Building a Robot Judge

Data Science for Decision-Making
ETH Zurich, Fall 2021

Welcome to the course!

Instructions before we begin:
(1) Turn on video and set audio to mute
(2) In Participants panel, set zoom name to "Full Name, School, Dept/Major"
(ex: "Leon Smith, ETH Computer Science")
(3) Say "hi" in the chat

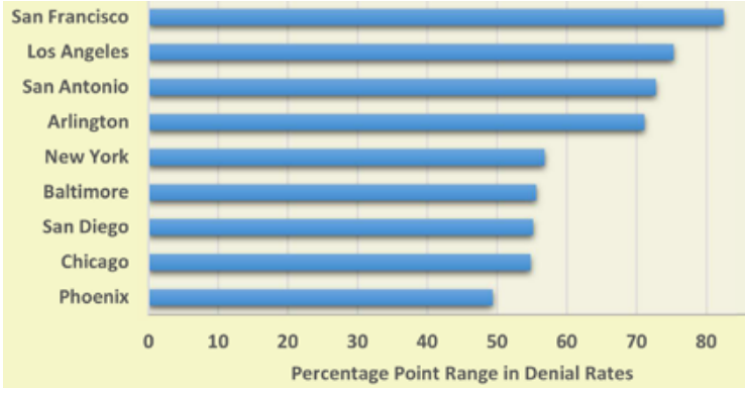
Building a Robot Judge

Data Science for Decision-Making

1. Course Overview and Introduction

What's the matter with human decision-making?

U.S. Asylum Courts: Disparities in Grant Rates



- In San Francisco, one judge grants 90.6% of asylum requests, while another judge grants just 2.9%!

Jailing Decisions Before/After Lunch Breaks

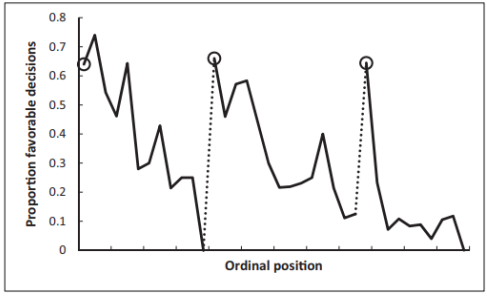
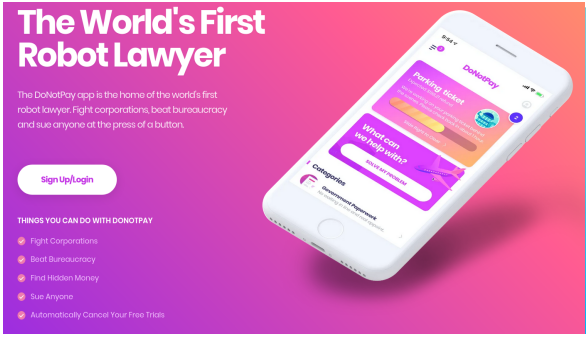


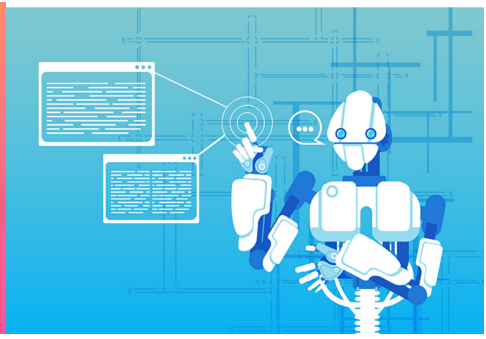
Fig. 1. Proportion of rulings in favor of the prisoners by ordinal position. Circled points indicate the first decision in each of the three decision sessions; tick marks on *x* axis denote every third case; dotted line denotes food break. Because unequal session lengths resulted in a low number of cases for some of the later ordinal positions, the graph is based on the first 95% of the data from each session.

Source: Danziger et al, PNAS 2011, Israel judges deciding on parole.

