Advanced Statistical Methods for Observational Studies

LECTURE 06

thick description

QUALIA HUNTING

pre-trial detention

pre-trial detention

- When someone gets picked up by the police there’s some time between when they are detained and when their case will be heard

pre-trial detention

- When someone gets picked up by the police there’s some time between when they are detained and when their case will be heard
- This can be weeks or years

pre-trial detention

- When someone gets picked up by the police there's some time between when they are detained and when their case will be heard
- This can be weeks or years
- There is an arraignment, usually within 24 hours, after being picked up where it's decided if the person will be either released on recognizance (RoR'ed) or required to be detained in some manner (e.g., money bail)

pre-trial detention

- When someone gets picked up by the police there's some time between when they are detained and when their case will be heard
- This can be weeks or years
- There is an arraignment, usually within 24 hours, after being picked up where it's decided if the person will be either released on recognizance (RoR'ed) or required to be detained in some manner (e.g., money bail)
- There are three parties involved: prosecutor, defendant/defense attorney, and judge

pre-trial detention

- When someone gets picked up by the police there's some time between when they are detained and when their case will be heard
- This can be weeks or years
- There is an arraignment, usually within 24 hours, after being picked up where it's decided if the person will be either released on recognizance (RoR'ed) or required to be detained in some manner (e.g., money bail)
- There are three parties involved: prosecutor, defendant/defense attorney, and judge
- The defendant and defense attorney have likely just met, possibly they've spoken for as many as 2-5 minutes before going into the decision meeting

pre-trial detention

- When someone gets picked up by the police there's some time between when they are detained and when their case will be heard
- This can be weeks or years
- There is an arraignment, usually within 24 hours, after being picked up where it's decided if the person will be either released on recognizance (RoR'ed) or required to be detained in some manner (e.g., money bail)
- There are three parties involved: prosecutor, defendant/defense attorney, and judge
- The defendant and defense attorney have likely just met, possibly they've spoken for as many as 2-5 minutes before going into the decision meeting
- This process takes about 5-15 minutes for each person under consideration

pre-trial detention

- When someone gets picked up by the police there's some time between when they are detained and when their case will be heard
- This can be weeks or years
- There is an arraignment, usually within 24 hours, after being picked up where it's decided if the person will be either released on recognizance (RoR'ed) or required to be detained in some manner (e.g., money bail)
- There are three parties involved: prosecutor, defendant/defense attorney, and judge
- The defendant and defense attorney have likely just met, possibly they've spoken for as many as 2-5 minutes before going into the decision meeting
- This process takes about 5-15 minutes for each person under consideration

- This is deeply insane

pre-trial detention

- If money-bail is set and the person cannot pay then he is remanded into custody for the duration

pre-trial detention

- If money-bail is set and the person cannot pay then he is remanded into custody for the duration
- This incarceration has largely been at Riker's Island

pre-trial detention

- If money-bail is set and the person cannot pay then he is remanded into custody for the duration
- This incarceration has largely been at Riker's Island

pre-trial detention

- If money-bail is set and the person cannot pay then he is remanded into custody for the duration
- This incarceration has largely been at Riker's Island

pre-trial detention

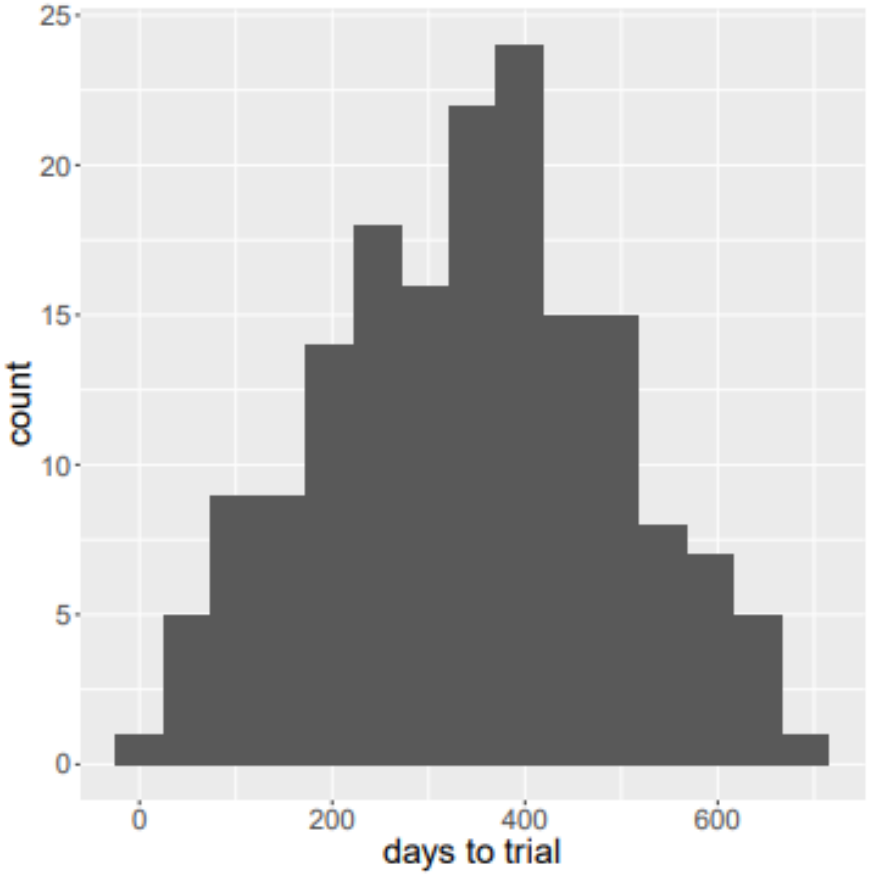
- If money-bail is set and the person cannot pay then he is remanded into custody for the duration
- This incarceration has largely been at Riker's Island

pre-trial detention

- This isn’t a quick process

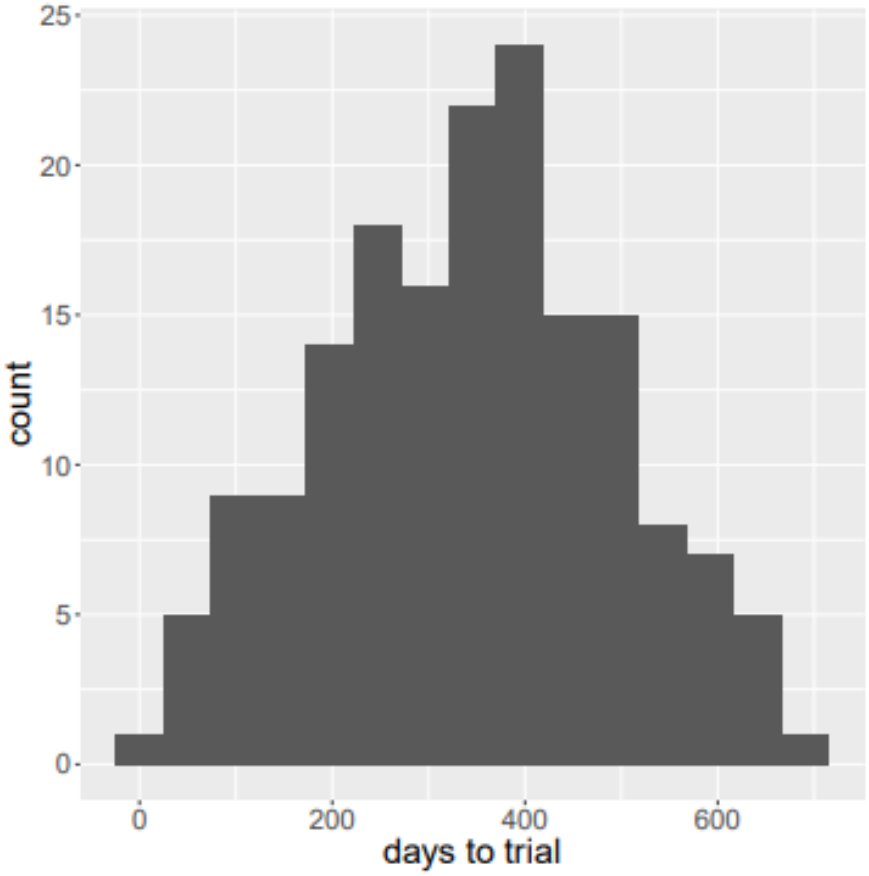
pre-trial detention

- This isn’t a quick process



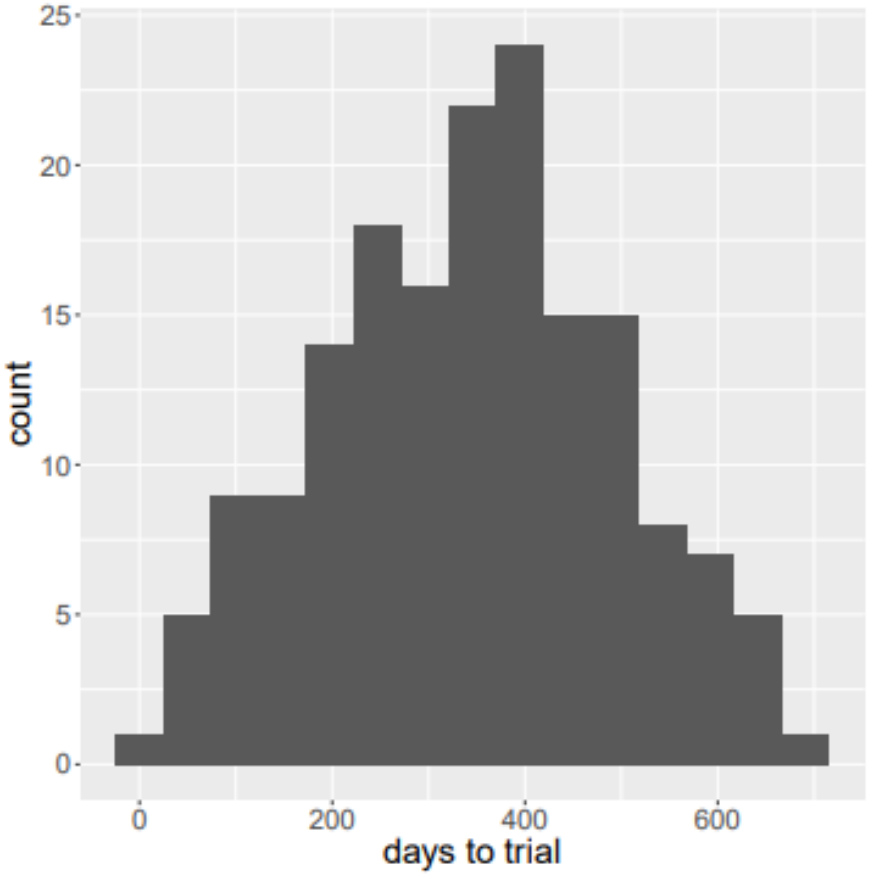
pre-trial detention

- This isn't a quick process
- If poor then you are likely to plead out



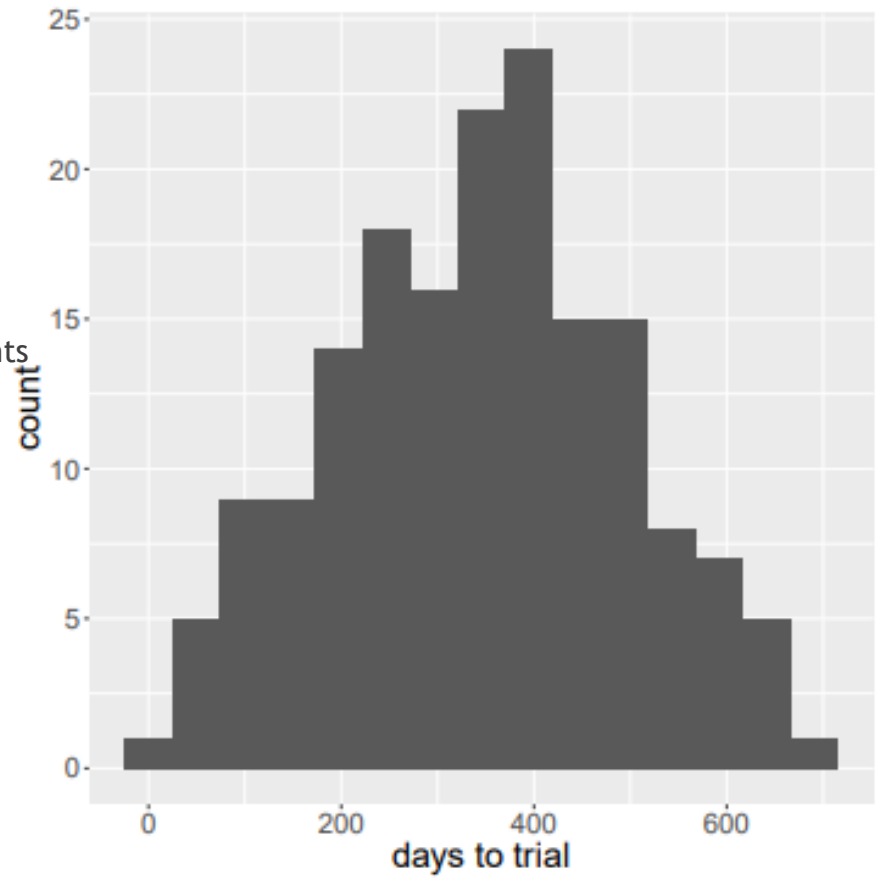
pre-trial detention

- This isn't a quick process
- If poor then you are likely to plead out
  - Get out of the violence



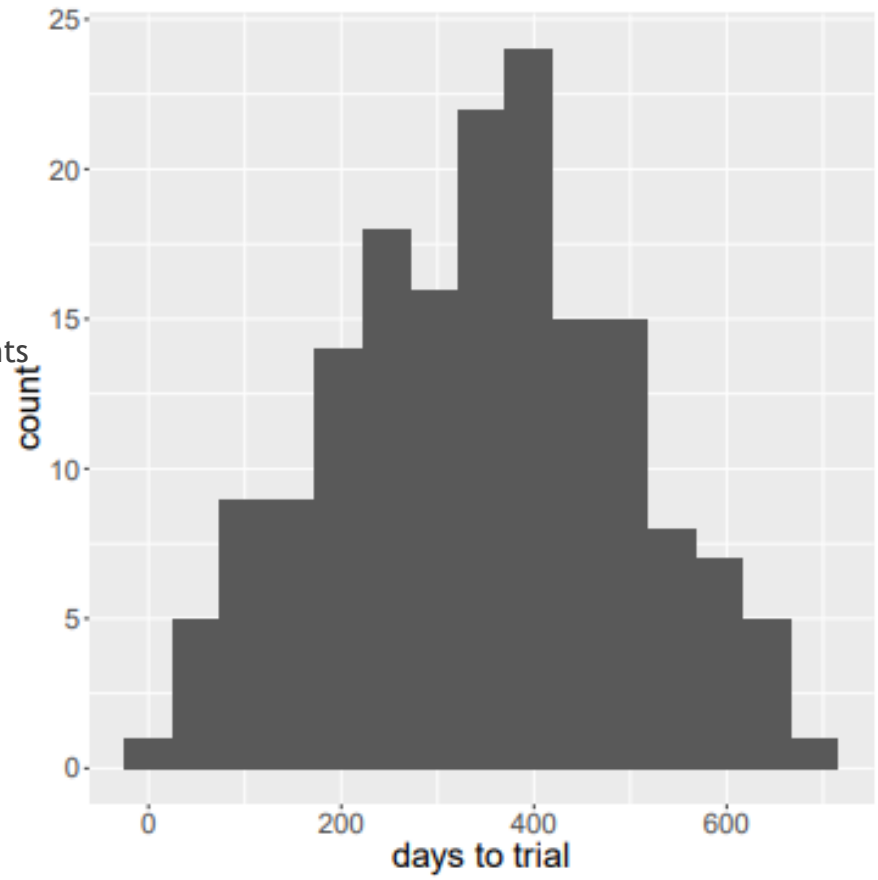
pre-trial detention

- This isn't a quick process
- If poor then you are likely to plead out
  - Get out of the violence
  - Get back to taking care of your dependents



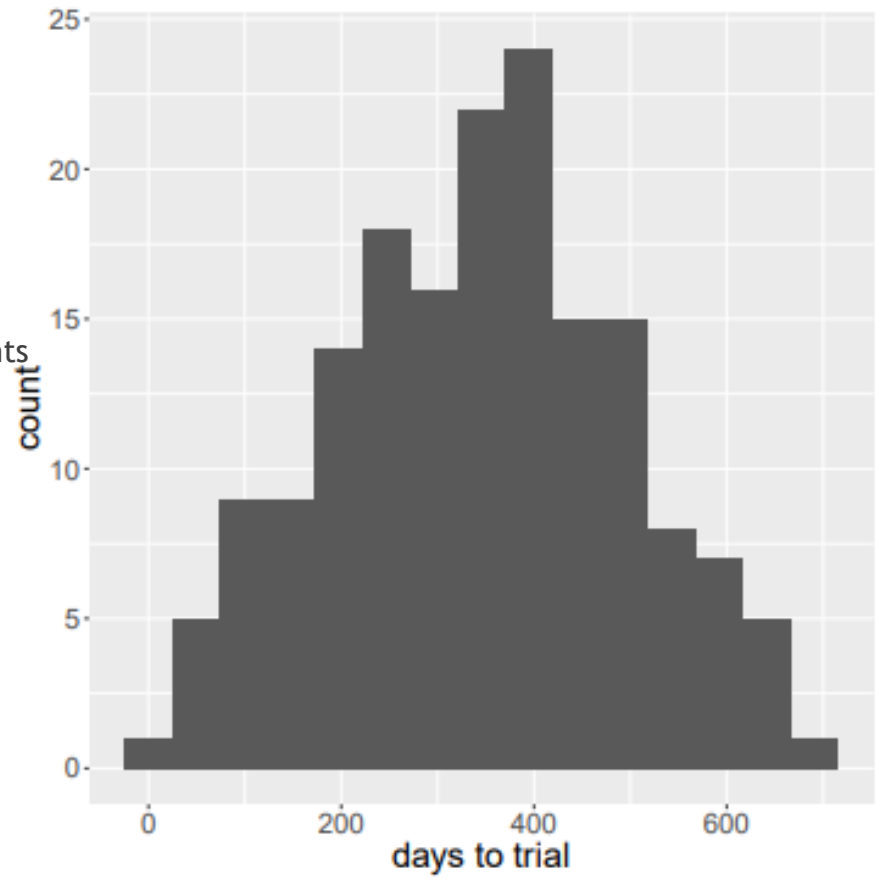
pre-trial detention

- This isn't a quick process
- If poor then you are likely to plead out
  - Get out of the violence
  - Get back to taking care of your dependents
  - Need to earn money



