Big Data for Public Policy

AI Policy

Elliott Ash & Malka Guillot

Outline

Example: Allocating Inspectors

Athey 2017; Glaeser et al, AER P&P 2016

- Governments can conserve resources by inspecting establishments that are likely to have violations, e.g.:
  - NYC's Firecast algorithm predicts fire risk and code violation
  - Glaeser et al.'s (2016) algorithm predicts health code violations in Boston restaurants (improved violation detection rates by 30%).

Example: Allocating Inspectors

Athey 2017; Glaeser et al, AER P&P 2016

- Governments can conserve resources by inspecting establishments that are likely to have violations, e.g.:
  - NYC's Firecast algorithm predicts fire risk and code violation
  - Glaeser et al.'s (2016) algorithm predicts health code violations in Boston restaurants (improved violation detection rates by 30%).
- Are these predictions sufficient for optimal allocation of inspectors?

Example: Allocating Inspectors

Athey 2017; Glaeser et al, AER P&P 2016

- Governments can conserve resources by inspecting establishments that are likely to have violations, e.g.:
  - NYC's Firecast algorithm predicts fire risk and code violation
  - Glaeser et al.'s (2016) algorithm predicts health code violations in Boston restaurants (improved violation detection rates by 30%).
- Are these predictions sufficient for optimal allocation of inspectors?

Yes, if:

- Costs of fixing problems are mostly homogeneous.
  - it could be that buildings with high fire risk also have old wiring that is costly to replace $\rightarrow$ better to inspect buildings with cheaper fixes.

Example: Allocating Inspectors

Athey 2017; Glaeser et al, AER P&P 2016

- Governments can conserve resources by inspecting establishments that are likely to have violations, e.g.:
  - NYC's Firecast algorithm predicts fire risk and code violation
  - Glaeser et al.'s (2016) algorithm predicts health code violations in Boston restaurants (improved violation detection rates by 30%).
- Are these predictions sufficient for optimal allocation of inspectors?

Yes, if:

- Costs of fixing problems are mostly homogeneous.
  - it could be that buildings with high fire risk also have old wiring that is costly to replace $\rightarrow$ better to inspect buildings with cheaper fixes.
- Establishments subject to inspection do not respond to the algorithm.
  - some firms may be more sensitive to penalties than others, or may be easier for some firms to game the predictors.

Example: Allocating Inspectors

Athey 2017; Glaeser et al, AER P&P 2016

- Governments can conserve resources by inspecting establishments that are likely to have violations, e.g.:
  - NYC's Firecast algorithm predicts fire risk and code violation
  - Glaeser et al.'s (2016) algorithm predicts health code violations in Boston restaurants (improved violation detection rates by 30%).
- Are these predictions sufficient for optimal allocation of inspectors?

Yes, if:

- Costs of fixing problems are mostly homogeneous.
  - it could be that buildings with high fire risk also have old wiring that is costly to replace $\rightarrow$ better to inspect buildings with cheaper fixes.
- Establishments subject to inspection do not respond to the algorithm.
  - some firms may be more sensitive to penalties than others, or may be easier for some firms to game the predictors.
  - some firms might know they have a low inspection due to a low violation probability (because of their neighborhood, for example), and reduce safety measures.

Example: Allocating Inspectors

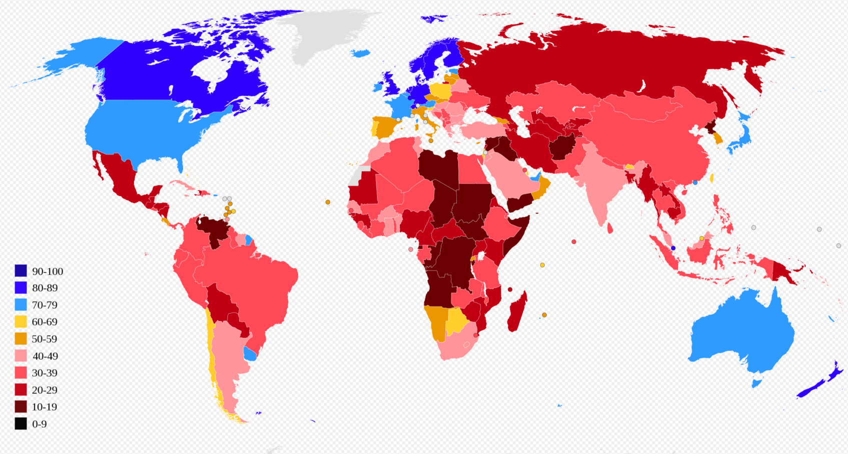
Athey 2017; Glaeser et al, AER P&P 2016

- Governments can conserve resources by inspecting establishments that are likely to have violations, e.g.:
  - NYC's Firecast algorithm predicts fire risk and code violation
  - Glaeser et al.'s (2016) algorithm predicts health code violations in Boston restaurants (improved violation detection rates by 30%).
- Are these predictions sufficient for optimal allocation of inspectors?

Yes, if:

- Costs of fixing problems are mostly homogeneous.
  - it could be that buildings with high fire risk also have old wiring that is costly to replace $\rightarrow$ better to inspect buildings with cheaper fixes.
- Establishments subject to inspection do not respond to the algorithm.
  - some firms may be more sensitive to penalties than others, or may be easier for some firms to game the predictors.
  - some firms might know they have a low inspection due to a low violation probability (because of their neighborhood, for example), and reduce safety measures.
- Overall, the inspection policy problem is a causal inference problem:
  - What is the expected improvement in overall quality of units (e.g., food poisoning rates) in the city under a new inspector allocation regime?

Motivation (Ash, Galletta, Giommoni 2021)



Corruption Perceptions Index, 2018

Global costs of corruption were $2.6 trillion in 2018, according to U.N. data.
Firms and individuals spend more than $1 trillion in bribes every year.

Setting

- In Brazil, local municipalities ($N = 5563$) play a central role in government services:
  - e.g., primary education, healthcare, housing, transportation.

