Evidence-based Policy Evaluation

Kosuke Imai

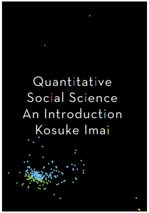
Harvard University

Data Analytics Colloquium
National Chung Hsing University University of Texas, Dallas

November 13, 2020

Quantitative Social Science

- Massive technological changes ⇝ Internet and computing revolution
- **Past**: only statisticians and methodologists analyzed data
- **Today**: EVERYONE is analyzing data
  Data are affecting our lives too!
- **Past**: government data, national survey data
- **Today**: more of old types of data and lots of new data
  - surveys
  - experiments
  - administrative records
  - social media data
  - GIS data
  - text, images, sounds, videos



- "Big (Social Science) Data" revolution
- We must learn and teach how to analyze data

Evidence-based Policy Evaluation

- QSS is also about analyzing data to solve problems in the society
- Evidence-based policy evaluation
  - evaluating the existing policies in place
  - informing policy-making

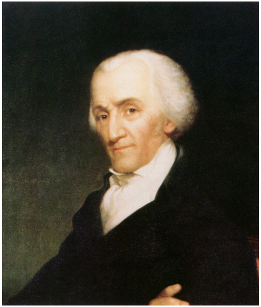
- Examples from my own research
  - Job training in Afghanistan

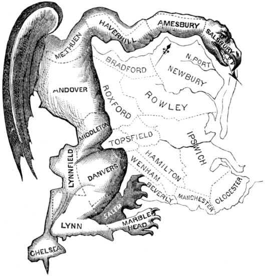
  - Indian national health insurance

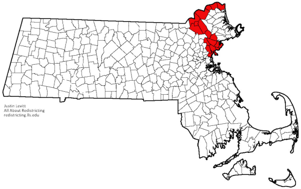
राष्ट्रीय स्वास्थ्य बीमा योजना
Rashtriya Swasthya Bima Yojana
XXX. XXXXXXX XXXXX
AGE: 30 Yrs
Gender: Male
0000 0000 00000 0000 0
राष्ट्रीय स्वास्थ्य बीमा योजना

- Today's talk
  1. Detecting gerrymandering in legislative redistricting
  2. Use of AI in judicial decision making

What is Gerrymandering?

