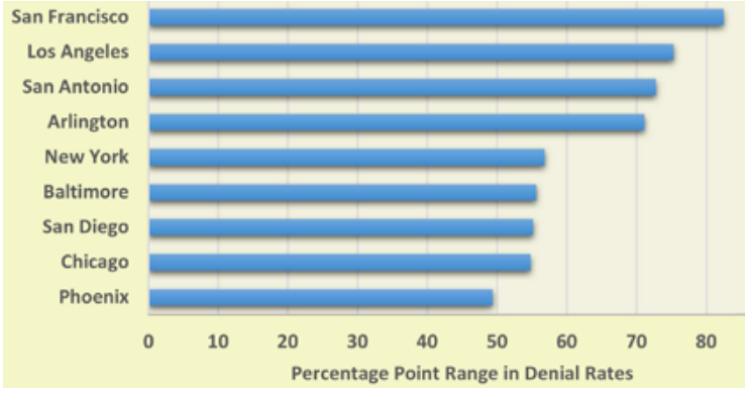
Building a Robot Judge

Data Science for Decision-Making
ETH Zurich, Fall 2023

1. Intro and Logistics

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

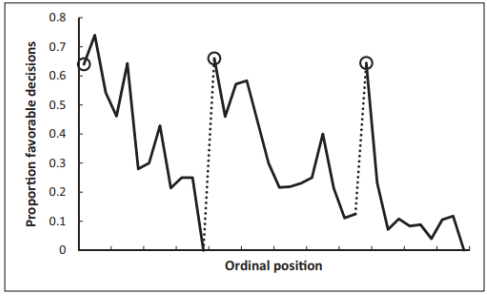
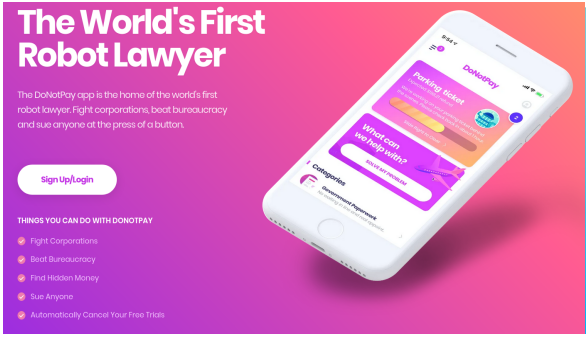


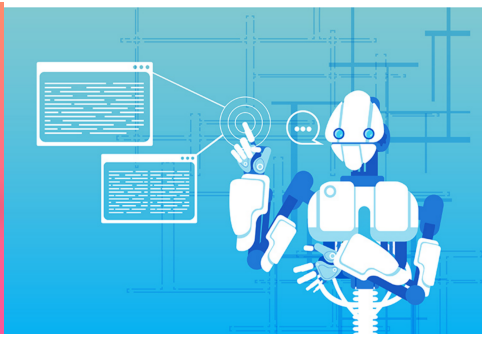
Fig. 1. Proportion of rulings in favor of the prisoners by ordinal position. Circled points indicate the first decision in each of the three decision sessions; tick marks on *x* axis denote every third case; dotted line denotes food break. Because unequal session lengths resulted in a low number of cases 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.

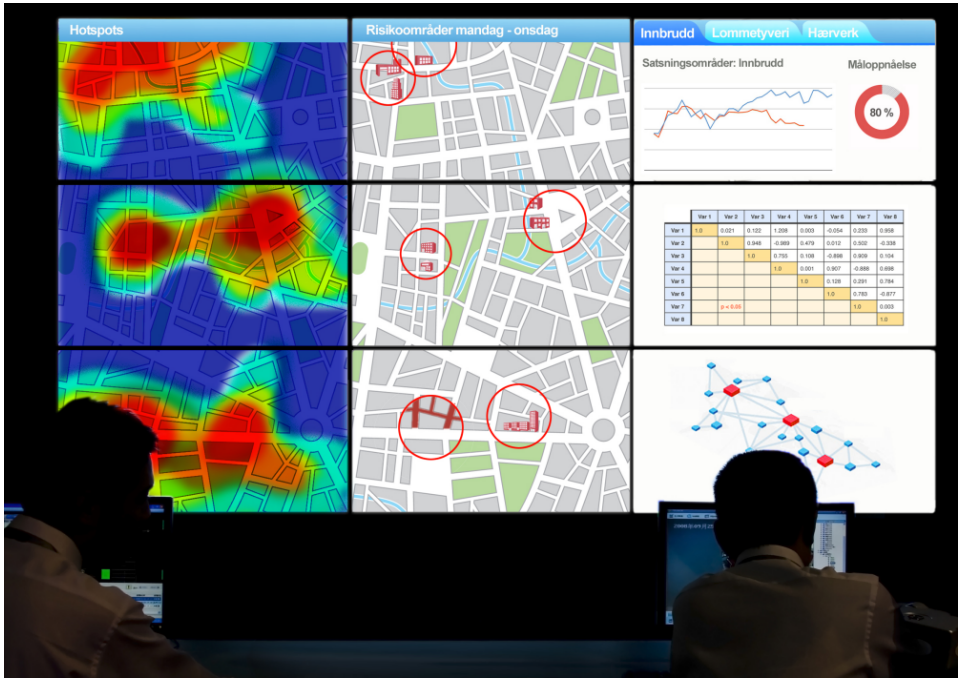
How about robot decision-making?

