

Dealing with network of texts



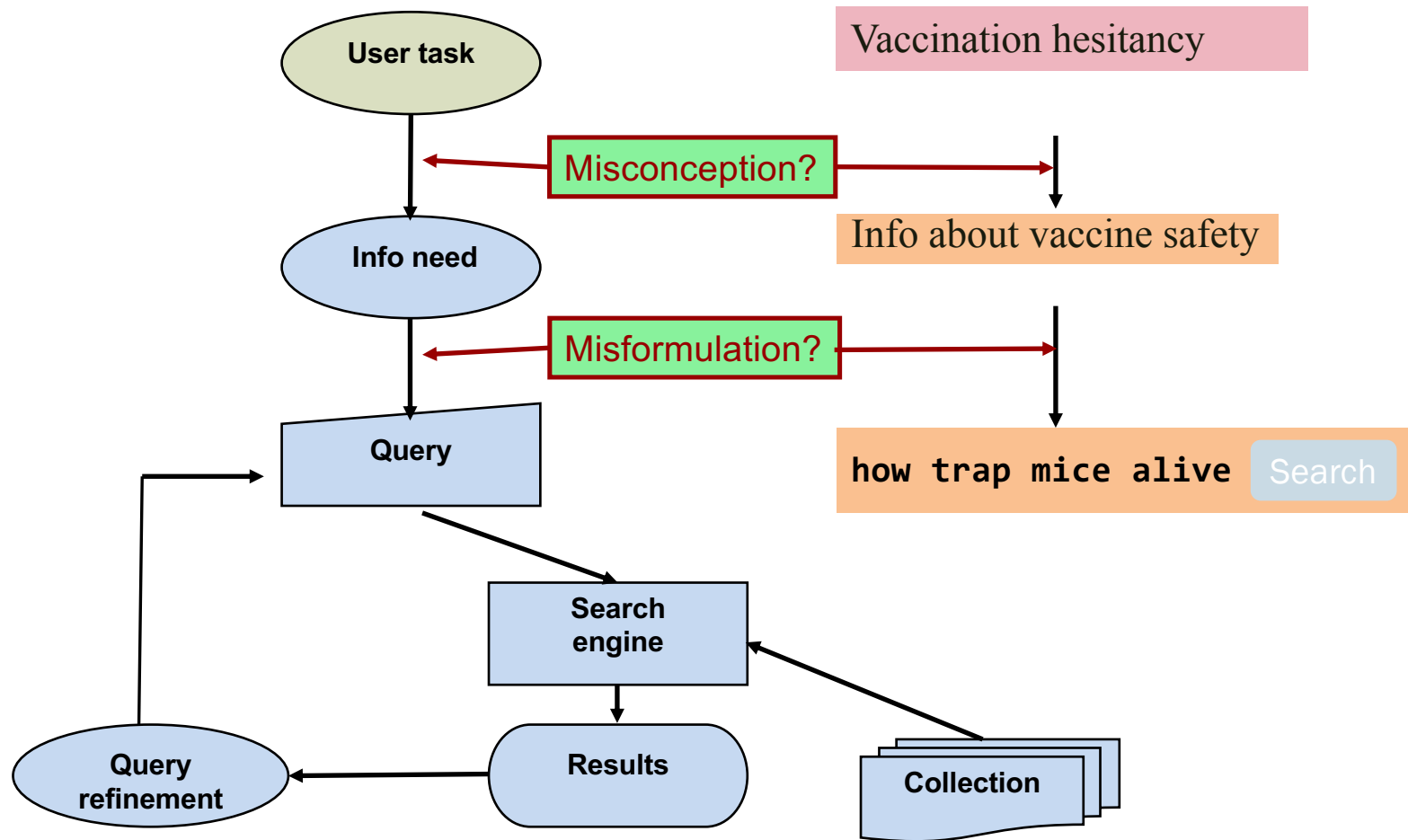
Semantic Networks: a definition

- WHAT graphical representations of knowledge based on meaningful relationships of written text, structured as a network of labeled nodes cognitively related to one another
- WHY GOAL: extract meanings
- HOW semantic networks connect words to words/hashtags/phrases, based on their co-occurrence
- WHO human and computerized methods, dealing with challenges such as co-reference resolution, synonym resolution, and ambiguity

COLLECT DATA: Information Retrieval IR

- Information Retrieval (IR) is **finding material** (usually documents) of an **unstructured** nature (usually text) that satisfies an **information need** from within **large collections** (usually stored on computers).
 - *These days we frequently think first of web search, but there are many other cases:*
 - E-mail search
 - Searching your laptop
 - Corporate knowledge bases
 - Legal information retrieval
 - Archival corpuses of legislations, parliamentary debates

The classic search model



How good are the retrieved docs?

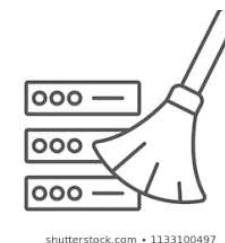


- *Precision* : “purity” Fraction of retrieved docs that are relevant to the user’s information need (reject irrelevant)



- *Recall* : “completeness” Fraction of relevant docs in collection that are retrieved (select relevant)

CLEAN DATA



Pre-processing starts the text preparation into a more structured representation.



- 1) **Tokenization:** Tokenization is used to identify all words in a given text.
- 2) **Data Filtering:** People use a lot of casual language on twitter. To improve this and make words more similar to generic words, such sets of repeated letters are replaced by two occurrences.

haaaaappy -> haappy.