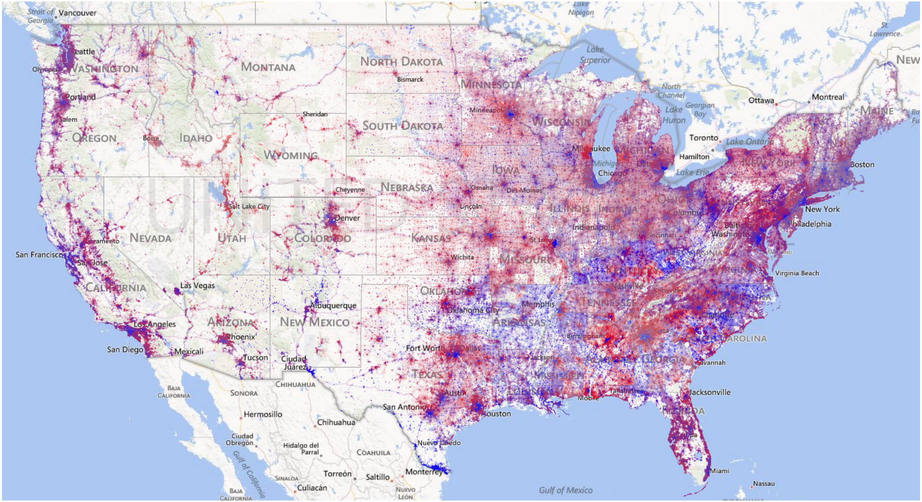




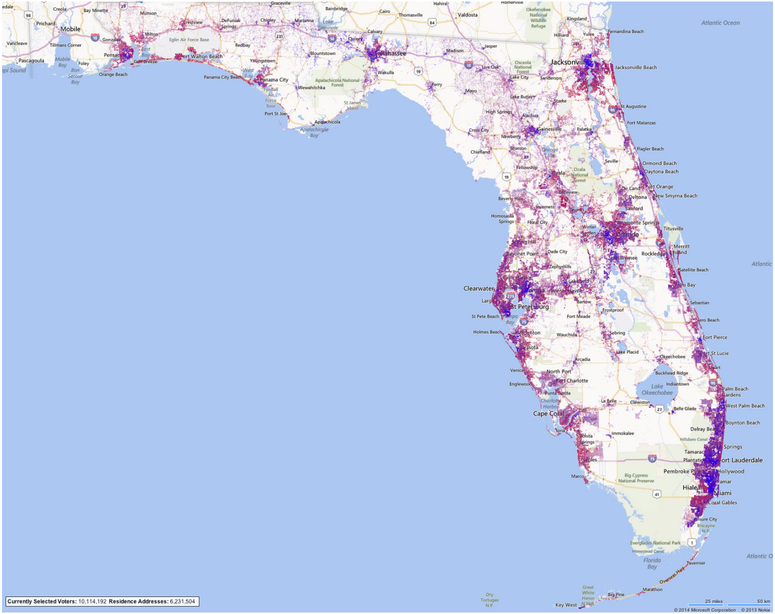

Justin Levitt
All About Redistricting
redistricting.lls.edu

- Elbridge Gerry (Massachusetts governor)
- Gerry+Salamander = Gerry-mander
- Partisan and racial gerrymandering

Elections During the Big-Data Era

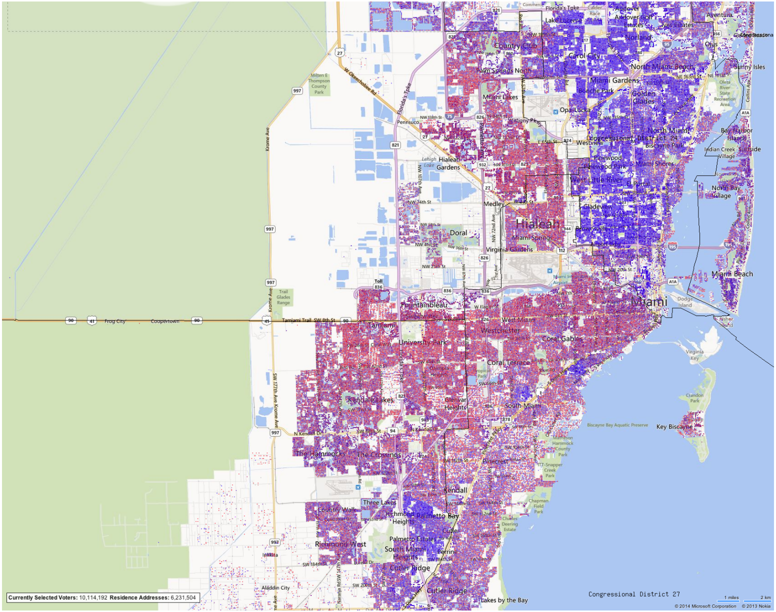


State Level

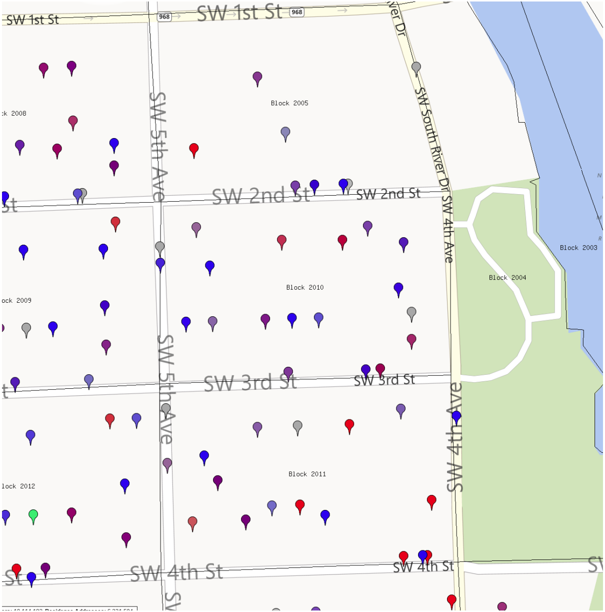


Currently Selected Voters: 10,114,192 **Residence Addresses:** 6,231,504

District Level



Household Level



Democrats

Republicans

Independents

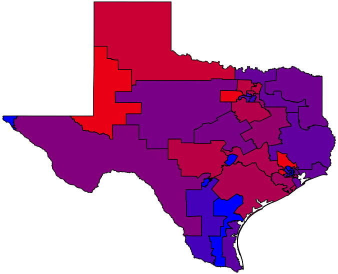
Mixed

- Registered voter list
  - Name, Address
  - Sex, Birthday
  - Partisanship
  - Race (South)
  - Turnout

