

Katie Keith

<https://kakeith.github.io/>
Twitter: @kakeith



Current:
PhD Computer Science



Next:
Post-Doc
Assistant Professor



Research Area: Text-as-data Methods



NLP
ML
Stats
Causal Inference

Text data



Text-as-data Methods

- 1. Dependency parsing**
- 2. Zero-shot language models**
- 3. Evaluation of causal intermediates**

Identifying civilians killed by police with distantly supervised entity-event extraction

Katherine A. Keith, Abram Handler, Michael Pinkham,
Cara Magliozzi, Joshua McDuffie, and Brendan O'Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

kkeith@cs.umass.edu, brenocon@cs.umass.edu

<http://slanglab.cs.umass.edu/>

Abstract

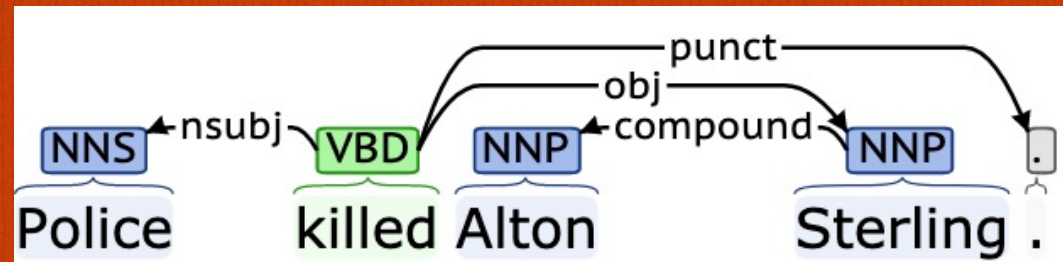
We propose a new, socially-impactful task for natural language processing: from a news corpus, extract names of persons who have been killed by police. We present a newly collected police fatality corpus, which we release publicly, and present a model to solve this problem that uses EM-based distant supervision with logistic regression and convolutional neural network classifiers. Our model out-

| Text | Person killed by police? |
|---|--------------------------|
| Alton Sterling was killed by police. | True |
| Officers shot and killed Philando Castile . | True |
| Officer Andrew Hanson was shot. | False |
| Police report Megan Short was fatally shot in apparent murder-suicide. | False |

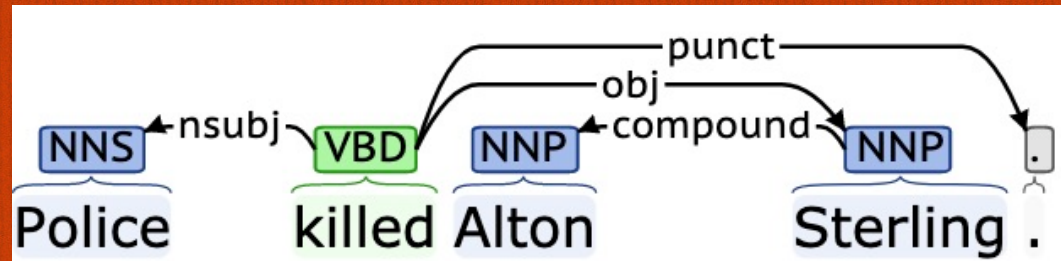
Table 1: Toy examples (with entities in bold) illustrating the problem of extracting from text names of persons who have been killed by police.

