Causal Inference

Indira Sen
13.06.24

With many credits to Qingyuan Zhao [1]

Universität Konstanz

[1] https://www.statslab.cam.ac.uk/~qz280/talk/ssrmp-2020/slides

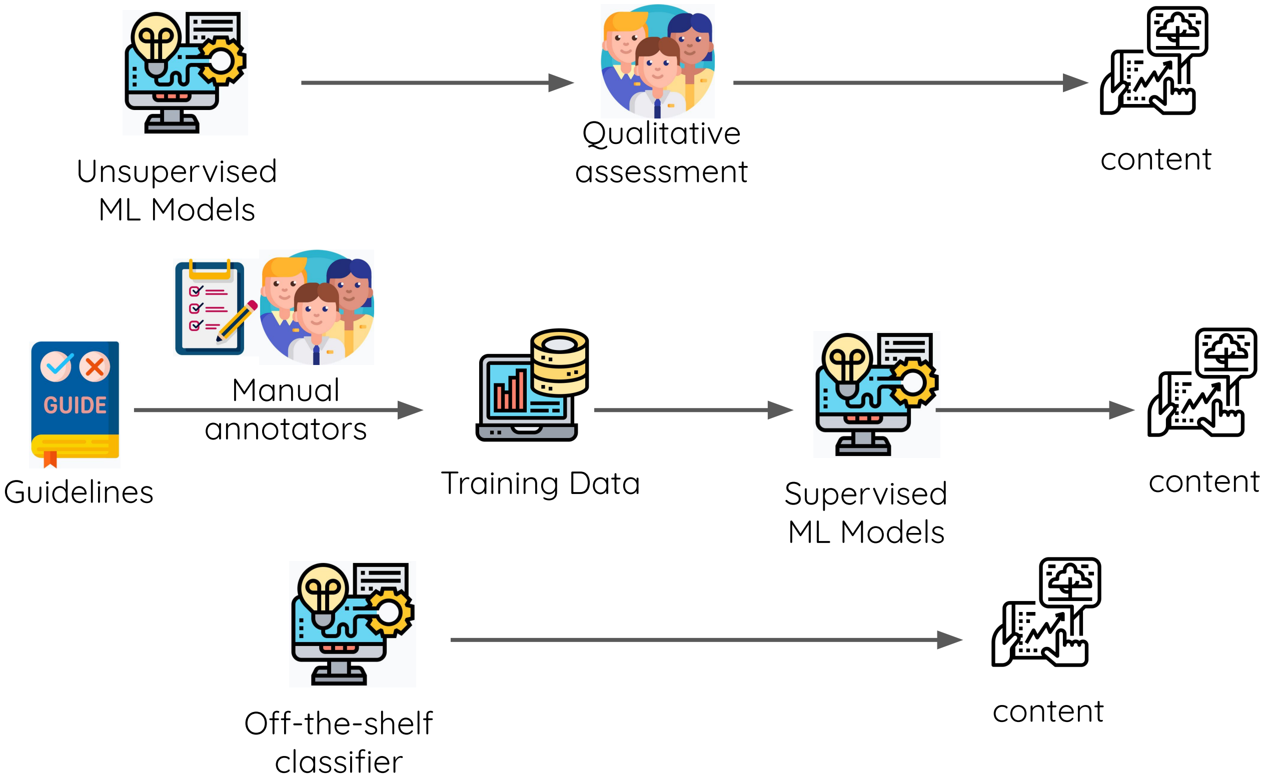
Agenda

- Recap (Text-as-data methods) + topic modeling
- Today: Causal Inference
  - Experiments
  - Causal inference with observational data
- Resources
- Hands-On

Recap

Text-as-Data Methods

- unsupervised
- supervised
- ‘off-the-shelf’



Today: Causal Inference

What is causal inference and why should you learn it?

- Causal inference is the study of how actions, interventions, or treatments affect outcomes of interest.

- Beyond prediction: explanations of phenomena

- Experiments, particularly randomized control trials (RCTs), considered gold standard of science

Some causal questions in the (C)SS

- Does fact-checking interventions like ‘community notes’ reduce the spread of misinformation?

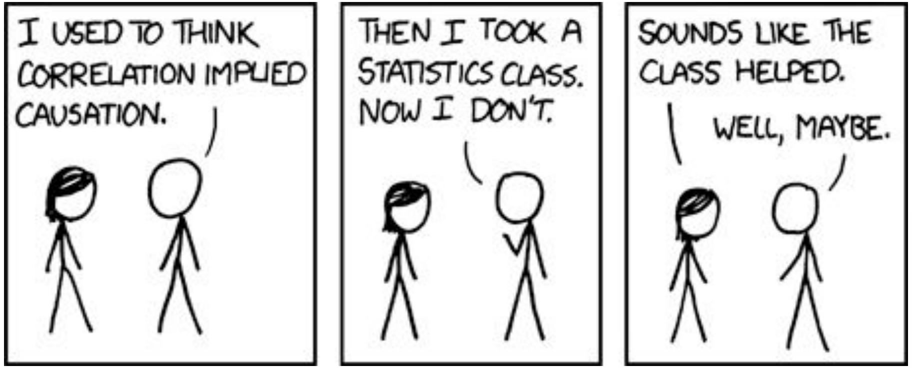
- Does in-group sanctioning reduce the spread of toxicity on social networks?

