# Finding Regions of Heterogeneity in Decision-Making via Expected Conditional Covariance

Justin Lim*[1], Christina X Ji*[1], Michael Oberst*[1], Saul Blecker[2], Leora Horwitz[2], David Sontag[1]

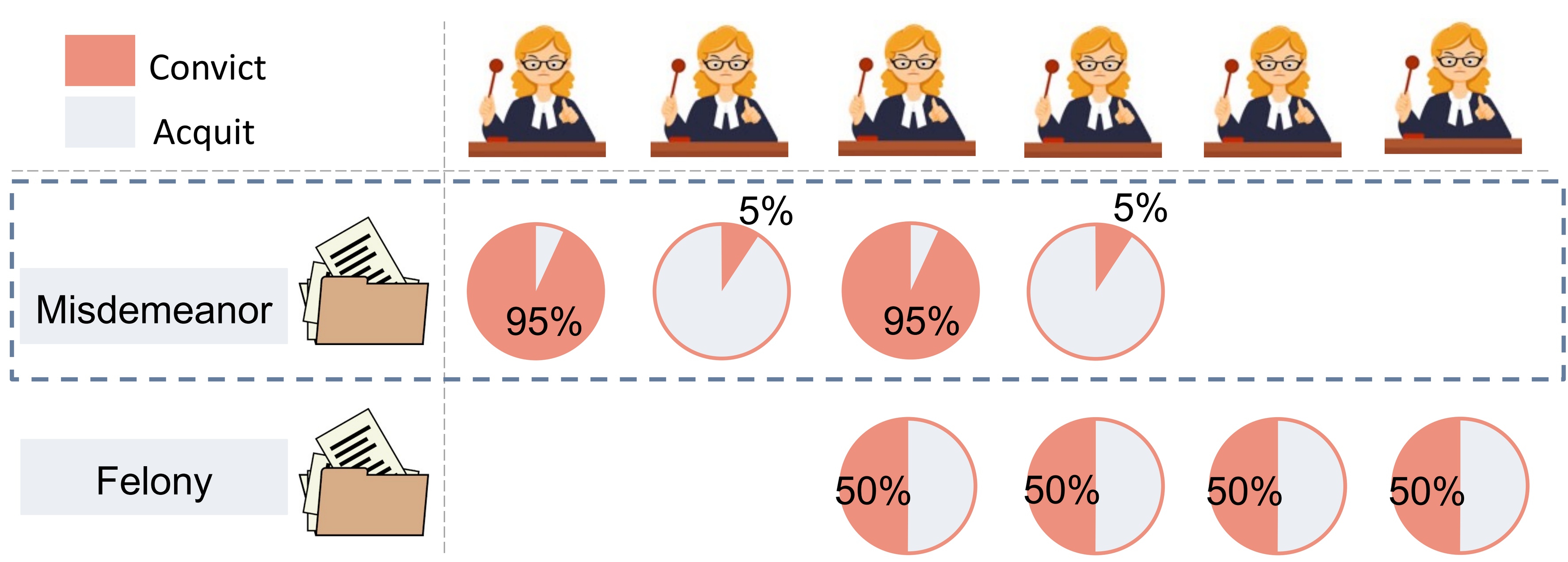
*equal contribution, [1]MIT CSAIL and IMES, [2]NYU Langone

## Goal

Characterize **the types of decisions** where the identity of the **decision-maker makes a substantial difference** in the ultimate decision

Individuals often make **different decisions** when faced with the **same context**, e.g.,
- **Judges** may vary in leniency towards certain offenses
- **Doctors** may vary in preference for how to start treatment for certain types of patients

**Illustrative Example:** Judges vary in leniency towards misdemeanor cases




**Challenge #1: What if judges simply see different types of misdemeanor cases?**
Need to adjust for potential confounding factors

**Challenge #2: Very few samples per judge**
Hard to reliably estimate the bias for any individual judge

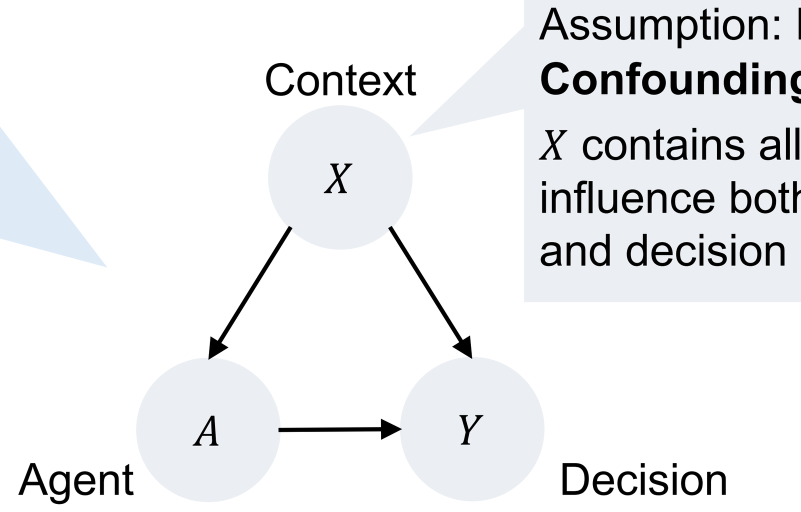
## Estimating heterogeneity as causal contrast

