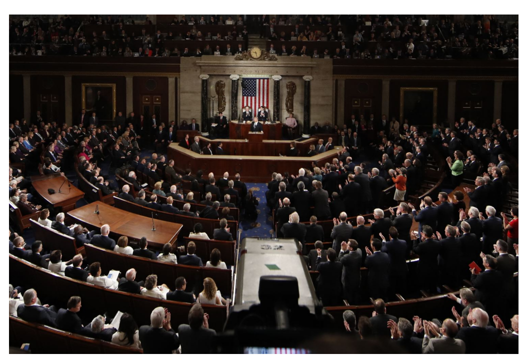
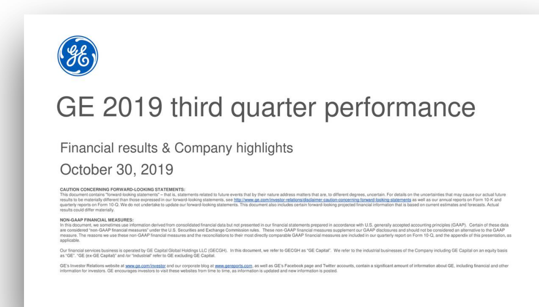
Social Data Science with Text

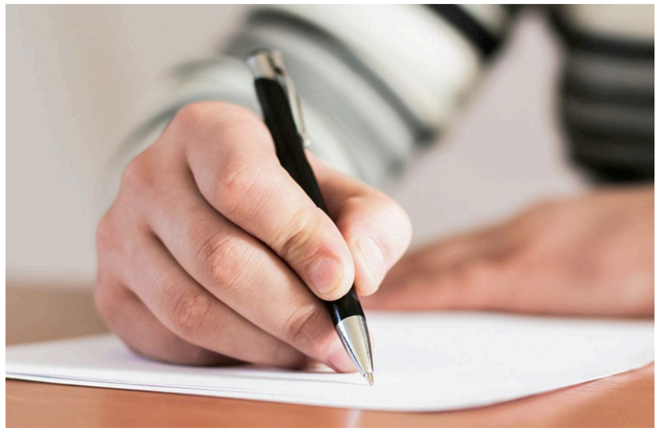
Katherine Keith
PhD Candidate
College of Information and Computer Sciences
University of Massachusetts Amherst

Social Data Science with Text

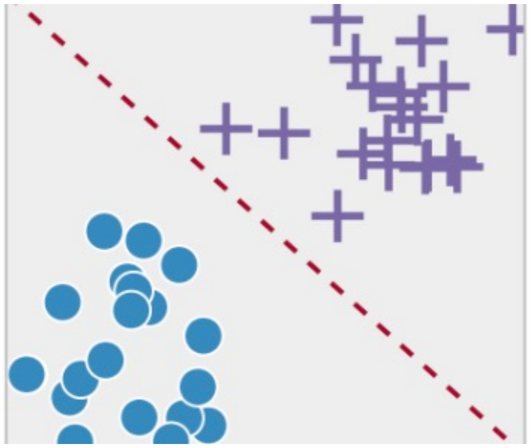




Social Data Science with Text



vs.



