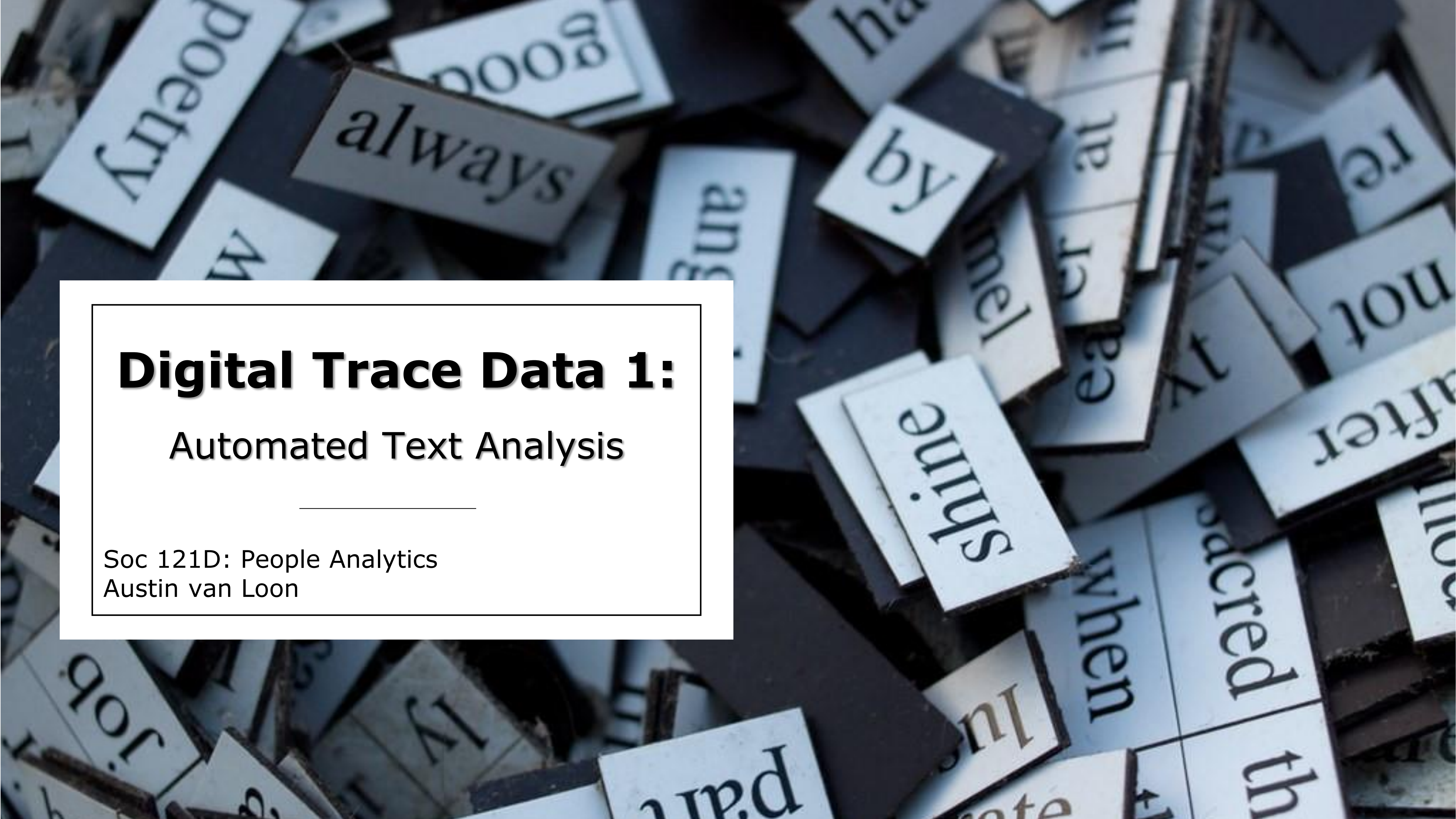


Digital Trace Data 1:

Automated Text Analysis

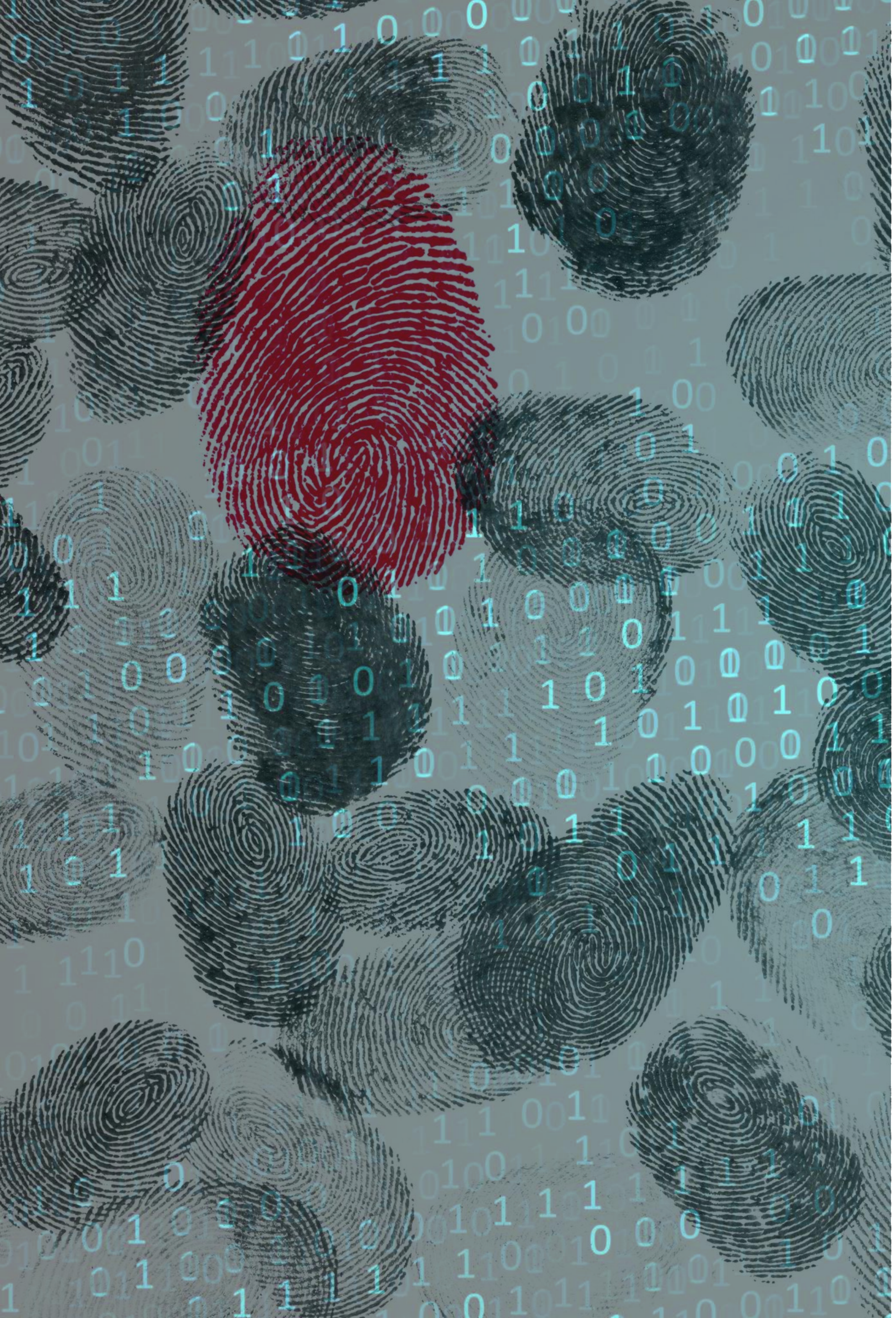
Soc 121D: People Analytics
Austin van Loon

Review Survey—Week 5

- Thanks for the honest responses!
- My dilemma with talking about P-values…
- More “applied” readings in the coming weeks
- Will keep up the group work
- Suggestions for how to make the instructions clearer for the pay equity analysis exercise?

Administrivia

- No word back yet on our room’s thermostat
- Methods module 2 is up
- I’m behind on grading discussion papers (will catch up this week)
- Aiming to publish the example final paper by the end of the day Thursday



Digital Footprints

- In modern society, nearly everything you do digitally is recorded
- This “digital footprint” records a lot of information about you
- Companies use the digital footprints of consumers to target advertisements, customize prices, and predict market trends
- Everyday organizational life produces a similar “digital footprint”. How could we (and should we) use this information to better manage our employees?

This Week

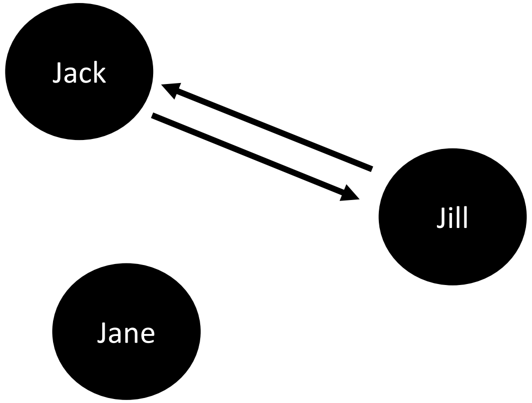
Automated text analysis (Tue)

- Examines the **content** of communication/text
- Measure style, themes, and tone of individual and group communication

Jack: Hey @Jill! How's it going?
Jill: @Jack, it's going terribly. Do you think we'll meet the deadline tomorrow?
Jack: If I'm being honest, I don't think so.
Jill: Jane's not going to happy about that…
Jack: No, she will not ☹
Jill: Check ya later, @Jack!

Network Analysis (Thur)

- Examines the **structure** of communication/behavior
- Measure individuals' position in network, group structure, and organization-level patterns



Text in Organizational Life

- Orders are given to subordinates through email
- Teams work together over Slack
- Future employees are recruited through written job ads
- In the modern economy, many organizations create a “digital exhaust”, which can be harnessed to improve how we manage employees

The Fundamental Issues of Text

- The Curse of Dimensionality
  - At its base, text is an **ordered** collection of **incommensurate** symbols
  - This makes the number of possible realizations of text huge (there are more possible tweets than atoms in the universe)
  - This means we **must** simplify our representation of text
- The complexities of meaning
  - Meaning is an elusive construct (People still argue about the meaning of passages in The Bible)
  - Meaning’s mapping to words isn’t one-to-one
    - The word “tie” could mean one of several things
    - “Austin van Loon is smart” and “the teacher of SOC 121D possesses great intelligence” might mean roughly the same thing, but have exactly zero words in common
  - Meaning is sensitive to small changes in the text (e.g., “love your neighbor” and “don’t love your neighbor” are 75% the same words and 0% the same meaning)

Three “Families” of Automated Text Analysis

- Term Frequency Analysis
    - “Dictionary” analysis
    - Differential language analysis (DLA)
- Latent document structure analysis (topic models)
- Semantic similarity analysis (word embeddings)

Term Frequency Analysis: “Dictionary” analysis

Work words	Social words	Negative words	Positive words
Work	Talk	Bad	Good
Desk	Recreation	Unsatisfactory	Great
Producing	Fun	Appalled	Happy
Report	Hang out	Unacceptable	Merry
Assignment	Weekend	Unpleasant	Jolly

Term Frequency Analysis: “Dictionary” analysis

Work words	Social words	Negative words	Positive words
Work	Talk	Bad	Good
Desk	Recreation	Unsatisfactory	Great
Producing	Fun	Appalled	Happy
Report	Hang out	Unacceptable	Merry
Assignment	Weekend	Unpleasant	Jolly

Dear Dan,

I saw the report you put on my desk Friday, and I was appalled. This was spotty work, and it is unacceptable. If you keep producing unsatisfactory work like this, we might need to sit down and have an unpleasant talk about your future at this company.

Merry Christmas,
Joe

Term Frequency Analysis: “Dictionary” analysis

Work words	Social words	Negative words	Positive words
Work	Talk	Bad	Good
Desk	Recreation	Unsatisfactory	Great
Producing	Fun	Appalled	Happy
Report	Hang out	Unacceptable	Merry
Assignment	Weekend	Unpleasant	Jolly

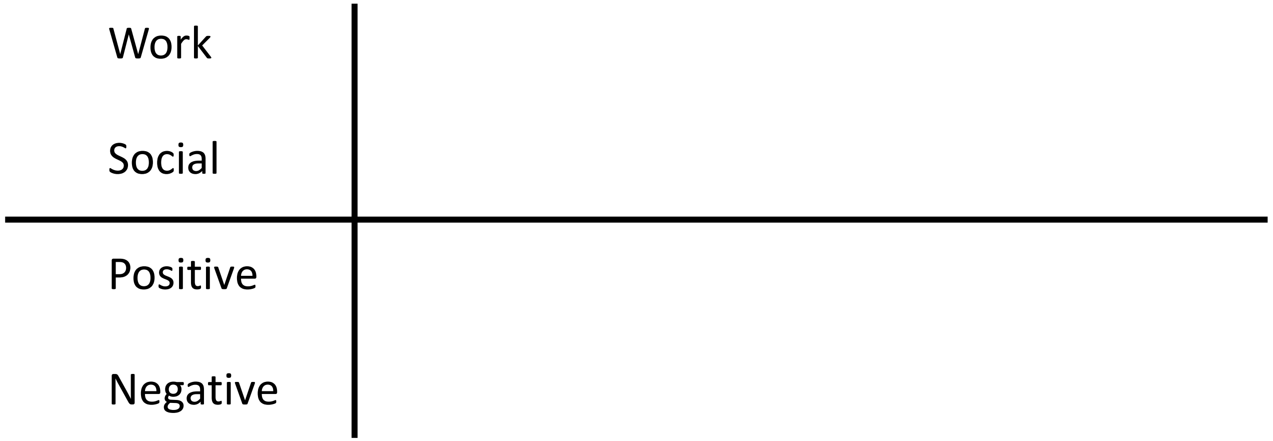
Dear Dan,

I saw the report you put on my desk Friday, and I was appalled. This was spotty work, and it is unacceptable. If you keep producing unsatisfactory work like this, we might need to sit down and have an unpleasant talk about your future at this company.