- Do non-anglophile names receive fewer replies to cold emails?

- Does deplatforming of extreme individuals improve community health?

- ?

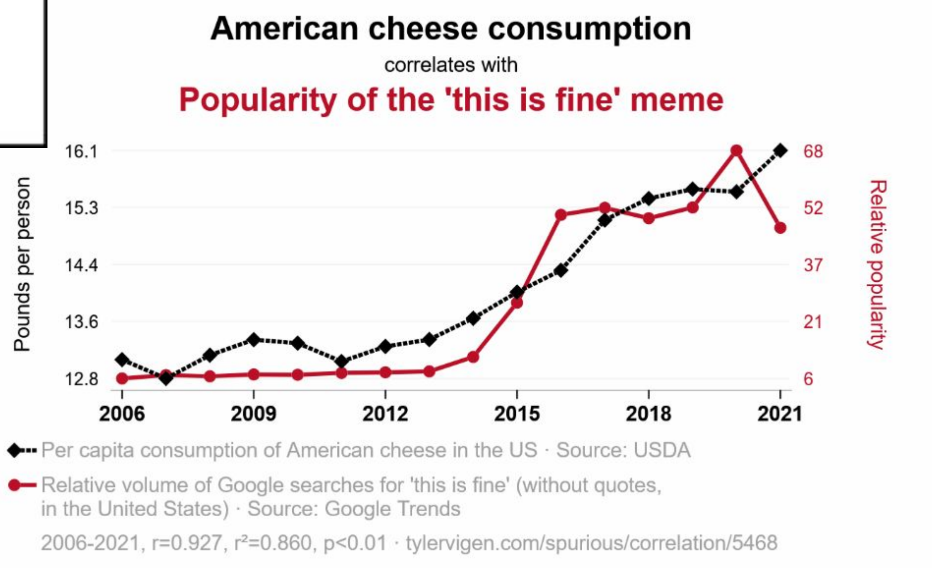
Causation vs. Correlation



Example: “Use of Causal Language” in the author guidelines of JAMA (Journal of American Medical Association):

“Causal language (including use of terms such as effect and efficacy) should be used only for randomised clinical trials. For all other study designs, methods and results should be described in terms of association or correlation and should avoid cause-and-effect wording.”

https://tylervigen.com/spurious-correlations



Causal Inference for Experiments

- In the lab
  - High internal validity
  - Might lack external validity

Logo
Social Psychological and Personality Science, Volume 4, Issue 5, September 2013, Pages 579-586

https://doi.org/10.1177/1948550612469233
Article

Does Posting Facebook Status Updates Increase or Decrease Loneliness? An Online Social Networking Experiment

Fenne große Deters[1] and Matthias R. Mehl[2]

Abstract

Online social networking is a pervasive but empirically understudied phenomenon. Strong public opinions on its consequences exist but are backed up by little empirical evidence and almost no causally conclusive, experimental research. The current study tested the psychological effects of posting status updates on Facebook using an experimental design. For 1 week, participants in the experimental condition were asked to post more than they usually do, whereas participants in the control condition received no instructions. Participants added a lab "Research Profile" as a Facebook friend allowing for the objective documentation of protocol compliance, participants' status updates, and friends' responses. Results revealed (1) that the experimentally induced increase in status updating activity reduced loneliness, (2) that the decrease in loneliness was due to participants feeling more connected to their friends on a daily basis, and (3) that the effect of posting on loneliness was independent of direct social feedback (i.e., responses) by friends.