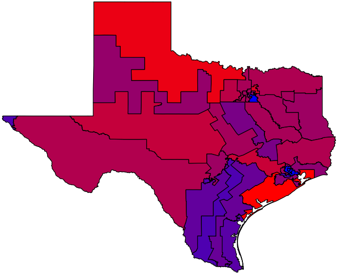
Today's Gerrymander (2003)

- Tom Delay (Republican majority leader)
- 16 seats (2002) ⇝ 21 seats (2004): total 32 seats
- US Supreme Court ruled racial gerrymander

Congressional Vote Share in Texas (2002)
16 Republican Seats

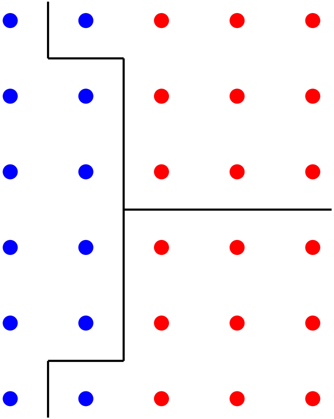


Congressional Vote Share in Texas (2004)
21 Republican Seats

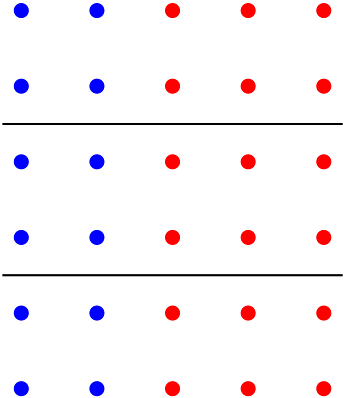


Gerrymandering Strategies

Packing



Cracking



Redistricting in America

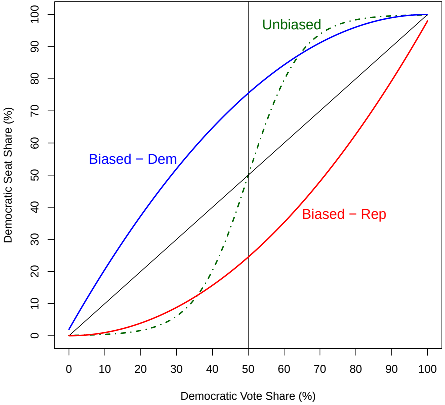
- Redistricting after every decennial census
- Congressional and state legislative districts
- Rules vary across states

- Basic rules
  - Federal level: equal population, voting rights act of 1965
  - State level: contiguity, compactness, preservation of administrative and community boundaries

- Who decides?
  1. State legislature (majority of states)
  2. Independent commission (6 states): California, Arizona, Washington, ...
  3. Until *Shelby County v Holder* (2013), Southern states with the history of racial discrimination were required to obtain “preclearance”
  4. Involvement by state and federal courts: courts decided redistricting in 12 states (2010)

Detecting Gerrymandering

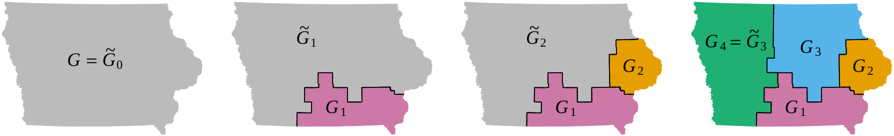
- Statistical measures of gerrymandering
  1. Based on “wasted” votes: efficiency gap
  2. Based on seat-vote curve: partisan symmetry
- Outlier analysis $\rightsquigarrow$ need for baseline distribution
- Must account for state specific geography and voter distribution



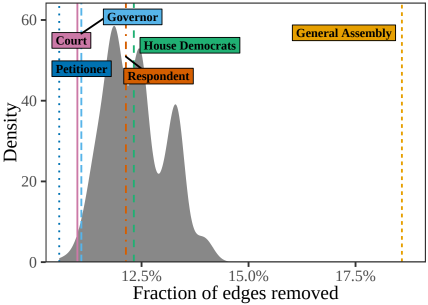
- It is impossible to count all possible redistricting plans
  - Number of ways to divide up an 8×8 checker board into 2 regions
  - 1.2×10^{11}
- Sampling: Markov chain Monte Carlo, Sequential Monte Carlo

