ABCDH
https://abcdh.eu/

Distant Reading in R

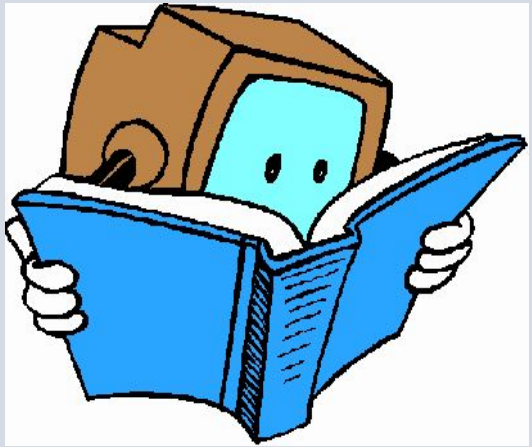
NLP for Distant Reading

Simone Rebora & Giovanni Pietro Vitali
simone.rebora@univr.it giovannipietrovitali@gmail.com

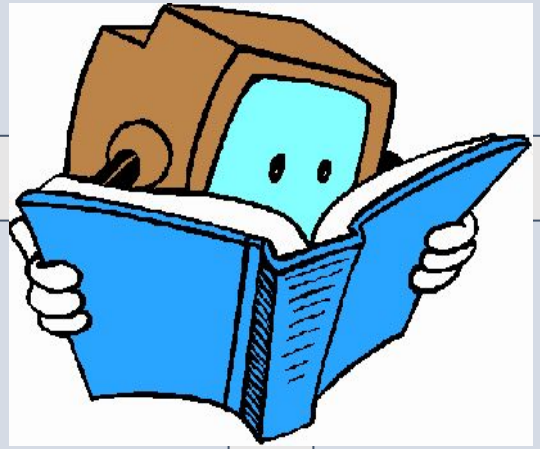
NATURAL LANGUAGE PROCESSING (NLP)

"[…] is a subfield of computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data"

(Wikipedia)



NATURAL LANGUAGE PROCESSING (NLP)



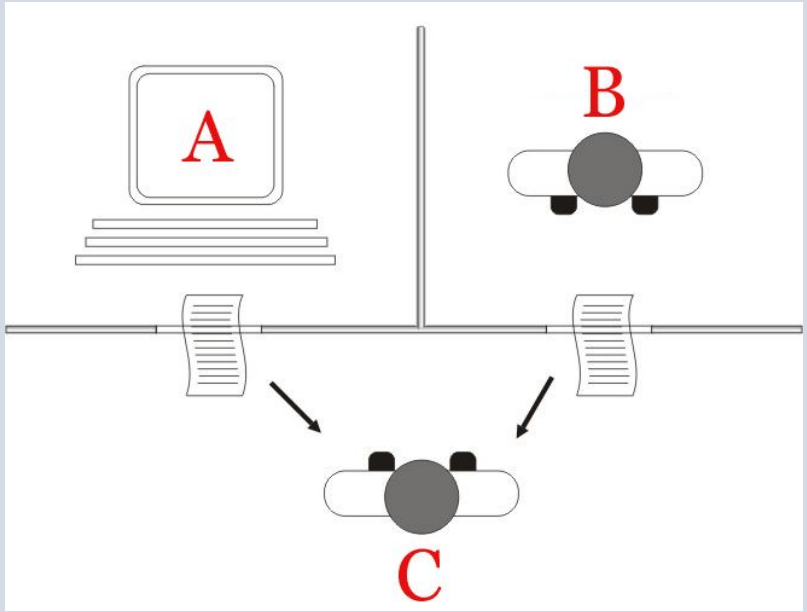
Sentiment Analysis

Topic Modeling

Stylometry

THE NLP “PROBLEM”

Can a computer (really) understand human language?



...see the Turing Test (and Artificial Intelligence)

THE NLP “PROBLEM”

Can a computer (really) understand human language?

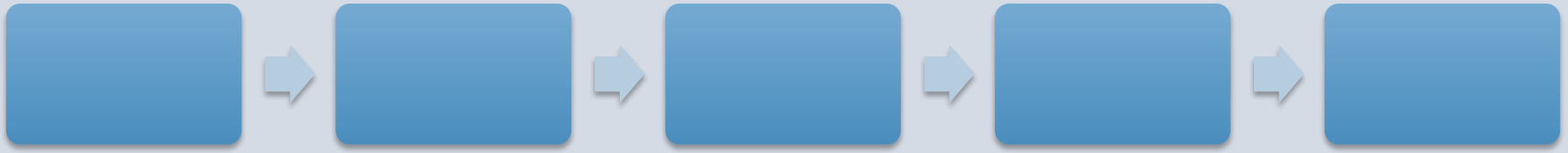
Issues:

- language ambiguity, idiomatic expressions, non-standard language…

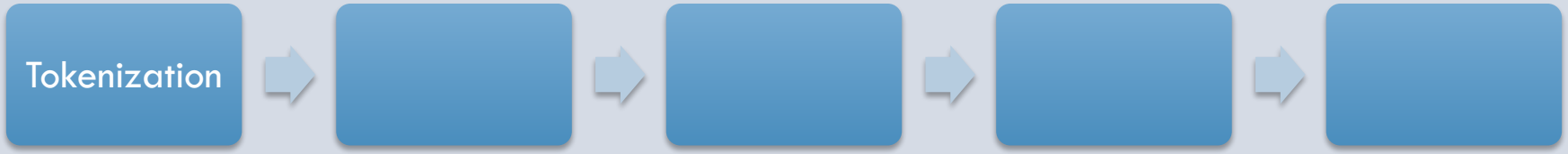
Solution:

- *divide et impera!*

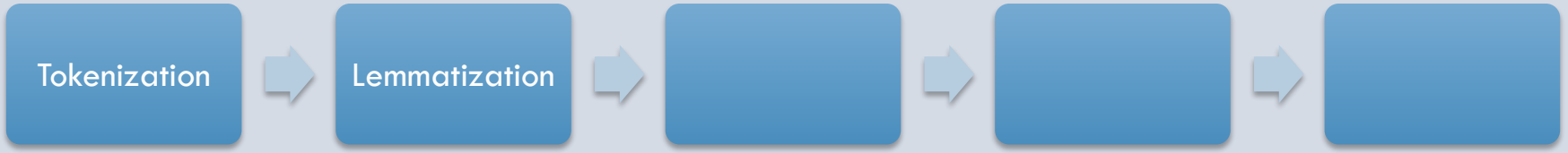
THE NLP “PIPELINE”