Merry Christmas,
Joe

Term Frequency Analysis: “Dictionary” analysis

Work words	Social words	Negative words	Positive words
Work	Talk	Bad	Good
Desk	Recreation	Unsatisfactory	Great
Producing	Fun	Appalled	Happy
Report	Hang out	Unacceptable	Merry
Assignment	Weekend	Unpleasant	Jolly



Work

Social

Positive

Negative

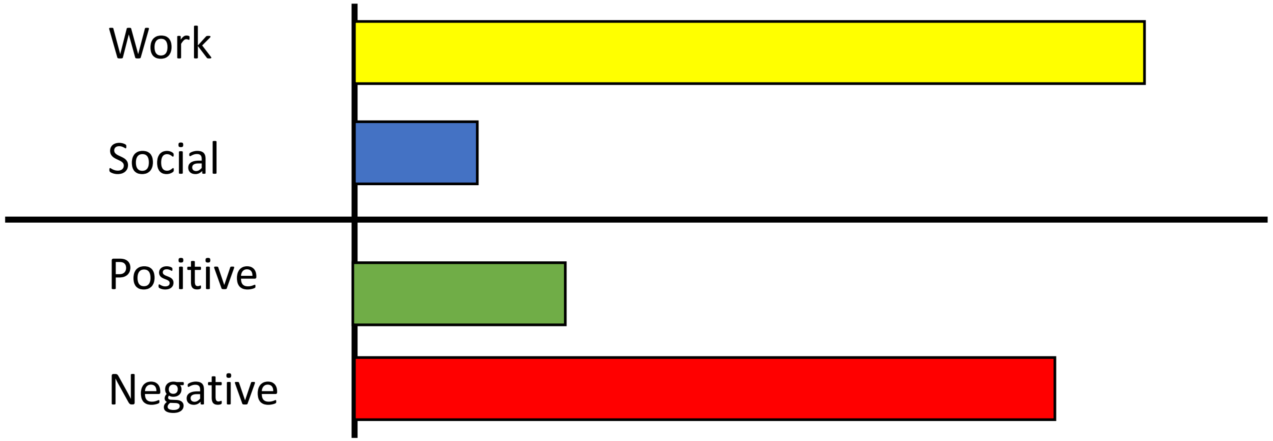
Dear Dan,

I saw the report you put on my desk Friday, and I was appalled. This was spotty work, and it is unacceptable. If you keep producing unsatisfactory work like this, we might need to sit down and have an unpleasant talk about your future at this company.

Merry Christmas,
Joe

Term Frequency Analysis: "Dictionary" analysis

Work words	Social words	Negative words	Positive words
Work	Talk	Bad	Good
Desk	Recreation	Unsatisfactory	Great
Producing	Fun	Appalled	Happy
Report	Hang out	Unacceptable	Merry
Assignment	Weekend	Unpleasant	Jolly



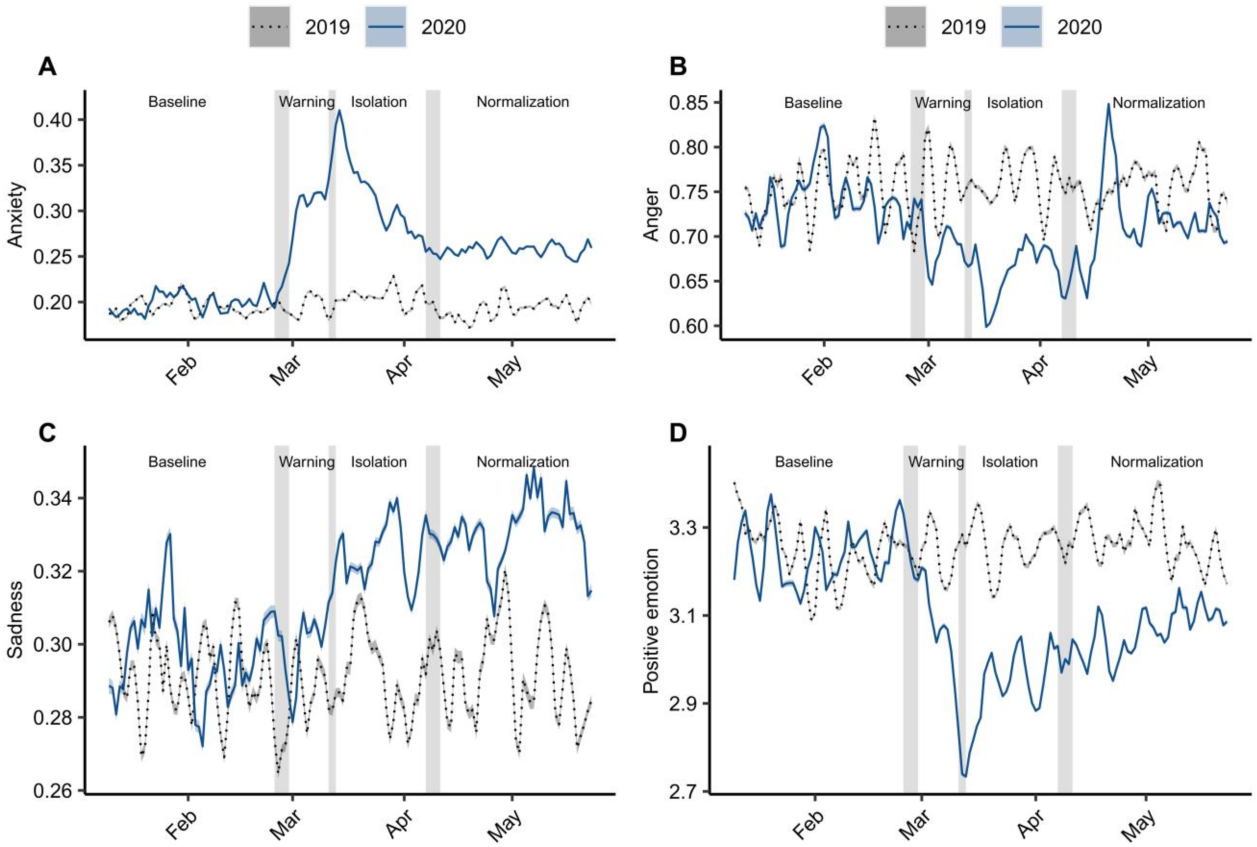
Dear Dan,

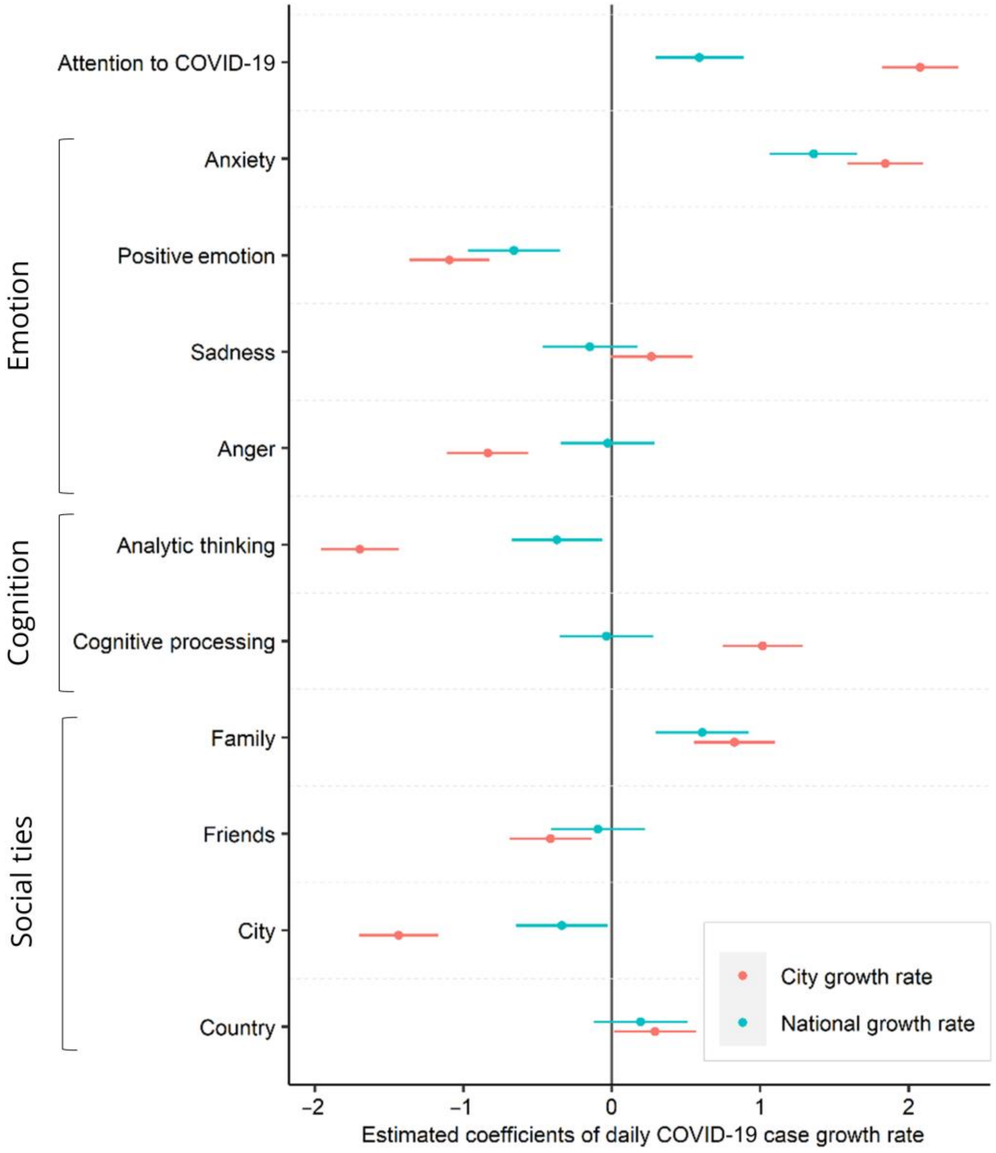
I saw the report you put on my desk Friday, and I was appalled. This was spotty work, and it is unacceptable. If you keep producing unsatisfactory work like this, we might need to sit down and have an unpleasant talk about your future at this company.

Merry Christmas,
Joe

Social media conversations reveal large psychological shifts caused by COVID-19’s onset across U.S. cities

ASHWINI ASHOKKUMAR AND JAMES W. PENNEBAKER





Group Activity (2-3 per group)

- Open the following two webpages:
  - tinyurl.com/PAmusicLIWC
  - liwc.app/demo
- At the first link, enter at least 6 songs (whose lyrics are primarily in English) into the table (song title, artist name, and genre), with each song belonging to one of these three genres: (1) pop, (2) rock/metal, (3) rap/R&B. Do at least two songs of each genre. Try not to repeat songs already in the table!
- Find the lyrics each song you entered and paste them into the box at the second site. Select "Entertainment" in the dropdown box.
- Paste the results of the LIWC analysis into the spreadsheet at the first site

Strengths and Weaknesses

Weaknesses

- Without deep reading, will likely lead one astray
- Doesn’t do much about the complexities of language (e.g., negation)

Strengths

- You can use this on text/corpus of any size
- Relatively easy to implement
- The analysis is very transparent

