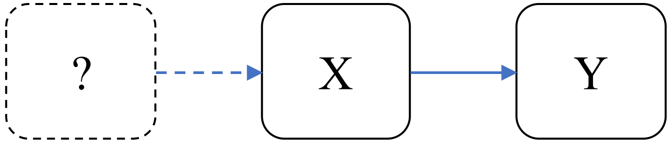
Machine Learning as A Tool for Hypothesis Generation

Jens Ludwig, Sendhil Mullainathan

QJE, 2024

Present by Long Zhen

Motivation



- Science is curiously asymmetric
  - Tested meticulously & originated intuitively
  - → idea generation is also an empirical activity but off stage
  - How to formalize?
- Two developments
  - Machine learning can find patterns that not noticed by human
  - Data on human behavior is exploding → machine readable
- → use ML to expand how hypotheses are generated

Motivation & Research question

- The key challenge:
  - One goal of science is generalization, which requires interpretability
  - The predictors produced by ML are "black boxes"

- This paper's RQ:
  - How to generate hypotheses that are both novel and testable using machine learning algorithms?

Contribution

- Literature on hypothesis generation
  - Prior literature generate hypothesis based on existing theory or economic intuition
  - This paper: propose a systematic procedure to generate hypo using ML
- Literature on machine learning in economic research
  - Prior literature:
    - New measures
    - New models
    - Causal inference tools
  - This paper: apply data-driven ML algorithms to a novel field

How?

- Two challenges:
  - Black box nature of most machine learning algorithms
    - Development in CS to create counterfactual explanations
  - Rorschach test problem – need independent subjects to inspect the outcomes
    - Use independent subjects to inspect
    - Whole new concepts that humans do not yet understand cannot be produced

- Apply to other settings:
  - Images, text, and time series are rich to explore potential hypos

A simple framework for discovery

- Criteria for hypotheses generated:
  - Novelty – orthogonalize to known factors
  - Testability – hard to define ex ante
    - Interpretability: let us generalize
    - Empirical plausibility: correlation between y – outcome of interest and $h(x)$ – hypo
- Human vs algorithm
  - Human:
    - interpretable but idiosyncratic and not necessarily replicable;
    - novel but noisy (Polanyi's paradox);
    - Not necessarily empirical plausible – over-fitting/ curse of dimensionality
  - $\rightarrow$ supervised learning: empirically plausible by construction $m(x)$
    - Not interpretable

- Related concepts:
  - Closed world problem: the fundamental laws are known, but drawing out predictions is computationally hard. E.g., protein
  - Open world problem: relation between x and y is unknown
    - ML: generate findings & hypos

An application

- Why in this US criminal justice setting?
  - Clear decision maker
  - Large samples
  - High-dimensional data
- Institutional background
  - Pretrial hearing: within 24-48 hours after arrest, a judge must decide on the bail
  - Based on the defendant’s risk of flight
  - Reality: judges systematically mispredict

Data

- Mecklenburg county, North Carolina, the second most populated county in the state
  - Representative sample √
  - The Mecklenburg County Sheriff's Office (MCSO):
    - arrest data in the last 3yrs. Demographics/charge/mug shots
  - The North Carolina Administrative Office of the Courts (NCAOC)
    - decision: detain/release/etc.
  - North Carolina Department of Public Safety
    - Defendant's prior convictions and incarceration spells
- → almost all info that judge has
- Jan 18, 2017~ Jan 17, 2020; 51,751 arrests

2017.1.18 2019.7.17 2020.1.18

70% training(70%)-plus-validation(30%), 30% test Out-of-sample

Step 0: ask human

- Ask human to label important features (HIT)

- Demographic-related: ethnicity/skin tone/age
- Psychology-related: trustworthiness/dominance/attractiveness/competence
  - Rate images on a 9-point scale

Step 1: predict judge decisions (y=1/0) using all x

- Predict judge behavior via ML
  - Gradient-boosted decision tree – structured data $m_s(x)$
  - CNN – unstructured data $m_u(x)$
  - → Combine $m_p(x) = [\hat{\beta}_s m_s(x) + \hat{\beta}_u m_u(x)]$
    - "stacking procedure" to form a single weighted-average prediction
    - (also tried fusion model, but not outperform this ensemble model)

- Do judges behave based on flight risk or cognitive error?
  - Rearrest ~ detention prediction
  - → reflects errors in the judicial decision-making process

Does the Algorithm Predict Judge Behavior after Controlling for Known Factors?

