SPIA 586a: Machine Learning for Policy Analysis
Week 9: Found Data

Brandon Stewart[1]

Princeton

March 24–26, 2025

[1]Many thanks to Ian Lundberg and the team of Dean Knox, Will Lowe and Jonathan Mummolo for slides on KLM.

Where We've Been and Where We're Going...

Where We've Been and Where We're Going...

- Last Week
  - designed data

Where We've Been and Where We're Going...

- Last Week
  - designed data
- This Week
  - found data

Where We've Been and Where We're Going...

- Last Week
  - designed data
- This Week
  - found data
- Next Week
  - privacy

Where We've Been and Where We're Going...

- Last Week
  - designed data
- This Week
  - found data
- Next Week
  - privacy
- Long Run
  - target tasks → data sources → guiding values

Admin

- Goals Requests:

Admin

- Goals Requests:
  - have a sense of where to go next
  - understand the full model workflow
  - understand data pipelines

Admin

- Goals Requests:
  - have a sense of where to go next
  - understand the full model workflow
  - understand data pipelines
- Today: (Brief) Introduction $\rightsquigarrow$ Case Studies and Code in Response to Interest

Three Data Archetypes

- Experimental Data
  we intervene, measure and analyze
- Designed Data
  we measure and analyze
- Found Data
  we analyze

What Do We Call This?

What Do We Call This?

- A lot of people talk about big data which typically is a catchall for large digital trace, administrative and commercial data.

What Do We Call This?

- A lot of people talk about big data which typically is a catchall for large digital trace, administrative and commercial data.
- Salganik likes talking about data as repurposed rather than found/designed to emphasizes that what we typically think of as "found" data is often designed by someone.

What Do We Call This?

- A lot of people talk about big data which typically is a catchall for large digital trace, administrative and commercial data.
- Salganik likes talking about data as repurposed rather than found/designed to emphasizes that what we typically think of as "found" data is often designed by someone.
- Key distinctions are:
  1) does the data measure what you want?

What Do We Call This?

- A lot of people talk about big data which typically is a catchall for large digital trace, administrative and commercial data.
- Salganik likes talking about data as repurposed rather than found/designed to emphasizes that what we typically think of as "found" data is often designed by someone.
- Key distinctions are:
    1) does the data measure what you want?
    2) do you understand how it was collected?

What Do We Call This?

- A lot of people talk about big data which typically is a catchall for large digital trace, administrative and commercial data.
- Salganik likes talking about data as repurposed rather than found/designed to emphasizes that what we typically think of as "found" data is often designed by someone.
- Key distinctions are:
  1) does the data measure what you want?
  2) do you understand how it was collected?
- Record linkage can be an important tool in bringing data together.

What Do We Call This?

- A lot of people talk about big data which typically is a catchall for large digital trace, administrative and commercial data.
- Salganik likes talking about data as repurposed rather than found/designed to emphasizes that what we typically think of as "found" data is often designed by someone.
- Key distinctions are:
  1) does the data measure what you want?
  2) do you understand how it was collected?
- Record linkage can be an important tool in bringing data together.
- Broadly speaking this week is about how to think about not getting exactly what you wanted.

Bigger Data $\neq$ Better Data

Salganik's 10 Characteristics

Salganik's 10 Characteristics
(1) Big — there is a lot of it!

Salganik's 10 Characteristics

(1) Big — there is a lot of it!

- Good for:

Salganik's 10 Characteristics
(1) Big — there is a lot of it!

- Good for: rare events,

Salganik's 10 Characteristics
(1) Big — there is a lot of it!

- Good for: rare events, heterogeneity,

Salganik's 10 Characteristics
(1) Big — there is a lot of it!

- Good for: rare events, heterogeneity, small differences

Salganik's 10 Characteristics
(1) Big — there is a lot of it!

- Good for: rare events, heterogeneity, small differences
- Bad for: carefully examining our data.

Salganik's 10 Characteristics

(1) Big — there is a lot of it!

- Good for: rare events, heterogeneity, small differences
- Bad for: carefully examining our data.

(2) Always-on — constant collection

Salganik's 10 Characteristics

(1) Big — there is a lot of it!

- Good for: rare events, heterogeneity, small differences
- Bad for: carefully examining our data.

(2) Always-on — constant collection

- Good for: retrospective data collection.

Salganik's 10 Characteristics

(1) Big — there is a lot of it!

- Good for: rare events, heterogeneity, small differences
- Bad for: carefully examining our data.

(2) Always-on — constant collection

- Good for: retrospective data collection.

(3) Nonreactive — less observer effects?

Salganik's 10 Characteristics

(1) Big — there is a lot of it!

- Good for: rare events, heterogeneity, small differences
- Bad for: carefully examining our data.

(2) Always-on — constant collection

- Good for: retrospective data collection.

(3) Nonreactive — less observer effects?

- some data collection is so passive that it doesn't have the same kind of observer effects.

Salganik's 10 Characteristics

(1) Big — there is a lot of it!

- Good for: rare events, heterogeneity, small differences
- Bad for: carefully examining our data.

(2) Always-on — constant collection

- Good for: retrospective data collection.

(3) Nonreactive — less observer effects?

- some data collection is so passive that it doesn't have the same kind of observer effects.
- e.g. Stephens-Davidowitz (2017) and google search data.

Salganik's 10 Characteristics

(1) Big — there is a lot of it!

- Good for: rare events, heterogeneity, small differences
- Bad for: carefully examining our data.

(2) Always-on — constant collection

- Good for: retrospective data collection.

(3) Nonreactive — less observer effects?

- some data collection is so passive that it doesn't have the same kind of observer effects.
- e.g. Stephens-Davidowitz (2017) and google search data.
- in my experience nonreactive data often comes with difficult construct validity.

Salganik's 10 Characteristics

Salganik's 10 Characteristics
(4) Incomplete

Salganik's 10 Characteristics
(4) Incomplete

- three types: participant demographics, behavior on other platforms, ways to operationalize theoretical concepts

Salganik's 10 Characteristics
(4) Incomplete

- three types: participant demographics, behavior on other platforms, ways to operationalize theoretical concepts
- construct validity concerns

Salganik's 10 Characteristics

(4) Incomplete

- three types: participant demographics, behavior on other platforms, ways to operationalize theoretical concepts
- construct validity concerns

(5) Inaccessible

Salganik's 10 Characteristics

(4) Incomplete

- three types: participant demographics, behavior on other platforms, ways to operationalize theoretical concepts
- construct validity concerns

(5) Inaccessible

- data in governments, companies etc.

Salganik's 10 Characteristics
(4) Incomplete

- three types: participant demographics, behavior on other platforms, ways to operationalize theoretical concepts
- construct validity concerns

(5) Inaccessible

- data in governments, companies etc.
- ethical and self-interest reasons

Salganik's 10 Characteristics

(4) Incomplete

- three types: participant demographics, behavior on other platforms, ways to operationalize theoretical concepts
- construct validity concerns

(5) Inaccessible

- data in governments, companies etc.
- ethical and self-interest reasons
- when available, huge challenges to open science

Salganik's 10 Characteristics

(4) Incomplete

- three types: participant demographics, behavior on other platforms, ways to operationalize theoretical concepts
- construct validity concerns

(5) Inaccessible

- data in governments, companies etc.
- ethical and self-interest reasons
- when available, huge challenges to open science

(6) Nonrepresentative

Salganik's 10 Characteristics

(4) Incomplete

- three types: participant demographics, behavior on other platforms, ways to operationalize theoretical concepts
- construct validity concerns

(5) Inaccessible

- data in governments, companies etc.
- ethical and self-interest reasons
- when available, huge challenges to open science

(6) Nonrepresentative

- what is it a sample of?

Salganik's 10 Characteristics

(4) Incomplete

- three types: participant demographics, behavior on other platforms, ways to operationalize theoretical concepts
- construct validity concerns

(5) Inaccessible

- data in governments, companies etc.
- ethical and self-interest reasons
- when available, huge challenges to open science

(6) Nonrepresentative

- what is it a sample of?
- risk of studying what is available instead of what is interesting.

Salganik's 10 Characteristics

Salganik's 10 Characteristics

(7) Drifting

Salganik's 10 Characteristics

(7) Drifting
(8) Algorithmically Confounded

Salganik's 10 Characteristics

(7) Drifting

(8) Algorithmically Confounded

(9) Dirty

Salganik's 10 Characteristics

(7) Drifting
(8) Algorithmically Confounded
(9) Dirty
(10) Sensitive

Research Strategies

- Counting Things
  - even counting things can use machine learning (think King, Pan, Roberts on Chinese censorship!)

Research Strategies

- Counting Things
  - even counting things can use machine learning (think King, Pan, Roberts on Chinese censorship!)
  - description is great and there is lots of basic descriptive info we don't have!

Research Strategies

- Counting Things
  - even counting things can use machine learning (think King, Pan, Roberts on Chinese censorship!)
  - description is great and there is lots of basic descriptive info we don't have!
- Forecasting and Nowcasting

Research Strategies

- Counting Things
  - even counting things can use machine learning (think King, Pan, Roberts on Chinese censorship!)
  - description is great and there is lots of basic descriptive info we don't have!
- Forecasting and Nowcasting
  - Google Flu but also cases like Blumenstock et al.

