Lecture 9: Natural Language Processing

Cognition and Communication, Monday, Nov. 8th 2021

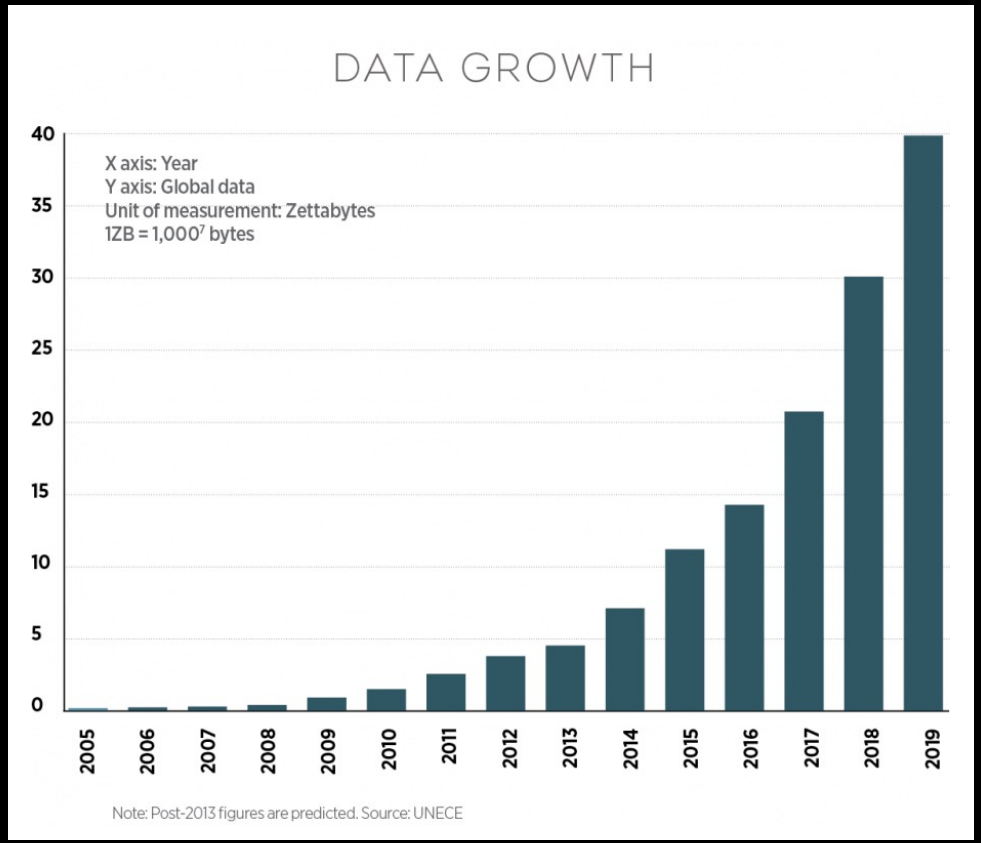
Kristian Tylén

agenda

- Natural Language Processing
  - Investigating human cognition in words/text
- Natural language processing:
  - The bag of word principle
- Dictionary approaches:
  - Sentiment analysis
- Preprocessing of text data
  - Tokenization, stop words, stemming, lemmatization
- Word embeddings
  - How to quantify semantics
- LDA: probabilistic topic models

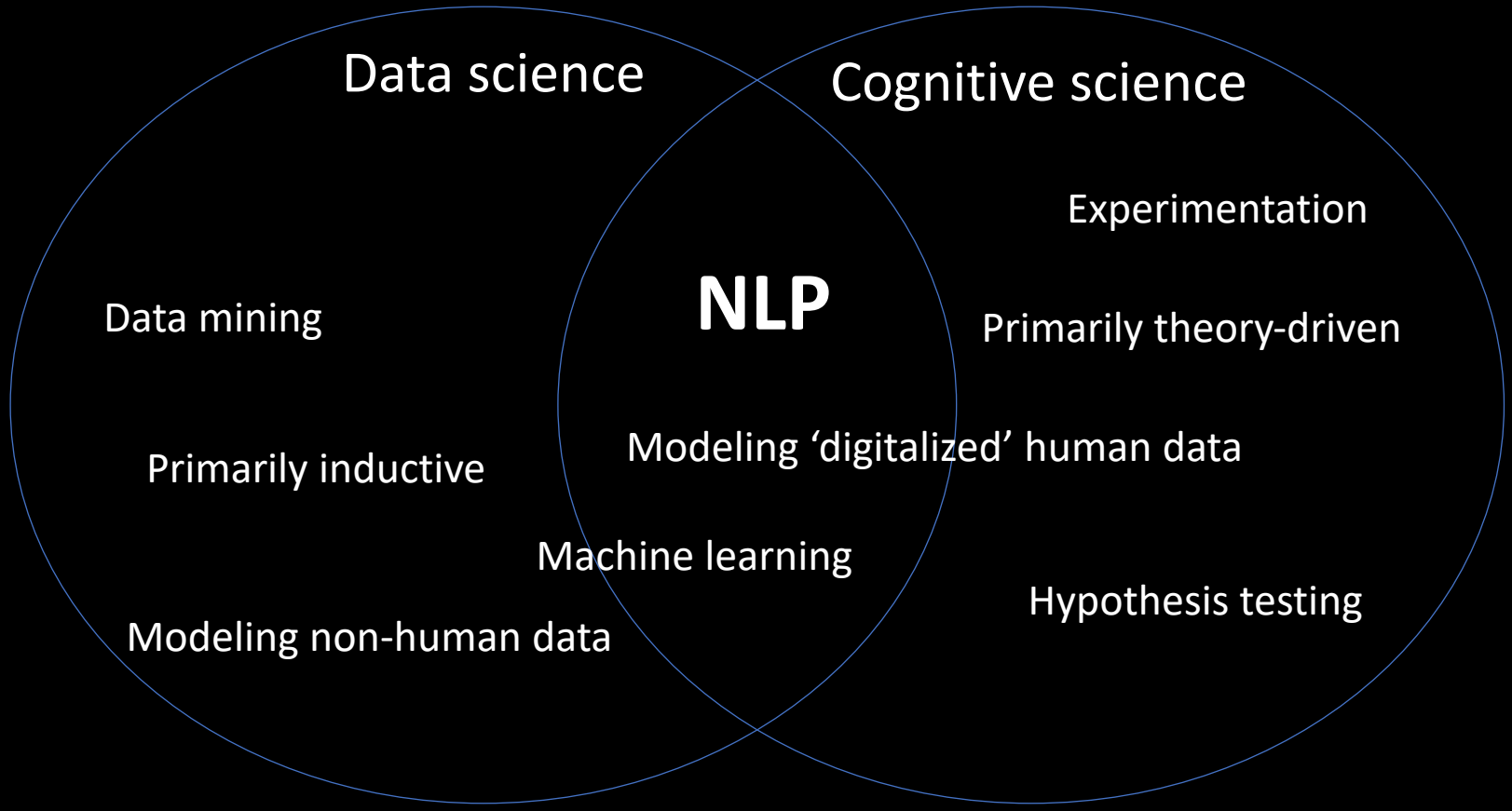
The ”era of big data”

- The digital revolution has led to accumulation of big data
- New ways of studying human mind and behavior
- Advantages:
  - Huge material: great statistical power
  - Very accessible
  - Largely un-elicited, ‘naturalistic’/ecological data
- Disadvantages:
  - Less control, more noise(?)
  - Difficult to control variables to test causal links
  - Ethical challenges (Cambridge Analytica, Google, Apple, FaceBook, TikTok)



Cognitive science versus data science

- Similarities and differences?



Natural Language Processing, *NLP*

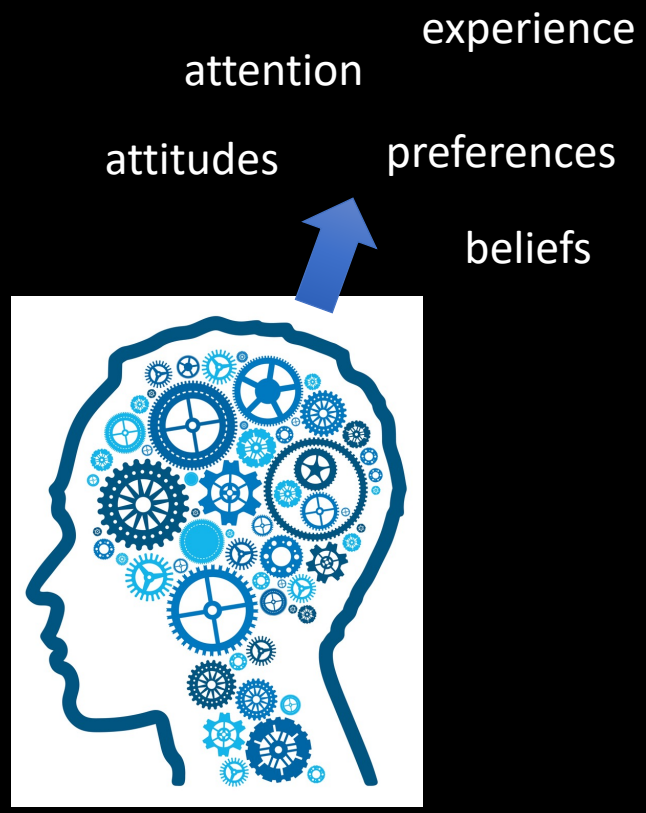
- How to program computers to process and analyze large amounts of natural language data
  - Machine learning perspective: how can we detect spam? How can we make automatic subtiteling of youtube videos
  - Cognitive science perspective: how can we derive variables that are informative of human cognitive processes

- Big text repositories:
  - Legal texts
  - News
  - SoMe
  - Literature

- Not just written text:
  - Speech synthesis
  - Speech recognition
  - Chat bot systems
  - Automatic analysis of voice features in clinical context

The psychological meaning of words

- ”The words we use in daily life reflect what we are paying attention to, what we are thinking about, what we are trying to avoid, how we are feeling, and how we are organizing and analyzing our worlds” (Tausczik & Pennebaker 2010:30)



Different tools for different “levels of processing”

- Lexical level: “Let the rock roll!”
- Syntactic level: Verb-imp, NP, Verb-inf
- Semantic level: geology? … or music?

