Causal Inference with Text

Sandeep Soni

04/18/2024

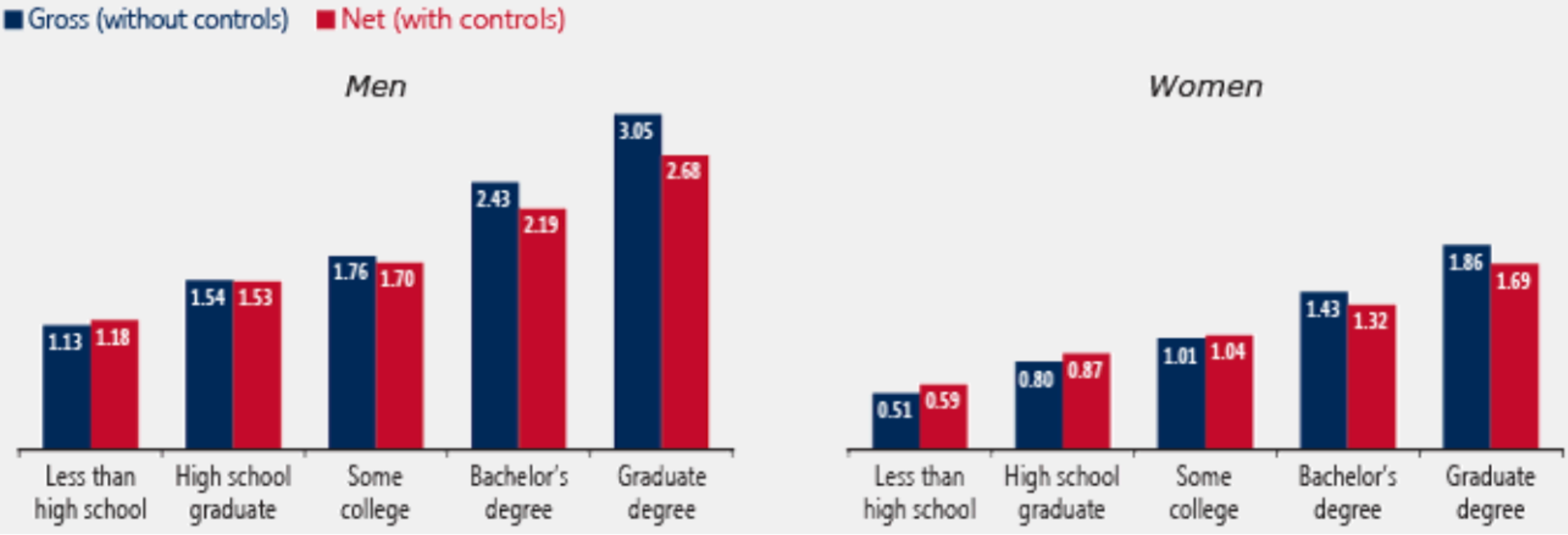
Association

If there is a statistical relationship between two variables, we say that they are correlated or associated

Example

- Higher education is associated with high earnings

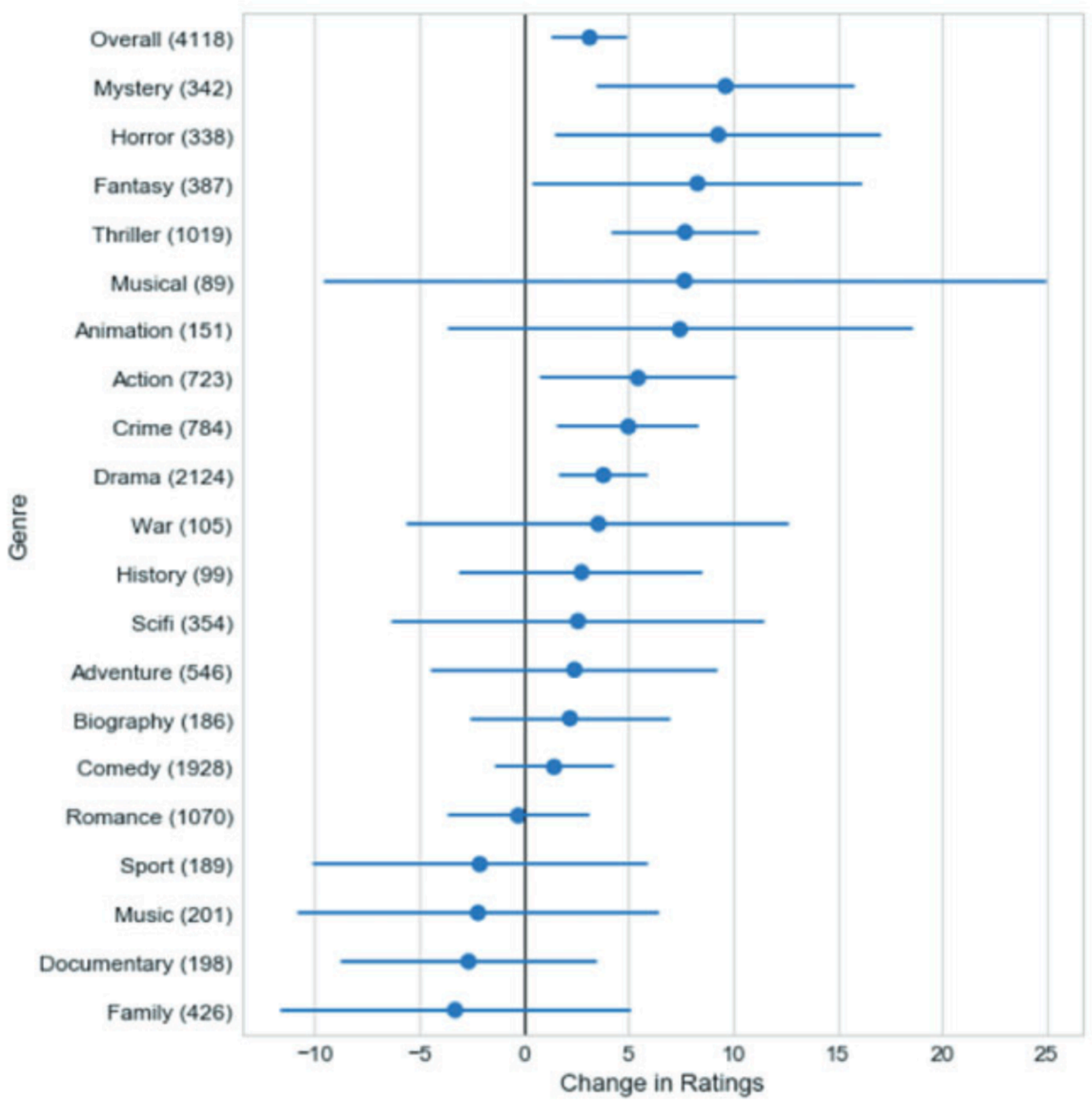
Estimated lifetime earnings by educational attainment (in millions of dollars)



