Applications of Data & Machine Learning in Economic Research: Part IV – Judge Decisions

BAI 30545 – Foundations of Economic Sciences

Julian Streyczek (Bocconi)

Previous session

Brand-new working paper on using satellite features to predict socio-economic outcomes

What Can Satellite Imagery and Machine Learning Measure?

Jonathan Proctor, Tamma Carleton, Trinetta Chong, Taryn Fransen, Simon Greenhill, Jessica Katz, Hikari Murayama, Luke Sherman, Jeanette Tseng, Hannah Druckenmiller & Solomon Hsiang

https://www.nber.org/papers/w34315, last accessed 2025-10-07

Previous session

Brand-new working paper on using satellite features to predict socio-economic outcomes

As in last week's paper:

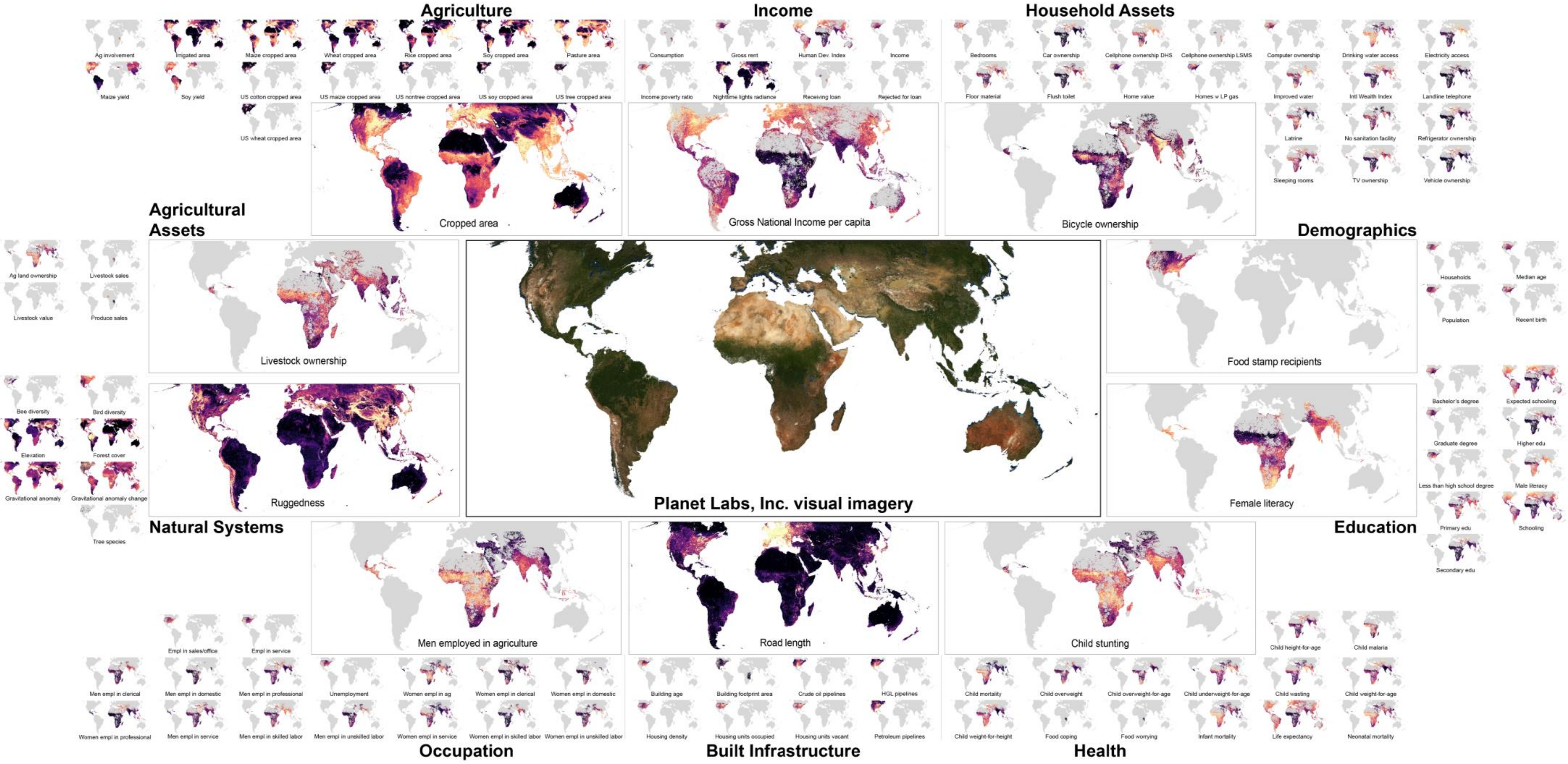
- Detect surface features from satellite images *once*
  => For each 1x1 km grid cell, get a *low-dimensional vector* of features describing the surface
  (Download as CSV on *mosaiks.org* – super easy!)
- Use simple (ridge) regression to predict socio-economic outcomes collected through surveys
  (GPD, wealth, literacy, etc.)
- Evaluate performance using R^2 metric
  (Share of variation in outcome variable explained through model)