THE NLP “PIPELINE”



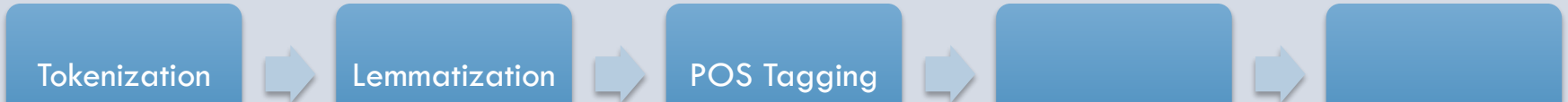
THE NLP “PIPELINE”

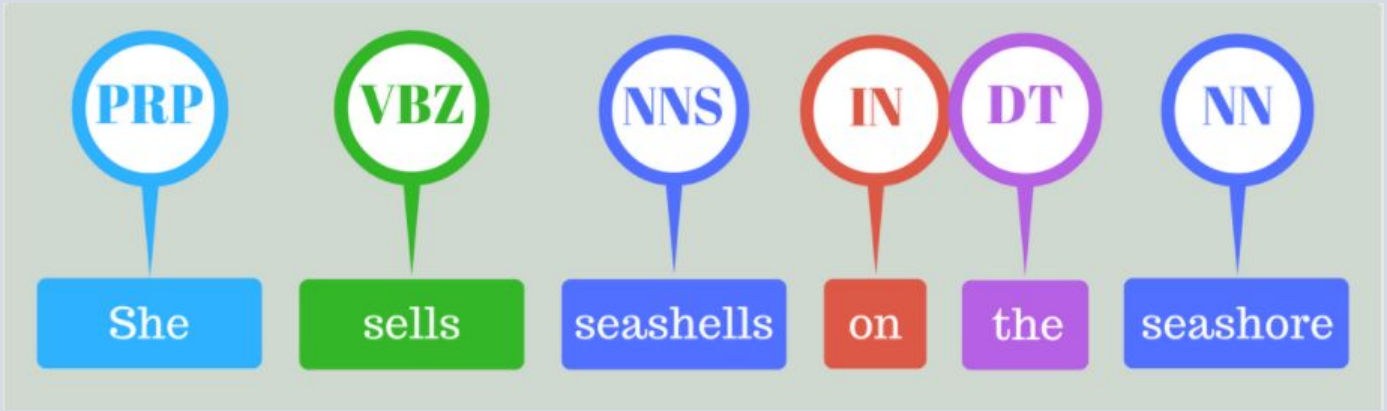


It’s a truth universally disputed

it be a truth universally disputed

THE NLP “PIPELINE”





THE NLP “PIPELINE”

(Herrmann, 2018)

	moot	TreeTag-ger	MarMoT	Percept-ron
ADJA	94,5	93	93,5	95
ADJD	85,37	79,67	78,05	76,42
ADV	81,2	72,93	75,56	68,42
NE	75,25	61,87	63,55	87,63
NN	92,81	93,46	92,32	91,67

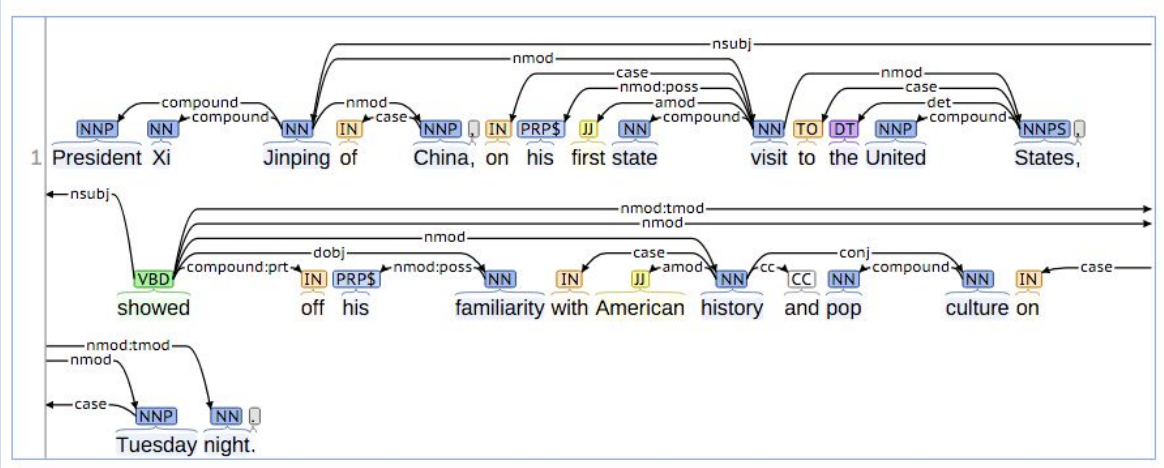
Tabelle 2: *Genauigkeit einiger STTS-Tags über Tagger (in Prozent)*

THE NLP “PIPELINE”

