Structurizing Misinformation Stories via Rationalizing Fact-Checks

Shan Jiang and Christo Wilson

Northeastern University

1

Background: misinformation

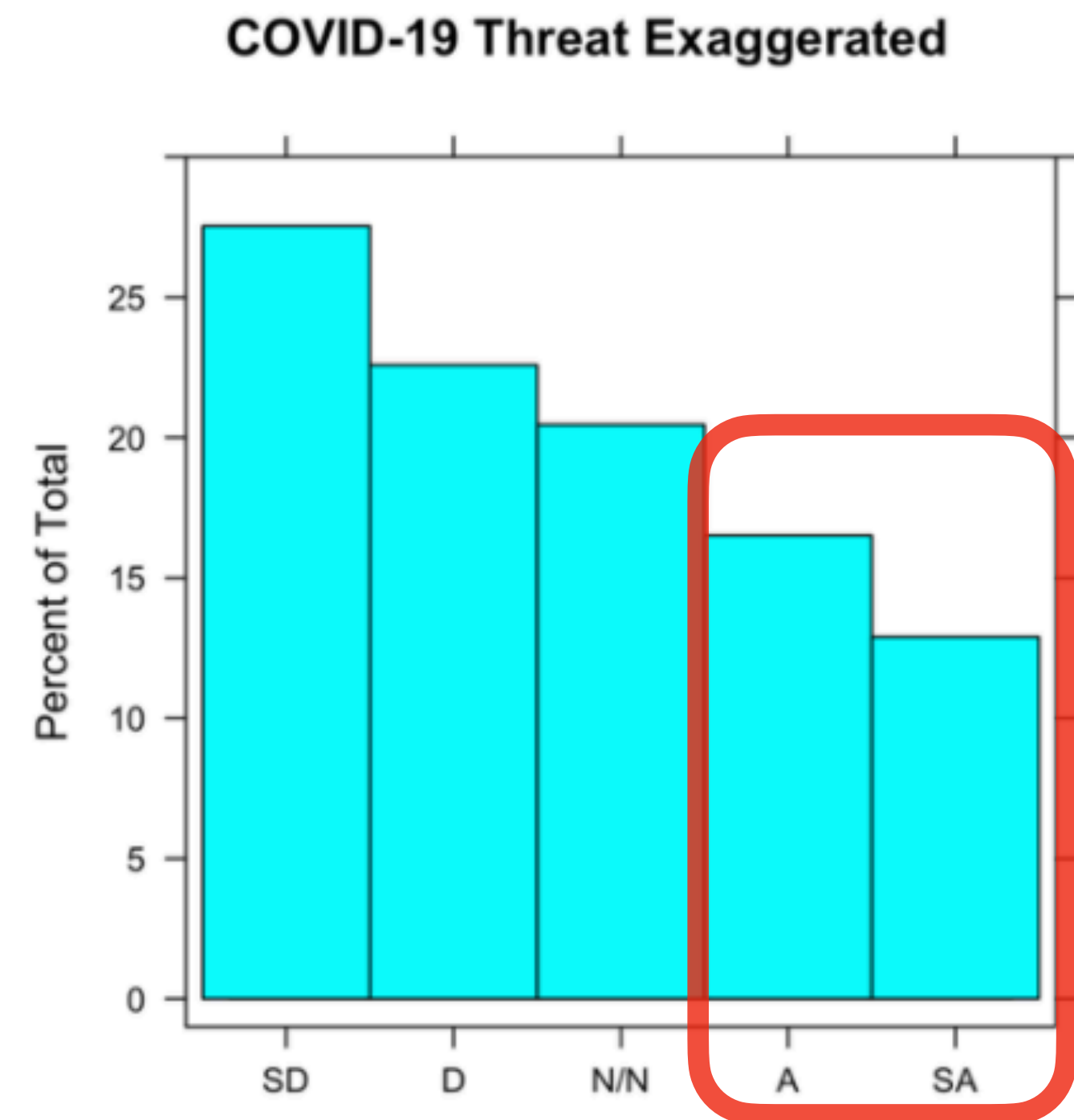**Structurizing Misinformation Stories via Rationalizing Fact-Checks**

bad / wrong / lack of / not / ...

Background: misinformation

COVID-19 Pandemic:

- **29% of respondents agree that the threat of COVID-19 is exaggerated [1].**

COVID-19 Threat Exaggerated

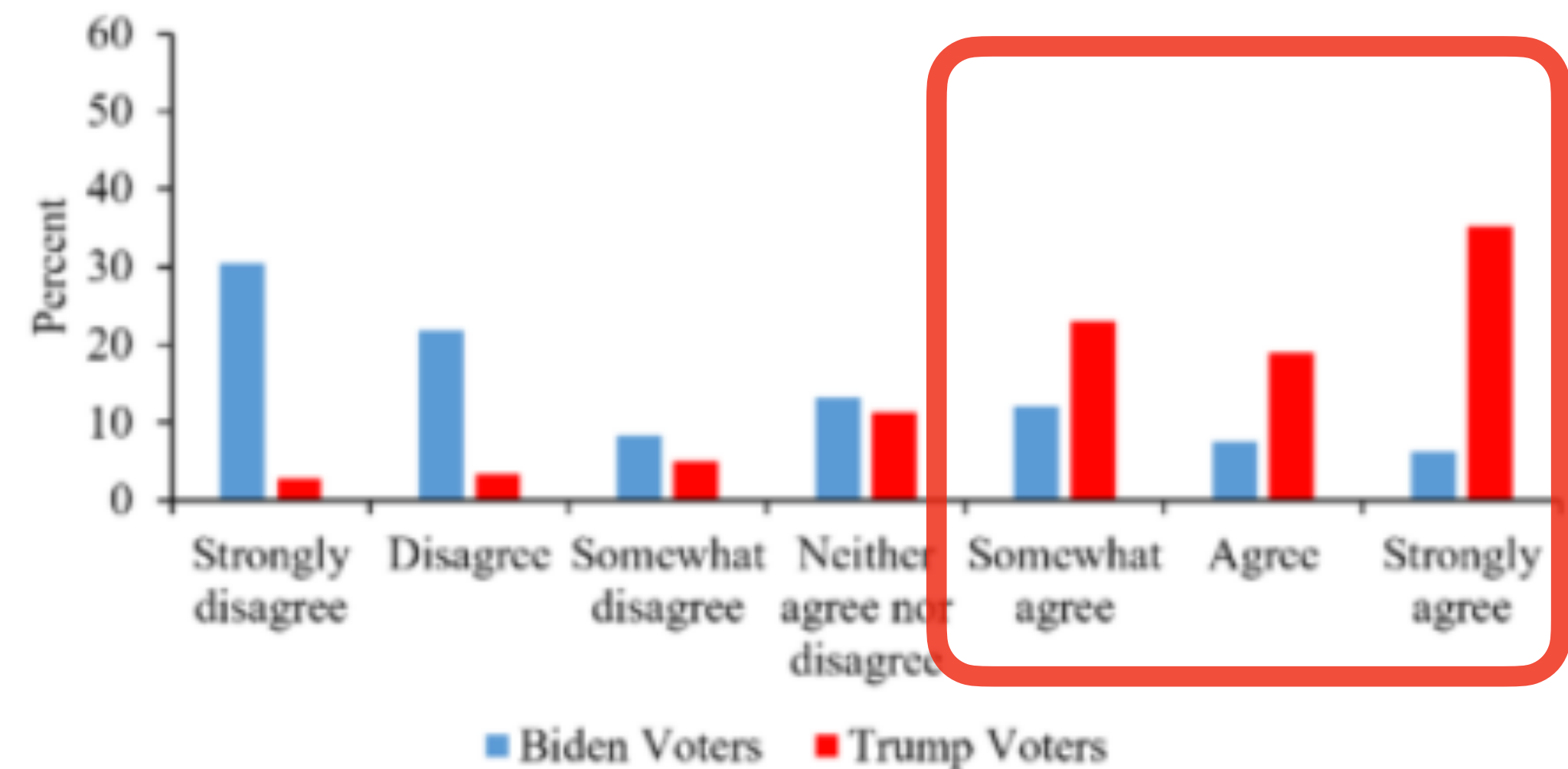[1] Uscinski, Joseph E., et al. "Why do people believe COVID-19 conspiracy theories?." *HKS Misinfo Review* 2020.

Background: **misinformation**

COVID-19 Pandemic:

- **29% of respondents agree that the threat of COVID-19 is exaggerated.**

2020 US Presidential Election:

- **77% of Trump voters believe that voter fraud is common [2].**

[2] Pennycook, G. and Rand, D. Examining false beliefs about voter fraud in the wake of the 2020 Presidential Election. *HKS Misinfo Review* 2021.

Background: misinformation types

Structurizing Misinformation Stories via Rationalizing Fact-Checks

bad / wrong / lack of / not / ...

Background: misinformation types

Structurizing Misinformation Stories via Rationalizing Fact-Checks

bad / wrong / lack of / not / …

urban legend

Background: misinformation types

Structurizing Misinformation Stories via Rationalizing Fact-Checks

bad / wrong / lack of / not / …

urban legend

satire

Background: misinformation types

Structurizing Misinformation Stories via Rationalizing Fact-Checks

bad / wrong / lack of / not / ...

urban legend

satire

conspiracy

Background: misinformation types

Structurizing Misinformation Stories via Rationalizing Fact-Checks

bad / wrong / lack of / not / ...

urban legend   **...**   **satire**   **...**   **conspiracy**

Background: misinformation types

Structurizing Misinformation Stories via

Rationalizing Fact-Checks

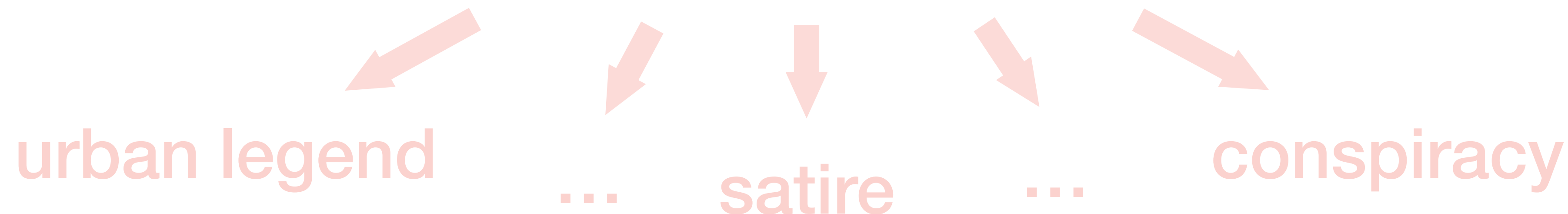bad / wrong / lack of / not / ...

urban legend   ... satire ... conspiracy

Background: **fact-checks**

Structurizing Misinformation Stories via Rationalizing **Fact-Checks**

bad / wrong / lack of / not / …

urban legend

… satire …

conspiracy

Background: fact-checks

Background: example fact-check

Snopes

Greta Thunberg is the 'Highest Paid Activist'?

…For instance, some started sharing a doctored photograph of Thunberg with alt-right boogeyman George Soros (the original photograph featured former Vice President Al Gore)…

Background: fact-check example

Greta Thunberg is the 'Highest Paid Activist'?

…For instance, some started sharing a doctored photograph of Thunberg with alt-right boogeyman George Soros (the original photograph featured former Vice President Al Gore)…

False

Background: phrases in fact-checks

Greta Thunberg is the 'Highest Paid Activist'?

…For instance, some started sharing a doctored photograph of Thunberg with alt-right boogeyman George Soros (the original photograph featured former Vice President Al Gore)…

False

Background: **misinformation types**

Snopes

Greta Thunberg is the 'Highest Paid Activist'?

…For instance, some started sharing a doctored photograph of Thunberg with alt-right boogeyman George Soros (the original photograph featured former Vice President Al Gore)…

False

Background: **misinformation types**

Greta Thunberg is the 'Highest Paid Activist'?

…For instance, some started sharing a **doctored photograph** of Thunberg with alt-right boogeyman George Soros (the original photograph featured former Vice President Al Gore)…

False

Background: **more fact-checks**

...a classic urban legend about...
...to substantiate   the tale   through...

...these doctored images have featured...
...was a   digitally altered   mashup of...

...this is a prank from...   ...a survey   scam   that...
...another   hoax   when...   ...posting bogus offers...

...that a media conspiracy sought...
...unsubstantiated and   baseless   stories...

...responsibility for the satirical nature of...
...just a bit of political   humor   from the...

Background: **more misinformation types**

...a classic **urban legend** about...
...to substantiate **the tale** through...

...these **doctored images** have featured...
...was a **digitally altered** mashup of...

...this is a **prank** from... ...a survey **scam** that...
...another **hoax** when... ...posting **bogus** offers...

...that a media **conspiracy** sought...
...unsubstantiated and **baseless** stories...