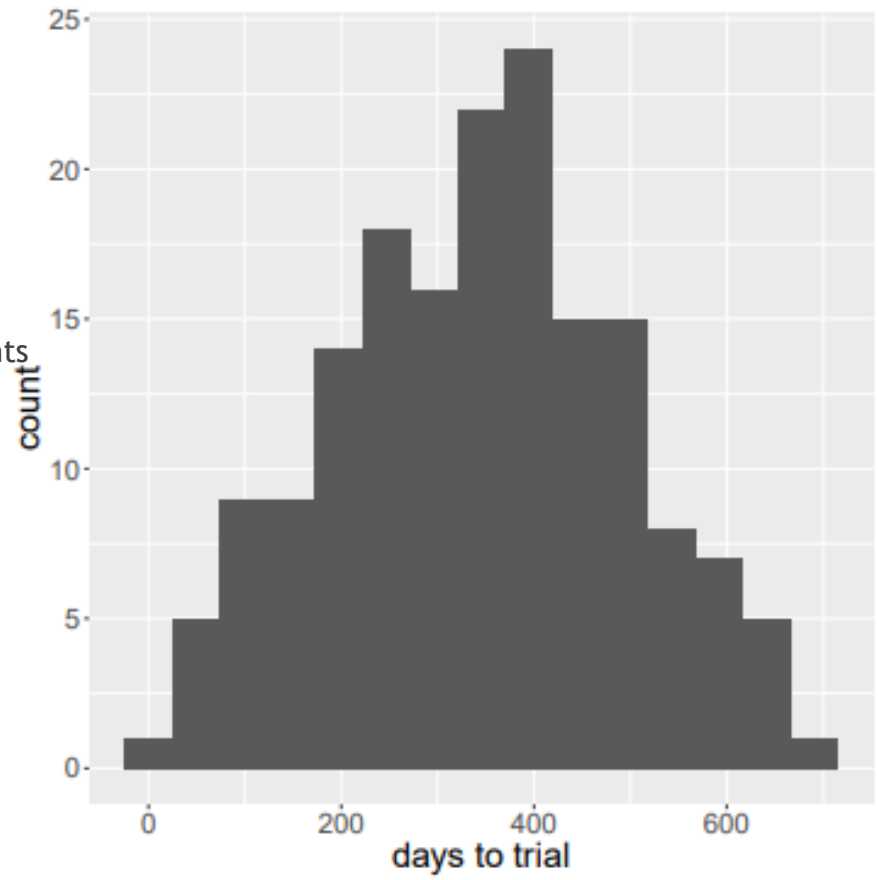
pre-trial detention

- This isn't a quick process
- If poor then you are likely to plead out
  - Get out of the violence
  - Get back to taking care of your dependents
  - Need to earn money
- Of the 80,000 cases in our data set…



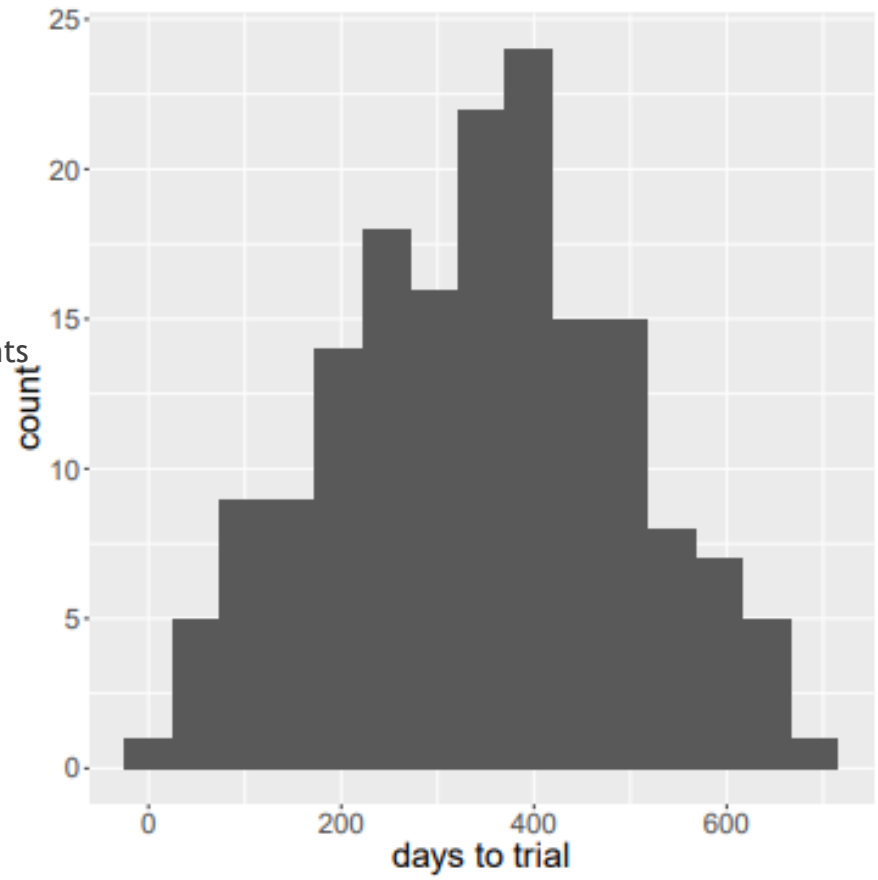
pre-trial detention

- This isn't a quick process
- If poor then you are likely to plead out
  - Get out of the violence
  - Get back to taking care of your dependents
  - Need to earn money
- Of the 80,000 cases in our data set only 200 of them went to trial



pre-trial detention

- This isn't a quick process
- If poor then you are likely to plead out
  - Get out of the violence
  - Get back to taking care of your dependents
  - Need to earn money
- Of the 80,000 cases in our data set only 200 of them went to trial
  - That is a quarter of one percent (0.25%)



the night of

to the defense

to the defense

- The Legal Aid Society provides free defense in the five boroughs of New York

to the defense

- The Legal Aid Society provides free defense in the five boroughs of New York
- They know this is a deeply unfair system, likely legally untenable

to the defense

- The Legal Aid Society provides free defense in the five boroughs of New York
- They know this is a deeply unfair system, likely legally untenable
- They've been collecting information for several years

to the defense

- The Legal Aid Society provides free defense in the five boroughs of New York
- They know this is a deeply unfair system, likely legally untenable
- They've been collecting information for several years
- In late-2016 they approached HRDAG and their lead statistician

THE LEGAL AID SOCIETY

MAKING THE CASE FOR HUMANITY

Human Rights Data Analysis Group hrdag.org

to the defense

- The Legal Aid Society provides free defense in the five boroughs of New York
- They know this is a deeply unfair system, likely legally untenable
- They've been collecting information for several years
- In late-2016 they approached HRDAG and their lead statistician

THE LEGAL AID SOCIETY
MAKING THE CASE FOR HUMANITY

Human Rights Data Analysis Group hrdag.org

- Kristian Lum: …

to the defense

- The Legal Aid Society provides free defense in the five boroughs of New York
- They know this is a deeply unfair system, likely legally untenable
- They've been collecting information for several years
- In late-2016 they approached HRDAG and their lead statistician

THE LEGAL AID SOCIETY
MAKING THE CASE FOR HUMANITY

Human Rights Data Analysis Group hrdag.org

- Kristian Lum: PhD statistician…

to the defense

- The Legal Aid Society provides free defense in the five boroughs of New York
- They know this is a deeply unfair system, likely legally untenable
- They've been collecting information for several years
- In late-2016 they approached HRDAG and their lead statistician

THE LEGAL AID SOCIETY
MAKING THE CASE FOR HUMANITY

Human Rights Data Analysis Group hrdag.org

- Kristian Lum: PhD statistician, Bayesian…

to the defense

- The Legal Aid Society provides free defense in the five boroughs of New York
- They know this is a deeply unfair system, likely legally untenable
- They've been collecting information for several years
- In late-2016 they approached HRDAG and their lead statistician

THE LEGAL AID SOCIETY
MAKING THE CASE FOR HUMANITY

Human Rights Data Analysis Group hrdag.org

- Kristian Lum: PhD statistician, Bayesian, and all around badass

what is the impact of money-bail?

what is the impact of money-bail?

- This is a tough question because pre-trial detention is determined by variables that are also linked to probability of guilt (e.g., “looks guilty”)

what is the impact of money-bail?

- This is a tough question because pre-trial detention is determined by variables that are also linked to probability of guilt (e.g., “looks guilty”)
- One way to do this is by finding a “natural experiment”

what is the impact of money-bail?

- This is a tough question because pre-trial detention is determined by variables that are also linked to probability of guilt (e.g., “looks guilty”)
- One way to do this is by finding a “natural experiment”
  - Almost all decisions have some sort of haphazardness in them

what is the impact of money-bail?

- This is a tough question because pre-trial detention is determined by variables that are also linked to probability of guilt (e.g., “looks guilty”)
- One way to do this is by finding a “natural experiment”
  - Almost all decisions have some sort of haphazardness in them
- Location – just over the border

what is the impact of money-bail?

- This is a tough question because pre-trial detention is determined by variables that are also linked to probability of guilt (e.g., “looks guilty”)
- One way to do this is by finding a “natural experiment”
  - Almost all decisions have some sort of haphazardness in them
- Location – just over the border
- Judges – before/after lunch

what is the impact of money-bail?

- This is a tough question because pre-trial detention is determined by variables that are also linked to probability of guilt (e.g., “looks guilty”)
- One way to do this is by finding a “natural experiment”
  - Almost all decisions have some sort of haphazardness in them
- Location – just over the border
- Judges – before/after lunch
- We used judges’ disposition:

what is the impact of money-bail?

- This is a tough question because pre-trial detention is determined by variables that are also linked to probability of guilt (e.g., “looks guilty”)
- One way to do this is by finding a “natural experiment”
  - Almost all decisions have some sort of haphazardness in them
- Location – just over the border
- Judges – before/after lunch
- We used judges’ disposition:
  - Some judges are “strict” and some are “lenient”

what is the impact of money-bail?

- This is a tough question because pre-trial detention is determined by variables that are also linked to probability of guilt (e.g., “looks guilty”)
- One way to do this is by finding a “natural experiment”
  - Almost all decisions have some sort of haphazardness in them
- Location – just over the border
- Judges – before/after lunch
- We used judges’ disposition:
  - Some judges are “strict” and some are “lenient”
  - Looked at historical records for a given judge, for a given charge, in a given location

what is the impact of money-bail?

- This is a tough question because pre-trial detention is determined by variables that are also linked to probability of guilt (e.g., “looks guilty”)
- One way to do this is by finding a “natural experiment”
  - Almost all decisions have some sort of haphazardness in them
- Location – just over the border
- Judges – before/after lunch
- We used judges’ disposition:
  - Some judges are “strict” and some are “lenient”
  - Looked at historical records for a given judge, for a given charge, in a given location
- This is known as an “instrumental variable” study design

near-far matching

near-far matching

- Instrumental variables has historically been an exercise in heavily parameterized modeling (e.g., two-stage least squares)

near-far matching

- Instrumental variables has historically been an exercise in heavily parameterized modeling (e.g., two-stage least squares)

Given the linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

where X_1 is an endogenous variable:

(1) Regress X_1 on Z_1, X_1, and X_3 to obtain $\hat{X}_1$

$$\hat{X}_1 = \gamma_0 + \gamma_1 Z_1 + \gamma_2 X_2 + \gamma_3 X_3 + v$$

where Z_1 is the instrumental variable

(2) Plug in the fitted values of $\hat{X}_1$ derived from equation (1) into the original linear regression equation:

$$Y = \beta_0 + \beta_1 \hat{X}_1 + \beta_2 X_2 + \beta_3 X_3 + v$$

where v is a composite error term that is uncorrelated with $\hat{X}_1, X_2$, and X_3

near-far matching

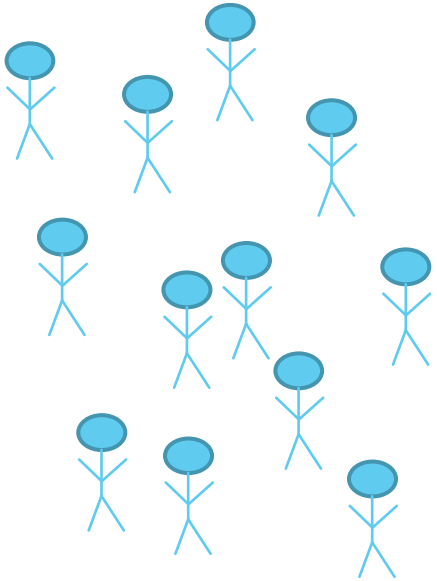
- Instrumental variables has historically been an exercise in heavily parameterized modeling (e.g., two-stage least squares)
- I trust randomized trials; can’t we do something more like RCTs?

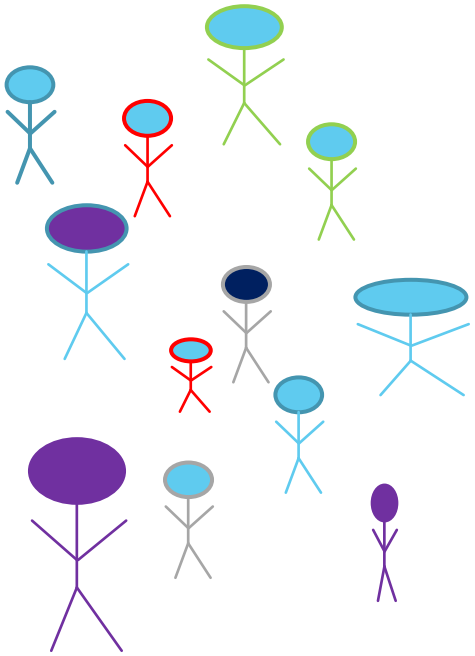
near-far matching

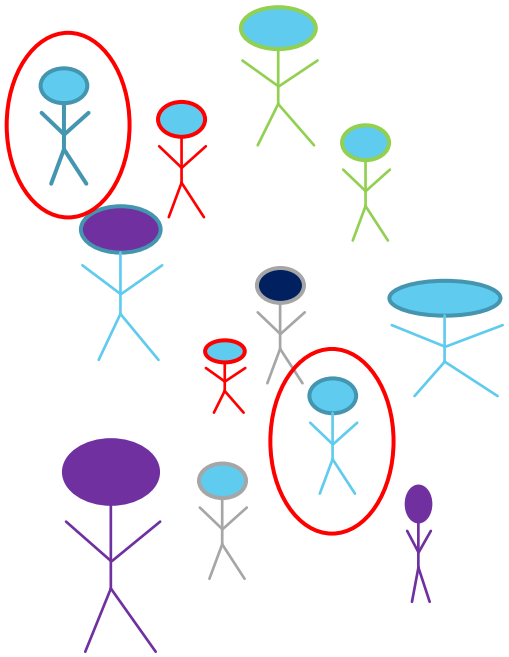
- Instrumental variables has historically been an exercise in heavily parameterized modeling (e.g., two-stage least squares)
- I trust randomized trials; can't we do something more like RCTs?
- Yeah, turns out you can:

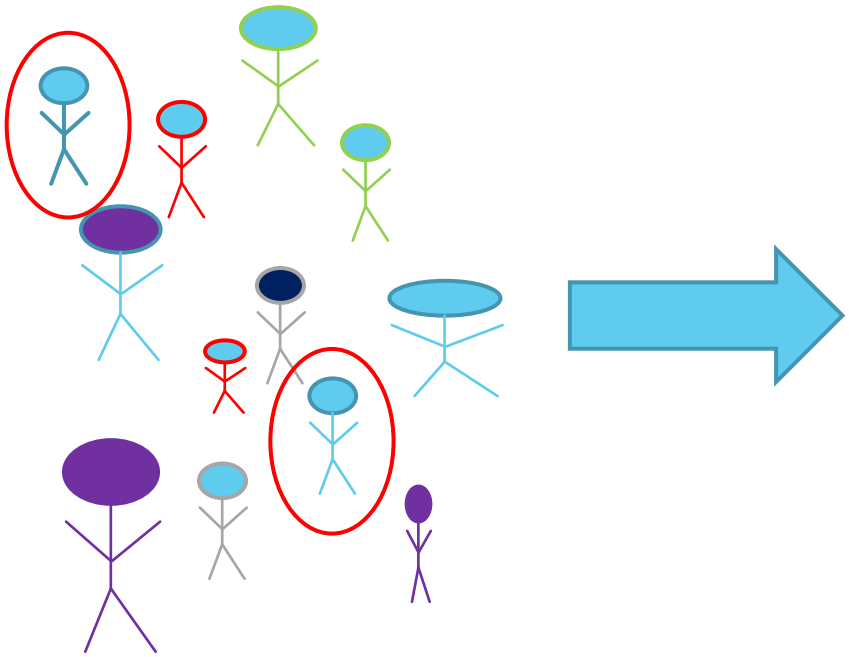
near-far matching

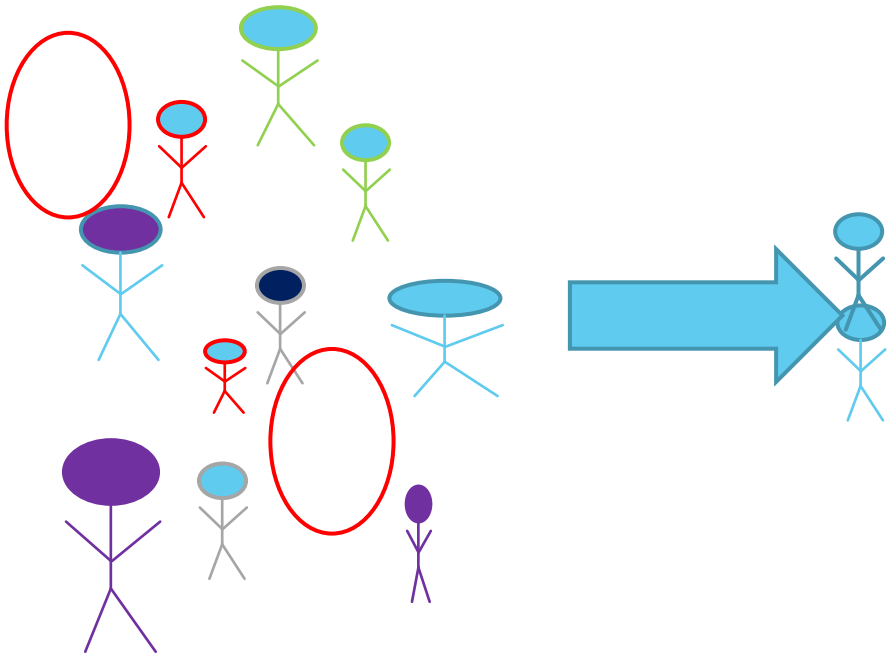
- Instrumental variables has historically been an exercise in heavily parameterized modeling (e.g., two-stage least squares)
- I trust randomized trials; can't we do something more like RCTs?
- Yeah, turns out you can: near-far matching

