Measurement Without Random Validation Samples

Melissa Dell Ashesh Rambachan

NBER Methods Lecture 2026

Slides posted on the NBER Methods Lecture website.

Measurement Bottleneck

- The researcher collects unstructured data X_i (e.g., documents, images, videos) and linked covariates W_i.
- Each piece of unstructured data expresses a construct, defined by an existing measurement process: $V_i^* = f^*(X_i)$. Downstream parameters are defined from (V_i^*, W_i).
- The measurement process $f^*(\cdot)$ is chosen by the researcher. It is whatever they would use absent time and resource constraints:
  - *Text*: write a careful rubric that defines the construct / adjudicates edge cases, and have experts label each document in short sessions.
  - *Environment*: send surveyors into remote forests at high frequency to measure tree cover; blanket fields with physical monitors to measure pollution.
  - *Economic outcomes*: field extensive in-person surveys to elicit consumption, income, ...

Measurement Bottleneck

- The researcher collects unstructured data X_i (e.g., documents, images, videos) and linked covariates W_i. Unstructured data expresses a construct, defined by an existing measurement process: $V_i^* = f^*(X_i)$.

- The researcher faces a measurement bottleneck. Applying the measurement process across the full sample is infeasible.
  - We cannot carefully read every document in the corpus.
  - We cannot send surveyors into remote forests at high frequency or blanket fields with physical monitors.
  - We cannot run extensive, in-person surveys in every place and period.

- **Question**: how can the researcher use ML/AI-generated measurements R_i to solve the measurement bottleneck?

Measurement with Validation Samples

- **Question**: how can the researcher use ML/AI-generated measurements R_i to solve the measurement bottleneck?

- **Solution so far**: collect R_i on all observations and collect V_i^* on a random validation sample to debias the ML/AI-generated measurements.
  - An old idea in the measurement error literature (e.g., Bound and Krueger, 1991; Lee and Sepanski, 1995; Chen et al., 2005, 2011).
  - Recently revived by ML/AI literature (e.g., Angelopoulos et al., 2023; Egami et al., 2024; Ludwig et al., 2026; Carlson and Dell, 2025).
  - Related: bias correction also possible w/ a small external audit, or structure on how the measurement is generated (Battaglia et al., 2025).

- Why is this attractive? Frontier ML/AI models are complicated (billions of parameters, trained on data you have never seen) $\rightarrow$ correct for errors without strong assumptions on their outputs.

Measurement with Validation Samples

- **Question**: how can the researcher use ML/AI-generated measurements R_i to solve the measurement bottleneck?
- **Solution so far**: collect R_i on all observations and collect V_i^* on a random validation sample to debias the ML/AI-generated measurements.
- **Payoff**: recover the estimand defined by (V_i^*, W_i) while paying measurement costs on only a subset of the sample.
- There are two requirements:
  1. Researcher must choose the existing measurement process $f^*(\cdot)$.
  2. Validation sample collected from researcher's own sample.

This Lecture

- **Question**: how can the researcher use ML/AI-generated measurements R_i to solve the measurement bottleneck?

- **Solution so far**: collect R_i on all observations and additionally collect V_i^* on a random validation sample to debias the ML/AI-generated measurements.

- **This lecture**: what happens when 1 or 2 fails?
  1. There is no existing measurement process that we would like to scale.
     - Most often in text applications: labeling documents with large language models.
  2. There is an existing measurement but only in another sample.
     - Most often in remote-sensing applications: existing measurements from surveys or censuses covering another region or period.

This Lecture

- **Question**: how can the researcher use ML/AI-generated measurements R_i to solve the measurement bottleneck?

- **Solution so far**: collect R_i on all observations and additionally collect V_i^* on a random validation sample to debias the ML/AI-generated measurements.

- **This lecture**: what happens when 1 or 2 fails?
  1. There is no existing measurement that we would like to scale.
     - Measurement error assumptions are assumptions about frontier model behavior.
     - Construct is latent, and ML/AI offers multiple measurements.
     - Revisit identification results using what is known about frontier model behavior.
  2. There is an existing measurement but only in another sample.

This Lecture

- **Question**: how can the researcher use ML/AI-generated measurements R_i to solve the measurement bottleneck?

- **Solution so far**: collect R_i on all observations and collect V_i^* on a random validation sample to debias the ML/AI-generated measurements.

- **This lecture**: what happens when these requirements fail?
  1. There is no existing measurement that we would like to scale.
  2. There is an existing measurement but only in another sample.
     - Identification turns on what is stable between the researcher's sample and the auxiliary sample.
     - Two candidate assumptions are incompatible (outside of knife-edge cases), and so researchers must choose.

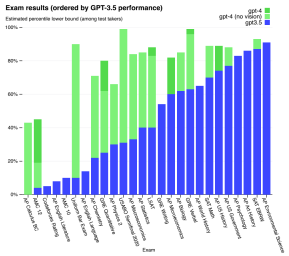
This Lecture

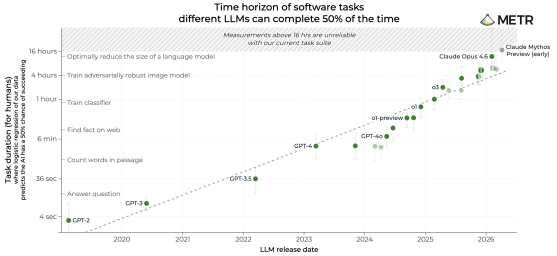
- **Question**: how can the researcher use ML/AI-generated measurements R_i to solve the measurement bottleneck?
- **Solution so far**: collect R_i on all observations and collect V_i^* on a random validation sample to debias the ML/AI-generated measurements.
- **This lecture**: what happens when these requirements fail?
  1. There is no existing measurement that we would like to scale.
  2. There is an existing measurement but only in another sample.

What Counts as Ground Truth?

- With text, the measurement process typically proceeds in two steps (e.g., Baker et al., 2016; Card et al., 2022; Ottonello et al., 2024).
  - Step 1 (Definition): draft a detailed rubric for the construct; label with it; surface edge cases and disagreements; iterate.
  - Step 2 (Labeling): trained annotators apply the rubric across the corpus.
- **Solution so far**: invest in Step 1 and then run Step 2 on a random subsample $\rightarrow$ recovers the estimand defined by V_i^*.
- Why should we privilege trained human annotators once the rubric is defined?

Impressive Capabilities of AI





- If frontier AI can accomplish these feats, why privilege trained human annotators once the rubric is defined?

Which Model? Which Prompt?

- Suppose we grant the point: just hand the rubric to a frontier model rather than a trained annotator.
- The rubric alone does not pin down the measurement R_i. Many choices remain:
  - Which model should I use?
  - Which prompting strategy should I use?
- Do these choices change the resulting measurements? (e.g., Sclar et al., 2024; Atreja et al., 2025)

Illustration: Congressional Legislation

- **Dataset**: 10,000 randomly sampled Congressional bills through 2018 (Adler and Wilkerson, 2020). For each bill, we observe:
  - Text description (X_i),
  - Party affiliation of sponsor, DW-NOMINATE score of sponsor, and whether bill originated in House or Senate (W_i),
  - Label for bill's policy area (e.g., defense, health, etc.) based on text description (V_i^*).

- Example: "`A bill to revise the boundary of Crater Lake National Park in the State of Oregon.`"

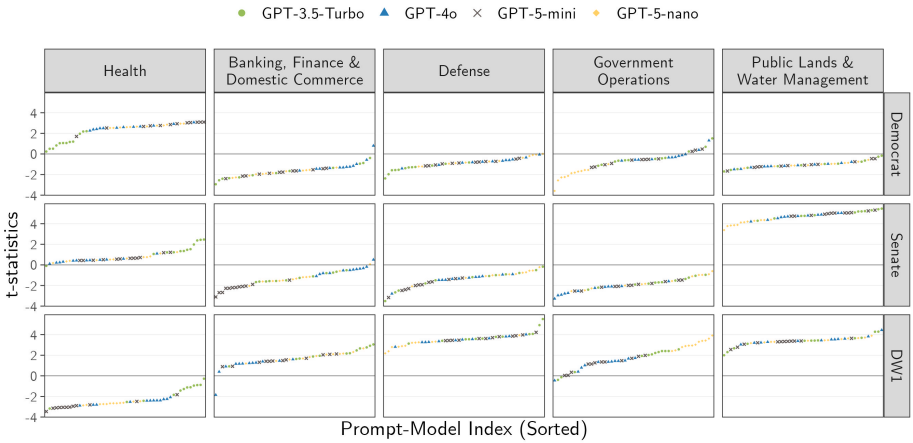
Illustration: Congressional Legislation

- **Dataset**: 10,000 randomly sampled Congressional bills through 2018 (Adler and Wilkerson, 2020). For each bill, we observe:
  - Text description (X_i),
  - Party affiliation of sponsor, DW-NOMINATE score of sponsor, and whether bill originated in House or Senate (W_i),
  - Label for bill's policy area (e.g., defense, health, etc.) based on text description (V_i^*).
- Using various prompt strategies with models from OpenAI's GPT family, label each bill's policy area (R_i) (Ludwig et al., 2026).
- Compare regression estimates across alternative LLM measurements:

$$1\{R_i = r\} = W_i'\beta^{\text{LLM}} + \tilde{\epsilon}_i$$

Do these choices matter?