Example

- Movies that are emotional rollercoasters are more likely to be successful.

- Change in sentiment over the length of the movie as a measure of emotional volatility
- IMDB ratings of movies as a measure of their success



Association

- Association between two variables can be tested using hypothesis tests
- Known associations can also be used to boost performance of predictive models

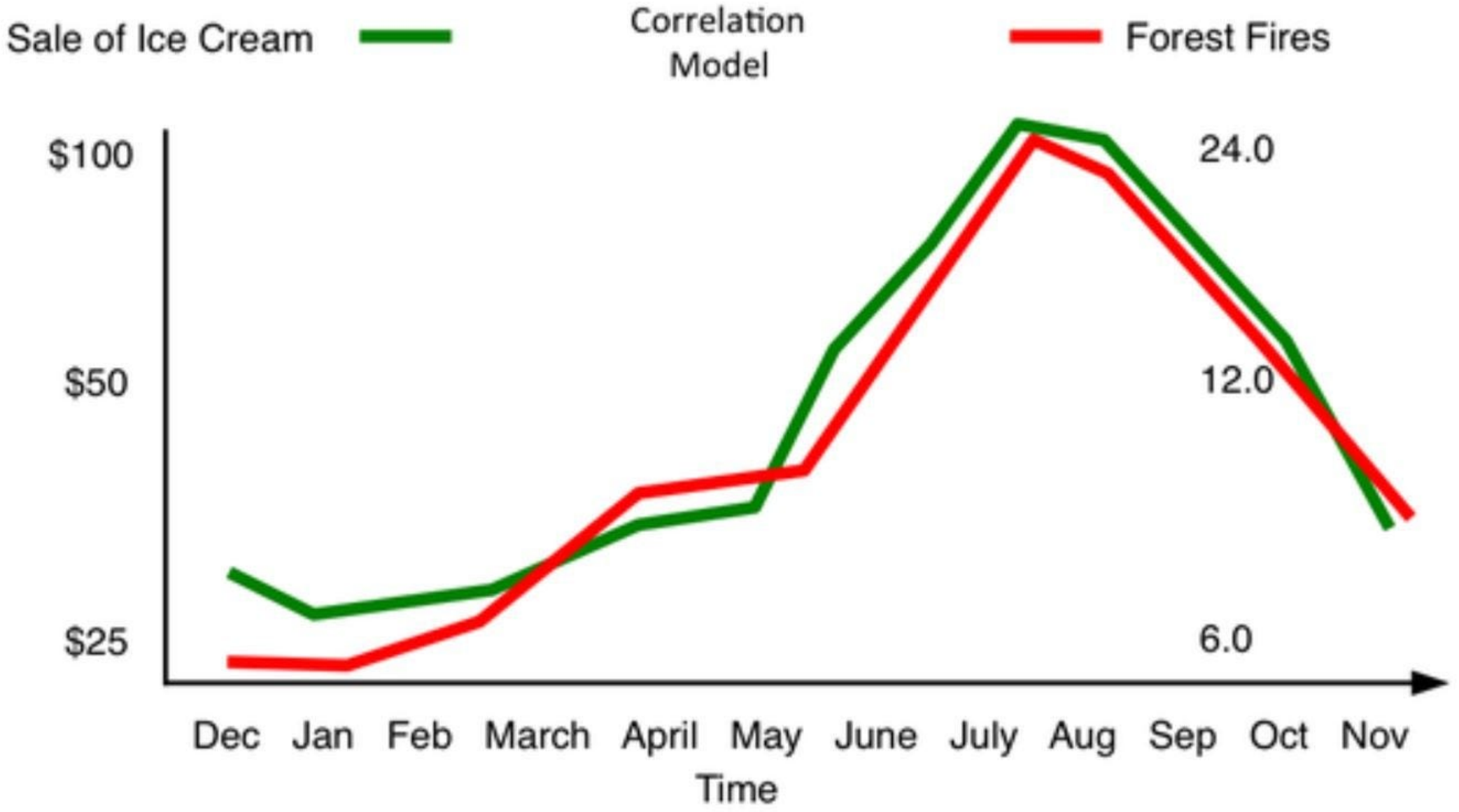
Association

Relationship Type	Operationalization	Examples
Association	$P(y \mid x)$	Emotional rollercoasters are more likely to be successful

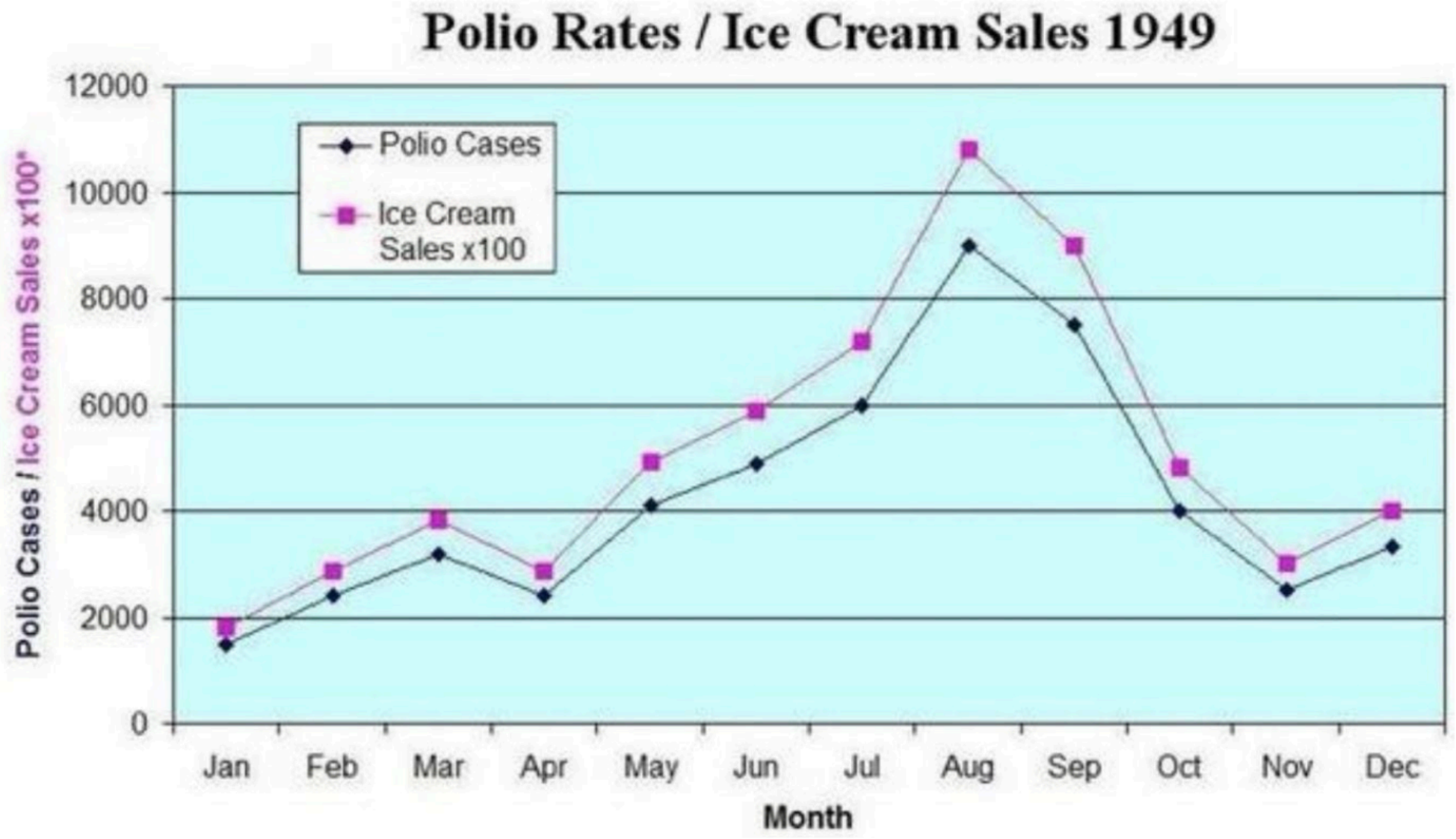
- Hypothesis: Stock markets are likely to be up at year end if an NFC team wins superbowl
- Data suggests 75% correlation
- Should we use this to predict stock price movement?

