Språk-
BANKEN

CENTRUM FÖR DIGITAL HUMANIORA

GÖTEBORGS
UNIVERSITET

Detecting language change for the digital humanities;

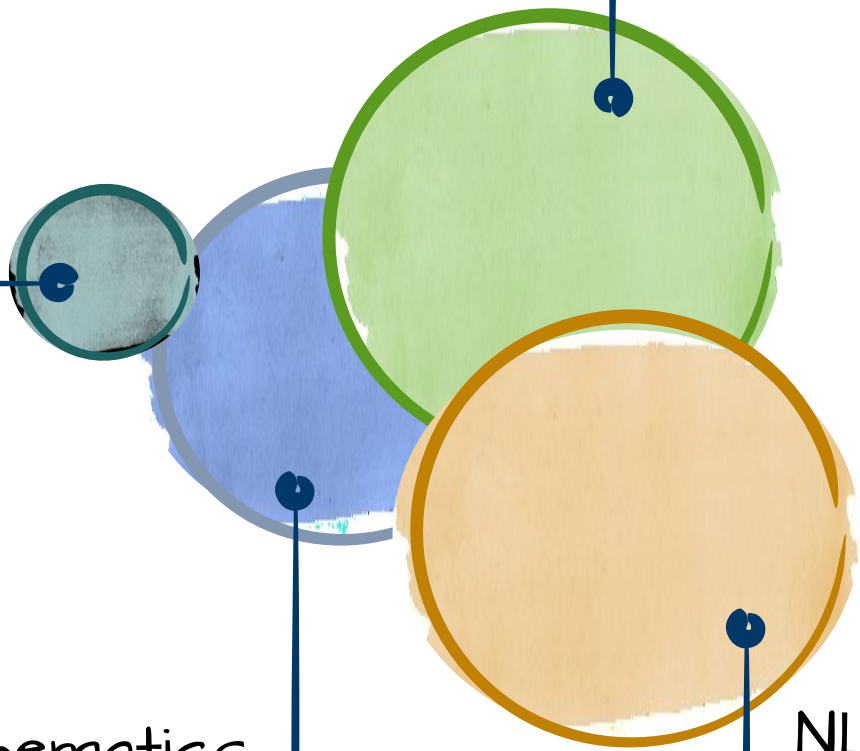
challenges and opportunities

Nina Tahmasebi, PhD

University of Gothenburg

6th Estonian Digital Humanities conference

Me



Computer science
(Phd + Postdoc)

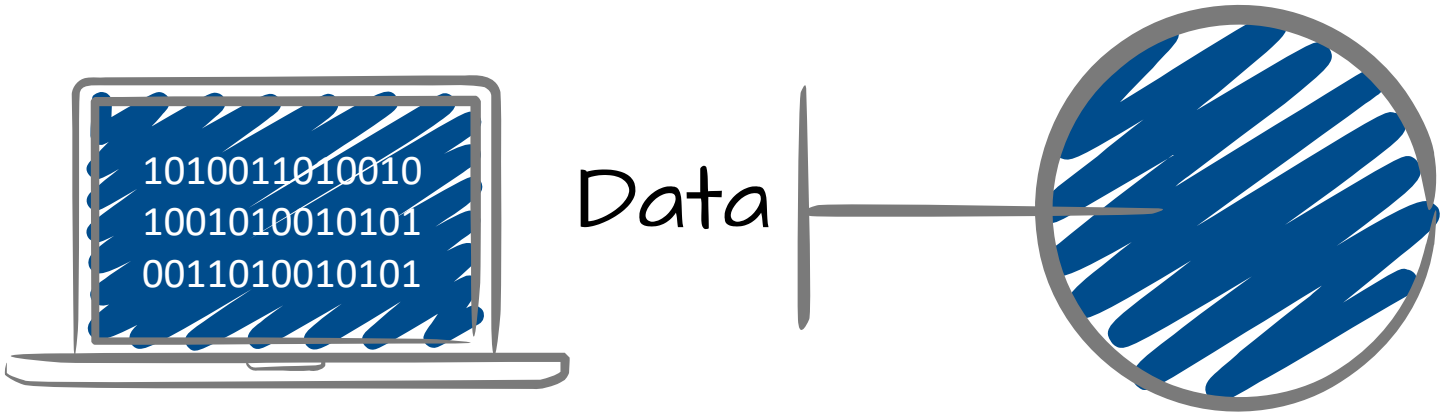
Electrical
Engineering

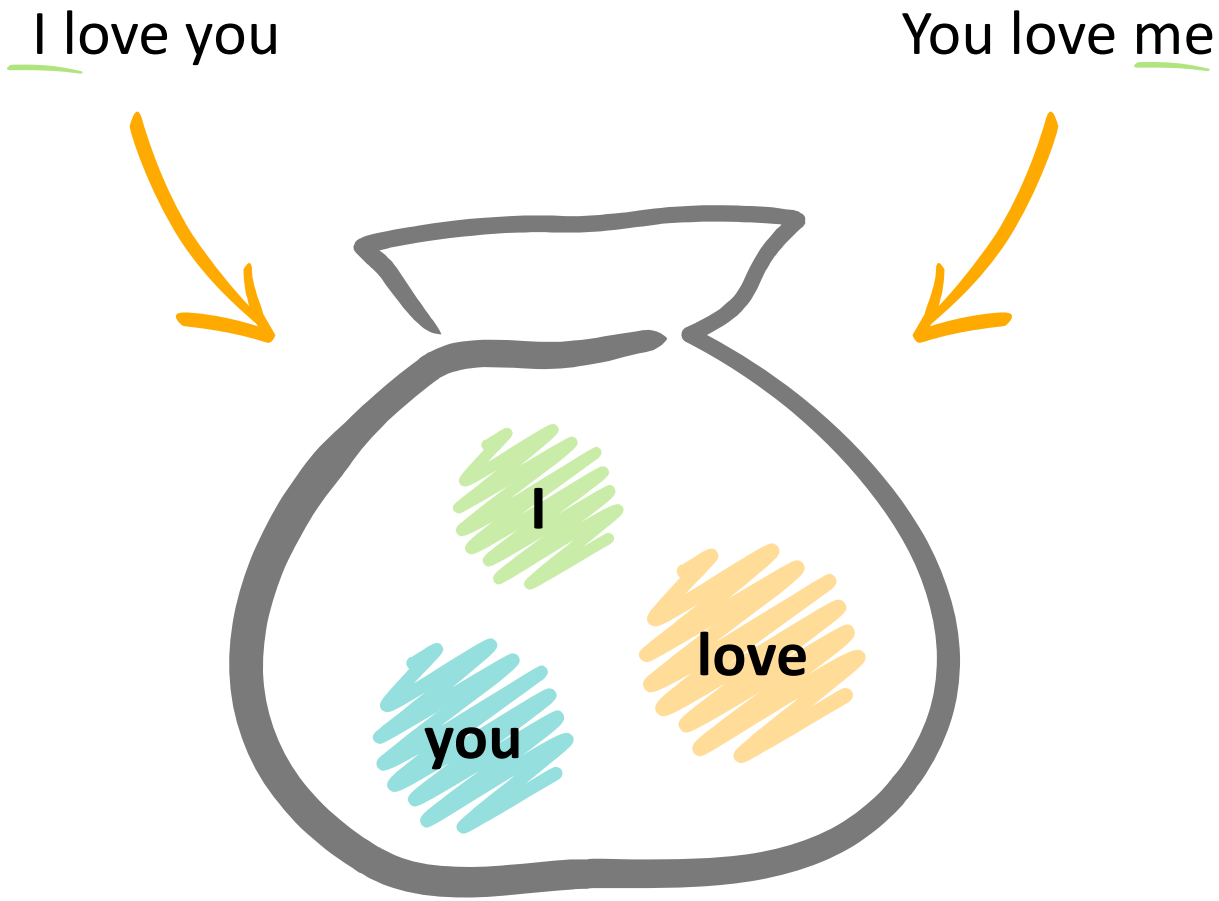
Mathematics
(B.Sc &
M.Sc.)

NLP /
Language
Technology
(Researcher)

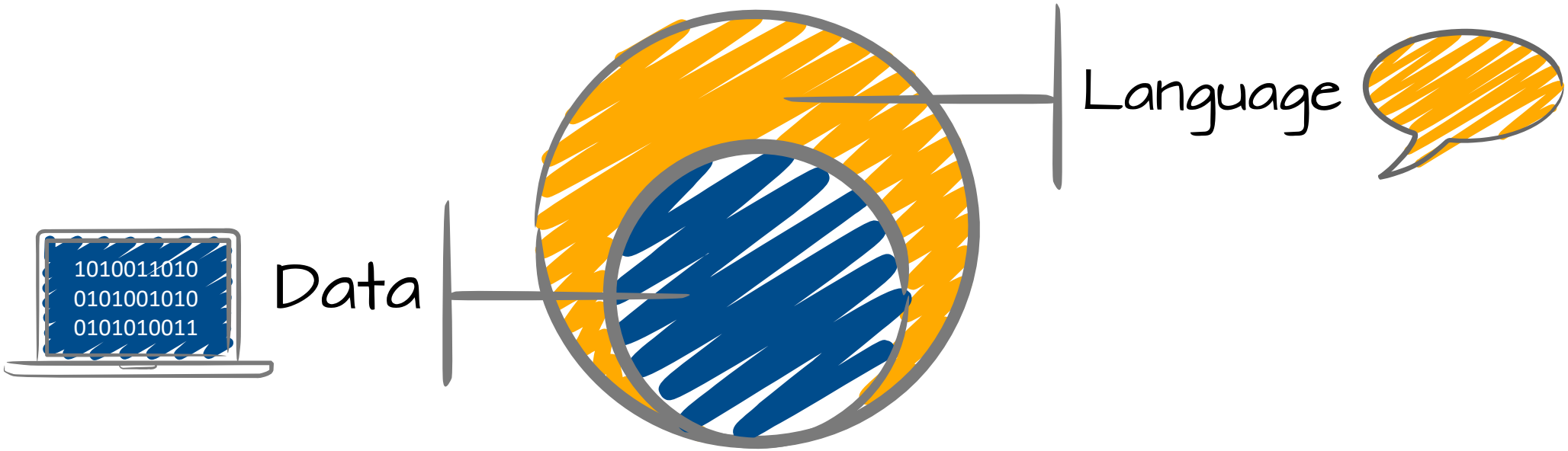
Digital humanities
(Postdoc at CDH)

Computer science




I love you
You love me
I
love
you

Language technology

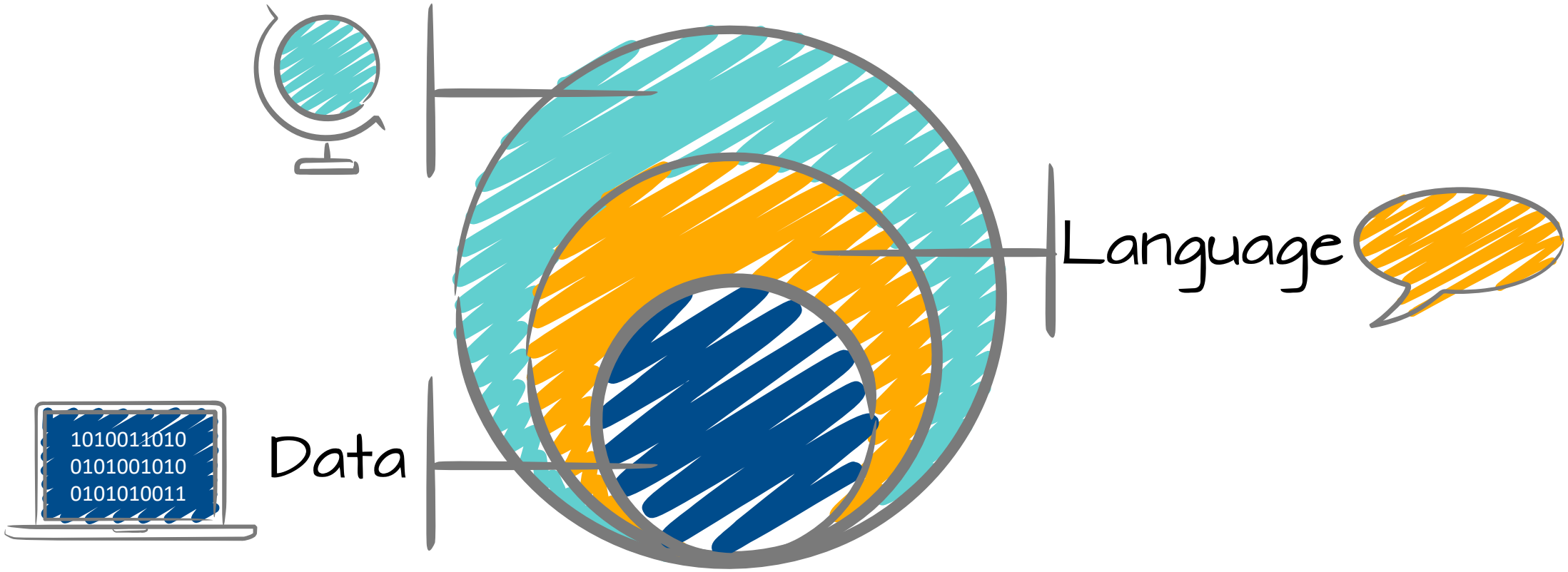



PIENSE
THINK
DENKE
PENSER
$24,000
$77,147
$21,600
Who is Stoker?
$1,000
Who is Bram Stoker?
$ 17,973
$5600


ALPHAGO
00:05:30
Google DeepMind
Challenge Match
LEE SEDOL
00:29:59


IBM
RS6000 SP

Digital humanities



Some terminology

Digital Humanities, Computer Science,

Language Technology

LT ←→ NLP

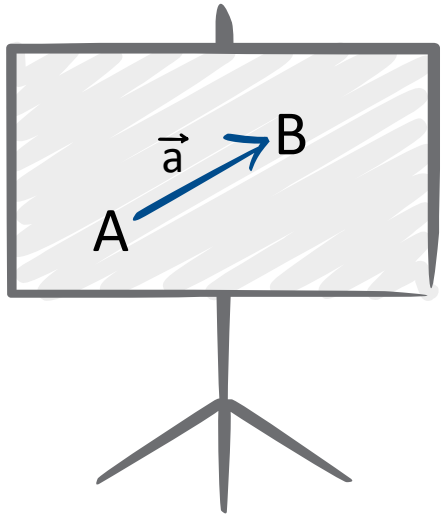
Data Science

Text and Resource

Long-term / diachronic

Token

Some terminology



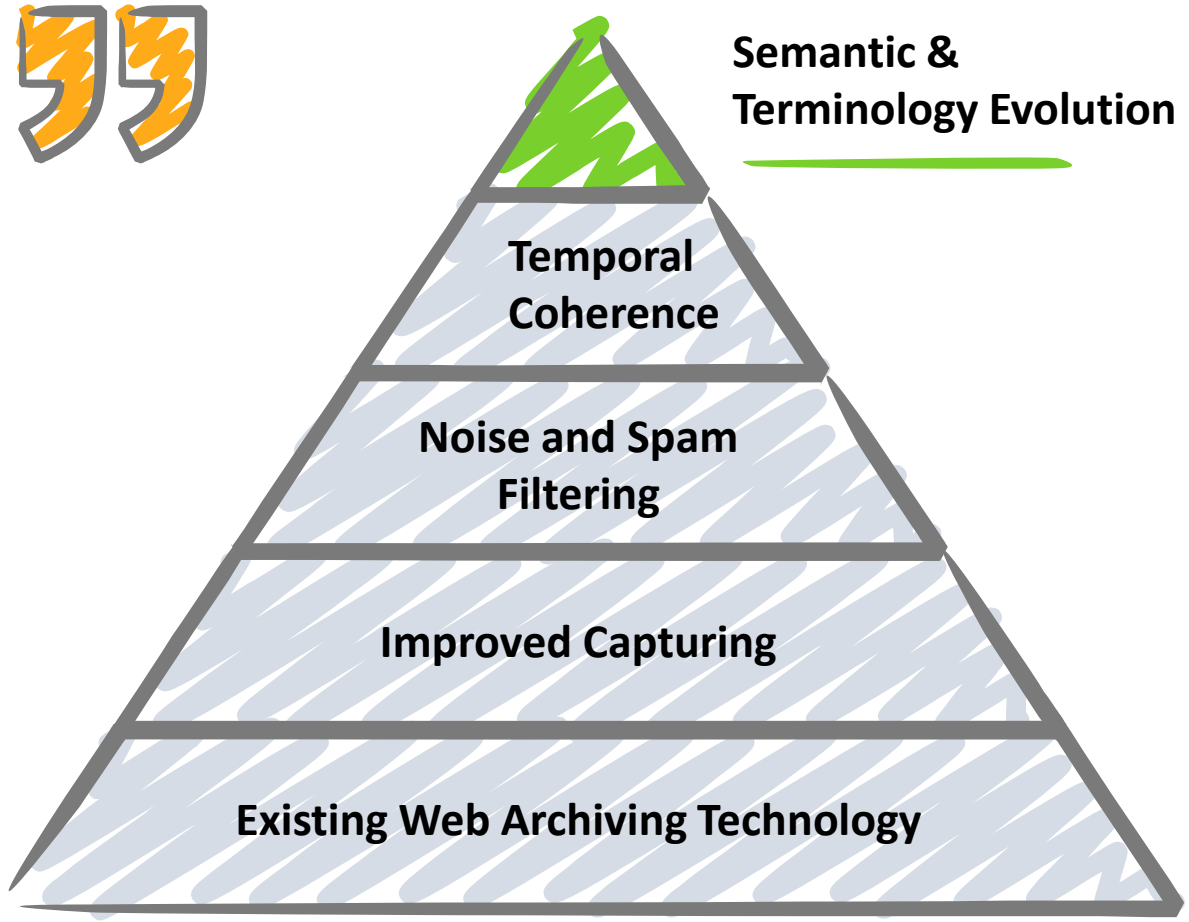
Vector (1, 4, 3) (=3 dimensions)