Setting

- In Brazil, local municipalities ($N = 5563$) play a central role in government services:
  - e.g., primary education, healthcare, housing, transportation.
- In 2003, Brazilian government introduced innovative anti-corruption program:
  - **Audit of public spending** in **randomly selected municipalities** (through public lottery).
  - team of 10-15 auditors spend two weeks in municipal offices.
  - they write a report, send to authorities for criminal penalties and make it public.

Setting

- In Brazil, local municipalities ($N = 5563$) play a central role in government services:
    - e.g., primary education, healthcare, housing, transportation.
- In 2003, Brazilian government introduced innovative anti-corruption program:
    - **Audit of public spending** in **randomly selected municipalities** (through public lottery).
    - team of 10-15 auditors spend two weeks in municipal offices.
    - they write a report, send to authorities for criminal penalties and make it public.
- This paper: train xgboost classifier to detect corruption from local budget data.

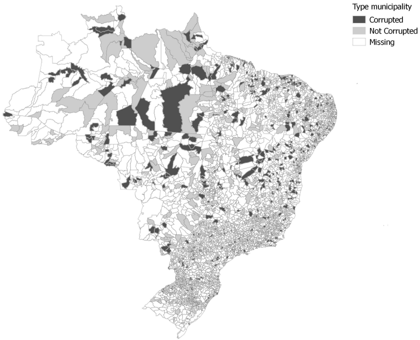
Model Performance in Test Set

	OLS (1)	Lasso (2)	Logistic (3)	**XGBoost (4)**
Accuracy	0.476	0.474	0.560	**0.723**
	(0.022)	(0.022)	(0.022)	**(0.012)**
AUC-ROC	0.487	0.507	0.568	**0.777**
	(0.016)	(0.012)	(0.016)	**(0.013)**
F1	0.685	0.538	0.545	**0.632**
	(0.031)	(0.050)	(0.054)	**(0.018)**

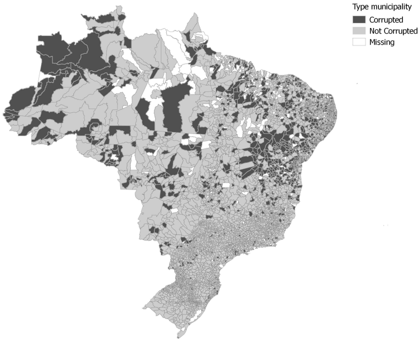
SE of mean across 5 folds in parentheses.

- AUC-ROC ("Area under the receiver operating curve") is a standard metric, ranging from 0.5 (guessing) ato 1.0 (perfect accuracy).
  - Interpretation: probability that a randomly sampled corrupt municipality is ranked more highly by predicted probability of corruption than a randomly sampled non-corrupt municipality.

Detecting Corruption in all of Brazil



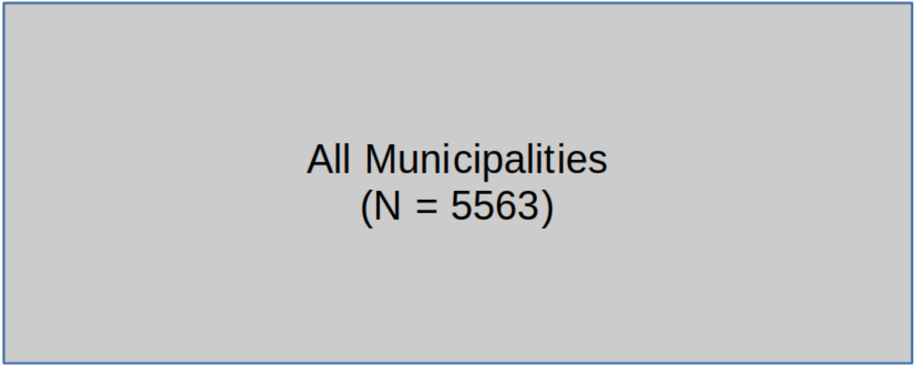
(a) Actual Corruption



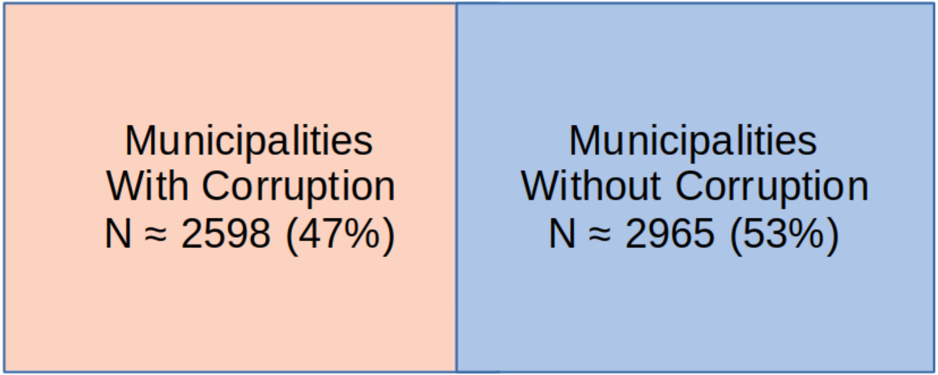
(b) Predicted Corruption

We regressed predicted corruption in pre-audit years on having an audit, and there was no difference in any specification (consistent with randomization of audits).

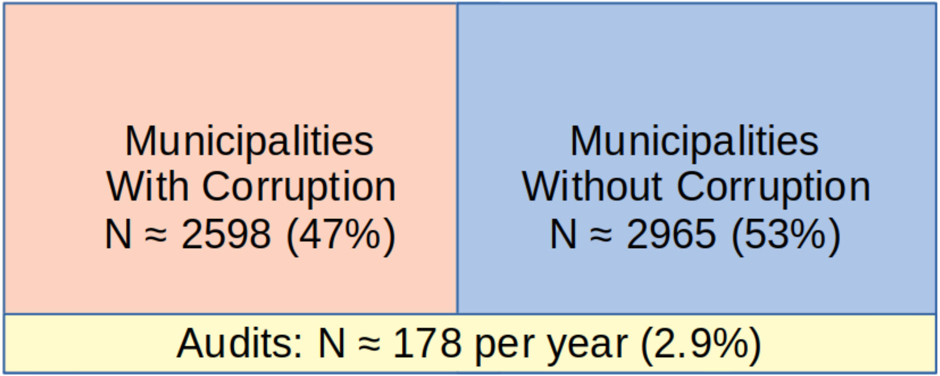
How effective are random audits?