Year	Super Bowl Winner	S&P 500 Index
1967	Packers (NFC)	20%
1968	Packers (NFC)	8%
1969	Jets (AFC)	-11%
1970	Chiefs (AFC)	0%
1971	Colts (AFC but pre-merger NFL)	11%
1972	Cowboys (NFC)	16%
1973	Dolphins (AFC)	-17%
1974	Dolphins (AFC)	-30%
1975	Steelers (AFC but pre-merger NFL)	32%
1976	Steelers (AFC but pre-merger NFL)	19%
1977	Raiders (AFC)	-12%

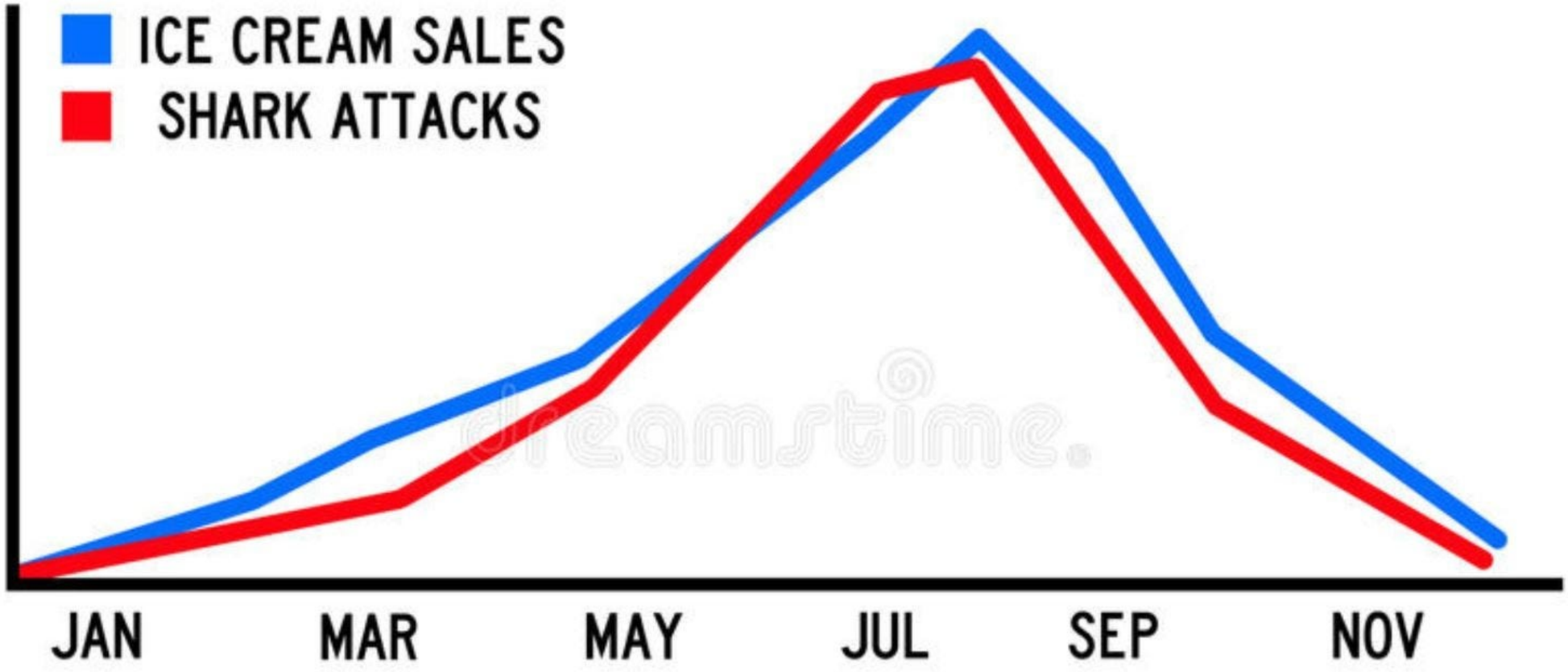
Two temporal variables are shown here. Are they related?



