Discovery and Hypothesis Generation with Unstructured Data

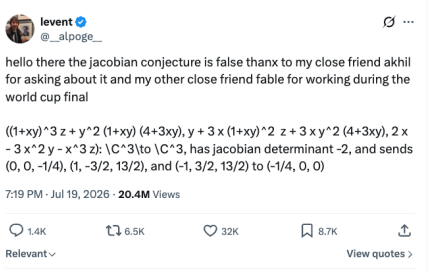
Melissa Dell Ashesh Rambachan

NBER Methods Lecture 2026

Slides posted on the NBER Methods Lecture website.

AI and Science

- Across many fields, AI and machine learning are accelerating scientific discovery.
- Algorithms are proposing and proving conjectures, predicting novel structures, and surfacing patterns that researchers did not see.



AI and Economics

- How might AI accelerate economics? There is no single answer.

- AI could accelerate existing workflows:
  - Empirical: automating empirical analysis and policy evaluation (e.g., Han, 2025; Korinek, 2025; Social Catalyst Lab, 2026).
  - Theory: formalizing arguments and suggesting new results and proofs, as in AI for mathematics (e.g., Chen et al., 2026; Garg, 2026; Bei et al., 2026; Tan and Syrgkanis, 2026).

AI and Economics

- How might AI accelerate economics? There is no single answer.
- AI could accelerate existing workflows.
- AI lets us formalize and accelerate a *new* type of problem: hypothesis generation (this lecture).
- This opportunity rests on three observations about how AI and machine learning interacts with economics.

AI and Economics

- In many structured domains we study, machine learning algorithms outpredict our best economic theories out of sample (Fudenberg et al., 2022; Liang, 2026).
  - Decision-making under risk & ambiguity (e.g., Peterson et al., 2021; Plonsky et al., 2025; Peysakhovich and Naecker, 2017; Enke and Shubatt, 2023; Mullainathan and Rambachan, 2025a).
  - Game theory (e.g., Wright and Leyton-Brown, 2010, 2017; Fudenberg and Liang, 2019; Zhu et al., 2024; Hirasawa et al., 2025).

- What signal have these machine learning algorithms discovered that researchers missed? How can we extract theoretical insight?

AI and Economics

- In many structured domains we study, machine learning algorithms outpredict our best economic theories out of sample. What signal have they discovered that researchers missed?

- But economics is not just interested in within-domain prediction and generation.

- We want to generalize to new settings, evaluate counterfactuals, and understand mechanisms:
  - Economic theories transfer to new domains and data-scarce settings where black-box algorithms may fail (Andrews et al., 2025; Ellis et al., 2026; Manning and Horton, 2026).
  - Interpretable insights are portable: they can be communicated, taught, and built upon (Ludwig and Mullainathan, 2024).

AI and Economics

- Machine learning algorithms outpredict our best economic theories out of sample.
- But economics is not just interested in within-domain prediction and generation.
- We face a pervasive lamppost problem: we search where the light is in our structured data.
  - Closed survey items, RCT outcomes, and the fields of an administrative extract we study capture only a small sliver of the domains we study.

AI and Economics

- Machine learning algorithms outpredict our best economic theories out of sample.
- But economics is not just interested in within-domain prediction and generation.
- And we face a pervasive lamppost problem: we search where the light is in our structured data.
- Across all subfields, unstructured data (text, images, audio, video) record features that we never thought to measure (e.g., Gentzkow et al., 2019; Dell, 2025).
  - Extremely incomplete list: policy uncertainty (Baker et al., 2016); urban life (Glaeser et al., 2018); political risk (Hassan et al., 2019); job tasks (Atalay et al., 2020); culture (Michalopoulos and Xue, 2021); race and gender (Adukia et al., 2023); open-ended surveys (Haaland et al., 2024); religiosity (Dube et al., 2026); firm perceptions (Sarkar, 2026); career transitions (Vafa et al., 2022; Athey et al., 2024); monetary policy (Hansen et al., 2018; Gorodnichenko et al., 2023); product characteristics (Compiani et al., 2025; Han et al., 2021), macro growth (Backer-Peral and Wittenbrink, 2026) ...

Discovery and Hypothesis Generation

- Three observations about how AI and machine learning interacts with economics:
  - Algorithms extract signal from data that researchers and our theories miss.
  - Economics seeks more than just within-domain prediction and generation.
  - Unstructured data record features of economic life we never thought to measure.

- Together, these observations suggest an exciting opportunity:
  1. Use ML and AI on unstructured data to discover what to measure and how it relates to the quantities we care about $\rightarrow$ "hypothesis generation."
  2. Analyze generated hypotheses using our empirical and theoretical toolkit.

Discovery and Hypothesis Generation

- Together, these observations suggest an exciting opportunity:
  1. Use ML and AI on unstructured data to discover what to measure and how it relates to the quantities we care about $\rightarrow$ "hypothesis generation."
  2. Analyze what we discover using our empirical and theoretical toolkit.

- Almost every empirical study begins with a "prescientific" step: deciding what to measure and which hypotheses to pursue.
  - There are many rigorous tools to test hypotheses with collected data.
  - But generating them is left to researcher intuition, inspiration, and creativity.

- We can now build ML and AI tools that assist or automate hypothesis generation from unstructured data (e.g., Ludwig and Mullainathan, 2024; Mullainathan and Rambachan, 2025b; Zhou et al., 2024; Movva et al., 2025).

Discovery and Hypothesis Generation

- Together, they suggest an exciting opportunity for economists:
  1. Use ML and AI on unstructured data to discover what to measure and how it relates to the quantities we care about — "hypothesis generation."
  2. Analyze what we discover using our conventional empirical and theoretical toolkit.

- We can now build ML and AI tools that assist or automate hypothesis generation from unstructured data (e.g., Ludwig and Mullainathan, 2024; Mullainathan and Rambachan, 2025b; Zhou et al., 2024; Movva et al., 2025).

- We will focus the discussion around three empirical examples.

Prediction Policy Problems

- Many high-stakes policy decisions hinge on a prediction (Kleinberg et al., 2015; Mullainathan, 2025).
  - Pretrial release: a judge jails or releases a defendant based on predicted flight / re-arrest risk.
  - Child protective services: an agency opens a case against a family based on predicted maltreatment risk to the child.

Prediction Policy Problems

- Many high-stakes policy decisions hinge on a prediction (Kleinberg et al., 2015; Mullainathan, 2025).

- In pretrial release, judges see more than any structured dataset (the defendant is in the room), yet ML algorithms outpredict judges (Kleinberg et al., 2018; Rambachan, 2024; Angelova et al., 2025).
  - Reranking releases by predicted risk could cut crime by 25% with no extra jailing — or cut the jail population by 42% with no extra crime.

- Judges appear to get these decisions wrong. But what exactly are they getting wrong?

Hypothesis Generation in Pretrial Release

- Ludwig and Mullainathan (2024) turn the camera on the judge and predict detention decisions in Mecklenburg County, North Carolina.

- The most important predictor is the mug shot, above and beyond:
  - structured information: current charge, prior record, demographics;
  - known facial features and incentivized human guesses of who gets detained.

- $\approx$78% of the predictive signal is unexplained by known facial features.

Hypothesis Generation in Pretrial Release

- Ludwig and Mullainathan (2024) turn the camera on the judge and predict detention decisions in Mecklenburg County, North Carolina.

- The most important predictor is the mug shot, above and beyond:
  - structured information: current charge, prior record, demographics;
  - known facial features and incentivized human guesses of who gets detained.

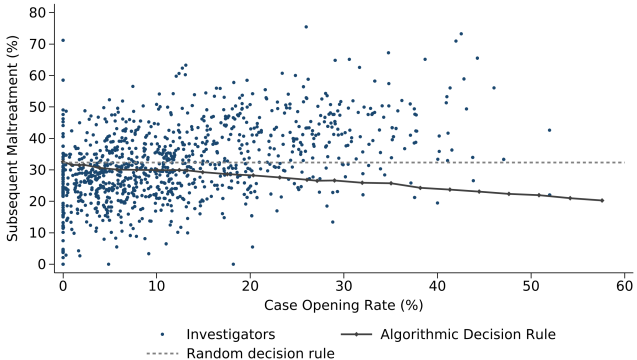
- Something unnamed in the mug shot systematically drives detention decisions. What is it?

Hypothesis Generation in Child Protective Services

- In child protective services (CPS), an investigator decides whether to open a case against a family based on predicted maltreatment risk to the child.

Hypothesis Generation in Child Protective Services

- Baron et al. (2026): in Michigan CPS, 36% of investigators outperform a machine learning algorithm trained on structured characteristics.



- The gap is not explained by how investigators use the structured information. Something else separates the best investigators. What is it?

Hypothesis Generation in Child Protective Services

- Investigators visit the home, interview the family, and contact teachers and doctors. Investigators record what they learn in their case notes.

Incriminating.

"Both Mr. Ramirez and Ms. Ramirez had scratches on them; Ms. Ramirez ended up having bruising all over her arms and her neck."

Exculpatory.

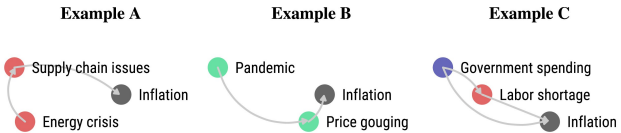
"Both children reported that they feel safe at home. The home was viewed to be appropriate. The water, and electricity observed to be in working order."

- Baron et al. (2026): can we generate hypotheses about what high- and low-performing investigators *do* and *observe* differently? Does this help explain heterogeneity across investigators?

Hypothesis Generation in Macroeconomic Expectations

- How people explain the macroeconomy shapes their expectations, and expectations drive their behavior.

- Andre et al. (2026) study the narratives people held during the 2021–22 inflation surge.
  - Narratives shape inflation expectations and how people interpret macroeconomic news.



- What narratives do people *actually* hold about the causes of inflation?

Hypothesis Generation in Macroeconomic Expectations

- To discover narratives, Andre et al. (2026) elicit open-ended survey responses.
  - Essential: an ex-ante menu presupposes the narratives one is trying to discover.
  - Painstaking: humans read every response and hand-code it into a DAG.

> *"Demand growing faster than supply, shortage of truck drivers, deliveries, Covid, Devaluation, rising wages, climate control affecting Farmers."*

> *"The federal reserve having to print more money. The dollar is weak trash and will only get weaker."*

- Collecting open-ended responses is increasingly popular (Haaland et al., 2024) and increasingly cheap with AI interviewers (Chopra and Haaland, 2026; Geiecke and Jaravel, 2026).