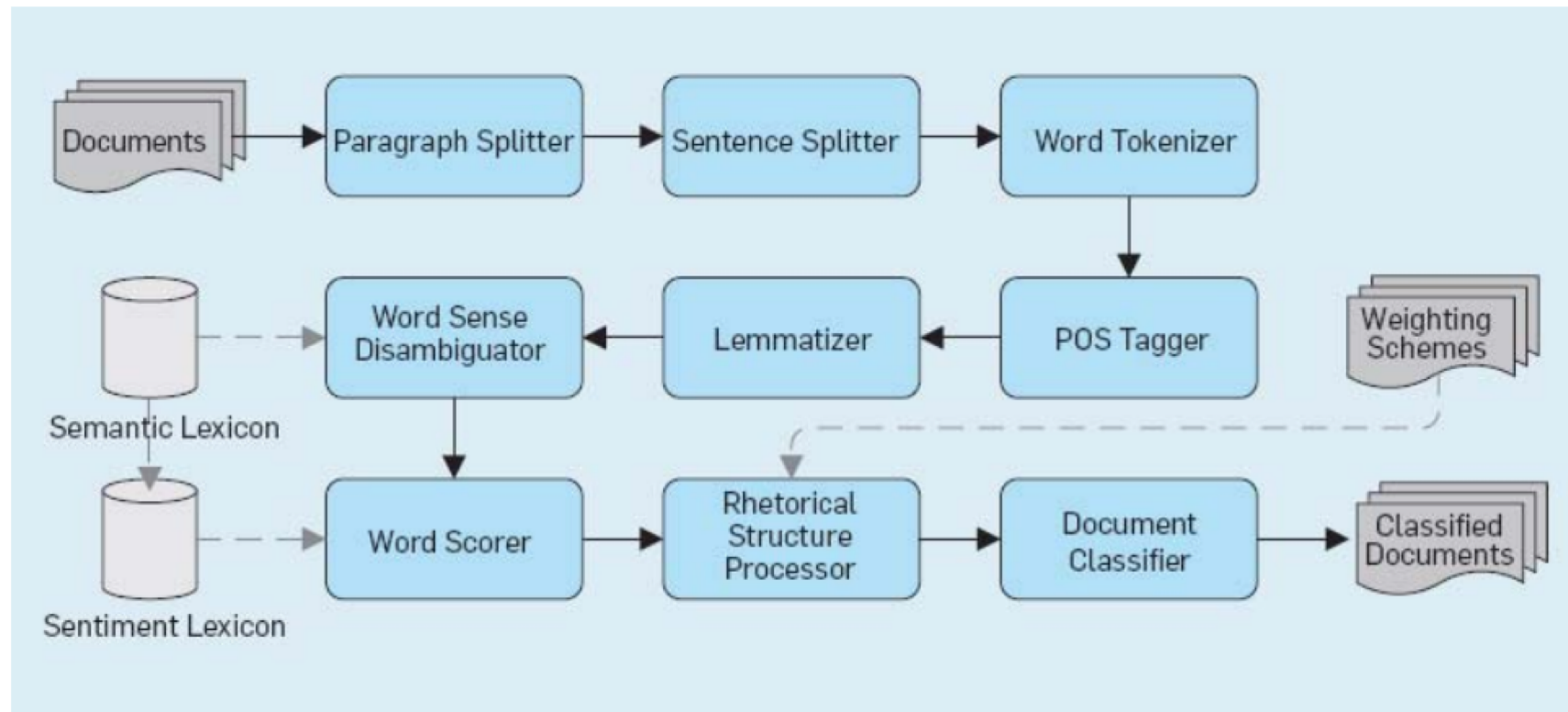
- 3) **Stop Word Removal:** Is used to eliminate that words that occurs frequently such as article, prepositions, conjunction and adverbs. These stop words depends on language of the text in questions. For example, words like the, and, before, while, and so on do not contribute to the sentiment.



- 4) **Stemming:** In information retrieval, stemming is the process of reducing a word to its root form.