Empirical Methods in Natural Language Processing (EMNLP), 2017

Syntactic dependency parsing

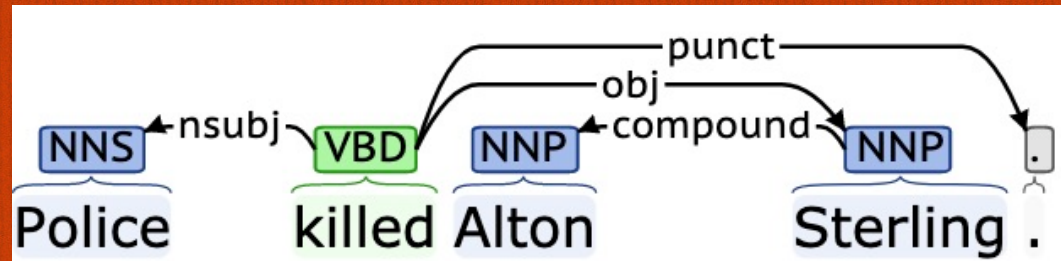


Syntactic dependency parsing

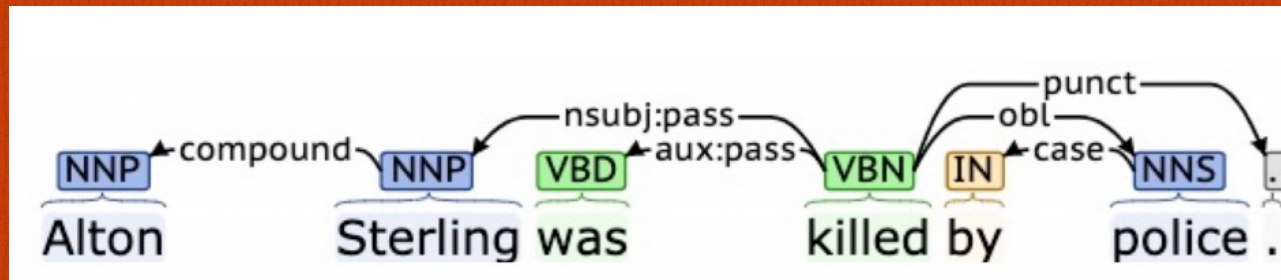


police -> nsubj -> killed -> obj -> Alton Sterling

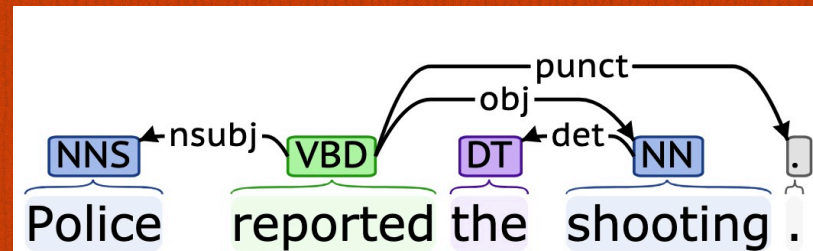
Syntactic dependency parsing



police -> nsubj -> killed -> obj -> Alton Sterling



Alton Sterling -> nsubj:pass -> killed -> obl -> police



police -> nsubj -> reported

<https://corenlp.run/>

Syntactic dependency parsing

- + Interpretability
- + Beyond bag-of-words (or lexical semantics)



Data available!
~1 million news documents

<http://slanglab.cs.umass.edu/PoliceKillingsExtraction/>



Text-as-data Methods

- 1. Dependency parsing**
- 2. Zero-shot language models**
- 3. Evaluation of causal intermediates**

Corpus-Level Evaluation for Event QA: The IndiaPoliceEvents Corpus Covering the 2002 Gujarat Violence

Andrew Halterman*

Massachusetts Institute of Technology
ahalt@mit.edu

Katherine A. Keith*

University of Massachusetts Amherst
kkeith@cs.umass.edu

Sheikh Muhammad Sarwar*

University of Massachusetts Amherst
smsarwar@cs.umass.edu

Brendan O'Connor

University of Massachusetts Amherst
brenocon@cs.umass.edu

Abstract

Automated event extraction in social science applications often requires corpus-level evaluations: for example, aggregating text predictions across metadata and unbiased estimates of recall. We combine corpus-level evaluation requirements with a real-world, social science setting and introduce the INDIAPOLICEEVENTS corpus—all 21,391 sentences from 1,257 English-language *Times of India* articles about events in the state of Gujarat during March 2002. Our trained annotators read and label every document for mentions of police activity events, allowing for unbiased recall evaluations. In contrast to other datasets with structured event representations, we gather annotations by posing natural questions, and evaluate off-the-shelf models for three different tasks: sentence classification,