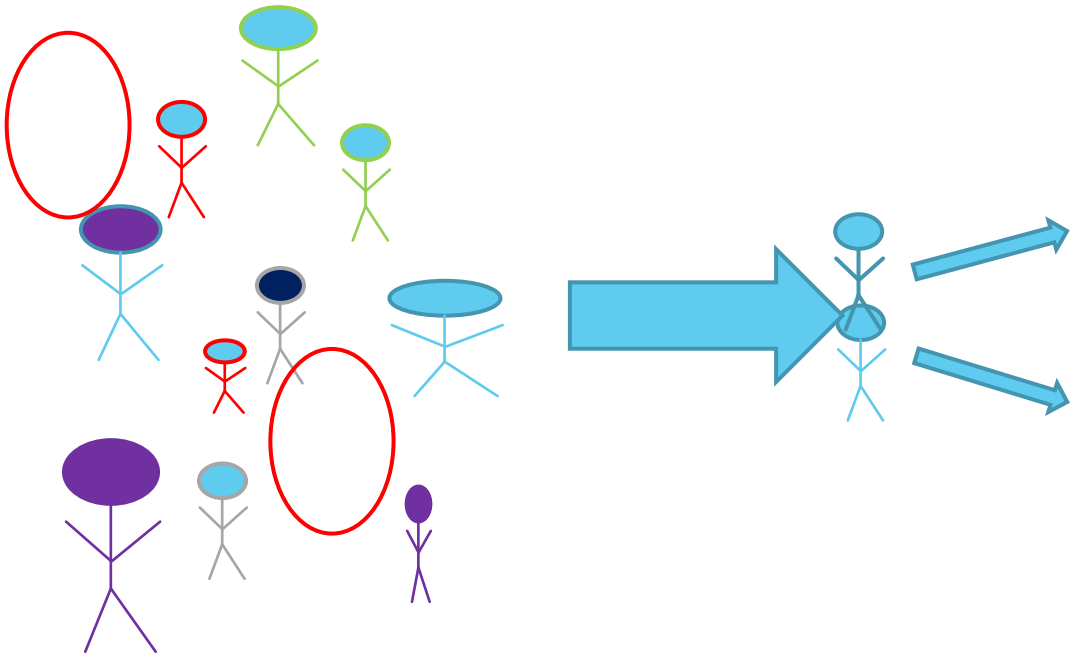


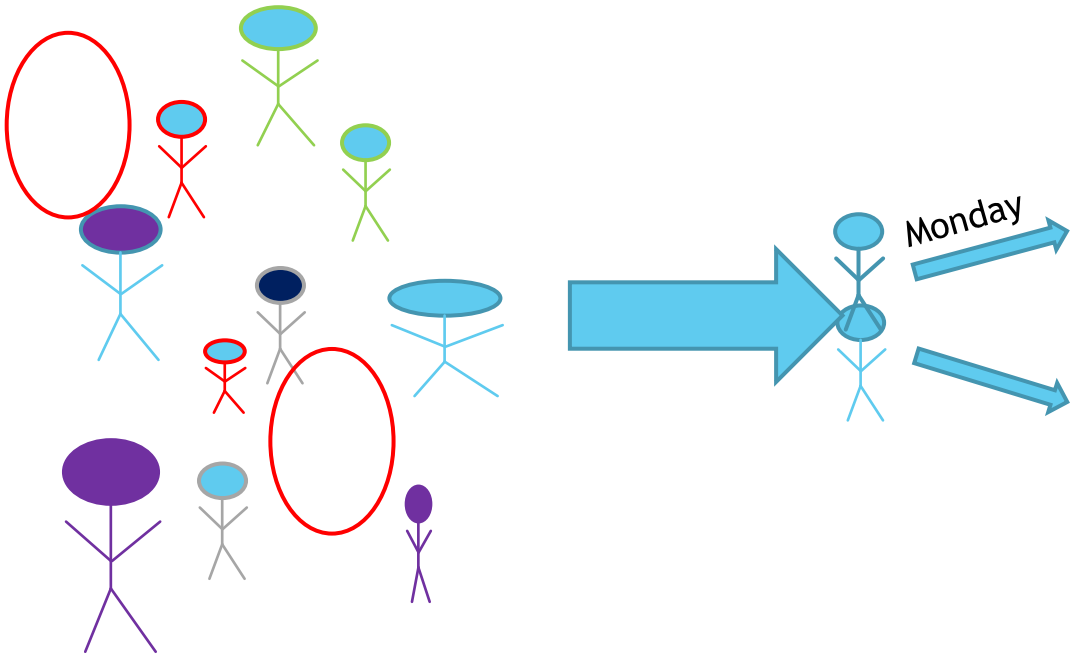


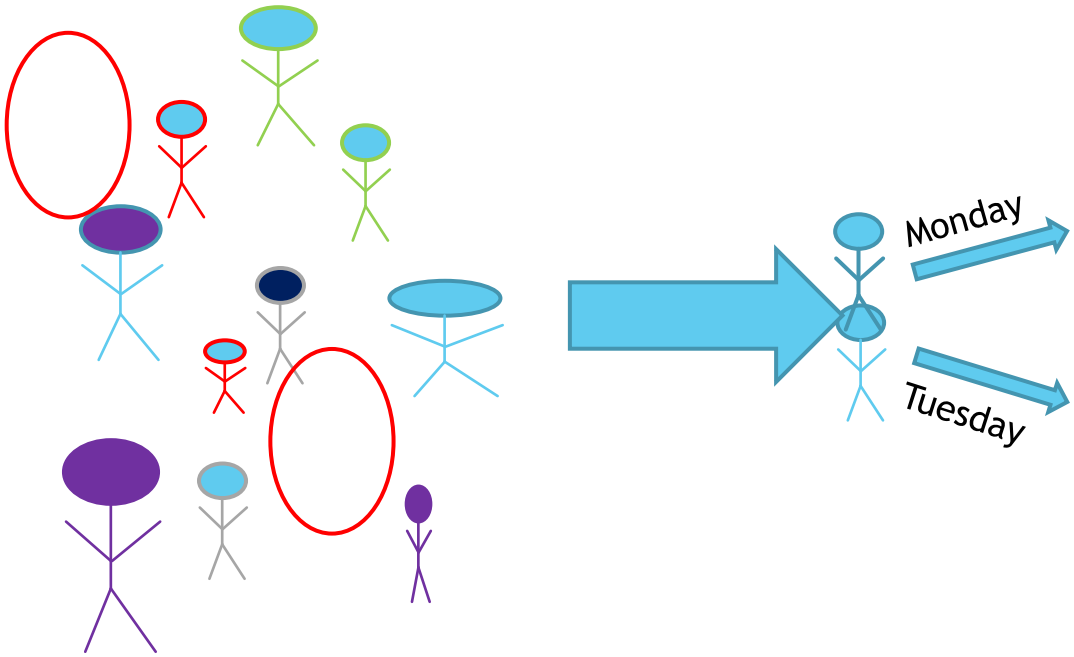


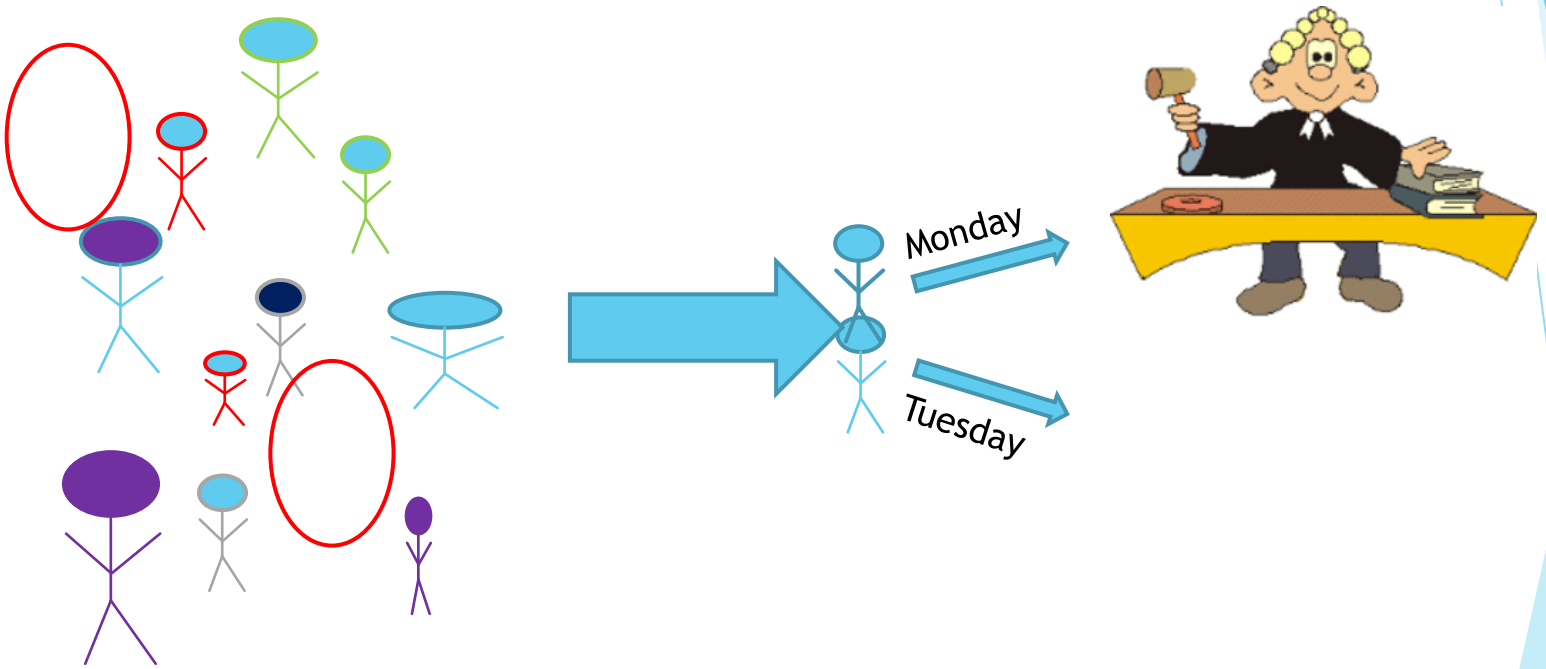


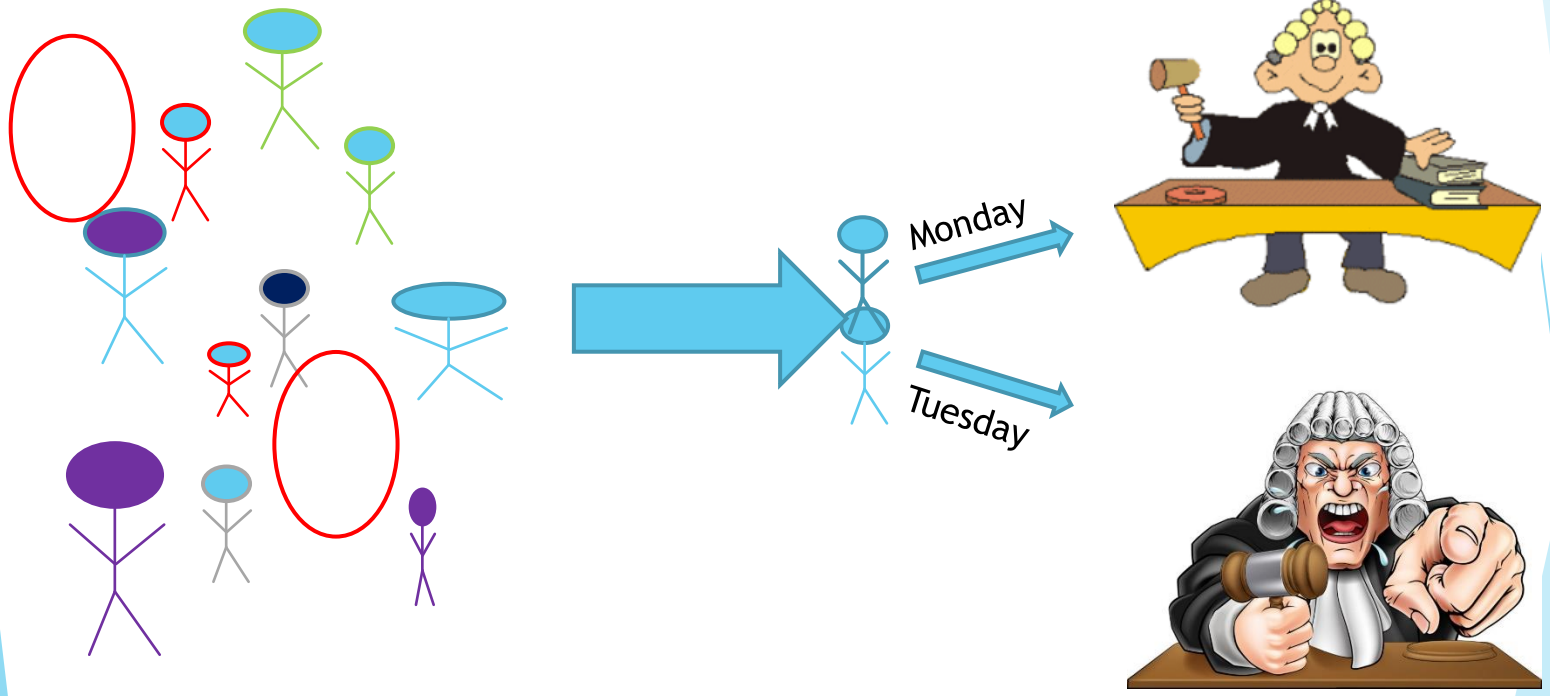


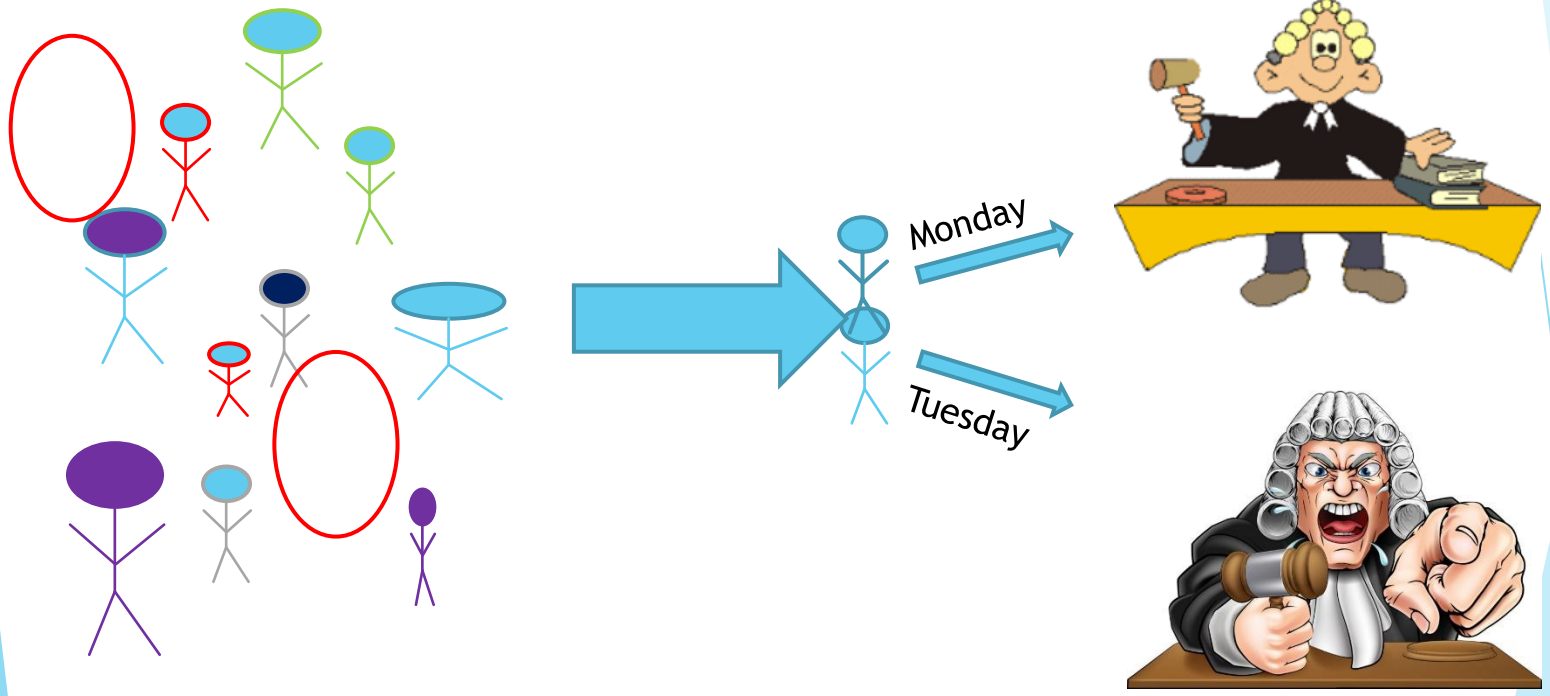



Monday

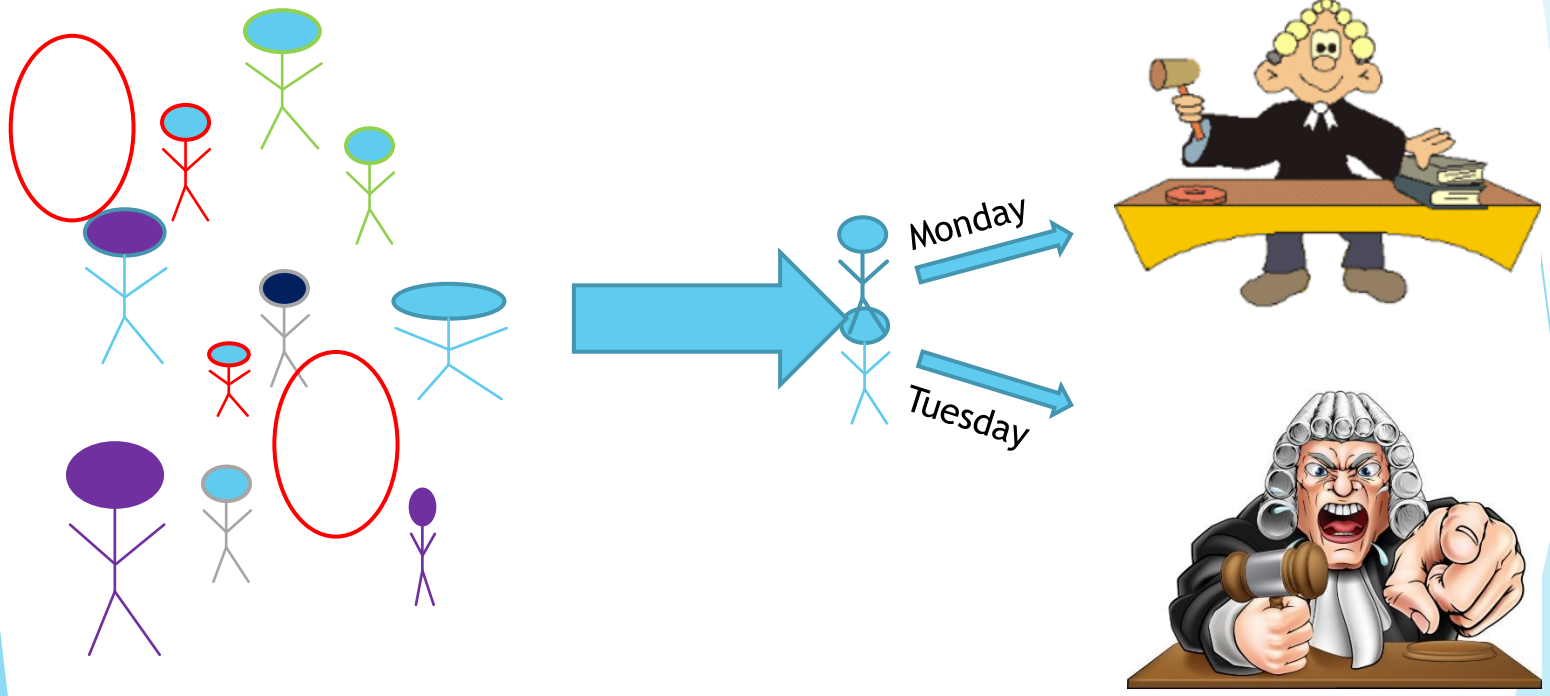

Monday
Tuesday


Monday
Tuesday


Monday
Tuesday



“near” in all the observed covariates.



“near” in all the observed covariates.

“far” in the assignment.

near-far matching

near-far matching

- Near-far matching is meant to recapture the “random” and “controlled” part of a randomized controlled experiment

near-far matching

- Near-far matching is meant to recapture the “random” and “controlled” part of a randomized controlled experiment
  - You can point to the randomness

near-far matching

- Near-far matching is meant to recapture the “random” and “controlled” part of a randomized controlled experiment
  - You can point to the randomness
  - You can “isolate” the randomness

near-far matching

- Near-far matching is meant to recapture the “random” and “controlled” part of a randomized controlled experiment
  - You can point to the randomness
  - You can “isolate” the randomness
- This is an improvement upon propensity score matching

enter:

enter: everybody

enter: everybody

- About two months into our project the floodgates opened:

enter: everybody

- About two months into our project the floodgates opened:
  - Leslie, E. and Pope, N. G. (2016). The unintended impact of pretrial detention on case outcomes: Evidence from nyc arraignments. Unpublished Working Paper.

enter: everybody

- About two months into our project the floodgates opened:
  - Leslie, E. and Pope, N. G. (2016). The unintended impact of pretrial detention on case outcomes: Evidence from nyc arraignments. Unpublished Working Paper.
  - Dobbie, W., Goldin, J., and Yang, C. (2016). The effects of pre-trial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. Technical report, National Bureau of Economic Research.

enter: everybody

- About two months into our project the floodgates opened:
  - Leslie, E. and Pope, N. G. (2016). The unintended impact of pretrial detention on case outcomes: Evidence from nyc arraignments. Unpublished Working Paper.
  - Dobbie, W., Goldin, J., and Yang, C. (2016). The effects of pre-trial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. Technical report, National Bureau of Economic Research.
  - Gupta, A., Hansman, C., and Frenchman, E. (2016). The heavy costs of high bail: Evidence from judge randomization. The Journal of Legal Studies, 45(2):471–505.

post-matching

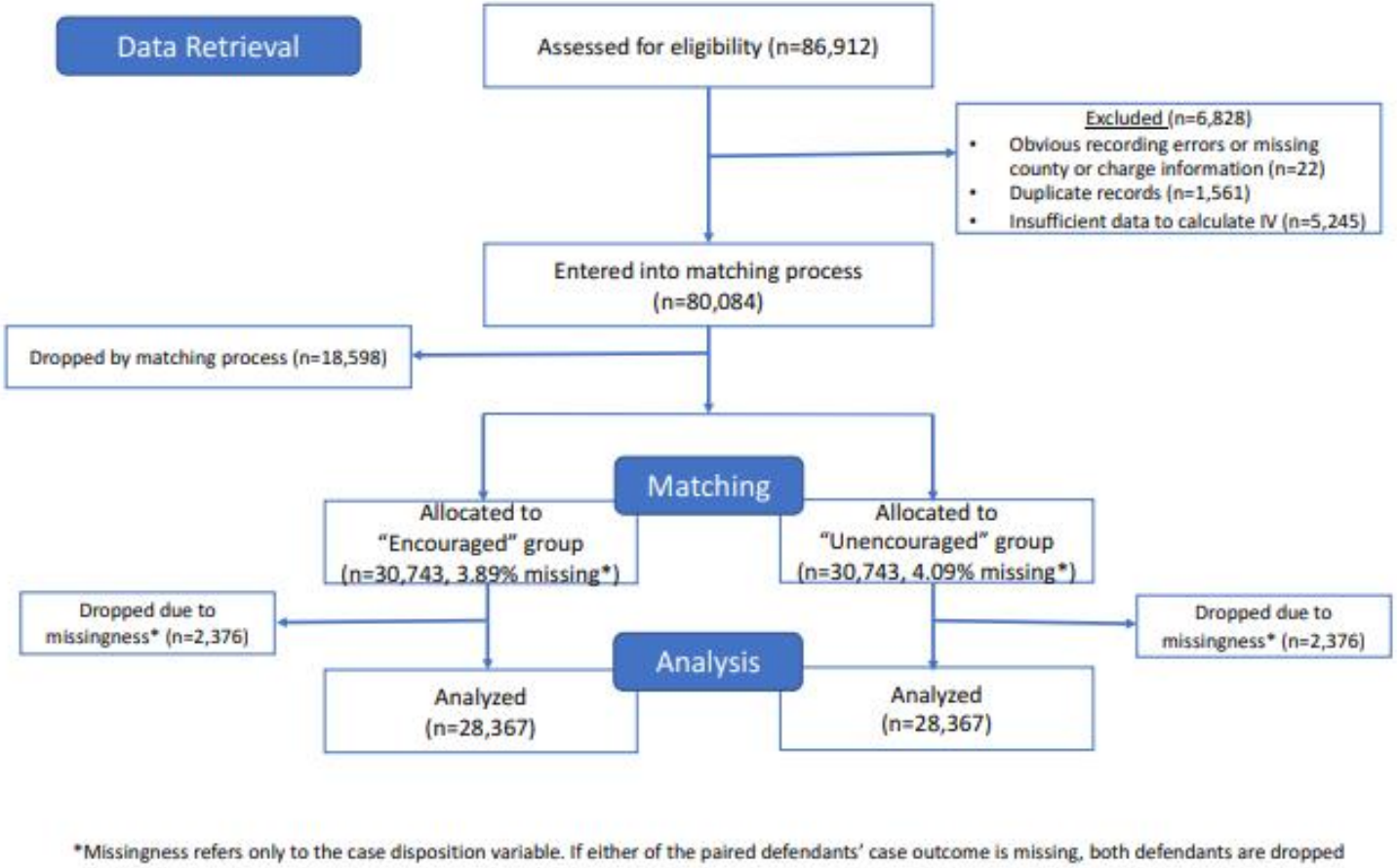


Figure 3: CONSORT flow diagram which shows the number of records dropped from the final analysis at each stage of the procedure.

pre-matching differences

	No Bail n=62826	Bail n=14057	Abs St Dif
Guilty	0.33	0.73	1.04
Bail Set	0.00	1.00	2.04
IV	0.01	-0.02	0.22
Age	31.93	35.94	0.32
White	0.29	0.28	0.02
Black	0.48	0.61	0.26
Non-Hispanic	0.63	0.66	0.08
Male	0.79	0.90	0.30
Prior Records 2014	0.38	1.36	0.70
Weekly Income	67.41	63.94	0.02
Any Income	0.15	0.14	0.02
Reported Employer	0.21	0.21	0.00
Reported Phone Number	0.18	0.19	0.04
Reported Address	0.89	0.93	0.12

Table 3: Table of pre-match standardized differences.

pre-matching differences

	No Bail n=62826	Bail n=14057	Abs St Dif
Guilty	0.33	0.73	1.04
Bail Set	0.00	1.00	2.04
IV	0.01	-0.02	0.22
Age	31.93	35.94	0.32
White	0.29	0.28	0.02
Black	0.48	0.61	0.26
Non-Hispanic	0.63	0.66	0.08
Male	0.79	0.90	0.30
Prior Records 2014	0.38	1.36	0.70
Weekly Income	67.41	63.94	0.02
Any Income	0.15	0.14	0.02
Reported Employer	0.21	0.21	0.00
Reported Phone Number	0.18	0.19	0.04
Reported Address	0.89	0.93	0.12