Topic modeling

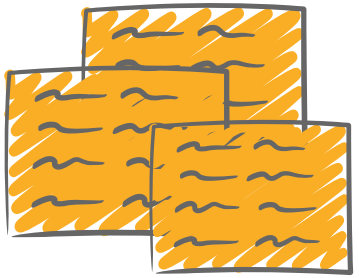
Language changes

LiWA – Living Web Archives

- dealing with terminology evolution
- preparing for evolution aware access support



Increasing amount of historical texts
in digital format

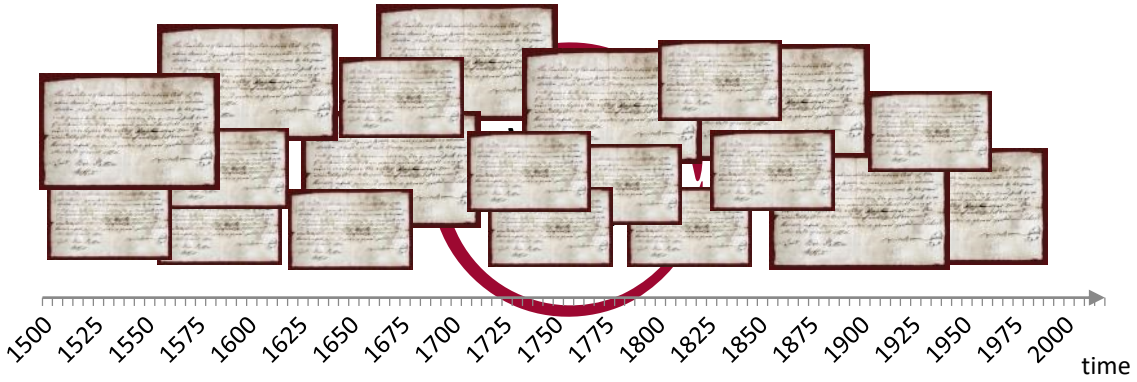


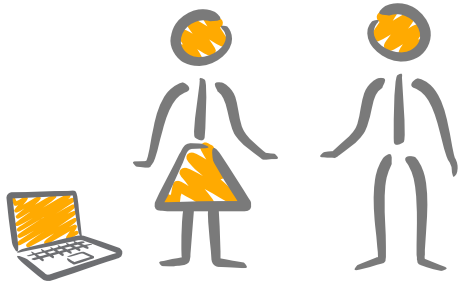
Easy digital access for anyone!
Not only scholars.

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

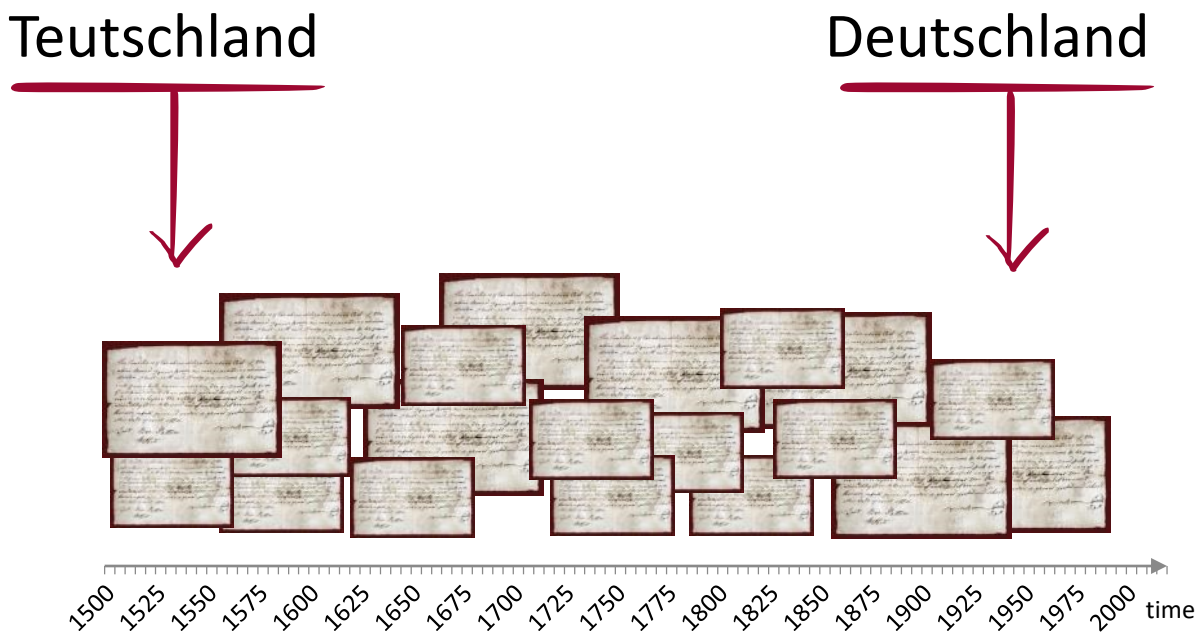
Information from primary sources
Not only modern interpretations.

Text-based
Digital Humanities

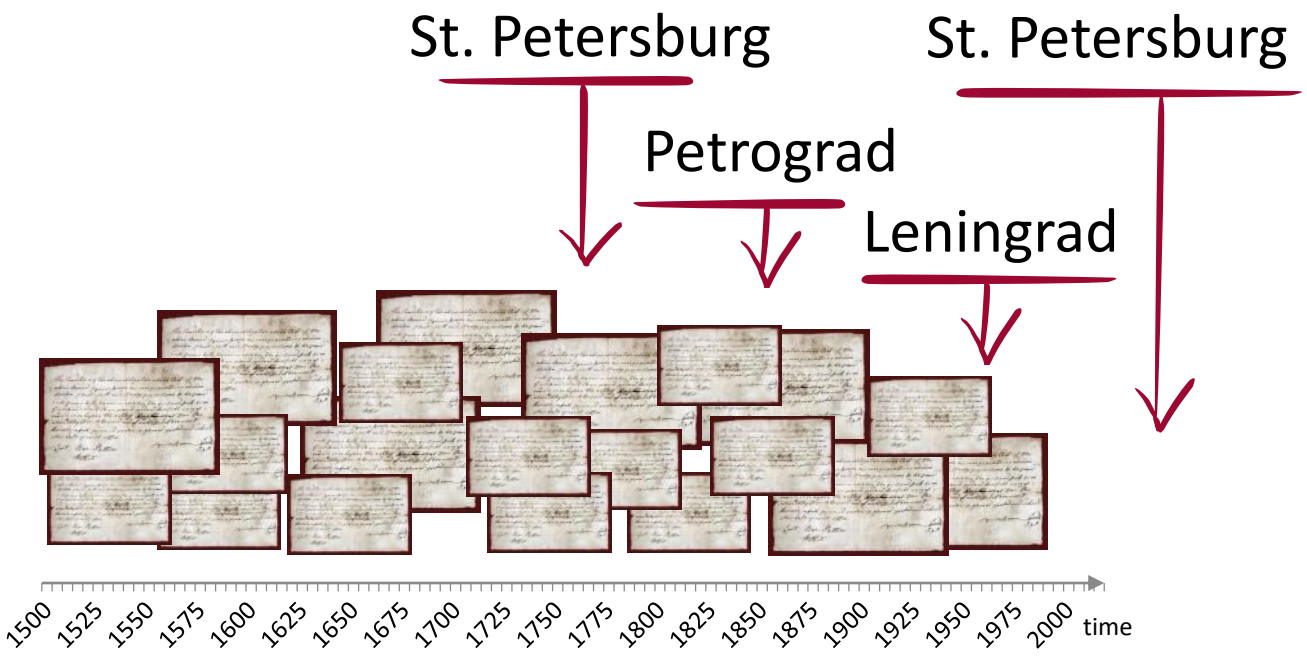


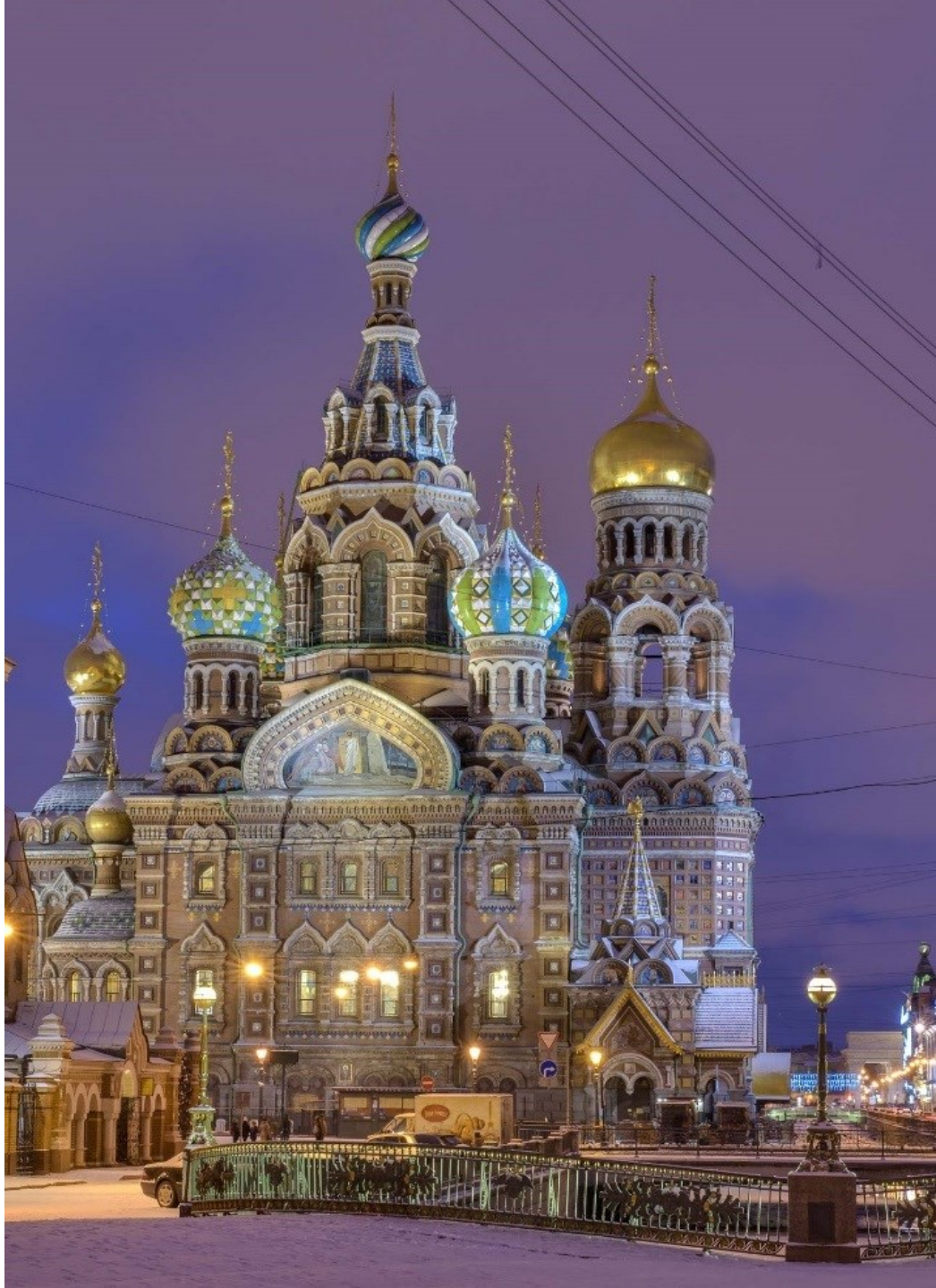


Spelling change

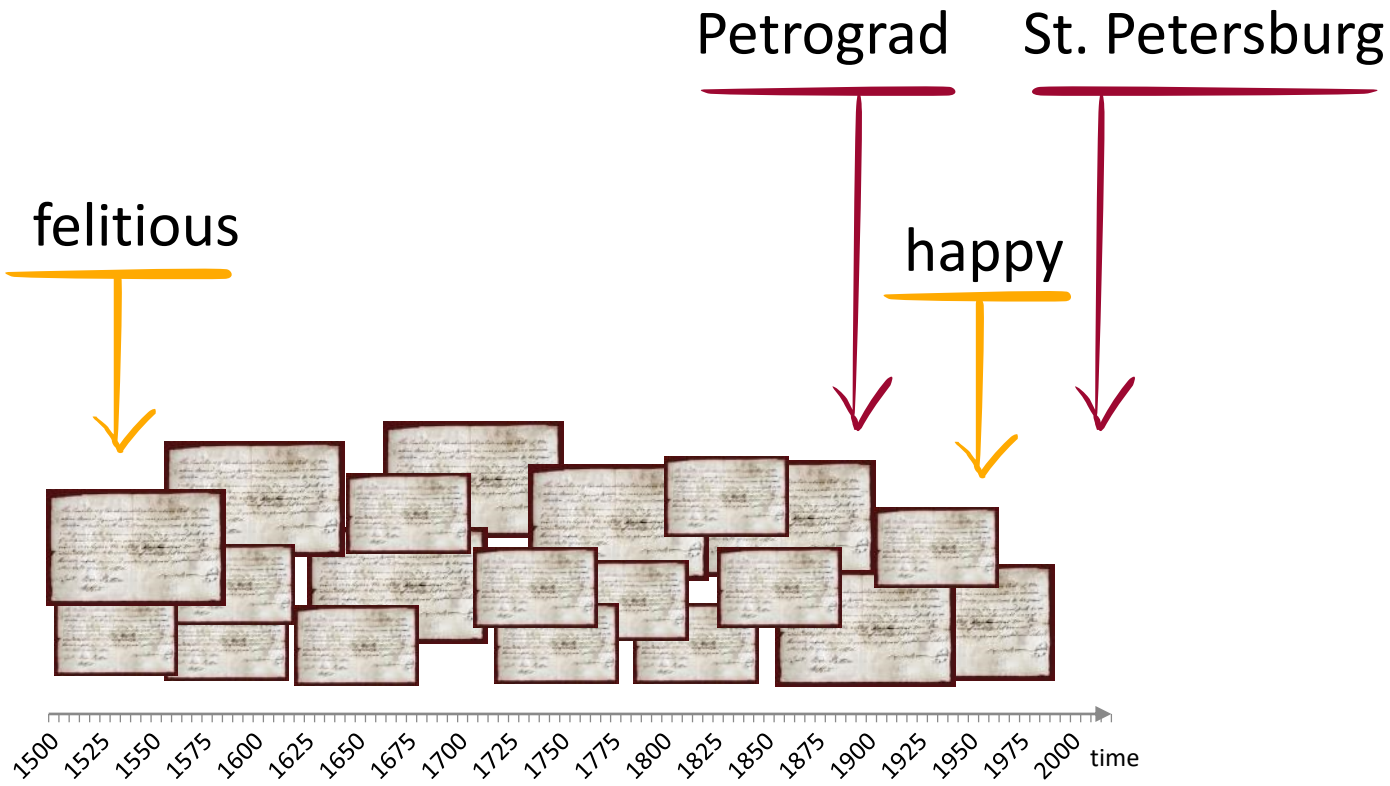


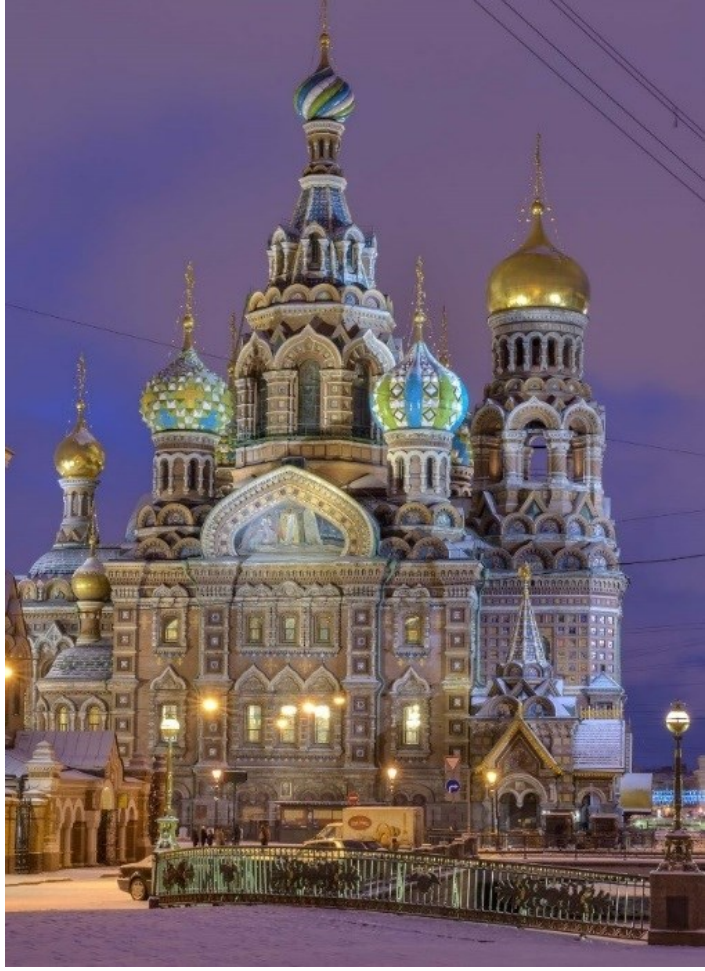
Lexical replacement:
Named entity change





Lexical replacement:





awesome

He was an **awesome** leader!

