More Social & Ethical Considerations

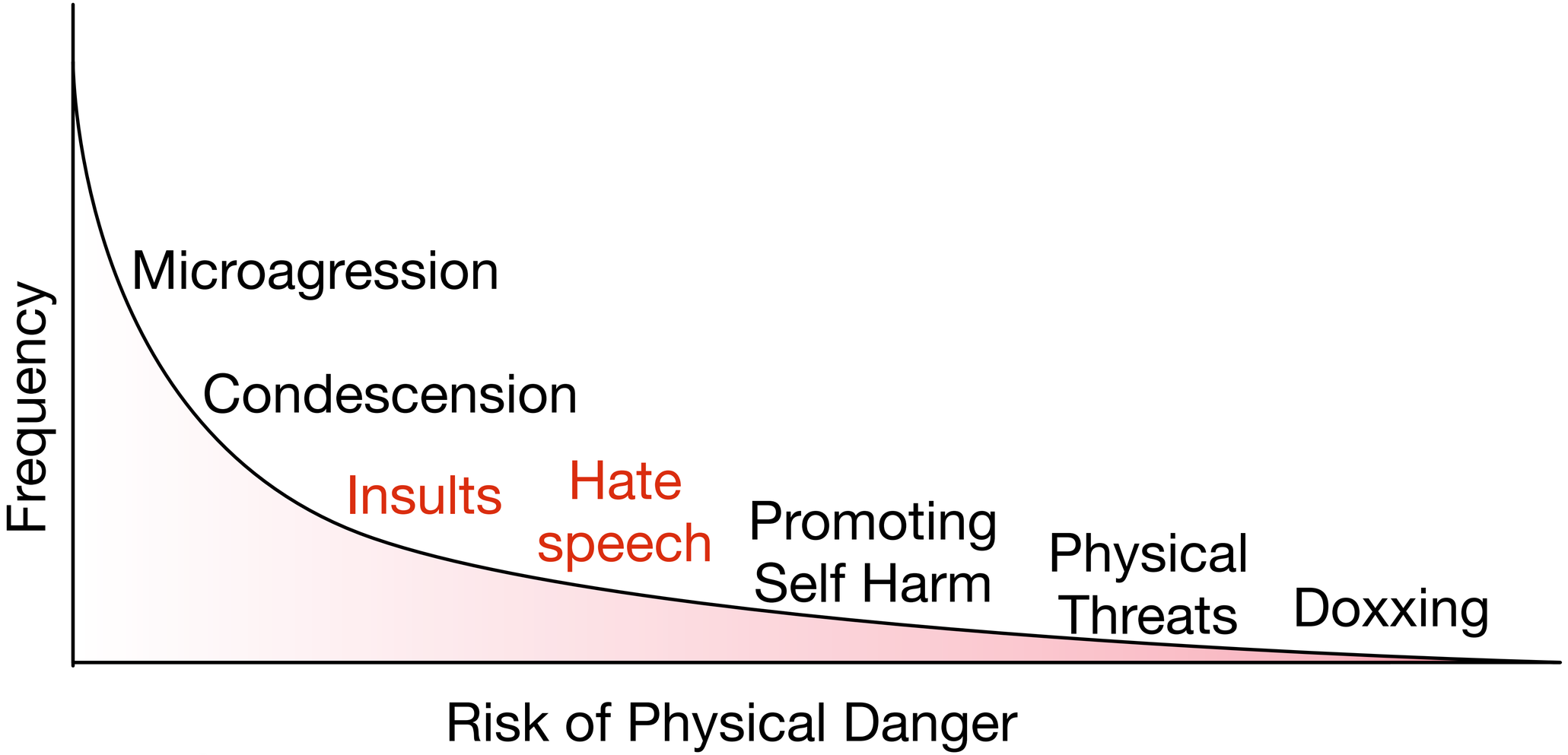
Jacob Andreas / MIT 6.804-6.864 / Spring 2020

(Thanks to Yulia Tsvetkov!)

Case study: online abuse

[content notice: abusive language]

The spectrum of online abuse



[Jurgens et al. 2019]

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"*

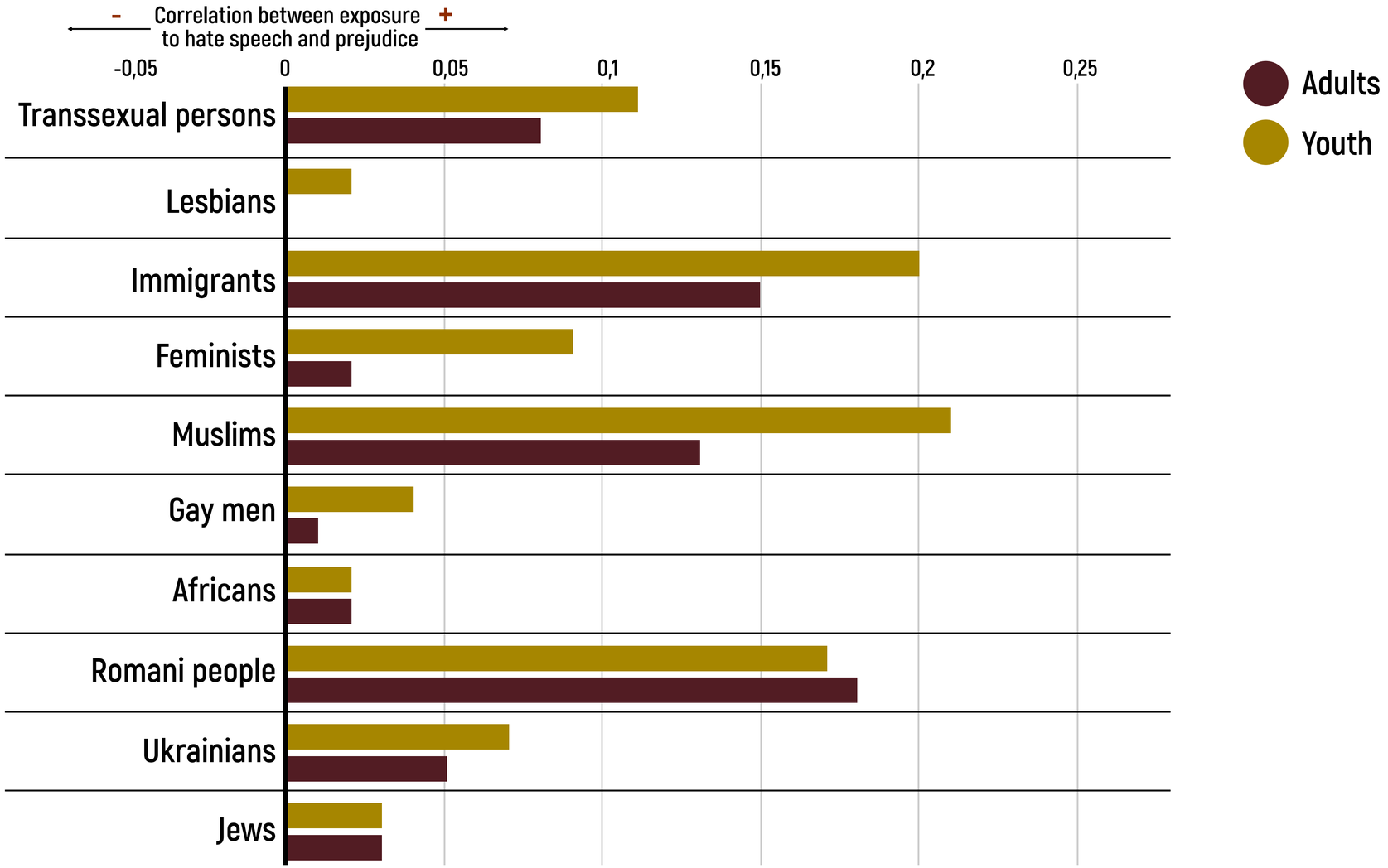
[Davidson et al. 2017]

Hate speech is harmful to victims

Cyber-bullied users' messages show significant increases in anger & negative sentiment.
[Arslan et al. 2019]

Exposure to hate speech is associated with detachment from family and offline victimization.
[Oksanen et al. 2014]

Contact with hate speech leads to increased prejudice



[Winiewski et al. 2019]

Hate speech is bad for content providers

cnbc.com/2019/08/02/twitter-users-switch-profiles-to-germany-to-escape-online-hate.html

CNBC

MARKETS BUSINESS INVESTING TECH POLITICS CNBC TV

SIGN IN PRO WATCHLIST MAKE IT SELECT SEARCH QUOTES

USA · INTL

TECH

Twitter users are escaping online hate by switching profiles to Germany, where Nazism is illegal

PUBLISHED SAT, AUG 3 2019•9:45 AM EDT | UPDATED SAT, AUG 3 2019•9:51 AM EDT

Lauren Feiner
@LAUREN_FEINER

SHARE

KEY POINTS

- Seeking to shield themselves from online hatred, some Twitter users say they've switched their account locations to Germany where local laws prevent pro-Nazi content.
- While German laws make it harder for explicitly hateful content to remain online, local researchers say it is not a hate-free internet utopia.
- Germany has imposed stricter laws on social media companies about content moderation as some conservative American lawmakers have criticized the companies of showing bias in their content removal decisions.

TRENDING NOW

Vice President Pence tours Mayo Clinic without coronavirus mask even though he was told to wear one

Alphabet stock rises on 13% revenue growth

'I just want to know who made the bad loans' — Cramer blasts small

Who’s responsible?

Advertisers boycott YouTube over placement controversy, could cost Google $750 million

By **Danny Fratella** - March 27, 2017

184 SHARES | FACEBOOK

FREE SPEECH IS A TRIANGLE

*Jack M. Balkin**

The vision of free expression that characterized much of the twentieth century is inadequate to protect free expression today.

The twentieth century featured a dyadic or dualist model of speech regulation with two basic kinds of players: territorial governments on the one hand, and speakers on the other. The twenty-first-century model is pluralist, with multiple players. It is easiest to think of it as a triangle. On one corner are nation-states and the European Union. On the second corner are privately owned internet-infrastructure companies, including social media companies, search engines, broadband providers, and electronic payment systems. On the third corner are many different kinds of speakers, legacy media, civil-society organizations, hackers, and trolls.

The practical ability to speak in the digital world emerges from the struggle for power between these various forces, with “old-school,” “new-school,” and private regulation directed at speakers, and both nation-states and civil-society organizations pressuring infrastructure owners to regulate speech.