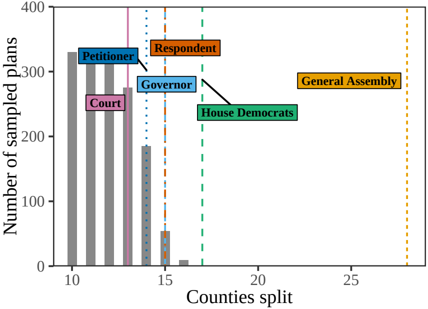
Sequential Monte Carlo (McCartan and Imai, 2020)

- Account for equal population, contiguity, and compactness
- Limit the number of splits of administrative units
- Specify the target distribution of redistricting plans
- Applicable to large states
- Pennsylvania: 9256 precincts, 67 counties, 18 districts
- Independent samples $\leftrightsquigarrow$ Markov chain Monte Carlo
- 1,500 sampled redistricting plans to approximate baseline distribution

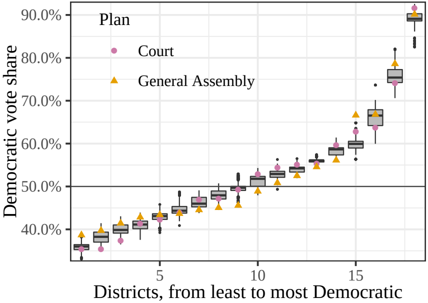


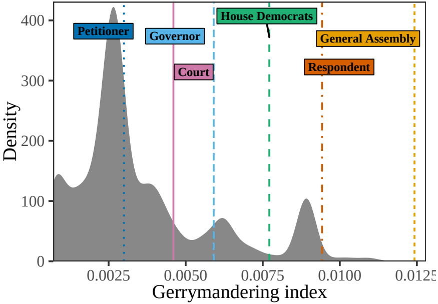
Compactness and Number of County Splits

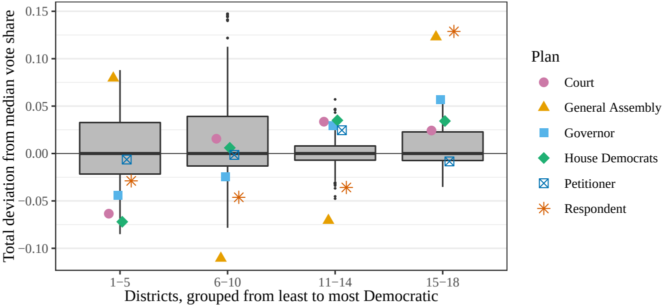




Voteshare Distribution and Gerrymandering Index







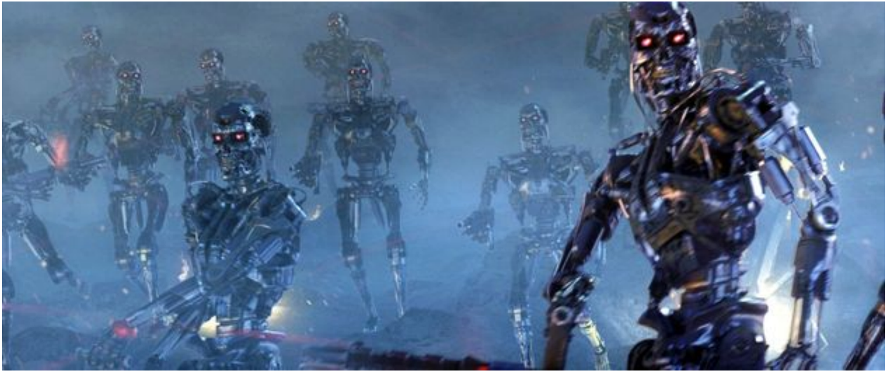
Concluding Remarks

- Political parties use data extensively
    - micro-targeting for voter mobilization
    - opinion polls for messaging
    - voter and election data for redistricting

- Using data analysis for detecting gerrymandering
    - outlier analysis by simulating redistricting plans
    - our algorithm is easy to use and widely applicable
    - R package redist publicly available so that anyone can evaluate redistricting plans

- Legislative redistricting in Taiwan?

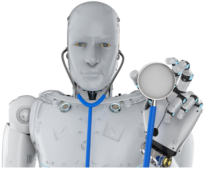
Rise of the Machines



- Statistics, machine learning, artificial intelligence in our daily lives
- Nothing new but accelerated due to technological advances
- Examples: factory assembly lines, home appliances, autonomous cars and drones, games (Chess, Go, Shogi), ...

