Identifying Bias in Data Using Two-Distribution Hypothesis Tests

William Yik, Limnanthes Serafini, Timothy Lindsey, and George D. Montañez

Introduction

- **Bias** in machine learning training data can be reflected in the model's output.
- **Statistical hypothesis testing** can help identify this bias by probabilistically ruling out proposed explanations.
- Since **training data** is often the source of bias, it should be analyzed directly, even before training the model.

Methodology

For an event or dataset x, we define the **kardis test statistic** as

$$\kappa(x) := r\frac{p(x)}{\nu(x)}$$

- **Complexity** $p(x)$: the probability of x under some distribution P, the null hypothesis
- **Specificity** $\nu(x)$: the observation's conformity to a pattern
- Normalizing constant r

Functional Specified Complexity Kardis:

$$\kappa(x) := |\mathcal{X}|\,(1+\ln|\mathcal{X}|)\frac{p(x)}{F_g(x)^{-1}}$$

- $r = |\mathcal{X}|\,(1+\ln|\mathcal{X}|)$, $\nu(x) = F_g(x)^{-1}$
- $\mathcal{X}$: space of possible events
- $F_g(x)$: proportion of events **more extreme** than x

The Level-α Property for $\kappa(x)$:

$$\Pr(\kappa(x) \leq \alpha) \leq \alpha$$

Allows us to **reject** hypotheses for a given significance level α.

Probability Boosting Factor:

$$s \geq \frac{\alpha\nu(x)}{rp(x)}$$

- Any **plausible explanation** of the data must boost the probability of observing x, $p(x)$, by at least a factor of s.
- The **closest plausible distribution/explanation** is the probability distribution which boosts $p(x)$ by at least s and is ”closest” to the null hypothesis by some metric.