Illustration: Congressional Legislation



- **Yes**: across models and prompts, t-statistics differ in magnitude, in significance, and even in sign (Ludwig et al., 2026).

Illustration: Immigration Speeches, Revisited

- **Dataset**: immigration-related speeches in the U.S. Congress, 1880–2020 (Card et al., 2022). For each speech, we observe:
  - Speech text (X_i),
  - Speaker's party, chamber, and Congress (W_i),
  - Tone label: proimmigration, neutral, or antiimmigration (V_i^*).

- Example: "`Let us not forget the part that immigrants played in this country. Who are you and I to discredit the immigrants?`"

Illustration: Immigration Speeches, Revisited

- **Dataset**: immigration-related speeches in the U.S. Congress, 1880–2020 (Card et al., 2022). For each speech, we observe:
  - Speech text (X_i),
  - Speaker's party, chamber, and Congress (W_i),
  - Tone label: proimmigration, neutral, or antiimmigration (V_i^*).

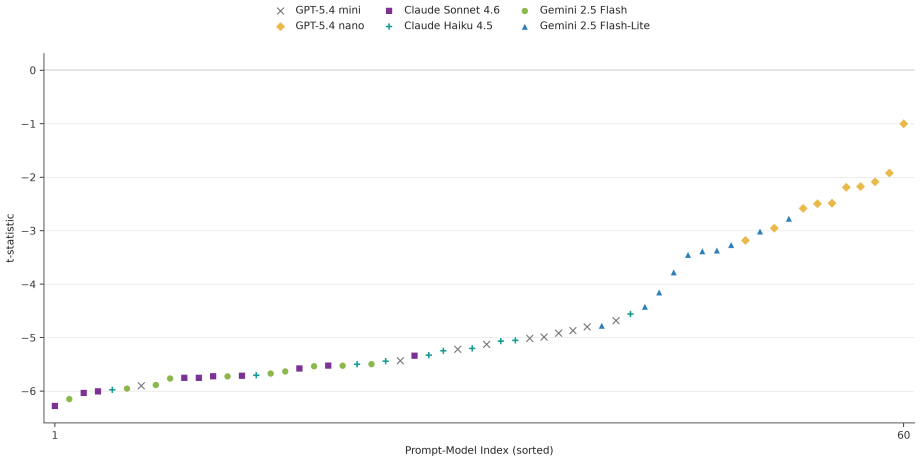
- Using various prompt strategies with models from OpenAI's GPT, Anthropic's Claude, and Google's Gemini families, label tone of each speech segment (R_i).

- Compare estimates of the partisan tone gap using alternative LLM measurements:

$$R_i = W_i' \beta^{\text{LLM}} + \tilde{\epsilon}_i$$

  Do these choices matter?

Illustration: Immigration Speeches, Revisited



- **Yes**: across models and prompts, the estimated partisan tone gap differs in magnitude and in significance.

The Language Model Is All There Is?

- Given the rubric, the remaining choices — which model? which prompt? — can move the estimates in magnitude, significance, and even sign (e.g., Ludwig et al., 2026; Yin et al., 2026).

- **One answer**: stop treating the model as a proxy. The construct *is* the model's output — $f^*(\cdot)$ is the measurement produced by a particular model under a particular prompt.
  - How do we manage the choice of model (GPT vs. Claude vs. Gemini)? The choice of prompt?
  - Do we revisit everything once the next model generation ships?

- **Complication**: the estimand is now indexed by the model-prompt choice. Are researchers using different models and prompts really studying different estimands?

The Language Model Is All There Is?

- Given the rubric, the remaining choices — which model? which prompt? — can move the estimates in magnitude, significance, and even sign (e.g., Ludwig et al., 2026; Yin et al., 2026).
- **One answer**: stop treating the model as a proxy. The construct *is* the model's output — $f^*(\cdot)$ is the measurement produced by a particular model under a particular prompt.
- **Complication**: the estimand is now indexed by the model-prompt choice. Are researchers using different models and prompts really studying different estimands?
- Better measurement harnesses may make these choices less consequential: Asirvatham et al. (2026) provides one such harness, with evidence of better robustness across semantically similar prompts.

Latent Constructs and Repeated Measurements

- **Alternative view**: the construct V_i^* is inherently latent.
  - There is no feasible measurement process that recovers it, human or machine.
  - Every label (whether by expert or AI) is an imperfect proxy for the construct.

- Under this view, humans vs. AI and ChatGPT vs. Claude vs. open-source model are potentially repeated measurements of the same latent construct V_i^*.

- Could we apply classic ideas from the measurement error literature? (e.g., Dawid and Skene, 1979; Hu and Schennach, 2008; Bonhomme et al., 2016)

Example: Identification with Repeated Measurements

- Let $k = 1, \ldots, K$ denote different model-prompt combinations, where R_{ik} is the associated measurement.
- Suppose $V_i^* \in \{0, 1\}$ with $\theta := \Pr(V_i^* = 1) \in (0, 1)$.
- **Assumption**: $R_{i1}, \ldots, R_{iK}$ are conditionally independent given V_i^*.
  - Measurements co-move only because of the latent construct.
  - Nests classical measurement error: with continuous measurements, $R_{ik} = V_i^* + \epsilon_{ik}$, noise independent across k and of V_i^*.
- Intuition: $2^K - 1$ observed frequencies vs. $2K + 1$ unknowns: at $K = 3$, seven equations and seven unknowns.

Example: Identification with Repeated Measurements

- **Assumption**: $R_{i1}, \ldots, R_{iK}$ are conditionally independent given V_i^*.
- **Additional assumption**: at least three measurements are informative and oriented, $\Pr(R_{ik} = 1 \mid V_i^* = 1) > \Pr(R_{ik} = 1 \mid V_i^* = 0)$.
- **Theorem** (informal): under conditional independence, informativeness, and orientation, the joint distribution of $(R_{i1}, \ldots, R_{iK})$ point identifies θ.
  - Celebrated result: Dawid and Skene (1979); Schennach (2021)
- Generalizations: multi-valued case: informativeness becomes a full-rank condition on three *blocks* of protocols (i.e., dependence allowed within blocks) and generalized orientation (Bonhomme et al., 2016; Schennach, 2021); general moment conditions (Chen et al., 2011; Schennach, 2020).

What Does This Mean for LLMs?

- Identification is driven by what we are willing to assume about language model behavior. Which assumptions are credible?

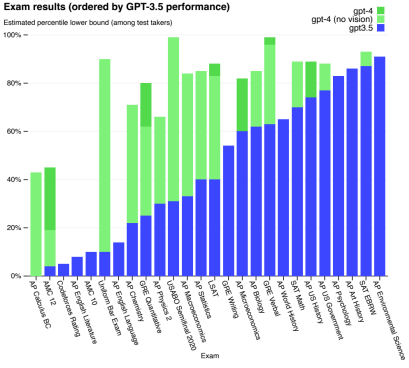
- In this simple case, the additional informativeness and orientation assumption is mild: at least three measurements satisfy $\Pr(R_{ik} = 1 \mid V_i^* = 1) > \Pr(R_{ik} = 1 \mid V_i^* = 0)$.
  - Presumably a researcher willing to use language model measurements is already assumes this.

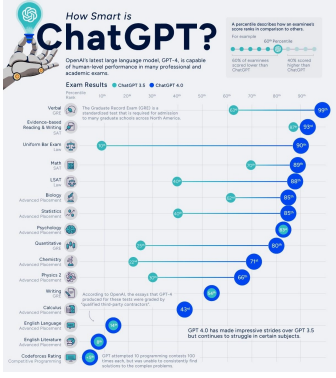
What Does This Mean for LLMs?