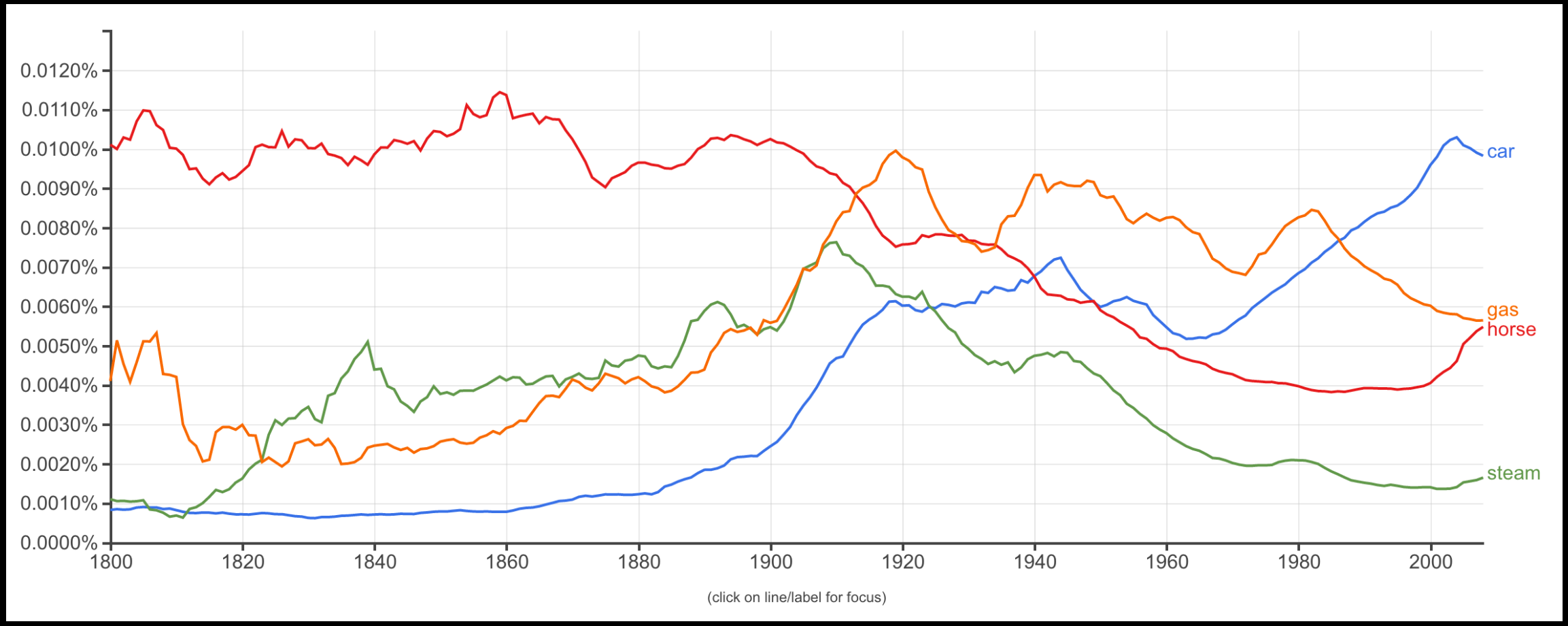
Frequency of words
Distribution in texts
Word length/LIX

Predicting the next word
Word co-occurances
Sentence complexity

Semantic relations
Sentiment
Topics
Other semantic dimensions

Word count approaches

- Even bare word counts can potentially inform questions of human cognition
- https://books.google.com/ngrams



Word Use in the Poetry of Suicidal and Nonsuicidal Poets

SHANNON WILTSEY STIRMAN, MA, AND JAMES W. PENNEBAKER, PHD

Objective: The purpose of this study was to determine whether distinctive features of language could be discerned in the poems of poets who committed suicide and to test two suicide models by use of a text-analysis program. **Method:** Approximately 300 poems from the early, middle, and late periods of nine suicidal poets and nine nonsuicidal poets were compared by use of the computer text analysis program, Linguistic Inquiry and Word Count (LIWC). Language use within the poems was analyzed within the context of two suicide models. **Results:** In line with a model of social integration, writings of suicidal poets contained more words pertaining to the individual self and fewer words pertaining to the collective than did those of nonsuicidal poets. In addition, the direction of effects for words pertaining to communication was consistent with the social integration model of suicide. **Conclusions:** The study found support for a model that suggests that suicidal individuals are detached from others and are preoccupied with self. Furthermore, the findings suggest that linguistic predictors of suicide can be discerned through text analysis. **Key words:** suicide, text analysis, poetry, social integration, LIWC.

TABLE 1. Means for LIWC Categories

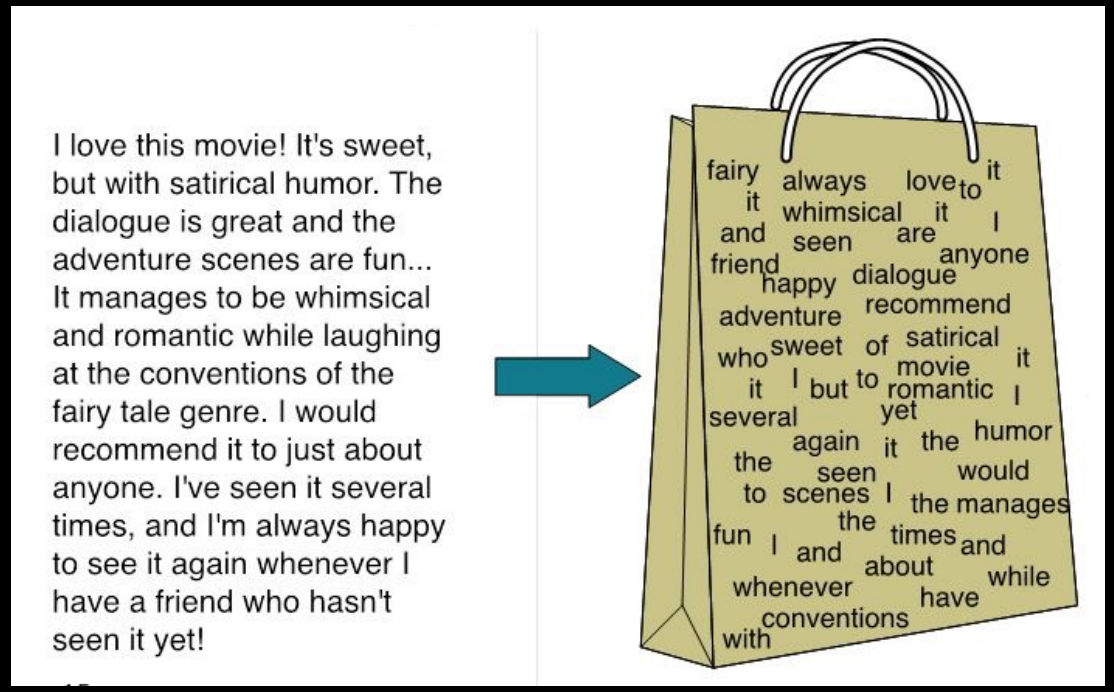
	Suicide Group			Control Group			Effects
	Early	Middle	Later	Early	Middle	Late	
Disengagement theory							
I (me, my)	4.0	3.4	4.0	2.5	1.6	2.5	S
We (us, our)	.73	1.3	.85	.69	.40	1.1	S,P**
Communication (talk, share)	1.2	1.1	1.0	.89	1.1	1.3	—
Hopelessness theory							
Negative emotion (hate, worthless)	2.2	1.8	1.7	2.3	2.1	1.7	—
Positive emotion (happy, love)	3.3	3.1	3.9	2.9	2.9	2.5	—
Death (dead, grave)	.52	.47	.69	.34	.43	.41	S**
Other findings							
Sexual words (lust, breast)	.60	.84	.47	.36	.36	.31	S

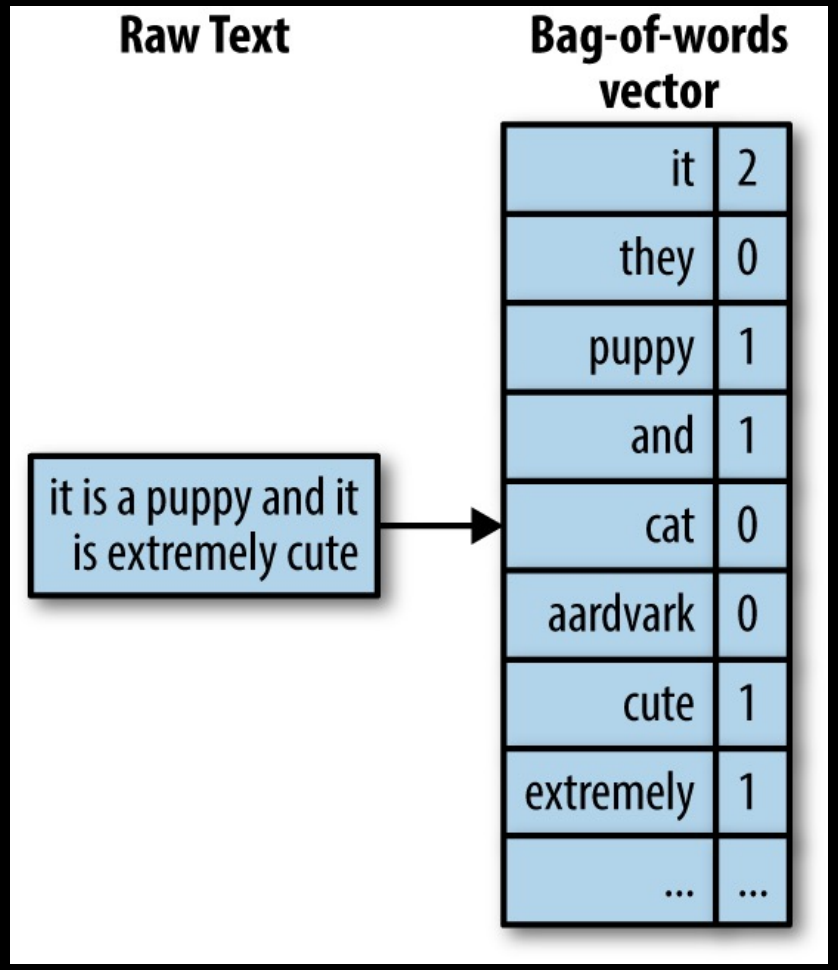
Note: Means reflect percentage of total words used in each poem within the relevant category. Effects refer to: S = suicide vs. nonsuicide main effect, P = phase of career main effect. All effects are significant $p \leq .05$, except ** $p \leq .08$.

Stirman, S. W., & Pennebaker, J. W. (2001). *Psychosomatic medicine*, *63*(4), 517-522.