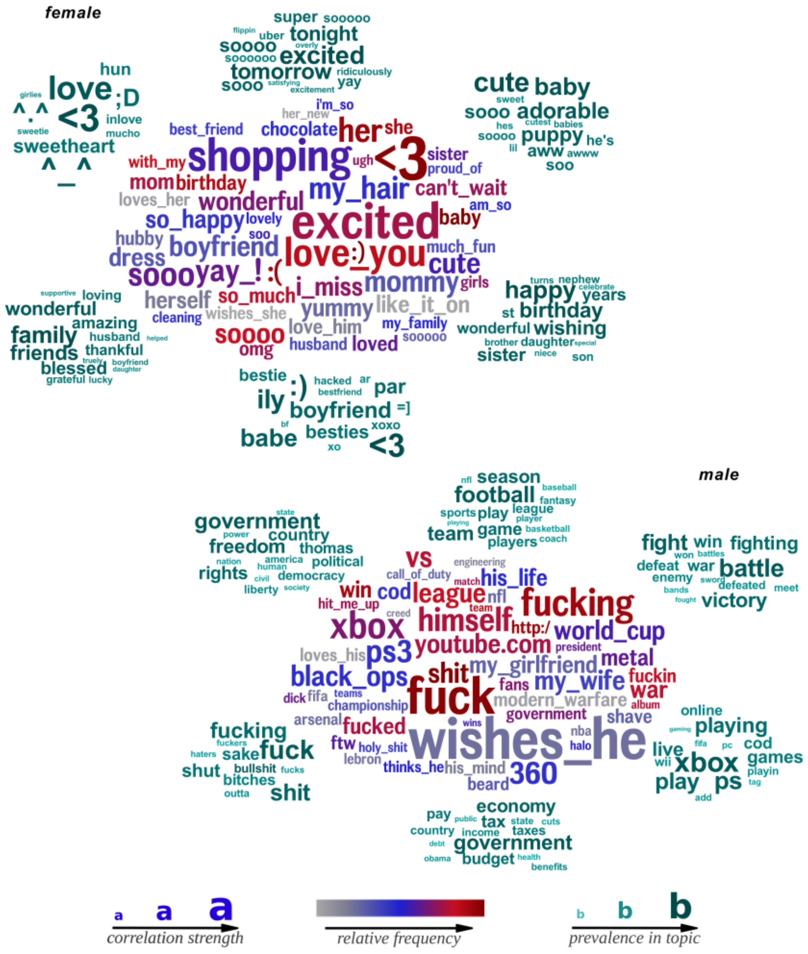
Differential Language Analysis (DLA)

- Say we already know the outcome of interest but want to know the language that maps onto that outcome
- DLA is a method to find differences in term frequency based on meta-data
- Basically, select the *K* most frequent words in the corpus and, for each, test whether it is significant associated with meta-data

Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, Lyle H. Ungar

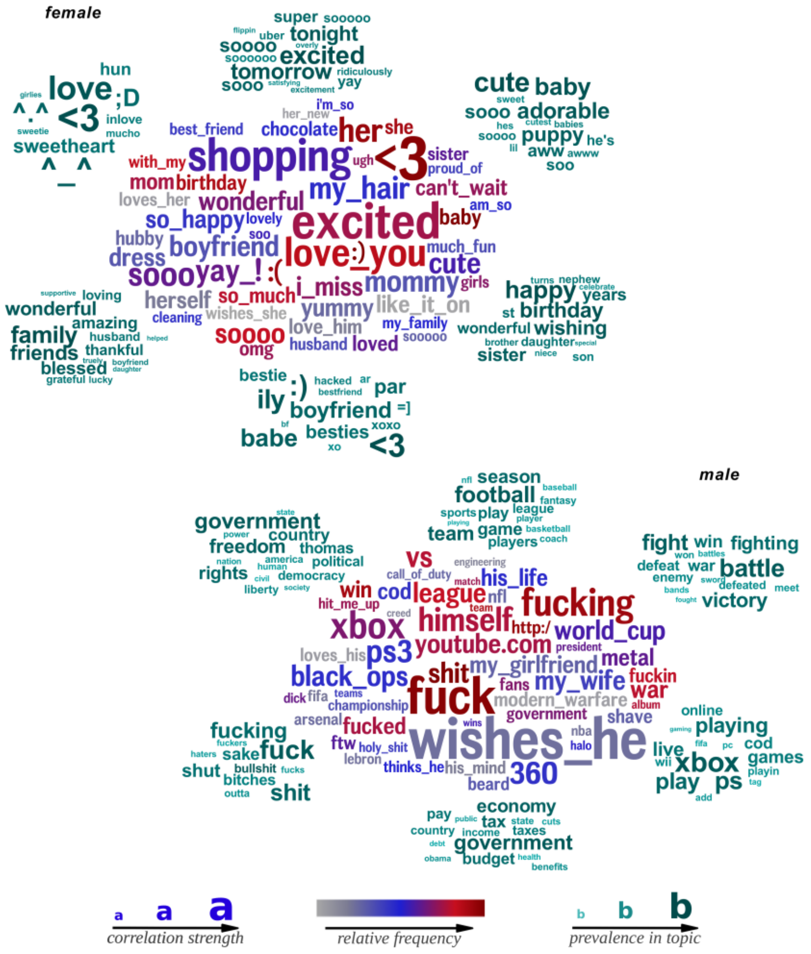
Published: September 25, 2013 • https://doi.org/10.1371/journal.pone.0073791

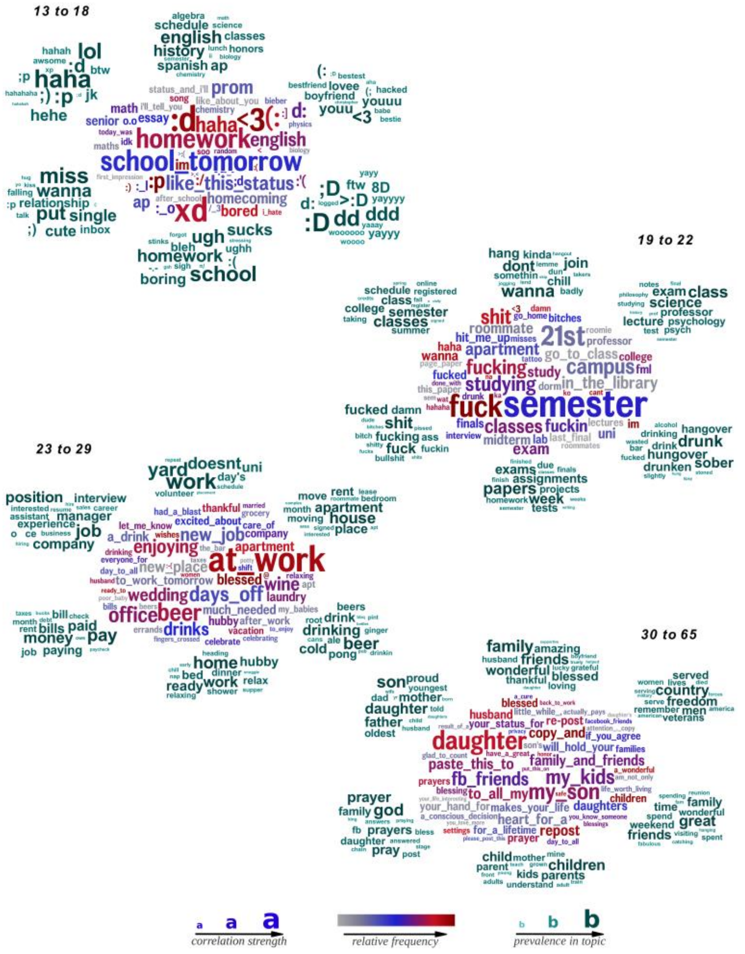


Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, Lyle H. Ungar

Published: September 25, 2013 • https://doi.org/10.1371/journal.pone.0073791

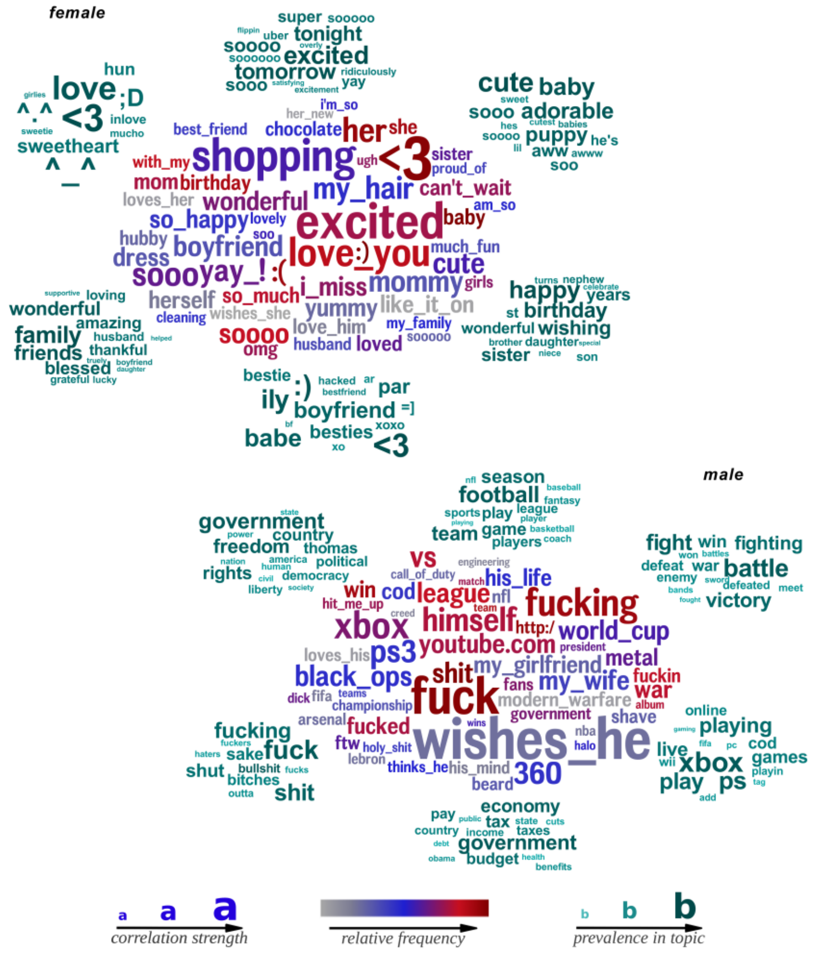


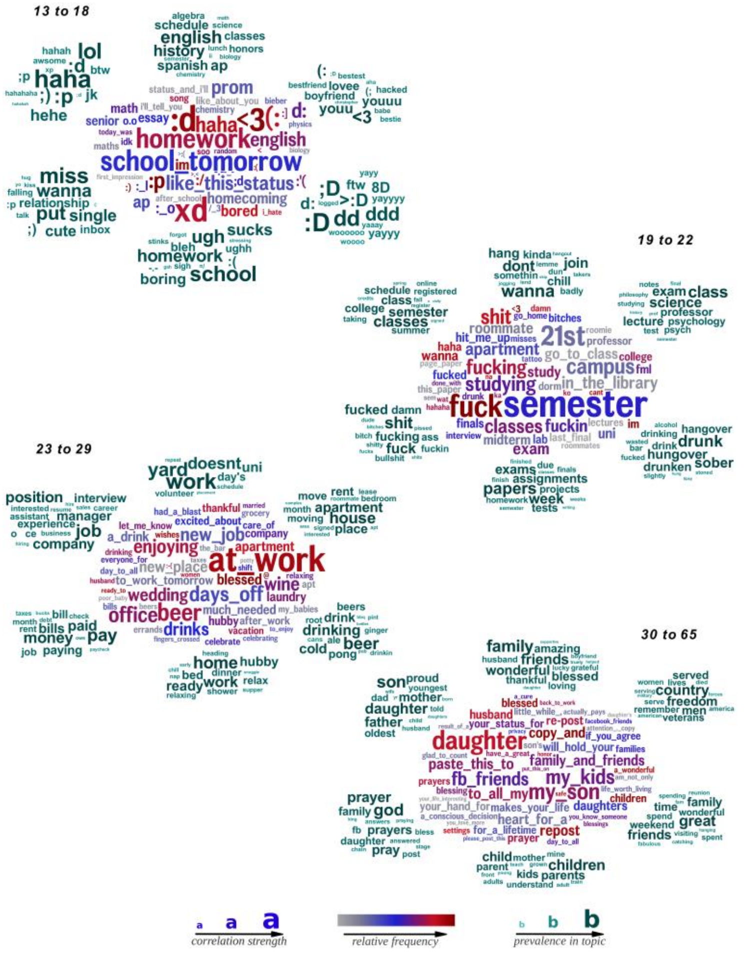


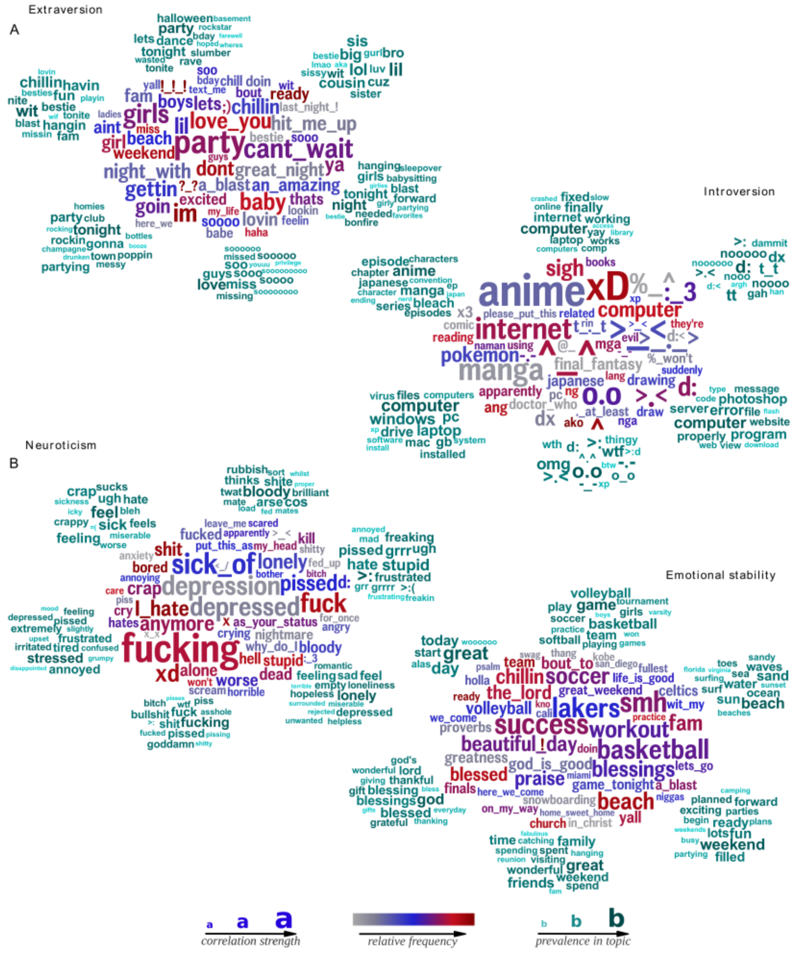
Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, Lyle H. Ungar

