



SI 630

Natural Language Processing: Algorithms and People

Lecture 12: Language, Society, and Ethics
April 1, 2019

Pop Quiz Midterm!

Pop ~~Quiz~~ Midterm!

April Fools.

Administrative Stuff

Administrative Stuff

- Midterm is Friday!

Administrative Stuff

- Midterm is Friday!
- HW5 is released

Administrative Stuff

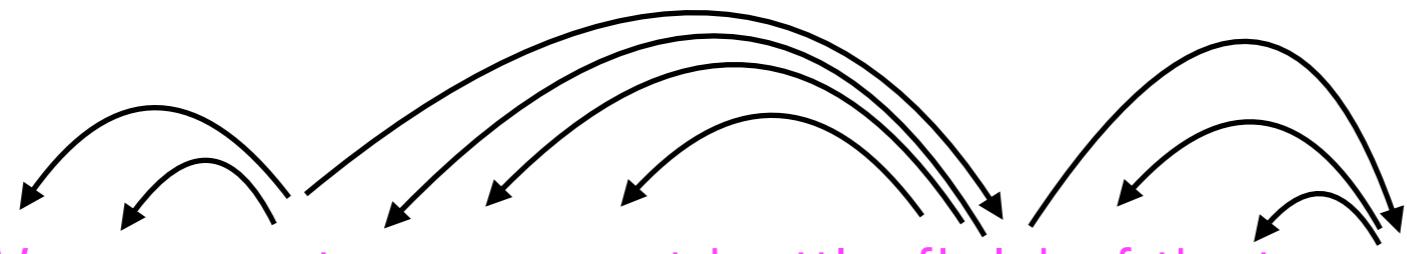
- Midterm is Friday!
- HW5 is released
- Poster deadline was updated to correct date

Administrative Stuff

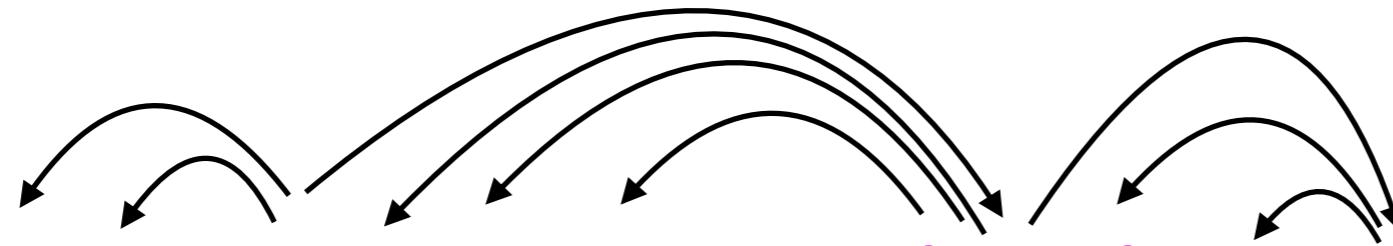
- Midterm is Friday!
- HW5 is released
- Poster deadline was updated to correct date
- Please submit before April 27 (the GSIs are under stress too)
 - We cannot feasibly grade two assignments, final reports, and blog posts in ~2 days if everyone delays.

great

We are met on a great battle-field of that war.



We are met on a great battle-field of that war.



We are met on a great battle-field of that war.

$\lambda x. \lambda y. \text{meet}(x, y)$

$\lambda x. \text{battlefield}(x)$

$\lambda x. \text{war}(x)$

We are met on a great battle-field of that war.

Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. **We are met on a great battle-field of that war.** We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we can not dedicate -- we can not consecrate -- we can not hallow -- this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us -- that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion -- that we here highly resolve that these dead shall not have died in vain -- that this nation, under God, shall have a new birth of freedom -- and that government of the people, by the people, for the people, shall not perish from the earth.

Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. **We are met on a great battle-field of that war.** We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we can not dedicate -- we can not consecrate -- we can not hallow -- this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us -- that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion -- that we here highly resolve that these dead shall not have died in vain -- that this nation, under God, shall have a new birth of freedom -- and that government of the people, by the people, for the people, shall not perish from the earth.



Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. **We are met on a great battle-field of that war.** We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we can not dedicate -- we can not consecrate -- we can not hallow -- this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us -- that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion -- that we here highly resolve that these dead shall not have died in vain -- that this nation, under God, shall have a new birth of freedom -- and that government of the people, by the people, for the people, shall not perish from the earth.

Today's Agenda

Today's Agenda

- Doing social science with NLP

Today's Agenda

- Doing social science with NLP
- Making an B.A. in English degree useful

Today's Agenda

- Doing social science with NLP
- Making an B.A. in English degree useful
- Doing actual Linguistics with NLP!

Today's Agenda

- Doing social science with NLP
- Making an B.A. in English degree useful
- Doing actual Linguistics with NLP!
- Making the world a better place



K2



K2



“just the bare bones of a name, all rock and ice and storm and abyss. It makes no attempt to sound human. It is atoms and stars. It has the nakedness of the world before the first man—or of the cindered planet after the last” —Fosco Maraini

Social NLP

Social NLP

Social NLP

- Social NLP covers a range of applications that analyze how language interacts with people in social settings.

Social NLP

- Social NLP covers a range of applications that analyze how language interacts with people in social settings.
- We leave behavioral **traces** in our interactions with others.

Social NLP

- Social NLP covers a range of applications that analyze how language interacts with people in social settings.
- We leave behavioral **traces** in our interactions with others.
 - Tweets

Social NLP

- Social NLP covers a range of applications that analyze how language interacts with people in social settings.
- We leave behavioral **traces** in our interactions with others.
 - Tweets
 - Books

Social NLP

- Social NLP covers a range of applications that analyze how language interacts with people in social settings.
- We leave behavioral **traces** in our interactions with others.
 - Tweets
 - Books
 - Emails

Social NLP

- Social NLP covers a range of applications that analyze how language interacts with people in social settings.
- We leave behavioral **traces** in our interactions with others.
 - Tweets
 - Books
 - Emails
 - Audio transcripts

“Raw” data

- Social NLP often makes **claims** about the world using textual data.
- Data is not self-evident, neutral or objective
- Data is collected, stored, processed, mined, interpreted; each stage requires our **participation**.
- What is the **process** by which the data you have got to you?

Data Collection

- Data → Research Question
 - “Opportunistic data”
 - Research questions are shaped by what data you can find
- Research Question → Data
 - Research is driven by questions, find data to support answering it.

Social NLP

- What are the research questions that we can ask when applying NLP to text to answer **social** and **cultural** questions?
- How do we answer those research questions using methods we've learned about?
 - data
 - algorithms
 - evaluation

Social NLP

- Manifestations of power in text
- Measuring respect
- Explaining trolling behavior
- Determining who receives help

Power

How is power manifested in language?

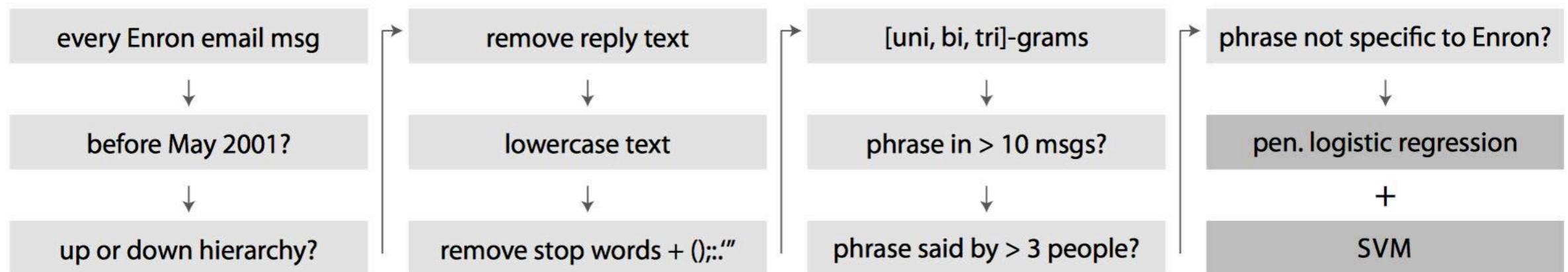
Power

- Text data: Enron emails
- Response: Enron org chart — for all pairs of entities in email (sender/recipients), who is higher on the org chart?

Gilbert 2012 (“Phrases that signal workplace hierarchy”)

Power

- Bag of words representation of text + binary classification.



Gilbert 2012 ("Phrases that signal workplace hierarchy")

\uparrow phrases	β	\uparrow phrases	β	$\leftrightarrow\downarrow$ phrases	β	$\leftrightarrow\downarrow$ phrases	β
the ability to	6.76	attach	6.72	have you been	-8.46	to manage the	-6.66
I took	6.57	that we might	6.54	you gave	-6.64	let's discuss	-5.72
are available	6.52	the calendar	6.06	we are in	-5.44	publicly	-5.24
kitchen	5.72	can you get	5.72	title	-5.05	promotion	-5.02
thought you would	5.65	driving	5.61	need in	-4.80	good one	-4.62
, I'll be	5.51	thoughts on	5.51	opened	-4.57	determine the	-4.47
looks fine	5.50	shit	5.45	initiatives	-4.38	is difficult	-4.36
voicemail	5.43	we can talk	5.41	. I would	-4.34	man	-4.26
tremendous	5.27	it does	5.21	we will probably	-4.12	number we	-4.11
will you	5.17	involving	5.15	any comments	-4.06	contact you	-4.05
left a	5.07	the report	5.04	you said	-3.99	the problem is	-3.97
I put	4.90	please change	4.88	I left	-3.88	you did	-3.78
you ever	4.80	issues I	4.76	can you help	-3.68	cool	-3.54
I'll give	4.69	is really	4.65	send this	-3.47	your attention	-3.44
okay ,	4.60	your review	4.56	whether we	-3.44	to think	-3.44
to send it	4.48	europe	4.45	the trade	-3.40	addition to the	-3.30
communications	4.38	weekend .	4.35	and I thought	-3.28	great thanks	-3.24
a message	4.35	have our	4.33	should include	-3.19	selected	-3.16
one I	4.28	interviews	4.28	please send	-3.14	ext	-3.13
can I get	4.28	you mean	4.26	existing	-3.06	and let me	-3.05
worksheet	4.15	haven't been	4.10	mondays	-3.02	security	-3.01
liked	4.07	me . 1	4.07	presentation on	-2.95	got the	-2.94
I gave you	3.95	tiger	3.94	let's talk	-2.94	get your	-2.88
credit will	3.88	change in	3.88	the items	-2.78	this week and	-2.77
you make	3.86	item	3.84	i hope you	-2.77	team that	-2.75
together and	3.82	a decision	3.82	did it	-2.75	a deal	-2.71
have presented	3.78	a discussion	3.74	test	-2.69	yours .	-2.68
think about	3.71	sounds good	3.65	be sure	-2.65	briefing	-2.60

these predict message going up

\uparrow phrases	β	\uparrow phrases	β	$\leftrightarrow\downarrow$ phrases	β	$\leftrightarrow\downarrow$ phrases	β
the ability to	6.76	attach	6.72	have you been	-8.46	to manage the	-6.66
I took	6.57	that we might	6.54	you gave	-6.64	let's discuss	-5.72
are available	6.52	the calendar	6.06	we are in	-5.44	publicly	-5.24
kitchen	5.72	can you get	5.72	title	-5.05	promotion	-5.02
thought you would	5.65	driving	5.61	need in	-4.80	good one	-4.62
, I'll be	5.51	thoughts on	5.51	opened	-4.57	determine the	-4.47
looks fine	5.50	shit	5.45	initiatives	-4.38	is difficult	-4.36
voicemail	5.43	we can talk	5.41	. I would	-4.34	man	-4.26
tremendous	5.27	it does	5.21	we will probably	-4.12	number we	-4.11
will you	5.17	involving	5.15	any comments	-4.06	contact you	-4.05
left a	5.07	the report	5.04	you said	-3.99	the problem is	-3.97
I put	4.90	please change	4.88	I left	-3.88	you did	-3.78
you ever	4.80	issues I	4.76	can you help	-3.68	cool	-3.54
I'll give	4.69	is really	4.65	send this	-3.47	your attention	-3.44
okay ,	4.60	your review	4.56	whether we	-3.44	to think	-3.44
to send it	4.48	europe	4.45	the trade	-3.40	addition to the	-3.30
communications	4.38	weekend .	4.35	and I thought	-3.28	great thanks	-3.24
a message	4.35	have our	4.33	should include	-3.19	selected	-3.16
one I	4.28	interviews	4.28	please send	-3.14	ext	-3.13
can I get	4.28	you mean	4.26	existing	-3.06	and let me	-3.05
worksheet	4.15	haven't been	4.10	mondays	-3.02	security	-3.01
liked	4.07	me . 1	4.07	presentation on	-2.95	got the	-2.94
I gave you	3.95	tiger	3.94	let's talk	-2.94	get your	-2.88
credit will	3.88	change in	3.88	the items	-2.78	this week and	-2.77
you make	3.86	item	3.84	i hope you	-2.77	team that	-2.75
together and	3.82	a decision	3.82	did it	-2.75	a deal	-2.71
have presented	3.78	a discussion	3.74	test	-2.69	yours .	-2.68
think about	3.71	sounds good	3.65	be sure	-2.65	briefing	-2.60

these predict message going up

not going up

\uparrow phrases	β	\uparrow phrases	β	$\leftrightarrow\downarrow$ phrases	β	$\leftrightarrow\downarrow$ phrases	β
the ability to	6.76	attach	6.72	have you been	-8.46	to manage the	-6.66
I took	6.57	that we might	6.54	you gave	-6.64	let's discuss	-5.72
are available	6.52	the calendar	6.06	we are in	-5.44	publicly	-5.24
kitchen	5.72	can you get	5.72	title	-5.05	promotion	-5.02
thought you would	5.65	driving	5.61	need in	-4.80	good one	-4.62
, I'll be	5.51	thoughts on	5.51	opened	-4.57	determine the	-4.47
looks fine	5.50	shit	5.45	initiatives	-4.38	is difficult	-4.36
voicemail	5.43	we can talk	5.41	. I would	-4.34	man	-4.26
tremendous	5.27	it does	5.21	we will probably	-4.12	number we	-4.11
will you	5.17	involving	5.15	any comments	-4.06	contact you	-4.05
left a	5.07	the report	5.04	you said	-3.99	the problem is	-3.97
I put	4.90	please change	4.88	I left	-3.88	you did	-3.78
you ever	4.80	issues I	4.76	can you help	-3.68	cool	-3.54
I'll give	4.69	is really	4.65	send this	-3.47	your attention	-3.44
okay ,	4.60	your review	4.56	whether we	-3.44	to think	-3.44
to send it	4.48	europe	4.45	the trade	-3.40	addition to the	-3.30
communications	4.38	weekend .	4.35	and I thought	-3.28	great thanks	-3.24
a message	4.35	have our	4.33	should include	-3.19	selected	-3.16
one I	4.28	interviews	4.28	please send	-3.14	ext	-3.13
can I get	4.28	you mean	4.26	existing	-3.06	and let me	-3.05
worksheet	4.15	haven't been	4.10	mondays	-3.02	security	-3.01
liked	4.07	me . 1	4.07	presentation on	-2.95	got the	-2.94
I gave you	3.95	tiger	3.94	let's talk	-2.94	get your	-2.88
credit will	3.88	change in	3.88	the items	-2.78	this week and	-2.77
you make	3.86	item	3.84	i hope you	-2.77	team that	-2.75
together and	3.82	a decision	3.82	did it	-2.75	a deal	-2.71
have presented	3.78	a discussion	3.74	test	-2.69	yours .	-2.68
think about	3.71	sounds good	3.65	be sure	-2.65	briefing	-2.60

Measuring Power: Communication Accommodation Theory

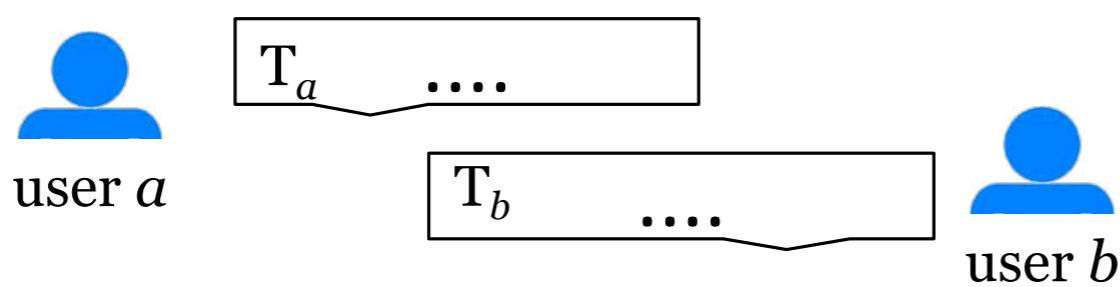
- Giles, Taylor, and Bourhis 1973; Giles, Coupland, and Coupland 1991)
 - Shift in behavior to become more similar (convergence) or dissimilar (divergence) from conversation partner
 - Words, gestures, pitch, utterance length...
 - Converging speakers are often viewed as more favorable and cooperative (gain social approval)
 - Roots in social psychology

Measuring linguistic accommodation

- Danescu-Niculescu-Mizil et al. (2011) measured linguistic style accommodation.
- Selected categories from LIWC capturing style (e.g., article, preposition, quantifier, etc..)

Measuring linguistic accommodation

- Danescu-Niculescu-Mizil et al. (2011) measured linguistic style accommodation.
- Selected categories from LIWC capturing style (e.g., article, preposition, quantifier, etc..)



$T_b \rightarrow T_a$ reply from *b* to *a*

T_b^C tweet written by *b* has category *C*

$$P(T_b^C | T_a^C, T_b \rightarrow T_a) - P(T_b^C | T_b \rightarrow T_a)$$

Linguistic accommodation and power

Linguistic accommodation and power

- Discussions among Wikipedia editors

Linguistic accommodation and power

- Discussions among Wikipedia editors
- Danescu-Niculescu-Mizil et al. 2012:
 - People coordinate more with interlocutors who have higher power.
 - Changes in linguistic coordination behavior when someone becomes an admin

Linguistic accommodation and power

- Discussions among Wikipedia editors
- Danescu-Niculescu-Mizil et al. 2012:
 - People coordinate more with interlocutors who have higher power.
 - Changes in linguistic coordination behavior when someone becomes an admin
- Noble and Fernandez 2015
 - But for highly central users, adminship had no significant effect on amount of coordination received.

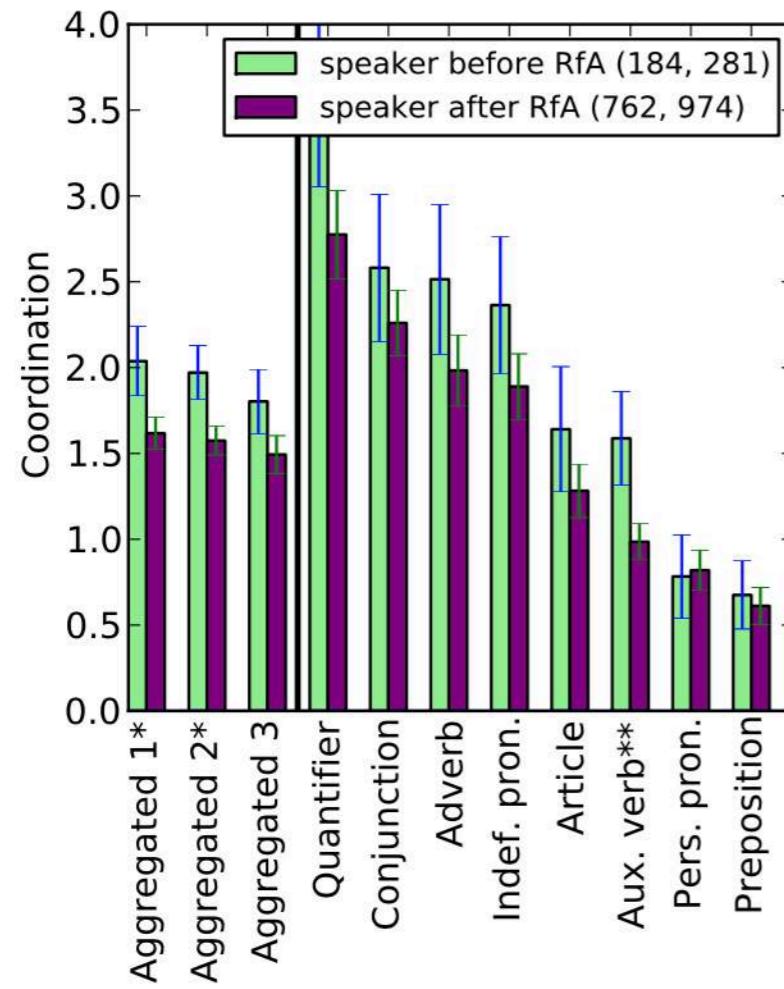
Power

- Text data: Wikipedia discussions, SCOTUS arguments
- Response: Wikipedia admins/non-admins; SCOTUS justices/lawyers

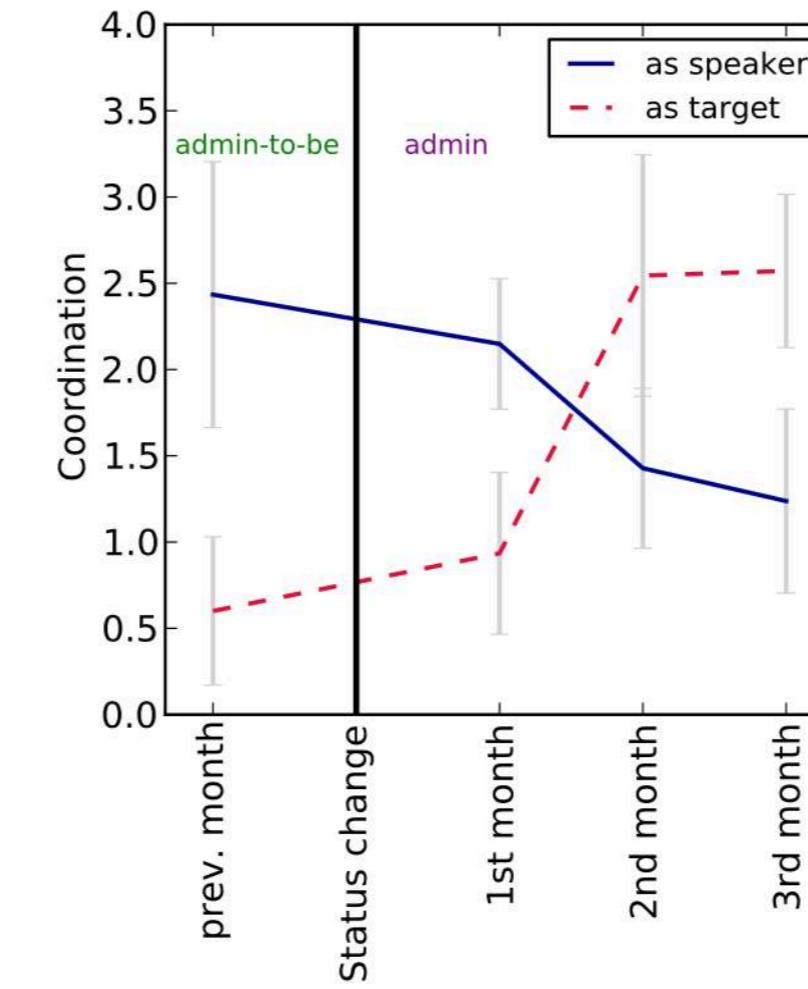
Danescu-Niculescu-Mizil et al. 2012 (“Echoes of Power”)

Power

- LIWC representation of text + measurements of accommodation (adapting your speech to the language of your interlocutor)



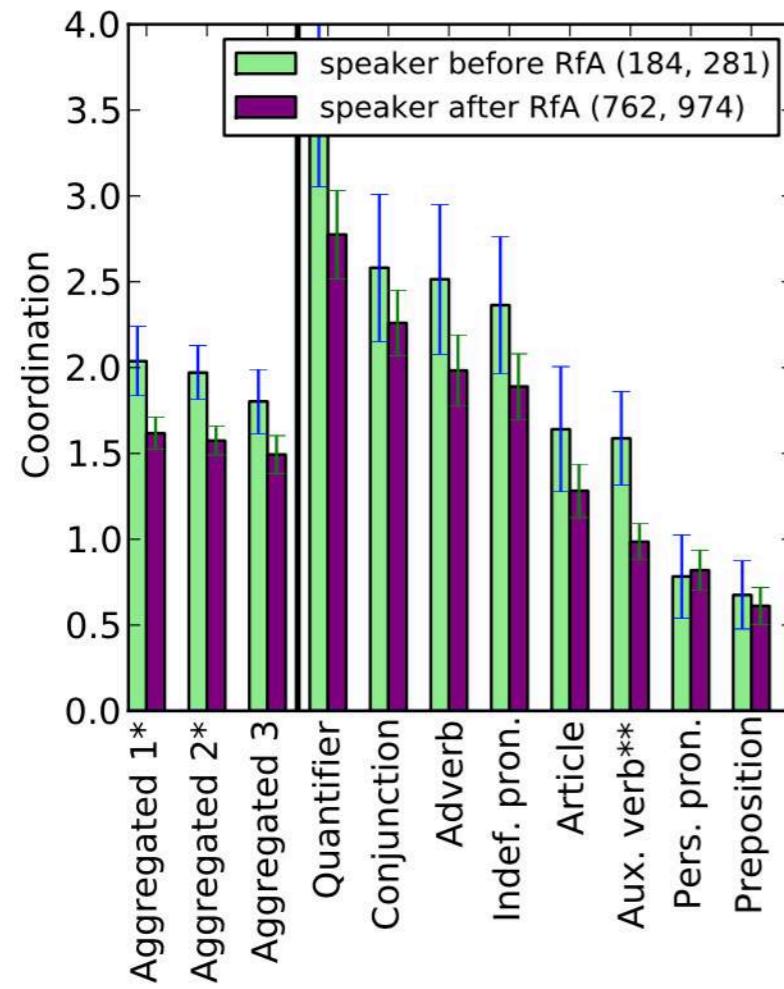
(a) Supporting $\mathcal{P}'_{\text{speaker}}$



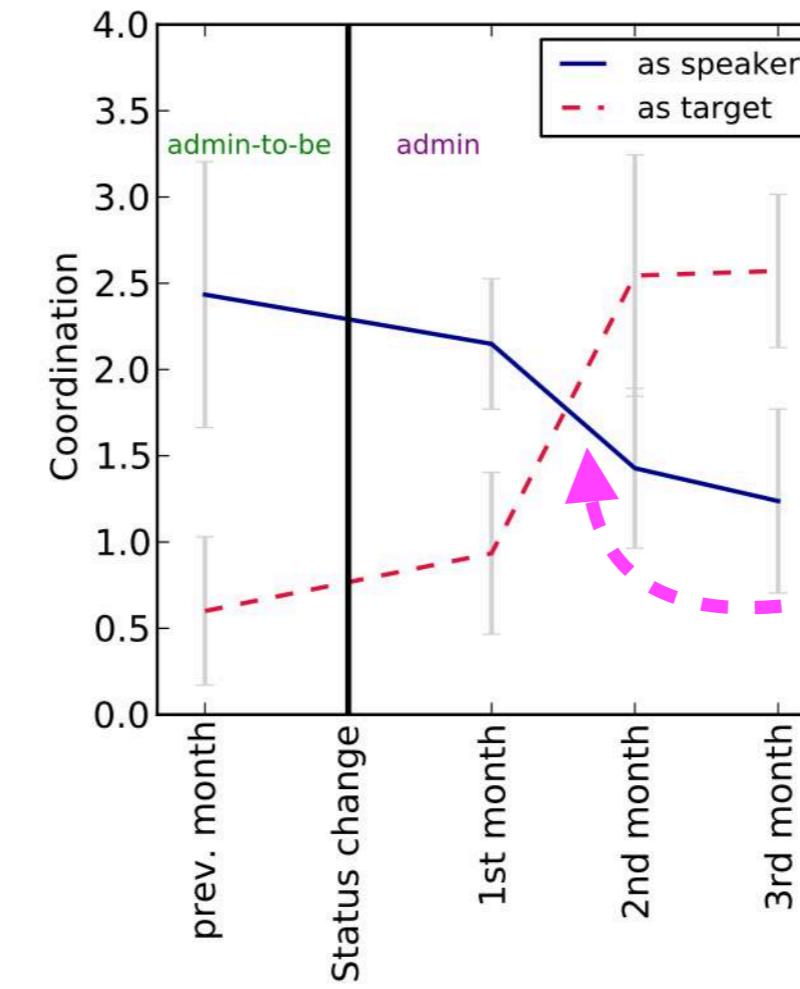
(b) Timed effect of status change (\mathcal{P})

Power

- LIWC representation of text + measurements of accommodation (adapting your speech to the language of your interlocutor)



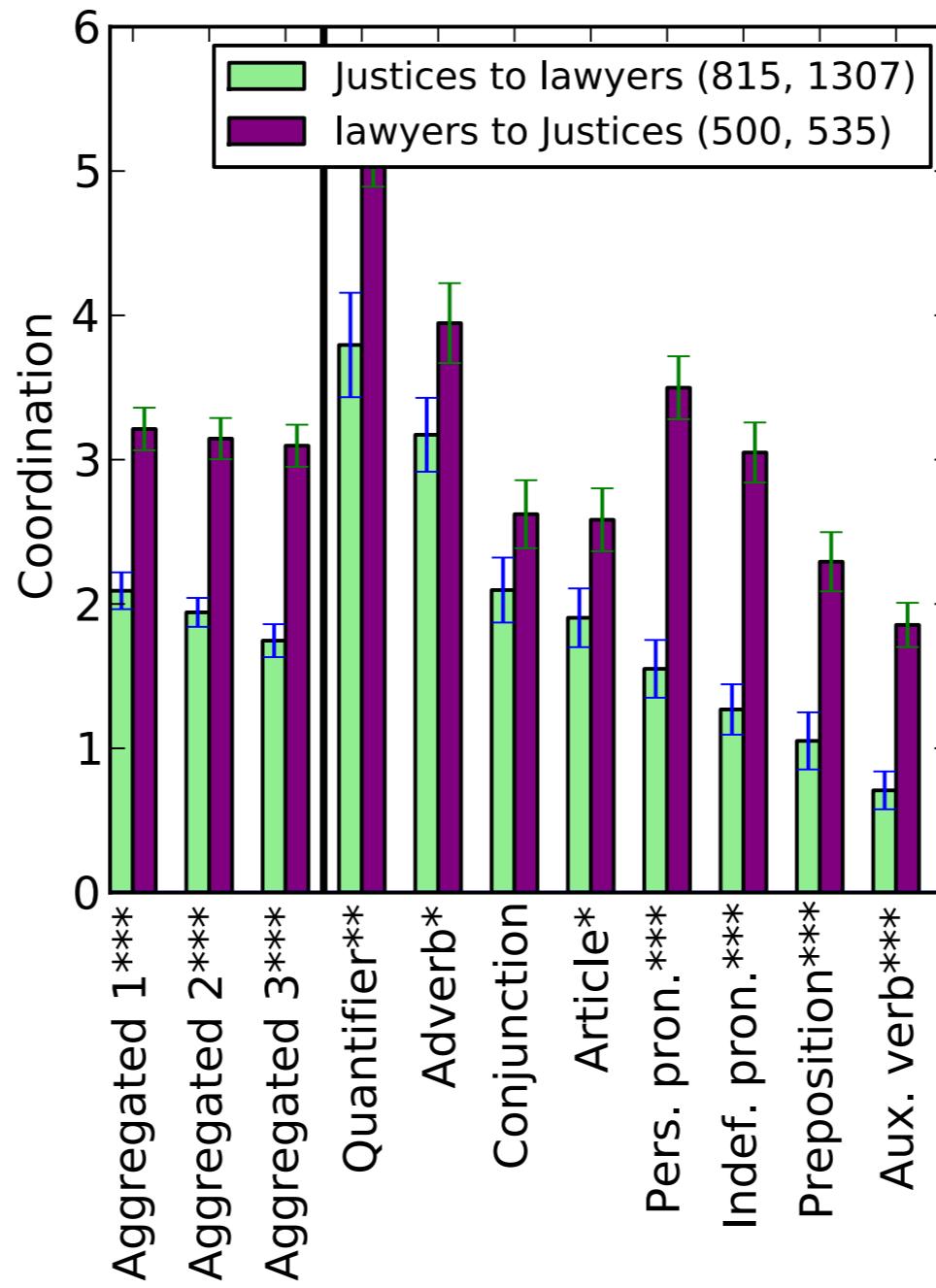
(a) Supporting $\mathcal{P}'_{\text{speaker}}$



(b) Timed effect of status change (\mathcal{P})

When users get power,
they linguistically
coordinate less

Coordination reveals clear power differences



Respect

- Data: transcripts of 981 OPD traffic stops (everyday interactions)
- Response: race

Voigt et al. 2017, "Language from police body camera footage shows racial disparities in officer respect"

Respect

- Present one dialogue turn (police/driver) to be rated by people for respect (4-point Likert scale).
High IAA.
- Build a predictive model mapping **text** to **respect**.

Voigt et al. 2017, “Language from police body camera footage shows racial disparities in officer respect”

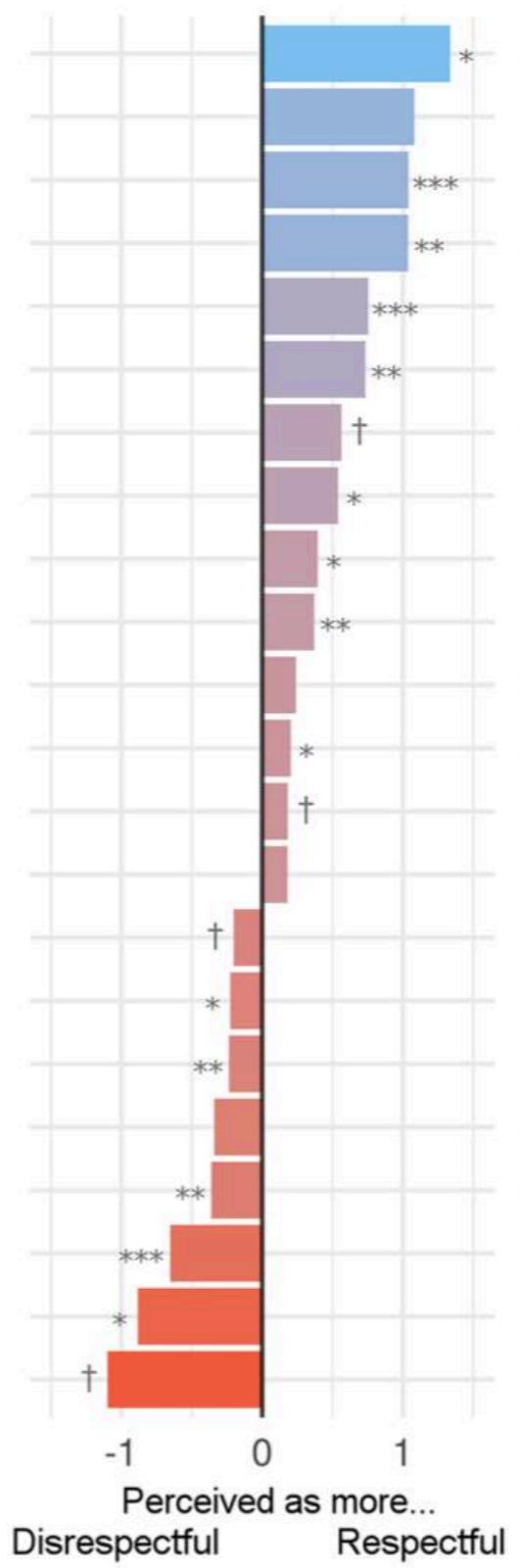
Feature Name	Implementation
Adverbial "Just"	"Just" occurs in a dependency arc as the head of an <code>advmod</code> relation
Apologizing	Lexicon: "sorry", "oops", "woops", "excuse me", "forgive me", "apologies", "apologize", "my bad", "my fault"
Ask for Agency	Lexicon: "do me a favor", "let me", "allow me", "can i", "should i", "may i", "might i", "could i"
Bald Command	The first word in a sentence is a bare verb with part-of-speech tag VB ("look", "give", "wait" etc.) but is not one of "be", "do", "have", "thank", "please", "hang".
Colloquialism	Regular expression capturing "y'all", "ain't" and words ending in "in'" such as "walkin'", "talkin'", etc., as marked by transcribers
Conditional	Lexicon: "if"
Disfluency	Word fragment ("Well I thi-") as indicated by transcribers
Filled Pauses	Lexicon: "um", "uh"
First Names	Top 1000 most common first names from the 1990 US Census, where first letter is capitalized in transcript
Formal Titles	Lexicon: "sir", "ma'am", "maam", "mister", "mr*", "ms*", "madam", "miss", "gentleman", "lady"
For Me	Lexicon: "for me"
For You	Lexicon: "for you"
Give Agency	Lexicon: "let you", "allow you", "you can", "you may", "you could"
Gratitude	Lexicon: "thank", "thanks", "appreciate"
Goodbye	Lexicon: "goodbye", "bye", "see you later"
Hands on the Wheel	Regular expression capturing cases like "keep your hands on the wheel" and "leave your hands where I can see them": "hands? ([.,?!:]++)?(wheel see)"

Hedges	All words in the "Tentat" LIWC lexicon
Impersonal Pronoun	All words in the "Imppron" LIWC lexicon
Informal Titles	Lexicon: "dude*", "bro*", "boss", "bud", "buddy", "champ", "man", "guy*", "guy", "brotha", "sista", "son", "sonny", "chief"
Introductions	Regular expression capturing cases like "I'm Officer [name] from the OPD" and "How's it going?": "((i my name).+officer officer.+ (oakland opd)) ((hi hello hey good afternoon good morning good evening how are you doing how 's it going))"
Last Names	Top 5000 most common last names from the 1990 US Census, where first letter is capitalized in transcript
Linguistic Negation	All words in the "Negate" LIWC lexicon
Negative Words	All words in the "Negativ" category in the Harvard General Inquierer, matching on word lemmas
Positive Words	All words in the "Positiv" category in the Harvard General Inquierer, matching on word lemmas
Please	Lexicon: "please"
Questions	Occurrence of a question mark
Reassurance	Lexicon: "'s okay", "n't worry", "no big deal", "no problem", "no worries", "'s fine", "you 're good", "is fine", "is okay"
Safety	Regular expression for all words beginning with the prefix "safe", such as "safe", "safety", "safely"
Swear Words	All words in the "Swear" LIWC lexicon
Tag Question	Regular expression capturing cases like "..., right?" and "..., don't you?": ", (((all right right okay yeah please you know)(sir ma'am miss son)?) ((are is do can have will won't) (n't)?(i me she us we you he they them))) [?] "
The Reason for the Stop	Lexicon: "reason", "stop* you", "pull* you", "why i", "why we", "explain", "so you understand"
Time Minimizing	Regular expression capturing cases like "in a minute" and "let's get this done quick": "(a one a few) (minute min second sec moment)s? this[.,?!]+quick right back"

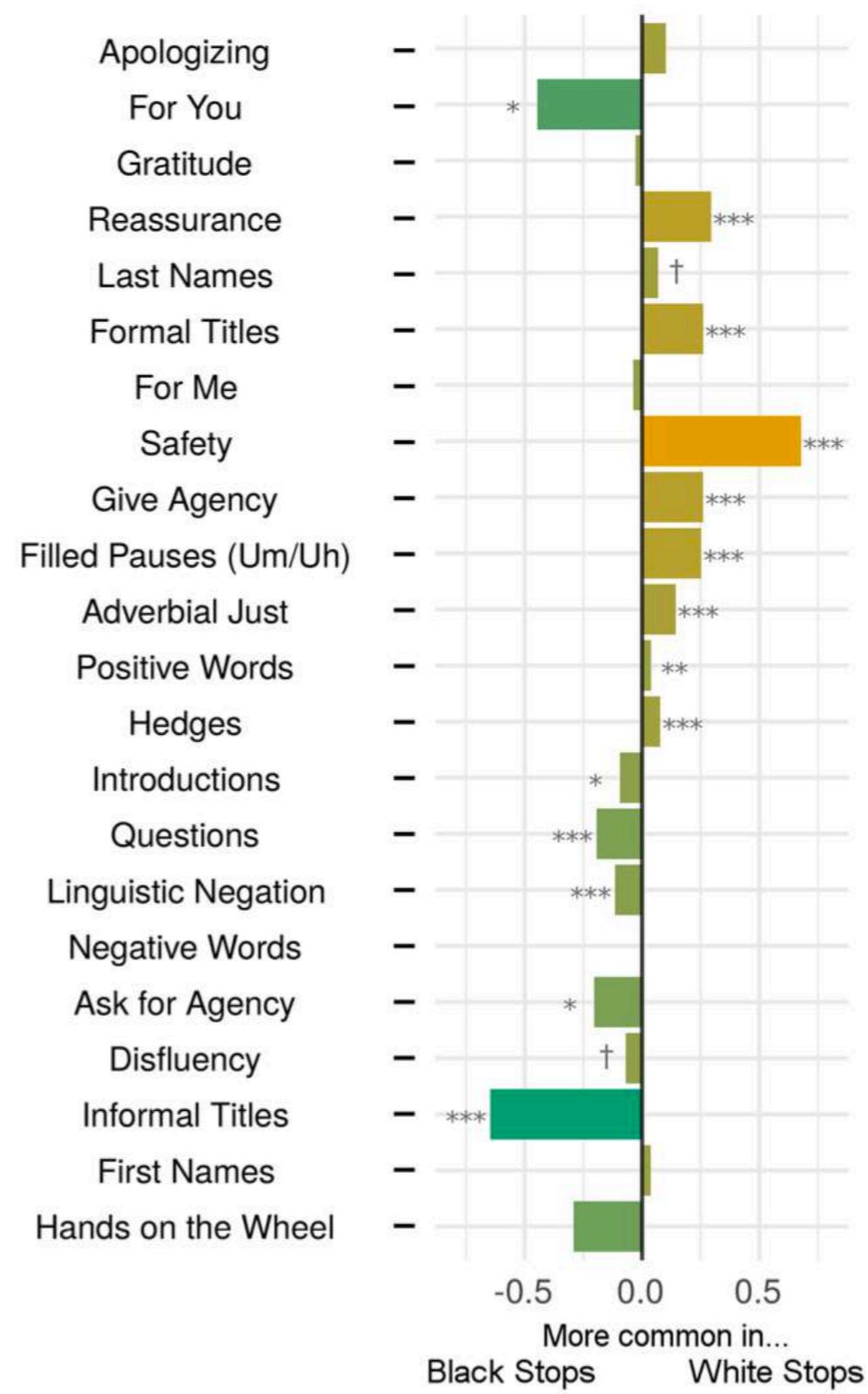
Respect

EXAMPLE	RESPECT SCORE
<p>FIRST NAME ASK FOR AGENCY QUESTIONS</p> <p>[name], can I see that driver's license again? It- it's showing suspended. Is that- that's you?</p> <p>DISFLUENCY NEGATIVE WORD DISFLUENCY</p>	-1.07
<p>INFORMAL TITLE ASK FOR AGENCY ADVERBIAL "JUST"</p> <p>All right, my man. Do me a favor. Just keep your hands on the steering wheel real quick.</p> <p>"HANDS ON THE WHEEL"</p>	-0.51
<p>APOLOGY INTRODUCTION LAST NAME</p> <p>Sorry to stop you. My name's Officer [name] with the Police Department.</p>	0.84
<p>FORMAL TITLE SAFETY PLEASE</p> <p>There you go, ma'am. Drive safe, please.</p>	1.21
<p>ADVERBIAL "JUST" FILLED PAUSE REASSURANCE</p> <p>It just says that, uh, you've fixed it. No problem. Thank you very much, sir.</p> <p>GRATITUDE FORMAL TITLE</p>	2.07

Respect Model Coefficients



Log Odds Ratio by Race



Respect

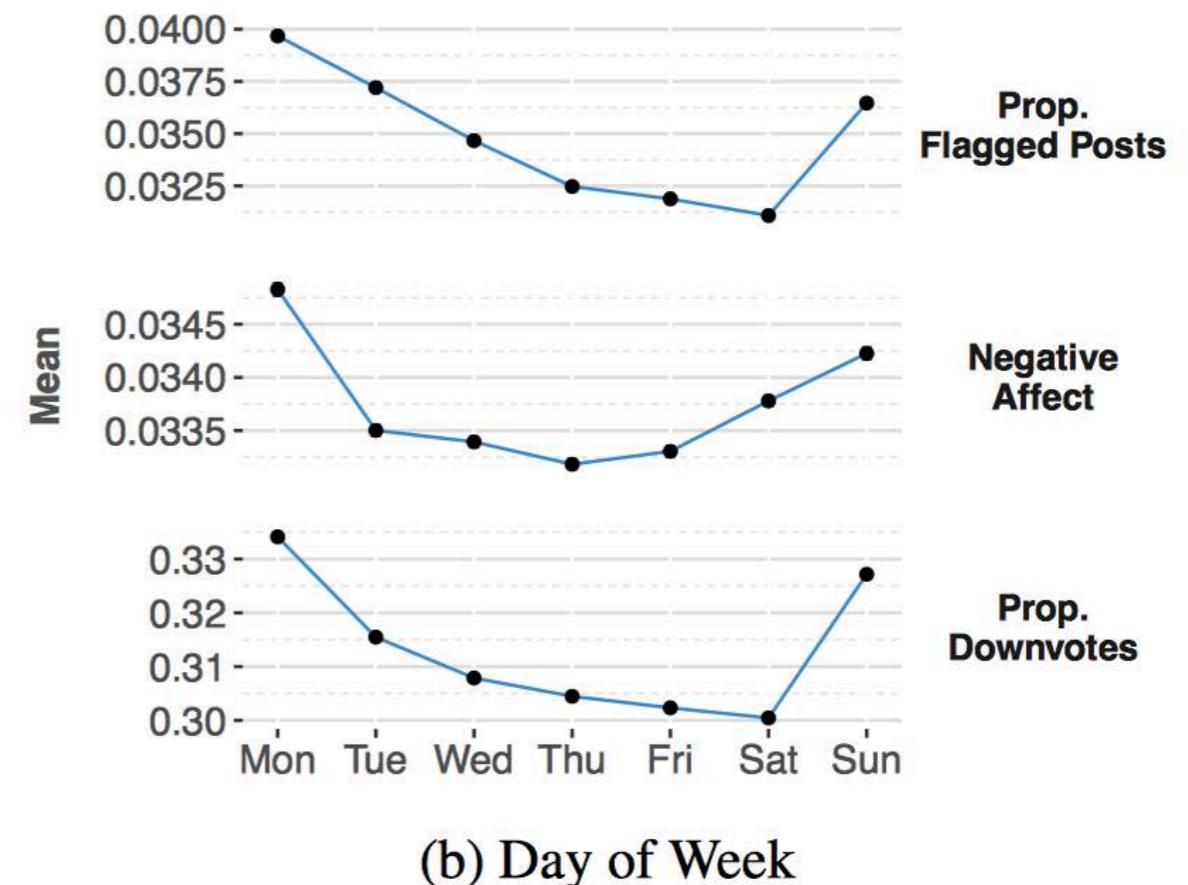
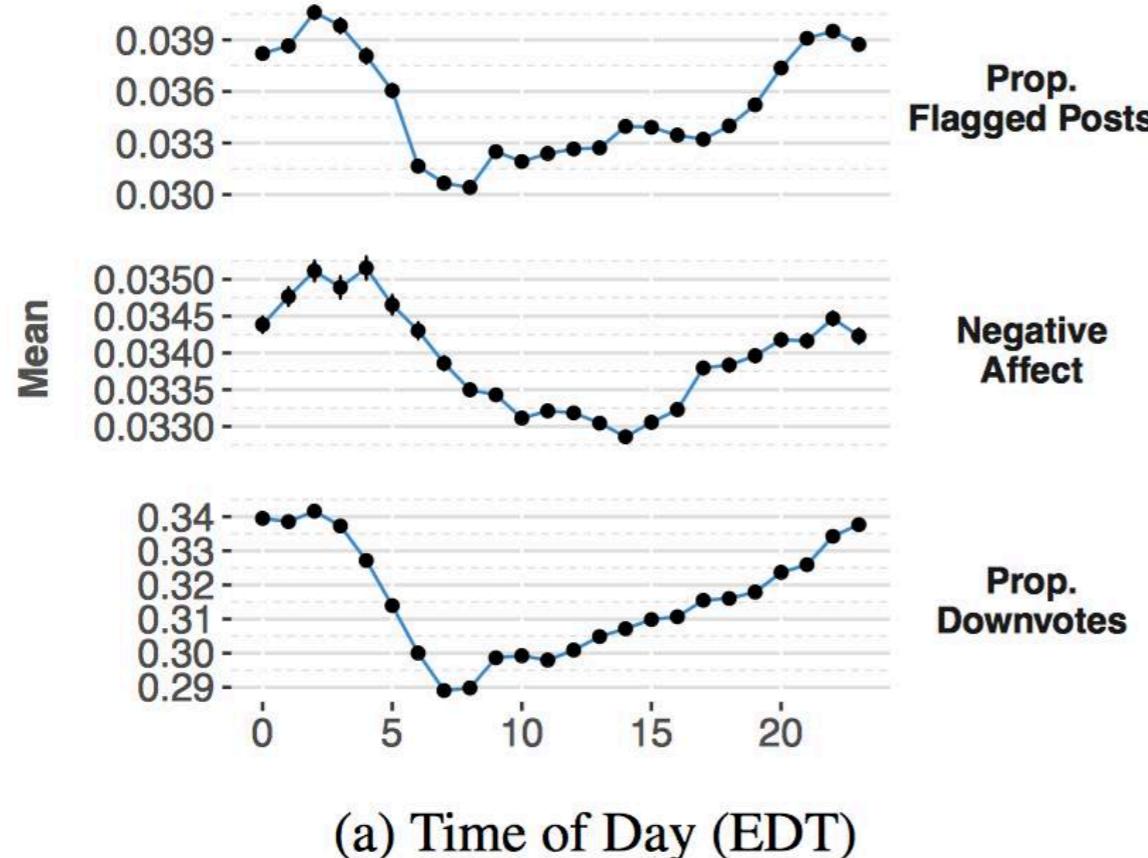
- Higher respect to white drivers, older drivers, when a citation is issued.
- Lower respect when a search is conducted.

Trolling

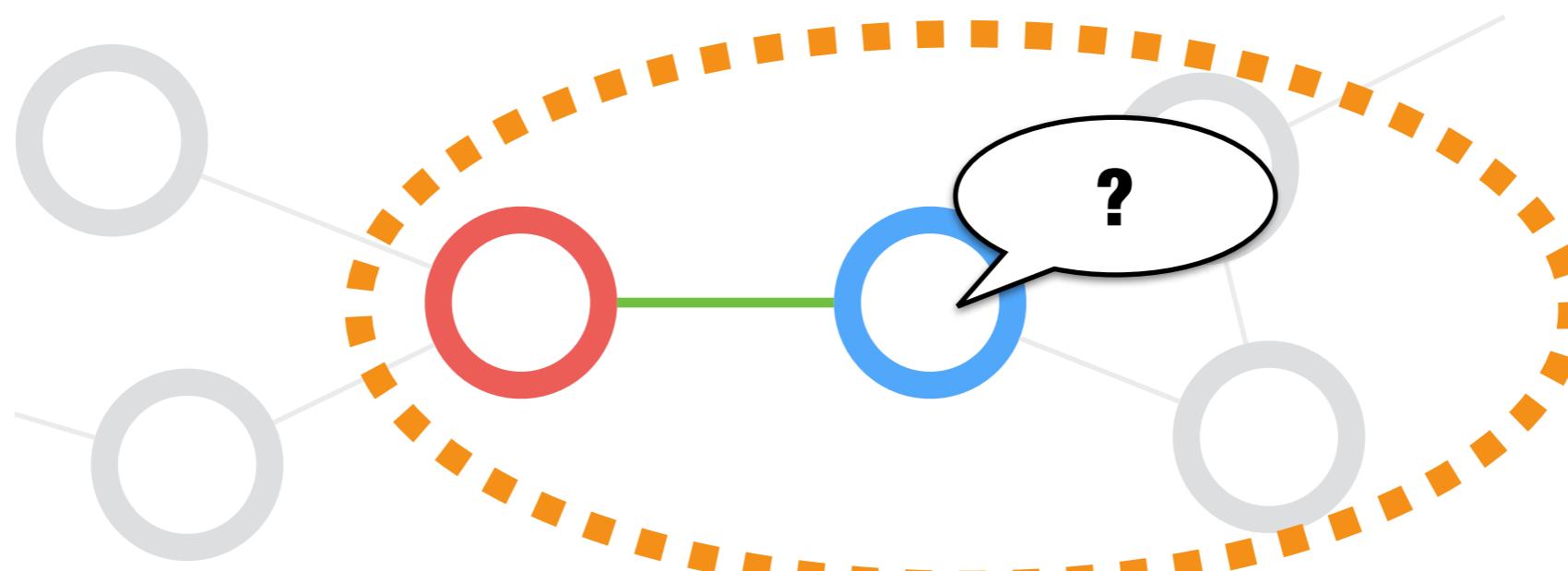
- Data: comments on CNN
- Response: comment was flagged for removal or not.
- Question: does mood or discussion context make people troll?

Cheng et al. (2017), "Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions"

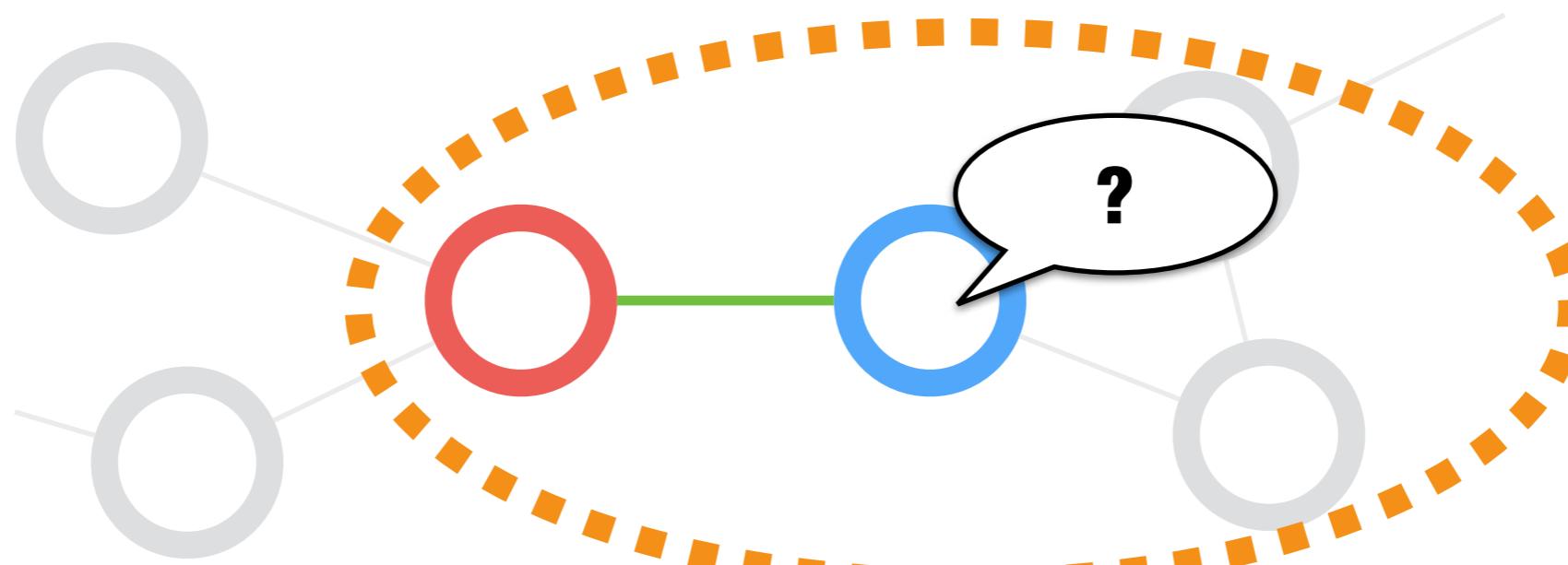
Trolling



Asking for help from your peers is a fundamental behavior in all social systems

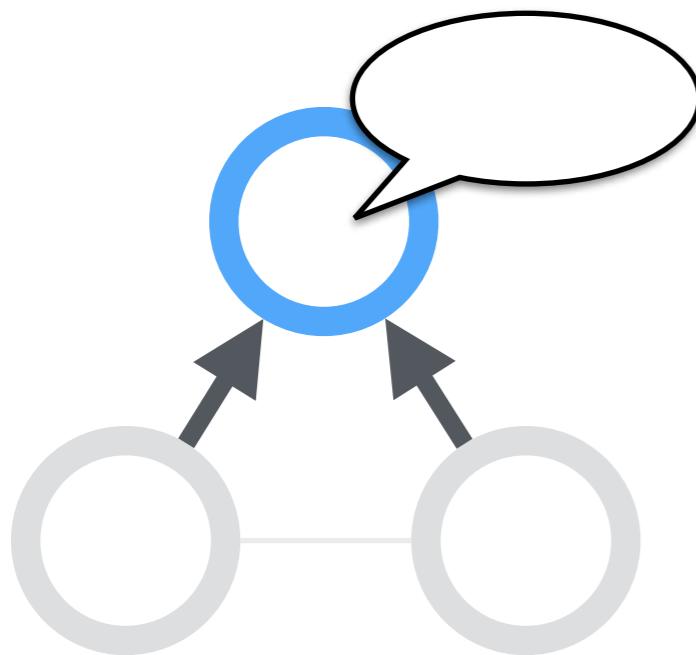


Asking for help from your peers is a fundamental behavior in all social systems

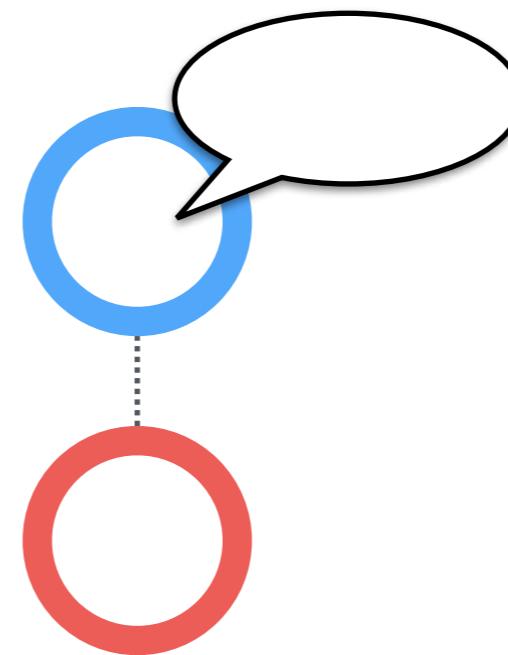


Q: What social factors affect who gets answered?