- Could we hand the responses to an AI and let it discover inflation narratives automatically?

Discovery and Hypothesis Generation

- Hypothesis generation is a *nascent* but *rapidly growing* space across fields:
  - Applications: pretrial release (Ludwig and Mullainathan, 2024); ECG biomarkers (Obermeyer et al., 2026); effective teaching (Workman, 2025); parental speech (Pernaudet et al., 2026); preference data (Movva et al., 2026); open-response surveys (Wang et al., 2026); cognitive experiments (Zhu et al., 2026).
  - Methods: describing how text corpora differ (Zhong et al., 2022, 2023); hypothesis generation with LLMs (Zhou et al., 2024; Batista and Ross, 2024; Manning et al., 2024; Tranchero et al., 2024) and sparse autoencoders (Movva et al., 2025; Peng et al., 2026); statistical inference (Modarressi et al., 2025; Carlson, 2026); morphing (Miller et al., 2019); anomaly generation (Mullainathan and Rambachan, 2025a); benchmarking (Liu et al., 2025).

- **Goal**: describe the structure of hypothesis generation procedures, the key choices required, and some challenges that arise.

Setup

- We observe data $x \in X$ and an outcome $y \in Y$: each unit x is a piece of unstructured data.
  - Ludwig and Mullainathan (2024): x is a mug shot and y is the detention decision.
  - Baron et al. (2026): x is the case note and y is the case-opening decision.
  - Andre et al. (2026): x is the open-ended response and y is the respondent's inflation expectations.

- When x is structured (e.g., choice menu, game), the same setup describes techniques to improve economic theory using machine learning (e.g., Fudenberg and Liang, 2019; Mullainathan and Rambachan, 2025a).

Hypotheses

- We observe unstructured data $x \in X$ and an outcome $y \in Y$.
- The researcher chooses a candidate space of hypotheses $\mathcal{H}$ to generate into.
  - Each hypothesis names a feature of the unstructured data: a function $h : X \to \mathbb{R}$ that measures the feature on each unit x.
  - Often an indicator $h(x) \in \{0, 1\}$: "does x exhibit this feature?"

Hypotheses

- The researcher chooses a candidate space of hypotheses $\mathcal{H}$: each names a measurable feature of the data.

- In unstructured data, we can describe $\mathcal{H}$ at a high level:
  - Ludwig and Mullainathan (2024): descriptions of facial features.
  - Baron et al. (2026): descriptions of narrative evidence themes.
  - Andre et al. (2026): descriptions of causal narratives.

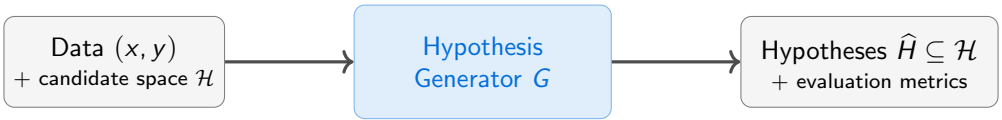
- But we cannot ex ante specify every hypothesis contained in $\mathcal{H}$—it is implicit.
  - That would mean enumerating everything we could ever discover from unstructured data.
  - A version of "Polanyi's Paradox": we recognize hypotheses when we see them, but we cannot list them in advance (Autor, 2014).

Hypothesis Generation Procedure

- We observe data $x \in X$ and an outcome $y \in Y$.
- The researcher chooses a candidate space of hypotheses $\mathcal{H}$: each hypothesis names a measurable feature of the data.
- A hypothesis generator G takes input data (x, y) and returns a collection of hypotheses $\widehat{H} \subseteq \mathcal{H}$, alongside evaluation metrics that assess their "quality."
  - Ludwig and Mullainathan (2024): (mug shots, detention decisions) $\rightarrow$ descriptions of facial features.
  - Baron et al. (2026): (case notes, case-opening decisions) $\rightarrow$ descriptions of narrative evidence themes.
  - Andre et al. (2026): (open responses, inflation expectations) $\rightarrow$ descriptions of causal narratives.

Hypothesis Generation Procedure

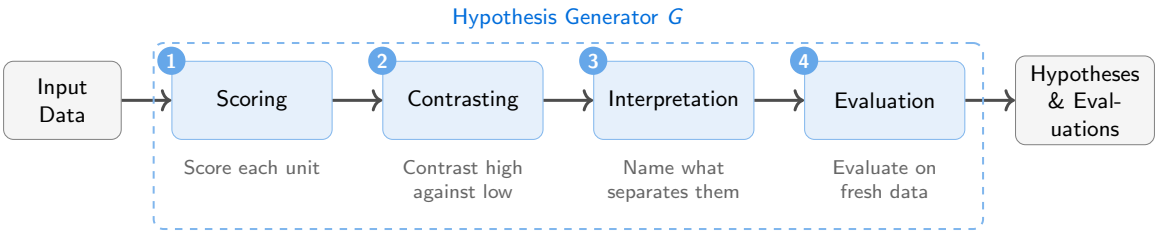
- A hypothesis generator G maps input data to a set of hypotheses and evaluations of their quality.



- For now, G is a black box: anything that maps data to named hypotheses qualifies.
  - It could be a researcher (status quo) or a frontier AI model (tomorrow?).

Hypothesis Generation Procedure

- Many procedures share a common structure. They typically organize G into four stages.

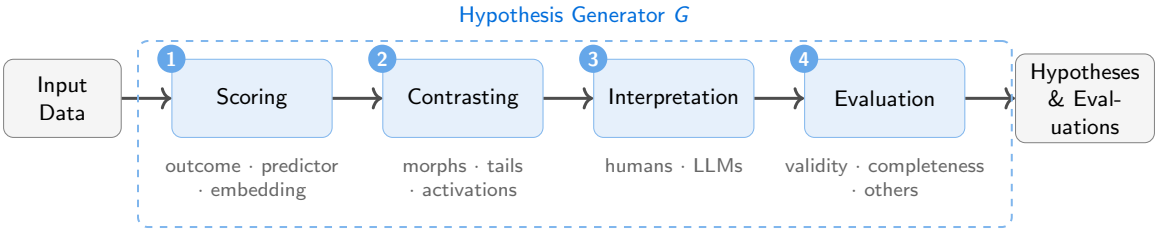


- (1) *score* each unit; (2) *contrast* high against low scores; (3) *interpret* what separates them; and (4) *evaluate* the result.

- We walk through each stage: how the three examples implement it, and the problems that arise.

Hypothesis Generation Procedure

- Many procedures share a common structure. They typically organize G into four stages.

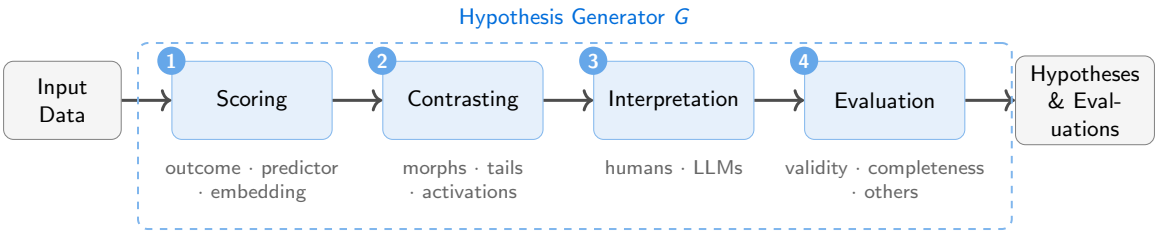


- Each stage admits many implementations, varying within and across data modalities (images, text, waveforms).

- To this point, there has been little systematic comparison of these choices. Which score, contrast, or interpreter generates "better" hypotheses?

Hypothesis Generation Procedure

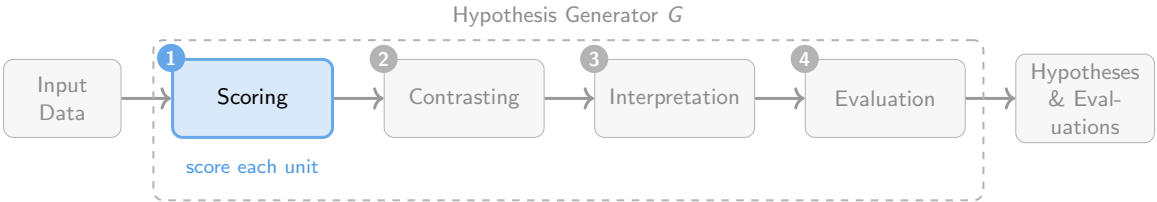
- Many procedures share a common structure. They typically organize G into four stages.



- Implementations can be judged the same way: do the generated hypotheses correlate with the outcome y on a held-out sample?
  - Within a domain, there's a verifiable "common task" (Donoho, 2024).

- One evaluation, but not the only one. What to demand of generated hypotheses is itself a design choice.

Stage 1: Scoring



Stage 1: Scoring

- Scoring constructs a score $r(x)$: a number or vector summarizing each unit of data x.
- The score turns unstructured data into something we can systematically compare: which differences in x move $r(x)$?

Stage 1: Scoring

- Scoring constructs a score $r(x)$: a number or vector summarizing each unit of data x. Three choices dominate existing work:

- The outcome itself, $r(x) = y$: score each case note by whether the case was opened (e.g., Zhong et al., 2023; Modarressi et al., 2025).
  - Simplest choice. The data are already split into high ($y = 1$) and low ($y = 0$).

Stage 1: Scoring

- Scoring constructs a score $r(x)$: a number or vector summarizing each unit of data x. Three choices dominate existing work:

- The outcome itself, $r(x) = y$: score each case note by whether the case was opened (e.g., Zhong et al., 2023; Modarressi et al., 2025).

- A learned predictor, $r(x) = \widehat{\mathbb{E}}[y \mid x]$: score each mug shot by its predicted detention risk (e.g., Ludwig and Mullainathan, 2024; Baron et al., 2026; Obermeyer et al., 2026).
  - Keeps only the variation in y explained by x.
  - We can counterfactually score units we never observed.

Stage 1: Scoring

- Scoring constructs a score $r(x)$: a number or vector summarizing each unit of data x. Three choices dominate existing work:

- The outcome itself, $r(x) = y$: score each case note by whether the case was opened (e.g., Zhong et al., 2023; Modarressi et al., 2025).

- A learned predictor, $r(x) = \widehat{\mathbb{E}}[y \mid x]$: score each mug shot by its predicted detention risk (e.g., Ludwig and Mullainathan, 2024; Baron et al., 2026; Obermeyer et al., 2026).