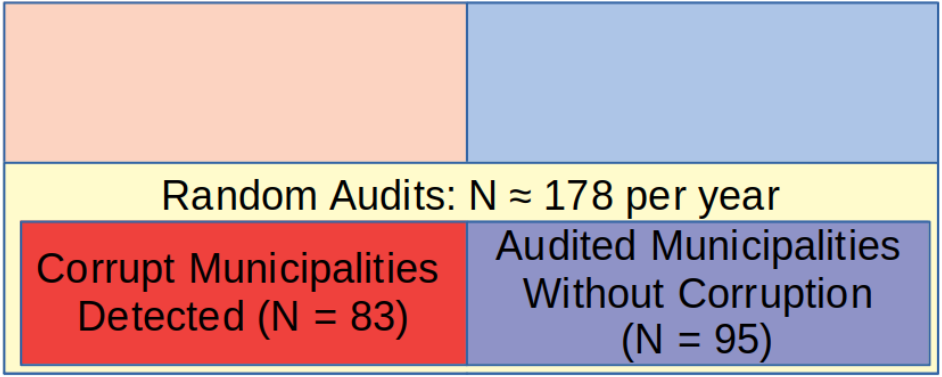
How effective are random audits?



How effective are random audits?



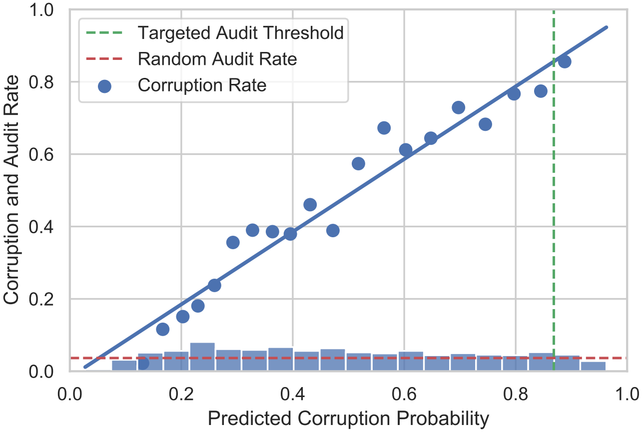
How effective are random audits?



- Under random audits, and assuming perfect detection conditional on audit, detection rate (per corrupt municipality) is equal to the audit rate (2.9%).

Targeting Audits by Corruption Risk

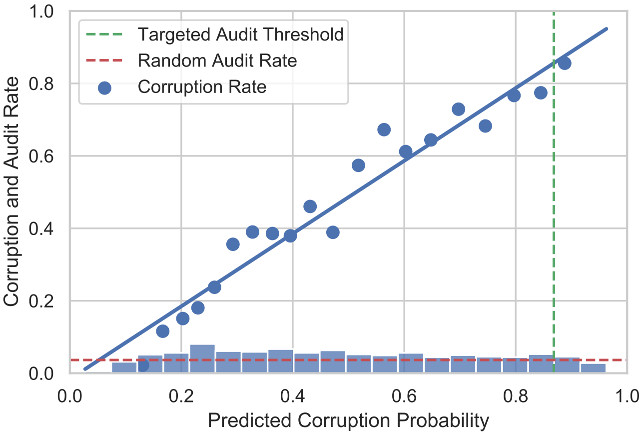
Targeting Audits by Corruption Risk



Rank municipalities by corruption risk:

- Apply model to budget data for each municipality to produce $\hat{y}_{it}$
- for each year t, get an ordinal ranking of the municipalities by predicted probability of corruption.

Targeting Audits by Corruption Risk



Proposed policy: Replace random audits with audits targeted by predicted corruption risk.

- Rather than sampling 200 municipalities uniformly from distribution, audit 200 with highest $\hat{y}_{it}$.

Performance of Targeting Audits

	Random Audits (1)	Targeted Audits (2)	
Corruption Rate, if Audited	0.4664	0.8563	(0.0163)
Audit Rate, if Corrupt	0.0365	0.0671	(0.0013)
$\hookrightarrow$ Ratio over Random Audits		1.836	(0.035)

Notes: Metrics for comparing the effectiveness of audit policies: random audits (column 1), targeting audits to the municipalities with the highest corruption risk (column 2). "Corruption Rate, if Audited" is the share of audited municipalities where narrow corruption is detected, for the respective policy. "Audit Rate, if Corrupt" is the expected probability of being audited, if narrow corrupt, under the various policies. Column 1 reports the observed rates in the data. In Column 2, statistics give the mean and standard error (in parentheses) across five values for the predicted corruption risk, produced using different training-set folds. "Ratio over Random Audits" is the "Audit Rate, if Corrupt" value for the indicated policy, divided by that value under random audits.

Claudio Ferraz
@claudferraz

1/3 I just came across this very interesting work by @ellliottt @sergallet and @T_Giommoni using Machine Learning to predict corrupt practices in Brazil's municipalities. They show that a ML prediction algorithm can be more effective than a random auditing....