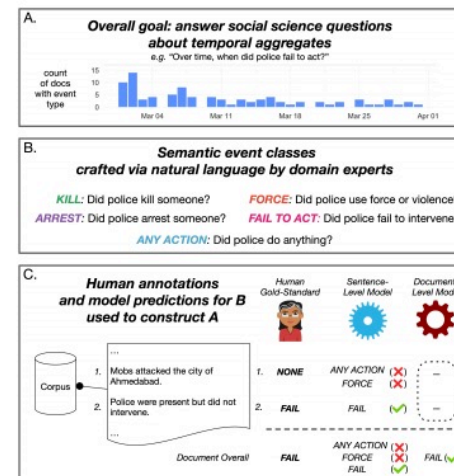


Figure 1: Motivation (A-B) and procedures (B-C) for this paper: **A.** Social scientists often use text data to an-

Findings of Association for Computational Linguistics (ACL), 2021

Data annotated with natural questions

1257 documents

21,391 sentences

From Times of India, March 2002

KILL: Did police kill someone?

FORCE: Did police use force or violence?

ARREST: Did police arrest someone?

FAIL TO ACT: Did police fail to intervene?

ANY ACTION: Did police do anything?

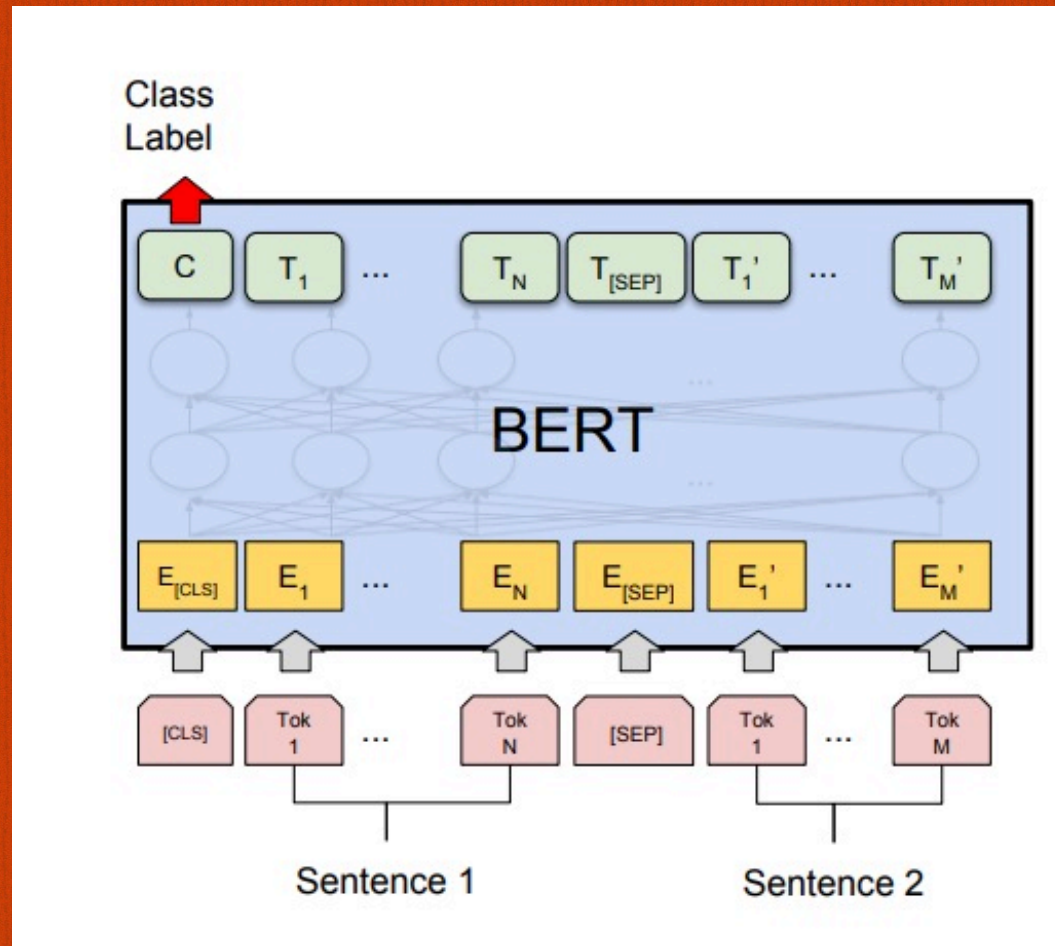
<https://github.com/slanglab/IndiaPoliceEvents>