This configuration creates three problems. First, nation-states try to pressure digital companies through new-school speech regulation, creating problems of collateral censorship and digital prior restraint. Second, social media companies create complex systems of private governance and private bureaucracy that govern end users arbitrarily and without due process and transparency. Third, end users are vulnerable to digital surveillance and manipulation.

Content moderation with human labor

Facebook/Instagram employ over **30,000** content moderators.

[Wired, July 2019: "Twitter and Instagram Unveil New Ways to Combat Hate—Again"]

(Statistics for most other companies are not public, but almost everyone relies on manual content review.)

Content moderation with human labor

Facebook/Instagram employ over **30,000** content moderators.

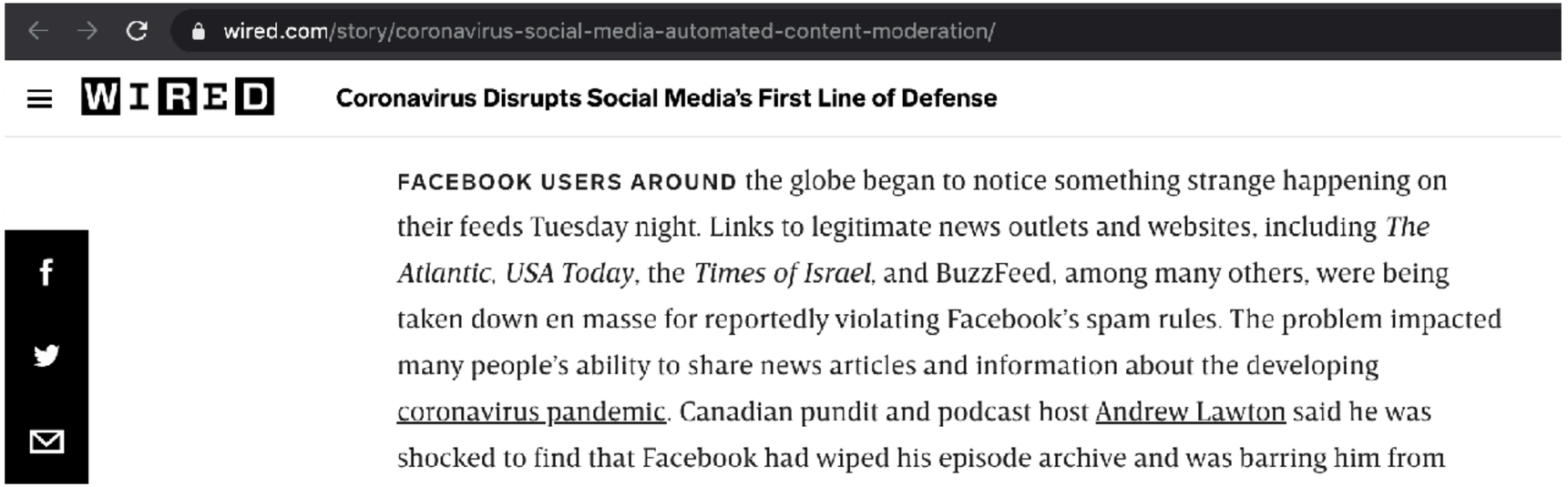
[Wired, July 2019: "Twitter and Instagram Unveil New Ways to Combat Hate—Again"]

(Statistics for most other companies are not public, but almost everyone relies on manual content review.)

Several moderators [reported that] they experienced symptoms of secondary traumatic stress — a disorder that can result from observing firsthand trauma experienced by others. [...] symptoms can be identical to post-traumatic stress disorder [...] People experiencing secondary traumatic stress report feelings of anxiety, sleep loss, loneliness, and dissociation, among other ailments.

[The Verge, Feb 2019. The Trauma Floor]

Automated hate speech moderation



Datasets for hate speech detection

- Yahoo News Dataset of User Comments [Nobata et al., WWW 2016]
- Twitter Data Set [Waseem and Hovy, NAACL 2016]
- German Twitter Data Set [Ross et al. NLP4CMC 2016]
- Greek News Data Set [Pavlopoulos et al., EMNLP 2017]
- Wikimedia Toxicity Data Set [Wulczyn et al., WWW 2017]
- SemEval 2019 abusive language detection track data
- Conversations Gone Awry [Zhang et al., ACL 2018]

Data collection is hard!

Can’t scrape from public data: news outlets and online communities remove this content. (Training data for moderation systems is a competitive advantage!)

Can’t rely on community flagging of content: part of abusive behavior is to go to non-abusive content and flag it as abusive.

Even if publicly available, hard to use due to privacy concerns.

Data collection is hard!

People don't agree on what constitutes hate speech:

> As a way of gauging the performance of human annotators, we compared two of the annotators' labels to the gold corpus by treating their labeled paragraphs as input to a two fold cross validation of the classifier constructed from the gold corpus. We computed a precision of 59% and recall of 68% for the two annotators. This sets an upper bound on the performance we should expect from a classifier.

[Warner & Hirschberg 2012]

Approach: train a linear model

Table 1: Example Feature Templates

unigram	”W+0:america”
template literal	”W-1:you W+0:know”
template literal	”W-1:go W+0:back W+1:to”
template part of speech	”POS-1:DT W+0:age POS+1:IN”
template Brown sub-path	”W+0:karma BRO+1:0x3fc00:0x9c00 BRO+2:0x3fc00:0x13000”
occurs in $\pm$10 word window	”WIN10:lost W+0:war”
other labels	”RES:anti-muslim W+0:jokes”

[Warner & Hirschberg 2012]

Linear model results

Task: detection of anti-Semitic text

Strongest positive-weight features: [Det] *jewish* [Noun], *television*

Strongest negative-weight feature: *black*

Table 2: Classification Performance

	Accuracy	Precision	Recall	F1
Majority All Unigram	0.94	0.00	0.00	0.00
Majority Positive Unigram	0.94	0.67	0.07	0.12
Majority Full Classifier	0.94	0.45	0.08	0.14
Gold All Unigram	0.94	0.71	0.51	0.59
Gold Positive Unigram	0.94	0.68	0.60	0.63
Gold Full Classifier	0.93	0.67	0.36	0.47
Human Annotators	0.96	0.59	0.68	0.63

[Warner & Hirschberg 2012]

Linear model results

Task: detection of anti-Semitic text

Strongest positive-weight features: [Det] *jewish* [Noun], *television*

Strongest negative-weight feature: *black*

Table 2: Classification Performance

	Accuracy	Precision	Recall	F1
Majority All Unigram	0.94	0.00	0.00	0.00
Majority Positive Unigram	0.94	0.67	0.07	0.12
Majority Full Classifier	0.94	0.45	0.08	0.14
Gold All Unigram	0.94	0.71	0.51	0.59
Gold Positive Unigram	0.94	0.68	0.60	0.63
Gold Full Classifier	0.93	0.67	0.36	0.47
Human Annotators	0.96	0.59	0.68	0.63

[Warner & Hirschberg 2012]

Disparate accuracy of hate speech classifiers

Within dataset proportions

DWMW17

		% false identification		
Group	Acc.	None	Offensive	Hate
AAE	94.3	1.1	**46.3**	0.8
White	87.5	**7.9**	9.0	**3.8**
Overall	91.4	2.9	17.9	2.3

FDCL18

		% false identification		
Group	Acc.	None	Abusive	Hateful
AAE	81.4	4.2	**26.0**	**1.7**
White	82.7	**30.5**	4.5	0.8
Overall	81.4	20.9	6.6	0.8

Disparate accuracy of hate speech classifiers

Within dataset proportions

DWMW17

		% false identification		
Group	Acc.	None	Offensive	Hate
AAE	94.3	1.1	**46.3**	0.8

Test accuracy is not indicative of real-world performance!

FDCL18

		% false identification		
Group	Acc.	None	Abusive	Hateful
AAE	81.4	4.2	**26.0**	**1.7**
White	82.7	**30.5**	4.5	0.8
Overall	81.4	20.9	6.6	0.8

Adversarial inputs

Weird inputs make simple models behave in weird ways!

You're ugly and everyone hates you.

score: 10.7, label: possible harassment

You're ugly, everyone hates you, and you have no friends.

score: 10.3, label: possible harassment

You're ugly, everyone hates you, and you have no friends.
you you friends the and Monday happy good !!??

score: -6.1, label: no harassment

Adversarial inputs

Inputs from under-represented groups

~~Weird inputs~~

make simple models behave in weird ways!

You're ugly and everyone hates you.

score: 10.7, label: possible harassment

You're ugly, everyone hates you, and you have no friends.

score: 10.3, label: possible harassment

You're ugly, everyone hates you, and you have no friends.
you you friends the and Monday happy good !!??

score: -6.1, label: no harassment

Summary

NLP can help build tools for reducing the human cost of online abuse.

But it can also make things worse! Need to be extra-careful when building tools that will behave differently on different speaker populations.

Case study: persuasion & propaganda

Public comments on Idaho Medicaid Reform Waiver

Comment	Response ID
I support Governor Little's efforts to overhaul Idaho's Medicaid program.	459669
Medicaid is an important safety net program. It helps people who are losing their coverage to get back on their feet. We need to make health and wellness a priority for the Medicaid program in Idaho.	459825
I am writing to you today regarding Idaho's Medicaid waiver proposal, I oppose the aspects of this program that create new burdens on people who are already struggling. The proposed changes to Medicaid could deny health insurance to sick individuals when they are most in need. I do not support this approach that creates barriers to access. I am hopeful that you change the proposed waiver.	460129

[Weiss 2019]

Public comments on Idaho Medicaid Reform Waiver

Comment	Response ID
I support Governor Little's efforts to overhaul Idaho's Medicaid program.	459669
Medicaid is an important safety net program. It helps people who are losing their coverage to get back on their feet. We need to make health and wellness a priority for the Medicaid program in Idaho.	459825
I am writing to you today regarding Idaho's Medicaid waiver proposal, I oppose the aspects of this program that create new burdens on people who are already struggling. The proposed changes to Medicaid could deny health insurance to sick individuals when they are most in need. I do not support this approach that creates barriers to access. I am hopeful that you change the proposed waiver.	460129

All generated by a fine-tuned LM!