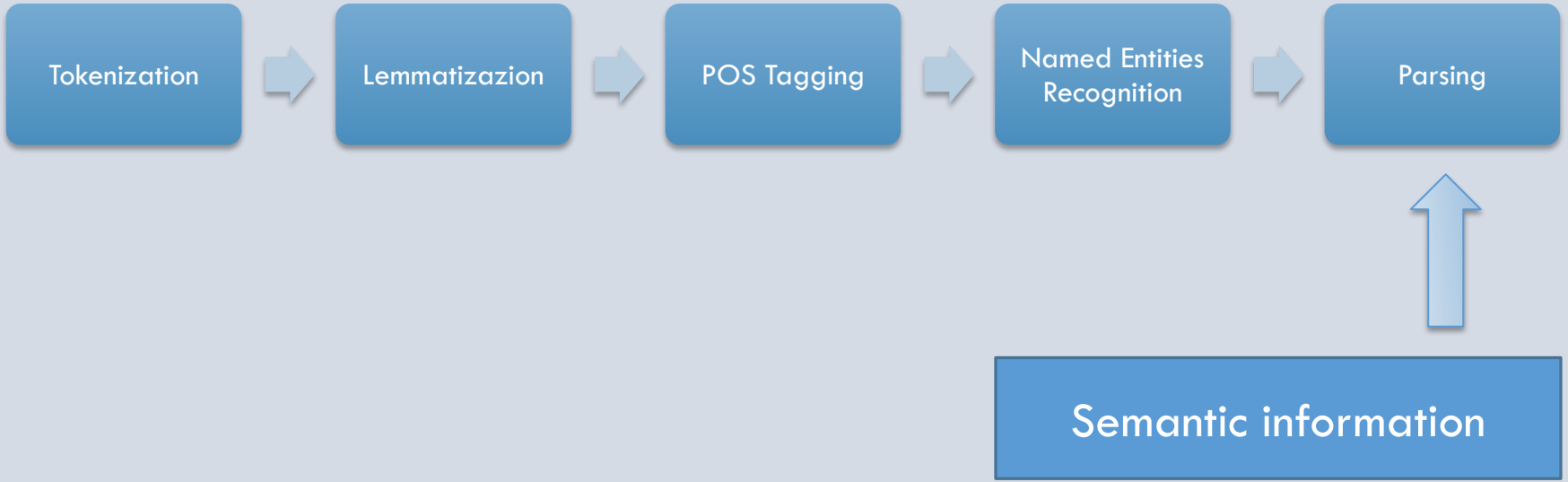
1 President Xi Jinping (Person) of China (Loc), on his first (ORDINAL) state visit to the United States (Location), showed off his familiarity with American (Misc) history and pop culture on Tuesday (Date) night (Time).

THE NLP “PIPELINE”

Tokenization → Lemmatizazion → POS Tagging → Named Entities Recognition → Parsing



THE NLP “PIPELINE”



UDPIPE

https://ufal.mff.cuni.cz/udpipe
https://cran.r-project.org/web/packages/udpipe/index.html

* i.e. a bottom-up approach, that lets the machine learn a task based on human-made examples, instead of following precise rules

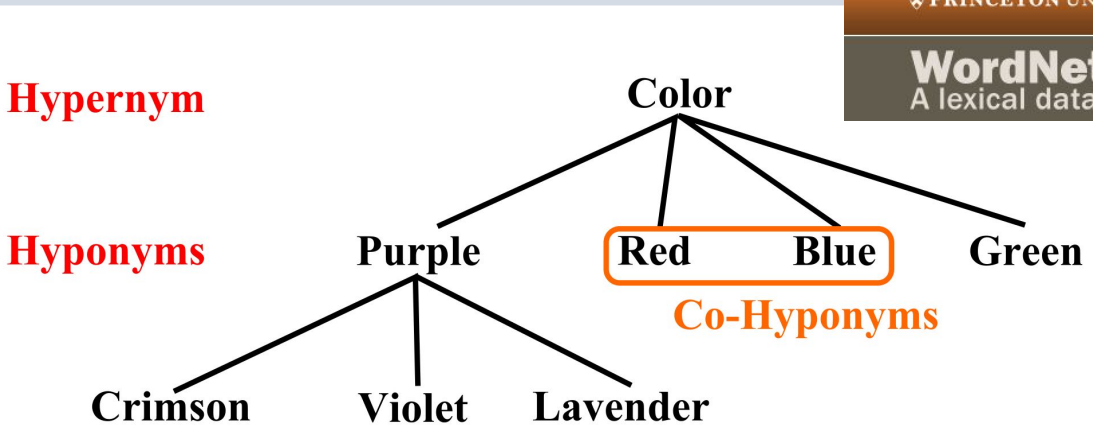
- An advanced NLP tool based on machine learning* and working on multiple languages
- UD stands for “universal dependencies”, as it tries to map all human languages into a single framework

NLP SEMANTICS

Tokenization → Lemmatizazion → POS Tagging → Named Entities Recognition → Parsing