Dictionary approaches

- Relies on the bag-of-word principle:
  - treats words as independent units
  - disregards the context/syntax in which they occur





Dictionary approaches

- Looks up words one-by-one in an digital ”dictionary” to retreive word properties
  - E.g. the word’s
    - Frequency
    - Abstractness
    - Familiarity
    - Imaginability
    - Sentiment
  - Depends on available dictionaries (which are often established by manual annotation of words)
  - … which are often somewhat limited in their repertoires

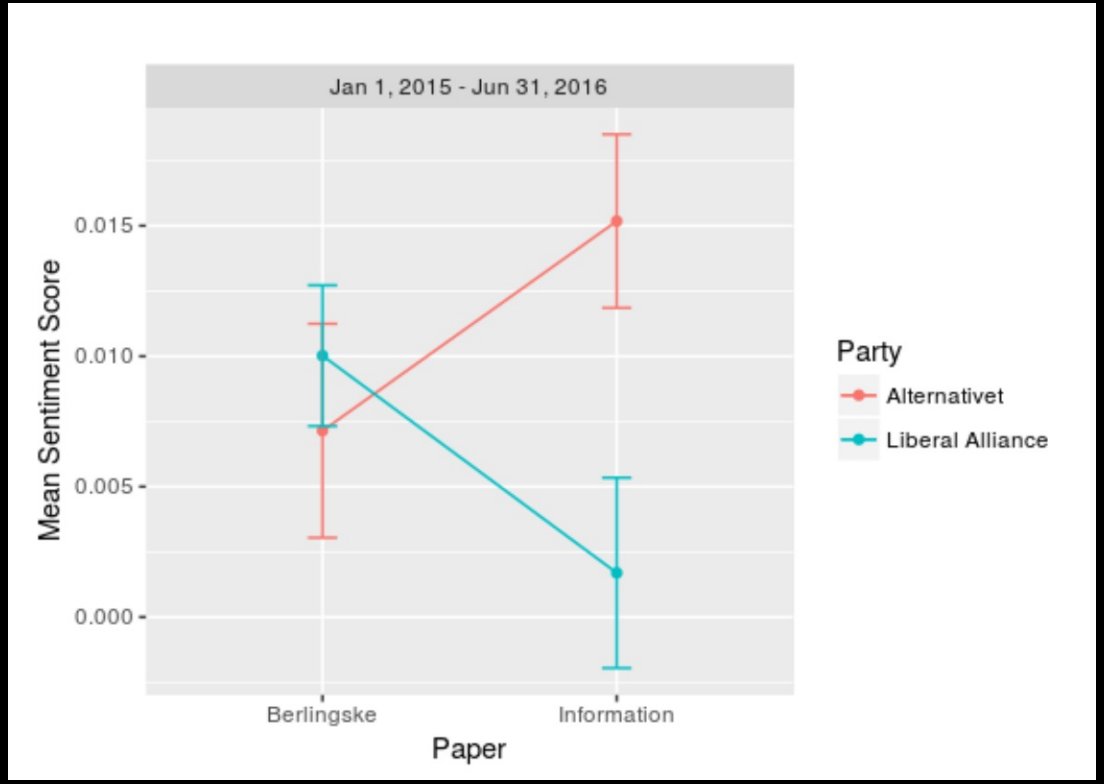
Sentiment analysis (or opinion mining)

- = the use of natural language processing to systematically identify, extract, quantify, and study affective states in texts

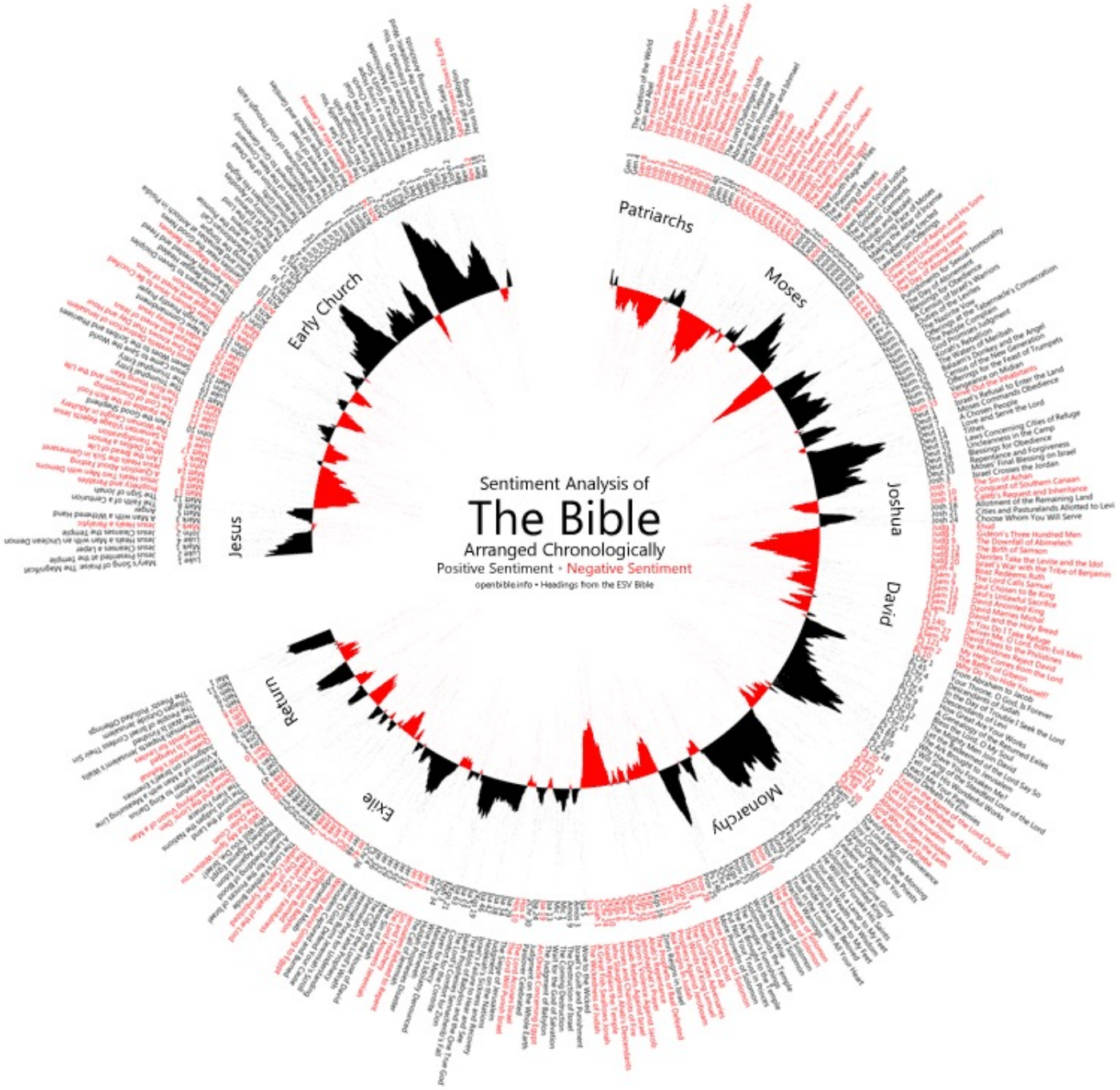
- Aims to ‘measure’ or identify the emotion or attitude of a speaker/writer with respect to some topic or event

- Can take the emotional temperature on discourse

- Looks up words in a sentiment dictionary and return the sentiment of the word as either…
  - Binary ‘positive’ or ‘negative’
  - Category of emotion: “angry”, “sad”, “pleasure”, “dominance” etc.
  - Continuous valence scale, e.g. -5 to +5
- … dependent on the dictionary



Enevoldsen, K. C., & Hansen, L. (2017)


Sentiment Analysis of
The Bible
Arranged Chronologically
Positive Sentiment • Negative Sentiment
openbible.info • Headings from the ESV Bible
Patriarchs
Moses
Joshua
David
Monarchy
Exile
Return
Early Church
Jesus

e-bible/

Preprocessing of text data

- Tokenization:
  - Splitting up texts in smaller chucks (most often individual words)
  - Cleaning up punctutations

- Regular expressions:
  - search for patterns (rather than individual word tokens) in texts
  - Uses ‘meta-characters’ such as ? * +
  - For example, we could use the pattern ”(red|blue|green|yellow)” to count color words
  - ”colou?r” matches both "color" and "colour”
  - ”ab*c” matches "ac", "abc", "abbc", "abbbc", and so on
  - ”\d” any digit
  - Good tools for R and Python

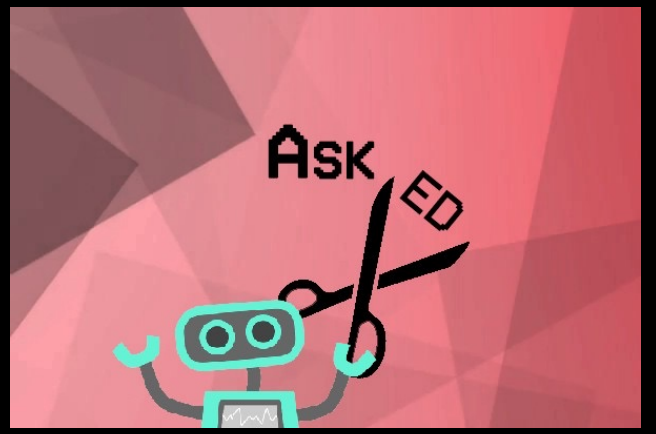
Preprocessing of text data

- Stop words
  - Sometimes we are not interested in including very frequent function words like “a”, “the”, “that”, “in”, “on”
  - Can be filtered using a ”stop word list”
  - Some standard lists comes with NLP packages:
    - Tidytext: list of 1149 words
    - TM: has 174
    - Quanteda: 175
  - Often relevant to modify or make custom stop word list