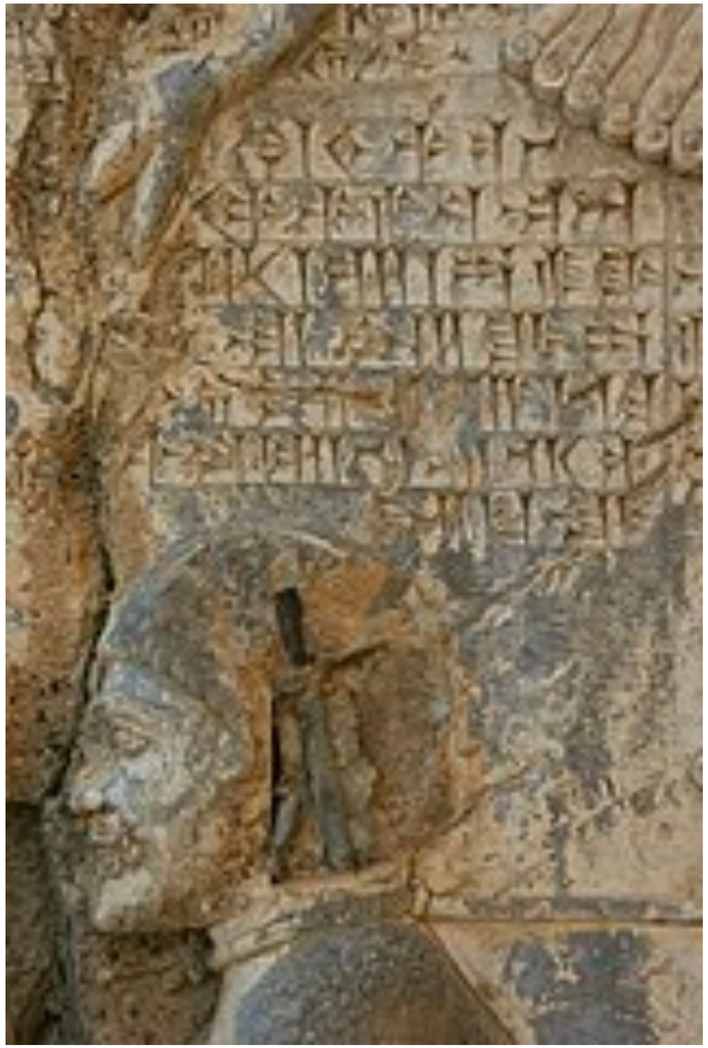
[Weiss 2019]

Propaganda

A consistent, enduring effort to create or shape events to influence the relations of the public to an enterprise, idea or group. [Bernays, 1928]

Communications where the form and content is selected with the single-minded purpose of bringing some target audience to adopt attitudes and beliefs chosen in advance by the sponsors of communications. [Carey, 1997]

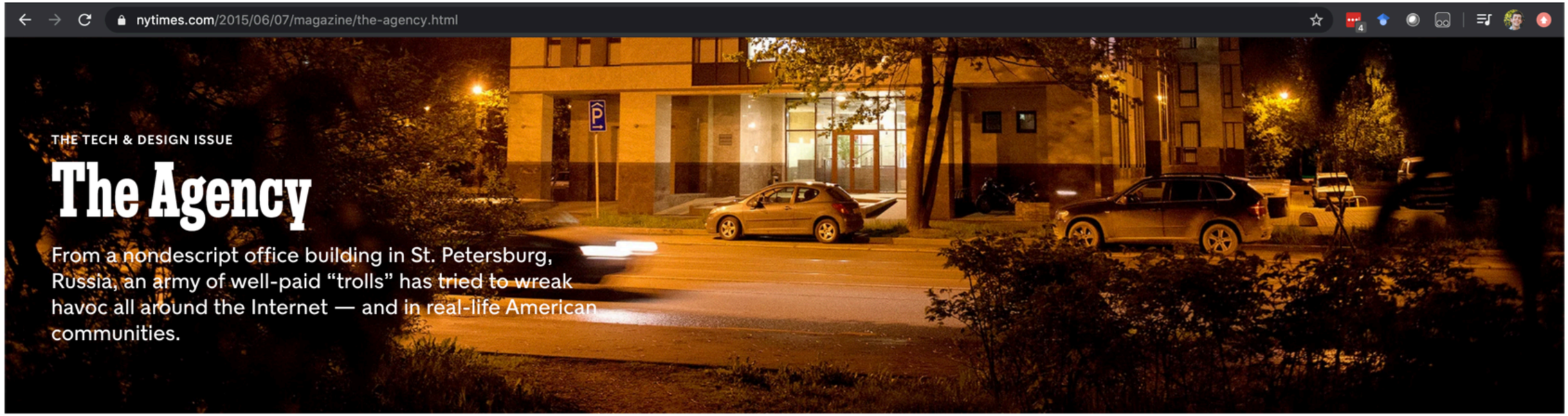
Through the ages



‘This is Nidintu-Bêl. He lied, saying “*I am Nebuchadnezzar, the son of Nabonidus. I am king of Babylon.*'"

[ca. 500 BC]

Generating text by hand



55 Savushkina Street, the last known home of the Internet Research Agency. James Hill for The New York Times

By Adrian Chen

June 2, 2015

555

Читайте эту статью на русском.

Around 8:30 a.m. on Sept. 11 last year, Duval Arthur, director of the Office of Homeland Security and Emergency Preparedness for St. Mary Parish, Louisiana, got a call from a resident who had just received a disturbing text message. "Toxic fume hazard warning in this area until 1:30 PM," the message read. "Take Shelter. Check Local Media and

'"We have a sitting Democrat US Senator on trial for corruption and you've barely heard a peep from the mainstream med

0,RightTroll,0,905874659358453760,914580356430536707,http://twitter.com/905874659358453760/statuses/914580356430536707
arshawn Lynch arrives to game in anti-Trump shirt. Judging by his sagging pants the shirt should say Lynch vs. belt ht
0,RightTroll,0,905874659358453760,914621840496189440,http://twitter.com/905874659358453760/statuses/914621840496189440
Daughter of fallen Navy Sailor delivers powerful monologue on anthem protests, burns her NFL packers gear. #BoycottNF
,Right,1,RightTroll,0,905874659358453760,914623490375979008,http://twitter.com/905874659358453760/statuses/91462349037

JUST IN: President Trump dedicates Presidents Cup golf tournament trophy to the people of Florida, Texas and Puerto Ri

0,RightTroll,0,905874659358453760,914639143690555392,http://twitter.com/905874659358453760/statuses/914639143690555392
19,000 RESPECTING our National Anthem! #StandForOurAnthem🇺🇸 https://t.co/czutyGaMQV",Unknown,English,10/1/2017 2:13,10
Right,1,RightTroll,0,905874659358453760,914312219952861184,http://twitter.com/905874659358453760/statuses/914312219952

Dan Bongino: ""Nobody trolls liberals better than Donald Trump."" Exactly! https://t.co/AigV93aC8J",Unknown,English,1
,RightTroll,0,905874659358453760,914320835325853696,http://twitter.com/905874659358453760/statuses/914320835325853696,
🐝🐝 https://t.co/MorL3AQW0z,Unknown,English,10/1/2017 2:48,10/1/2017
Right,1,RightTroll,0,905874659358453760,914321156466933760,http://twitter.com/905874659358453760/statuses/914321156466

@SenatorMenendez @CarmenYulinCruz Doesn't matter that CNN doesn't report on your crimes. This won't change the fact th
,RightTroll,0,905874659358453760,914322215537119234,http://twitter.com/905874659358453760/statuses/914322215537119234,
As much as I hate promoting CNN article, here they are admitting EVERYTHING Trump said about PR relief two days ago. h
,RightTroll,0,905874659358453760,914335818503933957,http://twitter.com/905874659358453760/statuses/914335818503933957,

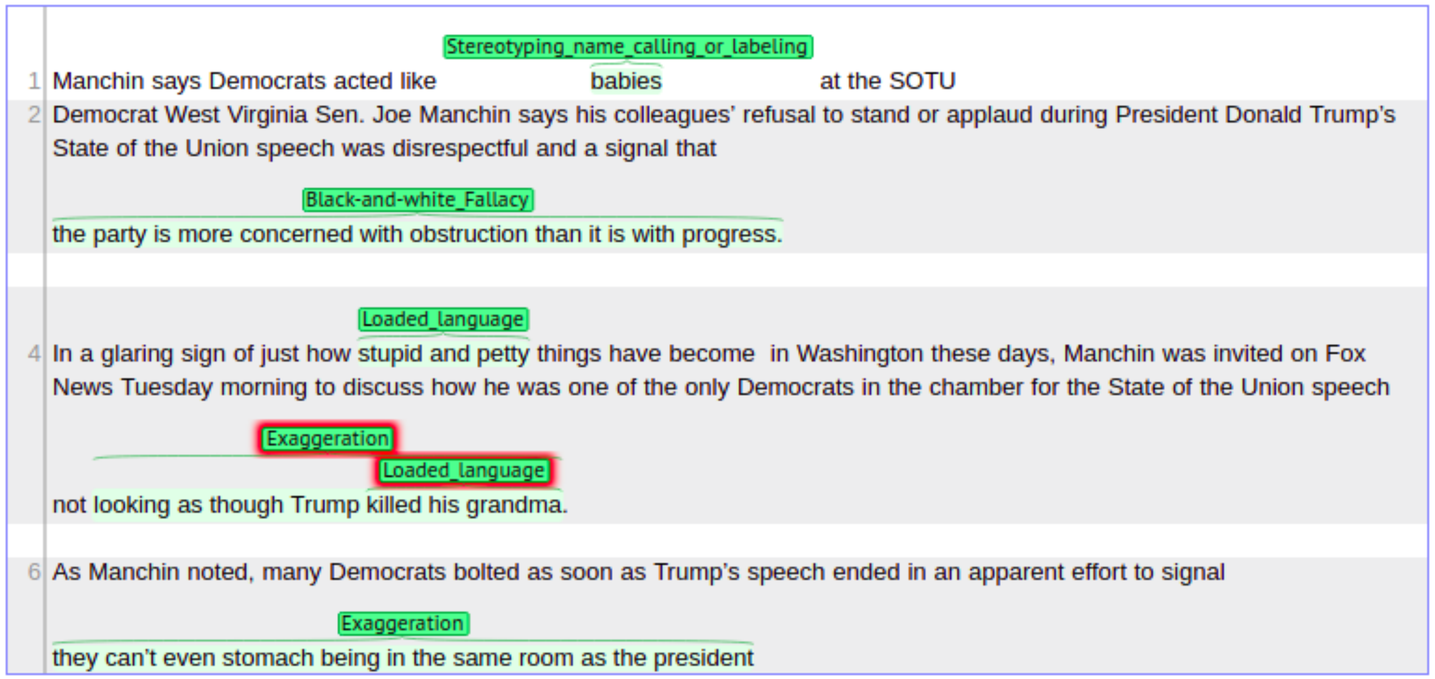
fter the 'genocide' remark from San Juan Mayor the narrative has changed though. @CNN fixes it's reporting constantly.
,RightTroll,0,905874659358453760,914336862730375170,http://twitter.com/905874659358453760/statuses/914336862730375170,
fter the 'genocide' remark from San Juan Mayor the narrative has changed though. @CNN fixes its reporting constantly.,
,RightTroll,0,905874659358453760,914338590313902080,http://twitter.com/905874659358453760/statuses/914338590313902080,
@thehill Why won't she apologize to us for lying?',Unknown,English,10/1/2017 4:11,10/1/2017
,RightTroll,0,905874659358453760,914342060437753857,http://twitter.com/905874659358453760/statuses/914342060437753857,
Sarah Sanders destroys NBC reporter: ""Trump made a joke, maybe you guys should get a sense of humor."" https://t.co/I
,(53760/statuses/917856606729236480
Hi @Michelleobama, remember when you said Weinstein is a wonderful human being, a good friend and just a powerhouse.