Machine Learning

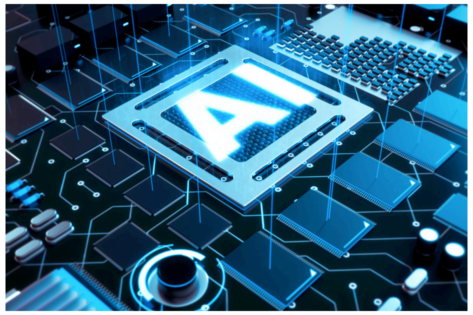
Natural Language Processing

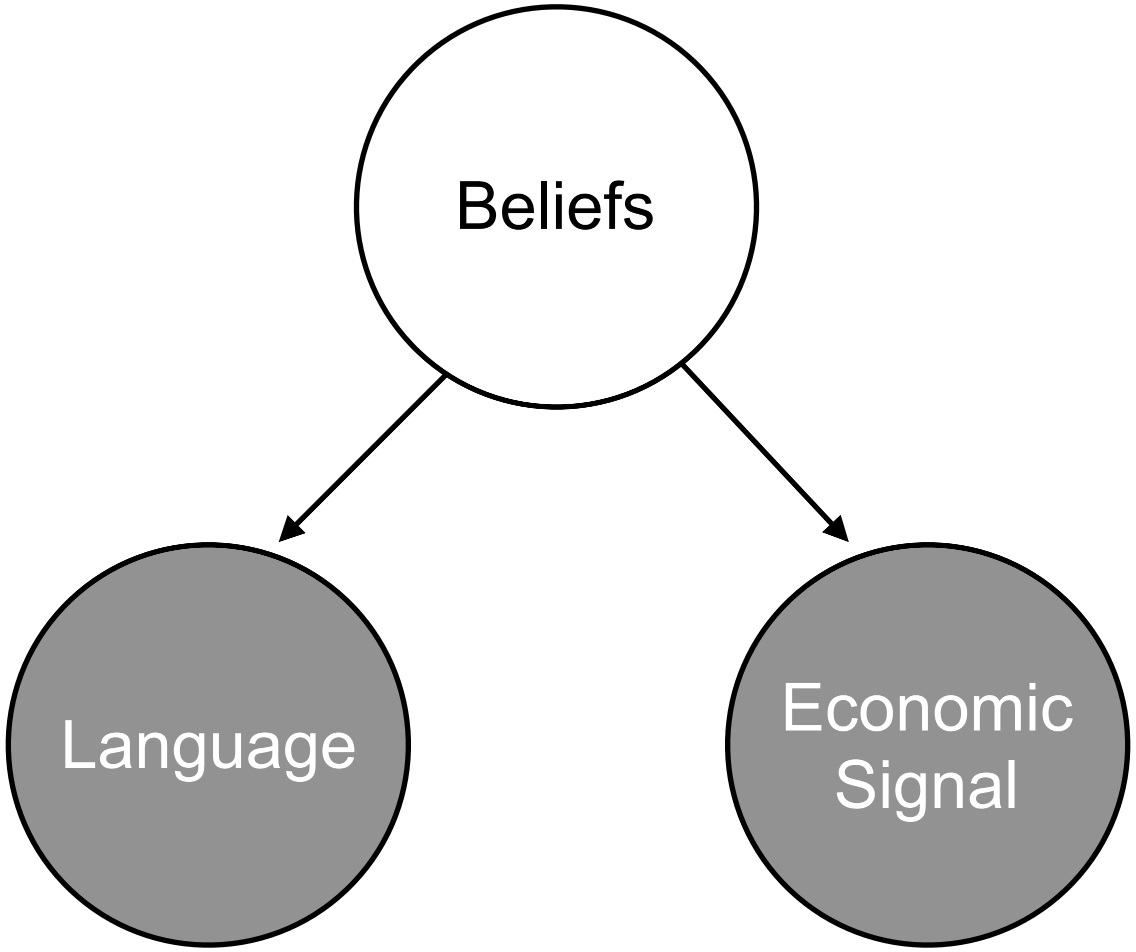
Social **Data Science** with Text

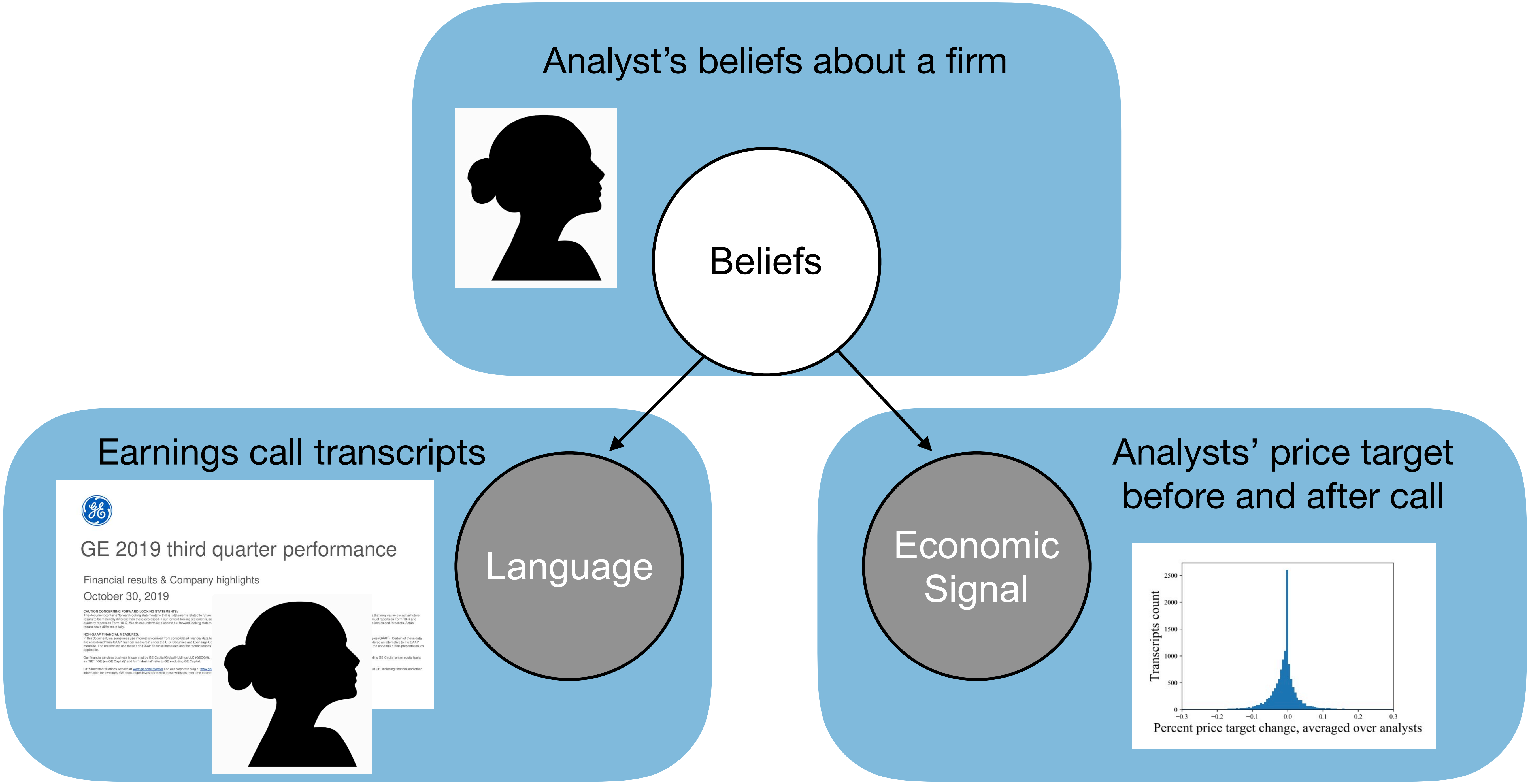
Interdisciplinary collaboration

Bloomberg Data Science Fellow 2019-2021
Bloomberg Research Intern 2018, 2020

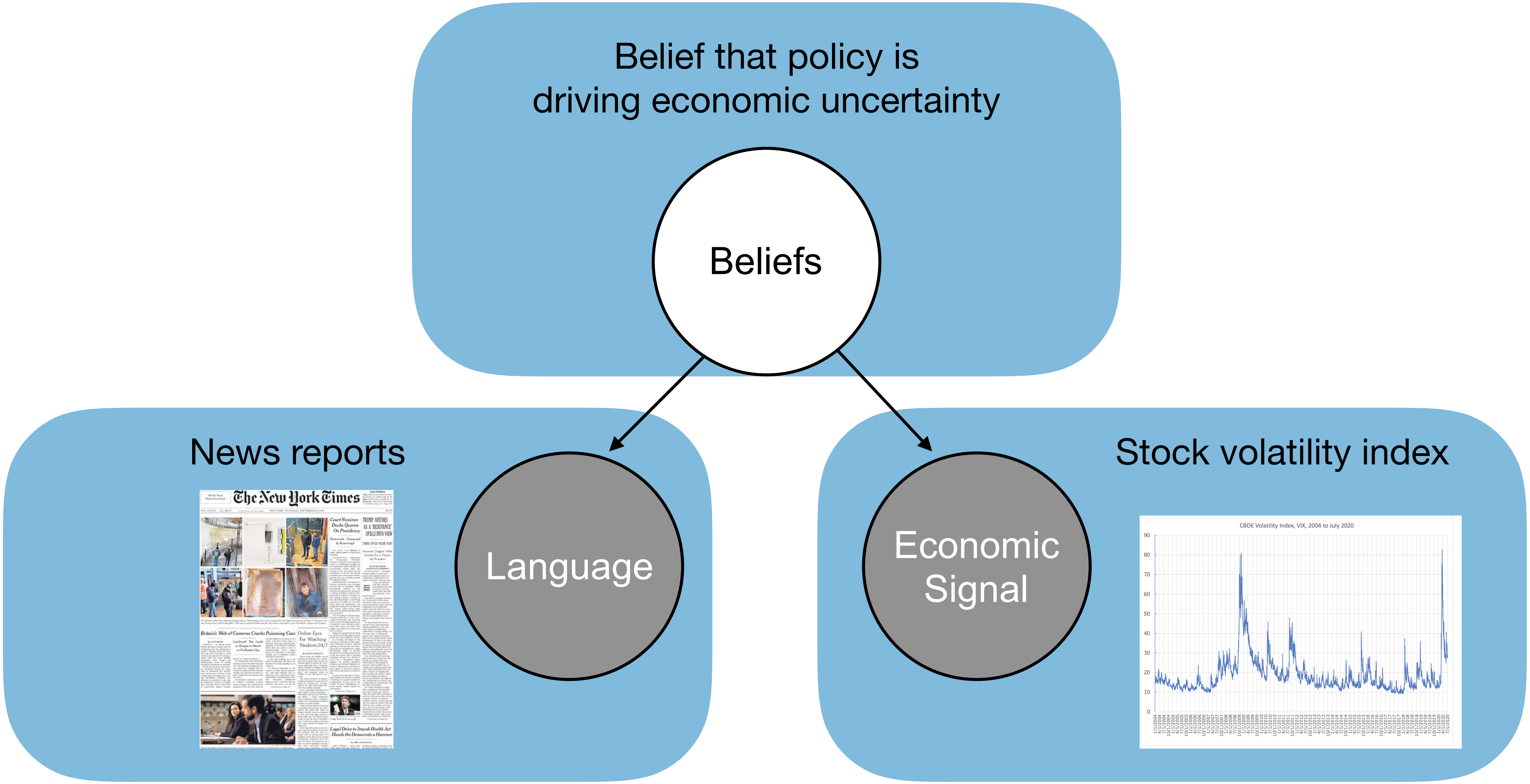
Interdisciplinary collaboration




Beliefs
Language
Economic Signal

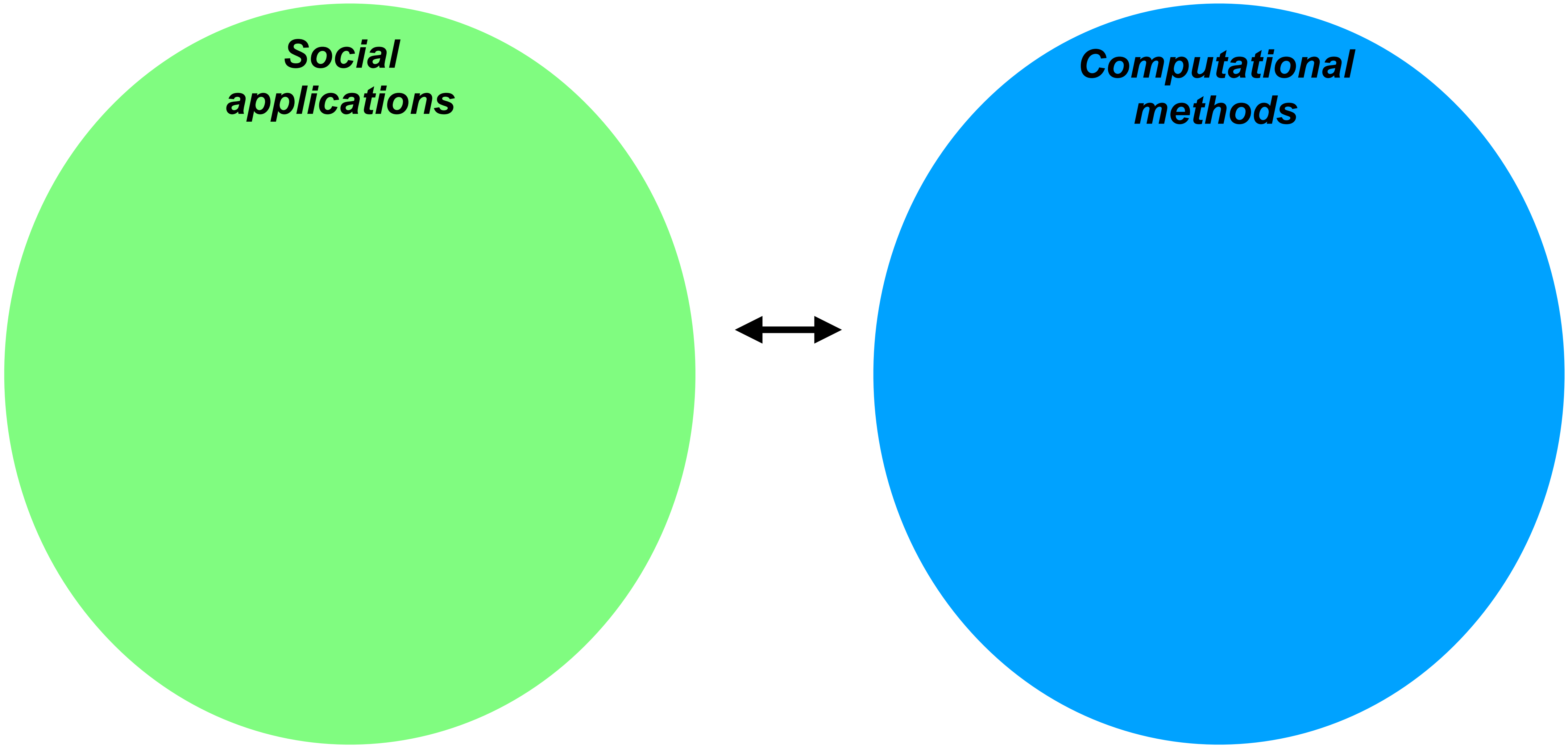


(***Keith*** *and Stent, "Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls." ACL, 2019)*