...responsibility for the **satirical nature** of...
...just a bit of political **humor** from the...
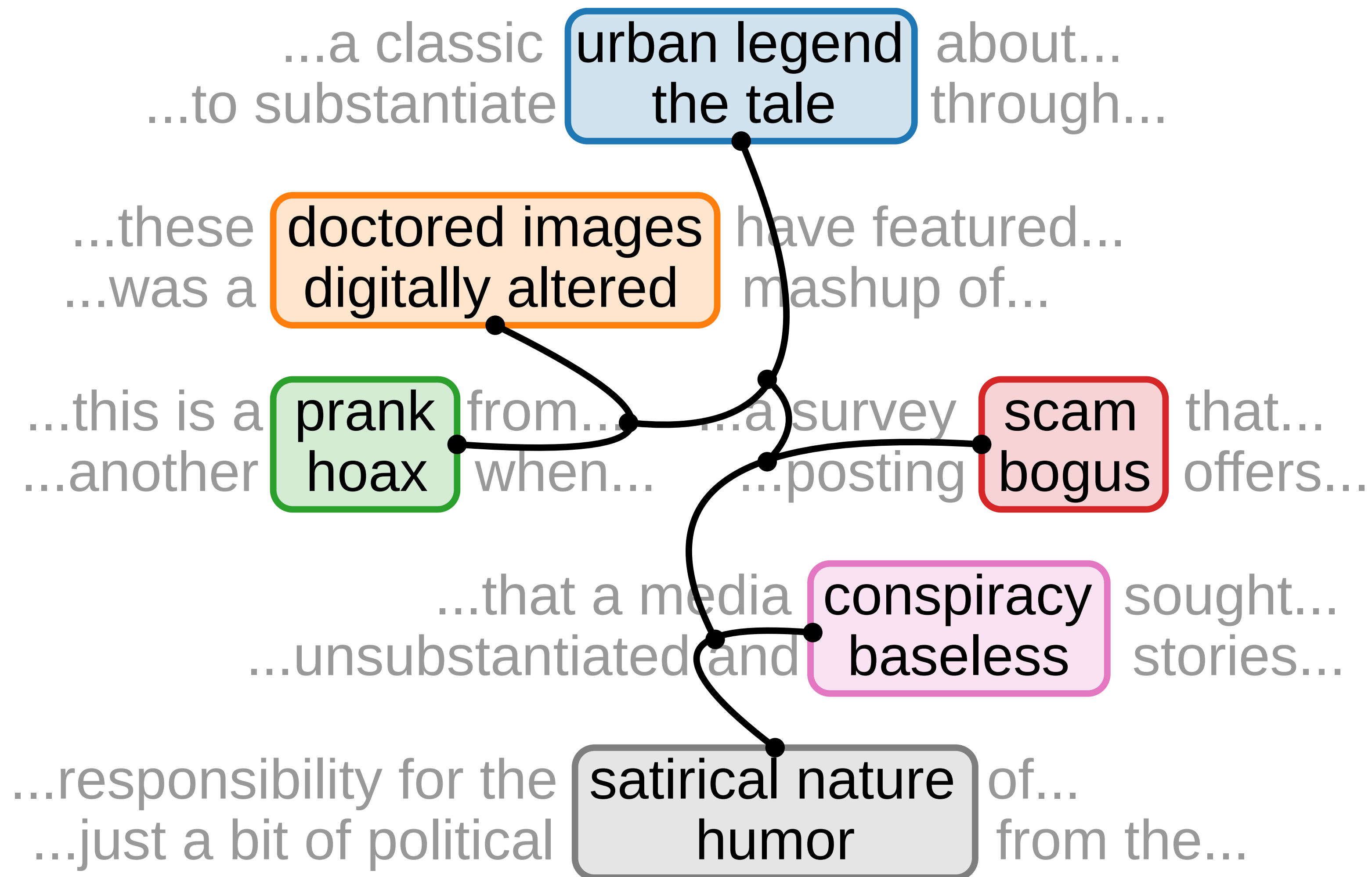
Background: rationalizing - *fact-checks*

…For instance, some started sharing a **doctored photograph** of Thunberg with alt-right boogeyman George Soros…   →   **False**

Background: rationalizing - *movie reviews*

…For instance, some started sharing a doctored photograph of Thunberg with alt-right boogeyman George Soros…   →   **False**

…Until then, the usual boring cliches of college life (booze, sex, more booze) fill the time, which is made to feel longer…   →   **Negative**

Background: rationalizing - *offensive content*

…For instance, some started sharing a doctored photograph of Thunberg with alt-right boogeyman George Soros…

False

…Until then, the usual boring cliches of college life (booze, sex, more booze) fill the time, which is made to feel longer…

Negative

… YOU FUCKING KIKE! DON'T EDIT THINGS YOU HAVE NO IDEA ABOUT! GIVE CREDIT WHERE IT'S DUE!…

Offensive

Background: **rationalizing**

Structurizing Misinformation Stories via

Rationalizing Fact-Checks

neural models that can "reason" about their predictions
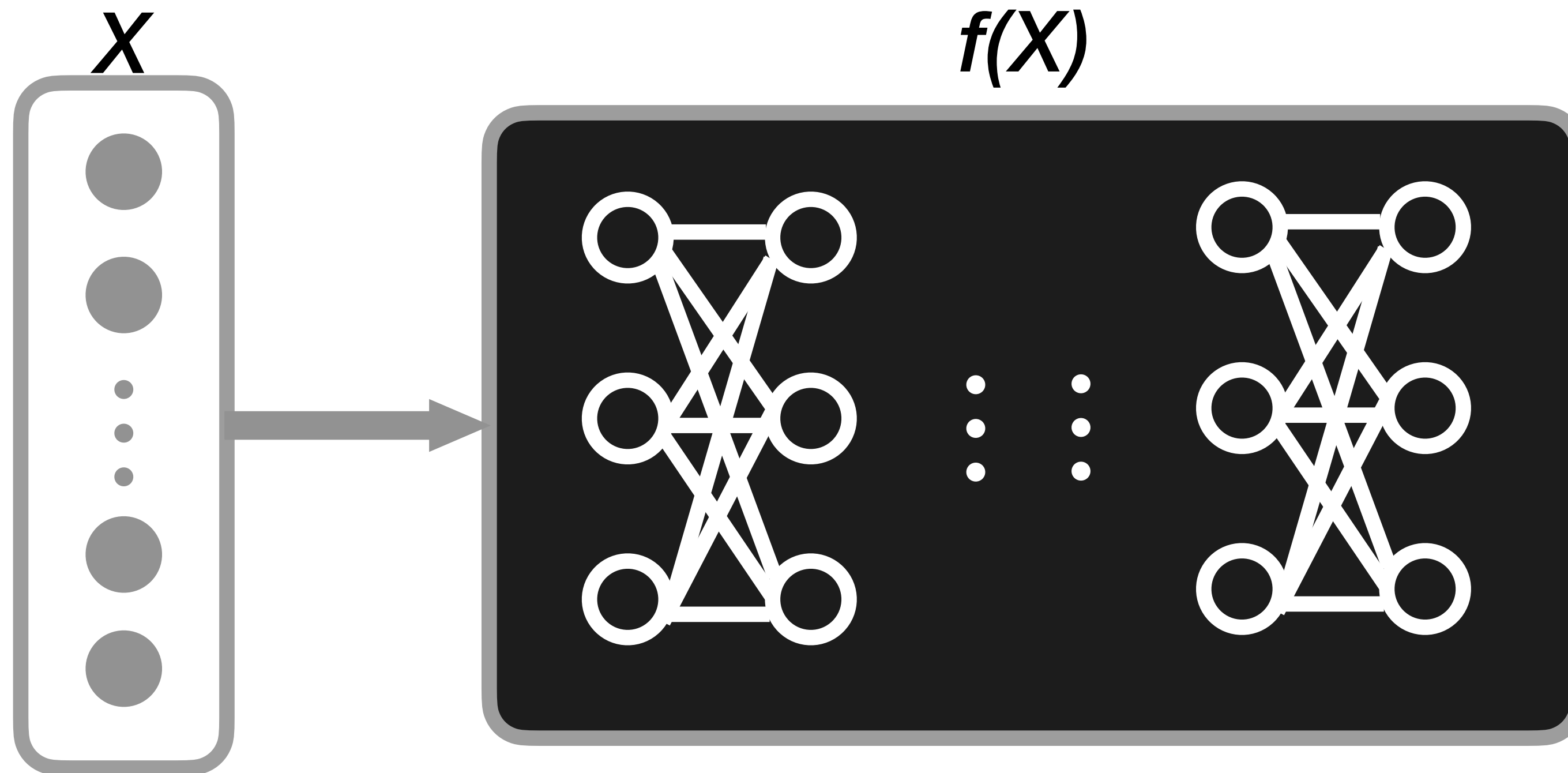
Methods: **neural models**

…For instance, some started sharing
a doctored photograph of Thunberg with
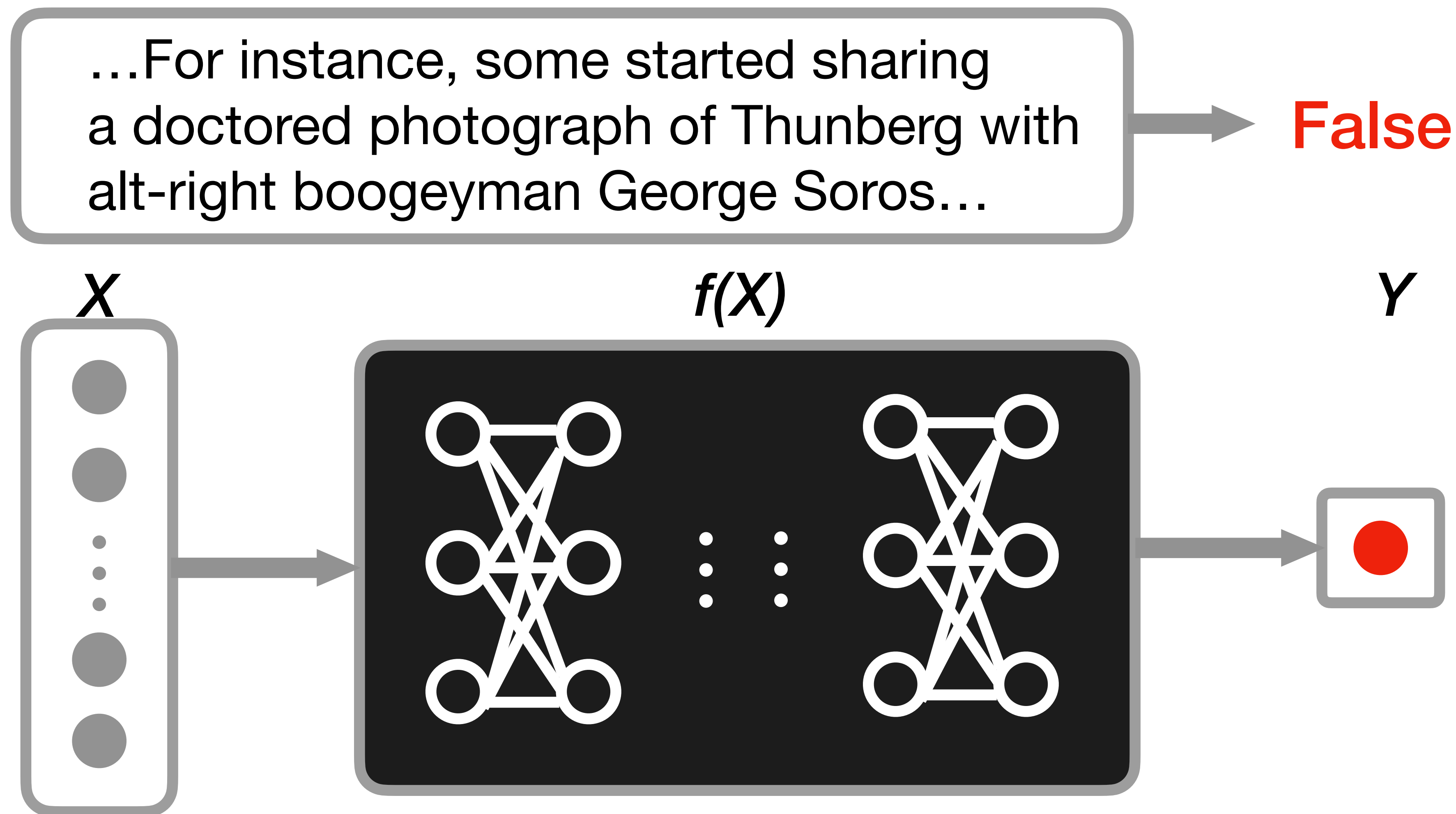alt-right boogeyman George Soros…
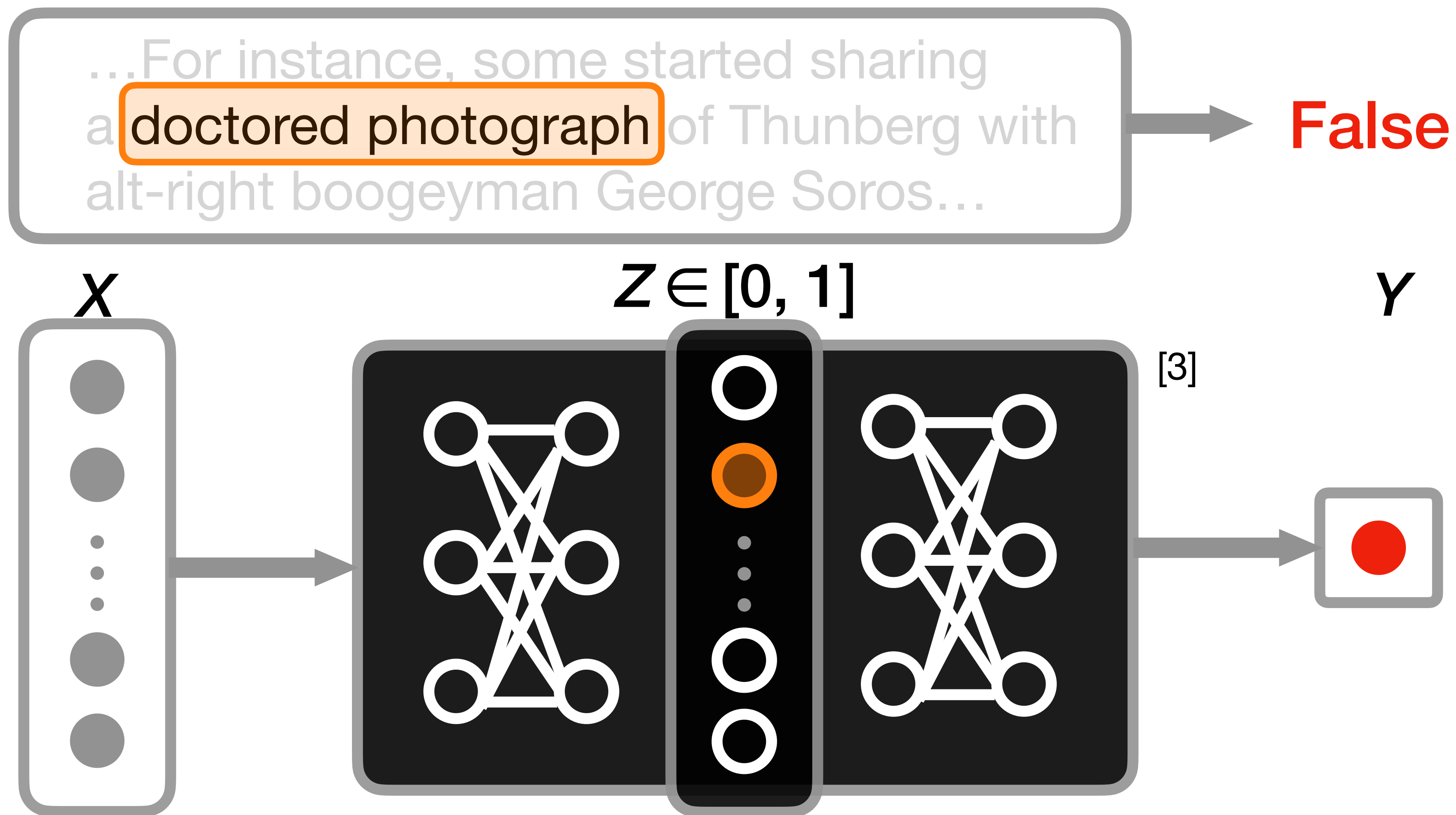
X