- An embedding, $r(x) \in \mathbb{R}^d$: place each response in a vector space where nearby points express similar things. (e.g., Mikolov et al., 2013; Pennington et al., 2014) ▸ primer
  - Not tied to any single outcome; a general-purpose description of x.
  - Familiar ancestor: *topic models* group documents by word co-occurrence (LDA) or by embedding similarity (BERTopic), and how strongly a document belongs to a group could be a score (Blei et al., 2003; Grootendorst, 2022).

Stage 1: Which Score?

- The outcome y and the predictor $\widehat{\mathbb{E}}[y \mid x]$ both score x by its relation to y; the predictor additionally attempts to smooth this relationship.

- Two reasons the smoothed score is usually preferred:
  1. y is noisy; $\widehat{\mathbb{E}}[y \mid x]$ attempts to preserve only the variation in y that x can explain.
  2. A predictor scores units that never appeared in the data, which later stages can exploit (e.g., scoring counterfactual images or waveforms).

Stage 1: Which Score?

- An embedding is different. Each unit becomes a point in a vector space; units that express similar content sit near each other.
  - Baron et al. (2026): Case notes about domestic violence cluster together, far from notes about housing conditions.
  - Each direction in the space is a potential score: $r(x)$ measures how strongly unit x aligns with that direction.
  - Still need to pick a direction and decide how to interpret it (later).

Stage 1: Which Score?

- An embedding is different. Each unit becomes a point in a vector space; units with similar content sit near each other.

- A given direction often behaves like an index. It typically bundles several concepts at once.
  - If we contrast units along a direction, then the differences mix many candidate hypotheses and are therefore hard to differentiate.
  - Jargon: with more concepts than dimensions, they are stored in superposition" — so individual dimensions, and typical directions, are polysemantic" (Olah et al., 2020; Elhage et al., 2022; Cunningham et al., 2024). ▸ more

- Sparse autoencoders (SAEs) attempt to "unbundle" the index, re-expressing the embedding as sparse dimensions that each track a single concept (Cunningham et al., 2024; Movva et al., 2025) ▸ more.

Stage 1: Which Score?

- Which score? The choice turns on at least three considerations:
  1. *Where the outcome enters*: built into the score ($\widehat{\mathbb{E}}[y \mid x]$); used only to select from a general-purpose space (embeddings/SAEs) (Movva et al., 2025); or not at all.
  2. *What the data allow*: predictors need enough (x, y) pairs to meaningfully train; pretrained embeddings are cheap but encode only what the underlying model learned about your domain.
  3. *What comes next*: a scalar score supports ranking and morphing; an embedding supports many contrasts at once. The score is chosen jointly with the later stages.

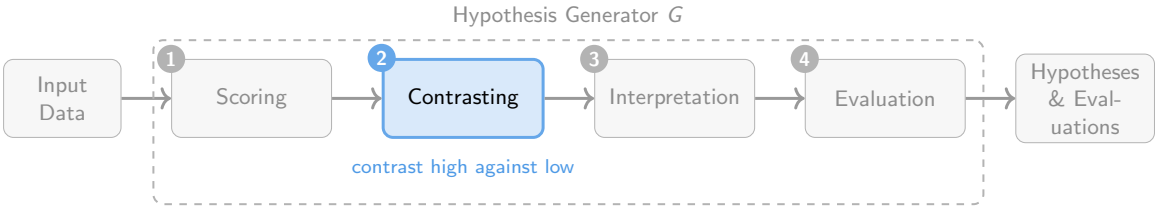
- Our examples illustrate different possible choices.

Stage 1: Which Score?

- Predictor versus embedding scores consume labeled data very differently.
  - A predictor is supervised from the start. Every step of training consumes (x, y) pairs.
  - Embeddings come pretrained, and features are learned from *unlabeled* x alone.

- So the routes may better suit different data regimes:
  - *Label-rich*: ample (x, y) pairs; train the predictor directly.
  - *Label-poor, data-rich*: millions of unlabeled x (e.g., many mug shots but few linked to detention decisions, many case notes but few linked to case openings).

Stage 2: Contrasting



Stage 2: Contrasting

- Contrasting uses the score to construct a set of examples $E \subseteq X$: units that *differ* along $r(x)$.

- Working presumption: hypotheses can be labeled by inspecting differences. If we can see what separates high scorers from low scorers, we can name it.

- The example sets take two broad forms:
  - *Morphed / counterfactual pairs*: two versions of the same unit, constructed to differ only in the score (i.e., a mug shot morphed to have higher detention risk).
  - *High- versus low-scoring sets*: real units drawn from the two tails of the score (i.e., case notes with high versus low predicted case opening rates).

Contrasting: Building the Example Sets

- Three common contrast strategies, each returning an example set E:
  1. Morphs: generate a counterfactual pair $E = \{x, x'\}$ by moving along the gradient ∇r inside a generative model of the data (Ludwig and Mullainathan, 2024; Mullainathan and Rambachan, 2025a; Obermeyer et al., 2026).
  2. Tails: pick real units from the two tails of the score, $E = \{X^-, X^+\}$, potentially holding other known structure fixed (e.g., Batista and Ross, 2024; Baron et al., 2026).
  3. Activations: pick real units on which a selected *SAE feature*'s activation is high versus low, $E = \{X^-, X^+\}$ (Movva et al., 2025; Workman, 2025; Wang et al., 2026).

Contrasting: Building the Example Sets

- Three common contrast strategies, each returning an example set E:
  1. Morphs: generate a counterfactual pair $E = \{x, x'\}$.
  2. Tails: pick real units from the two tails of the score, $E = \{X^-, X^+\}$.
  3. Activations: real units where a selected SAE feature's activation is high versus low, $E = \{X^-, X^+\}$.

- Each strategy leans on a different choice of score:
  - Morphs score units that do *not* exist in the data $\rightarrow$ a predictor (plus a generative model).
  - Tails rank observed units $\rightarrow$ any scalar score works, even the raw outcome.
  - Activations require SAE features $\rightarrow$ embedding + SAE route.

- Let's return to our three examples to see these strategies in action.

Contrasting: Morphed Pairs of Mug Shots

- Ludwig and Mullainathan (2024): the score is a convolutional neural network that predicts detention from the mug shot x alone:

$$r(x) = \widehat{\mathbb{E}}[\, y \mid x \,].$$

- Morphing: pairs of mug shots that are close to one another but differ on the score, producing the example set $E = \{x, x'\}$.

- Why? Two faces sampled at low and high $r(x)$ differ along many dimensions at once (age, hair, lighting, ...). We may struggle to interpret what drives differences in the score.

Contrasting: Morphed Pairs of Mug Shots

- Ludwig and Mullainathan (2024): the score is a convolutional neural network that predicts detention from the mug shot x alone:

$$r(x) = \widehat{\mathbb{E}}[\, y \mid x \,].$$

- The goal: pairs of mug shots that are close to one another but differ on the score, producing the example set $E = \{x, x'\}$.

- The idea: start from a mug shot and follow the score's gradient,

$$x' \leftarrow x + \eta \, \nabla r(x),$$

direction in pixel space that most increases predicted detention (Ludwig and Mullainathan, 2024).

Staying on the Manifold

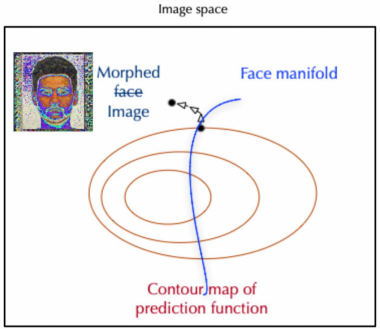
- Implement it literally, and the update $x + \eta \nabla r(x)$ produces something that is not a face.

original x

naive morph x'

Staying on the Manifold

- Implement it literally, and the update $x + \eta \nabla r(x)$ produces something that is not a face.
- In the space of all possible images, real faces occupy a low-dimensional manifold. Almost every direction steps off it.

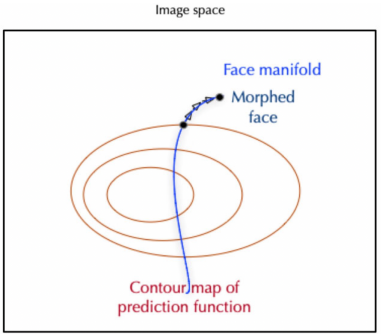


Staying on the Manifold

- Let's build a generative model of the mug-shot distribution and follow the gradient in its latent space, so every step decodes to a realistic face (Miller et al., 2019). ▸ GAN

Update in latent space $z' \leftarrow z + \eta \nabla r(g(z))$.

Then decode.



Sanity Check: Morphing Along a Known Feature

- Sanity check: apply the morphing algorithm to a score we already understand. Let's train $r(x)$ to predict age.
- Morphing along the age gradient recovers aging.

original x

morphed x' along the age predictor

Morphed Pairs

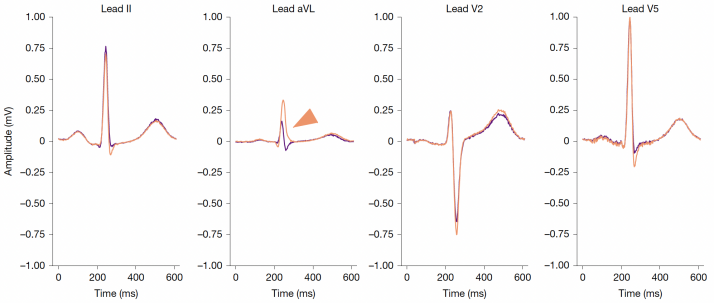
- Apply the morphing algorithm on the detention predictor: morph each face along $\nabla r(x)$ from low to high predicted detention, staying on the manifold.

- The output is many pairs (x, x'): faces differing along the score. This is the example set E.

Morphing Beyond Faces

- Nothing here is specific to faces and detention. Wherever we have a predictor and a good generative model, we can morph.

- Obermeyer et al. (2026) predict sudden cardiac death from ECGs, build a generative model of waveforms, and apply the same technique (Miller et al., 2019). ▸ VAE



Morphing Beyond Faces

- Nothing here is specific to faces and detention. Wherever we have a predictor and a good generative model, we can morph.
- Obermeyer et al. (2026) predict sudden cardiac death from ECGs, build a generative model of waveforms, and apply the same technique (Miller et al., 2019). ▸ VAE
- Images and waveforms have good generative models with latent spaces we can manipulate and then sample from.
- What about text?