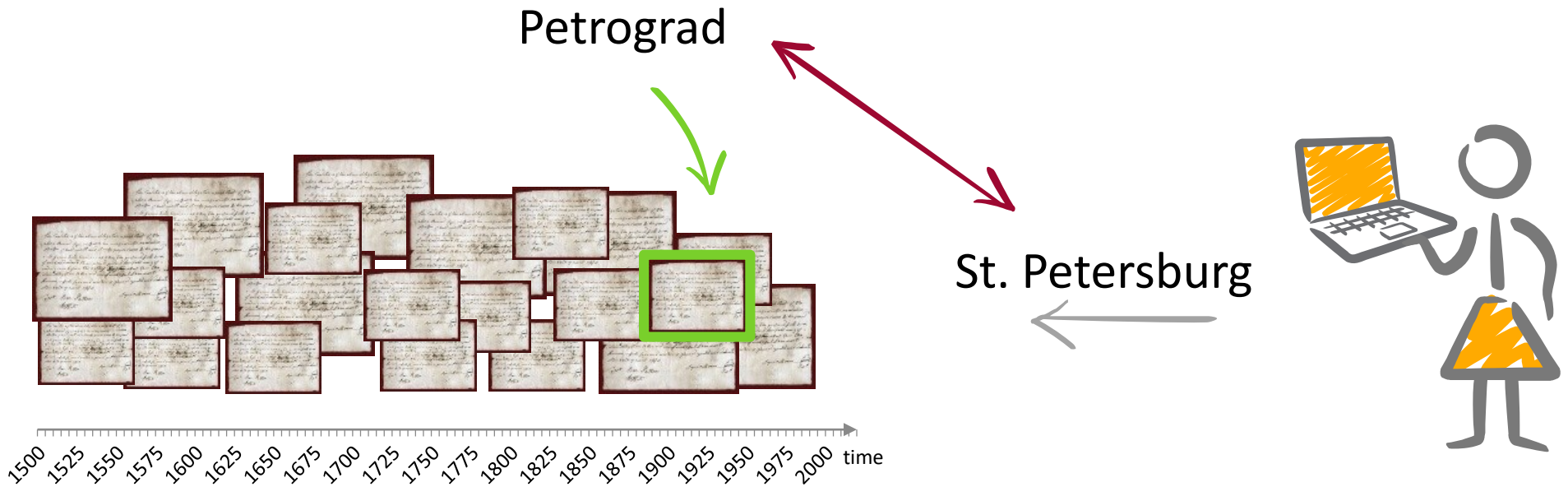
He was an **awesome** leader!

time

Kona → Qwinna → Qvinna → Kvinna

What is the problem?

Finding

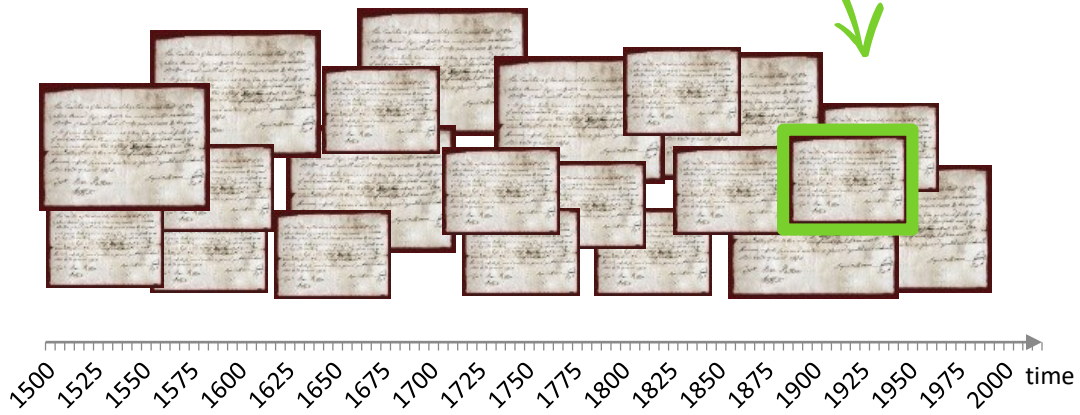


What is the problem?

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**”

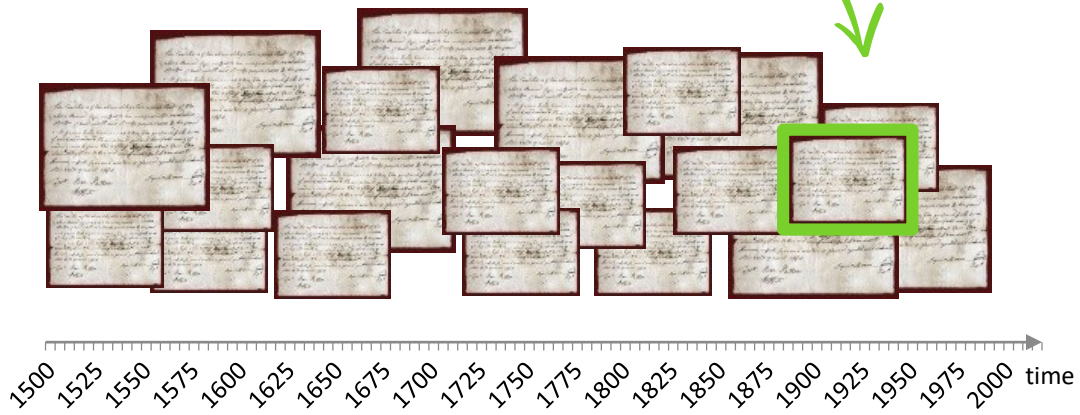
The Times, April 27th, 1787

What is the problem?

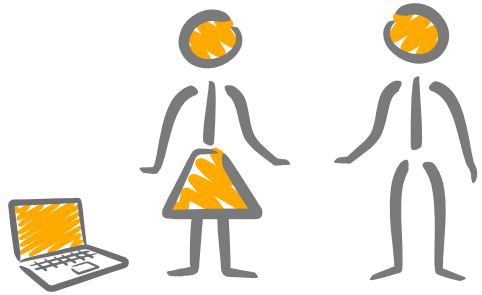
Finding

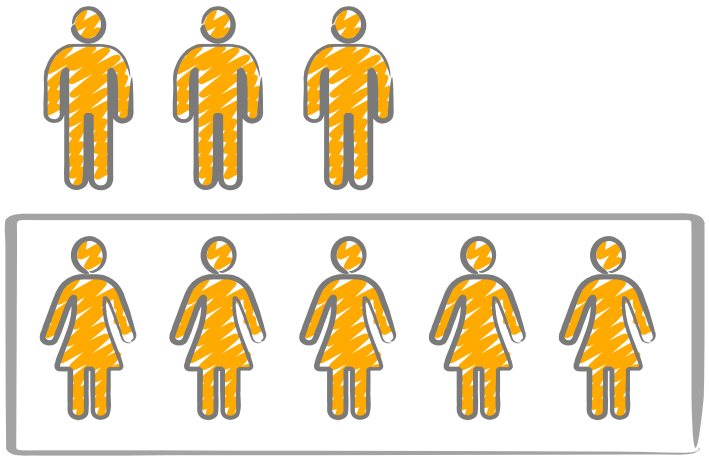
Interpreting

Petrograd



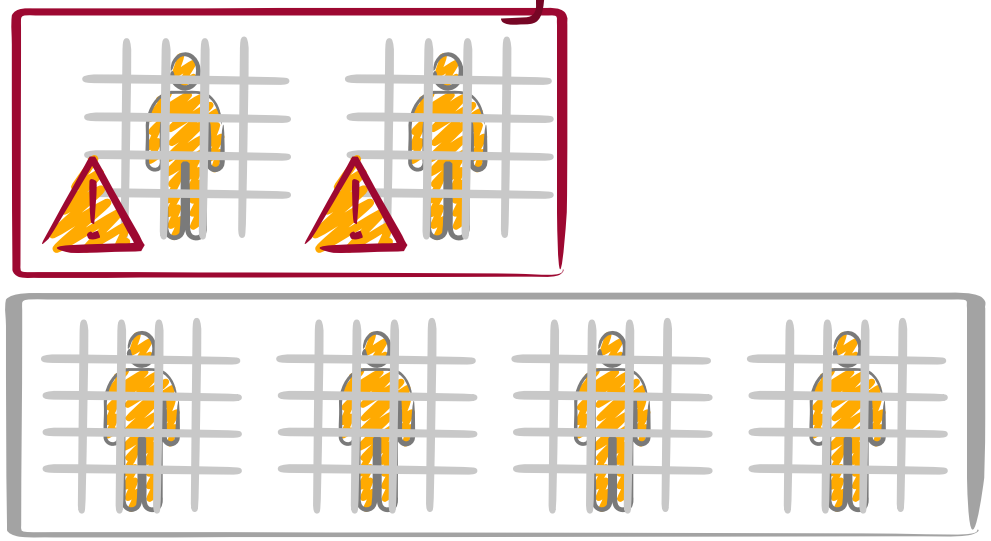
St. Petersburg





girl

Wolf 'varg'