Data annotated with natural questions

- + easy for annotators**
- + constraints on arguments (police being agents)**
- + predicates not specified in ontologies**
- + can use modern “every task in NLP is QA” paradigm**

Zero-Shot Classifiers

Entailment
Neutral
Contradiction



Police officials said nearly 2,537 people have so far been rounded up in the state.

Police arrested someone.

(Devlin et al. 2019)

Surprisingly positive results for a zero-shot learner with no training data

Zero-shot
BERT-like
model



| Event Class | <i>Task 1: Sent. Cls.</i> | |
|-------------|---------------------------|-------------|
| | Keyw. | R+MNLI |
| | F1 ↑ | F1 ↑ |
| KILL | 0.50 | 0.74 |
| ARREST | 0.48 | 0.62 |
| FAIL TO ACT | 0.05 | 0.48 |
| FORCE | 0.65 | 0.62 |
| ANY ACTION | 0.67 | 0.57 |



Text-as-data Methods

- 1. Dependency parsing**
- 2. Zero-shot language models**
- 3. Evaluation of causal intermediates**

Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates

Katherine A. Keith, David Jensen, and Brendan O'Connor
College of Information and Computer Sciences
University of Massachusetts Amherst
{kkeith, jensen, brenocon}@cs.umass.edu

Abstract

Many applications of computational social science aim to infer causal conclusions from non-experimental data. Such *observational* data often contains *confounders*, variables that influence both potential causes and potential effects. Unmeasured or *latent* confounders can bias causal estimates, and this has motivated interest in measuring potential confounders from observed text. For example, an individual's entire history of social media posts or the content of a news article could provide a rich measurement of multiple confounders. Yet, methods and applications for this problem are scattered across different communities and evaluation practices are inconsistent. This review is the first to gather and categorize these examples and provide a guide to data-processing and evaluation decisions. Despite increased attention on adjusting for confounding using text, there are still many open problems, which we highlight in this paper.

1 Introduction

In contrast to descriptive or predictive tasks, causal inference aims to understand how *intervening* on one variable affects another variable (Holland, 1986; Pearl, 2000; Morgan and Winship, 2015; Imbens and Rubin, 2015; Hernn and Robins, 2020).