Child Protective Services: Can We Morph Text?

- Baron et al. (2026) score each case note by predicted case opening, $r(x) = \widehat{\mathbb{E}}[y \mid x]$, built on pre-trained embeddings of the notes.

- Let's reuse the morphing recipe, with a language model as the generative model of text,

$$\text{Update in latent space } z' \leftarrow z + \eta \nabla r(g(z)).$$

Then decode.

Child Protective Services: Can We Morph Text?

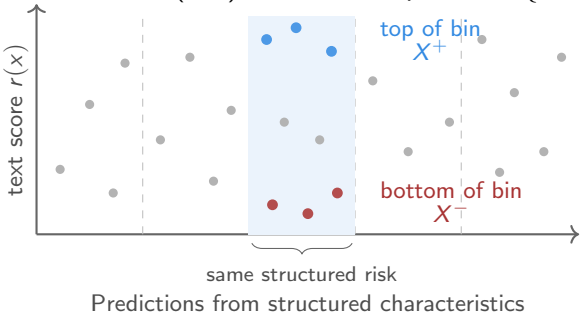
- Let's reuse the morphing recipe, with a language model as the generative model of text.
- But steering text along a trained predictor is challenging. Some obstacles:
  - General embeddings are polysemantic: moving along a direction shifts many features at once, not one (Elhage et al., 2022; Cunningham et al., 2024).
  - Sparse autoencoders isolate concepts, yet they *steer* poorly. Modifying a feature's activation does not cleanly change the text (Peng et al., 2026).
  - You could prompt: "rewrite this note." But that requires an English description, the very hypothesis we are trying to discover.

Child Protective Services: Can We Morph Text?

- But steering text along a trained predictor is challenging.
- This is an interesting problem. And the ingredients exist:
  - latent-space editing of known attributes (style, sentiment) (Wang et al., 2019); gradient-guided generation (Dathathri et al., 2020); steering SAE features (Bhattacharyya and Rooshenas, 2025); counterfactual text for causal estimation (Feldman et al., 2026).
- But we are not aware of a clear example of steering a language model along a trained outcome predictor to produce morphed text for hypothesis generation.
- So instead of generating counterfactual notes, Baron et al. (2026) select real ones.

Tails: Hold Structured Risk Fixed, Vary the Text

- Baron et al. (2026) find case notes that are similar on structured characteristics but differ on the score $r(x)$.
- Bin cases by structured risk predictions from the administrative data.
- Within a bin, a gap in $r(x)$ reflects the *text*, not the structured characteristics: the top notes (X^+) versus bottom notes (X^-) are the examples $E = \{X^-, X^+\}$.



Re-Analyzing Inflation Narratives

- Andre et al. (2026) collect open-ended explanations of why households expect the inflation they do; the outcome y is the inflation expectation.

- Let's convert each response into an embedding (e.g., using OpenAI's embedding models). But which direction should we choose? Recall that each direction may bundle many concepts.

- A sparse autoencoder (SAE) re-expresses the embedding as many features, each *intended* to track a single concept. ▸ more
  - HypotheSAEs: increasingly popular tool for hypothesis generation in text (Movva et al., 2025, 2026; Peng et al., 2026; Wang et al., 2026).

- The result is many candidate scores $r(x)$, one per SAE feature: "how strongly does feature j *activate* on response x?"

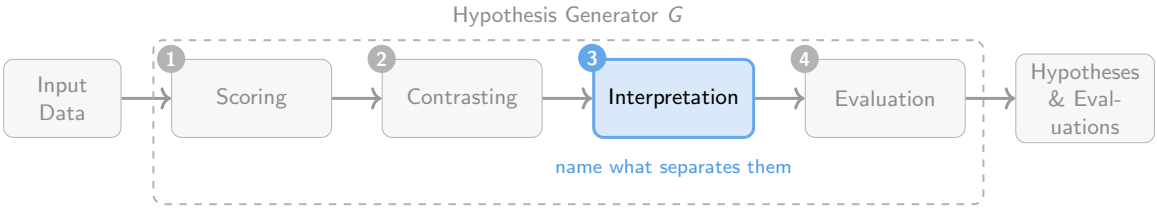
Re-Analyzing Inflation Narratives

- Which candidate scores deserve attention? Keep the ones that predict the outcome: the $\sim$20 features whose activation correlates with inflation expectations on a validation sample (Movva et al., 2025). ▸ details

- Notice that only now does the outcome enter.
  - An attractive option when unlabeled text is abundant but labeled pairs (x, y) are scarce.
  - The SAE builds candidate scores from unlabeled text, and the labels are used only for the final selection.

- Contrasting: for each kept feature j, the responses with high versus low *activation* form the example set $E = \{X^-, X^+\}$.

Stage 3: Interpretation



Stage 3: Interpretation

- Contrasting hands us examples E that differ on the score. How do we go from examples to hypotheses?
- How do people generate hypotheses? By searching for differences. Hypothesis generation procedures do the same.
- An interpreter ι examines the examples and returns hypotheses:

$$\widehat{H} = \iota(E).$$

- This is the stage most changed by AI and large language models. We have a meaningful choice over *who* the interpreter is.

Interpretation: Who Interprets?

- Existing work uses two interpreters:
  1. *Humans*: subjects inspect the examples and name the difference they see (Ludwig and Mullainathan, 2024; Obermeyer et al., 2026).
  2. *Large language models*: an LLM reads the examples and proposes candidate descriptions (Zhong et al., 2023; Baron et al., 2026; Movva et al., 2025).
- Both receive the same input (examples E) and return the same output: a description of what separates the high scorers from the low scorers.
- With multi-modal models, the same choice is in principle available in every modality: images, video, audio.

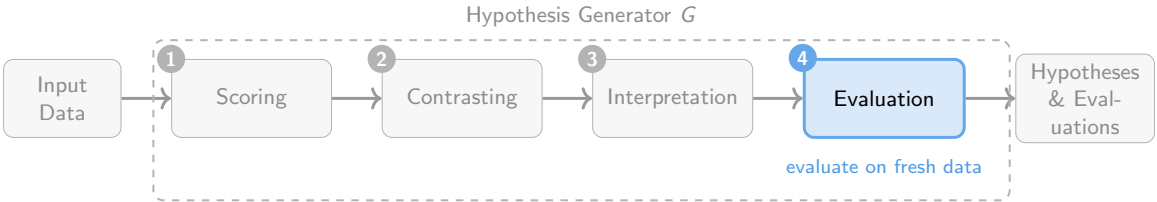
Interpretation: Which Interpreter?

- We are choosing from a class of interpreters $\mathcal{I}$, where each $\iota \in \mathcal{I}$ maps examples to hypotheses, $\iota : E \mapsto \widehat{H} \subseteq \mathcal{H}$.

- Different interpreters need *not* return the same hypothesis from the same examples. Humans and LLMs carry different "inductive biases" in what they notice from examples.

- Ample evidence says human and LLM intelligence differ:
  - LLMs can be jagged across tasks (Dell'Acqua et al., 2023), and humans mispredict where models succeed or fail (Vafa et al., 2024).
  - LLMs exhibit their own systematic reasoning patterns (McCoy et al., 2023; Dasgupta et al., 2022).

Stage 4: Evaluation



Stage 4: Evaluation

- The procedure returns named hypotheses $\widehat{H}$. Evaluation asks: how predictive of the outcome are they?

- So far evaluation is a standard supervised-learning problem: evaluate on a test set (Mullainathan and Spiess, 2017).
  - Generate hypotheses using only the training data; evaluate on a test set.
  - On the test set, each hypothesis is just a covariate $h(x)$.

- If hypothesis generation never touches the test set (firewall), it does *not* matter how the hypotheses were produced.
  - When LLMs are involved though, concerns surrounding "training leakage" apply (Sarkar and Vafa, 2024; Ludwig et al., 2025).
  - This is a verifiable "common task" for every implementation (Donoho, 2024).

Evaluation: Validity of Each Hypothesis

- What should we report on the test set?

- First the validity of each hypothesis: on its own, how correlated is $h(x)$ with the outcome y?

$$\mathbb{E}[y \mid h(x) = 1] - \mathbb{E}[y \mid h(x) = 0].$$

  This is a standard testing problem on the test set (Ludwig and Mullainathan, 2024; Baron et al., 2026; Movva et al., 2025).

  - *Caveat*: we need some way to measure $h(x)$ on the test set (next lecture).

- Researchers often additionally calculate, for example, the *incremental* R^2 or AUC contribution of $h(x)$ relative to structured covariates (Modarressi et al., 2025).

Evaluation: Many Hypotheses

- Hypothesis generation procedures propose *many* candidates. Corrections for multiple testing are essential in evaluation.
- How much can this matter? In one pipeline, 3,296 generated candidate hypotheses yield only 13 that survive correction (Zhong et al., 2023).
- Multiple testing corrections should be applied on the test set.

Evaluation: Many Hypotheses

- And the test set is not the endpoint: we often select the best performers for further downstream analysis and future work.

- Compute a test statistic $\widehat{V}(h_j)$ for each $h_j \in \widehat{H}$ on the test set summarizing how related are y and $h(x)$, and suppose

$$\big(\widehat{V}(h_1), \ldots, \widehat{V}(h_J)\big) \sim \mathcal{N}(V, \Sigma),$$
$$\text{selection rule: } \hat{\jmath} = \arg\max_j \widehat{V}(h_j) \text{ (or the top } k\text{)}.$$

- This is an inference-on-winners problem: conditional on being selected, the winner's test-set performance is a biased estimate of its true validity. Selective-inference corrections apply (Andrews et al., 2024).

Evaluation: Completeness of the Hypotheses

- Procedures return many hypotheses. So it is valuable to also ask: how much of the signal in the unstructured data do they capture?

- Build a benchmark B: a flexible predictor of y from the data itself.
  Completeness compares the hypotheses' predictive performance against it:

$$C(\widehat{H}) = \frac{\text{perf}(\widehat{H}) - \text{perf}(\text{naive})}{\text{perf}(B) - \text{perf}(\text{naive})}$$

(Fudenberg et al., 2022; Ludwig and Mullainathan, 2024; Modarressi et al., 2025; Baron et al., 2026)

- Originally proposed as a metric to evaluate the predictive performance of economic theory relative to machine learning algorithms (see Fudenberg et al., 2022).