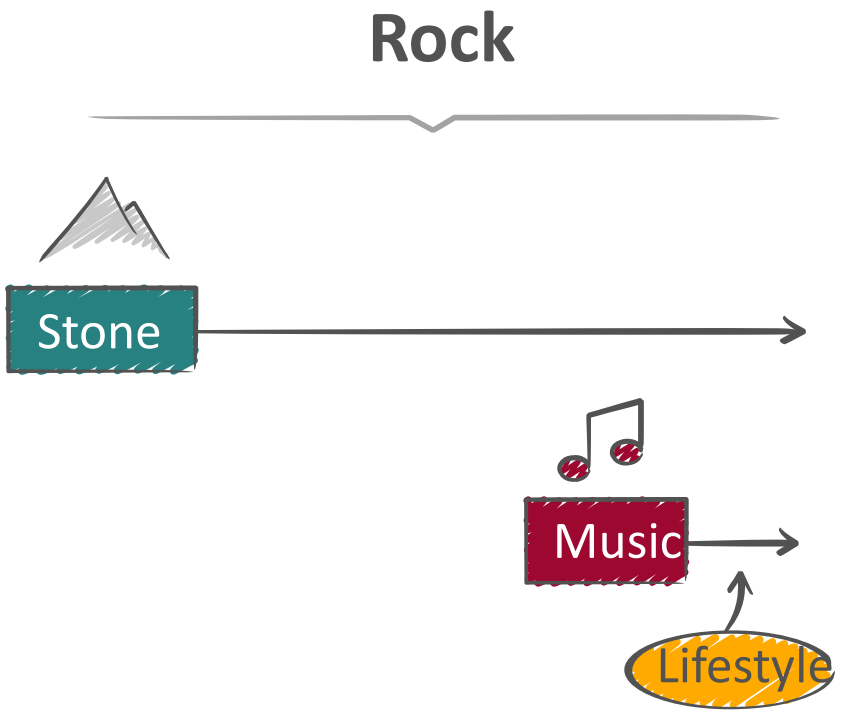


criminal

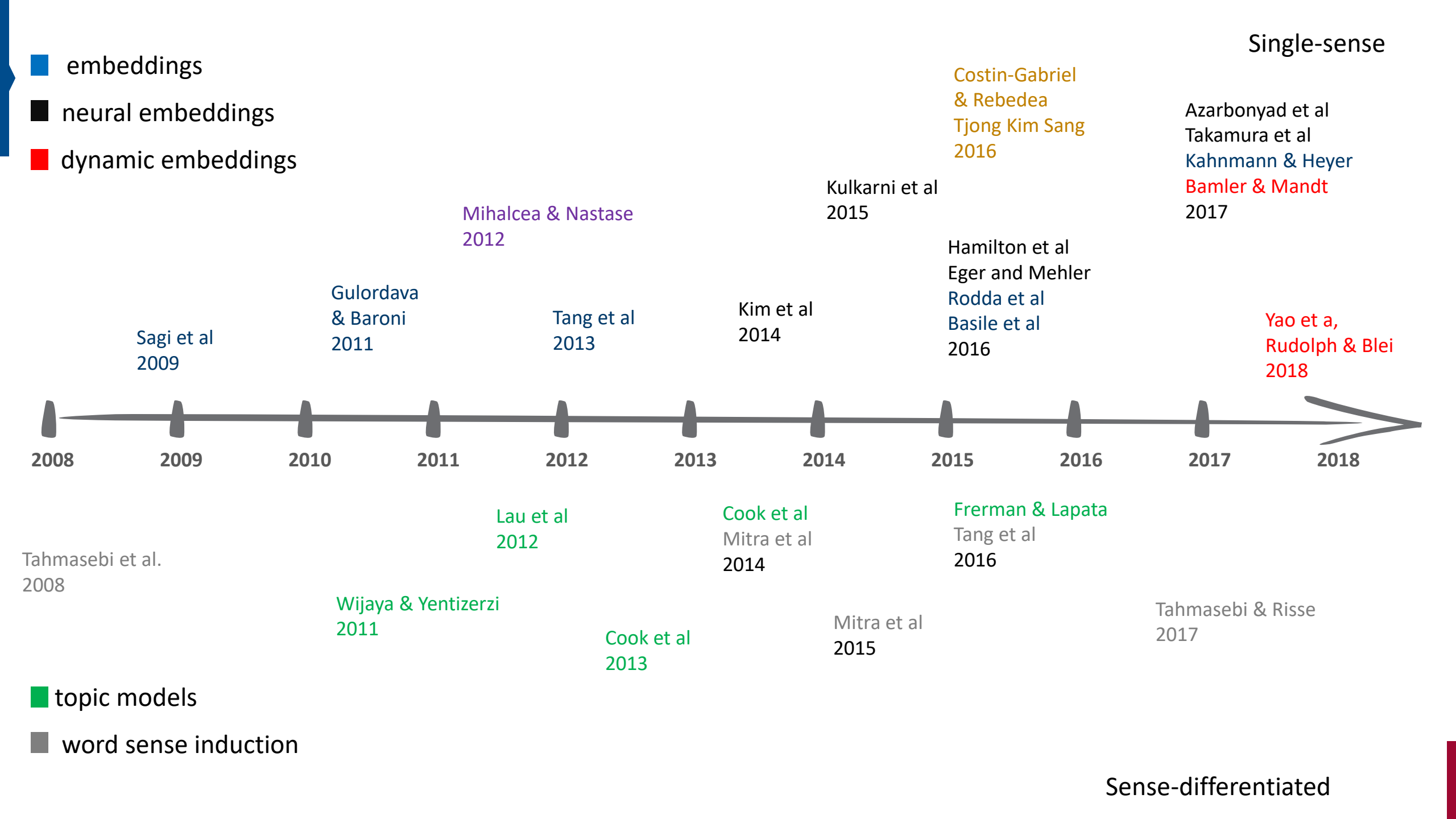
Aims

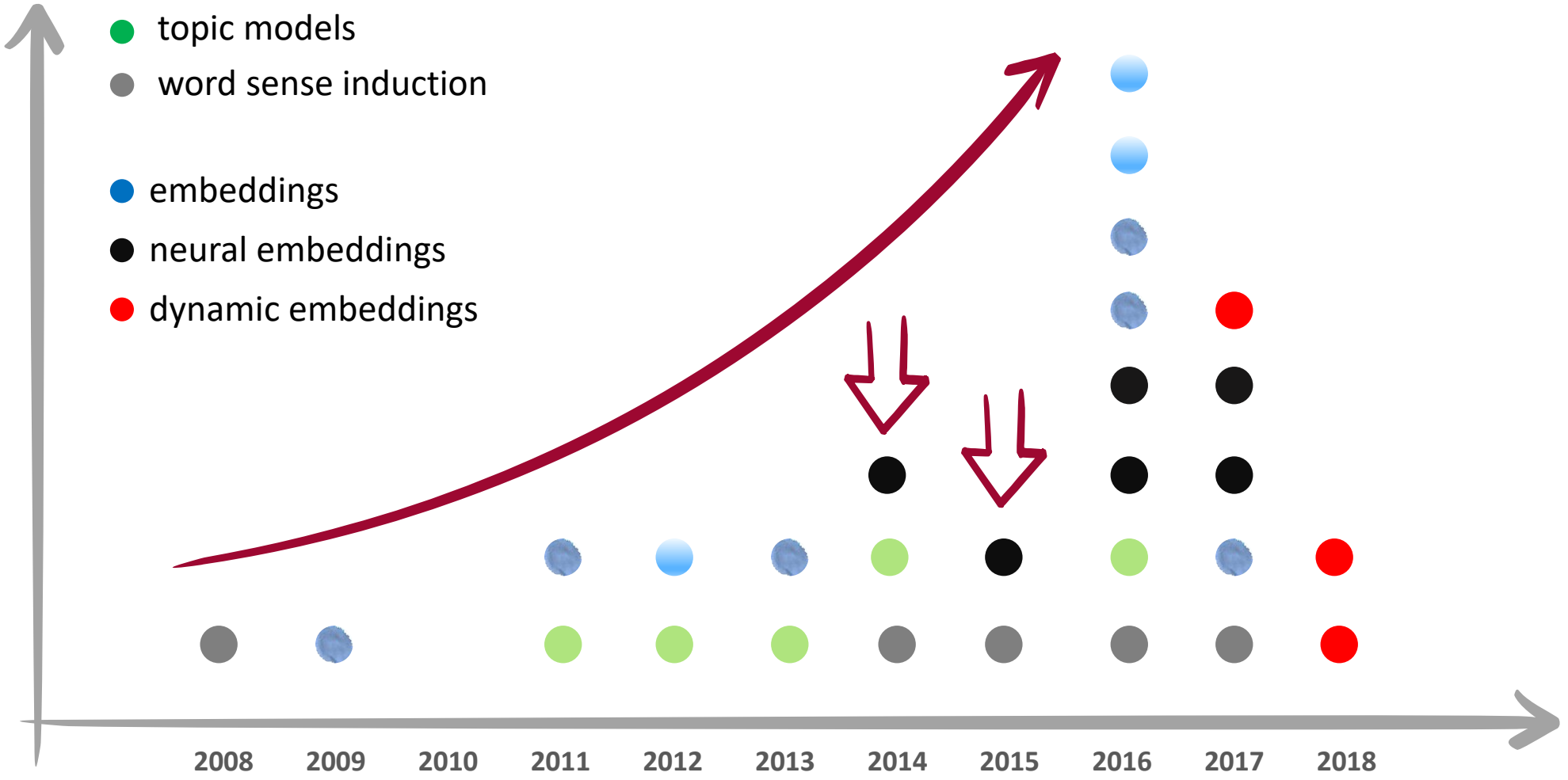
To find word sense changes **automatically** by

1. Modeling word senses
2. Comparing these over time

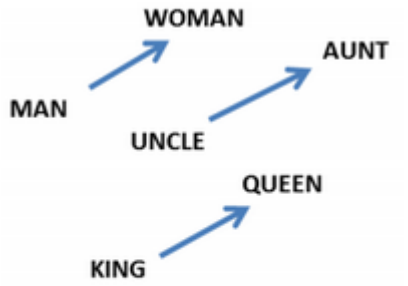


To find **what** changes, **how** it changed and **when** it changed

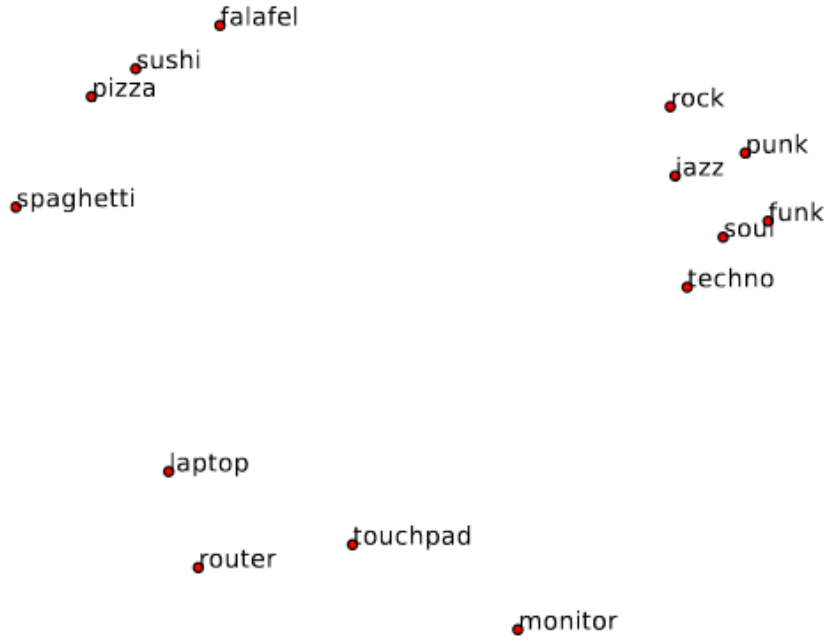

Single-sense
embeddings
neural embeddings
dynamic embeddings
Costin-Gabriel & Rebedea Tjong Kim Sang 2016
Azarbonyad et al
Takamura et al
Kahnmann & Heyer
Bamler & Mandt
2017
Kulkarni et al 2015
Mihalcea & Nastase 2012
Hamilton et al
Eger and Mehler
Rodda et al
Basile et al
2016
Gulordava & Baroni 2011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


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

(Neural) Word embeddings



From Mikolov *et al.* (2013a)



Word embeddings shown in 2D instead of 50-100000

Image: Nieto Pina and Johansson, RANLP'15

Word embedding-based models

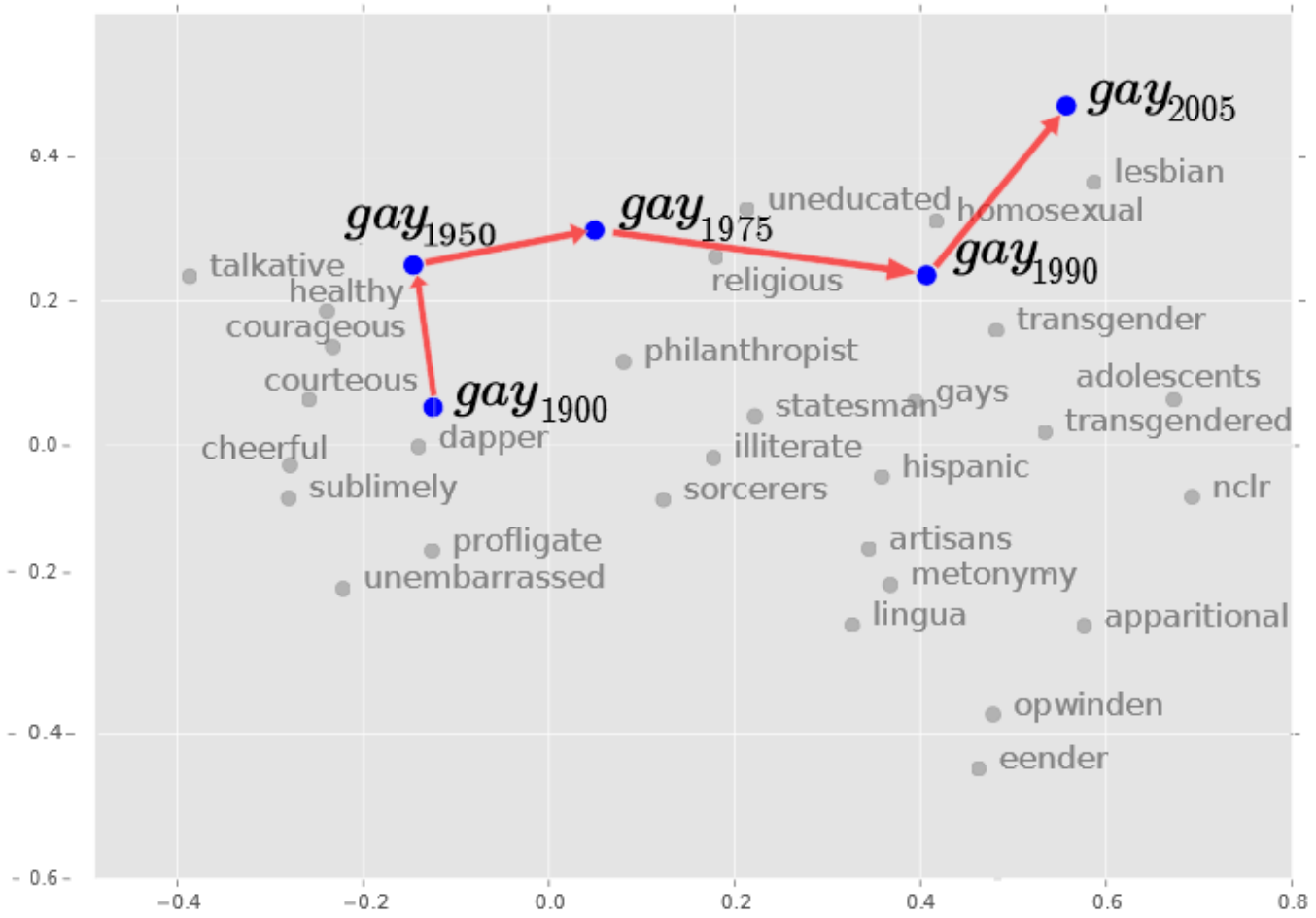


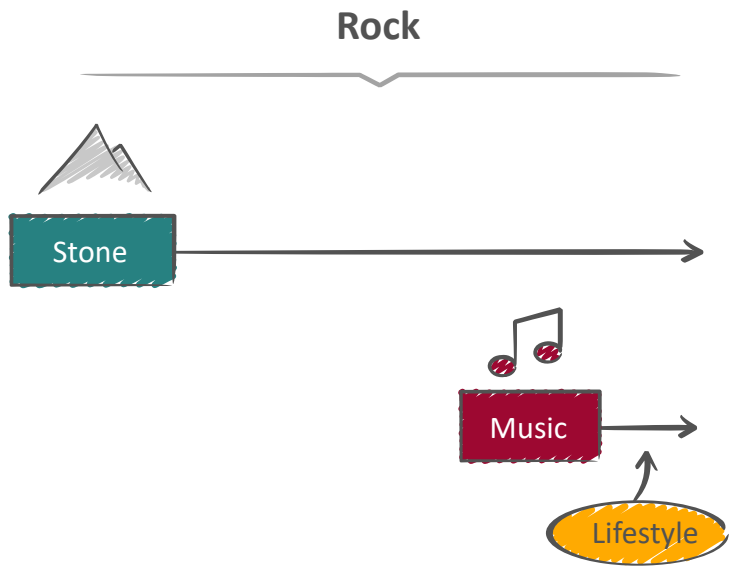
Image: Kulkarni et al. WWW'15

Downsides

Random in
- Initialization
- Order in which the training examples are seen

100 Million tokens per time span*

Typically learn one vector per word
→ Stable/less dominant **senses get lost**!

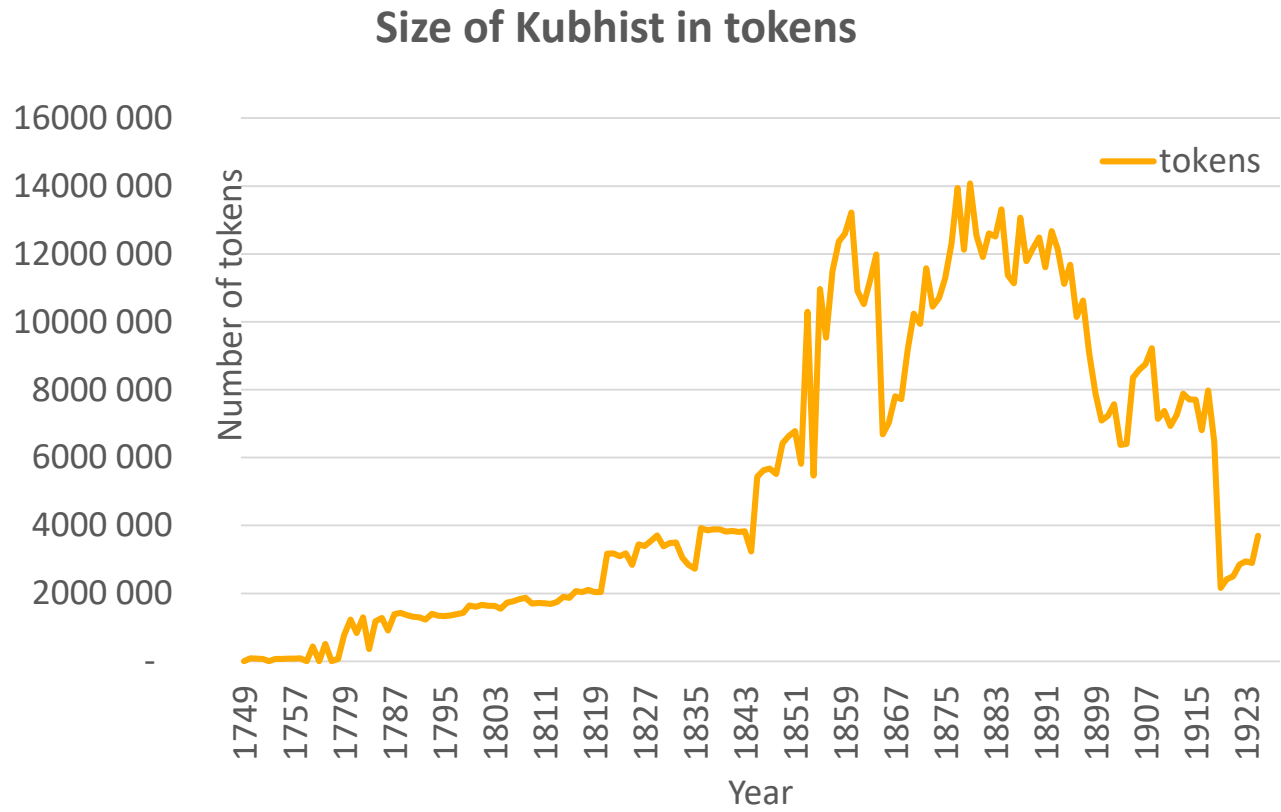


A Study on Word2Vec on a Historical Swedish Newspaper Corpus

Presented at DHN2018

Our study

- **Word2Vec (W2V)**
  a two-layer neural net (skip-gram)
- **KubHist***
  Swedish Newspapers 1749-1925
- **Trained yearly vectors**



* https://spraakbanken.gu.se/korp/?mode=kubhist

What did we do?