- Identification is driven by what we are willing to assume about language model behavior. Which assumptions are credible?
- In this simple case, the additional informativeness and orientation assumption is mild.
- Conditional independence is key. Is that credible for language models?
  - Co-movement of the measurements is only driven by the latent construct.
  - Errors must be unrelated across models and prompts conditional on the latent construct.
- Why might we worry? Substantial overlap in pre-training data, fine-tuning data, and architectures across models (Kleinberg and Raghavan, 2021; Bommasani et al., 2022).
- Ultimately this is an empirical question about frontier model behavior.

Benchmark Evaluations

- Quantitative basis for model behavior comes from benchmark evaluations: performance on standardized exams, domain-specific exams, tests of "general intelligence", coding, math problems, etc.





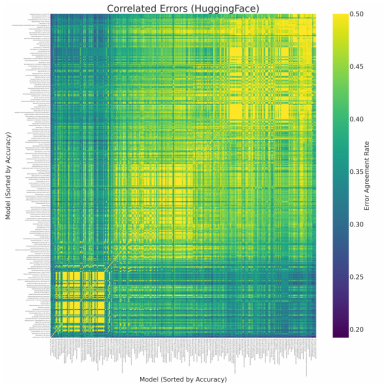
Evidence of Correlated Errors

- Why might errors be correlated across language models? Shared pre-training data, similar architectures, similar post-training data, ...

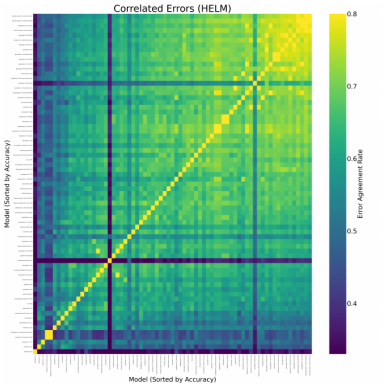
- Kim et al. (2025): across pairs of models on AI benchmarks, report how often two models give the same wrong answer conditional on both being wrong.
  - 349 models on a HuggingFace Leaderboard; 71 models on Stanford's Holistic Evaluation of Language Models (Helm).

- The authors find:
  1. Wrong-answer agreement far exceeds chance: 42% vs. 13% at random (HuggingFace); 60% vs. 33% (HELM).
  2. Better-performing models tend to have more correlated errors.

Evidence of Correlated Errors



(a) HuggingFace



(b) Helm

Same-wrong-answer agreement, models sorted by accuracy: nearly all pairs exceed chance, and more accurate models have more correlated errors (Kim et al., 2025, Figure 1)

From AI Benchmarks to Economic Applications

- This evidence focuses on AI benchmarks. What about the measurement problems that economists are interested in studying?

- Let's revisit our two running examples and assess whether measurements from different models / prompts are conditionally independent:
  1. Congressional legislation: policy topics (Ludwig et al., 2026).
  2. Immigration speeches: tone (Card et al., 2022).

- We use the authors' human-annotated labels as the reference ground-truth V_i^*.

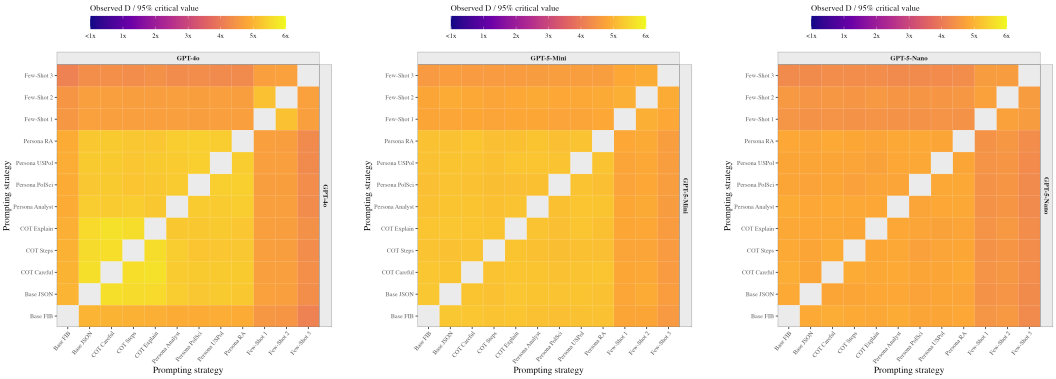
From AI Benchmarks to Economic Applications

- Let's revisit our two running examples and assess whether measurements from different models / prompts are conditionally independent.

- For each pair of measurements:
  1. If conditionally independent given the human label, the joint distribution of the pair factorizes into the product of marginals. Compute this benchmark.
  2. Compute the total variation (TV) distance between the observed joint distribution and the independence benchmark.
  3. Simulate critical values under the null by permuting one model's labels conditional on the authors' labels.

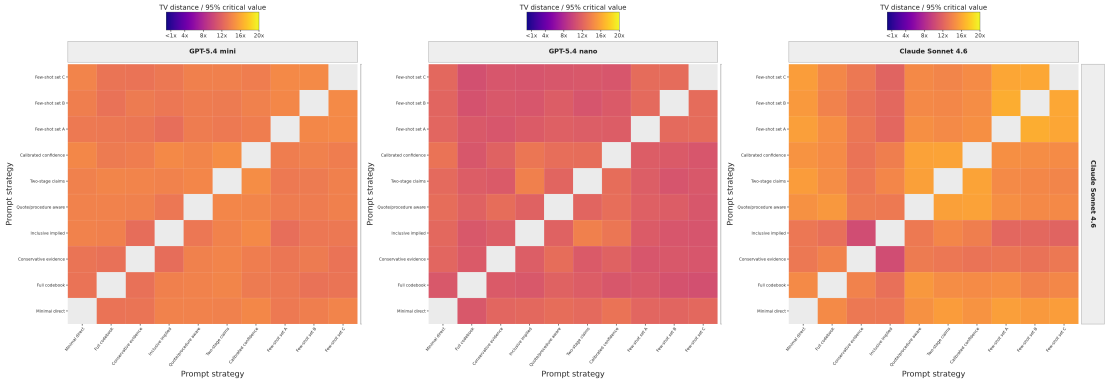
- Report observed TV / 95% critical value for each model pair: a simple summary of the strength of evidence against conditional independence (ratio > 1 rejects at the 5% level).

Illustration: Congressional Legislation



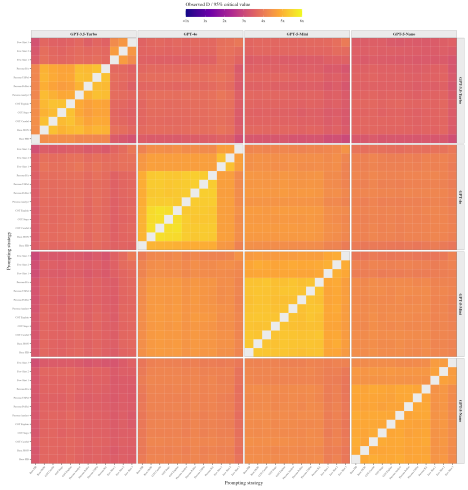
Within a model, each prompt pair sits at 3–6$\times$ the critical value.

Illustration: Immigration Speeches, Revisited

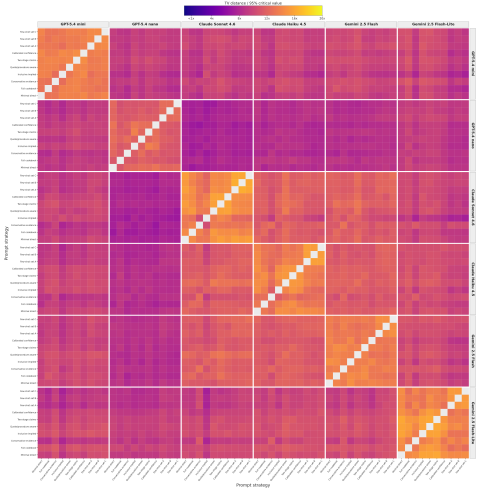


Within a model, each prompt pair sits at 8–20× the critical value.

Illustrations: Correlated Errors across Models



Congressional legislation



Immigration speeches

Evidence of Correlated Errors

- On benchmarks and in economic applications, errors are strongly correlated across models and prompts.
- Assuming language model labels across models and prompts are conditionally independent is not credible $\rightarrow$ identification results that rely on it should not be naively applied to LLM-generated measurements.
- But the errors are not arbitrarily correlated either. At least this is true for the benchmark evaluations conducted so far.
- Could we still make progress?

Relaxing Conditional Independence

- Errors are correlated but not perfectly. Benchmarks measure model behavior. Can we incorporate these evaluations into our econometric framework?