> **Sergio Galletta** @sergallet · May 1
> In a newly released WP, together with @ellliottt and @T_Giommoni, we show how ML techniques can be used to overcome data limitations when performing policy evaluation
>
> papers.ssrn.com/sol3/papers.cf...
> Show this thread

1:03 AM · Nov 29, 2020 · Twitter Web App

10 Likes

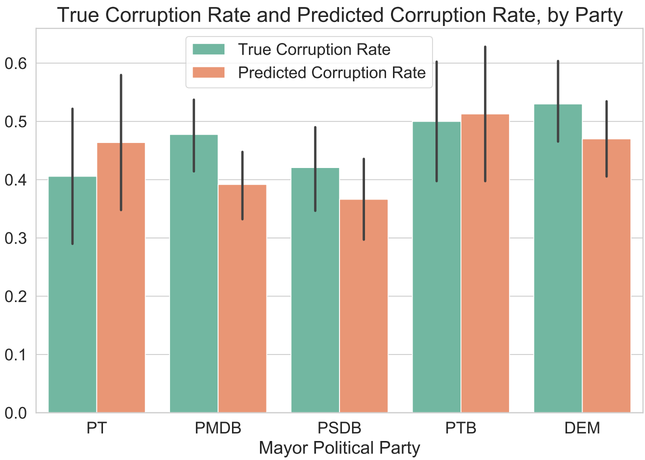
Claudio Ferraz @claudferraz · 9h
Replying to @claudferraz
2/3 But I think they miss an important point for the practical use of ML. The random audit was politically neutral and this is why it was credible to begin with. With a ML the estimated risk based on an algorithm can, in principle, be manipulated to target places or parties

1 5

Claudio Ferraz @claudferraz · 9h
3/3 So an important discussion is how to make these ML algorithms politically unbiased and how to gain credibility and convince government officials that using these types of algorithms for policy can generate important gains in the fight against corruption

What if the AI is biased toward one of the political parties?

Parties are treated differently by the algorithm



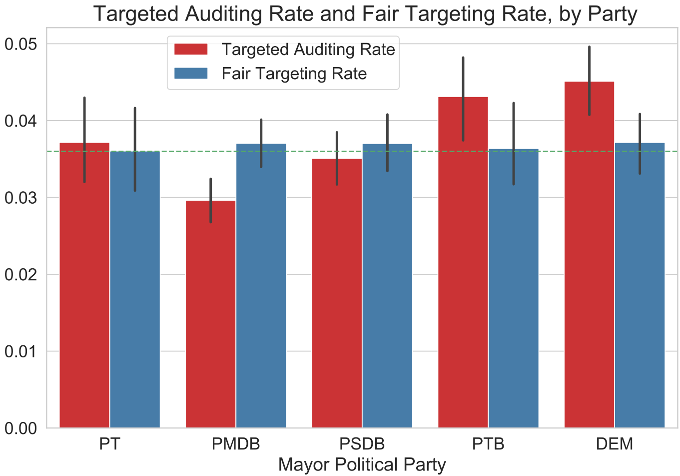
- This variation in true and predicted corruption means that some parties are audited more often under targeted audits than under random audits.

Politically Neutral Targeting Regime

e.g. Rambachan et al 2020:

- start with $\hat{y}_i$ for each municipality and the resulting corruption-risk ranking for all municipalities in a given year.
- produce separate rankings by party.
- within each party, audit the same share of municipalities.

Audit Allocation with Fair Targeting



- This variation in true and predicted corruption means that some parties are audited more often under targeted audits than under random audits.

Performance of Fair Targeting

	Random Audits (1)	Targeted Audits (2)		Fair Targeting (3)	
Corruption Rate, if Audited	0.4664	0.8563	(0.0163)	0.8364	(0.0173)
Audit Rate, if Corrupt	0.0365	0.0671	(0.0013)	0.0655	(0.0014)
↪ Ratio over Random Audits		1.836	(0.035)	1.793	(0.037)

Notes: Metrics for comparing the effectiveness of audit policies: random audits (column 1), targeting audits to the municipalities with the highest corruption risk (column 2), or targeting audits with highest corruption with the constraint that all political parties are audited at the same rate. "Political party" means the set of municipalities where that party controls the mayor's office and includes PT, PMDB, PSDB, PTB, and DEM (formerly PFL). "Corruption Rate, if Audited" is the share of audited municipalities where narrow corruption is detected, for the respective policy. "Audit Rate, if Corrupt" is the expected probability of being audited, if narrow corrupt, under the various policies. Column 1 reports the observed rates in the data. In Columns 2 and 3, statistics give the mean and standard error (in parentheses) across five values for the predicted corruption risk, produced using different training-set folds. "Ratio over Random Audits" is the "Audit Rate, if Corrupt" value for the indicated policy, divided by that value under random audits.

Mechanism Design Issues

- With repeated audits, there could be behavioral responses by local officials.
  - could produce significant errors favoring savvy mayors.
  - Would still deter corrupt fiscal actions that are not easily substitutable.

How much information to publicize about audit targeting?

How much information to publicize about audit targeting?

Option 1: Give **full information** about the policy and the associated model weights.

How much information to publicize about audit targeting?

Option 1: Give **full information** about the policy and the associated model weights.

- Would increase deterrence against corruption actions captured by the model, that are not substitutable.
- But would make gaming the system easier.

How much information to publicize about audit targeting?

Option 1: Give **full information** about the policy and the associated model weights.

- Would increase deterrence against corruption actions captured by the model, that are not substitutable.
- But would make gaming the system easier.

Option 2: Give **no information** about how targeting is done.

- This is "the industry approach", e.g., for how google/facebook detect violations.

How much information to publicize about audit targeting?

Option 1: Give **full information** about the policy and the associated model weights.

- Would increase deterrence against corruption actions captured by the model, that are not substitutable.
- But would make gaming the system easier.

Option 2: Give **no information** about how targeting is done.

- This is "the industry approach", e.g., for how google/facebook detect violations.
- mayors might learn how algorithm works over time.
- weights could be updated in response to behavioral responses

Mixing random and targeted audits

- Random audits could be maintained (along with targeted audits).
  - Preserves some deterrence incentive for all municipalities.
  - Results of random audits could be used to update algorithm parameters.