11 (10) words over time

nyhet 'news'
tidning 'newspaper'
politik 'politics'
telefon 'telephone'
telegraf 'telegraph'
kvinna 'woman'
man 'man'
glad, 'happy'
retorik ‘rhetoric'
resa 'travel'
musik 'music'

A = {happy, smiling, glad}
B = {happy, joyful, cheerful, excited}

Overlap = 1

Unique = 3+4-1 = 6

Jaccard similarity = 1/6

Some results I

Woman:

1912: 'kvinna': [valbarhet, valrätt, rösträtt, självförsörjande, sexuell, okunnig, högerparti, politisk, radikal, vänsterparti]

1908: 'kvinna': [österåsen, ung, **rösträtt**, ljusglimt, flicka, iförda, knäböjande, begåfvad, värnlös, jubla]

1895: 'kvinna': [qvinna, varelse, människa, öfvermåttan, flicka, reptil, gosse, förälskade, öfvergifven, högväxt]

1879: 'kvinna': [qvarlefva, vålnad, öfvade, rättskaffens, begåfvade, skenbart, skummande, vilde, **herskar**, mygga]

1868: 'kvinna': [piller, kvilken, mis, kade, klo, nde, äock, reään, äsom, bvilken]

1867: 'kvinna': [äes, kvrk, kunäe, mle, näo, nuvaranäe, äer, v«r«, uä, äig]

Some results II

Politics:

1925: 'politik': [näring, trygghet, kamp, arbetarrörelse, konservativ, nationell, strävan, europa, neutralitet, önskad]

1922: 'politik': [åskådning, socialistisk, ägnad, demokrati, utrikespolitisk, sakligt, situation, representativ, auktoritet, ärlig]

1900: 'politik': [enig, bvad, finlands, politisk, konstitutionel, revolution, armenien, citera, civiliserade, dementi]

1872: 'politik': [republikansk, opposition, kränka, reaktionär, neutral, republikan, tillbakavisa, changarniers, påfvedöme, horace]

1858: 'politik': [**asylrätt**, allians, frankrikes, konstitutionell, konflikt, försonlig, rysslands, press, makt, fördrag]

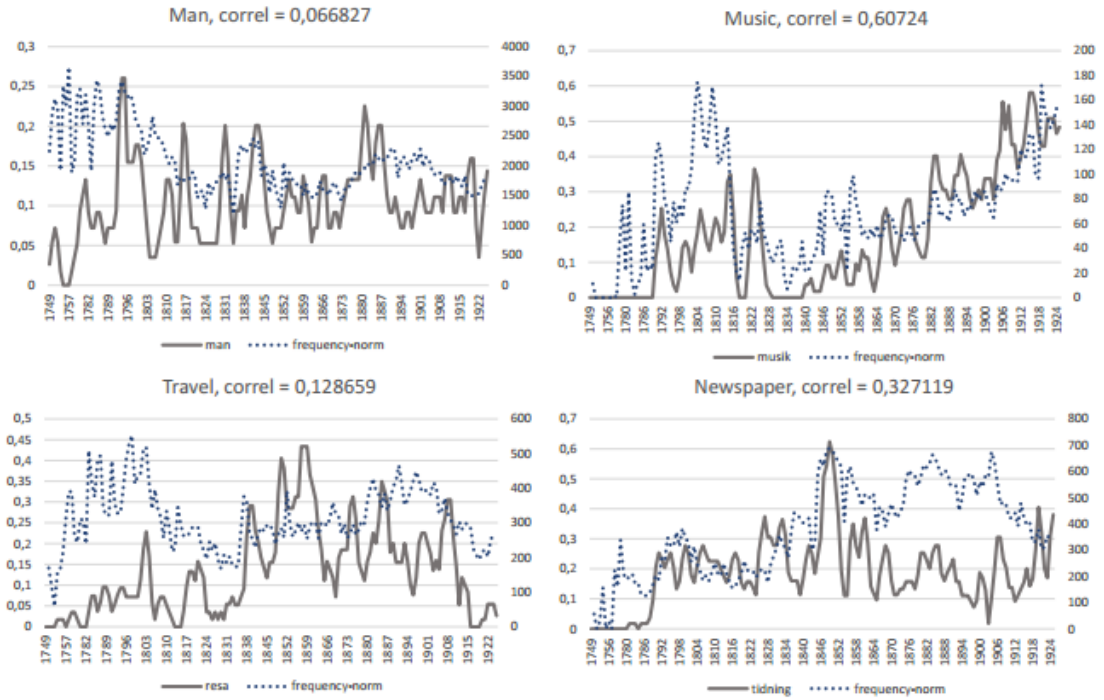
1844: 'politik': [tadla, allians, vägran, irländsk, frankrikes, bemedling, tribun, segra, ministeriell, fördrag]