Methods: **neural models**

…For instance, some started sharing a doctored photograph of Thunberg with alt-right boogeyman George Soros…

X

f(X)

Methods: **neural models**

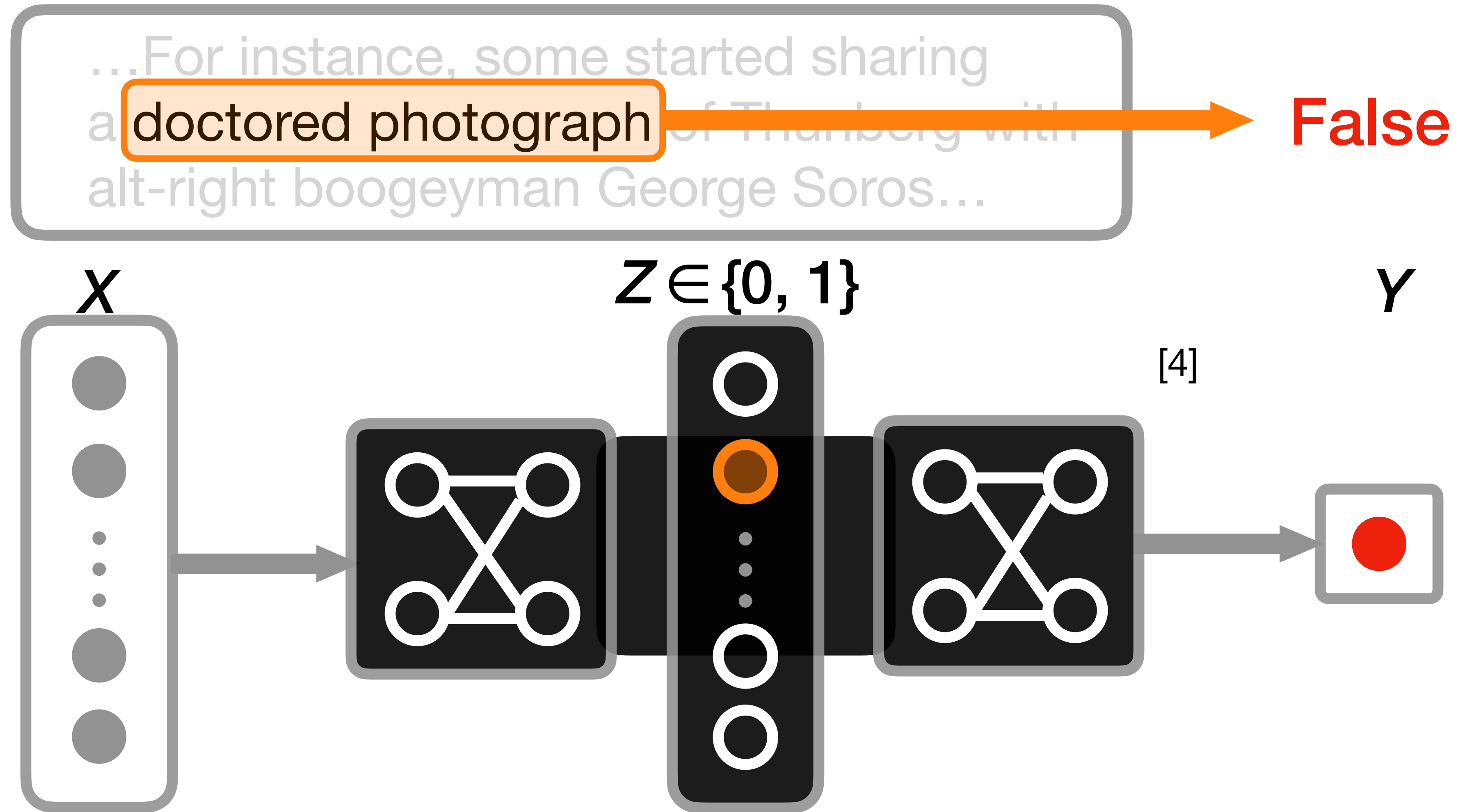…For instance, some started sharing a doctored photograph of Thunberg with alt-right boogeyman George Soros… **→** **False**

X   *f(X)*   *Y*

Methods: **rationalized models -** *soft*

…For instance, some started sharing
a doctored photograph of Thunberg with
alt-right boogeyman George Soros…

→ **False**

X $Z \in [0, 1]$ Y

[3]
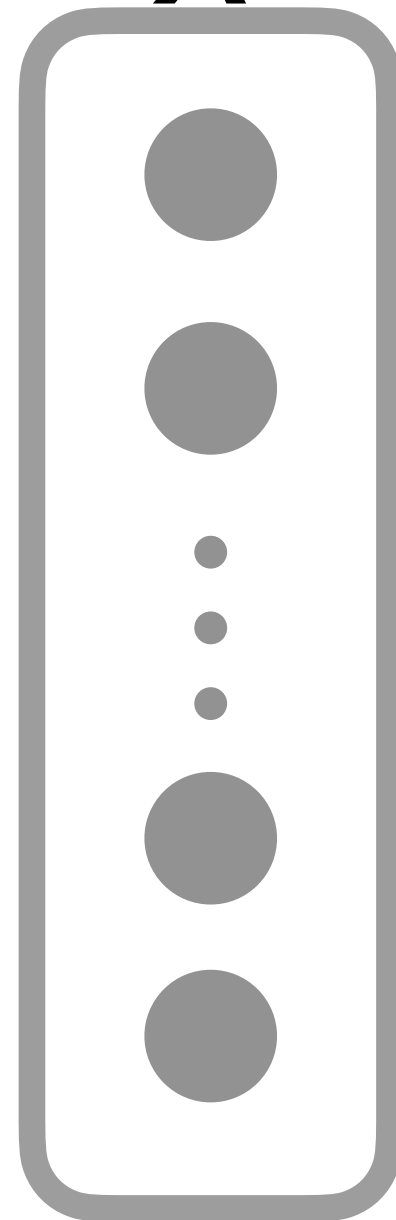
[3] Jain, et al. Learning to faithfully rationalize by construction. *ACL 2020*