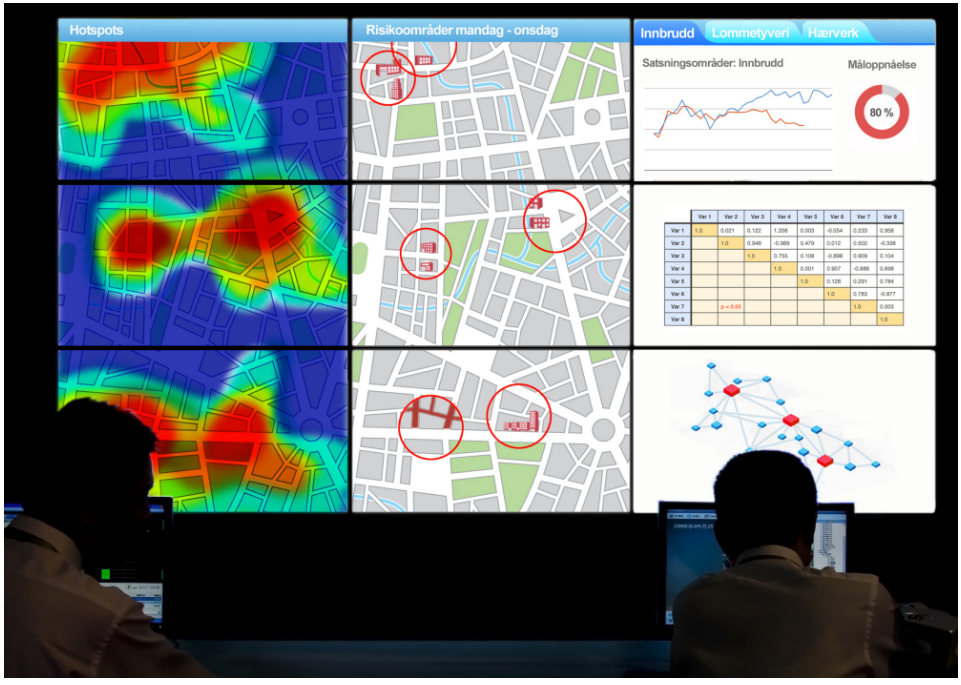
How about robot decision-making?

Your Court-Appointed Chatbot – Is Artificial Intelligence Threatening the Legal Profession?





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

Zoom Poll 1.1

Welcome to ***Building a Robot Judge***

- This course focuses on **machine learning** and **causal inference** for **decision-making**.
  - **expert** decision-making requiring **judgment** – not just legal but also medical, political, etc.

Welcome to ***Building a Robot Judge***

- This course focuses on **machine learning** and **causal inference** for **decision-making**.
  - **expert** decision-making requiring **judgment** – not just legal but also medical, political, etc.
- Engineering goals:
  - Develop tools for "building a robot judge" – machine prediction and support of expert decisions.

Welcome to ***Building a Robot Judge***

- This course focuses on **machine learning** and **causal inference** for **decision-making**.
  - **expert** decision-making requiring **judgment** – not just legal but also medical, political, etc.
- Engineering goals:
  - Develop tools for "building a robot judge" – machine prediction and support of expert decisions.
- Scientific goals:
  - Understand the factors underlying decisions of judges.

Welcome to ***Building a Robot Judge***

- This course focuses on **machine learning** and **causal inference** for **decision-making**.
  - **expert** decision-making requiring **judgment** – not just legal but also medical, political, etc.
- Engineering goals:
  - Develop tools for "building a robot judge" – machine prediction and support of expert decisions.
- Scientific goals:
  - Understand the factors underlying decisions of judges.
  - Assess the real-world impacts of decisions on society – e.g. defendants, patients.

Learning objectives

Learning objectives

1. **Implement and evaluate machine learning pipelines.**

Learning objectives

1. **Implement and evaluate machine learning pipelines.**
   - Evaluate (find problems in) existing machine learning pipelines.
   - Design a pipeline to solve a given ML problem.
   - Implement some standard pipelines in Python.

Learning objectives

1. **Implement and evaluate machine learning pipelines.**
   - Evaluate (find problems in) existing machine learning pipelines.
   - Design a pipeline to solve a given ML problem.
   - Implement some standard pipelines in Python.
2. **Implement and evaluate causal inference designs.**

Learning objectives

1. **Implement and evaluate machine learning pipelines.**
   - Evaluate (find problems in) existing machine learning pipelines.
   - Design a pipeline to solve a given ML problem.
   - Implement some standard pipelines in Python.
2. **Implement and evaluate causal inference designs.**
   - Evaluate (find problems in) causal claims.
   - Apply the standard research designs to produce causal evidence for a given empirical setting – or articulate why it is not possible.
   - Implement these research designs using regressions.