Result summary

The more frequent the term, the more stable the vectors

0.11-0.19 overlap between years

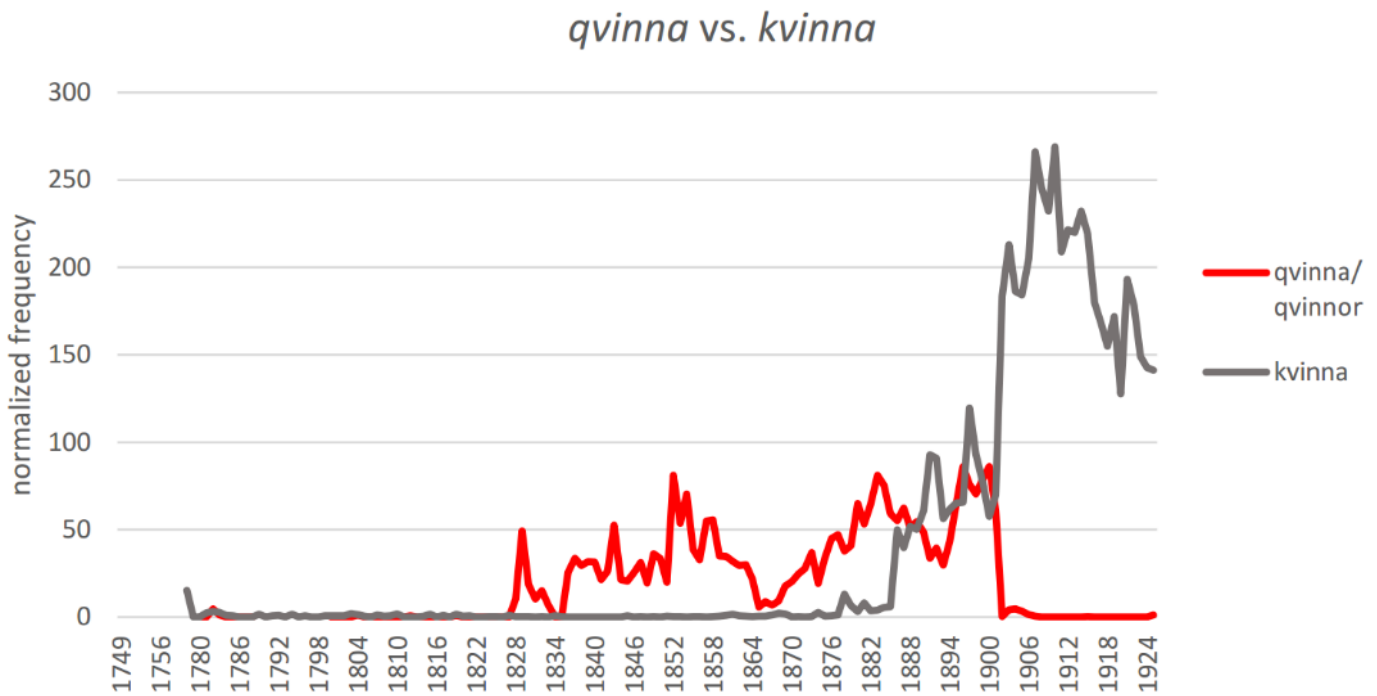
2-3 words in common each year



Next step

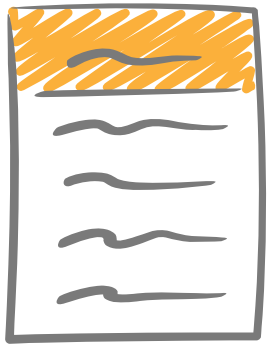
OCR errors

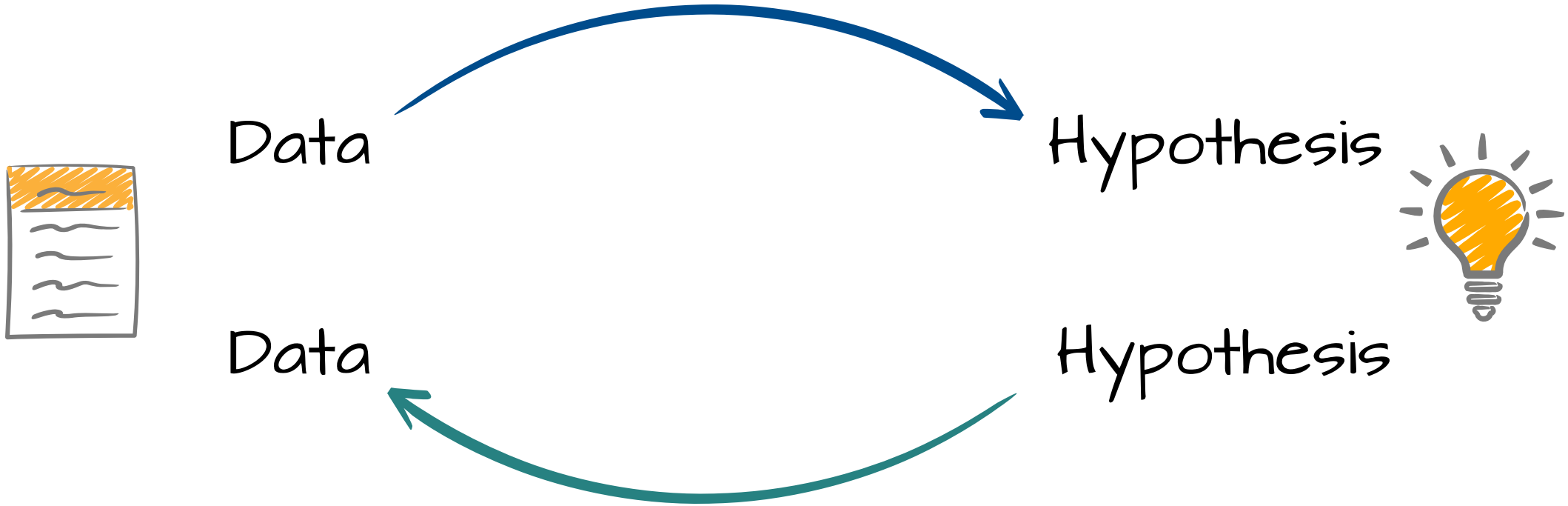
Spelling normalization



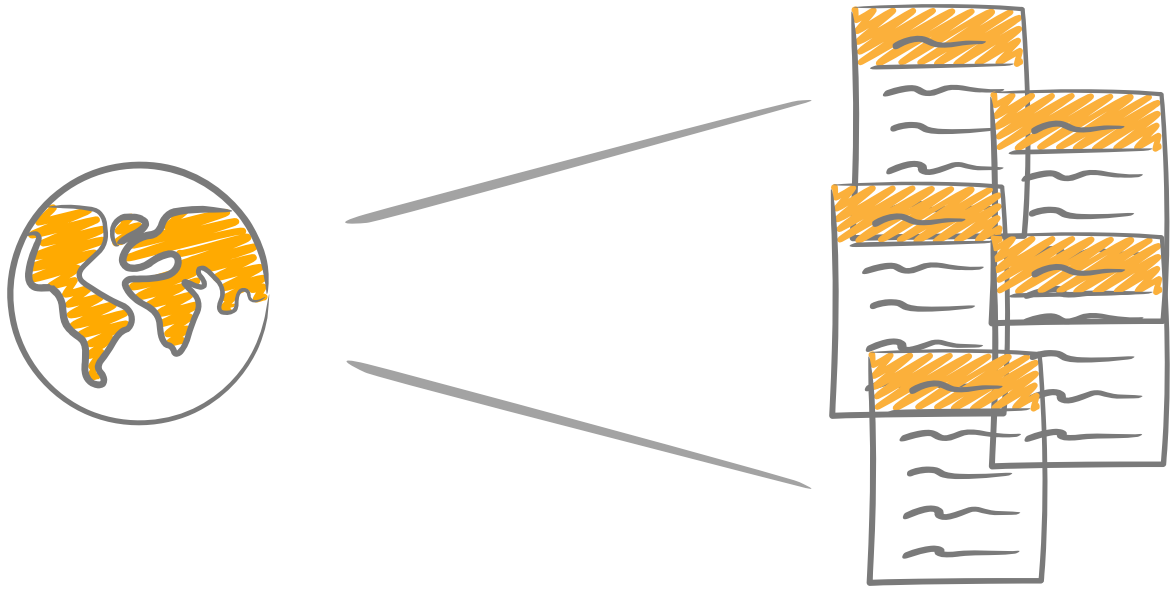
Research methodologies in DH

Digital, large-scale data

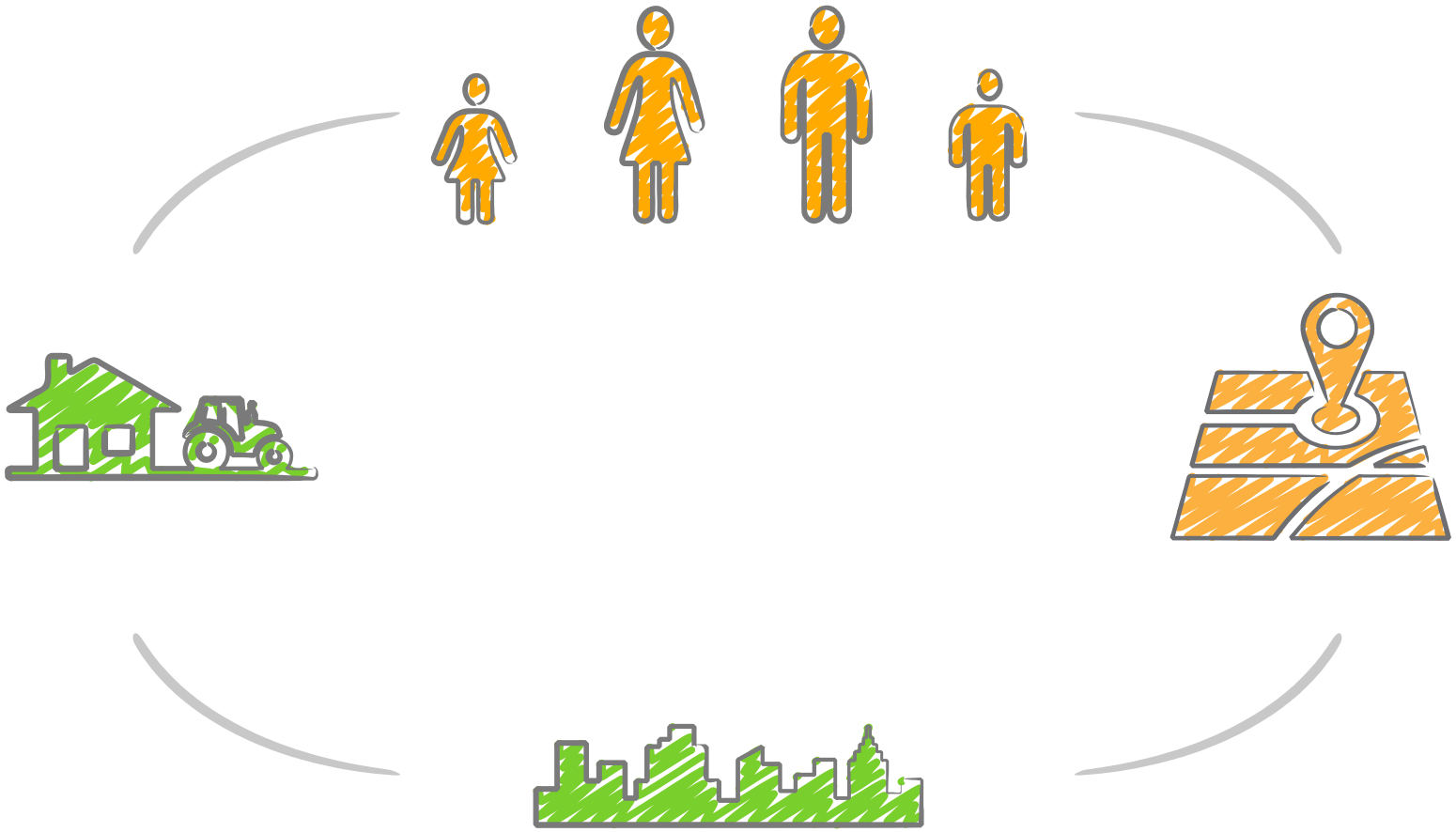



Data
Hypothesis
Data
Hypothesis

Questions



Representativeness



The Street light effect

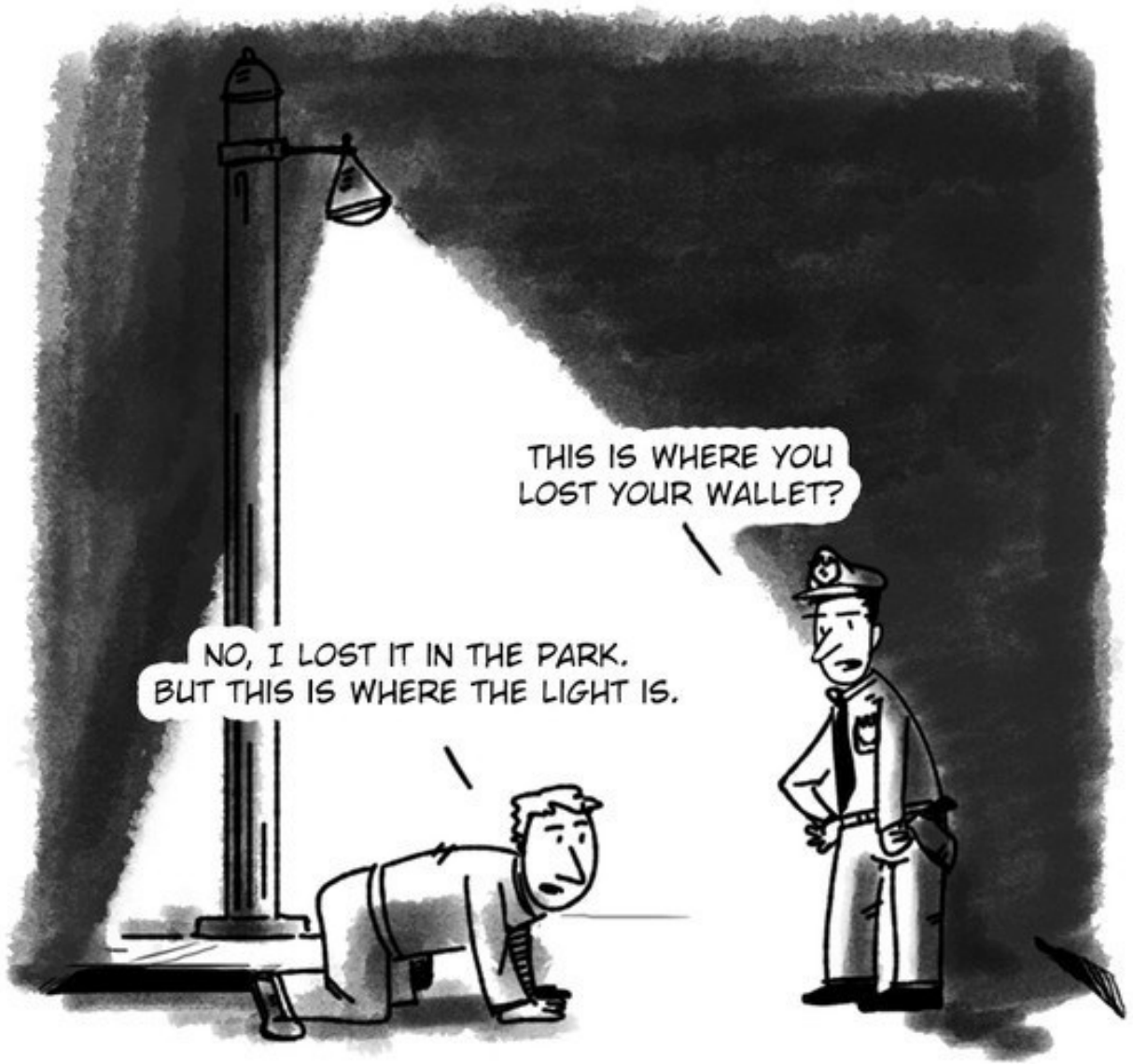
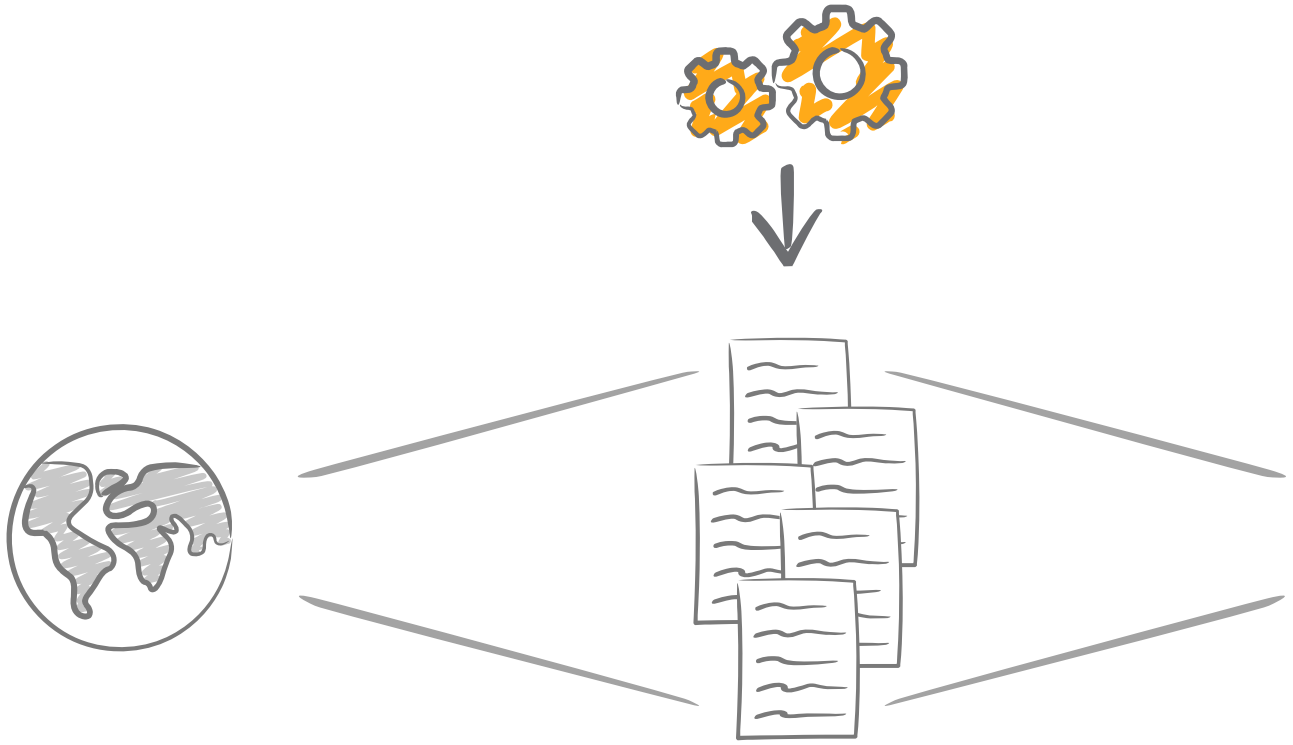
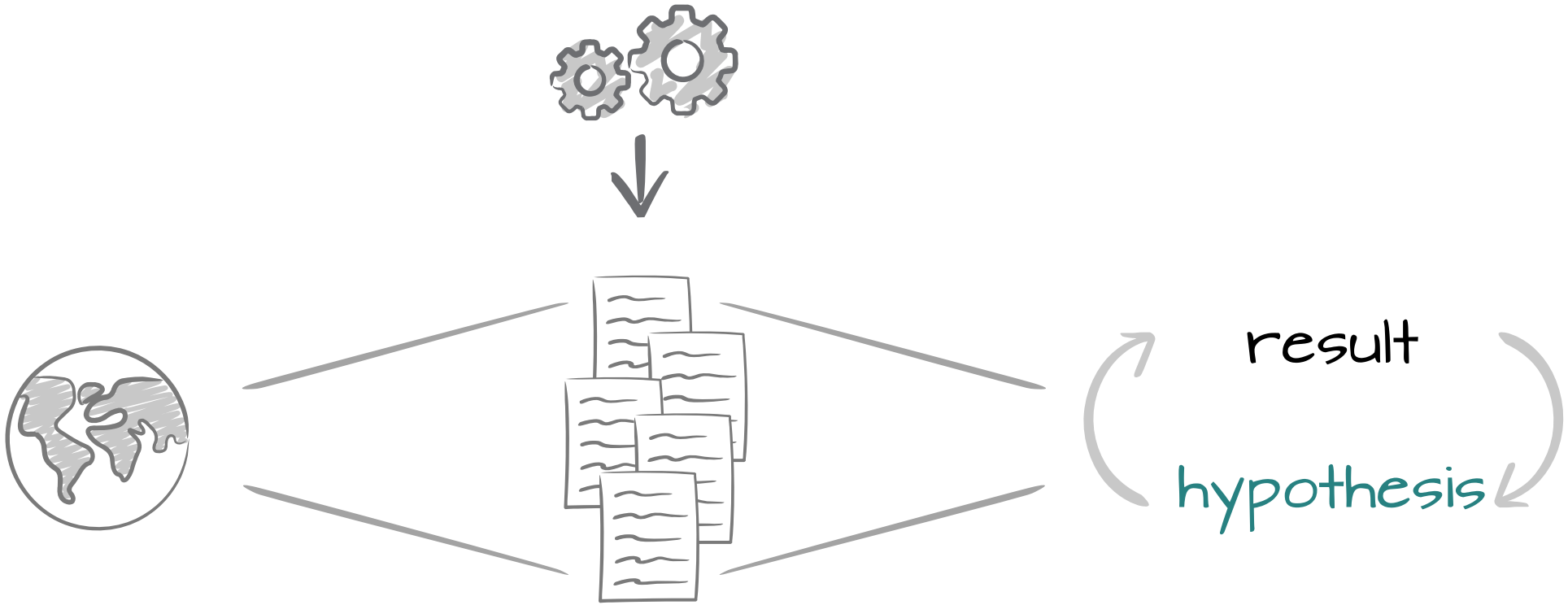


