statement

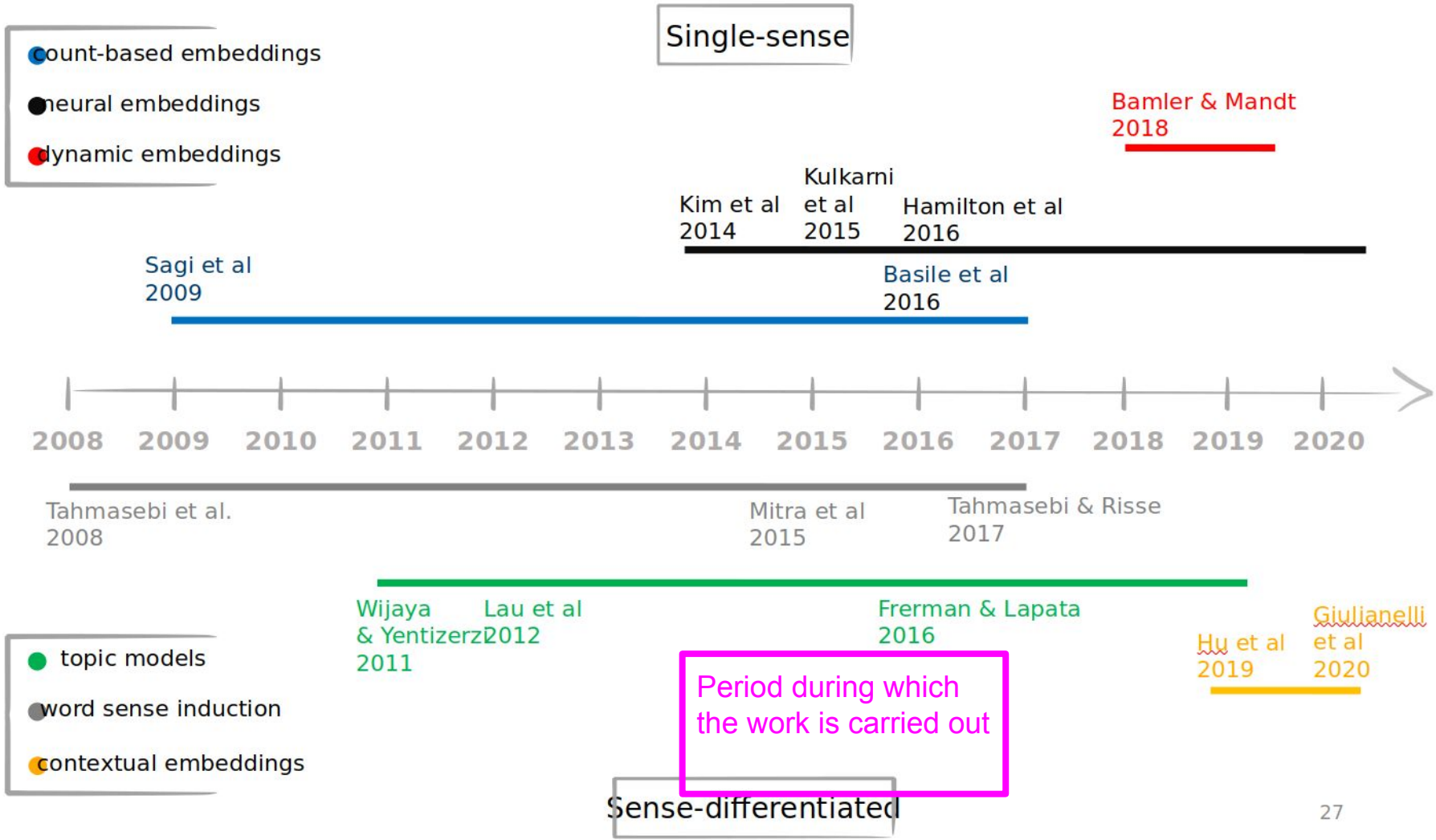
We want to detect cases of semantic change:

- that are relevant
  - within a context
  - within a theme
- that are minute
  - maybe within a specific context only
- that we can see examples of