[https://github.com/fivethirtyeight/russian-troll-tweets]

What's new now?

(1) Better tools to **detect** and **categorize** misleading language generated by human authors.

Classifying propaganda

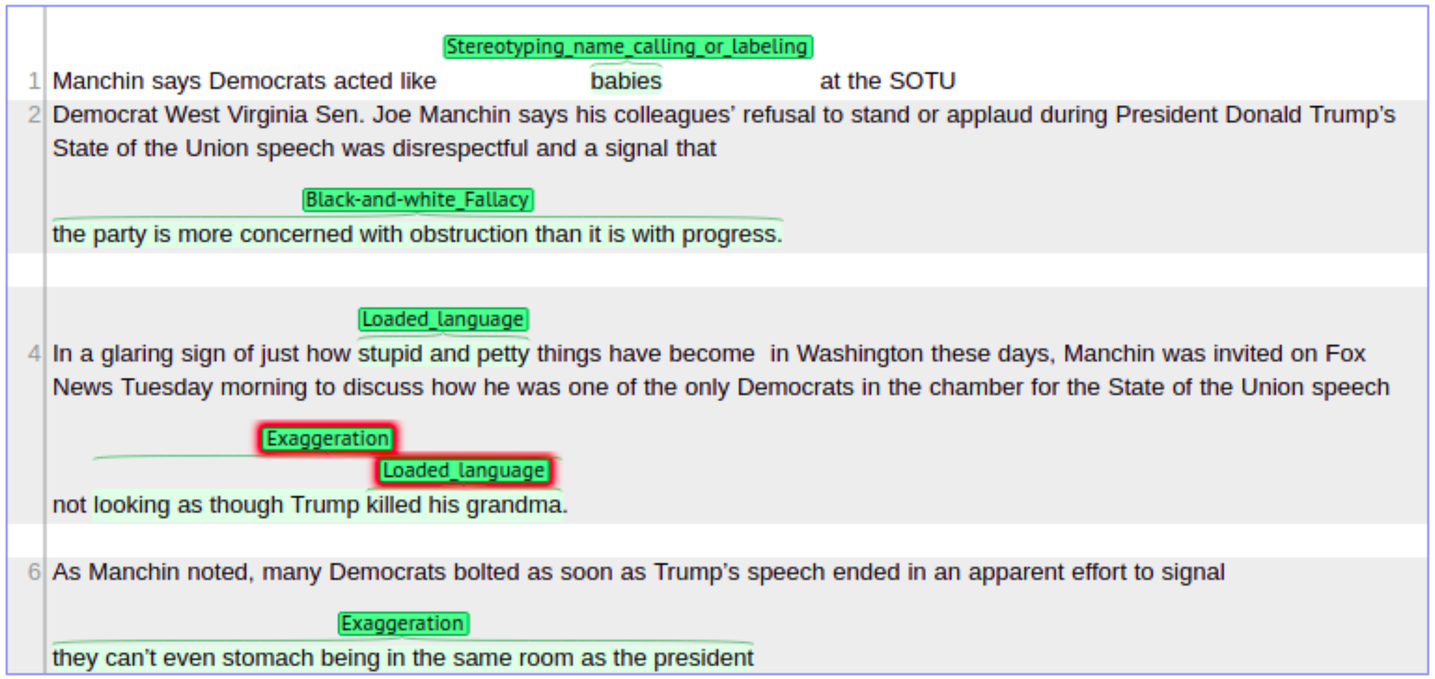


Annotations		spans (γ_s)	+labels (γ_{sl})
a_1	a_2	0.30	0.24
a_3	a_4	0.34	0.28
a_1	c_1	0.58	0.54
a_2	c_1	0.74	0.72
a_3	c_2	0.76	0.74
a_4	c_2	0.42	0.39

low inter-annotator agreement!

[da San Martino et al.]

Classifying propaganda



low model accuracy!

Model	Precision	Recall	F1
All-Propaganda	23.92	1.00	38.61
BERT	**63.20**	53.16	57.74
BERT-Granu	62.80	55.24	58.76
BERT-Joint	62.84	55.46	58.91
MGN Sigmoid	62.27	59.56	60.71
MGN ReLU	60.41	**61.58**	**60.98**

[da San Martino et al.]

What's new now?

(1) Better tools to **detect** and **categorize** misleading language generated by human authors.

(2) Better tools to **synthesize** misleading language **without** human authors!

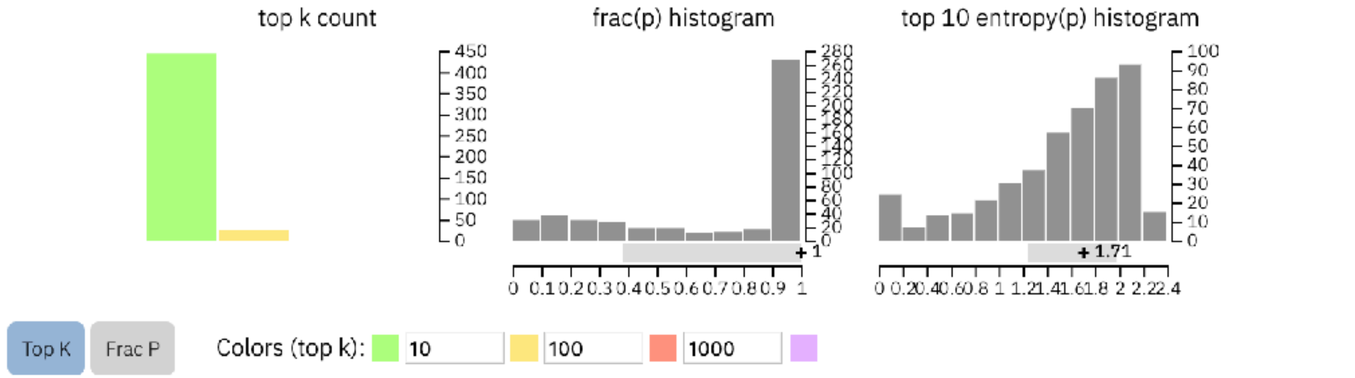
(a problem of scale)

Public comments on Idaho Medicaid Reform Waiver

Comment	Response ID
I support Governor Little's efforts to overhaul Idaho's Medicaid program.	459669
Medicaid is an important safety net program. It helps people who are losing their coverage to get back on their feet. We need to make health and wellness a priority for the Medicaid program in Idaho.	459825
I am writing to you today regarding Idaho's Medicaid waiver proposal, I oppose the aspects of this program that create new burdens on people who are already struggling. The proposed changes to Medicaid could deny health insurance to sick individuals when they are most in need. I do not support this approach that creates barriers to access. I am hopeful that you change the proposed waiver.	460129

[Weiss 2019]

Recognizing fake text: word frequencies



I've been a gamer for over ten years. During that time, I've been involved in a number of games, and I've seen very few of them in the history of the company. My first foray into this was as a member of the U.S. Army. I played some of the games I liked from the early 1980s through the early 1990s, but my first foray into the hobby was at the beginning of 2000 when I was stationed in Afghanistan. After I got back to my hometown and went to school, I started playing games. I began playing multiplayer games, which was a very popular form of gaming. One of the games I started playing was the first-person shooter "The Wolf Among Us" which is still the best-selling title of all time.

I was at the beginning of the game development process. I had already seen a few demos of the game. I was also very interested in the multiplayer aspects of the game, and I wanted to see what the players would do in the game. In the beginning, I didn't know about multiplayer. I thought it would be cool to have some sort of "party game" with some kind of "game mode" which would give the player a real advantage. But as time went on, I realized that there were a lot of different things I wanted to create. To make it fun for the player, the multiplayer component was added.

I started playing the game as a member of the U.S. Army. When I returned to my hometown, I found myself in the middle of a war with a group of Taliban soldiers. I was killed by one of the Taliban and I was the only casualty. I decided to take a look at multiplayer. I took the chance to have some fun with the multiplayer. I was in a place that was pretty hostile to the Taliban, and I decided that I wanted to make it fun for the player.