Algorithm-Assisted Human Decision Making

- But, humans still make many consequential decisions
- We have not yet outsourced these decisions to machines



  - this is true even when human decisions can be suboptimal
  - we may want to hold *someone*, rather than *something*, accountable

- Most prevalent system is algorithm-assisted human decision making
  - humans make decisions with the aid of algorithmic recommendations
  - routine decisions made by individuals in daily lives
  - consequential decisions made by judges, doctors, etc.

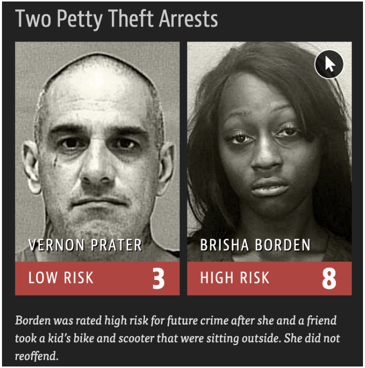
Questions and Contributions

- How do algorithmic recommendations influence human decisions?
  - Do they help human decision-makers achieve their goals?
  - Do they help humans improve the fairness of their decisions?

- Many have studied the accuracy and fairness of algorithms
  - Few have researched their impacts on human decisions
  - Little is known about how algorithmic bias interacts with human bias

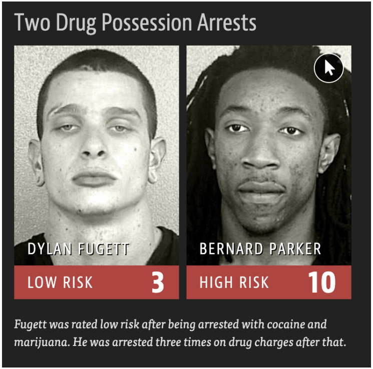
- Our contributions:
  1. experimental evaluation of algorithm-assisted human decision making
  2. principal fairness: new fairness notion based on causality
  3. real-world field experiment evaluating pretrial public safety assessment

Controversy over the COMPAS Score (Propublica)



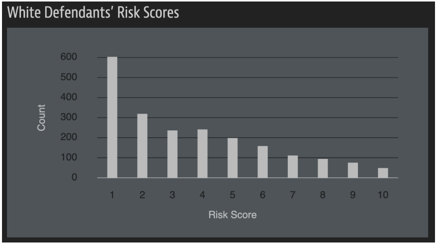
Two Petty Theft Arrests

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

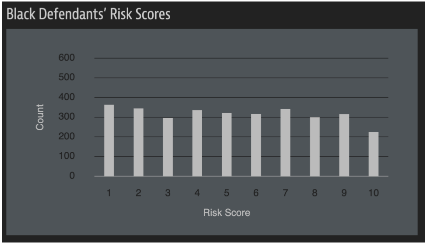


Two Drug Possession Arrests

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.



White Defendants' Risk Scores



Black Defendants' Risk Scores

Pretrial Public Safety Assessment (PSA)

- Algorithmic recommendations often used in US criminal justice system
- At the first appearance hearing, judges primarily make two decisions
  1. whether to release an arrestee pending disposition of criminal charges
  2. what conditions (e.g., bail and monitoring) to impose if released
- Goal: avoid predispositional incarceration while preserving public safety

- Judges are required to consider three risk factors along with others
  1. arrestee may fail to appear in court (FTA)
  2. arrestee may engage in new criminal activity (NCA)
  3. arrestee may engage in new violent criminal activity (NVCA)

- PSA as an algorithmic recommendation to judges
  - classifying arrestees according to FTA and NCA/NVCA risks
  - derived from an application of a machine learning algorithm to a training data set based on past observations
  - different from COMPAS score

A Field Experiment for Evaluating the PSA

- Dane County, Wisconsin
- PSA = weighted indices of ten factors
  1. two separate ordinal six-point risk scores for FTA and NCA
  2. one binary risk score for new violent criminal activity (NVCA)
  3. age as the single demographic factor: no gender or race
  4. nine factors drawn from criminal history (prior convictions and FTA)

- Judges may have other information about an arrestee
  - affidavit by a police officer about the arrest
  - defense attorney may inform about the arrestee's connections to the community (e.g., family, employment)