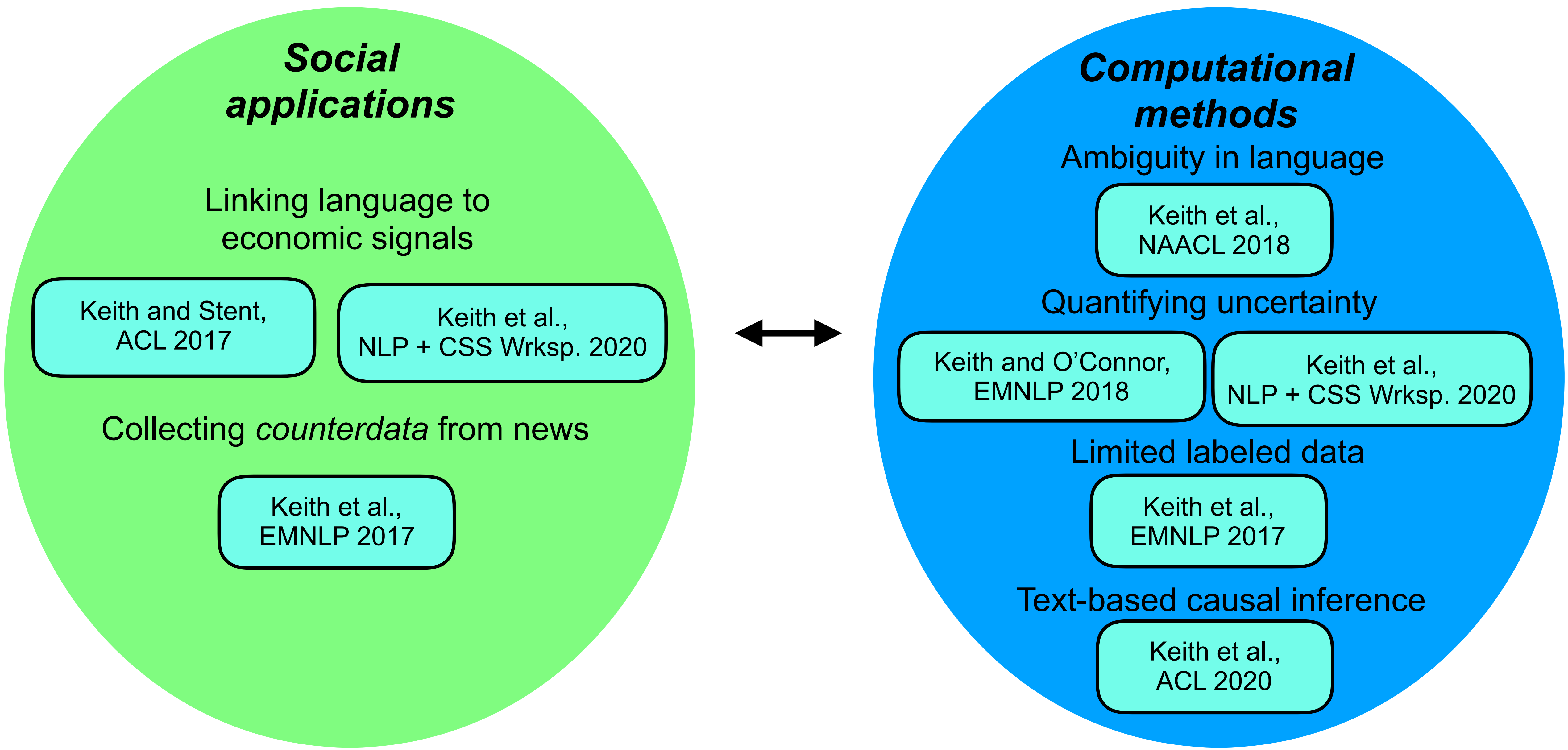


(***Keith*** *et al., "Uncertainty over Uncertainty: Investigating the Assumptions, Annotations, and Text Measurements of Economic Policy Uncertainty." NLP+CSS Workshop, 2020)*

My research philosophy: *Social data science with text* requires a rich symbiosis between *domain applications* and *computational methods*.



My Research Agenda



My Research Agenda

Social applications

Linking language to economic signals

Keith and Stent, ACL 2017

Keith et al., NLP + CSS Wrksp. 2020

Collecting *counterdata* from news

Keith et al., EMNLP 2017

↔

Computational methods

Ambiguity in language

Keith et al., NAACL 2018

Quantifying uncertainty

Keith and O'Connor, EMNLP 2018

Keith et al., NLP + CSS Wrksp. 2020

Limited labeled data

Keith et al., EMNLP 2017

Text-based causal inference

Keith et al., ACL 2020

Questions?

How can **data science** contribute to **social impact?**

How can **data science** contribute to **social impact?**

How can we improve outcomes in the world?

How can **data science** contribute to **social impact?**

How can we ***improve outcomes*** in the world?

values +
measurement

How can we ***improve outcomes*** in the world?

value:
police should
not kill civilians

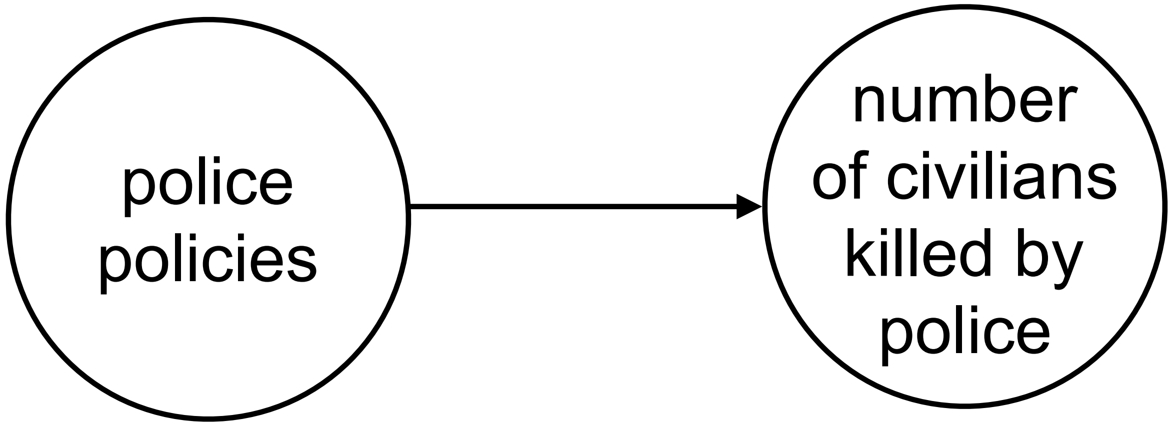
How can we ***improve outcomes*** in the world?

Source: Times Magazine

value:
police should
not kill civilians

How can we ***improve outcomes*** in the world?

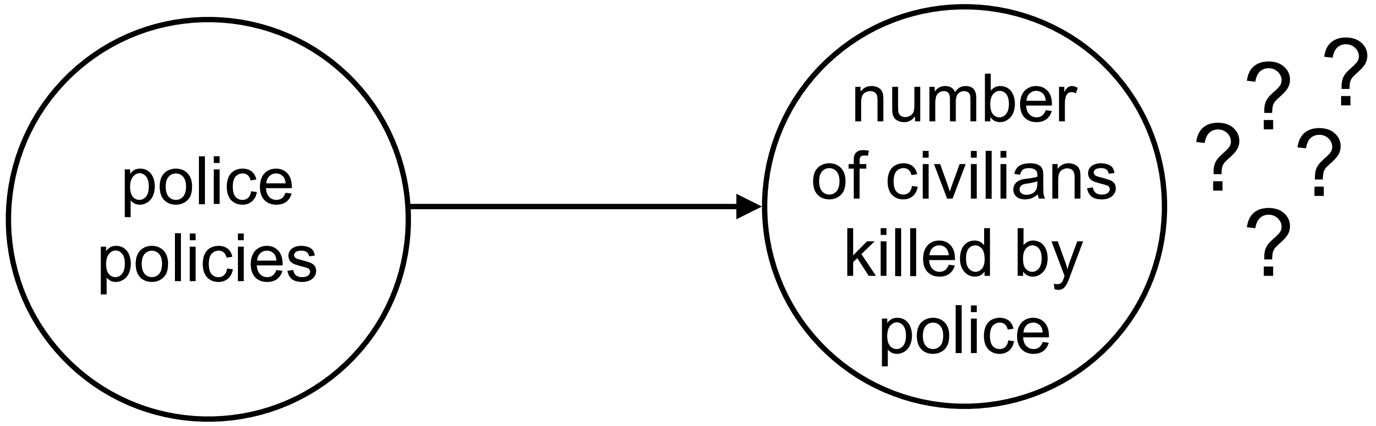
measurement:



value:
police should
not kill civilians

How can we ***improve outcomes*** in the world?

measurement:



U.S. federal government systematically undercounts or fails to count police fatalities

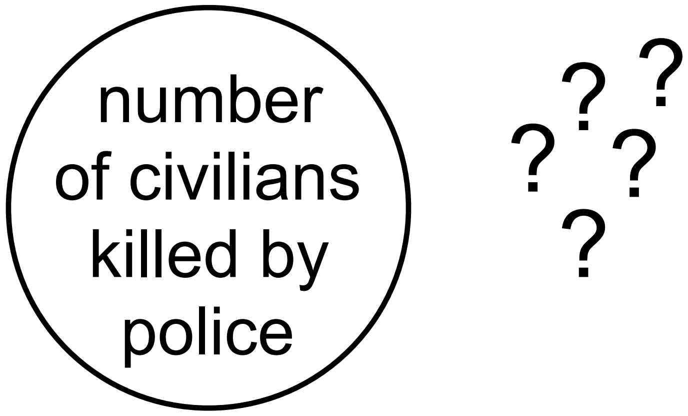
- **2013:** Obama signs *Death in Custody Reporting Act (DCRA)*
  - Requires police departments to report every time a citizen dies in custody
- **2019:** FBI begins *National Use of Force Data Collection*
  - Local law enforcement agencies are not required to participate and the data is not yet public

Counterdata: grassroots collection of missing datasets

(D'Ignazio and Klein, Data Feminism, 2020)

Counterdata: police fatalities from news reports

measurement:



Counterdata: police fatalities from news reports

measurement:

number of civilians killed by police

news reports

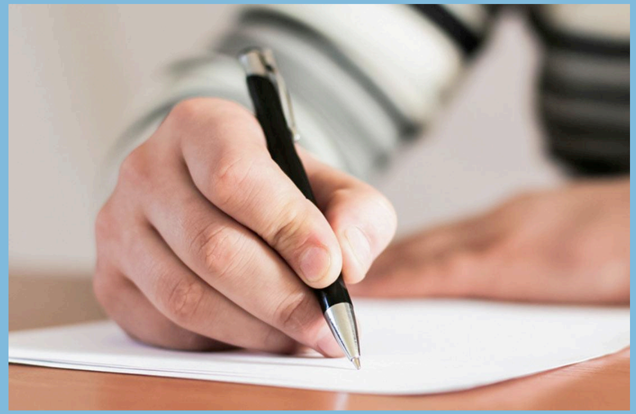
Issue: Cost of human time and emotional strain

Approach 1: Manual

FatalEncounters.org,

KilledByPolice.net,
The Guardian,

Washington Post



Approach 2: Automated

(**Keith** et al.,

EMNLP 2017)

Why is automatically detecting police fatality events hard?

Police killed PERSON.

Police killed PERSON.

Police officers spotted the butt of a handgun in Alton Sterling's front pocket and saw him reach for the weapon before opening fire, according to a Baton Rouge Police Department search warrant filed Monday that offers the first police account of the events leading up to his fatal shooting.

Police killed PERSON.

long-range dependencies

Police officers spotted the butt of a handgun in Alton Sterling's front pocket and saw him reach for the weapon before **opening fire**, according to a Baton Rouge Police Department search warrant filed Monday that offers the first police account of the events leading up to his fatal shooting.