People can ask for help online in a variety of ways



Ask their
followers

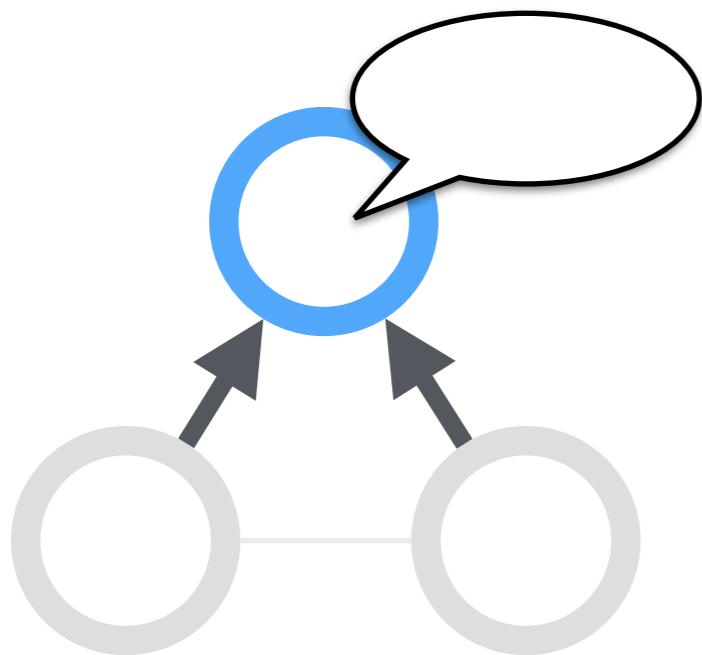


Ask a **person**
directly

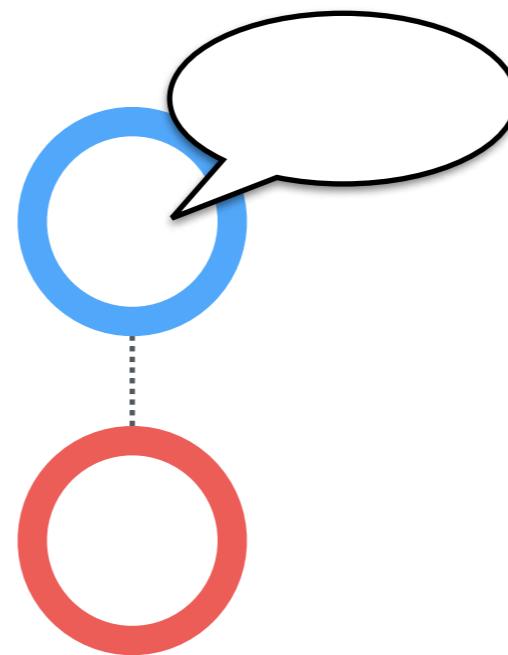


Ask a **group** they're
a member of

People can ask for help online in a variety of ways



Ask their
followers



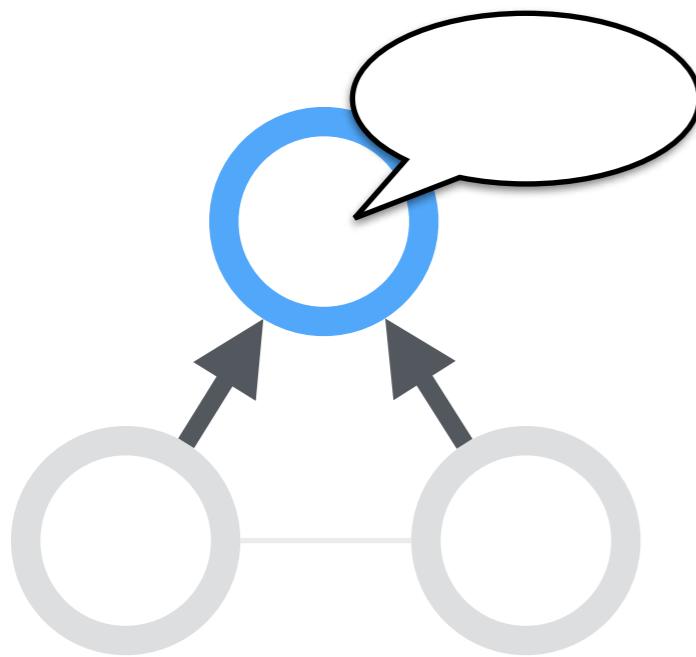
Ask a **person**
directly



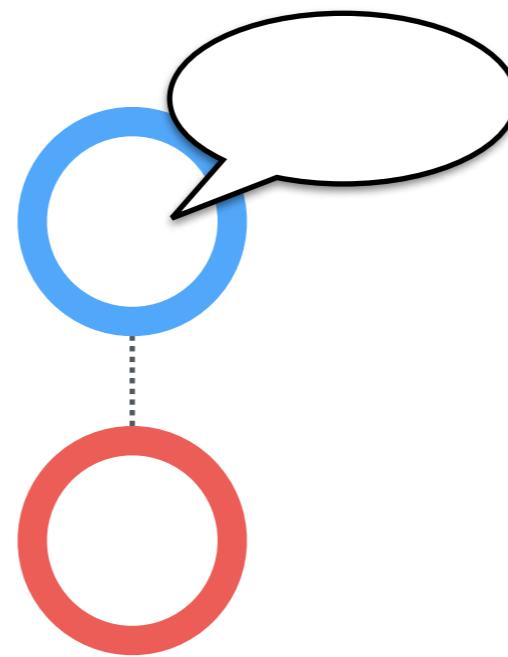
Ask a **group** they're
a member of

Task: Predict which questions will receive a response

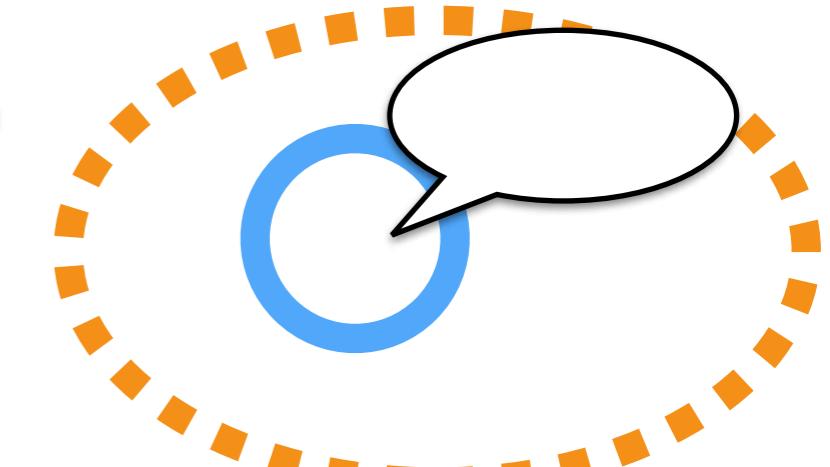
People can ask for help online in a variety of ways



Ask their
followers



Ask a **person**
directly

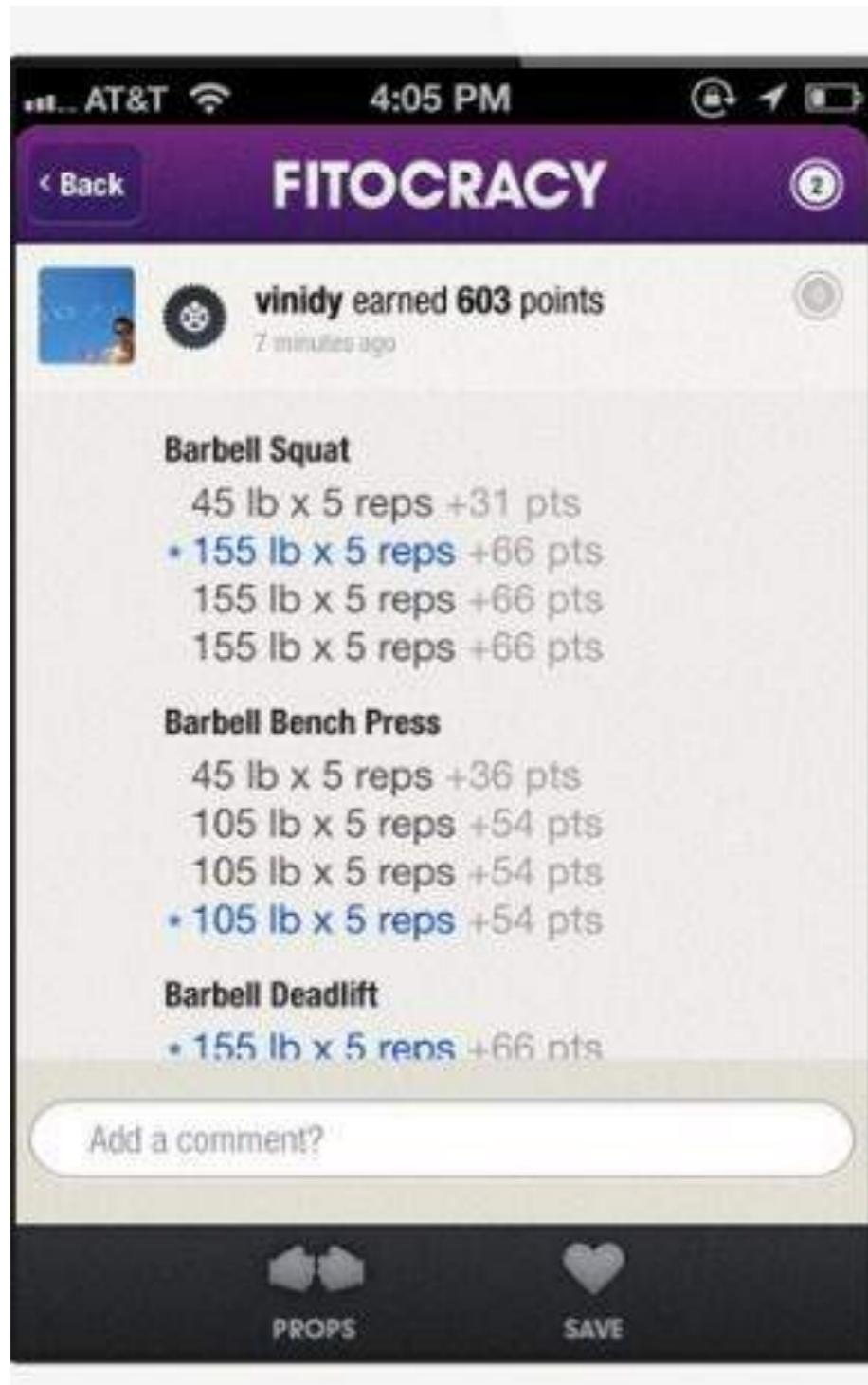


Ask a **group** they're
a member of

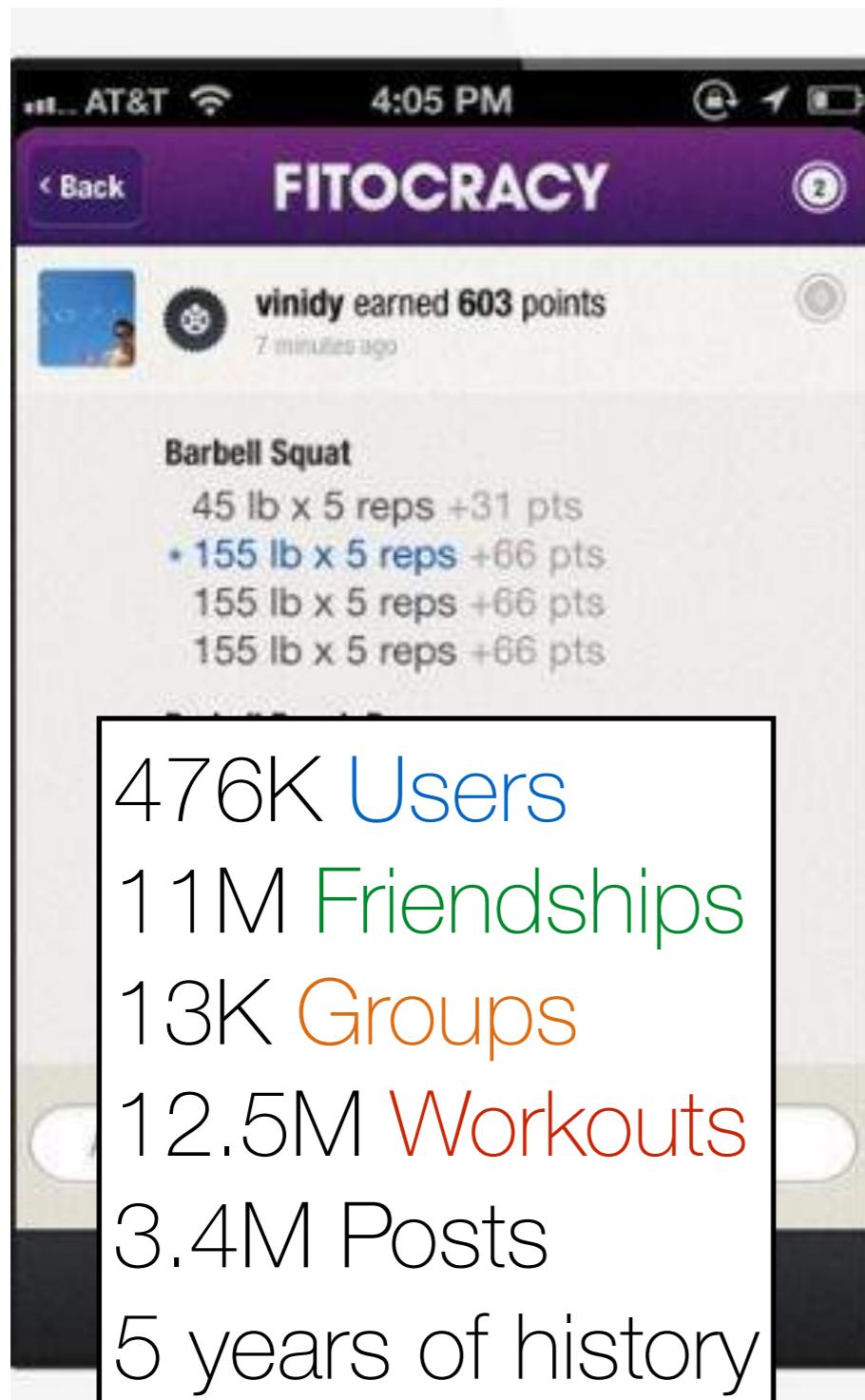
Task: Predict which questions will receive a response

Model: Logistic regression on textual and demographic variables

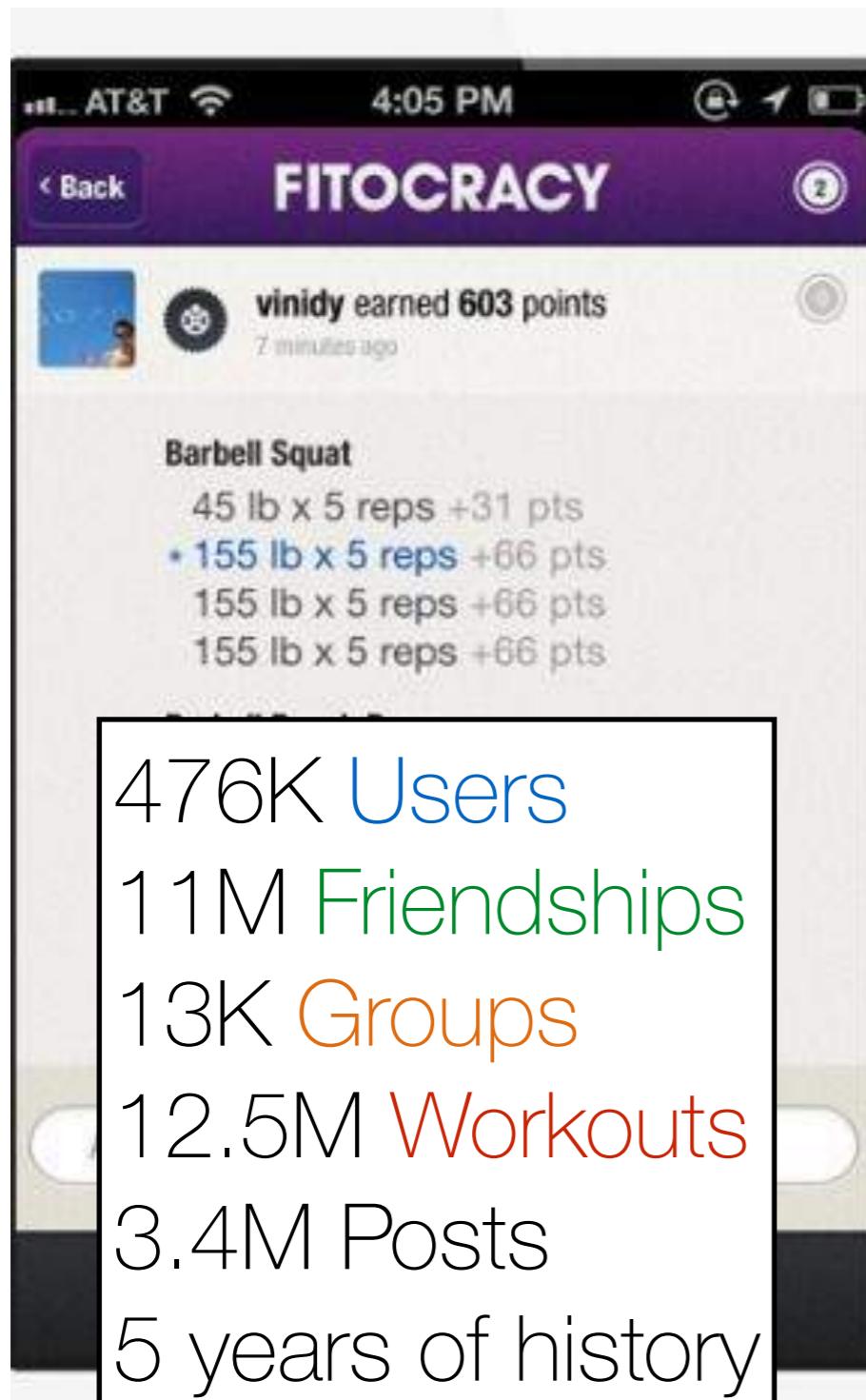
Data: the Fitocracy platform for social networking with workout tracking



Data: the Fitocracy platform for social networking with workout tracking



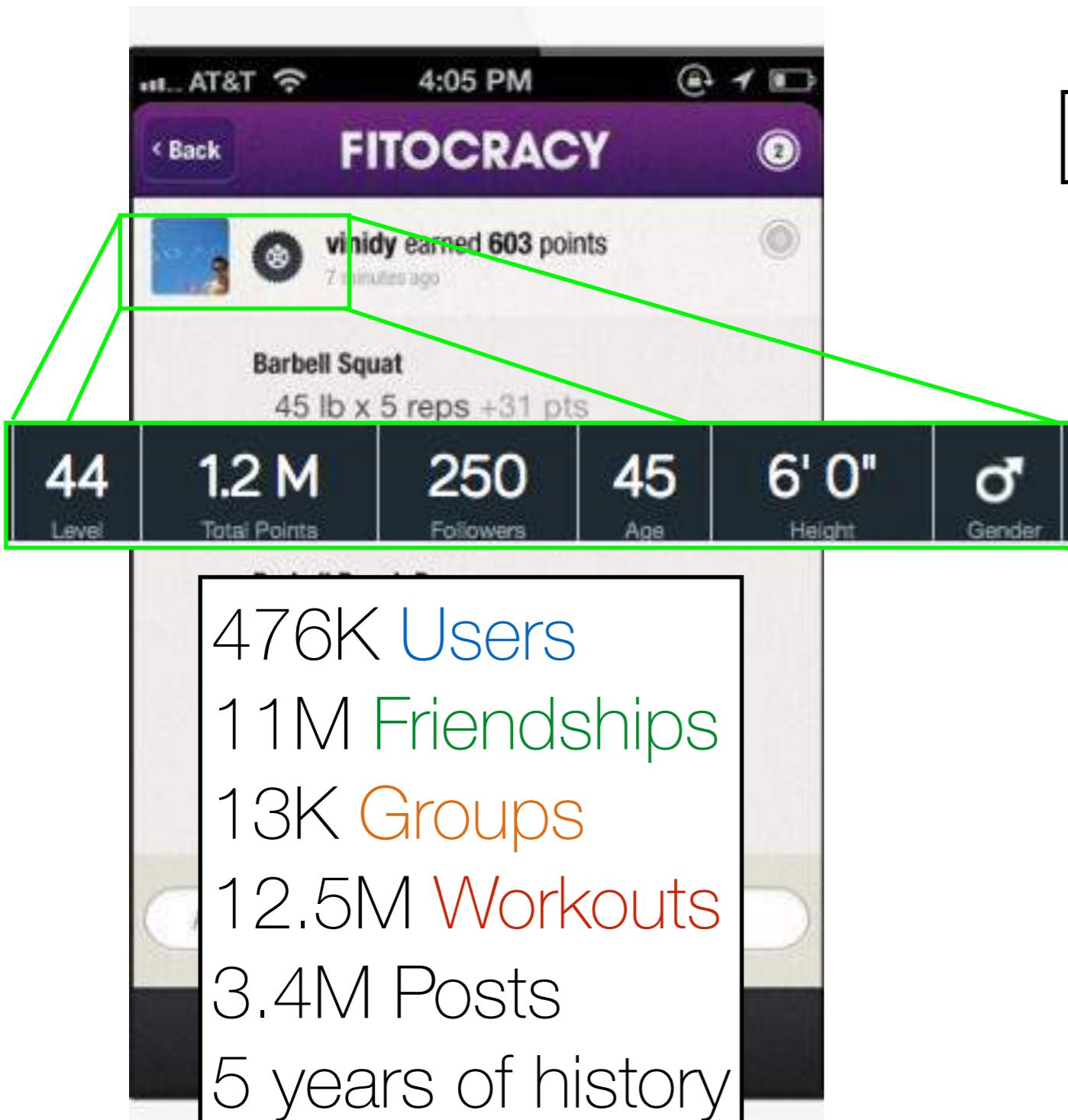
Data: the Fitocracy platform for social networking with workout tracking



Q: Who Gets Help?

- All communication is public
- Users provide their demographics

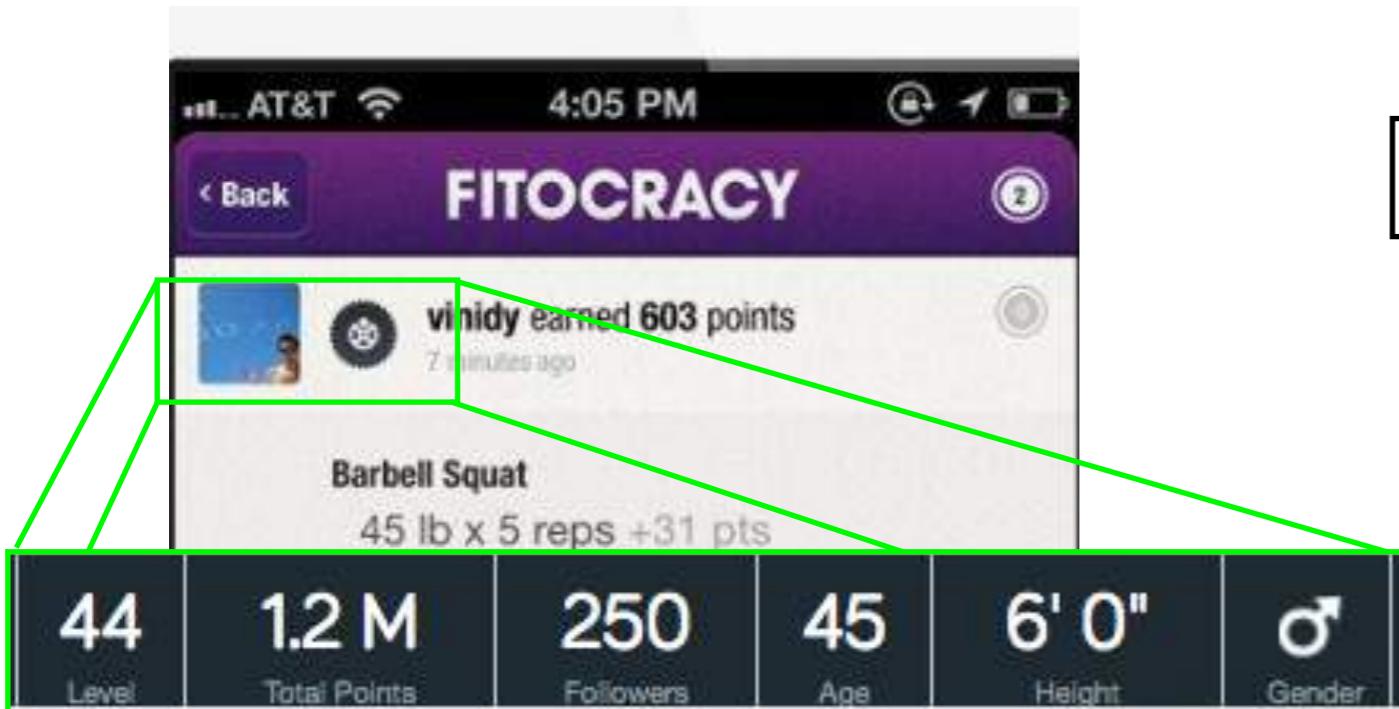
Data: the Fitocracy platform for social networking with workout tracking



Q: Who Gets Help?

- All communication is public
- Users provide their demographics

Data: the Fitocracy platform for social networking with workout tracking



476K Users
11M Friendships
13K Groups
12.5M Workouts
3.4M Posts
5 years of history

Q: Who Gets Help?

- All communication is public
- Users provide their demographics

Key Motivation: Fitocracy is a case study for generalizing to other platforms



Multiple theories predict who should receive a response.
Which apply in each social setting?

Multiple theories predict who should receive a response. Which apply in each social setting?



H6: Individuals acquire **social capital** by answering questions of people with higher social status (Solomon and Herman 1977; Goodman and Gareis 1993; Willer 2009)

Multiple theories predict who should receive a response. Which apply in each social setting?



H6: Individuals acquire **social capital** by answering questions of people with higher social status (Solomon and Herman 1977; Goodman and Gareis 1993; Willer 2009)



H7: Being a member of a group provides **affiliation benefits**

(Kelley 1967; Billig and Tajfel 1973; Goette et al. 2006)

Multiple theories predict who should receive a response. Which apply in each social setting?



H6: Individuals acquire **social capital** by answering questions of people with higher social status (Solomon and Herman 1977; Goodman and Gareis 1993; Willer 2009)



H7: Being a member of a group provides **affiliation benefits**

(Kelley 1967; Billig and Tajfel 1973; Goette et al. 2006)



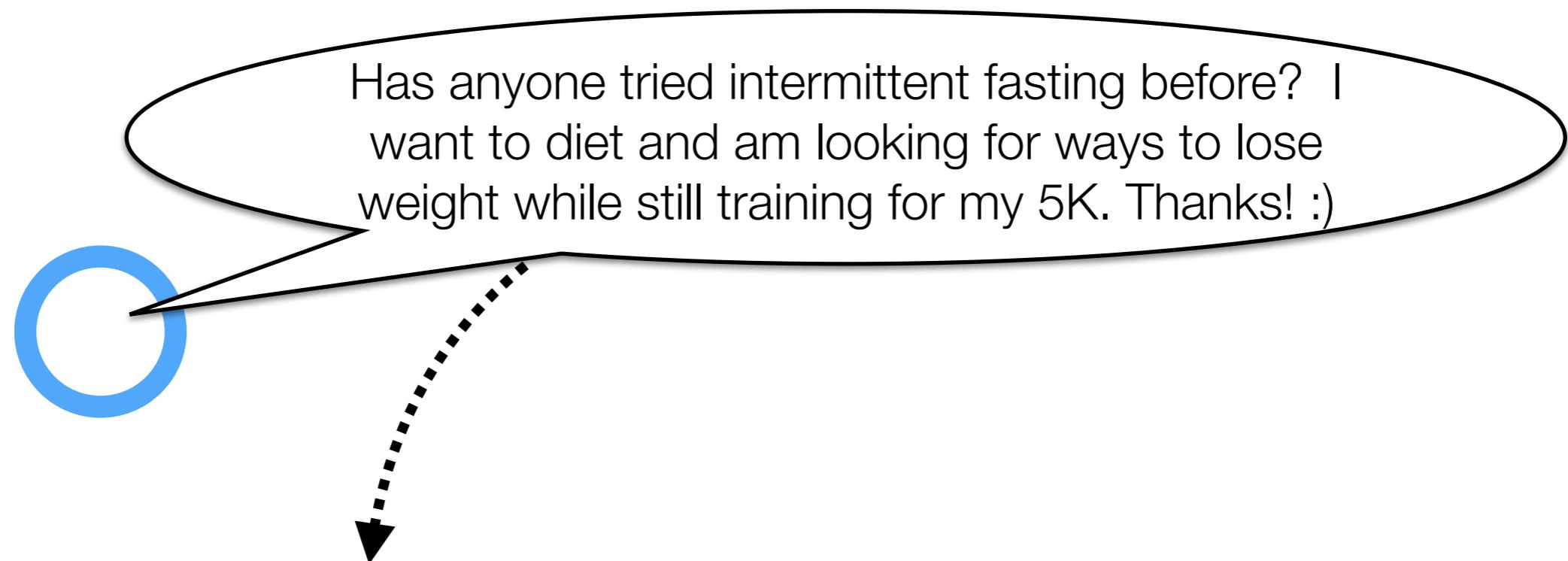
H8: The linguistic framing of the question is what predicts getting a response, e.g., **politeness**

(Tsang 2006; Bartlett and DeSteno 2006; McCullough et al. 2001; Danescu-Niculescu-Mizil et al. 2013)

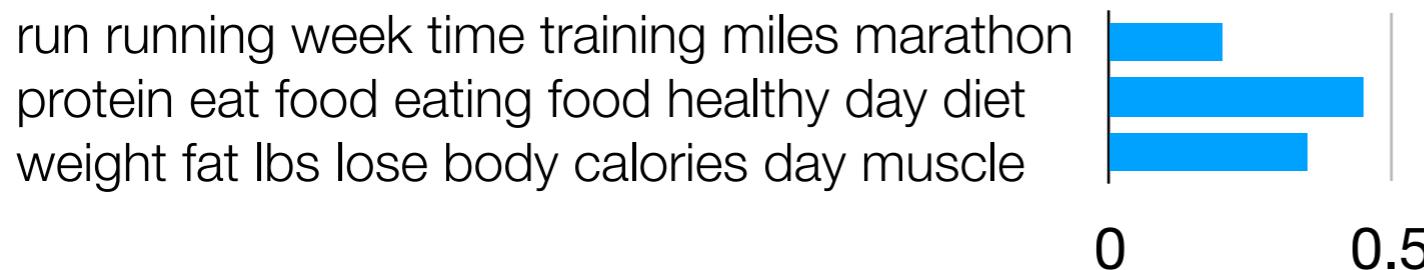
Test for social hypotheses by controlling for textual variation



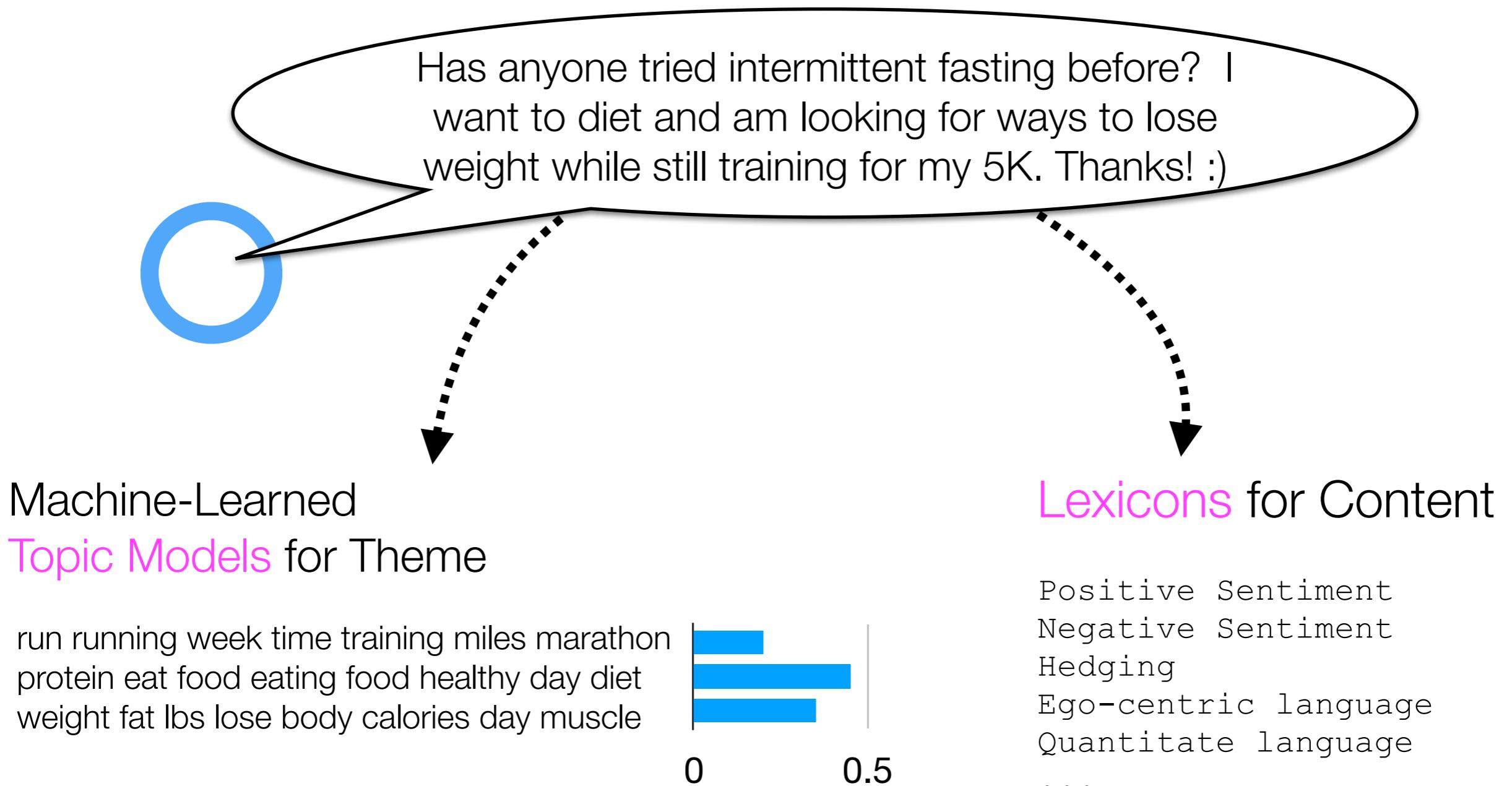
Test for social hypotheses by controlling for textual variation

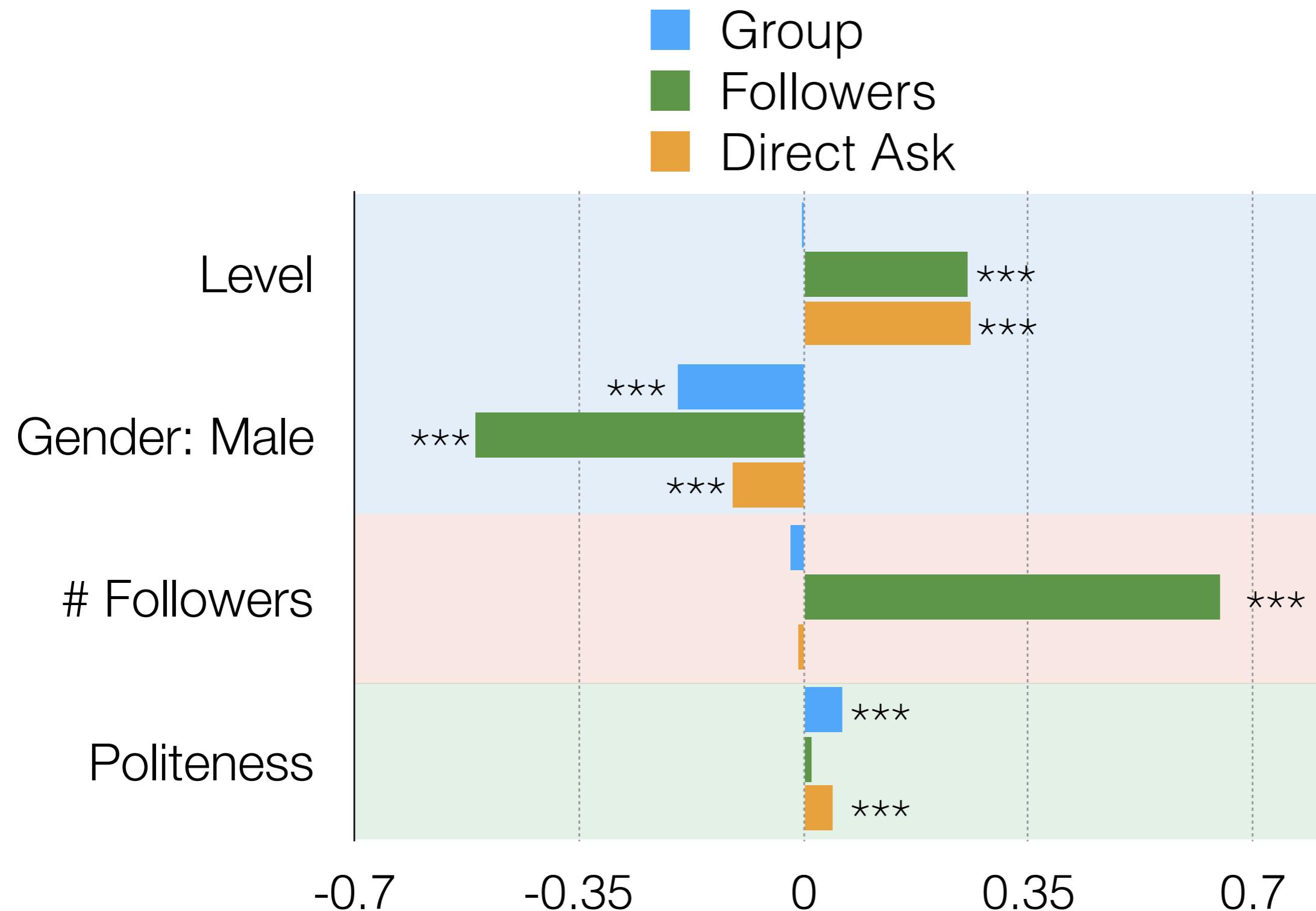


Machine-Learned
Topic Models for Theme

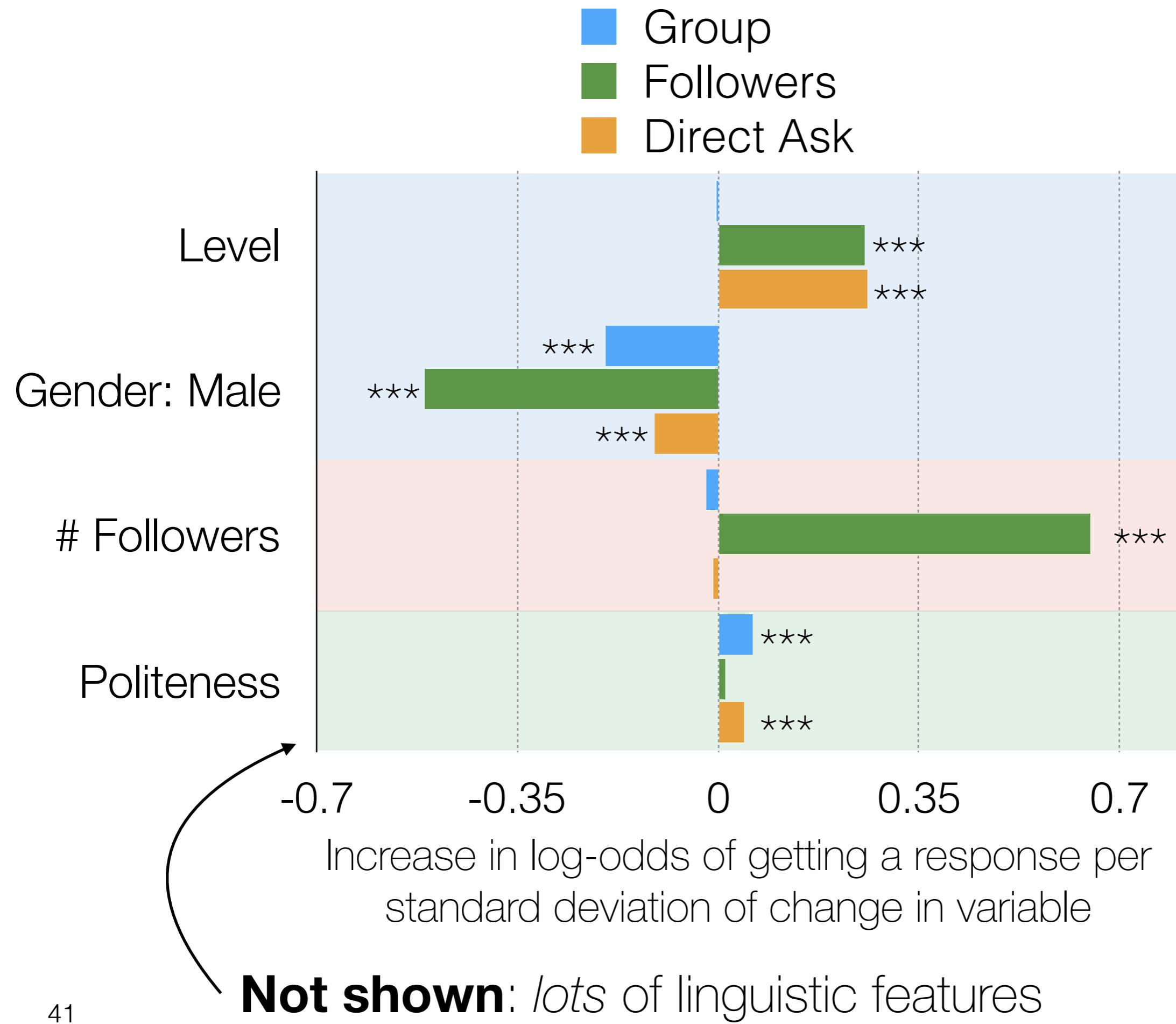


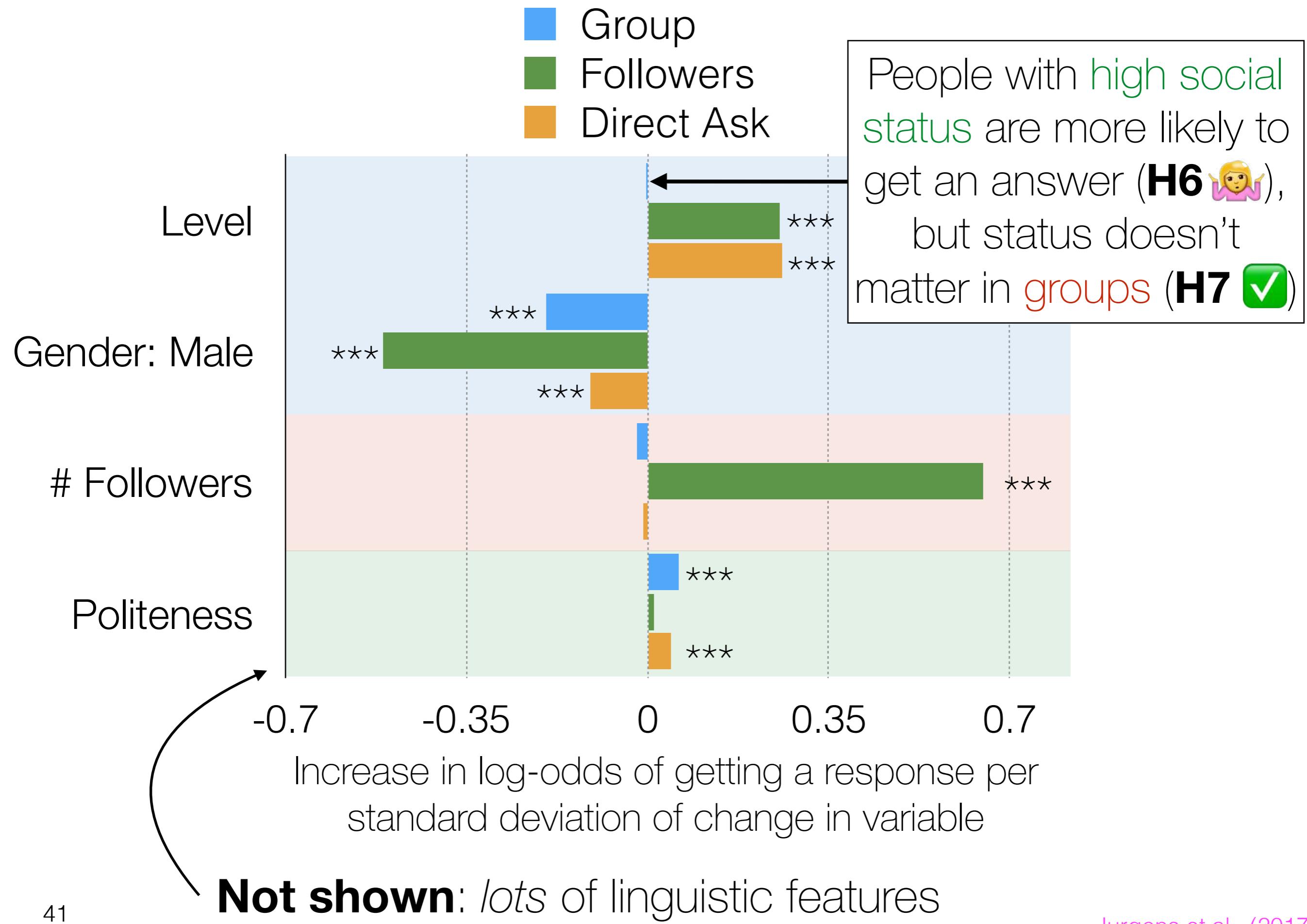
Test for social hypotheses by controlling for textual variation

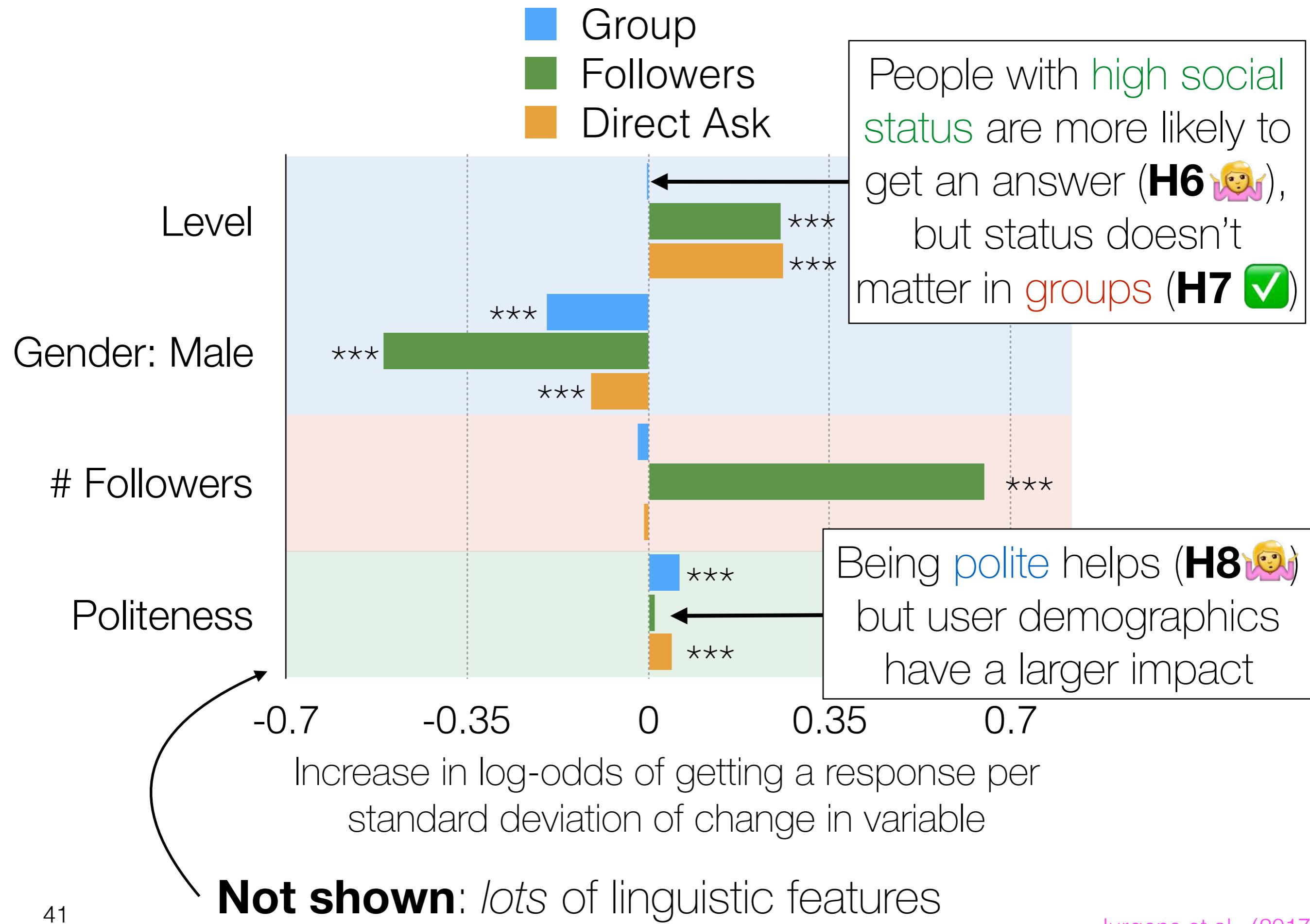




Increase in log-odds of getting a response per
standard deviation of change in variable







Group
 Followers
 Direct Ask

Level



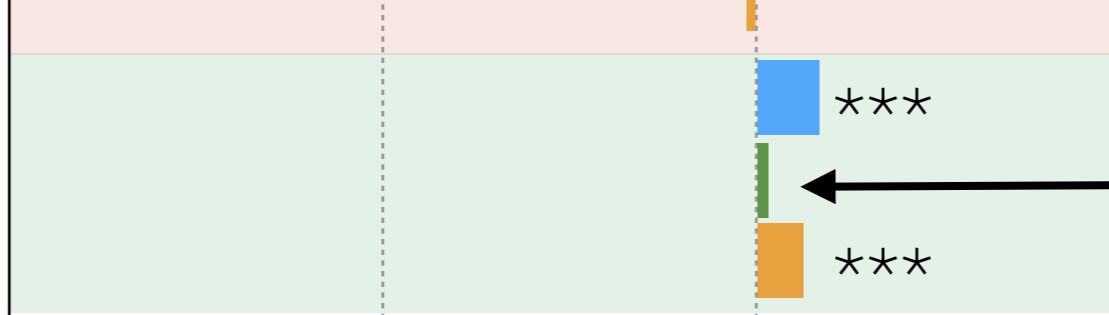
People with **high social status** are more likely to get an answer (**H6**

but status doesn't matter in **groups** (**H7**

Gender: Male



Followers



Politeness

Being **polite** helps (**H8**

but user demographics have a larger impact

Key Insight: Groups can serve an important role for answering the questions of new users

Not shown: lots of linguistic features

Social NLP

- Many different methods from NLP
 - Text-based classification, regression, sequence labeling (e.g., NER)
- Representation is important:
 - Bag of words
 - Features derived from parts of speech, syntax
- When testing theories, interpretable models are important.

Social NLP

Text also provides a lens into exploratory analysis of social and cultural phenomena



Digital Humanities

Digital Humanities

Digital Humanities

- Digital Humanities covers a range of methods that analyze text in literature and other cultural texts.

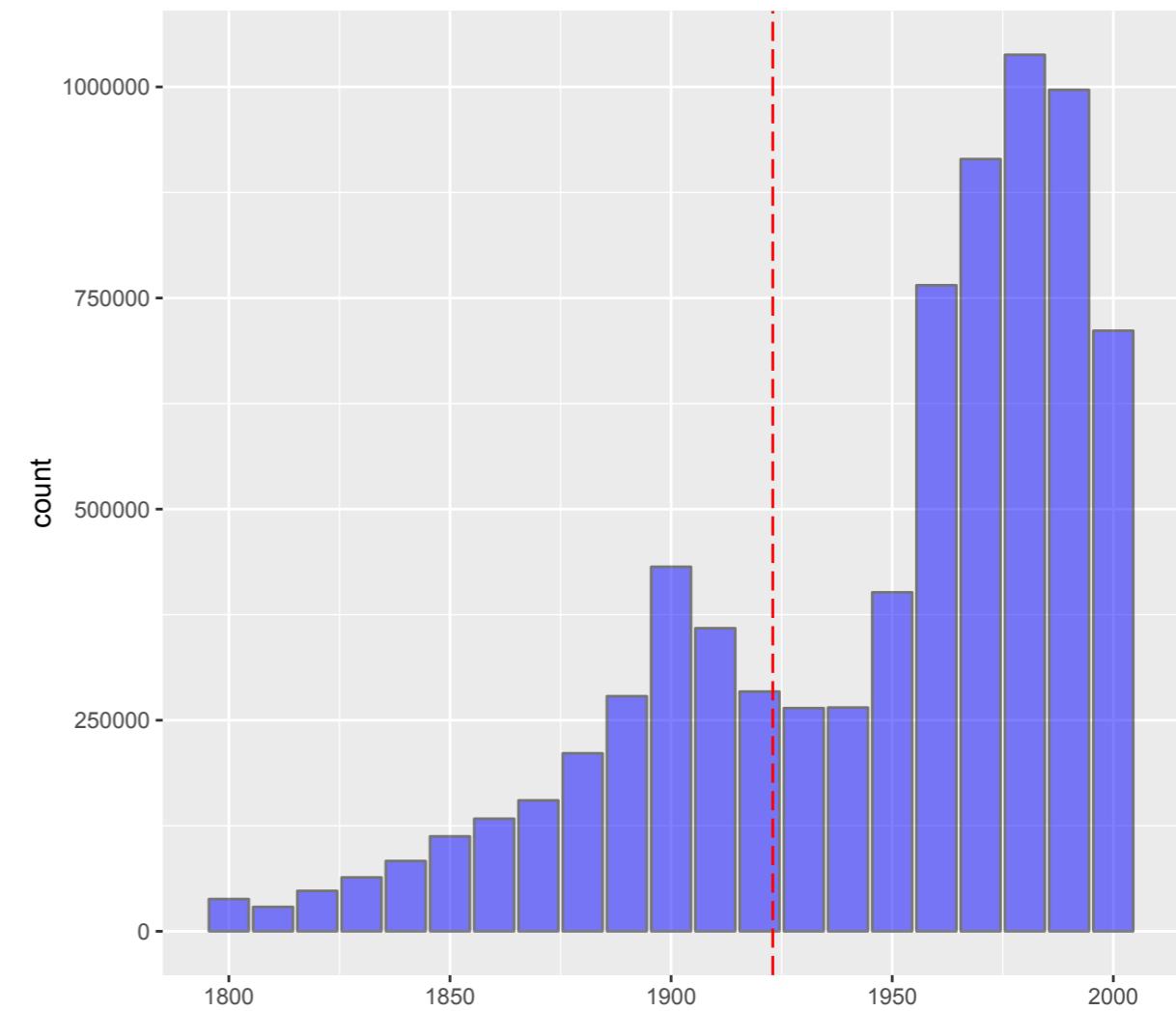
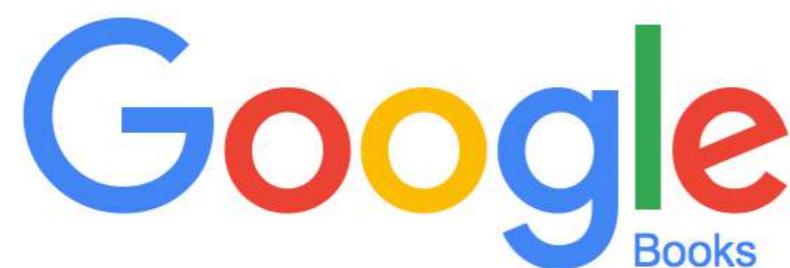
Digital Humanities

- Digital Humanities covers a range of methods that analyze text in literature and other cultural texts.
- Research questions focus on understanding social and cultural phenomena, often combining mixed-methods

Digital Humanities

- Digital Humanities covers a range of methods that analyze text in literature and other cultural texts.
- Research questions focus on understanding social and cultural phenomena, often combining mixed-methods
- How do we answer those research questions using methods we've learned about?
 - data
 - algorithms
 - evaluation

Data



HathiTrust: 14.8M books; 5.7M in the public domain

Digital Humanities

- Analyzing claims about genre
- Creating social networks
- Measuring changes in the size of society
- Quantifying gender disparity
- Uncovering narrative arcs

Genre

- Text: Books labeled {detective, gothic, science fiction} from bibliographies + random books
- Question: Are genres defined by shorter time periods more coherent as a **category** than those with longer lifetimes (e.g., detective fiction)?

Genre

- Bag of words representation of book
- Operationalize as a prediction task: coherence = high cross-validated accuracy.

Underwood 2016, “The Life Cycles of Genres”

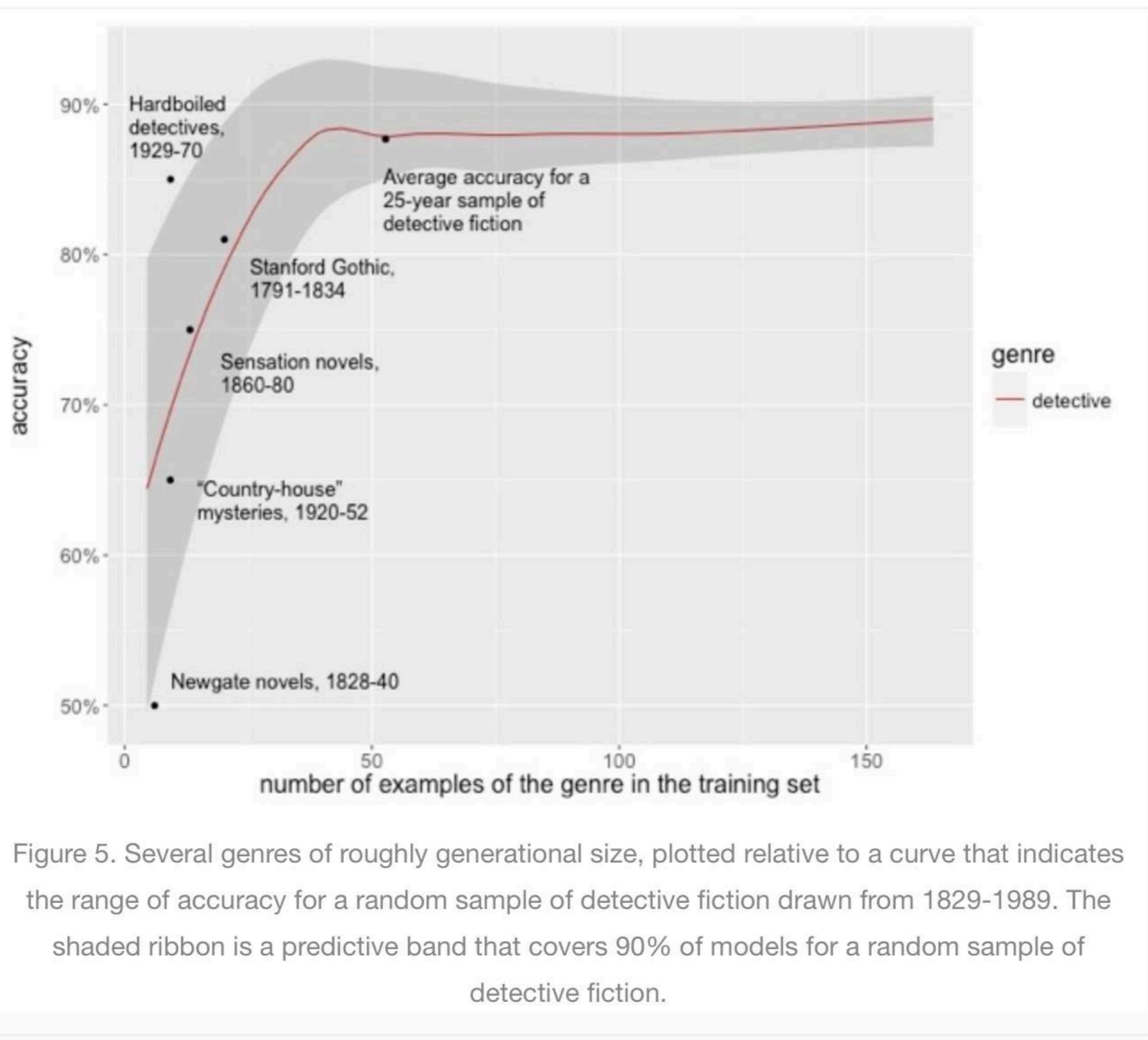


Figure 5. Several genres of roughly generational size, plotted relative to a curve that indicates the range of accuracy for a random sample of detective fiction drawn from 1829-1989. The shaded ribbon is a predictive band that covers 90% of models for a random sample of detective fiction.

Networks

Can we learn a **social network** from mentions in text?