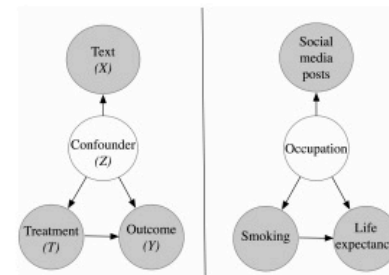
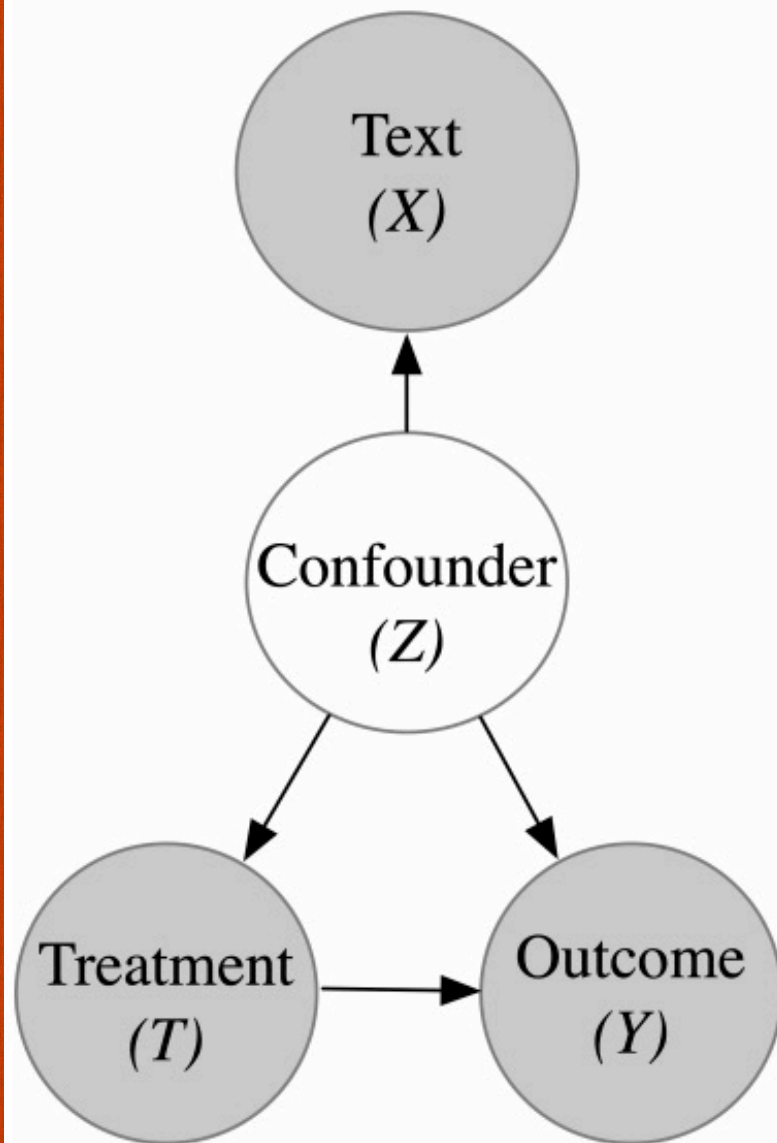