Research Strategies

- Counting Things
  - even counting things can use machine learning (think King, Pan, Roberts on Chinese censorship!)
  - description is great and there is lots of basic descriptive info we don't have!
- Forecasting and Nowcasting
  - Google Flu but also cases like Blumenstock et al.
- Approximating Experiments

Research Strategies

- Counting Things
  - even counting things can use machine learning (think King, Pan, Roberts on Chinese censorship!)
  - description is great and there is lots of basic descriptive info we don't have!
- Forecasting and Nowcasting
  - Google Flu but also cases like Blumenstock et al.
- Approximating Experiments
  - tools of causal inference we learned in Week 5

Research Strategies

- Counting Things
  - even counting things can use machine learning (think King, Pan, Roberts on Chinese censorship!)
  - description is great and there is lots of basic descriptive info we don't have!
- Forecasting and Nowcasting
  - Google Flu but also cases like Blumenstock et al.
- Approximating Experiments
  - tools of causal inference we learned in Week 5
  - natural experiments

The Economist Risk Estimator

- On March 11th, 2021 *The Economist* released their Covid-19 risk estimator. The goal is to provide estimates of hospitalization and death for U.S. patients if they contracted Covid on a particular day.

The Economist Risk Estimator

- On March 11th, 2021 *The Economist* released their Covid-19 risk estimator. The goal is to provide estimates of hospitalization and death for U.S. patients if they contracted Covid on a particular day.
- A key motivating factor here is heterogeneity—risk is very different across comorbidities!

The Economist Risk Estimator

- On March 11th, 2021 *The Economist* released their Covid-19 risk estimator. The goal is to provide estimates of hospitalization and death for U.S. patients if they contracted Covid on a particular day.
- A key motivating factor here is heterogeneity—risk is very different across comorbidities!
- Great writeup at: `https://www.economist.com/graphic-detail/2021/03/11/how-we-built-our-covid-19-risk-estimator` plus they provide code!

The Economist Risk Estimator

- On March 11th, 2021 *The Economist* released their Covid-19 risk estimator. The goal is to provide estimates of hospitalization and death for U.S. patients if they contracted Covid on a particular day.
- A key motivating factor here is heterogeneity—risk is very different across comorbidities!
- Great writeup at: `https://www.economist.com/graphic-detail/2021/03/11/how-we-built-our-covid-19-risk-estimator` plus they provide code!
- Uses many of the techniques we have discussed.

All images on following slides credit to the Economist Write-up

Data Collection Effort

- Covid-19 Research Database—group of healthcare companies pooling anonymised data on patients.

Data Collection Effort

- Covid-19 Research Database—group of healthcare companies pooling anonymised data on patients.
- 5 billion individual records with data for 250 million Americans.

Data Collection Effort

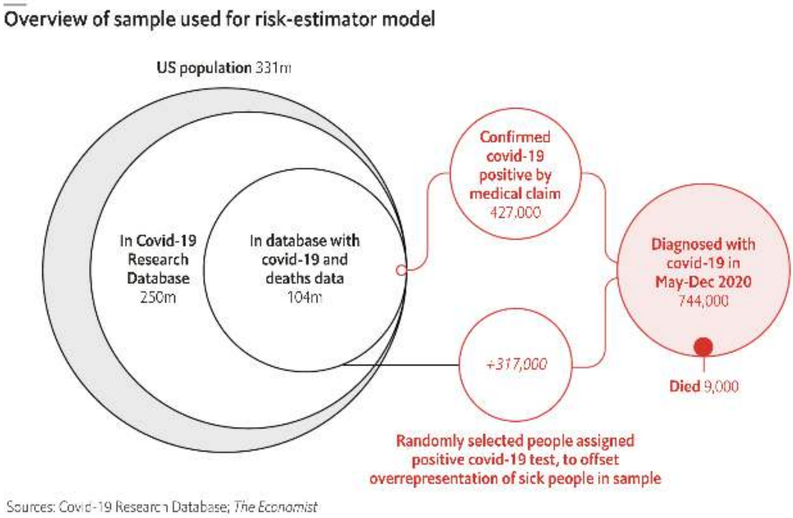
- Covid-19 Research Database—group of healthcare companies pooling anonymised data on patients.
- 5 billion individual records with data for 250 million Americans.
- 2.1 million people who have tested positive for covid.

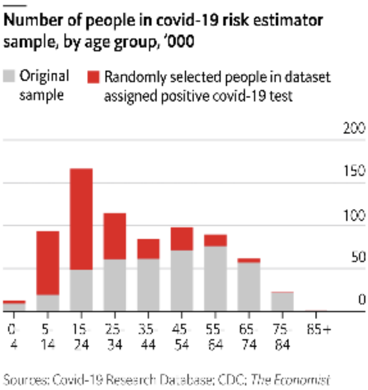
Data Collection Effort

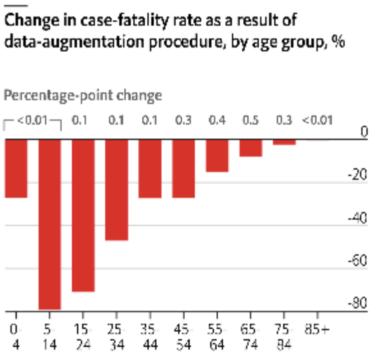
- Covid-19 Research Database—group of healthcare companies pooling anonymised data on patients.
- 5 billion individual records with data for 250 million Americans.
- 2.1 million people who have tested positive for covid.
- standardize variables to 30 different ailments and merged data to add age, gender, whether they were admitted to a hospital and whether they died.

Data Collection Effort

- Covid-19 Research Database—group of healthcare companies pooling anonymised data on patients.
- 5 billion individual records with data for 250 million Americans.
- 2.1 million people who have tested positive for covid.
- standardize variables to 30 different ailments and merged data to add age, gender, whether they were admitted to a hospital and whether they died.
- Aim: determine chances that someone diagnosed on December 1, 2020 would die by end of the year.

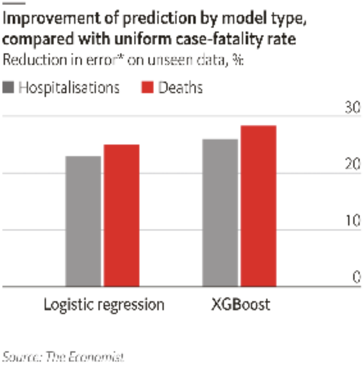

Overview of sample used for risk-estimator model
US population 331m
In Covid-19 Research Database 250m
In database with covid-19 and deaths data 104m
Confirmed covid-19 positive by medical claim 427,000
+317,000
Diagnosed with covid-19 in May-Dec 2020 744,000
Died 9,000
Randomly selected people assigned positive covid-19 test, to offset overrepresentation of sick people in sample
Sources: Covid-19 Research Database; The Economist

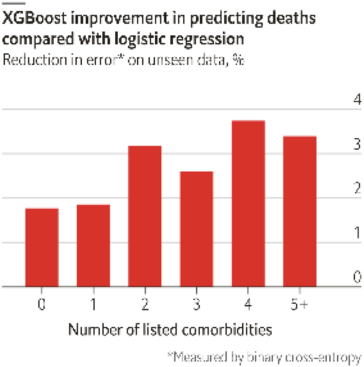

Number of people in covid-19 risk estimator sample, by age group, '000
Original sample
Randomly selected people in dataset assigned positive covid-19 test
200
150
100
50
0
0 4
5 14
15 24
25 34
35 44
45 54
55 64
65 74
75 84
85+
Sources: Covid-19 Research Database; CDC; The Economist

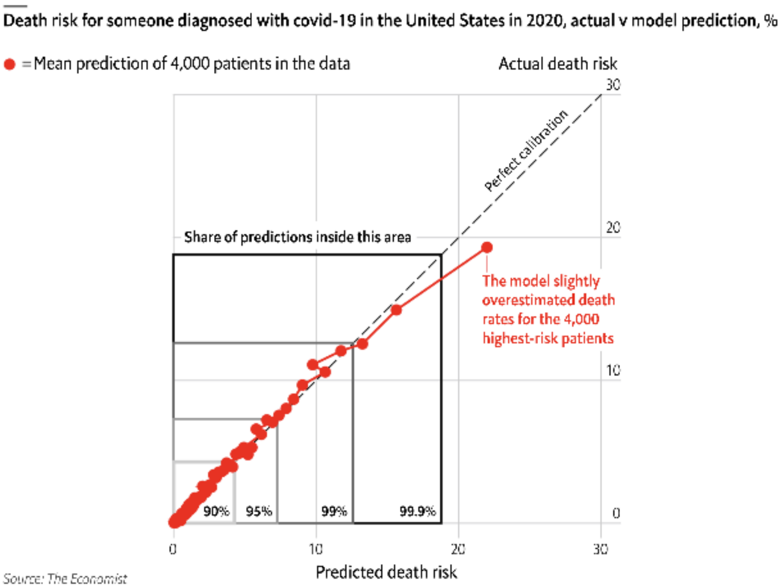

Change in case-fatality rate as a result of data-augmentation procedure, by age group, %
Percentage-point change
<0.01
0.1
0.1
0.1
0.3
0.4
0.5
0.3
<0.01
0
-20
-40
-60
-80
0 4
5 14
15 24
25 34
35 44
45 54
55 64
65 74
75 84
85+

and those involving large numbers of underlying conditions. In order to capture such complexity, we needed to allow for the possibility that comorbidities do not have constant effects that can simply be added together, but instead interact with each other, producing overall risk levels that are either higher or lower than the sum of their parts. As a result, we tested not just standard regression-based approaches such as such as logistic regression and proportional-hazards models, but also a range of machine-learning algorithms, ranging from the cerebral (dense neural networks) to the arboreal (random forests).