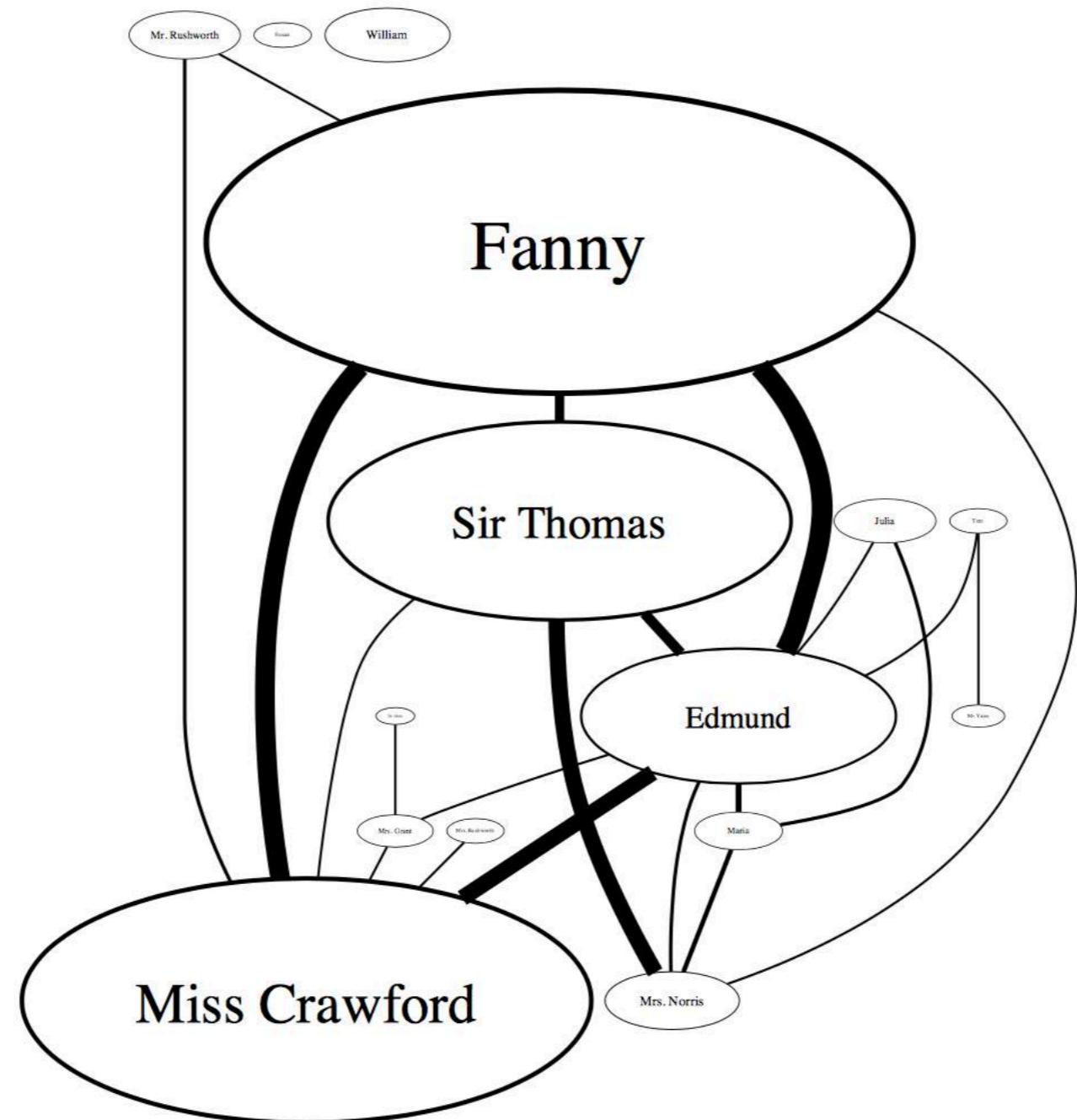
Networks

- Data: 60 novels from Project Gutenberg
- Conversational network:
 - Characters are in the same place at the same time
 - Characters take turns speaking
 - The characters are mutually aware of each other and each character's speech is mutually intended for the other to hear.

Networks

- Alias clustering: Tom, Tom Sawyer, Mr. Sawyer = TOM SAWYER)
- Quoted speech attribution (“Yes,” said TOM SAWYER)
- Network construction
 - Divide book into 10-paragraph sections, count number of sections with two characters
 - Count occurrences of one character mentioning another in dialogue

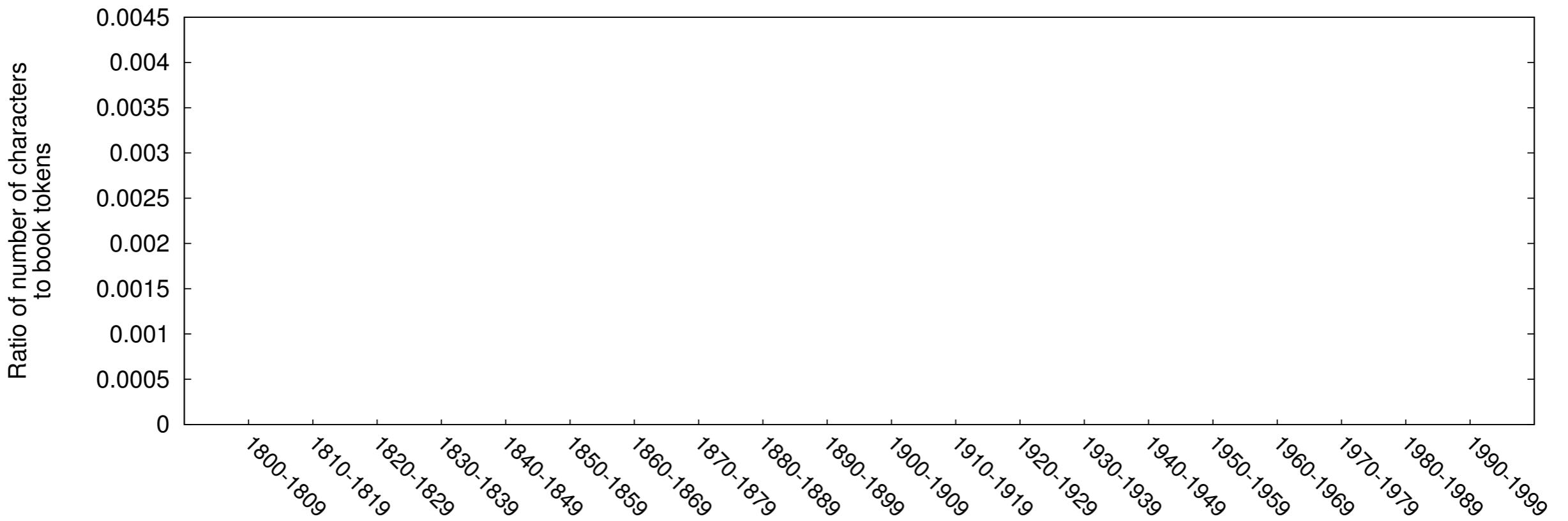
Networks



Literature is a lens on society at a point in time

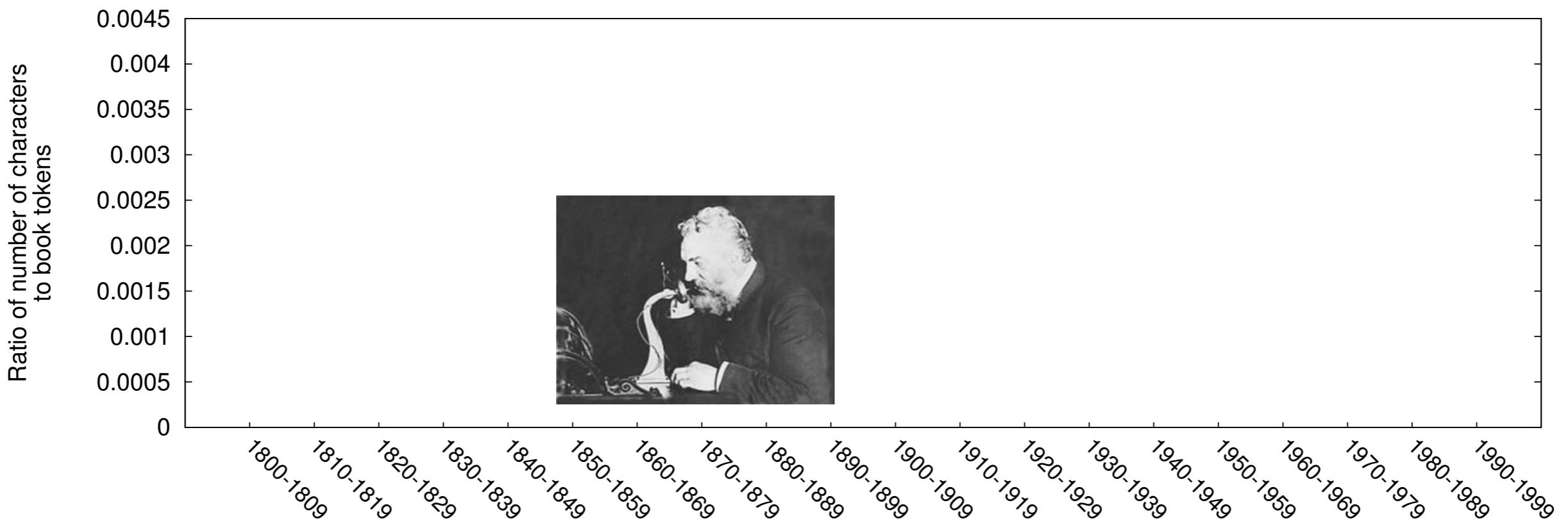
- Can we learn whether changes to daily life (technology, wars) have changed society?

New technologies have the ability to shape how we connect with others



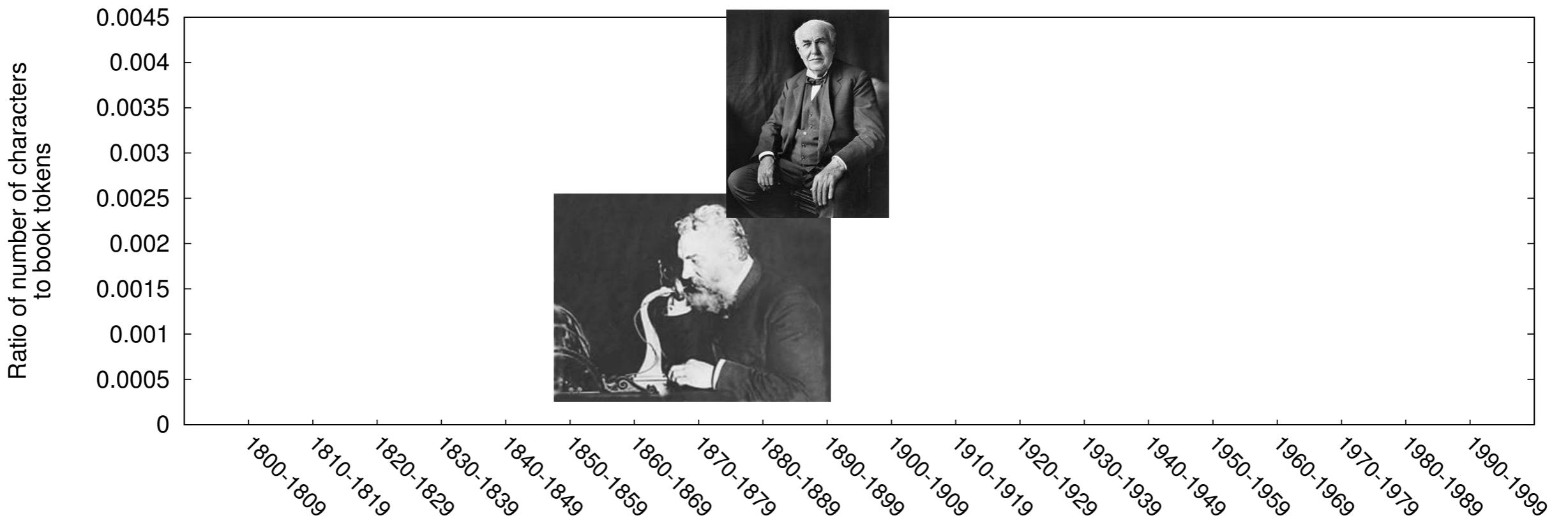
- Does the introduction of new technology change our ability to have a **larger social circle**?

New technologies have the ability to shape how we connect with others



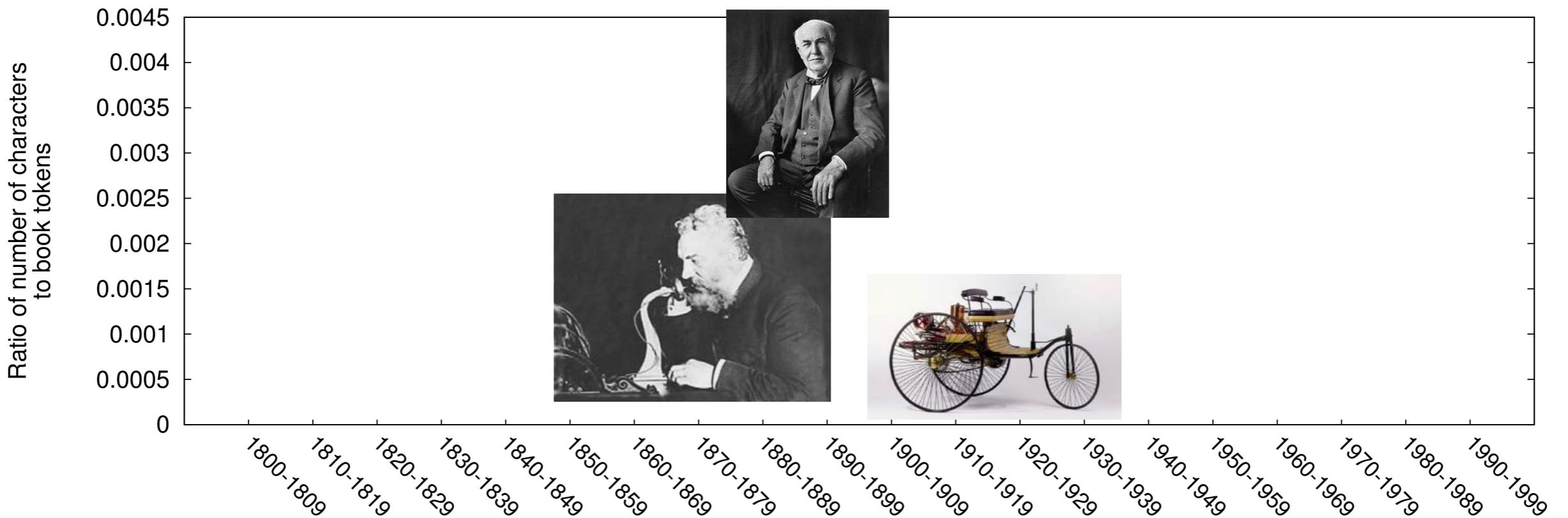
- Does the introduction of new technology change our ability to have a **larger social circle**?

New technologies have the ability to shape how we connect with others



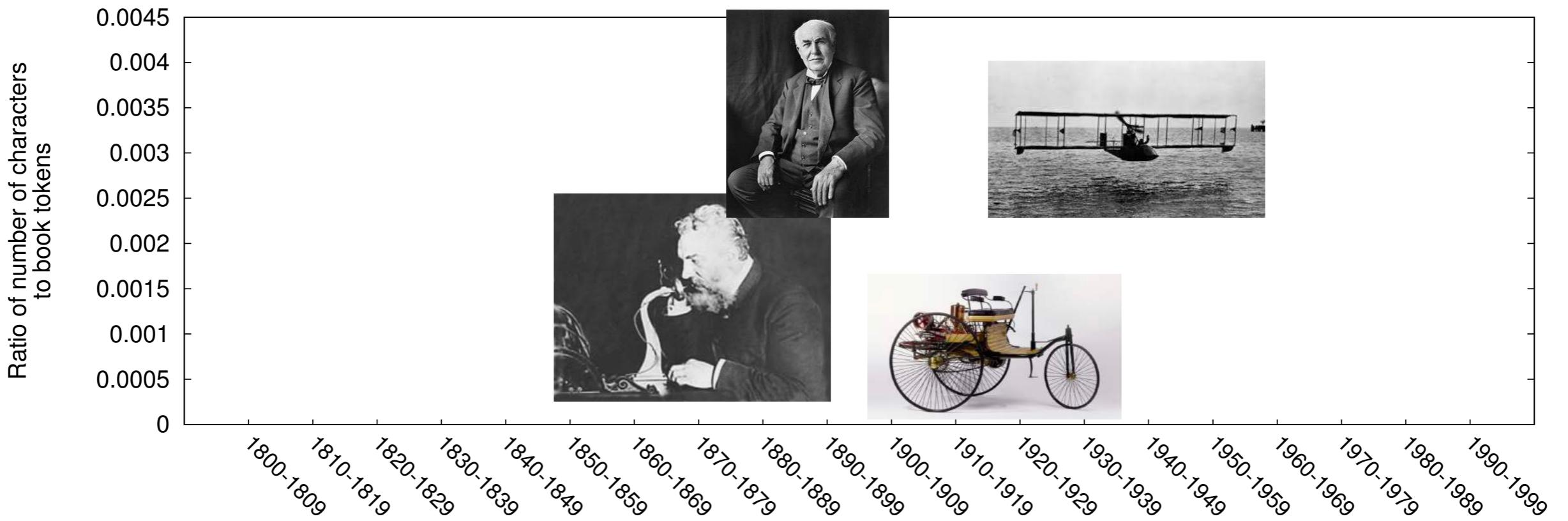
- Does the introduction of new technology change our ability to have a **larger social circle**?

New technologies have the ability to shape how we connect with others



- Does the introduction of new technology change our ability to have a **larger social circle**?

New technologies have the ability to shape how we connect with others



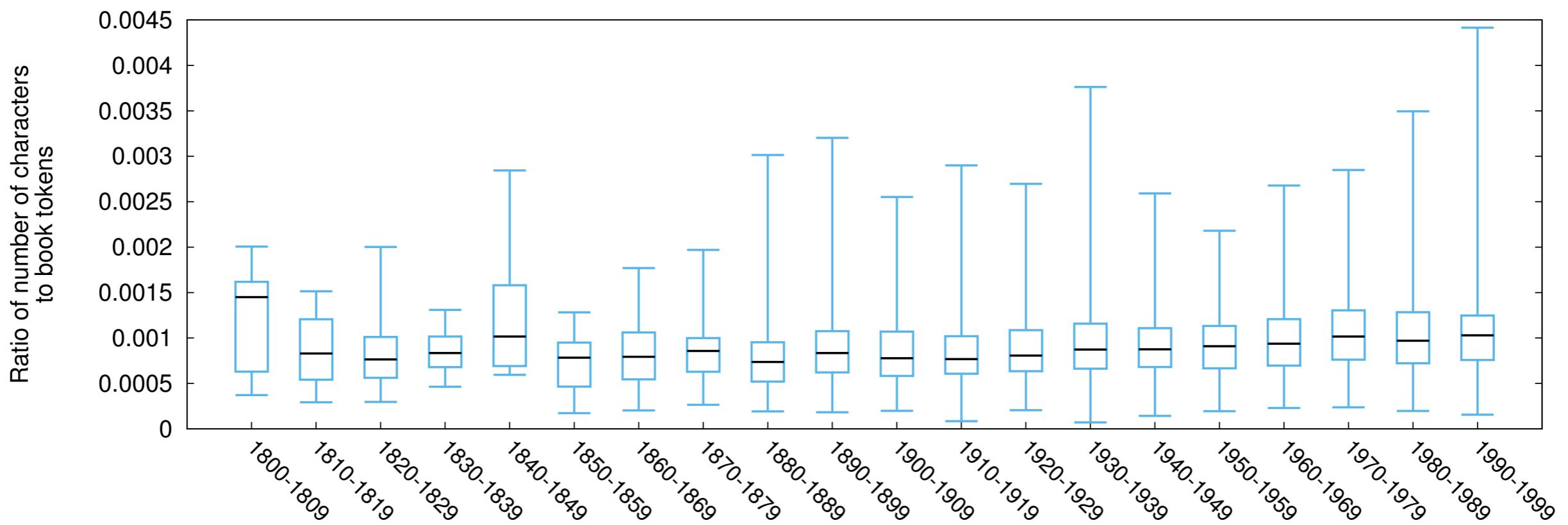
- Does the introduction of new technology change our ability to have a **larger social circle**?

How many people are described in a book?

How many people are described in a book?

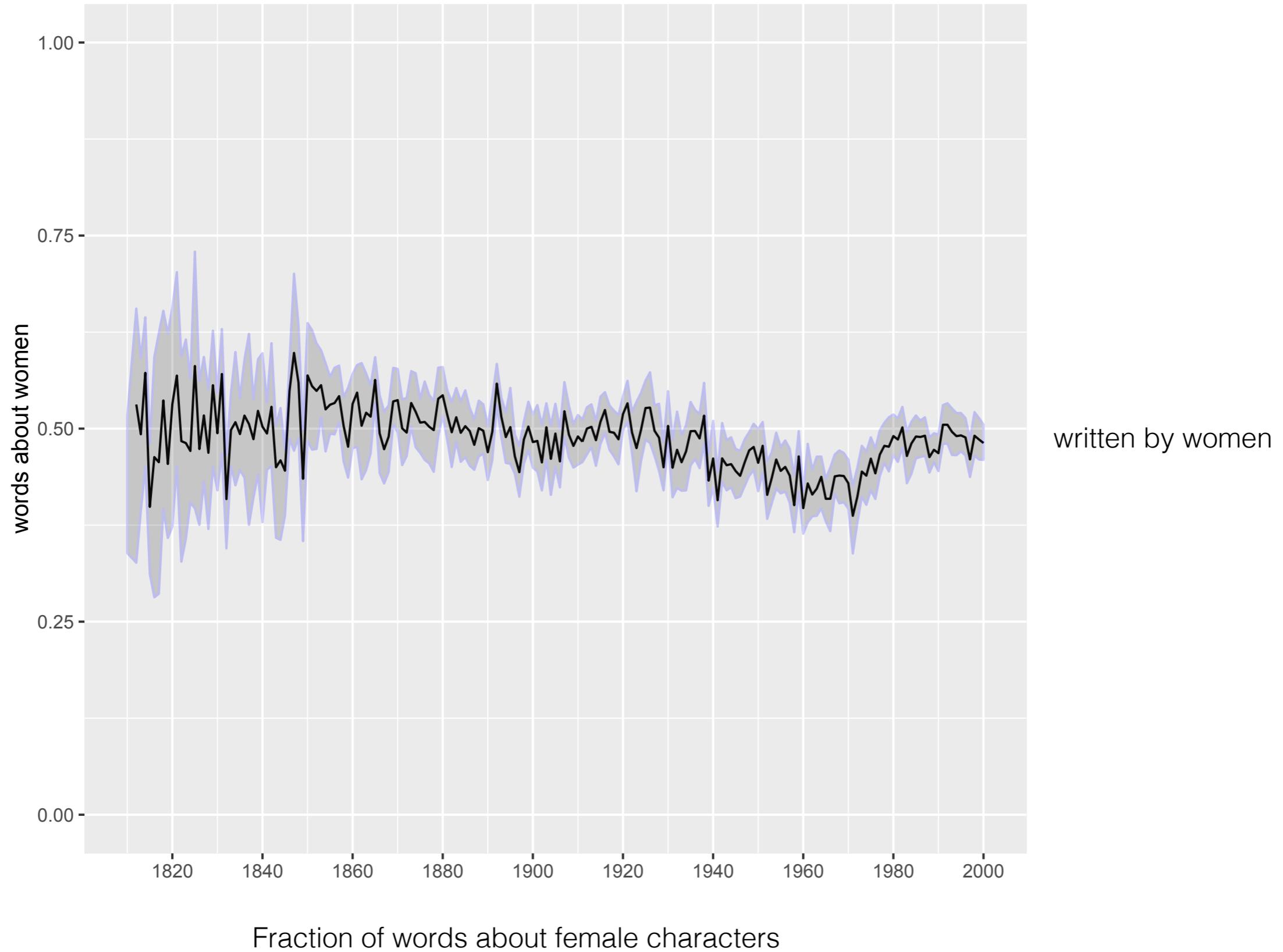
“Between him and Darcy there was a very steady friendship, in spite of great opposition of character. Bingley was endeared to Darcy by the easiness, openness, and ductility of his temper, though no disposition could offer a greater contrast to his own, and though with his own he never appeared dissatisfied. On the strength of Darcy's regard, Bingley had the firmest reliance, and of his judgement the highest opinion. In understanding, Darcy was the superior. Bingley was by no means deficient, but Darcy was clever. He was at the same time haughty, reserved, and fastidious, and his manners, though well-bred, were not inviting. In that respect his friend had greatly the advantage. Bingley was sure of being liked wherever he appeared, Darcy was continually giving offense.”

New technologies and other large societal events did little to change how big our societies get

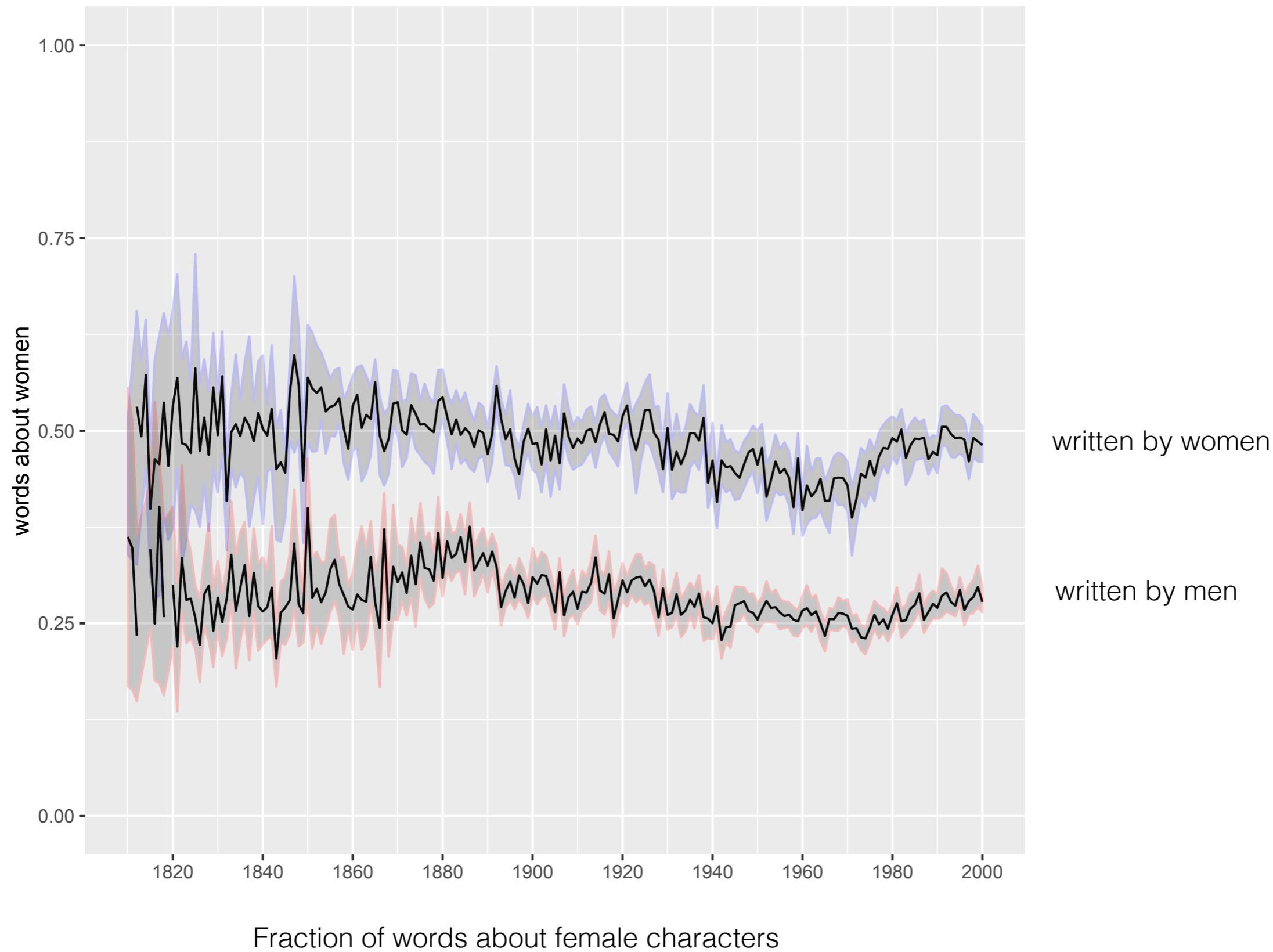


Gender in Literature

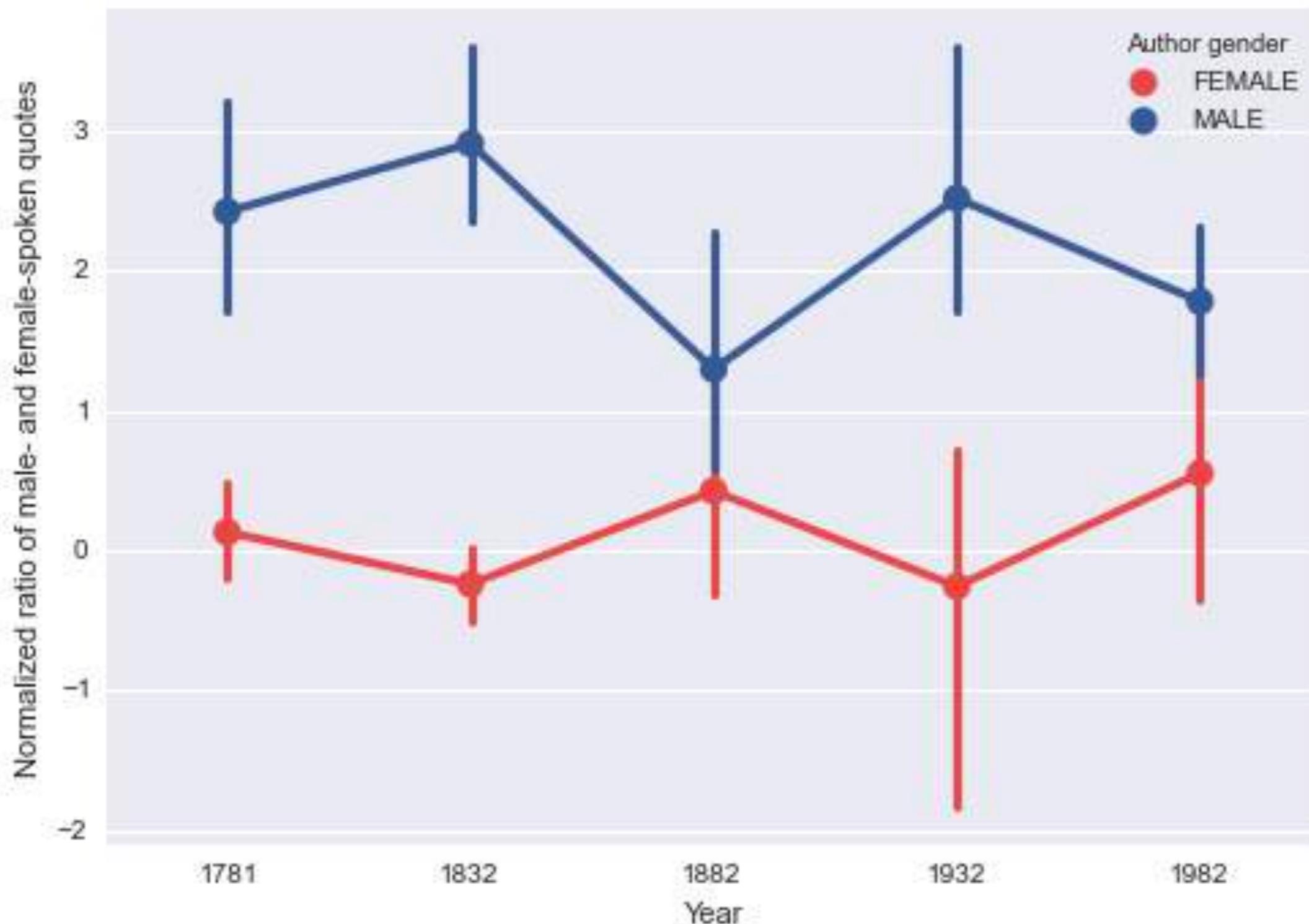
- How are men and women depicted in literature?



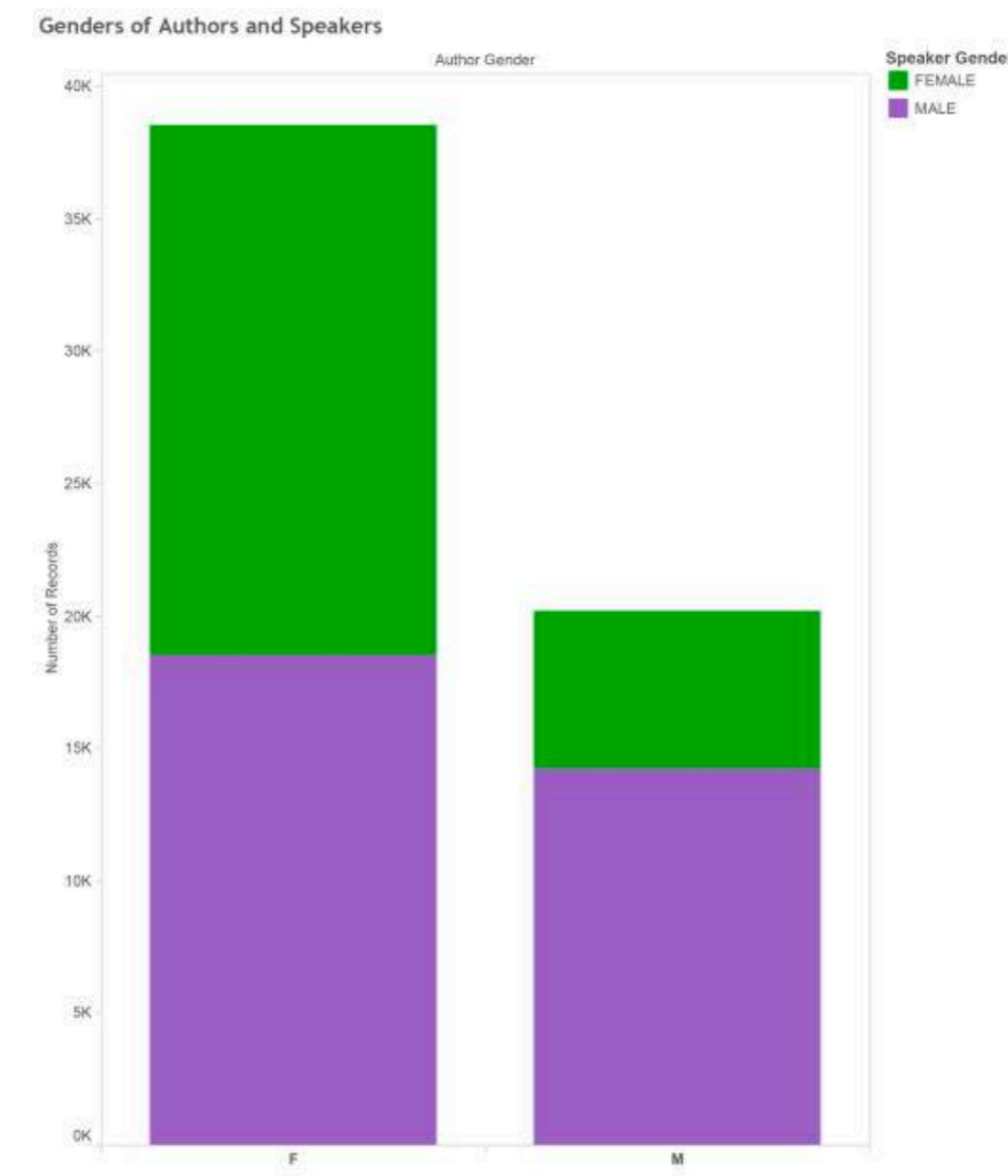
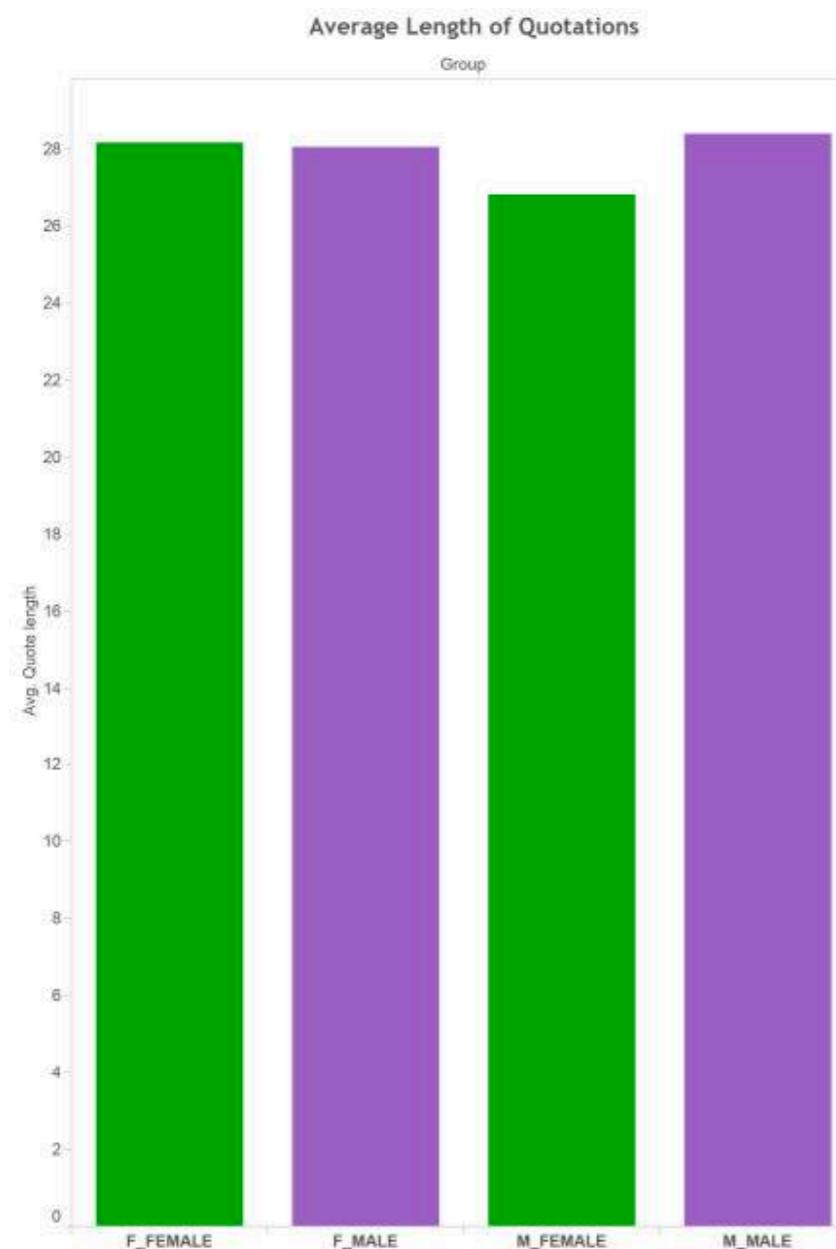
Ted Underwood and David Bamman (2016), “The Instability of Gender” (MLA);
“The Gender Balance of Fiction” (2017).



Ted Underwood and David Bamman (2016), “The Instability of Gender” (MLA);
“The Gender Balance of Fiction” (2017).

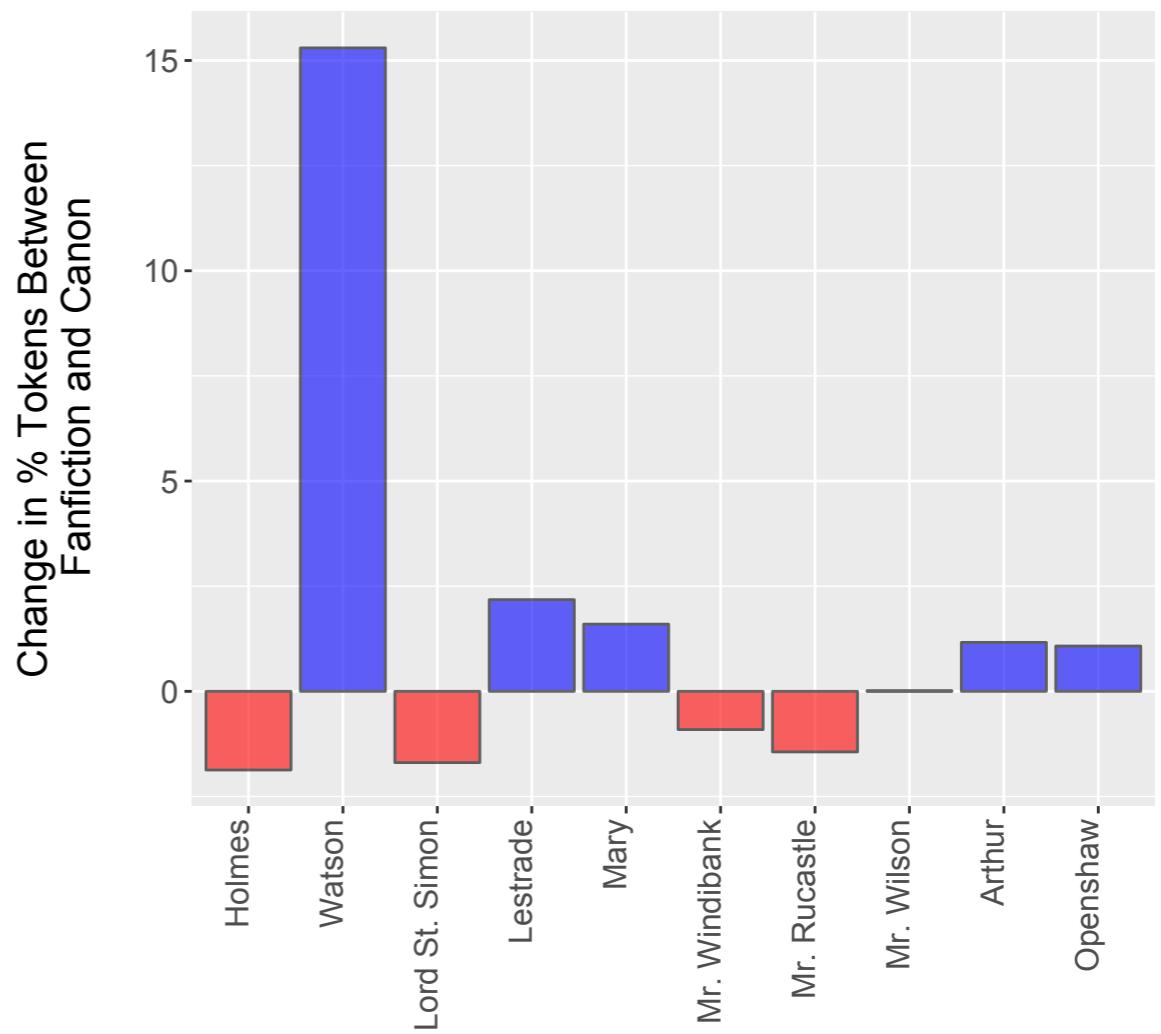


Women speak less frequently and less content



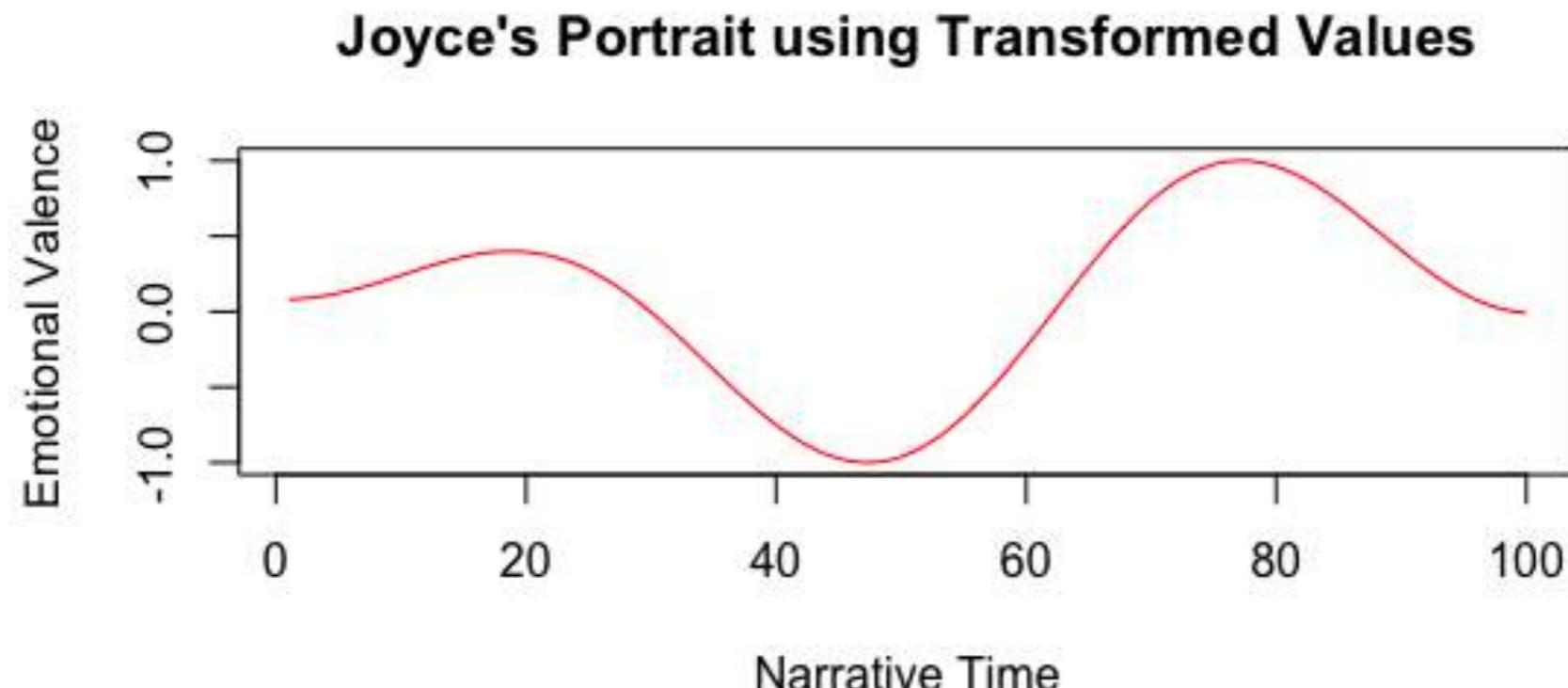
Fanfiction

Fanfiction allocates less attention to main protagonists and more attention to female characters



Computational Narratology

- Syuzhet is an employment of narrative
- Fabula is the chronological order of the events contained in the story
- How does the emotional arc of a book play out over time?



Digital Humanities

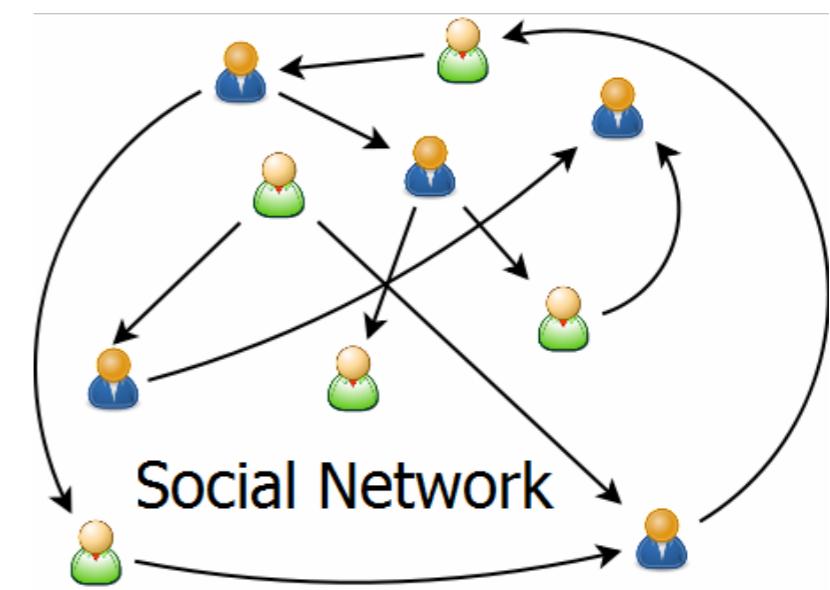
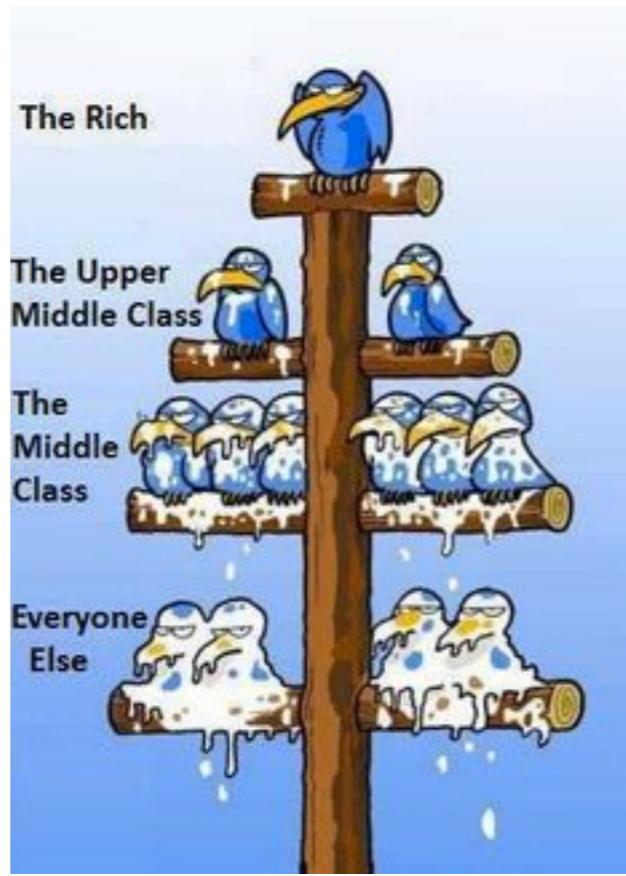
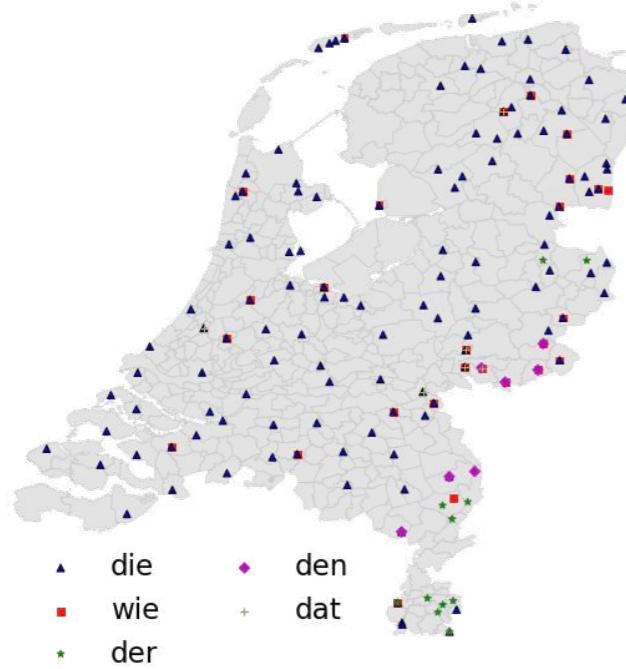
- NLP provides a computational tools to study social and cultural phenomena in literature
- Many fundamental questions about literature are unanswerable without infeasible human effort (e.g., how many people are in a book?)
 - Many opportunities for new insights!
- Strongly tied to theories — sometimes centuries old!



Computational Sociolinguistics

Sociolinguistics

Sociolinguistics is the descriptive study of the effect of any and all aspects of society, including cultural norms, expectations, and context, on the way language is used, and the effects of language use on society. (Wikipedia)



Variationist Sociolinguistics

- The most quantitative strand in sociolinguistics
- VARBRUL (logistic regression)
- Famous study of variation in New York City :
 - Labov (1966, 1972) asked sales people about products on the "fourth floor" (natural context instead of lab)
 - Postvocalic (r)
 - Three department stores (different social status rankings):
 - Saks Fifth Avenue (62%)
 - Macy's (51%)
 - Klein (21%)

Background reading

- Introductory books:
 - The guidebook to sociolinguistics, Bell 2014
 - An introduction to sociolinguistics, Holmes 2013
 - Introducing sociolinguistics, Meyerhoff, 2011
- Journals: Journal of Sociolinguistics, Language in Society, Language Variation and Change, etc..

Traditional data sources in the social sciences



Surveys



Observation

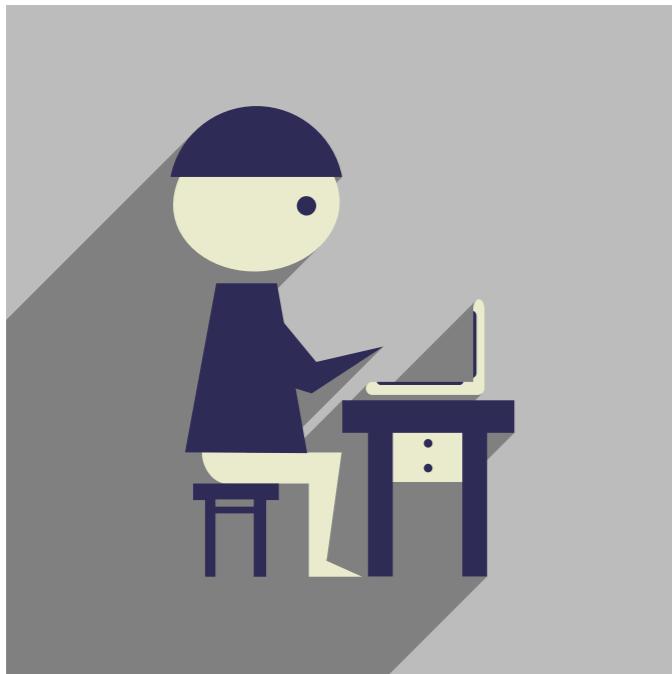


Interviews

Time consuming 😞

Observer's Paradox 😞 – Labov (1972): "the aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain these data by systematic observation".

Big social and cultural data



Big social and cultural data

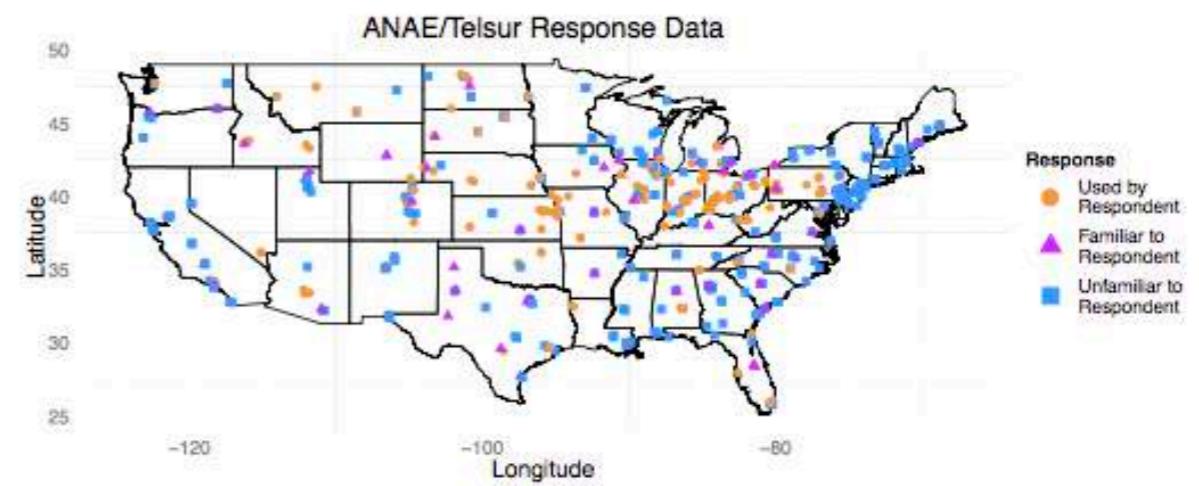
- Informal
- Large amounts of data
- Interaction patterns
- Over time
- Multimodal

Motivation

- NLP tools for processing informal language
- Testing and refining theories using large-scale naturalistic text data
- New analysis tools for sociolinguists and social scientists
- More fine-grained analyses of online behavior (e.g., user profiling, participation in campaigns)

Testing sociolinguistic theories

- Doyle (2014) analyzed the geographical distribution of dialectal variants based on Twitter data
- Compared with traditional sociolinguistic data collection methods.
- Achieved high correlations with data from sociolinguistic studies.



(a) ANAE/Telsur survey responses for *need+past participle*.



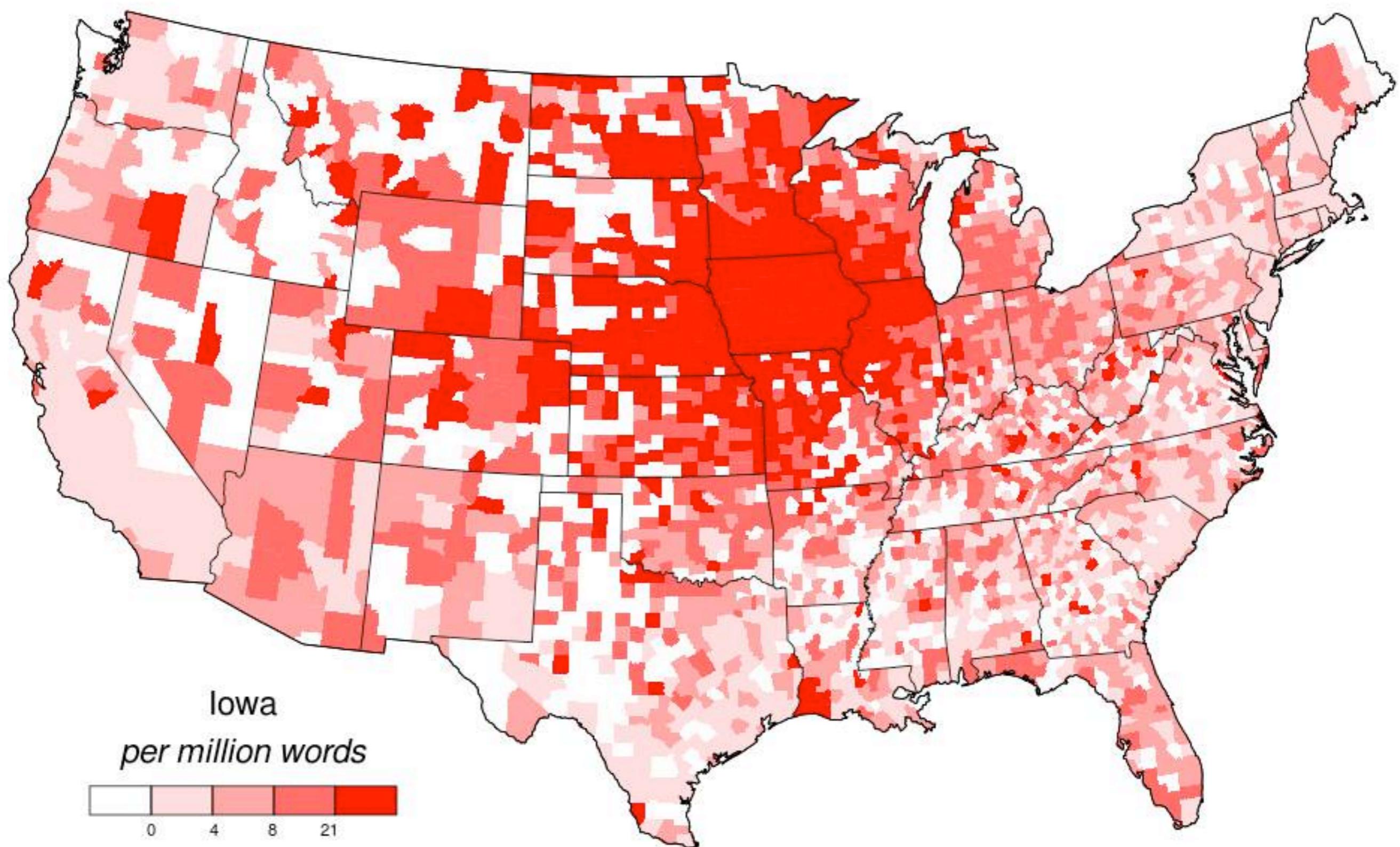
(b) SeeTweet search for “needs done”.