Learning objectives

1. **Implement and evaluate machine learning pipelines.**
   - Evaluate (find problems in) existing machine learning pipelines.
   - Design a pipeline to solve a given ML problem.
   - Implement some standard pipelines in Python.
2. **Implement and evaluate causal inference designs.**
   - Evaluate (find problems in) causal claims.
   - Apply the standard research designs to produce causal evidence for a given empirical setting – or articulate why it is not possible.
   - Implement these research designs using regressions.
3. **Understand how (not) to use data science tools (ML and CI) to support expert decision-making.**

Learning objectives

1. **Implement and evaluate machine learning pipelines.**
   - Evaluate (find problems in) existing machine learning pipelines.
   - Design a pipeline to solve a given ML problem.
   - Implement some standard pipelines in Python.
2. **Implement and evaluate causal inference designs.**
   - Evaluate (find problems in) causal claims.
   - Apply the standard research designs to produce causal evidence for a given empirical setting – or articulate why it is not possible.
   - Implement these research designs using regressions.
3. **Understand how (not) to use data science tools (ML and CI) to support expert decision-making.**
   - Explore the connections/distinctions between **prediction**, **inference**, and **decisions**.
   - Evaluate proposed policies/systems that use algorithms for decision support – along accuracy, bias, gaming, and other dimensions.
   - Read and critique research papers reporting on these policies/systems.
   - If you are signed up for the project: Implement/analyze such a system and write a paper about it.

Outline

Lecture Times

- Mondays, 1415h-16h
  - Zoom (Meeting ID 927 5461 2589)
- 10 minute break, 15h-1510h

Online Lecture Norms

Let's make the most of online learning!

- Live attendance at lectures is required – we will monitor attendance through group work, can bump up grades for good participation.
- Keep video on if connection allows.
- Stay muted when not talking.
- To make questions or comments, type in the chat (private or public) or use the "raise hand" function.

Online Course Materials

- Course Syllabus:
  - `https://bit.ly/BRJ_syll`
- Course Repo (slides, notebooks, and assignments):
  - `https://github.com/elliottash/robot_judge_2021`

Teaching Assistant Claudia Marangon

Claudia Marangon (claudia.marangon@gess.ethz.ch)

- weekly TA sessions to go over code notebooks and assignments.
  - not mandatory – attend if you are new to the tools.
  - we will also post recordings.
- can answer questions about lectures, notebooks, assignments, and projects.

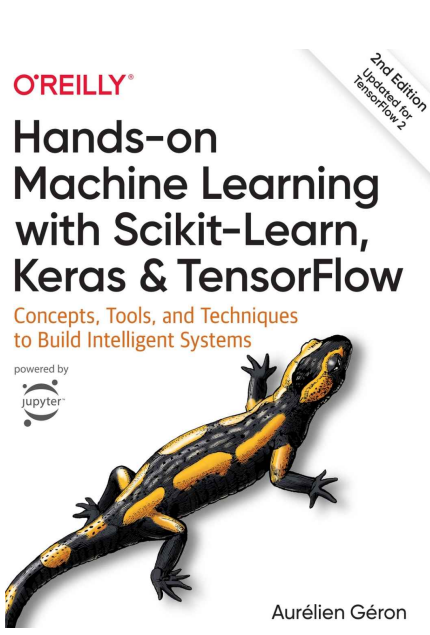
Course Communication

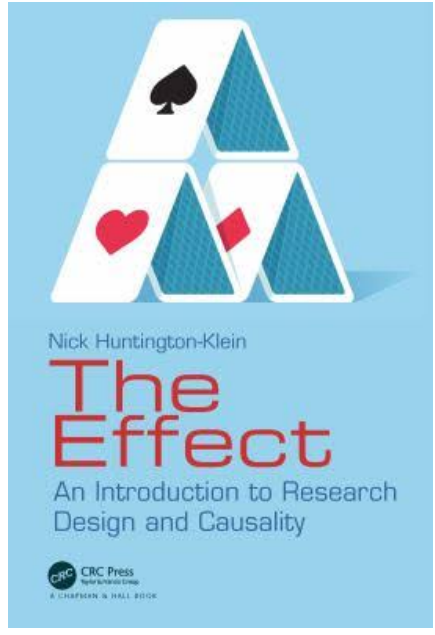
- Course communication will be done through eDoz.
- Questions welcome via email, to me or to the TA's.
- I will be available in the zoom 5 minutes early, during the mid-lecture break, and for 10 minutes after the end of lecture.
- Will schedule meetings with students doing projects.

Reading List

`https://bit.ly/BRJ-readings`

- There are a handful of required readings (highlighted).
  - for use in classroom activities
  - could ask about them in final assignment
  - first one, for today, is a short news article, can read it during the break.
- Other readings can be used as reference:
  - to complement the slides
  - to be used for reading response essays (more next week)
  - lit review for projects


O'REILLY
2nd Edition
Updated for TensorFlow 2
Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow
Concepts, Tools, and Techniques to Build Intelligent Systems
powered by
jupyter
Aurélien Géron


Nick Huntington-Klein
The Effect
An Introduction to Research Design and Causality
CRC Press
A CHAPMAN & HALL BOOK

Programming Material is Python-Oriented

- Python 3.7 is ideal for machine learning.
- You can use Anaconda or download the packages we need to a pip environment.