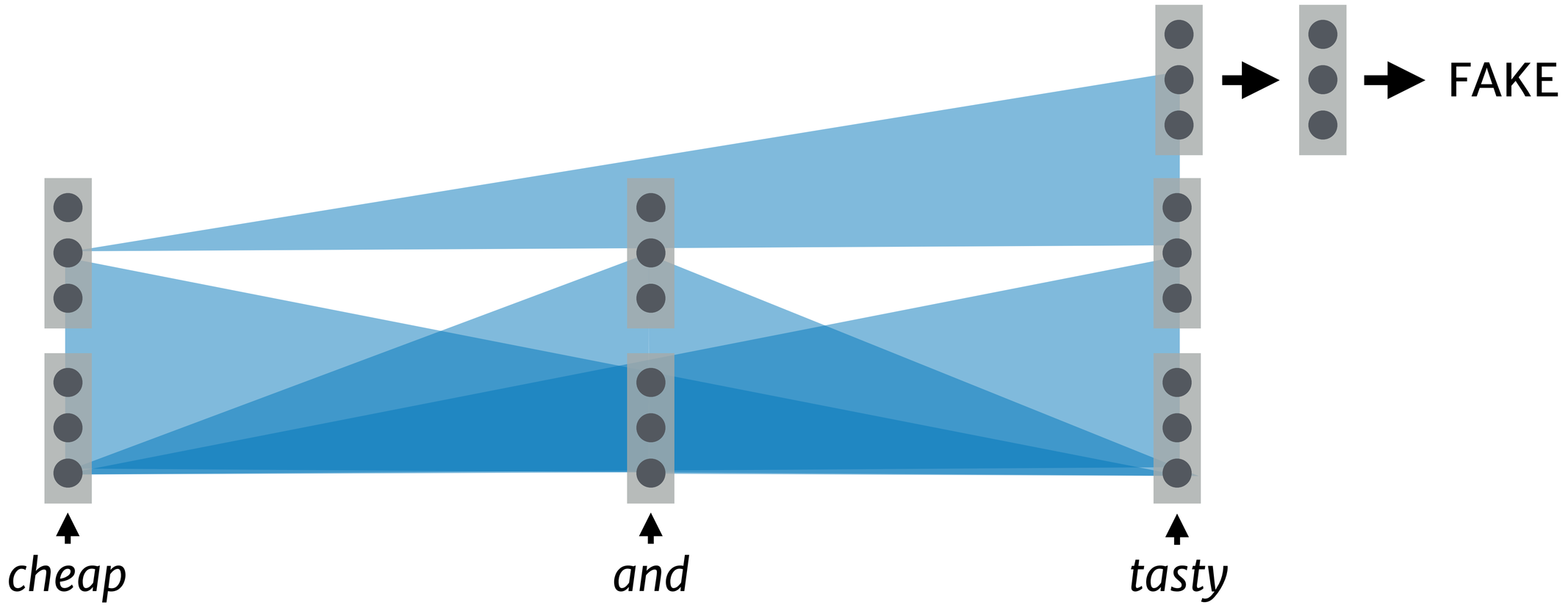
The game was designed to be a good way of showing off combat experience. It was supposed to be a combat-focused game, and I wanted to show off how well the players could play. The multiplayer was designed to be a nice way to show off that. The game is a multiplayer game, and the game is designed to be a fun and interesting multiplayer game.

[Strobelt et al. 2019]

Recognizing fake text: word frequencies

[Strobelt et al. 2019]

Recognizing fake text: Fine-tune a language model!



[Zellers et al. 2019]

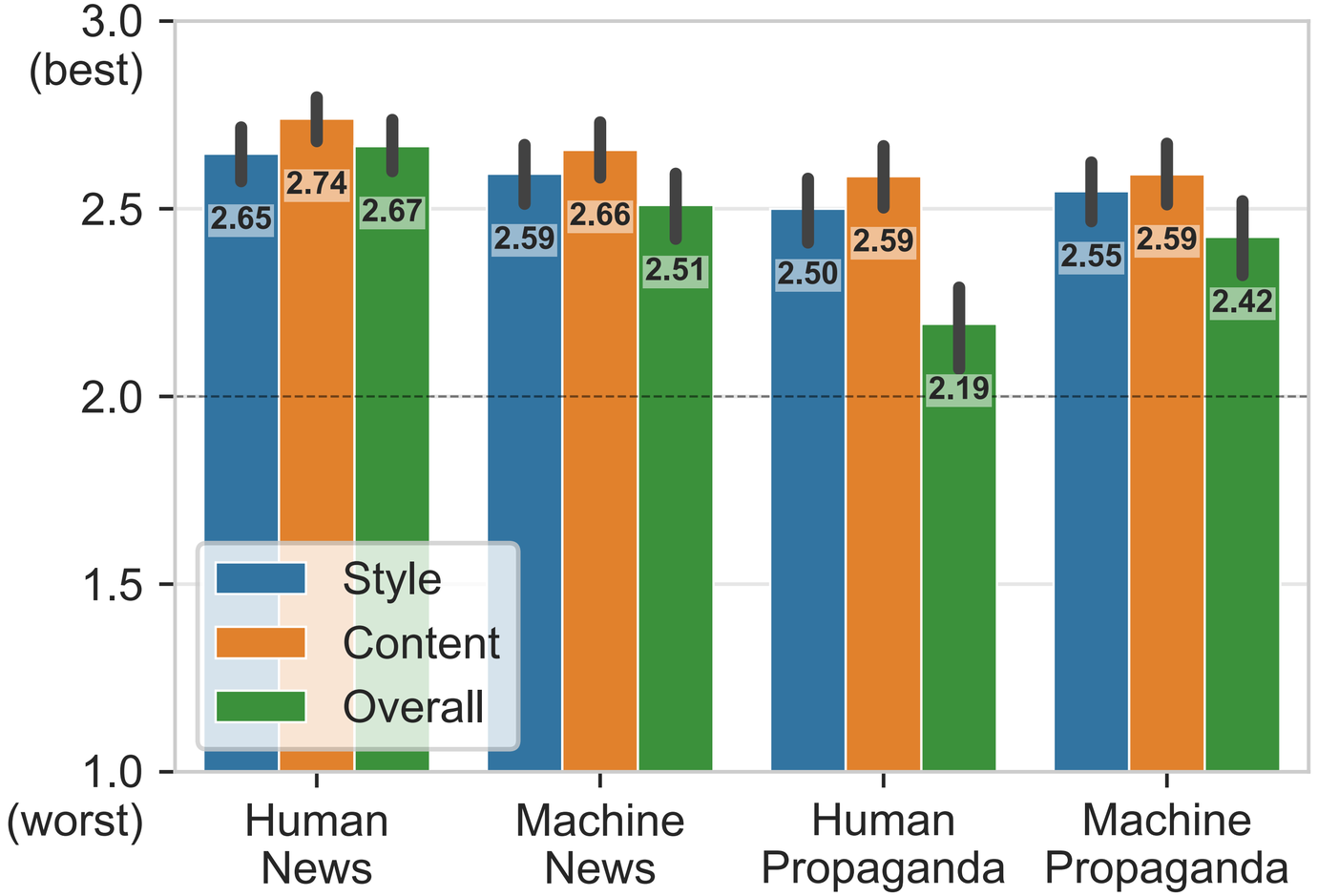
Recognizing fake text: Fine-tune a language model!

Some findings:

- Works best if you use representations from the same model that generated the fake language!
- If you don't have that, semi-supervised training on (fake sentences from your own LM, fake sentences from the LM you're trying to fight) is pretty effective.

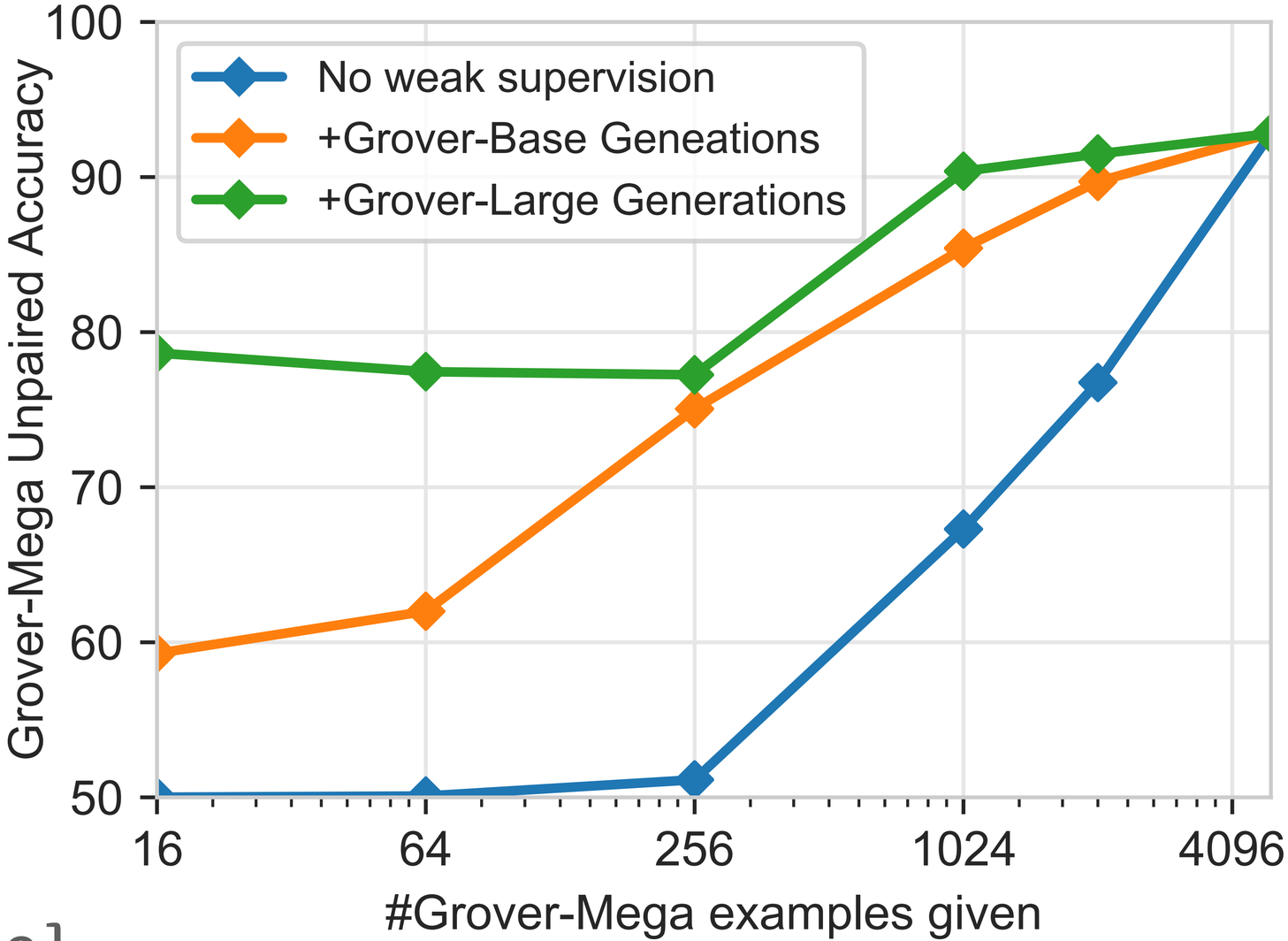
[Zellers et al. 2019]

Who's winning the arms race?



[Zellers et al. 2019]

Who's winning the arms race?



[Zellers et al. 2019]

Summary

NLP tools can help identify misleading / propagandistic information online, whether generated by humans or automated systems.