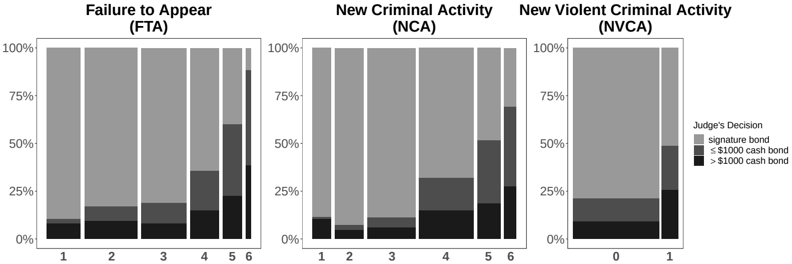
- Field experiment
  - clerk assigns case numbers sequentially as cases enter the system
  - PSA is calculated for each case using a computer system
  - if the first digit of case number is even, PSA is given to the judge
  - mid-2017 – 2019 (randomization), 2-year follow-up for half sample

PSA Provision, Demographics, and Outcomes

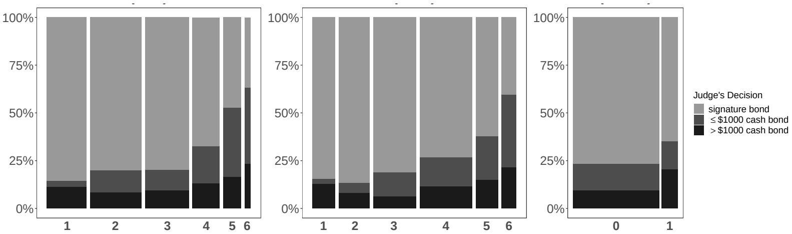
	no PSA			PSA			
	Signature	Cash bond		Signature	Cash bond		
	bond	small	large	bond	small	large	Total (%)
Non-white female	64	11	6	67	6	0	154 (8)
White female	91	17	7	104	17	10	246 (13)
Non-white male	261	56	49	258	53	57	734 (39)
White male	289	48	44	276	54	46	757 (40)
FTA committed	218	42	16	221	45	16	558 (29)
not committed	487	90	90	484	85	97	1333 (71)
NCA committed	211	39	14	202	40	17	523 (28)
not committed	494	93	92	503	90	96	1368 (72)
NVCA committed	36	10	3	44	10	6	109 (6)
not committed	669	122	103	661	120	107	1782 (94)
Total (%)	705	132	106	705	130	113	1891
	(37)	(7)	(6)	(37)	(7)	(6)	(100)