PRINCETON UNIVERSITY

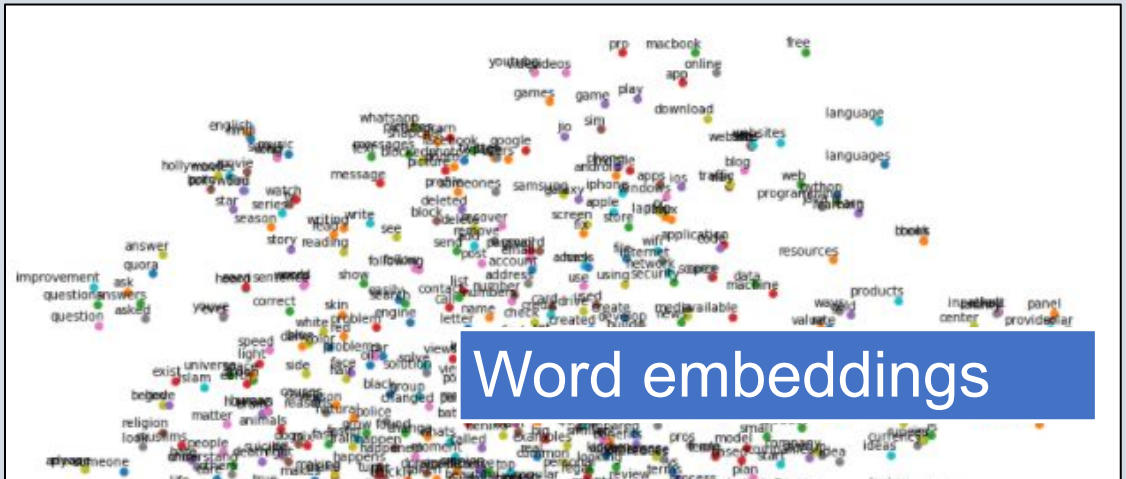
WordNet
A lexical database for English



Pros: well-structured and precise
Cons: very rigid, imposed from above

NLP SEMANTICS

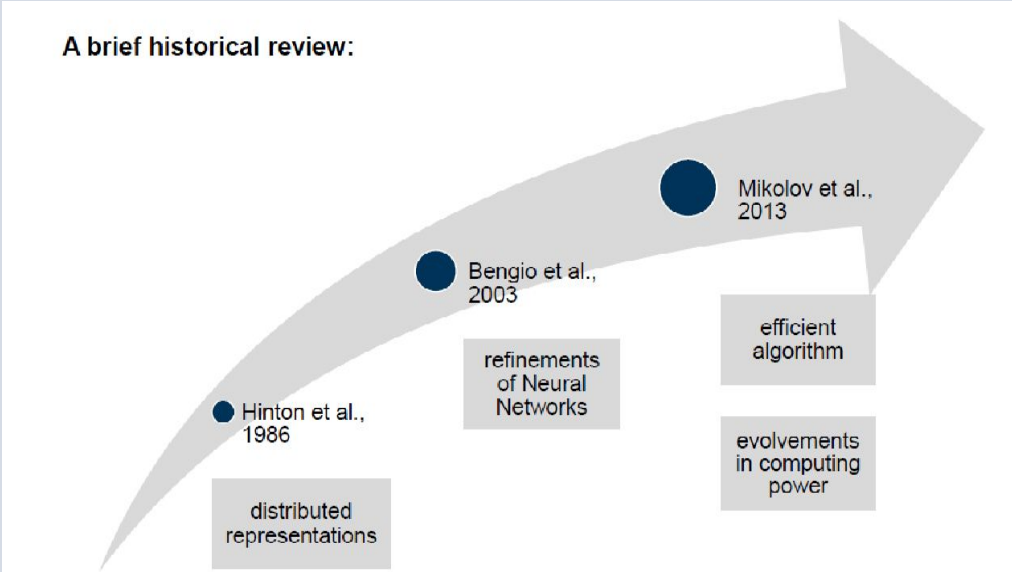
Tokenization → Lemmatizazion → POS Tagging → Named Entities Recognition → Parsing



Semantic information

DISTRIBUTIONAL SEMANTICS

- Instead of providing a pre-compiled description of semantic relationships...
- Let them emerge from the texts themselves (a huge amount of…)
- Fundamental concept: word collocations



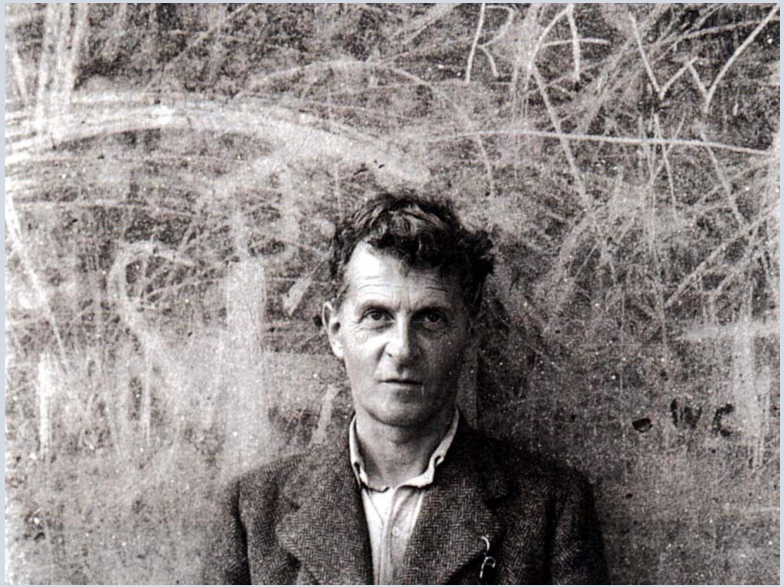
COLLOCATIONS

“The meaning of a word is its use in the language”

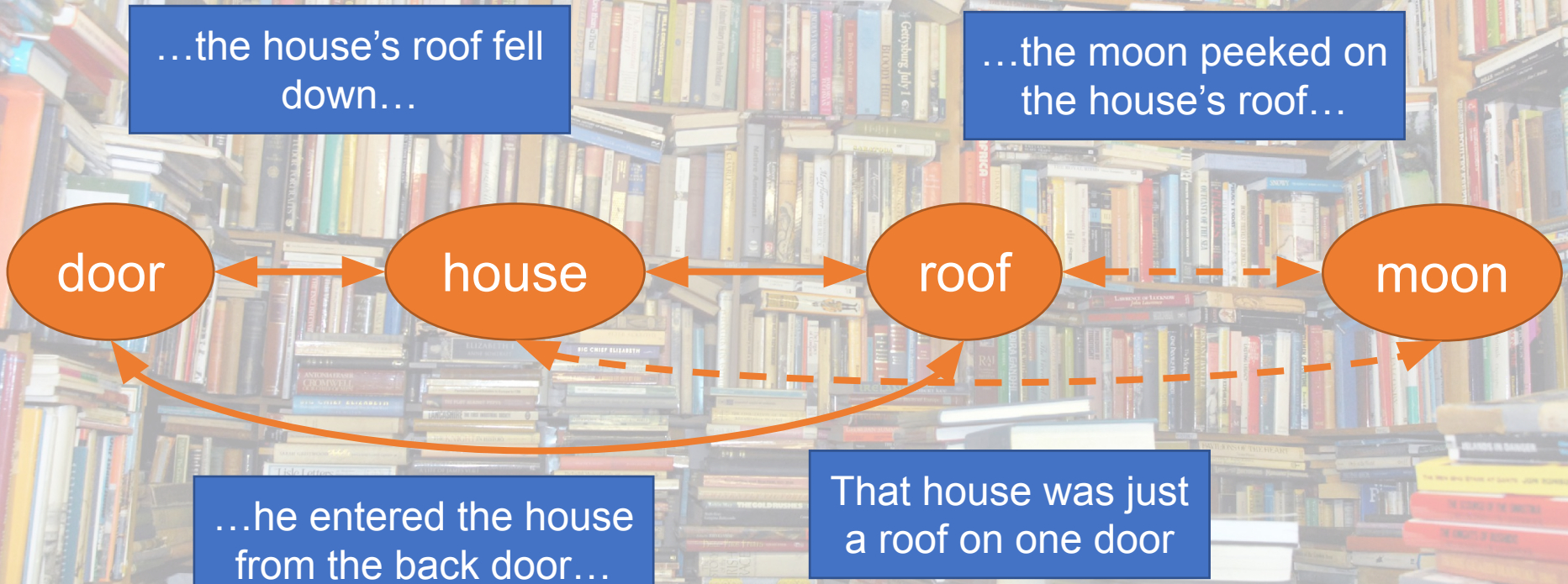
(Wittgenstein, 1953)