Evaluation: Judging the Procedure

- So far we evaluated one realized output: given $\widehat{H}$, compute validity and completeness.
- But a procedure G maps data to hypotheses and metrics. Imagine resampling the data: how would V and C vary across draws?
- There are two properties of the *procedure* we can in principle define:
  - Size-like property: if x is independent of y, how often does G still return hypotheses on the test set? $\Pr\left(\max_{h \in \widehat{H}} |\widehat{V}(h)| / \widehat{se} \geq c\right)$ under $x \perp\!\!\!\perp y$.
  - Completeness property: across draws of the data, how much of the unstructured signal does $\widehat{H}$ summarize? $\overline{C}(G) = \mathbb{E}\left[C(\widehat{H})\right]$.

Evaluation: Judging the Procedure

- So far we evaluated one realized output: given $\widehat{H}$, compute validity and completeness.
- But a procedure G maps data to hypotheses and metrics. Imagine resampling the data: how would V and C vary across draws?
- There are two properties of the *procedure* we can in principle define:
  - Size-like property.
  - Completeness property.
- Test-set evaluation is important for controlling size: under $x \perp\!\!\!\perp y$, every rejection is a false discovery, so setting c appropriately controls it using standard arguments.
- But splitting the data leaves less signal for generating the hypotheses in the first place. Are we therefore losing out on completeness?

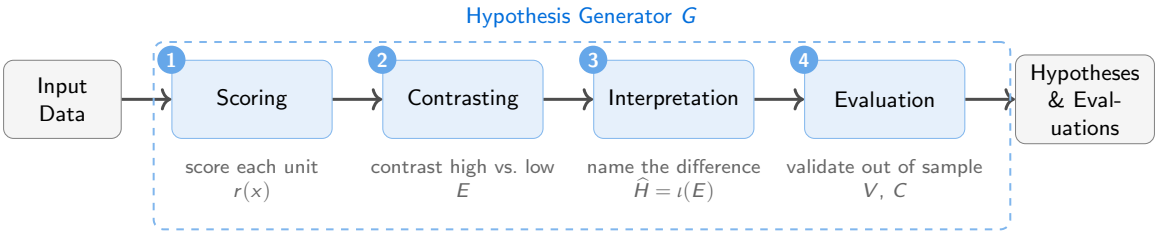
Evaluation: Judging the Procedure

- There are two properties of the procedure: a size-like property and a completeness property, defined over redraws of the data.

- This raises open questions:
  1. Are these the only desiderata? What about novelty relative to known features or existing theory? (Mullainathan and Rambachan, 2025a)
  2. Are there fundamental tradeoffs between them? In hypothesis testing, there is a size-power tradeoff. Are there analogues for hypothesis generators?

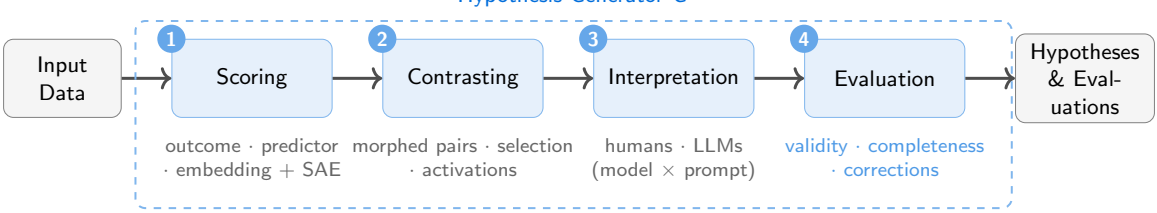
- Which circles back to where we started: as implementations proliferate, more systematic evaluation of alternative procedures is needed (both theoretical and empirical).

Hypothesis Generation Procedures



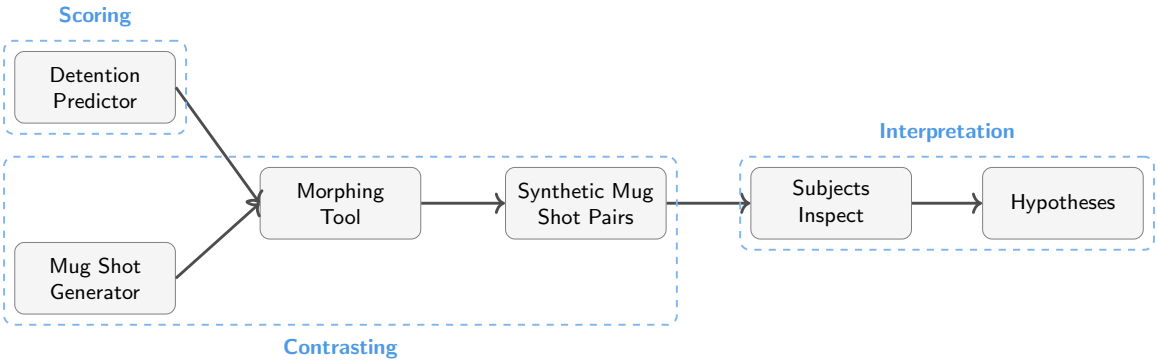
- Four stages: *score* each unit, *contrast* high against low, *interpret* what separates them, *evaluate* the result.

Hypothesis Generation Procedures



- Many defensible choices at every stage, and the choices are interconnected (e.g., morphing requires a predictor).
- So what did these procedures actually produce in our three examples?

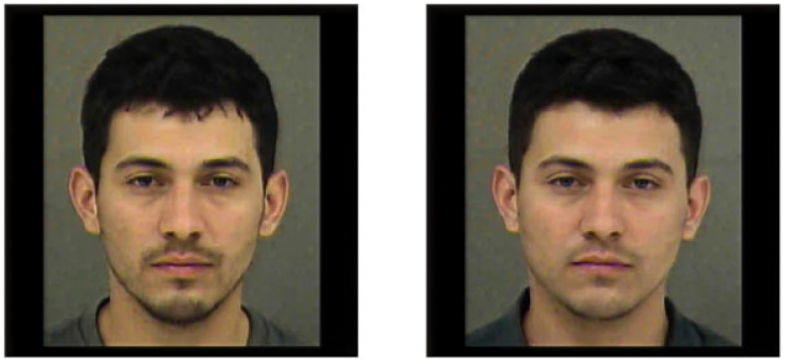
Ludwig & Mullainathan (2024): Pretrial Release



- The detention predictor is the score; the generator, morphing tool, and synthetic pairs build the contrasts; human subjects who inspect and name them are the interpreter.

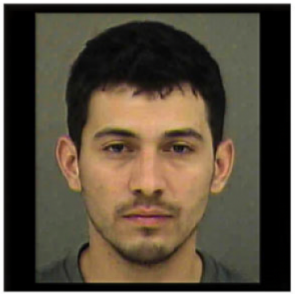
What Hypotheses Were Generated?

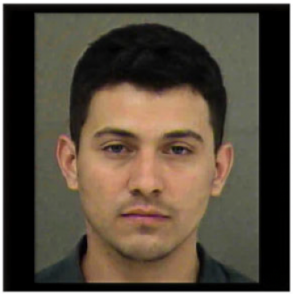
- A morphed pair: the same face, moved from low to high predicted detention risk. What changed?



What Hypotheses Were Generated?

- A morphed pair: the same face, moved from low to high predicted detention risk. Subjects name what they see.





morph: low → high predicted detention

what subjects named

- The named hypothesis: well-groomed defendants are less likely to be detained.

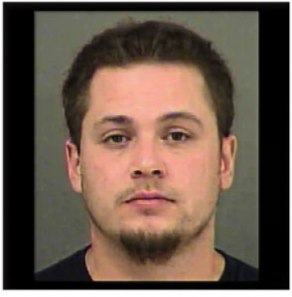
Validating the Hypothesis Out of Sample

- Is well-groomed real? Evaluate out of sample: fresh raters measure $h(x)$ on held-out photos, and then examine how it correlates with detention decisions (validity).

- Defendants in the top versus bottom quartile of well-groomedness differ in detention rates by 5.5 percentage points: 24% of the base rate, and larger than the gap between violent and nonviolent arrests (4.8pp).

Generating Another Hypothesis

- Morph again, now in a direction orthogonal to well-groomedness, searching for the residual signal.



morph: low → high predicted detention, orthogonal to well-groomed

what subjects named

- The second named hypothesis: heavy-faced defendants are less likely to be detained.

Generating Another Hypothesis

- The second hypothesis validates out of sample too: defendants in the top versus bottom quartile of heavy-facedness differ in detention rates by 7 percentage points (about 30% of the base rate).

- Ludwig and Mullainathan (2024) compute a completeness analogue: well-groomed, heavy-faced, and *all previously known* facial features together explain 27% of the variation in the mug-shot predictor.
  - $C(\widehat{H})$ with the predictor as the benchmark B: most of what the algorithm sees remains unnamed.

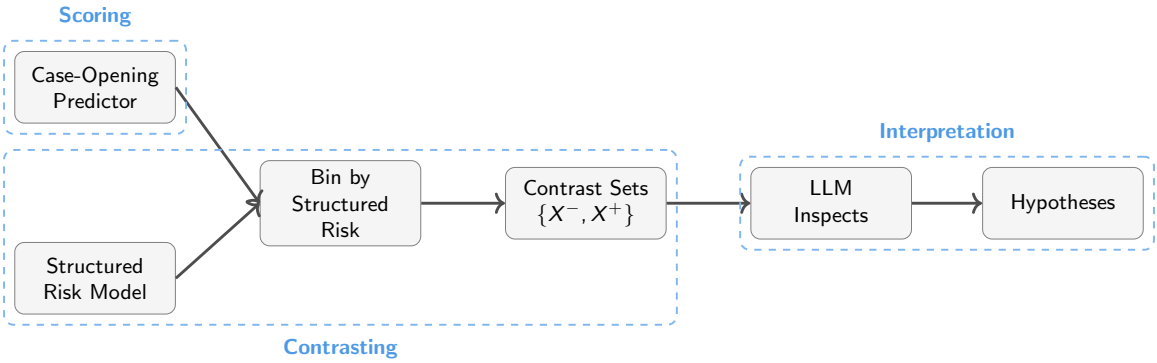
Ludwig & Mullainathan (2024): Takeaways

well-groomed

heavy-faced

- Ludwig and Mullainathan (2024) probe novelty: hypotheses are new to the literature *and* to practitioners.
- Unstructured data plus hypothesis generation uncovered a potential mechanism for judge error in pretrial release.
- The method is general: the same pipeline discovered a novel ECG marker of sudden cardiac death (Obermeyer et al., 2026).

Baron et al. (2026): Child Protective Services



- The case-opening predictor is the score; binning by structured risk and selecting within bins build the contrasts; the LLM that reads and names them is the interpreter.