Keywords
Facebook, loneliness, social inte

Does Posting Facebook Status Updates Increase or Decrease Loneliness? An Online Social Networking Experiment

Causal Inference for Experiments

- In the lab
  - High internal validity
  - Might lack external validity
- **In the “wild” or field experiments**

Polit Behav (2017) 39:629–649
DOI 10.1007/s11109-016-9373-5

CrossMark

ORIGINAL PAPER

Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment

Kevin Munger[1]

Published online: 11 November 2016

Abstract I conduct an experiment which examines the impact of group norm promotion and social sanctioning on racist online harassment. Racist online harassment de-mobilizes the minorities it targets, and the open, unopposed expression of racism in a public forum can legitimize racist viewpoints and prime ethnocentrism. I employ an intervention designed to reduce the use of anti-black racist slurs by white men on Twitter. I collect a sample of Twitter users who have harassed other users and use accounts I control (“bots”) to sanction the harassers. By varying the identity of the bots between in-group (white man) and out-group (black man) and by varying the number of Twitter followers each bot has, I find that subjects who were sanctioned by a high-follower white male significantly reduced their use of a racist slur. This paper extends findings from lab experiments to a naturalistic setting using an objective, behavioral outcome measure and a continuous 2-month data collection period. This represents an ac

[Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment]

Causal Inference for Experiments

- In the lab
  - High internal validity
  - Might lack external validity
- **In the “wild” or field experiments**

[Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment]

Causal Inference for Experiments

- In the lab
  - High internal validity
  - Might lack external validity
- **In the “wild” or field experiments**

Polit Behav (2017) 39:629–649
DOI 10.1007/s11109-016-9373-5

CrossMark

ORIGINAL PAPER

informs.
https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 70, No. 5, May 2024, pp. 3264–3280
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Motivating Experts to Contribute to Digital Public Goods: A Personalized Field Experiment on Wikipedia

Yan Chen,[a,b,*] Rosta Farzan,[c] Robert Kraut,[d] Iman YeckehZaare,[a] Ark Fangzhou Zhang[e]

[a] School of Information, University of Michigan, Ann Arbor, Michigan 48109; [b] Department of Economics, School of Economics and Management, Tsinghua University, Beijing 100084, China; [c] School of Computing and Information, University of Pittsburgh, Pittsburgh, Pennsylvania 15260; [d] School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213; [e] Google LLC, Mountain View, California 94043
*Corresponding author
Contact: yanchen@umich.edu, https://orcid.org/0000-0003-1287-4525 (YC); rfarzan@pitt.edu, https://orcid.org/0000-0003-0391-8354 (RF); robert.kraut@cmu.edu, https://orcid.org/0000-0002-8554-2137 (RK); oneweb@umich.edu, https://orcid.org/0000-0002-1788-2922 (IYZ); arkzhang@umich.edu, https://orcid.org/0000-0002-3282-5811 (AFZ)

Received: April 26, 2020
Revised: September 10, 2021; June 8, 2022
Accepted: July 11, 2022
Published Online in Articles in Advance: December 4, 2023

https://doi.org/10.1287/mnsc.2023.4852

Copyright: © 2023 The Author(s)

Abstract. We conducted a large-scale personalized field experiment to examine how match quality, recognition, and social impact influence domain experts' contributions to Wikipedia. Forty-five percent of the experts expressed willingness to contribute in the baseline condition, whereas 51% (a 13% increase over the baseline) expressed interest when they received a signal that an article matched their expertise. However, none of the treatments had a significant effect on actual contributions. Instead experts contributed longer and better comments when the actual match between a recommended Wikipedia article and an expert's expertise, measured by cosine similarity, was higher, when they had higher reputation, and when the original article was longer. These findings suggest that match quality between volunteers and tasks is critically important in encouraging contributions to digital public goods and likely to volunteering in general.

History: Accepted by David Simchi-Levi, behavioral economics and decision analysis.
Open Access Statement: This work is licensed under a Creative Commons Attribution 4.0 International License. You are free to copy, distribute, transmit and adapt this work, but you must attribute this work as “Management Science. Copyright © 2023 The Author(s). https://doi.org/10.1287/mnsc.2023.4852, used under a Creative Commons Attribution License: https://creativecommons.org/licenses/by/4.0/.”
Funding: This work was supported by the National Science Foundation through [Grant SES-1620319] awarded to Carnegie Mellon University and the University of Michigan.
Supplemental Material: The data files and online appendix are available at https://doi.org/10.1287/mnsc.2023.4852.

Keywords: digital public goods • match quality • machine learning • field experiment

1. Introduction

with both informational and emotional support (Wang et al. 2012).

In each of these cases, the peer-produced digital public goods have distinct characteristics. They are informa-... goods with free and open access to the general public. Second, these public goods are contributor dependent in the sense that matching to the right expert

tributions of labor and resources. Dedicated to the provision of free information, the Wikipedia community has developed the largest and most read encyclo-

[Motivating Experts to Contribute to Digital Public Goods:A Personalized Field Experiment on Wikipedia](#)

Causal Inference for Experiments

- In the lab
  - High internal validity
  - Might lack external validity
- In the “wild” or field experiments
- **Experiments are not always possible**
  - Ethical implications

PERSPECTIVE

Ethics in field experimentation: A call to establish new standards to protect the public from unwanted manipulation and real harms

Rose McDermott[a] and Peter K. Hatemi[b,c,1]