“You shall know a word by the company it keeps”

(Firth, 1957)

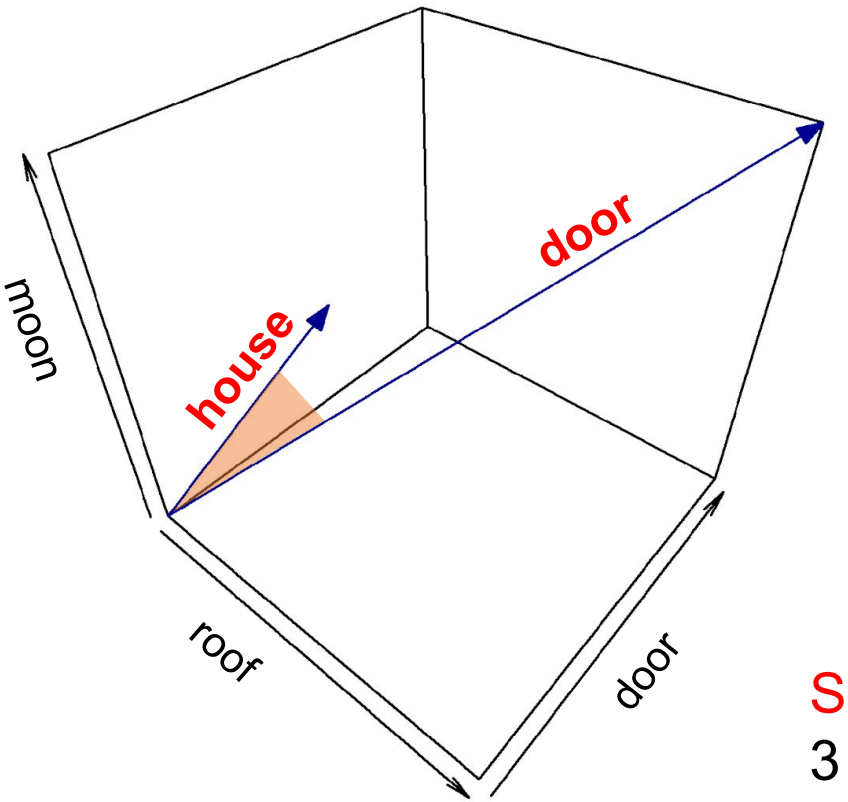


COLLOCATIONS



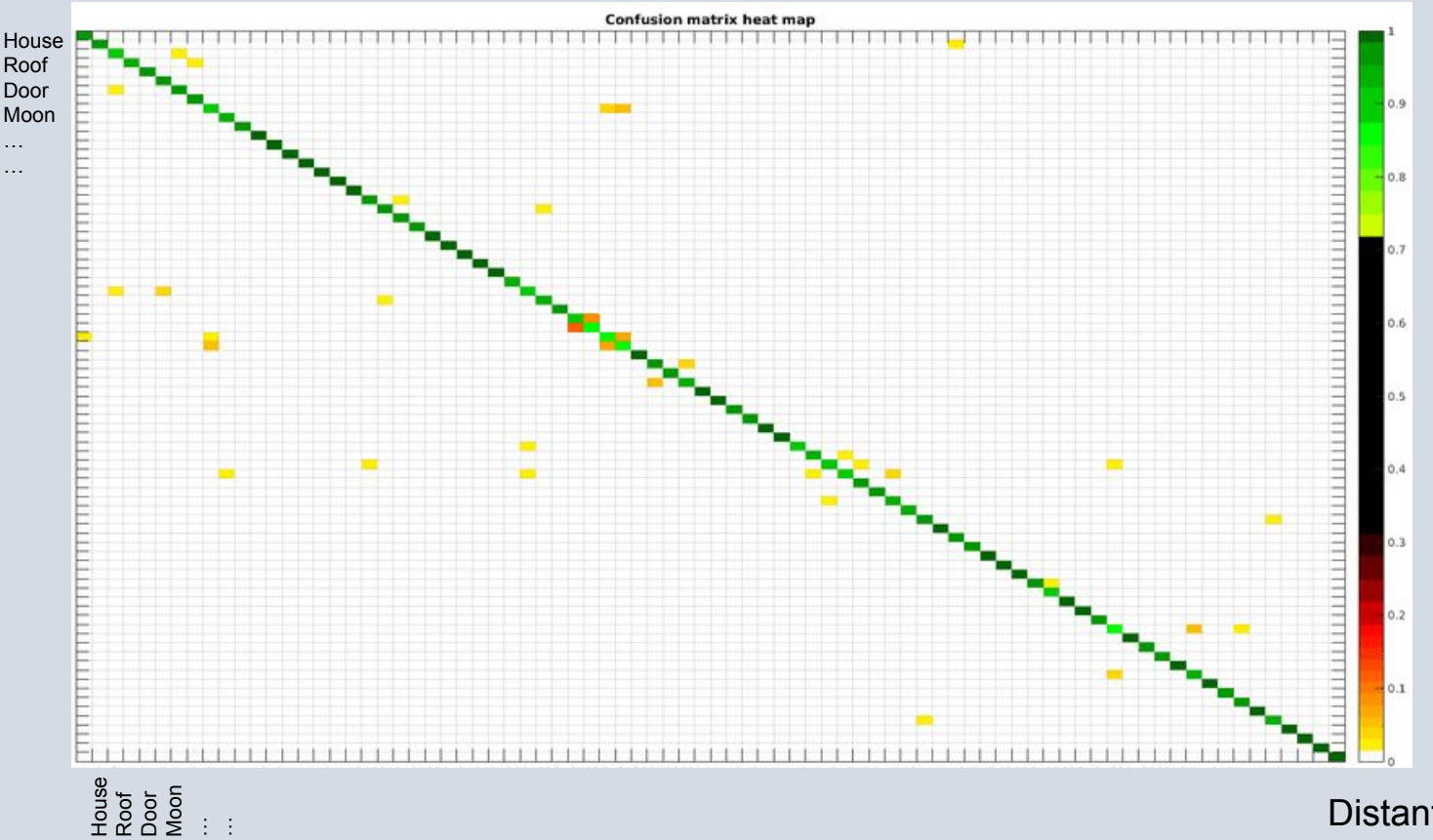
DISTRIBUTIONAL SEMANTICS AND WORD REPRESENTATION

	house	roof	door	moon
house	800	1,000	1,000	10