Methods: rationalized models - *hard*

...For instance, some started sharing

doctored photograph

of Thunberg with

alt-right boogeyman George Soros...
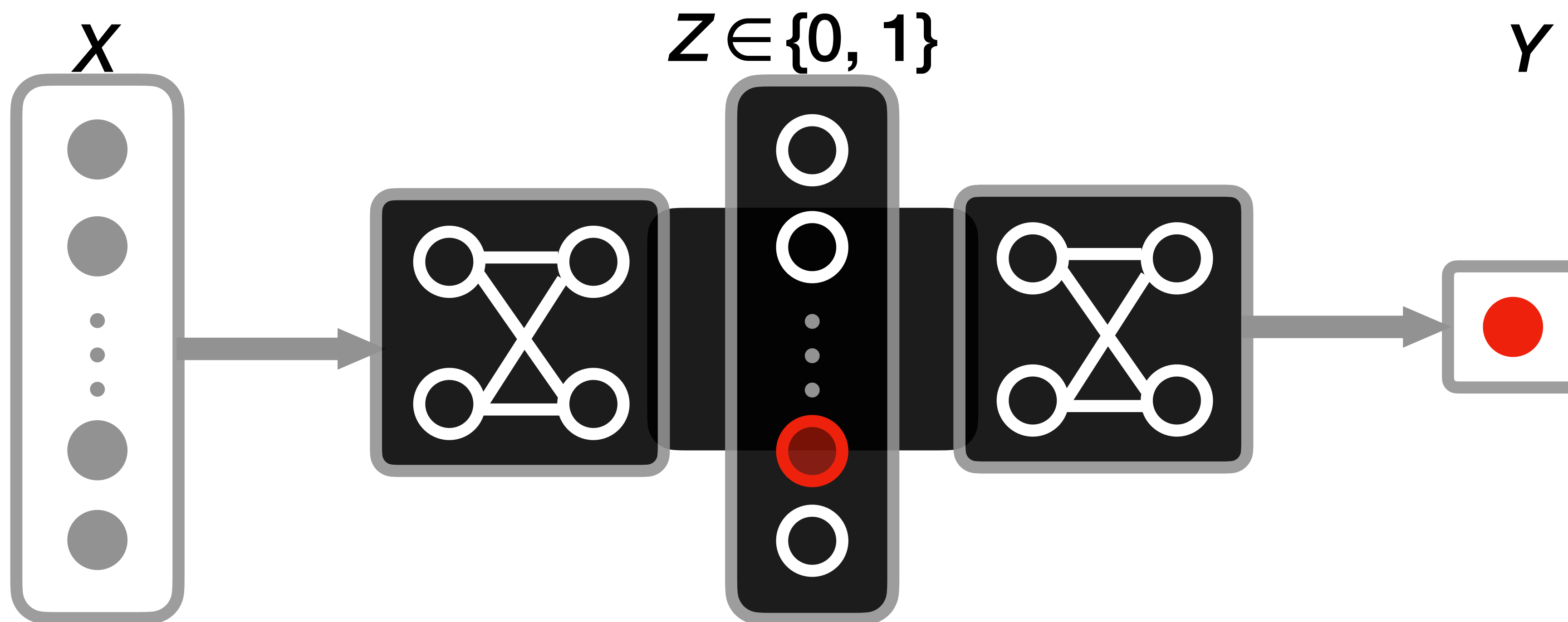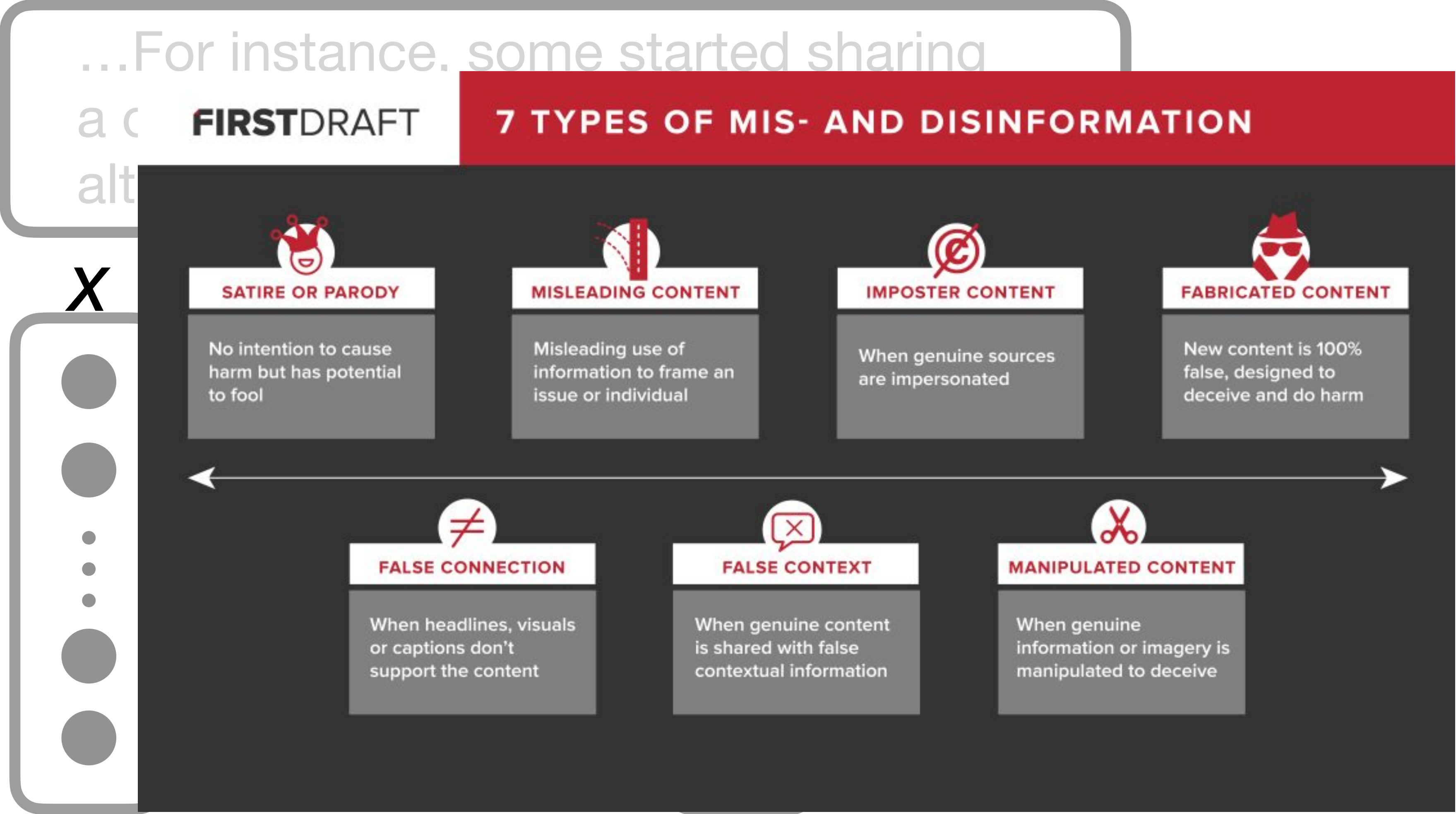
False

X

Z ∈ {0, 1}

Y

[4]

[4] Lei, et al. Rationalizing neural predictions. *EMNLP 2016*

Methods: **rationalized models** - *error*

…For instance, some started sharing
a doctored photograph of Thunberg with
alt-right boogeyman George Soros…

False

X $Z \in \{0, 1\}$ Y

Methods: **theorized misinformation types**

...For instance, some started sharing

a

alt

[5]

X

FIRSTDRAFT   **7 TYPES OF MIS- AND DISINFORMATION**

SATIRE OR PARODY
No intention to cause harm but has potential to fool

MISLEADING CONTENT
Misleading use of information to frame an issue or individual

IMPOSTER CONTENT
When genuine sources are impersonated

FABRICATED CONTENT
New content is 100% false, designed to deceive and do harm

FALSE CONNECTION
When headlines, visuals or captions don't support the content

FALSE CONTEXT
When genuine content is shared with false contextual information

MANIPULATED CONTENT
When genuine information or imagery is manipulated to deceive

[5] Wardle, Fake news. It's complicated. *First Draft 2017*

Methods: **domain knowledge**

…For instance, some started sharing a **doctored photograph** of Thunberg with alt-right boogeyman **George Soros**… → **False**

Z

D

Methods: **domain knowledge**

…For instance, some started sharing a **doctored photograph** of Thunberg with alt-right boogeyman **George Soros**… → **False**

Z

Domain Knowledge

L(Z, D)

D

Methods: domain knowledge

$L(Z, D)$⬇

doctored photograph ➡ **False**

George Soros ➡

$L(Z, D)$⬆

Z
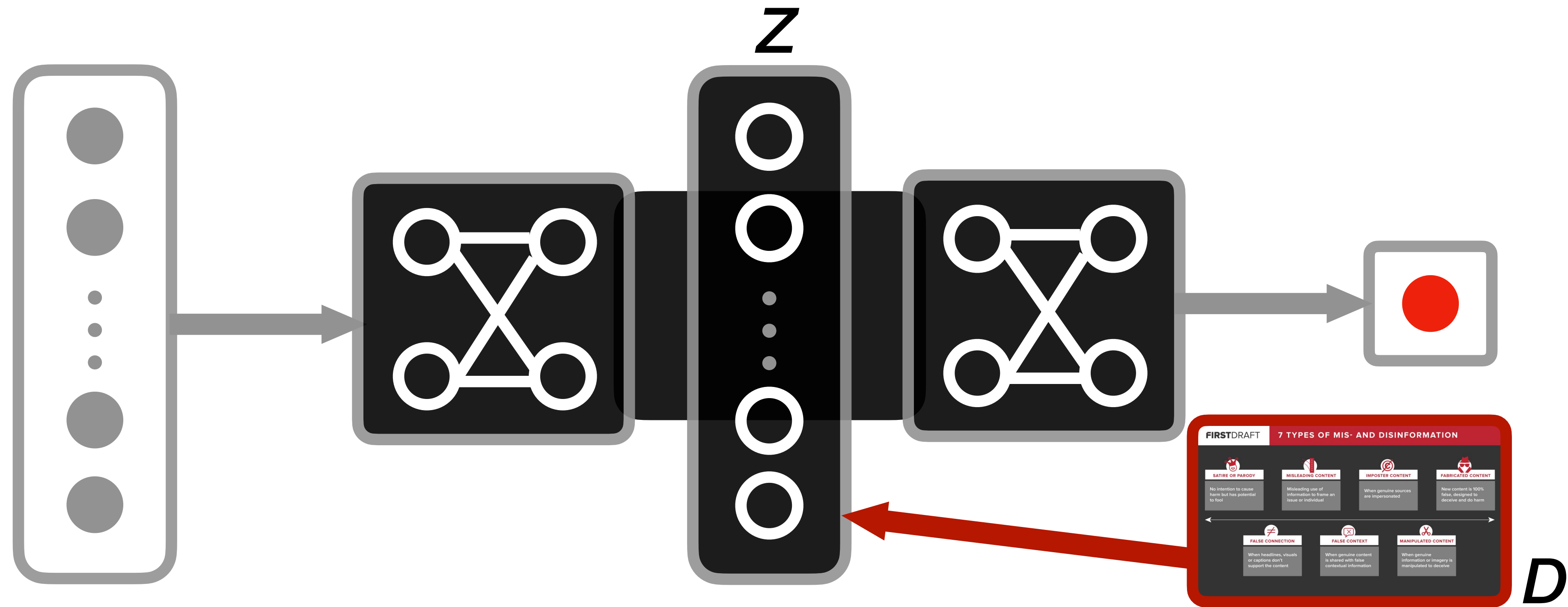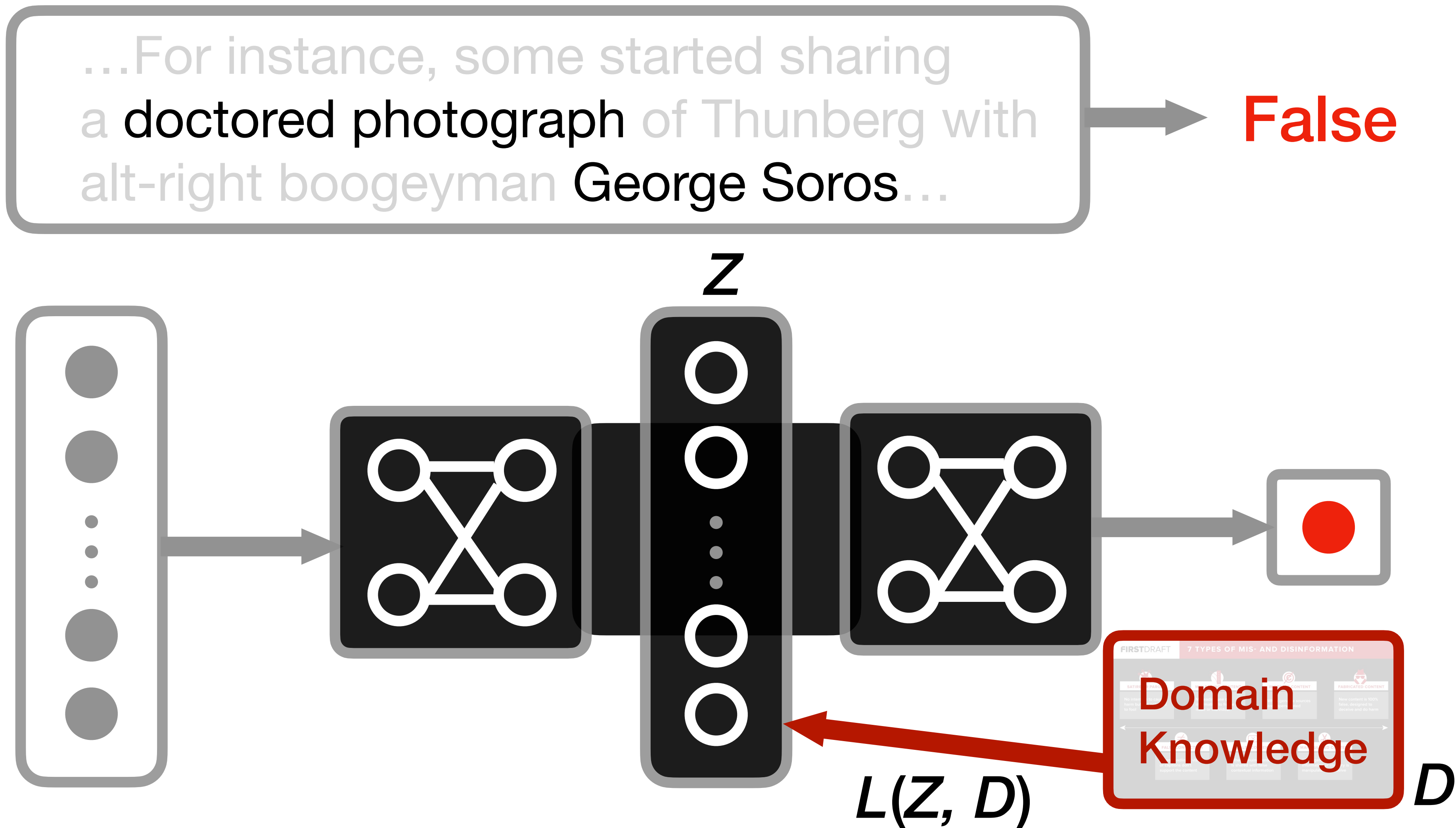
$L(Z, D)$ Domain Knowledge D
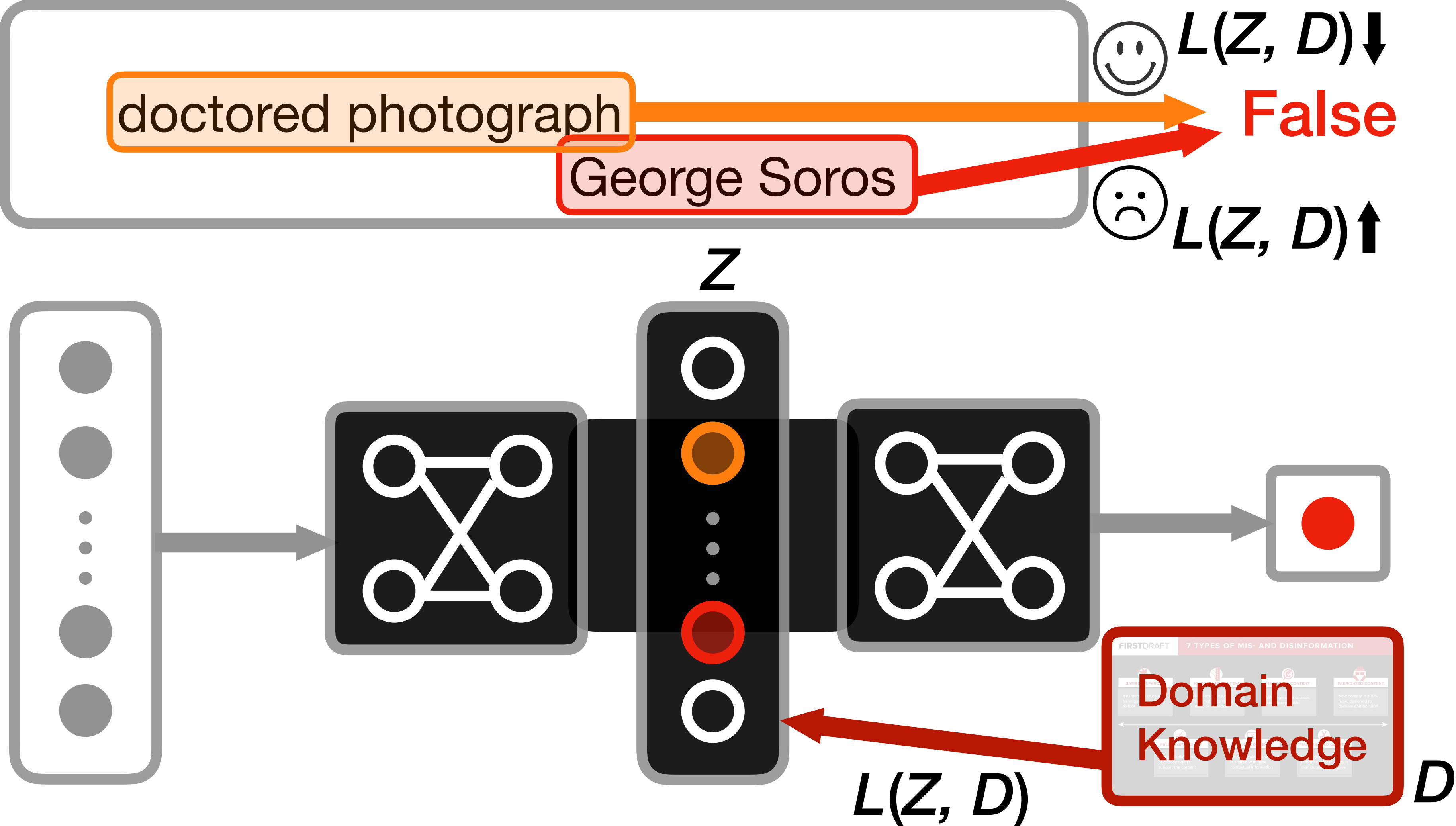
Methods: **experiments**