Police killed PERSON.

long-range dependencies

Police officers spotted the butt of a handgun in **Alton Sterling**'s front pocket and saw him reach for the weapon before **opening fire**, according to a Baton Rouge Police Department search warrant filed Monday that offers the first police account of the events leading up to **his** fatal shooting.

coreference

Police killed PERSON.

long-range dependencies

Police officers spotted the butt of a handgun in **Alton Sterling**'s front pocket and saw him reach for the weapon before **opening fire**, according to a Baton Rouge Police Department search warrant filed Monday that offers the first police account of the events leading up to **his fatal shooting**.

coreference

event coreference

Automatically detecting police fatality events

Domain knowledge vs. Machine Learning

Automatically detecting police fatality events

Domain knowledge vs. Machine Learning

Domain Knowledge: Keyword Matching

Input: sentences

*PERSON was **fatally shot** by **police**.*

***Officers** reported PERSON was **killed** in a car accident.*

Keyword matching

officers, police, cops, troopers, deputy, …

kill, killing, shoot, shooting, murder, homicide …

Issue: many false positives *(low precision)*

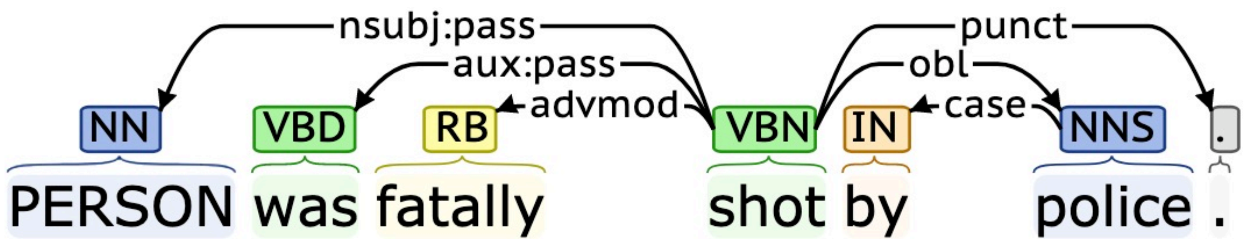
Classification

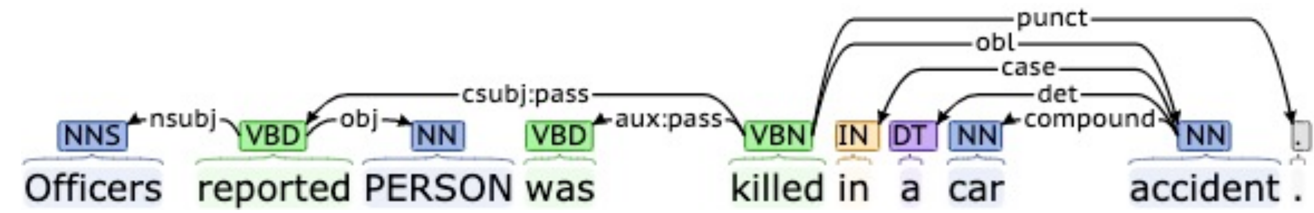
Yes

~~Yes~~

Domain Knowledge: Syntactic Dependency Parsing

Input: automatically infer dependency parse trees over sentences





Rules over dependency paths

PERSON <-nsubj:pass <- [kill, killing, shoot, shooting, murder, homicide …] ->obl -> [officers, police, cops, troopers, deputy, …]

Classification

Yes

No

Issue:
Difficult for a domain expert to list all possible rules
(low recall)

(e.g. Chen and Manning, EMNLP, 2014; Nivre et al. LREC, 2016; ***Keith*** *et. al, NAACL, 2018)*

Automatically detecting police fatality events

Domain knowledge vs. **Machine Learning**

Supervised Machine Learning

1. **Gather** input data

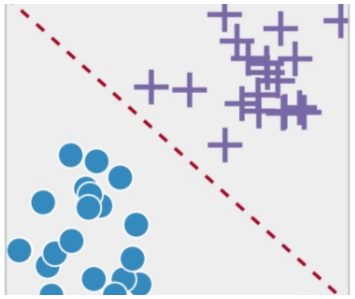
Police killed PERSON.

2. **Label** input da

Our work:

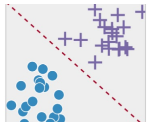
- *logistic regression* with bag of words features
- ***convolutional neural networks*** initialized with pre-trained word embeddings

3. **Train model:** statistical pattern matching between inputs and labels



4. **Inference:** (generalization) apply trained model on unseen inputs

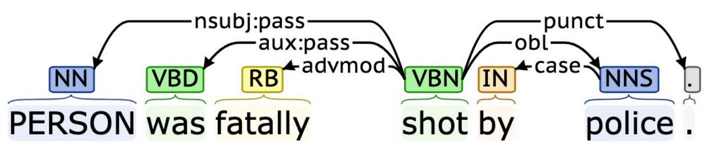
PERSON died in a police homicide.



Yes

Need to **evaluate tradeoffs** for methods of event extraction

kill, killing, shoot, shooting, murder, homicide …



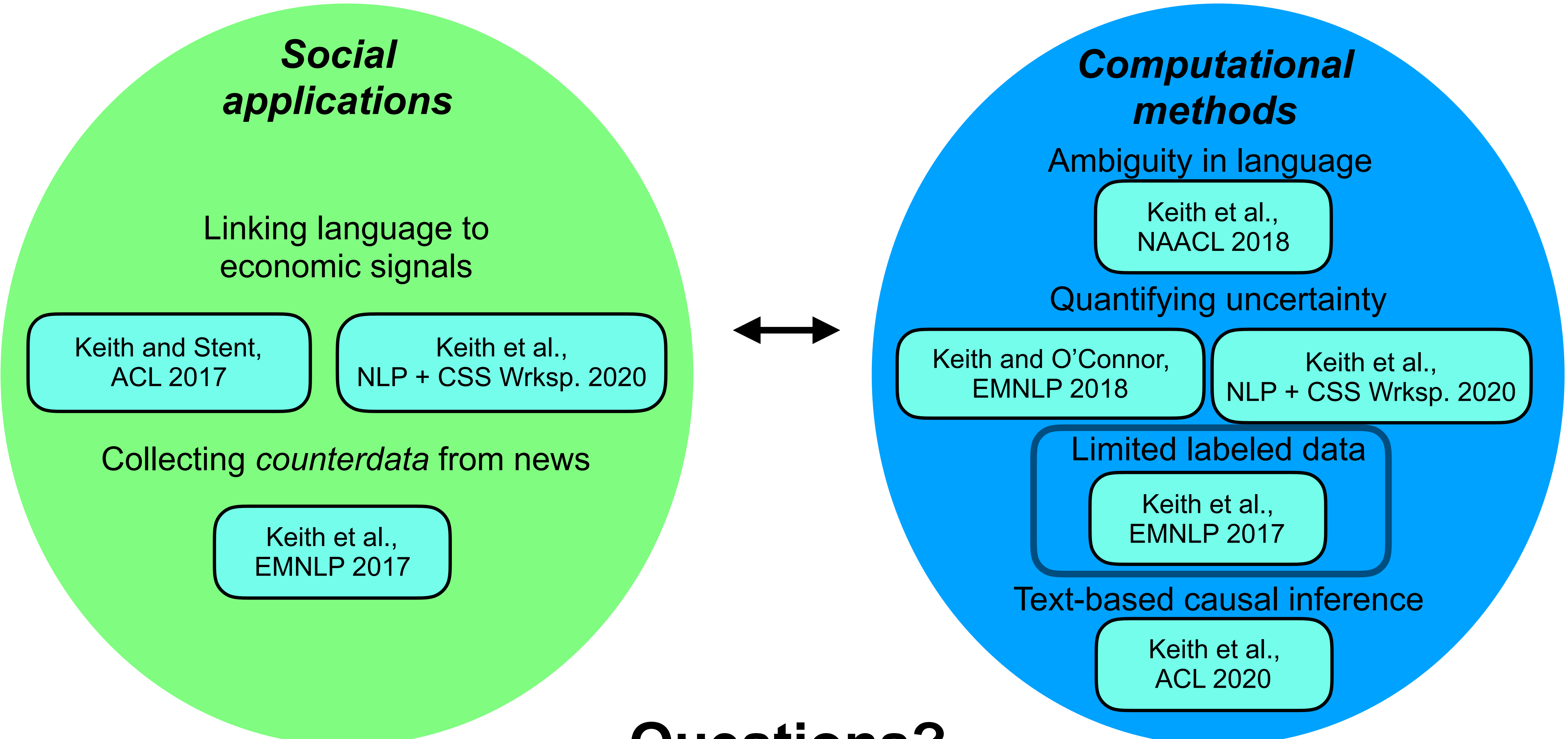
Event extraction methods can be used to collect ***counterdata***

number of civilians killed by police

Without these **measurements**, some questions can be nearly impossible to answer.

How can we ***improve outcomes*** in the world?

My Research Agenda



Questions?

Supervised Machine Learning

1. **Gather** input data

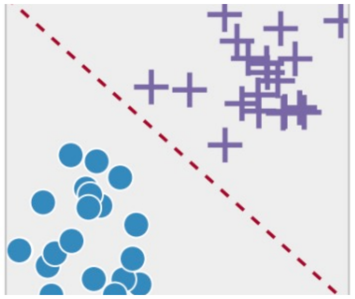
Police killed PERSON.

2. **Label** input data

Yes/No

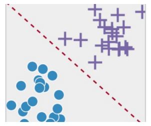
How do we design and
train models with few to no
labeled examples?

3. **Train mo**…rn
matching between inputs and labels



4. **Inference:** (generalization) apply trained model on unseen inputs

PERSON died in a police homicide.