- One strategy (Chen et al., 2026): replace conditional independence with what benchmarks can help evaluate:
  1. Marginal performance: how often each model-prompt combination is right?
  2. Joint response events: how often two model-prompt combinations err together?

Relaxing Conditional Independence

- Errors are correlated but not perfectly. Benchmarks measure model behavior. Can we incorporate these evaluations into our econometric framework?

- One strategy (Chen et al., 2026): replace conditional independence with restrictions on marginal performance and joint response events, informed by benchmarks.

- Downstream parameter is partially identified under restrictions on how correlated protocols may be, or how far performance degrades from benchmark to task, indexed by δ:

$$\mathcal{H}(\theta;\delta) = \left[\, \underline{\theta}(\delta),\ \overline{\theta}(\delta) \,\right],$$

  Identified set is an interval whose endpoints solve two linear programs.

Relaxing Conditional Independence

- One strategy (Chen et al., 2026): replace conditional independence with restrictions on marginal performance and joint response events, informed by benchmarks.

- Downstream parameter is partially identified under restrictions on how correlated protocols may be, or how far performance degrades from benchmark to task, indexed by δ:

$$\mathcal{H}(\theta; \delta) = \left[\underline{\theta}(\delta),\ \overline{\theta}(\delta) \right],$$

- Researchers can report sensitivity analyses on measurements that are informed by benchmark evaluations:
  - "Performance would have to degrade by δ^* relative to benchmarks before my conclusion is overturned."
  - "Model-prompts would have to at least as correlated as they are on benchmarks before my conclusion is overturned."

Takeaways

- **Step back**: cleanest framework remains the validation sample. Properly understood, its goal is to automate our measurements at their best.

- "Best" is *not* the compromise we settled for pre-AI. It is the measurement process we would actually defend: a documented adjudicated rubric decided upon and applied by domain experts.

- "Best" no longer has to cover the corpus. A small expert-labeled sample may be affordable, and AI does the scaling.

- The scarce input is constructing high-quality measurements, and that is what researchers must invest in.
  - We have seen this before: when running regressions became cheap, attention moved to the identification strategy (Angrist and Pischke, 2010).

Measurement Error as Model Behavior

- Outside of the validation-sample framework, assumptions on measurement error are unavoidable.

- Assumptions on measurement error are claims about frontier model behavior. An entire field already studies it at scale: benchmarking.
  - Conditional independence: assumed it $\rightarrow$ benchmarks falsified it $\rightarrow$ benchmarks provide external calibration.
  - What's the analogue of instruments?

Measurement Error as Model Behavior

- Outside of the validation-sample framework, assumptions on measurement error are unavoidable: conditional independence, instruments, bounded dependence.

- Assumptions on measurement error are claims about frontier model behavior. An entire field already studies it at scale: benchmarking.

- But today's AI benchmarks are not the ones we need, in two ways:
  - Wrong tasks: exams and coding, not economic measurement. We are bad at generalizing performance does to new tasks (Vafa et al., 2024).
  - Wrong properties: typically report accuracy, which does not fully characterize biases of downstream estimates.

- Benchmarking for economic measurement is an essential activity: it allows us to check and falsify measurement error assumptions.

This Lecture, Revisited

- **So far**: 1 there is no existing measurement to scale. The construct is latent, every identification assumption is a claim about frontier model behavior, and benchmarks are how we discipline it.

- **Next**: 2 there is an existing measurement but only in another sample.
  - Ground truth from a survey or census covering another region, period, or population.
  - Most often in remote sensing and digital-trace applications.

- The researcher must combine two samples, and identification turns on what is assumed stable across them.

When Ground Truth Is Expensive

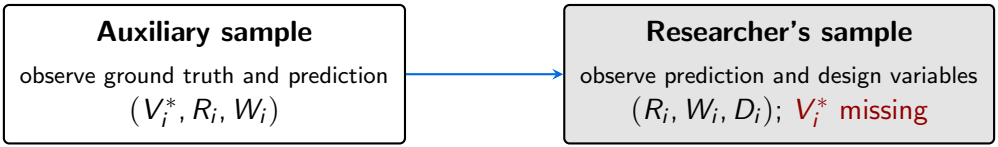
- The existing measurement is not in dispute: everyone agrees what a consumption survey or a forest-cover ground plot measures.
- The constraint is cost and coverage. Surveys and field measurements run in some places, in some years.
- By contrast, satellite images and digital traces are frequent, granular, already collected.
  - A growing literature measures poverty, consumption, deforestation, and crop burning at scale (e.g., Blumenstock et al., 2015; Jean et al., 2016; Huang et al., 2021; Ratledge et al., 2021; Aiken et al., 2025; Jack et al., 2025).
- How can we take advantage of these unstructured data to lower measurement costs in experiments and observational studies?

With Validation Data, We Are on Familiar Ground

- Researcher collects unstructured data X_i (satellite imagery and digital traces) for each unit.
- Each unit is associated with some construct V_i^* (i.e., the survey or field measurement) and a prediction R_i of the construct that is built from the unstructured data X_i.
- Remotely sensed predictions are imperfect proxies of V_i^* that suffer from various forms of measurement error (e.g., Alix-García and Millimet, 2023; Proctor et al., 2023, 2025).
- **Solution so far**: collect R_i on all observations and collect V_i^* on a random validation sample to debias the ML/AI-generated measurements (e.g., Kluger et al., 2025; Lu et al., 2025; Pelletier et al., 2026).

Labels Exist Somewhere Else

- We cannot always collect a validation sample: infeasible, or too costly.
- But labels frequently already exist: a government environmental survey, a census, or an earlier survey round.
  - Collected somewhere other than the researcher's sample, such as another region or another time period.



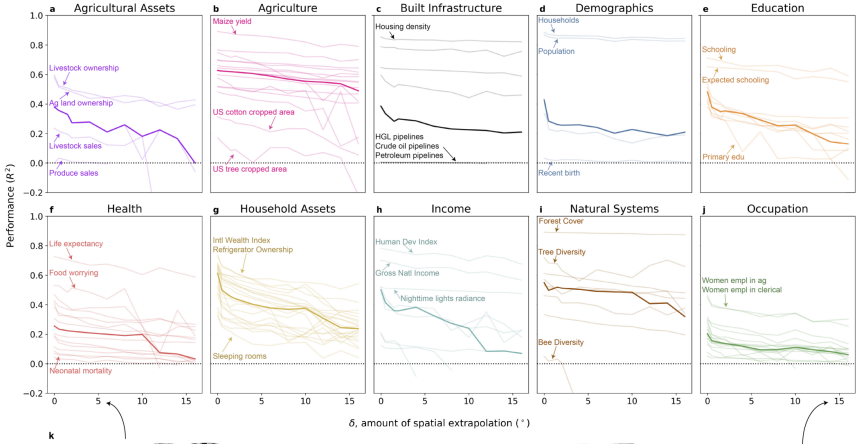
- Can we use these existing labels to correct the errors in R_i? A data combination problem.

Do Predictions Transfer?

- Suppose the auxiliary sample were a random validation sample. What would we expect?
  - Predictive performance should be stable as we transfer to another region or time period.

- Proctor et al. (2025): predictors of 115 socioeconomic and environmental variables, built from a single satellite mosaic.
  - `MOSAIKS`: a general-purpose embedding of satellite imagery, computed once and reused for every outcome (Rolf et al., 2021).
  - Each of the 115 variables is predicted by a ridge regression on that embedding.
  - Evaluate on a random split, then on checkerboards that push training and validation cells further apart geographically.
  - See also Aiken et al. (2023, 2026).

Do Predictions Transfer?



- **No**: median R^2 falls by half, $0.44 \rightarrow 0.22$ (Proctor et al., 2025, Figure 3).

Do Predictions Transfer?

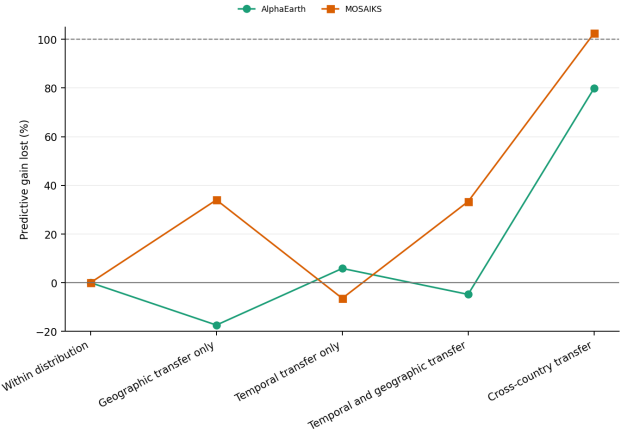
- Run the same exercise ourselves: train in one place and period, then transfer the prediction to another region or another period.