		Movie reviews (Zaidan et al., 2007)				Personal attacks (Carton et al., 2018)			
		$\Pr(z)$	$\%(z)$	$F_1(y)$	$\%(x)$	$\Pr(z)$	$\%(z)$	$F_1(y)$	$\%(x)$
h_0	**Hard rationalization**	0.37	2.7%	0.72	2.7%	0.17	32.5%	0.73	32.5%
h_1	w/ Domain knowledge	0.38	3.7%	0.72	3.7%	0.22	16.9%	0.73	16.9%
h_2	w/o Rationale regularization	0.31	99.9%	0.92	99.9%	0.19	99.9%	0.82	99.9%
h_3	w/ Adversarial components	0.33	2.5%	0.70	2.5%	0.22	14.9%	0.75	14.9%
s_0	**Soft rationalization**	0.58	3.7%	0.91	100%	0.35	16.9%	**0.82**	100%
s_1	w/ Domain knowledge	**0.62**	3.7%	**0.92**	100%	**0.39**	16.9%	**0.82**	100%
s_2	w/ Half rationales	0.64	1.9%	0.92	100%	0.46	8.4%	0.82	100%
s_3	w/ Double rationales	0.55	7.4%	0.92	100%	0.31	33.8%	0.82	100%

Methods: extracted misinformation types
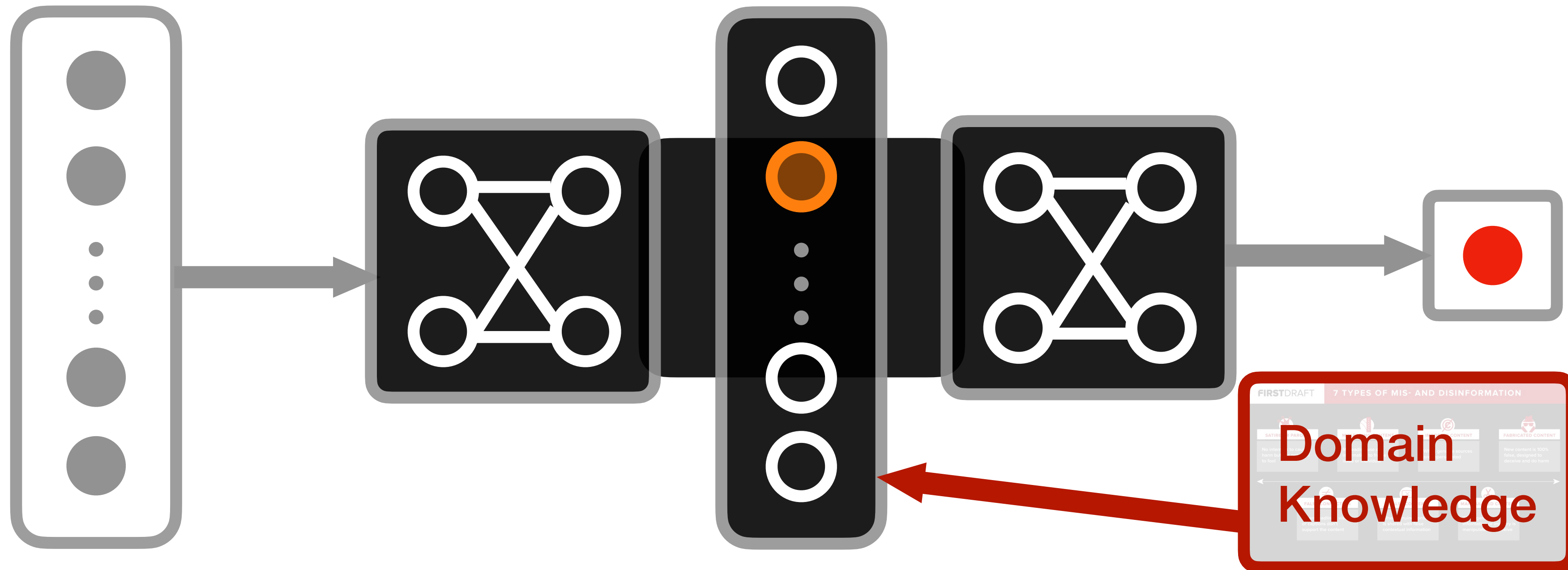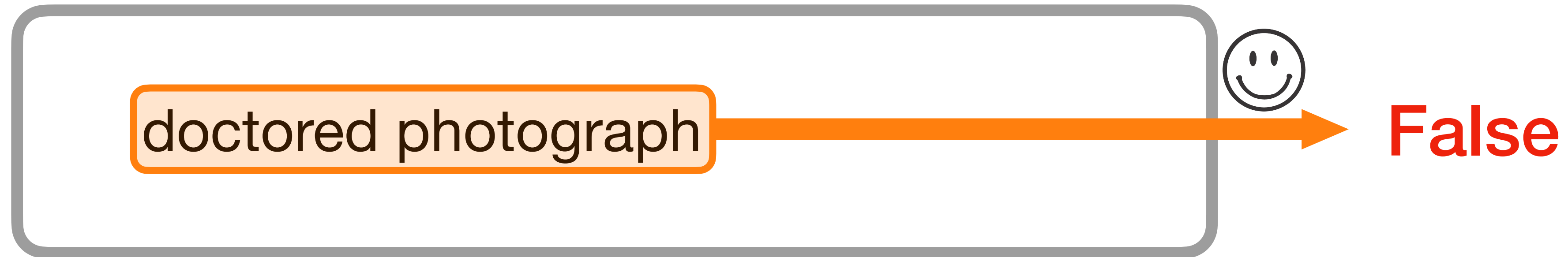
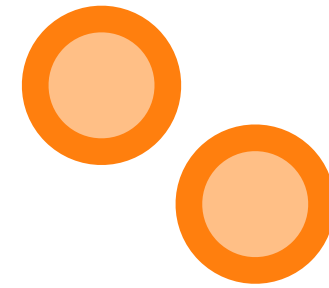# Methods: extracted misinformation types

doctored photograph
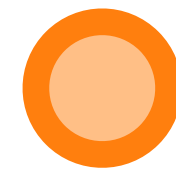
Methods: extracted misinformation types
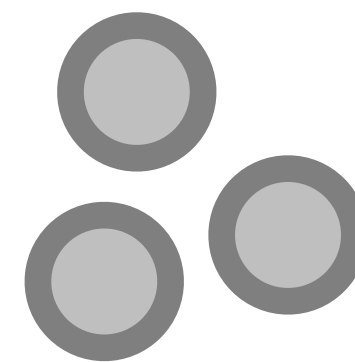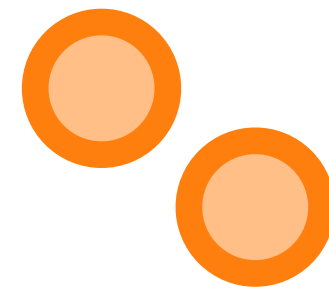
doctored photograph

digitally altered image

Methods: extracted misinformation types
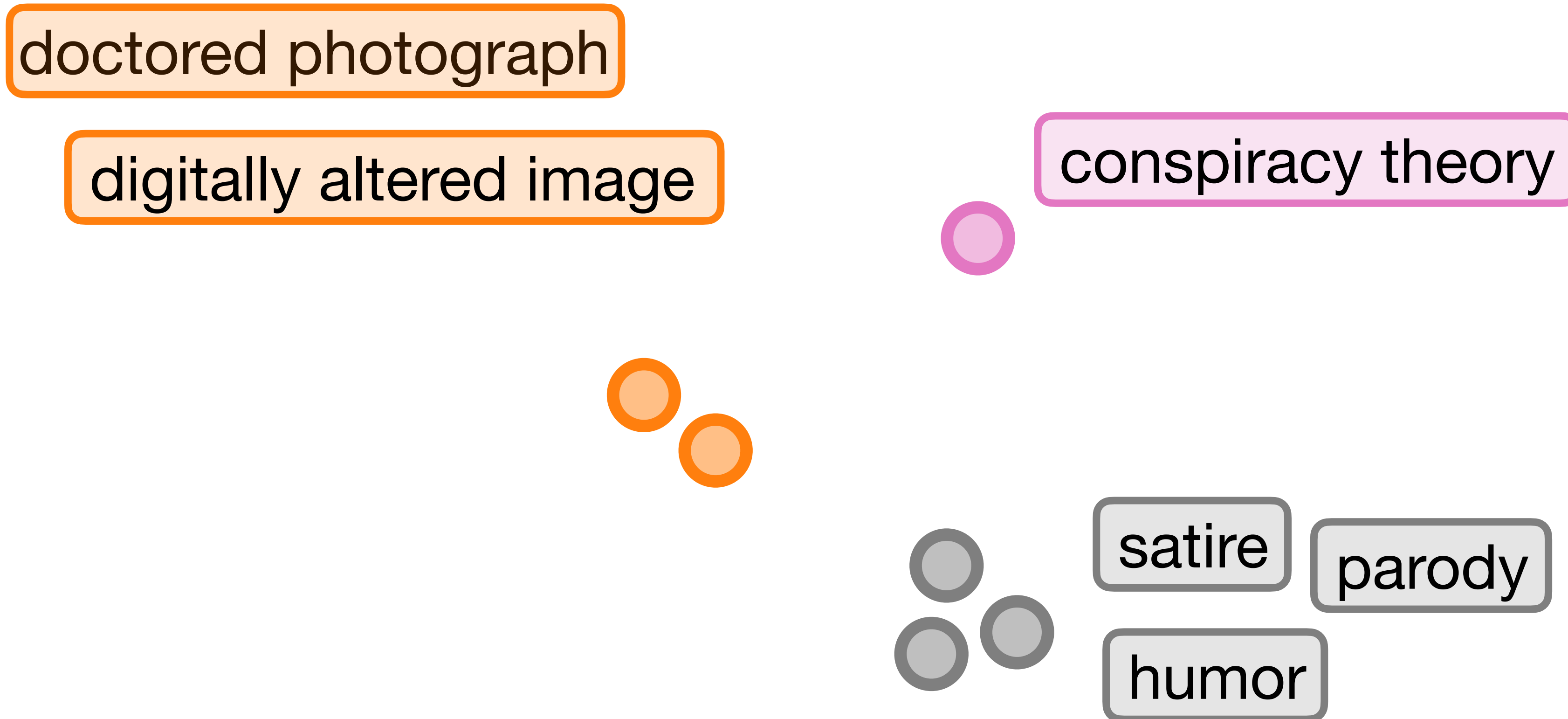
doctored photograph