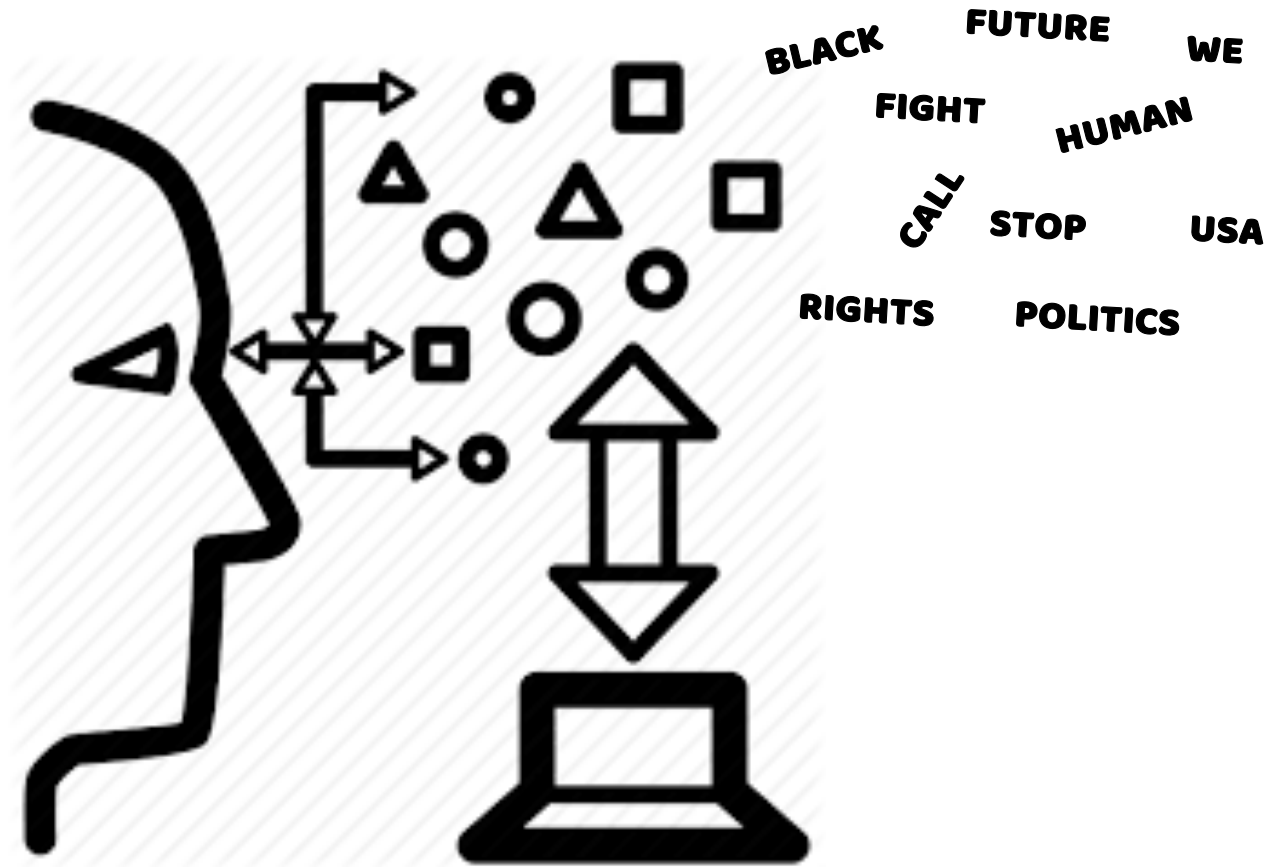
walking, walker, walked ->walk

Data preparation



PROCESS DATA

Dealing with textual data: from text to numbers





Words or Hashtags



- Top down semantic/sentiment classification: bag of words
 - Bottom up semantic/sentiment classification: human coding
 - Meta-semantic classification: pronouns, nouns, verbs, adjectives
 - Meta-semantic structural properties: word order, dropping
 - Semantic & grammar: future/past/present tense
- topical signifier : shared conversation marker,
 - can also represent the context of a tweet
 - flag an individual's community membership
 - indicate shared interests