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

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.

Logistics: See syllabus (sent by email)

- Course Repo (slides, notebooks, and assignments):
  - `https://github.com/elliottash/robot_judge_2023`
- Course communication through Moodle.

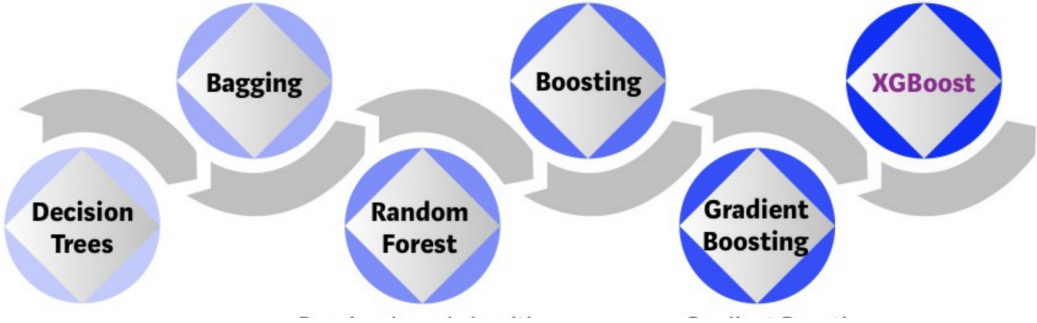
LO1: Implement and evaluate machine learning pipelines

LO1: 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 & Deep Learning
- Week 06 Machine Learning and Causal Inference
- Week 08: 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)
```

Implement and evaluate causal inference designs

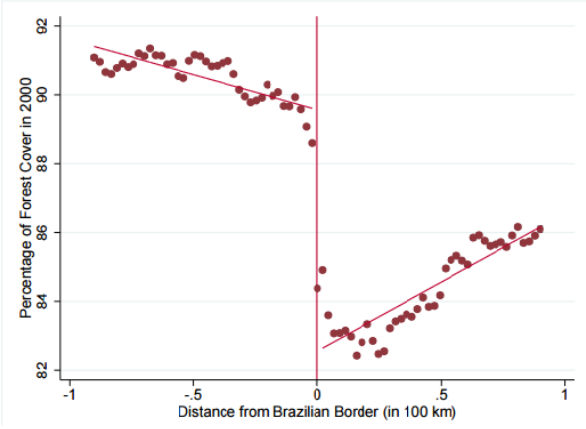
Implement and evaluate causal inference designs

- 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.

- Week 02 Causal Inference Essentials
- Week 04 Panel Data Models
- Week 06 Machine Learning and Causal Inference
- Week 07 Instrumental Variables

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

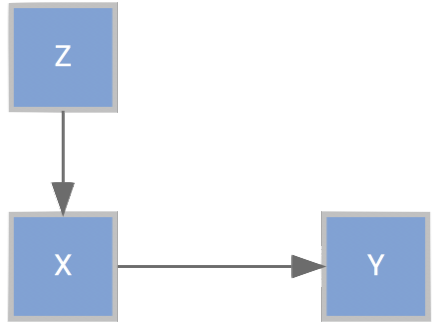
Burgess, Costa, and Olken, "The Brazilian Amazon's Double Reversal of Fortune"



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

Instrumental Variables

Before reading the course syllabus, had you ever heard of instrumental variables?



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 9-12
  - 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 $u(W^*(X); Y)$.

Further: 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.

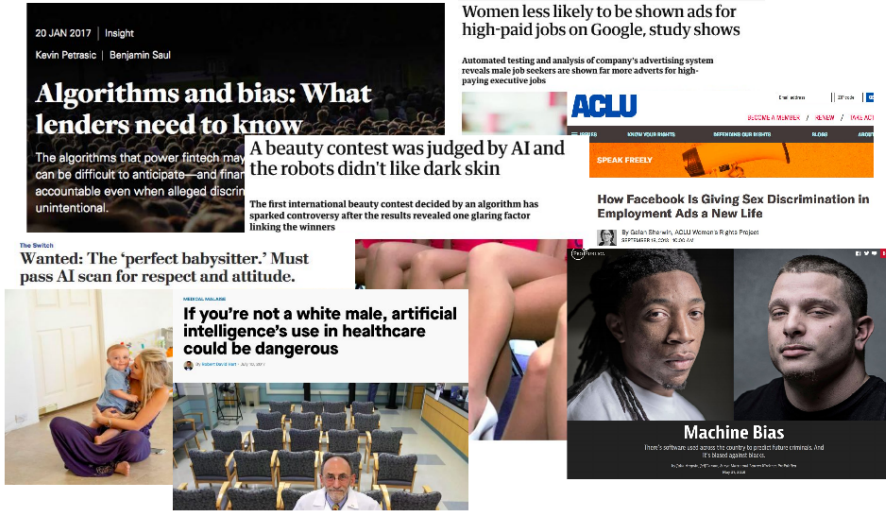
Examples

- **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.

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
  - model 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
  - model 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.