Additional contribution:

- Look at the **entire world** and **115 different outcomes** (!)

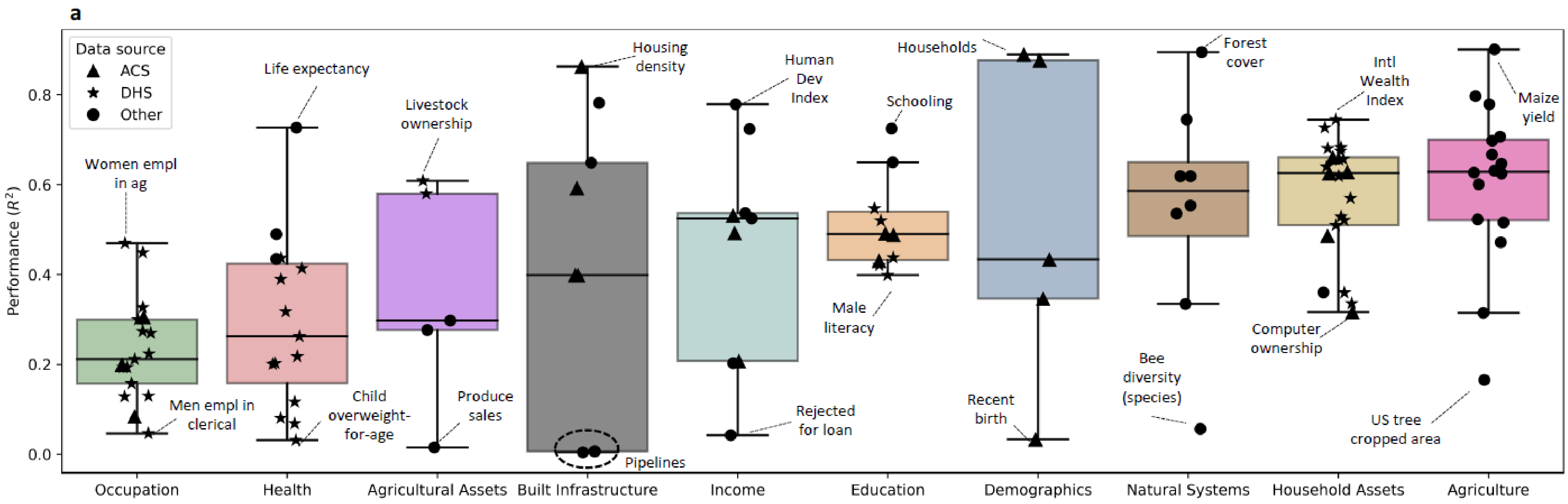
Previous session

Brand-new working paper on using satellite features to predict socio-economic outcomes



Previous session

Brand-new working paper on using satellite features to predict socio-economic outcomes



Introduction

Research Paper: Ludwig and Mullainathan (2024, QJE)

JOURNAL ARTICLE

Machine Learning as a Tool for Hypothesis Generation*

Jens Ludwig, Sendhil Mullainathan

The Quarterly Journal of Economics, Volume 139, Issue 2, May 2024, Pages 751–827,
https://doi.org/10.1093/qje/qjad055

Published: 10 January 2024 **Article history** ▾

Jens Ludwig, Sendhil Mullainathan: *"Machine Learning as a Tool for Hypothesis Generation"*,
The Quarterly Journal of Economics, Volume 139, Issue 2, May 2024, Pages 751–827,
https://doi.org/10.1093/qje/qjad055

Motivation

- *"Science is curiously asymmetric"*
  - Hypothesis **testing:** Formalized process using data and statistics
  - Hypothesis **generation:** Mysterious process using intuition and creativity

- What is ***"creativity"***?
  "Data" stored in researcher's mind, **"analyzed"** subconsciously
  => Can we (attempt to) formalize creativity as well?