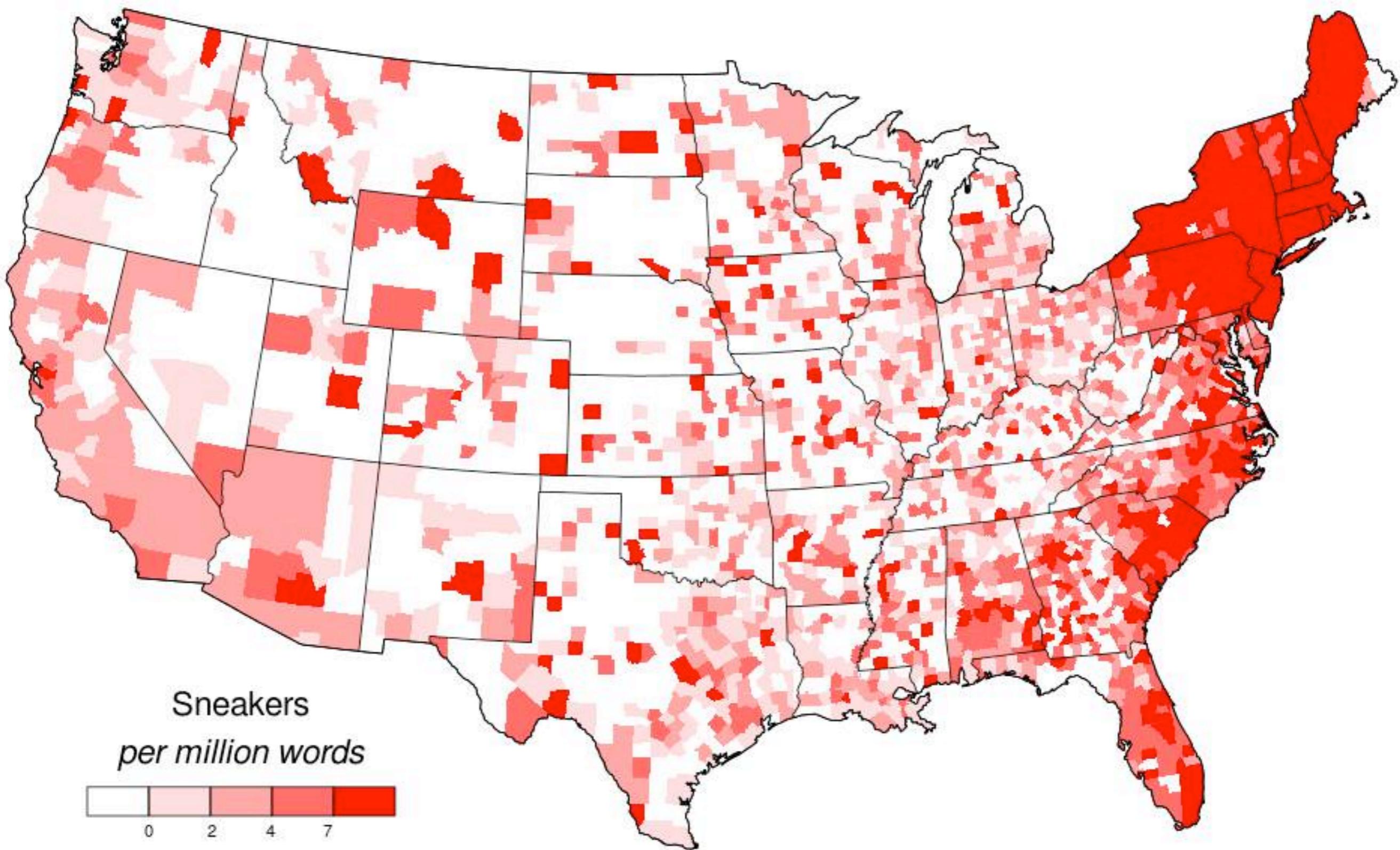
<https://www.worldtravelguide.net/wp-content/uploads/2017/04/Think-Switzerland-Country-Zermatt-Matterhorn-486574518-extravagantni-copy.jpg>

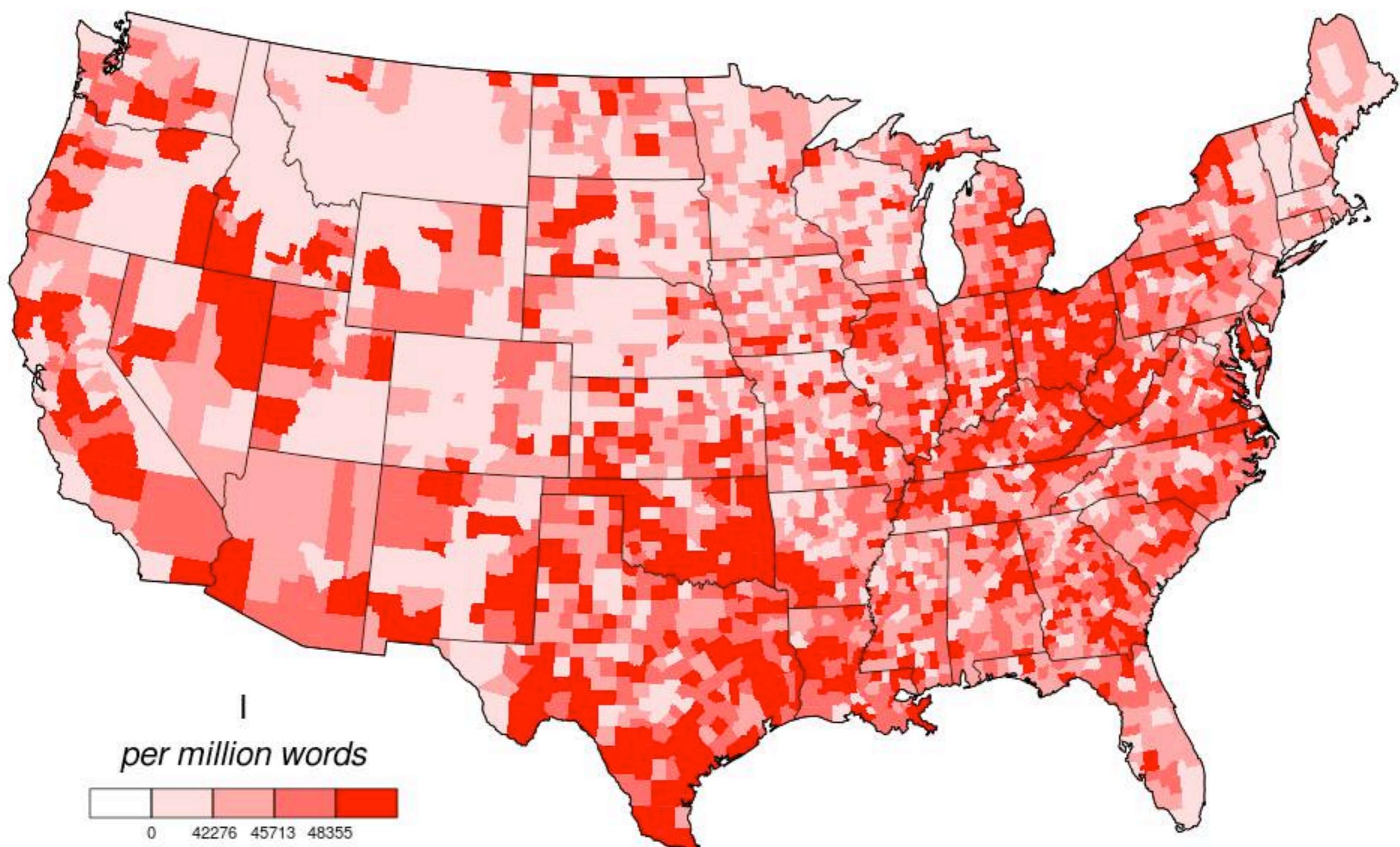


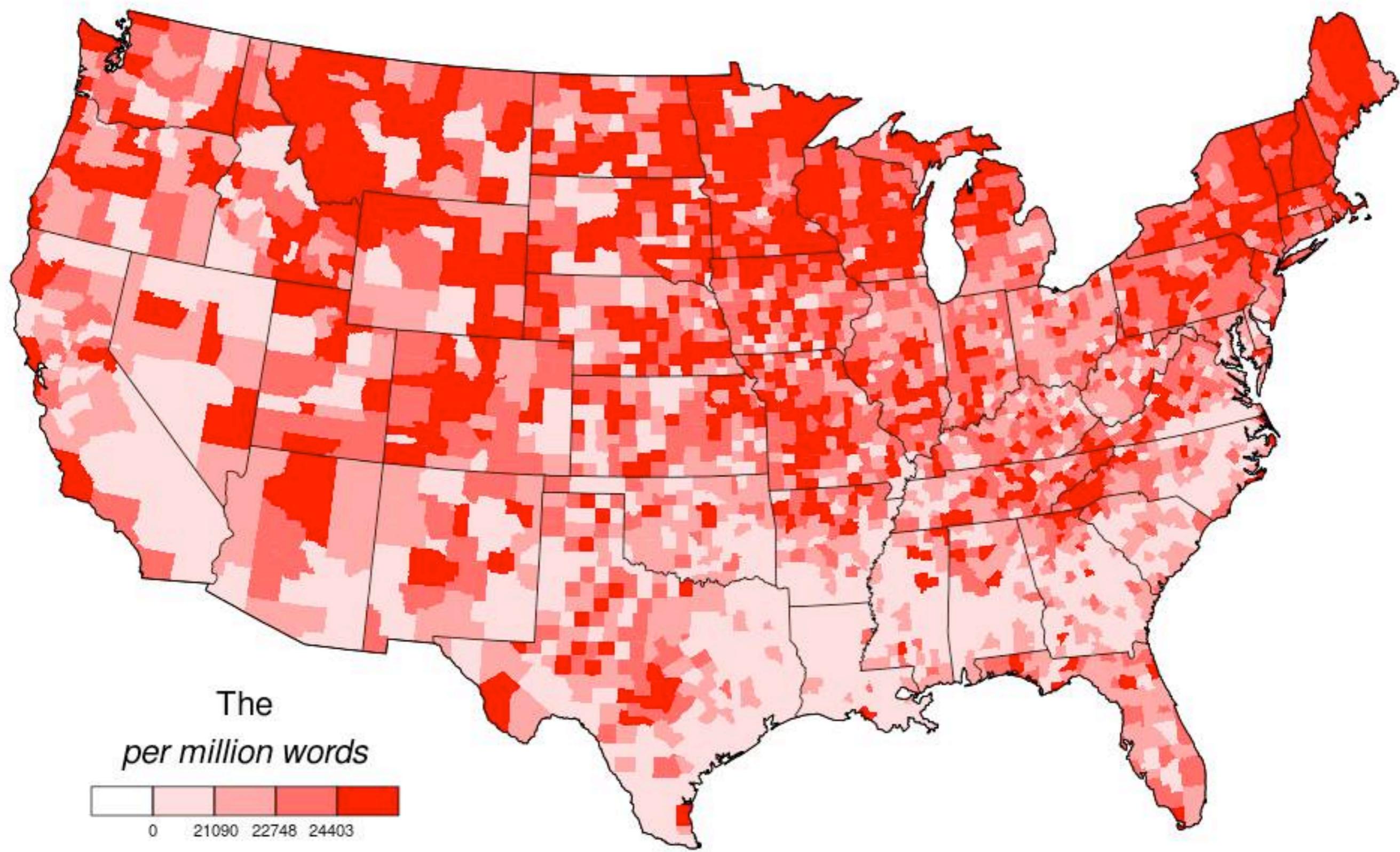
Are word uses regional?

- To investigate whether word frequencies tend to show regional patterns or are distributed at random, all 10,000 maps were subjected to a Moran's I global spatial autocorrelation analysis.
- All 10,000 words show significant levels of global spatial autocorrelation ($p < 0.000005$, Bonferroni corrected).









Isogloss analysis

- In dialectology, common patterns of regional variation are traditionally found following a two step procedure:
 1. Regional patterns are identified in maps for individual variables by drawing **isoglosses**.
 2. Common patterns of regional variation are identified by searching for **bundles of isoglosses**.

Using Spatial Analysis Tools for Linguistic Analysis (Dialectology)

Using Spatial Analysis Tools for Linguistic Analysis (Dialectology)

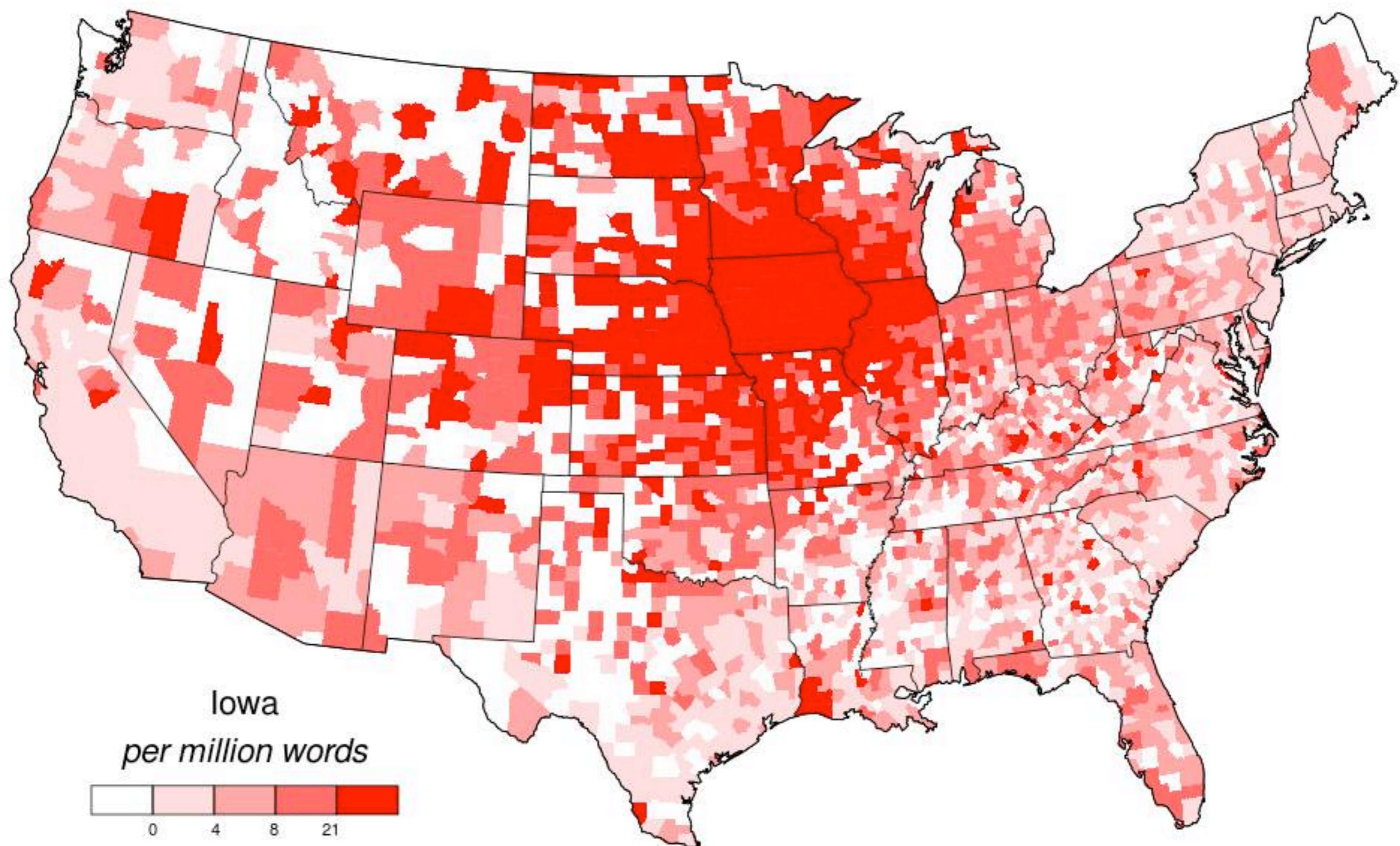
- To identify common patterns of regional variation in topic, the 8,973 content word maps were analyzed following a two step statistical analysis.

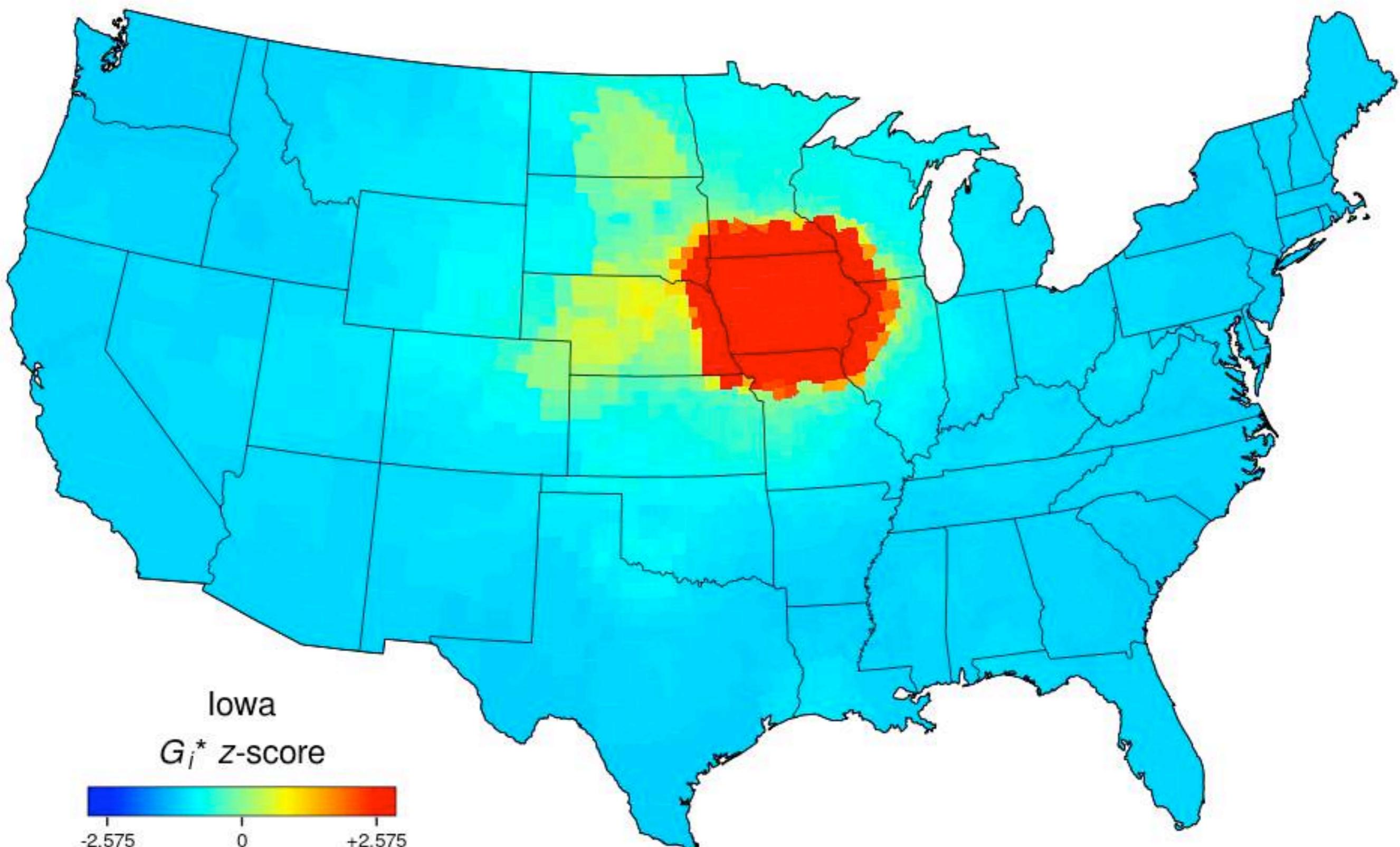
Using Spatial Analysis Tools for Linguistic Analysis (Dialectology)

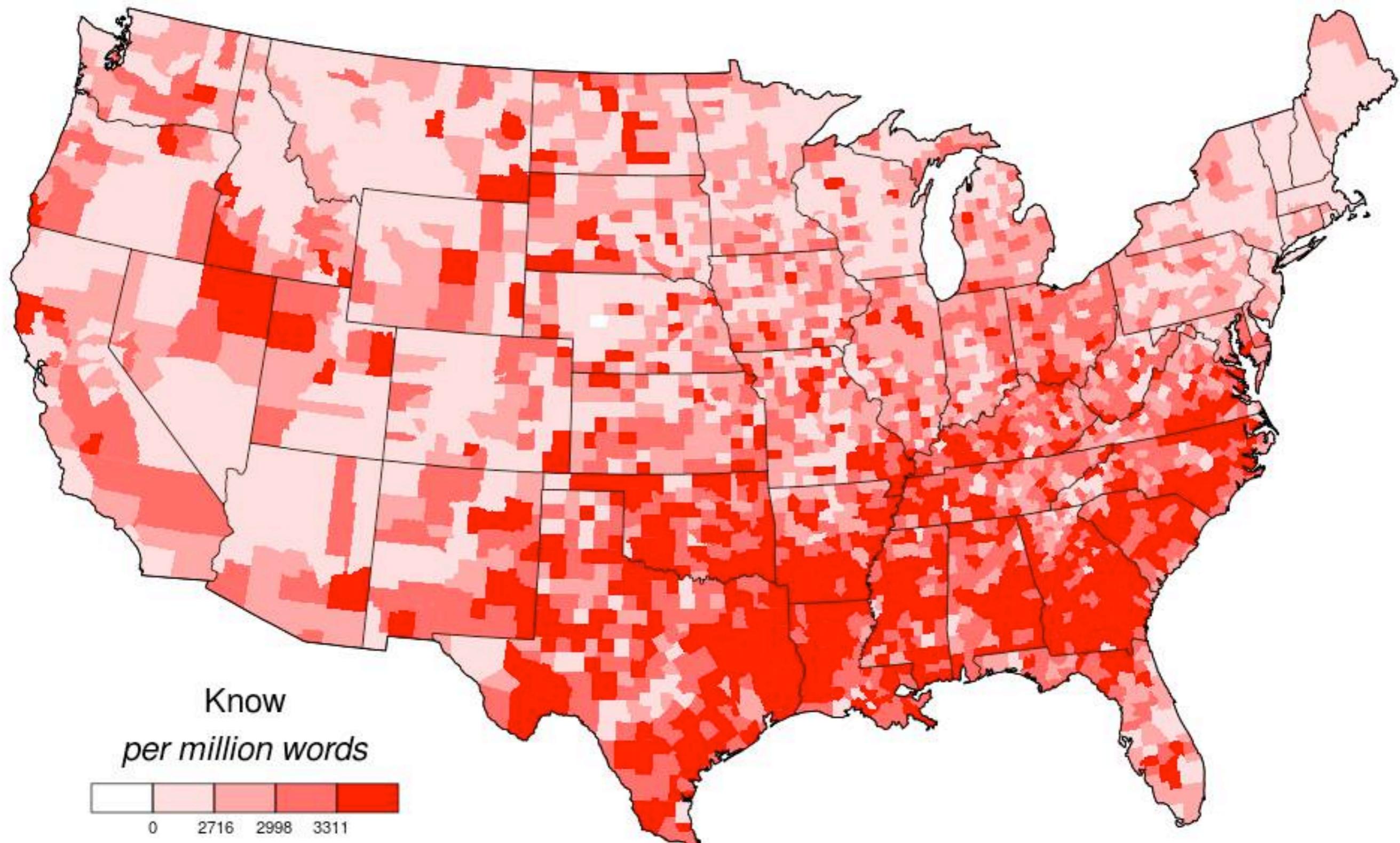
- To identify common patterns of regional variation in topic, the 8,973 content word maps were analyzed following a two step statistical analysis.
 1. Isoglosses: Getis-Ord Gi* Local Spatial Autocorrelation analysis to smooth frequency maps.

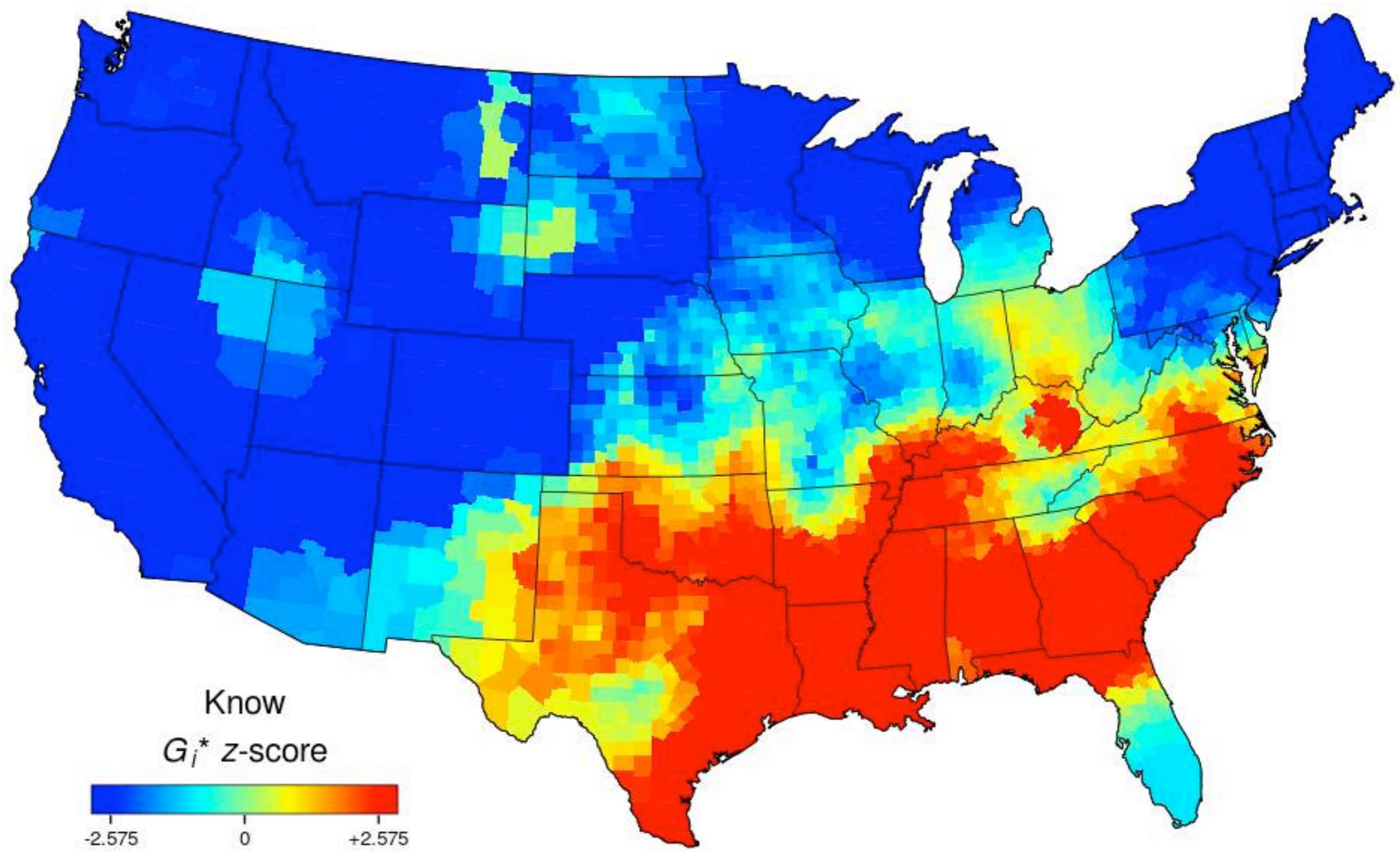
Using Spatial Analysis Tools for Linguistic Analysis (Dialectology)

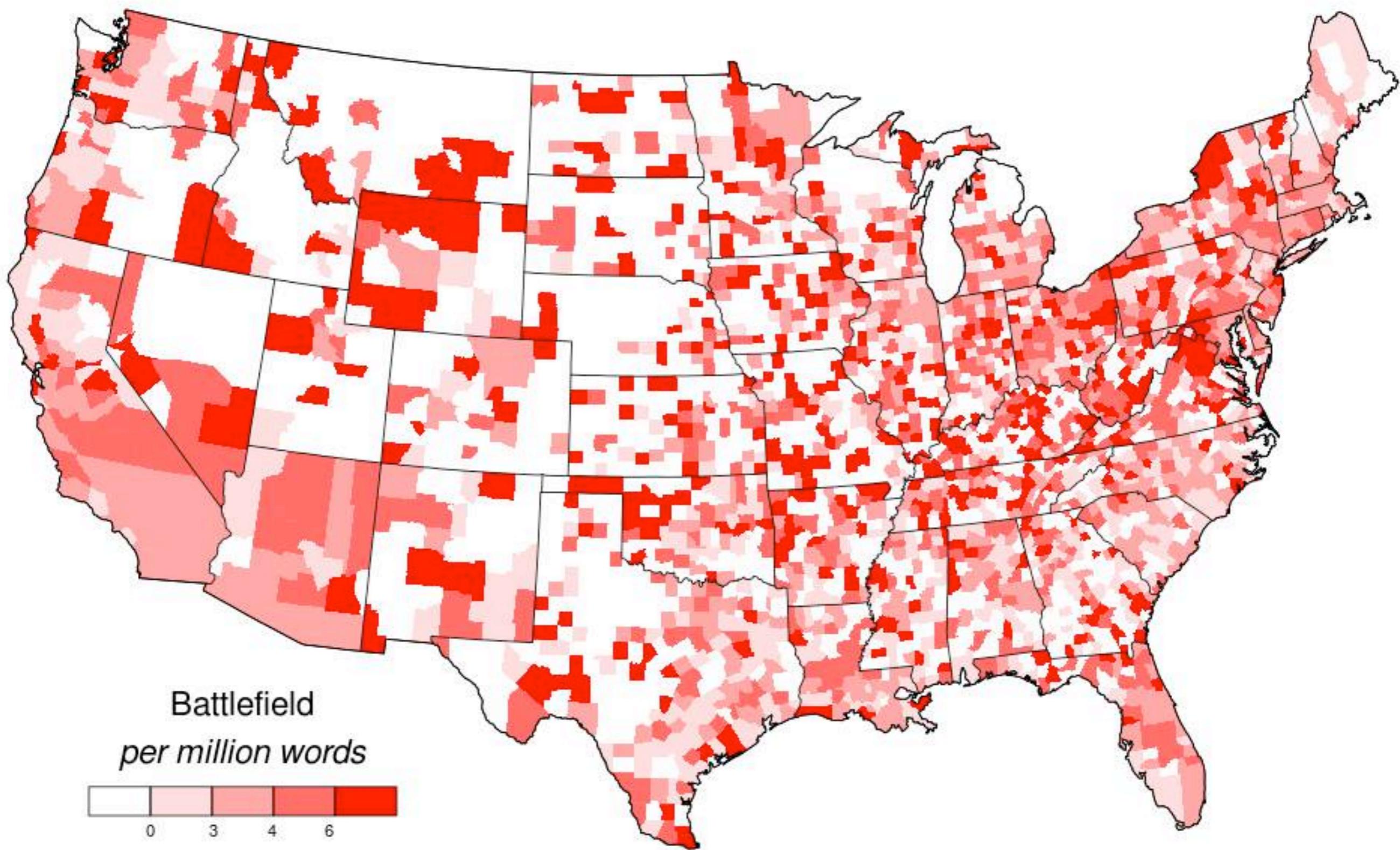
- To identify common patterns of regional variation in topic, the 8,973 content word maps were analyzed following a two step statistical analysis.
 1. Isoglosses: Getis-Ord Gi* Local Spatial Autocorrelation analysis to smooth frequency maps.
 2. Bundles of isoglosses: Principal Component Analysis of smoothed maps.

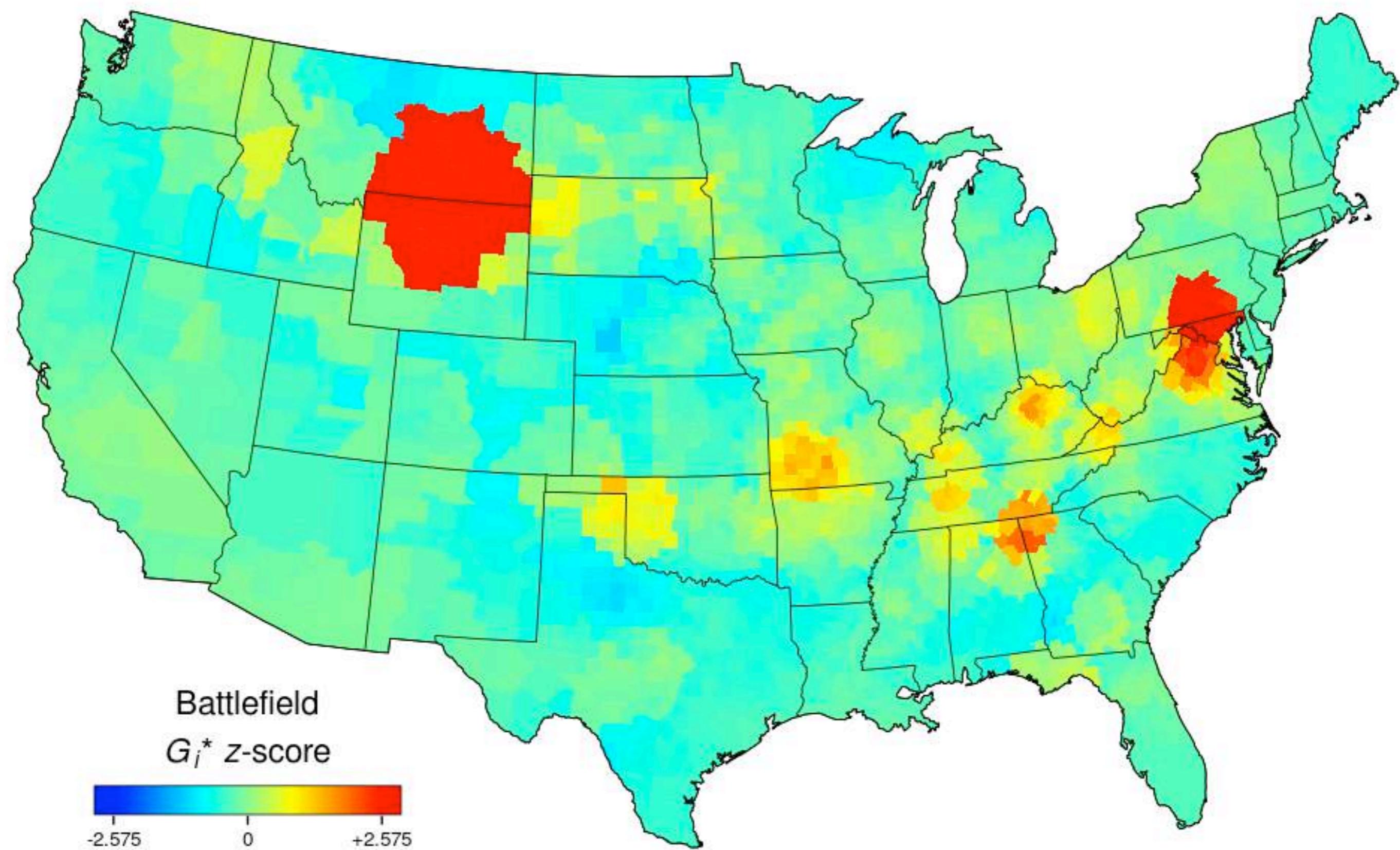






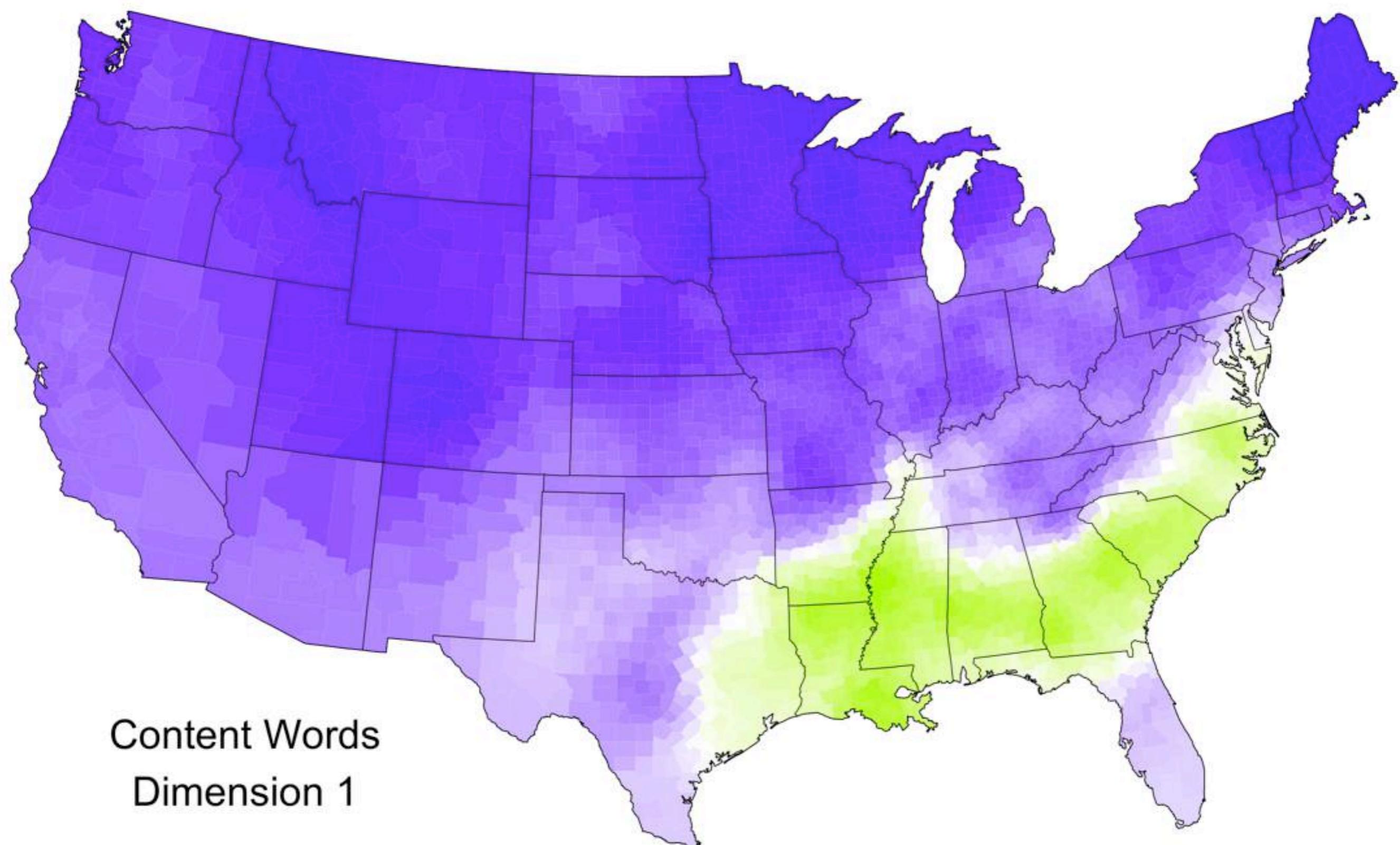




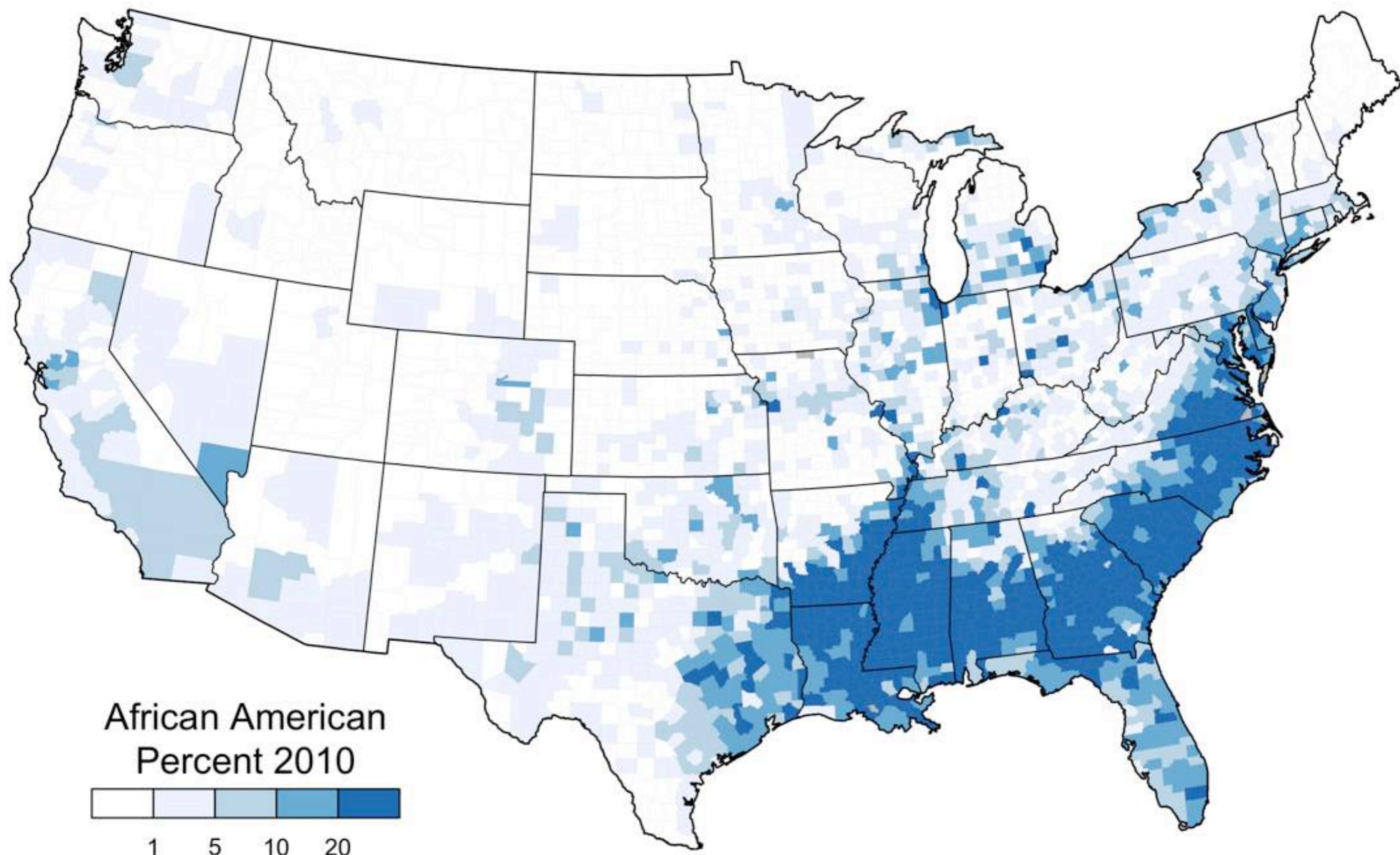


Using Spatial Analysis Tools for Linguistic Analysis (Dialectology)

- To identify common patterns of regional variation in topic, the 8,973 content word maps were analyzed following a two step statistical analysis.
 1. Isoglosses: Getis-Ord Gi* Local Spatial Autocorrelation analysis to smooth frequency maps.
 2. Bundles of isoglosses: Principal Component Analysis of smoothed maps.
- **Let's look at the first few**



Content Words
Dimension 1



Souteast: Social Interaction

Communication: *argue, arguing, claim, claiming, convo, cuss, diss, dissing, hush, lied, lien, lies, lyin, lying, ...*

Interpersonal Interaction: *act, actin, acting, aggravating, balling, beefing, exposing, flex, flexin, flexing, ...*

Family and Friends: *bestfriend, cuh, cuhh, cuzzo, daddy, dawg, fam, granny, homie, mama, moma, ...*

Sex and Relationships: *ass, azz, baby, babygirl, bae, baee, bitch, bitches, cheating, chick, cuff, cuffing, ...*

Rest and Relaxation: *chillin, chilling, coolin, cooling, crib, gn, goodmorning, goodnight, kickback, schleep, ...*

.

Non-Southeast: Evaluation + Home Leisure

Quantities: *amount, amounts, bunch, biggest, ended, ending, final, first, five, four, giant, hour, hours, ...*

Positive Qualities: *adorable, awesome, clever, coolest, easier, enjoyable, excited, exciting, fantastic, ...*

Negative Qualities: *awkward, creepy, disappointed, embarrassed, frustrating, inappropriate, insane, odd, ...*

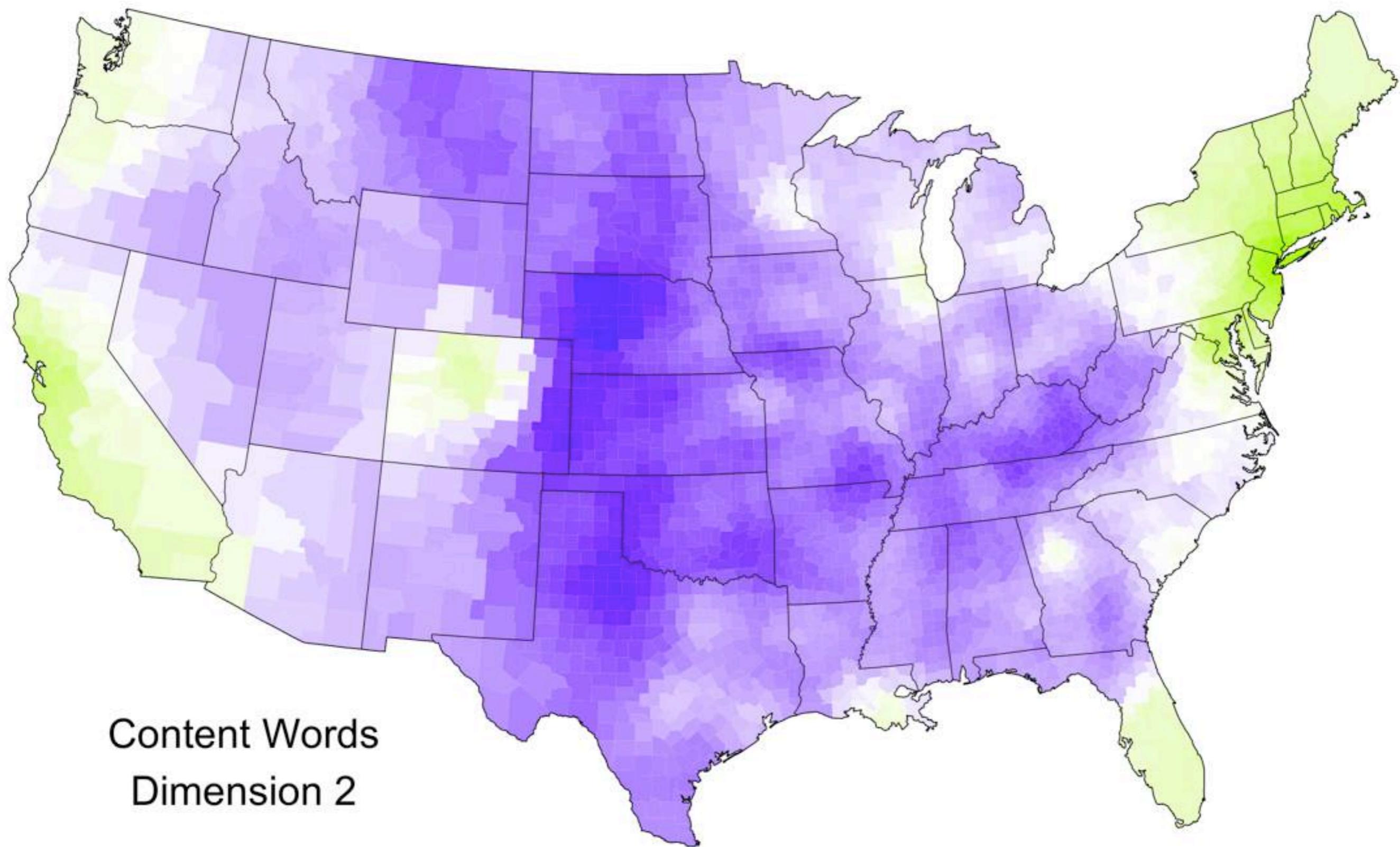
Stance: *accidentally, actually, apparently, awkwardly, definitely, easily, incredibly, insanely, legitimately, ...*

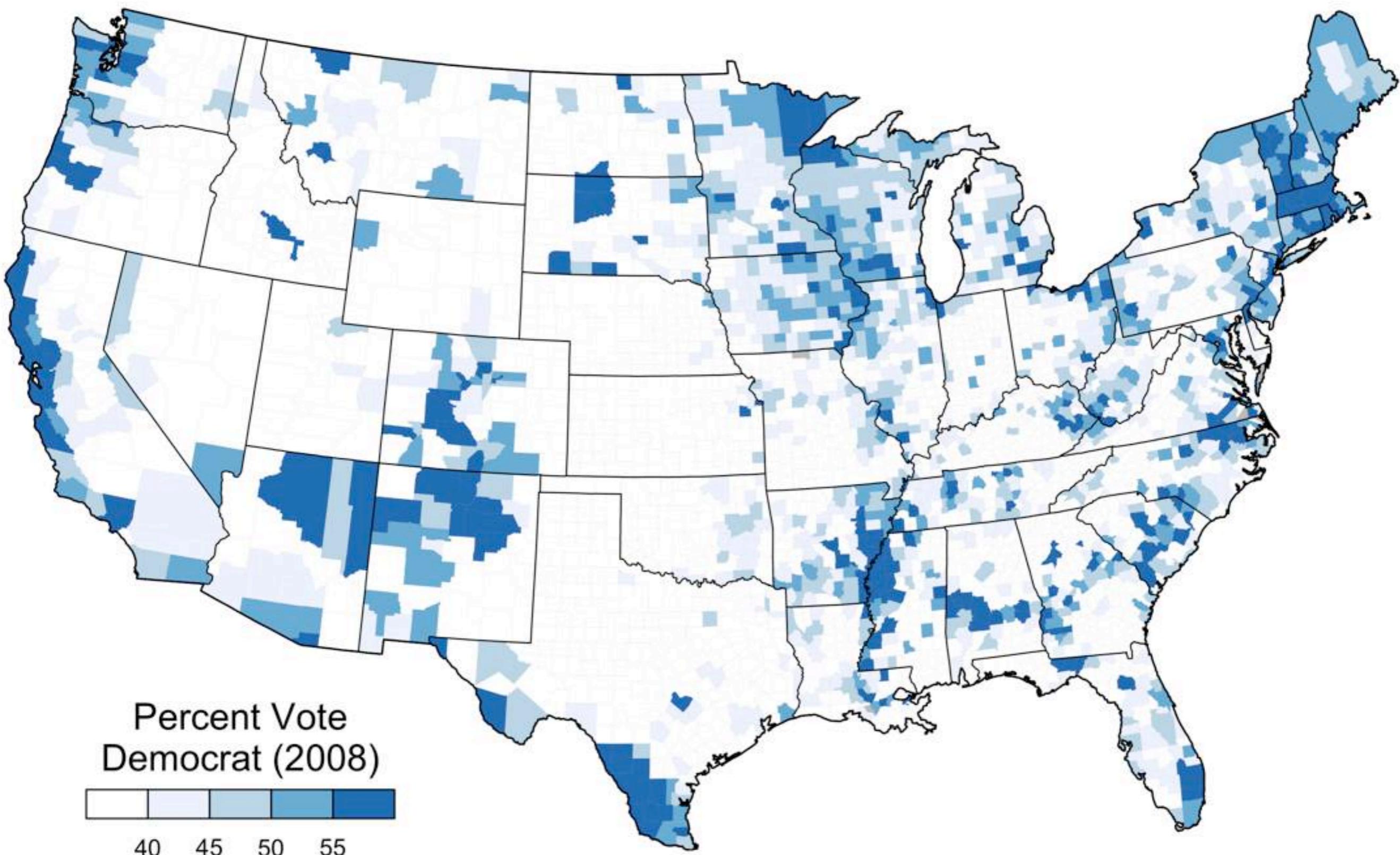
Pets and Animals: *animal, animals, bunny, cat, dog, dogs, kitten, kitty, pet, pup, puppies, spider, spiders*

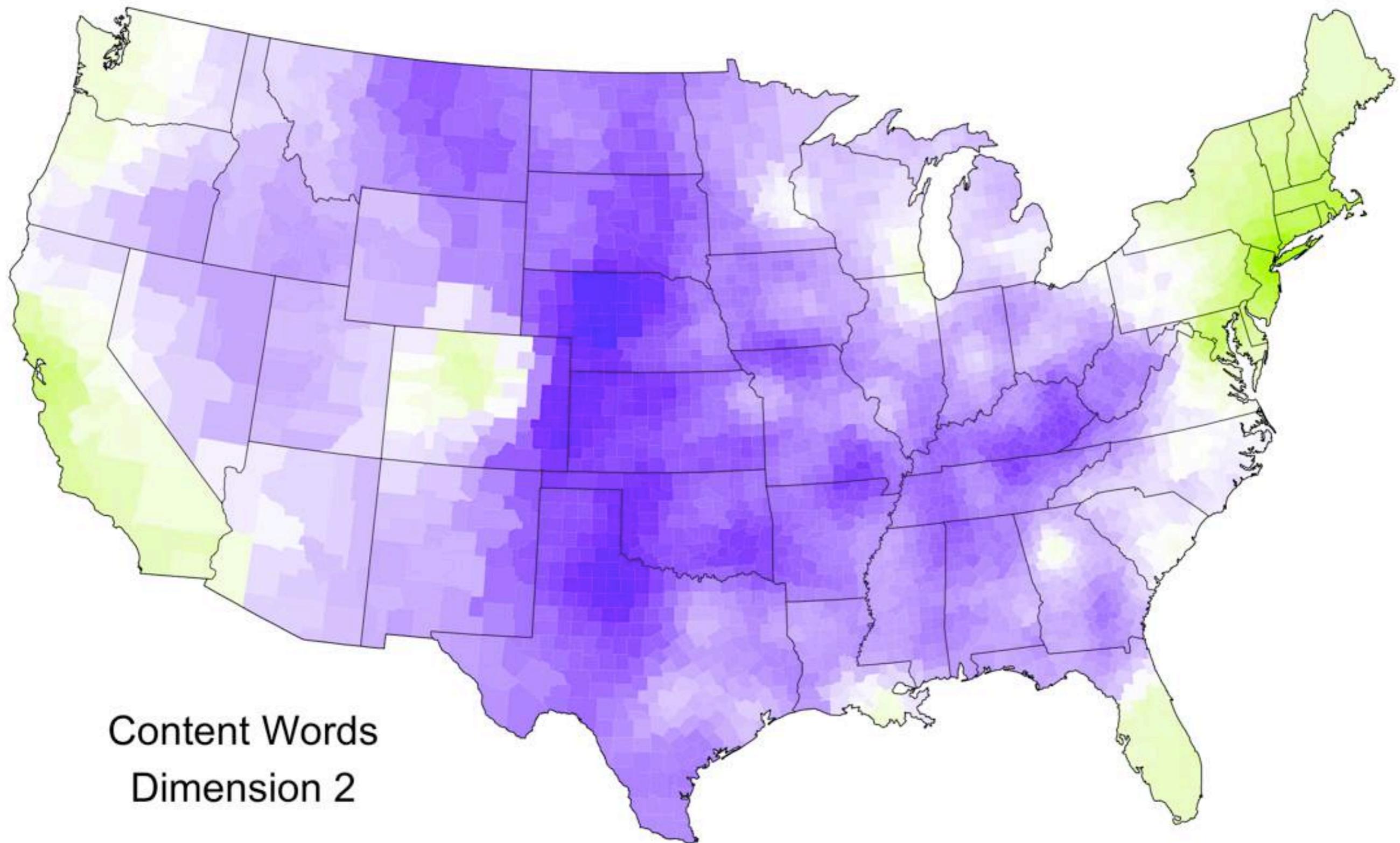
Media and Entertainment: *anchorman, book, concerts, hobbit, internet, potter, reading, stars, story, watched*

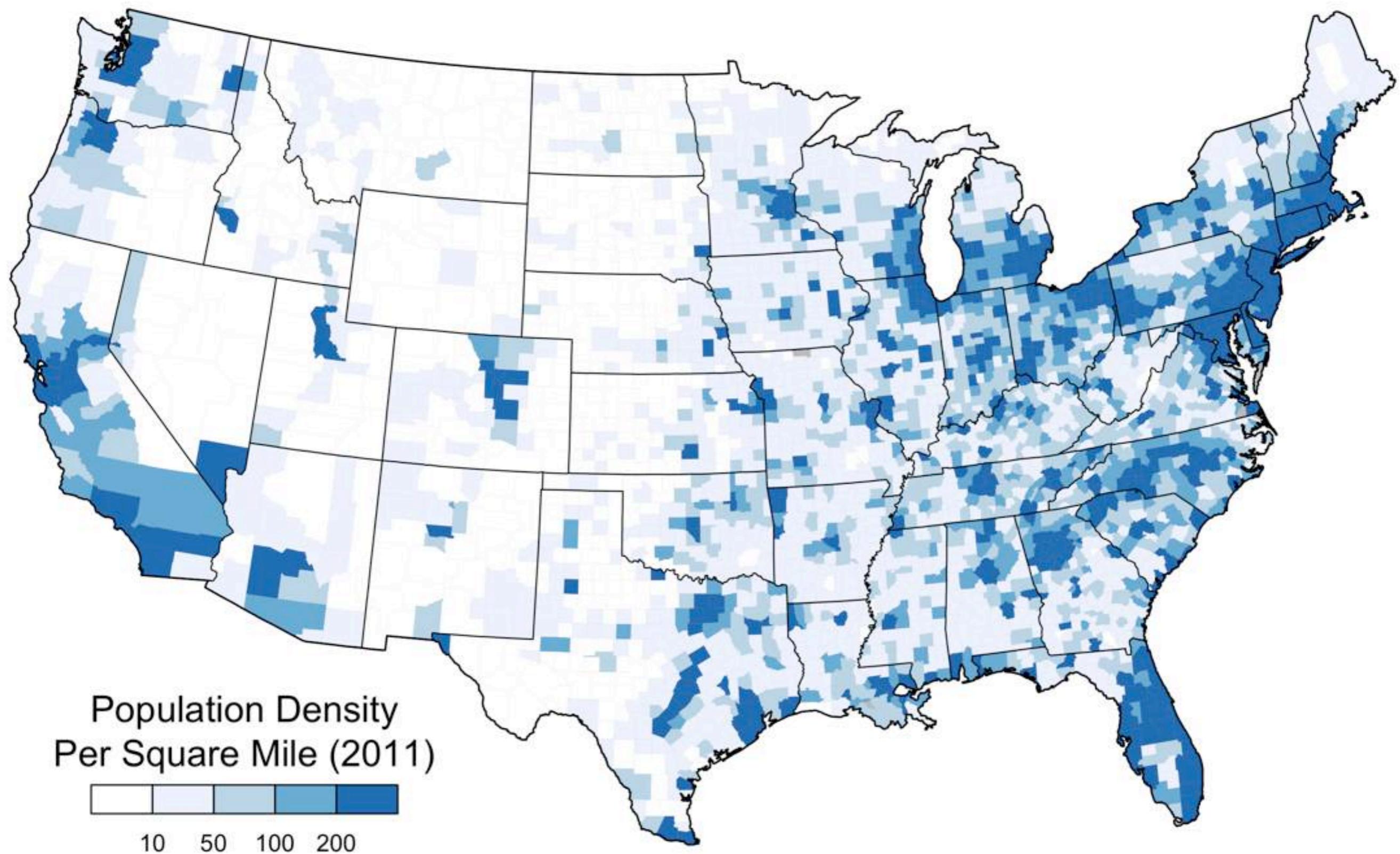
Spectator Sports: *derby, finals, league, marathon, medal, olympics, pitch, pro, reception, root, rooting, ...*

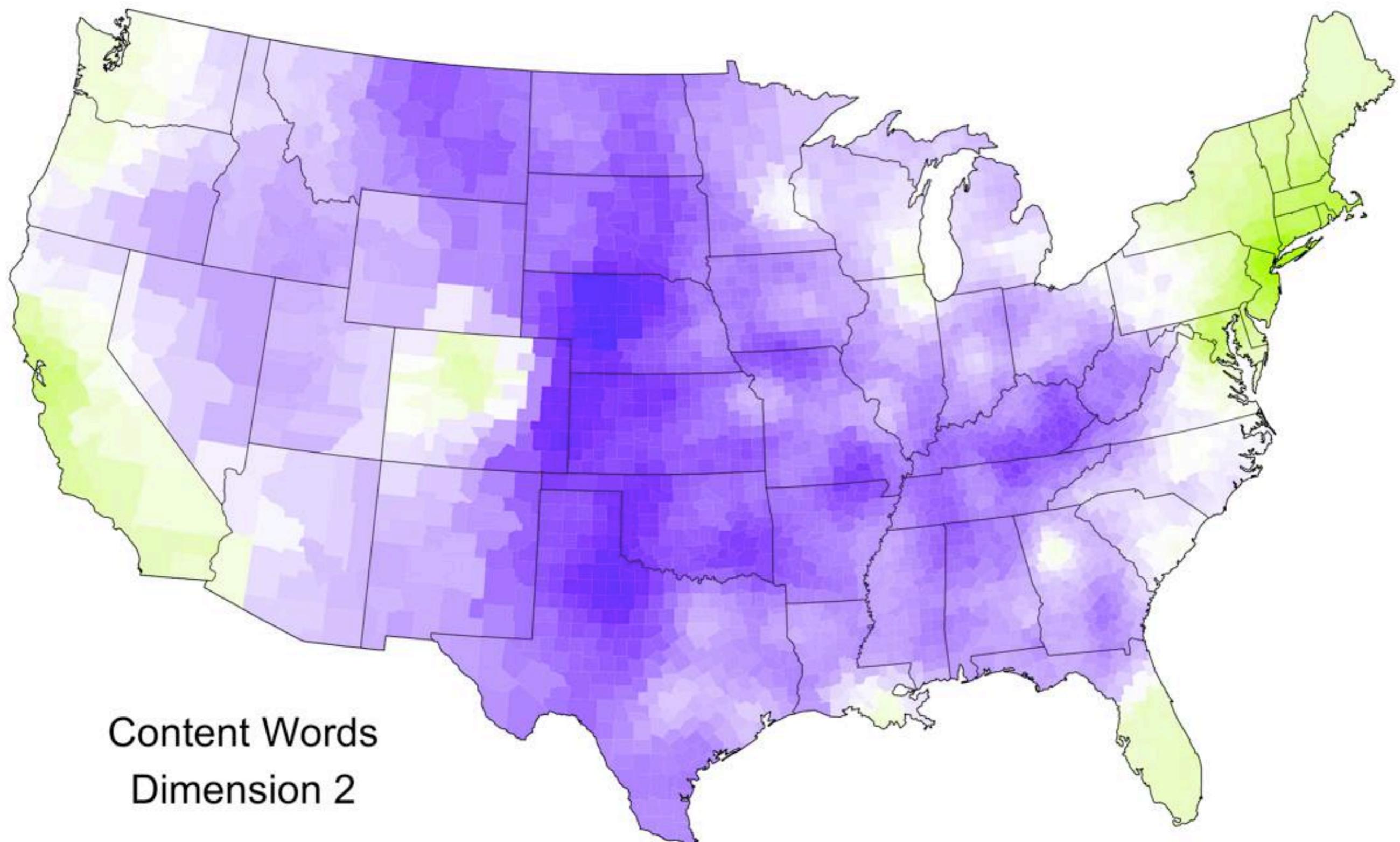
.