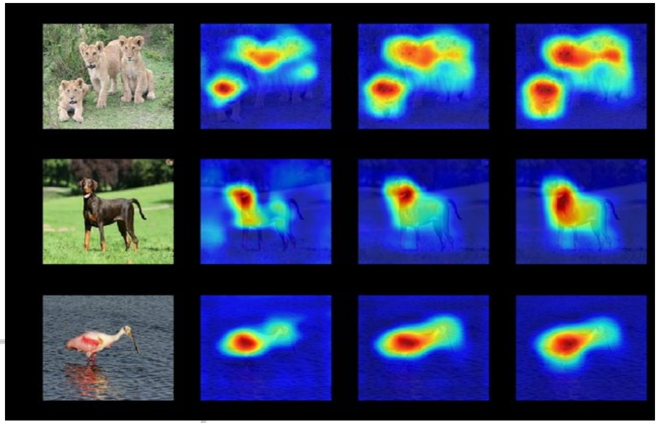
	Dependent variable: Judge detain decision						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Algo judge detain prediction	0.6963***					0.6262***	0.6171***
	(0.0383)					(0.0433)	(0.0434)
Male		0.1040***	0.0978***		0.0940***	0.0228*	0.0244**
		(0.0105)	(0.0106)		(0.0108)	(0.0117)	(0.0117)
Age		−0.0008**	−0.0009**		−0.0013***	−0.0015***	−0.0015***
		(0.0004)	(0.0004)		(0.0004)	(0.0004)	(0.0004)
Black		−0.0139	−0.0651***		−0.0618***	−0.0513***	−0.0521***
		(0.0098)	(0.0156)		(0.0156)	(0.0154)	(0.0154)
Trustworthiness				−0.0190***	−0.0135*	−0.0105	−0.0092
				(0.0070)	(0.0071)	(0.0070)	(0.0070)
Human guess							0.0852***
							(0.0265)
Constant	0.0576***	0.1868***	0.2780***	0.3054***	0.3928***	0.2429***	0.1981***
	(0.0106)	(0.0165)	(0.0272)	(0.0258)	(0.0381)	(0.0391)	(0.0415)
Naive-AUC	0.625	0.56	0.571	0.549	0.586	0.633	0.635
Observations	9,604	9,604	9,604	9,604	9,604	9,604	9,604
Adjusted R^2	0.0331	0.0101	0.0119	0.0049	0.0162	0.0370	0.0380

Is the Algorithm Rediscovering Known Facial Features?

	Dependent variable: Algorithmic judge detain prediction				
	(1)	(2)	(3)	(4)	(5)
Male	0.1186***	0.1179***	0.1153***	0.1138***	0.1140***
	(0.0025)	(0.0025)	(0.0025)	(0.0025)	(0.0025)
Age		0.0006***	0.0006***	0.0003***	0.0003***
		(0.0001)	(0.0001)	(0.0001)	(0.0001)
Black		0.0029	−0.0185***	−0.0168***	−0.0171***
		(0.0023)	(0.0037)	(0.0036)	(0.0036)
Asian		−0.0204*	−0.0232**	−0.0210*	−0.0216*
		(0.0115)	(0.0115)	(0.0114)	(0.0114)
Indigenous American		0.0103	0.0061	0.0135	0.0126
		(0.0241)	(0.0240)	(0.0238)	(0.0238)
Skin tone			−0.0441***	−0.0411***	−0.0417***
			(0.0059)	(0.0058)	(0.0058)
Attractiveness				−0.0055***	−0.0051***
				(0.0016)	(0.0016)
Competence				−0.0091***	−0.0087***
				(0.0017)	(0.0017)
Dominance				0.0037***	0.0030**
				(0.0012)	(0.0012)
Trustworthiness				−0.0048***	−0.0041**
				(0.0016)	(0.0016)
Human guess					0.0399***
					(0.0062)
Constant	0.1595***	0.1391***	0.1771***	0.2393***	0.2173***
	(0.0022)	(0.0039)	(0.0064)	(0.0089)	(0.0095)
Observations	9,604	9,604	9,604	9,604	9,604
Adjusted R^2	0.1954	0.1992	0.2038	0.2195	0.2228