roof	1,000	700	10	50
door	1,000	10	900	5
moon	10	50	5	500

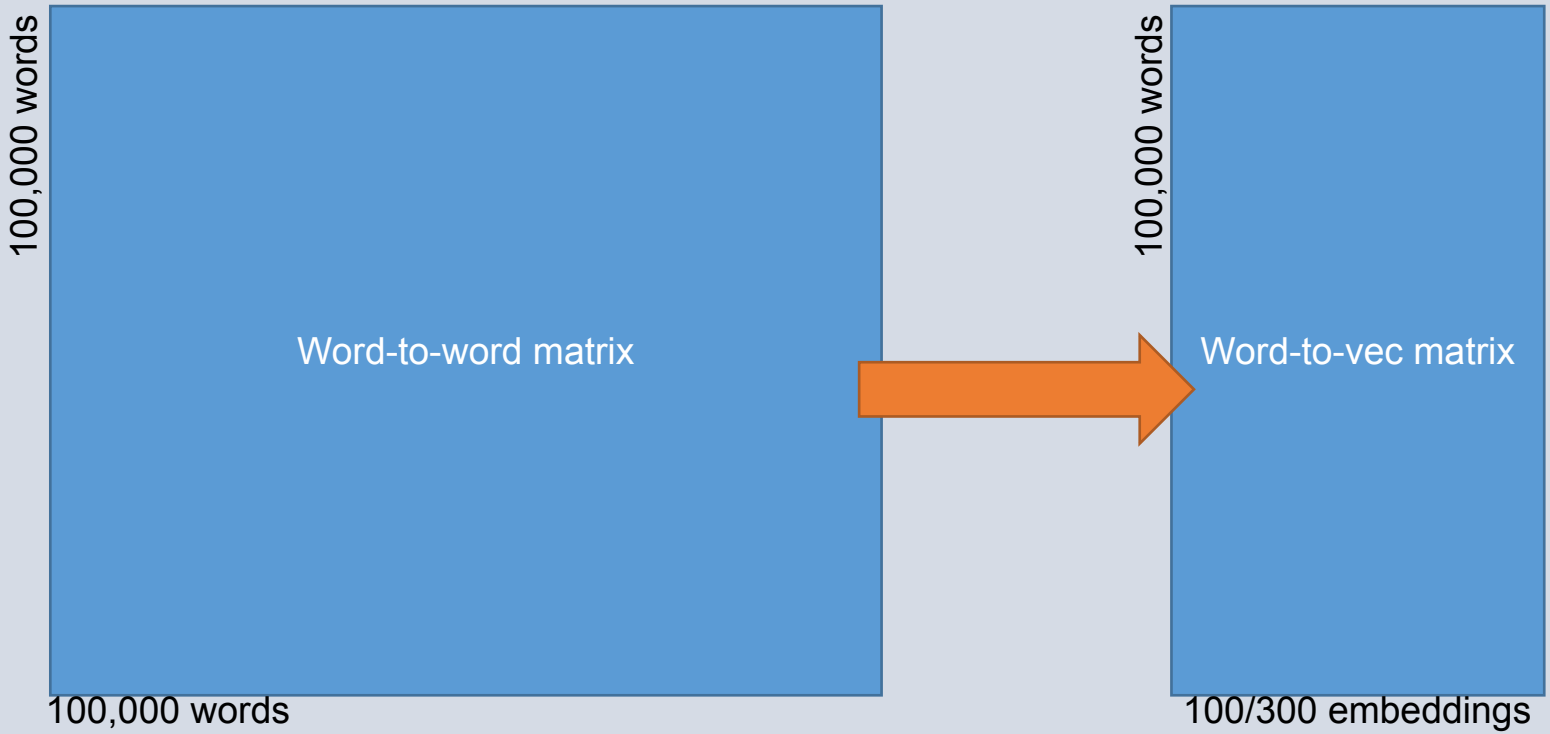


Simplification:
3 dimensions instead of thousands!!!

ISSUE: SPARSE MATRIXES

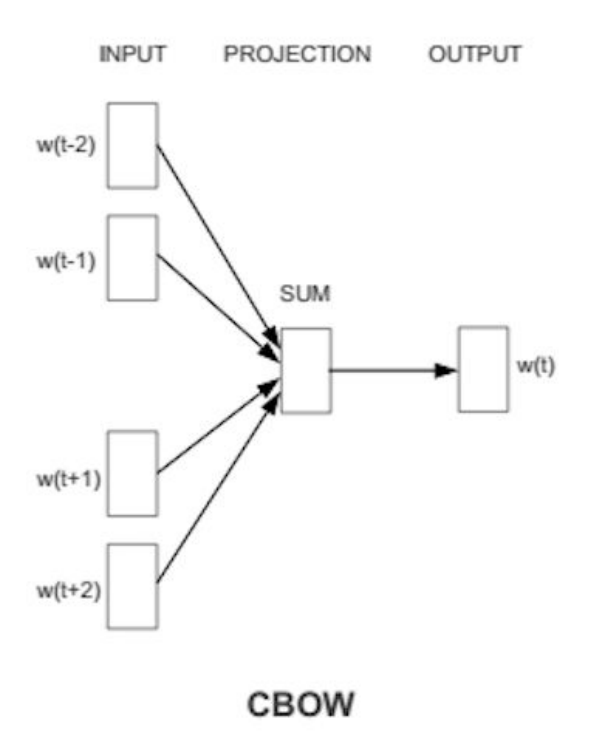


WORD EMBEDDINGS



(Mikolov et al. 2013)