Of all of these methods, the one that performed best is called XGBoost, a common implementation of a family of algorithms called "gradient-boosted trees". This tool often ranks high on leaderboards in data-science competitions. Gradient-boosted trees


Improvement of prediction by model type, compared with uniform case-fatality rate
Reduction in error* on unseen data, %
Hospitalisations
Deaths
30
20
10
0
Logistic regression
XGBoost
Source: The Economist


XGBoost improvement in predicting deaths compared with logistic regression
Reduction in error* on unseen data, %
4
3
2
1
0
0
1
2
3
4
5+
Number of listed comorbidities
*Measured by binary cross-entropy

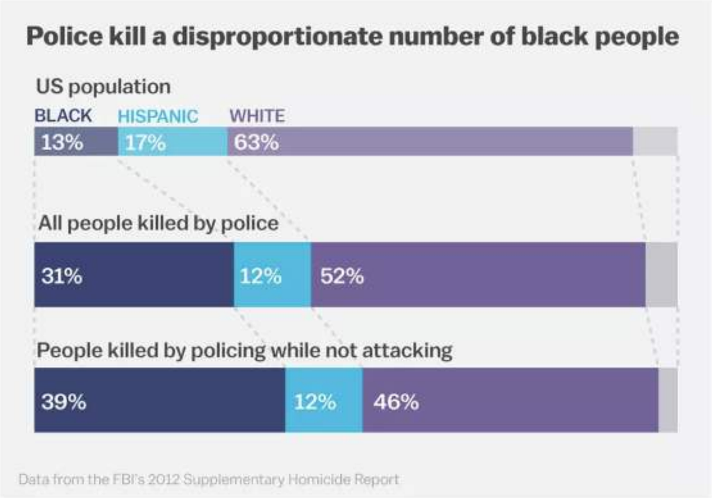


Death risk for someone diagnosed with covid-19 in the United States in 2020, actual v model prediction, %
= Mean prediction of 4,000 patients in the data
Actual death risk
30
20
10
0
Perfect calibration
Share of predictions inside this area
The model slightly overestimated death rates for the 4,000 highest-risk patients
90%
95%
99%
99.9%
0
10
20
30
Predicted death risk
Source: The Economist

Cleaning up after ourselves

Although gradient-boosted trees yielded the most accurate predictions overall, they do have a drawback. Because they do not impose assumptions about the relationships between variables and outcomes, they are just as likely to predict that adding a comorbidity or raising a person's age will decrease risk rather than increasing it, in the absence of evidence to the contrary. Sometimes, such counterintuitive behaviour can result from biases in the data, such as comorbidities potentially being more severe when they are the sole diagnosis than when they appear in conjunction with other illnesses. In other cases, it is the product of the model making poor guesses when presented with a type of patient that rarely or never appears in its training data.

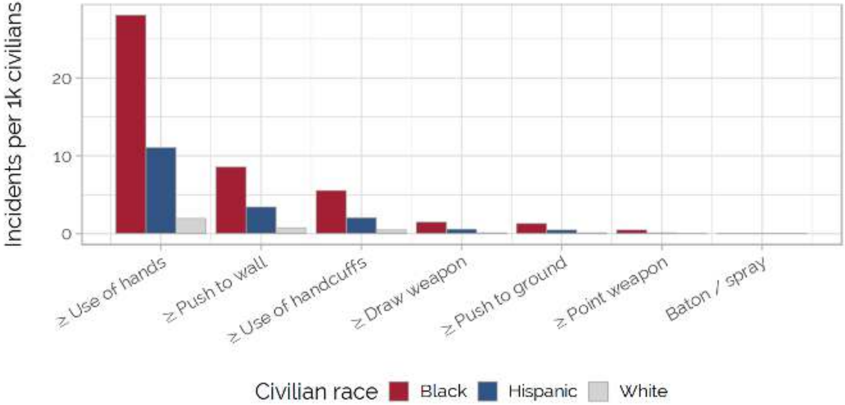
Police shootings per capita



Alvin Chang/Vox

Slide Credit: Knox, Lowe and Mummolo

NYPD Stop and Frisk, 2003-2013



Slide Credit: Knox, Lowe and Mummolo

Core Idea of KLM 2020

Core Idea of KLM 2020

- Administrative records are subject to extremely strong processes of selection and these cannot simply be shrugged off.

Core Idea of KLM 2020

- Administrative records are subject to extremely strong processes of selection and these cannot simply be shrugged off.
- Can we state a series of necessary assumptions and place bounds on quantities of interest?

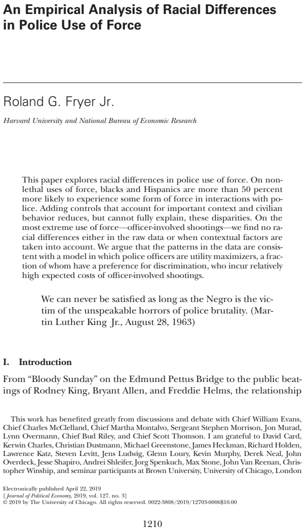
Core Idea of KLM 2020

- Administrative records are subject to extremely strong processes of selection and these cannot simply be shrugged off.
- Can we state a series of necessary assumptions and place bounds on quantities of interest?
- Much of the controversy here is around what is the quantity of interest?

Core Idea of KLM 2020

- Administrative records are subject to extremely strong processes of selection and these cannot simply be shrugged off.
- Can we state a series of necessary assumptions and place bounds on quantities of interest?
- Much of the controversy here is around what is the quantity of interest?
- Key Takeaway: Enormous challenges not in the methods but in the data.

It is the most surprising result of my career.
— Roland Fryer



An Empirical Analysis of Racial Differences in Police Use of Force

Roland G. Fryer Jr.

Harvard University and National Bureau of Economic Research

This paper explores racial differences in police use of force. On nonlethal uses of force, blacks and Hispanics are more than 50 percent more likely to experience some form of force in interactions with police. Adding controls that account for important context and civilian behavior reduces, but cannot fully explain, these disparities. On the most extreme use of force—officer-involved shootings—we find no racial differences either in the raw data or when contextual factors are taken into account. We argue that the patterns in the data are consistent with a model in which police officers are utility maximizers, a fraction of whom have a preference for discrimination, who incur relatively high expected costs of officer-involved shootings.

> We can never be satisfied as long as the Negro is the victim of the unspeakable horrors of police brutality. (Martin Luther King Jr., August 28, 1963)

I. Introduction

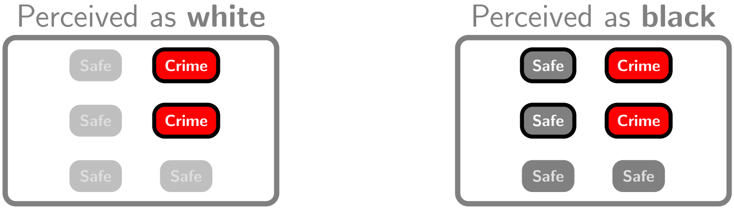
From “Bloody Sunday” on the Edmund Pettus Bridge to the public beatings of Rodney King, Bryant Allen, and Freddie Helms, the relationship

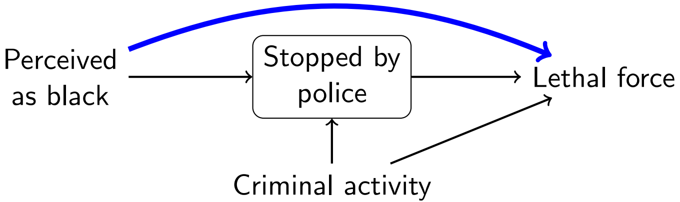
This work has benefited greatly from discussions and debate with Chief William Evans, Chief Charles McClelland, Chief Martha Montalvo, Sergeant Stephen Morrison, Jon Murad, Lynn Overmann, Chief Bud Riley, and Chief Scott Thomson. I am grateful to David Card, Kerwin Charles, Christian Dustmann, Michael Greenstone, James Heckman, Richard Holden, Lawrence Katz, Steven Levitt, Jens Ludwig, Glenn Loury, Kevin Murphy, Derek Neal, John Overdeck, Jesse Shapiro, Andrei Shleifer, Jorg Spenkuch, Max Stone, John Van Reenan, Christopher Winship, and seminar participants at Brown University, University of Chicago, London

Electronically published April 22, 2019
[*Journal of Political Economy*, 2019, vol. 127, no. 3]
© 2019 by The University of Chicago. All rights reserved. 0022-3808/2019/12703-0006$10.00

1210

Slide Credit: Ian Lundberg





Fryer 2019. Fuller critique by Knox et al. 2020 and Durlauf and Heckman 2020.