Yes

Distant supervision

(Craven and Kumlien, 1999; Mintz et al., 2009)

John Doe

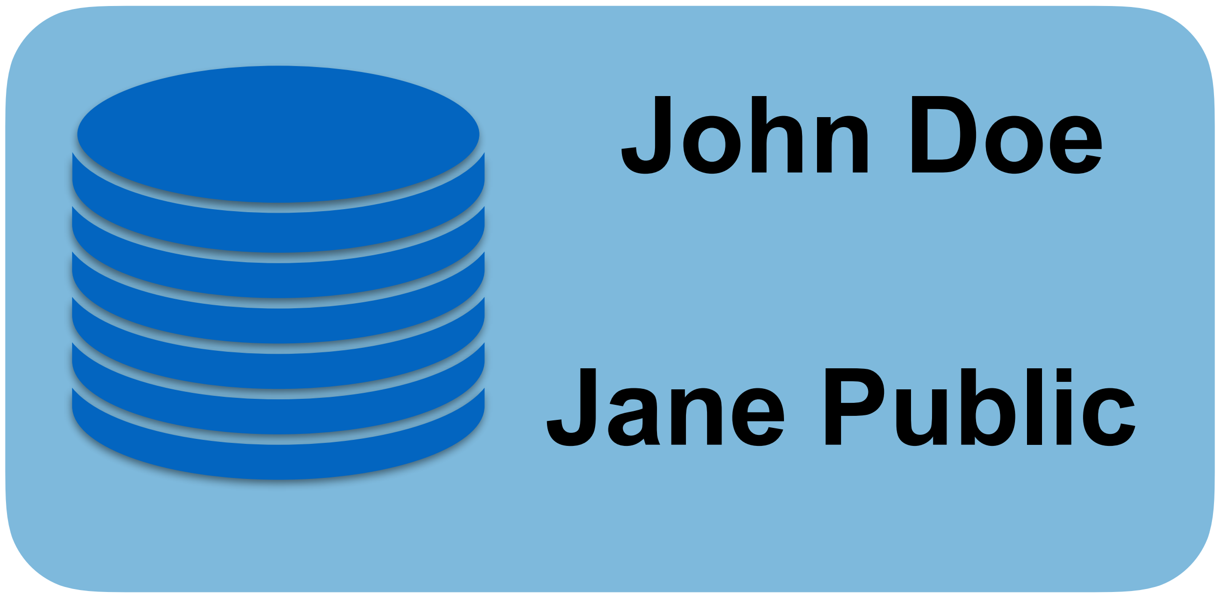
Jane Public

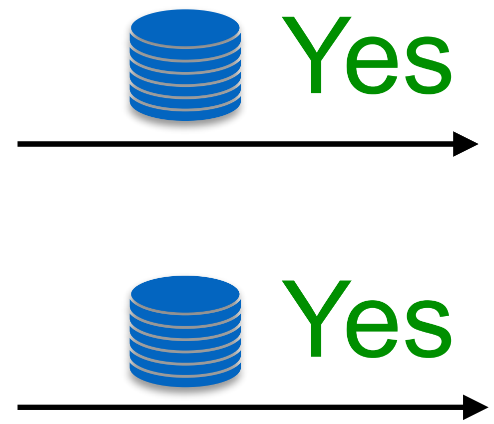
(e.g. Fatal Encounters)

Requirement: knowledge in an external database that is not currently aligned with text

Distant supervision
(Craven and Kumlien, 1999; Mintz et al., 2009)

1. Impute positive labels from an **external database**



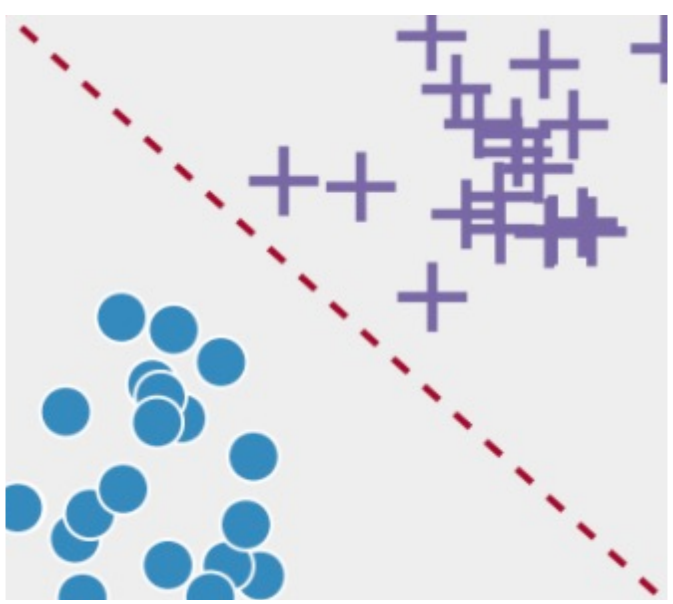


John Doe was killed by police.

Police fatally shot **Jane Public**.

(e.g. Fatal Encounters)

2. Use those labels to train a model



Distant supervision
(Craven and Kumlien, 1999; Mintz et al., 2009)

1. Impute positive labels from an **external database**

John Doe

Jane Public

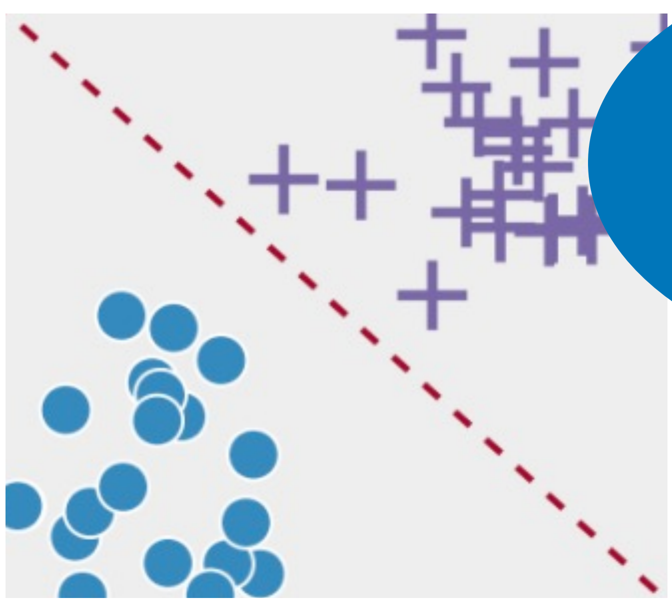
(e.g. Fatal Encounters)

Yes → **John Doe** was killed by police.

Yes → Police fatally shot **Jane Public**.

~~Yes~~ → Police were present during protests over **Jane Public**.

2. Use those labels to train a model



Issue: Many false positives (~30%)

Latent disjunction model

(*Keith et al., EMNLP 2017*)

For each sentence *i*

e_i Entity (person's name)

$y_{e_i} \in \{0, 1\}$ Entity label

$x_{\mathcal{M}(e_i)}$ Set of all sentences with entity

$z_i \in \{0, 1\}$ Label of whether the sentence indicates a police fatality event

Latent disjunction model

(*Keith et al., EMNLP 2017)*

For each sentence *i*

e_i Entity (person's name)

$y_{e_i} \in \{0, 1\}$ Entity label

$x_{\mathcal{M}(e_i)}$ Se...

$z_i \in \{0, 1\}$ Labe... sentence i... police fatality event

Expectati... ...rithm

Builds in assumption that *at least one sentence* must be a positive per entity.

E-Step:

$$q(z_i) := P(z_i \mid x_{\mathcal{M}(e_i)}, y_{e_i})$$

Key takeaway: Transforms sentences labels from binary (0 or 1) to a *probability*

...tep:

$$\max_\theta \sum_i \sum_{z \in \{0,1\}} q(z_i = z) \log P_\theta(z_i = z \mid x_i)$$

Parameters for sentence classifier

Iterate until conve...

Empirical evaluation

Police fatality data

Google News

	Train	Test
Document dates	Jan-Aug 2016	Sept-Dec 2016
Total Docs.	793,010	317,345
Total Entities	49,203	24,550

Data publicly available: http://slanglab.cs.umass.edu/PoliceKillingsExtraction/

Distant supervision vs. Latent disjunction model

Cheaper than standard supervision

Empirical results: Improves entity-level F1 by 18% on the test set

Reduces distantly-labeled *false positives*

Error Analysis

Logan Clarke was **shot** by a **campus police officer** after waving kitchen knives at fellow students outside the cafeteria at Hug High School in Reno, Nevada, on December 7.

Model prediction: Yes

True value: No

Error Analysis

Model has not learned to distinguish *killed* vs. *shot*

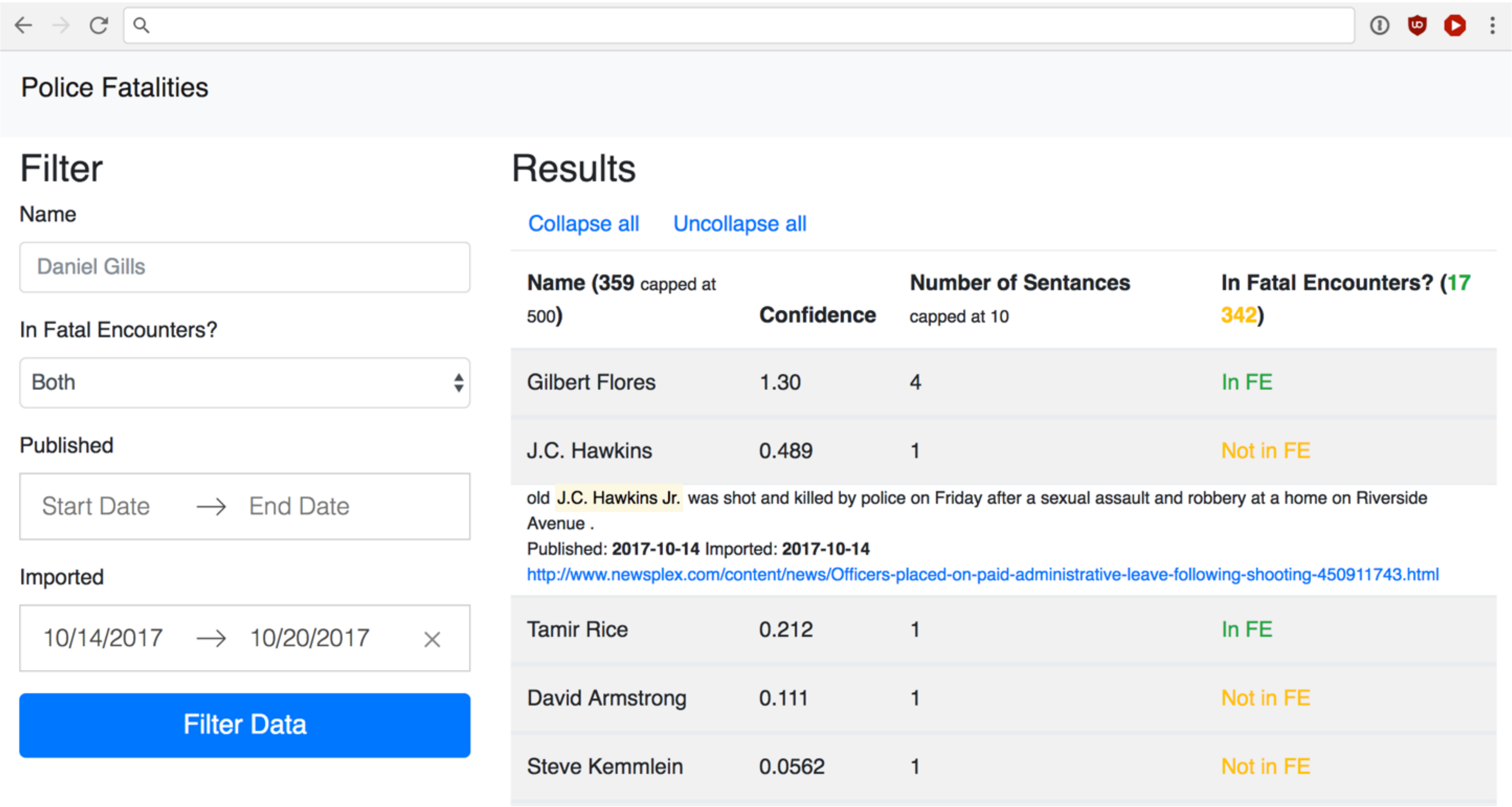
Logan Clarke was **shot** by a **campus police officer** after waving kitchen knives at fellow students outside the cafeteria at Hug High School in Reno, Nevada, on December 7.