- Two important **developments**:
  1. Exploding availability of **machine-readable data** on human behavior
     (text, video, prices, cellphones, etc.)
  2. Machine learning **algorithms** capable of **finding patterns** (that humans might miss)

This paper

Two levels:

- **Abstract:** Develop machine-learning-driven methodology to develop testable hypotheses about real-world patterns
- **Concrete:** Illustrate method using judges' decisions on whether to jail defendants awaiting trial
  - Key feature: Mug shots (pictures of defendants' faces)

We will proceed as follows:

1. Train a model that predicts how judges' jailing decisions depend on defendants' facial features
2. Have humans interact with the model to generate hypotheses *which* features matter

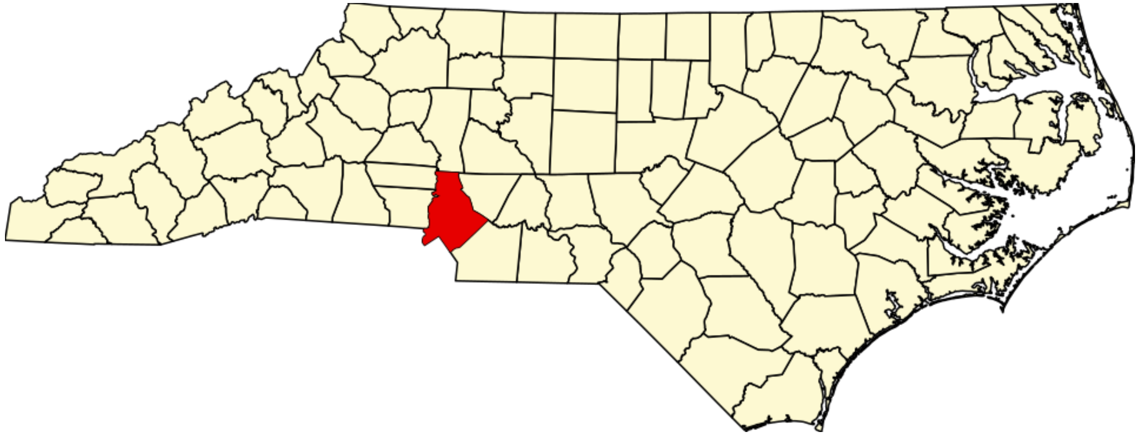
Background and Data

Setting

- "Pre-trial" hearings in the US:
  - **Setting:** After person ("defendant") is arrested for alleged crime, judge must decide within 24-48h whether defendant waits for trail at home or in jail
  - **Idea:** Jail if high risk of flight or committing another crime
  - **Consequential:**
    - Cases take several months
    - Jail time is major disruption for defendants and their families
  - **Existing research:** Judges' decisions systematically biased, based on:
    - Crime charged
    - Race
    - Weather
    - Recent performance of judges' favorite sports team
    - etc.

Data on judges' decisions in pre-trial hearings

- Mecklenburg County, North Carolina
  - 2nd-largest county in NC, home to largest city Charlotte (> 1m residents)
- 2017 through 2019
- Variables: For each criminal charge:
  - Charge characteristics (description)
  - Defendant characteristics (age, gender, race, etc.)
  - Defendant photo (400 x 480 pixels)
  - Defendant prior record (convictions, jail time, etc.)
  - Judge's pre-trail decision (detain vs. release)
- Training set: N=22,696 (train the model)
- Validation set: N=9,604 (evaluate the model)



https://it.m.wikipedia.org/wiki/File:Map_of_North_Carolina_highlighting_Mecklenburg_County.svg
https://en.wikipedia.org/wiki/Mecklenburg_County,_North_Carolina
last accessed 2025-10-07

Example data

- **Right:** Example mug shots
- **Note 1:** All photos shown are *synthetic* images (created from model used in paper that was trained on actual mug shots)
- **Note 2:** Use only data for non-Hispanic white males (homogenous sample, smaller role of racism and sexism)

Additional data

Augment images with human-generated annotations:

1. **Demographics:**
   - Age
   - Skin tone
2. Important **facial features** based on **previous scientific evidence**:
   - Trustworthiness
   - Dominance
   - Attractiveness
   - Competence

Part 1:
Modeling Judges' Decisions

Modeling Judges' Decisions

Idea

- **Objective:** Generate a model that captures judges' decisions
  - **Features:**
    - Criminal charges
    - Defendant characteristics + history