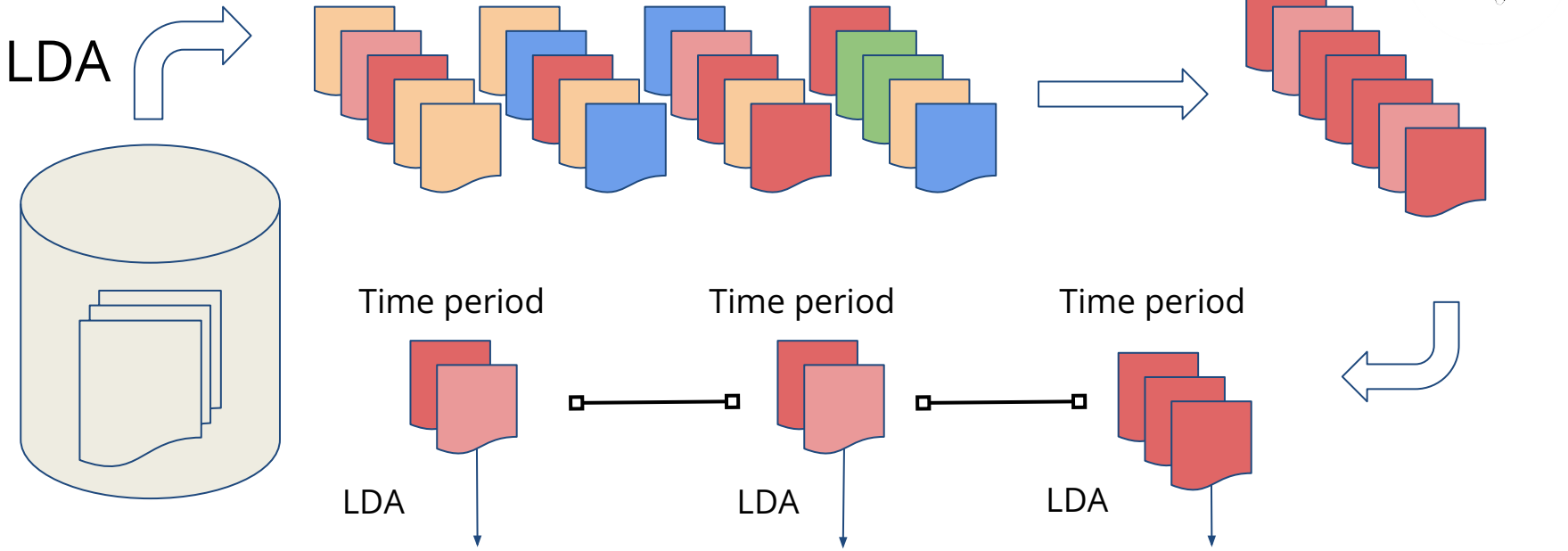
# Problem statement

Issues:

- people from the past are dead, so we can't ask them
- usual caveats of working with digitised, historical data:
  - quality
  - representativity (socio-cultural, geographical, availability, segmentation)
- working ‘in the wild’, meaning there is no evaluation data (unlike common NLP/ML scenarios)
- NLP methods aren't usually explainable nor do they allow for humanistic input/guiding


count-based embeddings
neural embeddings
dynamic embeddings
topic models
word sense induction
contextual embeddings
Single-sense
Sense-differentiated
Bamler & Mandt 2018
Kim et al 2014
Kulkarni et al 2015
Hamilton et al 2016
Sagi et al 2009
Basile et al 2016
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Tahmasebi et al. 2008
Mitra et al 2015
Tahmasebi & Risse 2017
Wijaya & Yentizerz 2011
Lau et al 2012
Frerman & Lapata 2016
Hu et al 2019
Giulianelli et al 2020
Period during which the work is carried out

# Method


LDA
Time period
Time period
Time period
LDA
LDA
LDA
lorem
ipsum
dolor
sit
amet
lorem
ipsum
sit
poena
poenam
lorem
ipsum
sit
poenam
amet

# LDA

LDA stands for latent Dirichlet allocation (Blei et al 2003):