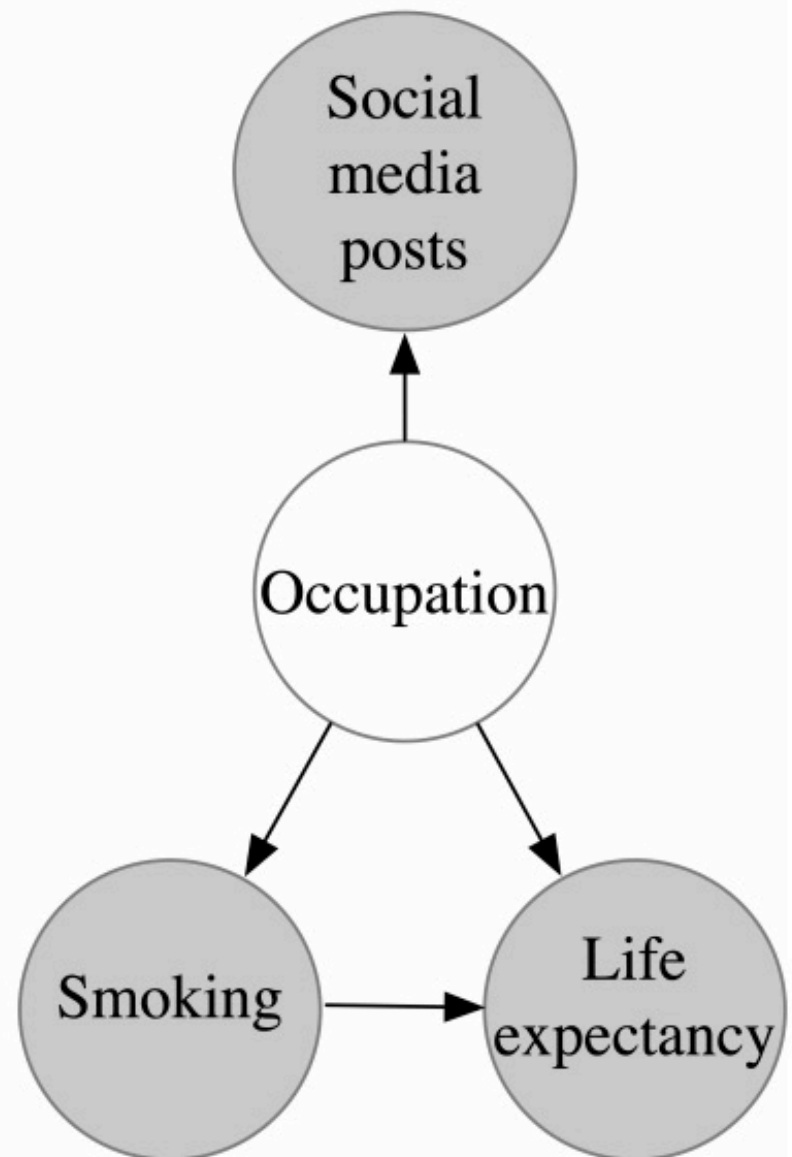
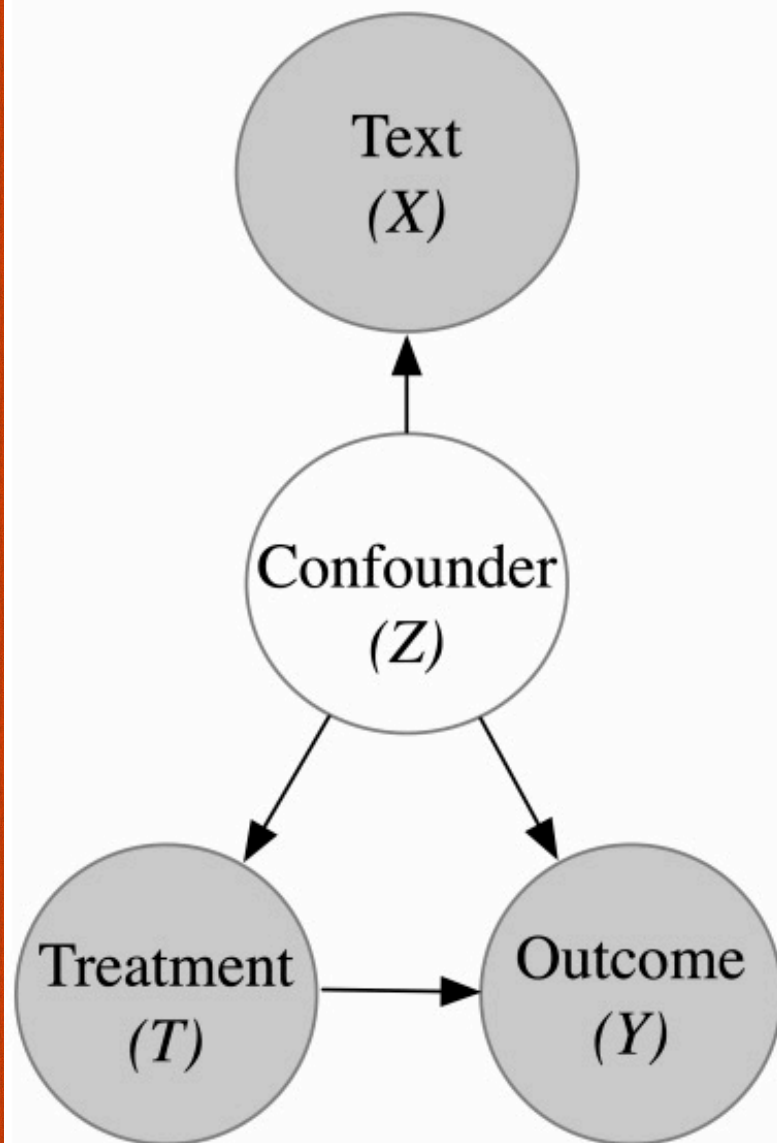
Figure 1: *Left*: A causal diagram for text that encodes causal confounders, the setting that is focus of this review paper. The major assumption is that latent confounders can be *measured* from text and those confounder measurements can be used in causal adjustments. *Right*: An example application in which practitioner does not have access to the confounding variable, *occupation*, in structured form but can measure confounders from unstructured text (e.g. an individual's social media posts).