Slide Credit: Ian Lundberg

Knox, Lowe and Mummolo 2020

Knox, Lowe and Mummolo 2020

1. Racial bias likely affects who police choose to investigate $\rightarrow$ which encounters appear in police data

Knox, Lowe and Mummolo 2020

1. Racial bias likely affects who police choose to investigate $\rightarrow$ which encounters appear in police data

2. Police administrative data are inherently post-treatment

Knox, Lowe and Mummolo 2020

1. Racial bias likely affects who police choose to investigate $\rightarrow$ which encounters appear in police data

2. Police administrative data are inherently post-treatment

3. Results statistically biased; bias often can't be "controlled away"

Knox, Lowe and Mummolo 2020

1. Racial bias likely affects who police choose to investigate $\rightarrow$ which encounters appear in police data

2. Police administrative data are inherently post-treatment

3. Results statistically biased; bias often can't be "controlled away"

4. Bias has a precise form, can derive informative bounds on the true causal effect of civilian race on police behavior

Knox, Lowe and Mummolo 2020

1. Racial bias likely affects who police choose to investigate → which encounters appear in police data

2. Police administrative data are inherently post-treatment

3. Results statistically biased; bias often can't be "controlled away"

4. Bias has a precise form, can derive informative bounds on the true causal effect of civilian race on police behavior

5. Prior work ignoring this feature substantially underestimates racial bias in use of force (Fryer, 2019)

Knox, Lowe and Mummolo 2020

1. Racial bias likely affects who police choose to investigate $\rightarrow$ which encounters appear in police data
2. Police administrative data are inherently post-treatment
3. Results statistically biased; bias often can't be "controlled away"
4. Bias has a precise form, can derive informative bounds on the true causal effect of civilian race on police behavior
5. Prior work ignoring this feature substantially underestimates racial bias in use of force (Fryer, 2019)
6. New research designs to avoid this pitfall

Slide Credit: Knox, Lowe and Mummolo

A causal mediation framework

A causal mediation framework

- Unit of analysis: police-civilian encounters:

A causal mediation framework

- Unit of analysis: police-civilian encounters:
  - "Encounter" $=$ sighting of individual by police officer

A causal mediation framework

- Unit of analysis: police-civilian encounters:
  - "Encounter" = sighting of individual by police officer
  - Counterfactual: substitution of individual of differing race into police-civilian encounter, holding circumstance and civilian behavior fixed

A causal mediation framework

- Unit of analysis: police-civilian encounters:
  - "Encounter" = sighting of individual by police officer
  - Counterfactual: substitution of individual of differing race into police-civilian encounter, holding circumstance and civilian behavior fixed
- Treatment (civilian is racial minority) $T_i \in \{0,1\}$
- Outcome (use of force) $Y_i \in \{0,1\}$
- Mediator (being stopped by police) $M_i \in \{0,1\}$

A causal mediation framework

- Unit of analysis: police-civilian encounters:
  - "Encounter" = sighting of individual by police officer
  - Counterfactual: substitution of individual of differing race into police-civilian encounter, holding circumstance and civilian behavior fixed
- Treatment (civilian is racial minority) $T_i \in \{0,1\}$
- Outcome (use of force) $Y_i \in \{0,1\}$
- Mediator (being stopped by police) $M_i \in \{0,1\}$
- Racial bias in police stops ($T_i \rightarrow M_i$)
  (e.g. Gelman, Fagan & Kiss 2007; Glaser 2014; Lerman & Weaver 2014; Goel, Rao & Shroff 2016)

Slide Credit: Knox, Lowe and Mummolo

Potential outcomes with mediation

Potential outcomes with mediation

Normally we consider $Y(t \in \{Y(1), Y(0)\}$; potential force given race (treatment)

Instead consider $Y(t, M(t))$

Potential outcomes with mediation

Normally we consider $Y(t \in \{Y(1), Y(0)\}$; potential force given race (treatment)

Instead consider $Y(t, M(t))$

$Y(1,1) =$ potential use of force if minority civilian stopped
$Y(1,0) =$ potential use of force if minority civilian <u>not</u> stopped
$Y(0,1) =$ potential use of force if white civilian stopped
$Y(0,0) =$ potential use of force if white civilian <u>not</u> stopped

Slide Credit: Knox, Lowe and Mummolo

A (really bad) missing data problem

A (really bad) missing data problem

- Problem: Only encounters where a stop is made ($M_i = 1$) appear in police data

A (really bad) missing data problem

- Problem: Only encounters where a stop is made ($M_i = 1$) appear in police data

- By conditioning on $M_i = 1$, analysis is inherently post-treatment

A (really bad) missing data problem

- Problem: Only encounters where a stop is made ($M_i = 1$) appear in police data
- By conditioning on $M_i = 1$, analysis is inherently post-treatment
- Lots of work on post-treatment bias, but usually:
  - observe all levels of post-treatment variable (e.g. Acharya, Blackwell & Sen 2016)
  - can quantify scale of missing data (e.g. Nyhan, Skovron & Titiunik 2017 with voting age population)

A (really bad) missing data problem

- Problem: Only encounters where a stop is made ($M_i = 1$) appear in police data
- By conditioning on $M_i = 1$, analysis is inherently post-treatment
- Lots of work on post-treatment bias, but usually:
  - observe all levels of post-treatment variable (e.g. Acharya, Blackwell & Sen 2016)
  - can quantify scale of missing data (e.g. Nyhan, Skovron & Titiunik 2017 with voting age population)
- Policing data: causal mediation problem (Fun)...
  but with missing data on mediator (Double fun!)...
  and missingness is of unknown scale (All the fun I can handle.)

Slide Credit: Knox, Lowe and Mummolo

Solution: principal stratification

If $T \rightarrow M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$		
$M_i(1) = 0$		

Slide Credit: Knox, Lowe and Mummolo

Solution: principal stratification

If $T \rightarrow M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$	"always stop" (serious crime)	
$M_i(1) = 0$		

Slide Credit: Knox, Lowe and Mummolo

Solution: principal stratification

If $T \to M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$	"always stop" (serious crime)	
$M_i(1) = 0$		"never stop" (inconspicuous)

Slide Credit: Knox, Lowe and Mummolo

Solution: principal stratification

If $T \rightarrow M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$	"always stop" (serious crime)	stop if black (jaywalking)
$M_i(1) = 0$		"never stop" (inconspicuous)

Slide Credit: Knox, Lowe and Mummolo

Solution: principal stratification

If $T \rightarrow M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$	"always stop" (serious crime)	stop if black (jaywalking)
$M_i(1) = 0$	stop if white ?	"never stop" (inconspicuous)

Slide Credit: Knox, Lowe and Mummolo

Solution: principal stratification

If $T \rightarrow M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$	"always stop" (serious crime)	stop if black (jaywalking)
$M_i(1) = 0$	stop if white ?	"never stop" (inconspicuous)

What do we get to see in police data?

Slide Credit: Knox, Lowe and Mummolo

Solution: principal stratification

If $T \rightarrow M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$	"always stop" (serious crime)	stop if black (jaywalking)
$M_i(1) = 0$	stop if white ?	"never stop" (inconspicuous)

What do we get to see in police data?
For black civilians...

Slide Credit: Knox, Lowe and Mummolo

Solution: principal stratification

If $T \rightarrow M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$	"always stop" (serious crime)	stop if black (jaywalking)
$M_i(1) = 0$	stop if white ?	"never stop" (inconspicuous)

What do we get to see in police data?
For white civilians...

Slide Credit: Knox, Lowe and Mummolo

Encounters (sightings) belong to one of four principal strata

always-stop	anti-black racial-stop	anti-white racial-stop	never-stop
$M_i(1) = M_i(0) = 1$	$M_i(1) = 1, M_i(0) = 0$	$M_i(1) = 0, M_i(0) = 1$	$M_i(1) = 0, M_i(0) = 0$

Slide Credit: Knox, Lowe and Mummolo

Within each, civilian is treated (black) or not (white)

always-stop	anti-black racial-stop	anti-white racial-stop	never-stop
$M_i(1) = M_i(0) = 1$	$M_i(1) = 1, M_i(0) = 0$	$M_i(1) = 0, M_i(0) = 1$	$M_i(1) = 0, M_i(0) = 0$
$D_i = 0$ $\quad$ $D_i = 1$	$D_i = 0$ $\quad$ $D_i = 1$	$D_i = 0$ $\quad$ $D_i = 1$	$D_i = 0$ $\quad$ $D_i = 1$

Slide Credit: Knox, Lowe and Mummolo

Very few potential outcomes appear in police data

always-stop	anti-black racial-stop	anti-white racial-stop	never-stop
$M_i(1) = M_i(0) = 1$	$M_i(1) = 1, M_i(0) = 0$	$M_i(1) = 0, M_i(0) = 1$	$M_i(1) = 0, M_i(0) = 0$

$$\begin{array}{cc|cc|cc|cc}
D_i = 0 & D_i = 1 & D_i = 0 & D_i = 1 & D_i = 0 & D_i = 1 & D_i = 0 & D_i = 1 \\
Y_i(1,1) & Y_i(1,1) & Y_i(1,1) & Y_i(1,1) & Y_i(1,1) & Y_i(1,1) & Y_i(1,1) & Y_i(1,1) \\
Y_i(1,0) & Y_i(1,0) & Y_i(1,0) & Y_i(1,0) & Y_i(1,0) & Y_i(1,0) & Y_i(1,0) & Y_i(1,0) \\
Y_i(0,1) & Y_i(0,1) & Y_i(0,1) & Y_i(0,1) & Y_i(0,1) & Y_i(0,1) & Y_i(0,1) & Y_i(0,1) \\
Y_i(0,0) & Y_i(0,0) & Y_i(0,0) & Y_i(0,0) & Y_i(0,0) & Y_i(0,0) & Y_i(0,0) & Y_i(0,0)
\end{array}$$

Slide Credit: Knox, Lowe and Mummolo

Very few potential outcomes appear in police data

always-stop $M_i(1) = M_i(0) = 1$		anti-black racial-stop $M_i(1) = 1, M_i(0) = 0$		anti-white racial-stop $M_i(1) = 0, M_i(0) = 1$		never-stop $M_i(1) = 0, M_i(0) = 0$	
$D_i = 0$	$D_i = 1$	$D_i = 0$	$D_i = 1$	$D_i = 0$	$D_i = 1$	$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$Y_i(1,1)$	$Y_i(1,1)$	$Y_i(1,1)$	$Y_i(1,1)$	$Y_i(1,1)$	$Y_i(1,1)$	$Y_i(1,1)$
$Y_i(1,0)$	$Y_i(1,0)$	$Y_i(1,0)$	$Y_i(1,0)$	$Y_i(1,0)$	$Y_i(1,0)$	$Y_i(1,0)$	$Y_i(1,0)$
$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$
$Y_i(0,0)$	$Y_i(0,0)$	$Y_i(0,0)$	$Y_i(0,0)$	$Y_i(0,0)$	$Y_i(0,0)$	$Y_i(0,0)$	$Y_i(0,0)$

Slide Credit: Knox, Lowe and Mummolo

Which causal effect?