Edited by Margaret Levi, Stanford University, Stanford, CA, and approved October 1, 2020 (received for review June 10, 2020)

In 1966, Henry Beecher published his foundational paper “Ethics and Clinical Research,” bringing to light unethical experiments that were routinely being conducted by leading universities and government agencies. A common theme was the lack of voluntary consent. Research regulations surrounding laboratory experiments flourished after his work. More than half a century later, we seek to follow in his footsteps and identify a new domain of risk to the public: certain types of field experiments. The nature of experimental research has changed greatly since the Belmont Report. Due in part to technological advances including social media, experimenters now target and affect whole societies, releasing interventions into a living public, often without sufficient review or controls. A large number of social science field experiments do not reflect compliance with current ethical and legal requirements that govern research with human participants. Real-world interventions are being conducted without consent or notice to the public they affect. Follow-ups and debriefing are routinely not being undertaken with the populations that experimenters injure. Importantly, even when ethical research guidelines are followed, researchers are following principles developed for experiments in controlled settings, with little assessment or protection for the wider societies within which individuals are embedded. We strive to improve the ethics of future work by advocating the creation of new norms, illustrating classes of field experiments where scholars do not appear to have recognized the ways such research circumvents ethical standards by putting people, including those outside the manipulated group, into harm’s way.

[Ethics in field experimentation: A call to establish new standards to protect the public from unwanted manipulation and real harms]

Causal Inference with Observational Data

- Also called ‘quasi-experimental’ approaches

- Broad category of approaches
  - Controlling for confounders
    - Matching strategies, including propensity score matching
  - Instrumental variables
  - Regression discontinuity design
  - Negative control (e.g. difference in differences).

Controlling for Confounders in Observational Studies

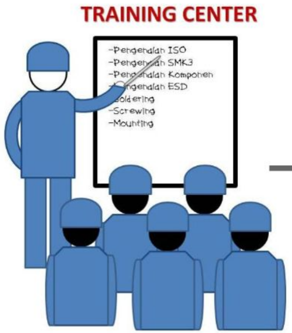
What is the average treatment effect of job training on participants' income in 1978?

Confounding Covariates

- Age
- Education
- Black
- Hispanic
- Married
- No Degree
- Income 1974
- Income 1975

Selected Intervention

(Non-Random Treatment Assignment)



Participants in the job training program earned on average $635 **less** in 1978 than those who did not participate.

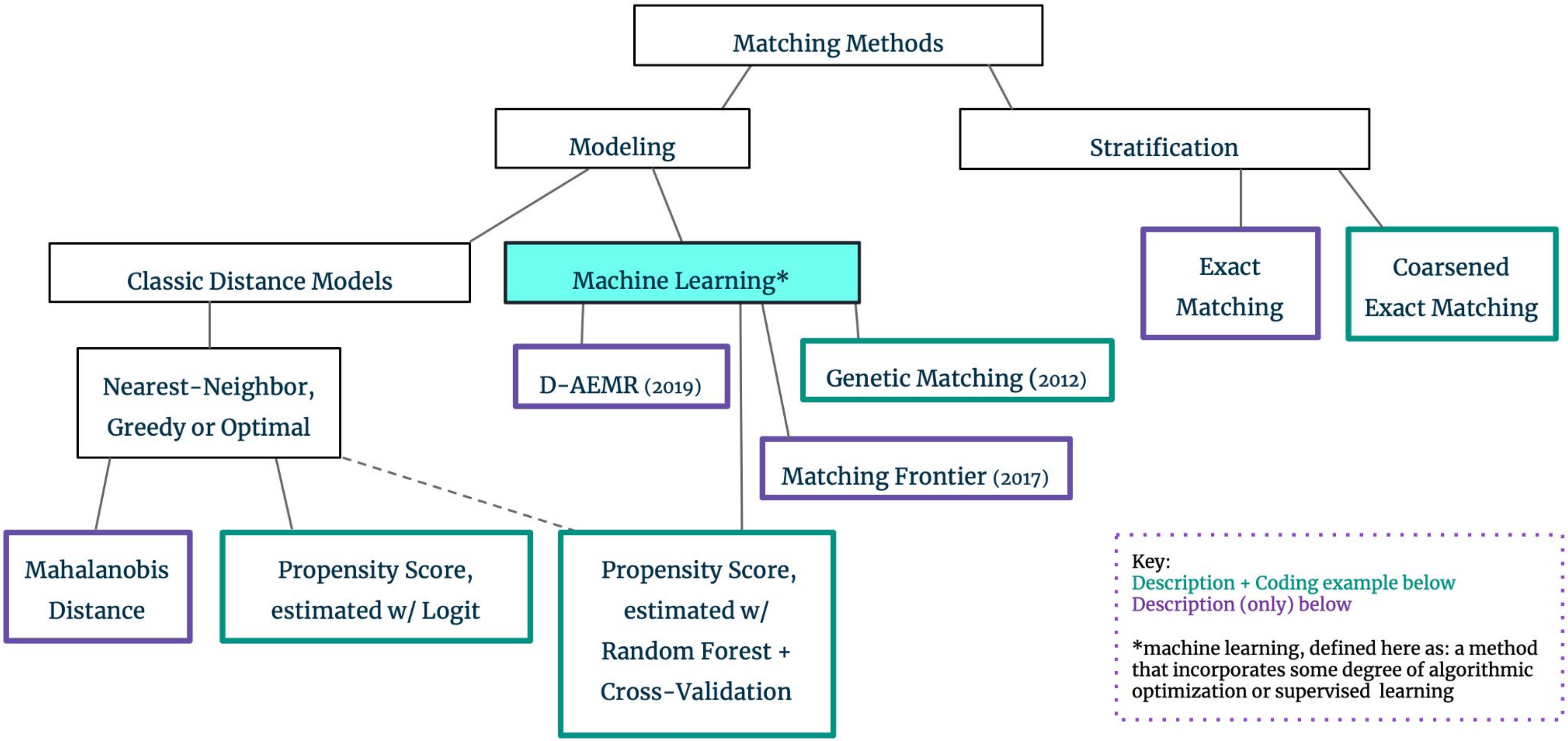
Average Treatment Effect on the Treated + Selection Bias