https://www.kdnuggets.com/2019/09/risk-ai-big-data.html



https://www.kdnuggets.com/2019/09/risk-ai-big-data.html

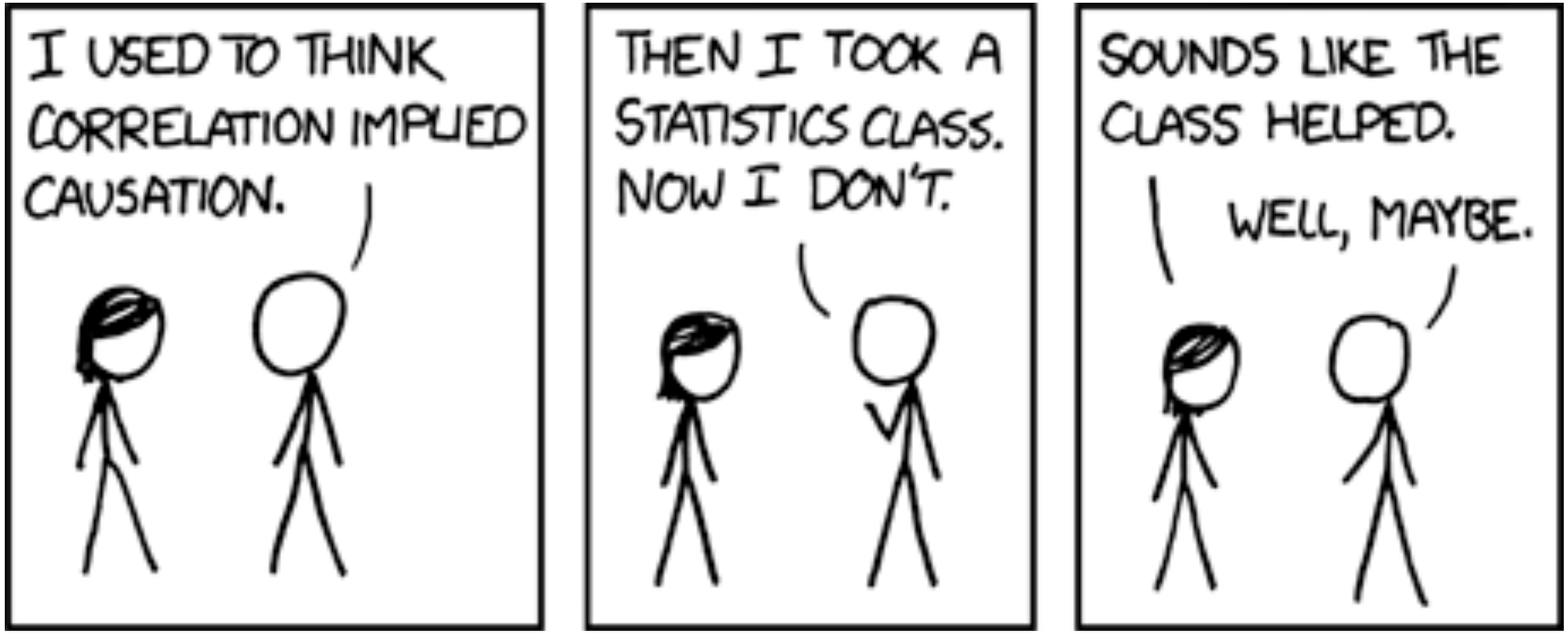


https://www.kdnuggets.com/2019/09/risk-ai-big-data.html

“So should we ban ice cream?”

What about such correlations?

- Two variables may be statistically associated but do not have a causal connection
- Such correlation is called spurious correlation


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Source: XKCD

“Correlation does not imply causation”

Causality

- Prediction: Assuming that two variables are associated, the objective is to predict one variable using another
- Counterfactual prediction: The objective is to predict what would happen to one variable in the absence of the other variable or a different variable.

Notation

- Let Y be an outcome of interest
- Let X be some treatment
- We want to find if X causes Y

Example

- Y={cured, not cured}
- X={pill given, pill not given}
- Question: Is the pill effective at curing the disease?

Experiment

- Randomly divide patients into two groups: one receiving treatment and other not receiving treatment
- Compare the recovery of patients from the two groups to measure the effect of treatment

Definitions

- Every patient is considered a unit
- Patients given a pill form the treatment group
- Patients not given a pill form the control group
- Because patients are randomly selected, such an experiment is called a randomized control trial

How to measure the effect on one unit?

$$P(Y_i \mid \mathrm{do}(X_i))$$

- We imagine the treatment to be a form of intervention
- So we assume that there is a function called do which can change the treatment on some unit

Treatment effect

$$Y_i \mid \mathrm{do}(X_i = 1) - Y_i \mid \mathrm{do}(X_i = 0)$$

- We can calculate the difference in outcome with and without the treatment to get a measure for the effect of the treatment

Average treatment effect

$$\text{ATE} = E[Y \mid \text{do}(X = 1) - Y \mid \text{do}(X = 0)]$$

- The average effect over all the units can give us the overall effect of the treatment

Why is causality difficult?

- Units are never stable
  - To perform do(X=1) and do(X=0), the same patient is measured under two different conditions; no way to isolate the effect
- Interventions are not always possible
- Randomization is not always possible

Potential outcomes framework

$$\text{ATE} = \frac{1}{N}\sum_{i=1}^{N} Y_i(1) - Y_i(0)$$

- Relax the assumption that we have the same unit both under the treatment and control conditions
- Compare units that have been treated to those that were not treated
- Assume SUTVA: treatment of one unit does not affect treatment or outcome of another unit

Association

Relationship Type	Operationalization	Examples
Association	$P(Y \mid X)$	Emotional rollercoasters are more likely to be successful
Causal	$P(Y \mid do(X))$	Emotional rollercoasters cause movies to be rated highly

How can NLP help?