Outline

- Algorithms influence various aspects of life:
  - selecting tax payers for audits
  - granting or denying immigration visas
  - security screening at airports
- Besides benefits, can have risks and harms.
- Public interest requires governance to reinforce benefits and minimize risks.

e.g., Incentive Responses

- Decisions today change features tomorrow.
- Take the case of ML-based credit scoring.
- Some strategic responses are benign/helpful:
  - e.g., pay back existing debts to improve score

e.g., Incentive Responses

- Decisions today change features tomorrow.
- Take the case of ML-based credit scoring.
- Some strategic responses are benign/helpful:
  - e.g., pay back existing debts to improve score
- Others could be costly manipulation
  - e.g., open more credit accounts to increase score, but at some risk
  - more generally, ML subjects can pay some cost and manipulate their features to improve their predicted label.

e.g., Incentive Responses

- Decisions today change features tomorrow.
- Take the case of ML-based credit scoring.
- Some strategic responses are benign/helpful:
  - e.g., pay back existing debts to improve score
- Others could be costly manipulation
  - e.g., open more credit accounts to increase score, but at some risk
  - more generally, ML subjects can pay some cost and manipulate their features to improve their predicted label.
- Milli et al, "The Social Cost of Strategic Classification" (2019)
  - model sequential decision of modeler and subject as Stackelberg Competition, a classic model from game theory on the interaction between duopolists.

Outline

“Fair ML” / “AI Fairness”

“Fair ML” / “AI Fairness”

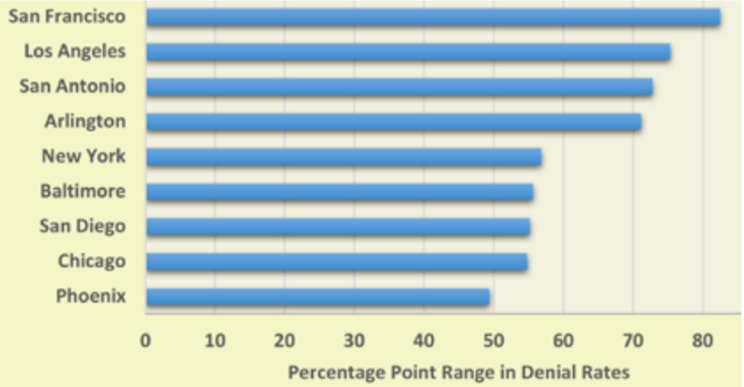
- “ML” or “AI” refer to statistical algorithms
  - can learning algorithms be fair or not?

“Fair ML” / “AI Fairness”

- “ML” or “AI” refer to statistical algorithms
  - can learning algorithms be fair or not?
- Rather: *fairness* is a property of *decisions*.
  - so *“AI Fairness”* should be understood as “*fairness of AI-supported decision-making*”.

Humans are Inconsistent

- Before getting into bias towards particular groups, it should be emphasized that humans are "biased" in the sense that some are more/less lenient:



- A robot judge would generate consistent decisions for same evidence, correcting individual-level leniencies across judges.

Data can be biased

- Education:
  - SAT scores might be used to guide college admissions, but some students get SAT prep courses

Data can be biased

- Education:
  - SAT scores might be used to guide college admissions, but some students get SAT prep courses
  - Teachers (grading essays) might be biased against some students $\rightarrow$ so will automated essay graders based on those grades.

Data can be biased

- Education:
  - SAT scores might be used to guide college admissions, but some students get SAT prep courses
  - Teachers (grading essays) might be biased against some students $\rightarrow$ so will automated essay graders based on those grades.
- Criminal risk scoring (Skeem and Lovenkamp 2016):
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories (**which were produced by biased system**).

Data can be biased

- Education:
  - SAT scores might be used to guide college admissions, but some students get SAT prep courses
  - Teachers (grading essays) might be biased against some students $\rightarrow$ so will automated essay graders based on those grades.
- Criminal risk scoring (Skeem and Lovenkamp 2016):
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories (**which were produced by biased system**).
  - similarly: we measure recidivism as "is re-arrested" rather than "commits more crimes". some people more likely to be re-arrested due to policing bias.

Data can be biased

- Education:
  - SAT scores might be used to guide college admissions, but some students get SAT prep courses
  - Teachers (grading essays) might be biased against some students $\rightarrow$ so will automated essay graders based on those grades.
- Criminal risk scoring (Skeem and Lovenkamp 2016):
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories (**which were produced by biased system**).
  - similarly: we measure recidivism as "is re-arrested" rather than "commits more crimes". some people more likely to be re-arrested due to policing bias.
  - selective labeling:
    - predictive policing – produces evidence of more crimes in the neighborhoods where police want to go.
    - only observe recidivism if released.

Data can be biased

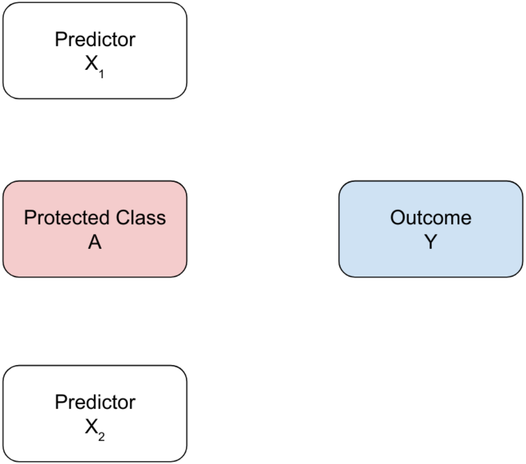
- Education:
  - SAT scores might be used to guide college admissions, but some students get SAT prep courses
  - Teachers (grading essays) might be biased against some students $\rightarrow$ so will automated essay graders based on those grades.
- Criminal risk scoring (Skeem and Lovenkamp 2016):
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories (**which were produced by biased system**).
  - similarly: we measure recidivism as "is re-arrested" rather than "commits more crimes". some people more likely to be re-arrested due to policing bias.
  - selective labeling:
    - predictive policing – produces evidence of more crimes in the neighborhoods where police want to go.
    - only observe recidivism if released.
- a subjective label, such as "harmful to self or others", when made by a human, could be biased (and so would teaching an ML model to reproduce that label)