$$q(x) \geq s \cdot p(x)$$

- **Interpretable** hypothesis test results

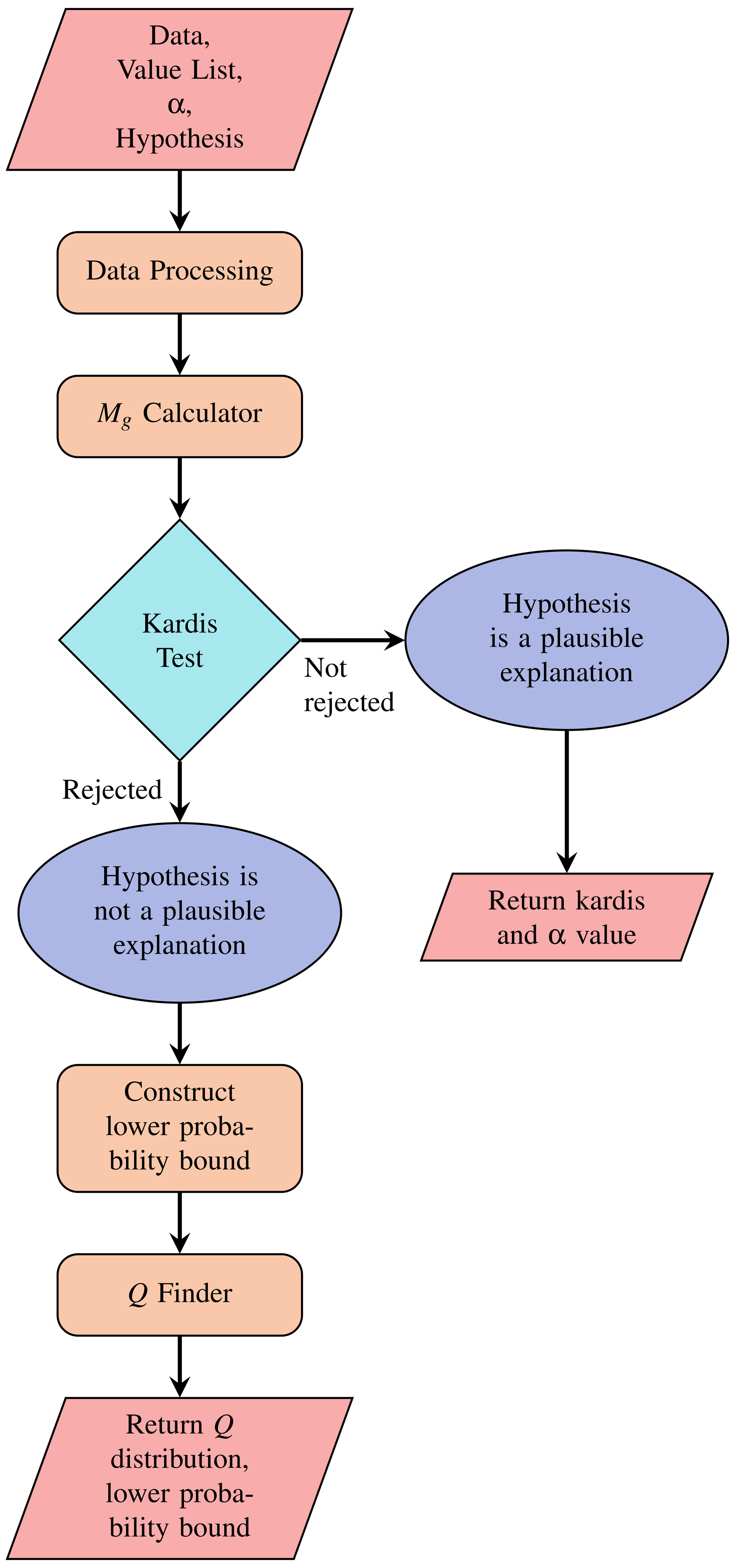
Experimental Setup

UCI Adult Dataset

- Contains information such as race, gender, and education level for over 40,000 US adults
- Target is the binary income label of "$\leq$ 50K" or "$>$ 50K"
- Investigating for **bias against women**
- **Null hypothesis**: the same process that generated the male income labels could plausibly explain the female income labels
  - Approximately 70% males labeled "$\leq$ 50K" and 30% labeled "$>$ 50K"

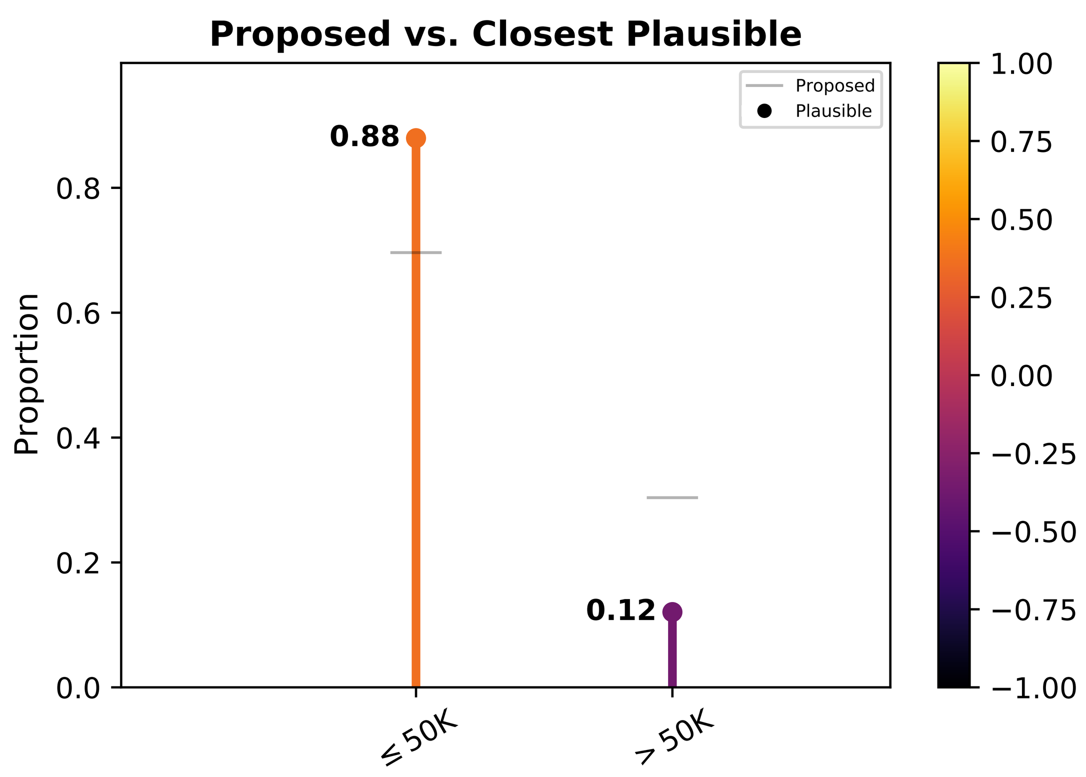
COMPAS Dataset

- Correctional Offender Management Profiling for Alternative Sanctions
- Assigns each person a low, medium, or high risk of recidivism
- Investigating for **bias against African Americans**
- **Null hypothesis**: the same process which generated the Caucasian risk score distribution could plausibly explain the African American risk score distribution
  - Approximately 11% of Caucasians labeled "high risk", 22% labeled "medium risk", and 67% labeled "low risk"