Programming Material is Python-Oriented

- Python 3.7 is ideal for machine learning.
- You can use Anaconda or download the packages we need to a pip environment.
- Econometrics:
  - applied economists use stata, which is closed-source statistical software
  - but we will have python versions of everything

Programming Material is Python-Oriented

- Python 3.7 is ideal for machine learning.
- You can use Anaconda or download the packages we need to a pip environment.
- Econometrics:
  - applied economists use stata, which is closed-source statistical software
  - but we will have python versions of everything
- See the syllabus for lists of packages.
- you can use a different programming language for the assignment if you want – if so, email me about it.

Course Workload

3 ECTS credits $\approx$ 90 hours of work

Course Workload

3 ECTS credits $\approx$ 90 hours of work

- 13 lectures, 1.75 hours each $=$ 23 hours
- 10 ungraded programming assignments, ~1 hour each $\approx$ 10 hours
- Required readings (three papers and a few short articles/snippets) $\approx$ 9 hours
- 3 response essays, ~6 hours each $\approx$ 18 hours
- Final assignment / take-home test, 4 hours
- $\approx$ **64 required hours.**

Course Workload

3 ECTS credits $\approx$ 90 hours of work

- 13 lectures, 1.75 hours each = 23 hours
- 10 ungraded programming assignments, ~1 hour each $\approx$ 10 hours
- Required readings (three papers and a few short articles/snippets) $\approx$ 9 hours
- 3 response essays, ~6 hours each $\approx$ 18 hours
- Final assignment / take-home test, 4 hours
- $\approx$ **64 required hours.**
- $\approx$ 26 hours at student discretion:
  - 12 optional TA sessions, 1 hour each $\approx$ 12 hours
  - leaves ~14 hours for additional study time

Course Projects

2 additional ECTS credits $\approx$ 60 additional hours of work

- About twice as much out-of-class work expected
  - previous course projects have turned into conference/journal publications.
  - two projects turned into funded Innoisuisse startups.

Course Projects

2 additional ECTS credits $\approx$ 60 additional hours of work

- About twice as much out-of-class work expected
  - previous course projects have turned into conference/journal publications.
  - two projects turned into funded Innoisuisse startups.
- Can be done individually or in small groups (preferably 2, up to 4 with good reason).

Course Projects

2 additional ECTS credits $\approx$ 60 additional hours of work

- About twice as much out-of-class work expected
  - previous course projects have turned into conference/journal publications.
  - two projects turned into funded Innoisuisse startups.
- Can be done individually or in small groups (preferably 2, up to 4 with good reason).
- Information session after Week 2 lecture (~10 minutes)
  - we have a list of potential topic ideas.

Question/Comment Padlets

- linked on the syllabus.
- first one: `https://padlet.com/eash44/hgzg8ogr1pqnw7ok`

Outline

Zoom Poll 1.2

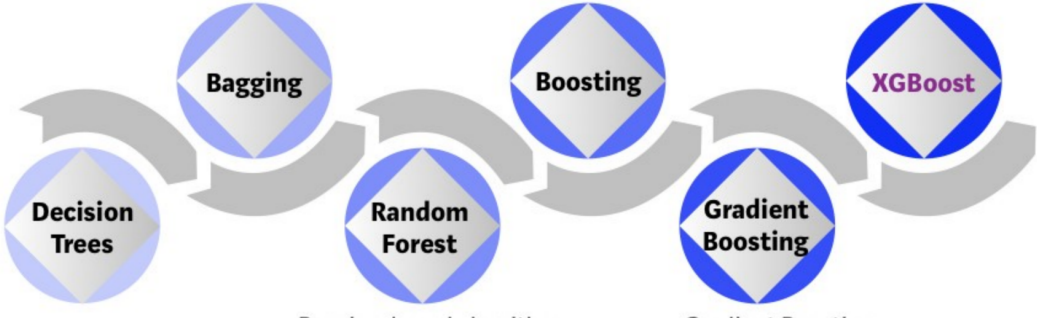
Implement and evaluate machine learning pipelines

Implement and evaluate machine learning pipelines

- Evaluate (find problems in) existing machine learning pipelines.
- Design a pipeline to solve a given ML problem.
- Implement some standard pipelines in Python.