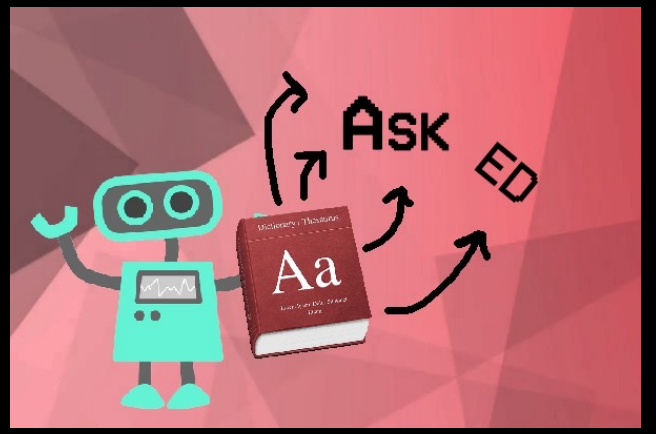
```
> stopwords(kind = "en")
  [1] "i"          "me"         "my"         "myself"     "we"
  [6] "our"        "ours"       "ourselves"  "you"        "your"
 [11] "yours"      "yourself"   "yourselves" "he"         "him"
 [16] "his"        "himself"    "she"        "her"        "hers"
 [21] "herself"    "it"         "its"        "itself"     "they"
 [26] "them"       "their"      "theirs"     "themselves" "what"
 [31] "which"      "who"        "whom"       "this"       "that"
 [36] "these"      "those"      "am"         "is"         "are"
 [41] "was"        "were"       "be"         "been"       "being"
 [46] "have"       "has"        "had"        "having"     "do"
 [51] "does"       "did"        "doing"      "would"      "should"
 [56] "could"      "ought"      "i'm"        "you're"     "he's"
 [61] "she's"      "it's"       "we're"      "they're"    "i've"
 [66] "you've"     "we've"      "they've"    "i'd"        "you'd"
 [71] "he'd"       "she'd"      "we'd"       "they'd"     "i'll"
 [76] "you'll"     "he'll"      "she'll"     "we'll"      "they'll"
 [81] "isn't"      "aren't"     "wasn't"     "weren't"    "hasn't"
 [86] "haven't"    "hadn't"     "doesn't"    "don't"      "didn't"
 [91] "won't"      "wouldn't"   "shan't"     "shouldn't"  "can't"
 [96] "cannot"     "couldn't"   "mustn't"    "let's"      "that's"
[101] "who's"      "what's"     "here's"     "there's"    "when's"
[106] "where's"    "why's"      "how's"      "a"          "an"
[111] "the"        "and"        "but"        "if"         "or"
[116] "because"    "as"         "until"      "while"      "of"
[121] "at"         "by"         "for"        "with"       "about"
[126] "against"    "between"    "into"       "through"    "during"
[131] "before"     "after"      "above"      "below"      "to"
[136] "from"       "up"         "down"       "in"         "out"
[141] "on"         "off"        "over"       "under"      "again"
[146] "further"    "then"       "once"       "here"       "there"
[151] "when"       "where"      "why"        "how"        "all"
[156] "any"        "both"       "each"       "few"        "more"
[161] "most"       "other"      "some"       "such"       "no"
[166] "nor"        "not"        "only"       "own"        "same"
[171] "so"         "than"       "too"        "very"
>
```

Preprocessing of text data

- Stemming
  - Cuts words to their stem (throw away suffixes and affixes)
  - “overcook”, “cooking,” and “cooked” -> “cook”
  - Be aware of potential overstemming: “university”, “universal”, “universities”, and “universe” could become “univers”



- Lemmatization:
  - replace word by its dictionary or canonical form
  - More computational heavy (e.g. involves part-of-speech parsing)
  - “is”, “was”, “were” -> “be”
  - “runs”, “running”, “ran” -> “run”

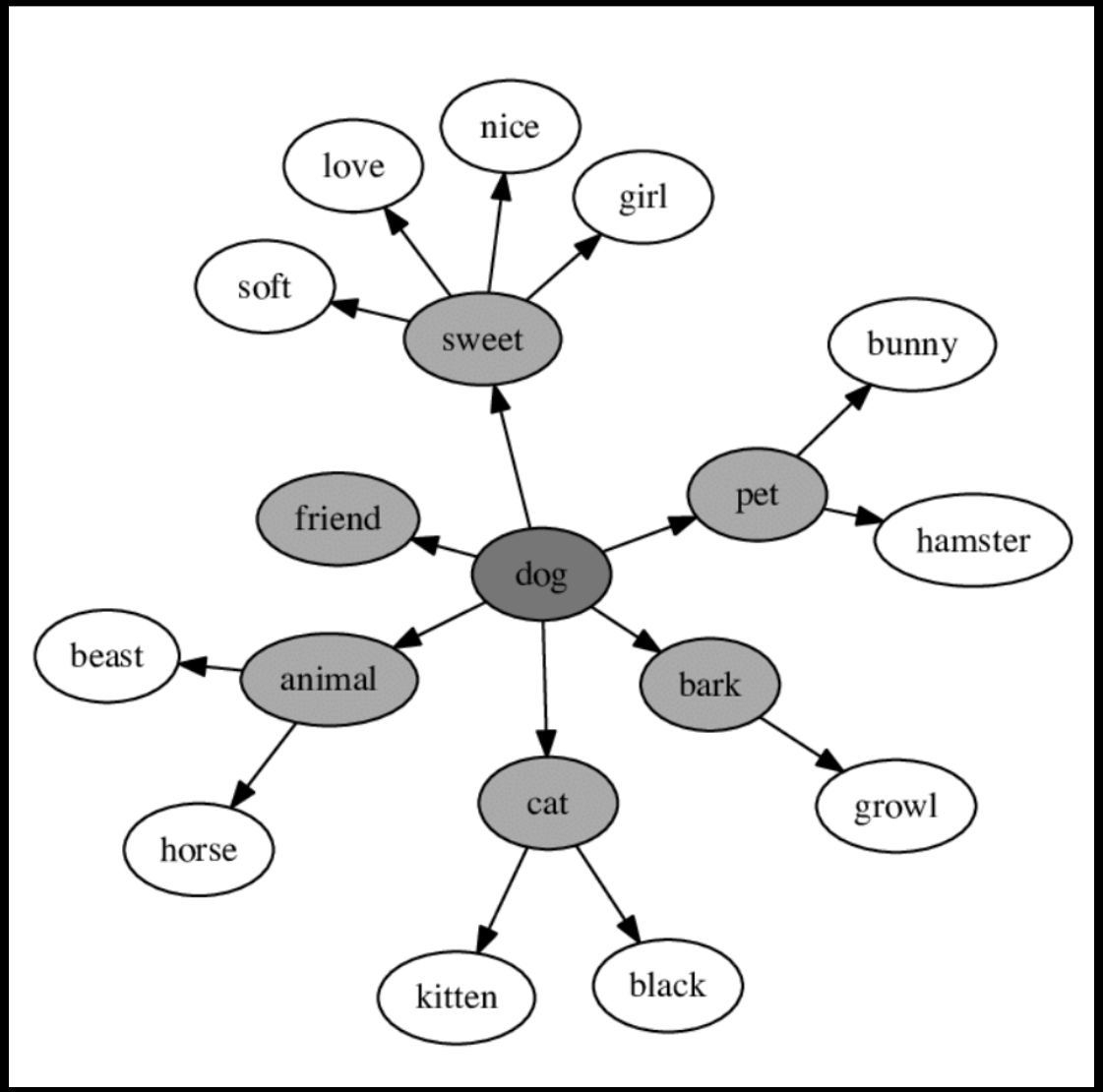


Word embeddings

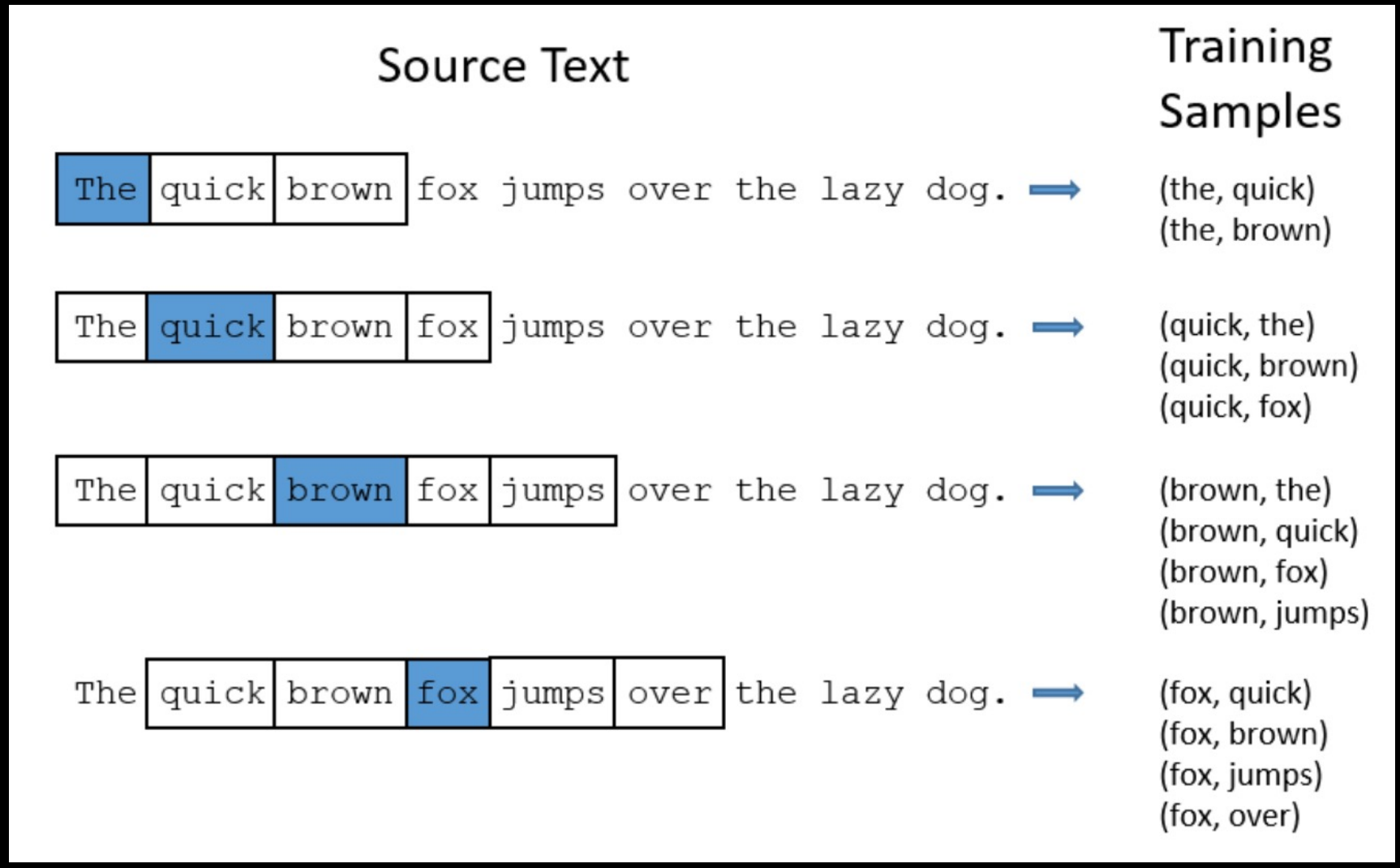
- A numeric representation of a word (a vector of values)
- Expression of the relation of the word to other words (the context)
- A way to quantify “semantics”

- From machine learning perspective:
  - Representations beyond the literal word:
    when we search Google for ”cognition”, we also get responses for ”psychology” or ”AI”
  - Better translations
  - Better word completions