Published: September 25, 2013 • https://doi.org/10.1371/journal.pone.0073791







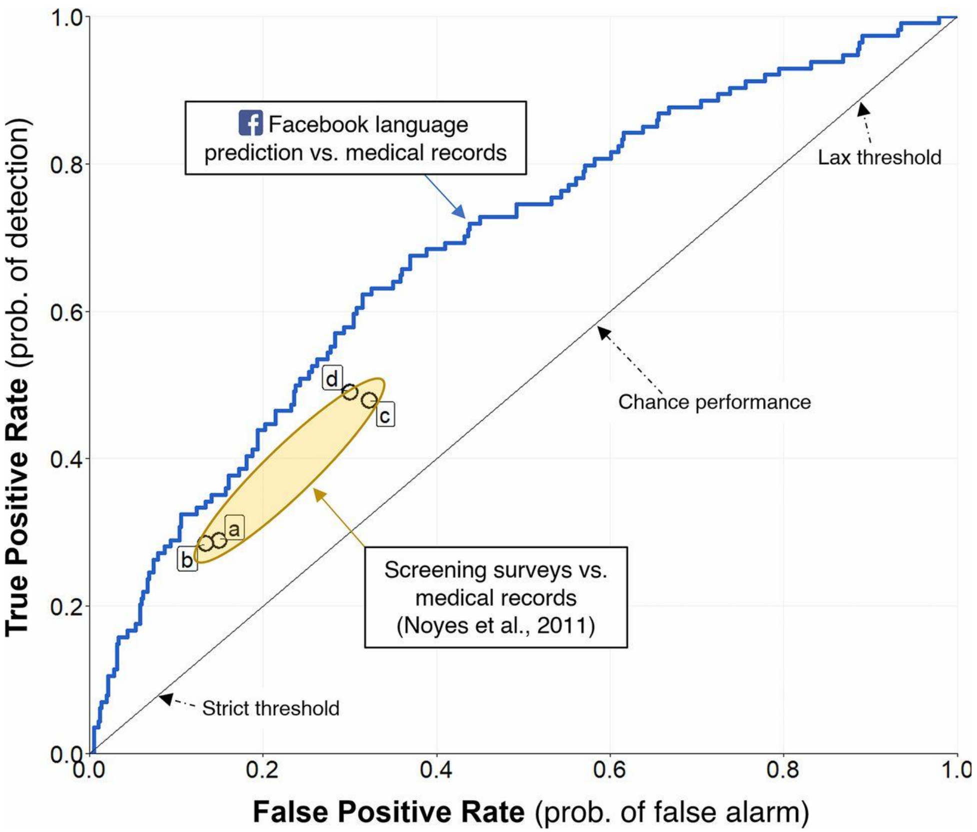
RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES |

Facebook language predicts depression in medical records

Johannes C. Eichstaedt, Robert J. Smith, Raina M. Merchant, +4 , and H. Andrew Schwartz Authors Info & Affiliations

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved September 11, 2018 (received for review February 26, 2018)

October 15, 2018 | 115 (44) 11203-11208 | https://doi.org/10.1073/pnas.1802331115



Strengths and Weaknesses

Weaknesses

- Takes a lot more data than dictionary analysis
- Without deep reading, can lead you astray

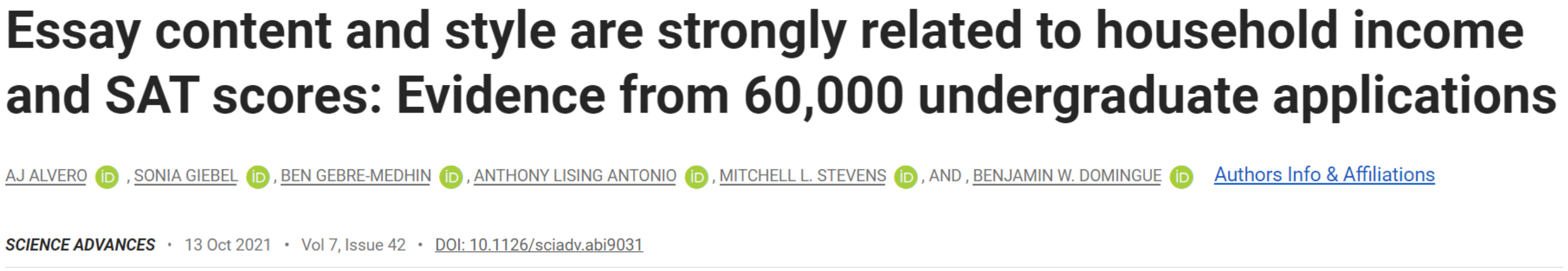
Strengths

- Generally better at prediction
- Meaning of language is learned from corpus

Break!

Latent Document Structure Analysis

- We don’t typically engage with text at the level of the words but at the level of the ideas in the document
- What if we want to get at something deeper than words?
- Analyses in this category (such as topic models) identify *themes* in the corpus (i.e., clusters of words that tend to co-occur)

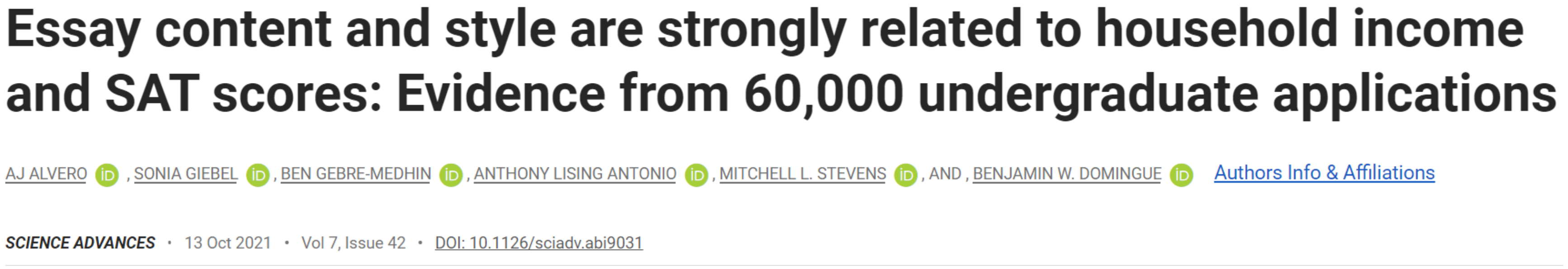


Essay content and style are strongly related to household income and SAT scores: Evidence from 60,000 undergraduate applications

AJ ALVERO , SONIA GIEBEL , BEN GEBRE-MEDHIN , ANTHONY LISING ANTONIO , MITCHELL L. STEVENS , AND , BENJAMIN W. DOMINGUE Authors Info & Affiliations

SCIENCE ADVANCES • 13 Oct 2021 • Vol 7, Issue 42 • DOI: 10.1126/sciadv.abi9031

Merged essay topic	Highest probability words	Frequent exclusive words	Excerpt from essay with highest topic score
Time management	time, work, help, get, school, abl, go	homework, manag, get, stress, done, stay, procrastin	"I do try hard to make sure that I complete my assignments by a certain time, but sometimes I have to stay up later than I expected to make sure I finish everything. This has affected my achievement by making me have to focus on one class than the other. This has proven to be a big challenge, but I plan to overcome it"
Helping others	peopl, help, can, make, way, differ, other	peopl, can, other, someon, everyon, differ, way	"When I am helping the students, I have to take charge, show them how each step is done to help them complete whatever it is that they are doing. If I see one of them is having trouble, then it is my duty as a leader to show them how to do it so they will understand how to do it the next time. Being a teacher's assistant is a hard job but it gives me responsibility skills that I will need in the future"
Tutoring groups	help, tutor, colleg, avid, also, go, need	avid, tutor, ffa, et, ag, via, tutori	"It has also taught me to seek help from tutors something that trained me to improve my homework and test taking abilities. Taking advantage of these educational opportunities made me feel empowered and grateful. Through these programs, I had the opportunity to learn valuable skills and tools to ease my transition and help me be successful in a four year college environment"
Preference words	also, like, thing, realli, subject, lot, alway	realli, lot, thing, good, favorit, influenc, enjoy	"My greatest talent or skill is acting. I absolutely love acting, and it is one of my greatest talents. Just recently I took an acting class and it is one of the best decisions I have ever made…My favorite monologue that I performed in my class was Charlie and the Chocolate Factory, it really matched me"
Education opportunity	colleg, educ, opportun, take, advantag, attend, school	advantag, educ, colleg, opportun, credit, graduat, prep	"I also took real college classes with other college students which are transferable if accepted by other colleges or universities. While it has been a major educational opportunity, it has also been a educational barrier which I have had to overcome. The more advanced high school education and full college courses have required me to put a large amount of my time and effort into my education"



Essay content and style are strongly related to household income and SAT scores: Evidence from 60,000 undergraduate applications

AJ ALVERO , SONIA GIEBEL , BEN GEBRE-MEDHIN , ANTHONY LISING ANTONIO , MITCHELL L. STEVENS , AND , BENJAMIN W. DOMINGUE Authors Info & Affiliations

SCIENCE ADVANCES • 13 Oct 2021 • Vol 7, Issue 42 • DOI: 10.1126/sciadv.abi9031

Merged essay topic	Highest probability words	Frequent exclusive words	Excerpt from essay with highest topic score
Seeking answers	question, book, like, research, read, answer, ask	telescop, astronom, map, probe, column, constel, encyclopedia	"Ever since the big bang took place, particles have been hovering around the universe for billions of years. Some of them now constituted me, but who knows where they were billions of years ago? Why couldn't they have come from Mars?"
Human nature	world, human, natur, passion, beyond, complex, explor	inher, manifest, notion, philosophi, nuanc, facet, myriad	"From a young age, I have found a fascination in the art of rhetoric and its influence on humanity…I believe as cognitively complex individuals we should maximize our ability as a collective species to understand the very nature of our surrounding"
China	chines, studi, student, also, time, china, school	china, provinc, hong, kong, chines, shanghai, wechat	"I served as the Chunhui emissary and participated in the voluntary activities in the 'Chunhui Action' in Qixingguan District in Bijie City in Guizhou province. Our team went to the poverty-stricken area in Qixingguan District and helped build"
Achievement words	result, provid, initi, began, becam, academ, effort	dilig, remain, util, attain, endeavor, initi, simultan	"Rather than taking the fundamental classes to proceed through high school, I chose to additionally push myself out of comfortability and undertake the strenuous task of taking Advanced Placement classes. Prior to entering my senior year, I have successfully passed a total of two honors and five Advanced Placement courses, all while managing both extracurricular activities and favorable pastimes"
Despite words	howev, one, may, rather, even, simpli, fact	simpli, rather, may, fact, truli, consid, howev	"To this day, I cannot begin such an ambitious project, though perhaps that is simply because it is so enterprising. Perhaps if the attempt was made to write something shorter and more reasonable I could have succeeded, could have written something to be remembered. But I never did, though I still have the chance. Maybe I will. Maybe today will be the day I decide to write"

D	“science”	“math”	“work”	…	“part-time”
0	2	3	1	…	0
1	2	1	8	…	5
2	5	6	0	…	0
…	…	…	…	…	…
N	0	0	12	…	15

D	T1	T2	…	T*K*
0	0.85	0.05	…	0.01
1	0.3	0.63	…	0.03
2	0.95	0.00	…	0.04
…	…	…	…	…
N	0.05	0.95	…	0.01

Strengths and Weaknesses

Weaknesses

- Takes more data than term frequency analysis
- Unclear how to know whether you have “enough” text or if the model has fit the data “well” or not
- Less transparent than term frequency analyis