- Identify confounders
- Design treatments
- Measure outcomes

Designing treatments

author	tweets	#retweets
natlsecuritycnn	t_1: FIRST ON CNN: After Petraeus scandal, Paula Broadwell looks to recapture 'normal life.' **http://t.co/qy7GGuYW**	$n_1 = 5$
	t_2: First on CNN: Broadwell photos shared with Security Clearance as she and her family fight media portrayal of her **[same URL]**	$n_2 = 29$
ABC	t_1: Workers, families take stand against Thanksgiving hours: **http://t.co/J9mQHiIEqv**	$n_1 = 46$
	t_2: Staples, Medieval Times Workers Say Opening Thanksgiving Day Crosses the Line **[same URL]**	$n_2 = 27$
cactus_music	t_1: I know at some point you've have been saved from hunger by our rolling food trucks friends. Let's help support them! **http://t.co/zg9jwA5j**	$n_1 = 2$
	t_2: Food trucks are the epitome of small independently owned LOCAL businesses! Help keep them going! Sign the petition **[same URL]**	$n_2 = 13$

Source: Tan et. al. (2014) The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter

Controlling for confounds

Adapting Text Embeddings for Causal Inference

Victor Veitch* **Dhanya Sridhar*** **David M. Blei**

Department of Statistics and Department of Computer Science
Columbia University

Abstract

Does adding a theorem to a paper affect its chance of acceptance? Does labeling a post with the author's gender affect the post popularity? This paper develops a method to estimate such causal effects from observational text data, adjusting for confounding features of the text such as the subject or writing quality. We assume that the text suffices for causal adjustment but that, in practice, it is prohibitively high-dimensional. To address this challenge

1 INTRODUCTION

This paper is about causal inference on text.

Example 1.1. Consider a corpus of scientific papers submitted to a conference. Some have theorems; others do not. We want to infer the causal effect of including a theorem on paper acceptance. The effect is confounded by the subject of the paper—more technical topics demand theorems, but may have different rates of acceptance. The data does not explicitly list the subject, but it does include each paper's abstract. We want to use the text to adjust for the subject and estimate the causal effect.

You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech

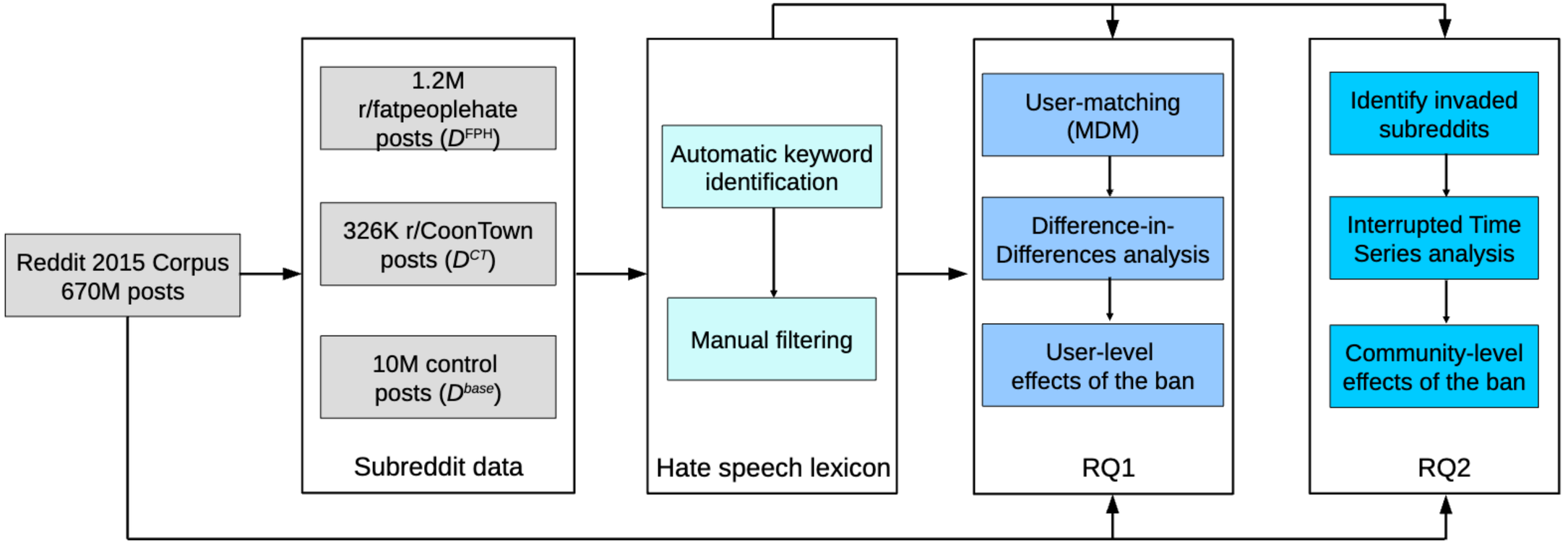
ESHWAR CHANDRASEKHARAN, Georgia Institute of Technology
UMASHANTHI PAVALANATHAN, Georgia Institute of Technology
ANIRUDH SRINIVASAN, Georgia Institute of Technology
ADAM GLYNN, Emory University
JACOB EISENSTEIN, Georgia Institute of Technology
ERIC GILBERT, University of Michigan