- Week 03 Machine Learning Essentials
- Week 05 Classification
- Week 07 Deep Learning Essentials
- Week 09: Encoders and Explanation

“Extreme Gradient Boosting”: Ingredients



Complicated in theory, easy in practice

```
from xgboost import XGBClassifier
model = XGBClassifier()

model.fit(X_train, y_train,
          early_stopping_rounds=10,
          eval_metric="logloss",
          eval_set=[(X_eval, y_eval)]
          )

y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
```

Predicting U.S. Asylum Court Decisions

Predicting U.S. Asylum Court Decisions

		Predicted	
		Denied	Granted
True	Denied	195,223	65,798
	Granted	73,269	104,406

Accuracy = 68.3%, F1 = 0.60

- Prediction App (Beta): `https://floating-lake-11821.herokuapp.com/`

Implement and evaluate causal inference designs

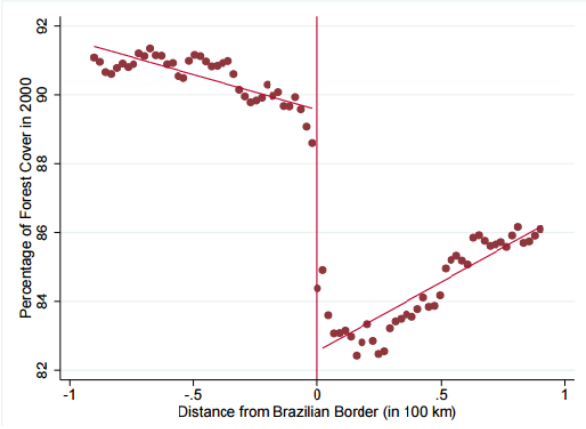
Implement and evaluate causal inference designs

- Evaluate (find problems in) causal claims.
- Apply the standard research designs to produce causal evidence for a given empirical setting – or articulate why it is not possible.
- Implement these research designs using Python or Stata.

- Week 02 Causal Inference Essentials
- Week 04 Panel Data Models
- Week 06 Machine Learning and Causal Inference
- Week 08 Instrumental Variables

http://www.tylervigen.com/spurious-correlations

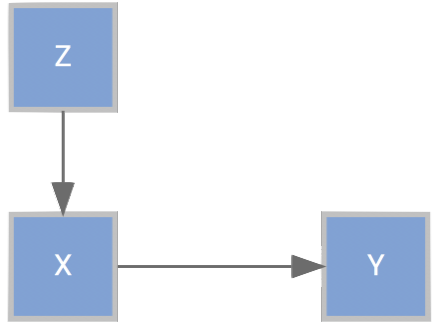
Burgess, Costa, and Olken, "The Brazilian Amazon's Double Reversal of Fortune"



Source: `https://economics.mit.edu/files/12732`

Instrumental Variables

Before reading the course syllabus, had you ever heard of instrumental variables?
– use zoom reactions for yes/no



Understand how (not) to use data science tools (ML and CI) to support expert decision-making

Understand how (not) to use data science tools (ML and CI) to support expert decision-making

- Appreciate the connections/distinctions between **prediction**, **inference**, and **decisions**.
- Evaluate proposed policies/systems that use algorithms for decision support – along accuracy, bias, gaming, and other dimensions.
- Read and critique research papers reporting on these policies/systems.

- Weeks 10-13
  - AI-supported decisions
  - AI fairness
  - Explanations
  - AI policy

The standard learning problem

- We have a dataset of predictors or features, represented as a big matrix X.
  - e.g., defendant characteristics, criminal history, etc.

The standard learning problem

- We have a dataset of predictors or features, represented as a big matrix X.
  - e.g., defendant characteristics, criminal history, etc.
- The outcome or label to predict, Y
  - e.g., whether a defendant will commit more crimes if released on bail.

The standard learning problem

- We have a dataset of predictors or features, represented as a big matrix X.
  - e.g., defendant characteristics, criminal history, etc.
- The outcome or label to predict, Y
  - e.g., whether a defendant will commit more crimes if released on bail.
- The label is a probabilistic function of the features:

$$Y = h(X)$$

A decision problem

Now consider a decision-maker who has to make a decision W, that will produce some value or benefit, conditional on the value of Y:

$$u(W; Y)$$

- e.g., whether to grant bail.
- the decision-maker only observes X, with $Y(X) = h(X) + \epsilon$.

A decision problem

Now consider a decision-maker who has to make a decision W, that will produce some value or benefit, conditional on the value of Y:

$$u(W; Y)$$

- e.g., whether to grant bail.
- the decision-maker only observes X, with $Y(X) = h(X) + \epsilon$.