Table 3: Table of pre-match standardized differences.

pre-matching differences

	No Bail n=62826	Bail n=14057	Abs St Dif
Guilty	0.33	0.73	1.04
Bail Set	0.00	1.00	2.04
IV	0.01	-0.02	0.22
Age	31.93	35.94	0.32
White	0.29	0.28	0.02
Black	0.48	0.61	0.26
Non-Hispanic	0.63	0.66	0.08
Male	0.79	0.90	0.30
Prior Records 2014	0.38	1.36	0.70
Weekly Income	67.41	63.94	0.02
Any Income	0.15	0.14	0.02
Reported Employer	0.21	0.21	0.00
Reported Phone Number	0.18	0.19	0.04
Reported Address	0.89	0.93	0.12

Table 3: Table of pre-match standardized differences.

pre-matching differences

	No Bail n=62826	Bail n=14057	Abs St Dif
Guilty	0.33	0.73	1.04
Bail Set	0.00	1.00	2.04
IV	0.01	-0.02	0.22
Age	31.93	35.94	0.32
White	0.29	0.28	0.02
Black	0.48	0.61	0.26
Non-Hispanic	0.63	0.66	0.08
Male	0.79	0.90	0.30
Prior Records 2014	0.38	1.36	0.70
Weekly Income	67.41	63.94	0.02
Any Income	0.15	0.14	0.02
Reported Employer	0.21	0.21	0.00
Reported Phone Number	0.18	0.19	0.04
Reported Address	0.89	0.93	0.12

Table 3: Table of pre-match standardized differences.

post-matching

post-matching

	Strict n=28367	Lenient n=28367	Abs St Dif
Guilty	0.41	0.40	0.03
Bail Set	0.21	0.16	0.12
IV	-0.07	0.07	1.18
Age	32.69	32.71	0.00
White	0.28	0.28	0.00
Black	0.52	0.52	0.00
Non-Hispanic	0.65	0.65	0.00
Male	0.81	0.81	0.00
Prior Records 2014	0.54	0.53	0.00
Weekly Income	53.00	52.75	0.00
Any Income	0.12	0.12	0.00
Reported Employer	0.17	0.17	0.00
Reported Phone Number	0.15	0.15	0.00
Reported Address	0.91	0.91	0.00

Table 5: Table of post-match standardized differences. Summary of data analyzed.

what happens next?

results and making a difference

results

results

Stratification	Stratum	Estimate	Lower	Upper	n	Signif
Aggregate	total	0.34	0.2	0.49	56734	*
County	New York	0.43	0.23	0.63	17010	*
	Kings	0.34	0.14	0.54	17936	*
	Bronx	-0.07	-0.52	0.33	7290	
	Queens	0.66	0.13	1.35	12174	*
	Richmond	0.88	0.11	2.89	2324	*
Crime Type	Felony	0.22	-0.12	0.58	8448	
	Misdemeanor	0.37	0.22	0.53	48286	*
Gender	Male	0.31	0.16	0.45	46118	*
	Female	0.65	0.12	1.27	10532	*

Table 6: Estimated causal impact of setting bail on judicial outcome

results

Stratification	Stratum	Estimate	Lower	Upper	n	Signif
Aggregate	total	0.34	0.2	0.49	56734	*
County	New York	0.43	0.23	0.63	17010	*
	Kings	0.34	0.14	0.54	17936	*
	Bronx	-0.07	-0.52	0.33	7290	
	Queens	0.66	0.13	1.35	12174	*
	Richmond	0.88	0.11	2.89	2324	*
Crime Type	Felony	0.22	-0.12	0.58	8448	
	Misdemeanor	0.37	0.22	0.53	48286	*
Gender	Male	0.31	0.16	0.45	46118	*
	Female	0.65	0.12	1.27	10532	*

Table 6: Estimated causal impact of setting bail on judicial outcome

If you had a 5% chance of “guilty” then bail would change that to 39%.

results

- If you had a 5% chance of “guilty” then bail would change that to 39%.

results

- If you had a 5% chance of “guilty” then bail would change that to 39%.
- The Legal Aid Society: “That is unconscionable...”

results

- If you had a 5% chance of “guilty” then bail would change that to 39%.
- The Legal Aid Society: “That is unconscionable. Let’s sue the city.”

results

- If you had a 5% chance of “guilty” then bail would change that to 39%.

- The Legal Aid Society: “That is unconscionable. Let’s sue the city.”
- Kristian/HRDA: “Cool. Let’s do this.”

results

- If you had a 5% chance of “guilty” then bail would change that to 39%.

- The Legal Aid Society: “That is unconscionable. Let’s sue the city.”
- Kristian/HRDA: “Cool. Let’s do this.”
- Marshall Project:

results

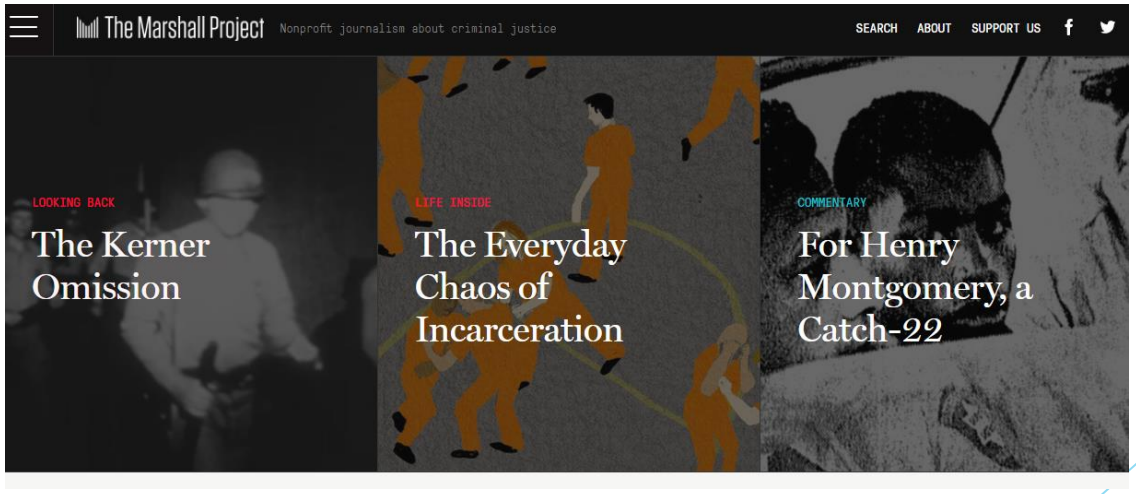
- If you had a 5% chance of “guilty” then bail would change that to 39%.

- The Legal Aid Society: “That is unconscionable. Let’s sue the city.”
- Kristian/HRDA: “Cool. Let’s do this.”
- Marshall Project: “Can we help?”

results

- If you had a 5% chance of “guilty” then bail would change that to 39%.