- Prior work does not name specific causal estimands (e.g. Antonovics and Knight, 2009; Fryer, 2019; Nix et al., 2017; Ross, 2018; Ridgeway, 2006)

- Without naming an estimand, we can't consider identifying assumptions or evaluate validity of an analysis

- There are many causal effects:
  - Average Treatment Effect (ATE) in the population
  - Effect among those who interact with police ($ATE_{M=1}$)
  - Effect among minorities who interact with police ($ATT_{M=1}$)

Slide Credit: Knox, Lowe and Mummolo

Causal estimands

Stratum	D_i	M_i	$M_i(0)$	$M_i(1)$	$Y_i(1,1)$	$Y_i(1,0)$	$Y_i(0,1)$	$Y_i(0,0)$	ATE	$ATE_{M=1}$	$ATT_{M=1}$	$CDE_{M=1}$
Always-Stop	1	1	1	1	1	0	1	0	0	0	0	0
Always-Stop	0	1	1	1	1	0	1	0	0	0		0
Racial Stop	1	1	0	1	1	0	1	0	1	1	1	0
Never-Stop	0	0	0	0	1	0	0	0	0			

Slide Credit: Knox, Lowe and Mummolo

Average Treatment Effect

$$ATE = \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))]$$

Stratum	D_i	M_i	$M_i(0)$	$M_i(1)$	$Y_i(1,1)$	$Y_i(1,0)$	$Y_i(0,1)$	$Y_i(0,0)$	ATE	$ATE_{M=1}$	$ATT_{M=1}$	$CDE_{M=1}$
Always-Stop	1	1	1	1	1	0	1	0	0	0	0	0
Always-Stop	0	1	1	1	1	0	1	0	0	0		0
Racial Stop	1	1	0	1	1	0	1	0	1	1	1	0
Never-Stop	0	0	0	0	1	0	0	0	0			

Slide Credit: Knox, Lowe and Mummolo

Average Treatment Effect Among the Stopped

$$ATE_{M=1} = \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0)) | M_i = 1]$$

Stratum	D_i	M_i	$M_i(0)$	$M_i(1)$	$Y_i(1,1)$	$Y_i(1,0)$	$Y_i(0,1)$	$Y_i(0,0)$	ATE	$ATE_{M=1}$	$ATT_{M=1}$	$CDE_{M=1}$
Always-Stop	1	1	1	1	1	0	1	0	0	0	0	0
Always-Stop	0	1	1	1	1	0	1	0	0	0		0
Racial Stop	1	1	0	1	1	0	1	0	1	1	1	0
Never-Stop	0	0	0	0	1	0	0	0	0			

Slide Credit: Knox, Lowe and Mummolo

Average Treatment Effect Among the Treated and Stopped

$$ATT_{M=1} = \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0)) | D_i = 1, M_i = 1]$$

Stratum	D_i	M_i	$M_i(0)$	$M_i(1)$	$Y_i(1,1)$	$Y_i(1,0)$	$Y_i(0,1)$	$Y_i(0,0)$	ATE	$ATE_{M=1}$	$ATT_{M=1}$	$CDE_{M=1}$
Always-Stop	1	1	1	1	1	0	1	0	0	0	0	0
Always-Stop	0	1	1	1	1	0	1	0	0	0		0
Racial Stop	1	1	0	1	1	0	1	0	1	1	1	0
Never-Stop	0	0	0	0	1	0	0	0	0			

Slide Credit: Knox, Lowe and Mummolo

Controlled Direct Effect Among the Stopped

$$CDE_{M=1} = \mathbb{E}[Y_i(1,1) - Y_i(0,1)|M_i = 1]$$

Stratum	D_i	M_i	$M_i(0)$	$M_i(1)$	$Y_i(1,1)$	$Y_i(1,0)$	$Y_i(0,1)$	$Y_i(0,0)$	ATE	$ATE_{M=1}$	$ATT_{M=1}$	$CDE_{M=1}$
Always-Stop	1	1	1	1	1	0	1	0	0	0	0	0
Always-Stop	0	1	1	1	1	0	1	0	0	0		0
Racial Stop	1	1	0	1	1	0	1	0	1	1	1	0
Never-Stop	0	0	0	0	1	0	0	0	0			

Slide Credit: Knox, Lowe and Mummolo

Assumption 1: Mandatory reporting

$$Y_i(d, 0) = 0$$

- If encounter not in the data, no force was applied
- Highly plausible for lethal/severe force
- Increasingly plausible for sub-lethal force given civilian oversight boards, cell phone cameras

Slide Credit: Knox, Lowe and Mummolo

Assumption 1: Mandatory reporting

always-stop
$M_i(1) = M_i(0) = 1$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$
$Y_i(1,0)$	$Y_i(1,0)$
$\boldsymbol{Y_i(0,1)}$	$Y_i(0,1)$
$Y_i(0,0)$	$Y_i(0,0)$

anti-black racial-stop
$M_i(1) = 1, M_i(0) = 0$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$
$Y_i(1,0)$	$Y_i(1,0)$
$Y_i(0,1)$	$Y_i(0,1)$
$Y_i(0,0)$	$Y_i(0,0)$

anti-white racial-stop
$M_i(1) = 0, M_i(0) = 1$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$Y_i(1,1)$
$Y_i(1,0)$	$Y_i(1,0)$
$\boldsymbol{Y_i(0,1)}$	$Y_i(0,1)$
$Y_i(0,0)$	$Y_i(0,0)$

never-stop
$M_i(1) = 0, M_i(0) = 0$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$Y_i(1,1)$
$Y_i(1,0)$	$Y_i(1,0)$
$Y_i(0,1)$	$Y_i(0,1)$
$Y_i(0,0)$	$Y_i(0,0)$

Slide Credit: Knox, Lowe and Mummolo

Assumption 1: Mandatory reporting

always-stop		anti-black racial-stop		anti-white racial-stop		never-stop	
$M_i(1) = M_i(0) = 1$		$M_i(1) = 1, M_i(0) = 0$		$M_i(1) = 0, M_i(0) = 1$		$M_i(1) = 0, M_i(0) = 0$	
$D_i = 0$	$D_i = 1$	$D_i = 0$	$D_i = 1$	$D_i = 0$	$D_i = 1$	$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$	$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$	$Y_i(1,1)$	$Y_i(1,1)$	$Y_i(1,1)$	$Y_i(1,1)$
0	**0**	**0**	**0**	**0**	**0**	**0**	**0**
$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$
0	**0**	**0**	**0**	**0**	**0**	**0**	**0**

Slide Credit: Knox, Lowe and Mummolo

Assumption 2: Mediator monotonicity

$$M_i(1) \geq M_i(0)$$

- No anti-white bias in stopping

Slide Credit: Knox, Lowe and Mummolo

Assumption 2: Mediator monotonicity

always-stop	anti-black racial-stop	anti-white racial-stop	never-stop
$M_i(1) = M_i(0) = 1$	$M_i(1) = 1, M_i(0) = 0$	$M_i(1) = 0, M_i(0) = 1$	$M_i(1) = 0, M_i(0) = 0$

always-stop: $M_i(1) = M_i(0) = 1$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$
0	**0**
$Y_i(0,1)$	$Y_i(0,1)$
0	**0**

anti-black racial-stop: $M_i(1) = 1, M_i(0) = 0$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$
0	**0**
$Y_i(0,1)$	$Y_i(0,1)$
0	**0**

anti-white racial-stop: $M_i(1) = 0, M_i(0) = 1$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$Y_i(1,1)$
0	**0**
$Y_i(0,1)$	$Y_i(0,1)$
0	**0**

never-stop: $M_i(1) = 0, M_i(0) = 0$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$Y_i(1,1)$
0	**0**
$Y_i(0,1)$	$Y_i(0,1)$
0	**0**

Slide Credit: Knox, Lowe and Mummolo

Assumption 2: Mediator monotonicity

always-stop
$M_i(1) = M_i(0) = 1$

$D_i = 0$ $D_i = 1$

$Y_i(1,1)$ $\boldsymbol{Y_i(1,1)}$
0 **0**
$Y_i(0,1)$ $Y_i(0,1)$
0 **0**

anti-black
racial-stop
$M_i(1) = 1, M_i(0) = 0$

$D_i = 0$ $D_i = 1$

$Y_i(1,1)$ $\boldsymbol{Y_i(1,1)}$
0 **0**
$Y_i(0,1)$ $Y_i(0,1)$
0 **0**

never-stop
$M_i(1) = 0, M_i(0) = 0$

$D_i = 0$ $D_i = 1$

$Y_i(1,1)$ $Y_i(1,1)$
0 **0**
$Y_i(0,1)$ $Y_i(0,1)$
0 **0**

Slide Credit: Knox, Lowe and Mummolo

Assumption 3: Relative non-severity of racial stops

$$E[Y_i(d,m)|D_i = d', M_i(1) = 1, M_i(0) = 1] \geq \mathbb{E}[Y_i(d,m)|D_i = d', M_i(1) = 1, M_i(0) = 0]$$