The decision-maker computes a prediction $\hat{Y}(X)$ and decides

$$W^*(X) = \arg\max_W u(W, \hat{Y}(X))$$

- after Y is observed, the payoff is $u(W^*(X); Y)$.

A decision problem

Now consider a decision-maker who has to make a decision W, that will produce some value or benefit, conditional on the value of Y:

$$u(W; Y)$$

- e.g., whether to grant bail.
- the decision-maker only observes X, with $Y(X) = h(X) + \epsilon$.

The decision-maker computes a prediction $\hat{Y}(X)$ and decides

$$W^*(X) = \arg\max_W u(W, \hat{Y}(X))$$

- after Y is observed, the payoff is $u(W^*(X); Y)$.

if X includes a defendant's choices – eg education, criminal record, hiring an attorney – then X becomes a function of $W^*(X)$:

- interactive decision problem $\rightarrow$ have to consider equilibria.

Examples

- **Bansak et al (*Science* 2018):**
  - assign refugees to locations using an algorithm that predicts higher employment.
  - paper demonstrates large gains relative to random assignment of refugees.

Examples

- **Bansak et al (*Science* 2018):**
  - assign refugees to locations using an algorithm that predicts higher employment.
  - paper demonstrates large gains relative to random assignment of refugees.
- **Kleinberg et al (*Quarterly Journal of Economics*, 2018):**
  - decide on bail/parole using an algorithm that predicts recidivism (whether defendant commits another crime)
  - algorithm could reduce both incarceration rates and recidivism.

Examples

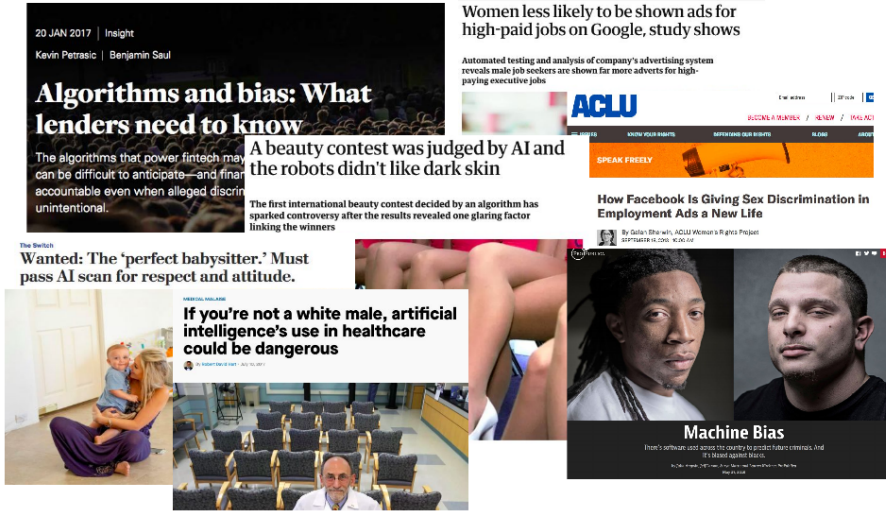
- **Bansak et al (*Science* 2018):**
  - assign refugees to locations using an algorithm that predicts higher employment.
  - paper demonstrates large gains relative to random assignment of refugees.
- **Kleinberg et al (*Quarterly Journal of Economics*, 2018):**
  - decide on bail/parole using an algorithm that predicts recidivism (whether defendant commits another crime)
  - algorithm could reduce both incarceration rates and recidivism.
- **Ash, Galletta, and Giommoni (2021):**
  - algorithm can predict fiscal corruption from budget data
  - could be used to double the detection rate of corruption relative to randomly assigned audits.

Activity (3 minutes)

Activity (3 minutes)

- Post in the menti:
  - `https://www.menti.com/drqu4gr71r`
  - An example of a decision or judgment that would be difficult to automate, and why.
  - Try to pick one that no one else picks.

Can AI decisions be biased?



Source: Hoda Heidari slides.

The AI Fairness Tradeoff

- Pros:
  - higher accuracy
  - lower cost
  - consistency – all defendants get the same decision for the same characteristics.