digitally altered image

satire

parody

humor

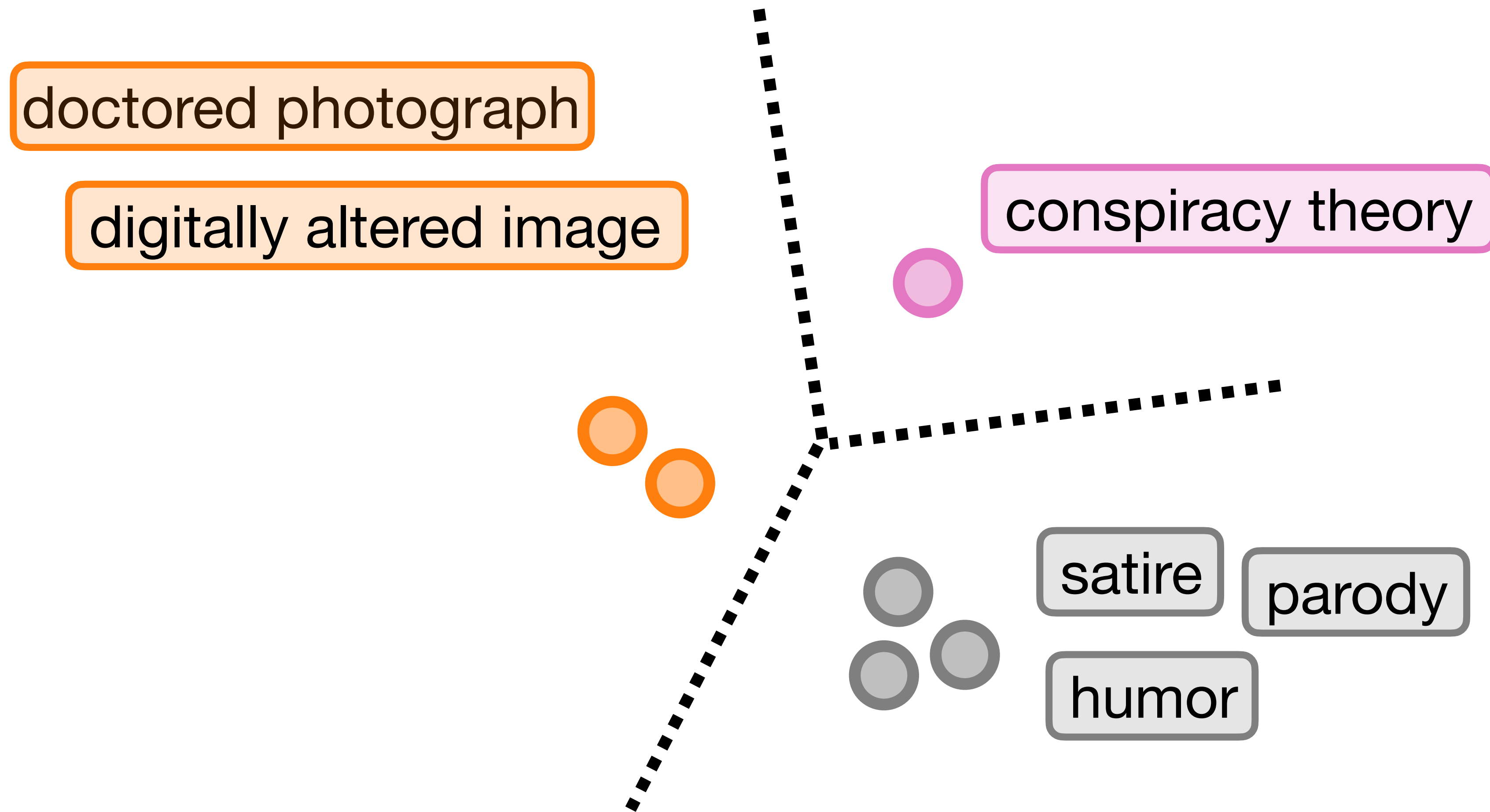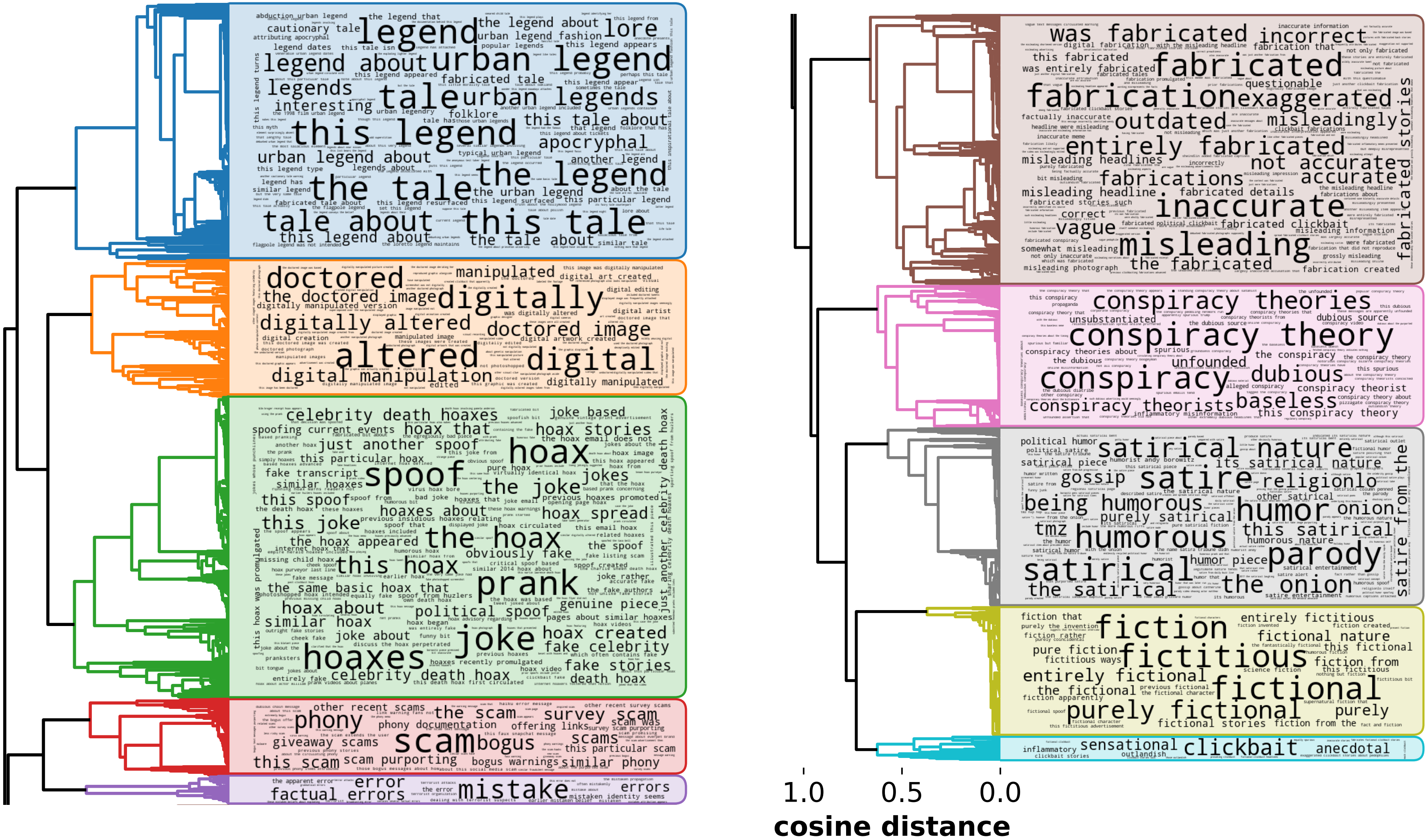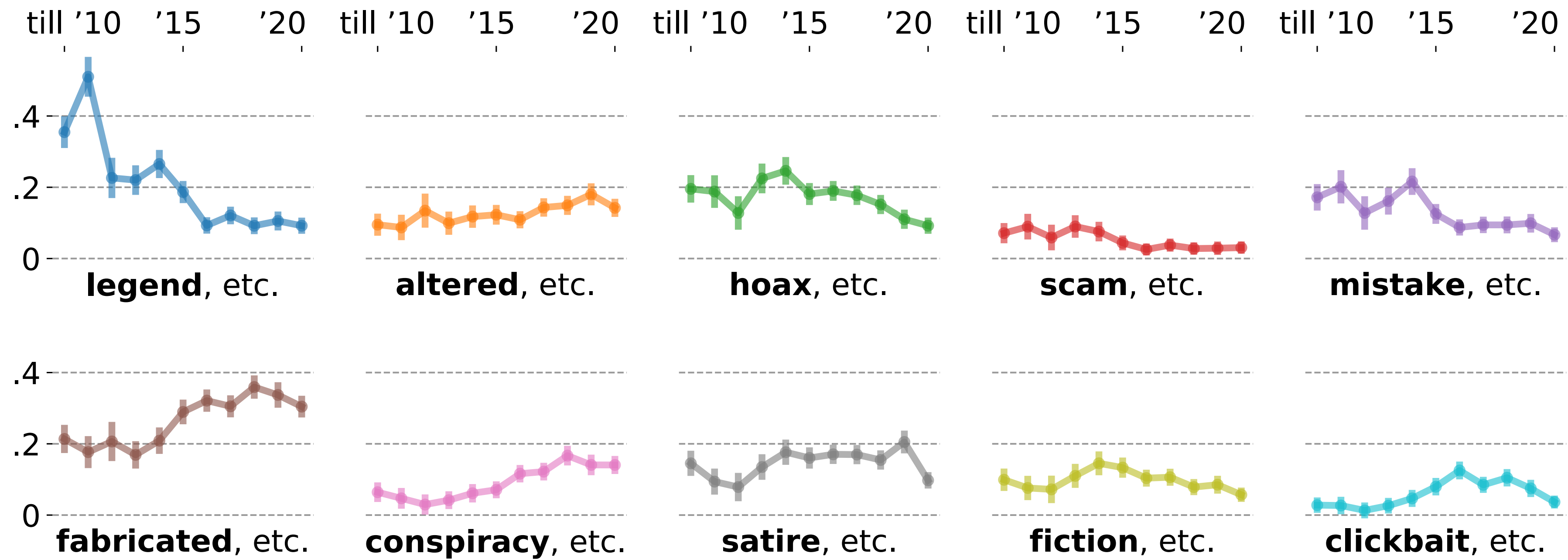# Methods: extracted misinformation types