Controlling for Confounders: Matching

***Matching**, def: any method that **strategically subsamples** dataset to **balance covariate distribution** in treated and control groups such that after matching both groups share an equal probability of treatment.*

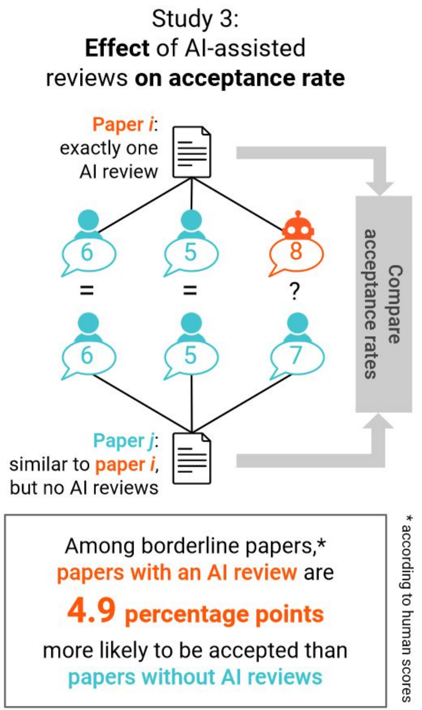
Non-Random Treatment Assignment $\xrightarrow[\text{to Subsample}]{\text{Matching Methods}}$ ***Average Treatment Effect on the Treated** + ~~**Selection Bias**~~*

https://humboldt-wi.github.io/blog/research/applied_predictive_modeling_19/matching_methods/



https://humboldt-wi.github.io/blog/research/applied_predictive_modeling_19/matching_methods/

Controlling for Confounders: Matching



The AI Review Lottery: Widespread AI-Assisted Peer Reviews Boost Paper Scores and Acceptance Rates

Giuseppe Russo Latona,* Manoel Horta Ribeiro,† Tim R. Davidson,†
Veniamin Veselovsky,† Robert West*
EPFL

Journals and conferences worry that peer reviews assisted by artificial intelligence (AI), in particular, large language models (LLMs), may negatively influence the validity and fairness of the peer-review system, a cornerstone of modern science. In this work, we address this concern with a quasi-experimental study of the prevalence and impact of AI-assisted peer reviews in the context of the 2024 International Conference on Learning Representations (ICLR), a large and prestigious machine-learning conference. Our contributions are threefold. Firstly, we obtain a lower bound for the prevalence of AI-assisted reviews at ICLR 2024 using the GPTZero LLM detector, estimating that at least 15.8% of reviews were written with AI assistance. Secondly, we estimate the impact of AI-assisted reviews on submission scores. Considering pairs of reviews with different scores assigned to the same paper, we find that in 53.4% of pairs the AI-assisted review scores higher than the human review ($p = 0.002$; relative difference in probability of scoring higher: +14.4% in favor of AI-assisted reviews). Thirdly, we assess the impact of receiving an AI-assisted peer review on submission acceptance. In a matched study, submissions near the acceptance threshold that received an AI-assisted peer review were 4.9 percentage points ($p = 0.024$) more likely to be accepted than submissions that did not. Overall, we show that AI-assisted reviews are consequential to the peer-review process and offer a discussion on future implications of current trends.