Compare **decisions of each agent** with **decisions of peers who see similar cases**

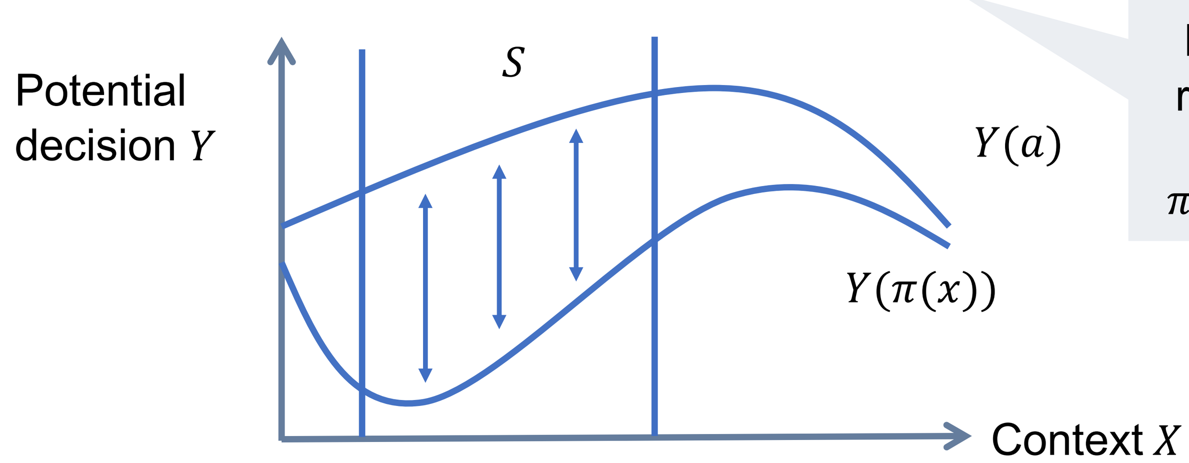
**Not Assumed: Overlap (Positivity)**
$P(A = a \mid X = x)$ does not have to be positive for all $x, a$.

We only assume **each context is seen by some agents**, not all agents

Assumption: **No Unmeasured Confounding**
$X$ contains all variables that influence both agent assignment and decision




**Conditional Relative Agent Bias:** $E[Y(a) - Y(\pi(x)) \mid A = a, X \in S]$




Expected decision of a random agent assigned proportionally to $\pi(x) \coloneqq p(A = a \mid X = x)$

## Causal objective for heterogeneity

Causal objective captures **aggregate bias across binary grouping $G$ of agents over region $S$ without agent-specific models**, an advantage when data for individual agents is scarce

We **construct an objective** using the **conditional relative agent bias** that was defined for estimating heterogeneity as causal contrast

Weighted sum over biases of agents $G(a) = 1$.

Conditional Relative Agent Bias

$$Q(S, G) \coloneqq \sum_{a; G(a)=1} p(A = a \mid X \in S) \cdot E[Y(a) - Y(\pi) \mid A = a, X \in S]$$

We can then **compute the region and grouping** that optimize this objective

- For a given region $S$, this objective is maximized by choosing $G(a) = 1$ if the bias of agent $a$ is non-negative on $S$.
- Find optimal region $S$ subject to minimum size constraint:

$$\max_S Q(S, G^*(S)) = \max_S \max_G Q(S, G) \quad \text{s.t.} \quad P(S) \geq \beta$$

**Theorem 1:** $Q(S, G)$ can be identified as $E[Cov(Y, G|X) \mid X \in S]$

## Algorithm for finding regions of heterogeneity

We propose an iterative algorithm that **optimizes the objective above**

Fit a model of the average treatment decision across all agents $f(x) = E[Y|X = x]$

Assign agents to groups using average residuals: $G(a) = 1$ if $E[Y - f(x)|A = a, X \in S] \geq 0$