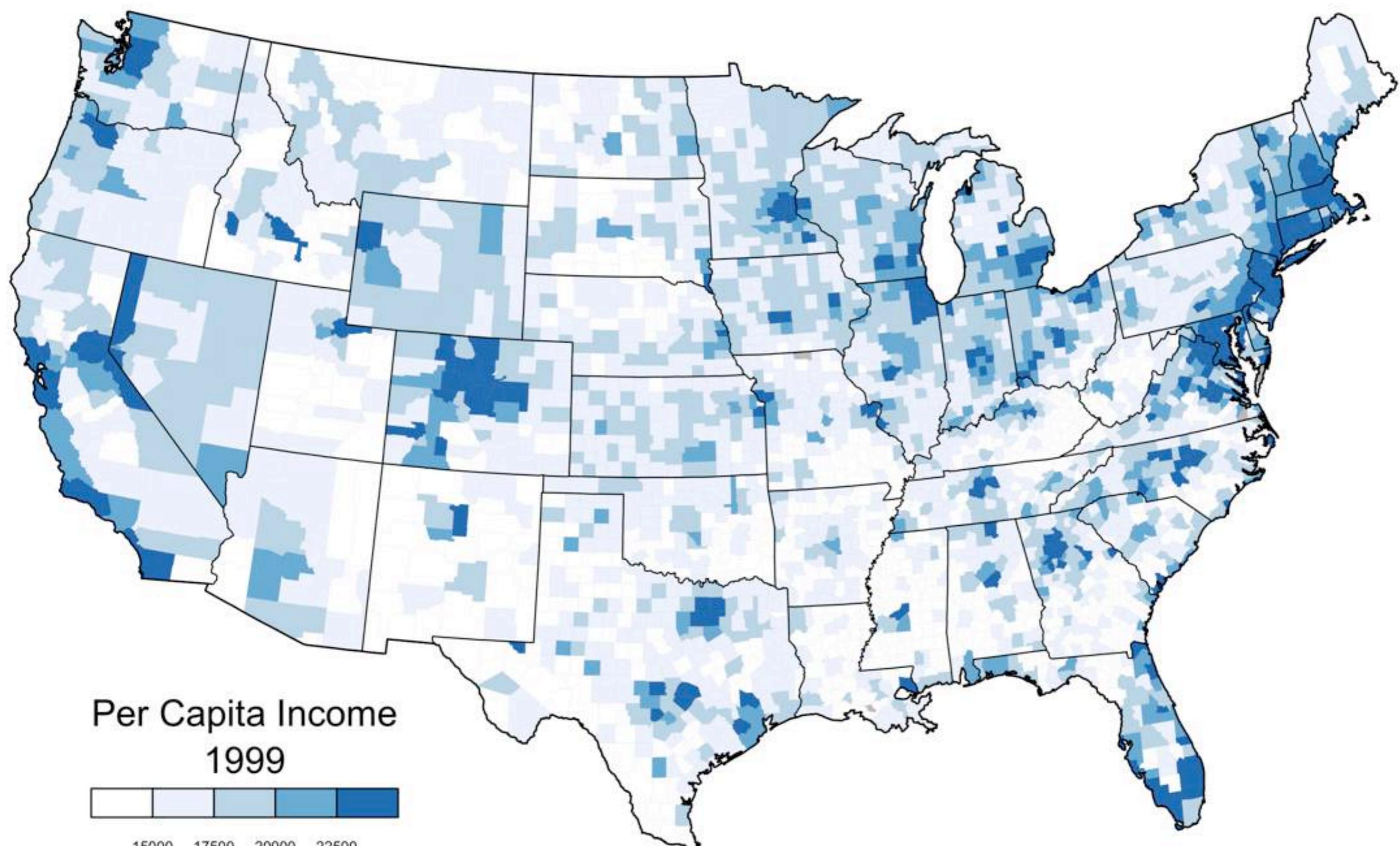












Coasts: Urban Lifestyle

Food and Dining: *avocado, bagel, bagels, barista, blueberry, bread, breakfast, brunch, burger, catering, ...*

Shopping, Fashion and Beauty: *copped, designer, eyebrows, errands, fitness, gym, ikea, mall, outlets, ...*

Alcohol and Marijuana: *ale, bar, blunt, blunts, bong, champagne, cider, ciroc, cocktail, dope, drinking, ...*

Urban Life: *bum, bus, neighborhood, parking, street, studio, suite, traffic, waterfront*

Foreign Nationalities: *asian, colombia, costa, dominican, french, greek, india, irish, italian, italy, jamaican, ...*

.

Inland: Rural Life + Introspection

Rural Life: *boots, cotton, country, cow, dirt, deer, donkey, dust, fence, hog, hunt, loaded, shoot, storm, ...*

Roads and Vehicles: *chevy, concrete, drive, drivin, ford, gas, roads, speed, tires, truck, trucks, wheeler, ...*

Sports: *ball, bandwagon, basketball, coach, coaches, compete, durant, football, fouls, halftime, heisman, ...*

Emotions and Mental States: *amazed, cherish, decide, fallin, figured, fix, fixing, forgive, forgiveness, judge, ...*

.

Everyone has a unique way of speaking



Gezellig bij Emily en Charlotte.

Translation: Having fun with Emily
and Charlotte.

Hiiiiii schatjesss!

Translation: Hiiiiii cutiesss!

@USER

Goodmorning

Saaie middag.

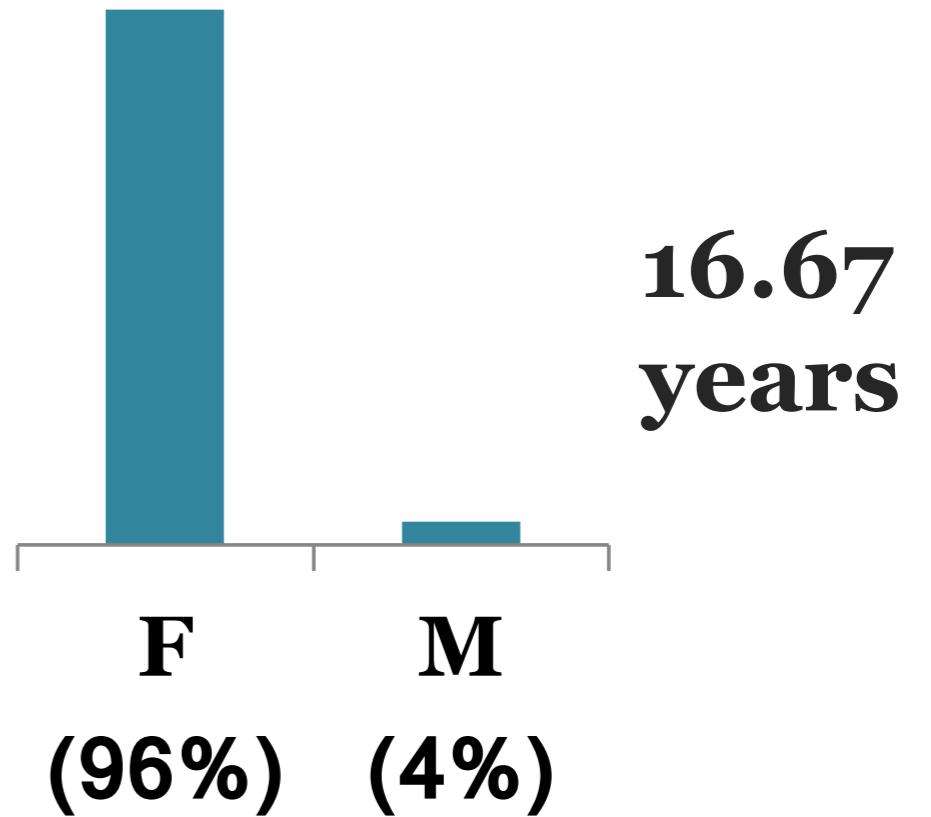
Translation: Boring afternoon

Gezellig bij Emily en Charlotte.
Translation: Having fun with Emily
and Charlotte.
Hiiiiii schatjesss!
Translation: Hiiiiii cutiesss!
@USER
Goodmorning
Saaie middag.
Translation: Boring afternoon

Female, 15

Gezellig bij Emily en Charlotte.
Translation: Having fun with Emily
and Charlotte.
Hiiiiii schatjesss!
Translation: Hiiiiii cutiesss!
@USER
Goodmorning
Saaie middag.
Translation: Boring afternoon

Female, 15



I'm walking on sunshine <3 #and don't
you feel good

lalaloveya <3

[LINK] Lekker nummer om mee op te
staan ^^

Translation: [LINK] Nice song to wake
up with ^^

Never thought it would mean so much
to me. This time I will prove the
world they were wrong about me.

I'm walking on sunshine <3 #and don't
you feel good
lalaloveya <3
[LINK] Lekker nummer om mee op te
staan ^^

Translation: [LINK] Nice song to wake
up with ^^

Never thought it would mean so much
to me. This time I will prove the
world they were wrong about me.

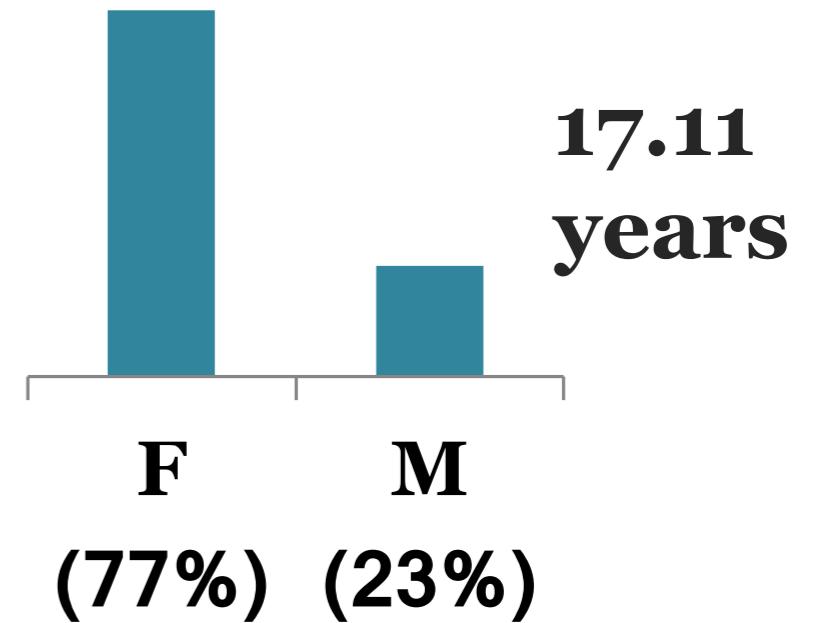
Male, 16

I'm walking on sunshine <3 #and don't
you feel good
lalaloveya <3
[LINK] Lekker nummer om mee op te
staan ^^

Translation: [LINK] Nice song to wake
up with ^^

Never thought it would mean so much
to me. This time I will prove the
world they were wrong about me.

Male, 16



Language & social identity

- Automatically inferring social variables
 - User profiling
 - Studying social and cultural phenomena
- New insights into language
 - Making NLP tools more robust

Social Language is *Everywhere*

- Twitter (Rao et al., 2010; Bamman et al., 2014; Fink et al., 2012; Bergsma and Van Durme, 2013; Burger et al., 2011, Rao et al., 2010; Nguyen et al., 2013, Eisenstein et al. 2010; ..),
- blogs (Mukherjee and Liu, 2010; Schler et al., 2005, Rosenthal and McKeown, 2011; Goswami et al., 2009),
- telephone conversations (Garera and Yarowsky, 2009)
- YouTube (Filippova, 2012)
- etc.. etc..

Gender in Language

Females

- pronouns
- emotion words
- emoticons
- CMC words

Males

- more numbers
- tech words
- links
- prepositions
- articles
- longer words

Gender and Power in Language

- Manifestations of power
 - Requests
 - Initiator (started the e-mail thread)
 - overt displays of power (I need the report by end of Friday)
 - etc...
- Power relations: superior and subordinate
- Gender variables:
 - Gender of author
 - Gender environment: based on gender of discourse participants (female, male, mixed)



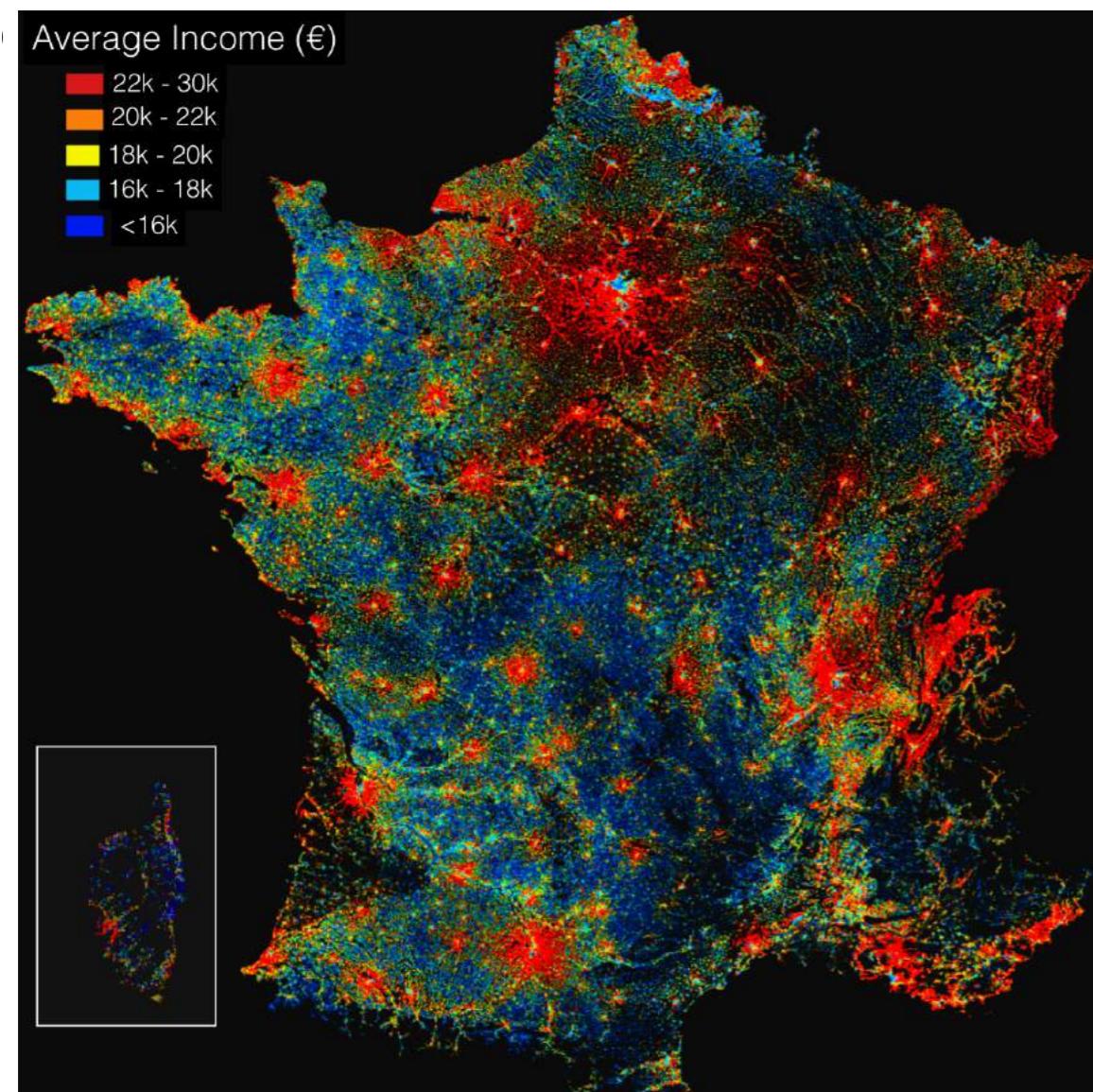
Both **gender** and **gender environment** influence how power is manifested

Gender and Prestige

- Movie review site.
 - Male dominated
 - Males tend to receive higher prestige (votes)
- Changes in style over time
 - Stylistic characteristics associated with gender
 - Females decreased their usage of some gendered stylistic characteristics (e.g., hedging)
 - ...'to be in sync with their male counterparts'?



Sociolinguistic variation with respect to socioeconomic status



- Compare language usage on Twitter with known socioeconomic indicators

Three kinds of Linguistic Markers

Three kinds of Linguistic Markers

- Standard negation: Ceci **n'est** pas une pipe

Three kinds of Linguistic Markers

- Standard negation: Ceci n'est pas une pipe

First negation participle is omitted in speech but required in text.

Idea: Measure what % of people omit it on twitter.

Three kinds of Linguistic Markers

- Standard negation: Ceci n'est pas une pipe

First negation participle is omitted in speech but required in text.
Idea: Measure what % of people omit it on twitter.

- Standard spelling:

Je mange une pomme. Je mange des pommes.

Three kinds of Linguistic Markers

- Standard negation: Ceci n'est pas une pipe

First negation participle is omitted in speech but required in text.
Idea: Measure what % of people omit it on twitter.

- Standard spelling:

Je mange une pomme. Je mange des pommes.

Plural ending is silent when pronounced.
Idea: Measure what % of plural words don't include the ending.

Three kinds of Linguistic Markers

- Standard negation: Ceci n'est pas une pipe

First negation participle is omitted in speech but required in text.
Idea: Measure what % of people omit it on twitter.

- Standard spelling:

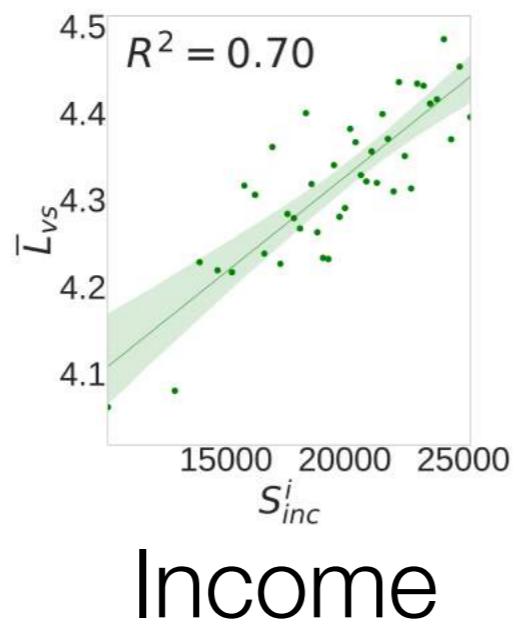
Je mange une pomme. Je mange des pommes.

Plural ending is silent when pronounced.
Idea: Measure what % of plural words don't include the ending.

- Vocabulary Size: how many unique words they used

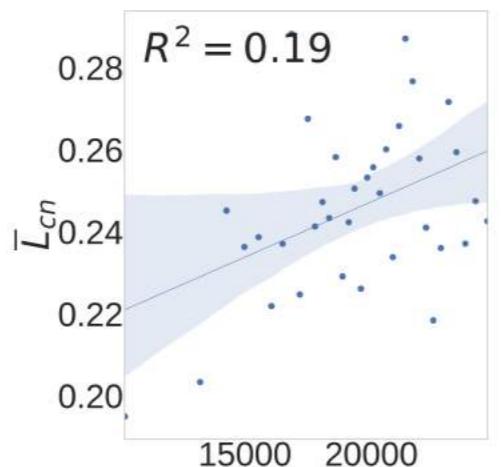
Results: Simple linguistic markers correlate with socioeconomic status

Vocabulary
Size

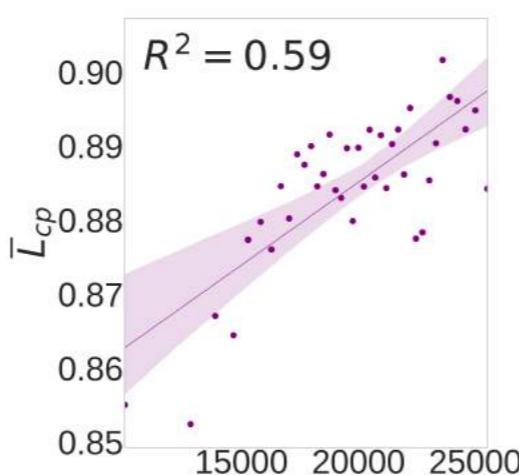


Results: Simple linguistic markers correlate with socioeconomic status

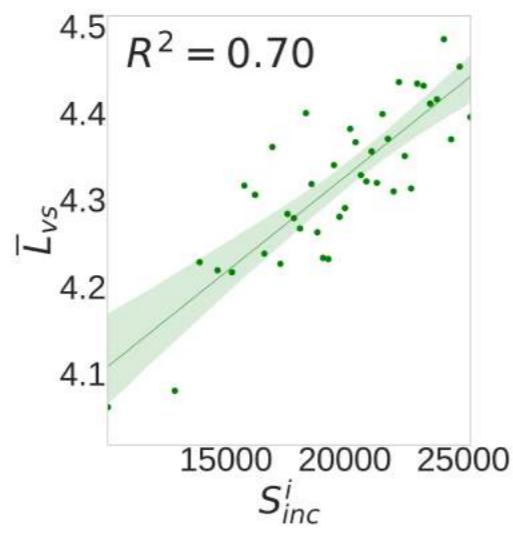
Correct
Negation



Correct
Pluralization



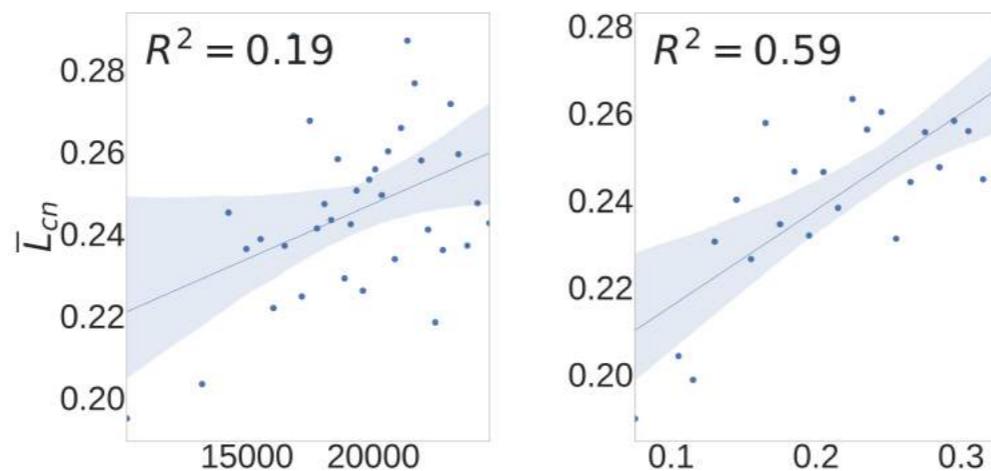
Vocabulary
Size



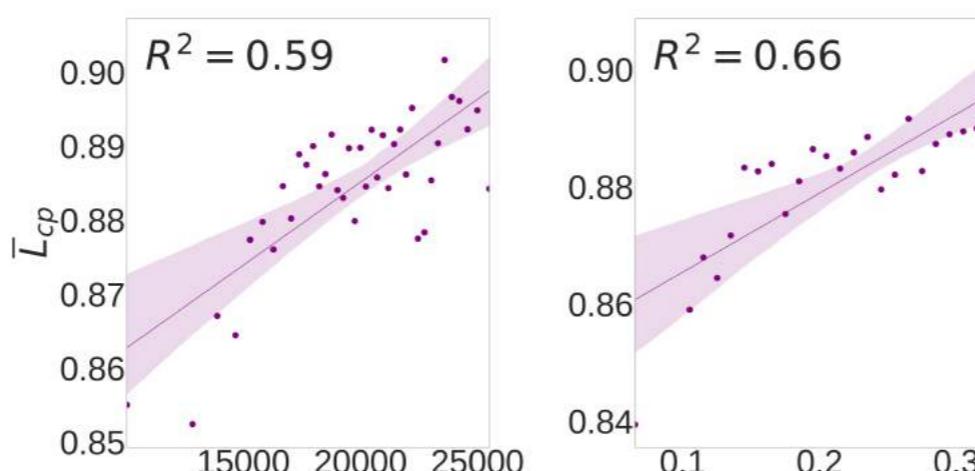
Income

Results: Simple linguistic markers correlate with socioeconomic status

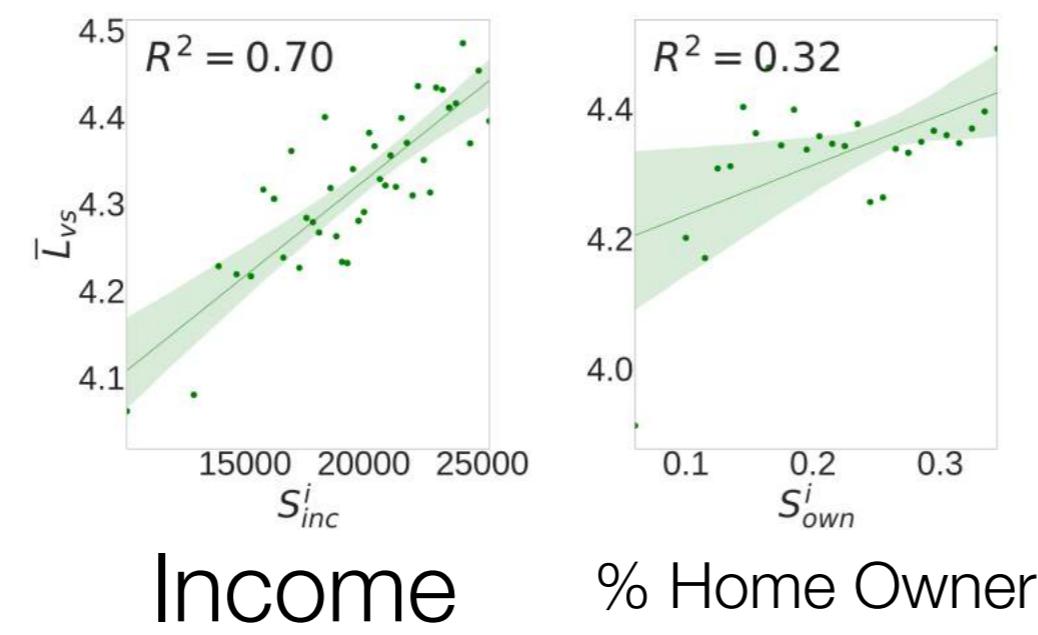
Correct
Negation



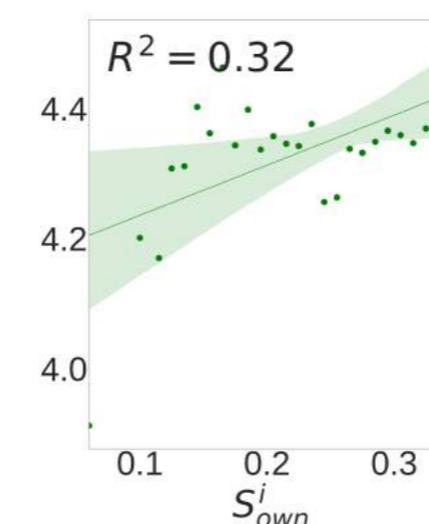
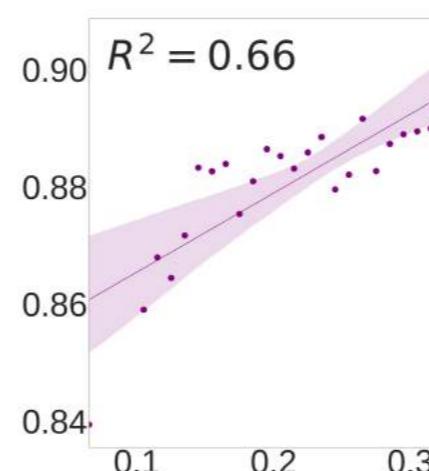
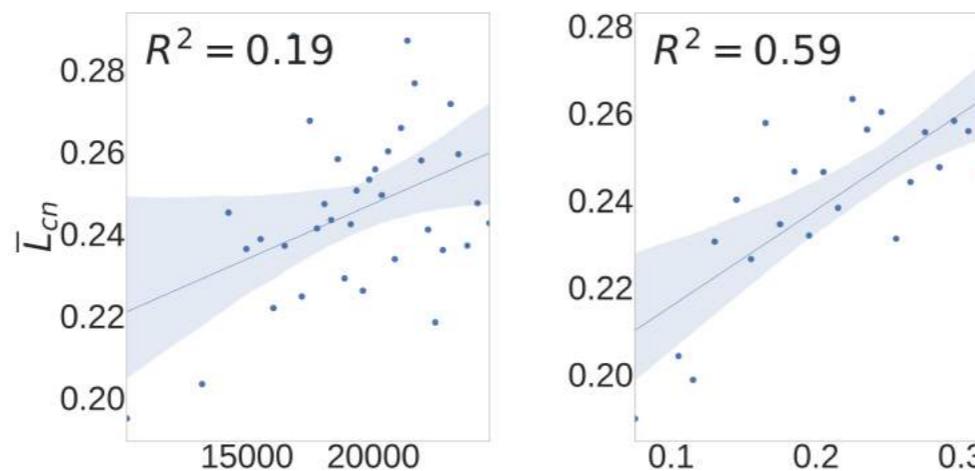
Correct
Pluralization



Vocabulary
Size

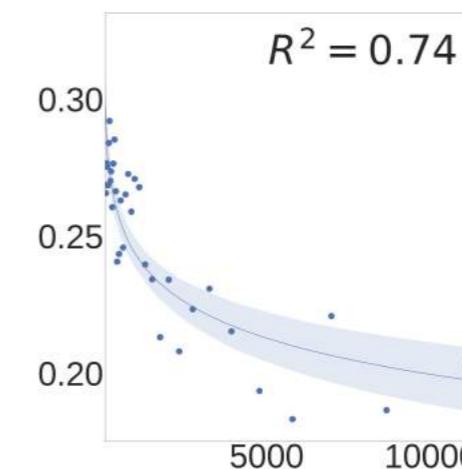
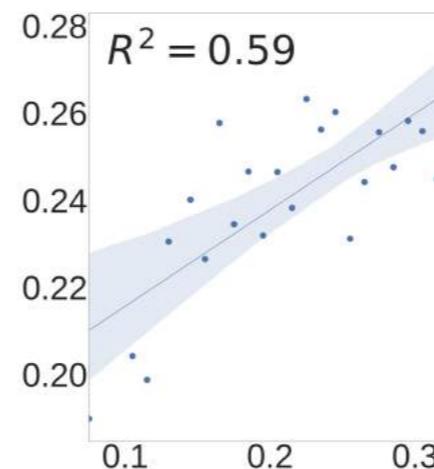
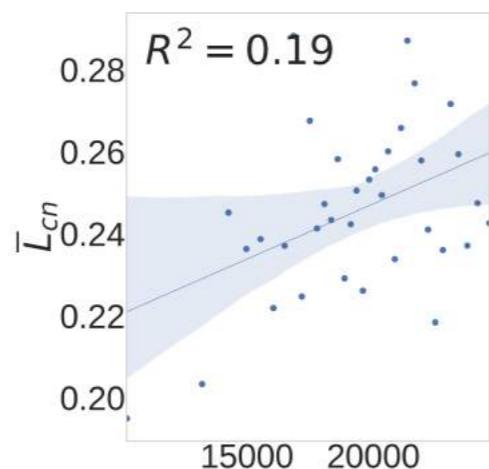


Income % Home Owner

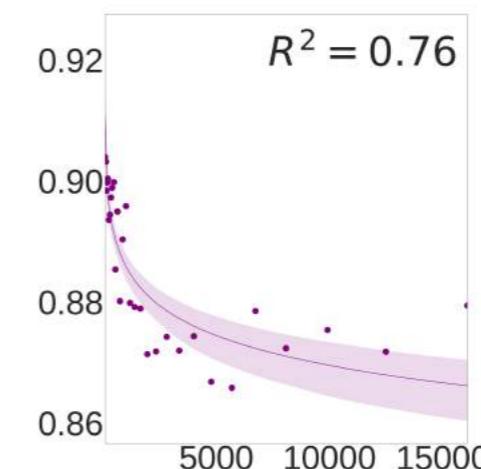
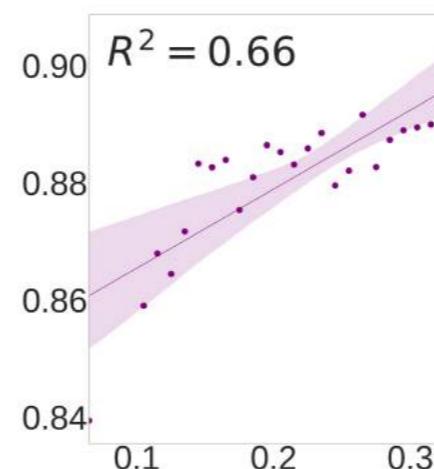
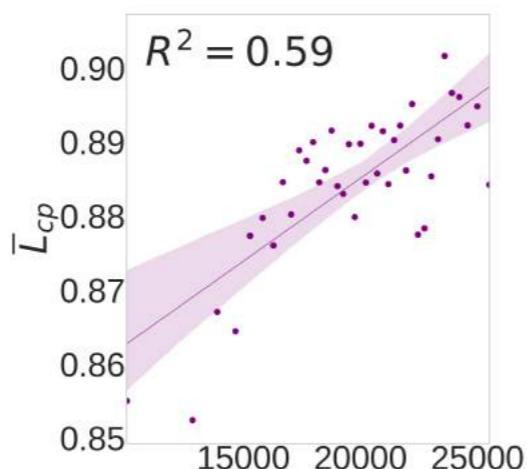


Results: Simple linguistic markers correlate with socioeconomic status

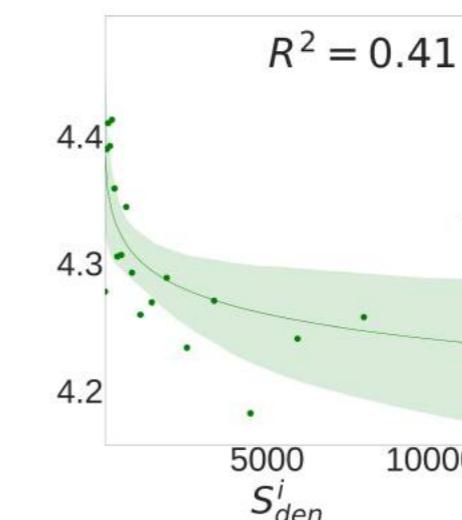
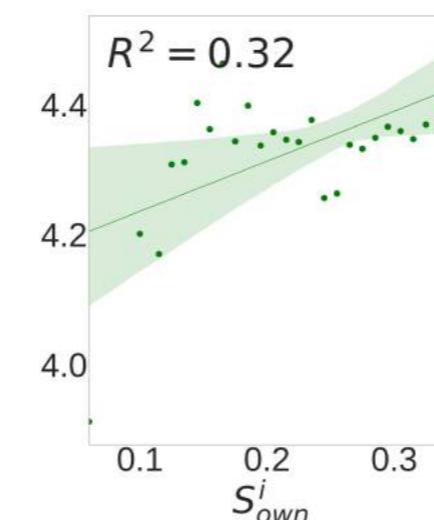
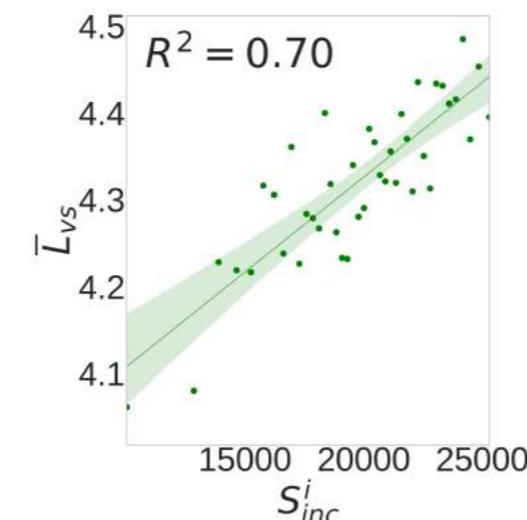
Correct
Negation



Correct
Pluralization



Vocabulary
Size

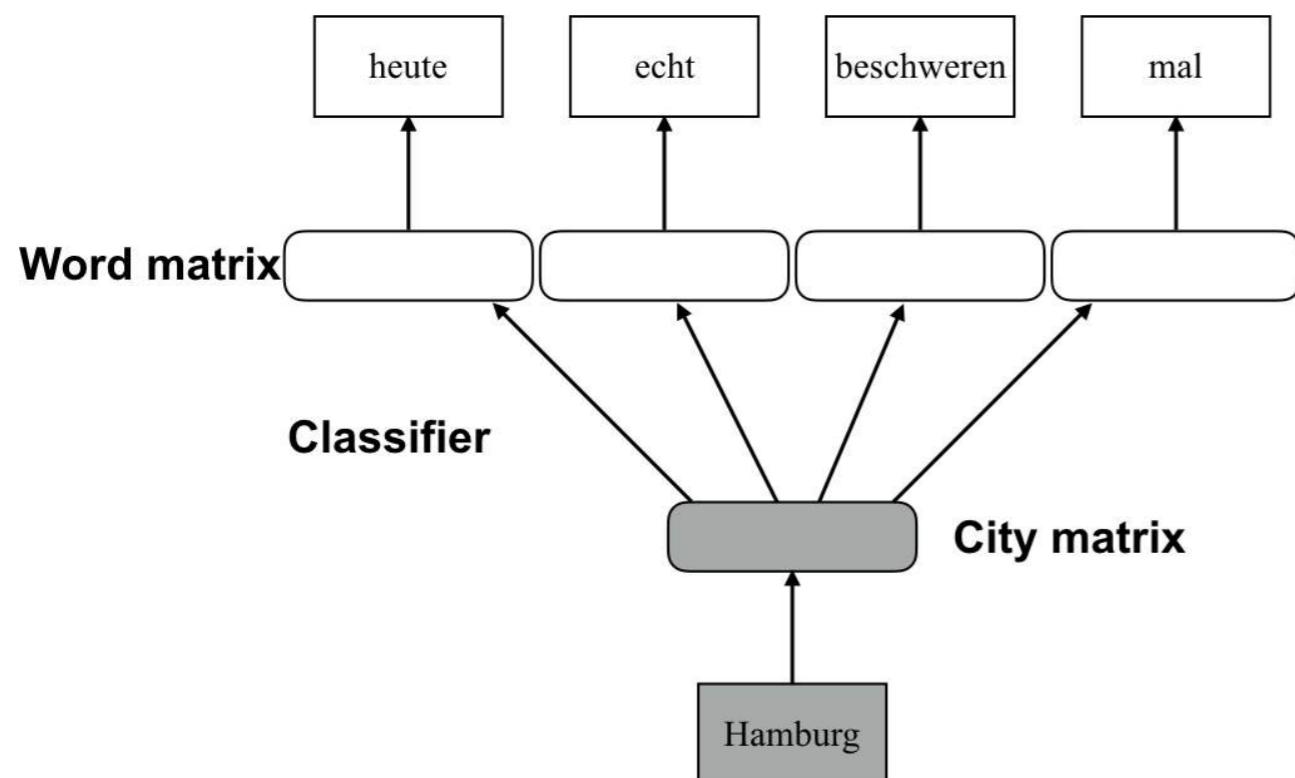


Income

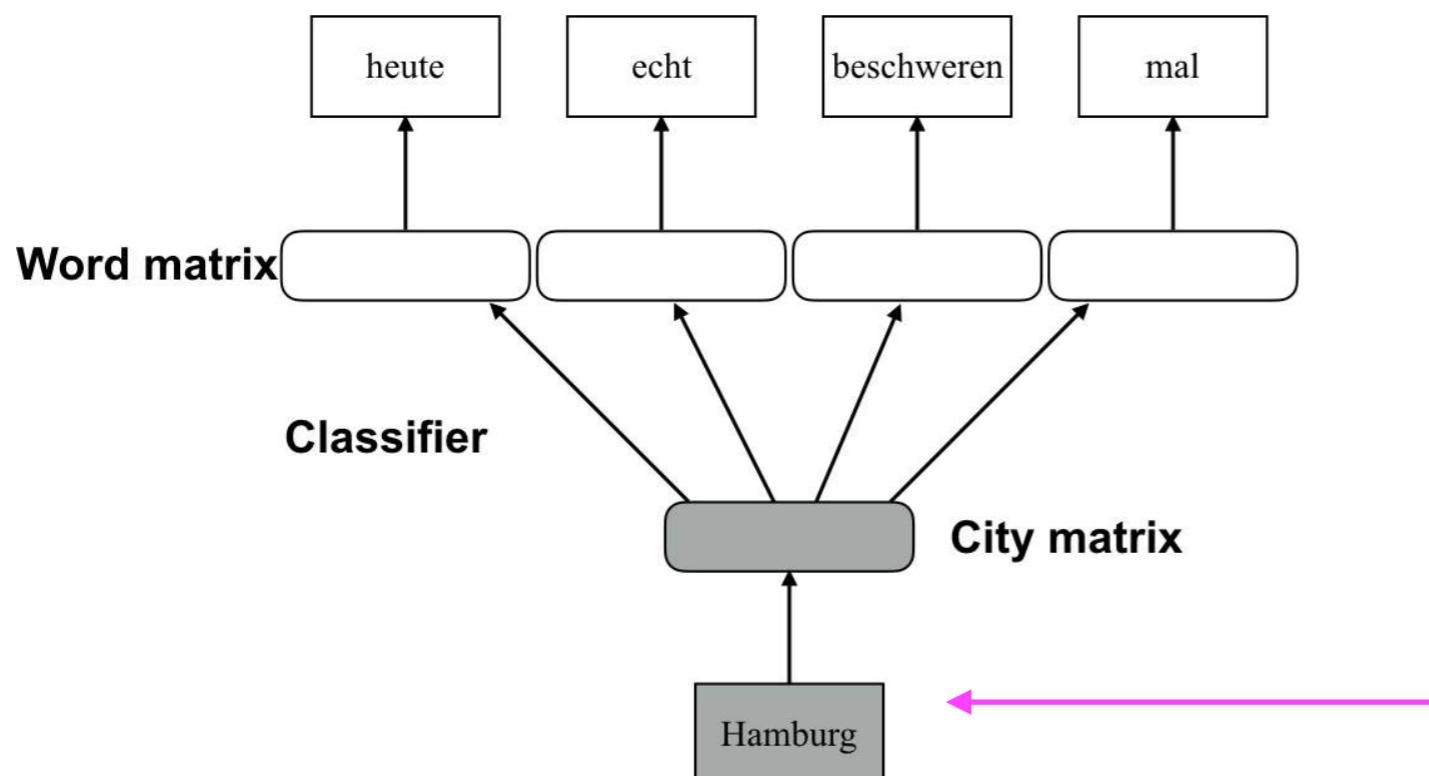
% Home Owner

Density

Capture dialectal regions with city2vec

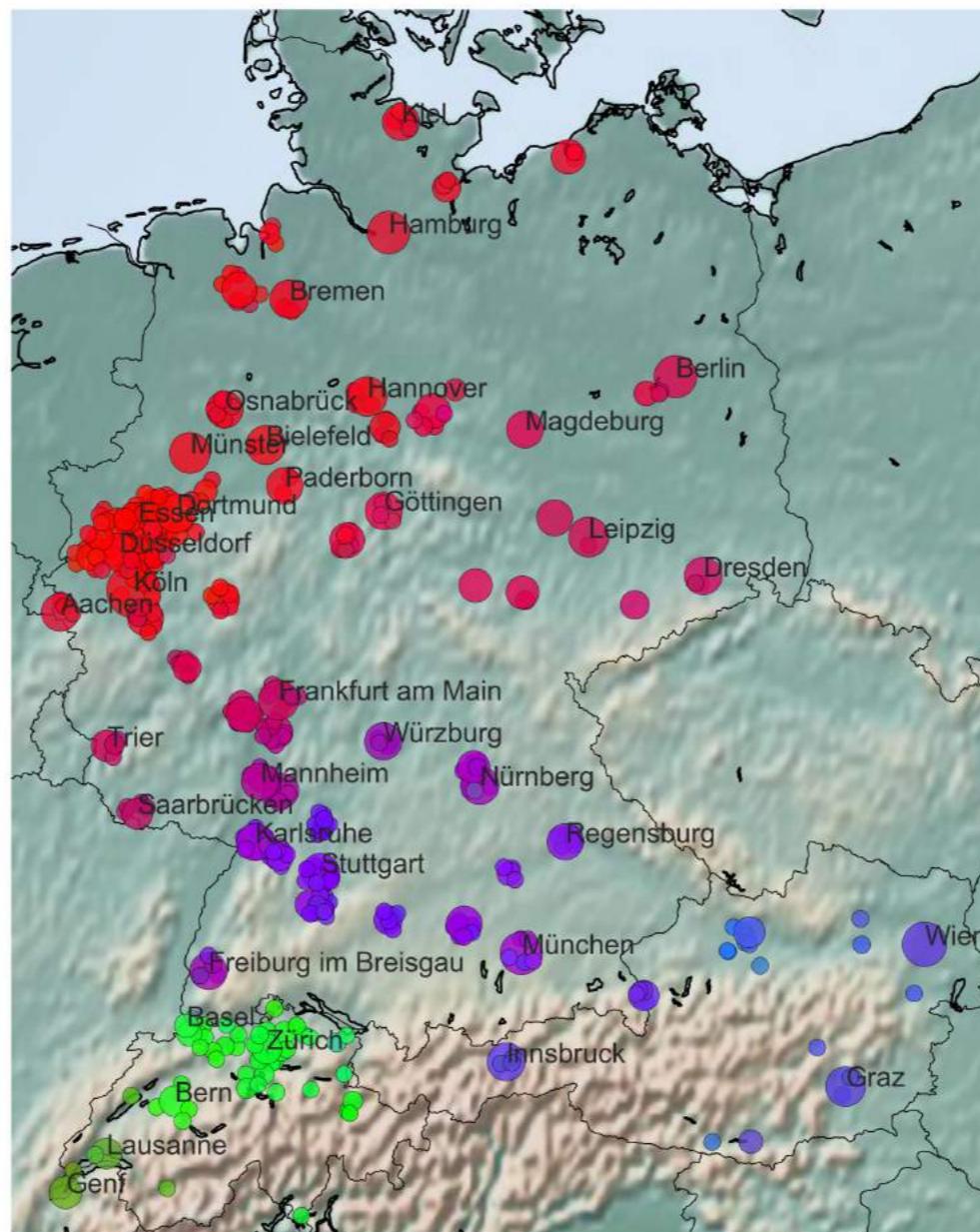


Capture dialectal regions with city2vec

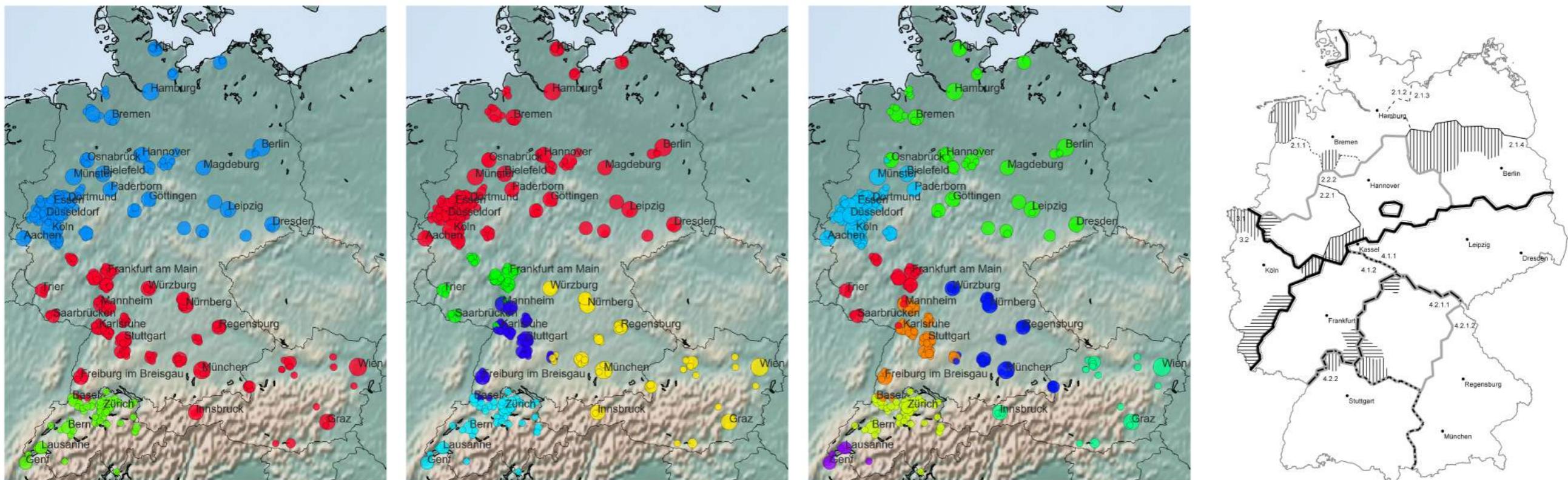


Run Principle Component Analysis (PCA) on the city vectors to discover the main ways in which cities vary

The first PCA component shows meaningful separation!



Clusters on PCA embeddings capture known regions





NLP for Computational Social Science

Exploration vs. prediction

Cox &
Forkum
©2003



www.CoxAndForkum.com

Natural Language Processing

- Focus on tasks: accuracy, f-score, precision, recall
- “[...] there has been an over-focus on numbers, on beating the state of the art.”
Manning, 2016

Social sciences & Humanities

- Interpretation (why?)
(theory, causality, interpretability)

Explanation vs. prediction

Explanatory modeling

- minimize bias
- model validation: R^2 , significance coefficients, etc.
- risks: type I and type II
- causal relationships
- variables: small number of variables, interpretable

Predictive modeling

- minimize bias + variance
- model validation: external test set
- risks: overfitting
- associations
- variables: Many variables, black box?

NLP for theory building and explanation

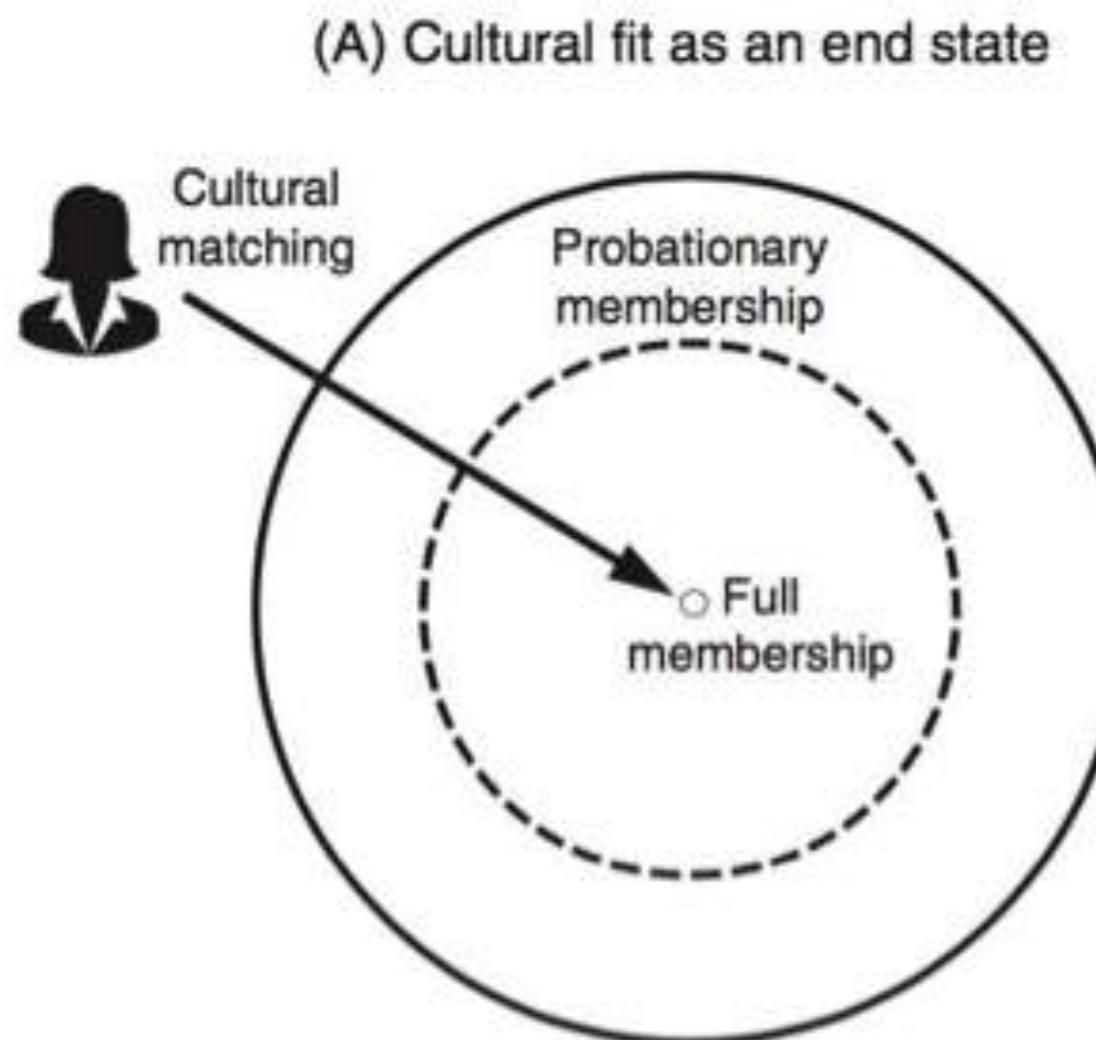
- ‘Traditional’ hypothesis testing but use NLP to operationalize variables
- Theory discovery using unsupervised methods
- Large-scale testing of existing theories using prediction models
- Theory discovery using black box(?) prediction models

Cultural fit in organisations

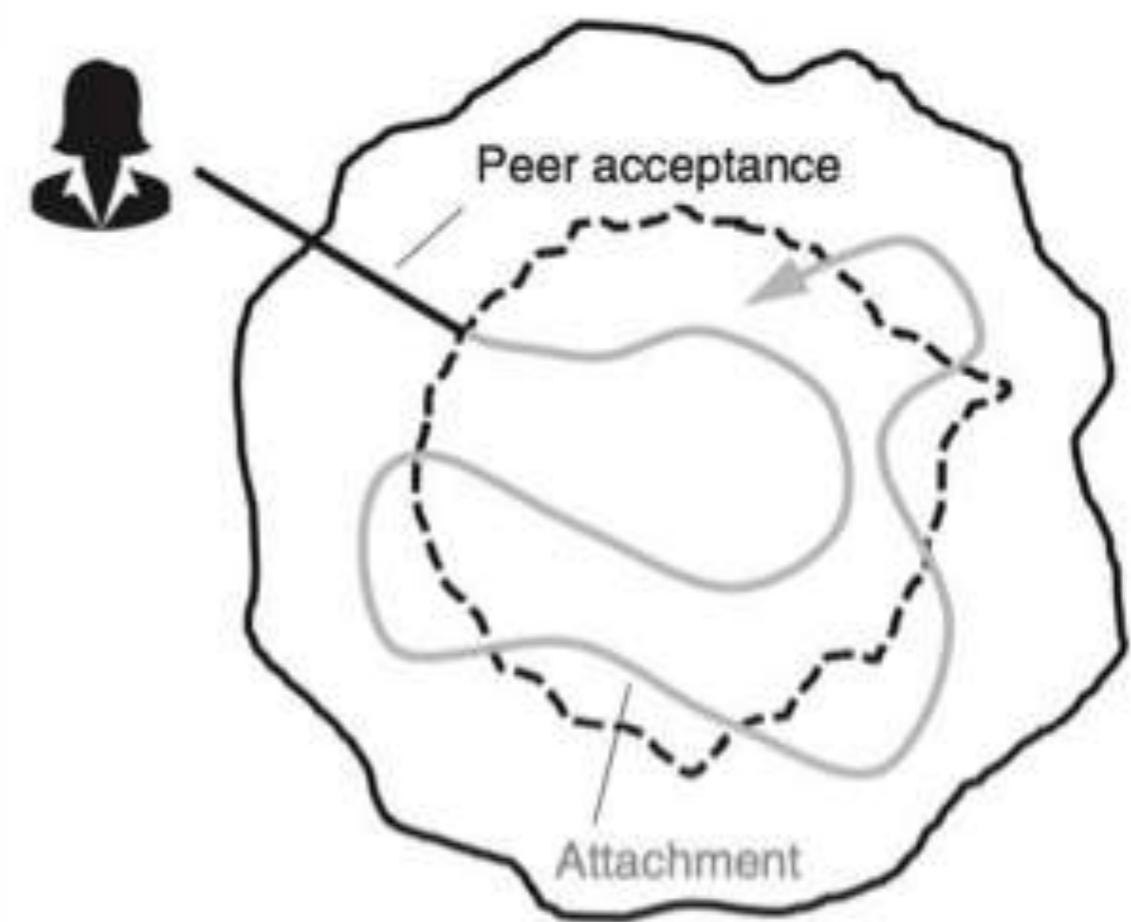
- Data: 10.24 million emails over five years. 601 employees of a midsized U.S. for-profit technology firm.
- How do people adapt in organisations? How does this affect career outcomes?
 - Previously: self reports
 - prone to bias
 - coarse categories
 - difficult to measure temporal variations
 - difficult to scale
 - Now: Measure based on language use

Cultural fit in organisations

enculturation trajectory: an individual's temporal pattern of cultural fit



(B) Enculturation as a process

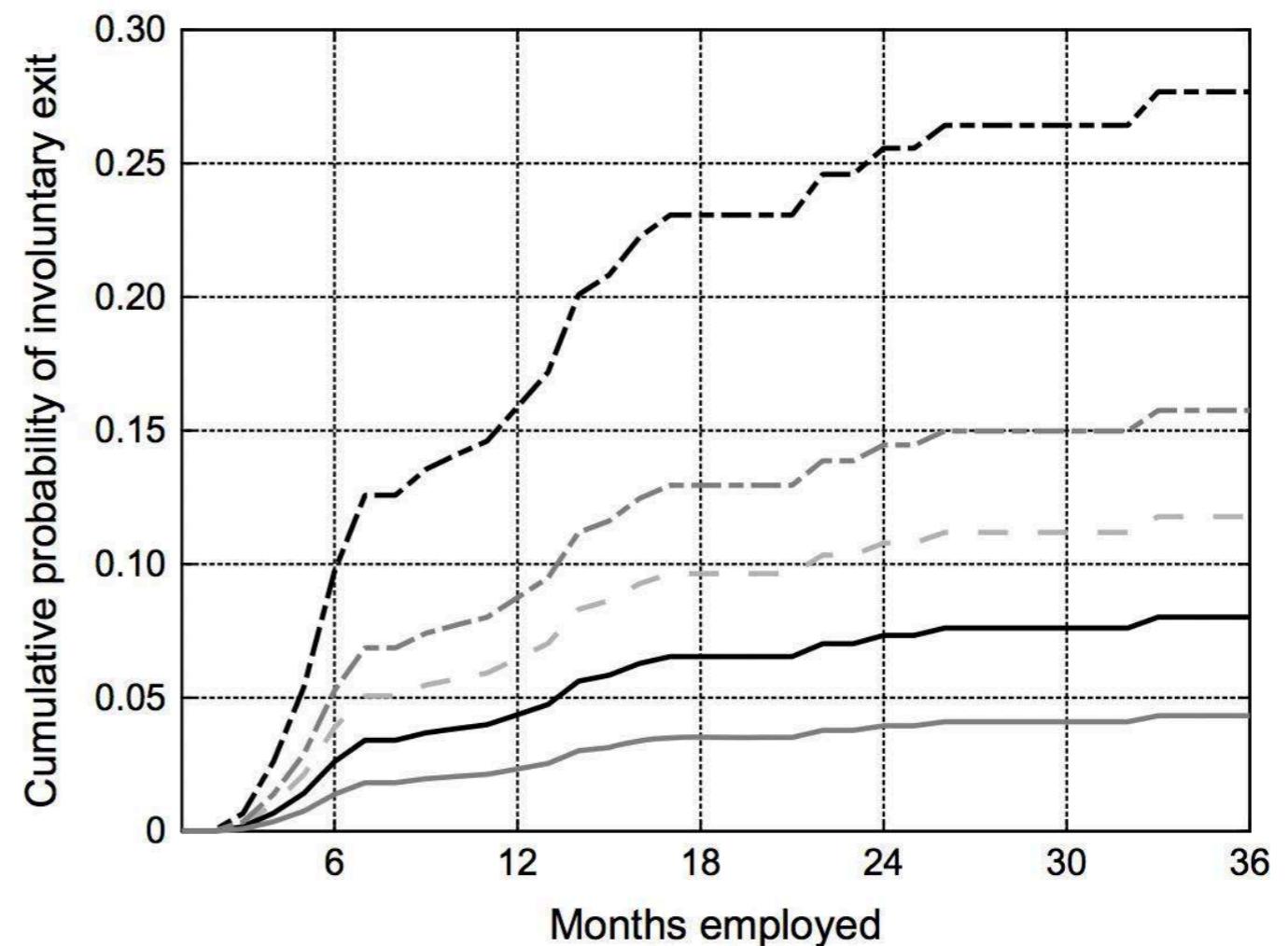
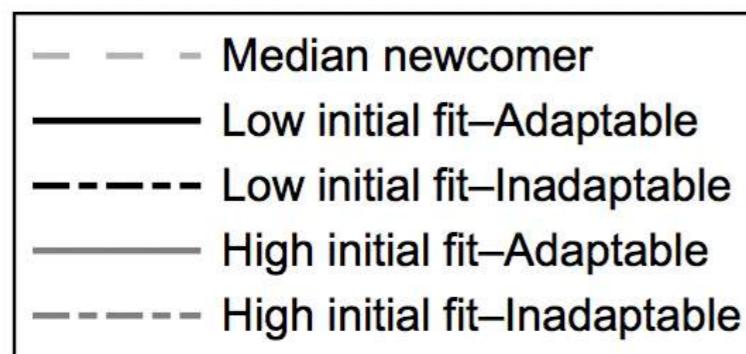


Cultural fit in organisations

- Cultural fit: linguistic alignment between an individual and her interaction partners in the organization.
 - Measure alignment between incoming and outgoing messages
 - Time windows: months
- LIWC (Linguistic Inquiry and Word Count):
 - Counts words in predefined categories (e.g., swear words, pronouns, insight, anxiety)

Cultural fit in organisations

- Slow enculturation rates in early stages: more likely to exit involuntarily
- Cultural fit can decline later in careers: sign of voluntary exit



Topic Modeling

- Assumptions:
 - A document is a mixture over topics
 - A topic: distribution over a vocabulary
 - Number of topics need to be specified beforehand
- Limitations:
 - Context? (bag of words), interpretative ability.
 - Sensitive to preprocessing steps.
 - Not all topics are meaningful.