- Level of force applied in always-stop encounters (serious crimes) $\geq$ level applied in racial stop encounters <u>on average</u>

Slide Credit: Knox, Lowe and Mummolo

Assumption 3: Relative non-severity of racial stops

always-stop

$M_i(1) = M_i(0) = 1$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$\mathbf{Y_i(1,1)}$
0	**0**
$Y_i(0,1)$	$Y_i(0,1)$
0	**0**

anti-black
racial-stop

$M_i(1) = 1, M_i(0) = 0$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$\mathbf{Y_i(1,1)}$
0	**0**
$Y_i(0,1)$	$Y_i(0,1)$
0	**0**

never-stop

$M_i(1) = 0, M_i(0) = 0$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$Y_i(1,1)$
0	**0**
$Y_i(0,1)$	$Y_i(0,1)$
0	**0**

Slide Credit: Knox, Lowe and Mummolo

Assumption 3: Relative non-severity of racial stops

always-stop		anti-black racial-stop	never-stop
$M_i(1) = M_i(0) = 1$		$M_i(1) = 1, M_i(0) = 0$	$M_i(1) = 0, M_i(0) = 0$
$D_i = 0$ $D_i = 1$		$D_i = 0$ $D_i = 1$	$D_i = 0$ $D_i = 1$
$Y_i(1,1)$ $\boldsymbol{Y_i(1,1)}$	$\geq$	$Y_i(1,1)$ $\boldsymbol{Y_i(1,1)}$	$Y_i(1,1)$ $Y_i(1,1)$
0 **0**		**0** **0**	**0** **0**
$Y_i(0,1)$ $Y_i(0,1)$		$Y_i(0,1)$ $Y_i(0,1)$	$Y_i(0,1)$ $Y_i(0,1)$
0 **0**		**0** **0**	**0** **0**

Slide Credit: Knox, Lowe and Mummolo

Assumption 4: Treatment ignorability

$$\begin{aligned} M_i(d) &\perp\!\!\!\perp D_i \\ Y_i(d, M(d)) &\perp\!\!\!\perp D_i | M_i(0) = m', M_i(1) = m'' \end{aligned}$$

- No omitted variables with respect to mediator or outcome
- More plausible in recent years (data on lat/lon, time, officer and suspect features, etc.)

Slide Credit: Knox, Lowe and Mummolo

Assumption 4: Treatment Ignorability

always-stop			anti-black racial-stop			never-stop	
$M_i(1) = M_i(0) = 1$			$M_i(1) = 1, M_i(0) = 0$			$M_i(1) = 0, M_i(0) = 0$	
$D_i = 0$	$D_i = 1$		$D_i = 0$	$D_i = 1$		$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$	$\geq$	$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$		$Y_i(1,1)$	$Y_i(1,1)$
0	**0**		**0**	**0**		**0**	**0**
$Y_i(0,1)$	$Y_i(0,1)$		$Y_i(0,1)$	$Y_i(0,1)$		$Y_i(0,1)$	$Y_i(0,1)$
0	**0**		**0**	**0**		**0**	**0**

Slide Credit: Knox, Lowe and Mummolo

Assumption 4: Treatment Ignorability

always-stop		anti-black racial-stop	never-stop
$M_i(1) = M_i(0) = 1$		$M_i(1) = 1, M_i(0) = 0$	$M_i(1) = 0, M_i(0) = 0$
$\boldsymbol{Y_i(1,1)}$		$\boldsymbol{Y_i(1,1)}$	$Y_i(1,1)$
0	$\geq$	**0**	**0**
$\boldsymbol{Y_i(0,1)}$		$\boldsymbol{Y_i(0,1)}$	$Y_i(0,1)$
0		**0**	**0**

Slide Credit: Knox, Lowe and Mummolo

Given assumptions 1-4, can we recover a causal estimate?

Given assumptions 1-4, can we recover a causal estimate?

- Consider the naïve estimator:

$$\hat{\Delta} = \hat{\mathbb{E}}[Y_i | D_i = 1, M_i = 1] - \hat{\mathbb{E}}[Y_i | D_i = 0, M_i = 1]$$

Given assumptions 1-4, can we recover a causal estimate?

- Consider the naïve estimator:

$$\hat{\Delta} = \hat{\mathbb{E}}[Y_i | D_i = 1, M_i = 1] - \hat{\mathbb{E}}[Y_i | D_i = 0, M_i = 1]$$

- Target the $ATE_{M=1}$ and $ATT_{M=1}$

Given assumptions 1-4, can we recover a causal estimate?

- Consider the naïve estimator:

$$\hat{\Delta} = \hat{\mathbb{E}}[Y_i|D_i = 1, M_i = 1] - \hat{\mathbb{E}}[Y_i|D_i = 0, M_i = 1]$$

- Target the $ATE_{M=1}$ and $ATT_{M=1}$
- Biased even without omitted variables.

Given assumptions 1-4, can we recover a causal estimate?

- Consider the naïve estimator:

$$\hat{\Delta} = \hat{\mathbb{E}}[Y_i|D_i = 1, M_i = 1] - \hat{\mathbb{E}}[Y_i|D_i = 0, M_i = 1]$$

- Target the $ATE_{M=1}$ and $ATT_{M=1}$
- Biased even without omitted variables.
- Bias is always nonpositive.

Given assumptions 1-4, can we recover a causal estimate?

- Consider the naïve estimator:

$$\hat{\Delta} = \hat{\mathbb{E}}[Y_i | D_i = 1, M_i = 1] - \hat{\mathbb{E}}[Y_i | D_i = 0, M_i = 1]$$

- Target the $ATE_{M=1}$ and $ATT_{M=1}$
- Biased even without omitted variables.
- Bias is always nonpositive.
- Bias remains unless there are no racial stops

Given assumptions 1-4, can we recover a causal estimate?

- Consider the naïve estimator:

$$\hat{\Delta} = \hat{\mathbb{E}}[Y_i | D_i = 1, M_i = 1] - \hat{\mathbb{E}}[Y_i | D_i = 0, M_i = 1]$$

- Target the $ATE_{M=1}$ and $ATT_{M=1}$
- Biased even without omitted variables.
- Bias is always nonpositive.
- Bias remains unless there are no racial stops
- Biased even if the goal is to estimate effect among the stopped.

Can we estimate racial bias with police administrative data?

Given assumptions 1-4, can we recover a causal estimate?

- Consider the naïve estimator:

$$\hat{\Delta} = \hat{\mathbb{E}}[Y_i | D_i = 1, M_i = 1] - \hat{\mathbb{E}}[Y_i | D_i = 0, M_i = 1]$$

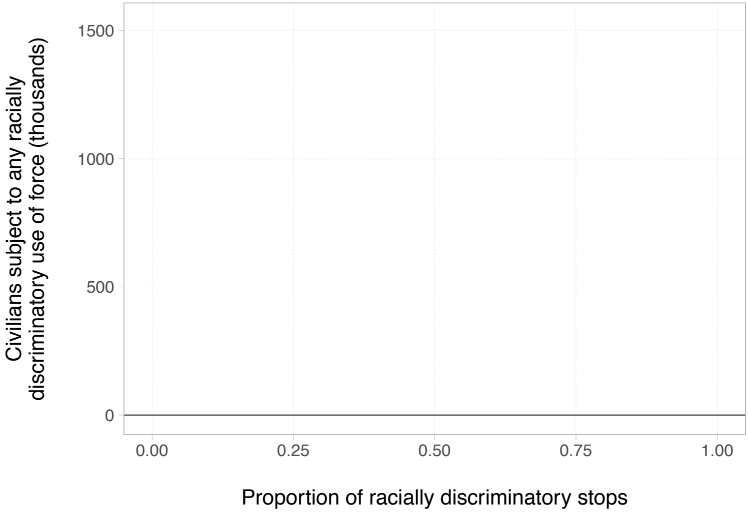
- Target the $ATE_{M=1}$ and $ATT_{M=1}$
- Biased even without omitted variables.
- Bias is always nonpositive.
- Bias remains unless there are no racial stops
- Biased even if the goal is to estimate effect among the stopped.

Can we estimate racial bias with police administrative data?

Solution: bounds on effects!