Only **official police** are in the database so campus police or security guards count as errors

Model prediction: Yes

True value: No

User Interface Prototype

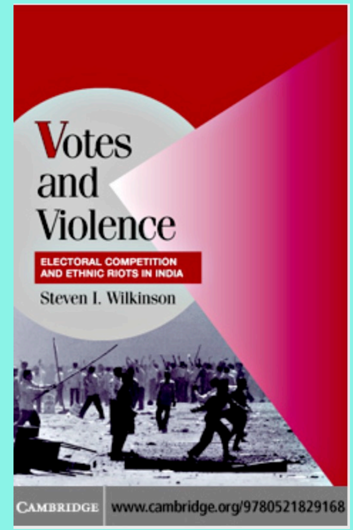


- Ongoing work: Fatal Encounters used our monitoring system for weekly updates
- Dozens of cases and updates found

My ongoing and future work on supporting *counterdata* collection

My ongoing and future work on supporting *counterdata* collection

Using news articles to automatically detect police actions during communal violence in India



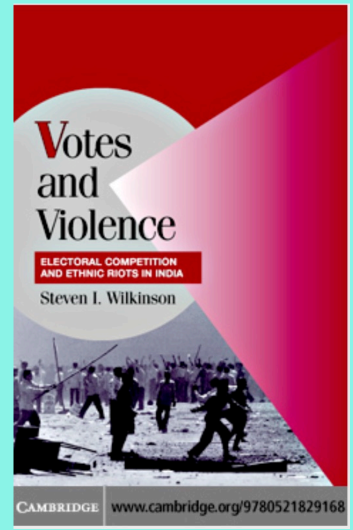
kaggle Open Data Research Grant

Andrew Halterman
Political Science, MIT

Sheikh Sarwar
Computer Science, UMass Amherst

My ongoing and future work on supporting *counterdata* collection

Using news articles to automatically detect police actions during communal violence in India

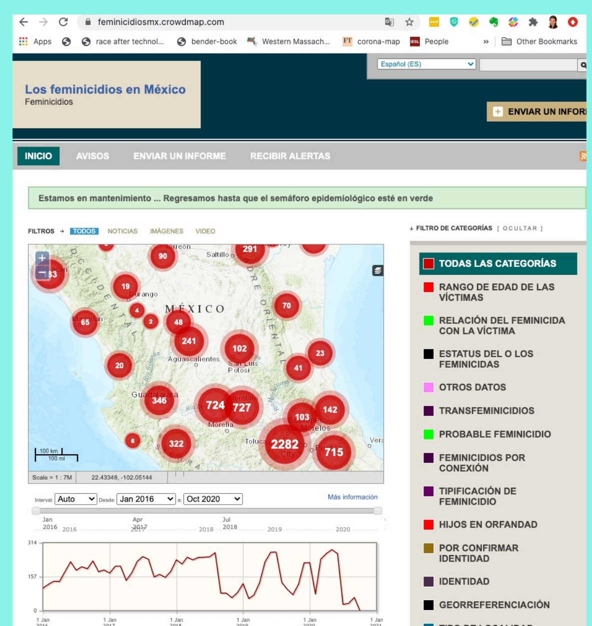


kaggle Open Data Research Grant

Andrew Halterman
Political Science, MIT

Sheikh Sarwar
Computer Science, UMass Amherst

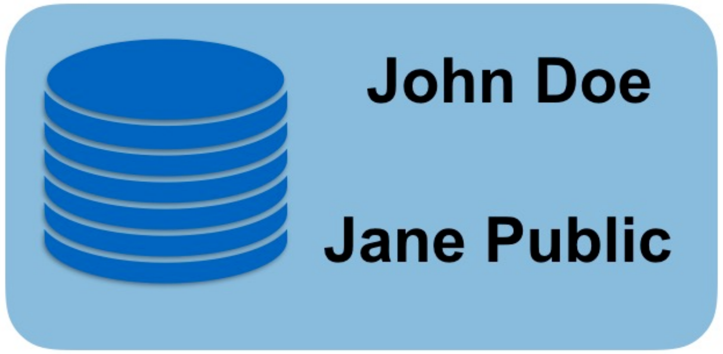
Building a broad set of NLP tools to augment human *counterdata* collection



e.g. Maria Salguero manually maps ***femicides*** in Mexico

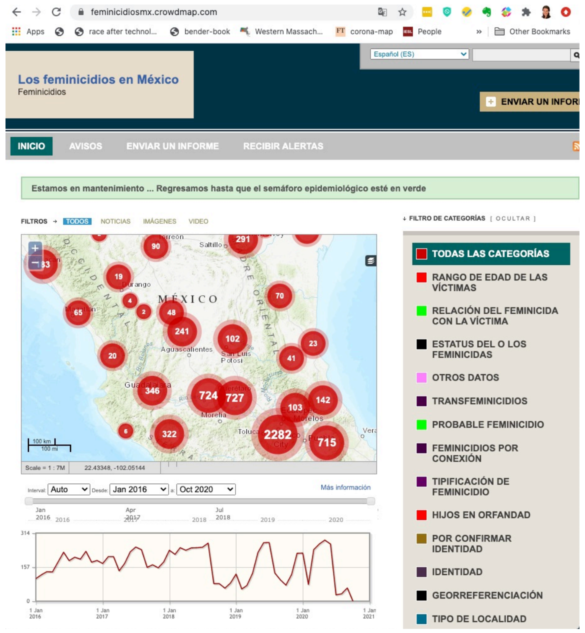
https://feminicidiosmx.crowdmap.com

We can reduce the **annotation burden** via methods such as distant supervision and its variants

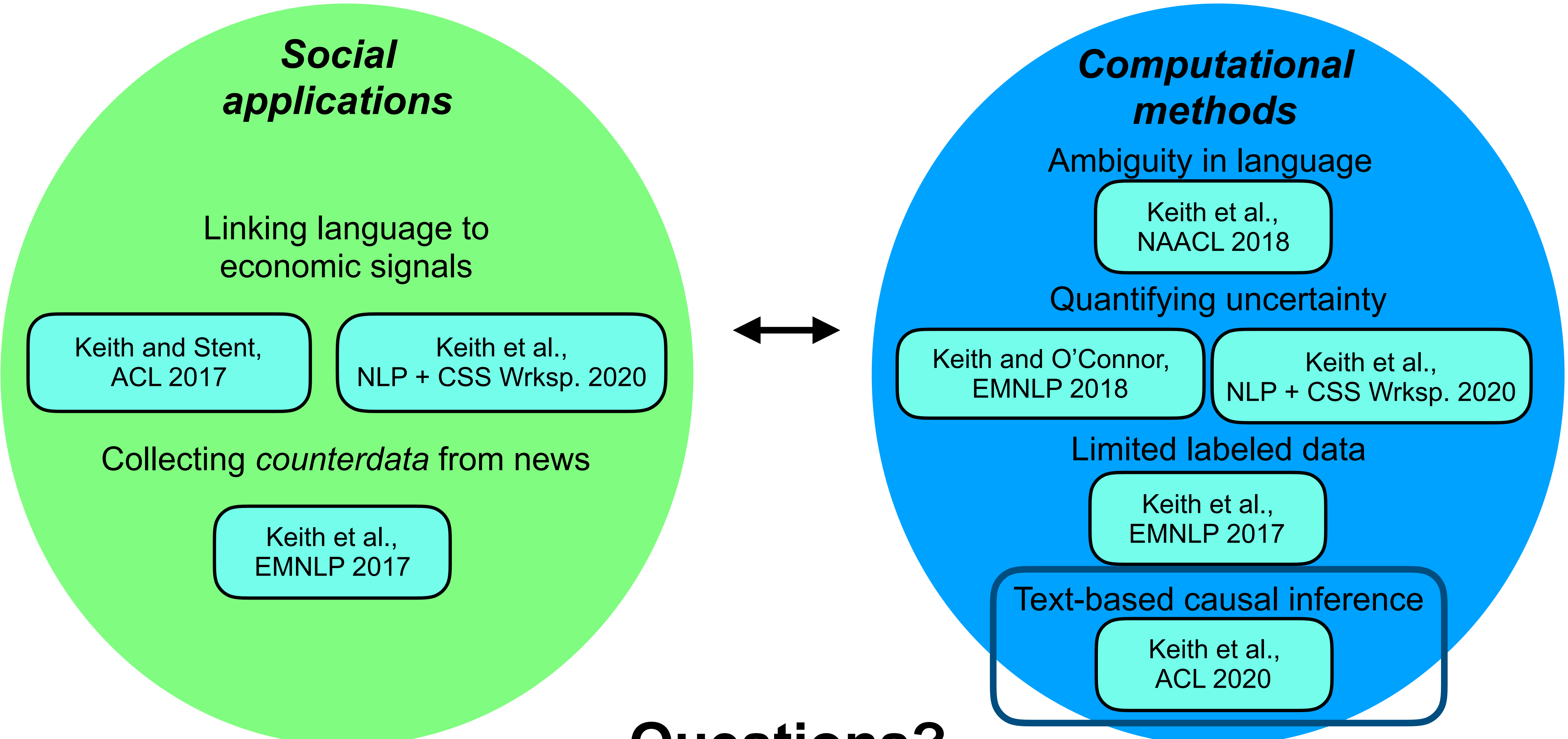


could help augment

Other counterdata collection efforts that are **currently done manually**



My Research Agenda



Questions?