What Hypotheses Were Generated?

- The LLM interpreter returns five evidence themes that appear consistently across the case notes:

Corroborated harm or injury. *"... bruising all over her arms and her neck."*

Absence, ambiguity, or contradiction of abuse evidence. *"Jordan denied that this hurt and does not have any visible injuries as a result."*

Exculpatory corroboration of adequate care. *"Both children reported that they feel safe at home. The home was viewed to be appropriate."*

Caregiver absence, unstable placement, or abandonment. *"Since the incident Rebecca has moved into her own apartment and is no longer living with the family."*

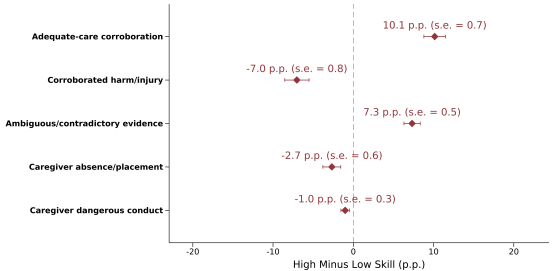
Caregiver involves child in dangerous or exploitative conduct. *"Maria admits to allowing Jordan to try her wine and putting the video on Snap Chat."*

Evaluating the Themes: Completeness

- The LLM interpreter returns five evidence themes that appear consistently across the case notes.
- On a held-out sample, Baron et al. (2026) report a completeness metric, with the embedding predictor as benchmark B and chance (AUC $= 0.5$) as the naive baseline.
- Five named, human-readable themes capture $\sim$84% of the predictive signal in the notes.

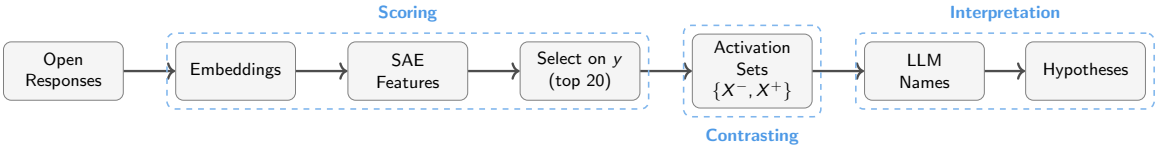
What Separates the Best Investigators?

- Higher-performing investigators are far more likely to record exculpatory evidence: +10.1pp adequate-care corroboration, +7.3pp ambiguous or contradictory evidence.



- The best investigators appear to report disconfirming evidence more often.
  - Classic corrective for confirmation bias (Nickerson, 1998; Lord et al., 1984) and used in interventions to promote system-2 thinking (Heller et al., 2017; Dube et al., 2025).

Reanalyzing Inflation Expectations



- The selected SAE features are the scores; the responses with high versus zero activation are the contrasts; the LLM that names them is the interpreter.

- Output: 20 named narrative concepts expressed in households' open responses.

Evaluating the Concepts Out of Sample

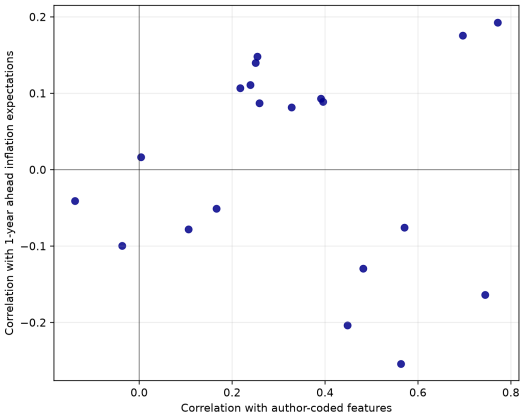
- On a held-out test set, predict inflation expectations y:

Model	Features	Test R^2
Structured covariates X	10	0.108
X + 20 SAE features	30	0.167
X + 20 named concepts	30	0.147
X + authors' hand-coded features	489	0.184

- The concepts add substantial signal over the structured covariates.

- Against the authors' 479 hand-coded features, the 20 generated hypotheses recover 51% of the hand-coding's predictive signal.
  - SAE activations recover 77% of the hand-coding's predictive signal.
  - LLM-generated hypotheses lose some fidelity relative to SAE features (Movva et al., 2025).

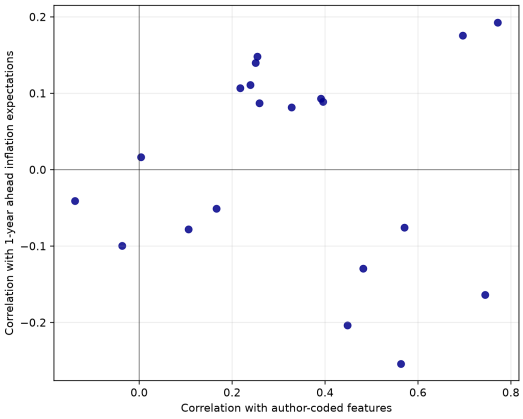
What Are the Concepts?



each dot: one of the 20 named concepts

- Each concept, two correlations: with inflation expectations (vertical), and with its closest hand-coded feature (horizontal).
- The right side recovers the hand-coding: hypotheses aligning with the authors' features, such as monetary policy, energy crisis, government spending, and supply chains.

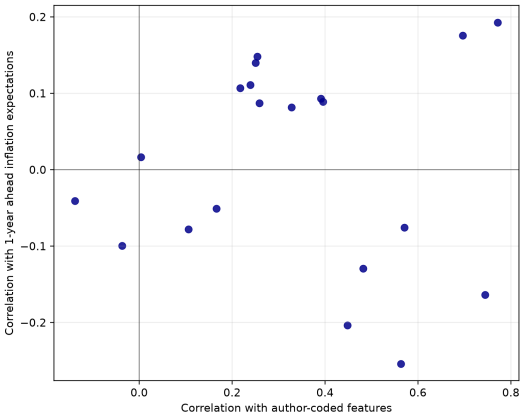
What Are the Concepts?



each dot: one of the 20 named concepts

- The right side recovers the authors' hand-coded categories.
- Moving left: concepts that predict expectations yet do not align with hand-coded features.
  - explicit blame of Biden, particularly by nickname.
  - pipeline shutdowns, reduced U.S. oil production.
  - government "giveaways."

What Are the Concepts?



each dot: one of the 20 named concepts

- The right side recovers the authors' hand-coded categories.
- Moving left are concepts that predict expectations yet do not fully align with hand-coded features.
- All of this runs on 2,951 open-ended responses, with 589 held out (i.e., an ordinary survey experiment).

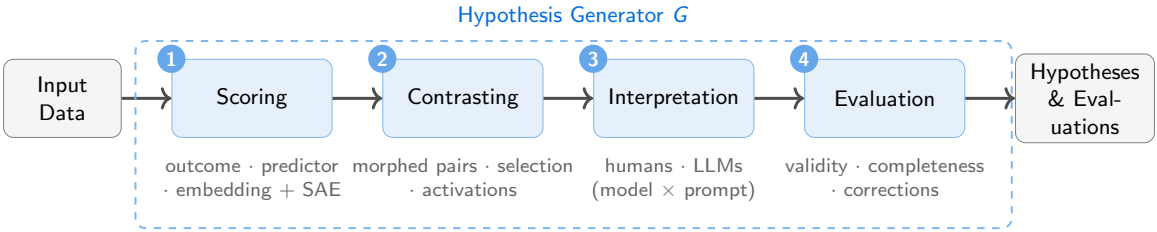
Hypothesis Generation

- We now have procedures that turn unstructured data into hypotheses (e.g., Ludwig and Mullainathan, 2024; Mullainathan and Rambachan, 2025b; Zhou et al., 2024; Movva et al., 2025).
  - Hypotheses: named, measurable features of the unstructured data.
  - Well-groomed/heavy-faced, exculpatory corroboration, inflation narratives.
  - Each carrying the claim that it predicts an outcome we care about (detention, case-opening, inflation expectations).

Hypothesis Generation

- We now have procedures that turn unstructured data into hypotheses (e.g., Ludwig and Mullainathan, 2024; Mullainathan and Rambachan, 2025b; Zhou et al., 2024; Movva et al., 2025).
- What happens next? Generated hypotheses can be explored with existing empirical tools.
- Hypothesis generation supplies fodder for downstream empirical work: which mechanisms to test, which interventions to trial, which causal effects we have overlooked.
  - In child protective services: would training lower-performing investigators to search for disconfirming evidence improve decisions? (Heller et al., 2017; Dube et al., 2025)
  - In inflation expectations, causal narratives about inflation are incorporated into a New-Keynesian model to quantify their potential effects on aggregate outcomes (Andre et al., 2026).

Hypothesis Generation Recipe



- The barrier to your entry is low:
  - (i) Simplest version: hand an LLM contrasting documents, name what differs, and validate out of sample.
  - (ii) Elaborate versions are largely off-the-shelf: pretrained embeddings, standard architectures for generative models, etc.

- The scarce input is your comparative advantage: economic setting with unstructured data and an outcome.

Methods Questions

- There are many open methodological problems:
  1. *Steering*: move unstructured data along a score while staying on the manifold.
  2. *Inference*: everything here was split-sample. Can we re-use data? Cross-validate? But the hypotheses themselves change across folds.
  3. *Evaluation*: which metrics matter (validity, completeness, novelty)? Are there tradeoffs between them?
  4. *Interpreters*: how should we understand the choice of interpreter and the space of models and prompts behind it?

- The frontier is moving rapidly: what was impractical a year ago may work now, so these tradeoffs are time-dependent $\rightarrow$ requires interdisciplinary engagement with CS/AI.

From Hypotheses to Measurement

- To analyze a generated hypothesis, we must measure it: turn the named concept into a variable across the full data and use those measurements reliably for downstream estimates.
- Whether the hypothesis came from a human or a machine, can we use AI to measure it at scale?
- That is where the next lectures pick up: estimation and inference with AI-generated data (e.g., Angelopoulos et al., 2023; Egami et al., 2024; Carlson and Dell, 2025; Ludwig et al., 2025).

Appendix Slides

What Is an Embedding?

back

- An embedding is a map $\phi : X \to \mathbb{R}^d$ that places units of data with similar content near each other (e.g., by cosine similarity).
- For text, geometry in the embedding space often encodes meaning:

$$\phi(\text{king}) - \phi(\text{man}) + \phi(\text{woman}) \approx \phi(\text{queen}).$$

(Mikolov et al., 2013; Pennington et al., 2014)

- Modern LLM embeddings extend this from words to whole sentences and documents.
- For scoring, any direction in the space is a potential score: project $\phi(x)$ onto it to measure how strongly unit x aligns with whatever that direction encodes.