Peer review is central to the modern scientific process and the current epistemic and social status of science [12, 43, 50]. The system is used by journals and conferences to ensure the validity and significance of research findings [47, 48] and by funding institutions to allocate grants [32, 33, 49]. Society treats peer-reviewe … peer-reviewed rese…

submissions being incorrectly judged [14]. Scientists' reliance on LLMs to write peer reviews (hereinafter called AI-assisted reviews) could thus decrease the peer-review system's reliability and harm its social and epistemic functions [19]. In response, multiple journals and conferences have already felt obliged … n the peer- … [9, 20, 30, 42].

[The AI Review Lottery: Widespread AI-Assisted Peer Reviews Boost Paper Scores and Acceptance Rates]

Propensity score matching is quite popular in text + CSS

Paper	Treatment	Outcome(s)	Confounder	Text data	Text rep.	Adjustment method
Johansson et al. (2016)	Viewing device (mobile or desktop)	Reader's experience	News content	News	Word counts	Causal-driven rep. learning
De Choudhury et al. (2016)	Word use in mental health community	User transitions to post in suicide community	Previous text written in a forum	Social media (Reddit)	Word counts	Stratified propensity score matching
De Choudhury and Kiciman (2017)	Language of comments	User transitions to post in suicide community	User's previous posts and comments received	Social media (Reddit)	Unigrams and bigrams	Stratified propensity score matching
Falavarjani et al. (2017)	Exercise (Foursquare checkins)	Shift in topical interest on Twitter	Pre-treatment topical interest shift	Social media (Twitter, Foursquare)	Topic models	Matching
Olteanu et al. (2017)	Current word use	Future word use	Past word use	Social media (Twitter)	Top unigrams and bigrams	Stratified propensity score matching
Pham and Shen (2017)	Group vs. individual loan requests	Time until borrowers get funded	Loan description	Microloans (Kiva)	Pre-trained embeddings + neural networks	A-IPTW, TMLE
Kiciman et al. (2018)	Alcohol mentions	College success (e.g. study habits, risky behaviors, emotions)	Previous posts	Social media (Twitter)	Word counts	Stratified propensity score matching
Sridhar et al. (2018)	Exercise	Mood	Mood triggers	Users' text on mood logging apps	Word counts	Propensity score matching
Saha et al. (2019)	Self-reported usage of psychiatric medication	Mood, cognition, depression, anxiety, psychosis, and suicidal ideation	Users' previous posts	Social media (Twitter)	Word counts + lexicons + supervised classifiers	Stratified propensity score matching
Sridhar and Getoor (2019)	Tone of replies	Changes in sentiment	Speaker's political ideology	Debate transcripts	Topic models + lexicons	Regression adjustment, IPTW, A-IPTW
Veitch et al. (2019)	Presence of a theorem	Rate of acceptance	Subject of the article	Scientific articles	BERT	Causal-driven rep. learning + Regression adjustment, TMLE
Roberts et al. (2020)	Perceived gender of author	Number of citations	Content of article	International Relations articles	Topic models + propensity score	Coarsened exact matching
Roberts et al. (2020)	Censorship	Subsequent censorship and posting rate	Content of posts	Social media (Weibo)	Topic models + propensity score	Coarsened exact matching