- The Legal Aid Society: “That is unconscionable. Let’s sue the city.”
- Kristian/HRDA: “Cool. Let’s do this.”
- Marshall Project: “Can we help?”



near-far matching

near-far matching

- Because of our study design, we can look at individuals and compare them

near-far matching

- Because of our study design, we can look at individuals and compare them
- Legal Aid has relationships with these people

near-far matching

- Because of our study design, we can look at individuals and compare them
- Legal Aid has relationships with these people
- Marshall Project has a reporter and a photo journalist assigned to this project

near-far matching

- Because of our study design, we can look at individuals and compare them
- Legal Aid has relationships with these people
- Marshall Project has a reporter and a photo journalist assigned to this project
- They interviewed 10 pairs of people

near-far matching

- Because of our study design, we can look at individuals and compare them
- Legal Aid has relationships with these people
- Marshall Project has a reporter and a photo journalist assigned to this project
- They interviewed 10 pairs of people
  - Nearly identical experiences up to arraignment

near-far matching

- Because of our study design, we can look at individuals and compare them
- Legal Aid has relationships with these people
- Marshall Project has a reporter and a photo journalist assigned to this project
- They interviewed 10 pairs of people
  - Nearly identical experiences up to arraignment
  - Far apart in their arraignment: one person had a strict judge, the other had lenient

near-far matching

- Because of our study design, we can look at individuals and compare them
- Legal Aid has relationships with these people
- Marshall Project has a reporter and a photo journalist assigned to this project
- They interviewed 10 pairs of people
  - Nearly identical experiences up to arraignment
  - Far apart in their arraignment: one person had a strict judge, the other had lenient
  - Finding out about their experiences during arraignment

near-far matching

- Because of our study design, we can look at individuals and compare them
- Legal Aid has relationships with these people
- Marshall Project has a reporter and a photo journalist assigned to this project
- They interviewed 10 pairs of people
  - Nearly identical experiences up to arraignment
  - Far apart in their arraignment: one person had a strict judge, the other had lenient
  - Finding out about their experiences during arraignment
  - Finding out about their experiences during pre-trial

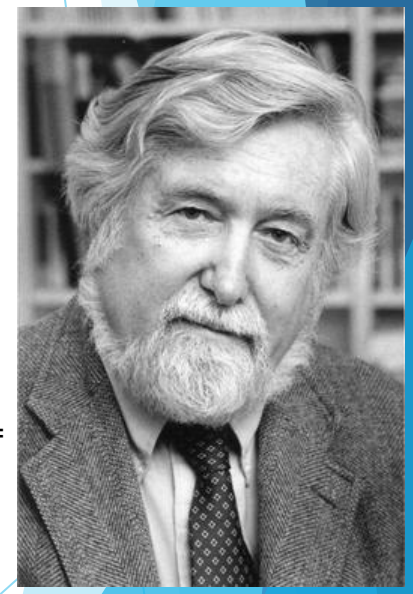
near-far matching

- Because of our study design, we can look at individuals and compare them
- Legal Aid has relationships with these people
- Marshall Project has a reporter and a photo journalist assigned to this project
- They interviewed 10 pairs of people
  - Nearly identical experiences up to arraignment
  - Far apart in their arraignment: one person had a strict judge, the other had lenient
  - Finding out about their experiences during arraignment
  - Finding out about their experiences during pre-trial
  - Finding out about their experiences in resolving the case

near-far matching

- Because of our study design, we can look at individuals and compare them
- Legal Aid has relationships with these people
- Marshall Project has a reporter and a photo journalist assigned to this project
- They interviewed 10 pairs of people
  - Nearly identical experiences up to arraignment
  - Far apart in their arraignment: one person had a strict judge, the other had lenient
  - Finding out about their experiences during arraignment
  - Finding out about their experiences during pre-trial
  - Finding out about their experiences in resolving the case

“thick description” =



Clifford Geertz

Did having bail set make you more likely to plead guilty?

Yes I believe so, because not being able to afford the bail money set, requires being remanded in county jail. Sitting in jail going back & forth to court with 2-3 months adjournments. After so many unjustified continuances time accumulating or frustrations wears out your resolve to fight for your freedom.

Did having bail set make you more likely to plead guilty?

At this point in my life No! Why? because this system has completed destroied my life, so really there's nothing left for them to take. We as a nation lock more people up for crime that have not even been committed and the blame is still on us.

Primary Contact:
Joshua Norkin
199 Water Street
New York, NY 10038
(212) 577-3509
jnorkin@legal-aid.org
www.legal-aid.org

MEMORANDUM OF SUPPORT

Comprehensive Bail Reform
A.09955 (Quart)

My bail was set at $25k, and I am a father of four, and only collect SSI disability. There is no way I could afford to take that money out of my kids mouth and still consider myself a decent man. - W.F., Legal Aid Client

Dear Legislator,

The Legal Aid Society supports A.09955 ("the Quart Bill"), and the New York State Assembly's efforts to reform our broken bail system. As you are well aware, Rikers Island and pretrial incarceration present an ongoing humanitarian crisis that requires urgent action.

In October 2017, Legal Aid released a ground-breaking report with the Human Rights Data Analysis Group and a Stanford University researcher on the impact that money bail has on our clients. The report definitively proved that New York's criminal justice system destroys the presumption of innocence. Specifically, the report found that "a strong causal relationship between setting bail and the outcome of a case for the [Legal Aid] clients. . . **setting bail results in a 34% increase in the chances that they will be found guilty.**" (Report found here: https://arxiv.org/pdf/1707.04666.pdf)

As part of this study, Legal Aid reached out to hundreds of our clients throughout New York to better understand how bail impacted their lives, their families, and their case. **Attached to this memorandum you will find 9 letters from Legal Aid clients detailing in vivid detail why the bail system must be reformed immediately.**

These stories are just one of many reasons why the Legal Aid Society strongly supports the Assembly's efforts to reform the system and the Quart Bill specifically. We remain opposed to the Governor's proposed bail legislation, which greatly expands the category of people eligible for remand, creates a presumption of detention, eliminates meaningful speedy trial release mechanisms, and permits courts to detain individuals for five days based on the mere request of the prosecutor—without any evidentiary showing or burden of proof.

Sincerely,

Justine Luongo
Attorney-in-Charge
Criminal Defense Practice
The Legal Aid Society

Sir Austin Bradford Hill

ENVIRONMENT AND DISEASE: ASSOCIATION OR CAUSATION?

Reprint with commentary

the criteria

The means of reasoning

- strength,
- consistency,
- specificity,
- temporality,
- biological gradient,
- plausibility,
- coherence,
- experiment,
- and analogy.

scorching hot takes still being served

- Phillips, C. and Goodman, K. (2004). The missed lessons of Sir Austin Bradford Hill. Epidemiologic Perspectives & Innovations, 1(1).
- Hofler, M. (2005). The Bradford Hill considerations on causality: a counterfactual perspective. Emerging Themes in Epidemiology, 2(1).
- Phillips, C. and Goodman, K. (2006). Causal criteria and counterfactuals; nothing more (or less) than scientific common sense. Emerging Themes in Epidemiology, 3(1).
- Hofler, M. (2006). Getting causal considerations back on the right track. Emerging Themes in Epidemiology, 3(1).

the criteria

- Strength: the observed association (e.g., correlation).
- Consistency: the relationship has been repeatedly observed by different persons, in different places, circumstances and times.
- Specificity: the association can be identified with particular people, locations, and outcomes.
- Temporality: cause precedes effect.
- Biological gradient: dose-response relationship.

the criteria

- Plausibility: biologically (physically, socially…) reasonable or probable.
- Coherence: does not conflict (too severely) with our present beliefs about related cause-and-effect relationships.
- Experiment: RCTs or observational studies to establish empirically rigorous associations.
- Analogy: the association is similar to existing associations, potentially sharing similar structures.

heuristics

- These are not presented as “rules” or “criteria” in the paper, rather they’re offered as ways of reasoning or moving toward taking action.
- This was a speech given to the first ever meeting of the occupational medicine group.
- A lot of times people talk about BH Criteria, they’ll introduce these as flawed. That’s fine. Impressionist paintings still help me think about light.

the criteria

- Several of these ideas have been modernized
  - Strength: sensitivity analysis
  - Consistency: reproducibility and meta-analysis
  - Specificity: treatment heterogeneity and transportability
  - Coherence: coherence statistics and the “known null” type designs
  - Analogy: transfer learning (in machine learning)

- It’s funny: he has a section in his paper where he rags on “t-tests.” Obscures decision making.

the criteria

- What is he trying to do?
- He’s trying to convince people.

Machine Learning

A NEW WAY OF REASONING ABOUT DATA

(PLAYER 2 HAS ENTERED THE GAME)

a principled prediction-problem ontology

modern algorithms

Blog \ Artifical intelligence \ Turing-NLG: A 17-billion-parameter language model by Microsoft

Turing-NLG: A 17-billion-parameter language model by Microsoft

February 13, 2020 | By Corby Rosset, Applied Scientist

How do we decide if this model is good?

do details help?

Pretraining T-NLG: Hardware and software breakthroughs

Any model with more than 1.3 billion parameters cannot fit into a single GPU (even one with 32GB of memory), so the model itself must be parallelized, or broken into pieces, across multiple GPUs. We took advantage of several hardware and software breakthroughs to achieve training T-NLG:

1. We leverage a NVIDIA DGX-2 hardware setup, with InfiniBand connections so that communication between GPUs is faster than previously achieved.

2. We apply tensor slicing to shard the model across four NVIDIA V100 GPUs on the NVIDIA Megatron-LM framework.

3. DeepSpeed with ZeRO allowed us to reduce the model-parallelism degree (from 16 to 4), increase batch size per node by fourfold, and reduce training time by three times. DeepSpeed makes training very large models more efficient with fewer GPUs, and it trains at batch size of 512 with only 256 NVIDIA GPUs compared to 1024 NVIDIA GPUs needed by using Megatron-LM alone. DeepSpeed is compatible with PyTorch.

The resulting T-NLG model has 78 Transformer layers with a hidden size of 4256 and 28 attention heads. To make results comparable to Megatron-LM, we pretrained the model with the same hyperparameters and learning schedule as Megatron-LM using autoregressive generation loss for 300,000 steps of batch size 512 on sequences of 1024 tokens. The learning schedule followed 3,200 steps of linear warmup up to a maximum learning rate of 1.5×10^{-4} and cosine decay over 500,000 steps, with FP16. We trained the model on the same type of data that Megatron-LM models were trained on.

the leaderboard

We also compared the performance of the pretrained T-NLG model on standard language tasks such as WikiText-103 perplexity (lower is better) and LAMBADA next word prediction accuracy (higher is better). The table below shows that we achieve the new state of the art on both LAMBADA and WikiText-103. Megatron-LM is the publicly released results from the NVIDIA Megatron model.

	LAMBADA (acc) strict	WikiText-103 (test adj. ppl)
Open AI GPT-2 1.5B	52.66 (63.24)*	17.48
Megatron-LM 8.3B	66.51	10.81
T-NLG 17B	67.98	10.21

*Open AI used additional processing (stopword filtering) to achieve higher numbers than the model achieved alone. Neither Megatron nor T-NLG use this stopword filtering technique.

the leaderboard

We also compared the performance of the pretrained T-NLG model on standard language tasks such as WikiText-103 perplexity (lower is better) and LAMBADA next word prediction accuracy (higher is better). The table below shows that we achieve the new state of the art on both LAMBADA and WikiText-103. Megatron-LM is the publicly released results from the NVIDIA Megatron model.

	LAMBADA (acc) strict	WikiText-103 (test adj. ppl)
Open AI GPT-2 1.5B	52.66 (63.24)*	17.48
Megatron-LM 8.3B	66.51	10.81
T-NLG 17B	67.98	10.21

*Open AI used additional processing (stopword filtering) to achieve higher numbers than the model achieved alone. Neither Megatron nor T-NLG use this stopword filtering technique.

≈ February 2019

the leaderboard

We also compared the performance of the pretrained T-NLG model on standard language tasks such as WikiText-103 perplexity (lower is better) and LAMBADA next word prediction accuracy (higher is better). The table below shows that we achieve the new state of the art on both LAMBADA and WikiText-103. Megatron-LM is the publicly released results from the NVIDIA Megatron model.

	LAMBADA (acc) strict	WikiText-103 (test adj. ppl)
Open AI GPT-2 1.5B	52.66 (63.24)*	17.48
Megatron-LM 8.3B	66.51	10.81
T-NLG 17B	67.98	10.21

*Open AI used additional processing (stopword filtering) to achieve higher numbers than the model achieved alone. Neither Megatron nor T-NLG use this stopword filtering technique.

≈ February 2019

≈ September 2019

the leaderboard

We also compared the performance of the pretrained T-NLG model on standard language tasks such as WikiText-103 perplexity (lower is better) and LAMBADA next word prediction accuracy (higher is better). The table below shows that we achieve the new state of the art on both LAMBADA and WikiText-103. Megatron-LM is the publicly released results from the NVIDIA Megatron model.

Basically every six months or so a group with a lot of money comes out with a bigger, better (?) language model

	LAMBADA (acc) strict	WikiText-103 (test adj. ppl)
Open AI GPT-2 1.5B	52.66 (63.24)*	17.48
Megatron-LM 8.3B	66.51	10.81
T-NLG 17B	67.98	10.21

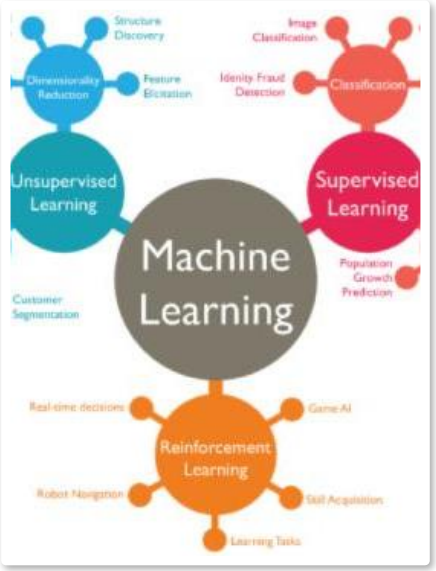
*Open AI used additional processing (stopword filtering) to achieve higher numbers than the model achieved alone. Neither Megatron nor T-NLG use this stopword filtering technique.

≈ February 2019

≈ September 2019

where we are at

prediction



setting the stage: prediction

setting the stage: prediction

- Machine learning is awesome

setting the stage: prediction

- Machine learning is awesome
  - Big wins

setting the stage: prediction

- Machine learning is awesome
  - Big wins
  - Fast wins

setting the stage: prediction

- Machine learning is awesome
  - Big wins
  - Fast wins
  - Extraordinarily complex algorithms

setting the stage: prediction

- Machine learning is awesome
  - Big wins
  - Fast wins
  - Extraordinarily complex algorithms
  - Everybody wants to use ML

setting the stage: prediction

- Machine learning is awesome
  - Big wins
  - Fast wins
  - Extraordinarily complex algorithms
  - Everybody wants to use ML

- In this talk I will (slightly incorrectly) use “prediction” and “machine learning” interchangeably.

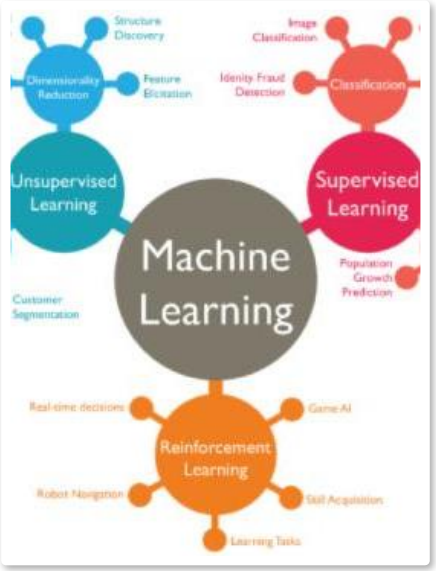
setting the stage: prediction

- Machine learning is awesome
  - Big wins
  - Fast wins
  - Extraordinarily complex algorithms
  - Everybody wants to use ML

- In this talk I will (slightly incorrectly) use “prediction” and “machine learning” interchangeably. I recognize this doesn’t acknowledge all the of the complexity of ML.

the common task framework

WTF is “the CTF”



the common task framework

- Much of the rapid success in prediction has come from the innovative infrastructure provided by the common task framework (CTF).

the common task framework

- Much of the rapid success in prediction has come from the innovative infrastructure provided by the common task framework (CTF). The CTF is the intellectual-engine of modern prediction.

the common task framework

- Much of the rapid success in prediction has come from the innovative infrastructure provided by the common task framework (CTF). The CTF is the intellectual-engine of modern prediction. One of the best places to read about the CTF is David Donoho's "50 Years of Data Science."

the common task framework

- Much of the rapid success in prediction has come from the innovative infrastructure provided by the common task framework (CTF). The CTF is the intellectual-engine of modern prediction. One of the best places to read about the CTF is David Donoho's "50 Years of Data Science."

the common task framework

- Much of the rapid success in prediction has come from the innovative infrastructure provided by the common task framework (CTF). The CTF is the intellectual-engine of modern prediction. One of the best places to read about the CTF is David Donoho's "50 Years of Data Science."

- The "Netflix Prize" is the most famous example of the CTF.

the common task framework

- Much of the rapid success in prediction has come from the innovative infrastructure provided by the common task framework (CTF). The CTF is the intellectual-engine of modern prediction. One of the best places to read about the CTF is David Donoho's "50 Years of Data Science."

- The "Netflix Prize" is the most famous example of the CTF.

the common task framework

- Much of the rapid success in prediction has come from the innovative infrastructure provided by the common task framework (CTF). The CTF is the intellectual-engine of modern prediction. One of the best places to read about the CTF is David Donoho's "50 Years of Data Science."

- The "Netflix Prize" is the most famous example of the CTF.

- The Common Task Framework:

the common task framework

- Much of the rapid success in prediction has come from the innovative infrastructure provided by the common task framework (CTF). The CTF is the intellectual-engine of modern prediction. One of the best places to read about the CTF is David Donoho's "50 Years of Data Science."

- The "Netflix Prize" is the most famous example of the CTF.