Image: http://first-the-trousers.com/hello-world/

method + data = results



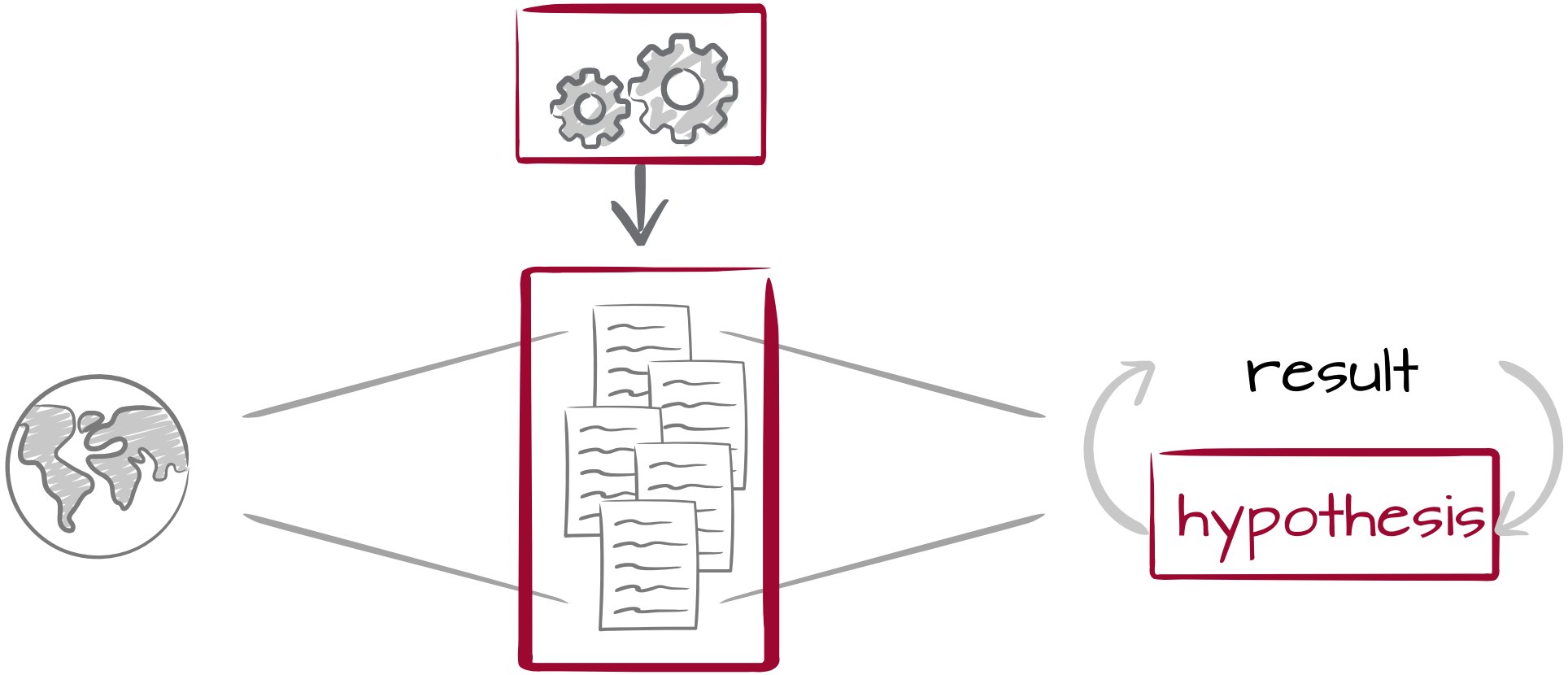
result


result
hypothesis

1 Data

2 Method / Preprocessing

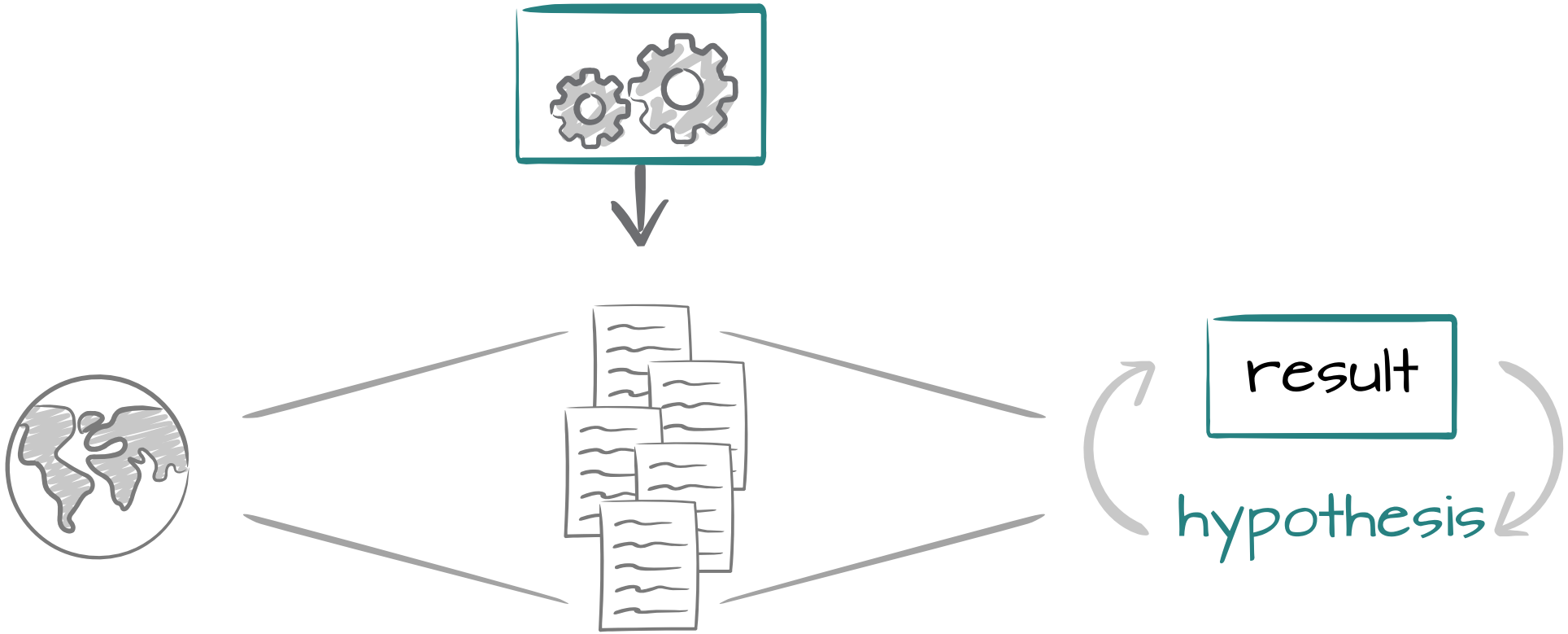
3 Hypothesis



Reject

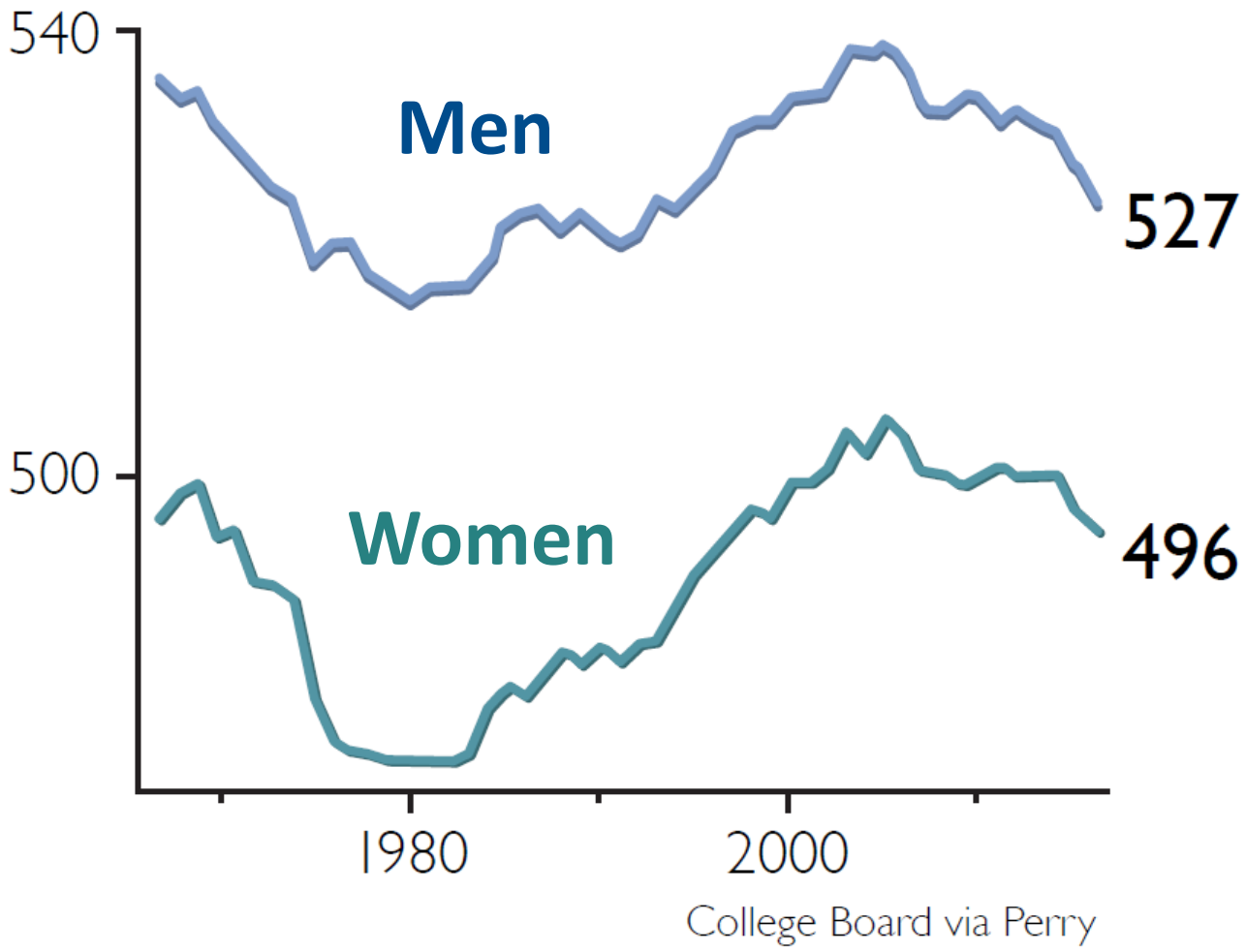
1 Method

2 Correct interpretation of the results



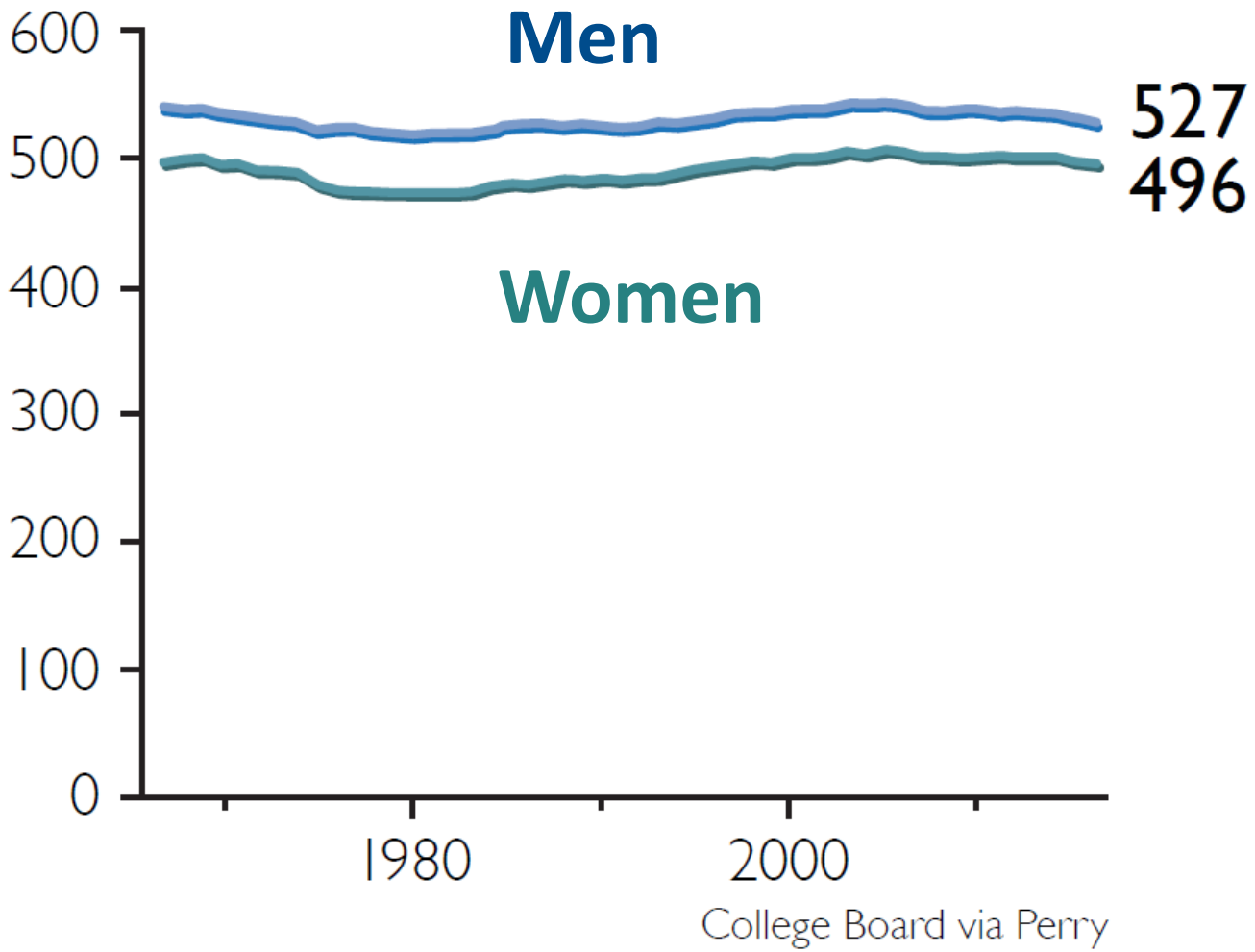
Accept

Math results, average difference



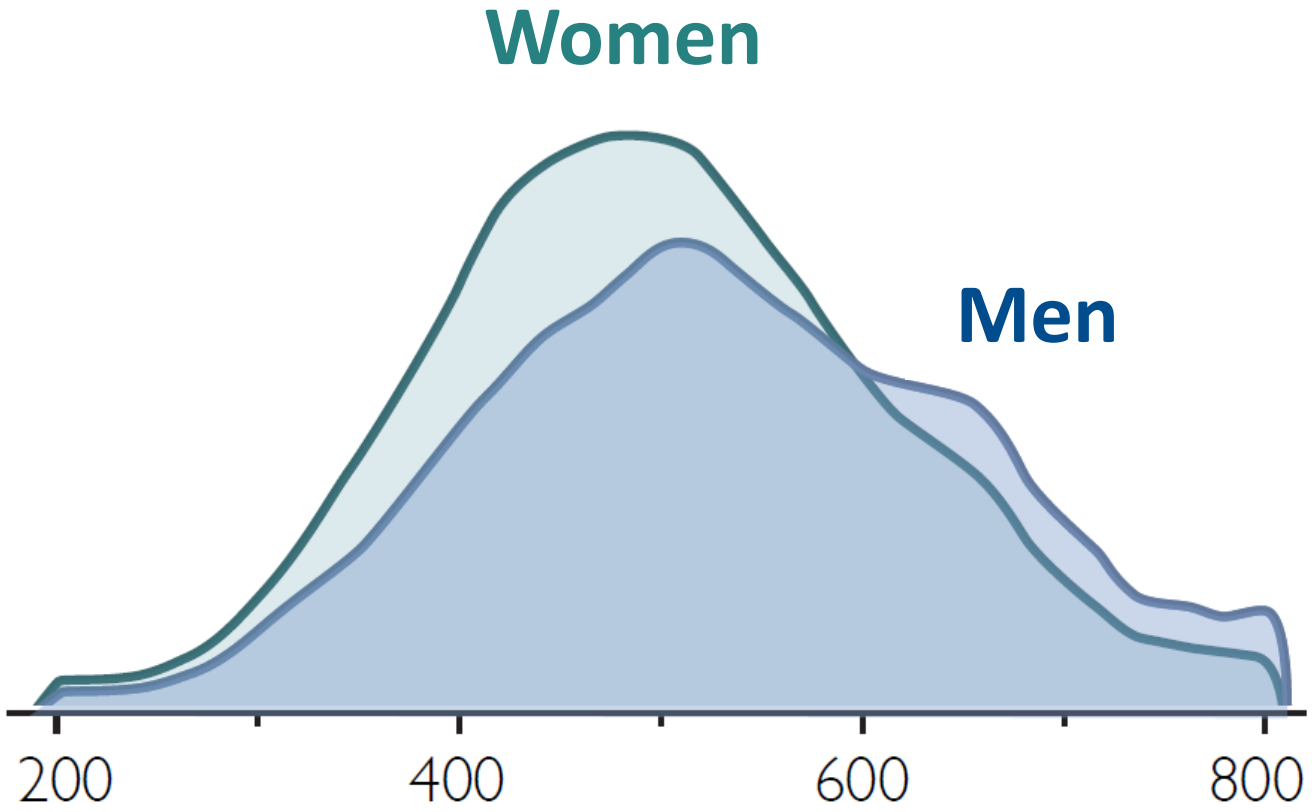
Source: Factfullness

Math results, average difference



Source: Factfullness

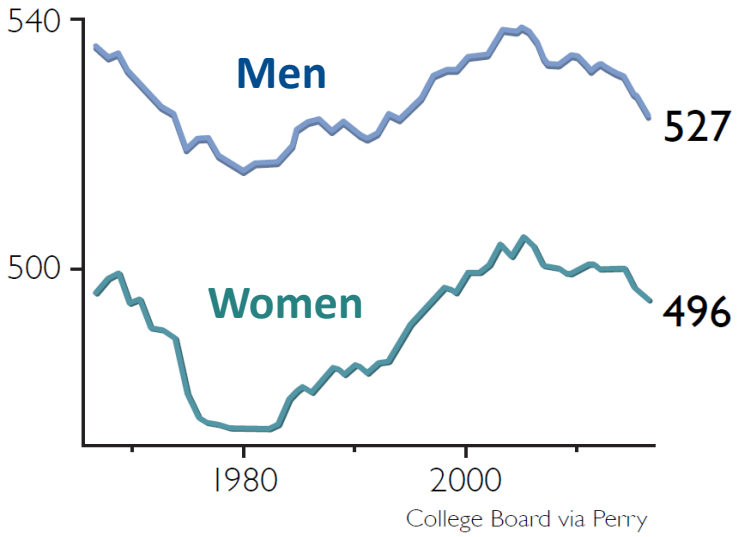
Range of math scores

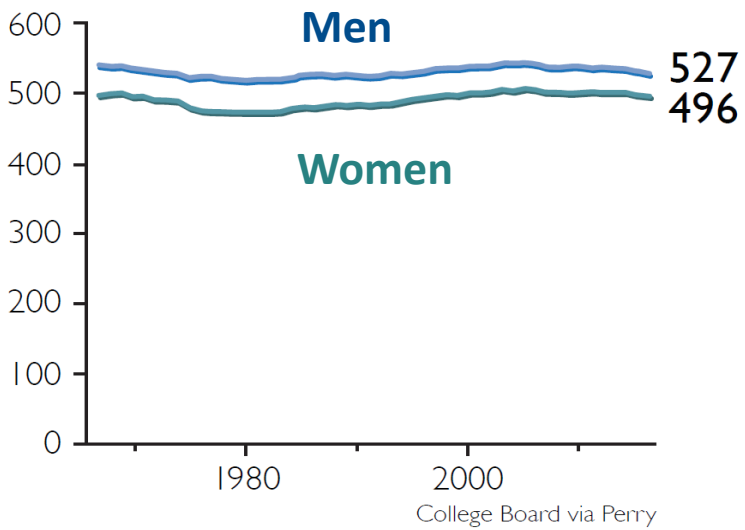


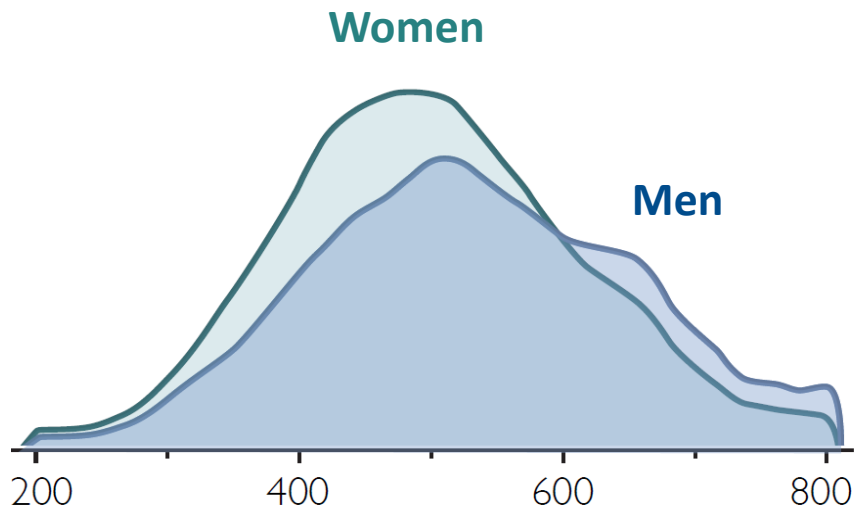
NUMBER OF INDIVIDUALS WITH DIFFERENT MATH SCORES 2016

Source: Factfullness

Comparison of the same data







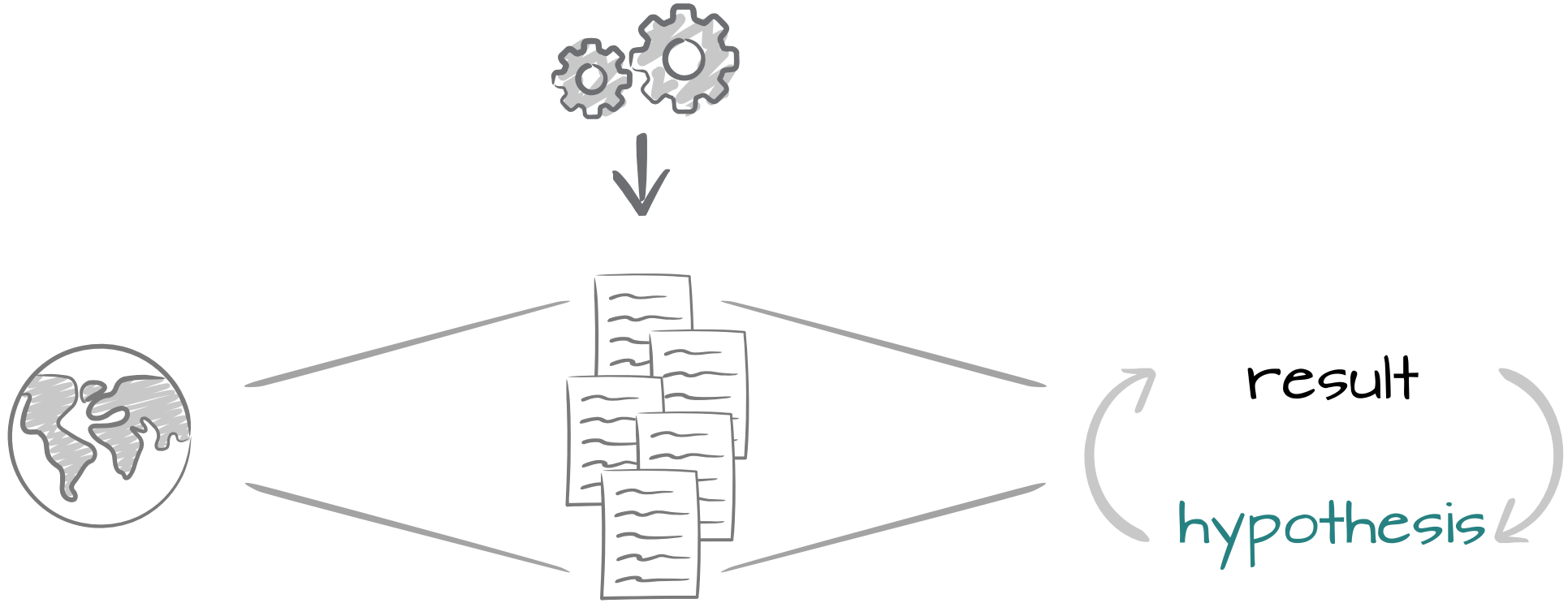
NUMBER OF INDIVIDUALS WITH DIFFERENT MATH SCORES 2016

Source: Factfullness

1 Method

2 Correct interpretation of the results

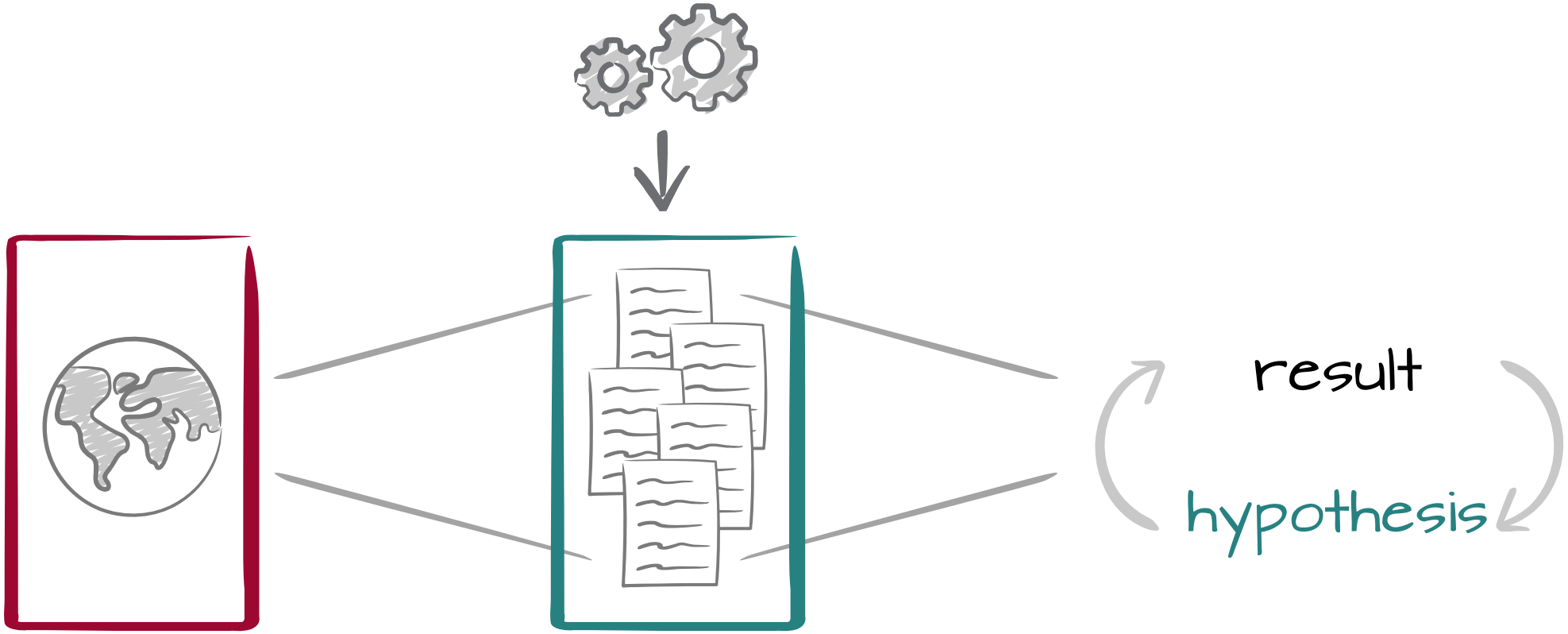
3 Where do the results live?



1 Method

2 Correct interpretation of the results

3 Where do the results live?



NLP pipeline: From text to result