The AI Fairness Tradeoff

- Pros:
  - higher accuracy
  - lower cost
  - consistency – all defendants get the same decision for the same characteristics.
- Cons:
  - systematic biases – for example those in training data – could be replicated or amplified.
  - ignores special circumstances / mitigating factors
  - lack of transparency / accountability
  - issues of privacy / surveillance
  - risks of gaming the system

The AI Fairness Tradeoff

- Pros:
  - higher accuracy
  - lower cost
  - consistency – all defendants get the same decision for the same characteristics.
- Cons:
  - systematic biases – for example those in training data – could be replicated or amplified.
  - ignores special circumstances / mitigating factors
  - lack of transparency / accountability
  - issues of privacy / surveillance
  - risks of gaming the system
- Active research area on addressing these issues
  - methods for diagnosing bias / data problems
  - modal explanation methods to open the blackbox

The AI Fairness Tradeoff

- Pros:
  - higher accuracy
  - lower cost
  - consistency – all defendants get the same decision for the same characteristics.
- Cons:
  - systematic biases – for example those in training data – could be replicated or amplified.
  - ignores special circumstances / mitigating factors
  - lack of transparency / accountability
  - issues of privacy / surveillance
  - risks of gaming the system
- Active research area on addressing these issues
  - methods for diagnosing bias / data problems
  - modal explanation methods to open the blackbox
- Further: algorithms can also be used to **detect** systematic bias, to **understand** it – and therefore to help **reduce** it.

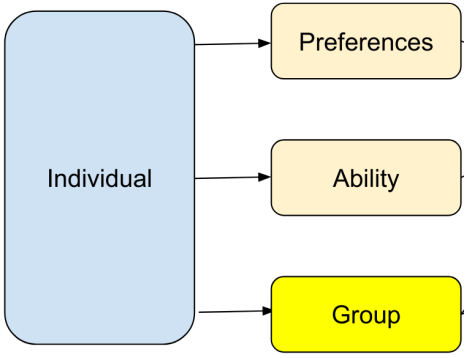
Outline

Motivation

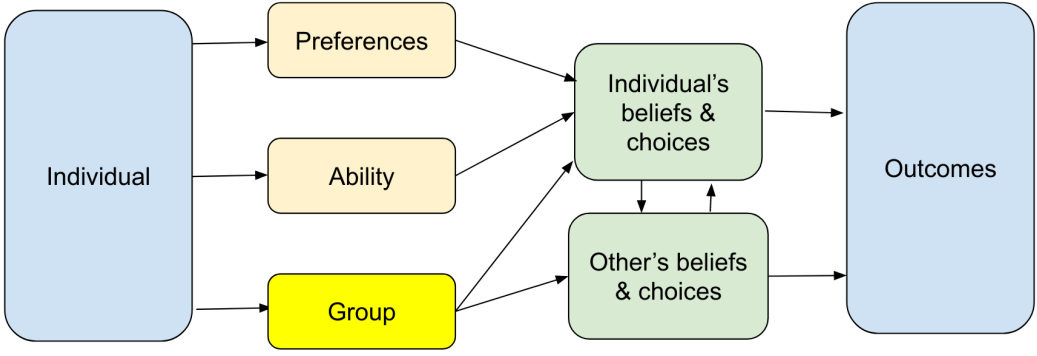
- Systematic and persistent differences in labor-market and justice-system outcomes across groups – e.g. men/women, across racial/ethnic groups.
- Anecdotally, there are clear examples of prejudice or biased treatment.

Motivation

- Systematic and persistent differences in labor-market and justice-system outcomes across groups – e.g. men/women, across racial/ethnic groups.
- Anecdotally, there are clear examples of prejudice or biased treatment.
- But disparate outcomes on average are also explained in part by differences in characteristics or choices across groups.
- If we want to reduce bias/prejudice, we have to distinguish it from other factors.



- There could be group differences in preferences and ability.
  - these could be correlated with group identity.



- Individual choices, especially around education and skills investments, depend on preferences, ability, and group identity.
- But they might also affect others' choices.

Note: This graph could be drawn in many other ways.

Gender Wage Gap

- There is clearly a long history of discrimination towards women.
  - but to address discrimination, we have to be open to other explanations for gender wage gaps
  - $\rightarrow$ helps policymakers allocate scarce resources.

Gender Wage Gap

- There is clearly a long history of discrimination towards women.
  - but to address discrimination, we have to be open to other explanations for gender wage gaps
  - $\rightarrow$ helps policymakers allocate scarce resources.

Non-discrimination reasons for the wage gap:

- women have different innate abilities
- women prefer less risky jobs, or shorter commute times, or fewer hours
- women take breaks from their career for childbearing, reducing skills/earnings.