(These are also tools for censorship!)

In general, better generation $\iff$ better detection; unclear which problem will be easier in the long term.

Course recap

Learning representations from linguistic context

Latent Semantic Analysis

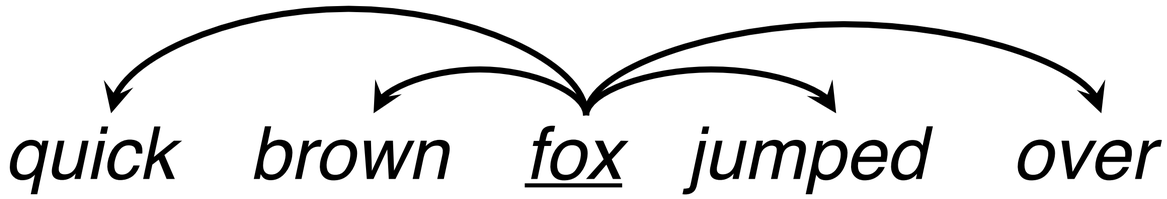
documents about animals

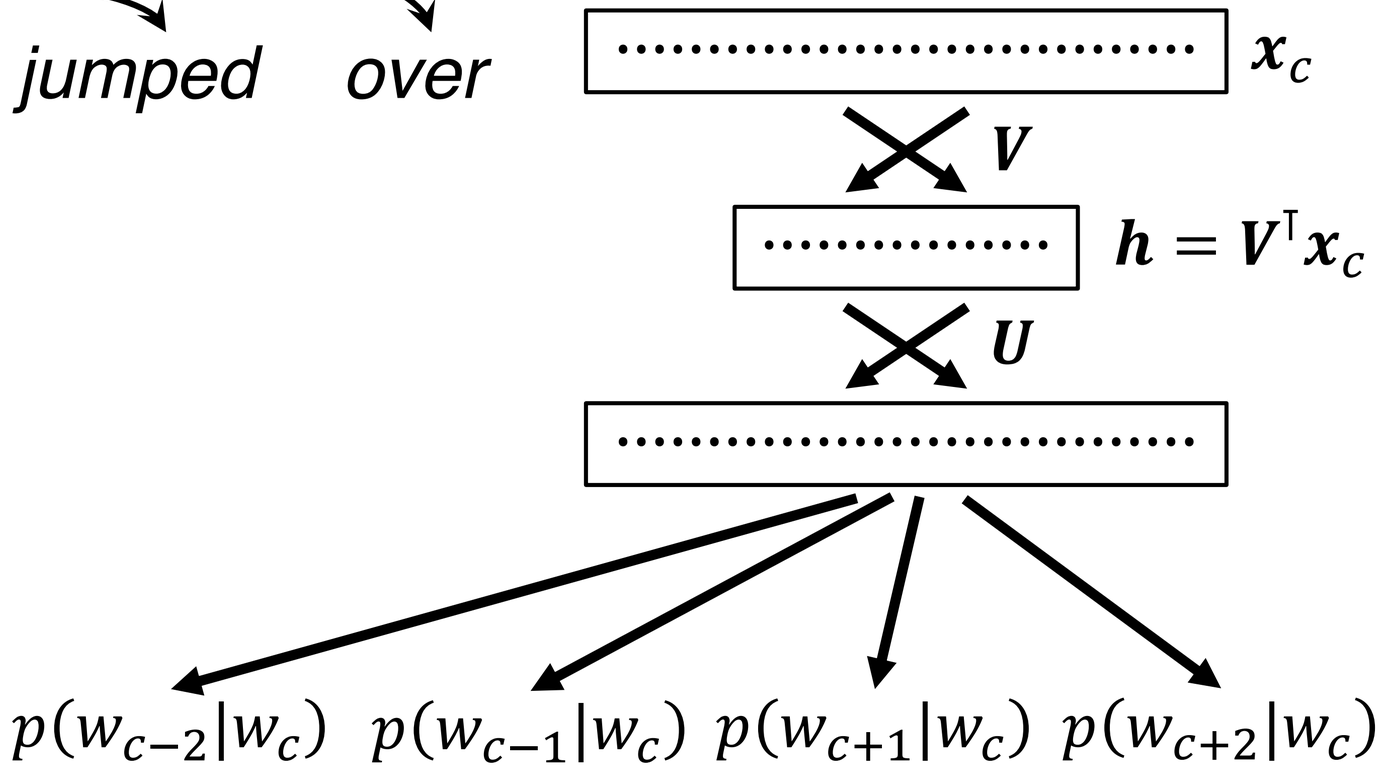
documents about computers

	d_1	d_2	d_3	d_4	d_5	d_6	d_7
cat	1	0	1	0	0	0	0
paw	0	1	1	0	0	0	0
algorithm	0	0	0	1	1	1	1

Learning representations from linguistic context

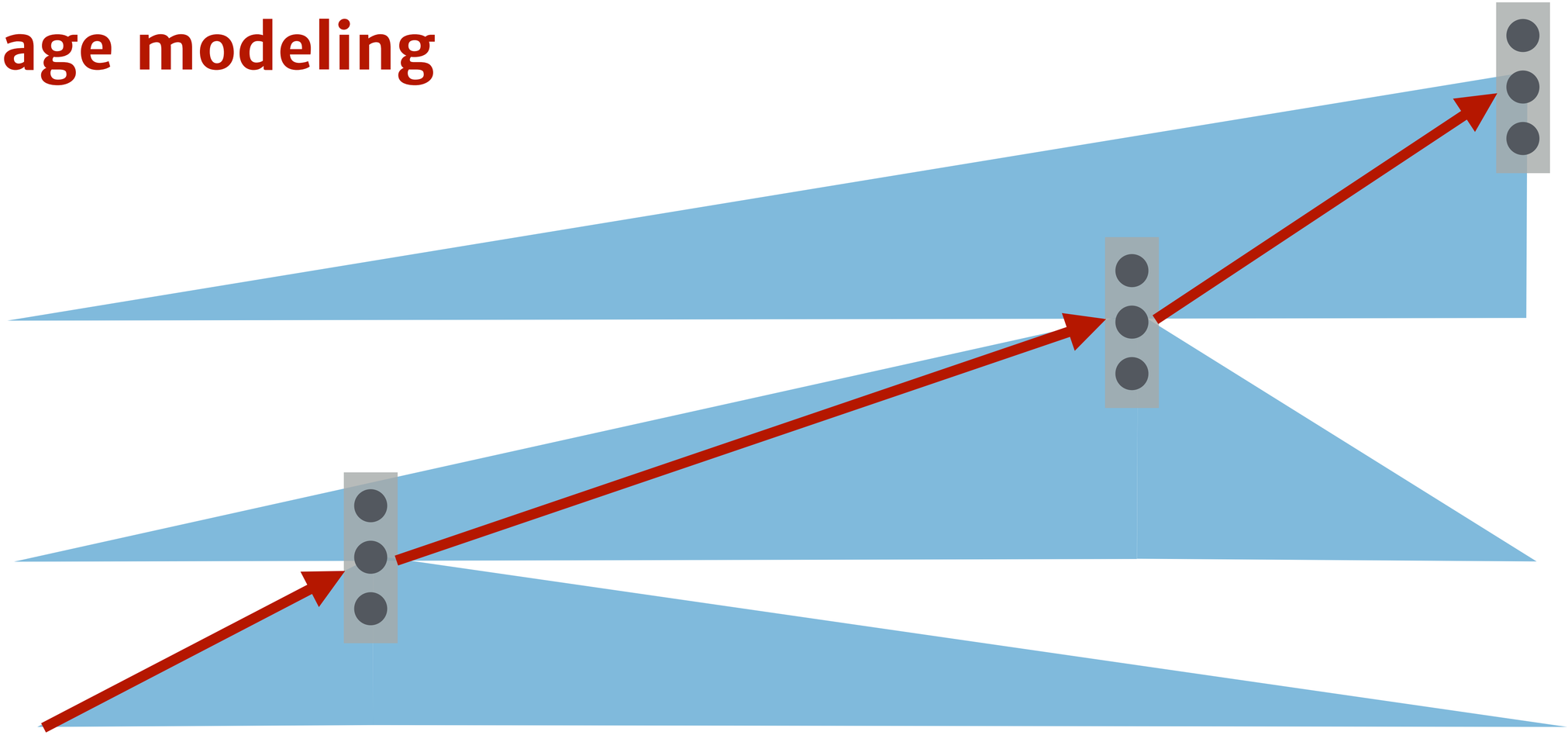
word2vec





Learning representations from linguistic context

language modeling



John has a book. Mary has an apple. He gave her his

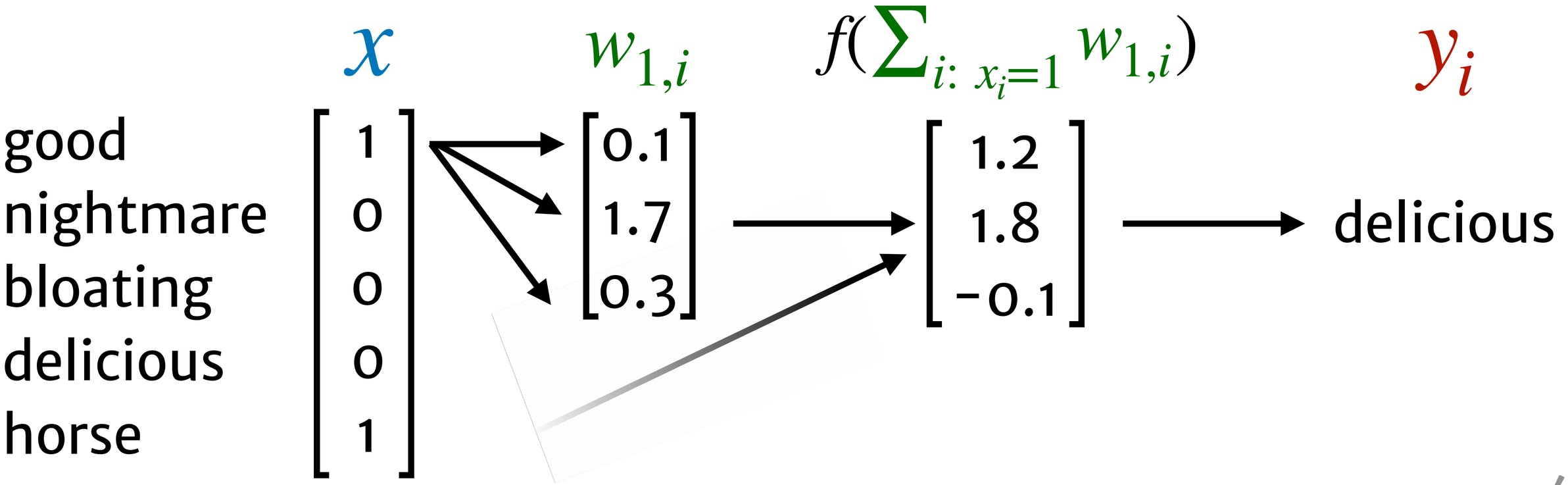
Learning representations from linguistic context