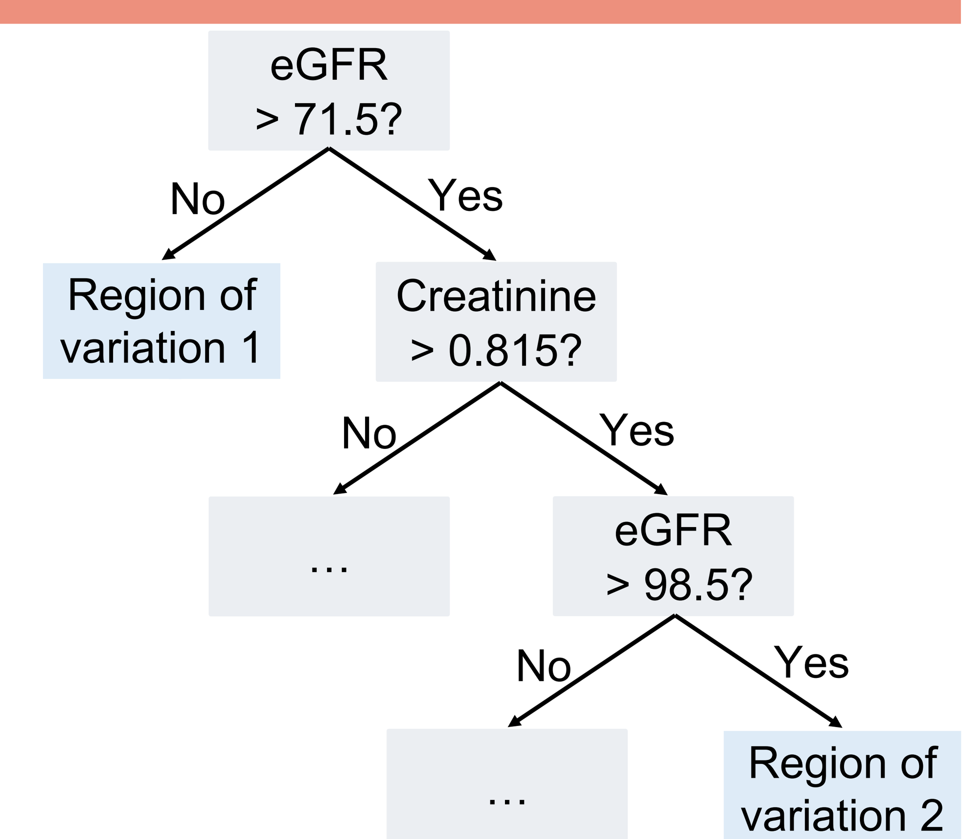
Identify region as samples in upper quantile of predictions from fitting $h(x)$ to $(y - f(x))\, G(a)$

## Example: Initial Treatment for Type 2 Diabetes

**Region** discovered by our algorithm **aligns with clinical knowledge**

Set-up:
- Predict metformin (typical recommendation from American Diabetes Association) vs other common first-line treatments[1,2]
- 3,576 patients and 176 group practices (agents)

Conclusions:
- Region 1: guidelines lacking where metformin is contraindicated[3,4,5]
- Region 2: no contraindications. Identifying why some doctors prescribe other medications can **help standardize practice**