- unsupervised algorithm
- processes sets of documents (what a document is is for you to choose)
- each document contains a mixture of topics, in different probabilities
- each topic is a distribution over words

Example:

- Document #1 is 83% likely to be topic #21 **AND** 5% to be topic #2 **AND** …
- Topic #21 is *elephant (.3) visit (.2) lion (.2) zebra (.15) …*
- Topic #2 is …

# Method

Advantages:

- easy enough
- focussed on a topic of interest
- intellectual overview at each step of the process

Limitations:

- takes some time
- not fully-automated → as human input is required, any iteration needs to be evaluated manually (see eg Tahmasebi and Hengchen 2019 for more on this)
- domain knowledge is required



Tahmasebi, N. and Hengchen, S., 2019. The Strengths and Pitfalls of Large-Scale Text Mining for Literary Studies. *Samlaren: Tidskrift för forskning om svensk och annan nordisk litteratur*, *140*, pp.198-227.

# Case study: *Vooruit*, a Flemish socialist newspaper

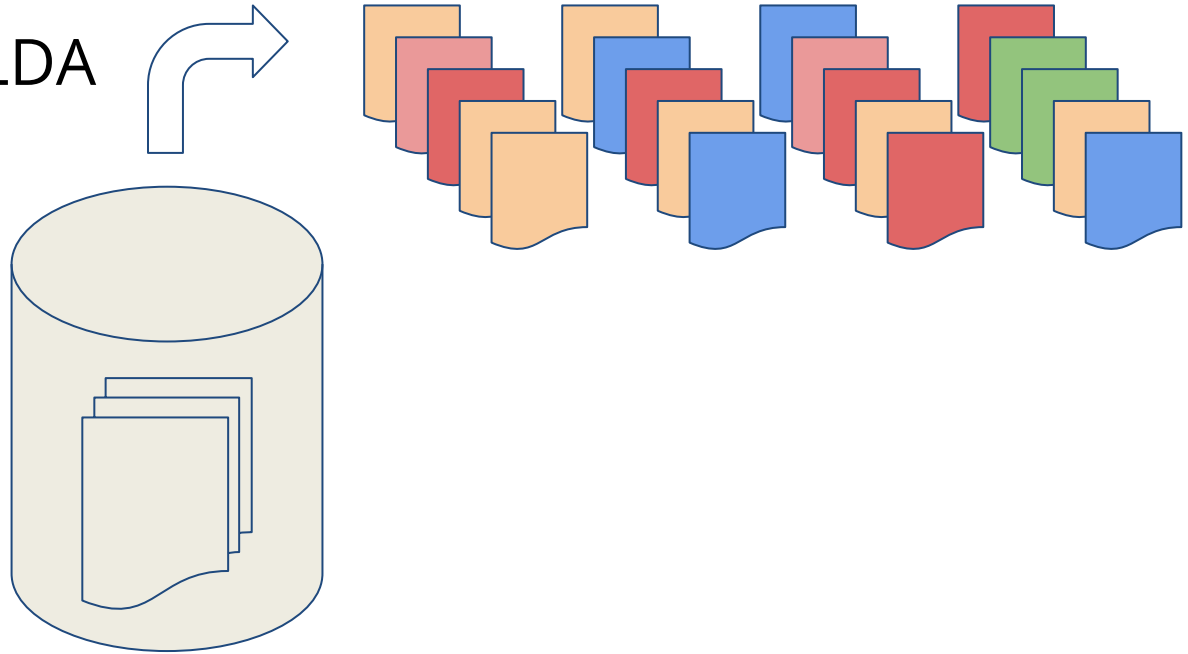
- Daily from the region of Ghent
- 1884 → 1950
- 445M words
- Digitised by two institutions:
  - 1884 → 1889 and 1911 → 1950: KBR
  - 1890 → 1910: Amsab-ISG

```
$ head freq-cleaned-end.txt
16252324 de
10392677 van
8623741 en
8138596 het
6000559 te
5187610 een
```

```
$ tail freq-cleaned-end.txt
1 ƒ4418
1 \t4
1 ƒ4
1 ƒ355
1 ƒ31
1 ƒ284
```