Data can be biased

- Education:
  - SAT scores might be used to guide college admissions, but some students get SAT prep courses
  - Teachers (grading essays) might be biased against some students $\rightarrow$ so will automated essay graders based on those grades.
- Criminal risk scoring (Skeem and Lovenkamp 2016):
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories (**which were produced by biased system**).
  - similarly: we measure recidivism as "is re-arrested" rather than "commits more crimes". some people more likely to be re-arrested due to policing bias.
  - selective labeling:
    - predictive policing – produces evidence of more crimes in the neighborhoods where police want to go.
    - only observe recidivism if released.
- a subjective label, such as "harmful to self or others", when made by a human, could be biased (and so would teaching an ML model to reproduce that label)

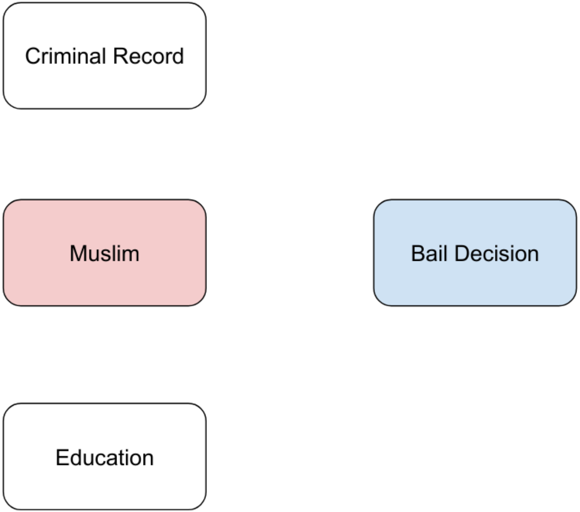
These types of problems cannot be fixed by ML.
But ML can help diagnose them.

Overview: Fairness in Decision-Making

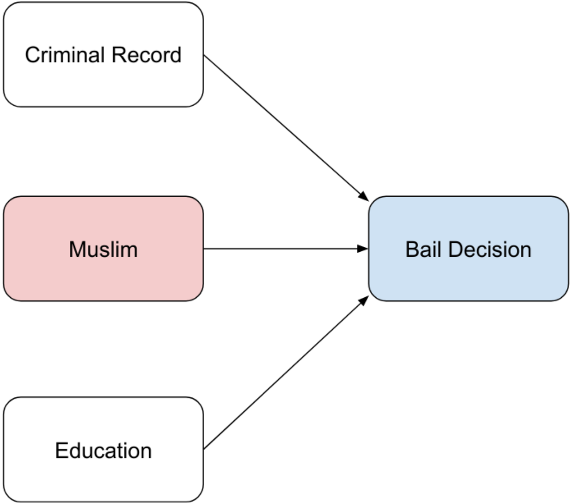


- $A \in \{0,1\}$= protected class, X= other predictors, Y= outcome.
- let $\hat{Y}(X,A)$ be our model predictions.

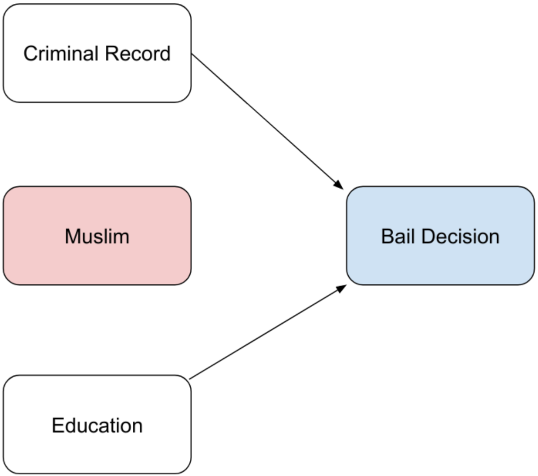
For example:



Standard Approach: Use All Data

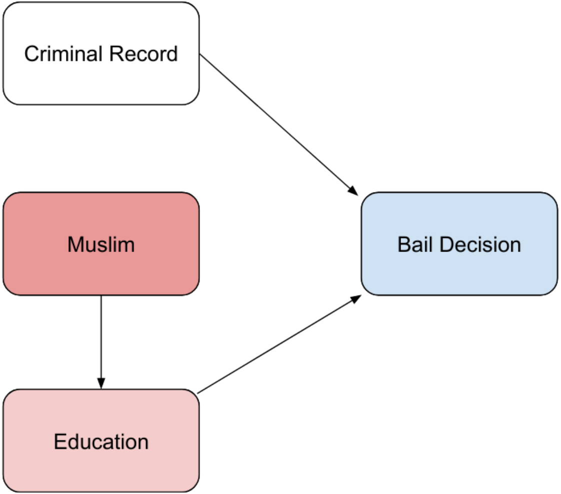


Fairness through Unawareness



- **Fairness through unawareness:** protected attributes are not explicitly used in the prediction process.
  - that is, $\hat{Y}(X,0) = \hat{Y}(X,1)$, $\forall X$.

Problem: Indirect Discrimination



- sensitive factors are implicitly being used by the model, to the extent that they are correlated with included predictors.
  - e.g., muslims have lower education than rest of population.

A deeper problem: Unobserved confounders

A deeper problem: Unobserved confounders

- There is a deeper problem with this approach:
  - protected attributes (e.g. race) are confounded with many unobserved factors, which are correlated with outcomes and predictors.

A deeper problem: Unobserved confounders

- There is a deeper problem with this approach:
  - protected attributes (e.g. race) are confounded with many unobserved factors, which are correlated with outcomes and predictors.
  - as we have seen in our causal inference lectures, resulting estimates for fairness formulas/criteria will be statistically biased.

A deeper problem: Unobserved confounders

- There is a deeper problem with this approach:
  - protected attributes (e.g. race) are confounded with many unobserved factors, which are correlated with outcomes and predictors.
  - as we have seen in our causal inference lectures, resulting estimates for fairness formulas/criteria will be statistically biased.
- Counterfactual fairness (e.g. Kusner et al 2018): "had the protected attributes (e.g., race) of the individual been different, other things being equal, the decision would have remained the same."