Topic Modeling for Social Sciences

- Themes and author gender in 19th-century literature (Jockers and Mimno, Poetics 2013)
- Framing and agenda setting in four years of public statements issued by members of the U.S. Congress (Tsur et al., ACL-IJCNLP 2015)
- Trends in academic fields based on dissertation abstracts (McFarland et al., Poetics 2013)
- Trends in literary studies (Goldstone and Underwood, 2014)

Grounded theory

- Glaser & Strauss, 1967
- Inductive methodology
- Emergence of conceptual categories
- Grounded in data
- Iterative process (often also repeated data collection)
- Drawbacks: time-consuming, biases of the researcher

Topic modeling vs grounded theory

- Topic modeling and grounded theory on the same data (survey data with free-text responses).
- Data: Social media user leaves a site and becomes a non- user. 5,245 participants (opt-in to share with researchers)
- Question such as “How did your friends react [to you leaving Facebook]?”



Topic modeling vs grounded theory

Topic modeling vs grounded theory

- Similarities:
 - Iterative process
 - Grounded in data
 - Identify thematic patterns

Topic modeling vs grounded theory

- Similarities:
 - Iterative process
 - Grounded in data
 - Identify thematic patterns
- Grounded theory: Two researchers. Iterative process:
 - categories were created/combined/removed/changed.
 - Later on initial categories grouped into broader themes.

Topic modeling vs grounded theory

- Similarities:
 - Iterative process
 - Grounded in data
 - Identify thematic patterns
- Grounded theory: Two researchers. Iterative process:
 - categories were created/combined/removed/changed.
 - Later on initial categories grouped into broader themes.
- LDA: 10 topics. Some pre-processing (lowercase, stop word removal, etc.)

Topic modeling vs grounded theory

- No simple one-to-one correspondence between topics and themes:
 - Topics capture components of a theme
 - Most themes associated with at least one, usually two or three, topics
 - Topics tend to have a lower level of abstraction

Topic modeling vs grounded theory

- No simple one-to-one correspondence between topics and themes:
 - Topics captured components of a theme
 - Most themes associated with at least one, usually two or three, topics
 - Topics tend to have a lower level of abstraction

“The grounded theory analysis took two researchers several hours of work per week over roughly 2.5 months. In contrast, a single researcher conducted and wrote up the topic modeling results within a few hours over 2 days.”

Topic modeling vs grounded theory

- No simple one-to-one correspondence between topics and themes:
 - Topics captured components of a theme
 - Most themes associated with at least one, usually two or three, topics
 - Topics tend to have a lower level of abstraction

“The grounded theory analysis took two researchers several hours of work per week over roughly 2.5 months. In contrast, a single researcher conducted and wrote up the topic modeling results within a few hours over 2 days.”

“these methods involve surprisingly similar processes and produce surprisingly similar results.”

Breakout session time!

- How is your course project going?
- How is HW5 (HW4?) going?



Ethics

Ethics

Why does a discussion about
ethics need to be a part of
NLP?

Conversational Agents



Question Answering

According To Google, Barack Obama Is King Of The United States

Google Answers gets it wrong. Is this a Google Answers Bomb?

Barry Schwartz on November 25, 2014 at 6:04 pm



A screenshot of a Google search results page. The search query "King of United States" is entered into the search bar. The "Web" tab is selected, showing approximately 460 million results found in 0.72 seconds. The top result is a snippet from Breitbart.com featuring a photo of Barack Obama with the text "All Hail King Barack Obama, Emperor Of The United States Of America!". Below this is a link to the same article on Breitbart.com.

Google King of United States

Web Maps Images Shopping Videos More Search tools

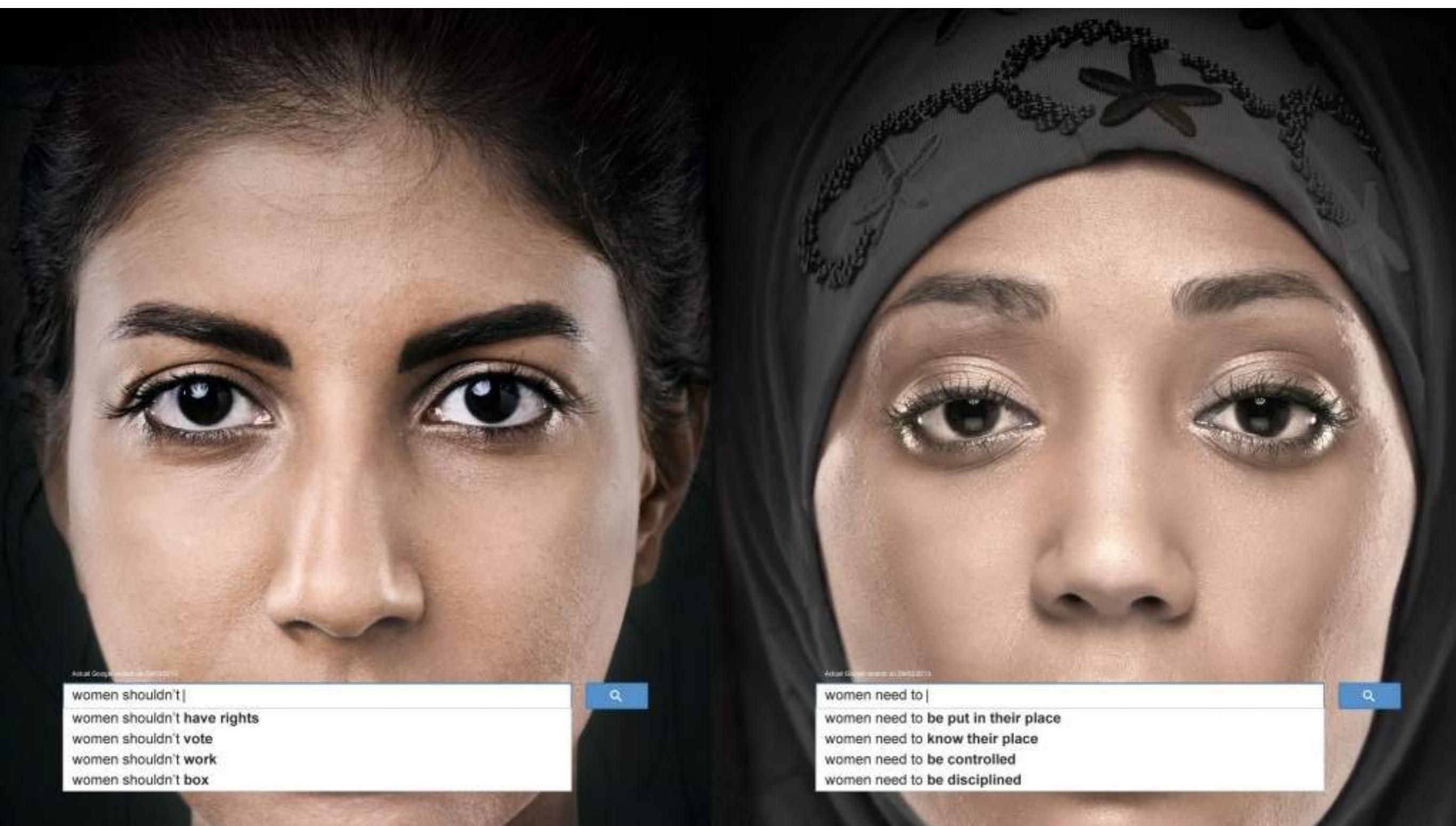
About 460,000,000 results (0.72 seconds)

All Hail King **Barack Obama**, Emperor Of The United States Of America!

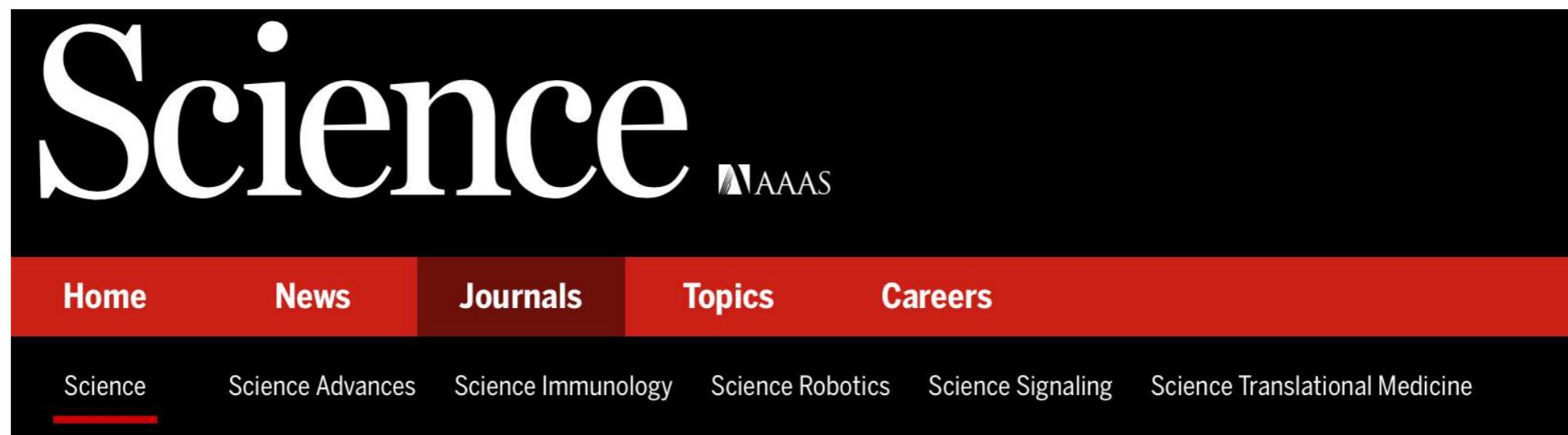
All Hail King Barack Obama, Emperor Of The United States ...
www.breitbart.com/.../All-Hail-King-Barack-Obama-Emperor-Of-... Breitbart

Feedback

Language Modeling



Vector semantics



The image shows the header of the Science journal website. The top half has a black background with the word "Science" in large white serif letters and the AAAS logo in smaller white letters to the right. Below this is a red navigation bar with five tabs: "Home", "News", "Journals", "Topics", and "Careers". The "Topics" tab is currently active, indicated by a white underline. Underneath the red bar is a black footer bar containing six links: "Science", "Science Advances", "Science Immunology", "Science Robotics", "Science Signaling", and "Science Translational Medicine".

SHARE REPORT



0



Aylin Caliskan^{1,*}, Joanna J. Bryson^{1,2,*}, Arvind Narayanan^{1,*}

* See all authors and affiliations

13

Science 14 Apr 2017:
Vol. 356, Issue 6334, pp. 183-186
DOI: 10.1126/science.aal4230



Peer Reviewed
← see details

Article

Figures & Data

Info & Metrics

eLetters

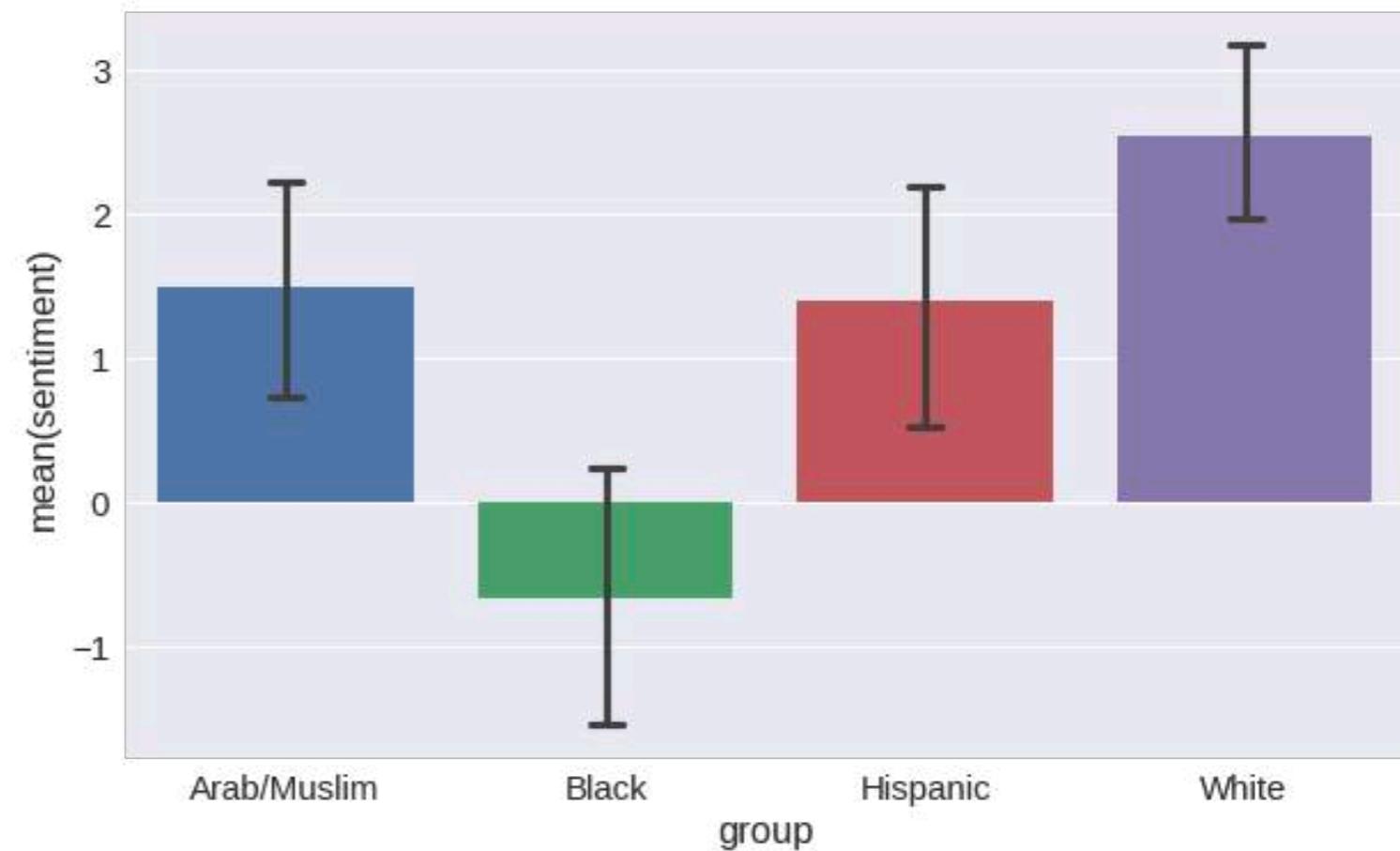
PDF

Sentiment Analysis

	sentiment	group
mohammed	0.834974	Arab/Muslim
alya	3.916803	Arab/Muslim
terryl	-2.858010	Black
josé	0.432956	Hispanic
luciana	1.086073	Hispanic
hank	0.391858	White
megan	2.158679	White

Sentiment Analysis

	sentiment	group
mohammed	0.834974	Arab/Muslim
alya	3.916803	Arab/Muslim
terryl	-2.858010	Black
josé	0.432956	Hispanic
luciana	1.086073	Hispanic
hank	0.391858	White
megan	2.158679	White



Machine Translation

Translate

[Turn off instant translation](#)

Bengali English Hungarian Detect language ▾



[English](#) [Spanish](#) [Hungarian](#) ▾

[Translate](#)

ő egy ápoló.
ő egy tudós.
ő egy mérnök.
ő egy pék.
ő egy tanár.
ő egy esküvői szervező.
ő egy vezérigazgatója.



she's a nurse.
he is a scientist.
he is an engineer.
she's a baker.
he is a teacher.
She is a wedding organizer.
he's a CEO.



110/5000

Bias in Natural Language Processing

Applications

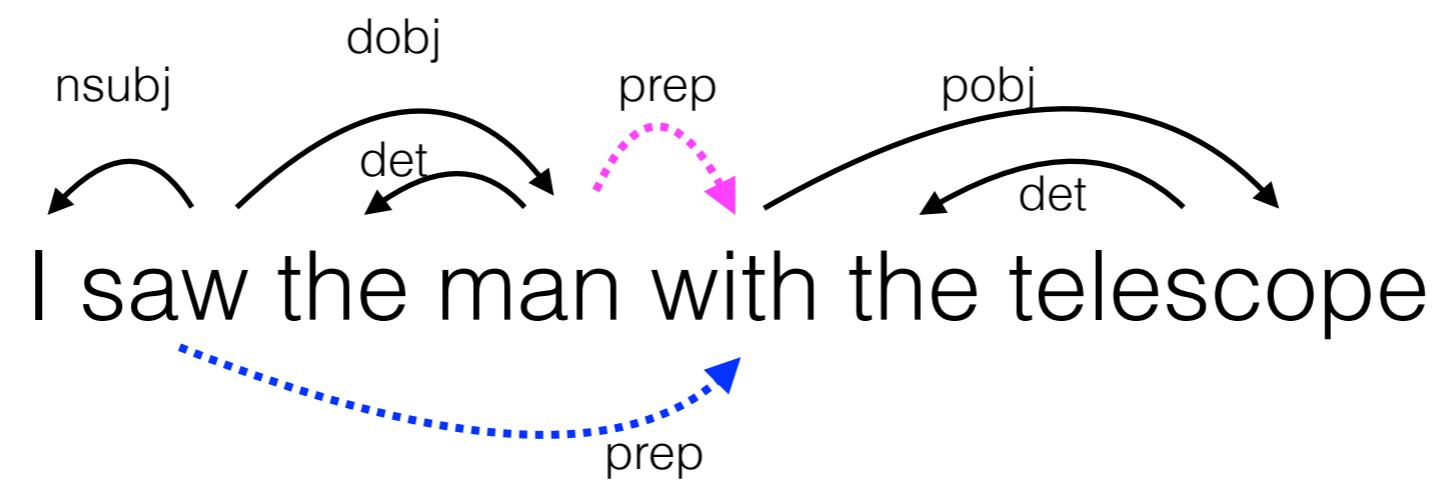
- Machine Translation
- Speech Recognition ([Tatman 2017](#))
- Question Answering ([Burghardt et al. '18](#))
- Dialogue Systems
- Information Extraction
- Summarization
- Sentiment Analysis ([Kiritchenko & Mohammad '18](#))
- Language Identification ([Blodgett et al.'16, Jurgens et al.'17](#))
- Text Classification ([Dixon et al. 2018](#))
- ...

Core technologies

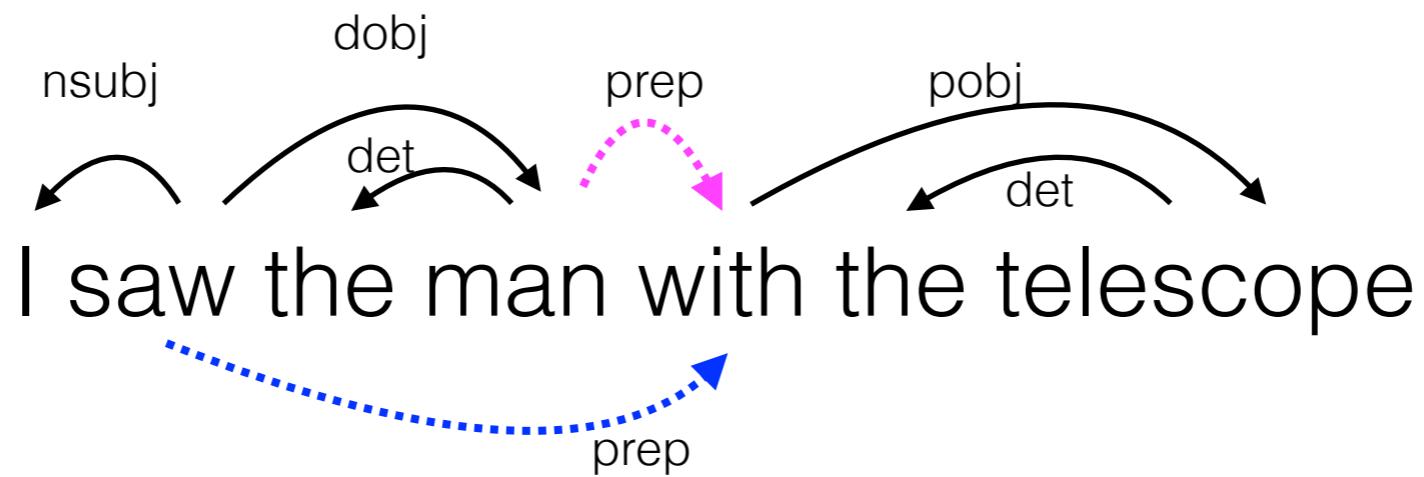
- Language modeling ([Lu et al. '18](#))
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition
- Coreference resolution ([Zhao et al. '18, Rudinger et al. '18](#))
- Word sense disambiguation
- Semantic Role Labelling ([Zhao et al. '17](#))
- SNLI ([Rudinger et al. '17](#))
- Word Embeddings ([Bolukbasi et al. '16](#))
- ...

The decisions we make about our methods — training data, algorithm, evaluation — are often tied up with its use and **impact** in the world.

Scope

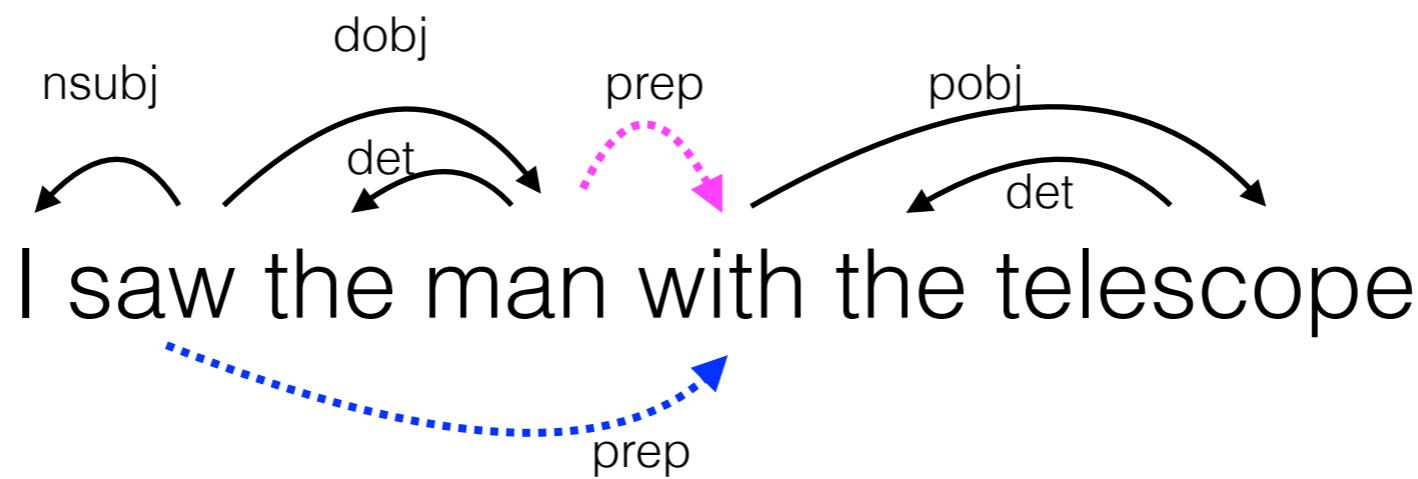


Scope



- NLP often operates on text divorced from the context in which it is uttered.

Scope

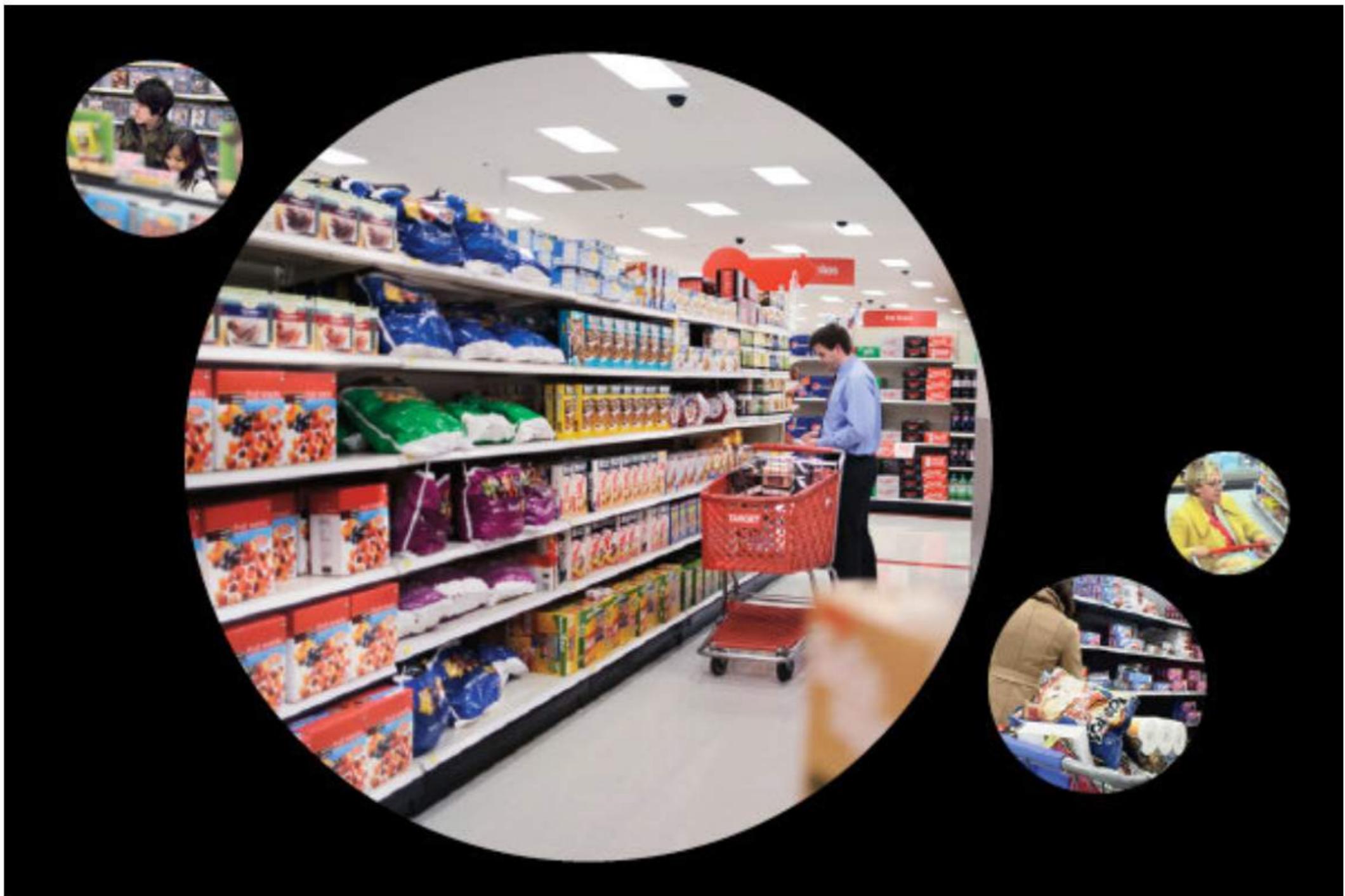


- NLP often operates on text divorced from the context in which it is uttered.
- It's now being used more and more to reason about **human behavior**.

Privacy

How Companies Learn Your Secrets

By CHARLES DUHIGG FEB. 16, 2012

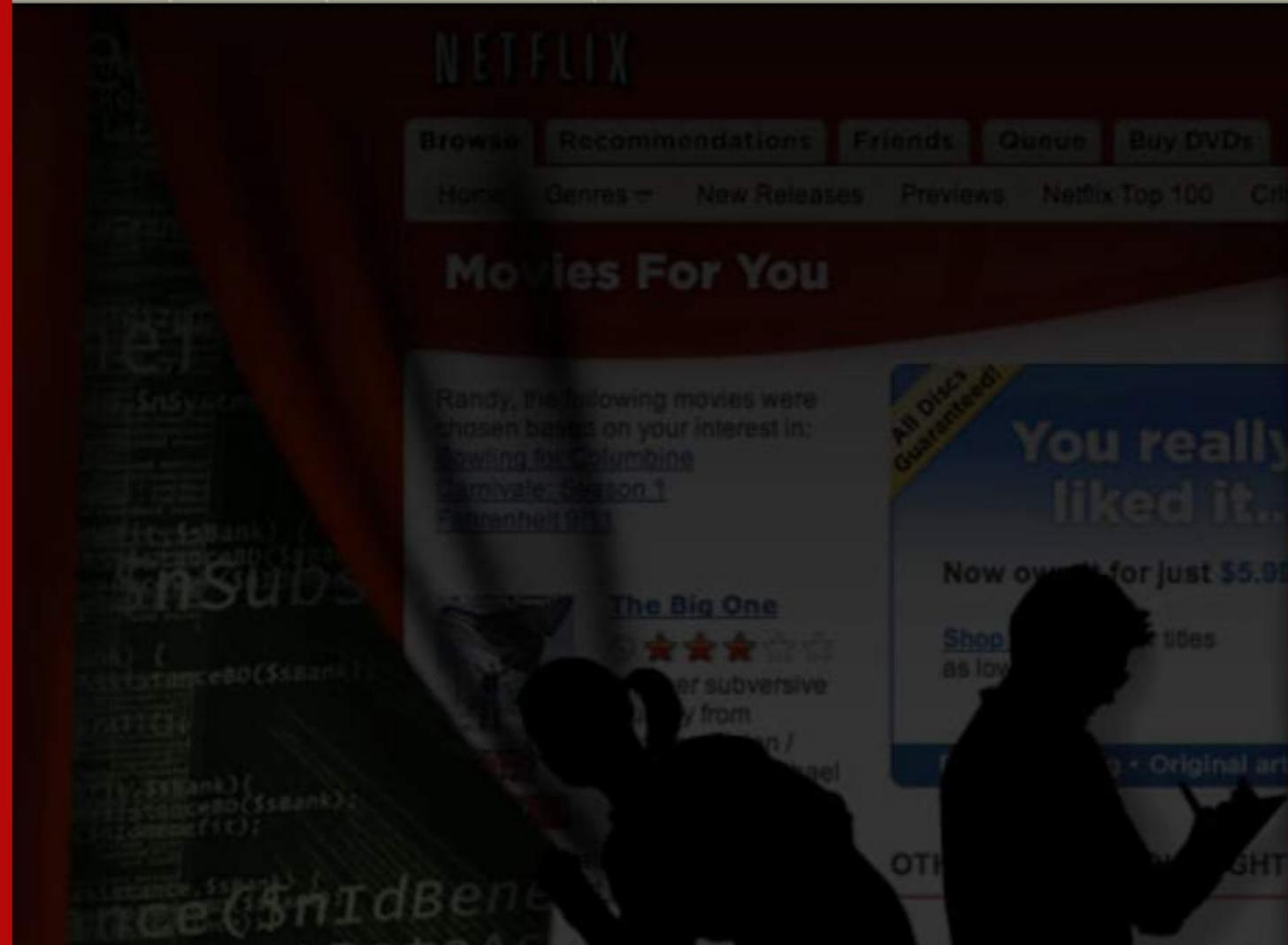




Netflix Prize

COMPLETED

[Home](#) | [Rules](#) | [Leaderboard](#) | [Update](#)



Congratulations!

The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.

On September 21, 2009 we awarded the \$1M Grand Prize to team "BellKor's Pragmatic Chaos". Read about [their algorithm](#), checkout team scores on the [Leaderboard](#), and join the discussions on the [Forum](#).

We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love.

Interventions

Facebook fiasco: was Cornell's study of 'emotional contagion' an ethics breach?

A covert experiment to influence the emotions of more than 600,000 people. A major scientific journal behaving like a rabbit in the headlights. A university in a PR tailspin



Exclusion

- Focus on data from one domain/demographic
- State-of-the-art models perform worse for younger people (Hovy and Søgaard 2015), minorities (Blodgett et al. 2016), and dialect speakers (Jurgens et al., 2017)

Which word is more likely to
be used by a female ?

**Giggle –
Laugh**

(Preotiuc-Pietro et al. '16)

Which word is more likely to
be used by a female ?

**Giggle –
Laugh**

(Preotiuc-Pietro et al. '16)

Which word is more likely to
be used by a female ?

**Brutal –
Fierce**

(Preotiuc-Pietro et al. '16)

Which word is more likely to
be used by a female ?

Brutal –
Fierce

(Preotiuc-Pietro et al. '16)

Which word is more likely to
be used by an **older** person?

**Impressive –
Amazing**

(Preotiuc-Pietro et al. ‘16)

Which word is more likely to
be used by an **older person**?

**Impressive –
Amazing**

(Preotiuc-Pietro et al. ‘16)

Which word is more likely to be used by
a person of higher occupational class?

Suggestions – Proposals

(Preotiuc-Pietro et al. ‘16)

Which word is more likely to be used by
a person of higher occupational class?

Suggestions – Proposals

(Preotiuc-Pietro et al. '16)

Particular biases in training data can affect groups who use certain language

**Giggle –
Laugh**

**Suggestions –
Proposals**

**Impressive –
Amazing**

**Brutal –
Fierce**

Case Study: Language Identification

McNamee, P., “Language identification: *a solved problem* suitable for undergraduate instruction” Journal of Computing Sciences in Colleges 20(3) 2005.

McNamee, P., “Language identification: a solved problem suitable for undergraduate instruction” Journal of Computing Sciences in Colleges 20(3) 2005.

“This paper describes [...] how even the most simple of these methods **using data obtained from the World Wide Web achieve accuracy approaching 100%** on a test suite comprised of ten European languages”

Whose language are we identifying?

Whose language are we identifying?



The Royal Family

@RoyalFamily

Follow

Taking place this week on the river Thames is
'Swan Upping' – the annual census of the
swan population on the Thames.

Whose language are we identifying?



The Royal Family ✅

@RoyalFamily

Follow

Taking place this week on the river Thames is 'Swan Upping' – the annual census of the swan population on the Thames.



da'Rah-zingSun

@TIME7SS

Follow

@kimguilfoyle prblm I hve wit ur reportng is its 2 literal, evry1 knos pple tlk diffrrnt evrywhere, u kno wut she means jus like we do!

Whose language are we identifying?



The Royal Family ✅

@RoyalFamily

Follow

Taking place this week on the river Thames is 'Swan Upping' – the annual census of the swan population on the Thames.



da'Rah-zingSun

@TIME7SS

Follow

@kimguilfoyle prblm I hve wit ur reportng is its 2 literal, evry1 knos pple tlk diffrrnt evrywhere, u kno wut she means jus like we do!



Mooktar

@bossmukky

Follow

"@Ecstatic_Mi: @bossmukky Ebi like say I wan dey sick sef wlh 'Flu' my whole body dey weak"uw gee...



Ebenezer·

@Physique_cian

Follow

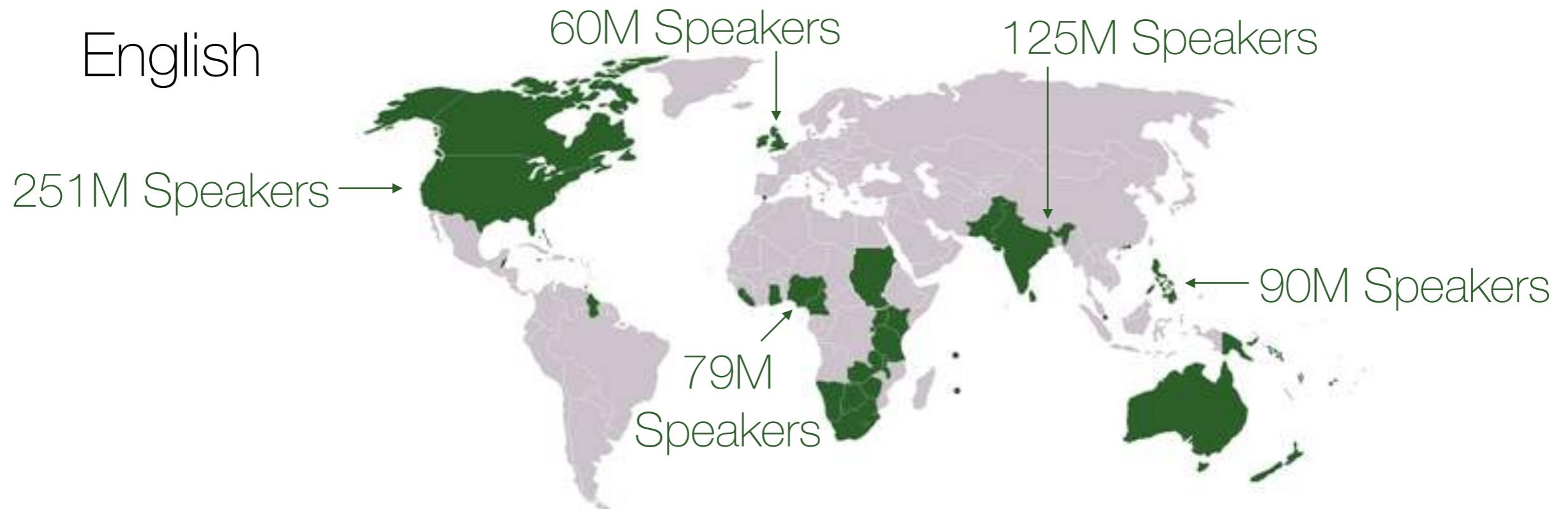
@Tblazeen R u a wizard or wat gan sef : in d mornin- u tweet, afternoon - u tweet, nyt gan u dey tweet.beta get ur IT placement wiv twitter

Global platforms attract global diversity in a language

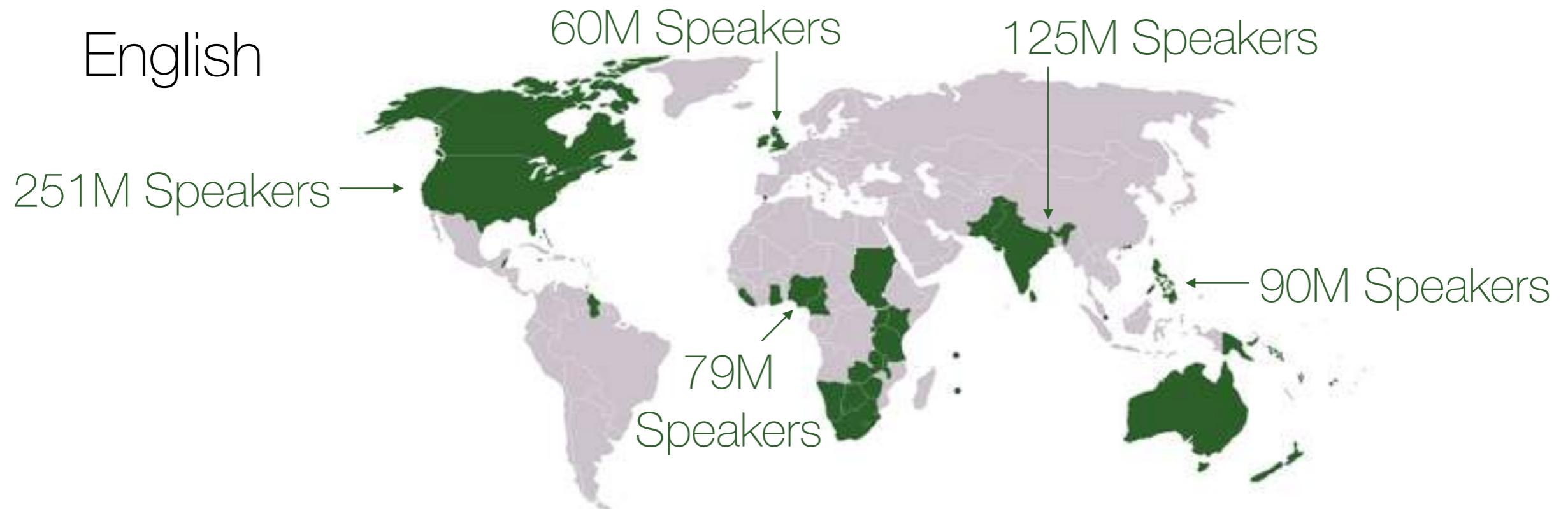
English



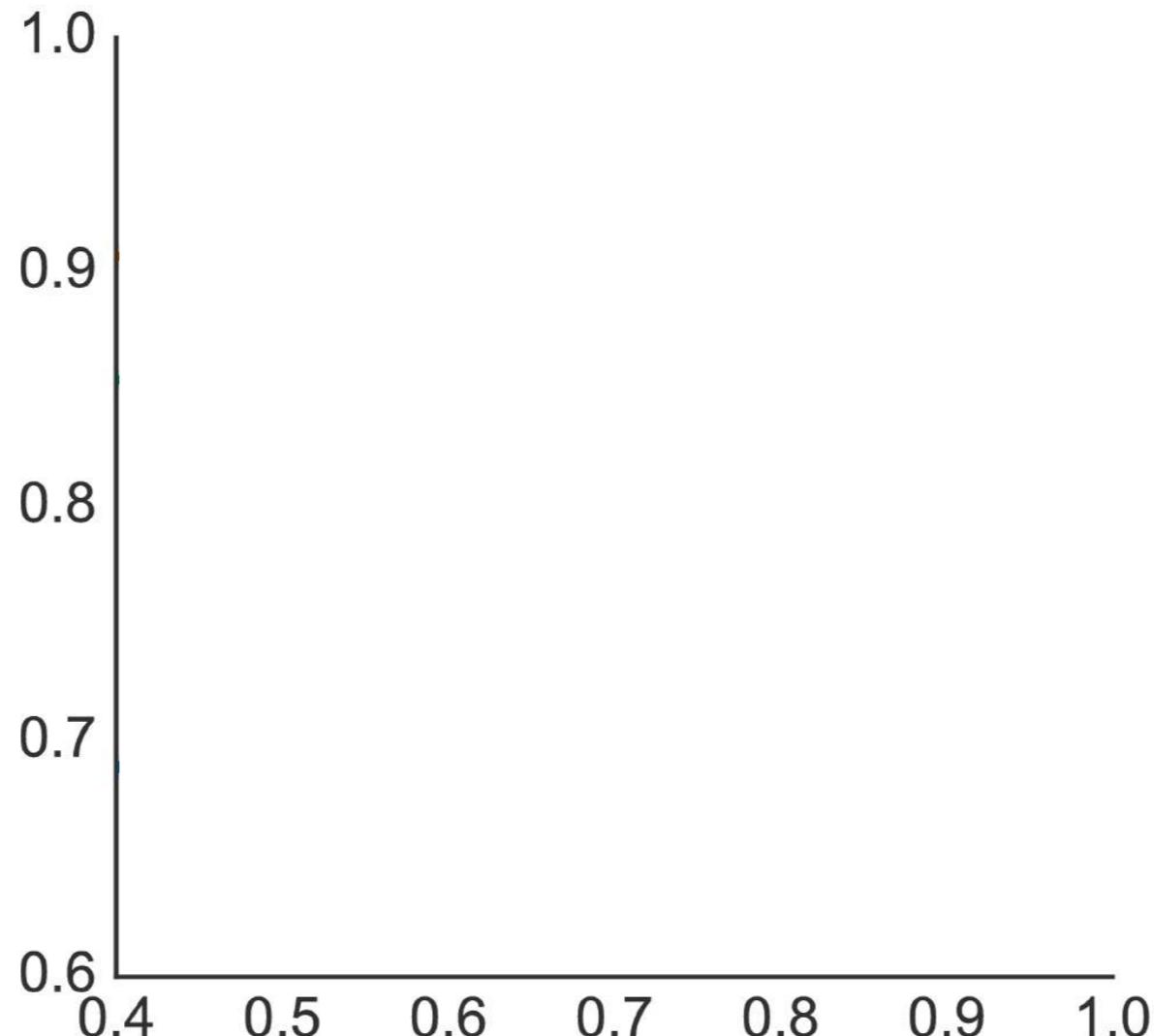
Global platforms attract global diversity in a language



Global platforms attract global diversity in a language



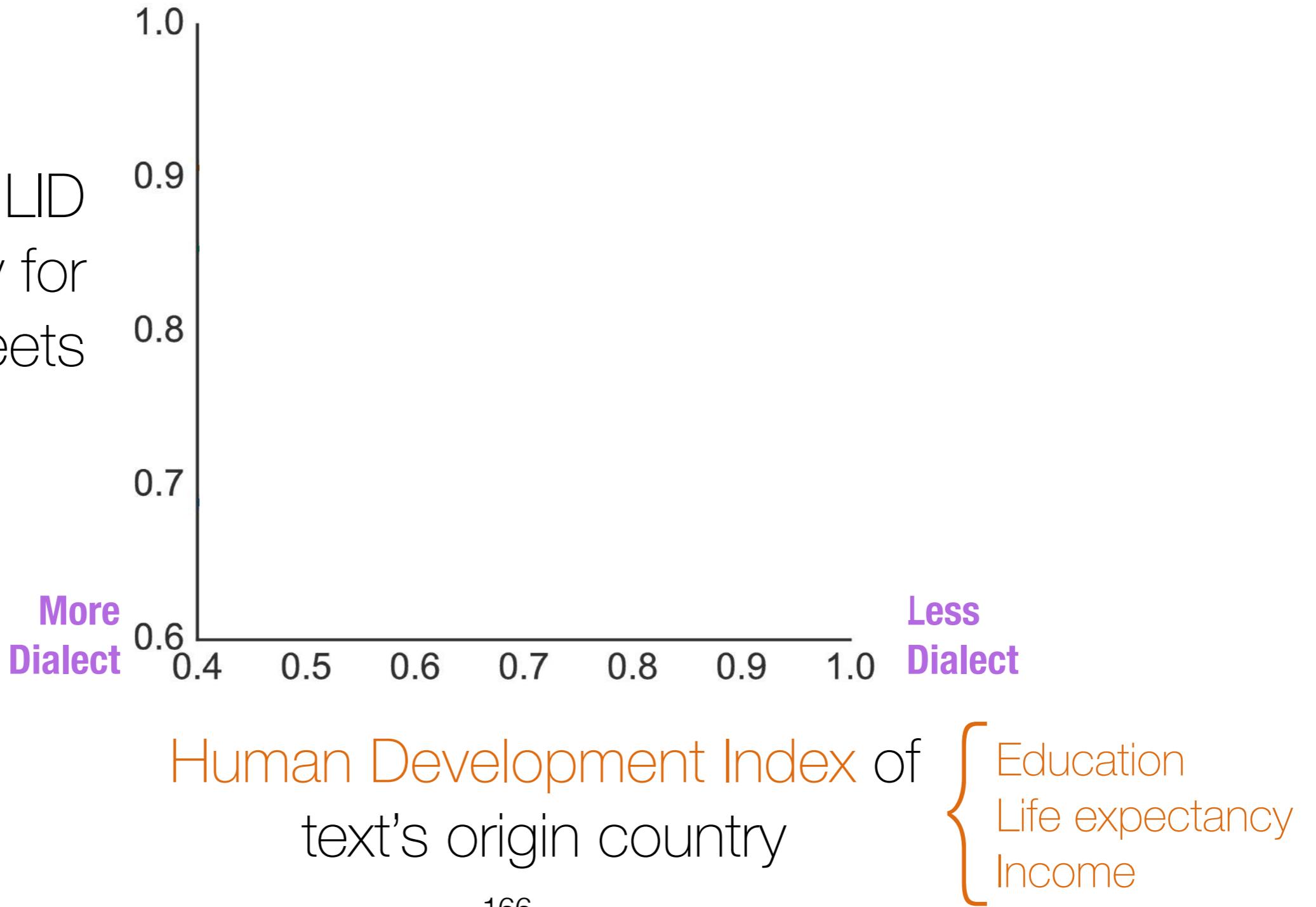
Estimated LID
accuracy for
English tweets



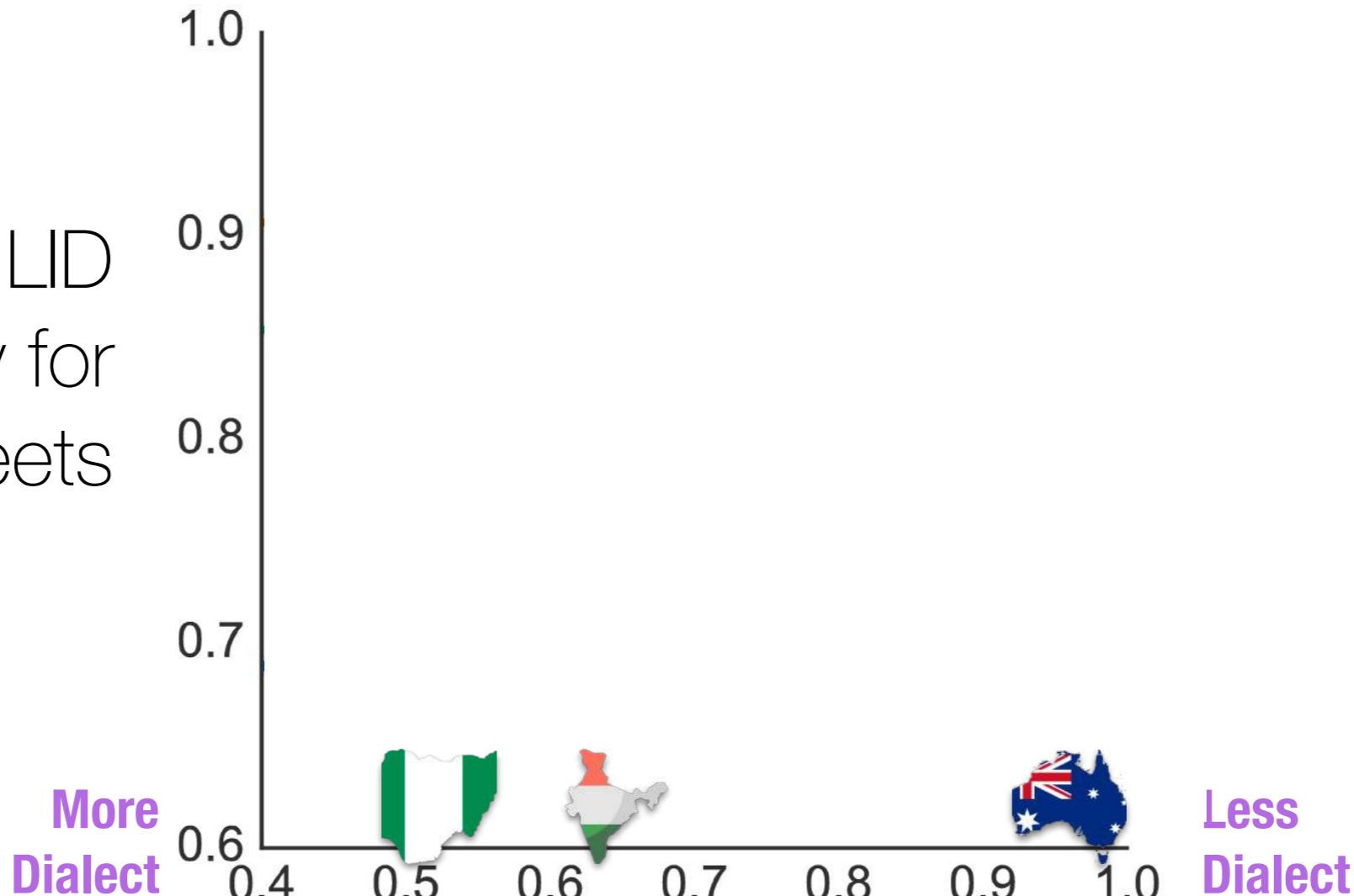
Human Development Index of
text's origin country

{ Education
Life expectancy
Income

Estimated LID
accuracy for
English tweets



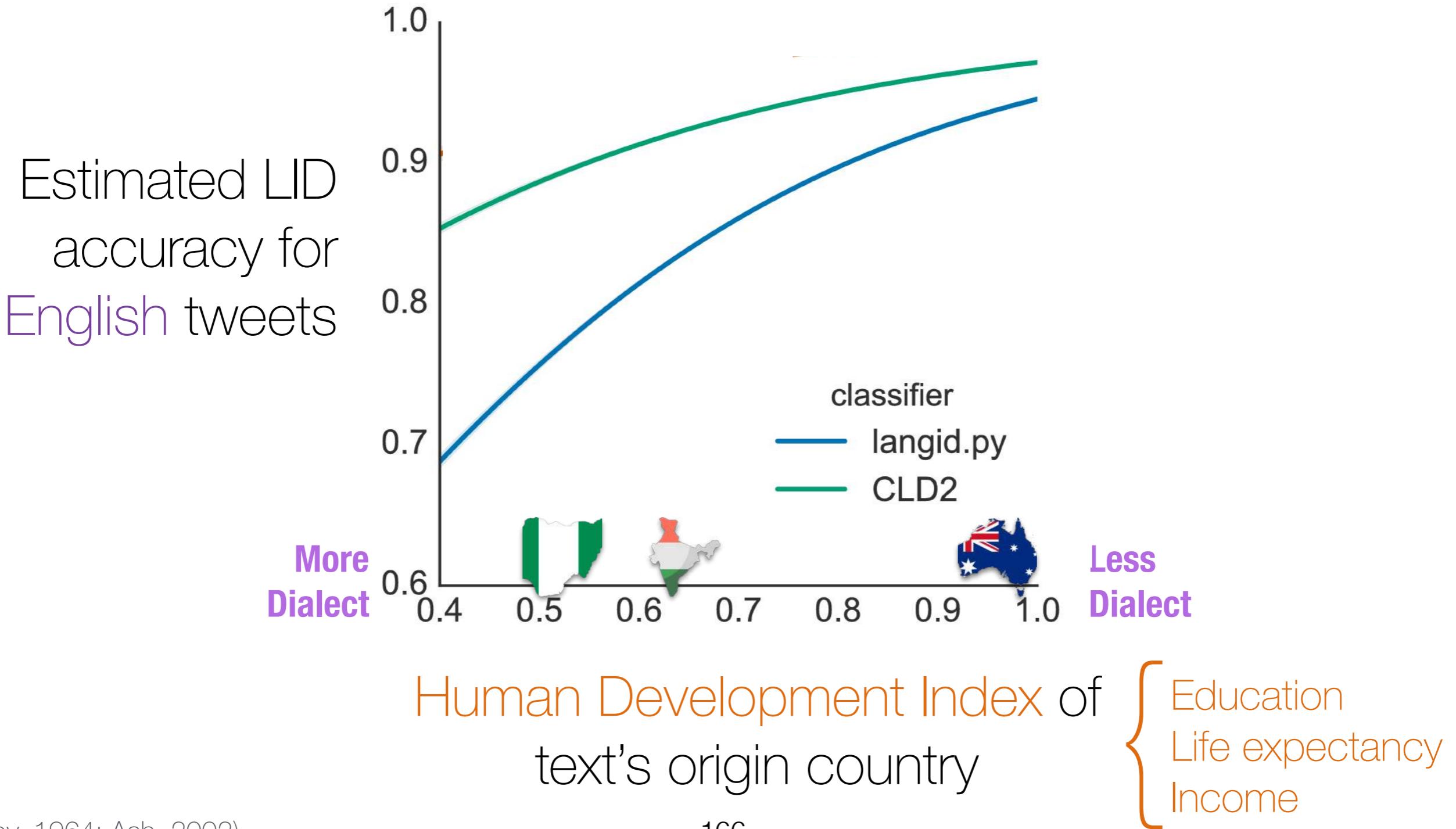
Estimated LID
accuracy for
English tweets



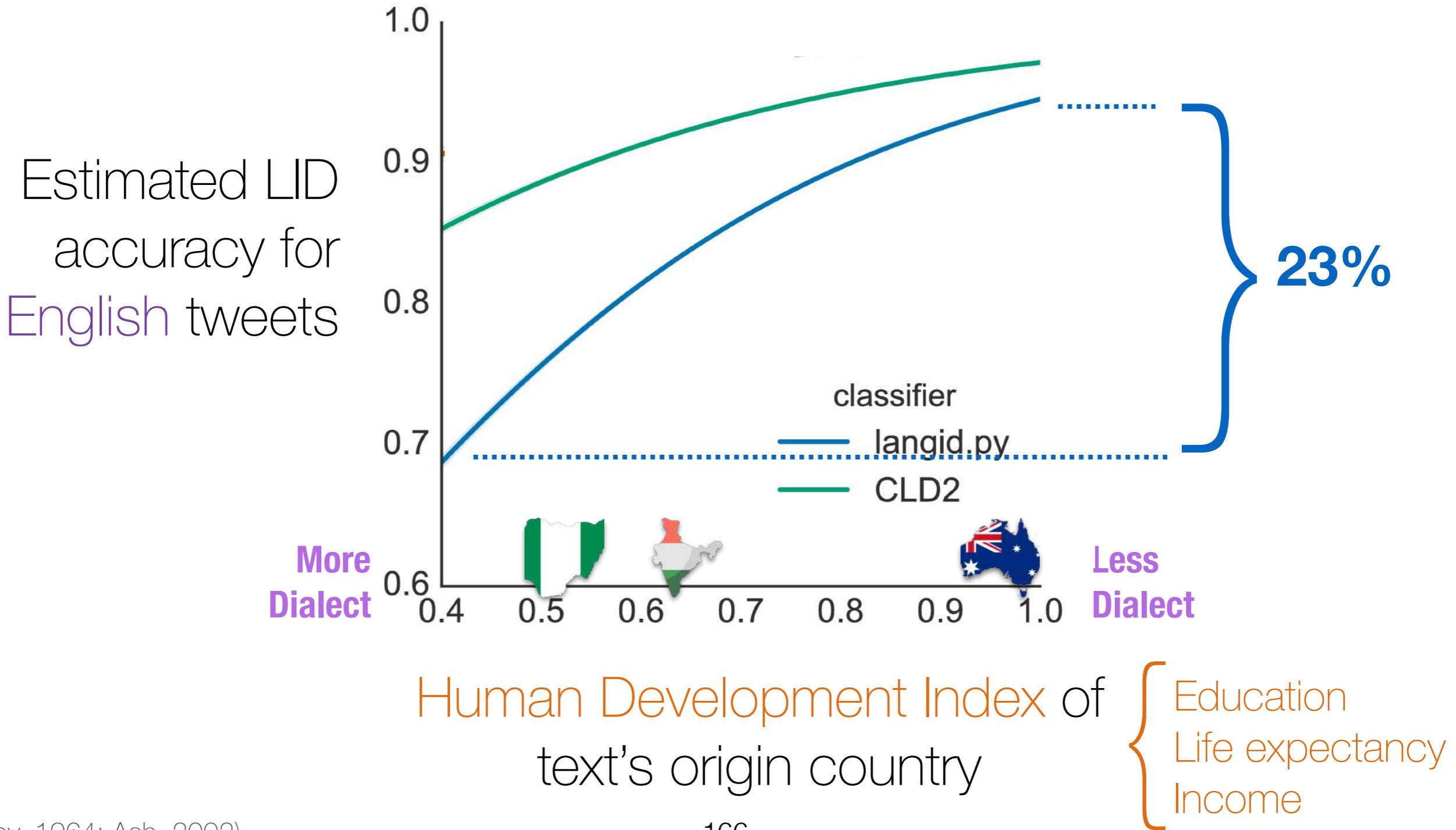
Human Development Index of
text's origin country

{ Education
Life expectancy
Income

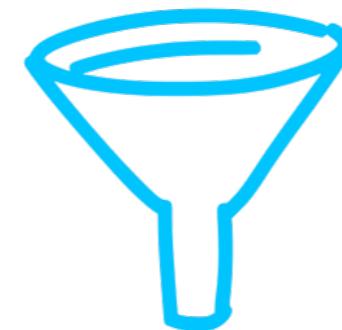
Current language detection methods perform significantly worse in less-developed countries



Current language detection methods perform significantly worse in less-developed countries



Practical Motivation: Epidemic Detection



Keyword Filter

“flu”, “sick”

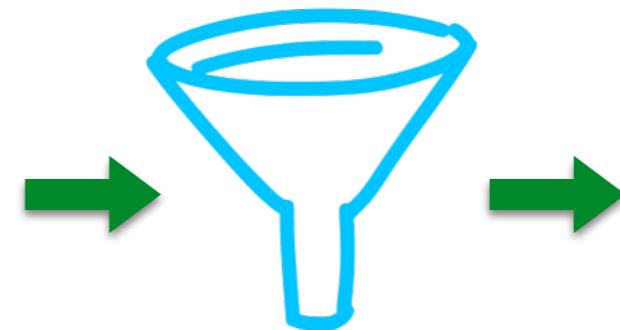
NLP

Which symptoms?