- CogSci perspective:
  - Proximation of “meaning” representation in the human mind (Spreading activation/associative neural network theory)

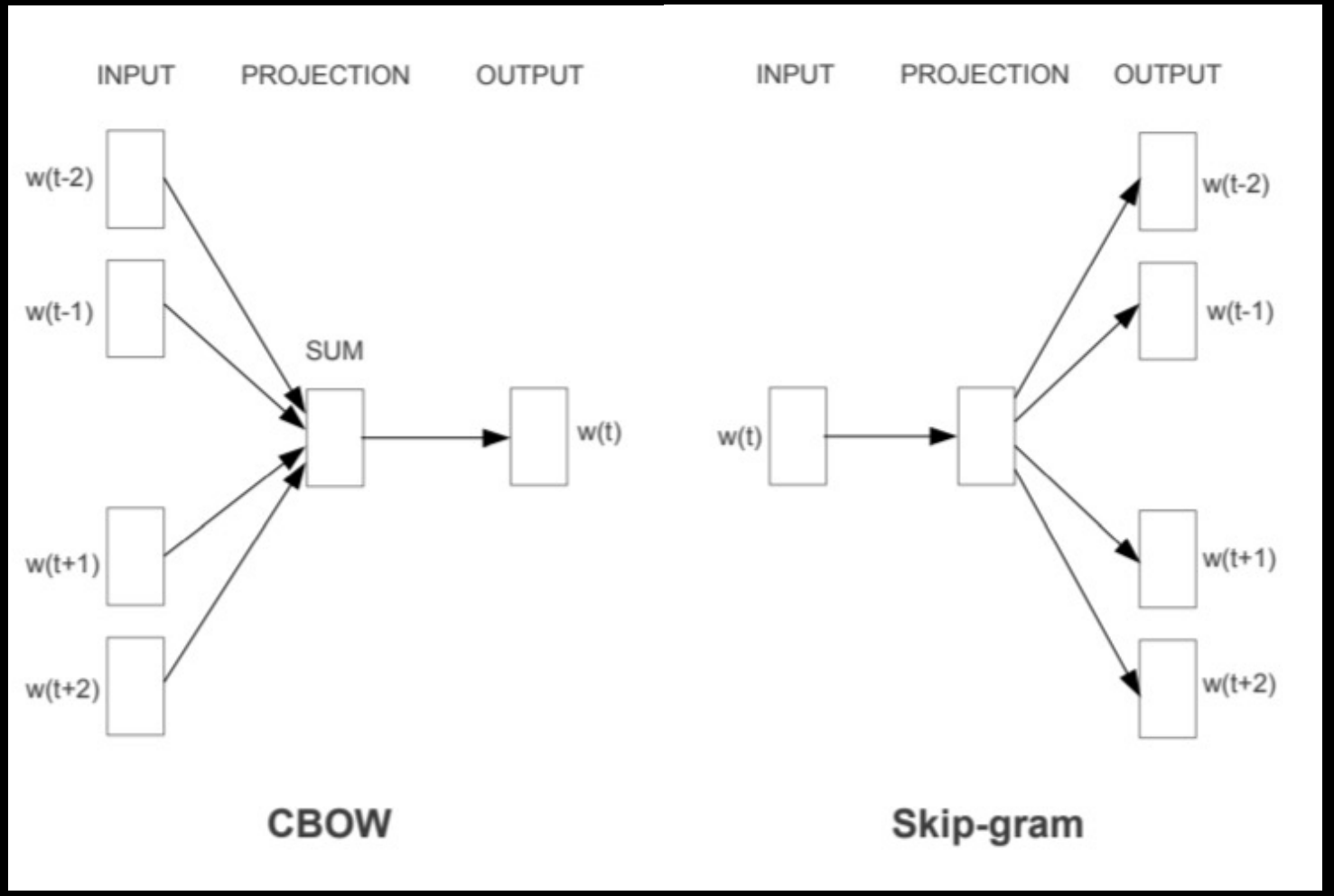


Word embeddings

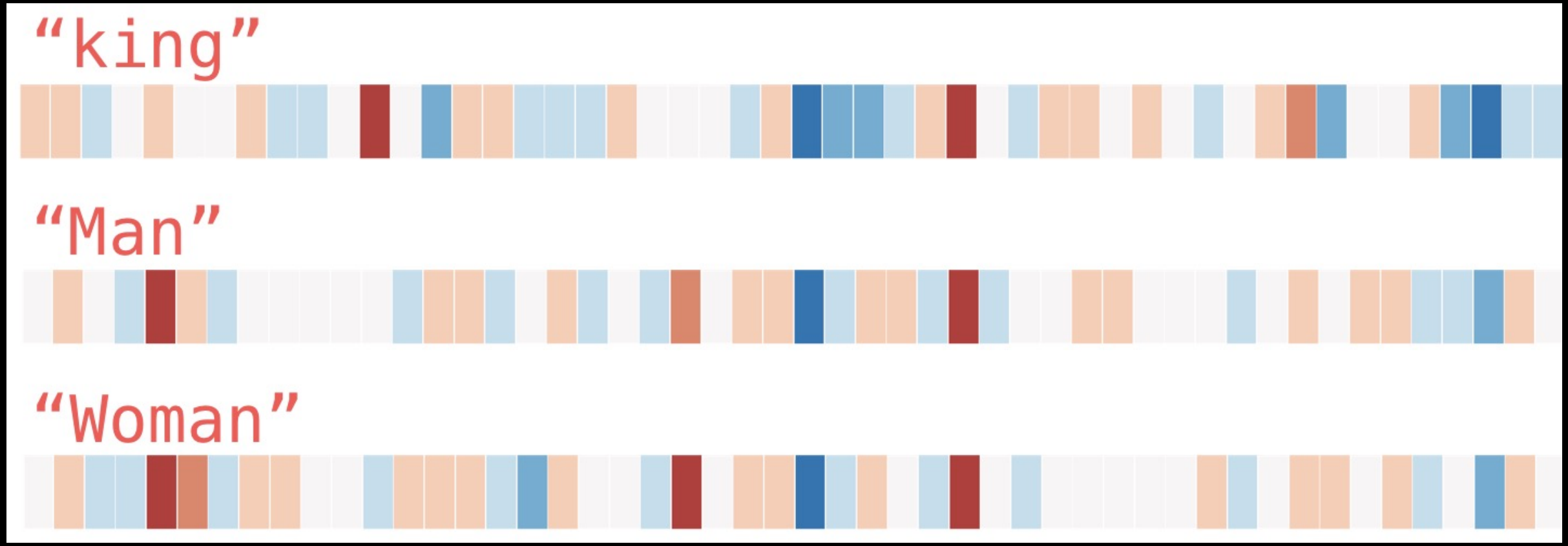


Word embeddings

- *Continous bag of words*, CBOW: a word is predicted by its context
  - faster to train than the skip-gram,
  - better accuracy for frequent words
- *Skipgram*: the context is predicted by a word
  - works well with a small amount of training data
  - represents well even rare words or phrases

