Housekeeping

• Friday’s meeting will take place in Blount Hall in the Traitel Building at the Hoover Institution. We will focus on a discussion of the first case study. Please come prepared having read the case materials closely.
• If you are not yet assigned to a section, please get in touch with Hilary as soon as possible.
• Reminder: We will be hosting a discussion with Reid Hoffman and Nicole Wong tonight at Cemex at 7pm. There are still seats available for students if you wish to attend.
1. Deciding to use an algorithm in the real-world
2. Interrogating the idea of efficiency gains
3. One approach: simulating outcomes
4. Second approach: implementing and measuring outcomes
5. When are algorithms useful for public policy?
6. Who should decide? How? When?
California’s SB 10

BAIL REFORM
POVERTY IS NOT A CRIME

SB10 = MASS INCARCERATION
NO SB10
In 2018, California legislators debated and voted on a bill that would:

- Eliminate cash bail and replace the system with “risk assessments” that would determine whether defendants are released and under what pre-trial, non-monetary conditions
- Require each county to set in place a system (either its own or one from a third party provider) to make this risk assessment
- Review the system in 2023 to check for bias

How would you vote? Why? What criteria would you use to make your decision?
Criteria

1. Does it work? (Efficiency)
2. Is it fair? (Fairness)
3. Can people understand how it works? (Transparency)
4. Can people appeal its judgment? (Due Process)
5. Does it use information that an individual might reasonably expect to remain private? (Privacy)
A Victory for Fair Treatment?

• “Today, California reforms its bail system so that rich and poor alike are treated fairly...”
  - Governor Jerry Brown, Aug. 28, 2018
Criteria

1. Does it work? (Efficiency)
2. Is it fair? (Fairness)
3. Can people understand how it works? (Transparency)
4. Can people appeal its judgment? (Due Process)
5. Does it use information that an individual might reasonably expect to remain private? (Privacy)

Mehran focused on issues of fairness, and whether there might be a trade off between accuracy and fairness.

Today, we are going to focus on efficiency – how would we know if we achieve better outcomes by using an algorithm?
1. Deciding to use an algorithm in the real-world
2. Interrogating the idea of efficiency gains
3. One approach: simulating outcomes
4. Second approach: implementing and measuring outcomes
5. When are algorithms useful for public policy?
6. Who should decide? How? When?
Efficiency Gains

What is the promise of using algorithmic decision-making to make bail decisions?

• It might reduce crime by more effectively/efficiently keeping dangerous criminals off of the streets

• It might reduce bias by limiting the degree to which judges take extraneous or protected characteristics into account in making their judgments

How do we know if the use of an algorithm more effectively keeps dangerous criminals off of the street?
Two Questions about Efficiency

1. Does the algorithm make more accurate decisions?
   • Issue of predictive validity
   • How it compares to other decision-making approaches

2. When introduced into the judicial process, does the use of an algorithm translate into better outcomes with respect to crime and recidivism?
   • Question of real-world impact
   • Depends a great deal on implementation
More Accurate Decisions than What?
More Accurate Decisions than Who?
Better Outcomes?

- **Observed Status**: County uses algorithmic risk assessment
- **Observed Outcome**: County experiences lower rates of crime and recidivism

Big question: did the use of the algorithm *cause* the outcome? Perhaps it would have happened anyway? The challenge is an “unobserved counterfactual”.
Counterfactual Approach

**Observed Status**
County uses algorithmic risk assessment

**Observed Outcome**
County experiences lower rates of crime and recidivism

**Counterfactual**
County does not use algorithmic risk assessment

**Unobserved Outcome**
Crime and recidivism rate?
Strategies of Causal Inference

- Gold standard: randomized controlled trial (RCT)
Is trying to get as close as possible to a clean measure of the counterfactual. We need to estimate the impact of one variable on another, *ceteris paribus* (“other things equal”). Strategies include:

- **Regression**: comparing the treated and control group with the same observed characteristics
- **Instrumental variables**: exploits partial or incomplete random assignment to identify the effect of the treatment
- **Regression discontinuity**: take advantage of arbitrary rules that occur in the world
- **Difference in differences**: leverage fact that treatment and control groups may move in parallel in absence of treatment
Today’s Agenda

1. Deciding to use an algorithm in the real-world
2. Interrogating the idea of efficiency gains
3. One approach: simulating outcomes
4. Second approach: implementing and measuring outcomes
5. When are algorithms useful for public policy?
6. Who should decide? How? When?
Beyond Predictive Accuracy

• We typically evaluate machine learning algorithms in terms of their accuracy. We do this by making predictions “out-of-sample”: using part our data to train and part to test.
• But there is a problem in the risk assessment case: we only observe outcomes for those individuals who are released by judges. We do not know whether those who are not released would reoffend or not.
• The implication is that the algorithm may perform well for some test data, but perhaps not in the real world.
• This is especially plausible in the bail case if judges have access to information about defendants that the algorithm does not (e.g. judges see gang tattoos on young people and only release those without them, who are less likely to reoffend).
Comparing Algos & Judges

• The right question is whether the algorithm improves upon judges’ decisions. But how can we evaluate this when we don’t know whether a jailed defendant would have committed a crime if released?
• Kleinberg et al (2017) focus on defendants who were not jailed, and ask whether the algorithm could do better in those cases (by jailing additional defendants)
• The results are pretty compelling:
  • The riskiest 1% of defendants, when released, fail to appear at a rate of 56% and 62% are rearrested – yet judges release nearly half of them
  • Stricter judges tend to jail defendants from throughout the risk distribution; if they jailed those with higher risk scores, they could jail half as many people and achieve the same predicted level of crime
A Popular Algorithm Is No Better at Predicting Crimes Than Random People

The COMPAS tool is widely used to assess a defendant’s risk of committing more crimes, but a new study puts its usefulness into perspective.