Step 2: algorithm-human communication

- Saliency map: use gradient to highlight specific pixels
- build a model of the data distribution – morph



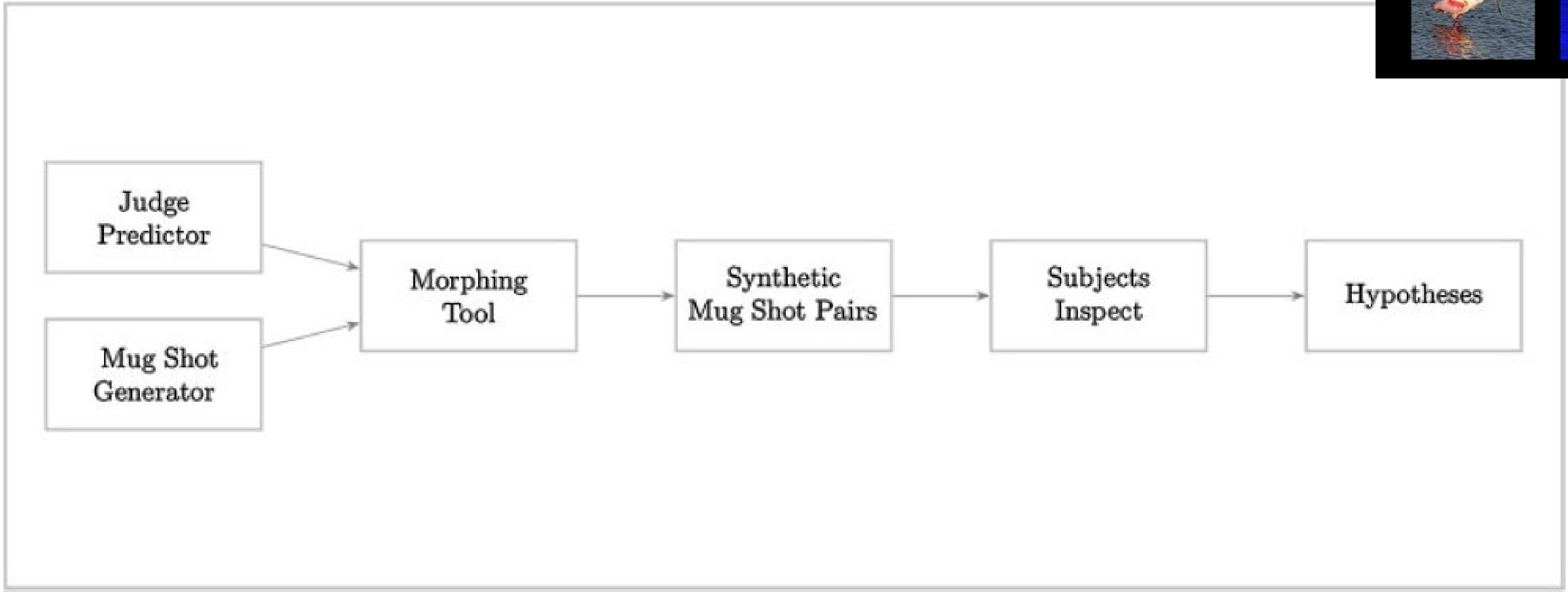
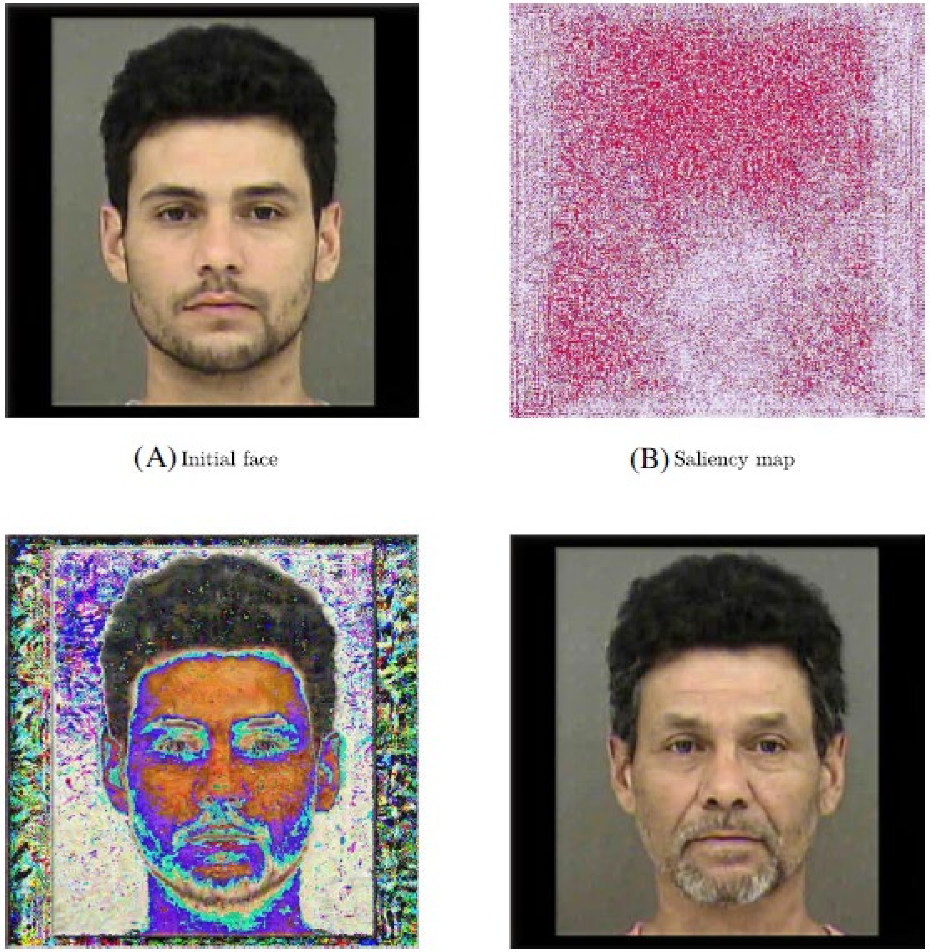