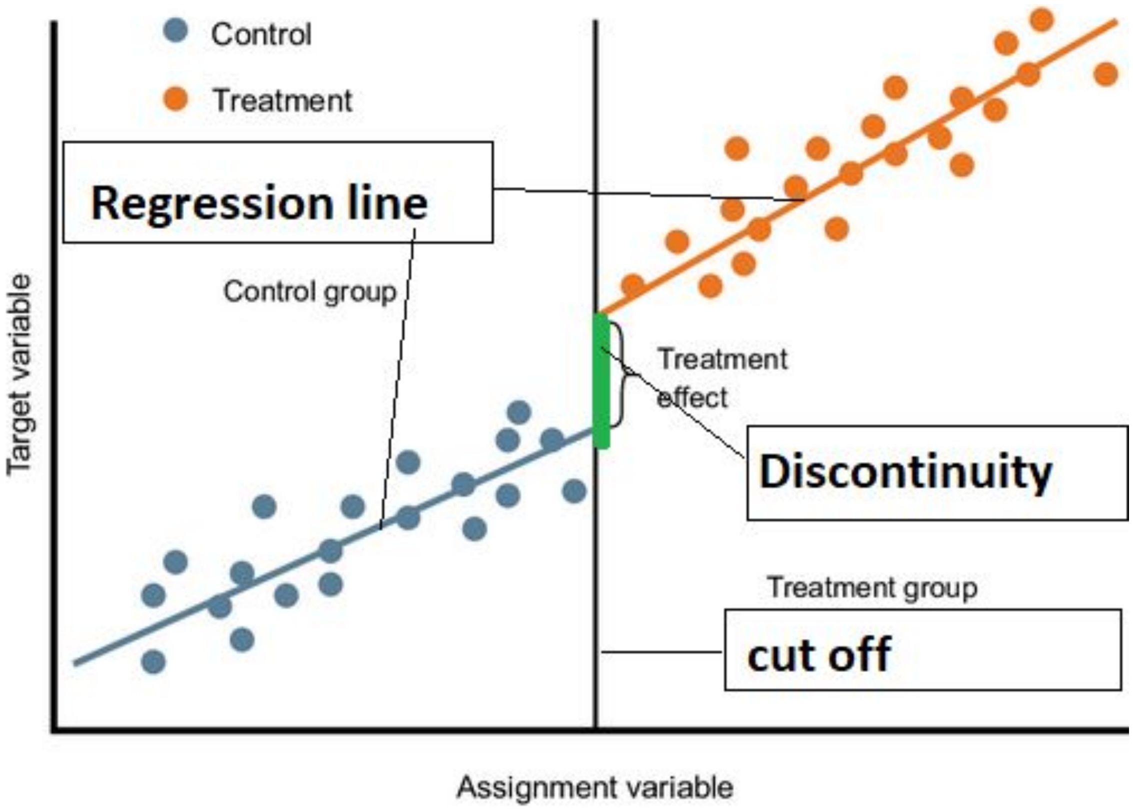
Table 1: Example applications that infer the causal effects of treatment on outcome by measuring confounders (unobserved) from text data (observed). In doing so, these applications choose a representation of text (text rep.) and a method to adjust for confounding.

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

Regression Discontinuity

“pretest-posttest design that aims to determine the causal effects of interventions by assigning a cutoff or threshold above or below which an intervention is assigned.”

https://en.wikipedia.org/wiki/Regression_discontinuity_design



Regression Discontinuity

“pretest–posttest design that aims to determine the causal effects of interventions by assigning a cutoff or threshold above or below which an intervention is assigned.”

https://en.wikipedia.org/wiki/Regression_discontinuity_design

[Identifying Platform Effects in Social Media Data]()

Unofficial corrected version. Published version appears in *Proceedings of the Tenth International AAAI Conference on Web and Social Media* (ICWSM 2016), pp. 241–249, available at http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13163.

Identifying Platform Effects in Social Media Data

Momin M. Malik[1] and **Jürgen Pfeffer**[1,2]
[1]Institute for Software Research
School of Computer Science
Carnegie Mellon University
[2]Bavarian School of Public Policy
Technical University of Munich

Abstract

Even when external researchers have access to social media data, they are not privy to decisions that went into platform design—including the measurement and testing that goes into deploying new platform features, such as recommender systems, that seek to shape user behavior towards desirable ends. Finding ways to identify platform effects is thus important both for generalizing findings, as well as understanding the nature of platform usage. One approach is to find temporal data covering the introduction of a new feature; observing differences in behavior before and after allow us to estimate the effect of the change. We investigate platform effects using two such datasets, the Netflix Prize dataset and the Facebook New Orleans data, in which we observe seeming discontinuities in user behavior but that we know or suspect are the result of a change in platform design. For the Netflix Prize, we estimate user ratings changing by an average of about 3% after the change, and in Facebook New Orleans, we find that the introduction of the ‘People You May Know’ feature locally

non-embedded researchers having access to the data (Savage and Burrows 2007; Lazer et al. 2009; Huberman 2012; boyd and Crawford 2012), but also that even when researchers have access, without full knowledge of the platform engineering and the decisions and internal research that went into design decisions, the data can be systematically misleading.

One way to study and quantify platform effects as an external researcher is to look for available data that include a significant platform change. Making the assumption that, in absence of the exogenous shock (the change) the previous ‘trend’ would have remained the same, we can apply the observational inference method of *regression discontinuity design* (Imbens and Lemieux 2008; Lee and Lemieux 2010; Li 2013). While not as certain as experimental design, observational inference methods are the best available way for outside researchers to understand the effects of platform design.

Regression Discontinuity

"pretest–posttest design that aims to determine the causal effects of interventions by assigning a cutoff or threshold above or below which an intervention is assigned."

https://en.wikipedia.org/wiki/Regression_discontinuity_design

Identifying Platform Effects in Social Media Data

Unofficial corrected version. Published version appears in *Proceedings of the Tenth International AAAI Conference on Web and Social Media* (ICWSM 2016), pp. 241–249, available at http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13163.