- Two targets:
  1. **Living standards**: a 14-component index of housing, infrastructure, assets, and financial access (Kenya DHS 2022). Train on 1,064 clusters; evaluate on held-out clusters, a western block, an earlier Kenya survey, and Ethiopia.
  2. **Forest cover**: train on Uganda cells, 2017–2020; evaluate in a western region and in 2021–2024, never used in fitting (Hansen et al., 2013).

- Report the predictive gain lost in each new environment:

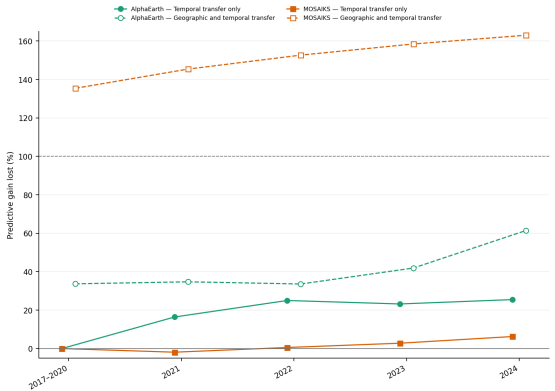
$$0\% = \text{does as well as in its own development sample,}$$
$$100\% = \text{no better than the sample mean.}$$

Living Standards: Transferring Across Regions and Countries



- Within Kenya transfer: modest losses. Across country transfer: roughly 80% of the gain is gone, and at worst the prediction does worse than the sample mean.

Forest Cover: Transferring Across Years and Regions



- Losses grow with temporal distance, and compound when geography shifts too.
- Prediction quality is not stable under transfer: using it elsewhere requires additional assumptions.

Two Samples

- Suppose the researcher has two samples, indexed by $S_i \in \{\mathrm{e}, \mathrm{a}\}$:
  1. Empirical sample ($S_i = \mathrm{e}$): sample of interest with only (W_i, R_i) observed.
  2. Auxiliary sample ($S_i = \mathrm{a}$): where the existing measurement was collected and (W_i, R_i, V_i^*) are observed.

- Both samples observe the prediction R_i and covariates W_i. Only the auxiliary sample observes the ground truth V_i^*.

- The target is a parameter of the empirical sample: e.g., a base rate $\Pr(V_i^* = 1 \mid S_i = \mathrm{e})$, or a treatment effect of D_i on V_i^*.

- How can we combine these two samples?

- NB: Distinct from classical data combination problems, which combines unmatched observations from the same population (Fan et al., 2014; D'Haultfœuille et al., 2025a,b).

Outcome Stability

- **Assumption** (outcome stability):

$$V_i^* \perp\!\!\!\perp S_i \mid (R_i, W_i).$$

  Predictions of $V_i^* \mid R_i, W_i$ can transfer across the samples, but the distribution of (R_i, W_i), and the base rate of V_i^* may differ (e.g., Chen et al., 2008; Athey et al., 2026; Kallus and Mao, 2025; Proctor et al., 2023).

- Example: a poverty score built from mobile phone metadata (Blumenstock et al., 2015; Aiken et al., 2025).
  - Plausible claim: the same score means the same expected consumption in both samples.

Outcome Stability

- **Assumption** (outcome stability): $V_i^* \perp\!\!\!\perp S_i \mid (R_i, W_i)$.
- **Strategy**: learn $\mu_{\mathrm{a}}(r, w) = \Pr(V_i^* = 1 \mid R_i = r, W_i = w, S_i = \mathrm{a})$ in the auxiliary sample, then transfer it over the empirical sample. For $\theta = \Pr(V_i^* = 1)$,

$$\theta = \mathbb{E}\left[\mu_{\mathrm{a}}(R_i, W_i) \mid S_i = \mathrm{e}\right].$$

- **Additional assumption** (support): $P_{\mathrm{e}}^{R,W} \ll P_{\mathrm{a}}^{R,W}$ — every (R_i, W_i) the empirical sample visits is one the auxiliary sample visits.

Outcome Stability

- **Assumption** (outcome stability): $V_i^* \perp\!\!\!\perp S_i \mid (R_i, W_i)$.
- Special case: no covariates, $V_i^* \in \{0, 1\}$ and $R_i \in [0, 1]$, so R_i is a predicted probability. Outcome stability is then

$$\mathbb{E}\left[V_i^* \mid R_i = r, S_i = \mathrm{e}\right] = \mathbb{E}\left[V_i^* \mid R_i = r, S_i = \mathrm{a}\right] \quad \text{for all } r.$$

- Interpretation: the predictor has the same calibration curve in both samples.
  - In algorithmic fairness, this is "sufficiency," with sample membership S_i playing the role of the protected attribute (Chouldechova, 2017; Kleinberg et al., 2017; Barocas et al., 2023).
  - In distribution shift, this is "covariate shift."

Measurement Stability

- **Assumption** (measurement stability):

$$R_i \perp\!\!\!\perp S_i \mid (V_i^*, W_i).$$

The measurement channel $R_i \mid V_i^*, W_i$ can transfer across the samples, but base rates and predictions may differ (e.g., Lipton et al., 2018; Rambachan et al., 2026).

- Example: crop burning, measured from satellite imagery (Jack et al., 2025; Rambachan et al., 2026).
  - The outcome generates the signal: a burned field reflects differently.
  - Plausible claim: burning looks the same from orbit in both samples.

Measurement Stability

- **Assumption** (measurement stability): $R_i \perp\!\!\!\perp S_i \mid (V_i^*, W_i)$.

- **Strategy**: learn $p_v(w) = \Pr(R_i = 1 \mid V_i^* = v, W_i = w, S_i = \mathrm{a})$ in the auxiliary sample, then invert the mixture $m_{\mathrm{e}}(w) = \Pr(R_i = 1 \mid W_i = w, S_i = \mathrm{e})$. For $\theta = \Pr(V_i^* = 1)$,

$$\theta(w) = \frac{m_{\mathrm{e}}(w) - p_0(w)}{p_1(w) - p_0(w)}, \qquad \theta = \mathbb{E}\left[\theta(W_i) \mid S_i = \mathrm{e}\right].$$

  (Rambachan et al., 2026)

- **Additional assumptions**: coverage, $\Pr(V_i^* = v \mid S_i = \mathrm{a}) > 0$ and $P_{\mathrm{e}}^W \ll P_{\mathrm{a},v}^W$ for $v = 0, 1$; relevance, $p_1(w) \neq p_0(w)$.

Measurement Stability

- **Assumption** (measurement stability): $R_i \perp\!\!\!\perp S_i \mid (V_i^*, W_i)$.
- Special case: no covariates, $V_i^* \in \{0, 1\}$ and $R_i \in \{0, 1\}$. Measurement stability is then

$$\Pr(R_i = 1 \mid V_i^* = v, S_i = \mathrm{e}) = \Pr(R_i = 1 \mid V_i^* = v, S_i = \mathrm{a}) \quad \text{for } v = 0, 1.$$

- The predictor has the same true positive rate and false positive rate in both samples.
  - In algorithmic fairness, this is equalized odds if $R_i \in \{0, 1\}$ and separation if $R_i \in [0, 1]$ (Hardt et al., 2016; Pleiss et al., 2017; Barocas et al., 2023).
  - In distribution shift, this is "label shift."

Relationship between Stability Assumptions

- Suppose **outcome stability** holds:
  $\Pr(V_i^* = 1 \mid R_i = r, S_i = \mathrm{e}) = \Pr(V_i^* = 1 \mid R_i = r, S_i = \mathrm{a})$.

- Write each side by Bayes' rule and rearrange for the measurement channel:

$$\Pr(R_i = r \mid V_i^* = v, S_i = \mathrm{e}) = \Pr(R_i = r \mid V_i^* = v, S_i = \mathrm{a}) \times \underbrace{\frac{\Pr(V_i^* = v \mid S_i = \mathrm{a})}{\Pr(V_i^* = v \mid S_i = \mathrm{e})} \cdot \frac{\Pr(R_i = r \mid S_i = \mathrm{e})}{\Pr(R_i = r \mid S_i = \mathrm{a})}}_{\text{depends on the base rates}}$$

- Measurement stability asks the two measurement channels to be equal, so it needs that factor to be one.

Relationship between Stability Assumptions

- Suppose instead **measurement stability** holds:
  $\Pr(R_i = r \mid V_i^* = v, S_i = \mathrm{e}) = \Pr(R_i = r \mid V_i^* = v, S_i = \mathrm{a})$.