# Unsupervised approach validated by a human

LDA

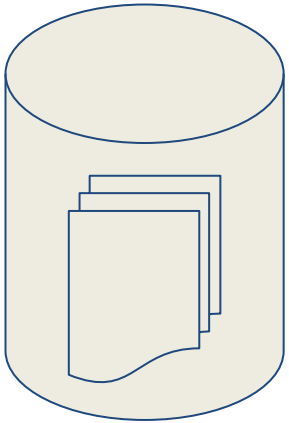


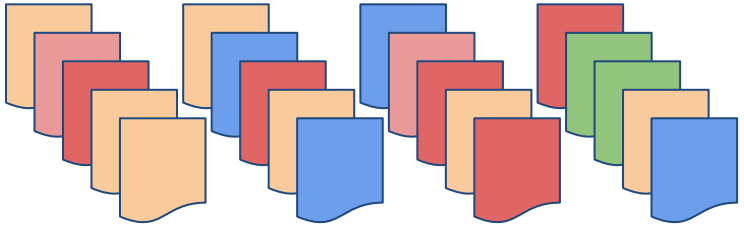
#24 0.03662 koncert orkest radio muziek gramofoonplaten dansmuziek concert programma hilversum lichte dagblad gesproken parijs gramofoonmuziek leiding gramofoon berichten zang londen

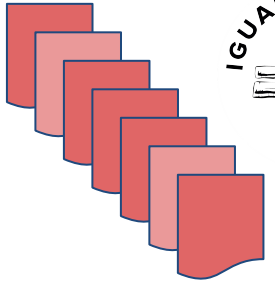
#29 0.20378 arbeiders moeten staking jaar werden patroons werk loonen groote toestand leden algemeene plaats nieuwe belgië belgische vergadering werklieden tusschen

## Unsupervised approach validated by a human

LDA







(human) choice of a historically-relevant topic

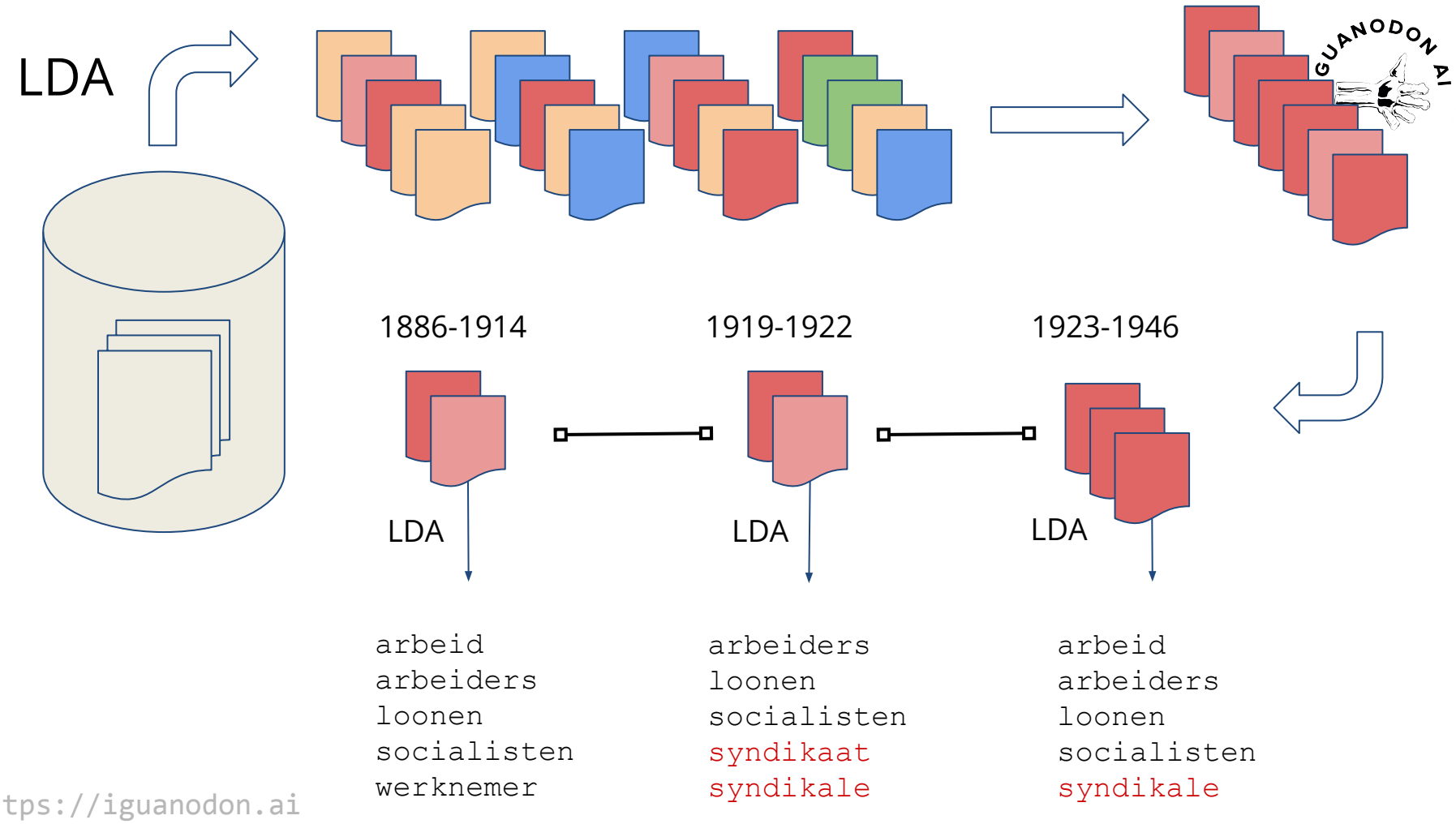
```
#29 0.20378 arbeiders moeten staking jaar werden patroons werk loonen
groote toestand leden algemeene plaats nieuwe belgië belgische vergadering
werklieden tusschen
```