Strengths

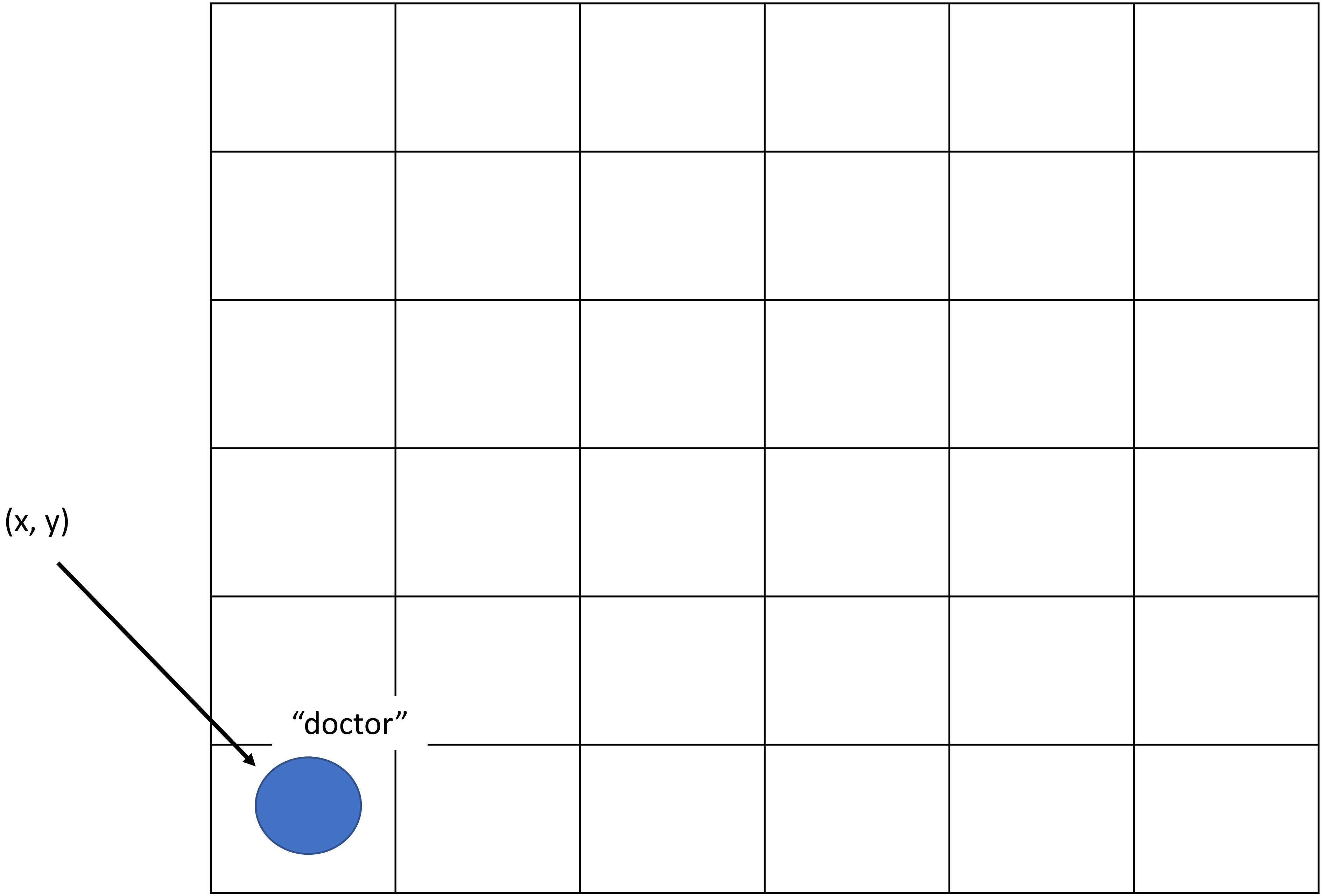
- Word meanings are tuned to the specific corpus
- Considers context to help deal with complexities of language (e.g., “March” with “soldier” can be treated differently than “March” with “February”)

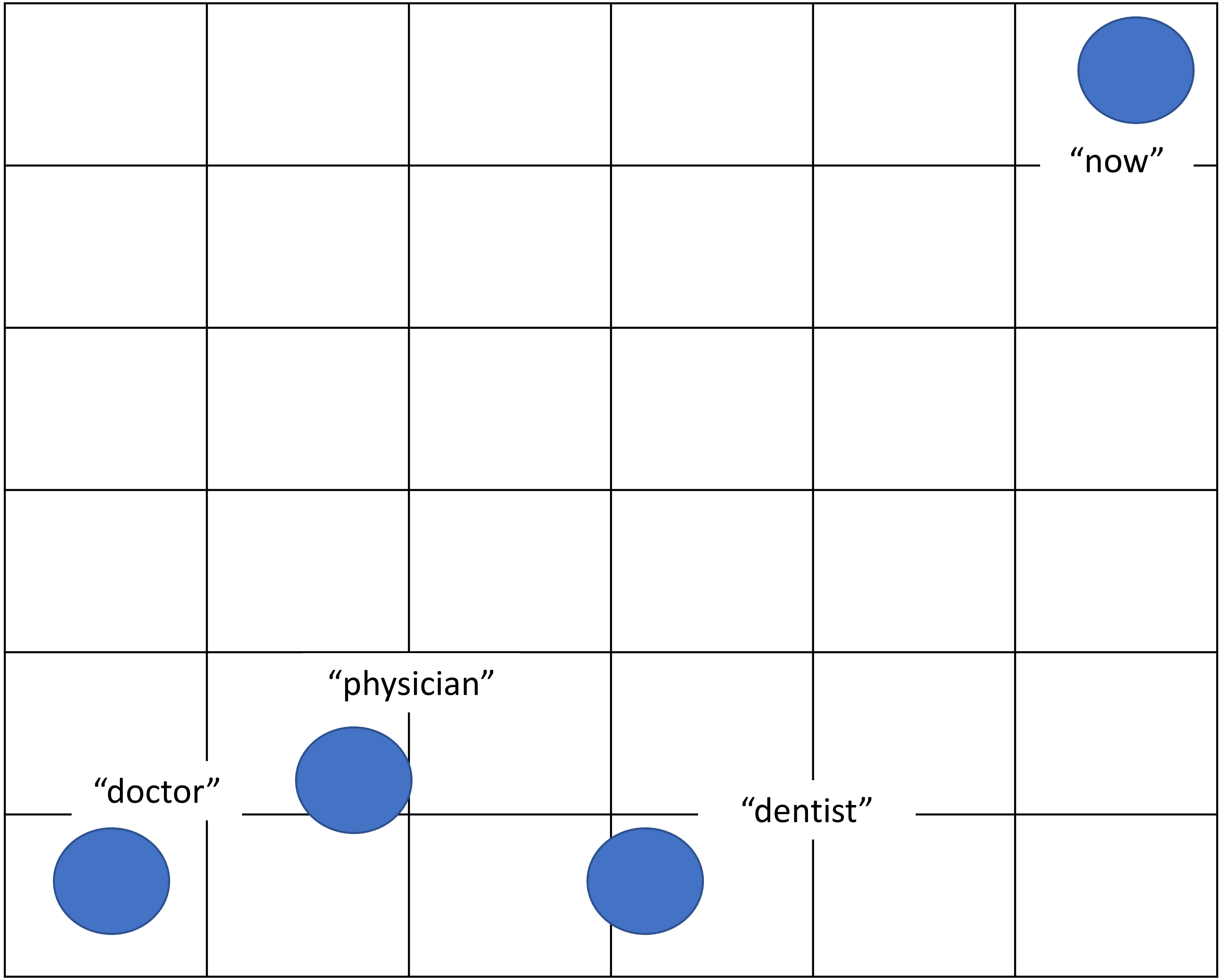
Semantic Similarity Analysis (*AKA* word embeddings)

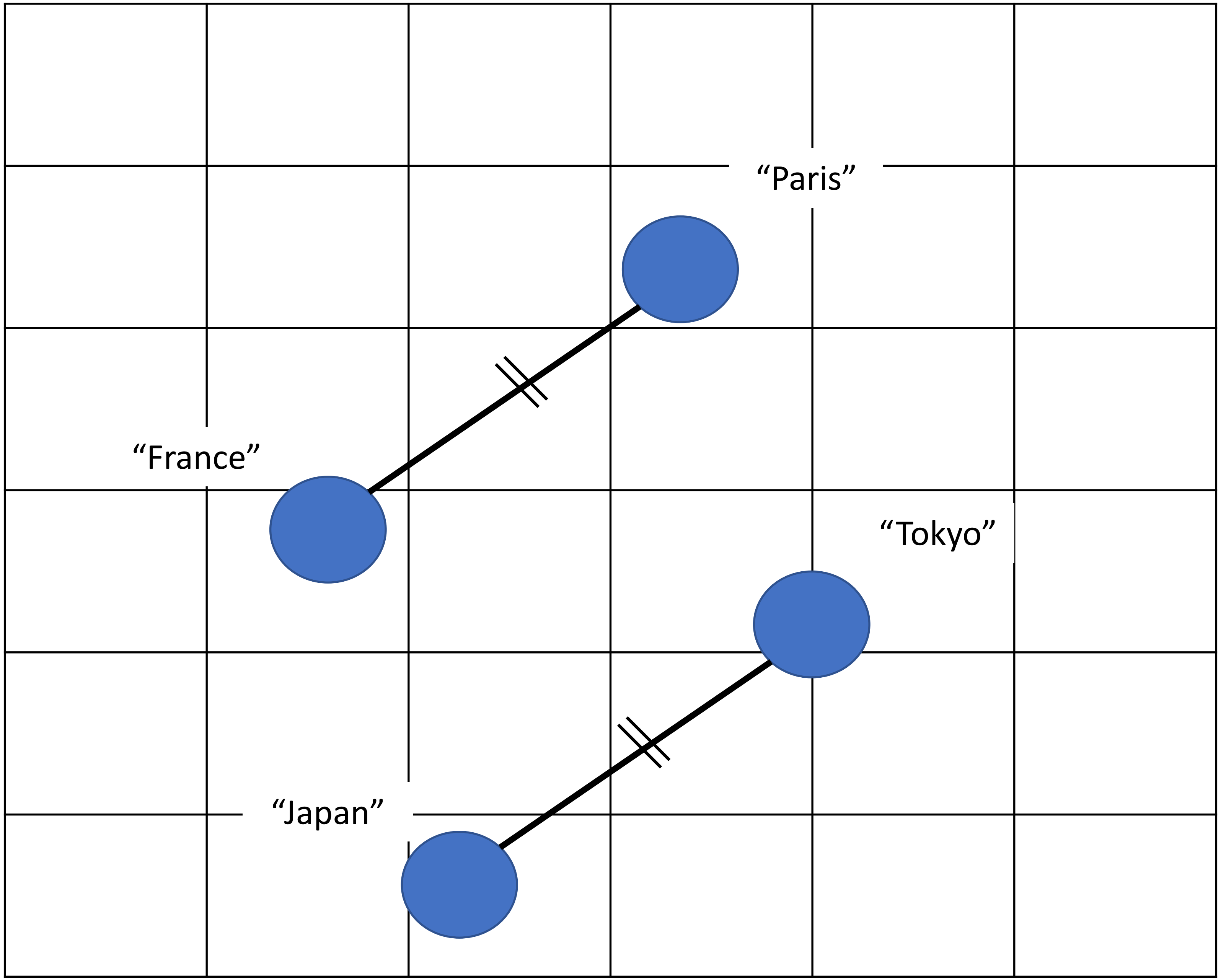
- Are the “hottest” thing in automated text analysis right now
- Conceptually, we estimate a space where words that are “similar in meaning” to each other are relatively close together (this is often done with deep learning)
- How do we define what “similar in meaning” is?