FIGURE IV

Hypothesis Generation Pipeline

How to morph?



(A) Initial face (B) Saliency map (C) Naive age-morphed image (D) Morphs from our procedure

- Take age as an example

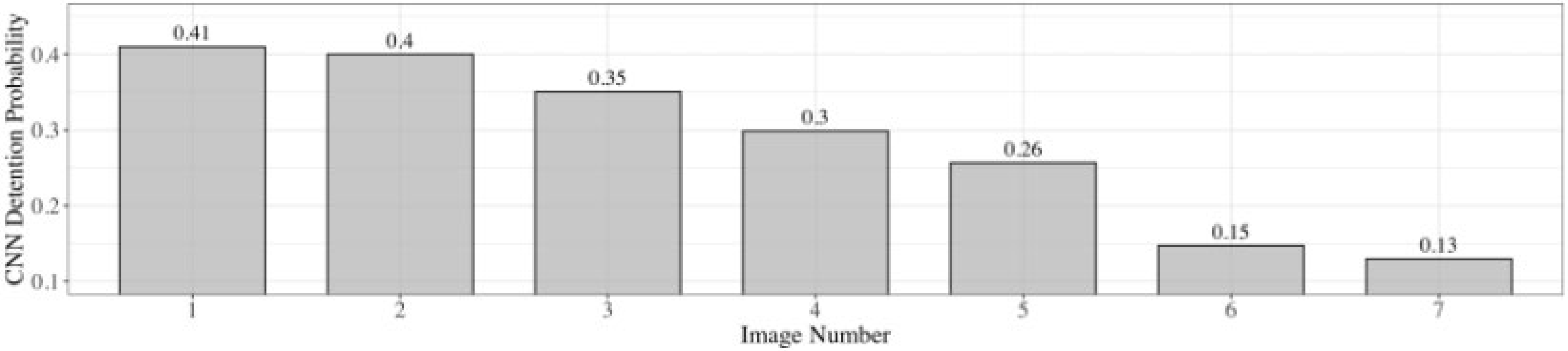
- Use the mug shot to predict age
- Get the saliency map
- change pixels in the direction of the gradient of the predicted outcome
- → change age to detention decision
  - → create the counterfactual