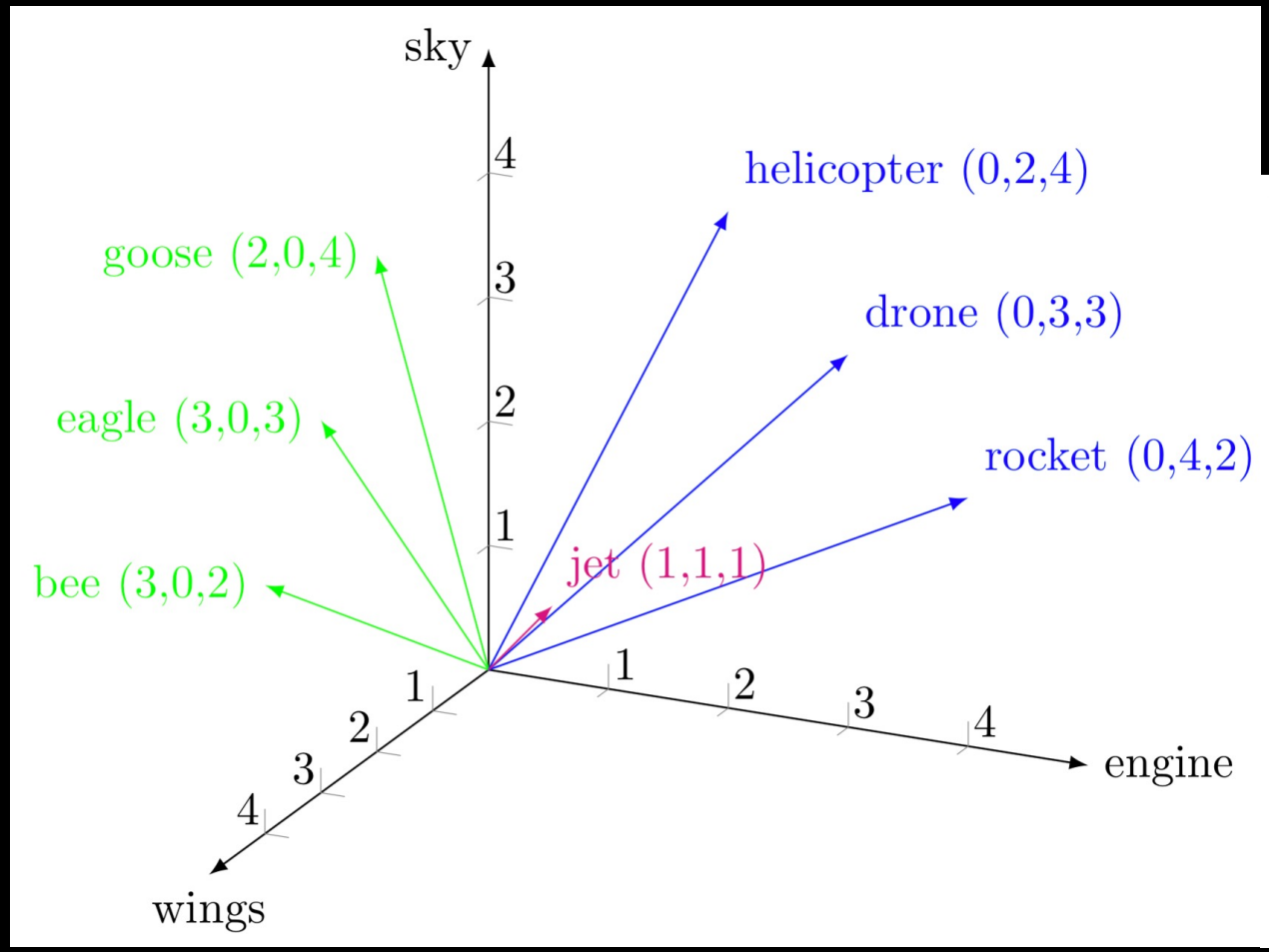


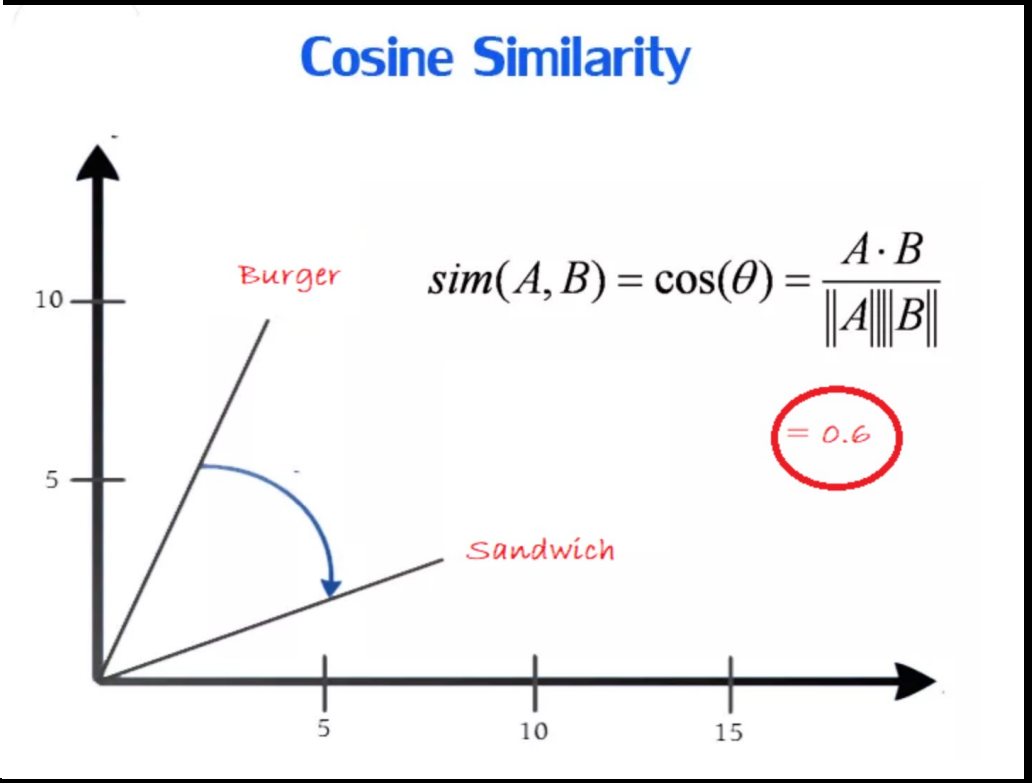
Word embeddings



Word embeddings

- calculate the “distance” between the meaning of two words in multidimensional space





Word embeddings

- Quiring the model

```
model.most_similar(positive=["king","woman"], negative=["man"])
```

```
[('queen', 0.8523603677749634),
 ('throne', 0.7664333581924438),
 ('prince', 0.7592144012451172),
 ('daughter', 0.7473883032798767),
 ('elizabeth', 0.7460219860076904),
 ('princess', 0.7424570322036743),
 ('kingdom', 0.7337411642074585),
 ('monarch', 0.721449077129364),
 ('eldest', 0.7184862494468689),
 ('widow', 0.7099430561065674)]
```

nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-020-0924-8

Cultural influences on word meanings revealed through large-scale semantic alignment

Bill Thompson[1]✉, Seán G. Roberts[2,3] and Gary Lupyan[4]

- Similarities and differences between 1010 meanings in 41 languages

- "If the structure of language vocabularies mirrors the structure of natural divisions that are universally perceived, then the meanings of words in different languages should closely align (…)
- …By contrast, if shared word meanings are a product of shared culture, history and geography, they may differ between languages in substantial but predictable ways."

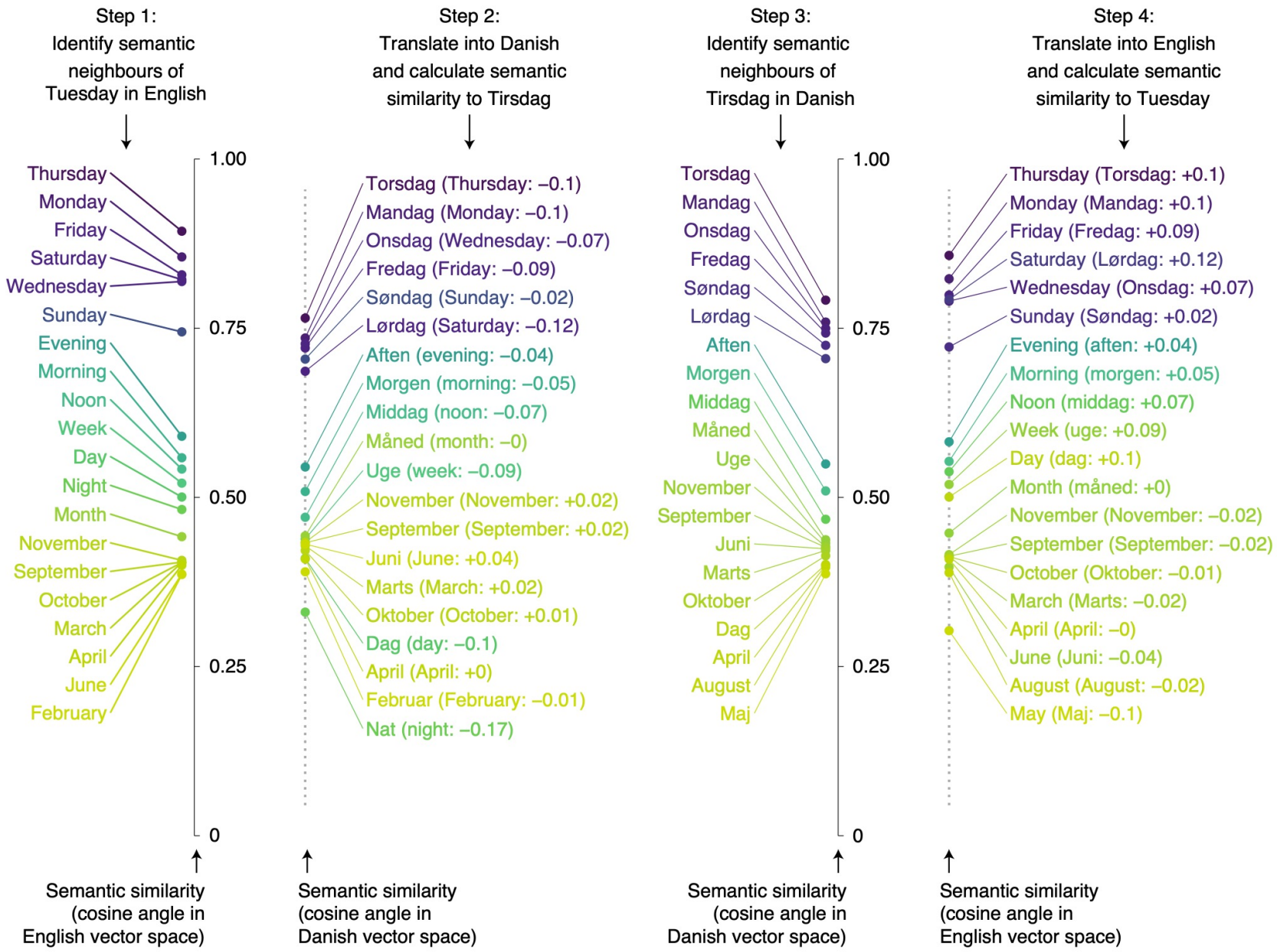
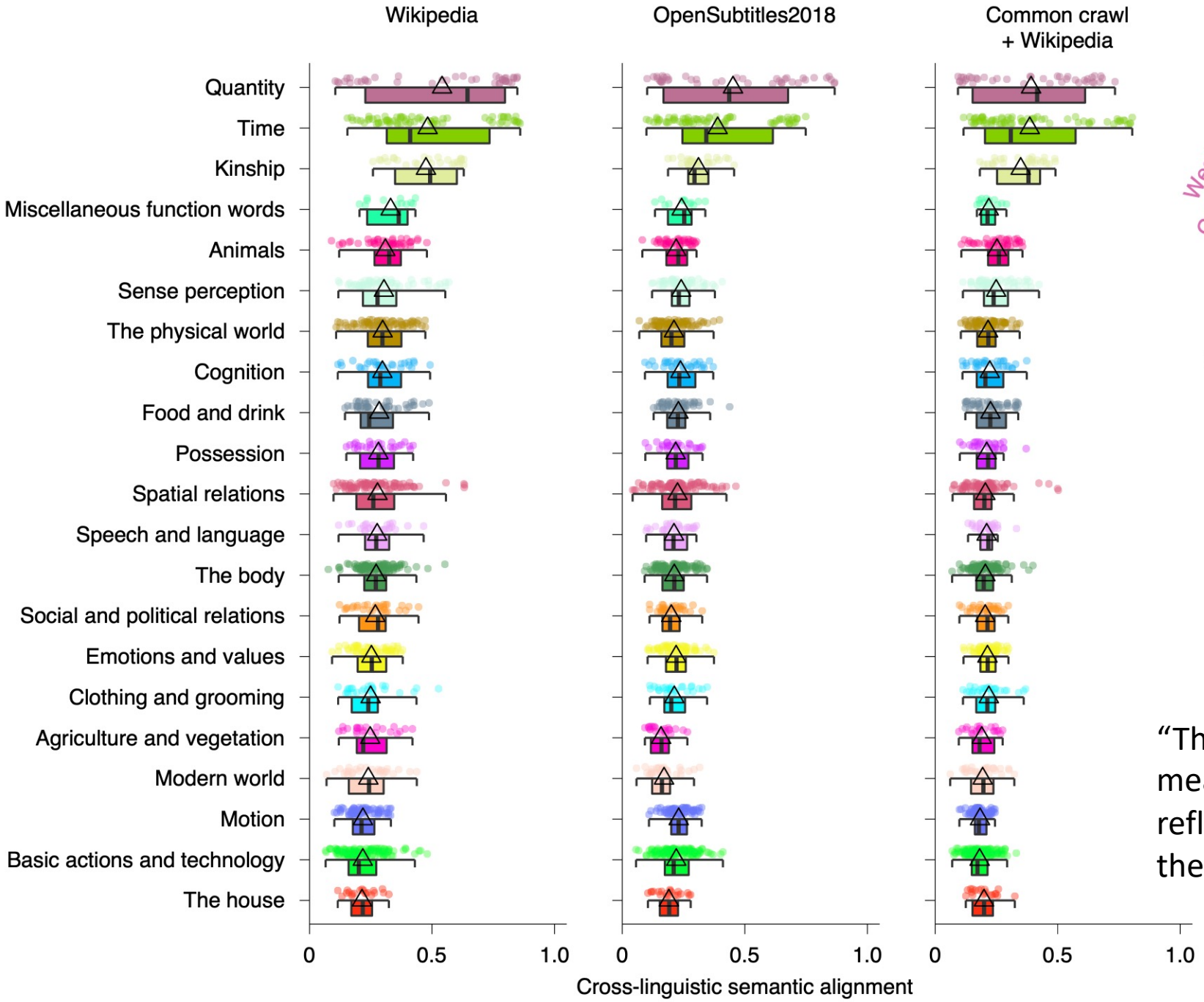
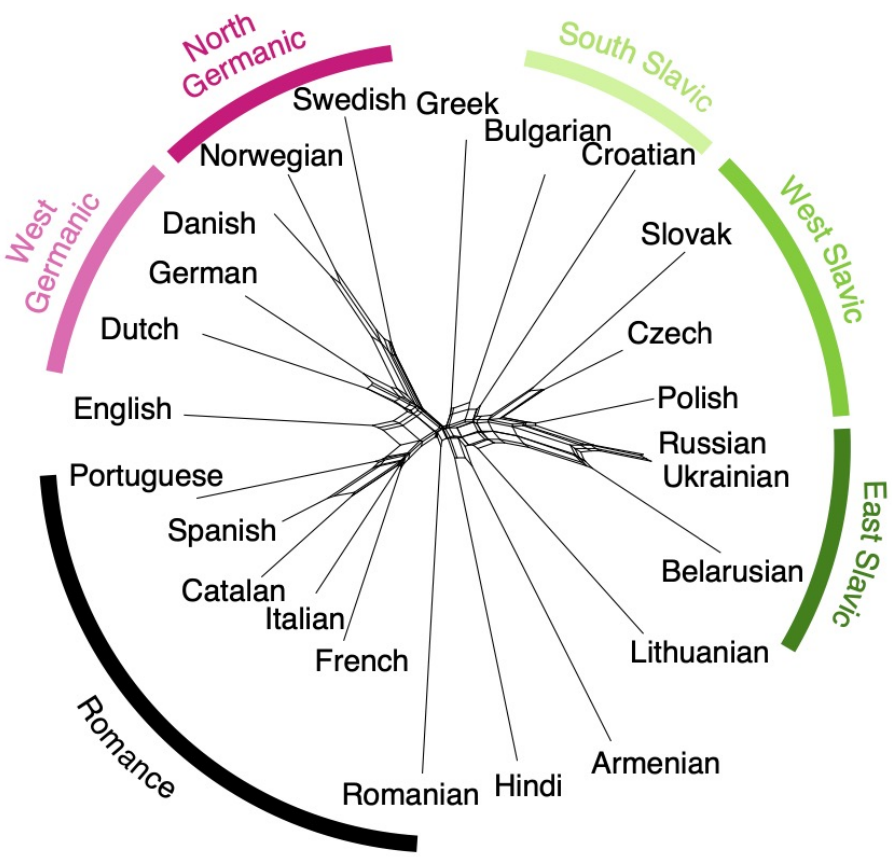