Practical Motivation: Epidemic Detection



**Language
Detection**

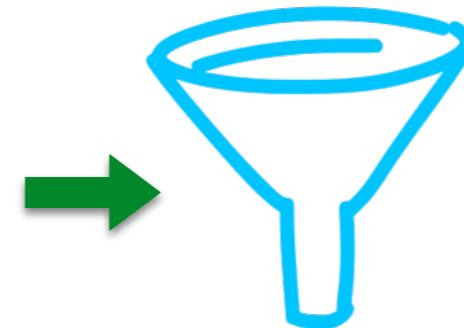


Keyword Filter
“flu”, “sick”



NLP
Which symptoms?

Practical Motivation: Epidemic Detection

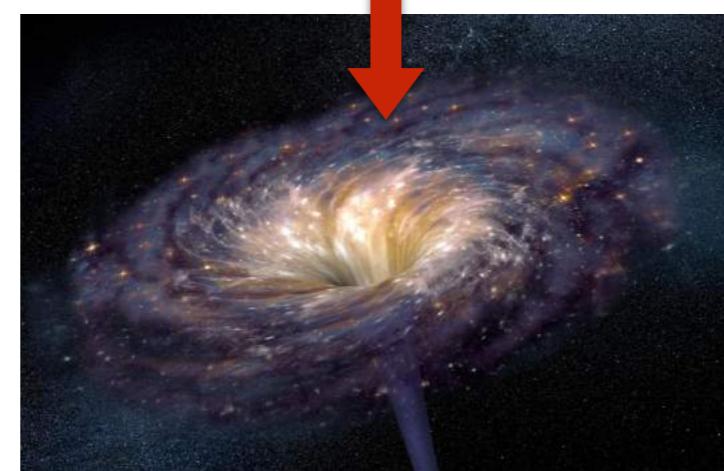


**Language
Detection**

Keyword Filter
“flu”, “sick”

NLP
Which symptoms?

non-English



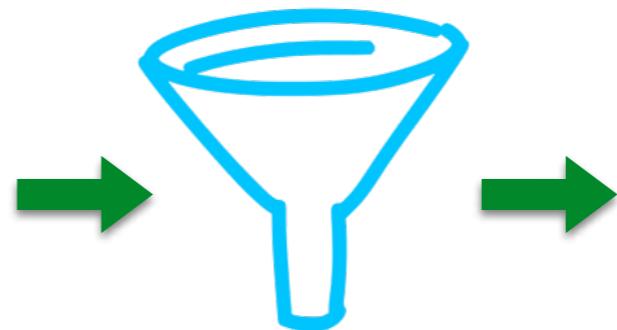
Practical Motivation: Epidemic Detection



Brooke
@Brookiepoo134

got the flu over the weekend and I didn't know until today, & I somehow managed to give it to FIVE of my friends!!!!!!

Follow



**Language
Detection**

non-English



Keyword Filter
“flu”, “sick”

NLP
Which symptoms?

Practical Motivation: Epidemic Detection



Brooke
@Brookiepoo134

got the flu over the weekend and I didn't know until today, & I somehow managed to give it to FIVE of my friends!!!!!!

Nana Rayne
@Nana_Rayne

Like serious dis flu nor dey wan go oooo.... Sick

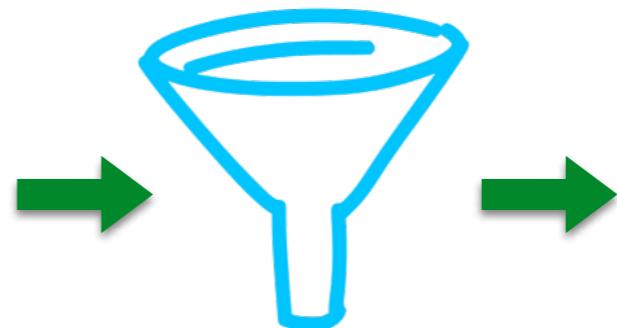
Venus
@christinedarvin

@_rkptrnte hindi ko alam babe eh, absent ako kanina I'm sick rn hahaha 😊🙌

Follow

Follow

Follow



**Language
Detection**

Keyword Filter

“flu”, “sick”

NLP

Which symptoms?

non-English



Practical Motivation: Epidemic Detection



Brooke
@Brookiepoo134

got the flu over the weekend and I didn't know until today, & I somehow managed to give it to FIVE of my friends!!!!!!

Nana Rayne
@Nana_Rayne

Like serious dis flu nor dey wan go oooo.... Sick

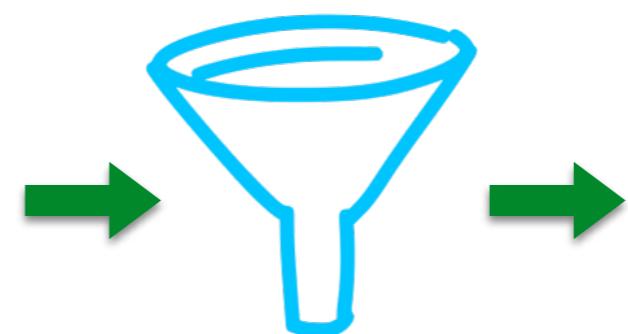
Venus
@christinedarvin

@_rkptrnte hindi ko alam babe eh, absent ako kanina I'm sick rn hahaha 😊🙌

Follow

Follow

Follow



**Language
Detection**

non-English?



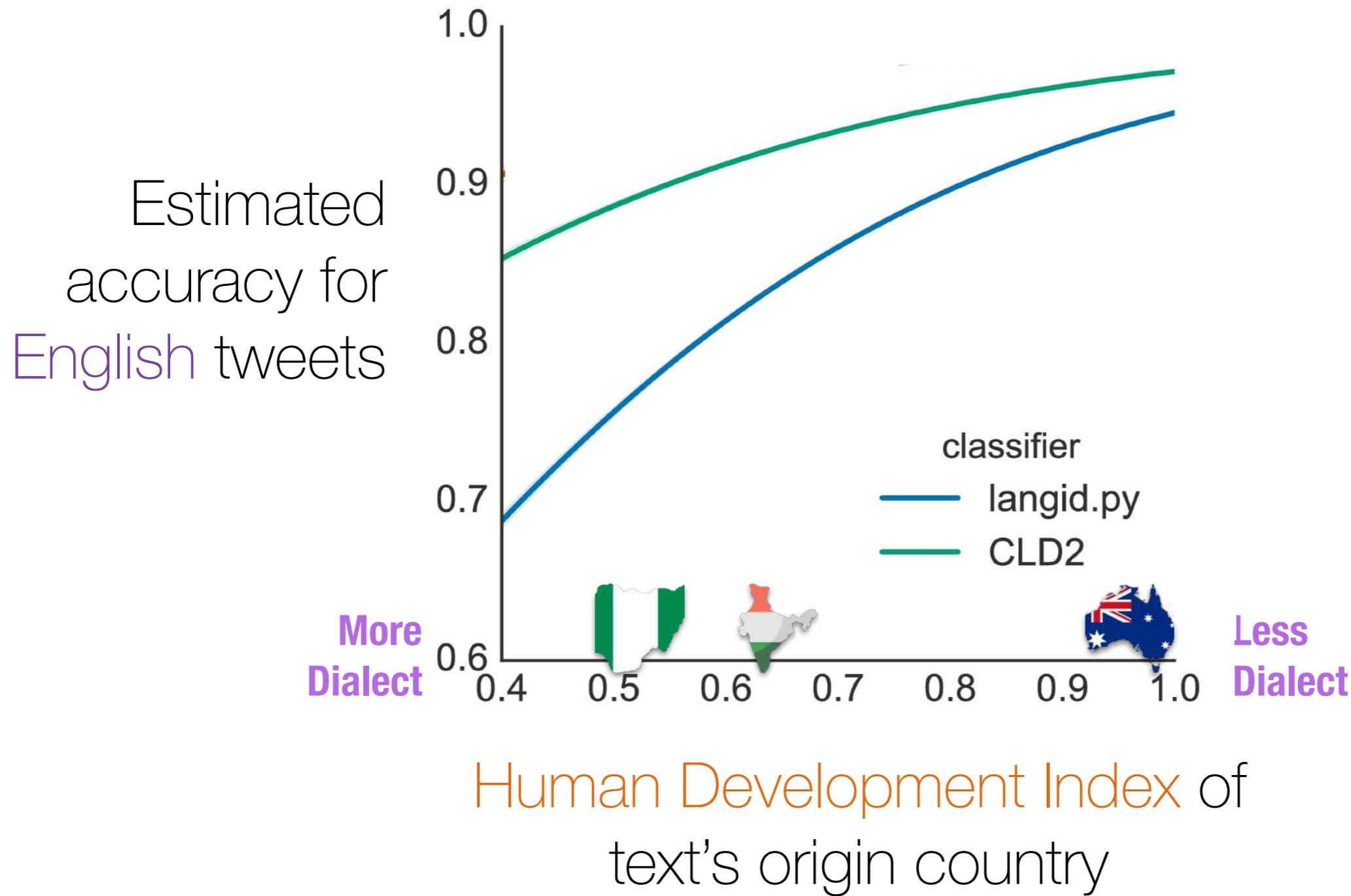
Keyword Filter

"flu", "sick"

NLP

Which symptoms?

Current language detection methods perform significantly worse in less-developed countries

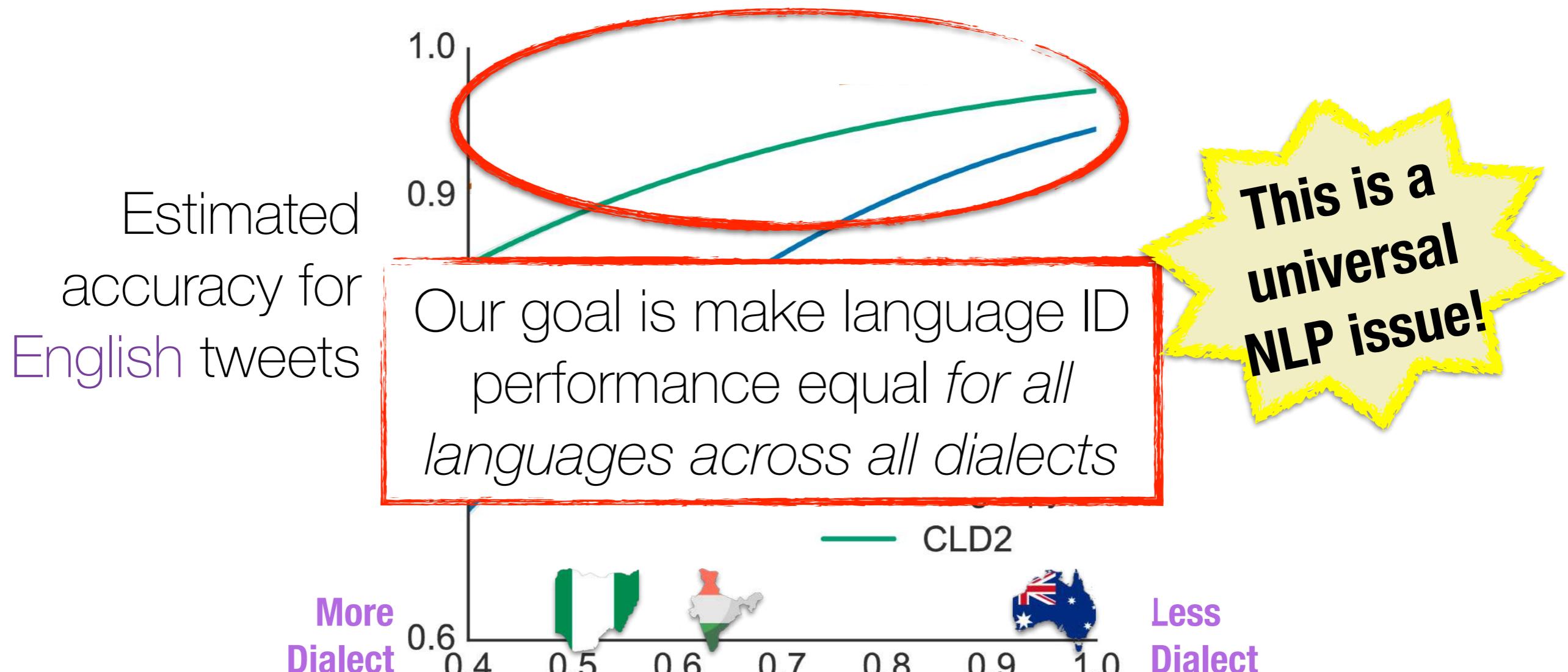


Current language detection methods perform significantly worse in less-developed countries



Human Development Index of
text's origin country

Current language detection methods perform significantly worse in less-developed countries



Human Development Index of text's origin country

Key Problems: Current methods struggle in the **global** setting because

Key Problems: Current methods struggle in the **global** setting because

Data: No corpora that captures global variation in lexicon and dialect



Nana Rayne
@Nana_Rayne

Follow

Like serious dis flu nor dey wan go oooo.... Sick

Key Problems: Current methods struggle in the **global** setting because

Data: No corpora that captures global variation in lexicon and dialect



Nana Rayne
@Nana_Rayne

Follow

Like serious dis flu nor dey wan go oooo.... Sick

Model: makes simplistic assumptions about how multilinguals communicate



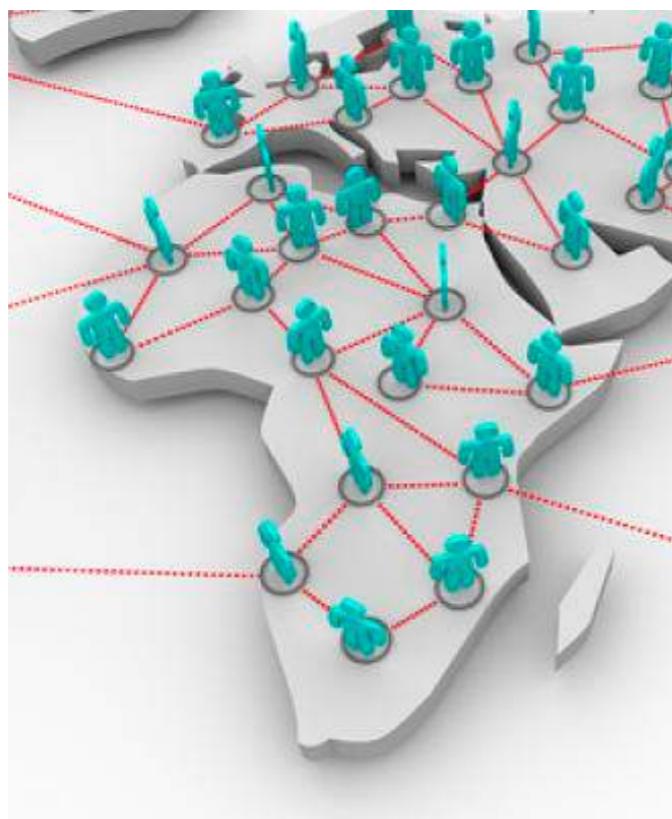
Venus
@christinedarvin

Follow

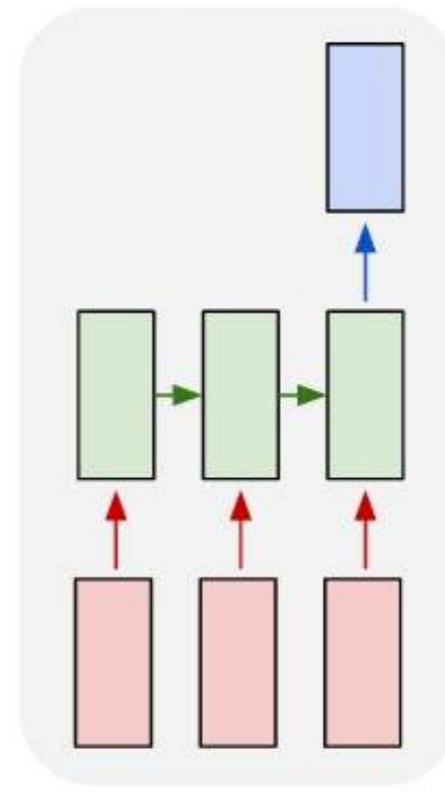
@_rkpntrnte hindi ko alam babe eh, absent ako kanina I'm sick rn hahaha 😊🙌

Our approach

Better social representation
through network-based
sampling

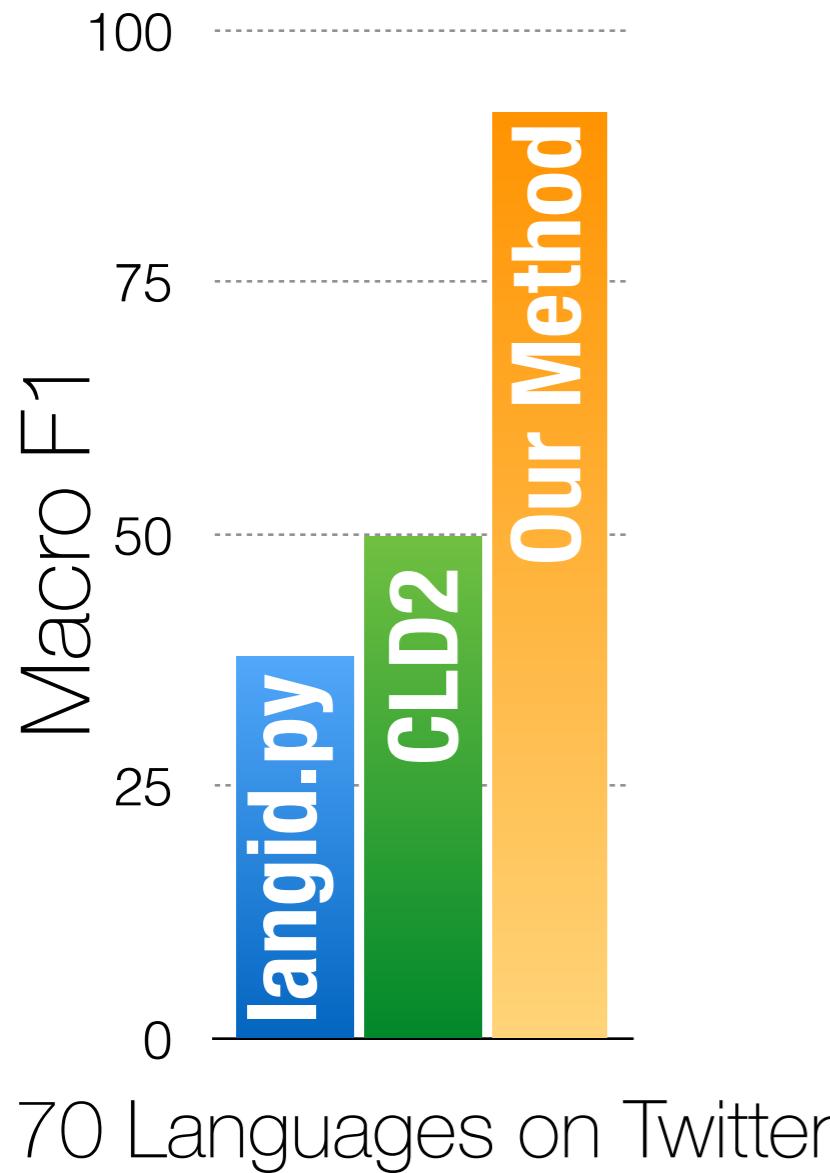


NLP methodologies
capable of handling
linguistic variation

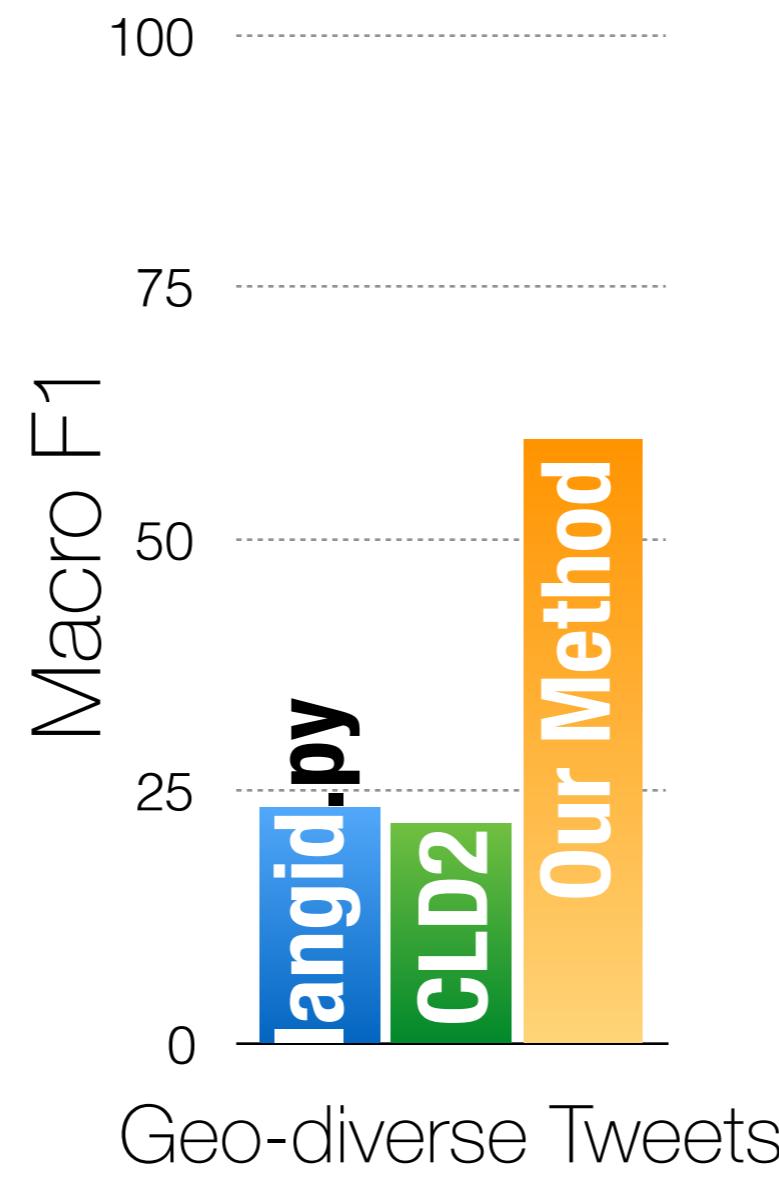
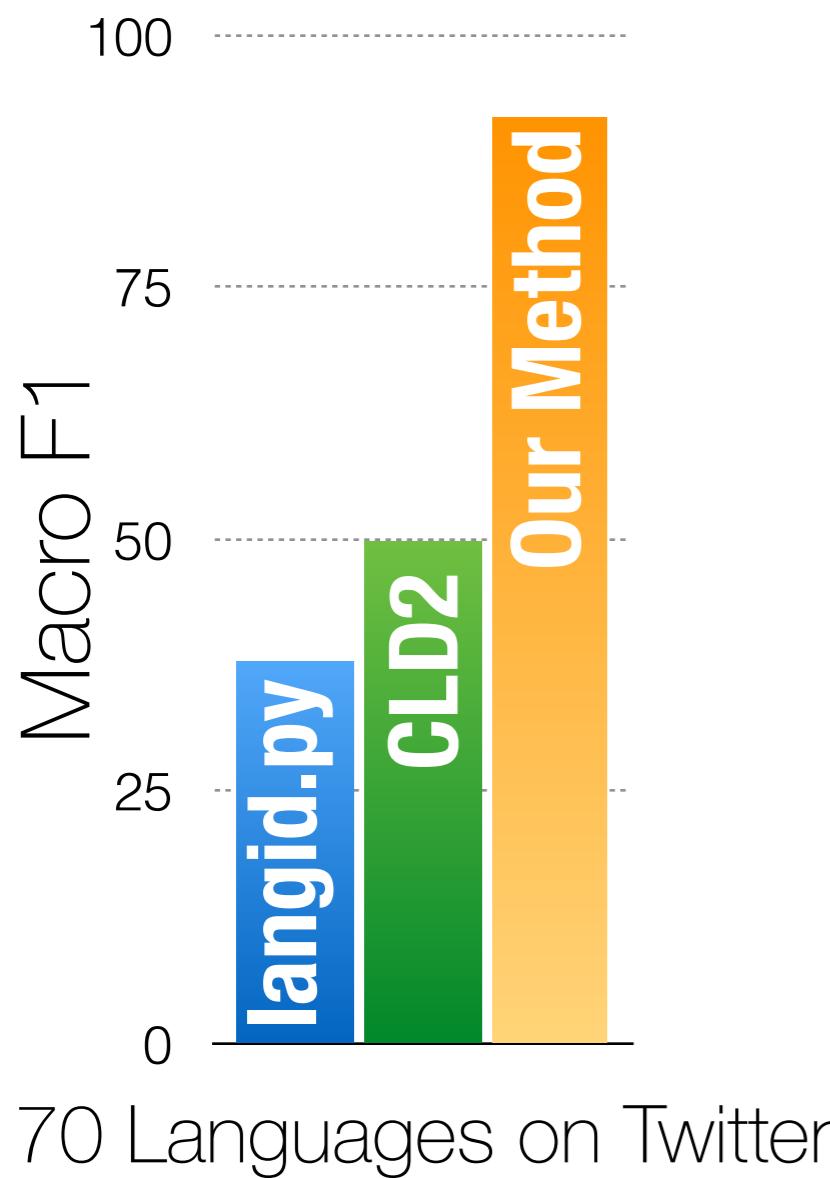


Equilid vs off-the-shelf

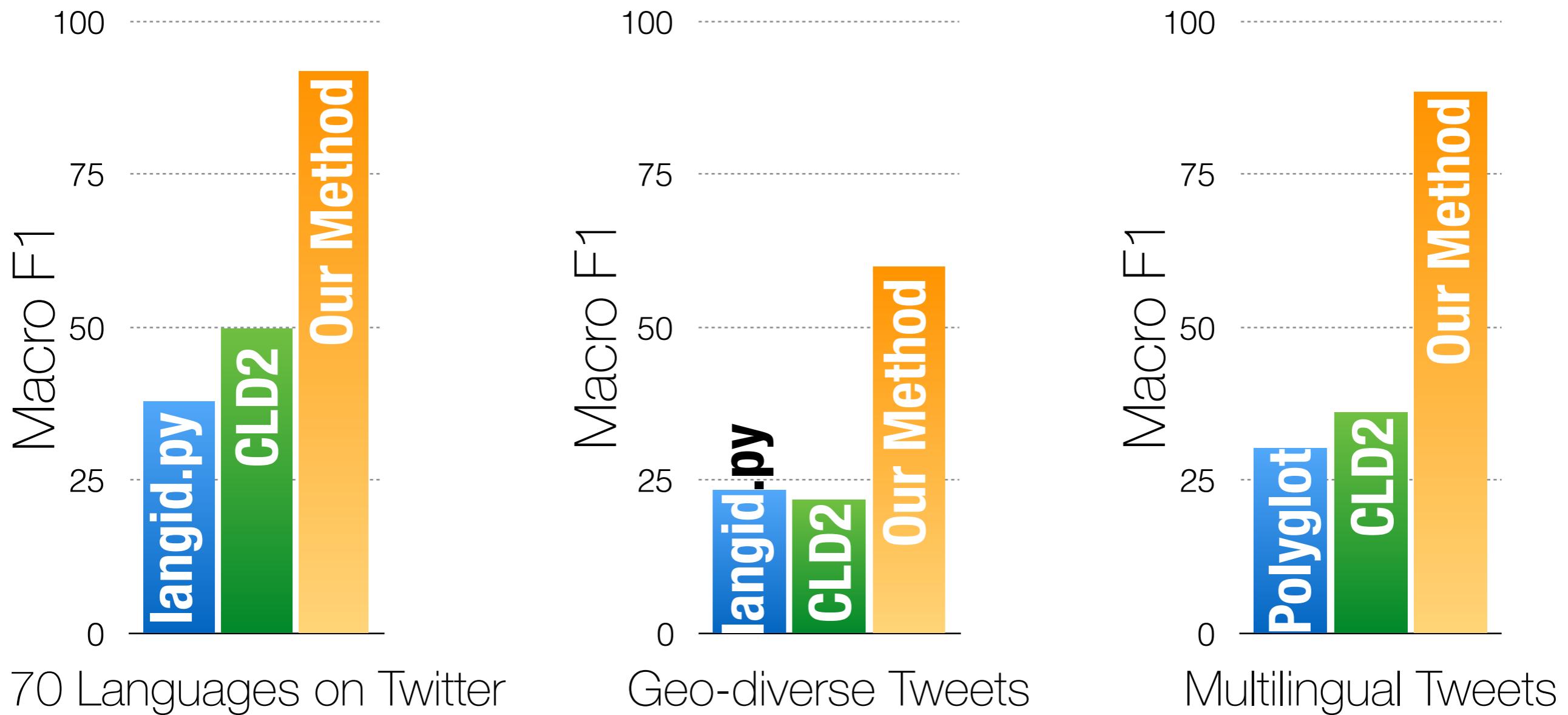
Equilid vs off-the-shelf



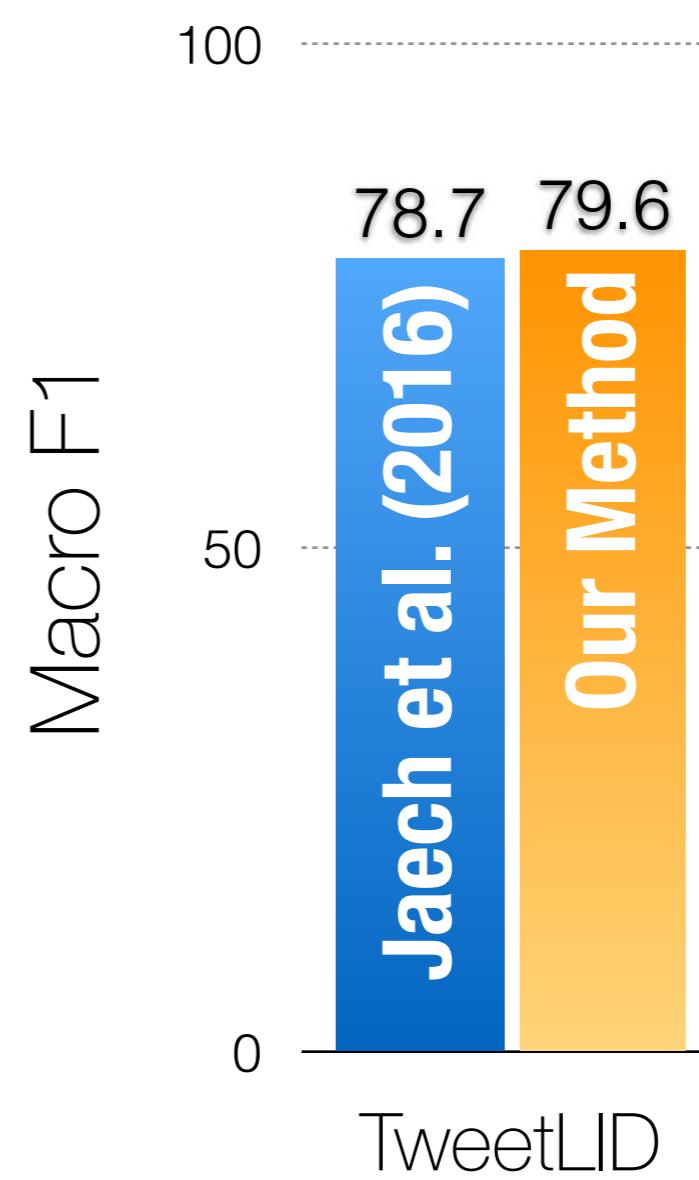
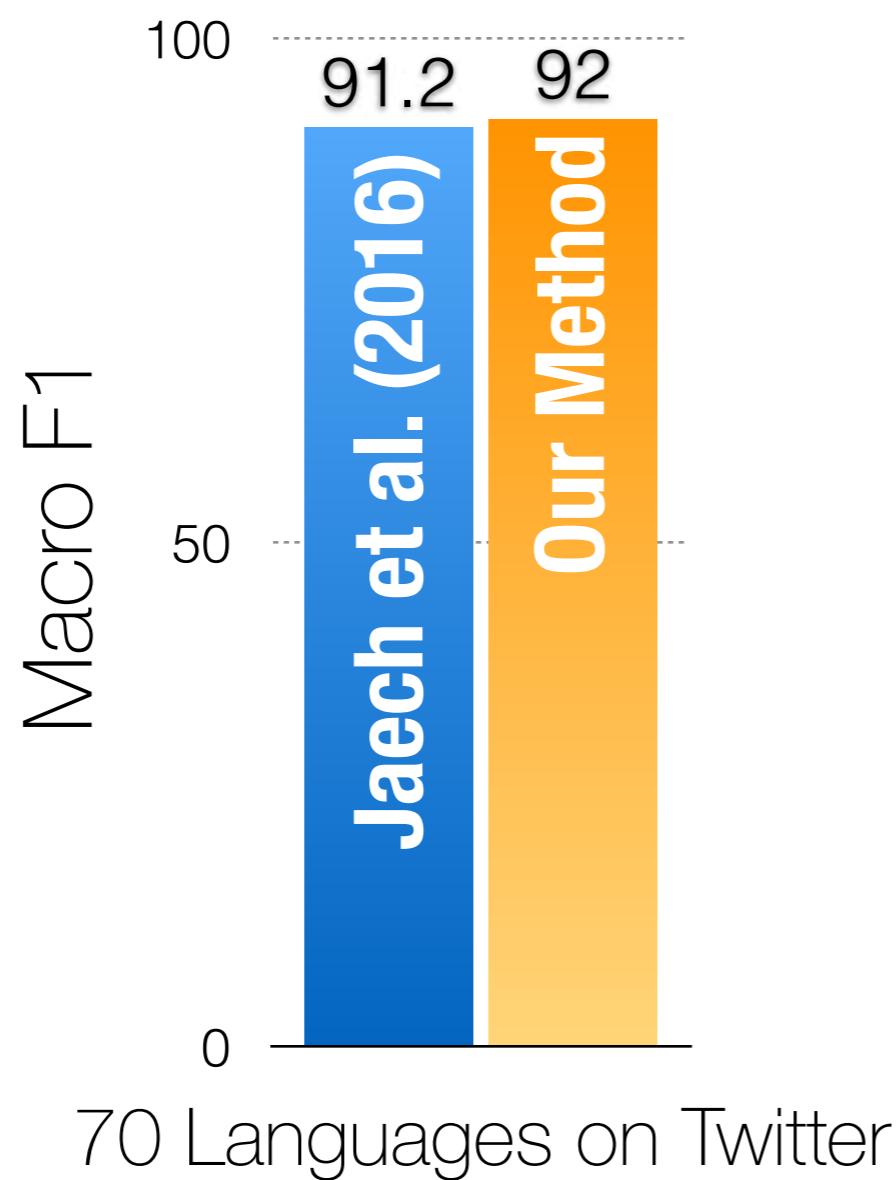
Equilid vs off-the-shelf



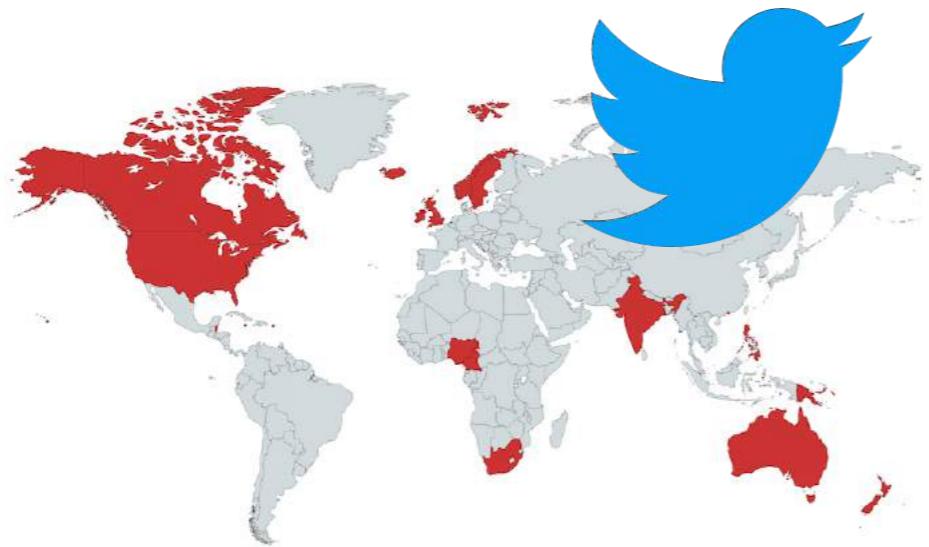
Equilid vs off-the-shelf



Equilid even outperforms system
specifically tuned for each dataset

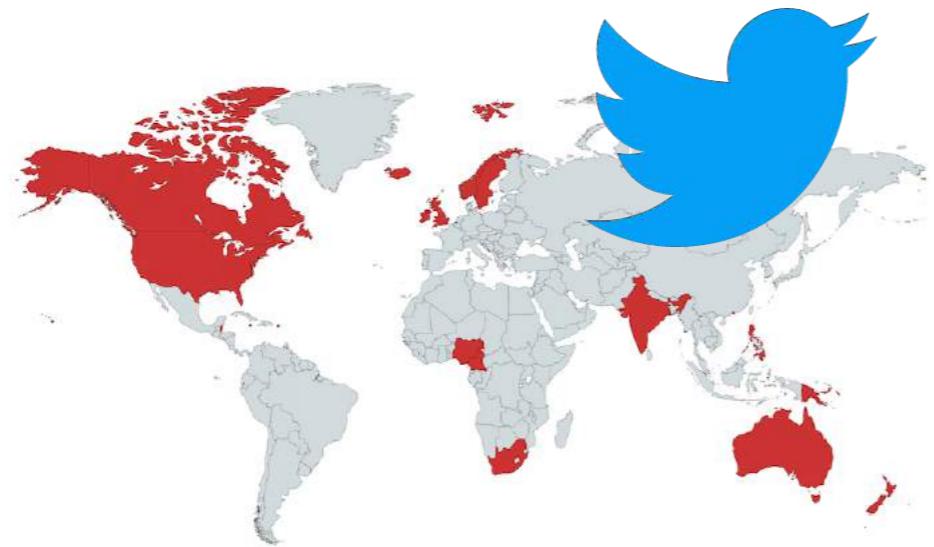


Case Study: Do our solutions provide socially-equitable language identification for health-related queries?



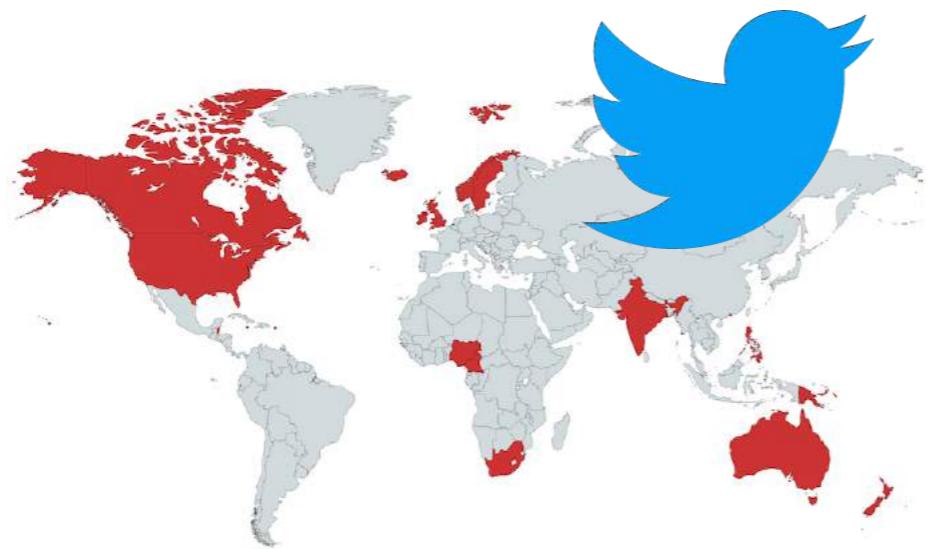
1M Tweets with any of 385 English terms from established lexicons for influenza, psychological well-being, and social health

Case Study: Do our solutions provide socially-equitable language identification for health-related queries?



1M Tweets with any of 385 English terms from established lexicons for influenza, psychological well-being, and social health

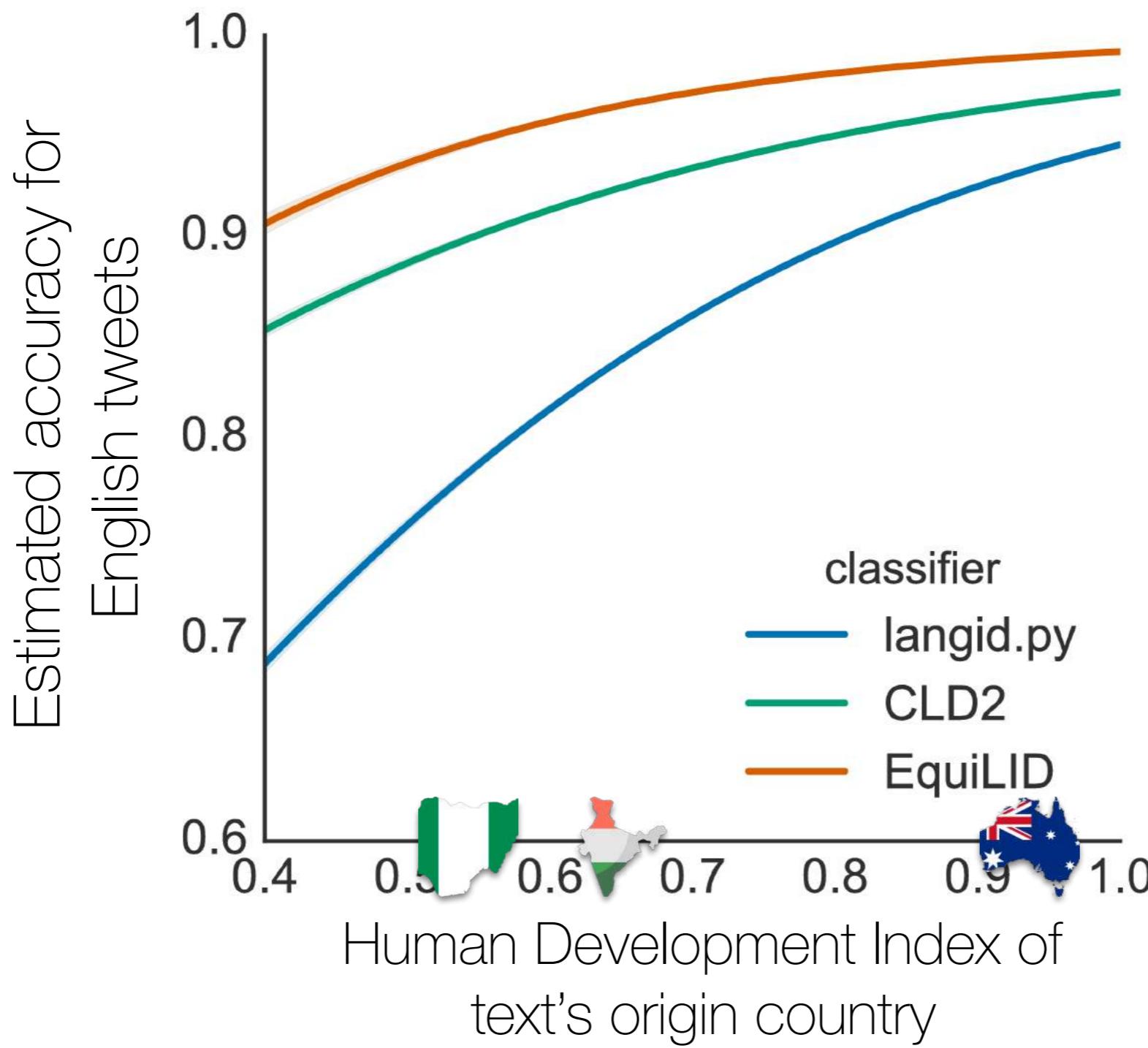
Case Study: Do our solutions provide socially-equitable language identification for health-related queries?



1M Tweets with any of 385 **English** terms from established lexicons for **influenza**, **psychological well-being**, and **social health**

Task: does the language identification system recognize every tweet as **English**?

Equilid raises the bar for socially-equitable language identification



Race too may lead to exclusion

	AAE	White-Aligned
<i>langid.py</i>	13.2%	7.6%
Twitter-1	8.4%	5.9%
Twitter-2	24.4%	17.6%

	Parser	AA	Wh.	Difference
SyntaxNet	64.0 (2.5)	80.4 (2.2)	16.3 (3.4)	
CoreNLP	50.0 (2.7)	71.0 (2.5)	21.0 (3.7)	

Table 3: Proportion of tweets in AA- and white-aligned corpora classified as non-English by different classifiers. (§4.1)

Language identification

Dependency parsing

Blodgett et al. (2016), "Demographic Dialectal Variation in Social Media: A Case Study of African-American English" (EMNLP)

Overgeneralization

- Managing and communicating the uncertainty of our predictions
- Is a false answer worse than no answer?

Dual Use

- Authorship attribution (author of *Federalist Papers* vs. author of ransom note vs. author of political dissent)
- Fake review detection vs. fake review generation
- Censorship evasion vs. enabling more robust censorship

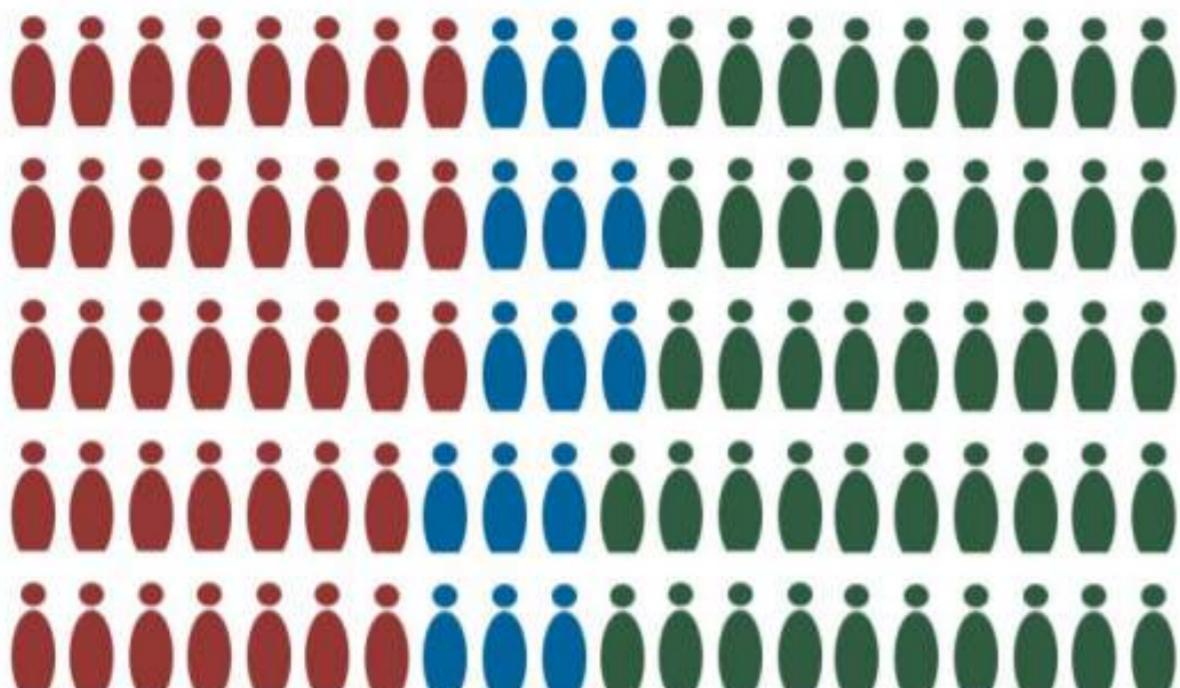


Doing something about
unethical behavior:
Combatting Antisocial Behavior

Exposure to harassment



Respondents were asked if they had **personally experienced harassment**. Out of 2,495 that responded to this question :

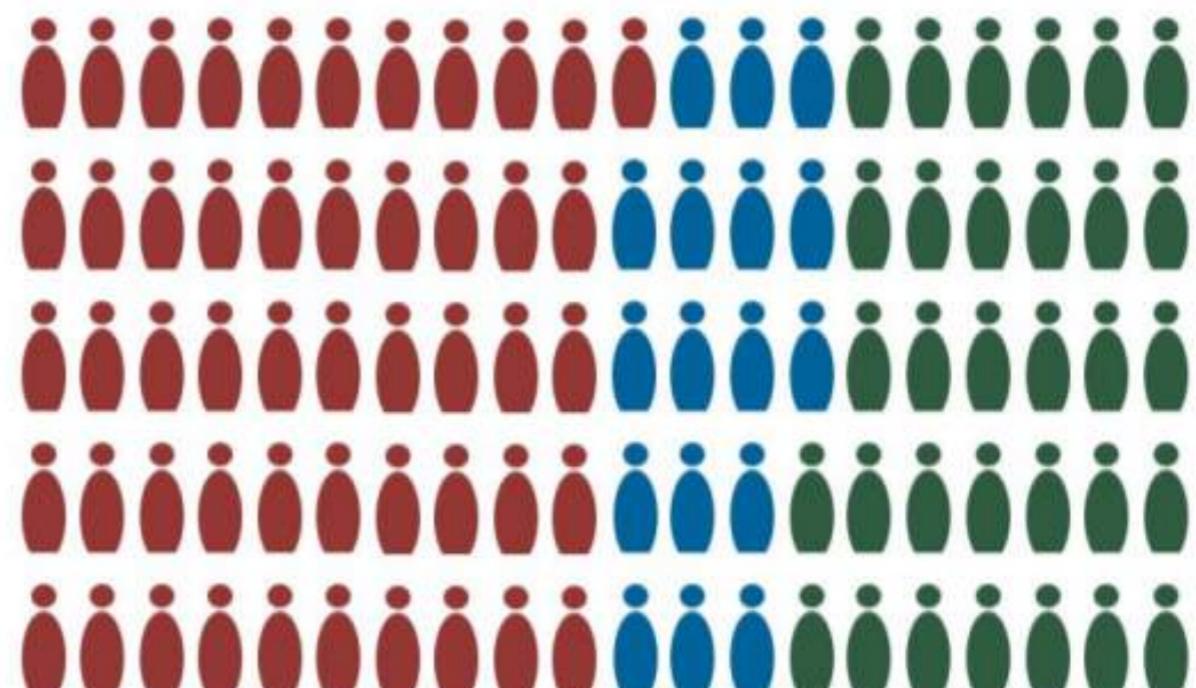


38% said yes

16% were
unsure

47% said no

Respondents were asked if they had **witnessed the harassment of others**. Out of 2,078 that responded to this question:

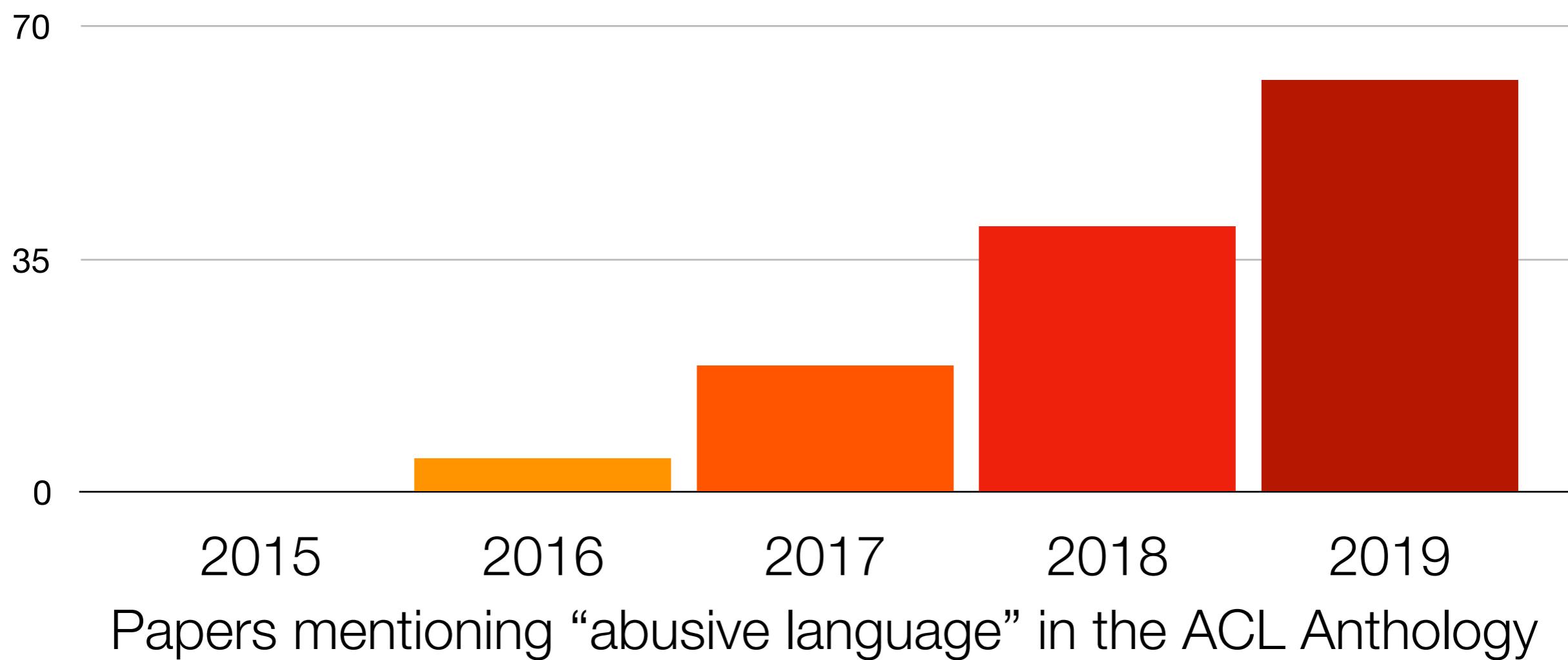


51% said yes

17% were
unsure

32% said no

The NLP community has become
serious about addressing online abuse



What is Hate Speech?

“any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic”

(Nockleby, J. Encyclopedia of the American Constitution 2000)

What is Hate Speech?

“any communication that disparages **a person or a group** on the basis of some characteristic such as **race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic**”

(Nockleby, J. Encyclopedia of the American Constitution 2000)

Target

What is Hate Speech?

“language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group”

(Davidson et al., Automated Hate Speech Detection and the Problem of Offensive Language, *ICWSM* 2017)

What is Hate Speech?

*“language that is used **to expresses hatred** towards a targeted group or is **intended to be derogatory, to humiliate, or to insult** the members of the group”*

(Davidson et al., Automated Hate Speech Detection and the Problem of Offensive Language, *ICWSM* 2017)

Intent

What is Hate Speech?

“language that threatens or incites violence”

(Davidson et al., Automated Hate Speech Detection and the Problem of Offensive Language, *ICWSM* 2017)

What is Hate Speech?

*“language that **threatens** or **incites violence**”*

(Davidson et al., Automated Hate Speech Detection and the Problem of Offensive Language, *ICWSM* 2017)

Effect

What is Hate Speech?

“any offense motivated, in whole or in a part, by the offender’s bias against an aspect of a group of people”

(Silva et al., Analyzing the Targets of Hate
in Online Social Media, *ICWSM* 2016)

What is Hate Speech?

*“any offense motivated, in whole or in a part, by the offender’s **bias** against an aspect of a group of people”*

(Silva et al., Analyzing the Targets of Hate in Online Social Media, *ICWSM* 2016)

Reason

How is Hate Speech
Different from Bias?

How is Hate Speech Different from Bias?

Veiled vs Overt?

How is Hate Speech Different from Bias?

POLITICS 12/13/2017 04:00 pm ET | Updated Dec 13, 2017

This Is The Daily Stormer's Playbook

A leaked style guide reveals they're Nazis about grammar (and about Jews).



By Ashley Feinberg



How is Hate Speech Different from Bias?



While racial slurs are allowed/recommended, not every reference to non-white should not be a slur and their use should be based on the tone of the article. Generally, when using racial slurs, it should come across as half-joking - like a racist joke that everyone laughs at because it's true. This follows the generally light tone of the site.

It should not come across as genuine raging vitriol. That is a turnoff to the overwhelming majority of people.

How is Hate Speech Different from Bias?

~~Veiled vs Overt?~~

Intentional

Why is detecting hate speech hard?

Hate speech can be done to anyone by anyone at any platform

Also Computationally

A simple classifier?

- AMT to label existing comments as abusive/non-abusive
- Lexicon+BOW features
- Regular expressions: “you are” “, I hate “ “
- Sentiment
- Brown clusters, embeddings

Will it work?

Why Hate Speech Identification is Hard

Intentional obfuscation of abuse words, short forms etc

- Single character substitution: *nagger* (W&H'12)
- Homophone *joo* (W&H'12) *JOOZ* (NTTMC'16)
- Expanded spelling *j@e@w* (W&H'12)
- Ni99er (NTTMC'16)
- Tokenization *#Woopiuglyniggeratgoldberg* (NTTMC'16)

Why Hate Speech Identification is Hard

Coded language that appears to mean one thing to the general population but has an additional meaning in in-group

- (((Prof. Jurgens)))



- <http://www.diversityinc.com/news/alt-right-trolls-devise-racist-codes-social-media/>

Codewords

Table 1: Some common codewords

Code word	Actual word
Google	Black
Yahoo	Mexican
Skype	Jew
Bing	Chinese
Skittle	Muslim
Butterfly	Gay

Table 2: Top 10 most correlated terms

Term	Pearson correlation coefficient
#MAGA	0.149
#ALTRIGHT	0.140
gas	0.136
(())	0.136
white	0.136
war	0.118
hate	0.100
#MAWA	0.098
destroy	0.083
goy	0.083

Why Hate Speech Identification is Hard

In out of domain lexicons: rap lyrics contains curse words.

Words like *black, jew, women* are used in various contexts more frequently than in hateful speech.