      Tabular
    - Defendant photo

      Image
  - **Outcome:** Detain vs. do not detain

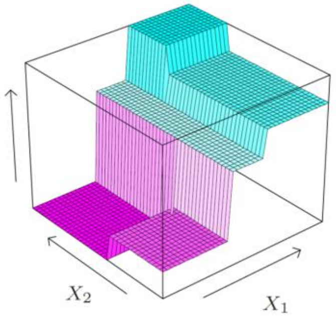
    (= wait for trial in jail vs. at home)

- **Important distinction:**
  - We do *not* model which facial characteristics predict crime
  - We model which facial characteristics predict judge behavior

Stacked model

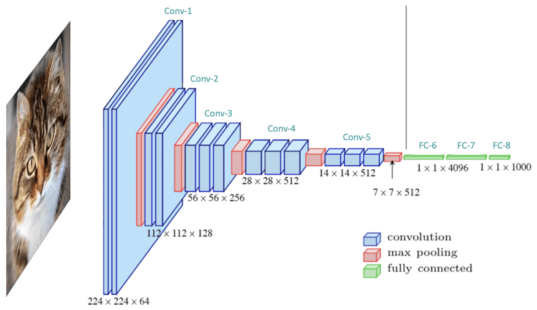
- "Stack" (average) **2 models,** one for each type of data:
  Each predicts judges' decision based only on its own data, predictions are averaged in the end:

Tabular → Gradient-Boosted Random Forest



→ Prediction 1

Image → Convolutional Neural Network



→ Prediction 2

Final Prediction

Results

- **Training:** Train model using the training set (N=22,696) and evaluate on the test set (N=9,604)
- **Measure of fit:** R^2
  (share of variation in detention decisions captured by model)
- **Results:**
  - **Full model:** $R^2 = 0.11$
  - **Only images:** $R^2 = 0.03$

    => Faces alone explain around 27% $\left(\frac{0.03}{0.11} \approx 0.273\right)$
- **Interpretation:** Faces matter for judges' detention decisions, although they should not
- **Next questions:**
  - Do these features correlate with actual criminal behavior? Paper: No! (skip here)
  - Which facial features predict detention?

Which facial features predict detention?

- **Table:** Correlation of actual judge detention decisions with model ("algo") prediction and other features
- **Method:** Linear Regression
- **How to read:**
  - Higher number = stronger correlation
  - Stars = Statistical significance (no stars: treat as zero)
- **Results:**
  - Model highly correlated with actual decision (R2 = 0.033)
  - Demographics and psychological features alone have less predictive power (R2 = 0.016)
  - Model captures more than demo. and psy. features
- => Our model discovered something new!

	Dependent variable: Judge detain decision		
	(1)	(5)	(6)
Algo judge detain prediction	0.6963^{***}		0.6262^{***}
Male		0.0940^{***}	0.0228^{*}
Age		-0.0013^{***}	-0.0015^{***}
Black		-0.0618^{***}	-0.0513^{***}
Asian		-0.0754	-0.0623
Indigenous American		0.0670	0.0585
Skin tone		-0.1004^{***}	-0.0747^{***}
Attractiveness		-0.0053	-0.0019
Competence		-0.0207^{***}	-0.0150^{**}
Dominance		0.0095^{*}	0.0071
Trustworthiness		-0.0135^{*}	-0.0105
Constant	0.0576^{***}	0.3928^{***}	0.2429^{***}
Observations	9,604	9,604	9,604
Adjusted R^2	0.0331	0.0162	0.0370

Part 2:
Generating Hypotheses

Need for generating new hypotheses

Taking stock:

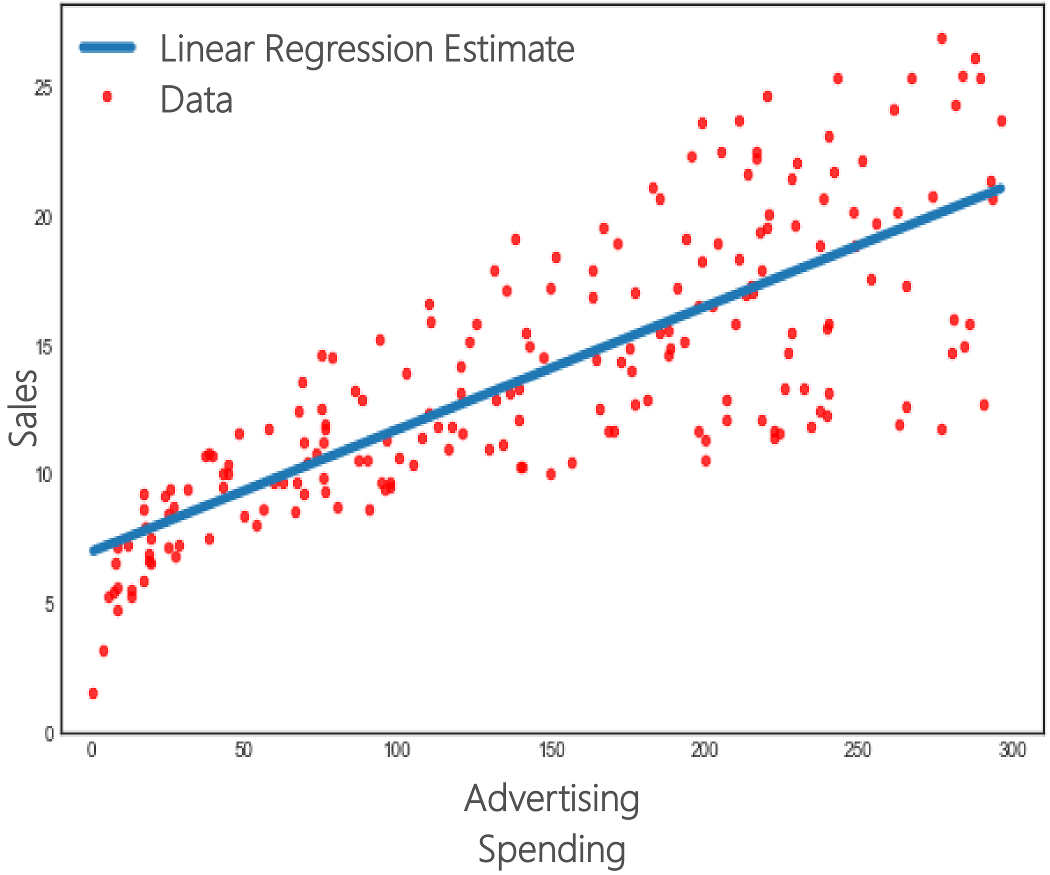
- We trained a model that predicts judges' detention decisions from tabular and image data
- We found that **facial features** matter much
- But we don't know *which* features:
  Even after accounting for known explanations, the model captures ***something*** else
  "We replaced one black-box (judge decision) with another black-box (model)"
- What is this ***something***?

Next step:

- We need to generate **new hypotheses** about which facial features matter

Naive Approach: Idea

- Given a model, its parameters, and a specific data point, we generally know how to change the data to increase the output
- **Right:** Given any data point (y,x) and our model (line), we see that we can *increase y* by *increasing x*
- **Question:** Can we do the same with our image model?
  => Given an image (x) and detention probability (y), can we *slightly* change *("morph")* facial features in x to decrease detention probability?



Matteo Courthoud: https://matteocourthoud.github.io/course/ml-econ/01_regression, last accessed 15/09/2025

Naive Approach: Result

That didn't work: The morphed face is not a "face"

Initial face

Morphed face with lower detention probability

What is a „face“?

- **Need:** An approach that, at the same time:
  1. Changes facial features to decrease detention risk
  2. Ensures the resulting image is still a „face“
- **Solution for 2:** *Generative Adversarial Network (GAN)*
  - Established method for generating synthetic data that look similar to some training set
  - Two models learn from each other by competing:
    - **Generator** tries to generate real-looking data (morphed faces)
    - **Classifier** tries to tell real from fake
- **Result:** Good *Generator* of synthetic faces, good *Classifier* of faces vs. non-faces

Morphing approach (1/2)

Combine the 2 previous ideas:

1. Change facial features such that detention probability slowly decreases
2. Restrict to “faces” generated by our *GAN*

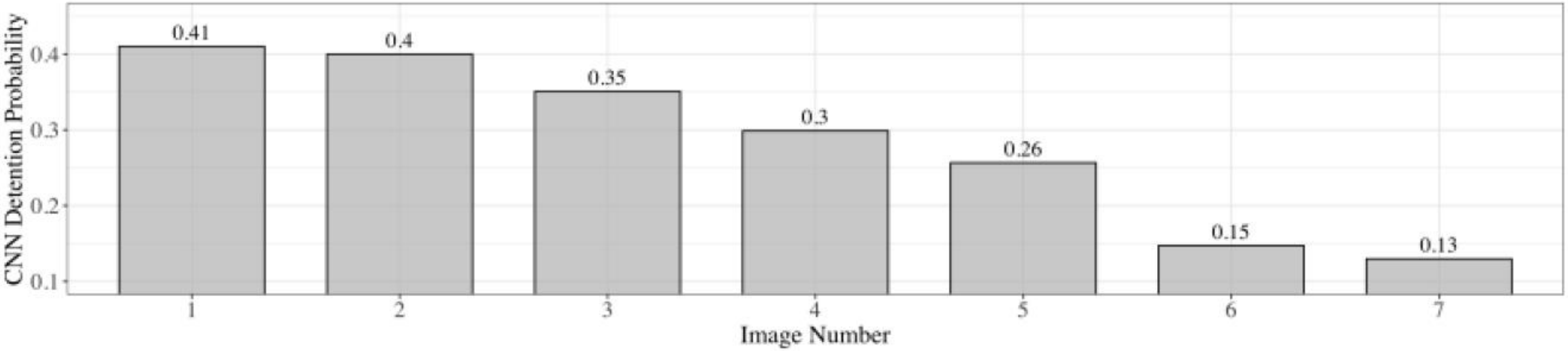
Initial face
Predicted detention probability = 41%

Morphed face
Predicted detention probability = 13%

Generating Hypotheses

Morphing approach (2/2)

(B) Transformations of the face along selected steps of the morphing process



(C) Detention probabilities for images in panel (b)

Generating New Hypotheses (1/2)

- Looking at the pictures on the right,
  can you come up with hypotheses about
  what makes faces on the **right *less likely* to be detained?**

Generating New Hypotheses (1/2)

- Looking at the pictures on the right,
  can you come up with hypotheses about
  what makes faces on the **right *less likely* to be detained?**
- Same question asked to survey participants
- Main result:
  - tidyness
  - well-groomedness
  - neatness
  - hair length

(A) A word cloud of the comments

Generating New Hypotheses (2/2)

- We have generated a new hypothesis:

„Defendants with a well-groomed face are less likely to be detained by judges"

- We can now test it:
  - Let survey participants label well-groomedness of faces on 1-9 scale
  - Check whether well-groomedness explains variation in judge decision that was not previously explained

Testing New Hypotheses

- **Table:** Correlation of well-groomedness with predicted probability that judge detains
- **Result:** Well-groomedness explains variation that we previously missed,
  even if we include other important variables
- **Magnitude:** Well-groomedness alone has R2 of 0.0247 compared to full model (R2 of 0.2361)
  => Explains around $\frac{0.0247}{0.2361} \approx 11\%$
- **Implication:** Novel result! Not mentioned in existing literature

	Dependent variable: Algorithmic judge detain prediction	
	(1)	(5)
Well-groomed	-0.0172^{***}	-0.0158^{***}
Male		0.1153^{***}
Age		0.0002^{**}
Black		-0.0165^{***}
Asian		-0.0153
Indigenous American		0.0181
Skin tone		-0.0437^{***}
Attractiveness		0.0006
Competence		-0.0062^{***}
Dominance		0.0036^{***}
Trustworthiness		-0.0024
Constant	0.3348^{***}	0.2568^{***}
Observations	9,604	9,604
Adjusted R^2	0.0247	0.2361

Iteration

We can repeat this process:

1. Generate synthetic data:
   Morph faces to decrease detention probability,
   holding well-groomedness fixed
   (-> force the model to explore *other* facial features)
2. Generate hypothesis:
   Show before / after images to humans,
   ask them to describe most obvious feature that may predict detention probability
3. Test hypothesis:
   Check whether feature actually correlates with detention probability

Result: New feature „heavy-facedness“

- Defendants on the right have **„heavier“ face** (bigger, wider, rounder)
- Explains 14% of variation in detention on its own
- **Another previously unknown result!**

Discussion

Discussion

Conclusion

- **New procedure for hypothesis generation** relying on interactions between humans and algorithm:
  1. For a given prediction problem, build
     a. a predictor
        (predict outcome given features)
     b. a data morphing procedure
        (decide which morphs are allowed)
  2. Generate pairs of initial and synthetic data, morphed such that prediction probability increases / decreases
  3. Generate hypotheses about features by having humans label key differences between initial and synthetic data

- **Showed that** defendant's face matters a lot for judges' jailing decisions
  (over and above skin color, race, etc.)

- Existing research alone cannot explain **which** facial features matters

- Show relevance of two previously unknown facial features:
  - Well-groomed
  - Heavy-faced