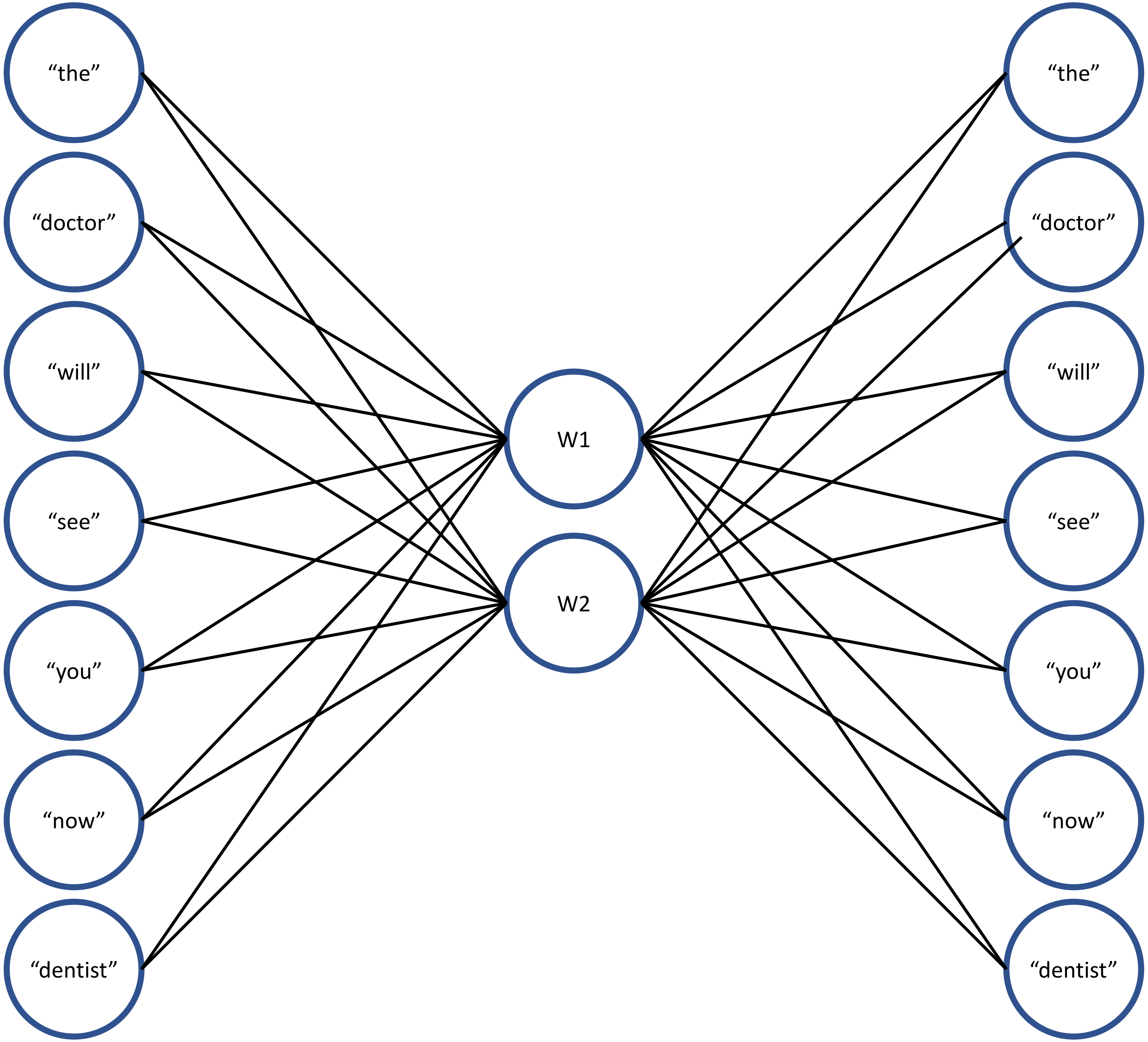
Word Embeddings

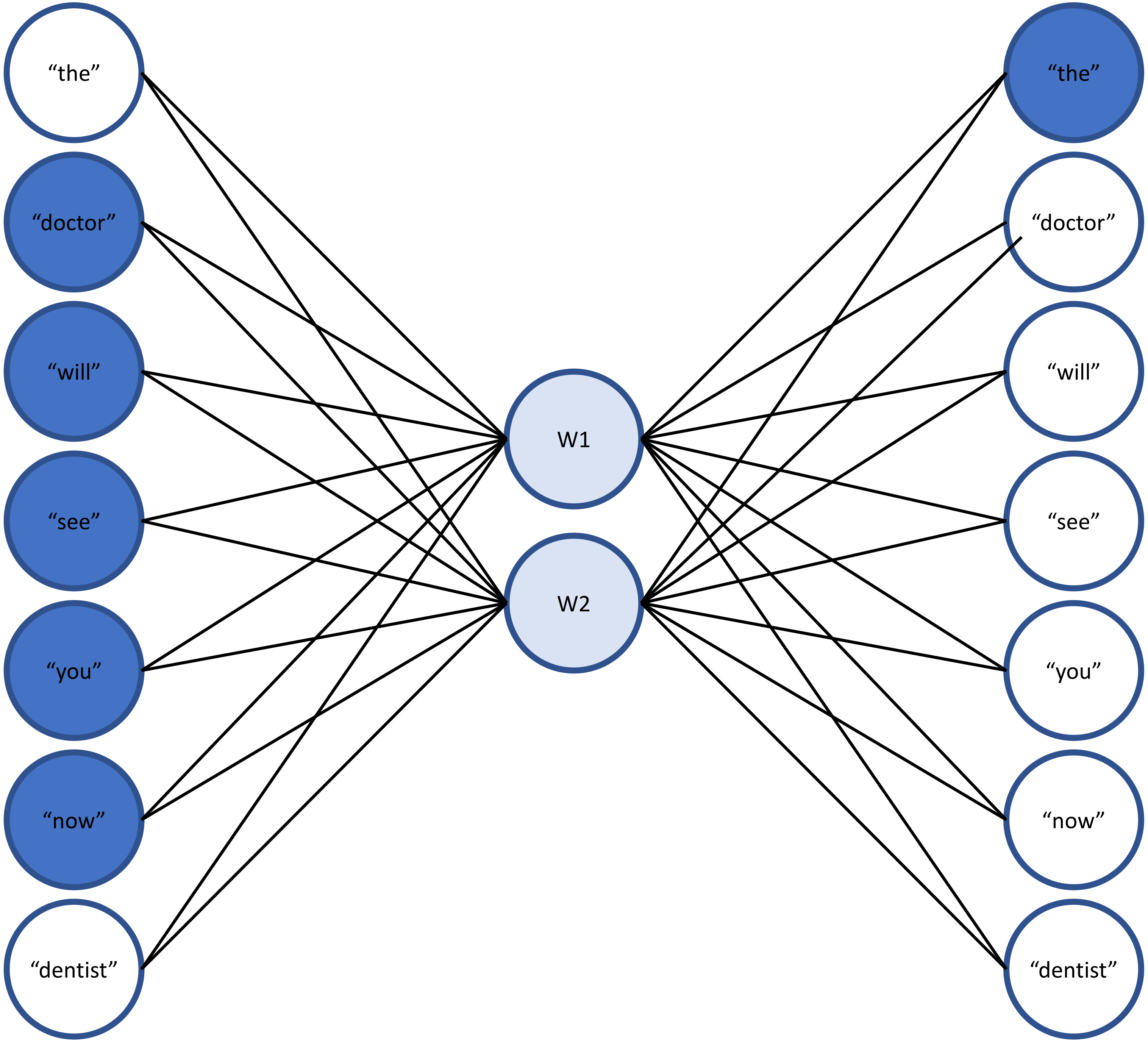
(From Tuesday)

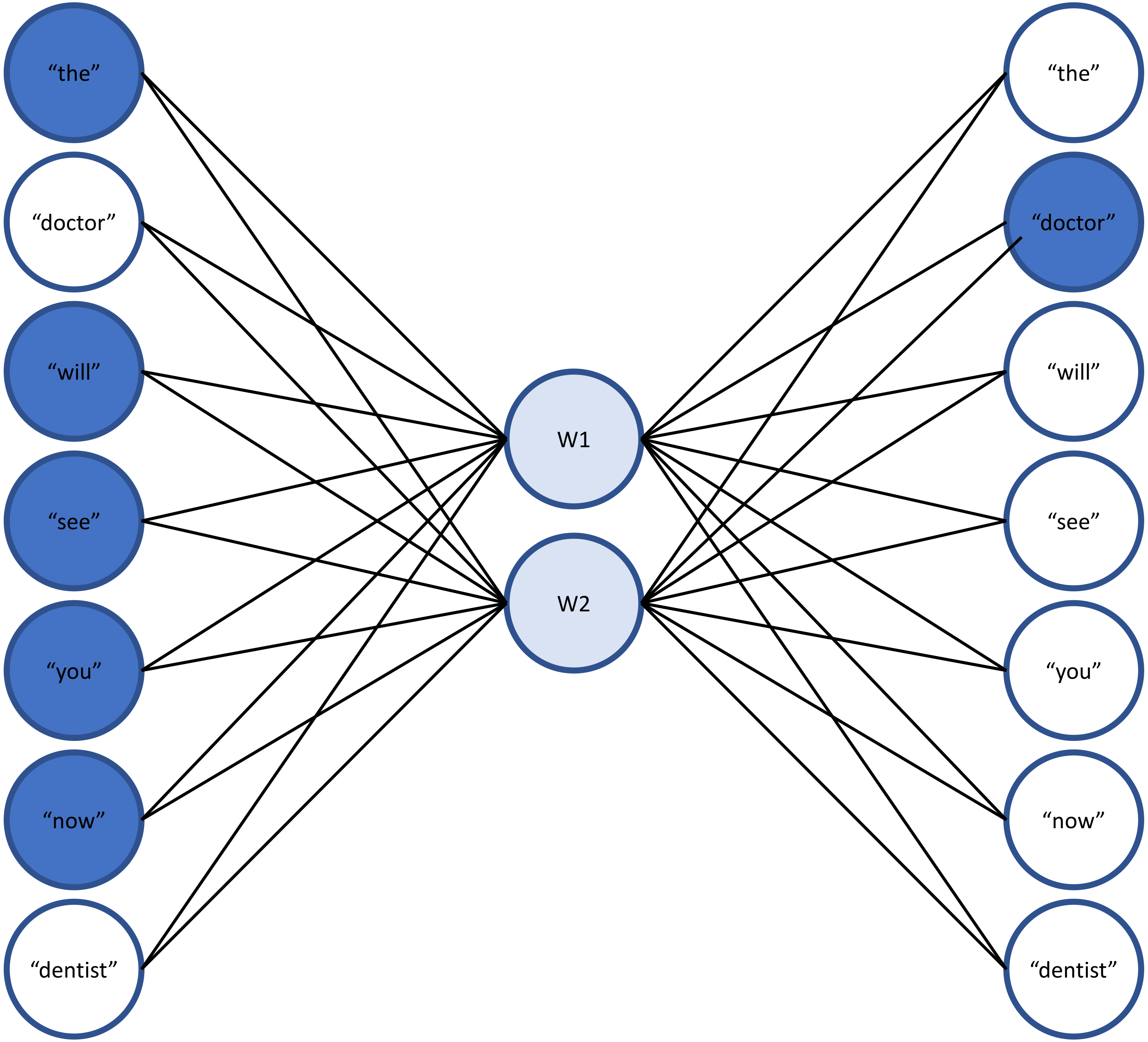

(x, y)
“doctor”

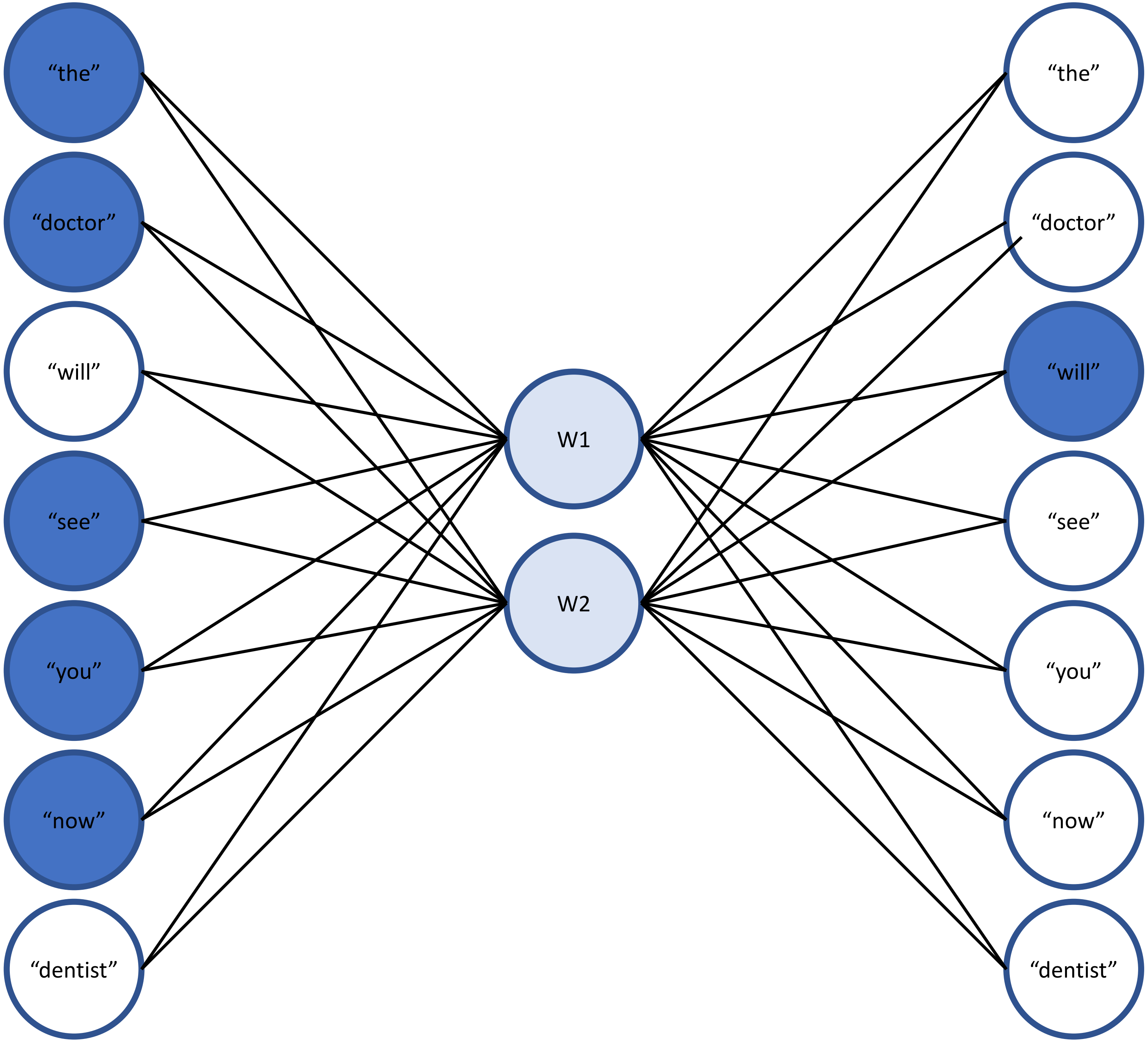

“now”
“physician”
“doctor”
“dentist”