strongly biased estimates and thus invalid causal conclusions.

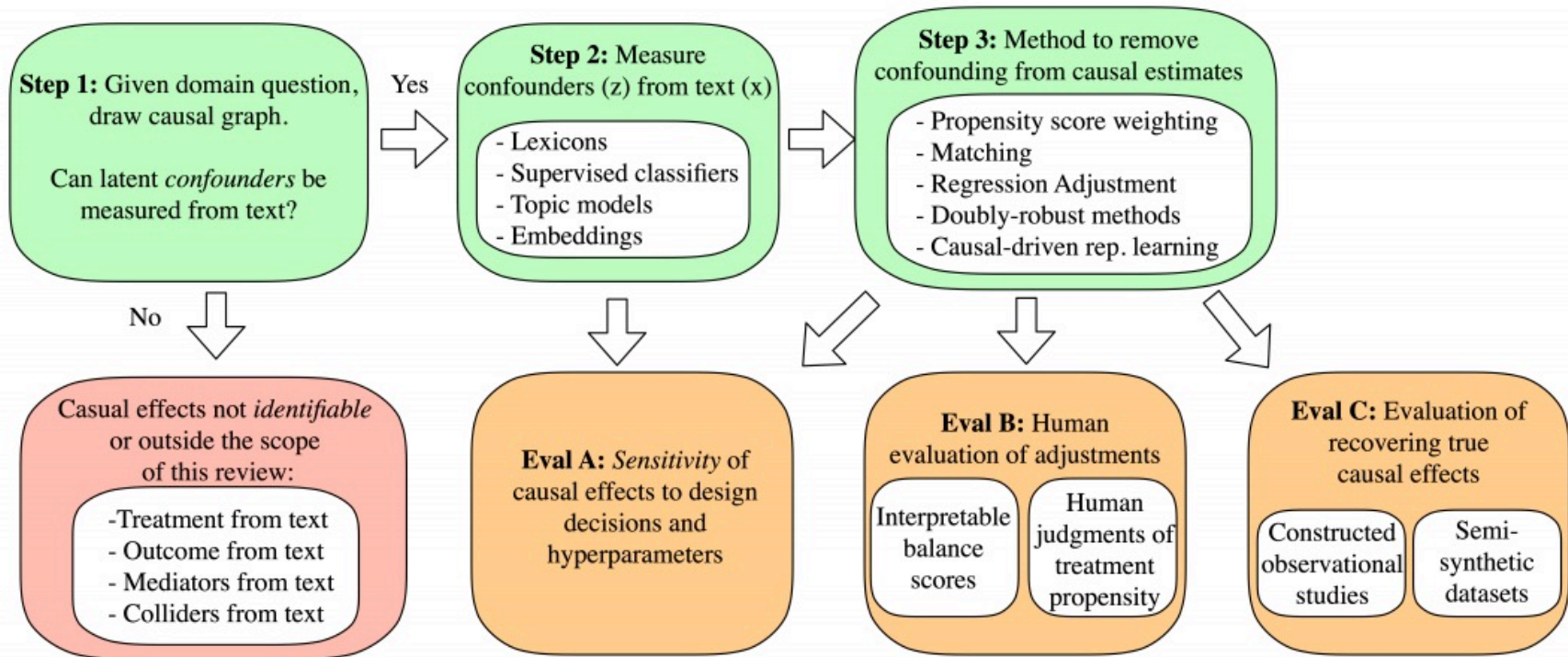
To eliminate confounding bias, one approach is to perform randomized controlled trials (RCTs) in which researchers randomly assign treatment. Yet, in many research areas such as healthcare, educa-