A deeper problem: Unobserved confounders

- There is a deeper problem with this approach:
  - protected attributes (e.g. race) are confounded with many unobserved factors, which are correlated with outcomes and predictors.
  - as we have seen in our causal inference lectures, resulting estimates for fairness formulas/criteria will be statistically biased.
- Counterfactual fairness (e.g. Kusner et al 2018): "had the protected attributes (e.g., race) of the individual been different, other things being equal, the decision would have remained the same."
  - e.g., had a defendant been from a different race, he would have had different education, different residence location, etc.

Outline

AI-Supported Policymaking

AI Governance

- AI Fairness: Overview
- AI Governance

What can and should AI decide?

Challenges to developing standards

- Collective decision processes
  - tradeoffs among various stakeholders
  - distortions from lobbying
  - technical issues $\rightarrow$ politicians and voters have low information

Challenges to developing standards

- Collective decision processes
  - tradeoffs among various stakeholders
  - distortions from lobbying
  - technical issues $\rightarrow$ politicians and voters have low information
- Global coordination needed for digital tech
  - accounting for different cultures and contexts

Challenges to developing standards

- Collective decision processes
  - tradeoffs among various stakeholders
  - distortions from lobbying
  - technical issues $\rightarrow$ politicians and voters have low information
- Global coordination needed for digital tech
  - accounting for different cultures and contexts
- How to assign responsibility for risks/harms
  - creator / owner / operator / user?
  - how to understand / determine intentions
  - balance accountability with innovation and growth

Governance Strategies

- Industry-driven approach;
  - Reduces regulatory red tape, could help innovation
  - No central authority to enforce best-practices
  - Expands the power of large corporations
  - Negative externalities, tendency to concentration

Governance Strategies

- Industry-driven approach;
  - Reduces regulatory red tape, could help innovation
  - No central authority to enforce best-practices
  - Expands the power of large corporations
  - Negative externalities, tendency to concentration
- Regulator-driven approach:
  - could reduce externalities and concentration
  - significant technical knowledge/skills needed to be effective
  - could limit innovation and expansion of digital economy
  - could collude with industry leaders

Transparency

- Closed-source algorithms result in "black box justice" and could be abused by insiders.
- But open-source algorithms are prone to gaming: savvy attorneys could "trick" the algorithm.

Transparency

- Closed-source algorithms result in “black box justice” and could be abused by insiders.
- But open-source algorithms are prone to gaming: savvy attorneys could “trick” the algorithm.
- Understanding the code/model not the same as understanding behavior
  - ML processes not understandable by non-experts
  - Sometimes even experts don’t understand the model

“An Economic Approach to Regulating Algorithms”

Rambachan, Kleinberg, Ludwig, and Mullainathan (2020)

- Algorithmic decision-making has two components:
  - (1) training a prediction function, and (2) a decision rule based on the predictions.

Result 1 (social planner):

- the equity preferences of the social planner should not change the training procedure for the prediction function.
  - i.e., there should be no limit on the use of sensitive attributes.
- instead, should use different decision thresholds for different groups.

“An Economic Approach to Regulating Algorithms”

Rambachan, Kleinberg, Ludwig, and Mullainathan (2020)

- Algorithmic decision-making has two components:
  - (1) training a prediction function, and (2) a decision rule based on the predictions.

Result 1 (social planner):

- the equity preferences of the social planner should not change the training procedure for the prediction function.
  - i.e., there should be no limit on the use of sensitive attributes.
- instead, should use different decision thresholds for different groups.

Result 2 (private actors):

- key factor is disclosure of decision process (data, ML training, and decision rule), which, unlike human decision-making, allows prejudicial treatment to be detected.

“An Economic Approach to Regulating Algorithms”

Rambachan, Kleinberg, Ludwig, and Mullainathan (2020)

- Algorithmic decision-making has two components:
  - (1) training a prediction function, and (2) a decision rule based on the predictions.

Result 1 (social planner):

- the equity preferences of the social planner should not change the training procedure for the prediction function.
  - i.e., there should be no limit on the use of sensitive attributes.
- instead, should use different decision thresholds for different groups.

Result 2 (private actors):

- key factor is disclosure of decision process (data, ML training, and decision rule), which, unlike human decision-making, allows prejudicial treatment to be detected.
- without disclosure, algorithms will be just as biased as humans.

"An Economic Approach to Regulating Algorithms"

Rambachan, Kleinberg, Ludwig, and Mullainathan (2020)

- Algorithmic decision-making has two components:
  - (1) training a prediction function, and (2) a decision rule based on the predictions.

Result 1 (social planner):

- the equity preferences of the social planner should not change the training procedure for the prediction function.
  - i.e., there should be no limit on the use of sensitive attributes.
- instead, should use different decision thresholds for different groups.

Result 2 (private actors):

- key factor is disclosure of decision process (data, ML training, and decision rule), which, unlike human decision-making, allows prejudicial treatment to be detected.
- without disclosure, algorithms will be just as biased as humans.
- with disclosure, discrimination decreases relative to humans, and government should impose no constraints on the use of sensitive attributes as predictors.

“An Economic Approach to Regulating Algorithms”

Rambachan, Kleinberg, Ludwig, and Mullainathan (2020)

- Algorithmic decision-making has two components:
  - (1) training a prediction function, and (2) a decision rule based on the predictions.

Result 1 (social planner):

- the equity preferences of the social planner should not change the training procedure for the prediction function.
  - i.e., there should be no limit on the use of sensitive attributes.
- instead, should use different decision thresholds for different groups.

Result 2 (private actors):

- key factor is disclosure of decision process (data, ML training, and decision rule), which, unlike human decision-making, allows prejudicial treatment to be detected.
- without disclosure, algorithms will be just as biased as humans.
- with disclosure, discrimination decreases relative to humans, and government should impose no constraints on the use of sensitive attributes as predictors.
  - caveat: disclosure must include the data and ML training process, not just the decision rule.