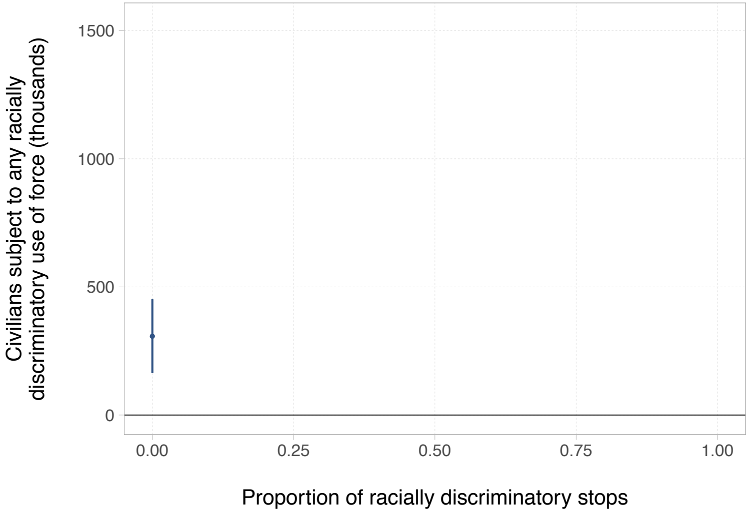
Slide Credit: Knox, Lowe and Mummolo

Bounds on race effects, black vs. white



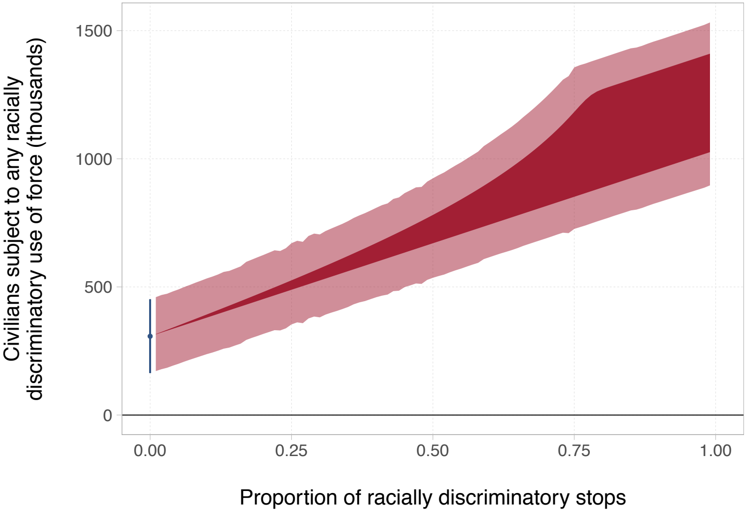
Slide Credit: Knox, Lowe and Mummolo

Bounds on race effects, black vs. white



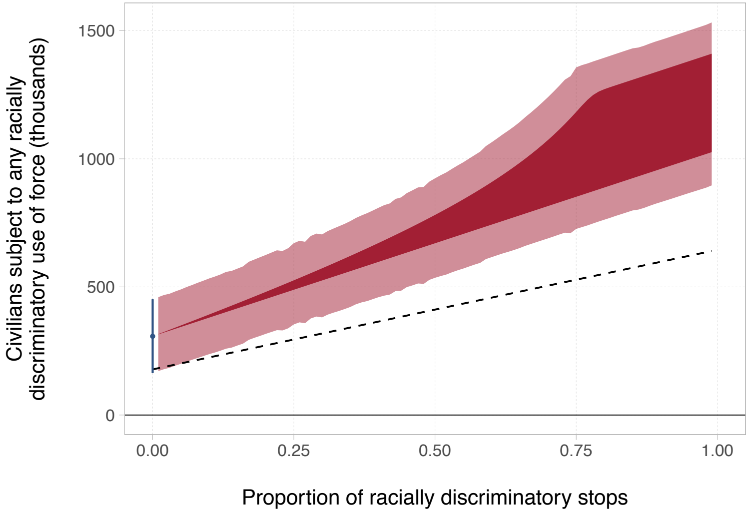
Slide Credit: Knox, Lowe and Mummolo

Bounds on race effects, black vs. white



Slide Credit: Knox, Lowe and Mummolo

Bounds on race effects, black vs. white



Slide Credit: Knox, Lowe and Mummolo

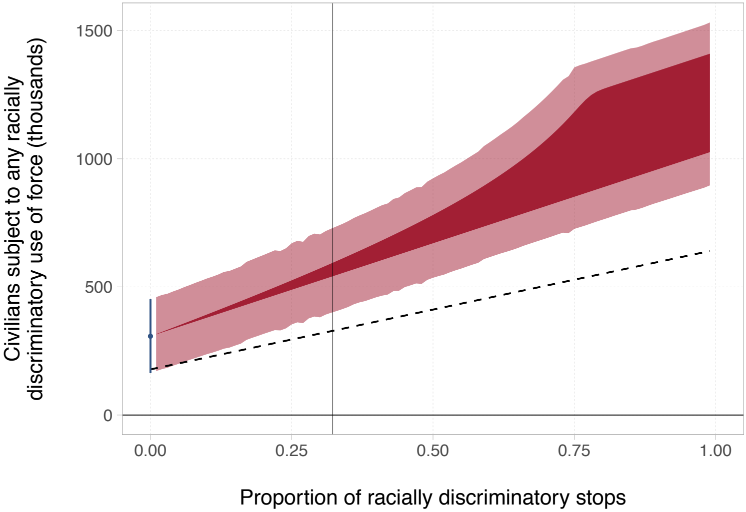
What is ρ?

What is the share of minority stops that would not have happened if civilians had been white?

- Can be anywhere in $[0, 1)$. If $\rho = 0$, bias disappears.
- Two prior studies estimate this using data on "Stop, Question and Frisk" in NYC
- Gelman, Fagan & Kiss (2007) and Goel, Rao and Schroff (2016)
- Studies take totally different approaches
- Results imply ρ is at least .32 or .34, respectively
- We use $\rho =$.32 to be conservative

Slide Credit: Knox, Lowe and Mummolo

Bounds on race effects, black vs. white



Slide Credit: Knox, Lowe and Mummolo

Improving Research Design

Improving Research Design

- Only partially identified. Can't get the population ATE.

Improving Research Design

- Only partially identified. Can't get the population ATE.
- Option 1: Identify situations with race-blind contact with police (e.g. rules for DUI stops; traffic stops and night; traffic accidents?)

Improving Research Design

- Only partially identified. Can't get the population ATE.
- Option 1: Identify situations with race-blind contact with police (e.g. rules for DUI stops; traffic stops and night; traffic accidents?)
- Option 2: Gather data on the non-stopped (e.g. traffic cameras, body camera footage)

Slide Credit: Knox, Lowe and Mummolo

Takeaways

Takeaways

- This isn't really a story about machine learning as much as it is a story about data.

Takeaways

- This isn't really a story about machine learning as much as it is a story about data.
- The key though is that there is no modeling that would help you get around this. You have to think about the data!

Takeaways

- This isn't really a story about machine learning as much as it is a story about data.
- The key though is that there is no modeling that would help you get around this. You have to think about the data!
- Key tools for thinking about data:
  1) Measurement $\rightsquigarrow$ does this capture what I want?

Takeaways

- This isn't really a story about machine learning as much as it is a story about data.
- The key though is that there is no modeling that would help you get around this. You have to think about the data!
- Key tools for thinking about data:
  1) Measurement $\rightsquigarrow$ does this capture what I want?
  2) Representation $\rightsquigarrow$ is this a sample from the right population?

Takeaways

- This isn't really a story about machine learning as much as it is a story about data.
- The key though is that there is no modeling that would help you get around this. You have to think about the data!
- Key tools for thinking about data:
  1) Measurement $\rightsquigarrow$ does this capture what I want?
  2) Representation $\rightsquigarrow$ is this a sample from the right population?
- Both questions require a clear sense of the task/goal to answer!

1. Repurposing Big, Found Data
2. Case Study: COVID Risk Estimator
3. Case Study: Policing Data
4. Case Study: Respect and Formality in Policing
5. Case Study: Failure to Appear Case

Case Study: Respect and Formality in Policing

Case Study: Respect and Formality in Policing

Example	Respect Score
First Name, Ask For Agency, Questions [name], can I see that driver's license again? It- it's showing suspended. Is that- that's you? Disfluency, Negative Word, Disfluency	-1.07
Informal Title, Ask For Agency, Adverbial "Just" All right, my man. Do me a favor. Just keep your hands on the steering wheel real quick. "Hands On The Wheel"	-0.51
Apology, Introduction, Last Name Sorry to stop you. My name's Officer [name] with the Police Department.	0.84
Formal Title, Safety, Please There you go, ma'am. Drive safe, please.	1.21
Adverbial "Just", Filled Pause, Reassurance It just says that, uh, you've fixed it. No problem. Thank you very much, sir. Gratitude, Formal Title	2.07

Fig. 3. Sample sentences with automatically generated Respect scores. Features in blue have positive coefficients in the model and connote respect, such as offering reassurance ("no problem") or mentioning community member well-being ("drive safe"). Features in red have negative coefficients in the model and connote disrespect, like informal titles ("my man"), or disfluencies ("that- that's").

Case Study: Respect and Formality in Policing

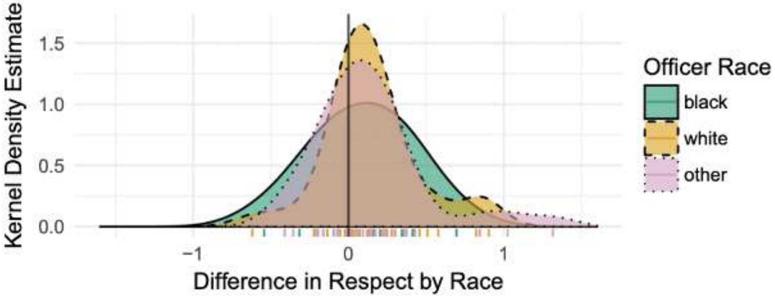


Fig. 4. Kernel density estimate of individual officer-level differences in Respect when talking to white as opposed to black community members, for the 90 officers in our dataset who have interactions with both blacks and whites. More positive numbers on the *x* axis represent a greater positive shift in Respect toward white community members.