Superposition & Polysemanticity

back

- Embeddings compress unstructured data, packing far more concepts than they have dimensions. To do so, they store concepts in superposition: concepts share dimensions (Elhage et al., 2022).
- The symptom: individual dimensions are *polysemantic*. One dimension corresponds to several unrelated concepts at once, so no single axis is one clean theme (Olah et al., 2020).
- This is the sense in which a raw direction behaves like an index: contrast units along it, and the differences you see mix several candidate hypotheses.
- This is a documented property of dense representations, and the motivation for sparse methods that aim to recover *monosemantic* features (Cunningham et al., 2024).

Sparse Autoencoders (SAEs)

back

- A sparse autoencoder (SAE) re-expresses a dense embedding $e = \phi(x) \in \mathbb{R}^d$ as a sparse, higher-dimensional code $z(x) \in \mathbb{R}^M$ ($M \gg d$), then reconstructs it:

$$z(x) = \mathrm{ReLU}(W_{\mathrm{enc}}\, e + b), \qquad \hat{e} = W_{\mathrm{dec}}\, z(x),$$

  trained to reconstruct e under a sparsity penalty on z (few features active per unit).

- Each dimension of $\mathbb{R}^M$ aims to be *monosemantic* and associated with one interpretable concept (Cunningham et al., 2024; O'Neill et al., 2024).

- Each feature $z_j(x)$ is a potential score. Its activation is high on units exhibiting its concept, and it can be selected for outcome-relevance (Movva et al., 2025).

HypotheSAEs: SAE and Feature Selection

back

- A k-sparse autoencoder re-expresses the dense embedding $e = \phi(x) \in \mathbb{R}^d$ as a wider but sparse code $z(x) \in \mathbb{R}^M$, $M \gg d$, keeping the top k features per response:

$$z(x) = \mathrm{ReLU}\big(\mathrm{TopK}(W_{\mathrm{enc}}(e - b_{\mathrm{pre}}) + b_{\mathrm{enc}})\big), \qquad \widehat{e} = W_{\mathrm{dec}}\, z(x).$$

- *Picking the features (the outcome enters).* Fit an ℓ_1-regularized regression of the outcome on the sparse code, tuning λ to keep $\sim$20 nonzero coefficients:

$$\min_{\beta} \ \tfrac{1}{n} \sum_i \big(y_i - \beta^\top z(x_i)\big)^2 + \lambda \|\beta\|_1.$$

Each selected feature is a potential score, $r(x) = z_j(x)$ (Movva et al., 2025).

What Is a GAN? back

- A generative adversarial network learns a generator g mapping a low-dimensional latent vector u to a realistic unit $g(u)$, trained so its outputs are indistinguishable from real data (Goodfellow et al., 2014).

- The decoded images $\{g(u)\}$ trace out (an approximation of) the data manifold — the thin set of pixel combinations that are actually faces.

- "Stepping in latent space" means updating $u \mapsto u + t\,\delta$ and then decoding. Every $g(u + t\delta)$ is a realistic face, so following ∇r in u-space produces on-manifold counterfactuals.

What Is a VAE? back

- A variational autoencoder learns an encoder $x \mapsto u$ and a decoder $u \mapsto \hat{x}$, trained so decoded latents reconstruct the data and the latent space is smooth (Kingma and Welling, 2014).
- As with a GAN, the decoder traces out the data manifold: every decoded u is a realistic waveform.
- Unlike a GAN, a VAE comes with an encoder: embed a real ECG into latent space, step along ∇r, decode (Miller et al., 2019).

References I

Adukia, Anjali, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, and Teodora Szasz, "What We Teach About Race and Gender: Representation in Images and Text of Children's Books," *The Quarterly Journal of Economics*, 2023, *138* (4), 2225–2285.

Andre, Peter, Ingar Haaland, Christopher Roth, Mirko Wiederholt, and Johannes Wohlfart, "Narratives about the Macroeconomy," *Review of Economic Studies*, 2026, pp. 1–40.

Andrews, Isaiah, Drew Fudenberg, Lihua Lei, Annie Liang, and Chaofeng Wu, "The Transfer Performance of Economic Models," 2025. Working paper.

__ **, Toru Kitagawa, and Adam McCloskey**, "Inference on Winners," *The Quarterly Journal of Economics*, 2024, *139* (1), 305–358.

Angelopoulos, Anastasios N., Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic, "Prediction-Powered Inference," *Science*, 2023, *382* (6671), 669–674.

Angelova, Victoria, Will S. Dobbie, and Crystal S. Yang, "Algorithmic Recommendations and Human Discretion," *Review of Economic Studies*, 2025. Forthcoming.

Atalay, Enghin, Phai Phongthiengtham, Sebastian Sotelo, and Daniel Tannenbaum, "The Evolution of Work in the United States," *American Economic Journal: Applied Economics*, 2020, *12* (2), 1–34.

Athey, Susan, Herman Brunborg, Tianyu Du, Ayush Kanodia, and Keyon Vafa, "LABOR-LLM: Language-Based Occupational Representations with Large Language Models," 2024.

References II

Autor, David H., "Polanyi's Paradox and the Shape of Employment Growth," NBER Working Paper 20485, National Bureau of Economic Research 2014.

Backer-Peral, Veronica and Benjamin Wittenbrink, "Cataloging Growth: A Re-Evaluation of 1900–1990," 2026.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis, "Measuring Economic Policy Uncertainty," *The Quarterly Journal of Economics*, 2016, *131* (4), 1593–1636.

Baron, Jason, Will Dobbie, Richard Lombardo, Ashesh Rambachan, and Joseph Ryan, "Human Decisions and Machine Predictions in Multistage Systems," 2026. Working paper, June 29, 2026.

Batista, Rafael M. and James Ross, "Words that Work: Using Large Language Models to Generate and Refine Hypotheses from Text," 2024. Working paper.

Bei, Xiaohui, Jiajun Ma, Zhan Jing, Hongfei Fu, and Zhihao Gavin Tang, "EconCSLib: A Lean Library for Computational Economics and AI-Assisted Research," 2026.

Bhattacharyya, Sumanta and Pedram Rooshenas, "Steered Generation via Gradient Descent on Sparse Features," 2025.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, March 2003, *3* (null), 993–1022.

References III

Carlson, Jacob, "Making Interpretable Discoveries from Unstructured Data: A High-Dimensional Multiple Hypothesis Testing Approach," 2026.

__ **and Melissa Dell**, "A Unifying Framework for Robust and Efficient Inference with Unstructured Data," 2025. Revised February 2026.

Chen, Ruize, Ben Eltschig, Scott Duke Kominers, Ken Ono, and Jujian Zhang, "We Can't Agree to Disagree, Formally: Aumann's Theorem and Assumption Accounting in Lean," 2026. SSRN Working Paper 6837298.

Chopra, Felix and Ingar Haaland, "Conducting Qualitative Interviews with AI," 2026. Working paper, March 9, 2026.

Compiani, Giovanni, Ilya Morozov, and Stephan Seiler, "Demand Estimation with Text and Image Data," 2025.

Cunningham, Hoagy, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey, "Sparse Autoencoders Find Highly Interpretable Features in Language Models," in "International Conference on Learning Representations (ICLR)" 2024.

Dasgupta, Ishita, Andrew K. Lampinen, Stephanie C. Y. Chan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill, "Language Models Show Human-like Content Effects on Reasoning," 2022.

References IV

Dathathri, Sumanth, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu, "Plug and Play Language Models: A Simple Approach to Controlled Text Generation," in "International Conference on Learning Representations (ICLR)" 2020.

Dell, Melissa, "Deep Learning for Economists," *Journal of Economic Literature*, 2025, *63* (1), 5–58.

Dell'Acqua, Fabrizio, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine C. Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani, "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality," Technology & Operations Mgt. Unit Working Paper 24-013, Harvard Business School 2023.

Donoho, David, "Data Science at the Singularity," *Harvard Data Science Review*, 2024, *6* (1).

Dube, Oeindrila, Joshua Blumenstock, and Michael Callen, "Measuring Religion from Behavior: Violence, Climate Shocks and Religious Adherence in Afghanistan," *Journal of Political Economy*, 2026. Forthcoming.

__ **, Sandy Jo MacArthur, and Anuj K Shah**, "A Cognitive View of Policing," *The Quarterly Journal of Economics*, 02 2025, *140* (1), 745–791.

Egami, Naoki, Musashi Hinck, Brandon M. Stewart, and Hanying Wei, "Using Large Language Model Annotations for the Social Sciences: A General Framework of Using Predicted Variables in Downstream Analyses," 2024. Working paper, November 17, 2024.

References V

Elhage, Nelson, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carson Denison, Roger Grosse, Nicholas Joseph et al., "Toy Models of Superposition," *Transformer Circuits Thread*, 2022.

Ellis, Keaton, Shachar Kariv, and Erkut Ozbay, "Predicting and Understanding Individual-Level Choice Under Risk," 2026. Working paper.

Enke, Benjamin and Cassidy Shubatt, "Quantifying Lottery Choice Complexity," NBER Working Paper 31677, National Bureau of Economic Research 2023. Revised January 2026.

Feldman, Omri, Amar Venugopal, Jann Spiess, and Amir Feder, "Causal Effect Estimation with Latent Textual Treatments," 2026.

Fudenberg, Drew and Annie Liang, "Predicting and Understanding Initial Play," *American Economic Review*, 2019, *109* (12), 4112–4141.

__, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan, "Measuring the Completeness of Economic Models," *Journal of Political Economy*, 2022, *130* (4), 956–990.

Garg, Nikhil, "EconCSLib: AI-Assisted Lean Formalization for Economics & Computation research," 2026.

Geiecke, Friedrich and Xavier Jaravel, "Conversations at Scale: Robust AI-led Interviews," 2026. Working paper, February 7, 2026.

References VI

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, "Text as Data," *Journal of Economic Literature*, 2019, *57* (3), 535–574.

Glaeser, Edward L., Scott Duke Kominers, Michael Luca, and Nikhil Naik, "Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life," *Economic Inquiry*, 2018, *56* (1), 114–137.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Nets," in "Advances in Neural Information Processing Systems (NeurIPS)," Vol. 27 2014.

Gorodnichenko, Yuriy, Tho Pham, and Oleksandr Talavera, "The Voice of Monetary Policy," *American Economic Review*, 2023, *113* (2), 548–584.