- By Bayes' rule again, we can re-arrange for the outcome channel:

$$\Pr(V_i^* = 1 \mid R_i = r, S_i = \mathrm{e}) = \Pr(V_i^* = 1 \mid R_i = r, S_i = \mathrm{a}) \times \underbrace{\frac{\Pr(V_i^* = 1 \mid S_i = \mathrm{e})}{\Pr(V_i^* = 1 \mid S_i = \mathrm{a})} \cdot \frac{\Pr(R_i = r \mid S_i = \mathrm{a})}{\Pr(R_i = r \mid S_i = \mathrm{e})}}_{\text{depends on the base rates}}$$

- Outcome stability asks the two outcome channels to be equal, so it also needs that same factor to be one.

Relationship between Stability Assumptions

- Suppose instead **measurement stability** holds:
  $\Pr(R_i = r \mid V_i^* = v, S_i = \mathrm{e}) = \Pr(R_i = r \mid V_i^* = v, S_i = \mathrm{a})$.

- By Bayes' rule again, we can re-arrange for the outcome channel:

$$\Pr(V_i^* = 1 \mid R_i = r, S_i = \mathrm{e}) = \Pr(V_i^* = 1 \mid R_i = r, S_i = \mathrm{a}) \times \underbrace{\frac{\Pr(V_i^* = 1 \mid S_i = \mathrm{e})}{\Pr(V_i^* = 1 \mid S_i = \mathrm{a})} \cdot \frac{\Pr(R_i = r \mid S_i = \mathrm{a})}{\Pr(R_i = r \mid S_i = \mathrm{e})}}_{\text{depends on the base rates}}$$

- But differing base rates are exactly what we expect! The auxiliary sample was collected somewhere else, at some other time.
  - A government forest survey run in another part of Uganda.
  - Mobile phone carrier data from Malawi matched to census year, with the predictor deployed during COVID.

The Two Assumptions Are Incompatible

- **Theorem** (informal): suppose $0 < \Pr(V_i^* = 1 \mid S_i = s) < 1$ for $s = \mathrm{e}, \mathrm{a}$. Then outcome stability and measurement stability can both hold only if
  - (i) the base rates agree, $\Pr(V_i^* = 1 \mid S_i = \mathrm{e}) = \Pr(V_i^* = 1 \mid S_i = \mathrm{a})$; or
  - (ii) the base rates differ and R_i perfectly reveals V_i^*: there is a function h with $V_i^* = h(R_i)$ almost surely.

The Two Assumptions Are Incompatible

- **Theorem** (informal): with interior base rates, outcome stability and measurement stability can both hold only if (i) base rates agree, or (ii) R_i perfectly reveals V_i^*.

- Neither case is reasonable in practice:
  - Equal base rates: the auxiliary sample was collected somewhere else, at some other time — another part of Uganda, a census year before COVID.
  - Perfect revelation: empirical benchmarks show that predictions are far from perfect.

- Incompatibility results are canonical in algorithmic fairness: sufficiency and separation cannot both hold when base rates differ across groups (Kleinberg et al., 2017; Chouldechova, 2017; Pleiss et al., 2017).
  - Stability assumptions are two definitions of what it means for a predictor to behave the "same" way across samples.

The Two Assumptions Are Incompatible

- **Theorem** (informal): with interior base rates, outcome stability and measurement stability can both hold only if (i) base rates agree, or (ii) R_i perfectly reveals V_i^*.
- Neither case is reasonable in practice.
- Incompatibility results are canonical in algorithmic fairness: sufficiency and separation cannot both hold when base rates differ across groups (Hardt et al., 2016; Kleinberg et al., 2017; Chouldechova, 2017; Pleiss et al., 2017).
- **Upshot**: the two stability are alternatives. Researchers must choose and defend the choice.

Using the Wrong Assumption

- The two bridges are alternatives, so the researcher must choose. What does it cost to choose wrong?

- Target the empirical base rate $\theta_{\mathrm{e}} = \Pr(V_i^* = 1 \mid S_i = \mathrm{e})$, and write θ_{a} for its auxiliary counterpart.

- With $0 < \theta_{\mathrm{a}} < 1$, summarize how good the prediction is by its reliability in the auxiliary sample,

$$\kappa = \frac{\mathrm{Var}_{\mathrm{a}}\left\{\mathbb{E}\left[V_i^* \mid R_i, S_i = \mathrm{a}\right]\right\}}{\mathrm{Var}_{\mathrm{a}}\left(V_i^*\right)} \in [0, 1].$$

  - The share of the variation in V_i^* the prediction accounts for.
  - For binary R_i, this is $\mathrm{Corr}_{\mathrm{a}}(V_i^*, R_i)^2$.

Using the Wrong Assumption

- Reliability $\kappa = \mathrm{Var}_{\mathrm{a}}\{\mathbb{E}[V_i^* \mid R_i, S_i = \mathrm{a}]\}/\mathrm{Var}_{\mathrm{a}}(V_i^*) \in [0, 1]$; target θ_{e}.

- **Theorem** (informal): suppose measurement stability holds, but the researcher imputes under outcome stability. Then,

$$\theta_{\mathrm{OS}}^{\dagger} = \mathbb{E}\left[\mathbb{E}\left[V_i^* \mid R_i, S_i = \mathrm{a}\right] \mid S_i = \mathrm{e}\right] = \theta_{\mathrm{a}} + \kappa\left(\theta_{\mathrm{e}} - \theta_{\mathrm{a}}\right).$$

- **Theorem** (informal): outcome stability holds, but the researcher *inverts* under measurement stability. Then, for binary R_i with $\kappa > 0$,

$$\theta_{\mathrm{MS}}^{\dagger} = \theta_{\mathrm{a}} + \frac{\theta_{\mathrm{e}} - \theta_{\mathrm{a}}}{\kappa}.$$

- The choice matters exactly when the samples differ, which is exactly when the auxiliary sample was needed.

Using the Correct Assumption

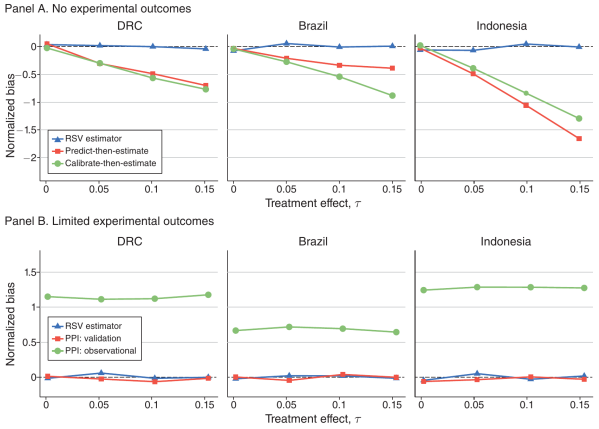
- Alsharif et al. (2026): semi-synthetic exercises illustrating a deforestation experiment in the DRC, Brazil, and Indonesia
  - R_i is prediction from MOSAIKS embeddings: held-out AUC of 0.979–0.998.
  - Measurement stability holds by construction.
  - Distribution shift: a more heavily forested area is targeted for intervention than auxiliary sample.

- Consider two data environments:
  1. *No experimental outcomes*: 1,000 experimental units, 1,000 auxiliary-sample units.
  2. *Limited experimental outcomes*: outcomes additionally retained for a 250-unit validation subsample.

- Example: What happens when measurement stability holds?

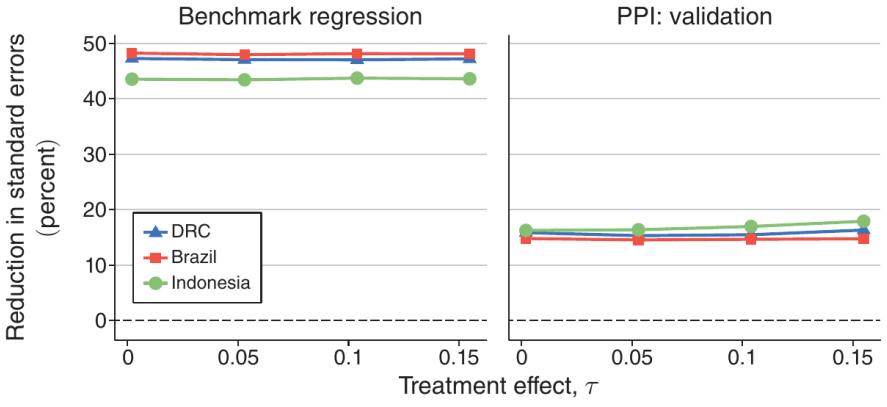
Using the Correct Assumption

- **Identification**: exploiting measurement stability is approximately unbiased. Alternative identifying strategies yield severely biased estimates.