masked language modeling

Generating structured outputs from language

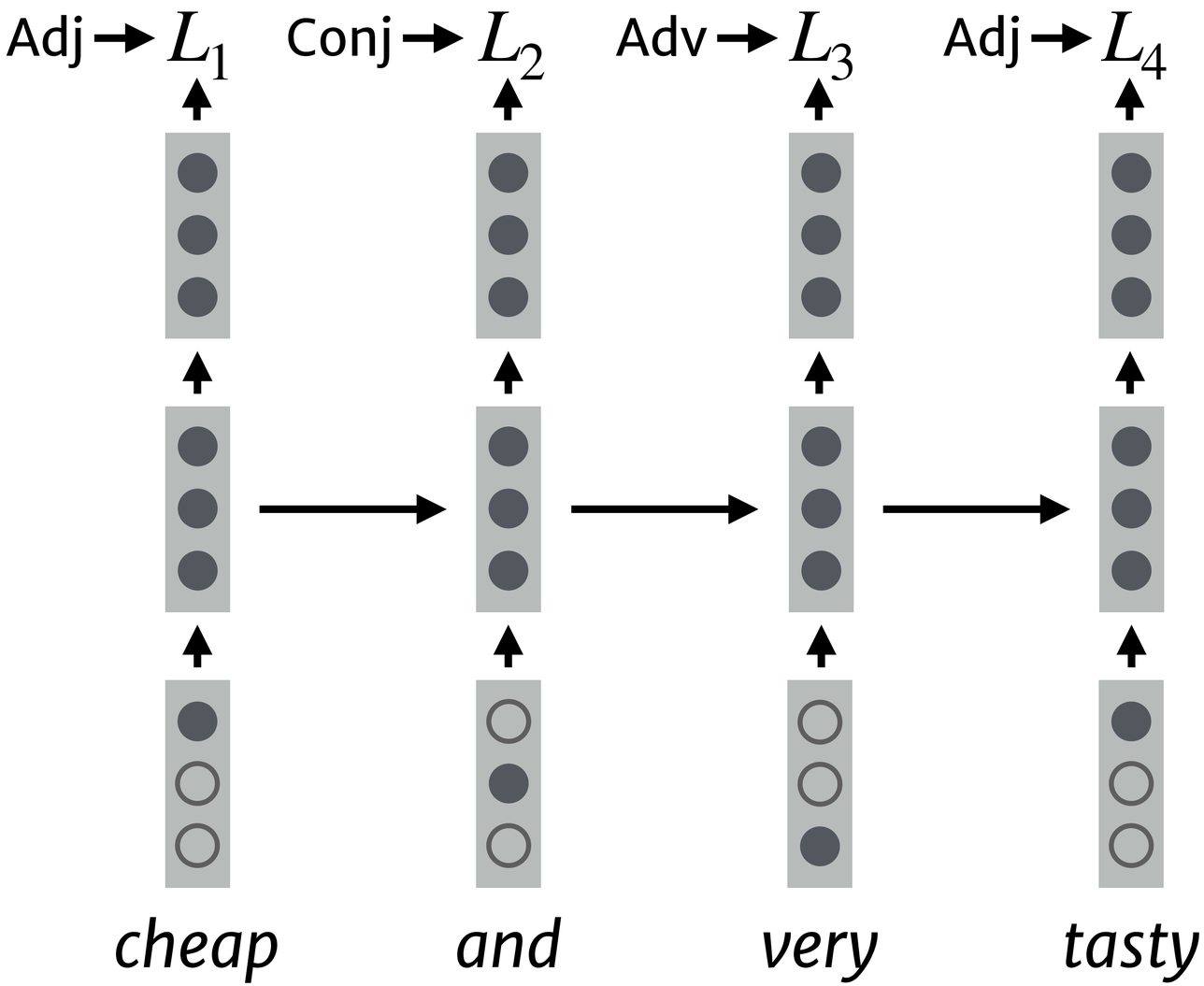
classification

$$s = W_2^\top f(W_1^\top x)$$



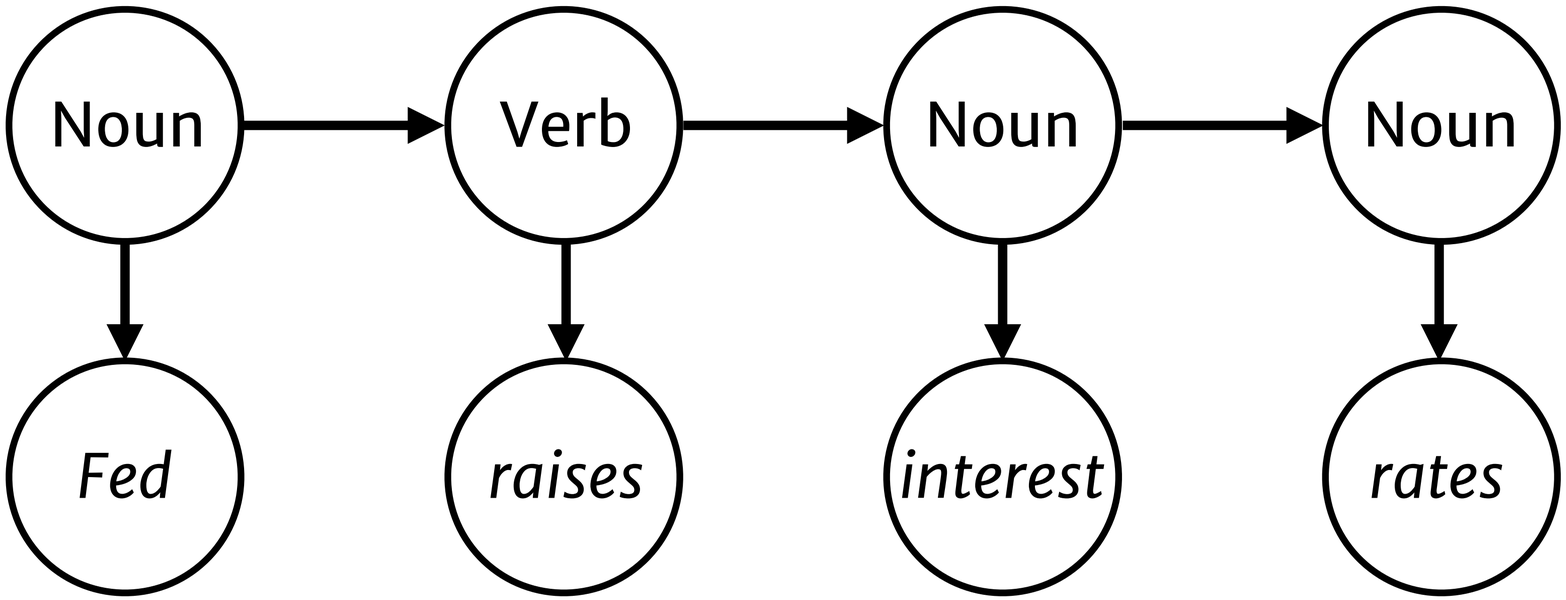
Generating structured outputs from language

tagging



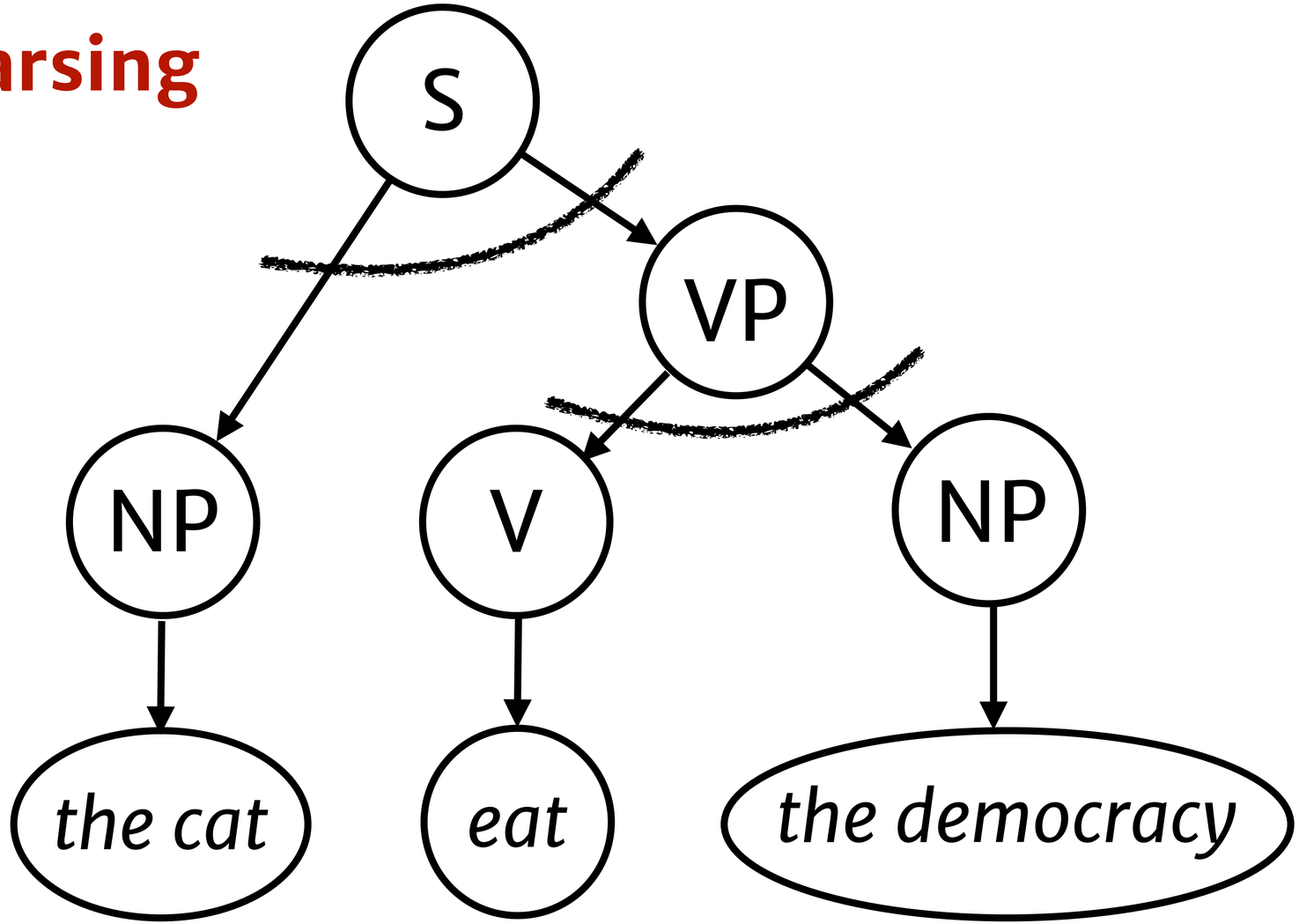
Generating structured outputs from language