Outline

Perception Tasks

- Content identification (Shazam, reverse image search)
- Face recognition
- Medical diagnosis from scans
- Speech to text
- Deepfakes

Perception Tasks

- Content identification (Shazam, reverse image search)
- Face recognition
- Medical diagnosis from scans
- Speech to text
- Deepfakes

High accuracy causes risk of privacy violations.

Perception Tasks

- Content identification (Shazam, reverse image search)
- Face recognition
- Medical diagnosis from scans
- Speech to text
- Deepfakes

High accuracy causes risk of privacy violations.

Systems are sometimes more accurate/effective for some groups, e.g. most-frequent customers.

Perception Tasks

- Content identification (Shazam, reverse image search)
- Face recognition
- Medical diagnosis from scans
- Speech to text
- Deepfakes

High accuracy causes risk of privacy violations.

Systems are sometimes more accurate/effective for some groups, e.g. most-frequent customers.

Overall, problems seem straightforward to solve.

Human Judgment Annotation Tasks

- Spam detection
- Detection of copyrighted material
- Automated essay grading
- Hate speech detection
- Content recommendation

Human Judgment Annotation Tasks

- Spam detection
- Detection of copyrighted material
- Automated essay grading
- Hate speech detection
- Content recommendation

These tasks are subjective, so some error is inevitable.
But human judgments are correlated enough that predictions are useful.

Human Judgment Annotation Tasks

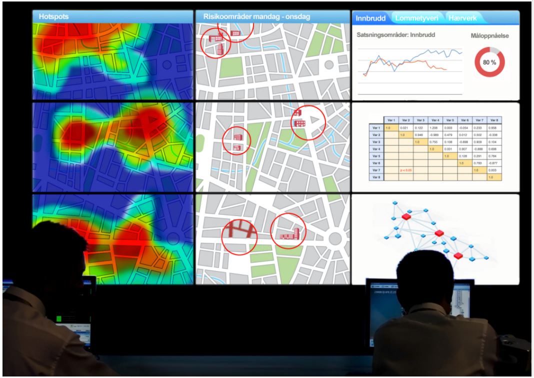
- Spam detection
- Detection of copyrighted material
- Automated essay grading
- Hate speech detection
- Content recommendation

These tasks are subjective, so some error is inevitable.
But human judgments are correlated enough that predictions are useful.

Labels are past behavior, so model is stable and incentive responses are constrained.

- compare: predicting how someone will score on these predictions in the future.

Predictive Policing



Predictive policing poses discrimination risk, thinktank warns

Machine-learning algorithms could replicate or amplify bias on race, sexuality and age

▲ One officer said human biases including more stop and searches of black men were likely to be introduced into algorithm data sets. Photograph: Carl Court/Getty Images

Predicting future choices and social outcomes

- Predicting criminal recidivism
- Predictive policing
- Predicting terrorist risk
- Predicting at-risk kids

Predicting future choices and social outcomes

- Predicting criminal recidivism
- Predictive policing
- Predicting terrorist risk
- Predicting at-risk kids

These systems are risky and can have unintended consequences.

Predicting future choices and social outcomes

- Predicting criminal recidivism
- Predictive policing
- Predicting terrorist risk
- Predicting at-risk kids

These systems are risky and can have unintended consequences.

Predictions influence availability of labels and subsequent behavior.

Predicting future choices and social outcomes

- Predicting criminal recidivism
- Predictive policing
- Predicting terrorist risk
- Predicting at-risk kids

These systems are risky and can have unintended consequences.

Predictions influence availability of labels and subsequent behavior.

Outcomes are in future so models lack external validity.

Predicting future choices and social outcomes

- Predicting criminal recidivism
- Predictive policing
- Predicting terrorist risk
- Predicting at-risk kids

These systems are risky and can have unintended consequences.

Predictions influence availability of labels and subsequent behavior.

Outcomes are in future so models lack external validity.

Errors are costly. Strong incentive responses.

Overview of problems relevant to ML fairness

1. Accuracy issues:
    - model stability
    - selective labeling

Overview of problems relevant to ML fairness

1. Accuracy issues:
   - model stability
   - selective labeling
2. Equity issues:
   - (relative) error rate
   - (relative) costs of errors

Overview of problems relevant to ML fairness

1. Accuracy issues:
   - model stability
   - selective labeling
2. Equity issues:
   - (relative) error rate
   - (relative) costs of errors
3. Social problems from introducing system:
   - externalities (e.g. privacy violations)
   - asymmetric information (AI company knows your preferences (price point)$\rightarrow$ they have information advantage and can capture more surplus).

Overview of problems relevant to ML fairness

1. Accuracy issues:
   - model stability
   - selective labeling
2. Equity issues:
   - (relative) error rate
   - (relative) costs of errors
3. Social problems from introducing system:
   - externalities (e.g. privacy violations)
   - asymmetric information (AI company knows your preferences (price point)$\rightarrow$ they have information advantage and can capture more surplus).
4. Incentive responses:
   - subjects try to manipulate features to game system
   - systems (e.g. essay grading) perceived as biased/unfair are discouraging.

Overview of problems relevant to ML fairness

1. Accuracy issues:
   - model stability
   - selective labeling
2. Equity issues:
   - (relative) error rate
   - (relative) costs of errors
3. Social problems from introducing system:
   - externalities (e.g. privacy violations)
   - asymmetric information (AI company knows your preferences (price point) $\rightarrow$ they have information advantage and can capture more surplus).
4. Incentive responses:
   - subjects try to manipulate features to game system
   - systems (e.g. essay grading) perceived as biased/unfair are discouraging.

Activity: Identify examples of each problem in the setting of algorithmic hiring.

Additional issues with using AI for predicting social outcomes
Narayanan slides

- Hunger for personal data
- Transfer of power from domain experts & workers to unaccountable tech companies
- Veneer of objectivity
- Lack of explainability
- . . . others?