Fig. 1 | High alignment between English ('Tuesday') and Danish ('Tirsdag'). A schematic of the algorithm for computing semantic alignment. The colour denotes semantic similarity in the first language; similar colour ordering on both sides of the plot indicates a high level of alignment.





"These results provide evidence that the meanings of common words vary in ways that reflect the culture, history and geography of their users" (2020:1).

nature human behaviour

ARTICLES

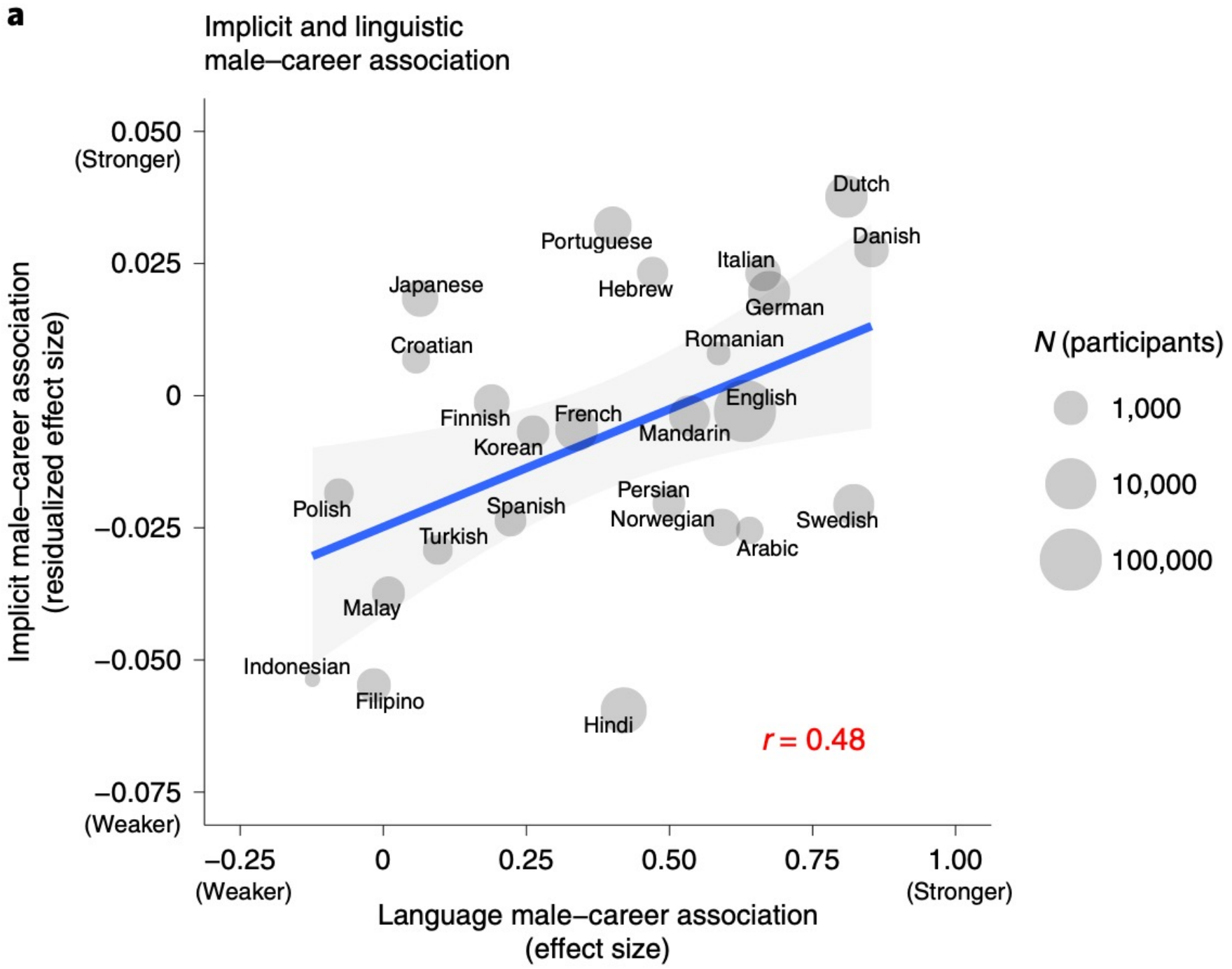
https://doi.org/10.1038/s41562-020-0918-6

Gender stereotypes are reflected in the distributional structure of 25 languages

Molly Lewis[1,2] ✉ and Gary Lupyan[3]

Cultural stereotypes such as the idea that men are more suited for paid work and women are more suited for taking care of the home and family, may contribute to gender imbalances in science, technology, engineering and mathematics (STEM) fields, among other undesirable gender disparities. Might these stereotypes be learned from language? Here we examine whether gender stereotypes are reflected in the large-scale distributional structure of natural language semantics. We measure gender associations embedded in the statistics of 25 languages and relate these to data on an international dataset of psychological gender associations ($N = 656{,}636$). People's implicit gender associations are strongly predicted by gender associations encoded in the statistics of the language they speak. These associations are further related to the extent that languages mark gender in occupation terms (for example, 'waiter'/'waitress'). Our pattern of findings is consistent with the possibility that linguistic associations shape people's implicit judgements.

- Tests the relation between responses on the Implicit Association Test (IAT) and word embeddings for:
  - Man – carreer
  - Woman - family


a
Implicit and linguistic
male–career association
Implicit male–career association
(residualized effect size)
0.050
(Stronger)
0.025
0
−0.025
−0.050
−0.075
(Weaker)
Dutch
Danish
Portuguese
Italian
Hebrew
Japanese
German
Romanian
Croatian
English
Finnish
French
Mandarin
Korean
Persian
Polish
Spanish
Norwegian
Swedish
Turkish
Arabic
Malay
Indonesian
Filipino
Hindi
r = 0.48
−0.25
(Weaker)
0
0.25
0.50
0.75
1.00
(Stronger)
Language male–career association
(effect size)
N (participants)
1,000
10,000
100,000