→ workers' rights, Blank (1999)'s "sociocultural change", one of his 'motivations' for semantic change

# Case study

Human choice of time periods:

- Historically-aware choice:
  - Relevant for chosen theme
- Pragmatic choices:
  - Enough data for intrinsically valid LDA model


LDA
GUANODON AI
1886-1914
1919-1922
1923-1946
LDA
LDA
LDA
arbeid
arbeiders
loonen
socialisten
werknemer
arbeiders
loonen
socialisten
syndikaat
syndikale
arbeid
arbeiders
loonen
socialisten
syndikale

# Case study: validation

*syndikaat* (modern: syndicaat):

- consortium, group of people with the same (professional goal). ≈ society, trade association
- trade union

# Case study: validation

*syndikaat* (modern: syndicaat):

- consortium, group of people with the same (professional goal). ≈ society, trade association
- trade union..
  - first recorded use is in 1914:

> *"Vereniging van vaklieden, vakbond"* in het socialistisch syndikaat van pelswerkers (1914; WNT pels)

Philippa, M., et al (2003-2009). Etymologisch Woordenboek van het Nederlands. Amsterdam University Press.

# Case study: validation

Concordance analysis

Table III.6: Distribution in % of the received meanings of *syndikaat*

| | Trade Union | Unclear | Trade Association | Duplicates |
|---|---|---|---|---|
| Subset 1 | 64 | 8 | 23 | 5 |
| Subset 2 | 81.7 | 3.4 | 7.7 | 7.2 |
| Subset 3 | 85.3 | 0.9 | 8.3 | 5.5 |

# Case study: conclusion

- Method seems to be working
- Method is unsupervised but requiring supervision
- Method is not a "put data in, get results out" pipeline

Caveats:

- Only shown to work in a specific context
- … with one example
- No standard evaluation set, but human evaluation

# More on computational approaches to semantic change

Tahmasebi, N., Borin, L., Jatowt, A., Hengchen, S. (Eds). Computational approaches to semantic change. 2021, Language Science Press.
https://langsci-press.org/catalog/book/303