Judge's Decision Is Positively Correlated with PSA

(a) Treatment Group

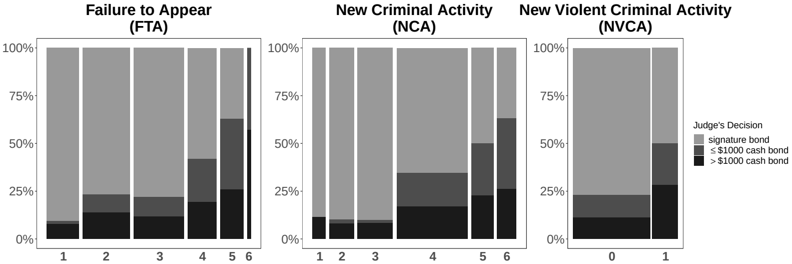


(b) Control Group

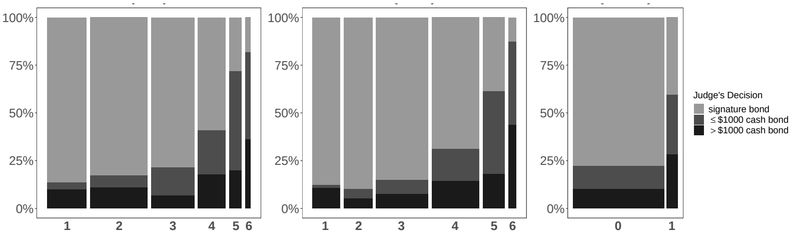


Racial Differences between Non-white and White Males

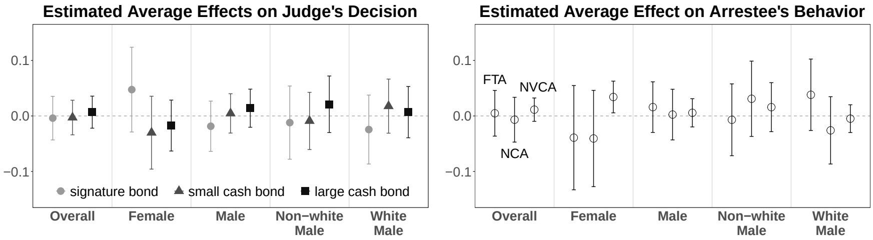
(a) Non-White Males



(b) White Males



Intention-to-Treat Analysis of PSA Provision

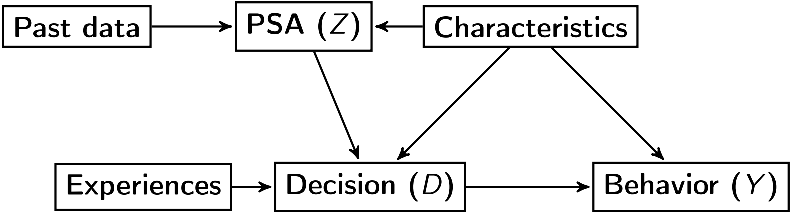

(a) Estimated effects on judge's decisions

(b) Estimated effects on outcomes

- Difference-in-means estimator
- Insignificant effects on judge's decisions
- Possible effect on NVCA outcome for females
- Need to explore causal heterogeneity based on risk-levels

Causal Inference

- Causal model



- Potential outcomes $\rightsquigarrow$ Fundamental problem of causal inference
  - $D(Z=1)$: Judge's decision with PSA
  - $D(Z=0)$: Judge's decision without PSA
  - $Y(D=1)$: Arestee's behavior if detained
  - $Y(D=0)$: Arestee's behavior if released
- Causal effects for different risk levels
  - Preventable case: $\mathbb{E}[D(1)-D(0) \mid Y(1)=0, Y(0)=1]$
  - Safe case: $\mathbb{E}[D(1)-D(0) \mid Y(1)=0, Y(0)=0]$
  - Risky case: $\mathbb{E}[D(1)-D(0) \mid Y(1)=1, Y(0)=1]$

Principal Fairness (Imai and Jiang, 2020)

- Literature focuses on the fairness of algorithmic recommendations
- We focus on the fairness of human decision

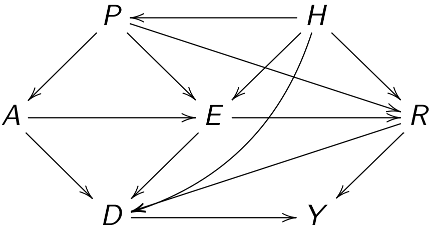
- Principal fairness: decision D should not (statistically) depend on a protected attribute A (e.g., race and gender) within a risk level R

$$D \quad \underbrace{\perp\!\!\!\perp}_{\textit{independent}} \quad A \underbrace{\ \mid\ }_{\textit{given}} R$$

- Existing statistical fairness definitions do not take into account how a decision affects individuals
  1. Overall parity: $D \perp\!\!\!\perp A$
  2. Calibration: $Y \perp\!\!\!\perp A \mid D$
  3. Accuracy: $D \perp\!\!\!\perp A \mid Y$
- These three criteria may not hold simultaneously

Relationships with the Existing Statistical Fairness Criteria

- All groups are created equal: There exist a set of covariates W such that the principal strata are conditionally independent of the protected attribute given W, i.e., $R \perp\!\!\!\perp A \mid W$.

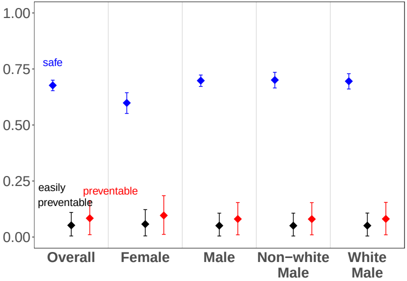


- H: historical processes
- P: parents' characteristics
- E: socio-economic factors

- Under this assumption, principal fairness implies all the other criteria

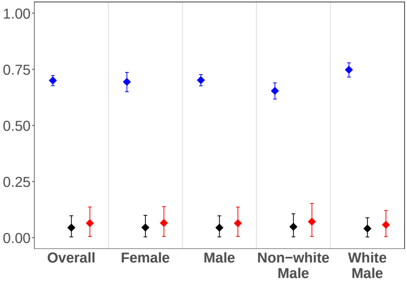
Estimated Proportion of Principal Strata