MACHINE LEARNING & WORD EMBEDDINGS



"try to guess the …. word in this sentence"

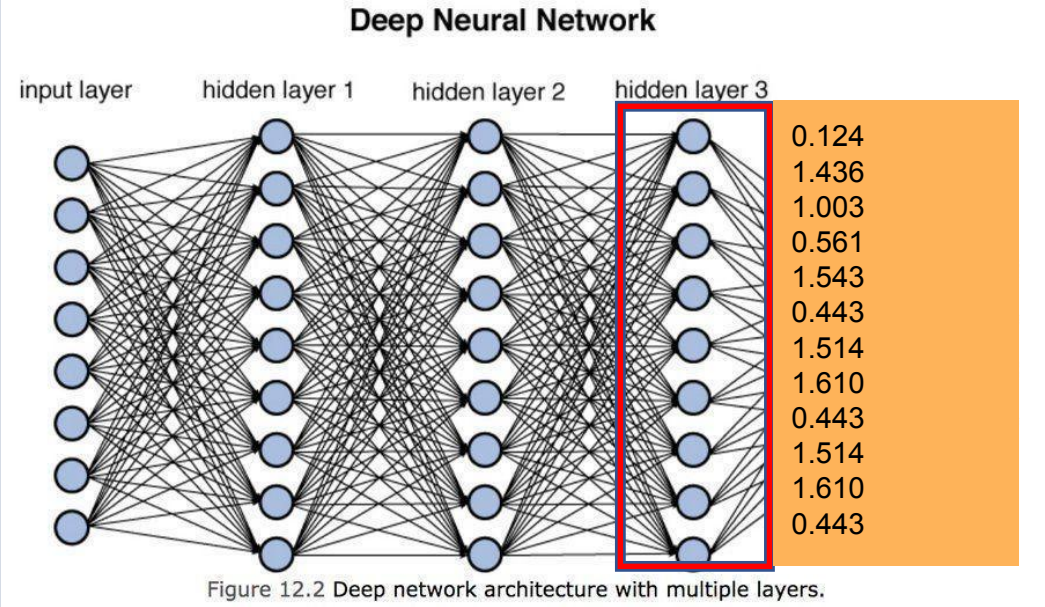
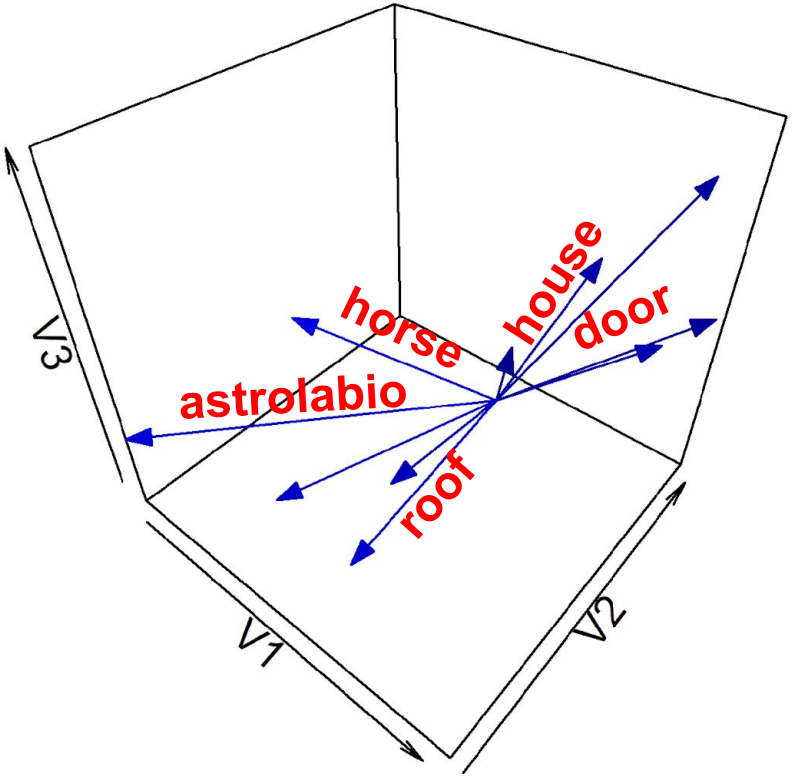


Figure 12.2 Deep network architecture with multiple layers.

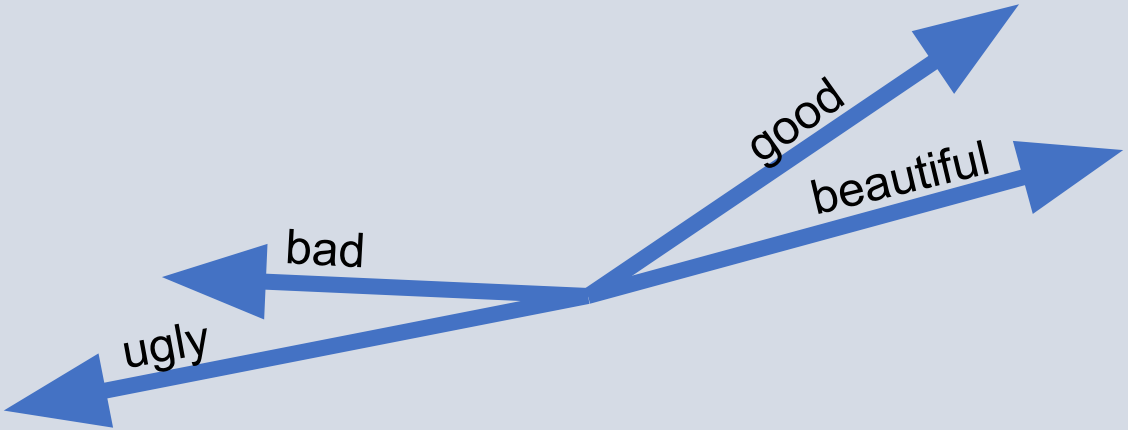
WORD EMBEDDINGS

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	...
house	2,854762725	4,334456292	10,82693154	3,434692738	12,65838366	7,457011416	9,208898264	12,02768361	10,09581653	3,244203478	0,612354558	...
roof	8,580630919	5,718236314	26,80190382	2,590861111	9,222107371	11,72907656	14,38706916	6,861195597	1,010863595	22,81134719	6,273506782	...
door	13,44130848	0,666201187	3,361132364	16,13986721	24,81150914	10,64126357	22,23819846	8,194863507	2,187735831	0,803091312	6,75278449	...
moon	20,26732886	14,68581974	13,37060907	4,162869764	1,166874822	5,738900901	5,920910053	7,247616555	9,122643759	7,82971818	2,000019266	...