1. Repurposing Big, Found Data
2. Case Study: COVID Risk Estimator
3. Case Study: Policing Data
4. Case Study: Respect and Formality in Policing
5. Case Study: Failure to Appear Case

HUMAN DECISIONS AND MACHINE PREDICTIONS*

Jon Kleinberg
Himabindu Lakkaraju
Jure Leskovec
Jens Ludwig
Sendhil Mullainathan

Can machine learning improve human decision making? Bail decisions provide a good test case. Millions of times each year, judges make jail-or-release decisions that hinge on a prediction of what a defendant would do if released. The concreteness of the prediction task combined with the volume of data available makes this a promising machine-learning application. Yet comparing the algorithm to judges proves complicated. First, the available data are generated by prior judge decisions. We only observe crime outcomes for released defendants, not for those judges detained. This makes it hard to evaluate counterfactual decision rules based on algorithmic predictions. Second, judges may have a broader set of preferences than the variable the algorithm predicts; for instance, judges

Kleinberg et. al

Kleinberg et. al

- Question: given the structure of the system as it is—can we make better decisions than judges about who to detain with machine learning?

Kleinberg et. al

- Question: given the structure of the system as it is—can we make better decisions than judges about who to detain with machine learning?
  - Straightforward, but uninteresting task: predict who the judge will release and who they won't.

Kleinberg et. al

- Question: given the structure of the system as it is—can we make better decisions than judges about who to detain with machine learning?
  - Straightforward, but uninteresting task: predict who the judge will release and who they won't.
  - Hard task: predict who would commit a crime if they were released

Kleinberg et. al

- Question: given the structure of the system as it is—can we make better decisions than judges about who to detain with machine learning?
  - Straightforward, but uninteresting task: predict who the judge will release and who they won't.
  - Hard task: predict who would commit a crime if they were released
  - Problem: we don't have Failure to Appear (FTA) for people who are detained.

Kleinberg et. al

- Question: given the structure of the system as it is—can we make better decisions than judges about who to detain with machine learning?
  - Straightforward, but uninteresting task: predict who the judge will release and who they won't.
  - Hard task: predict who would commit a crime if they were released
  - Problem: we don't have Failure to Appear (FTA) for people who are detained.
- An interesting and complex case. It is "easier" because one counterfactual is known—if the defendant is detained they cannot fail to appear.

Kleinberg et. al

- Question: given the structure of the system as it is—can we make better decisions than judges about who to detain with machine learning?
  - Straightforward, but uninteresting task: predict who the judge will release and who they won't.
  - Hard task: predict who would commit a crime if they were released
  - Problem: we don't have Failure to Appear (FTA) for people who are detained.
- An interesting and complex case. It is "easier" because one counterfactual is known—if the defendant is detained they cannot fail to appear.
- Three requirements: being clear about the link between predictions and decisions; specifying the scope of payoff functions; and constructing unbiased decision counterfactuals.

"It is telling that in our application, much of the work happens after the prediction function was estimated."

Data and Process

Data and Process

- 1.4 million cases between November 2008 and November 2013

Data and Process

- 1.4 million cases between November 2008 and November 2013
- Train, Test, and "Pure Hold-out"

Data and Process

- 1.4 million cases between November 2008 and November 2013
- Train, Test, and “Pure Hold-out”
- Pure Hold-out: all of some judges, last six months, some from all remaining judges

Data and Process

- 1.4 million cases between November 2008 and November 2013
- Train, Test, and "Pure Hold-out"
- Pure Hold-out: all of some judges, last six months, some from all remaining judges
- Train Gradient-Boosted Trees, 5-fold CV to select hyper parameters, retrain on all data afterwards.

Data and Process

- 1.4 million cases between November 2008 and November 2013
- Train, Test, and "Pure Hold-out"
- Pure Hold-out: all of some judges, last six months, some from all remaining judges
- Train Gradient-Boosted Trees, 5-fold CV to select hyper parameters, retrain on all data afterwards.
- Don't train on any demographic other than age. Does this help? hurt?

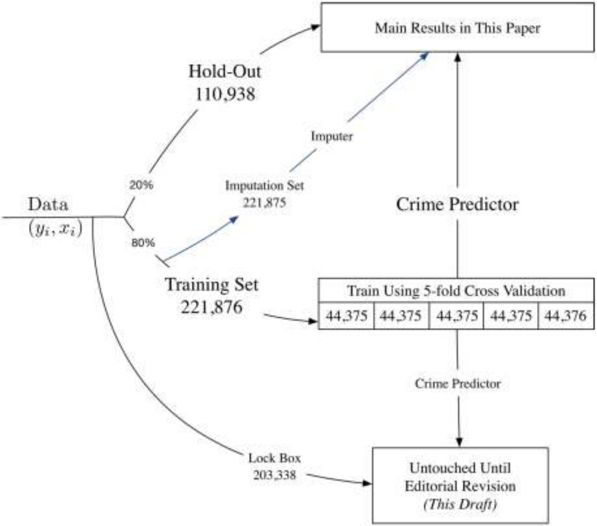


FIGURE I

Partition of New York City Data (2008–13) into Data Sets Used for Prediction and Evaluation

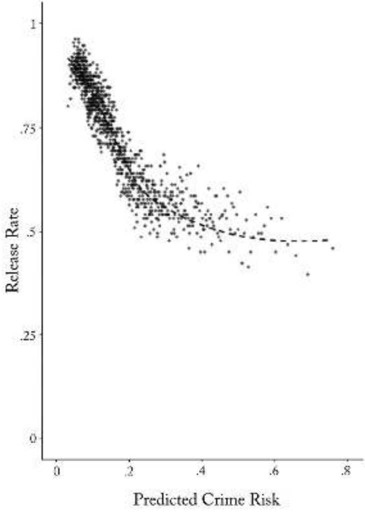


FIGURE II

How Machine Predictions of Crime Risk Relate to Judge Release Decisions and Actual Crime Rates

TABLE II

COMPARING LOGISTIC REGRESSION TO MACHINE-LEARNING PREDICTIONS OF CRIME RISK

Predicted risk percentile	ML/ logit overlap	Average observed crime rate for cases identified as high risk by:				
		Both ML and logit	ML only	Logit only	All ML cases	All logit cases
1%	30.6%	0.6080	0.5440	0.3996	0.5636	0.4633
		(0.0309)	(0.0209)	(0.0206)	(0.0173)	(0.0174)
5%	59.9%	0.4826	0.4090	0.3040	0.4531	0.4111
		(0.0101)	(0.0121)	(0.0114)	(0.0078)	(0.0077)
10%	65.9%	0.4134	0.3466	0.2532	0.3907	0.3589
		(0.0067)	(0.0090)	(0.0082)	(0.0054)	(0.0053)
25%	72.9%	0.3271	0.2445	0.1608	0.3048	0.2821
		(0.0038)	(0.0058)	(0.0049)	(0.0032)	(0.0031)

Notes. The table shows the results of fitting a machine-learning (ML) algorithm or a logistic regression to our training data set, to identify the highest-risk observations in our test set. Both machine-learning and logistic regression models are trained on the outcome of failure to appear (FTA). Each row presents statistics for the top part of the predicted risk distribution indicated in the first column: top 25% (N = 20,423); 10% (8,173); 5% (4,087); and 1% (818). The second column shows the share of cases in the top X% of the predicted risk distribution that overlap between the set identified by ML and the set identified by logistic regression. The subsequent columns report the average crime rate observed among the released defendants within the top X% of the predicted risk distribution as identified by both ML and logit, ML only, and logit only, and all top X% identified by ML (whether or not they are also identified by logistic regression) and top X% identified by logit.

5. In practice algorithms would be decision aids, not decision makers. Our calculations simply highlight the scope of the potential gains. Understanding the determinants of compliance with prediction tools is beyond the scope of this article, though recent work has begun to focus on it (Dietvorst, Simmons, and Masey 2015; Yeomans et al. 2016; Logg 2017).

28. This logistic regression uses the same set of covariates as are given to the machine-learning algorithm, using a standard linear additive functional form. Interestingly, estimating a model that adds all two-way interactions between the predictors causes the logistic regression model to overfit the training data set and predict poorly out of sample, even in a pared-down model with just 5 or 10 covariates. Many of these interactions seem to be picking up statistical noise in the training set, although the machine-learning algorithm is able to find the interactivity that reflects real signal.

Kleinberg et al.

- A lot more to how they set this up including, e.g. leveraging random judge assignment.

Kleinberg et al.

- A lot more to how they set this up including, e.g. leveraging random judge assignment.
- Key takeaway: even for a relatively straightforward problem, there is a lot to think about!

Kleinberg et al.

- A lot more to how they set this up including, e.g. leveraging random judge assignment.
- Key takeaway: even for a relatively straightforward problem, there is a lot to think about!
- This is why a third of the class is about sources of data—huge considerations that arise from how we came to observe the data we had.

Kleinberg et al.

- A lot more to how they set this up including, e.g. leveraging random judge assignment.
- Key takeaway: even for a relatively straightforward problem, there is a lot to think about!
- This is why a third of the class is about sources of data—huge considerations that arise from how we came to observe the data we had.
- Big data alone is not sufficient we have to understand the task and the process!

This Week in Review

This Week in Review

- Found or Repurposed Data
- Knox, Lowe and Mummolo 2020
- Voigt et al. 2017

Next Week:

Salganik Chapter 6 of *Bit by Bit*: "Ethics"
Lundberg et. al 2019 "Privacy, Ethics, and Data Access: A Case Study of the Fragile Families Challenge." *Socius*