- The Common Task Framework:
  - Starts with a curated data set

the common task framework

- Much of the rapid success in prediction has come from the innovative infrastructure provided by the common task framework (CTF). The CTF is the intellectual-engine of modern prediction. One of the best places to read about the CTF is David Donoho's "50 Years of Data Science."

- The "Netflix Prize" is the most famous example of the CTF.

- The Common Task Framework:
  - Starts with a curated data set
  - Available to everybody

the common task framework

- Much of the rapid success in prediction has come from the innovative infrastructure provided by the common task framework (CTF). The CTF is the intellectual-engine of modern prediction. One of the best places to read about the CTF is David Donoho's "50 Years of Data Science."

- The "Netflix Prize" is the most famous example of the CTF.

- The Common Task Framework:
  - Starts with a curated data set
  - Available to everybody
  - A task is described (e.g., given $\boldsymbol{X}_i$ for some new *i* predict Y_i)

the common task framework

- Much of the rapid success in prediction has come from the innovative infrastructure provided by the common task framework (CTF). The CTF is the intellectual-engine of modern prediction. One of the best places to read about the CTF is David Donoho's "50 Years of Data Science."

- The "Netflix Prize" is the most famous example of the CTF.

- The Common Task Framework:
  - Starts with a curated data set
  - Available to everybody
  - A task is described (e.g., given $\boldsymbol{X}_i$ for some new *i* predict Y_i)
  - A metric for performance is given (e.g., $MSE(y_i, \hat{y}_i)$)

the common task framework

- Much of the rapid success in prediction has come from the innovative infrastructure provided by the common task framework (CTF). The CTF is the intellectual-engine of modern prediction. One of the best places to read about the CTF is David Donoho's "50 Years of Data Science."

- The "Netflix Prize" is the most famous example of the CTF.

- The Common Task Framework:
  - Starts with a curated data set
  - Available to everybody
  - A task is described (e.g., given $\boldsymbol{X}_i$ for some new *i* predict Y_i)
  - A metric for performance is given (e.g., $MSE(y_i, \hat{y}_i)$)
  - A hold out data set is reserved for evaluating performance

the common task framework

- The CTF is fast because it is low friction:

the common task framework

- The CTF is fast because it is low friction:
  - No link to sampling,

the common task framework

- The CTF is fast because it is low friction:
  - No link to sampling,
  - No link to a data generating function,

the common task framework

- The CTF is fast because it is low friction:
  - No link to sampling,
  - No link to a data generating function,
  - Ground truth is directly observable in the evaluation data,

the common task framework

- The CTF is fast because it is low friction:
  - No link to sampling,
  - No link to a data generating function,
  - Ground truth is directly observable in the evaluation data,
  - Relative performance is assessed by a targeted metric.

the common task framework

- The CTF is fast because it is low friction:
  - No link to sampling,
  - No link to a data generating function,
  - Ground truth is directly observable in the evaluation data,
  - Relative performance is assessed by a targeted metric.

- This is brilliant.

the common task framework

- The CTF is fast because it is low friction:
  - No link to sampling,
  - No link to a data generating function,
  - Ground truth is directly observable in the evaluation data,
  - Relative performance is assessed by a targeted metric.

- This is brilliant. This is stunning.

the common task framework

- The CTF is fast because it is low friction:
  - No link to sampling,
  - No link to a data generating function,
  - Ground truth is directly observable in the evaluation data,
  - Relative performance is assessed by a targeted metric.

- This is brilliant. This is stunning. This is a new type of "epistemology" – a way of knowing what is correct.

the common task framework

- The CTF is fast because it is low friction:
  - No link to sampling,
  - No link to a data generating function,
  - Ground truth is directly observable in the evaluation data,
  - Relative performance is assessed by a targeted metric.

- This is brilliant. This is stunning. This is a new type of "epistemology" – a way of knowing what is correct.

- This has allowed us to innovate faster and with fewer restrictions.

the common task framework

- The CTF is fast because it is low friction:
  - No link to sampling,
  - No link to a data generating function,
  - Ground truth is directly observable in the evaluation data,
  - Relative performance is assessed by a targeted metric.

- This is brilliant. This is stunning. This is a new type of "epistemology" – a way of knowing what is correct.

- This has allowed us to innovate faster and with fewer restrictions.

- This has led to wildly complex algorithms which do not rely upon mathematical descriptions of how the algorithms take in variation from covariates and makes use of that variation to variation in the prediction space.

the common task framework

- The CTF is fast because it is low friction:
  - No link to sampling,
  - No link to a data generating function,
  - Ground truth is directly observable in the evaluation data,
  - Relative performance is assessed by a targeted metric.

- This is brilliant. This is stunning. This is a new type of "epistemology" – a way of knowing what is correct.

- This has allowed us to innovate faster and with fewer restrictions.

- This has led to wildly complex algorithms which do not rely upon mathematical descriptions of how the algorithms take in variation from covariates and makes use of that variation to variation in the prediction space. ("**black box algorithms**")

what if something goes wrong?

the common task framework: errors post-deployment

- Without a mathematical understanding of how an algorithm links variation in X to variation in Y, we have limited options:

the common task framework: errors post-deployment

- Without a mathematical understanding of how an algorithm links variation in X to variation in Y, we have limited options:
  - Sample more data and retrain.

the common task framework: errors post-deployment

- Without a mathematical understanding of how an algorithm links variation in X to variation in Y, we have limited options:
  - Sample more data and retrain.
  - Brute force patch.

the common task framework: errors post-deployment

- Without a mathematical understanding of how an algorithm links variation in X to variation in Y, we have limited options:
  - Sample more data and retrain.
  - Brute force patch.
  - Switch to new algorithm(s).

the common task framework: errors post-deployment

- Without a mathematical understanding of how an algorithm links variation in X to variation in Y, we have limited options:
  - Sample more data and retrain.
  - Brute force patch.
  - Switch to new algorithm(s).
  - Add a human into the loop.

the common task framework: errors post-deployment

- Without a mathematical understanding of how an algorithm links variation in X to variation in Y, we have limited options:
  - Sample more data and retrain.

  - Brute force patch.
  - Switch to new algorithm(s).
  - Add a human into the loop.

the common task framework: errors post-deployment

- Without a mathematical understanding of how an algorithm links variation in X to variation in Y, we have limited options:
  - Sample more data and retrain.
    If it's black box then we don't know what parts of the covariate space are extrapolation and where to sample for maximal benefit.
  - Brute force patch.
  - Switch to new algorithm(s).
  - Add a human into the loop.

the common task framework: errors post-deployment

- Without a mathematical understanding of how an algorithm links variation in X to variation in Y, we have limited options:
  - Sample more data and retrain.
    If it's black box then we don't know what parts of the covariate space are extrapolation and where to sample for maximal benefit.
  - Brute force patch.

  - Switch to new algorithm(s).
  - Add a human into the loop.

the common task framework: errors post-deployment

- Without a mathematical understanding of how an algorithm links variation in X to variation in Y, we have limited options:
  - Sample more data and retrain.
    If it's black box then we don't know what parts of the covariate space are extrapolation and where to sample for maximal benefit.
  - Brute force patch.
    Google Image Search tried to refit but couldn't get the issue resolved.
  - Switch to new algorithm(s).

  - Add a human into the loop.

the common task framework: errors post-deployment

- Without a mathematical understanding of how an algorithm links variation in X to variation in Y, we have limited options:
  - Sample more data and retrain.
    If it's black box then we don't know what parts of the covariate space are extrapolation and where to sample for maximal benefit.
  - Brute force patch.
    Google Image Search tried to refit but couldn't get the issue resolved. They just stopped returning results.
  - Switch to new algorithm(s).

  - Add a human into the loop.

the common task framework: errors post-deployment

- Without a mathematical understanding of how an algorithm links variation in X to variation in Y, we have limited options:
  - Sample more data and retrain.
    If it's black box then we don't know what parts of the covariate space are extrapolation and where to sample for maximal benefit.
  - Brute force patch.
    Google Image Search tried to refit but couldn't get the issue resolved. They just stopped returning results.
  - Switch to new algorithm(s).
  - Add a human into the loop.

the common task framework: errors post-deployment

- Without a mathematical understanding of how an algorithm links variation in X to variation in Y, we have limited options:
  - Sample more data and retrain.
    If it's black box then we don't know what parts of the covariate space are extrapolation and where to sample for maximal benefit.
  - Brute force patch.
    Google Image Search tried to refit but couldn't get the issue resolved. They just stopped returning results.
  - Switch to new algorithm(s).
    This is not a real solution. This just postpones the question.
  - Add a human into the loop.

the common task framework: errors post-deployment

- Without a mathematical understanding of how an algorithm links variation in X to variation in Y, we have limited options:
  - Sample more data and retrain.
    If it's black box then we don't know what parts of the covariate space are extrapolation and where to sample for maximal benefit.
  - Brute force patch.
    Google Image Search tried to refit but couldn't get the issue resolved. They just stopped returning results.
  - Switch to new algorithm(s).
    This is not a real solution. This just postpones the question.
  - Add a human into the loop.

the common task framework: errors post-deployment

- Without a mathematical understanding of how an algorithm links variation in X to variation in Y, we have limited options:
  - Sample more data and retrain.
    If it's black box then we don't know what parts of the covariate space are extrapolation and where to sample for maximal benefit.

  - Brute force patch.
    Google Image Search tried to refit but couldn't get the issue resolved. They just stopped returning results.

  - Switch to new algorithm(s).
    This is not a real solution. This just postpones the question.

  - Add a human into the loop.
    Boeing 737 Super Max: the algorithm kept fighting with the pilot.

the common task framework: errors post-deployment

- Without a mathematical understanding of how an algorithm links variation in X to variation in Y, we have limited options:
  - Sample more data and retrain.
    If it's black box then we don't know what parts of the covariate space are extrapolation and where to sample for maximal benefit.

  - Brute force patch.
    Google Image Search tried to refit but couldn't get the issue resolved. They just stopped returning results.

  - Switch to new algorithm(s).
    This is not a real solution. This just postpones the question.

  - Add a human into the loop.
    Boeing 737 Super Max: the algorithm kept fighting with the pilot. They've now added an override to give the pilot complete control after a major disagreement.

have we lost control of our analytics?

old thinking becomes new

- While any particular black-box algorithm is extraordinarily hard to understand…

old thinking becomes new

- While any particular black-box algorithm is extraordinarily hard to understand, what we realized is if we understand how they're developed then we can understand when they can be deployed.

old thinking becomes new

- While any particular black-box algorithm is extraordinarily hard to understand, what we realized is if we understand how they're developed then we can understand when they can be deployed.

- The rest of the talk I'll offer a framework for guiding our thinking about whether or not a particular prediction problem is suitable for an algorithm that was developed purely under the CTF.

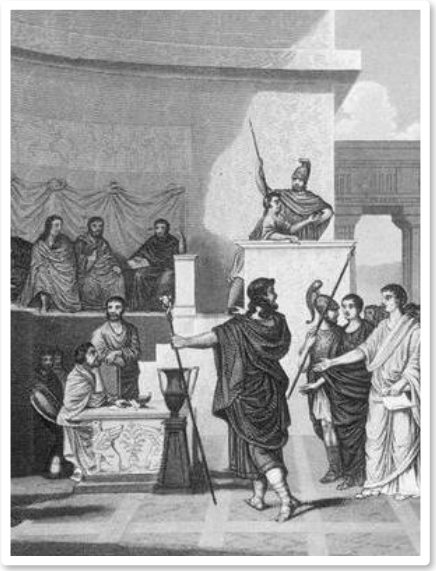
old thinking becomes new

- While any particular black-box algorithm is extraordinarily hard to understand, what we realized is if we understand how they're developed then we can understand when they can be deployed.

- The rest of the talk I'll offer a framework for guiding our thinking about whether or not a particular prediction problem is suitable for an algorithm that was developed purely under the CTF. That is, are the conditions right for a black box algorithm to be deployed into the real-world, where there are real consequences?

the principled prediction problem ontology

a framework for prediction



the principled prediction-problem ontology

the principled prediction-problem ontology

- When describing a prediction problem there are four features that must be satisfied in order for a black box algorithm to be deployed:

the principled prediction-problem ontology

- When describing a prediction problem there are four features that must be satisfied in order for a black box algorithm to be deployed:
  1.
  2.
  3.
  4.

the principled prediction-problem ontology

- When describing a prediction problem there are four features that must be satisfied in order for a black box algorithm to be deployed:
  1. [**measurement**] Can measure a function of the outcomes.
  2.
  3.
  4.

the principled prediction-problem ontology

- When describing a prediction problem there are four features that must be satisfied in order for a black box algorithm to be deployed:
  1. [**measurement**] Can measure a function of the outcomes.
  2. [**adaptability**] Can update on a useful timescale.
  3.
  4.

the principled prediction-problem ontology

- When describing a prediction problem there are four features that must be satisfied in order for a black box algorithm to be deployed:
  1. [**measurement**] Can measure a function of the outcomes.
  2. [**adaptability**] Can update on a useful timescale.
  3. [**resilience**] Have a high tolerance for accumulated errors.
  4.

the principled prediction-problem ontology

- When describing a prediction problem there are four features that must be satisfied in order for a black box algorithm to be deployed:
  1. [**measurement**] Can measure a function of the outcomes.
  2. [**adaptability**] Can update on a useful timescale.
  3. [**resilience**] Have a high tolerance for accumulated errors.
  4. [**agnosis**] Have a high tolerance for disagreement with beliefs (e.g., scientific knowledge, ethical concerns, intuitions).

the principled prediction-problem ontology

- When describing a prediction problem there are four features that must be satisfied in order for a black box algorithm to be deployed:
  1. [**measurement**] Can measure a function of the outcomes.
  2. [**adaptability**] Can update on a useful timescale.
  3. [**resilience**] Have a high tolerance for accumulated errors.
  4. [**agnosis**] Have a high tolerance for disagreement with beliefs (e.g., scientific knowledge, ethical concerns, intuitions).

  I'll refer to these collectively as MARA.

the principled prediction-problem ontology

- When describing a prediction problem there are four features that must be satisfied in order for a black box algorithm to be deployed:
  1. [**measurement**] Can measure a function of the outcomes.
  2. [**adaptability**] Can update on a useful timescale.
  3. [**resilience**] Have a high tolerance for accumulated errors.
  4. [**agnosis**] Have a high tolerance for disagreement with beliefs (e.g., scientific knowledge, ethical concerns, intuitions).

  I'll refer to these collectively as MARA.

  Given its dependence on stakeholders, we abbreviate: MARA(s)

MARA

examples

2) For some positive real number 'b',
$b - 1$, $b + 4$, $3b + 2$. What is the

a) 16
b) 20
c) 24
d) 28
e) 40

MARA:

MARA: example 1

MARA: example 1 – ad placement

MARA: example 1 – ad placement

1. measurement
2. adaptability
3. resilience
4. agnosis

MARA: example 1 – ad placement

Amazon wants to sell products A or B.

1. measurement
2. adaptability
3. resilience
4. agnosis

MARA: example 1 – ad placement

Amazon wants to sell products A or B.

Can show one ad at a time to one customer.

1. measurement
2. adaptability
3. resilience
4. agnosis

MARA: example 1 – ad placement

Amazon wants to sell products A or B.

Can show one ad at a time to one customer.

Which ad should they show?

1. measurement
2. adaptability
3. resilience
4. agnosis

MARA: example 1 – ad placement

Amazon wants to sell products A or B.

Can show one ad at a time to one customer.

Which ad should they show?

Can observe if people buy A or B.

1. measurement
2. adaptability
3. resilience
4. agnosis

MARA: example 1 – ad placement

Amazon wants to sell products A or B.

Can show one ad at a time to one customer.

Which ad should they show?

Can observe if people buy A or B.

Tons of customers come through and are not rapidly changing their purchasing habits.

1. measurement
2. adaptability
3. resilience
4. agnosis

MARA: example 1 – ad placement

Amazon wants to sell products A or B.

Can show one ad at a time to one customer.

Which ad should they show?

Can observe if people buy A or B.

Tons of customers come through and are not rapidly changing their purchasing habits.

This ad placement is helpful but many people will still buy regardless.

1. measurement
2. adaptability
3. resilience
4. agnosis

MARA: example 1 – ad placement

Amazon wants to sell products A or B.

Can show one ad at a time to one customer.

Which ad should they show?

Can observe if people buy A or B.

Tons of customers come through and are not rapidly changing their purchasing habits.

This ad placement is helpful but many people will still buy regardless.

No real concerns beyond making that sale.

1. measurement
2. adaptability
3. resilience
4. agnosis

MARA: example 1 – ad placement

Amazon wants to sell products A or B.

Can show one ad at a time to one customer.

Which ad should they show?

Can observe if people buy A or B.

Tons of customers come through and are not rapidly changing their purchasing habits.

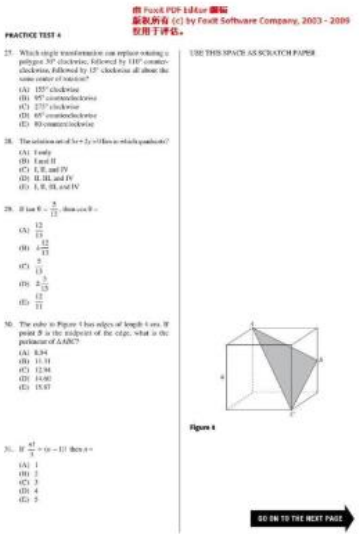
This ad placement is helpful but many people will still buy regardless.

No real concerns beyond making that sale.

1. measurement
   yes
2. adaptability
   yes
3. resilience
   yes
4. agnosis
   yes

MARA: example 2 – educational

Have a “mechanical math tutor” that can adaptively select practice questions for a student.