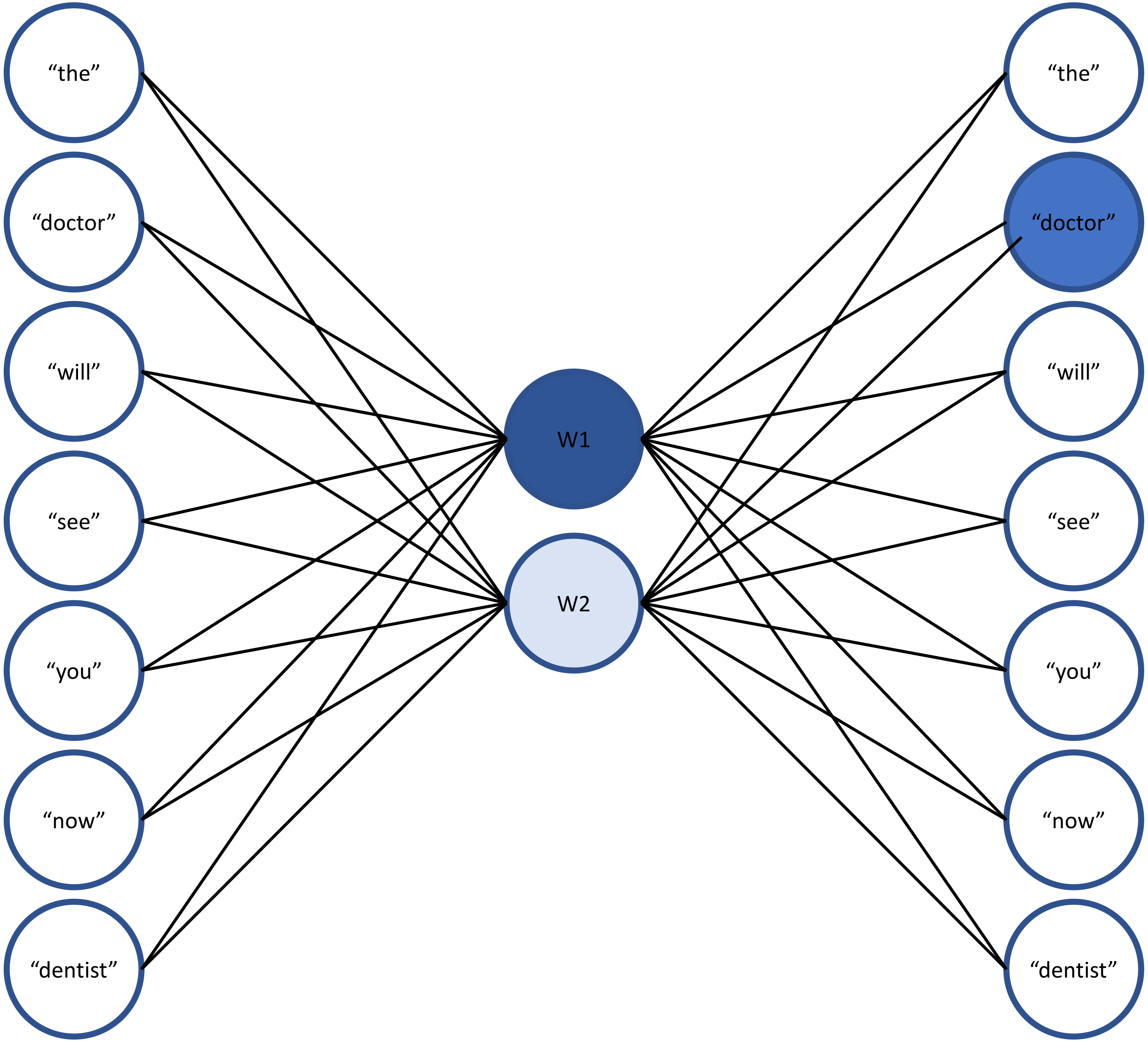

“Paris”
“France”
“Tokyo”
“Japan”

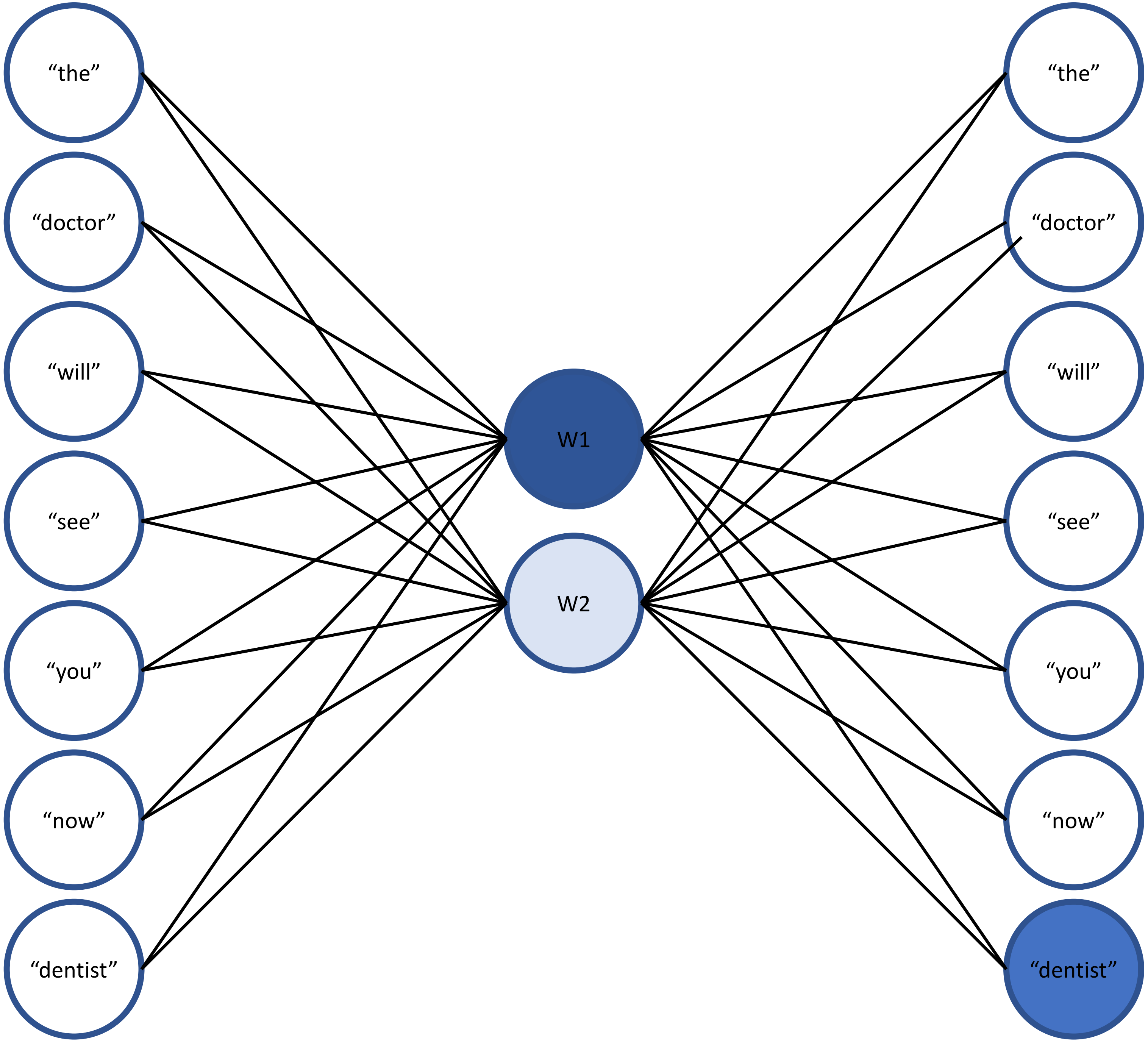

“the”
“doctor”
“will”
“see”
“you”
“now”
“dentist”
W1
W2
“the”
“doctor”
“will”
“see”
“you”
“now”
“dentist”

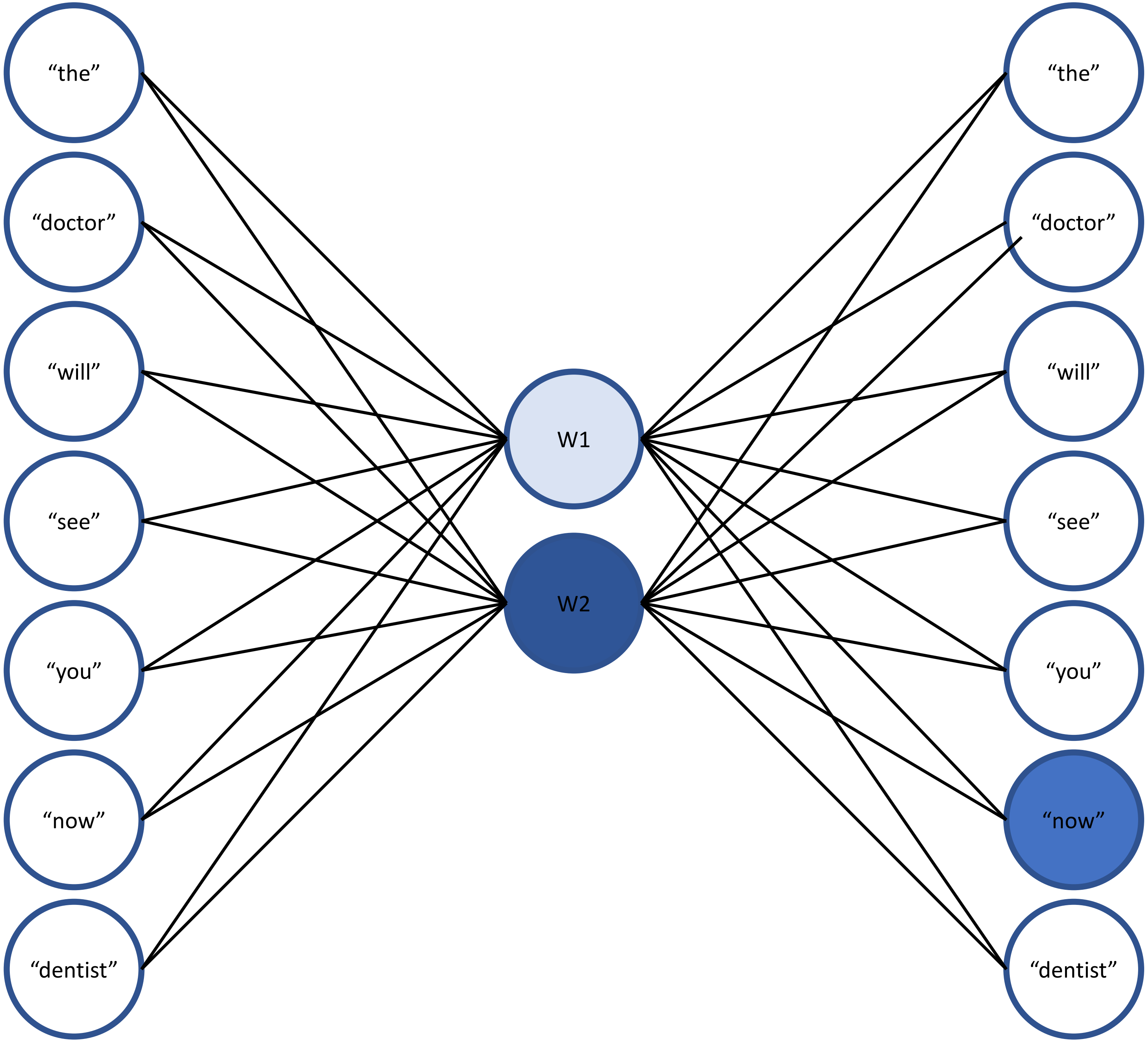

“the”
“doctor”
“will”
“see”
“you”
“now”
“dentist”
W1
W2
“the”
“doctor”
“will”
“see”
“you”
“now”
“dentist”


"the"
"doctor"
"will"
"see"
"you"
"now"
"dentist"
W1
W2
"the"
"doctor"
"will"
"see"
"you"
"now"
"dentist"


“the”
“doctor”
“will”
“see”
“you”
“now”
“dentist”
W1
W2
“the”
“doctor”
“will”
“see”
“you”
“now”
“dentist”


“the”
“doctor”
“will”
“see”
“you”
“now”
“dentist”
W1
W2
“the”
“doctor”
“will”
“see”
“you”
“now”
“dentist”