How can **data science** contribute to **social impact?**

How can we improve outcomes in the world?

How can **data science** contribute to **social impact?**

How can we ***improve outcomes*** in the world?

values +
measurement

How can **data science** contribute to **social impact?**

How can we improve outcomes in the world?

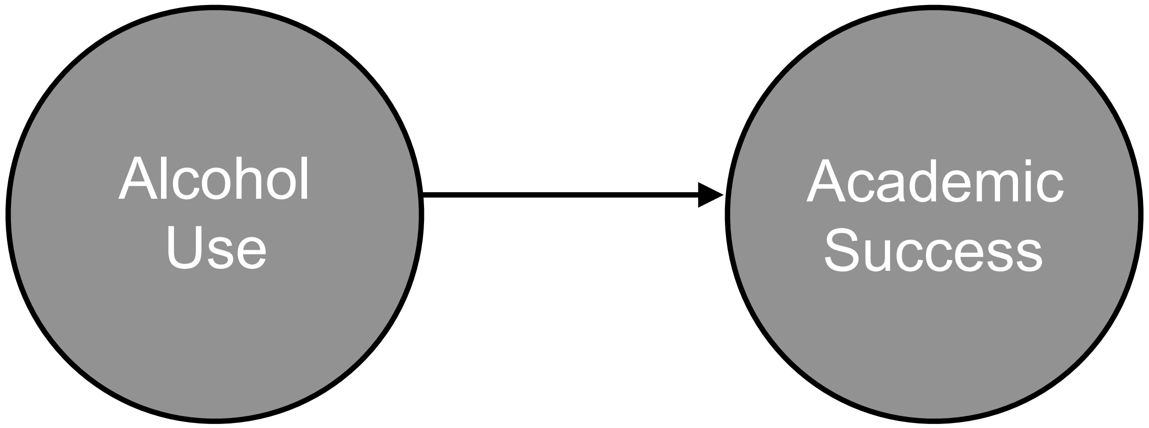
causal

Causality

How can we improve outcomes in the world?

How can we improve outcomes in the world?

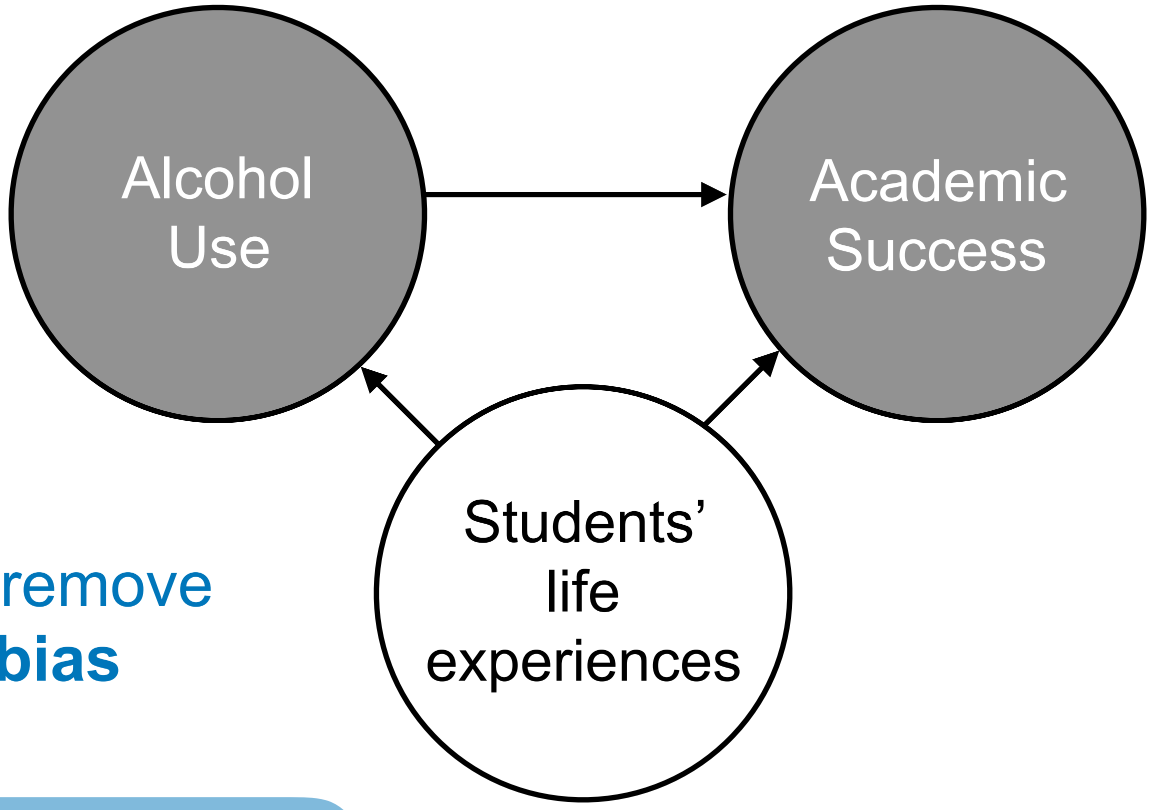
For college students, what is the effect of alcohol use on academic success?



(Kiciman et al. Using longitudinal social media analysis to understand the effects of early college alcohol use. ICWSM, 2020)

How can we improve outcomes in the world?

For college students, what is the effect of alcohol use on academic success?



Problem: need to remove **confounding bias**

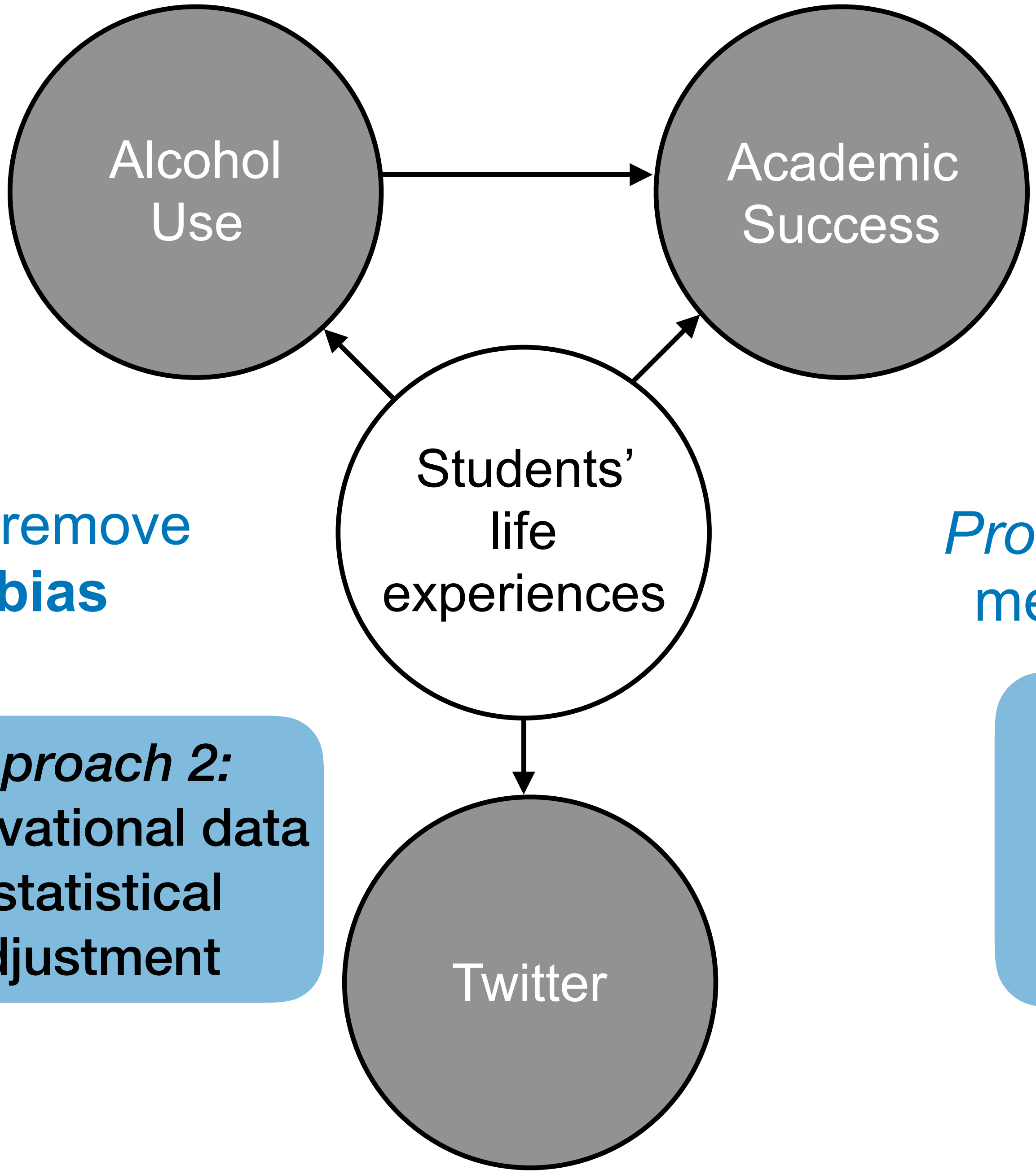
Approach 1: Intervention

Approach 2: observational data + statistical adjustment

(Kiciman et al. Using longitudinal social media analysis to understand the effects of early college alcohol use. ICWSM, 2020)

How can we improve outcomes in the world?

For college students, what is the effect of alcohol use on academic success?