Foxit PDF Editor 编辑
版权所有 (c) by Foxit Software Company, 2003 - 2009
仅用于评估。

PRACTICE TEST 4

27. Which single transformation can replace rotating a polygon 30° clockwise, followed by 110° counter-clockwise, followed by 15° clockwise all about the same center of rotation?
(A) 155° clockwise
(B) 95° counterclockwise
(C) 235° clockwise
(D) 65° counterclockwise
(E) 80 counterclockwise

28. The solution set of $5x + 2y > 10$ lies in which quadrant(s)?
(A) I only
(B) I and II
(C) I, II, and IV
(D) II, III, and IV
(E) I, II, III, and IV

29. If $\tan\theta = -\frac{5}{12}$, then $\cos\theta =$
(A) $\frac{12}{13}$
(B) $\pm\frac{12}{13}$
(C) $\frac{5}{13}$
(D) $\pm\frac{5}{13}$
(E) $\frac{12}{11}$

30. The cube in Figure 4 has edges of length 4 cm. If point B is the midpoint of the edge, what is the perimeter of $\triangle ABC$?
(A) 8.94
(B) 11.31
(C) 12.94
(D) 14.00
(E) 15.07

31. If $\frac{x^2}{3} = (x-1)^2$ then $x =$
(A) 1
(B) 2
(C) 3
(D) 4
(E) 5

USE THIS SPACE AS SCRATCH PAPER

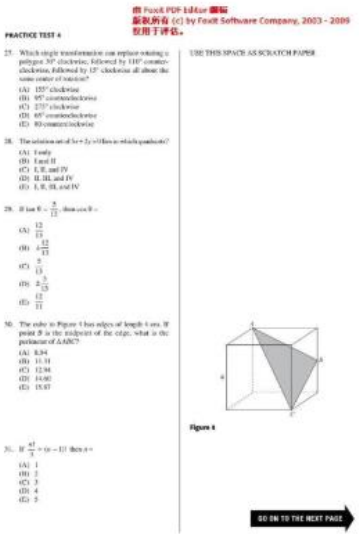
Figure 4

GO ON TO THE NEXT PAGE

MARA: example 2 – educational

Have a “mechanical math tutor” that can adaptively select practice questions for a student.

Can look at if scores improve, and/or time spent interacting with the tutor.



PRACTICE TEST 4

27. Which single transformation can replace rotating a polygon 30° clockwise, followed by 110° counter-clockwise, followed by 15° clockwise all about the same center of rotation?
(A) 155° clockwise
(B) 95° counterclockwise
(C) 235° clockwise
(D) 65° counterclockwise
(E) 80 counterclockwise

28. The solution set of $5x + 2y > 10$ lies in which quadrants?
(A) I only
(B) I and II
(C) I, II, and IV
(D) II, III, and IV
(E) I, II, III, and IV

29. If $\tan\theta = -\frac{5}{12}$, then $\cos\theta =$
(A) $\frac{12}{13}$
(B) $\pm\frac{12}{13}$
(C) $\frac{5}{13}$
(D) $\pm\frac{5}{13}$
(E) $-\frac{12}{13}$

30. The cube in Figure 4 has edges of length 4 cm. If point B is the midpoint of the edge, what is the perimeter of $\triangle ABC$?
(A) 8.94
(B) 11.31
(C) 12.94
(D) 14.00
(E) 15.07

31. If $\frac{x^2}{3} = (x-1)^2$ then $x =$
(A) 1
(B) 2
(C) 3
(D) 4
(E) 5

USE THIS SPACE AS SCRATCH PAPER

Figure 4

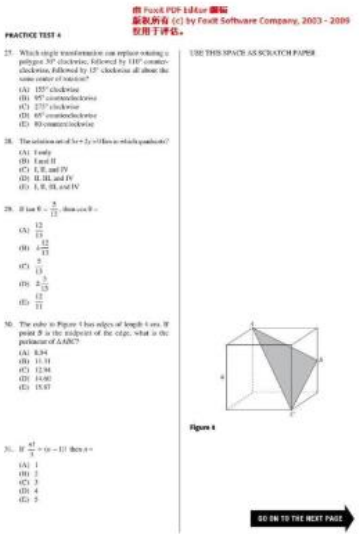
GO ON TO THE NEXT PAGE

MARA: example 2 – educational

Have a “mechanical math tutor” that can adaptively select practice questions for a student.

Can look at if scores improve, and/or time spent interacting with the tutor.

The student is likely to learn and adapt, but not so fast that an algorithm can’t be adapted as well.

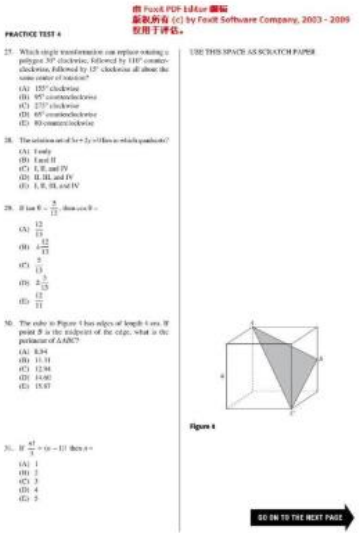


MARA: example 2 – educational

Have a “mechanical math tutor” that can adaptively select practice questions for a student.

Can look at if scores improve, and/or time spent interacting with the tutor.

The student is likely to learn and adapt, but not so fast that an algorithm can’t be adapted as well.



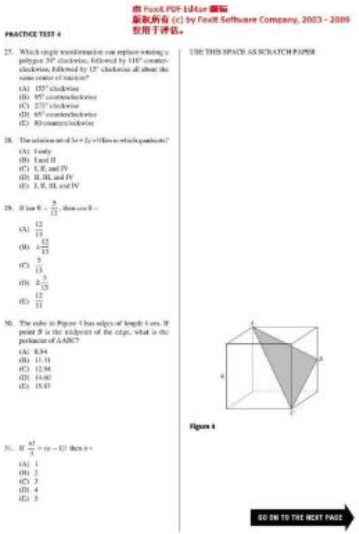
Practice problems are still useful, we’re just looking to be more efficient.

MARA: example 2 – educational

Have a “mechanical math tutor” that can adaptively select practice questions for a student.

Can look at if scores improve, and/or time spent interacting with the tutor.

The student is likely to learn and adapt, but not so fast that an algorithm can’t be adapted as well.

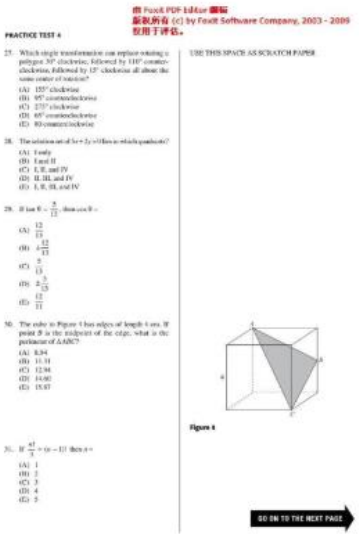


Practice problems are still useful, we’re just looking to be more efficient.

As long as the student is benefiting we’re ok.

MARA: example 2 – educational

Have a “mechanical math tutor” that can adaptively select practice questions for a student.



Can look at if scores improve, and/or time spent interacting with the tutor.

The student is likely to learn and adapt, but not so fast that an algorithm can’t be adapted as well.

Practice problems are still useful, we’re just looking to be more efficient.

As long as the student is benefiting we’re ok. In fact, some might argue that learning is so complicated that our beliefs are so bad as to mislead.

MARA: example 2 – educational

Have a “mechanical math tutor” that can adaptively select practice questions for a student.

Can look at if scores improve, and/or time spent interacting with the tutor.

The student is likely to learn and adapt, but not so fast that an algorithm can’t be adapted as well.

1. measurement
   yes
2. adaptability
   yes
3. resilience
   yes
4. agnosis
   yes

Practice problems are still useful, we’re just looking to be more efficient.

As long as the student is benefiting we’re ok. In fact, some might argue that learning is so complicated that our beliefs are so bad as to mislead.

MARA: example 3 – high-frequency finance

Rapidly want to trade.

MARA: example 3 – high-frequency finance

Rapidly want to trade.

Often estimating the covariance matrix of assets.
But can approximate with if we earn money.

MARA: example 3 – high-frequency finance

Rapidly want to trade.

Often estimating the covariance matrix of assets. But can approximate with if we earn money.

Winning and losing on a bet is part of finance. Well-studied problem.

MARA: example 3 – high-frequency finance

Rapidly want to trade.

Often estimating the covariance matrix of assets. But can approximate with if we earn money.

Winning and losing on a bet is part of finance. Well-studied problem.

Ethics?

MARA: example 3 – high-frequency finance

Rapidly want to trade.

Often estimating the covariance matrix of assets. But can approximate with if we earn money.

Winning and losing on a bet is part of finance. Well-studied problem.

Ethics? Lolz.

MARA: example 3 – high-frequency finance

Rapidly want to trade.

Often estimating the covariance matrix of assets. But can approximate with if we earn money.

Winning and losing on a bet is part of finance. Well-studied problem.

Ethics? Lolz.

1. measurement
   yes
2. adaptability
   yes
3. resilience
   yes
4. agnosis
   yes

MARA: example 3 – high-frequency finance

Rapidly want to trade.

Often estimating the covariance matrix of assets. But can approximate with if we earn money.

Winning and losing on a bet is part of finance. Well-studied problem.

Ethics? Lolz.

1. measurement
   yes
2. adaptability
   yes
3. resilience
   yes
4. agnosis
   yes

But what if we change stakeholders?

MARA: example 3 – family finance

Watch to see if the algorithm earns money.

MARA: example 3 – family finance

Watch to see if the algorithm earns money.

Error tolerance is likely less than for a high-frequency trading firm.

MARA: example 3 – family finance

Watch to see if the algorithm earns money.

Error tolerance is likely less than for a high-frequency trading firm.

The biggest issue is that the family may want to have the algorithm explained to them: (a)

(b)

MARA: example 3 – family finance

Watch to see if the algorithm earns money.

Error tolerance is likely less than for a high-frequency trading firm.

The biggest issue is that the family may want to have the algorithm explained to them: (a) does it match their beliefs about the future of the markets (e.g., bearish) (b)

MARA: example 3 – family finance

Watch to see if the algorithm earns money.

Error tolerance is likely less than for a high-frequency trading firm.

The biggest issue is that the family may want to have the algorithm explained to them: (a) does it match their beliefs about the future of the markets (e.g., bearish) (b) is it making “ethical” investments (e.g., not buying guns).

MARA: example 3 – family finance

Watch to see if the algorithm earns money.

Error tolerance is likely less than for a high-frequency trading firm.

The biggest issue is that the family may want to have the algorithm explained to them: (a) does it match their beliefs about the future of the markets (e.g., bearish) (b) is it making “ethical” investments (e.g., not buying guns).

1. measurement
   yes
2. adaptability
   yes
3. resilience
   no
4. agnosis
   no

MARA: example 3 – family finance

Watch to see if the algorithm earns money.

Error tolerance is likely less than for a high-frequency trading firm.

The biggest issue is that the family may want to have the algorithm explained to them: (a) does it match their beliefs about the future of the markets (e.g., bearish) (b) is it making “ethical” investments (e.g., not buying guns).

1. measurement
   yes
2. adaptability
   yes
3. resilience
   no
4. agnosis
   no

Changing stakeholders can change the loadings of the problem-features.

MARA: example 4 – elections

United States elections happen every four years.

MARA: example 4 – elections

United States elections happen every four years.

I’m very nervous about who will win
the election in 2020.

MARA: example 4 – elections

United States elections happen every four years.

I’m very nervous about who will win
the election in 2020.

MARA: example 4 – elections

United States elections happen every four years.

I'm very nervous about who will win
the election in 2020.

The forces that generate voters change dramatically
from election to election

MARA: example 4 – elections

United States elections happen every four years.

I’m very nervous about who will win
the election in 2020.

The forces that generate voters change dramatically
from election to election (e.g., candidate personalities,
trade wars with powerful countries, real wars, bad hair cuts).

MARA: example 4 – elections

United States elections happen every four years.

I'm very nervous about who will win
the election in 2020.

The forces that generate voters change dramatically
from election to election (e.g., candidate personalities,
trade wars with powerful countries, real wars, bad hair cuts).

An election prediction is a one-shot prediction. You can't adapt the algo.

MARA: example 4 – elections

United States elections happen every four years.

I'm very nervous about who will win the election in 2020.

The forces that generate voters change dramatically from election to election (e.g., candidate personalities, trade wars with powerful countries, real wars, bad hair cuts).

An election prediction is a one-shot prediction. You can't adapt the algo.

1. measurement
   yes
2. adaptability
   no
3. resilience
   -
4. agnosis
   -

MARA: example 5 – twitter & text

MARA: example 5 – twitter & text

Want to understand which politicians are talking about issues related to the #MeToo movement.

MARA: example 5 – twitter & text

Want to understand which politicians are talking about issues related to the #MeToo movement.

There are thousands of tweets that need to be analyzed.

MARA: example 5 – twitter & text

Want to understand which politicians are talking about issues related to the #MeToo movement.

There are thousands of tweets that need to be analyzed.

We could have undergrads hand code all the tweets, but that's costly and error-prone.

MARA: example 5 – twitter & text

Want to understand which politicians are talking about issues related to the #MeToo movement.

There are thousands of tweets that need to be analyzed.

We could have undergrads hand code all the tweets, but that's costly and error-prone.

Want to use an algorithm to do the labeling quickly and efficiently.

MARA: example 5 – twitter & text

Want to understand which politicians are talking about issues related to the #MeToo movement.

There are thousands of tweets that need to be analyzed.

We could have undergrads hand code all the tweets, but that's costly and error-prone.

Want to use an algorithm to do the labeling quickly and efficiently.

But…. we'll never know if the label is correct.

MARA: example 5 – twitter & text

Want to understand which politicians are talking about issues related to the #MeToo movement.

There are thousands of tweets that need to be analyzed.

We could have undergrads hand code all the tweets, but that's costly and error-prone.

Want to use an algorithm to do the labeling quickly and efficiently.

But…. we'll never know if the label is correct. The whole point was to understand the tweets.

MARA: example 5 – twitter & text

Want to understand which politicians are talking about issues related to the #MeToo movement.

There are thousands of tweets that need to be analyzed.

We could have undergrads hand code all the tweets, but that's costly and error-prone.

Want to use an algorithm to do the labeling quickly and efficiently.

But…. we'll never know if the label is correct. The whole point was to understand the tweets. The point of deploying the algorithm is directly in competition with observing outcomes.

MARA: example 5 – twitter & text

Want to understand which politicians are talking about issues related to the #MeToo movement.

There are thousands of tweets that need to be analyzed.

We could have undergrads hand code all the tweets, but that's costly and error-prone.

Want to use an algorithm to do the labeling quickly and efficiently.

But…. we'll never know if the label is correct. The whole point was to understand the tweets. The point of deploying the algorithm is directly in competition with observing outcomes.

1. measurement
   no
2. adaptability
   -
3. resilience
   -
4. agnosis
   -

MARA: example 5 – twitter & text

Want to understand which politicians are talking about issues related to the #MeToo movement.

There are thousands of tweets that need to be analyzed.

We could have undergrads hand code all the tweets, but that's costly and error-prone.

1. measurement
   no
2. adaptability
   -
3. resilience
   -
4. agnosis
   -

Want to use an algorithm to do the labeling quickly and efficiently.

But…. we'll never know if the label is correct. The whole point was to understand the tweets. The point of deploying the algorithm is directly in competition with observing outcomes.

This is common in text mining, but also "triaging" settings.

MARA: example 6 – recidivism

Should a criminal be allowed out early?

MARA: example 6 – recidivism

Should a criminal be allowed out early?
If stakeholders are judges and politicians:

MARA: example 6 – recidivism

Should a criminal be allowed out early?
If stakeholders are judges and politicians:
(i) Can't view outcomes with a hard cutoff point.
(ii) Crime patterns may stay the same.
(iii) We get things wrong, but maybe algorithms can do better.
(iv) Serious ethical issues.
What if algorithm tells you to lock up black men?

MARA: example 6 – recidivism

Should a criminal be allowed out early?
If stakeholders are judges and politicians:
(i) Can't view outcomes with a hard cutoff point.
(ii) Crime patterns may stay the same.
(iii) We get things wrong, but maybe algorithms can do better.
(iv) Serious ethical issues.
What if algorithm tells you to lock up black men?

If stakeholders include the accused as well:

MARA: example 6 – recidivism

Should a criminal be allowed out early?
If stakeholders are judges and politicians:
(i) Can't view outcomes with a hard cutoff point.
(ii) Crime patterns may stay the same.
(iii) We get things wrong, but maybe algorithms can do better.
(iv) Serious ethical issues.
What if algorithm tells you to lock up black men?

If stakeholders include the accused as well:
(i) Won't observe if the algorithm got it wrong if locked up.
(ii) One-shot prediction.
(iii) Massive consequences to an error.
(iv) Need to be able to explain why the algorithm is depriving the accused person of his/her freedom.

MARA: example 6 – recidivism

Should a criminal be allowed out early?
If stakeholders are judges and politicians:
(i) Can't view outcomes with a hard cutoff point.
(ii) Crime patterns may stay the same.
(iii) We get things wrong, but maybe algorithms can do better.
(iv) Serious ethical issues.
What if algorithm tells you to lock up black men?

If stakeholders include the accused as well:
(i) Won't observe if the algorithm got it wrong if locked up.
(ii) One-shot prediction.
(iii) Massive consequences to an error.
(iv) Need to be able to explain why the algorithm is depriving the accused person of his/her freedom.

These analyses led to "algorithmic fairness."

MARA: example 6 – recidivism

Should a criminal be allowed out early?
If stakeholders are judges and politicians:
(i) Can't view outcomes with a hard cutoff point.
(ii) Crime patterns may stay the same.
(iii) We get things wrong, but maybe algorithms can do better.
(iv) Serious ethical issues.
What if algorithm tells you to lock up black men?