- Not a face?
  - Use GAN

Create detention decision morphs

- Ask subjects to articulate the differences

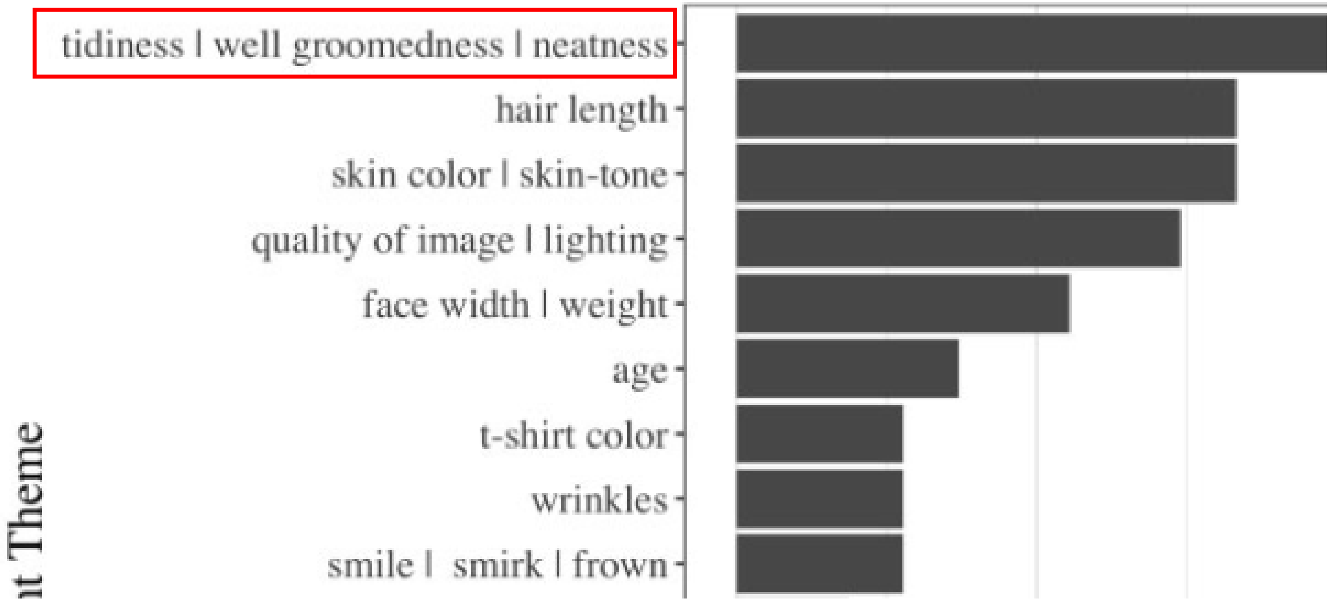
(B) Transformations of the face along selected steps of the morphing process