Failure to Appear (FTA)

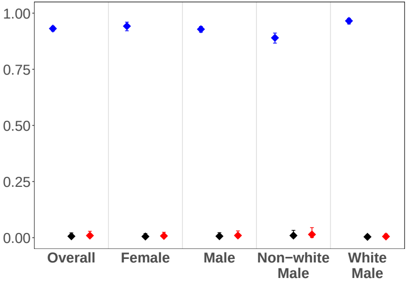


- safe cases: appear in court regardless of decision
- preventable cases: appear in court only when decision is harsh
- estimation of counterfactuals $\rightsquigarrow$ need statistics!

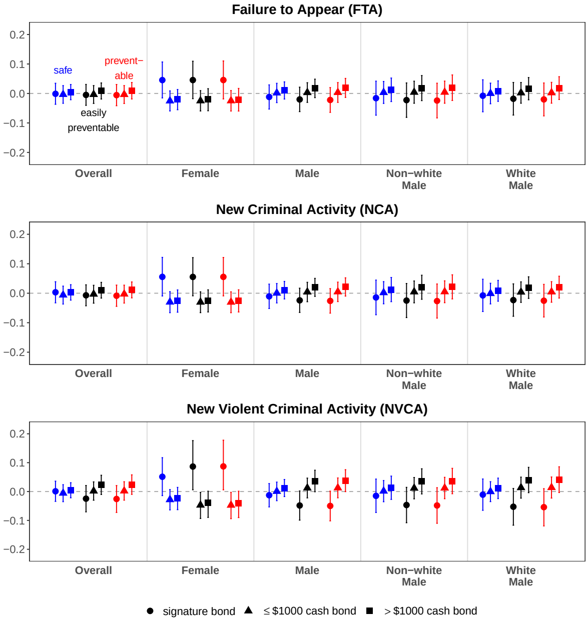
New Criminal Activity (NCA)



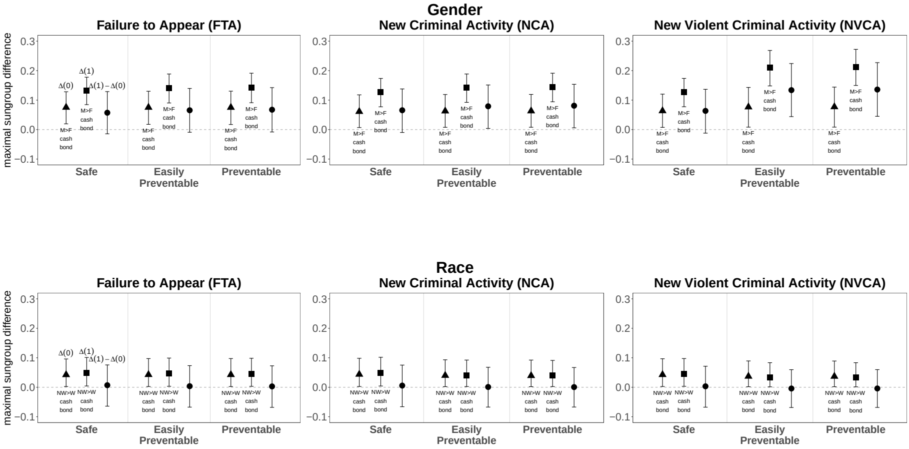
New Violent Criminal Activity (NVCA)



Estimated Average Principal Causal Effects



Principal Fairness



Concluding Remarks

- We offer a set of statistical methods for experimentally evaluating algorithm-assisted human decision making
- Field experiment for assessing the pretrial public safety assessment
  - most existing research uses observational data or hypothetical survey experiment
  - first field experiment since the small 1981–82 Philadelphia experiment about a new bond guideline
  - more ongoing experiments in this and several other counties
- Development of an open-source software package
- Ongoing research
  - extension to multi-dimensional decision (e.g., monitoring conditions)
  - role of incarceration
  - optimal PSA
  - effects of PSA on judges and arrestees over time

Importance of Quantitative Social Science

- Data analysis matters!
- It affects our policies and livelihood

- Statistics are not just for natural sciences and business
- Social scientists, policy makers, and journalists must analyze data
- Quantitative social science = Social science + Statistics
    - both are important
    - use data analysis to solve problems in the society