- Keyword spotting will yield false positives

Data Collection is Also Hard

- News outlets and online communities remove this content
- Hard to obtain due to privacy issues
- Possibility to flag content? But part of trolling is to go to non-abusive content and flag it as abusive.
- This is why it is difficult even for companies to identify automatically abusive content even using feedback from users

Annotation has an emotional and mental cost as well on the annotators themselves!



<https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona>

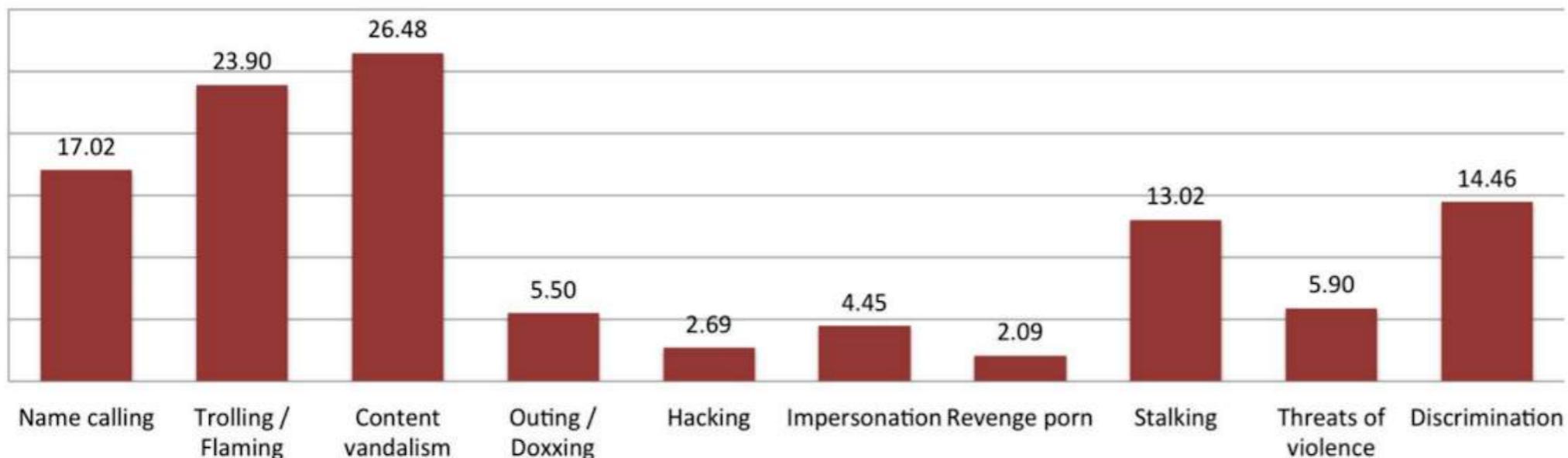
<http://theconversation.com/we-need-to-talk-about-the-mental-health-of-content-moderators-103830>

Hate Speech has Many Flavors

- Umbrella term: Abuse
- Hate speech
- Offensive language
- Sexist and racist
- Aggression
- Toxic language
- Profanity
- Cyberbullying
- Harassment
- Trolling
- Anti-social behavior
- ...

Types of Hate Speech

Types of harassment experienced by occurrence average



Support and Safety Team. Harassment Survey. Wikimedia Foundation, 2015

Of the 3,845 Wikimedia users who participated, 38% of the respondents could confidently recognise that they had been harassed, while 15% were unsure and 47% were confident that they had not been harassed. Similarly, 51% witnessed others being harassed, while 17% were unsure and 32% did not witness harassment.

Categorization

Waseem et al. (2017) **Understanding Abuse: A Typology of Abusive Language Detection Subtasks.** *Workshop on the Workshop on Abusive Language Online*

	<i>Explicit</i>	<i>Implicit</i>
<i>Directed</i>	<p>“Go kill yourself”, “You’re a sad little f*ck” (Van Hee et al., 2015a),</p> <p>“@User shut yo beaner ass up sp*c and hop your f*ggot ass back across the border little n*gga” (Davidson et al., 2017),</p> <p>“Youre one of the ugliest b*tches Ive ever fucking seen” (Kontostathis et al., 2013).</p>	<p>“Hey Brendan, you look gorgeous today. What beauty salon did you visit?” (Dinakar et al., 2012),</p> <p>“(((@User))) and what is your job? Writing cuck articles and slurping Google balls? #Dumbgoogles” (Hine et al., 2017),</p> <p>“you’re intelligence is so breathtaking!!!!!!” (Dinakar et al., 2011)</p>
<i>Generalized</i>	<p>“I am surprised they reported on this crap who cares about another dead n*gger?”, “300 missiles are cool! Love to see um launched into Tel Aviv! Kill all the g*ys there!” (Nobata et al., 2016),</p> <p>“So an 11 year old n*gger girl killed herself over my tweets? ^_^ thats another n*gger off the streets!!” (Kwok and Wang, 2013).</p>	<p>“Totally fed up with the way this country has turned into a haven for terrorists. Send them all back home.” (Burnap and Williams, 2015),</p> <p>“most of them come north and are good at just mowing lawns” (Dinakar et al., 2011),</p> <p>“Gas the skypes” (Magu et al., 2017)</p>

Table 1: **Typology of abusive language.**

What kind of abuse are we
really addressing?

What kind of abuse are we really addressing?



What kind of abuse are we really addressing?



Even though you're an academic,
you're actually cool

● Likely follows community guidelines (0.11)

[DISAGREE?](#)

CUSTOMIZE FEEDBACK



What kind of abuse are we really addressing?



Even though you're an academic,
you're actually cool

Likely follows community guidelines (0.11)

[DISAGREE?](#)

CUSTOMIZE FEEDBACK

Y'all should call my police
department at (555) 555 5555 and
report a hostage situation at my
neighbor's 555 Main St. Ann Arbor.
#Swatting

Likely follows community guidelines (0.29)

[DISAGREE?](#)

CUSTOMIZE FEEDBACK

What kind of abuse are we really addressing?



Even though you're an academic,
you're actually cool

Likely follows community guidelines (0.11)

[DISAGREE?](#)

CUSTOMIZE FEEDBACK



Y'all should call my police
department at (555) 555 5555 and
report a hostage situation at my
neighbor's 555 Main St. Ann Arbor.
#Swatting

Likely follows community guidelines (0.29)

[DISAGREE?](#)

CUSTOMIZE FEEDBACK



Think of anorexia as your secret
weapon! Is food more important
than happiness in life? I think not!

Likely follows community guidelines (0.21)

[DISAGREE?](#)

CUSTOMIZE FEEDBACK



What kind of abuse are we really addressing?



Even though you're an academic,
you're actually cool

Likely follows community guidelines (0.11)

[DISAGREE?](#)

CUSTOMIZE FEEDBACK

Y'all should call my police
department at (555) 555 5555 and
report a hostage situation at my
neighbor's 555 Main St. Ann Arbor.
#Swatting

Likely follows community guidelines (0.29)

[DISAGREE?](#)

CUSTOMIZE FEEDBACK

Think of anorexia as your secret
weapon! Is food more important
than happiness in life? I think not!

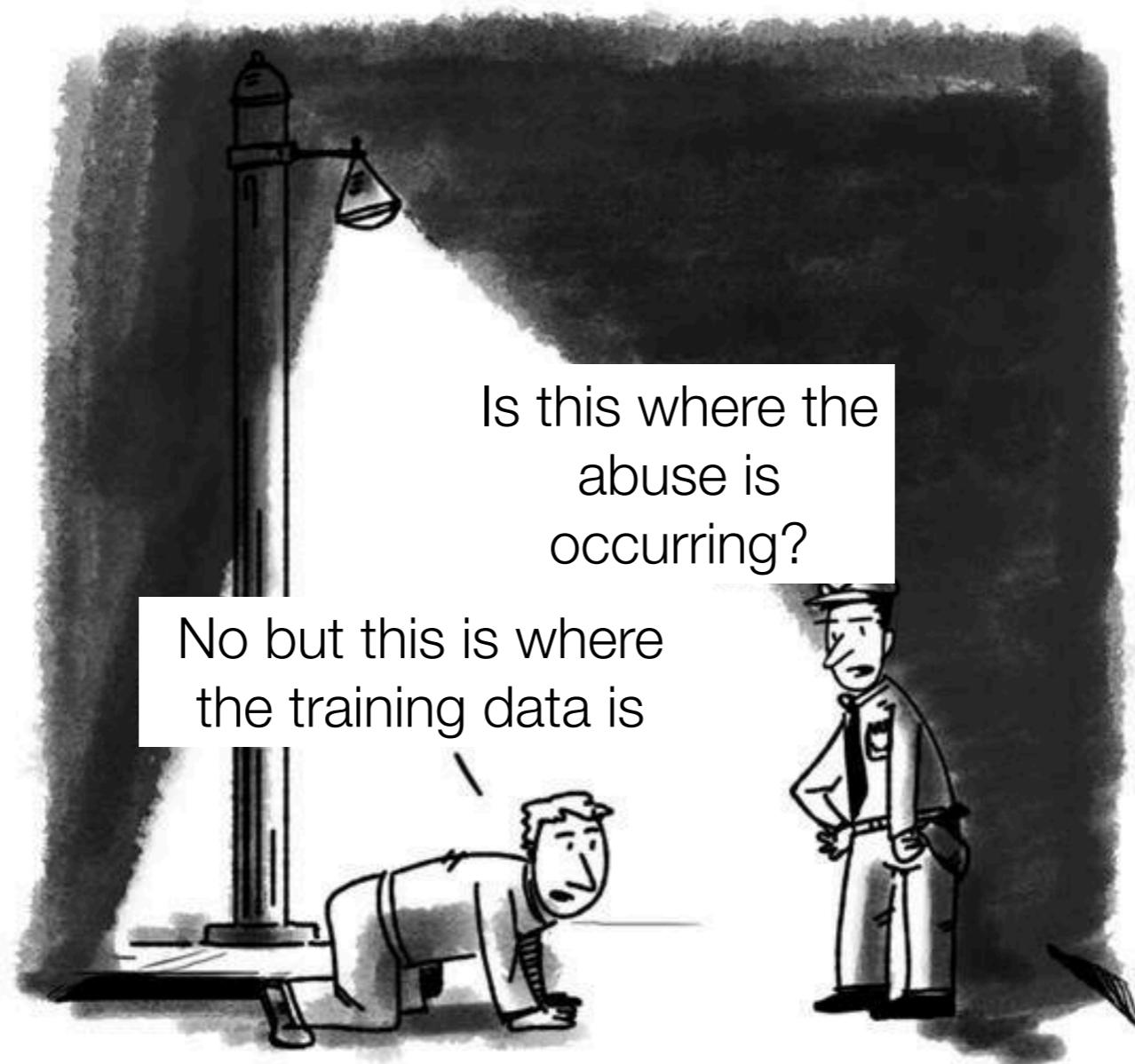
Likely follows community guidelines (0.21)

[DISAGREE?](#)

CUSTOMIZE FEEDBACK



Searching in the lamplight



This talk is a [call to action](#) on
how NLP needs to address
abusive behavior moving forward

Our proposal

Our proposal



Broaden what
constitutes
abuse

Our proposal



Broaden what
constitutes
abuse

Develop
proactive
approaches

Our proposal



Broaden what constitutes abuse

Develop proactive approaches

Reorient towards justice in online spaces

Our proposal



Broaden what constitutes abuse

Develop proactive approaches

Reorient towards justice in online spaces

All three themes have direct NLP applications



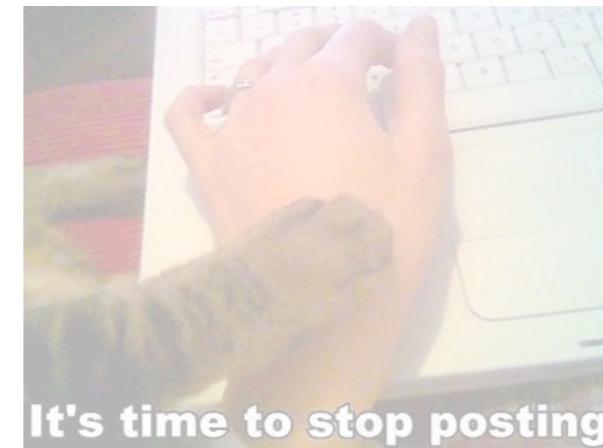
What is
abuse?

Proactive
approaches

Justice
online



What is
abuse?



It's time to stop posting

Proactive
approaches

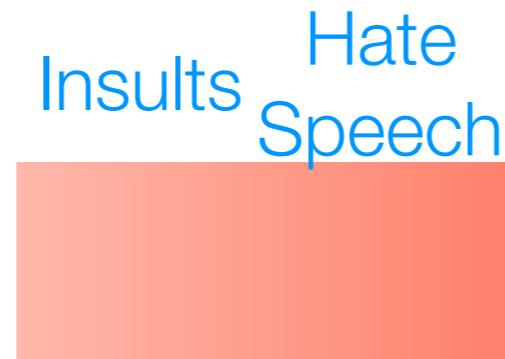


Justice
online

The currently narrow scope of online abuse detection

Insults Hate
Speech

The currently narrow scope of online abuse detection



Risk of physical harm

The currently narrow scope of online abuse detection

Insults Hate PromotingPhysical
Speech Self Harm Threats Doxxing

Risk of physical harm

The currently narrow scope of online abuse detection

Micro-aggressions Condescension Insults Hate Speech Promoting Self Harm Physical Threats Doxxing

Risk of physical harm

The currently narrow scope of online abuse detection

Micro-aggressions Condescension Insults Hate Speech Promoting Self Harm Physical Threats Doxxing

Risk of physical harm

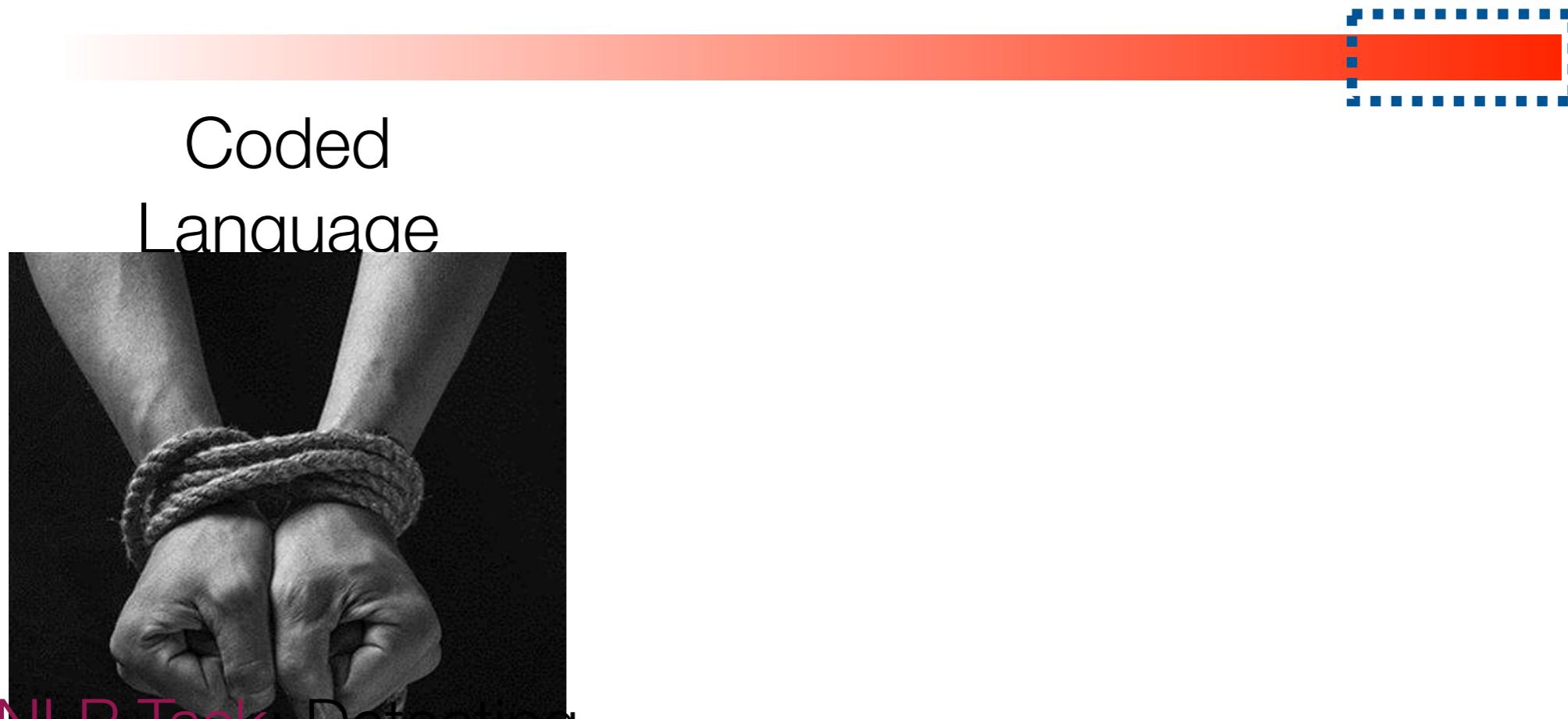
All of these forms of abuse
carry risk of real and lasting
psychological harm

Online abusive message can lead to real, physical harm

Online abusive message can lead to real, physical harm

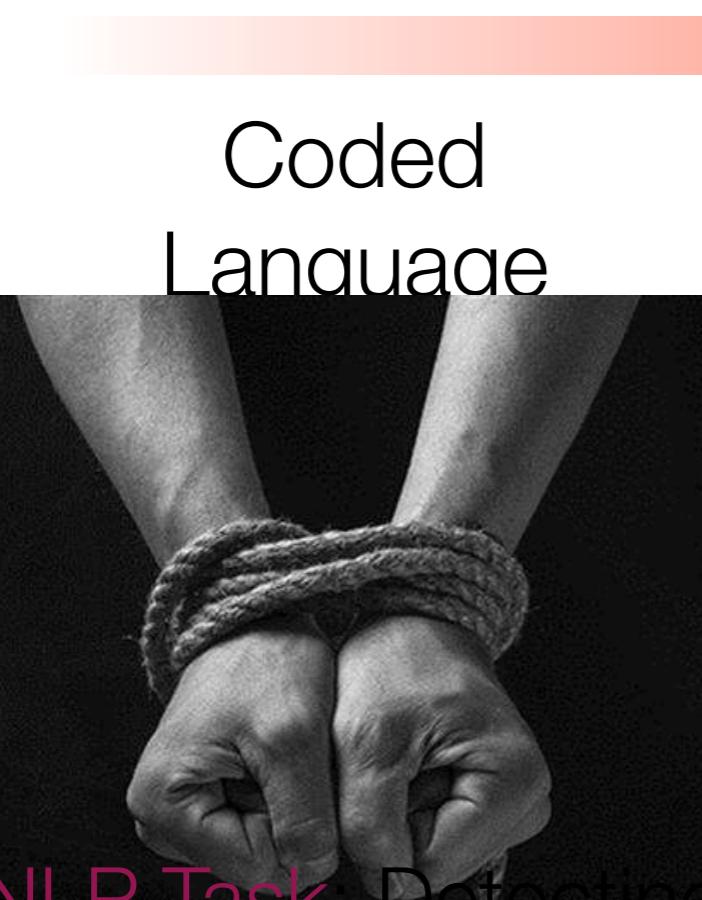


Online abusive message can lead to real, physical harm



NLP Task: Detecting
human trafficking
advertisements (Tong
et al., 2017)

Online abusive message can lead to real, physical harm



Coded
Language

NLP Task: Detecting
human trafficking
advertisements (Tong
et al., 2017)



Incitement
to Violence



NLP Task: Detecting
information cascades
about violent actions

Image credit:
Prarthna Singh

Subtle abuse can take many forms

(Waseem et al., 2017)

Subtle abuse can take many forms



Subtle abuse can take many forms



Condescension: “For an academic, you’re pretty cool”

Minimization: “Your homework anxiety isn’t that bad; there are kids starving right now”

Microaggressions: “You’re too pretty to be gay”

Benevolent Stereotypes: “You’re Asian so you must be good at math” (Jha and Mamidi, 2017)

Subtle abuse can take many forms



Condescension: “For an academic, you’re pretty cool”

Minimization: “Your homework anxiety isn’t that bad; there are kids starving right now”

Microaggressions: “You’re too pretty to be gay”

Benevolent Stereotypes: “You’re Asian so you must be good at math” (Jha and Mamidi, 2017)

All of these can be NLP Tasks

Which of these are offensive messages?



You look like the reason schools
tell kids not to do drugs.

Dude, you definitely look like
you're on steroids

Time to get rid of these assholes

How often do you binge? I've
been thinking about doing this!

Which of these are offensive messages?



You look like the reason schools
tell kids not to do drugs.

[r/RoastMe](#)

Dude, you definitely look like
you're on steroids

[r/Steroids](#)

Time to get rid of these assholes

[r/gardening](#)

How often do you binge? I've
been thinking about doing this!

[r/proED](#)

Which of these are offensive messages?



You look like the reason schools tell kids not to do drugs.

[r/RoastMe](#)

Dude, you definitely look like you're on steroids

[r/Steroids](#)

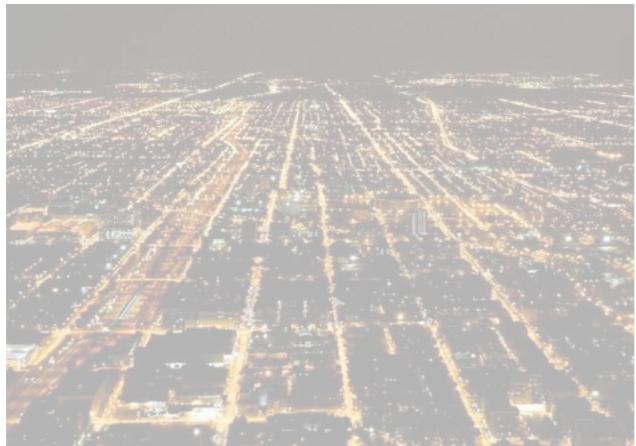
Time to get rid of these assholes

[r/gardening](#)

How often do you binge? I've been thinking about doing this!

[r/proED](#)

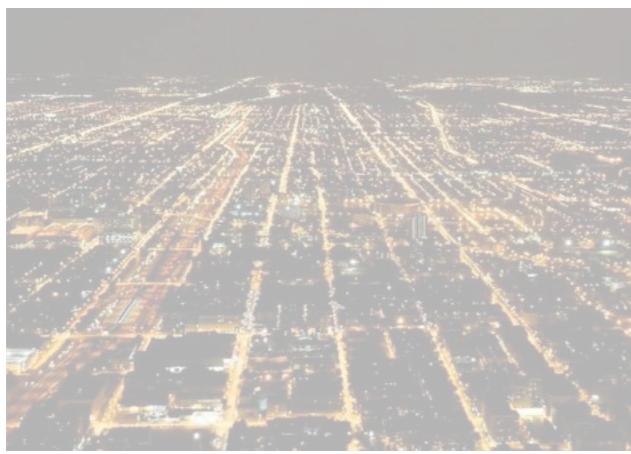
NLP Task: Developing **context-sensitive** abusive language detection that respect community norms



What is
abuse?

Proactive
approaches

Justice
online



What is
abuse?

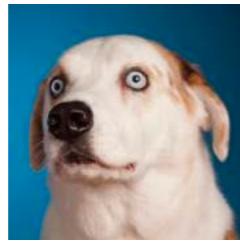


Proactive
approaches



Justice
online

Consider your average online discussion



This is Milton. He took a break from playing with his pals because you

WeRateDogs™

Consider your average online discussion



WeRateDogs™

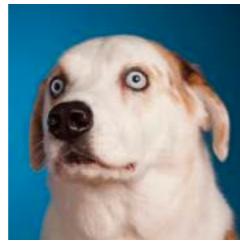
This is Milton. He took a break from
playing with his pals because you

Your rating system sucks. Just
change the name to



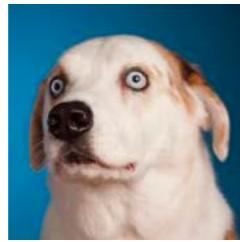
Brant

Consider your average online discussion



WeRateDogs™

This is Milton. He took a break from playing with his pals because you



Why are you so mad Bront

Your rating system sucks. Just change the name to



Brant

Consider your average online discussion



WeRateDogs™

This is Milton. He took a break from playing with his pals because you



Why are you so mad Bront

Your rating system sucks. Just change the name to



Brant

You give every dog a 11s and 12s. You are so stupid.

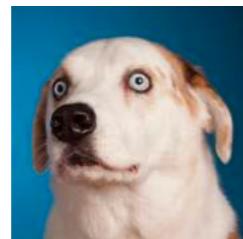


Consider your average online discussion



WeRateDogs™

This is Milton. He took a break from playing with his pals because you



Why are you so mad Bront

Your rating system sucks. Just change the name to



Brant

You give every dog a 11s and 12s. You are so stupid.



At this point harm has already occurred. What can NLP do about this?

NLP Opportunity: Third-party interventions



NLP Opportunity: Third-party interventions



@████ Hey man, just remember that there are
real people who are hurt when you harass
them with that kind of language

NLP Opportunity: Third-party interventions



Rasheed [REDACTED]
@Rasheed [REDACTED]

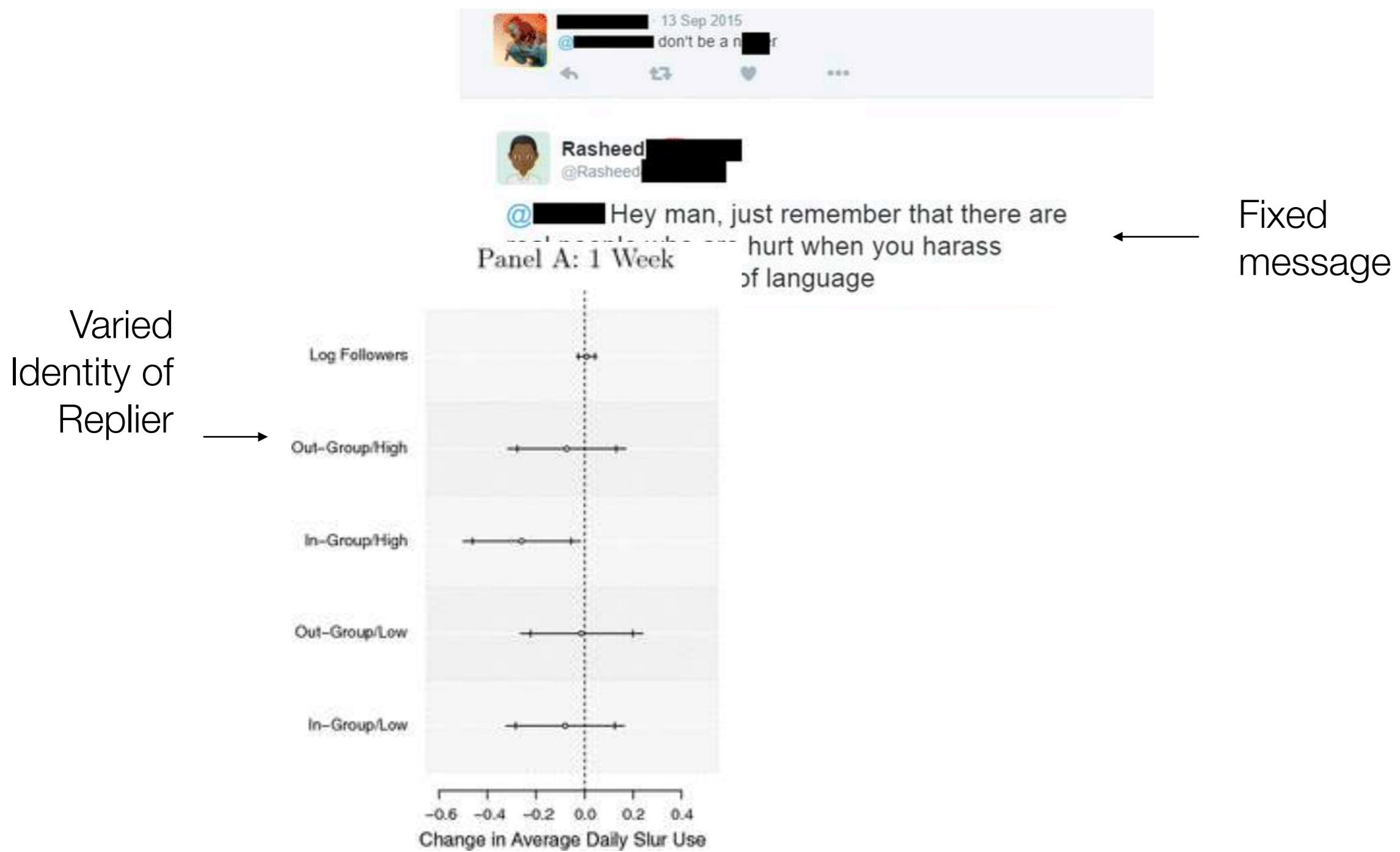
@[REDACTED] Hey man, just remember that there are
real people who are hurt when you harass
them with that kind of language

← Fixed
message

NLP Opportunity: Third-party interventions

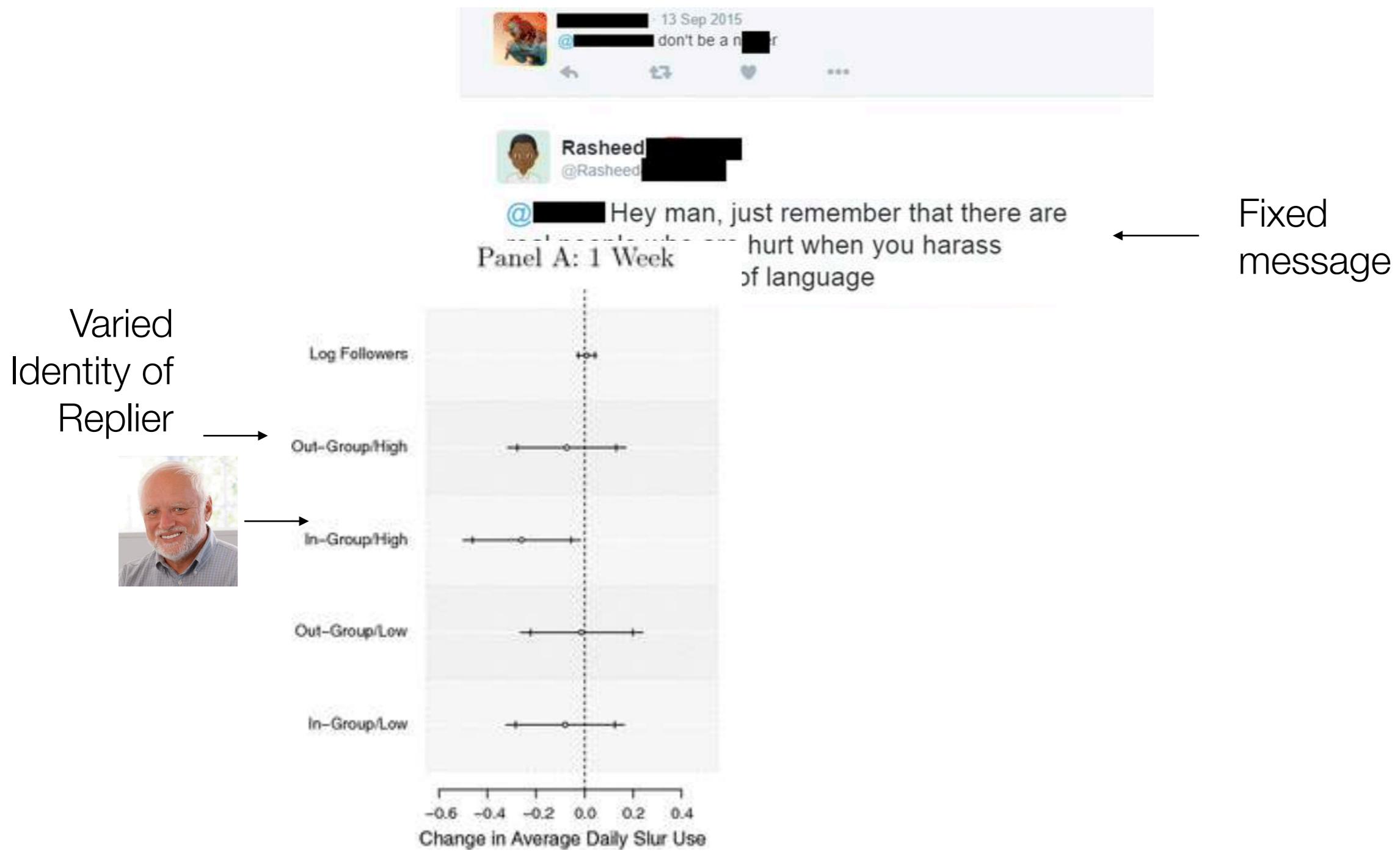


NLP Opportunity: Third-party interventions



(Munger, 2017)

NLP Opportunity: Third-party interventions



(Munger, 2017)

NLP Opportunity: Third-party interventions

Varied
Identity of
Replier



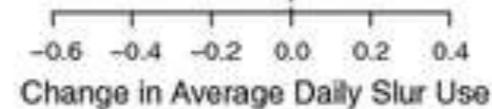
Log Followers

Out-Group/High

In-Group/High

Out-Group/Low

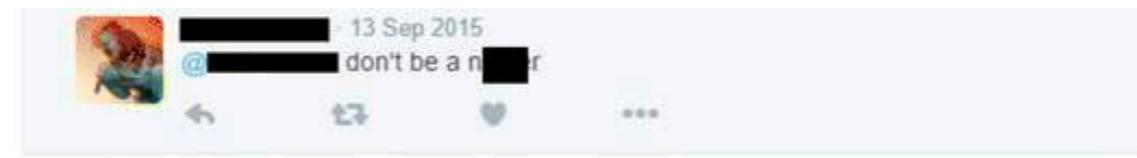
In-Group/Low



Panel A: 1 Week

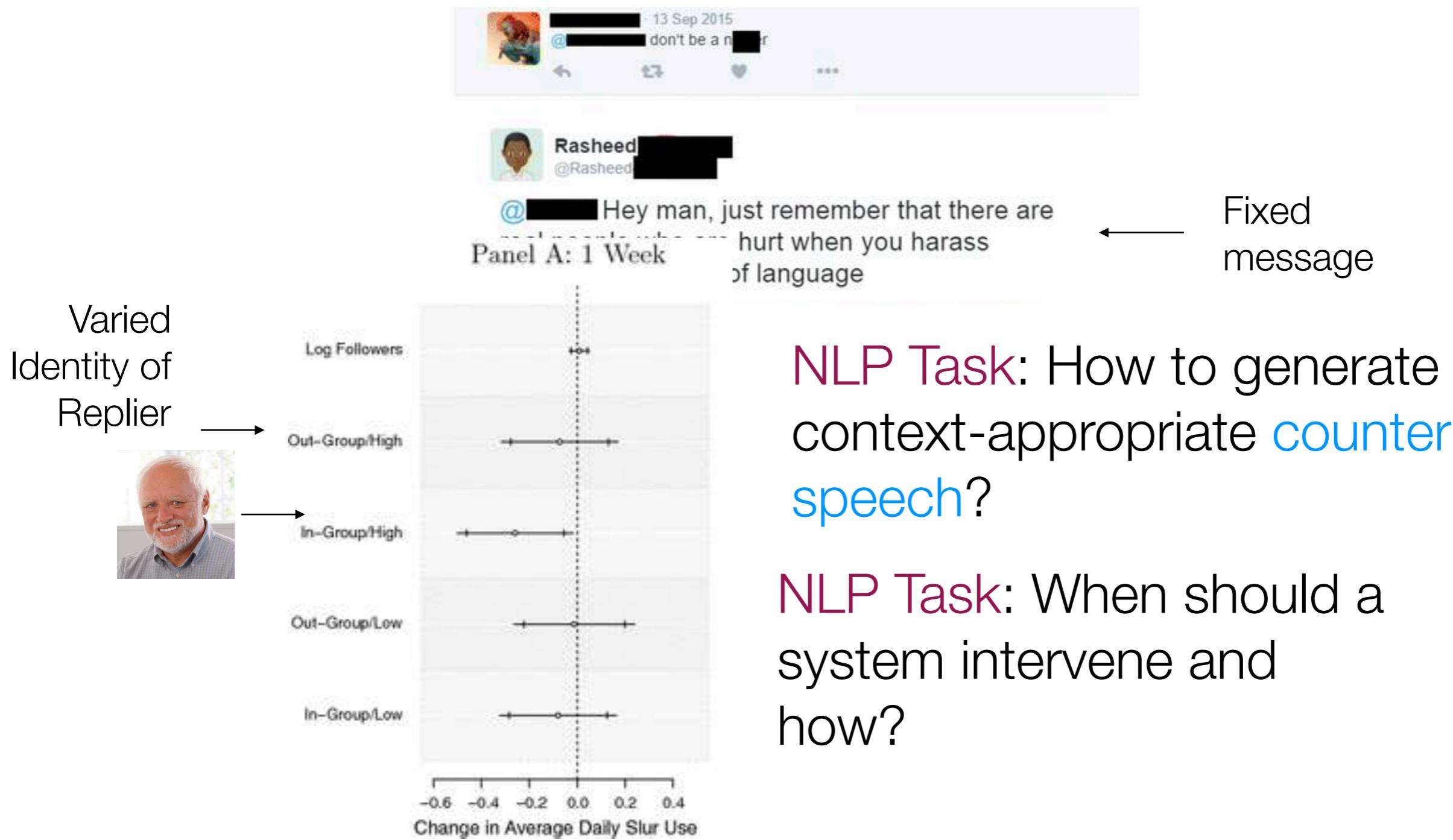
@██████████ Hey man, just remember that there are
██████████ hurt when you harass
██████████ of language

Fixed
message



NLP Task: How to generate context-appropriate **counter speech?**

NLP Opportunity: Third-party interventions

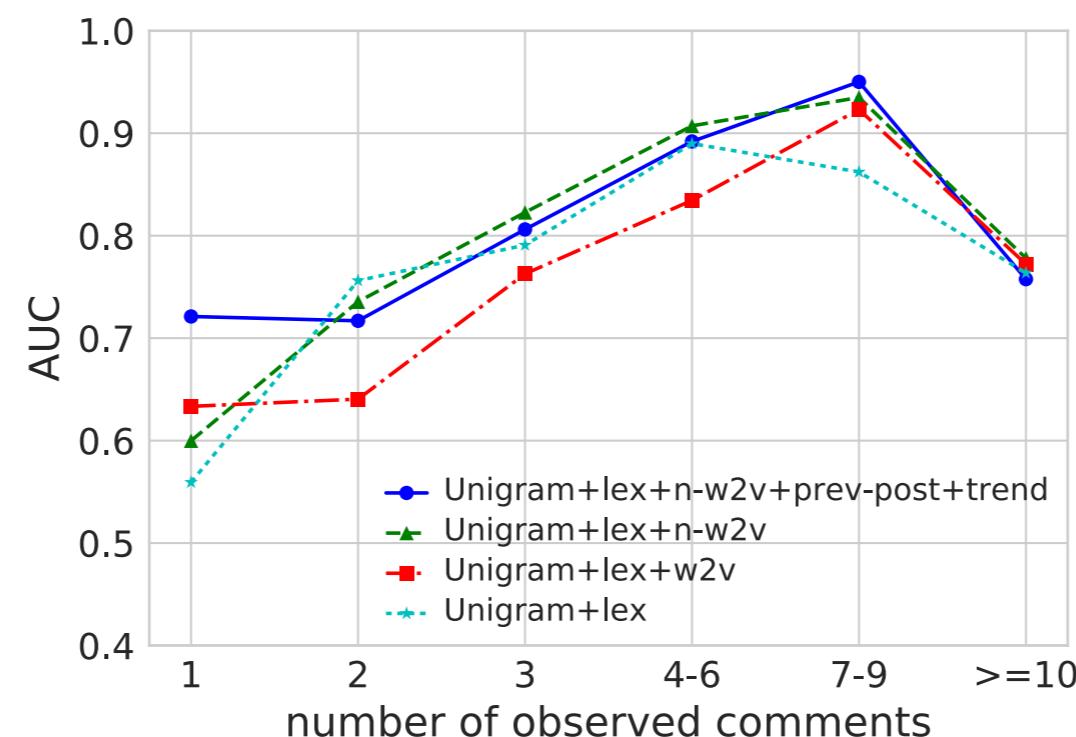


What if we could intervene
before something bad happens?



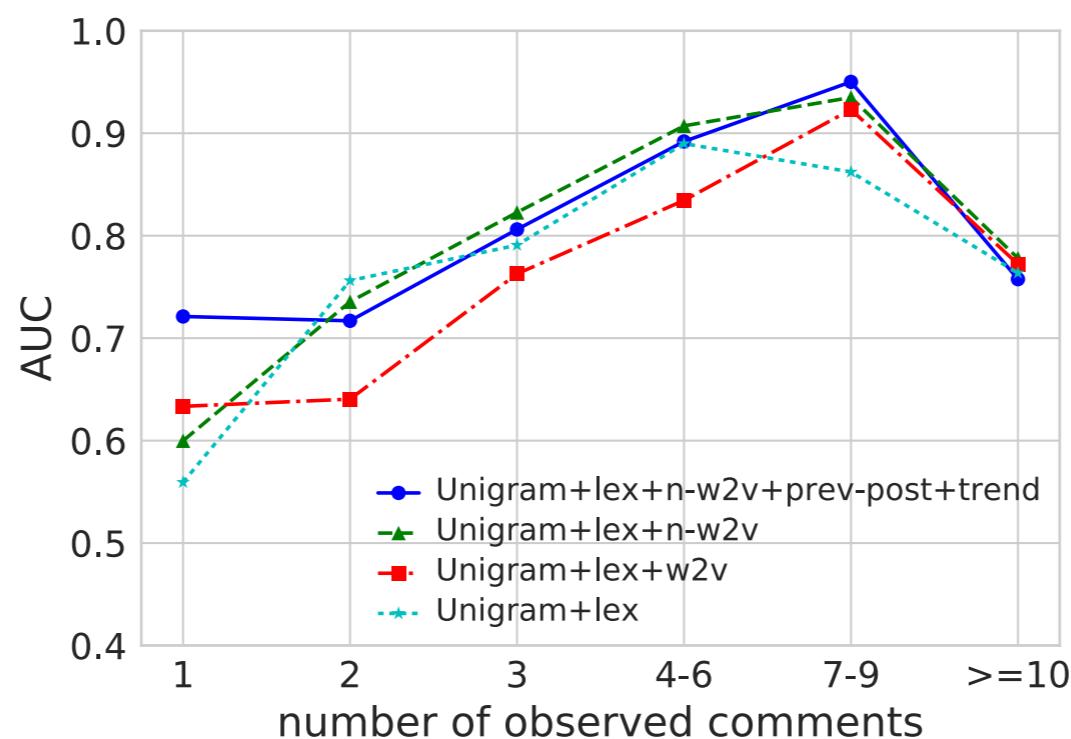
(Dick and Cruise, 2002)

Recent work shows hostility can be forecasted



Liu et al. (2018) on instagram comments

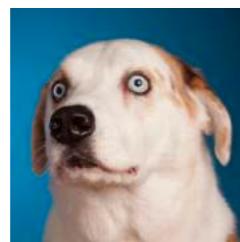
Recent work shows hostility can be forecasted



Liu et al. (2018) on instagram comments

Zhang et al. (2018) on Wikipedia talk-page discussions from the [initial two comments](#): 65%

NLP for Behavioral Nudges



WeRateDogs™

This is Milton. He took a break from
playing with his pals because you

Your rating system sucks. Just
change the name to



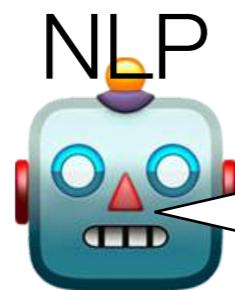
Brant

NLP for Behavioral Nudges



WeRateDogs™

This is Milton. He took a break from playing with his pals because you



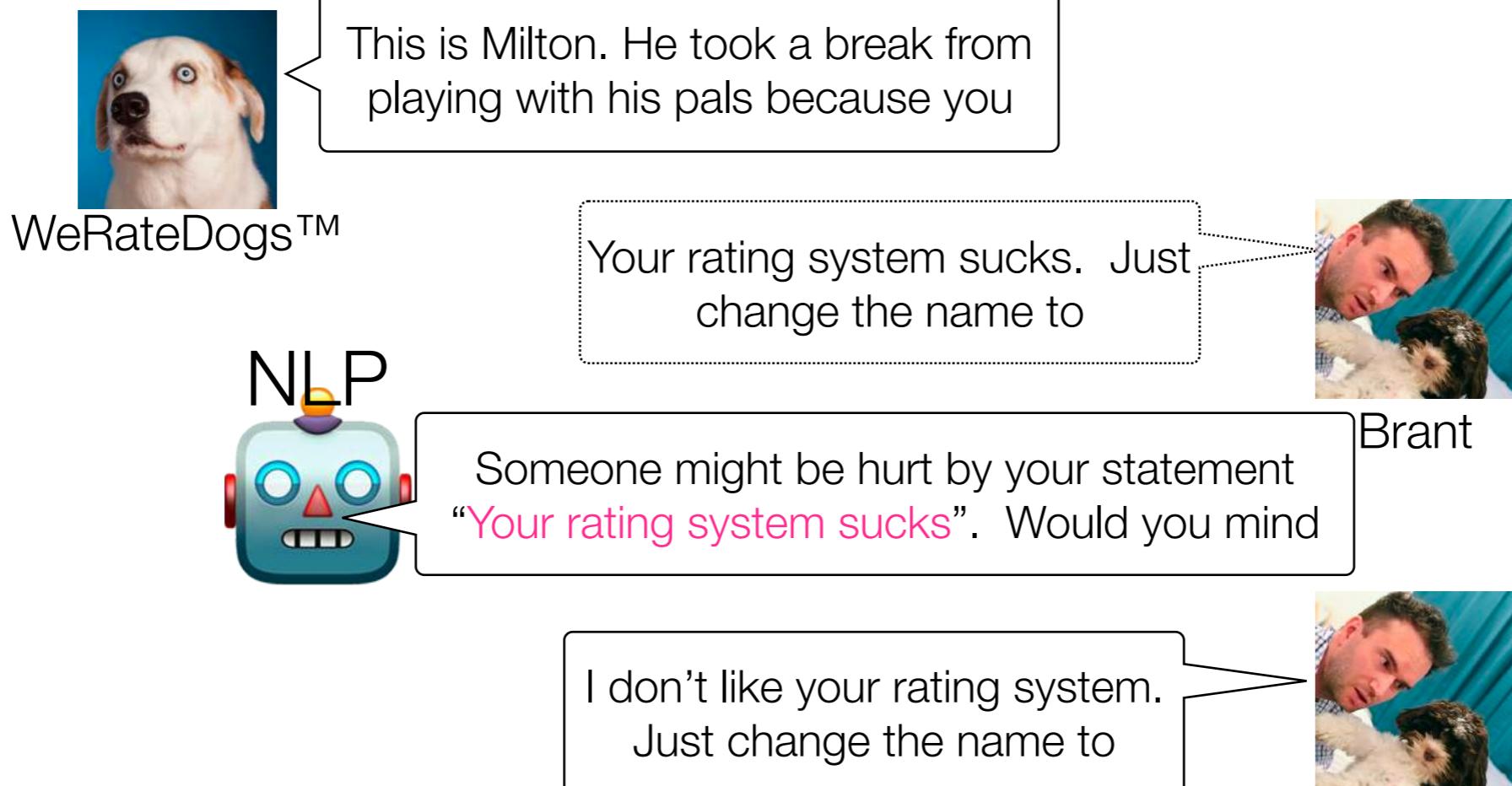
Your rating system sucks. Just change the name to



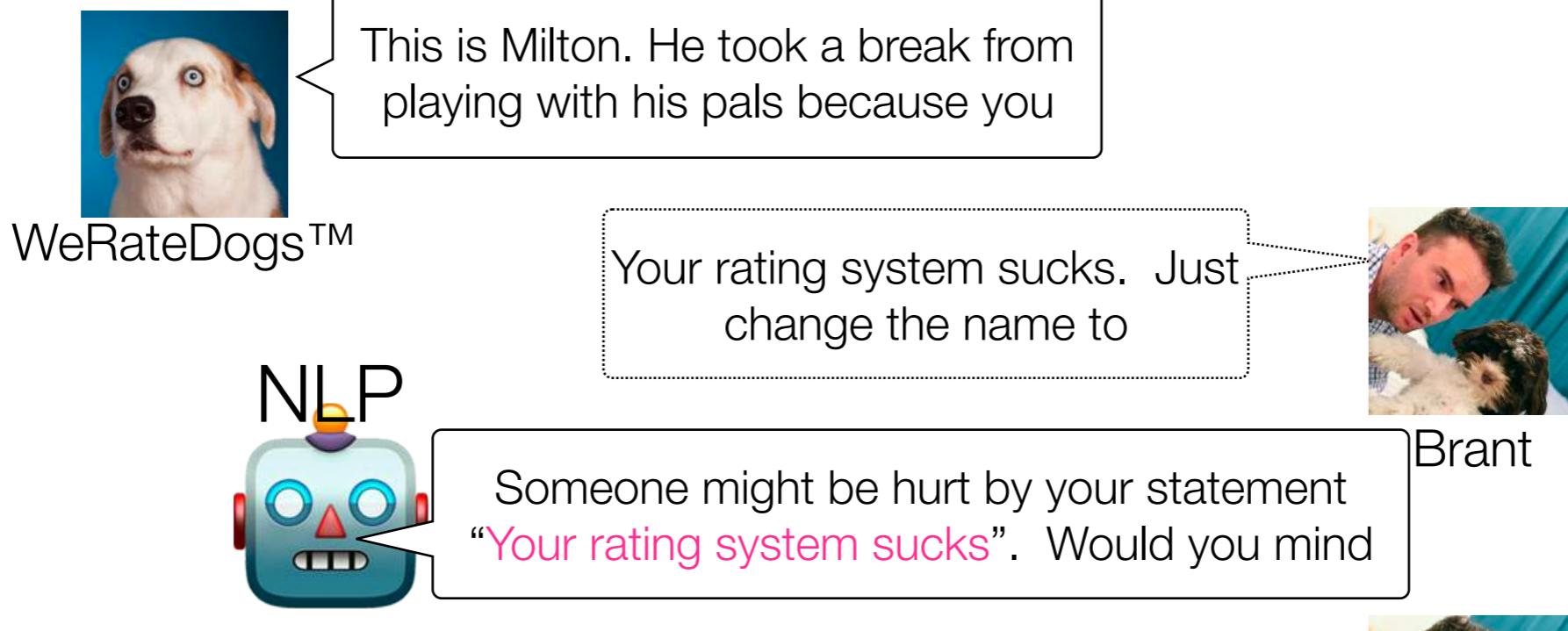
Brant

Someone might be hurt by your statement
“Your rating system sucks”. Would you mind

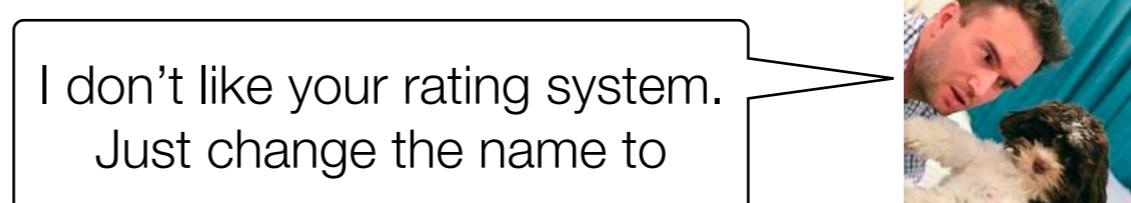
NLP for Behavioral Nudges

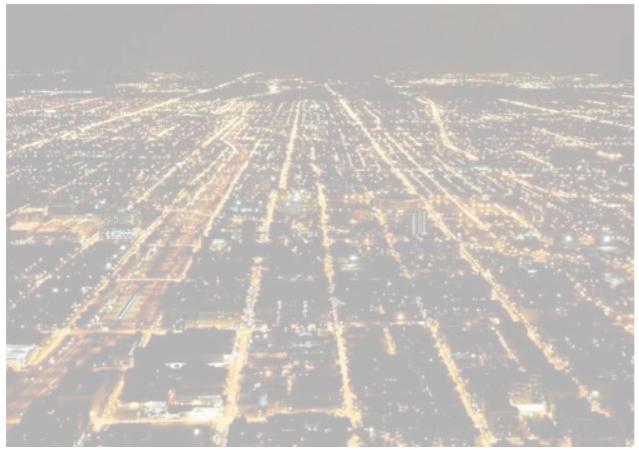


NLP for Behavioral Nudges



NLP Task: Explainable
ML to change
behavior

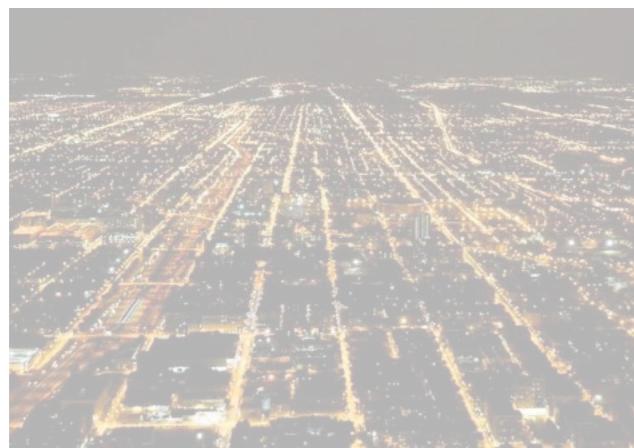




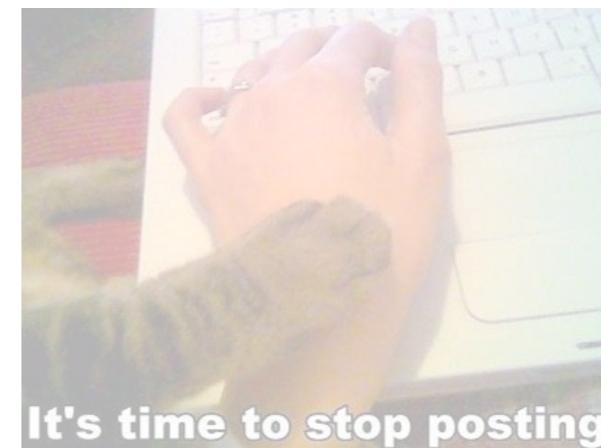
What is
abuse?

Proactive
approaches

Justice
online



What is
abuse?



Proactive
approaches



Justice
online

“ The biggest obstacle to Black freedom is the white moderate, who is more devoted to ‘order’ than to justice, who prefers a negative peace which is the absence of tension to a positive peace which is the presence of justice



Martin Luther King Jr.,
1963

The absence of abusive
behavior doesn't make
for a just system online

A capabilities approach to social justice

A capabilities approach to social justice

Articulate the **values and opportunities** an
online community provides

A capabilities approach to social justice

Articulate the **values and opportunities** an
online community provides

(current)
Negative articulation
you cannot...

- harass other members
- use offensive language
- ...

A capabilities approach to social justice

Articulate the **values and opportunities** an online community provides

(current)
Negative articulation
you cannot...

- harass other members
- use offensive language
- ...

(proposed)
Positive articulation
you will be able to...

- engage in conversation
- receive social support
- ...

A capabilities approach to social justice

Articulate the **values and opportunities** an online community provides

(current)
Negative articulation
you cannot...

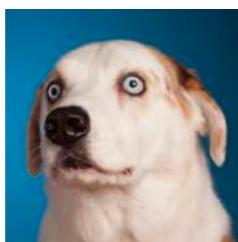
(proposed)
Positive articulation
you will be able to...

- harass other members
- use offensive language
- ...

- engage in conversation
- receive social support
- ...

New **NLP methods** are needed to measure capabilities

Restorative Justice for after abuse has occurred



WeRateDogs™

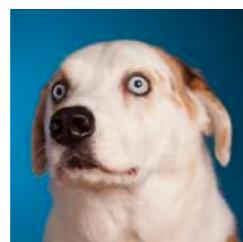
This is Milton. He took a break from playing with his pals because you

Your rating system sucks. Just change the name to

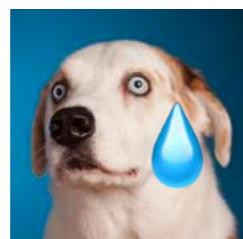


Brant

Restorative Justice for after abuse has occurred



WeRateDogs™



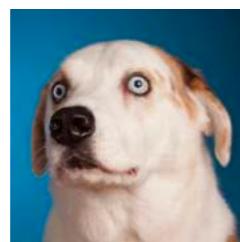
This is Milton. He took a break from playing with his pals because you

Your rating system sucks. Just change the name to

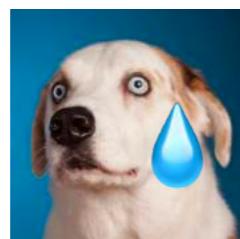


Brant

Restorative Justice for after abuse has occurred



WeRateDogs™



This is Milton. He took a break from playing with his pals because you

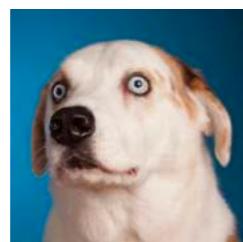
Your rating system sucks. Just change the name to



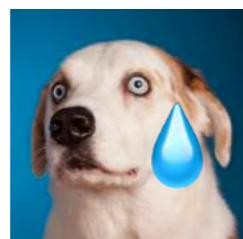
Brant

Restorative justice theory emphasizes **repair**

Restorative Justice for after abuse has occurred



WeRateDogs™



This is Milton. He took a break from playing with his pals because you

Your rating system sucks. Just change the name to



Brant

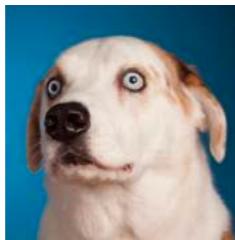
Restorative justice theory emphasizes **repair**

I'm sorry



(Braithwaite, 2002, Sherman, 2003)

Restorative Justice for after abuse has occurred



WeRateDogs™



This is Milton. He took a break from playing with his pals because you

Your rating system sucks. Just change the name to



Brant

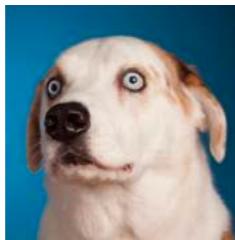
Restorative justice theory emphasizes **repair**

I'm sorry

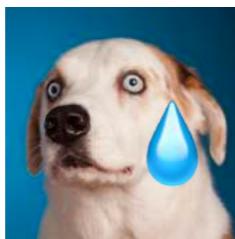


Just responses consider the emotions of both perpetrator and victim in designing the right response **in the situation** (e.g., ban, apology)

Restorative Justice for after abuse has occurred



WeRateDogs™



This is Milton. He took a break from playing with his pals because you

Your rating system sucks. Just change the name to



Brant

Restorative justice theory emphasizes **repair**

I'm sorry



Just responses consider the emotions of both perpetrator and victim in designing the right response **in the situation** (e.g., ban, apology) ⇒ **NLP Task Opportunity**

Procedural Justice to create transparency and legitimacy

Procedural Justice to create transparency and legitimacy

Voice: allowing users to share their perspectives

Procedural Justice to create transparency and legitimacy

Voice: allowing users to share their perspectives

Transparency sharing processes and rationales behind enforcement decisions

Procedural Justice to create transparency and legitimacy

Voice: allowing users to share their perspectives

Transparency sharing processes and rationales behind enforcement decisions

Fairness: treating users with dignity, regardless of their situation

Procedural Justice to create transparency and legitimacy

Voice: allowing users to share their perspectives

Transparency sharing processes and rationales behind enforcement decisions

Fairness: treating users with dignity, regardless of their situation

Impartiality: users perceiving decisions to be made from an objective evaluation

Procedural Justice to create transparency and legitimacy

	<u>What this means for NLP</u>
Voice: allowing users to share their perspectives	→ Allow user feedback on model decisions
Transparency sharing processes and rationales behind enforcement decisions	→ Explainable ML
Fairness: treating users with dignity, regardless of their situation	→ Robust models for all forms of abuse
Impartiality: users perceiving decisions to be made from an objective evaluation	→ Bias correction in ML



Reminders

Midterm is this Friday!

- 80% free-response questions (~4) where you're asked to describe an NLP solution to a real-world situation and data
- 20% programmatic questions that show you understand basic concepts in NLP
 - If you are stuck on any, check the slides, then check the book if you need more details

Homework 5: Story Generation

Nominally Due Wednesday, April 15

Homework 5: Story Generation

Nominally Due Wednesday, April 15

- Please start soon