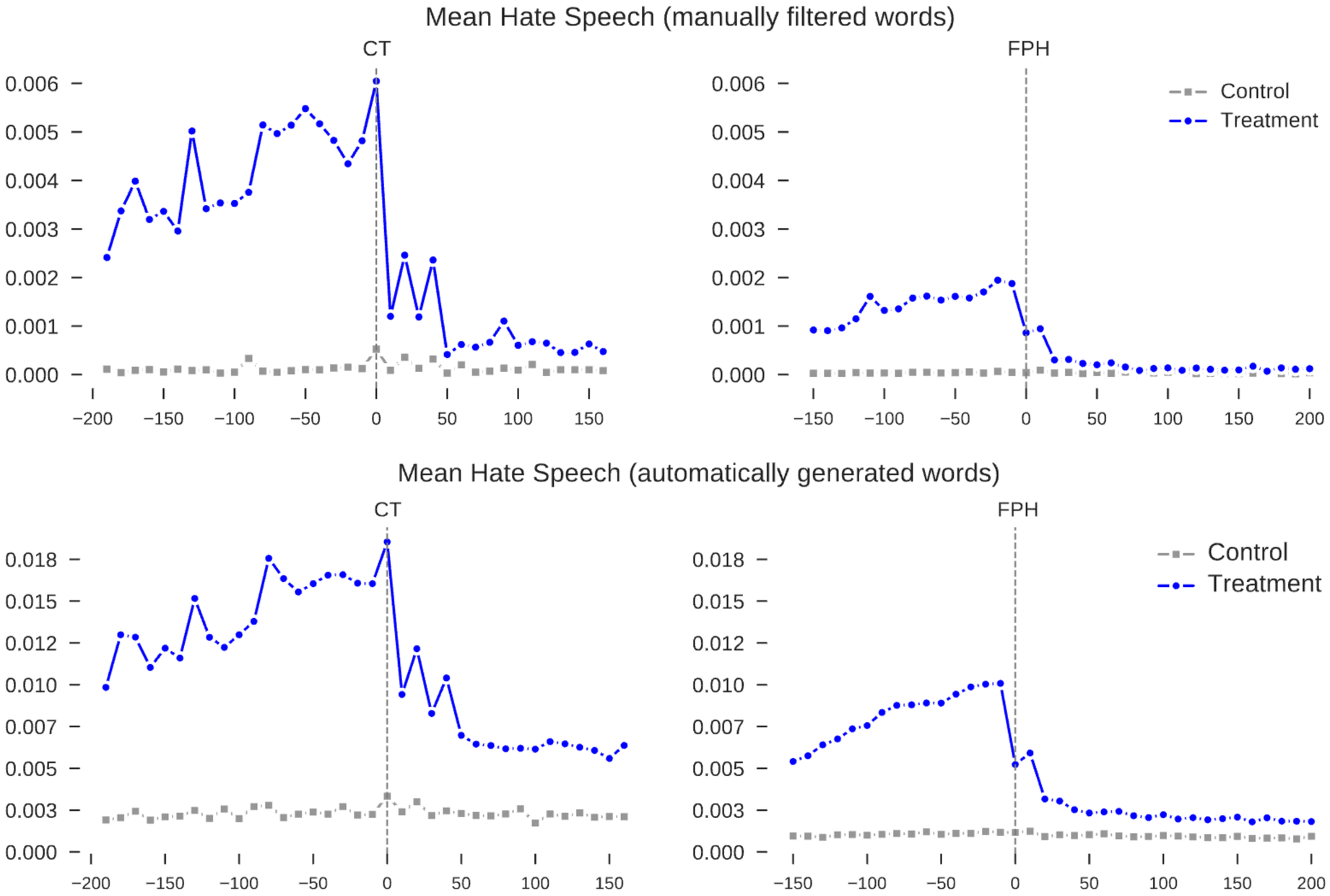
In 2015, Reddit closed several subreddits—foremost among them r/fatpeoplehate and r/CoonTown—due to violations of Reddit's anti-harassment policy. However, the effectiveness of banning as a moderation approach remains unclear: banning might diminish hateful behavior, or it may relocate such behavior to different parts of the site. We study the ban of r/fatpeoplehate and r/CoonTown in terms of its effect on both participating users and affected subreddits. Working from over 100M Reddit posts and comments, we generate hate speech lexicons to examine variations in hate speech usage via causal inference methods. We find that the *ban worked for Reddit*. More accounts than expected discontinued using the site; those that stayed drastically decreased their hate speech usage—by at least 80%. Though many subreddits saw an influx of r/fatpeoplehate and r/CoonTown "migrants," those subreddits saw no significant changes in hate speech usage. In other words, other subreddits did not inherit the problem. We conclude by reflecting on the apparent success of the ban, discussing implications for online moderation, Reddit and internet communities more broadly.