Grootendorst, Maarten, "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure," 2022. arXiv:2203.05794.

Haaland, Ingar K., Christopher Roth, Stefanie Stantcheva, and Johannes Wohlfart, "Understanding Economic Behavior Using Open-ended Survey Data," NBER Working Paper 32421, National Bureau of Economic Research 2024.

Han, Sukjin, "Mining Causality: AI-Assisted Search for Instrumental Variables," 2025. Working paper, October 22, 2025.

References VII

__ , **Eric H. Schulman, Kristen Grauman, and Santhosh Ramakrishnan**, "Shapes as Product Differentiation: Neural Network Embedding in the Analysis of Markets for Fonts," 2021.

Hansen, Stephen, Michael McMahon, and Andrea Prat, "Transparency and Deliberation within the FOMC: A Computational Linguistics Approach," *The Quarterly Journal of Economics*, 2018, *133* (2), 801–870.

Hassan, Tarek A., Stephan Hollander, Laurence van Lent, and Ahmed Tahoun, "Firm-Level Political Risk: Measurement and Effects," *The Quarterly Journal of Economics*, 2019, *134* (4), 2135–2202.

Heller, Sara B., Anuj K. Shah, Jonathan Guryan, Jens Ludwig, Sendhil Mullainathan, and Harold A. Pollack, "Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago," *The Quarterly Journal of Economics*, 02 2017, *132* (1), 1–54.

Hirasawa, Toshihiko, Michihiro Kandori, and Akira Matsushita, "Using Big Data and Machine Learning to Uncover How Players Choose Mixed Strategies," 2025. University of Tokyo Market Design Center Discussion Paper UTMD-084.

Kingma, Diederik P. and Max Welling, "Auto-Encoding Variational Bayes," in "International Conference on Learning Representations (ICLR)" 2014.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human Decisions and Machine Predictions," *The Quarterly Journal of Economics*, 2018, *133* (1), 237–293.

References VIII

__ , **Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer**, "Prediction Policy Problems," *American Economic Review: Papers and Proceedings*, 2015, *105* (5), 491–495.

Korinek, Anton, "AI Agents for Economic Research," NBER Working Paper 34202, National Bureau of Economic Research 2025.

Liang, Annie, "Using Machine Learning to Generate, Clarify, and Improve Economic Models," 2026. Working paper, April 13, 2026.

Liu, Haokun, Sicong Huang, Jingyu Hu, Yangqiaoyu Zhou, and Chenhao Tan, "HypoBench: Towards Systematic and Principled Benchmarking for Hypothesis Generation," 2025.

Lord, Charles G., Mark R. Lepper, and Elizabeth Preston, "Considering the Opposite: A Corrective Strategy for Social Judgment," *Journal of Personality and Social Psychology*, 1984, *47* (6), 1231–1243.

Ludwig, Jens and Sendhil Mullainathan, "Machine Learning as a Tool for Hypothesis Generation," *The Quarterly Journal of Economics*, 2024, *139* (2), 751–827.

__ , __ , **and Ashesh Rambachan**, "Large Language Models: An Applied Econometric Framework," 2025.

Manning, Benjamin S. and John J. Horton, "General Social Agents," 2026. Also NBER Working Paper 34937.

__ , **Kehang Zhu, and John J. Horton**, "Automated Social Science: Language Models as Scientist and Subjects," NBER Working Paper 32381, National Bureau of Economic Research 2024.

References IX

McCoy, R. Thomas, Shunyu Yao, Dan Friedman, Matthew Hardt, and Thomas L. Griffiths, "Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve," 2023.

Michalopoulos, Stelios and Melanie Meng Xue, "Folklore," *The Quarterly Journal of Economics*, 2021, *136* (4), 1993–2046.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean, "Distributed Representations of Words and Phrases and their Compositionality," in "Advances in Neural Information Processing Systems (NeurIPS)" 2013.

Miller, Andrew C., Ziad Obermeyer, John P. Cunningham, and Sendhil Mullainathan, "Discriminative Regularization for Latent Variable Models with Applications to Electrocardiography," in "Proceedings of the 36th International Conference on Machine Learning (ICML)" 2019.

Modarressi, Iman, Jann Spiess, and Amar Venugopal, "Causal Inference on Outcomes Learned from Text," 2025.

Movva, Rajiv, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson, "Sparse Autoencoders for Hypothesis Generation," in "Proceedings of the 42nd International Conference on Machine Learning (ICML)" 2025.

References X

__ , **Smitha Milli, Sewon Min, and Emma Pierson**, "What's in My Human Feedback? Learning Interpretable Descriptions of Preference Data," in "International Conference on Learning Representations (ICLR)" 2026.

Mullainathan, Sendhil, "Economics in the Age of Algorithms," *AEA Papers and Proceedings*, 2025, *115*, 1–23. AEA Distinguished Lecture.

__ **and Ashesh Rambachan**, "From Predictive Algorithms to Automatic Generation of Anomalies," 2025.

__ **and** __ , "Science in the Age of Algorithms," 2025. Working paper, October 10, 2025.

__ **and Jann Spiess**, "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 2017, *31* (2), 87–106.

Nickerson, Raymond S., "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises," *Review of General Psychology*, 1998, *2* (2), 175–220.

Obermeyer, Ziad, Alexander Schubert, James Ross, Sendhil Mullainathan, and Markus Lingman, "An ECG Biomarker for Sudden Cardiac Death Discovered with Deep Learning," *Nature*, 2026. Published online 24 June 2026.

Olah, Chris, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter, "Zoom In: An Introduction to Circuits," *Distill*, 2020.

O'Neill, Charles, Christine Ye, Kartheik Iyer, and John F. Wu, "Disentangling Dense Embeddings with Sparse Autoencoders," 2024.

References XI

Peng, Kenny, Rajiv Movva, Jon Kleinberg, Emma Pierson, and Nikhil Garg, "Position: Use Sparse Autoencoders to Discover Unknowns," in "Proceedings of the 43rd International Conference on Machine Learning (ICML)" 2026.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning, "GloVe: Global Vectors for Word Representation," in "Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)" 2014, pp. 1532–1543.

Pernaudet, Julie, John A. List, Arnoldo Müller-Molina, Majid Ahmadi, Imrul Huda, Ajay Sailopal, and Dana Suskind, "Leveraging Artificial Intelligence and Field Experiments to Explore Novel Features of Parental Speech and Foster Child Development," 2026. University of Chicago Becker Friedman Institute Working Paper 2026-80.

Peterson, Joshua C., David D. Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L. Griffiths, "Using Large-Scale Experiments and Machine Learning to Discover Theories of Human Decision-Making," *Science*, 2021, *372* (6547), 1209–1214.

Peysakhovich, Alexander and Jeffrey Naecker, "Using Methods from Machine Learning to Evaluate Behavioral Models of Choice under Risk and Ambiguity," *Journal of Economic Behavior & Organization*, 2017, *133*, 373–384.

References XII

Plonsky, Ori, Reut Apel, Eyal Ert, Moshe Tennenholtz, David Bourgin, Joshua C. Peterson, Daniel Reichman, Thomas L. Griffiths, Stuart J. Russell, Evan C. Carter, James F. Cavanagh, and Ido Erev, "Predicting Human Decisions with Behavioral Theories and Machine Learning," 2025. Working paper, March 28, 2025.

Rambachan, Ashesh, "Identifying Prediction Mistakes in Observational Data," *Quarterly Journal of Economics*, 2024, *139* (3), 1665–1711.

Sarkar, Suproteem K., "Economic Representations," 2026. Working paper, June 25, 2026.

__ **and Keyon Vafa**, "Lookahead Bias in Pretrained Language Models," 2024. SSRN Working Paper 4754678, June 28, 2024.

Social Catalyst Lab, "Project APE: Autonomous Policy Evaluation," `https://ape.socialcatalystlab.org` 2026.

Tan, Jiyuan and Vasilis Syrgkanis, "CausalForge: A Formally Grounded, Self-Improving Agentic Framework for Automated Research in Causal Inference," 2026.

Tranchero, Matteo, Cecil-Francis Brenninkmeijer, Arul Murugan, and Abhishek Nagaraj, "Theorizing with Large Language Models," NBER Working Paper 33033, National Bureau of Economic Research 2024.

References XIII

Vafa, Keyon, Ashesh Rambachan, and Sendhil Mullainathan, "Do Large Language Models Perform the Way People Expect? Measuring the Human Generalization Function," in "Proceedings of the 41st International Conference on Machine Learning (ICML)" 2024.

__ **, Emil Palikot, Tianyu Du, Ayush Kanodia, Susan Athey, and David M. Blei**, "CAREER: A Foundation Model for Labor Sequence Data," 2022.

Wang, Jenny S., Aliya Saperstein, and Emma Pierson, "In Your Own Words: Computationally Identifying Interpretable Themes in Free-Text Survey Data," 2026.

Wang, Ke, Hang Hua, and Xiaojun Wan, "Controllable Unsupervised Text Attribute Transfer via Editing Entangled Latent Representation," in "Advances in Neural Information Processing Systems (NeurIPS)" 2019.

Workman, Ben, "Inside the Black Box: Using Machine Learning to Discover Characteristics of Effective Teaching," 2025. Working paper.

Wright, James R. and Kevin Leyton-Brown, "Beyond Equilibrium: Predicting Human Behavior in Normal-Form Games," in "Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-10)" 2010.

__ **and** __ , "Predicting Human Behavior in Unrepeated, Simultaneous-Move Games," *Games and Economic Behavior*, 2017, *106*, 16–37.

References XIV

Zhong, Ruiqi, Charlie Snell, Dan Klein, and Jacob Steinhardt, "Describing Differences between Text Distributions with Natural Language," in "Proceedings of the 39th International Conference on Machine Learning (ICML)" 2022.

__ , **Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt**, "Goal Driven Discovery of Distributional Differences via Language Descriptions," in "Advances in Neural Information Processing Systems (NeurIPS)" 2023.

Zhou, Yangqiaoyu, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan, "Hypothesis Generation with Large Language Models," 2024.

Zhu, Jian-Qiao, Hanbo Xie, Dilip Arumugam, Robert C. Wilson, and Thomas L. Griffiths, "Using Reinforcement Learning to Train Large Language Models to Explain Human Decisions," in "International Conference on Learning Representations (ICLR)" 2026.

__ , **Joshua C. Peterson, Benjamin Enke, and Thomas L. Griffiths**, "Capturing the Complexity of Human Strategic Decision-Making with Machine Learning," 2024.