Text-mining method

Dimensions

Filtering: Function words

Filtering: Stopwords

Part-of-speech tagging

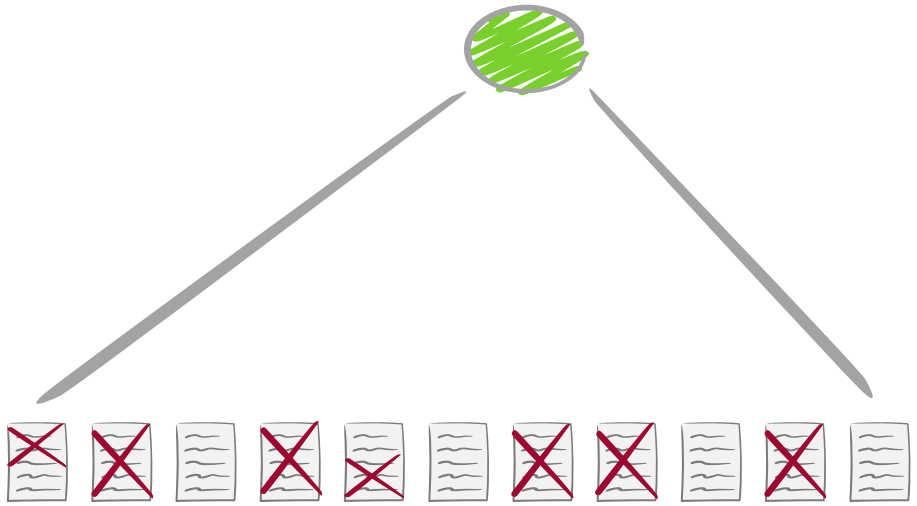
Lemmatization

Tokenization

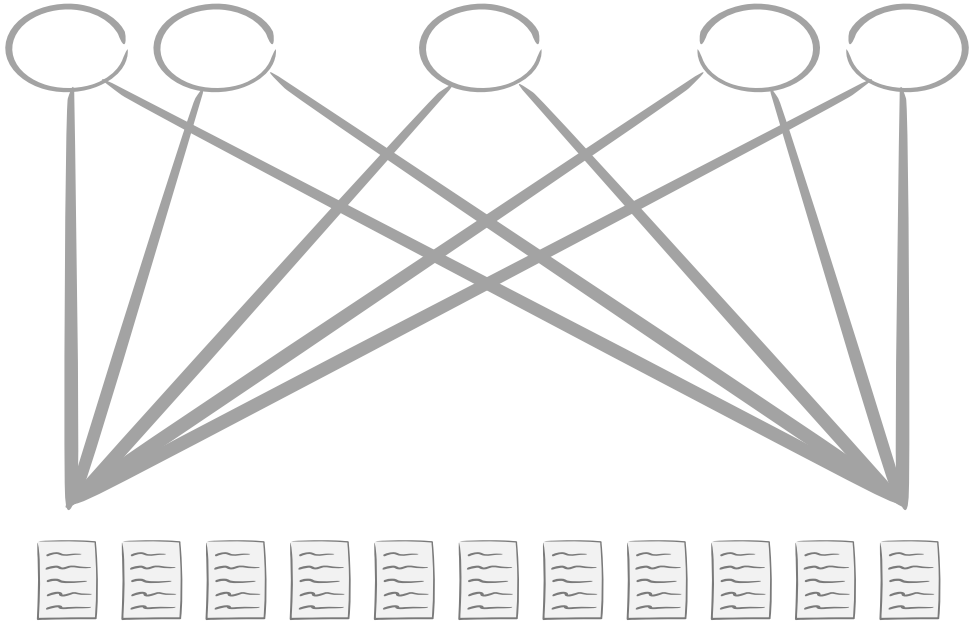
I like the room but not the sheets.

I like	room	sheets.	**(after stop word filtering)**
I like	room	sheet.	**(after lemmatization)**
	room	sheet.	**(only nouns)**
	room		**(frequency filtering)**
like			**(only verbs)**

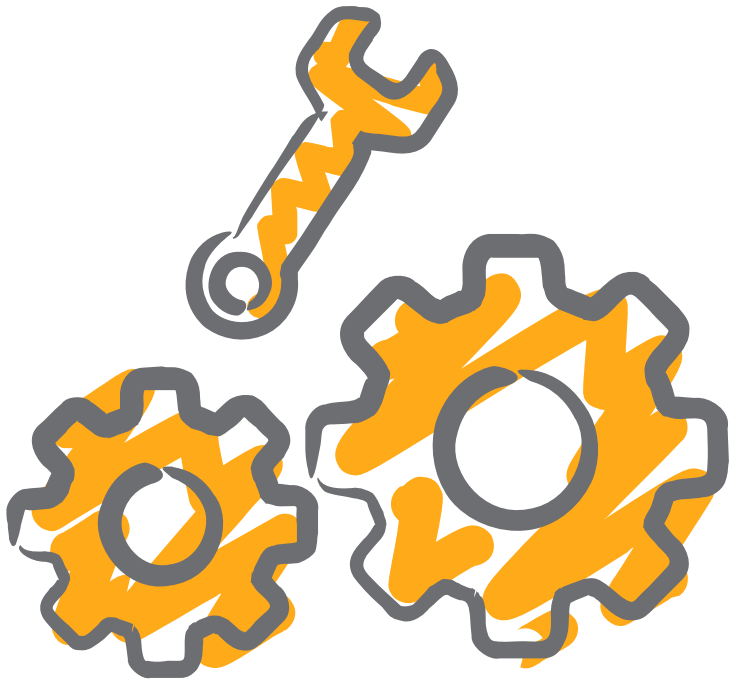
Viewpoint on the data



Viewpoint on the data



Choosing a method









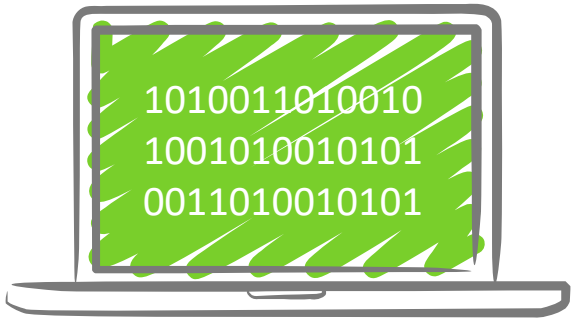


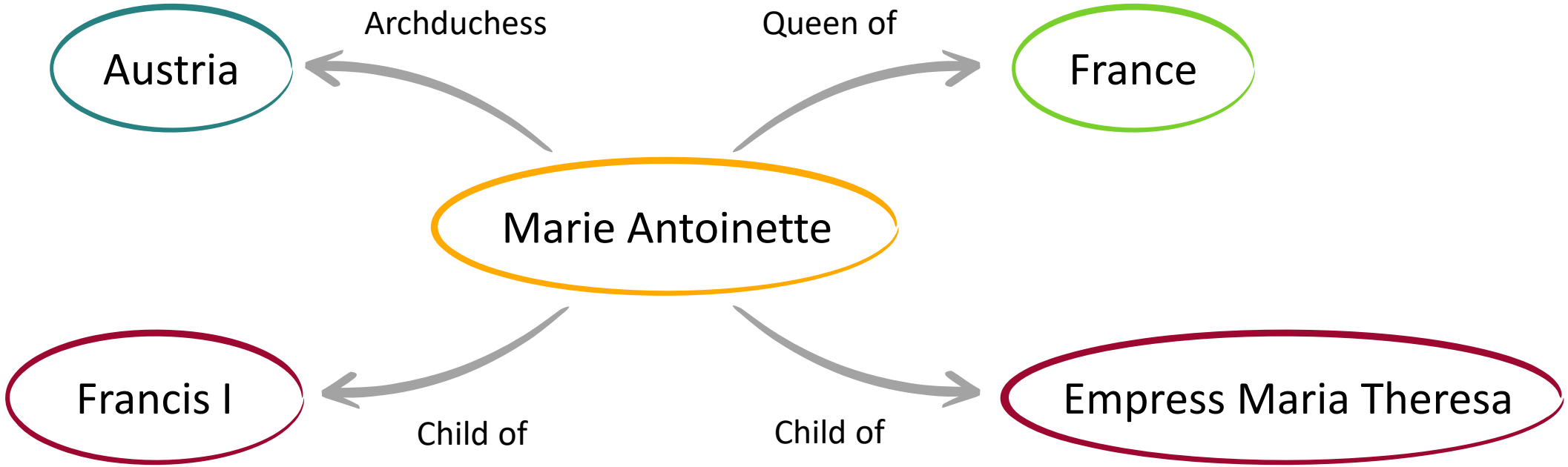




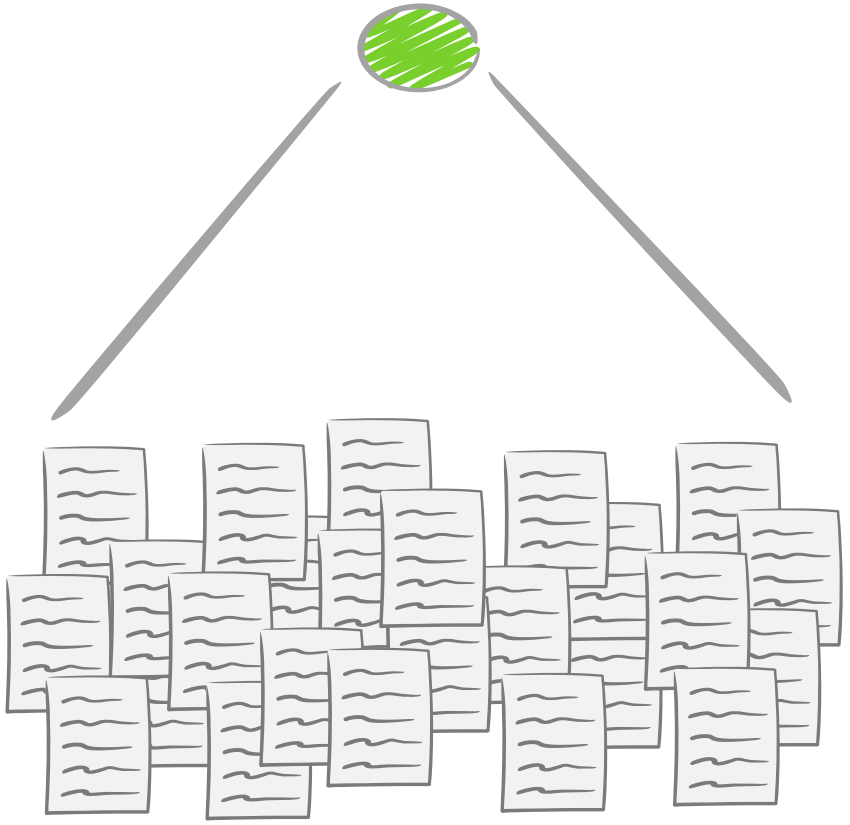



Results



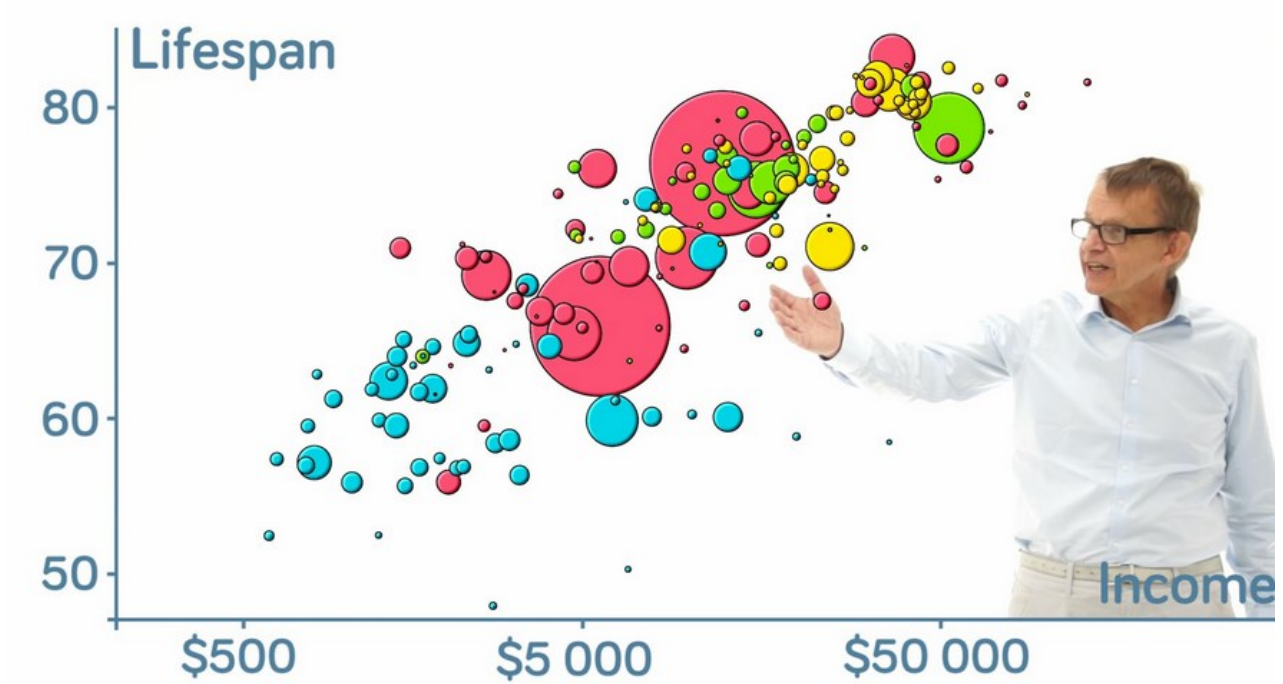

Austria
Archduchess
Queen of
France
Marie Antoinette
Francis I
Child of
Child of
Empress Maria Theresa

Evaluation



You can't
understand the
world
without numbers...

... and you cannot
understand it
only with numbers.



Factfullness

Prof. Hans Rosling

Språk-
BANKEN

CENTRUM FÖR DIGITAL HUMANIORA

GÖTEBORGS
UNIVERSITET

Thank you for listening!

Nina.tahmasebi@gu.se

nina@tahmasebi.se