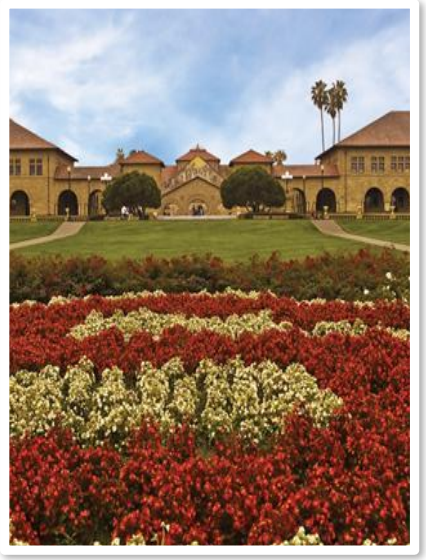
1. measurement
   maybe
2. adaptability
   yes
3. resilience
   no
4. agnosis
   no

If stakeholders include the accused as well:
(i) Won't observe if the algorithm got it wrong if locked up.
(ii) One-shot prediction.
(iii) Massive consequences to an error.
(iv) Need to be able to explain why the algorithm is depriving the accused person of his/her freedom.

These analyses led to "algorithmic fairness."

reasoning

outcome- vs model- reasoning



MARA: two types of reasoning

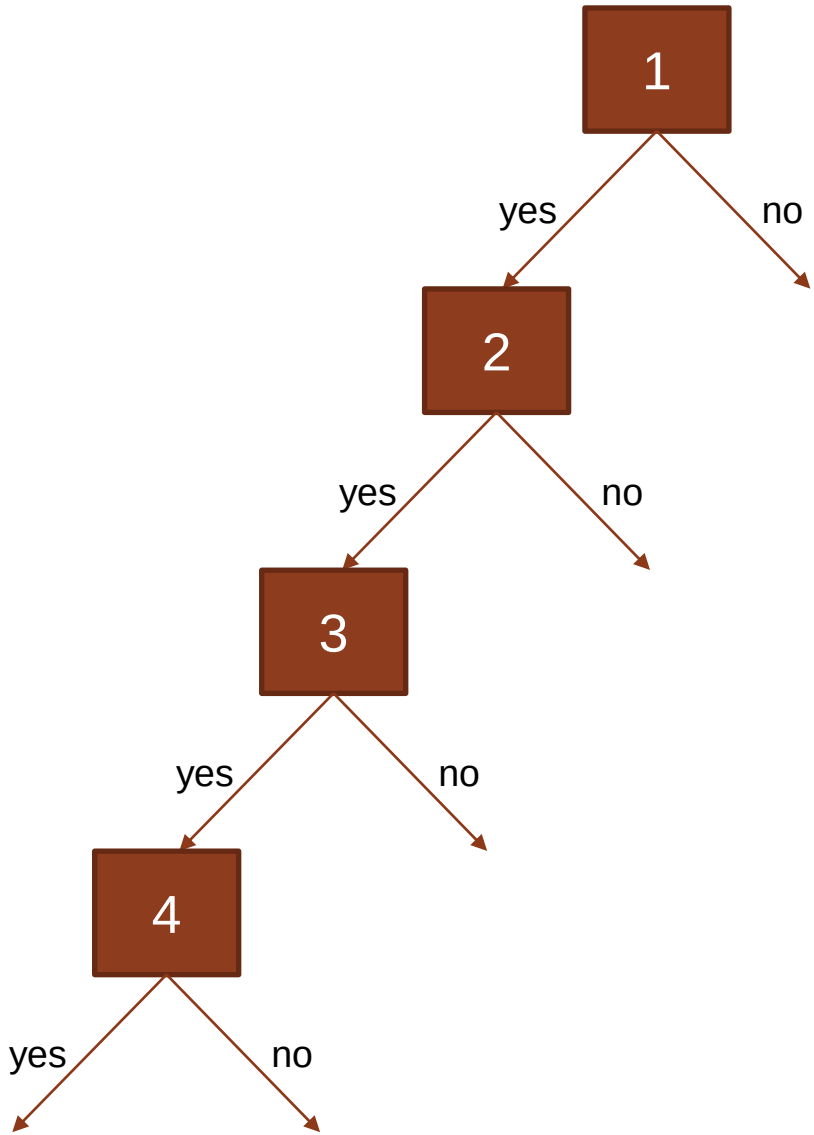
- There are two ways to reason about an algorithm in this framework:
  1. Outcome reasoning:
  2. Model reasoning:

MARA: two types of reasoning

- There are two ways to reason about an algorithm in this framework:
  1. Outcome reasoning:
     If the stakeholders agree that the problem satisfy all the MARA problem-features then we can deploy the algorithm and then debate if it is working after watching its behavior in the real-world.

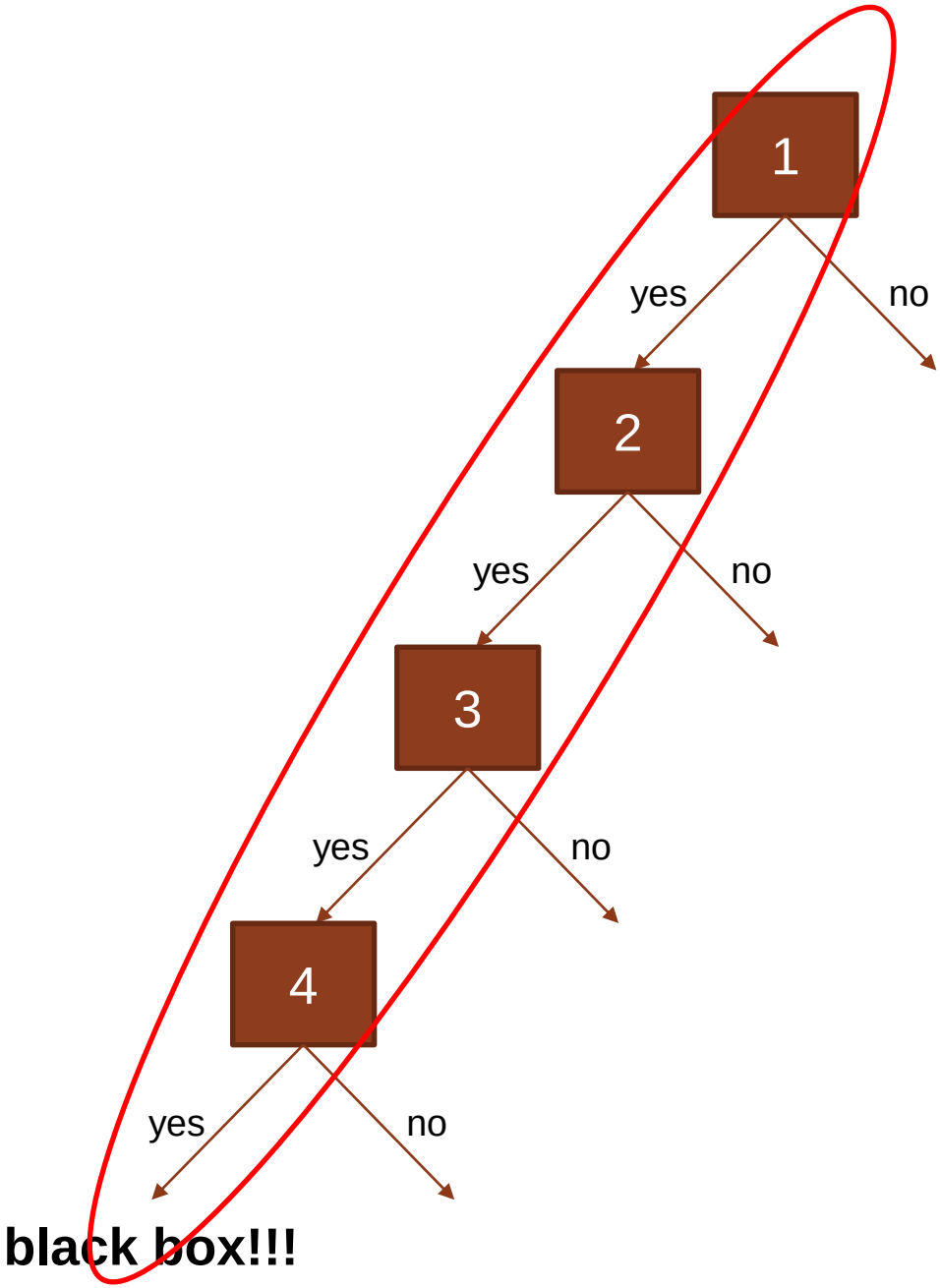
  2. Model reasoning:

MARA



black box!!!

MARA



MARA: two types of reasoning

- There are two ways to reason about an algorithm in this framework:
  1. Outcome reasoning:
     If the stakeholders agree that the problem satisfy all the MARA problem-features then we can deploy the algorithm and then debate if it is working after watching its behavior in the real-world.

  2. Model reasoning:

MARA: two types of reasoning

- There are two ways to reason about an algorithm in this framework:
  1. Outcome reasoning:
     If the stakeholders agree that the problem satisfy all the MARA problem-features then we can deploy the algorithm and then debate if it is working after watching its behavior in the real-world. This is perfect for a black box algorithm.

  2. Model reasoning:

MARA: two types of reasoning

- There are two ways to reason about an algorithm in this framework:
  1. Outcome reasoning:
     If the stakeholders agree that the problem satisfy all the MARA problem-features then we can deploy the algorithm and then debate if it is working after watching its behavior in the real-world. This is perfect for a black box algorithm.

  2. Model reasoning:
     If any problem-feature fails to satisfy then we need an algorithm that can be reasoned about independently of the data used to fit the algorithm (e.g., parameters).

MARA: two types of reasoning

- There are two ways to reason about an algorithm in this framework:
  1. Outcome reasoning:
     If the stakeholders agree that the problem satisfy all the MARA problem-features then we can deploy the algorithm and then debate if it is working after watching its behavior in the real-world. This is perfect for a black box algorithm.

  2. Model reasoning:
     If any problem-feature fails to satisfy then we need an algorithm that can be reasoned about independently of the data used to fit the algorithm (e.g., parameters). This requires a mathematical foundation that allows for deductive reasoning about the future behavior of the algorithm.

MARA: two types of reasoning

- There are two ways to reason about an algorithm in this framework:
  1. Outcome reasoning:
     If the stakeholders agree that the problem satisfy all the MARA problem-features then we can deploy the algorithm and then debate if it is working after watching its behavior in the real-world. This is perfect for a black box algorithm.

  2. Model reasoning:
     If any problem-feature fails to satisfy then we need an algorithm that can be reasoned about independently of the data used to fit the algorithm (e.g., parameters). This requires a mathematical foundation that allows for deductive reasoning about the future behavior of the algorithm.

Model reasoning:

MARA: two types of reasoning

- There are two ways to reason about an algorithm in this framework:
  1. Outcome reasoning:
     If the stakeholders agree that the problem satisfy all the MARA problem-features then we can deploy the algorithm and then debate if it is working after watching its behavior in the real-world. This is perfect for a black box algorithm.

  2. Model reasoning:
     If any problem-feature fails to satisfy then we need an algorithm that can be reasoned about independently of the data used to fit the algorithm (e.g., parameters). This requires a mathematical foundation that allows for deductive reasoning about the future behavior of the algorithm.

Model reasoning: consider a linear regression with $\hat{\beta}_3 < 0$ but every randomized controlled study finds X_3 causes Y to increase.

MARA: time

MARA: time

- Outcome reasoning happens after the deployment of the algorithm into the real-world.

MARA: time

- Outcome reasoning happens after the deployment of the algorithm into the real-world. The stakeholders must be comfortable with bearing the consequences of its errors as it accumulates errors.

MARA: time

- Outcome reasoning happens after the deployment of the algorithm into the real-world. The stakeholders must be comfortable with bearing the consequences of its errors as it accumulates errors.

- In contrast, model reasoning allows the stakeholders to discuss and argue about the suitability of the algorithm for the particular problem before the algorithm begins to accumulate errors in the real-world.

MARA: time

- Outcome reasoning happens after the deployment of the algorithm into the real-world. The stakeholders must be comfortable with bearing the consequences of its errors as it accumulates errors.

- In contrast, model reasoning allows the stakeholders to discuss and argue about the suitability of the algorithm for the particular problem before the algorithm begins to accumulate errors in the real-world.

- Getting stakeholder buy-in is absolutely vital because errors will occur.

MARA: time

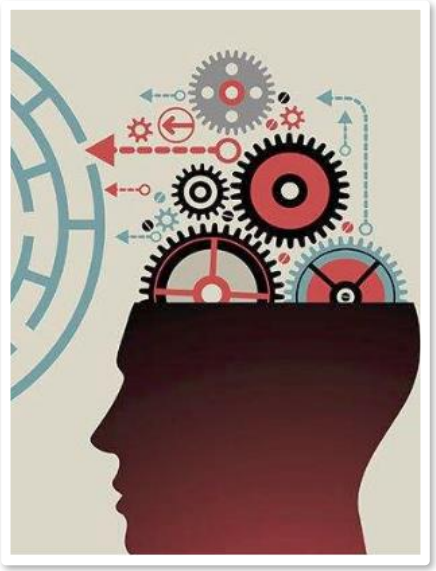
- Outcome reasoning happens after the deployment of the algorithm into the real-world. The stakeholders must be comfortable with bearing the consequences of its errors as it accumulates errors.

- In contrast, model reasoning allows the stakeholders to discuss and argue about the suitability of the algorithm for the particular problem before the algorithm begins to accumulate errors in the real-world.

- Getting stakeholder buy-in is absolutely vital because errors will occur. Whether through outcome-reasoning or model reasoning, achieving buy-in means the people who “own” the problem were part of the decision making and bear responsibility of the consequences.

insights

final thoughts



MARA: insights

MARA: insights

- A problem that satisfies all four MARA problem-features is likely best served by outcome reasoning…

MARA: insights

- A problem that satisfies all four MARA problem-features is likely best served by outcome reasoning; all algorithms are able to be assessed using outcome reasoning so restricting to model reasoning limits the space of algorithms you can deploy.

MARA: insights

- A problem that satisfies all four MARA problem-features is likely best served by outcome reasoning; all algorithms are able to be assessed using outcome reasoning so restricting to model reasoning limits the space of algorithms you can deploy.

- Any given algorithm can become model reasoning.

MARA: insights

- A problem that satisfies all four MARA problem-features is likely best served by outcome reasoning; all algorithms are able to be assessed using outcome reasoning so restricting to model reasoning limits the space of algorithms you can deploy.

- Any given algorithm can become model reasoning. This strongly implies a direction for research.

MARA: insights

- A problem that satisfies all four MARA problem-features is likely best served by outcome reasoning; all algorithms are able to be assessed using outcome reasoning so restricting to model reasoning limits the space of algorithms you can deploy.

- Any given algorithm can become model reasoning. This strongly implies a direction for research.

- Though I tried to keep the four problem-features distinct in my examples, in practice they will interact with each other.

MARA: insights

- A problem that satisfies all four MARA problem-features is likely best served by outcome reasoning; all algorithms are able to be assessed using outcome reasoning so restricting to model reasoning limits the space of algorithms you can deploy.

- Any given algorithm can become model reasoning. This strongly implies a direction for research.

- Though I tried to keep the four problem-features distinct in my examples, in practice they will interact with each other. And the joint failure to satisfy multiple problem-features is likely to cause more pronounced issues with a black box algorithm.

MARA: insights

- A problem that satisfies all four MARA problem-features is likely best served by outcome reasoning; all algorithms are able to be assessed using outcome reasoning so restricting to model reasoning limits the space of algorithms you can deploy.

- Any given algorithm can become model reasoning. This strongly implies a direction for research.

- Though I tried to keep the four problem-features distinct in my examples, in practice they will interact with each other. And the joint failure to satisfy multiple problem-features is likely to cause more pronounced issues with a black box algorithm. Similarly, the four MARA are likely best thought of as continuous, rather than as binary.

MARA: insights

- A problem that satisfies all four MARA problem-features is likely best served by outcome reasoning; all algorithms are able to be assessed using outcome reasoning so restricting to model reasoning limits the space of algorithms you can deploy.

- Any given algorithm can become model reasoning. This strongly implies a direction for research.

- Though I tried to keep the four problem-features distinct in my examples, in practice they will interact with each other. And the joint failure to satisfy multiple problem-features is likely to cause more pronounced issues with a black box algorithm. Similarly, the four MARA are likely best thought of as continuous, rather than as binary.

- The MARA identify problems that are similar to the Common Task Framework.

MARA: insights

- A problem that satisfies all four MARA problem-features is likely best served by outcome reasoning; all algorithms are able to be assessed using outcome reasoning so restricting to model reasoning limits the space of algorithms you can deploy.

- Any given algorithm can become model reasoning. This strongly implies a direction for research.

- Though I tried to keep the four problem-features distinct in my examples, in practice they will interact with each other. And the joint failure to satisfy multiple problem-features is likely to cause more pronounced issues with a black box algorithm. Similarly, the four MARA are likely best thought of as continuous, rather than as binary.

- The MARA identify problems that are similar to the Common Task Framework. MARA warrants “extrapolation” of black-boxes into unseen data settings.

MARA: insights

- A problem that satisfies all four MARA problem-features is likely best served by outcome reasoning; all algorithms are able to be assessed using outcome reasoning so restricting to model reasoning limits the space of algorithms you can deploy.

- Any given algorithm can become model reasoning. This strongly implies a direction for research.

- Though I tried to keep the four problem-features distinct in my examples, in practice they will interact with each other. And the joint failure to satisfy multiple problem-features is likely to cause more pronounced issues with a black box algorithm. Similarly, the four MARA are likely best thought of as continuous, rather than as binary.

- The MARA identify problems that are similar to the Common Task Framework. MARA warrants “extrapolation” of black-boxes into unseen data settings. It also anchors on stakeholders.

MARA: insights

- A problem that satisfies all four MARA problem-features is likely best served by outcome reasoning; all algorithms are able to be assessed using outcome reasoning so restricting to model reasoning limits the space of algorithms you can deploy.

- Any given algorithm can become model reasoning. This strongly implies a direction for research.

- Though I tried to keep the four problem-features distinct in my examples, in practice they will interact with each other. And the joint failure to satisfy multiple problem-features is likely to cause more pronounced issues with a black box algorithm. Similarly, the four MARA are likely best thought of as continuous, rather than as binary.

- The MARA identify problems that are similar to the Common Task Framework. MARA warrants “extrapolation” of black-boxes into unseen data settings. It also anchors on stakeholders.

fin.