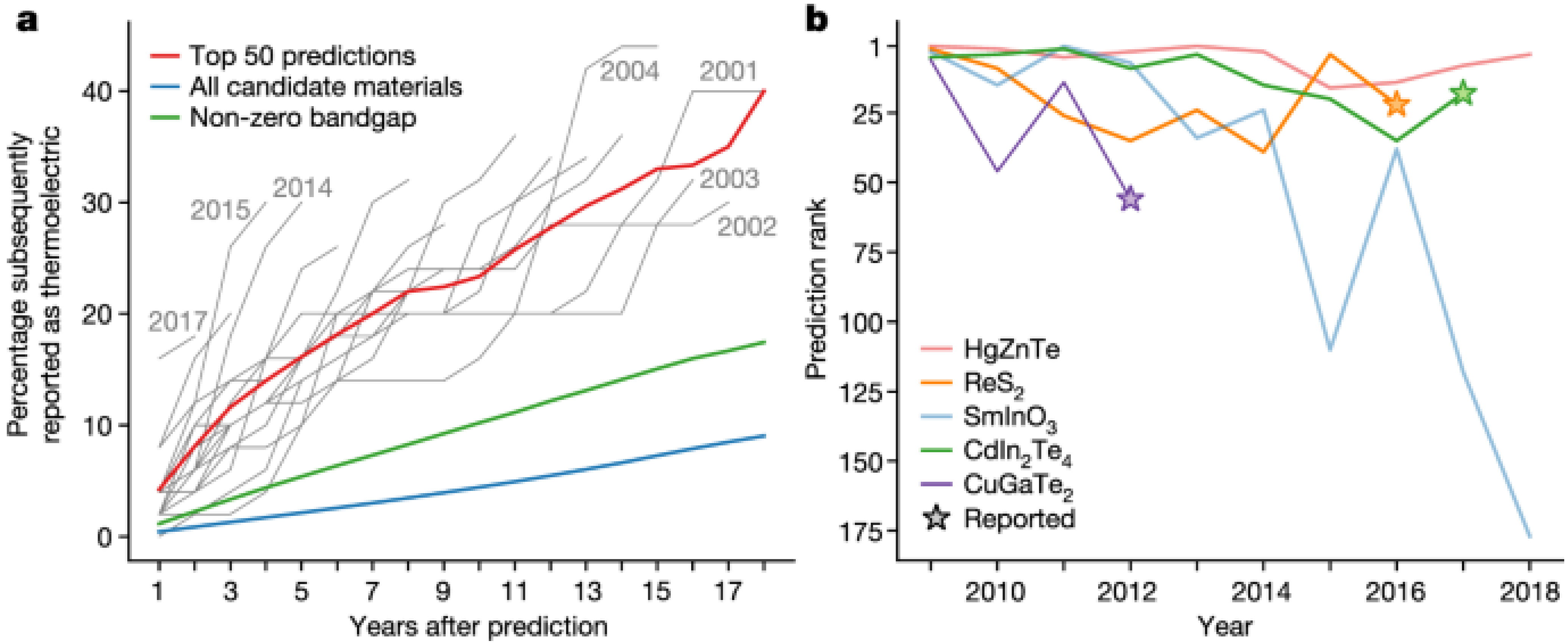

"the"
"doctor"
"will"
"see"
"you"
"now"
"dentist"
W1
W2
"the"
"doctor"
"will"
"see"
"you"
"now"
"dentist"

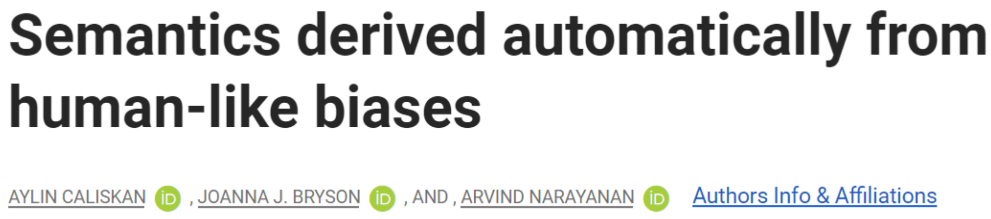

"the"
"doctor"
"will"
"see"
"you"
"now"
"dentist"
W1
W2
"the"
"doctor"
"will"
"see"
"you"
"now"
"dentist"

Letter | Published: 03 July 2019

Unsupervised word embeddings capture latent knowledge from materials science literature

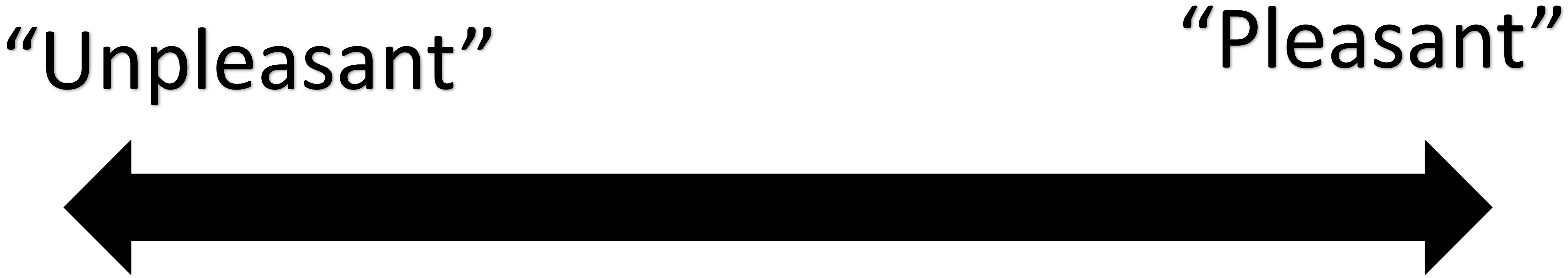
Vahe Tshitoyan ✉, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova,
Kristin A. Persson, Gerbrand Ceder ✉ & Anubhav Jain ✉





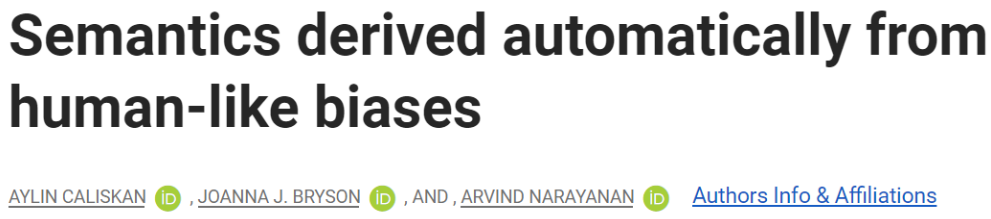
Semantics derived automatically from language corpora contain human-like biases

AYLIN CALISKAN , JOANNA J. BRYSON , AND , ARVIND NARAYANAN Authors Info & Affiliations



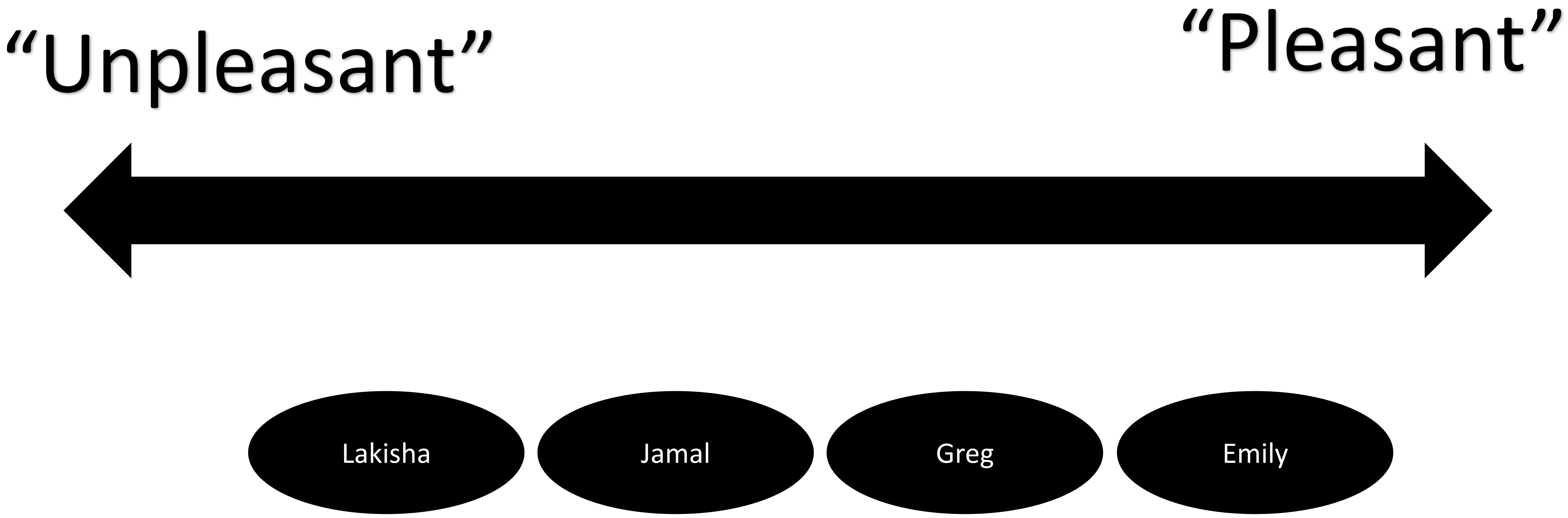
“Unpleasant”

“Pleasant”



Semantics derived automatically from language corpora contain human-like biases

AYLIN CALISKAN , JOANNA J. BRYSON , AND , ARVIND NARAYANAN Authors Info & Affiliations



“Unpleasant”

“Pleasant”

Lakisha

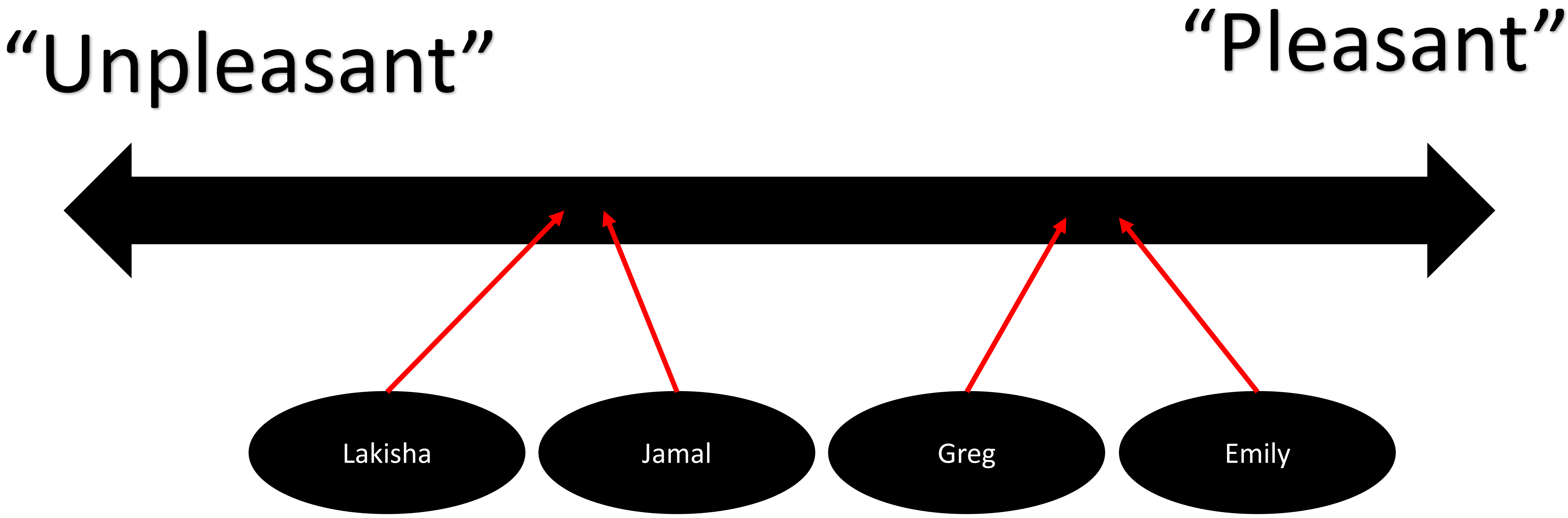
Jamal

Greg

Emily

Semantics derived automatically from language corpora contain human-like biases

AYLIN CALISKAN, JOANNA J. BRYSON, AND, ARVIND NARAYANAN Authors Info & Affiliations



“Unpleasant”

“Pleasant”

Lakisha

Jamal

Greg

Emily

Strengths and Weaknesses

Weaknesses

- Takes a lot more data than latent document structure analysis (to train from scratch)
- Complete and utter “black box”
- Captures both objective information and socio-cultural biases

Strengths

- Really starts to get at the “meaning” of words
- As deep learning progresses, so do the quality of the models
- Best linguistic tool for prediction we have now
- There are high-quality, pre-trained models you can use (that are free and online!)

(Some) Organizational Applications

- By analyzing term frequencies in employee’s emails, can we predict how engaged employees are on average?
- By analyzing themes in application cover letters and resumes, can we identify promising candidates?
- By examining word embedding representations of our job ads, can we get an objective take on what job seekers might infer about the job?
- Can you think of others?

Please fill out the following “network survey”:

tinyurl.com/PAnetworkSurvey

(Please do so as honestly as you can)

Once you’re done you’re free to go. Have a great rest of your day!