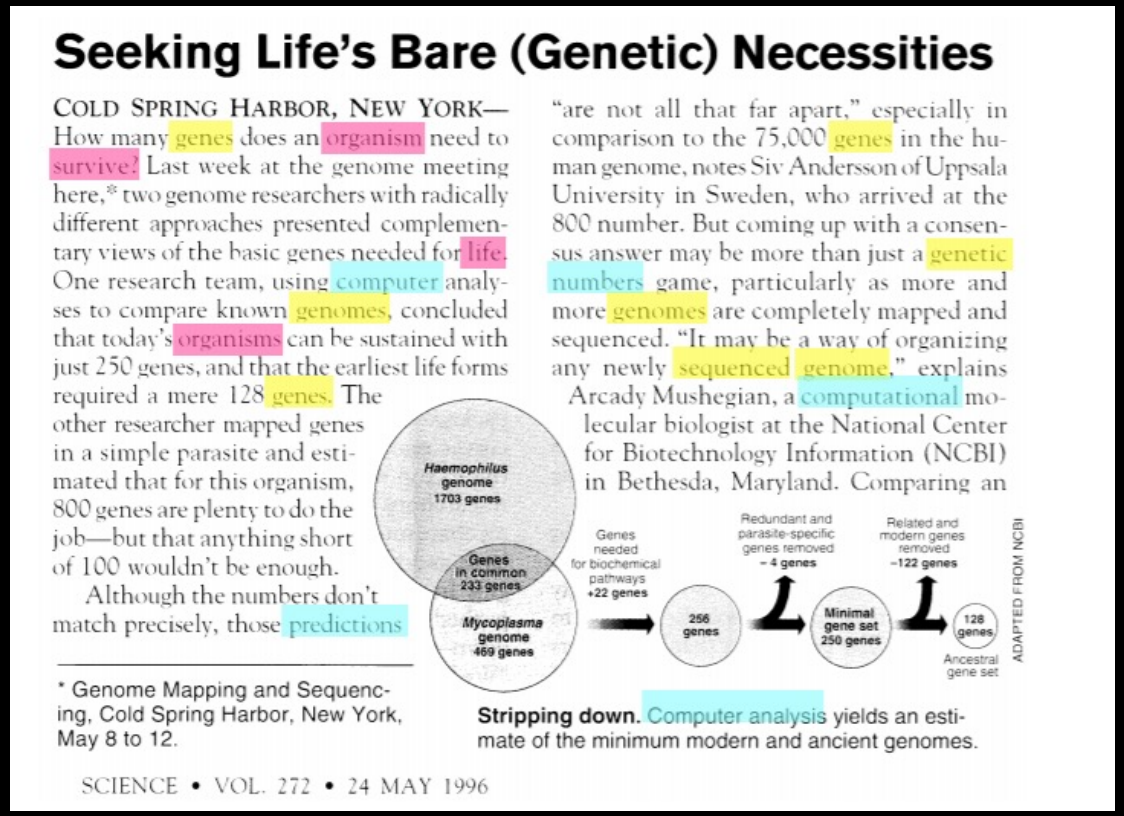
Probabalistic topic modelling

Daddy's bad knee bla bla bla
psychotherapy bla bla bla are
you getting any food? bla bla
bla Daddy's bad knee bla bla
bla television show bla bla bla

- Topic models are designed for:
  - Uncovering hidden thematic structues (topics)
  - Annotate documents according to thematic structure
  - Use annotation to organize, summarise or search
  - Inductive, explorative buttom-up approach

Latent Dirichlet Allocation (LDA)

- *Latent Dirichlet Allocation* (LDA) is a probabilisitc un-supervised method (Blei, 2007)
- Probabilistic: the output is a probability distribution
- Un-supervised: We don't tell it what to find (as with dictionaries)
- Uses mixed membership: each word is in multiple topics, and each topic in multiple documents
- It's also a generative model: It assumes that documents are generated in a certain way



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

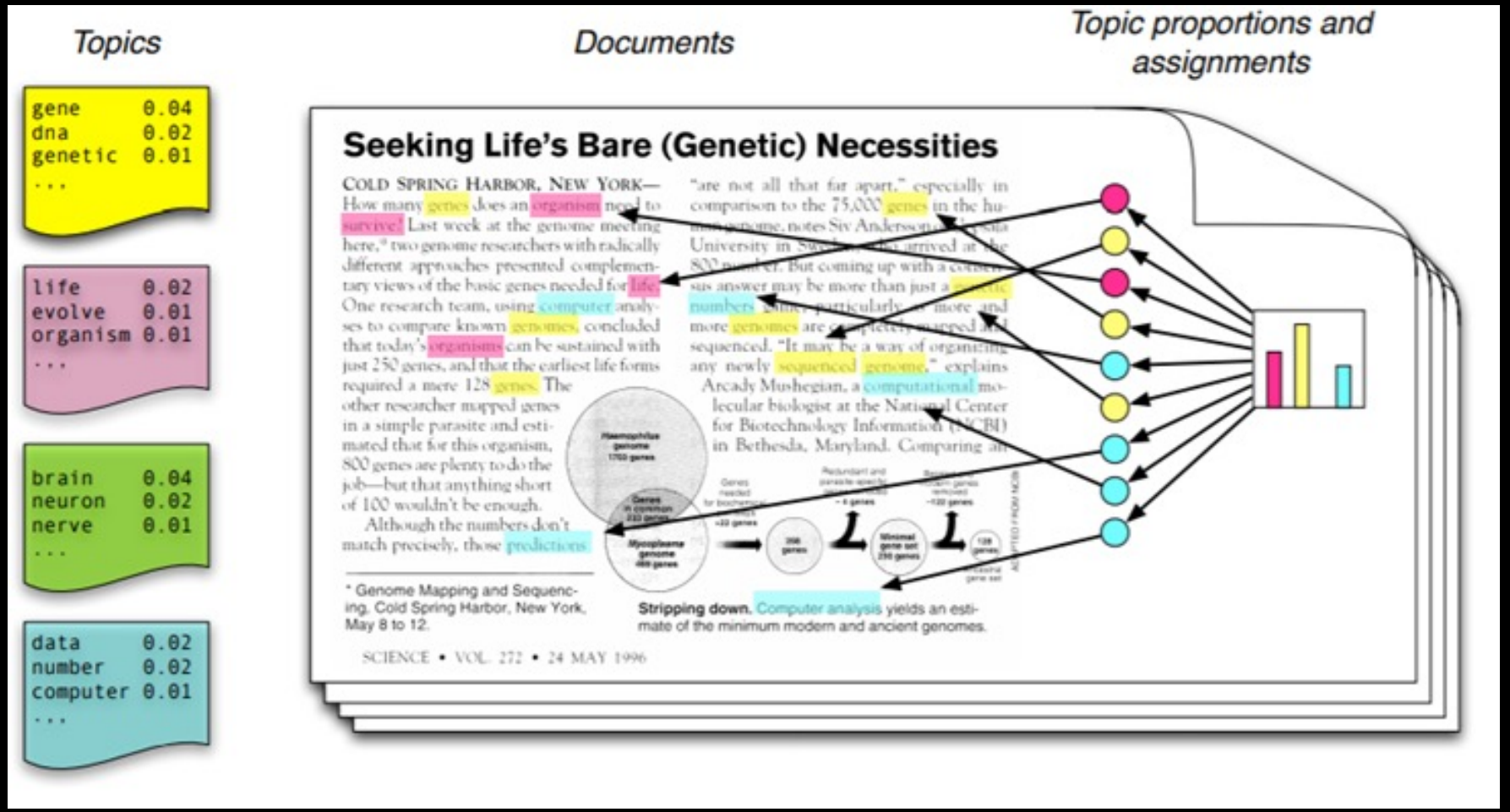
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

The generative process

- LDA assumes that a text is created by combining words from different topics
  - That is, your sweet mum has a certain number of topics from which she creates her text
- Each topic is a distribution over a fixed vocabulary
  - Some words are more probable to belong to one topic than another
- LDA tries from the surface word level to backward engineer/infer the topics
  - Reconstruct the probability distributions that are more likely given the data


Topics
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Documents
Topic proportions and assignments
Seeking Life's Bare (Genetic) Necessities
COLD SPRING HARBOR, NEW YORK—
How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.
Although the numbers don't match precisely, those predictions
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson
University in
arrived at the 800 number. But coming up with a
sis answer may be more than just a
numbers game, particularly
more and more genomes are
sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

Why does it work?

LDA builds on a trade off between two goals:

1) For each document minimize the number of topics
2) For each topic assign high probability to few words

These goals are conflicting

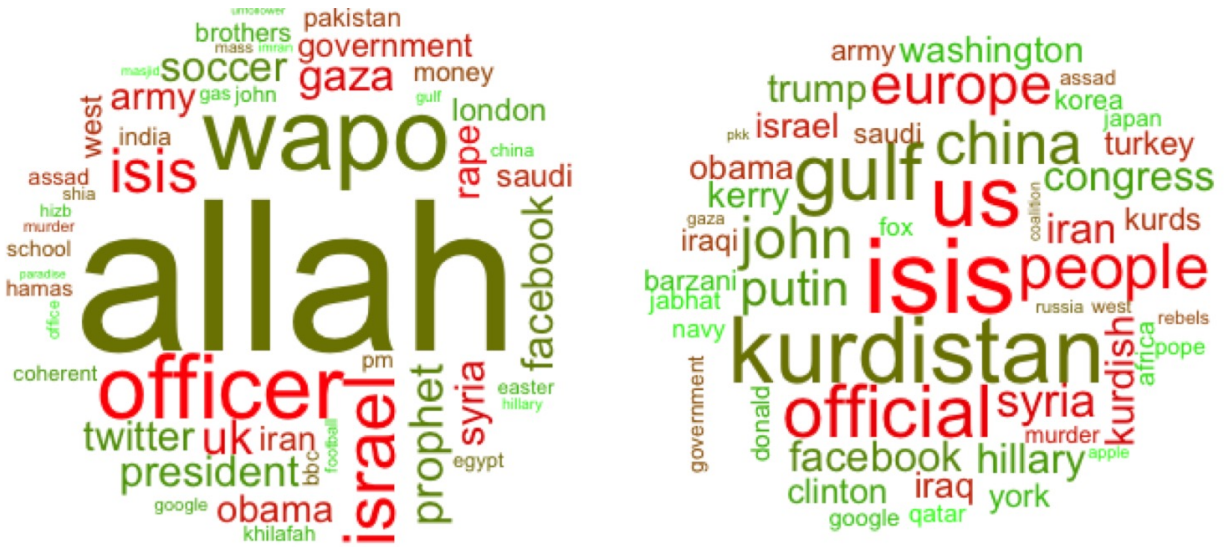
- Assigning 1 topic to a document makes 2) hard
- Assigning 1 word to a topic makes 1) hard

- Trading these goals off creates groups of tightly *co-occuring* words

A Semantic Graph-based Approach for Radicalisation Detection on Social Media

Hassan Saif,[1] Thomas Dickinson,[1] Leon Kastler,[2] Miriam Fernandez,[1] and Harith Alani[1]

[1] Knowledge Media Institute, The Open University, United Kingdom
{h.saif, thomas.dickinson, m.fernandez, h.alani}@open.ac.uk
[2] University of Koblenz Landau, Germany
lkastler@uni-koblenz.de



(a) pro-ISIS (b) anti-ISIS

Fig. 4: Word clouds of the top-50 named-entities published by pro-ISIS and anti-ISIS users, the colour indicates the sentiment attached to the entity - with red being negative, and green being positive.

Take home

- The growing field of NLP provide us with tools for studying human cognition in large data sets
- Word count methods can reveal e.g. attentional patterns
- Dictionary approaches like sentiment analysis can uncover patterns of emotion or attitude to topics
- Sometimes we need to preprocess texts through…
  - Tokenization
  - Stop word removal
  - Stemming or lemmatization
- Word embeddings (such as word2vec) allow us to quantitatively study aspects of semantics/meaning in text
- And LDA topic models can reveal the hidden topics organizing large text collections