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

Name what differs in image pairs - hypo

(A) A word cloud of the comments



Step 3: new hypothesis evaluation

TABLE IV

CORRELATION BETWEEN WELL-GROOMED AND THE ALGORITHM'S PREDICTION

	Dependent variable: Algorithmic judge detain prediction					
	(1)	(2)	(3)	(4)	(5)	(6)
Well-groomed	-0.0172^{***}	-0.0188^{***}	-0.0184^{***}	-0.0185^{***}	-0.0158^{***}	-0.0153^{***}
	(0.0011)	(0.0010)	(0.0010)	(0.0010)	(0.0012)	(0.0012)
Male		0.1201^{***}	0.1192^{***}	0.1166^{***}	0.1153^{***}	0.1154^{***}
		(0.0024)	(0.0024)	(0.0024)	(0.0025)	(0.0025)
Age			0.0003^{***}	0.0002^{***}	0.0002^{**}	0.0002^{**}
			(0.0001)	(0.0001)	(0.0001)	(0.0001)
Black			0.0050^{**}	-0.0168^{***}	-0.0165^{***}	-0.0168^{***}
			(0.0023)	(0.0036)	(0.0036)	(0.0036)
Asian			-0.0138	-0.0165	-0.0153	-0.0160
			(0.0113)	(0.0113)	(0.0113)	(0.0113)
Indigenous American			0.0211	0.0169	0.0181	0.0172
			(0.0237)	(0.0236)	(0.0236)	(0.0236)
Skin tone				-0.0449^{***}	-0.0437^{***}	-0.0440^{***}
				(0.0058)	(0.0058)	(0.0058)
Attractiveness					0.0006	0.0008
					(0.0016)	(0.0016)
Competence					-0.0062^{***}	-0.0060^{***}
					(0.0017)	(0.0017)
Dominance					0.0036^{***}	0.0031^{**}
					(0.0012)	(0.0012)

Iteration

- Generate new hypo orthogonalized to well-groomedness
  - use training data to build predictors of detention risk, $m(x)$, and the facial features to orthogonalize against, $h_1(x)$;
  - pick a point on the GAN latent space of faces;
  - collect the gradients with respect to $m(x)$ and $h_1(x)$;
  - use the Gram-Schmidt process to move within the latent space toward higher predicted detention risk $m(x)$, but orthogonal to $h_1(x)$; and
  - show new morphed image pairs to subjects, have them name a new feature.

TABLE VI

DO WELL-GROOMED AND HEAVY-FACED CORRELATE WITH JUDGE DECISIONS?

	Dependent variable: Judge detain decision						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Heavy-faced	−0.0234***		−0.0226***	−0.0223***		−0.0218***	−0.0111***
	(0.0036)		(0.0036)	(0.0036)		(0.0037)	(0.0037)
Well-groomed		−0.0198***	−0.0185***		−0.0124**	−0.0100*	−0.0022
		(0.0043)	(0.0043)		(0.0051)	(0.0051)	(0.0051)
Algo judge detain prediction							0.5842***
							(0.0449)
Male				0.0918***	0.0959***	0.0928***	0.0269**
				(0.0107)	(0.0108)	(0.0108)	(0.0118)
Age				−0.0011***	−0.0013***	−0.0012***	−0.0014***
				(0.0004)	(0.0004)	(0.0004)	(0.0004)
Black				−0.0645***	−0.0624***	−0.0643***	−0.0535***
				(0.0156)	(0.0156)	(0.0156)	(0.0154)
Asian				−0.0737	−0.0726	−0.0701	−0.0620
				(0.0488)	(0.0489)	(0.0488)	(0.0484)
Indigenous American				0.0490	0.0683	0.0524	0.0501
				(0.1019)	(0.1021)	(0.1019)	(0.1010)
Skin tone				−0.1062***	−0.1038***	−0.1076***	−0.0801***
				(0.0250)	(0.0251)	(0.0250)	(0.0249)
Attractiveness				−0.0084	0.0004	−0.0045	−0.0025
				(0.0067)	(0.0070)	(0.0070)	(0.0070)

Conclusion

- This paper presents a new semi-automated procedure for hypothesis generation. They apply this procedure to a social issue – bailing decision – to generate two hypothesis.
- Three conditions to apply this procedure:
  - A behavior that can statistically predict
  - Unstructured, high-dimensional data
  - Can morph the input data e.g., GAN; Bi-Encoder

Extension

- Textual/ audio hypothesis?