Dealing with textual data: from text to numbers



Theory Driven



Human Coding



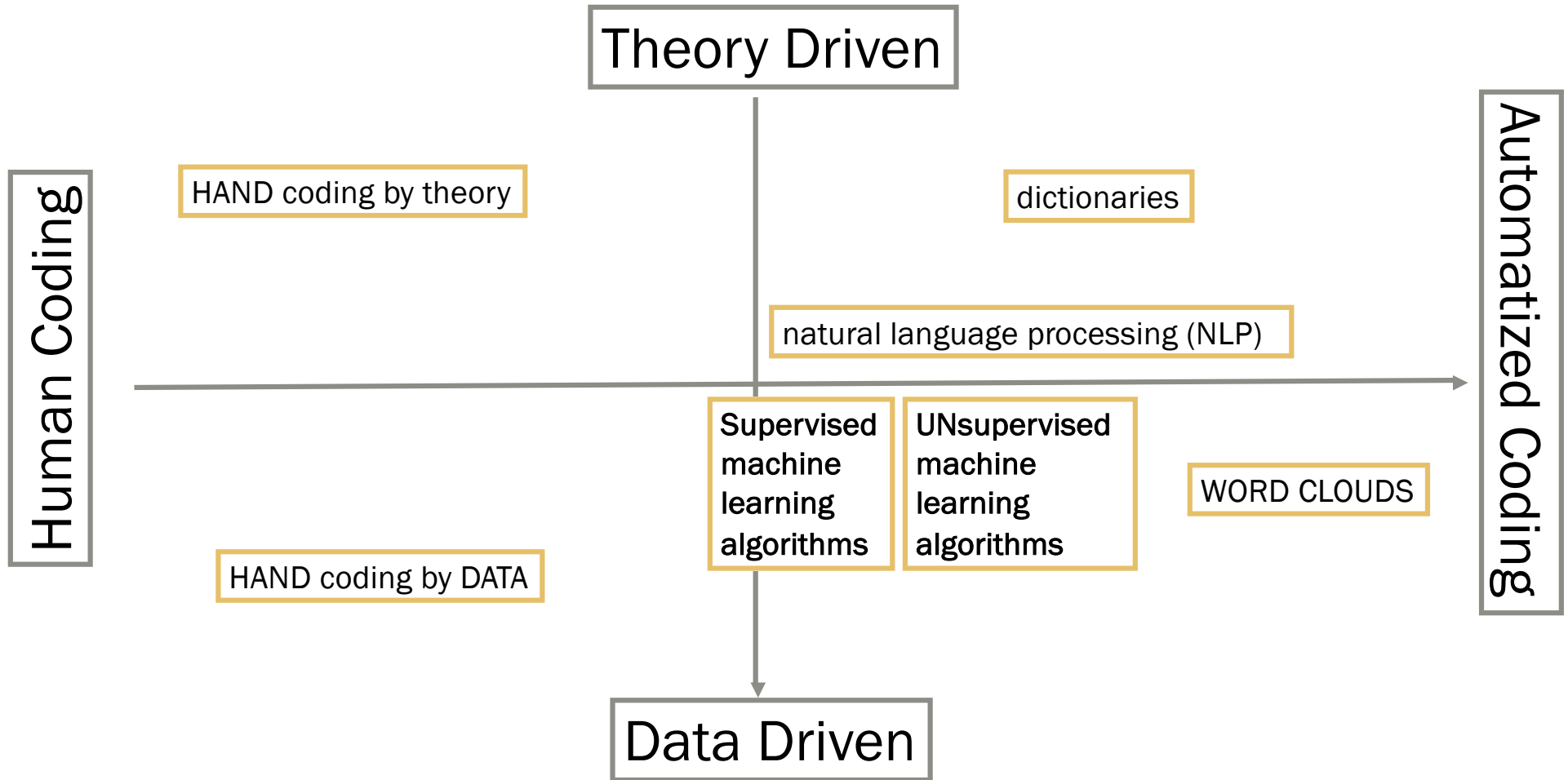
Automatized Coding



Data Driven

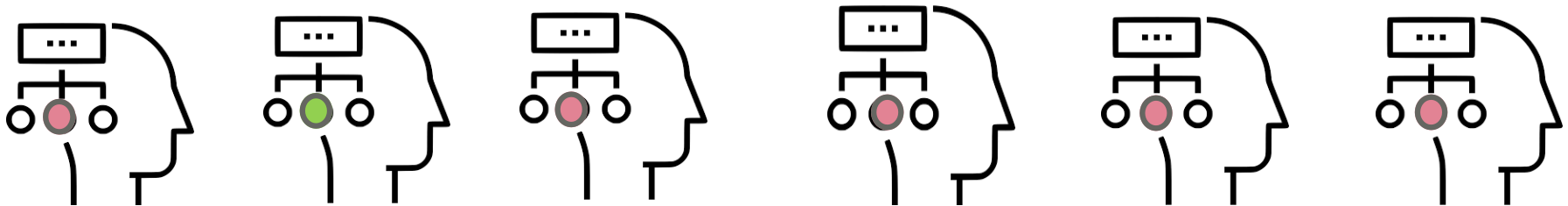


Dealing with textual data: from text to numbers



Human Coding

- *top down (coding by theory)*: initial coding scheme developed from the from pre-existing theory or assumptions
- *bottom up (grounded theory)*: initial coding scheme developed from the data
- *THE SUBJECTIVITY ISSUE: intercoder & intracoder reliability*
 - a classification procedure is reliable when it is consistent: Different people should code the same text in the same way



Dictionaries

- A **sentiment analysis dictionary** contains information about the emotions or polarity expressed by words, phrases, or concepts. In practice, a **dictionary** usually provides one or more scores for each word. We can then use them to compute the overall **sentiment** of an input sentence based on individual words.
- top down
- create you own dictionary
- Use a dictionary developed by other scientists
- LIWC, bing (in R), WordNet (Miller, 1990)

LIWC... *Psychometrics of Word Usage* ^{\$109.74}

The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods

Journal of Language and Social Psychology
29(1) 24-54

© 2010 SAGE Publications
DOI: 10.1177/0261927X09351676
<http://jls.sagepub.com>



Yla R. Tausczik¹ and James W. Pennebaker¹

Abstract

We are in the midst of a technological revolution whereby, for the first time, researchers can link daily word use to a broad array of real-world behaviors. This article reviews several computerized text analysis methods and describes how Linguistic Inquiry and Word Count (LIWC) was created and validated. LIWC is a transparent text analysis program that counts words in psychologically meaningful categories. Empirical results using LIWC demonstrate its ability to detect meaning in a wide variety of experimental settings, including to show attentional focus, emotionality, social relationships, thinking styles, and individual differences.

https://s3-us-west-2.amazonaws.com/downloads.liwc.net/LIWC2015_OperatorManual.pdf

LIWC

Summary Variable	Informal Speech	informal
Analytical Thinking	Swear words	swear
Clout	Netspeak	netspeak
Authentic	Assent	assent
Emotional Tone	Nonfluencies	nonfl
	Fillers	filler

With the exception of the summary variables and words per sentence, all LIWC2015 output variables are expressed as percentage of total words.

All Punctuation ⁵	Allpunc
Periods	Period
Commas	Comma
Colons	Colon
Semicolons	SemiC
Question marks	QMark
Exclamation marks	Exclam
Dashes	Dash
Quotation marks	Quote
Apostrophes	Apostro
Parentheses (pairs)	Parenth
Other punctuation	OtherP

Language Metrics	
Words per sentence ¹	WPS
Words>6 letters	Sixltr
Dictionary words	Dic
Function Words	function
Total pronouns	pronoun
Personal pronouns	ppron
1st pers singular	i
1st pers plural	we
2nd person	you
3rd pers singular	shehe
3rd pers plural	they
Impersonal pronouns	ipron
Articles	article
Prepositions	prep
Auxiliary verbs	auxverb
Common adverbs	adverb
Conjunctions	conj
Negations	negate

Grammar Other	
Regular verbs	verb
Adjectives	adj
Comparatives	compare
Interrogatives	interrog
Numbers	number
Quantifiers	quant

LIWC

People experiencing physical or emotional pain tend to use more first-person singular pronouns (Rude, Gortner, & Pennebaker, 2004).