Identifying Platform Effects in Social Media Data

Momin M. Malik[1] and **Jürgen Pfeffer**[1,2]
[1]Institute for Software Research
School of Computer Science

Even when ext
data, they are
design—incluc
deploying new
tems, that seek
Finding ways
both for gener
nature of platf
data covering
differences in
the effect of the
two such datas
New Orleans d
ities in user be
sult of a chang
estimate user r
ter the change,
introduction of

ers having access to the data (Sav-
Lazer et al. 2009; Huberman 2012;
)12), but also that even when re-
without full knowledge of the plat-
e decisions and internal research that
ons, the data can be systematically

l quantify platform effects as an ex-
ook for available data that include
hange. Making the assumption that,
nous shock (the change) the previ-
emained the same, we can apply the
method of *regression discontinuity*
nieux 2008; Lee and Lemieux 2010;
certain as experimental design, ob-
thods are the best available way for
nderstand the effects of platform de-

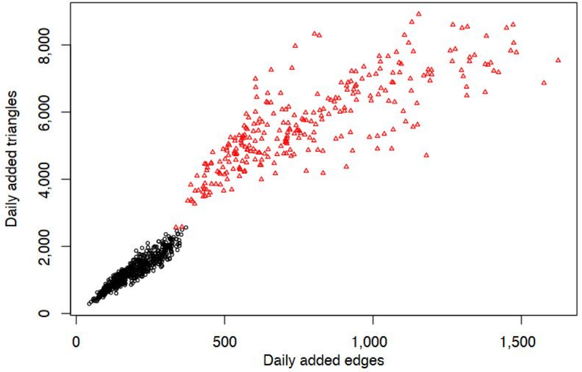


Figure 7: The daily added edges and triangles have a close relationship in the Facebook data. Black circles are time points before 2008-03-26, and red triangles are time points afterwards.

Differences in Differences?

Instrumental Variables?

Comparison of different approaches

Approach	Key assumptions	Strengths	Limitations
Regression-based approach[1]	Error term is uncorrelated with all regressors	Easy to implement. Well-developed literature on mediating/moderating and nested models.	No clear distinction between treatment and covariates.
Propensity score approach	Strong ignorability (i.e., selection on observables only)	Estimates causal effect at a given time. Explicit consideration of all variables that relate to treatment assignment. Sensitivity analysis for violation of strong ignorability assumption.	Diagnostics for adequacy of propensity score model and methods for estimating mediating/moderating /nested effects are still in early stages. Requires large sample sizes.
RD approach	Strong ignorability	Allows estimation of treatment effect if treatment assignment changes discontinuously on the basis of some Z.	Requires making some assumptions and extrapolations for control units in the range of Z, where we do not have any control units, and vice versa.
Dummy endogenous variable approach	Error terms of selection and outcome equations are linearly related and bivariate normal	Allows error terms of selection and outcome equations to be correlated.	Linearity and bivariate normality assumptions are not testable. Both researchers and managers have difficulty in conceptualizing and understanding these assumptions.
IV approach	Exclusion restriction[2]	Allows estimation of causal effects when treatment variable is endogenous.	Exclusion restrictions are not testable and rarely justifiable. Large standard errors if sample sizes are small or instruments are weak. Assumes a constant treatment effect for all individuals.

From Association to Causation Via a Potential Outcomes Approach

Resources

- Book: “Causal Inference: What If” Miguel A. Hernan, James M. Robins. Available online for free: https://www.hsph.harvard.edu/miguel-hernan/wp-content/uploads/sites/1268/2024/04/hernanrobins_WhatIf_26apr24.pdf
- Overview of causal methods for text:
  - Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond
  - https://github.com/causaltext/causal-text-papers
- Matching: https://humboldt-wi.github.io/blog/research/applied_predictive_modeling_19/matching_methods/
- Propensity scores intro: https://osf.io/preprints/socarxiv/ncvqs

Try Yourself:

- Propensity score matching: https://github.com/konosp/propensity-score-matching/blob/main/propensity_score_matching_v2.ipynb
- Causal inference and text, tutorial by Dhanya Sridhar:
  - https://nlp-css-201-tutorials.github.io/nlp-css-201-tutorials/docs/sridhar_umsi_nlp_tutorial.pdf
  - https://colab.research.google.com/drive/101rhkpnQInEkyPysdEmZhF2oZvEavmi6?usp=sharing#scrollTo=JFRl8xDNTqXo
- DoWhy: A python package for causal inference: https://github.com/py-why/dowhy

Next week(s)

- 17.06: Reproducible research pipeline + Ethics
- 24.06: First Project Guidance Session (in-person, same room)
- 27.06: Deadline for submitting project proposal
- 1.07: Project Guidance and Discussion
- 8.07: Project Guidance and Discussion
- 15.07: Project Guidance and Discussion
- 22.07 1:30-3:00 PM [online]: Midway Presentations
- 29.07: Project Guidance and Discussion
- 09.08 10-12 AM [online]: Final Presentations
- 26.08: Final Reports Due