Problem: need to remove **confounding bias**

Approach 1: Intervention

Approach 2: observational data + statistical adjustment

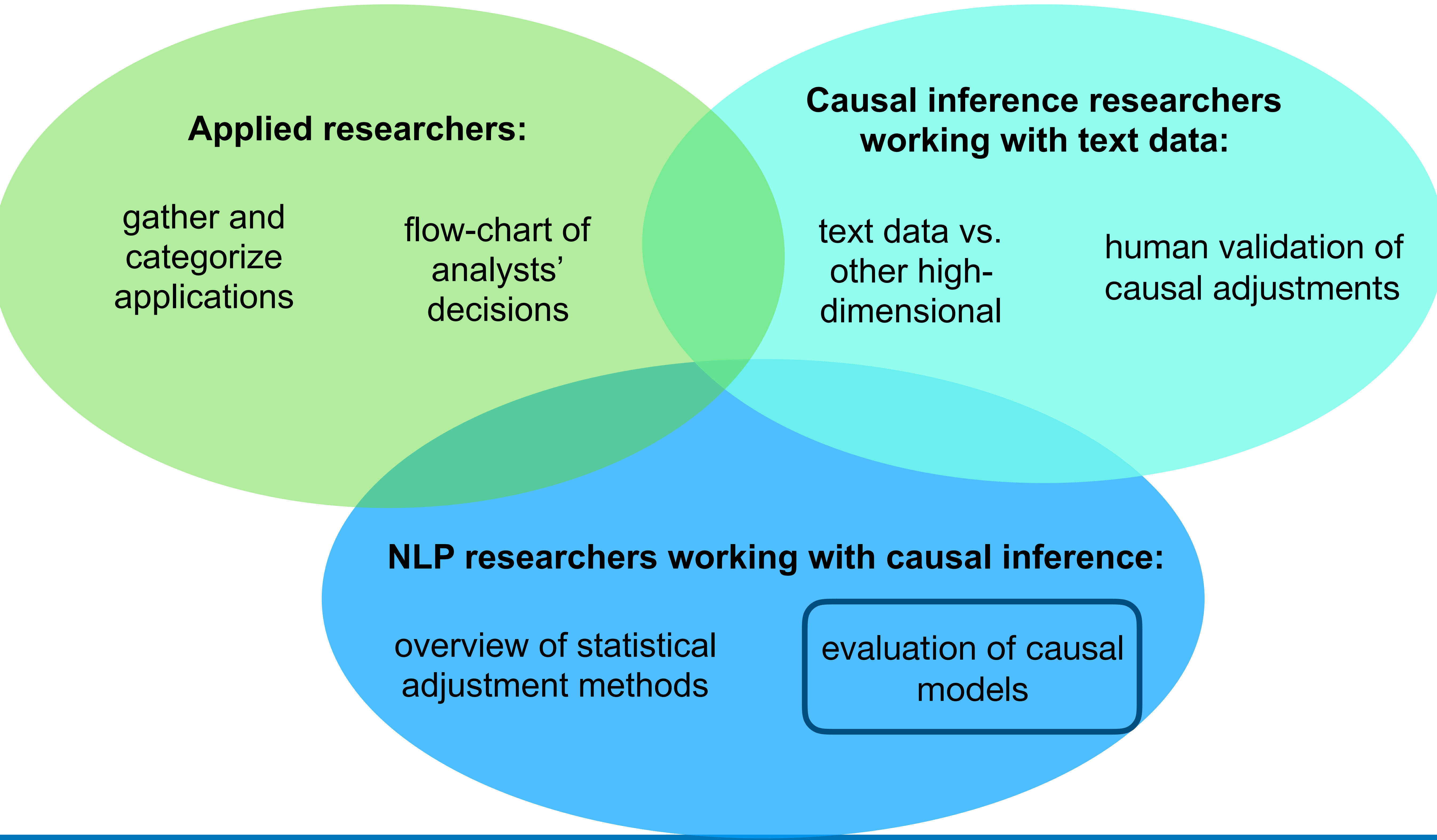
Problem: cannot directly measure confounders

One approach: Use text as a **surrogate** for confounders

(Kiciman et al. Using longitudinal social media analysis to understand the effects of early college alcohol use. ICWSM, 2020)

How does one use text to adjust for confounding?

(***Keith*** *et al. "Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates." ACL, 2020)*



Future work: evaluating text-based causal methods

Problem Type	Evaluation
Predictive	Predictive performance (e.g. accuracy) on a held-out test set
Causal	Estimated vs. true causal effects

Difficult to obtain!

Future work: evaluating text-based causal methods

(A) Constructed observational studies

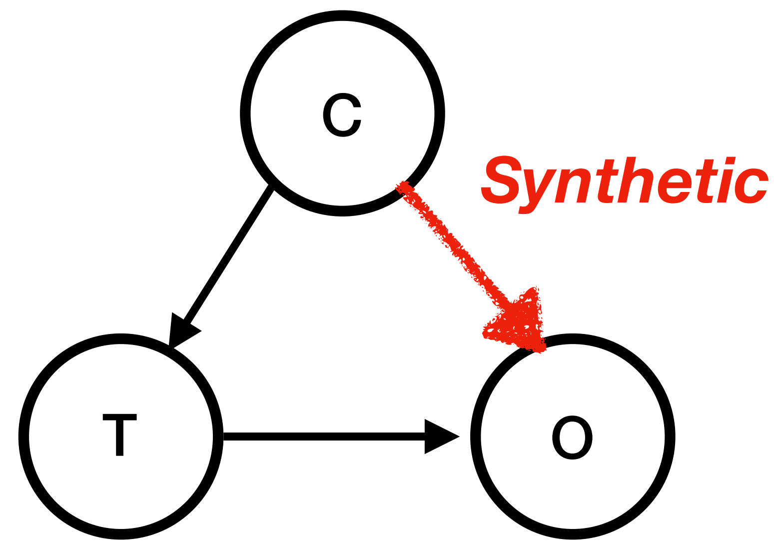
Randomized

↔

Non-randomized

In other social sciences:
(LaLonde (1986); Shadish et al. (2008); Glynn and Kashin (2013))

(B) Semi-synthetic datasets



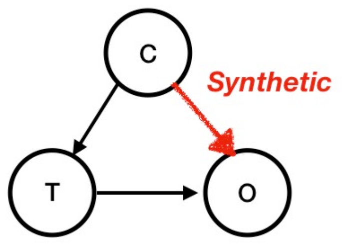
With text to remove confounding:
(Johansson et al. 2016; Veitch et al. 2019; Roberts et al. 2020)

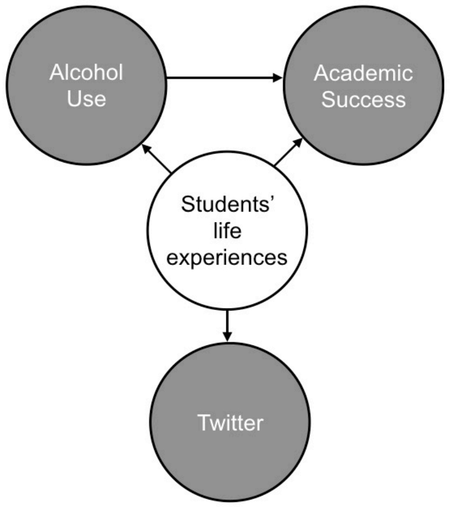
Many **open problems** in text-based causal inference

needed to answer

Socially impactful text-based causal questions

Causal evaluation





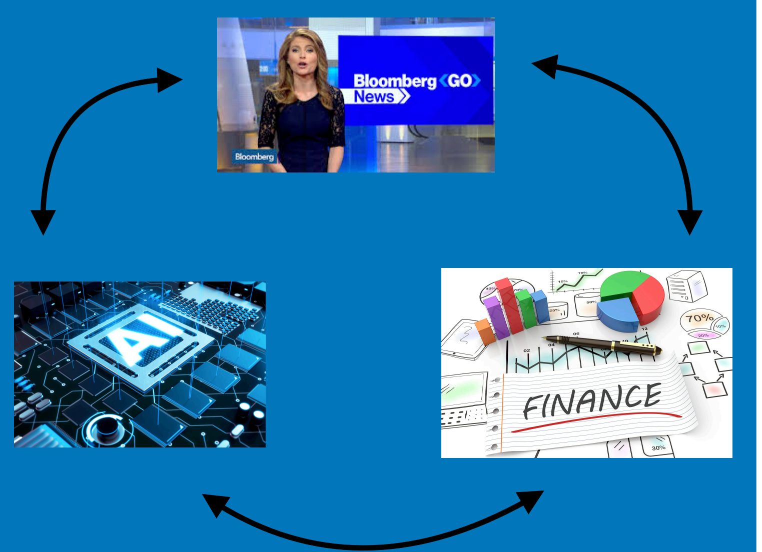
How can **data science** contribute to **social impact?**

How can we ***improve outcomes*** in the world?

causal

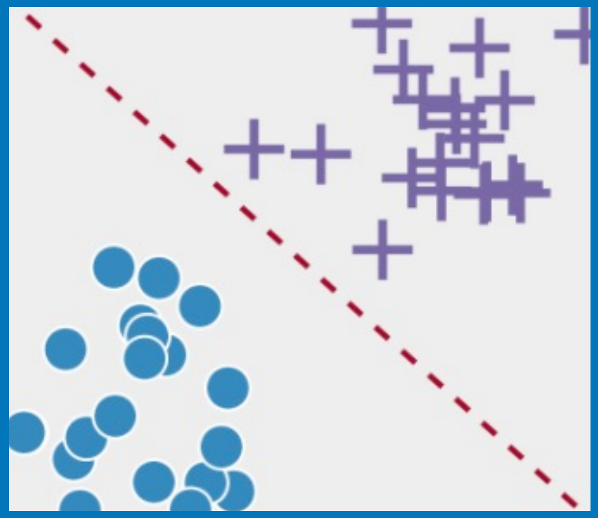
values + measurement

(1) Interdisciplinary collaborations



(2) Gathering additional text data sources

(3) Improving computational methods



Thank you to my co-authors!









My Research Agenda

Social applications

Linking language to economic signals

- Keith and Stent, ACL 2017
- Keith et al., NLP + CSS Wrksp. 2020

Collecting *counterdata* from news

- Keith et al., EMNLP 2017

↔

Computational methods

Ambiguity in language

- Keith et al., NAACL 2018

Quantifying uncertainty

- Keith and O'Connor, EMNLP 2018
- Keith et al., NLP + CSS Wrksp. 2020

Limited labeled data

- Keith et al., EMNLP 2017

Text-based causal inference

- Keith et al., ACL 2020

Thank you! Questions?