Depressed patients are more likely to use more first-person singular and more negative emotion words than participants who have never been depressed in emotional writings (Rude et al., 2004)

When people sit in front of a mirror use more words such as “I” and “me” than when the mirror is not present (Davis & Brock, 1975)

Colons	Colon
Semicolons	SemiC
Question marks	QMark
Exclamation marks	Exclam

Other punctuation	OtherP
-------------------	--------

Language Metrics	
Words per sentence ¹	WPS
Words>6 letters	Sixltr
Dictionary words	Dic
Function Words	function
Total pronouns	pronoun
Personal pronouns	ppron
1st pers singular	i
1st pers plural	we
2nd person	you
3rd pers singular	shehe
3rd pers plural	they
Impersonal pronouns	ipron
Articles	article
Prepositions	prep
Auxiliary verbs	auxverb
Common adverbs	adverb

Regular verbs	verb
Adjectives	adj
Comparatives	compare
Interrogatives	interrog
Numbers	number

“we” can signal a sense of group identity, such as when couples are asked to evaluate their marriages to an interviewer, the more the participants use “we,” the better their marriage (Simmons, Gordon, & Chambless, 2005)

Psycho-social index

Social Words	social
Family	family
Friends	friend
Female referents	female
Male referents	male

Core Drives and Needs	drives
Affiliation	affiliation
Achievement	achieve
Power	power
Reward focus	reward
Risk/prevention focus	risk
Time Orientation⁴	
Past focus	focuspast
Present focus	focuspresent
Future focus	focusfuture
Relativity	relativ

Affect Words	affect
Positive emotion	posemo
Negative emotion	negemo
Anxiety	anx
Anger	anger
Sadness	sad

Personal Concerns	
Work	work
Leisure	leisure
Home	home
Money	money
Religion	relig
Death	death

Cognition & perception

Cognitive Processes²	cogproc
Insight	insight
Cause	cause
Discrepancies	discrep
Tentativeness	tentat
Certainty	certain
Differentiation ³	differ
Perpetual Processes	percept
Seeing	see
Hearing	hear
Feeling	feel
Biological Processes	bio
Body	body
Health/illness	health
Sexuality	sexual
Ingesting	ingest

Cognitive Processes²	cogproc
Insight	insight
Cause	cause
Discrepancies	discrep
Tentativeness	tentat
Certainty	certain
Differentiation ³	differ

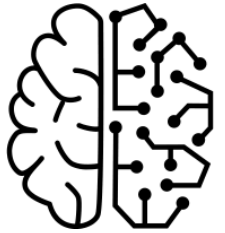
Natural language processing (NLP)

Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

- tokenization
- grammatical role POS (part of speech) tagging (subj, obj..)
- stemming
- thesauri
- shallow parsing : identifies constituent parts of sentences (nouns, verbs, adjectives, etc.)

the hand-coding of a set of rules, coupled with a dictionary lookup

Machine learning



Supervised machine learning algorithms apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

Unsupervised machine learning algorithms are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.

Content Analysis

- Detect systematic patterns in communication
 - -> *topic identification*

→ opinions

Sentiment Analysis

- extract, quantify, and study affective states and subjective information


→ attitudes

refers to the use of [natural language processing](#), [text analysis](#), [computational linguistics](#), and [biometrics](#) to systematically identify




ANALYSE DATA

- -> frequency
- -> correlations
- -> source comparison
- -> networks: centrality measures, community detection etc



THE RISE OF #CLIMATEACTION IN THE TIME OF THE FRIDAYSFORFUTURE MOVEMENT: A SEMANTIC NETWORK ANALYSIS

Caterina Suitner, Leonardo Badia, Damiano Clementel, Laura
Iacovissi, Matteo Migliorini, Bruno Gabriel Salvador Casara,
Domenico Solimini, Magdalena Formanowicz, Tomaso Erseghe



Theoretical framework

- Collective action-> any action addressing a goal that surpasses individuals interest (Van Zomeren et al., 2008)
- two central psychological predictors of protest engaging:
 - *affiliation (or identity)*
 - *empowerment*
 - + *future orientation: the tendency to foreseeing future events was positively associated to pro-environment behaviors (Sarigo 'llu ;2009)*

Data collection

- Posts on the social media site Twitter.
- English language
- March 1st, 2017 to April 19th, 2017
- March 1st, 2018 to April 19th, 2018
- March 1st, 2019 to April 19th, 2019
- The specific choice of intervals permits capturing the semantic of climate change discourses around two main events, namely the U.S. withdrawal from Paris Agreement in June 2017, and the first Strike for Climate on the 15th of March 2018

effectively used tweets to $N_{2017} = 3459$, $N_{2018} = 4031$, and $N_{2019} = 3931$.