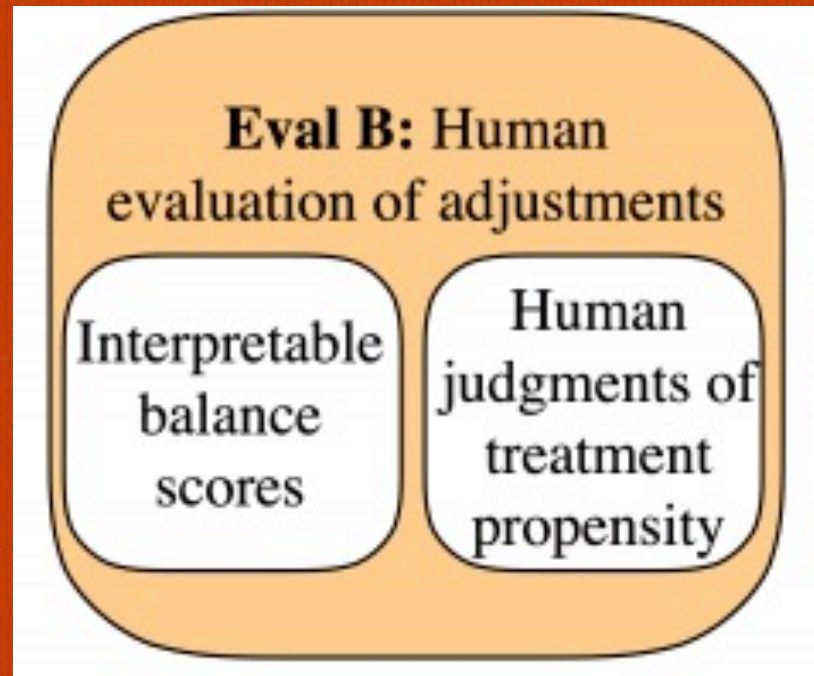
Association for Computational Linguistics (ACL), 2020





Applied analysis flowchart





Text for treated unit

Text for control unit

**To Crowdworker:
How similar are these two
texts (1-5)?**

Two shameless plugs!



First Workshop on Causal Inference & NLP

Home

Call for Papers

Workshop Program

Organizing Team

Curated Research

[Submit Paper](#)

November 10/11, 2021 at EMNLP 2021

The CI+NLP workshop brings together domain and methodological experts across academic communities, with an interest in the intersection of causal inference and natural language processing.

Together, we will explore how current and new NLP methodology can contribute to establishing and evaluating causal relationships, and how substantive expertise from different domains can contribute to answering causal questions with text.

Subscribe to our [mailing list](#) to stay updated!

Invited Speakers

- [Susan Athey](#) (Stanford)
- [David Blei](#) (Columbia)
- [Cristian Danescu-Niculescu-Mizil](#) (Cornell)

<https://causaltext.github.io/2021/>

Submissions due: Aug 5




Diaries of Social Data Research

By Katherine A. Keith & Lucy Li

Large-scale data has become a major component of research about human behavior and society. But how are interdisciplinary collaborations that use large-scale social data formed and maintained? What obstacles are encountered on the journey from idea conception to publication? In this podcast, we investigate these questions by probing the "research diaries" of scholars who have published papers in computational social science and adjacent fields. We unmask the research process with the hope of normalizing the challenges of finding research research with little fanfare.

Listen on  Spotify

 Message

<https://anchor.fm/diaries-soc-data-research>

In search of guests!

Thanks!



Katie Keith

<https://kakeith.github.io/>

Twitter: @kakeith



- 1. Dependency parsing**
- 2. Zero-shot language models**
- 3. Evaluation of causal intermediates**