“Imagine you’re a judge and your court has purchased this software; the people behind it say they have big data and algorithms, and their software says the defendant is high-risk,” says Farid. “Now imagine I said: Hey, I asked 20 random people online if this person will recidivate and they said yes. How would you weight those two pieces of data? I bet you’d weight them differently. But what we’ve shown should give the courts some pause.” (A spokesperson from Equivant declined a request for an interview.)
1. Deciding to use an algorithm in the real-world
2. Interrogating the idea of efficiency gains
3. One approach: simulating outcomes
4. Second approach: implementing and measuring outcomes
5. When are algorithms useful for public policy?
6. Who should decide? How? When?
Algorithms in the Real World

Why might the effects we estimate in simulations not be realized in the real world?

There are many reasons, but we will focus on two today:

• Implementation

• Endogenous response
A critical question is how algorithmic decision-making interacts with human judgment.

“For algorithms to add value, we need people to actually use them; that is, to actually pay attention to them…” (Kleinberg et al)
• I made you read this super boring and dry report: Risk-Based Pretrial Release Recommendation and Supervision Guidelines.

• Why? What is this exercise about?
  • Examines the use of a pre-trial risk assessment tool in real life
  • Intervenes in a new way to increase the likelihood that pre-trial risk assessment is actually used by pre-trial service agencies and judges
  • Randomly assigns the intervention across 29 pre-trial services agencies
  • Measures outcomes including: (a) pre-trial officer release recommendations, judicial release decisions, etc. and (b) pre-trial outcomes (court appearance, community safety, etc.)

• One of the very few RCTs of algorithmic decision-making in criminal justice.
Decisions About Implementation

• Suppose you have a really awesome machine learning algorithm that predicts recidivism and you want to make it available to judges. What decisions do you need to make if you want it to be used?

1. When in the judicial process to provide the information?
2. Who to provide the information to (e.g. pre-trial services agencies, judges, etc.)?
3. How to provide the information?
4. How much the information should be weighed in making the decision?

All of these choices could have HUGE effects on the outcomes.
### Ex. How to Present the Information

<table>
<thead>
<tr>
<th>Method 1</th>
<th>Method 2</th>
<th>Method 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Low</td>
<td>![Traffic Light]</td>
</tr>
<tr>
<td>1</td>
<td>Low</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Below Average</td>
<td>![Green Light]</td>
</tr>
<tr>
<td>3</td>
<td>Average</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Above Average</td>
<td>![Question Mark]</td>
</tr>
<tr>
<td>5</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>High</td>
<td></td>
</tr>
</tbody>
</table>

Which approach do you prefer? Why?
Do you have any concerns about the use of a Praxis to guide decisions?
Measuring Outcomes

The study of the deployment of the Praxis tool demonstrates that:

1. Officers trained on the tool were 2.3 times more likely to recommend pre-trial release on PR or UA bond
2. Judges released defendants 1.9 times more often when pre-trial recommendations were informed by the Praxis tool
3. Defendants assigned to the Praxis treatment are significantly less likely to experience a pre-trial failure (e.g. to appear, to commit a new crime)

This is good, but... potential omitted variables, tiny sample size, non-independence of observations, etc.
Why might the effects we estimate in simulations not be realized in the real world?

There are many reasons, but we will focus on two today:

- Implementation
- Endogenous response
Ex. Indian Health Service

Figure 1
Presence of Regular ANM, random checks

Ex. Using Algorithms to Optimize

- Setting: Assignment of entering freshmen at Air Force Academy to peer groups designed to maximize the performance of lowest ability students.
- Application of ML: Harness historical data on freshmen year performance to design “optimal” peer groups – critical insight was that low ability students benefit from being with peers who have high verbal SAT scores.
- Experimental Design: Half of incoming students are assigned to peer groups randomly; the other half are assigned via the optimal matching algorithm that places low and high ability students together.
- Empirical Results: Low ability students do significantly worse in optimally matched peer group.
Results

Predicted GPA of Bottom Third Treatment and Control
(excludes predicted peer effects)

Actual GPA of Bottom Third Treatment and Control
Interpretation

• What do you think happened?

• Carrell et al (2013) argue that, when sorted into groups with only low and high ability individuals, each group segregated into its own “study” groups – eliminating the beneficial effect of being assigned to the same peer group.

• Can you imagine any endogenous response of introducing a risk assessment algorithm into the criminal justice system?
1. Look for policy problems that hinge on prediction
2. Make sure you’re comfortable with the outcome you are predicting
3. Check for bias
4. Verify your algorithm in an experiment on data it hasn’t seen, preferably in the real world
5. Remember there’s still a lot we don’t know, especially about algorithms as decision aids and the endogenous response