Keyword identification

- sole hashtag #climatechange to identify the most relevant hashtags connected to the climate issue in 2017, 2018, and 2019, separately.
- 20 most frequent hashtags of each year
- <http://www.trendsmap.com/historical>
- top ranked neutral hashtags #climatechange, #climate, #sdgs, #sustainability, #environment, #globalwarming
- <http://www.trendsmap.com/historical>
-

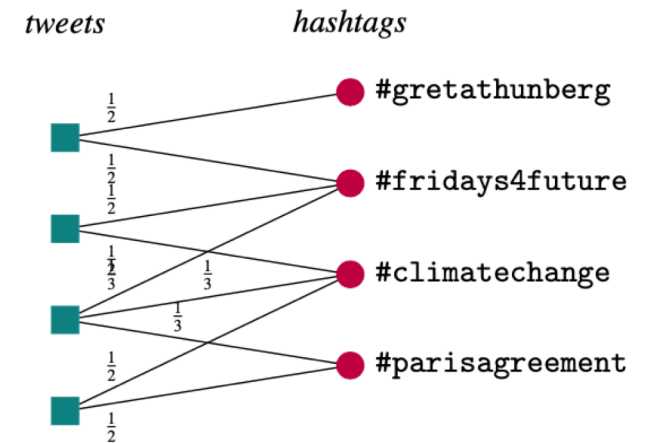
SEMANTIC CODING

- **Affiliation.** The LIWC score for the category *affiliation* (e.g., ally, friend, social) was used for measuring the in- group community orientation within the text. This proved to be a reliable index of implicit motives for affiliation (Schultheiss, 2013).
- **Group-identity salience.** The frequency of personal pro- nouns can be used to assess the salience of group member- ship. In particular, the first person plural pronouns (i.e., we) mark the sense of belonging (Zhang, 2010).
- **Empowerment.** We computed the empowerment scores aggregating with a mean the LIWC scores for the categories *power, achieve, reward, insight* and *cause*.(see Decter-Frain and Frimer, 2016; Pietraszkiewicz et al., 2019)
- **Temporal perspective.** The orientation of tweet to the past or future was measured using the specific LIWC categories of *past* (e.g., ago, did) and *future focus* (e.g., will, soon).

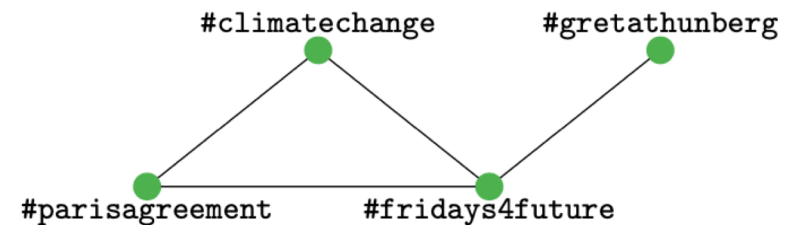
Network building

- tweets carry the semantics content
- while hashtags (the topics) may reveal those inter-dependencies that constitute the implicit holistic information
- bipartite graph linking each tweet to those hashtags that appear in the tweet.
- Projection activates a link only between those hashtags that appear together in a tweet at least once