## Semi-synthetic experiment

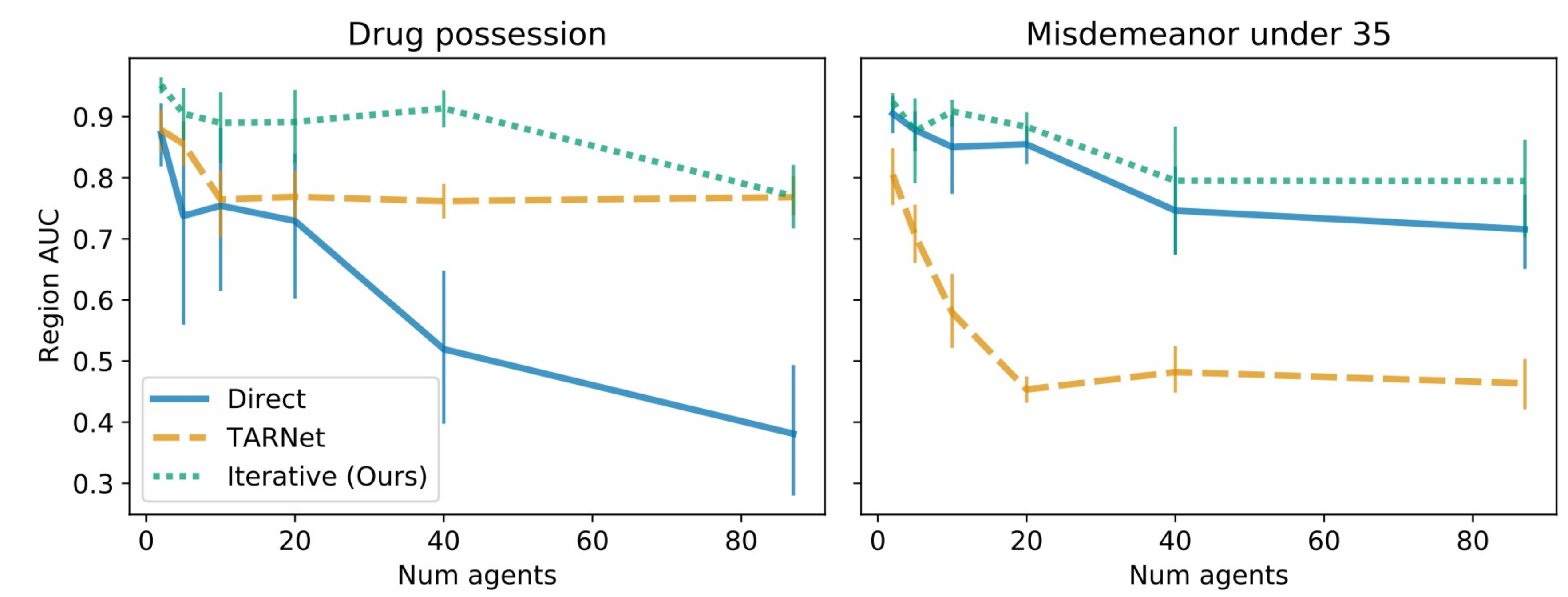
**Our algorithm outperforms baselines** in scenarios with **many agents** and **few samples per agent**

Dataset: Predictions of recidivism using COMPAS dataset collected from Mechanical Turk agents based on 5 risk factors[6,7]

Semi-synthetic data with **ground truth regions of heterogeneity**:
- Region 1: Drug possession charges
- Region 2: Misdemeanor charges for individuals under 35

4,550 samples are divided among 2, 5, 10, 20, 40, and 87 synthetic agents who are randomly assigned to one of two policies:
- Base policy: Learned on real agent predictions
- Alternative policy: Add **systematic preference** towards recidivism **in region**

Results:
- Region AUC evaluates classification with respect to true region




Baselines:
- **Direct:** Estimate $E[Y|A, X]$ and $E[Y \mid X]$. Identify region where agent is most informative, i.e. model with agents most outperforms model without agents.
- **TARNet[8]:** Predict $E[Y|A, X]$ using shared representation with separate prediction heads per agent. Identify region with largest variation in counterfactual outcomes across agents, i.e. where $\mathrm{Var}_A[E[Y|A, X]]$ is largest.

## Conclusion

**Finding regions of variation** can help **improve decision-making guidelines, increase fairness,** and **drive better outcomes**

- Heterogeneity in decision-making can be measured as a **causal contrast**
- Regions of heterogeneity can be found using an **iterative algorithm**

References

[1]American Diabetes Association. 2010. Standards of medical care in diabetes—2010. Diabetes care 33, Supplement 1 (2010), S11–S61.

[2]George Hripcsak, et al. 2016. Characterizing treatment pathways at scale using the OHDSI network. Proceedings of the National Academy of Sciences 113, 27 (2016), 7329–7336.

[3]AA Tahrani, GI Varughese, JH Scarpello, and FWF Hanna. 2007. Metformin, heart failure, and lactic acidosis: is metformin absolutely contraindicated? Bmj 335, 7618 (2007), 508–512.

[4]Kidney Disease: Improving Global Outcomes (KDIGO) CKD-MBDWork Group et al. 2009. KDIGO clinical practice guideline for the diagnosis, evaluation, prevention, and treatment of Chronic Kidney Disease-Mineral and Bone Disorder (CKD-MBD). Kidney international. Supplement 113 (2009), S1–S130.

[5]Guideline Development Group, Henk Bilo, Luis Coentrão, Cécile Couchoud, Adrian Covic, Johan De Sutter, Christiane Drechsler, Luigi Gnudi, David Goldsmith, James Heaf, et al. 2015. Clinical practice guideline on management of patients with diabetes and chronic kidney disease stage 3b or higher (eGFR< 45 mL/min). Nephrology Dialysis Transplantation 30, suppl_2 (2015), ii1–ii142.

[6]Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. Science advances 4, 1 (Jan. 2018), eaao5580.

[7]Zhiyuan (Jerry) Lin, Jongbin Jung, Sharad Goel, and Jennifer Skeem. 2020. The Limits of Human Predictions of Recidivism. Science Advances 6, 7 (2020).

[8]Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In International Conference on Machine Learning. PMLR, 3076–3085.