Gender Wage Gap

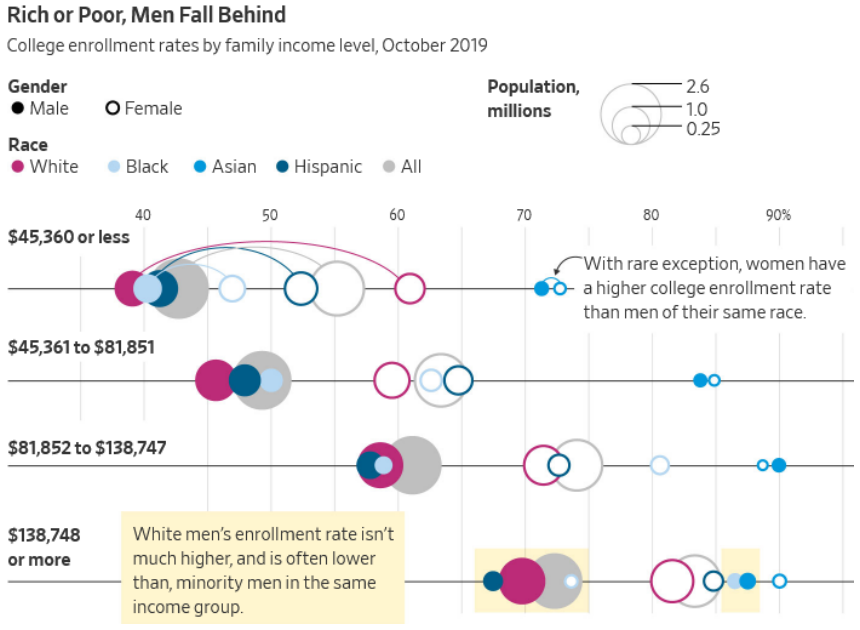
- There is clearly a long history of discrimination towards women.
  - but to address discrimination, we have to be open to other explanations for gender wage gaps
  - $\rightarrow$ helps policymakers allocate scarce resources.

Non-discrimination reasons for the wage gap:

- women have different innate abilities
- women prefer less risky jobs, or shorter commute times, or fewer hours
- women take breaks from their career for childbearing, reducing skills/earnings.

Responses to discrimination:

- women might stay out of the labor force to avoid discrimination



- Is this discrimination in college admissions? What else could it be?

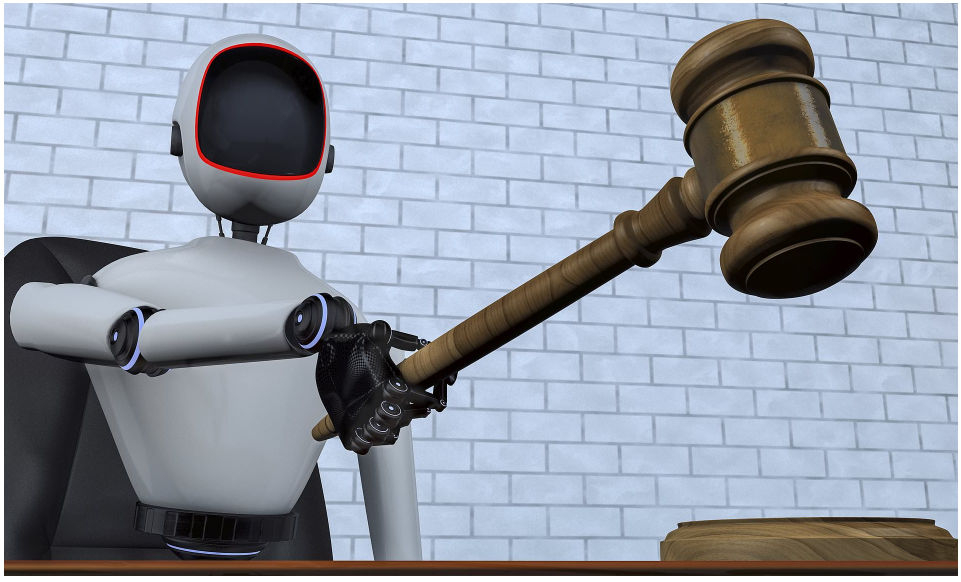
Activity: Breakout Rooms (8 minutes)

- Discuss "Biased algorithms are easier to fix than biased people" by Sendhil Mullainathan in *New York Times* (`bit.ly/nyt-bias`).
  - Think of another task where fixing biases in an algorithm is probably easier than fixing it in humans.
  - Can you think of the opposite case — a task where fixing biases in humans is easier than fixing biases in algorithms?

First Homework Assignment

Homework Assignments Page: `http://bit.ly/BRJ_HW`

- Write a short fake news article (~300 words) about a fake AI technology supporting/replacing expert decisions, such as by doctors or judges.
  - This is a completion grade – have fun with it!
  - See EduFlow page for submission instructions (due by midnight this Sunday).



Meeting Adjourned!