- Reddit banned a few subreddits after numerous complaints
- Question: Was the ban effective?
- Bigger question: Is censoring an effective moderation strategy?

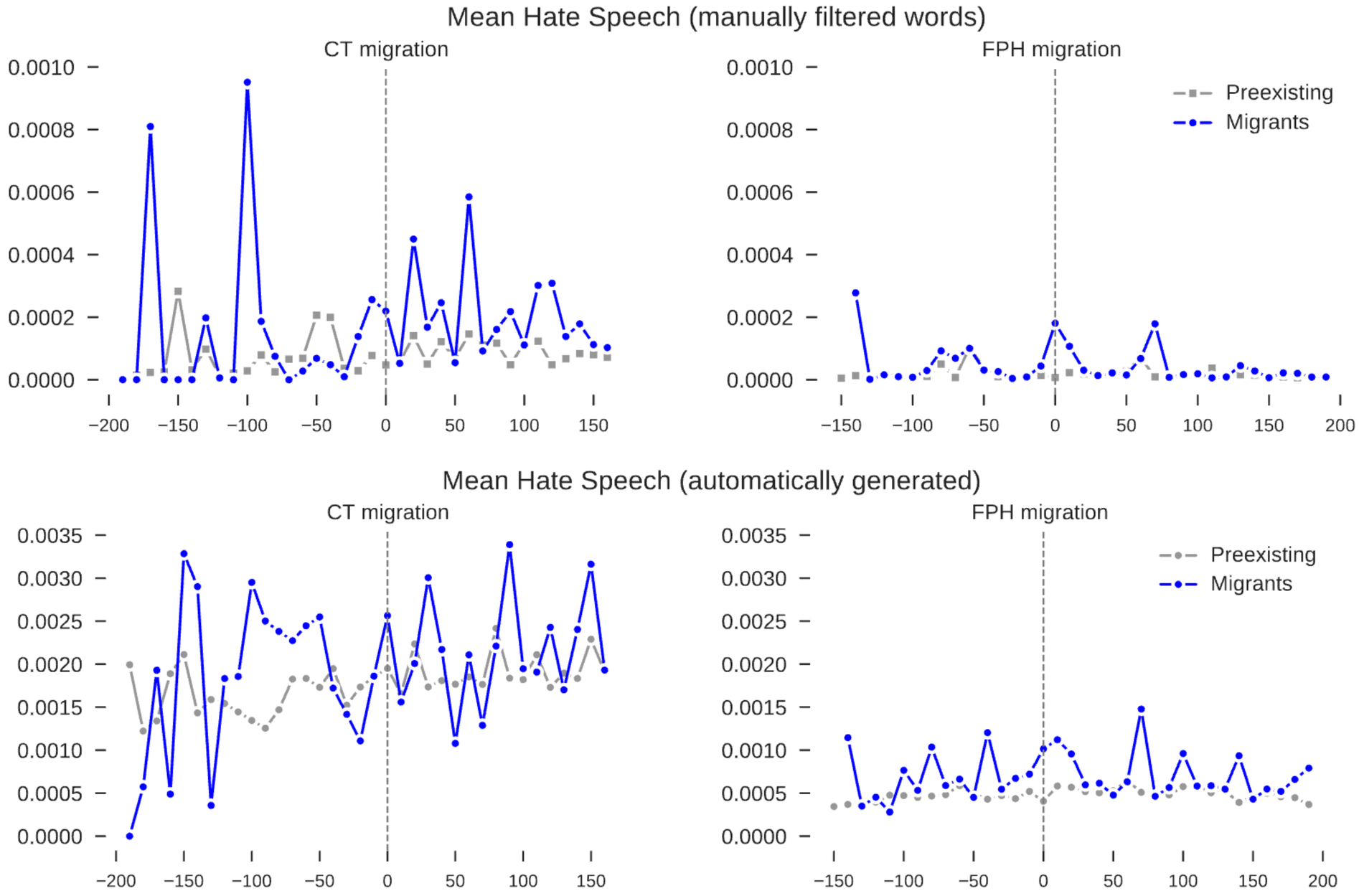
RQ1: What effect did Reddit's ban have on the contributors to banned subreddits?
RQ1a: How were their activity levels affected?
RQ1b: How did their hate speech usage change, if at all?

RQ2: What effect did the ban have on subreddits that saw an influx of banned subreddit users?
RQ2a: To which subreddits did the contributors to banned subreddits migrate after the ban?
RQ2b: How did hate speech usage by migrants change in these subreddits, if at all?
RQ2c: How did hate speech usage by preexisting users change in these subreddits, if at all?


Reddit 2015 Corpus
670M posts
1.2M r/fatpeoplehate posts (D^{FPH})
326K r/CoonTown posts (D^{CT})
10M control posts (D^{base})
Subreddit data
Automatic keyword identification
Manual filtering
Hate speech lexicon
User-matching (MDM)
Difference-in-Differences analysis
User-level effects of the ban
RQ1
Identify invaded subreddits
Interrupted Time Series analysis
Community-level effects of the ban
RQ2



Effect of ban on individual users



Effect of the ban on other communities