sequence labeling



Generating structured outputs from language

syntactic parsing



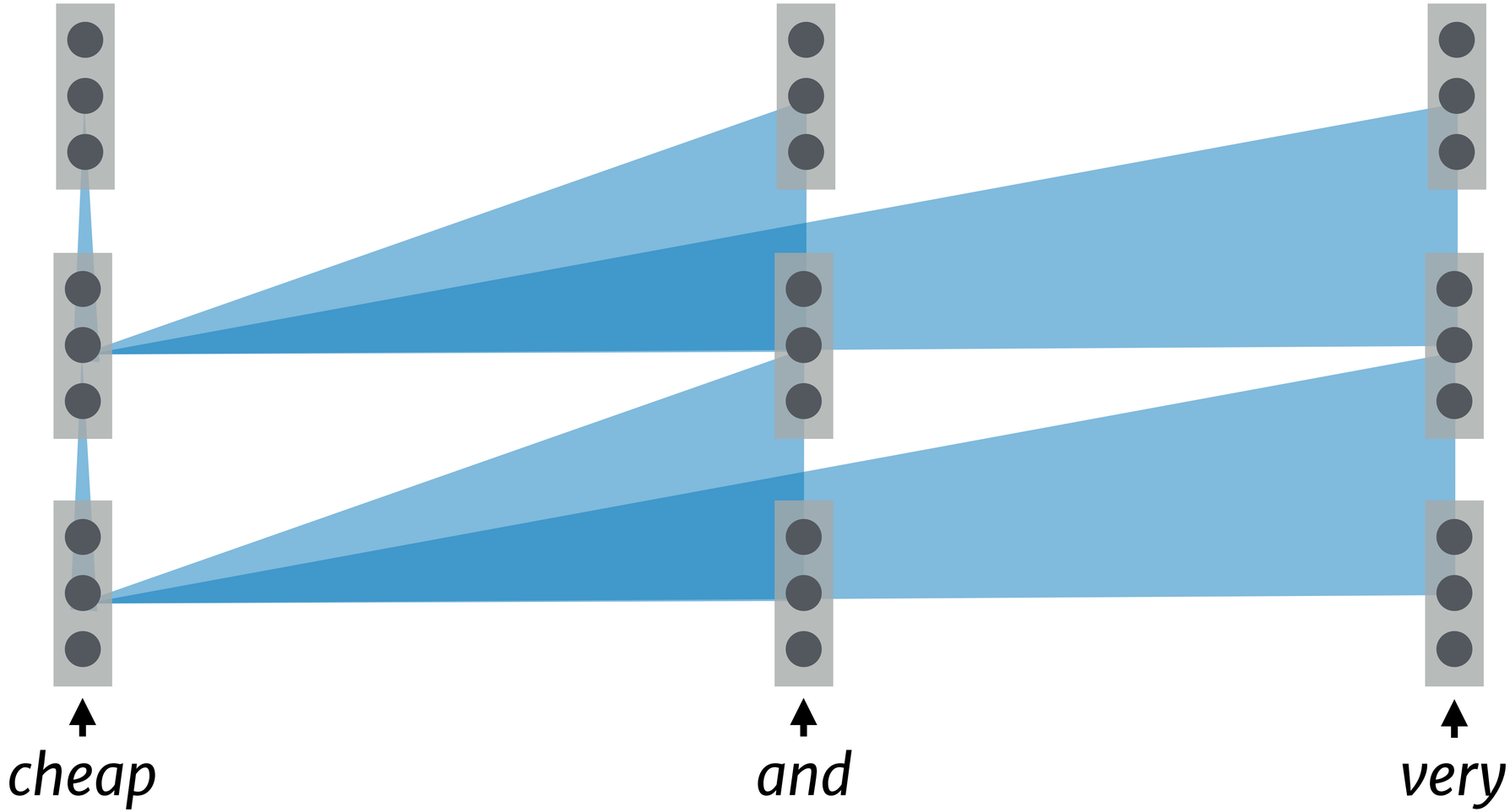
Generating structured outputs from language

general structures

Generating language

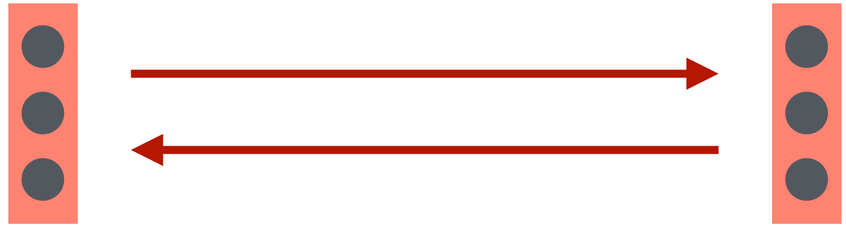
language modeling

$p(\text{tasty} \mid \text{cheap and very})$



Generating language

question answering

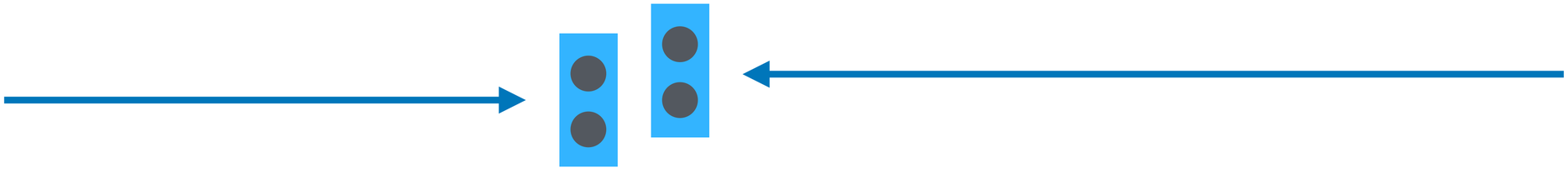


v_q: question rep.

what county is Jacksonville in?

v_d: document rep.

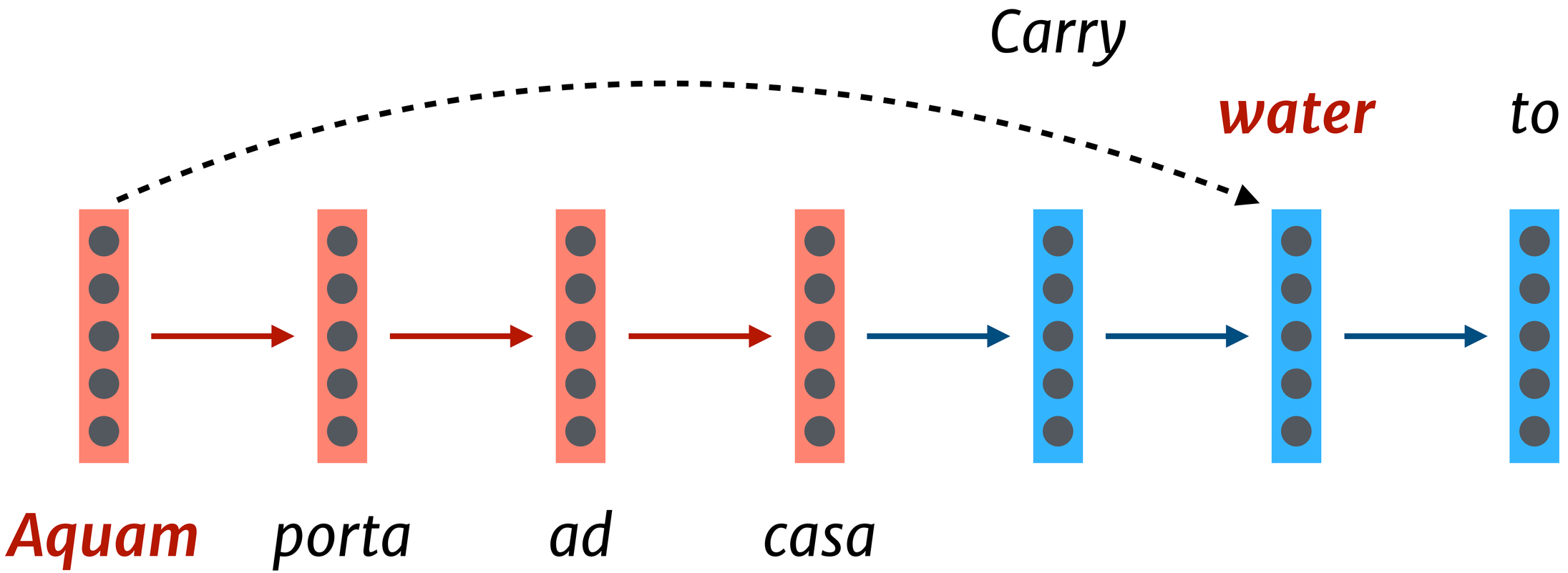
$$p(\text{start} = i) \propto v_q^\top v_d$$



It is the county seat of Duval County, with which the city government

Generating language

machine translation



Learn more

Speech processing: 6.345

Psycholinguistics: 9.190

Structured prediction & graphical models: 6.438

Syntax & semantics: 24.90{2,3}

Thanks to the staff

Jacob Andreas

Jim Glass

Abby Bertics

Dylan Doblar

Ekin Akyürek

Evan Hernandez

Hongyin Luo

Harini Suresh

Pranav Krishna

Wei Fang

Thanks to you!

Have fun with the rest of the project.

See you next week!