Results

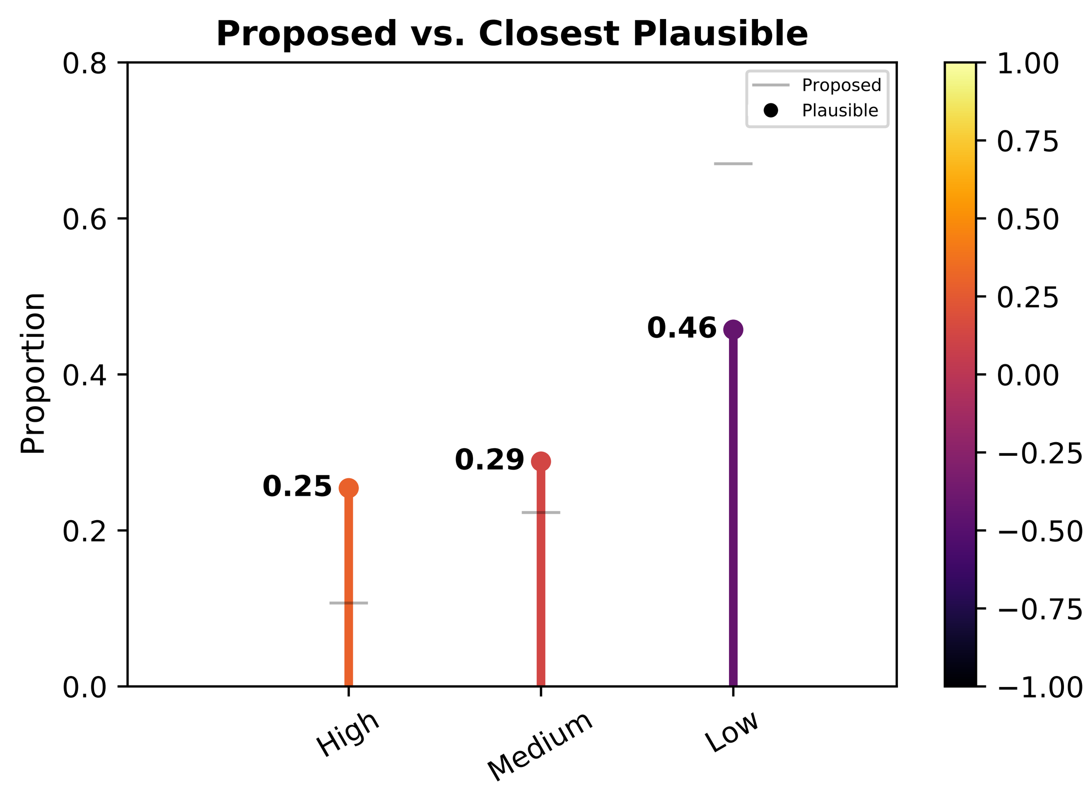
UCI Adult Dataset



- At a significance level of $\alpha = 0.05$, we **reject** the null hypothesis
- $\kappa(x) = 2.47 \times 10^{-755}$, $s \geq 2.02 \times 10^{753}$
- Confirms known biases in dataset and adds **extra degree of interpretability**

COMPAS Dataset

- At a significance level of $\alpha = 0.05$, we **reject** the null hypothesis
- $\kappa(x) = 8.99 \times 10^{-392}$, $s \geq 5.57 \times 10^{389}$
- Any plausible explanation must **significantly boost** the proportion of African Americans assigned a "high" recidivism risk score



Significance

- **Isolating** sources of bias
- Rejecting **whole sets** of hypotheses instead of just one
- **No assumptions** concerning sample or test statistic distribution
- **Fast** enough to run on a **consumer-grade** personal laptop

Conclusion

- Statistical hypothesis test for identifying bias in training data using the **specified complexity kardis**
- Test constructs **probabilistic lower bound** to rule out whole sets of hypotheses.
- Test returns a **closest plausible distribution**, providing a unique degree of clarity

Acknowledgments

This research was supported in part by the National Science Foundation under Grant No. 1950885. Any opinions, findings, or conclusions are those of the authors alone, and do not necessarily reflect the views of the National Science Foundation. The authors thank Xanda Schofield, Erin Talvitie, Melissa O'Neill, and Lucas Bang for helpful feedback and suggestions for this work, and thank their friends, family, and the members of the AMISTAD Lab for their continued support.

AMISTAD

Contact us at `https://www.cs.hmc.edu/~montanez/amistad.html`