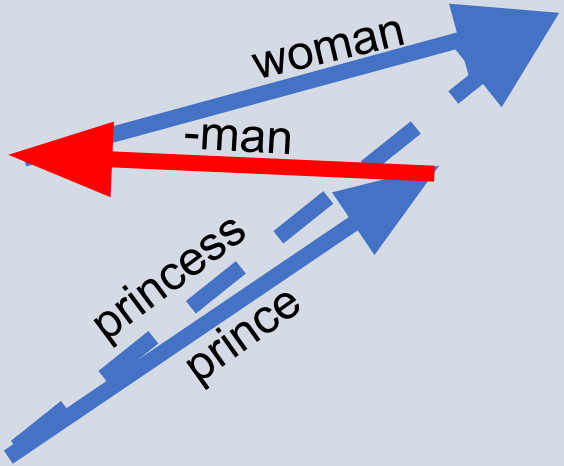


Simplification:
3 dimensions instead of hundreds!!!

WORD EMBEDDINGS' PROPERTIES



WORD EMBEDDINGS' PROPERTIES



WORD EMBEDDINGS' PROPERTIES

Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

(Mikolov et al. 2013)

APPLICATIONS

PNAS Proceedings of the National Academy of Sciences of the United States of America

RESEARCH ARTICLE

How quantifying the shape of stories predicts their success

Olivier Toubia, Jonah Berger, and Jehoshua Eliashberg

+ See all authors and affiliations

PNAS June 29, 2021 118 (26) e2011695118; https://doi.org/10.1073/pnas.2011695118

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved May 4, 2021 (received for review June 7, 2020)

Significance

APPLICATIONS

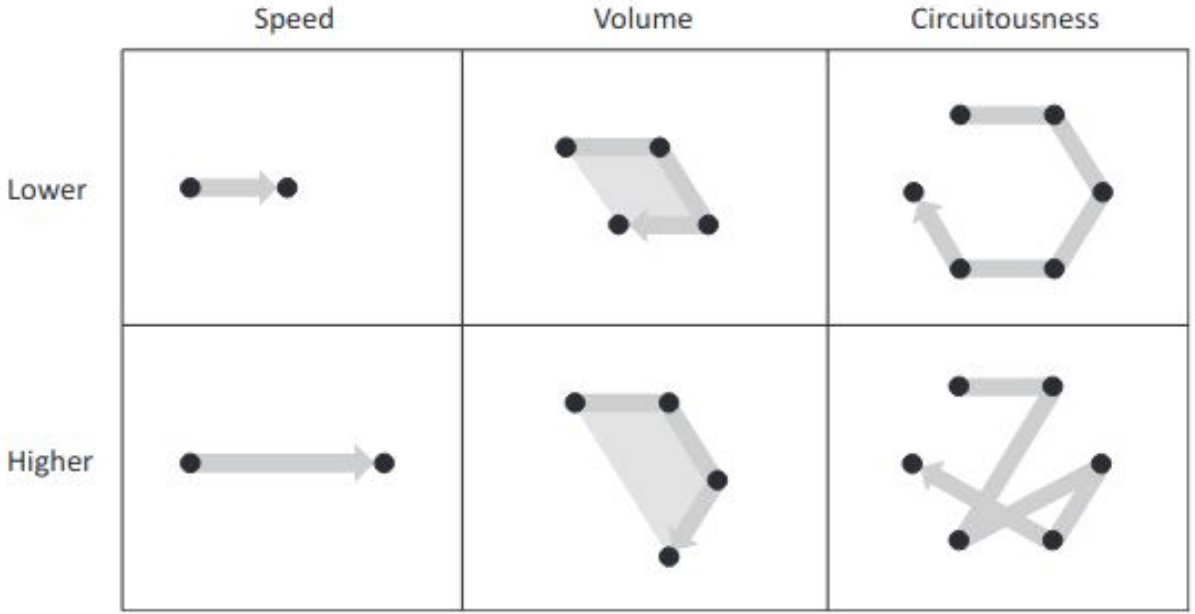


Fig. 1. Stylized illustration of the measures. Note that higher speed means more distance was covered in the same number of periods. Higher volume means that more ground was covered in the same number of periods. Higher circuitousness means that a less direct route was taken between a set of points.

CRITICAL ASPECTS

"...the best way for acquiring human-level semantics is to have machines learn through (physical) experience: if we want to teach a system the true meaning of "bumping into a wall," we simply have to bump it into walls repeatedly."

(Kiela et al. 2016)

CRITICAL ASPECTS

“Our findings show that these embeddings are lacking in basic features of perceptual meaning. These results suggest that distributional meaning (as operationalized by modern distributional models) may miss out on fundamental elements of semantics.”

(Lucy & Gauthier 2017)

CRITICAL ASPECTS

"The notion of similarity is challenging to define precisely. Existing word similarity data sets typically contain a broad range of semantic relations such as synonymy, antonymy, hypernymy, co-hypernymy, meronymy and topical relatedness. […] When human judges annotate word pairs for similarity, the distinctions in meaning they are asked to make are often very subtle, especially in the absence of context. For instance, the normalized similarity scores provided by 13 annotators for the pair "tiger–cat" range from 0.5 to 0.9" (Batchkarov et al. 2016)