(a) bipartite network



(b) projection

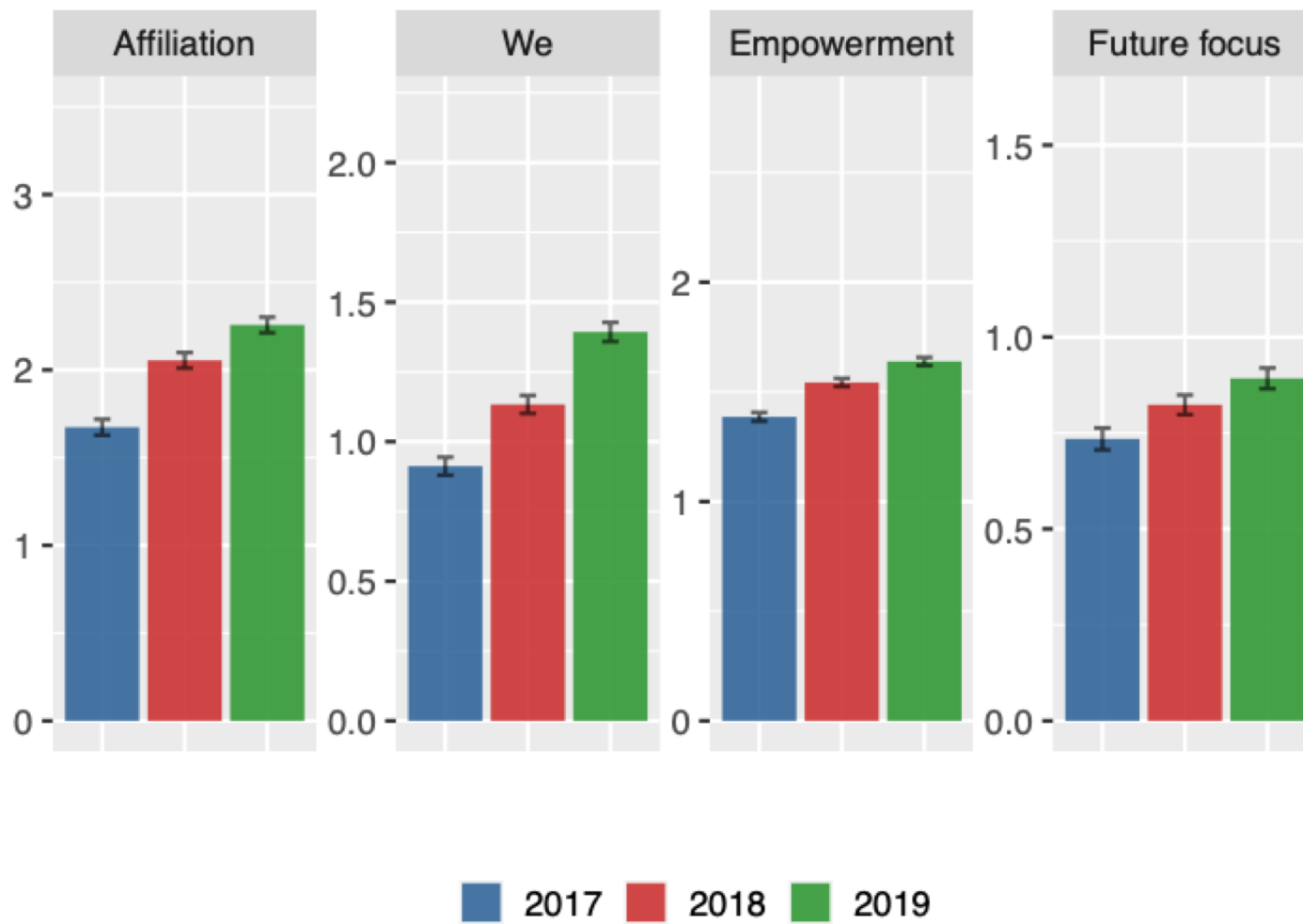




Community detection

- Louvain modularity (Blondel et al., 2008; Lancichinetti and Fortunato, 2009; Fortunato, 2010) is used to extract hashtags communities from the projected network
- A tweet will then be assigned to the community it is most similar to.

#	Community name	Descriptive hashtags	Brief description
1	climate action	#climateaction, #actonclimate, #energy, #science, #cdnpoli, #renewableenergy, #renewables, #greennewdeal, #climatestrike	calls to action related to climate change
2	nature	#nature, #earthday, #conservation, #biodiversity, #oceans, #ecology, #trees, #forests, #wildlife	photos ad videos about naturalistic environments and animals
3	recycling	#innovation, #circulareconomy, #plastic, #sustainabledevelopment, #recycling, #ecofriendly, #recycle	business solutions for the circular economy, and recycling techniques
4	work life	#leadership, #employment, #creativity, #partnerships, #decentwork, #career	professional-life and working environment aspects
5	developments goals	#globalgoals, #education, #parisagreement, #un, #2030agenda, #community, #migration, #teachsdgs	2030 Global Goals for Sustainable Development
6	green economy	#green, #eco, #sugarcane, #ecofashion, #sustainablefashion, #vegetarian	promoting green and eco-friendly products
7	international politics	#trump, #epa, #resist, #coal, #p2, #environmentaljustice, #tcot, #usa, #2a, #oil, #theresistance, #eu	political topics
8	digitalization	#ai, #iot, #dataviz, #data, #bigdata, #digital, #smartcity, #digitaltransformation, #smarthome	methods and procedures for the digital transformation and innovations
9	pollution and health	#health, #pollution, #airpollution, #cities, #healthforall, #publichealth, #wellbeing, #airquality, #worldhealthday	topics of air pollution and public health
10	lifestyle	#weather, #travel, #coffee, #worldmetday, #europe, #spring, #thursdaythoughts, #london, #sxsw, #snow, #summer, #noaa, #greenland	big variety of free-time-related topics
11	food	#agriculture, #food, #zerohunger, #foodsecurity, #regenerativeagriculture, #insect, #urbanfarming, #learn, #foodtech	food issues and food technologies
12	Australia	#auspol, #extinctionrebellion, #climatecrisis, #greatbarrierreef, #stopadani, #australia, #extinction, #factsmatter, #ausvotes, #actnowforfuture, #brisbane	climate collective actions in Australia
13	women	#genderequality, #women, #womensday, #gender, #internationalwomensday, #iwd2018, #sdg5, #unea4, #localgov, #solvedifferent, #women4climate	gender-related topics
14	green technology	#earth, #carbon, #jobs, #blockchain, #emissions, #cleantech, #engineering, #startups, #ghg, #electric, #natural, #paris, #life, #mining, #crypto	technological and sustainable innovations
15	architecture	#architecture, #fashion, #design, #construction, #greenbuilding, #building, #webinar, #steamdrills, #5star, #innovative, #free, #interiordesign	architecture topics
16	other	#agenda2030, #brexit, #news, #healthcare, #fracking, #ocean, #photography, #art, #wednesdaywisdom, #infrastructure, #climatejustice, #tourism, #mentalhealth	mixed topics