doctored photograph

digitally altered image

conspiracy theory

satire

parody

humor

Methods: clustering misinformation types

doctored photograph

digitally altered image

conspiracy theory

satire

parody

humor

Results: clustering misinformation types

cosine distance

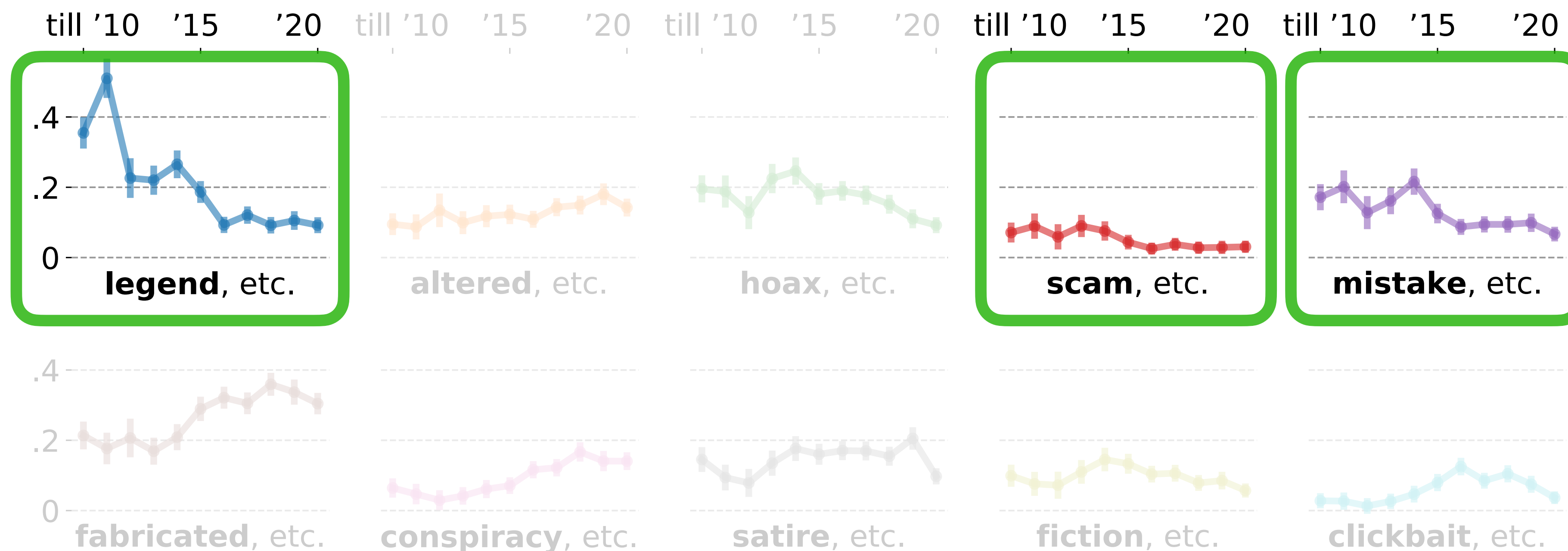# Results: prevalence over time

Results: **prevalence over time**
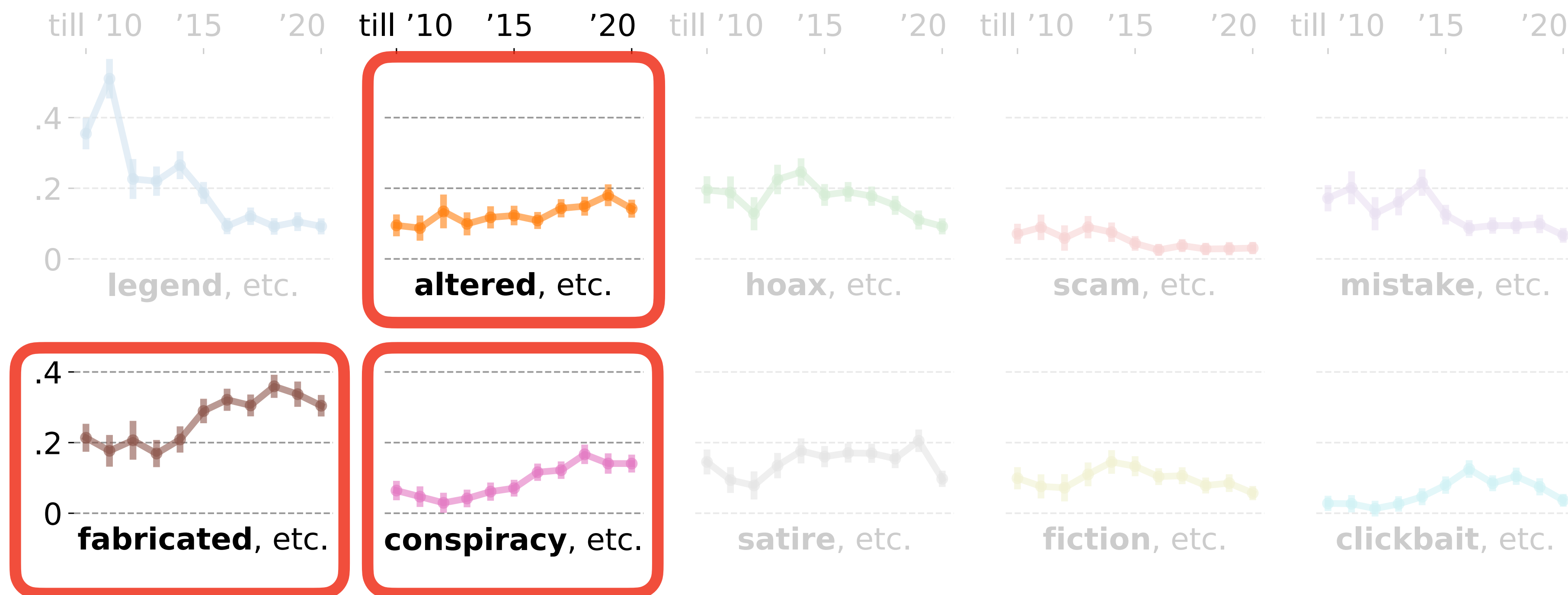
Over the last 10+ years:

- **More digital alteration, more fabrication, and more conspiracies.**
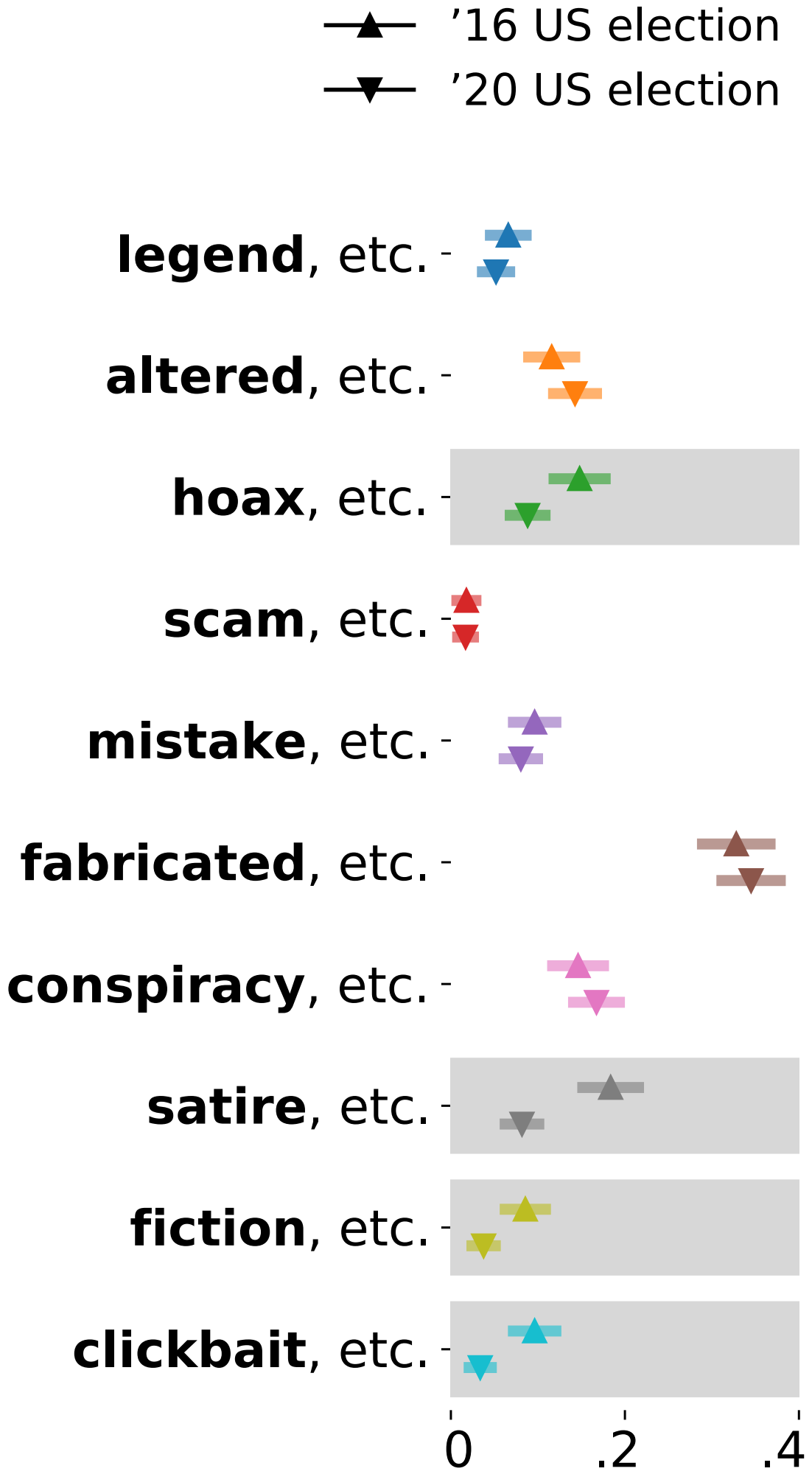
Results: **prevalence over time**

Over the last 10+ years:

- **Less legends / tales, less scams / phony, and less mistakes.**

Results: prevalence between major events

2016 US Election:

2020 US Election:

Results: prevalence between major events

▲— '16 US election
▼— '20 US election

2016 US Election:

- More hoaxes.
- More satires.

2020 US Election:

- Less hoaxes.
- Less satires.

legend, etc.
altered, etc.
hoax, etc.
scam, etc.
mistake, etc.
fabricated, etc.
conspiracy, etc.
satire, etc.
fiction, etc.
clickbait, etc.

0   .2   .4

Results: prevalence between major events

▲— '16 US election
▼— '20 US election

legend, etc.

altered, etc.

hoax, etc.

scam, etc.

mistake, etc.

fabricated, etc.

conspiracy, etc.

satire, etc.

fiction, etc.

clickbait, etc.

0   .2   .4
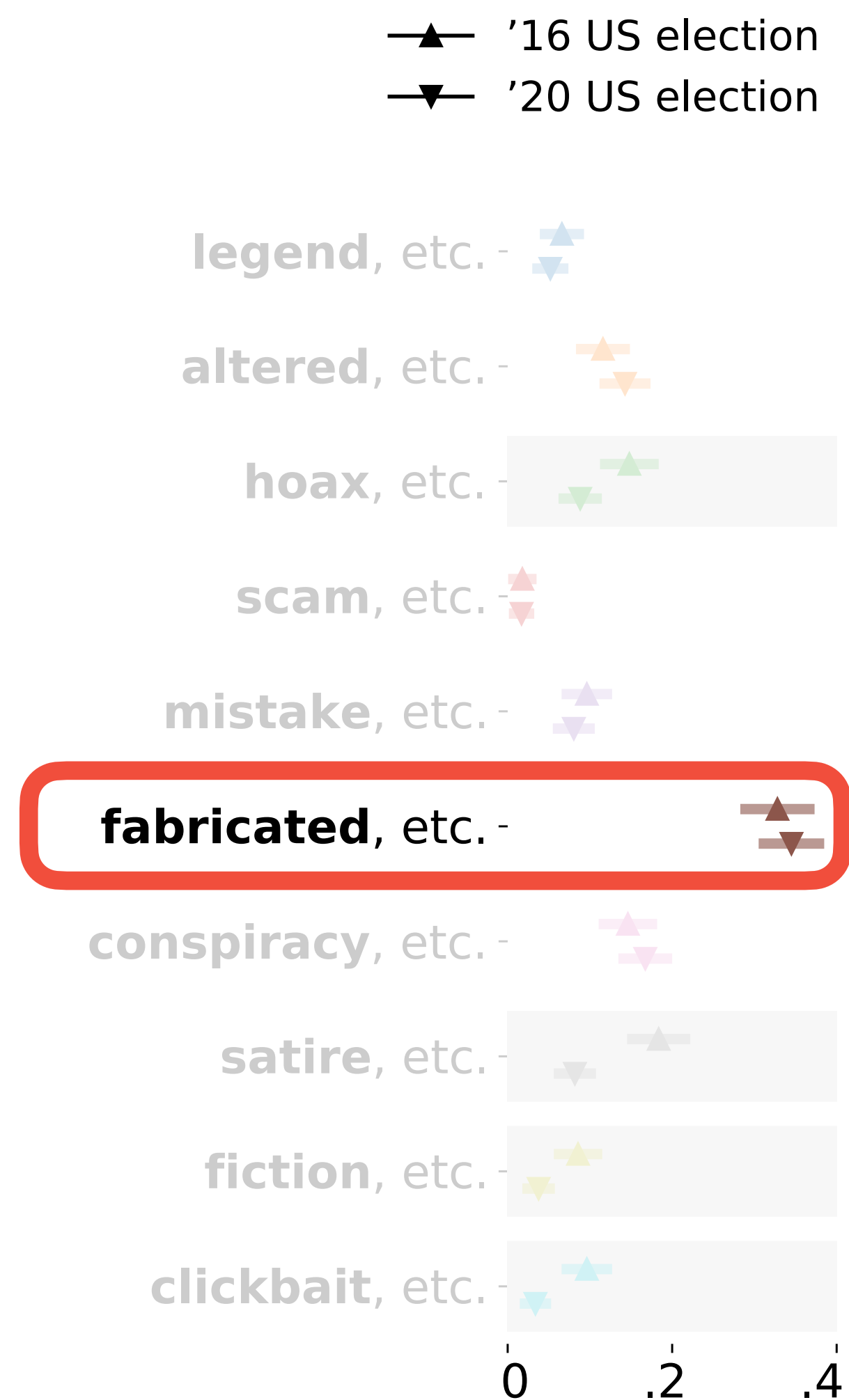
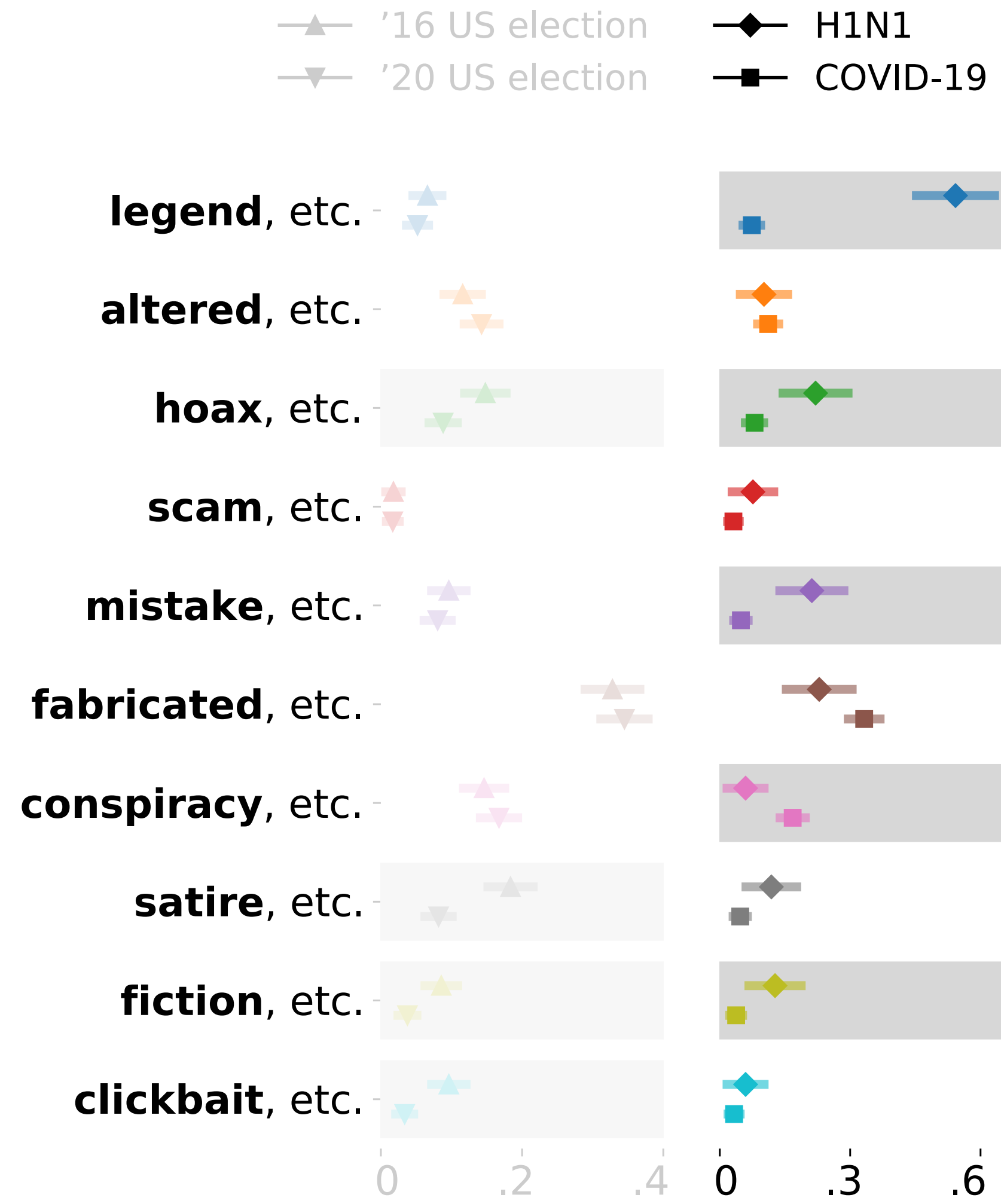## 2016 US Election:

- More hoaxes.
- More satires.

2020 US Election:

- Less hoaxes.
- Less satires.

Both Elections:

- Most prevalent - fabricated content.

Results: prevalence between major events

'16 US election   ◆ H1N1
'20 US election   ■ COVID-19

legend, etc.
altered, etc.
hoax, etc.
scam, etc.
mistake, etc.
fabricated, etc.
conspiracy, etc.
satire, etc.
fiction, etc.
clickbait, etc.

0   .2   .4   0   .1   .3   .6

H1N1 Pandemic:

COVID-19 Pandemic:

Results: prevalence between major events

Legend:
- ▲ '16 US election
- ▼ '20 US election
- ◆ H1N1
- ■ COVID-19

Categories (top to bottom): legend, etc.; altered, etc.; hoax, etc.; scam, etc.; mistake, etc.; fabricated, etc.; conspiracy, etc.; satire, etc.; fiction, etc.; clickbait, etc.
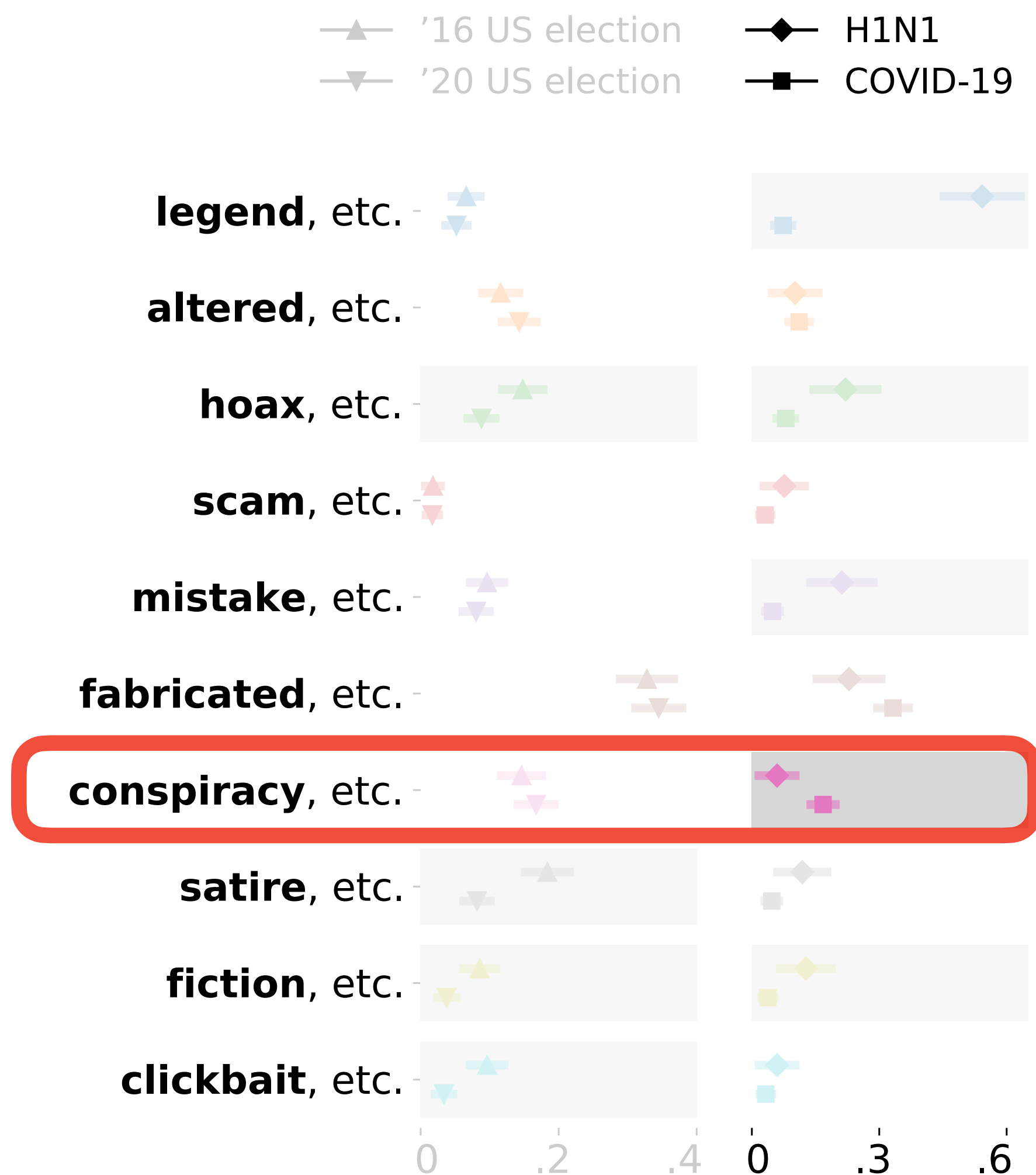
Axis: 0 .2 .4 | 0 .3 .6
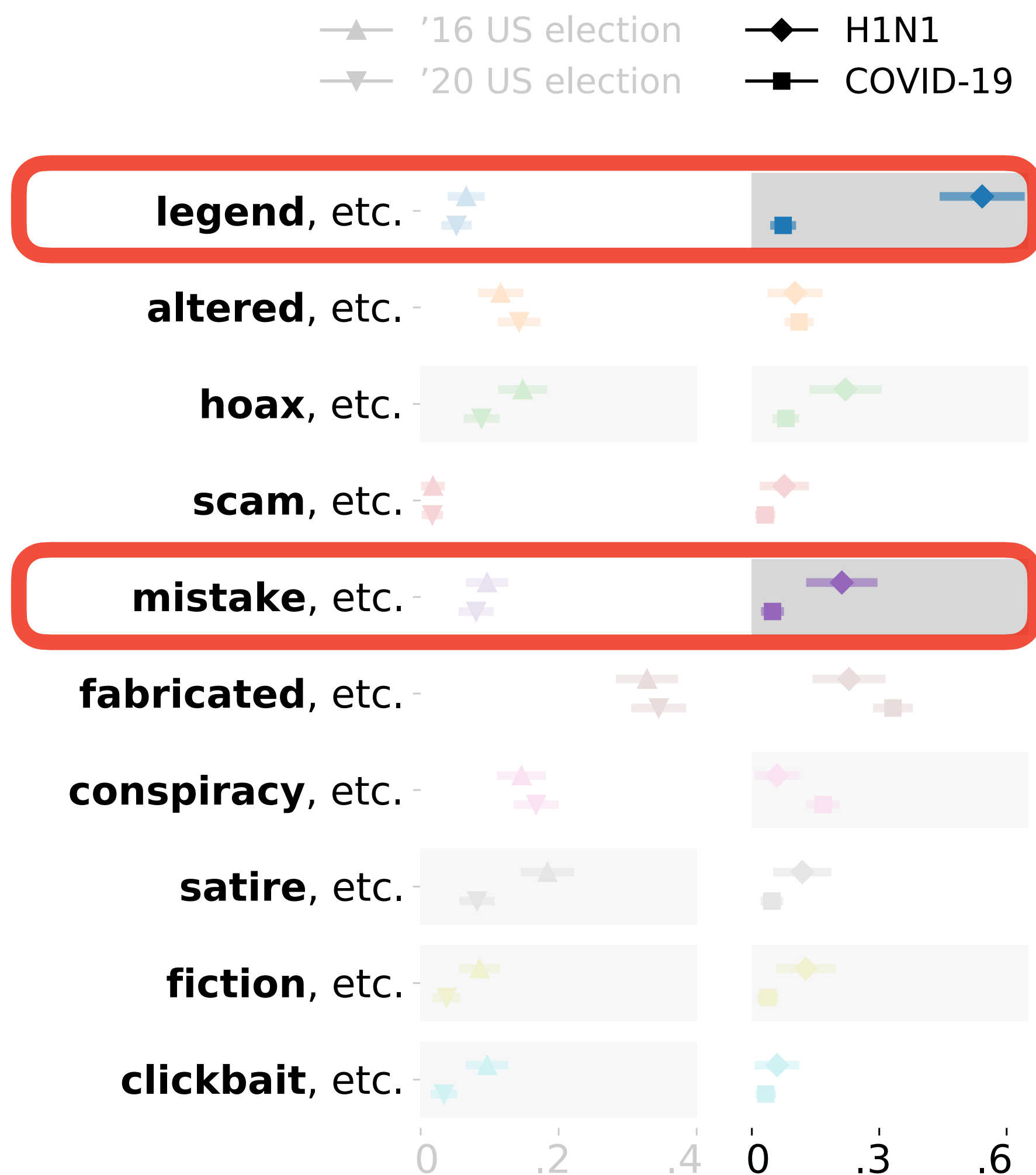
H1N1 Pandemic:

- More legends.
- More mistakes.

COVID-19 Pandemic:

- Less legends.
- Less mistakes.

Results: prevalence between major events

Legend:
- '16 US election
- '20 US election
- H1N1
- COVID-19

Categories: legend, etc. · altered, etc. · hoax, etc. · scam, etc. · mistake, etc. · fabricated, etc. · conspiracy, etc. · satire, etc. · fiction, etc. · clickbait, etc.

H1N1 Pandemic:

- More legends.
- More mistakes.
- Less conspiracy.

COVID-19 Pandemic:

- Less legends.
- Less mistakes.
- More conspiracy.

Limitations: representation

- **Representation: US vs. international.**

Limitations: **proxy validity**

- Representation: US vs. international.
- Proxy: semantics of misinformation types.

Data and code: https://factcheck.shanjiang.me

Thanks!

Shan Jiang and Christo Wilson

{sjiang or cbw}@ccs.neu.edu

Northeastern University