Using the Correct Assumption

- **Precision**: exploiting the auxiliary sample through measurement stability substantially tightens standard errors.



Using the Correct Assumption

- Example: what happens when measurement stability holds?
  - (i) Exploiting measurement stability is approximately unbiased. Alternative identifying strategies yield severely biased estimates.
  - (ii) Exploiting the auxiliary sample substantially tightens standard errors.

- Rambachan et al. (2026) provide two further empirical illustrations to an anti-poverty program and crop-burning experiment in which measurement stability holds.

- There are large returns to exploiting the right stability assumption: both in terms of bias and precision. How should researchers choose?

How Should Researchers Choose?

- There are large returns to exploiting the right stability assumption: both in terms of bias and precision. How should researchers choose?

- Ask what generated the data: does the outcome produce the generated-measurement, or does the generated-measurement merely predict the outcome?

- Post-outcome measurement: the realized outcome generates the measurement. For example, a burned field reflects differently, a new building appears in the imagery.
  - The mapping is physical, so it has a reason to hold elsewhere: start from measurement stability (Rambachan et al., 2026).
  - Threats are engineering: sensor, season, product vintage, processing pipeline changing across samples.

How Should Researchers Choose?

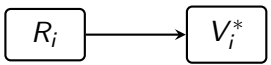
- Post-outcome measurement: the outcome generates the trace, the mapping is physical → start from measurement stability.

- Pre-outcome prediction: the signal is correlated / merely forecasts the outcome — phone metadata, baseline characteristics, short-run surrogates.
  - There is no physical mechanism to appeal to, and the link is a correlation: start from outcome stability.
  - Threats: treatment, institutions, or behavior changing how the signal predicts the outcome.

How Should Researchers Choose?

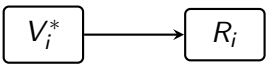
- Another way to see the choice: which way does the causal arrow run? (Athey et al., 2026; Rambachan et al., 2026).

outcome stability



Pre-outcome
surrogates

measurement stability



Post-outcome
remote sensing

Testable Implications and Sensitivity Analyses

- Measurement stability implies the empirical distribution of the prediction is a mixture:

$$\Pr(R_i = r \mid W_i = w, S_i = \mathrm{e}) = \sum_v \Pr(R_i = r \mid V_i^* = v, W_i = w, S_i = \mathrm{a}) \times \Pr(V_i^* = v \mid W_i = w, S_i = \mathrm{e}).$$

  Everything but the weights are observed, so the model has testable restrictions (Rambachan et al., 2026):

- By contrast, outcome stability restricts $\Pr(V_i^* = 1 \mid R_i, W_i, S_i = \mathrm{e})$ but V_i^* is never observed when $S_i = \mathrm{e}$ (Athey et al., 2026).

- Replace exact stability assumptions with a bounded violation, and report the identified set or the breakdown violation that would overturn the conclusion (Rambachan et al., 2026; Fan et al., 2026).

Stability as Model Behavior

- Ultimately, outcome stability and measurement stability are assumptions about how an ML/AI measurement behaves when it is moved to another place or period.
- So we need an empirical answer: evaluation and benchmarking.
- Existing benchmarks evaluate remote-sensing predictors for their predictive performance (Rolf et al., 2021; Proctor et al., 2025), but that is not the only metric of interest:
  - A predictor can be shed inaccurate and still deliver an unbiased effect, or be quite accurate and fail to deliver an unbiased estimate.
  - Predictive performance does not inform which stability assumption is plausible.
- The answer may vary depending on outcome, by modality, and by the kind of transfer — across space, time, institutions.

AI and Economics

- AI and machine learning are accelerating discovery across the sciences. What might that look like in economics?

- These lectures offered one answer to this question: ML/AI let us attack the lamppost problem, which limits economics in two ways:
  1. We collect limited measurements from the settings we study.
  2. Collecting high-quality measurements is costly, which limits the questions we can study at scale.

AI as a Tool for Measurement

- **1** We collect limited structured measurements → ML/AI provide the opportunity to expand our aperture.
  - Use ML and AI on unstructured data to discover what to measure and how it relates to the quantities we care about — hypothesis generation.
  - The "prescientific" step can now be assisted, and sometimes automated. This is a nascent and rapidly growing space.

- **2** Measurement is costly → scale the measurements we would already defend.
  - *With a random validation sample*: collect ground truth on a subsample and debias. Credible measurement is a missing data problem.
  - *Without a random validation sample*: many proxies for the same latent construct, or measurements transferred from another sample. Credible measurement rests on assumptions about frontier model behavior.

Measurement in the Age of AI

- Causal inference $\Rightarrow$ unlocked new ways to see the world, and opened questions we could not study before.

- And we built an econometric toolkit to go with it. Every method paired with an assumption you defend:
  - Difference-in-differences $\Leftrightarrow$ defend parallel trends.
  - Instrumental variables $\Leftrightarrow$ defend exclusion, independence, monotonicity.

- ML/AI as tools for measurement $\Rightarrow$ unlock new ways to see the world and open up new questions we could not study before.

- We are building new toolkits that enable creative empirical research using ML/AI + new data modalities while preserving rigorous downstream statistical conclusions.

Appendix Slides

References I

Adler, E Scott and John Wilkerson, *Congressional Bills Project, NSF 00880066 and 00880061*, `http://congressionalbills.org/download.html` (accessed July 5, 2024), 2020.

Aiken, Emily, Esther Rolf, and Joshua Blumenstock, "Fairness and Representation in Satellite-Based Poverty Maps: Evidence of Urban-Rural Disparities and Their Impacts on Downstream Policy," in "Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence" 2023, pp. 5888–5896.

__ **, Joshua E. Blumenstock, Sveta Milusheva, and M. Merritt Smith**, "Predicting Well-Being with Mobile Phone Data: Evidence from Four Countries," *AEA Papers and Proceedings*, 2026, *116*, 178–183.

__ **, Suzanne Bellue, Joshua E. Blumenstock, Dean Karlan, and Christopher Udry**, "Estimating Impact with Surveys versus Digital Traces: Evidence from Randomized Cash Transfers in Togo," *Journal of Development Economics*, 2025, *175*, 103477.

Alix-García, Jennifer M. and Daniel L. Millimet, "Remotely Incorrect? Accounting for Nonclassical Measurement Error in Satellite Data on Deforestation," *Journal of the Association of Environmental and Resource Economists*, 2023, *10* (5), 1335–1367.

Alsharif, Haya, Ashesh Rambachan, Rahul Singh, and Davide Viviano, "Causal Inference with Satellite Imagery: A Comparison of Methods for Forest Conservation Data," *AEA Papers and Proceedings*, 2026, *116*, 87–91.

References II

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

Angrist, Joshua D. and Jörn-Steffen Pischke, "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics," *Journal of Economic Perspectives*, 2010, *24* (2), 3–30.

Asirvatham, Hemanth, Elliott Mokski, and Andrei Shleifer, "GPT as a Measurement Tool," NBER Working Paper 34834, National Bureau of Economic Research 2026.

Athey, Susan, Raj Chetty, Guido W. Imbens, and Hyunseung Kang, "The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely," *The Review of Economic Studies*, 2026, *93* (4), 2284–2312.

Atreja, Shubham, Joshua Ashkinaze, Lingyao Li, Julia Mendelsohn, and Libby Hemphill, "What's in a Prompt? A Large-Scale Experiment to Assess the Impact of Prompt Design on the Compliance and Accuracy of LLM-Generated Text Annotations," in "Proceedings of the Nineteenth International AAAI Conference on Web and Social Media" 2025, pp. 122–145.

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

References III

Barocas, Solon, Moritz Hardt, and Arvind Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*, Cambridge, MA: MIT Press, 2023.

Battaglia, Laura, Timothy Christensen, Stephen Hansen, and Szymon Sacher, "Inference for Regression with Variables Generated by AI or Machine Learning," Cowles Foundation Discussion Paper 2421, Cowles Foundation for Research in Economics, Yale University 2025.

Blumenstock, Joshua E., Gabriel Cadamuro, and Robert On, "Predicting Poverty and Wealth from Mobile Phone Metadata," *Science*, 2015, *350* (6264), 1073–1076.