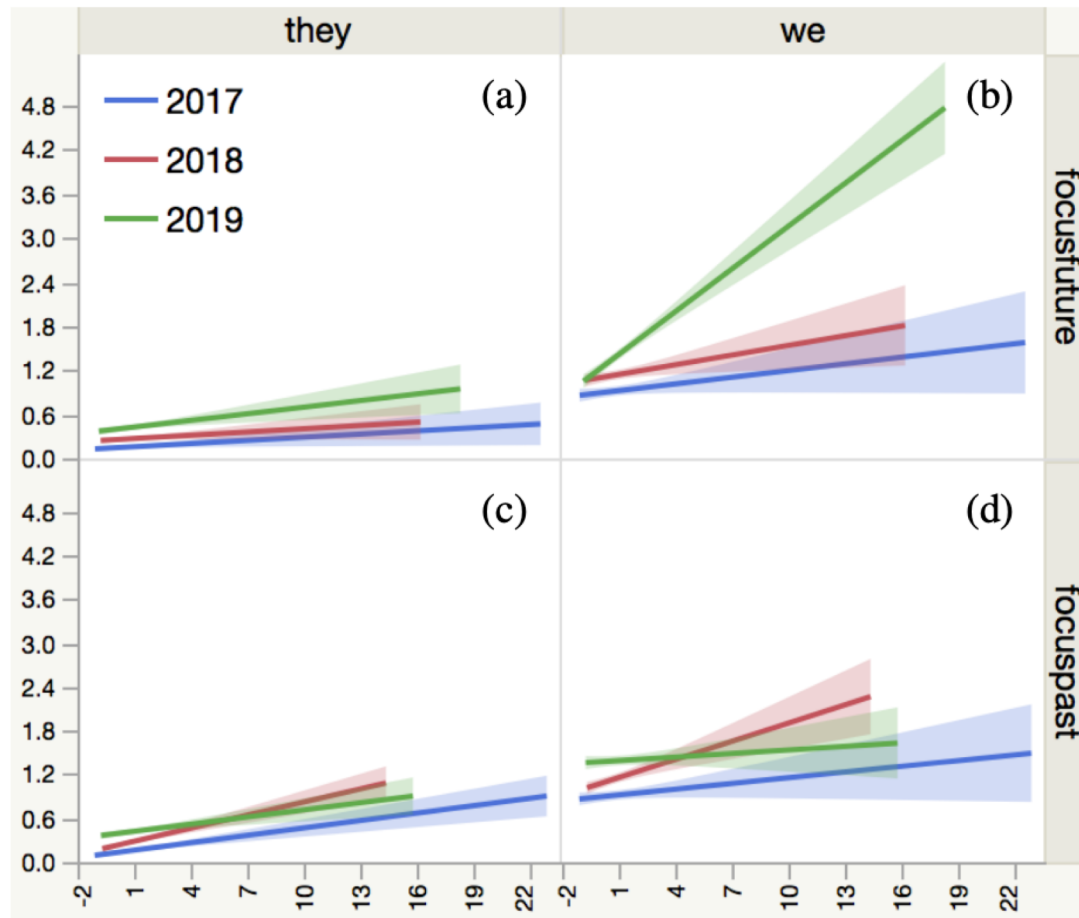
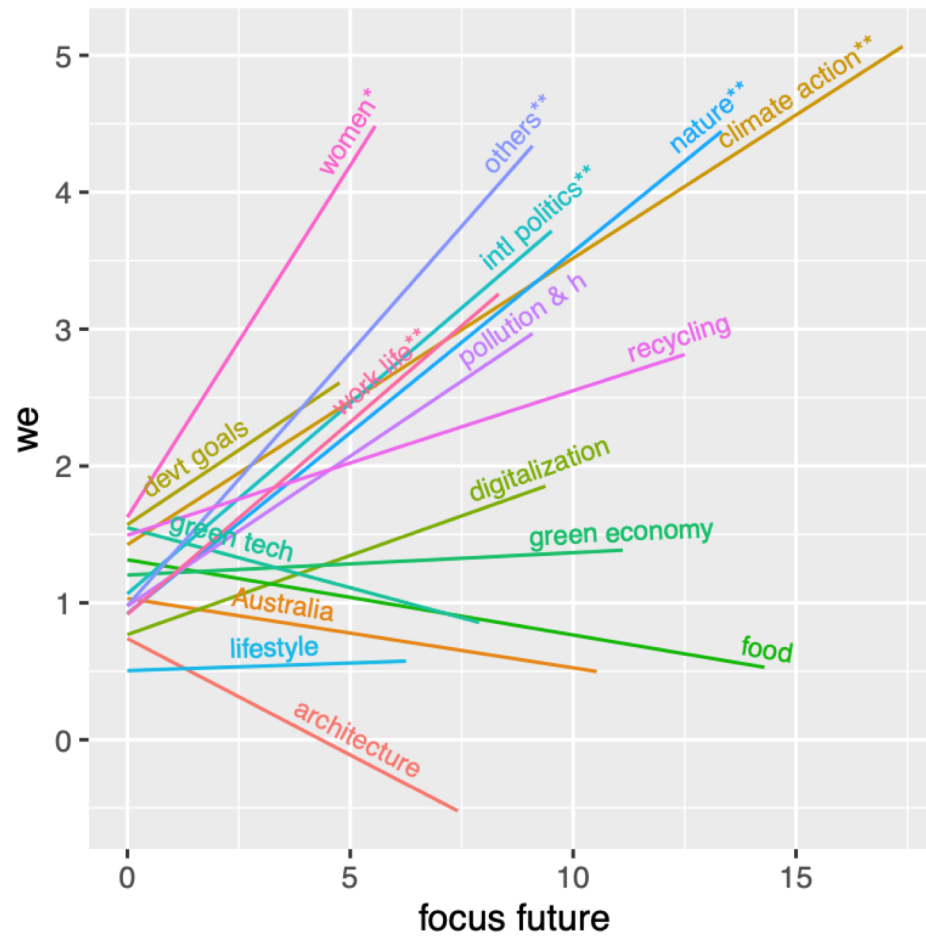



Figure 8: Linear regressions with confidence intervals over the three considered years for we/they versus past/future focus markers.

Linear regression of first person plural pronouns (we) as a function of future-framed wording (focus future) by community: an asterisk denotes a $p < 0.05$ significance of the slope coefficient, two asterisks a $p < 0.01$ significance.



- 
- increased relevance and the traits of affiliation, empowerment in the online discourse about climate action
 - the projection of the in- group into the future
 - this association is particularly striking in the community of climate action in which future words are the most used and clearly associated with the pronoun we.
 - identification of linguistic markers that specifically characterize the evolution of a collective action over time