Bommasani, Rishi, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang, "Picking on the Same Person: Does Algorithmic Monoculture Lead to Outcome Homogenization?," in "Advances in Neural Information Processing Systems," Vol. 35 2022.

Bonhomme, Stéphane, Koen Jochmans, and Jean-Marc Robin, "Non-parametric Estimation of Finite Mixtures from Repeated Measurements," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2016, *78* (1), 211–229.

Bound, John and Alan B. Krueger, "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?," *Journal of Labor Economics*, 1991, *9* (1), 1–24.

References IV

Card, Dallas, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky, "Computational Analysis of 140 Years of US Political Speeches Reveals More Positive but Increasingly Polarized Framing of Immigration," *Proceedings of the National Academy of Sciences*, 2022, *119* (31), e2120510119.

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

Chen, Xiaohong, Ashesh Rambachan, and Elie Tamer, "Partial Identification from LLM Prompts," 2026. June 25, 2026 version.

__ , Han Hong, and Alessandro Tarozzi, "Semiparametric Efficiency in GMM Models with Auxiliary Data," *The Annals of Statistics*, 2008, *36* (2), 808–843.

__ , __ , and Denis Nekipelov, "Nonlinear Models of Measurement Errors," *Journal of Economic Literature*, 2011, *49* (4), 901–937.

__ , __ , and Elie Tamer, "Measurement Error Models with Auxiliary Data," *The Review of Economic Studies*, 2005, *72* (2), 343–366.

Chouldechova, Alexandra, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," *Big Data*, 2017, *5* (2), 153–163.

References V

Dawid, A. P. and A. M. Skene, "Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm," *Journal of the Royal Statistical Society Series C: Applied Statistics*, 1979, *28* (1), 20–28.

D'Haultfœuille, Xavier, Christophe Gaillac, and Arnaud Maurel, "Linear Regressions with Combined Data," NBER Working Paper 34507, National Bureau of Economic Research 2025.

__ , __ , and __ , "Partially Linear Models under Data Combination," *The Review of Economic Studies*, 2025, *92* (1), 238–267.

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.

Fan, Yanqin, Carlos A. Manzanares, Hyeonseok Park, and Yuan Qi, "A Sensitivity Analysis of the Surrogate Index Approach for Estimating Long-Term Treatment Effects," 2026.

__ , Robert Sherman, and Matthew Shum, "Identifying Treatment Effects under Data Combination," *Econometrica*, 2014, *82* (2), 811–822.

fei Lee, Lung and Jungsywan H. Sepanski, "Estimation of Linear and Nonlinear Errors-in-Variables Models Using Validation Data," *Journal of the American Statistical Association*, 1995, *90* (429), 130–140.

References VI

Hansen, Matthew C., Peter V. Potapov, Rebecca Moore, Matt Hancher, Svetlana A. Turubanova, Alexandra Tyukavina, David Thau, Stephen V. Stehman, Scott J. Goetz, Thomas R. Loveland, Anil Kommareddy, Alexey Egorov, Louise Chini, Christopher O. Justice, and John R. G. Townshend, "High-Resolution Global Maps of 21st-Century Forest Cover Change," *Science*, 2013, *342* (6160), 850–853.

Hardt, Moritz, Eric Price, and Nati Srebro, "Equality of Opportunity in Supervised Learning," in "Advances in Neural Information Processing Systems 29" 2016.

Hu, Yingyao and Susanne M. Schennach, "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 2008, *76* (1), 195–216.

Huang, Luna Yue, Solomon M. Hsiang, and Marco Gonzalez-Navarro, "Using Satellite Imagery and Deep Learning to Evaluate the Impact of Anti-Poverty Programs," NBER Working Paper 29105, National Bureau of Economic Research 2021.

Jack, B. Kelsey, Seema Jayachandran, Namrata Kala, and Rohini Pande, "Money (Not) to Burn: Payments for Ecosystem Services to Reduce Crop Residue Burning," *American Economic Review: Insights*, 2025, *7* (1), 39–55.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Alampay Davis, David B. Lobell, and Stefano Ermon, "Combining Satellite Imagery and Machine Learning to Predict Poverty," *Science*, 2016, *353* (6301), 790–794.

References VII

Kallus, Nathan and Xiaojie Mao, "On the Role of Surrogates in the Efficient Estimation of Treatment Effects with Limited Outcome Data," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2025, *87* (2), 480–509.

Kim, Elliot Myunghoon, Avi Garg, Kenny Peng, and Nikhil Garg, "Correlated Errors in Large Language Models," in "Proceedings of the 42nd International Conference on Machine Learning," Vol. 267 of *Proceedings of Machine Learning Research* PMLR 2025, pp. 30038–30066.

Kleinberg, Jon and Manish Raghavan, "Algorithmic Monoculture and Social Welfare," *Proceedings of the National Academy of Sciences*, 2021, *118* (22), e2018340118.

_ , Sendhil Mullainathan, and Manish Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," in "8th Innovations in Theoretical Computer Science Conference," Vol. 67 of *Leibniz International Proceedings in Informatics* Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik 2017, pp. 43:1–43:23.

Kluger, Dan M., Kerri Lu, Tijana Zrnic, Sherrie Wang, and Stephen Bates, "Prediction-Powered Inference with Imputed Covariates and Nonuniform Sampling," 2025.

Lipton, Zachary C., Yu-Xiang Wang, and Alexander J. Smola, "Detecting and Correcting for Label Shift with Black Box Predictors," in "Proceedings of the 35th International Conference on Machine Learning," Vol. 80 of *Proceedings of Machine Learning Research* PMLR 2018, pp. 3122–3130.

References VIII

Lu, Kerri, Dan M. Kluger, Stephen Bates, and Sherrie Wang, "Regression Coefficient Estimation from Remote Sensing Maps," *Remote Sensing of Environment*, 2025, *330*, 114949.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan, "Large Language Models: An Applied Econometric Framework," *Annual Review of Economics*, 2026.

Ottonello, Pablo, Wenting Song, and Sebastian Sotelo, "An Anatomy of Firms' Political Speech," NBER Working Paper 32923, National Bureau of Economic Research 2024.

Pelletier, Johanne, Mira Korb, Solomon Alemu, Manex B. Yonis, Travis J. Lybbert, and Matthieu Stigler, "Causal Inference with Predicted Outcomes: Correcting prediction error bias in satellite-based impact evaluation," *Journal of Development Economics*, 2026, *179*, 103655.

Pleiss, Geoff, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger, "On Fairness and Calibration," in "Advances in Neural Information Processing Systems 30" 2017, pp. 5680–5689.

Proctor, Jonathan, Tamma Carleton, and Sandy Sum, "Parameter Recovery Using Remotely Sensed Variables," NBER Working Paper 30861, National Bureau of Economic Research 2023.

_ , _ , Trinetta Chong, Taryn Fransen, Simon Greenhill, Jessica Katz, Hikari Murayama, Luke Sherman, Jeanette Tseng, Hannah Druckenmiller, and Solomon Hsiang, "What Can Satellite Imagery and Machine Learning Measure?," NBER Working Paper 34315, National Bureau of Economic Research 2025.

References IX

Rambachan, Ashesh, Rahul Singh, and Davide Viviano, "Program Evaluation with Remotely Sensed Outcomes," 2026.

Ratledge, Nathan, Gabriel Cadamuro, Brandon de la Cuesta, Matthieu Stigler, and Marshall Burke, "Using Satellite Imagery and Machine Learning to Estimate the Livelihood Impact of Electricity Access," NBER Working Paper 29237, National Bureau of Economic Research 2021.

Rolf, Esther, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang, "A Generalizable and Accessible Approach to Machine Learning with Global Satellite Imagery," *Nature Communications*, 2021, *12*, 4392.

Schennach, Susanne M., "Mismeasured and Unobserved Variables," in Steven N. Durlauf, Lars Peter Hansen, James J. Heckman, and Rosa L. Matzkin, eds., *Handbook of Econometrics*, Vol. 7A, Elsevier, 2020, pp. 487–565.

__ , "Measurement Systems," cemmap Working Paper CWP12/21, Centre for Microdata Methods and Practice 2021.

Sclar, Melanie, Yejin Choi, Yulia Tsvetkov, and Alane Suhr, "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying about Prompt Formatting," in "International Conference on Learning Representations" 2024.

References X

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," Vol. 235 of *Proceedings of Machine Learning Research* 2024, pp. 48919–48937.

Yin, Michelle, Hoa Vu, and Claudia Persico, "How (un)Stable Are LLM Occupational Exposure Scores? Evidence from Multi-Model Replication," NBER Working Paper 35110, National Bureau of Economic Research 2026.