It’s COMPASlicated

The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks

Michelle Bao
baom@stanford.edu

Angela Zhou
az434@cornell.edu

Samantha Zottola
szottola@prainc.com

Brian Brubach
bb100@wellesley.edu

Sarah Desmarais
sdesmarais@prainc.com

Aaron Horowitz
ahorowitz@aclu.org

Kristian Lum
kristianl@twitter.com

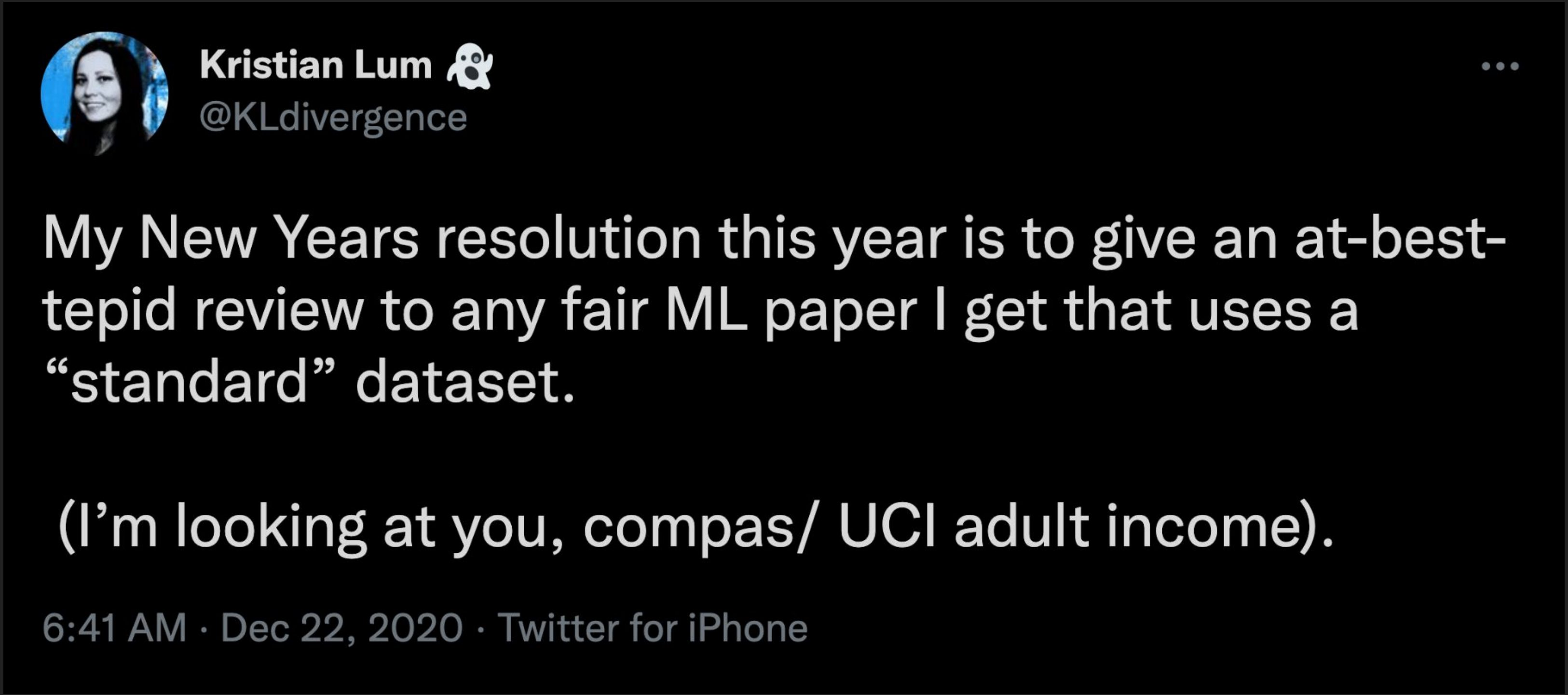
Suresh Venkatasubramanian
suresh_venkatasubramanian@brown.edu

Machine Bias

There’s software used across the country to predict future criminals. And it’s biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016



Kristian Lum 👻
@KLdivergence

My New Years resolution this year is to give an at-best-tepid review to any fair ML paper I get that uses a “standard” dataset.

(I’m looking at you, compas/ UCI adult income).

6:41 AM · Dec 22, 2020 · Twitter for iPhone

Data Biases and Errors in Pretrial RAIs

$(X, A, Y) \sim D$

	Y	A	X
Variables	Rearrest during pretrial period; failure to appear (FTA) in court	Race; sex	Criminal history summaries (past arrests; past convictions)
Statistical or Societal Bias	Label bias; measurement bias	Observed instead of self-reported; instability of discrete categories	Racial bias; encoding choices in categories

Limits of Algorithmic Fairness in Real-World Application

- Fairness goes beyond prediction and error
- Judges (and others) apply algorithm results
  - Their interpretation matters
- Fair predictions may be applied unfairly

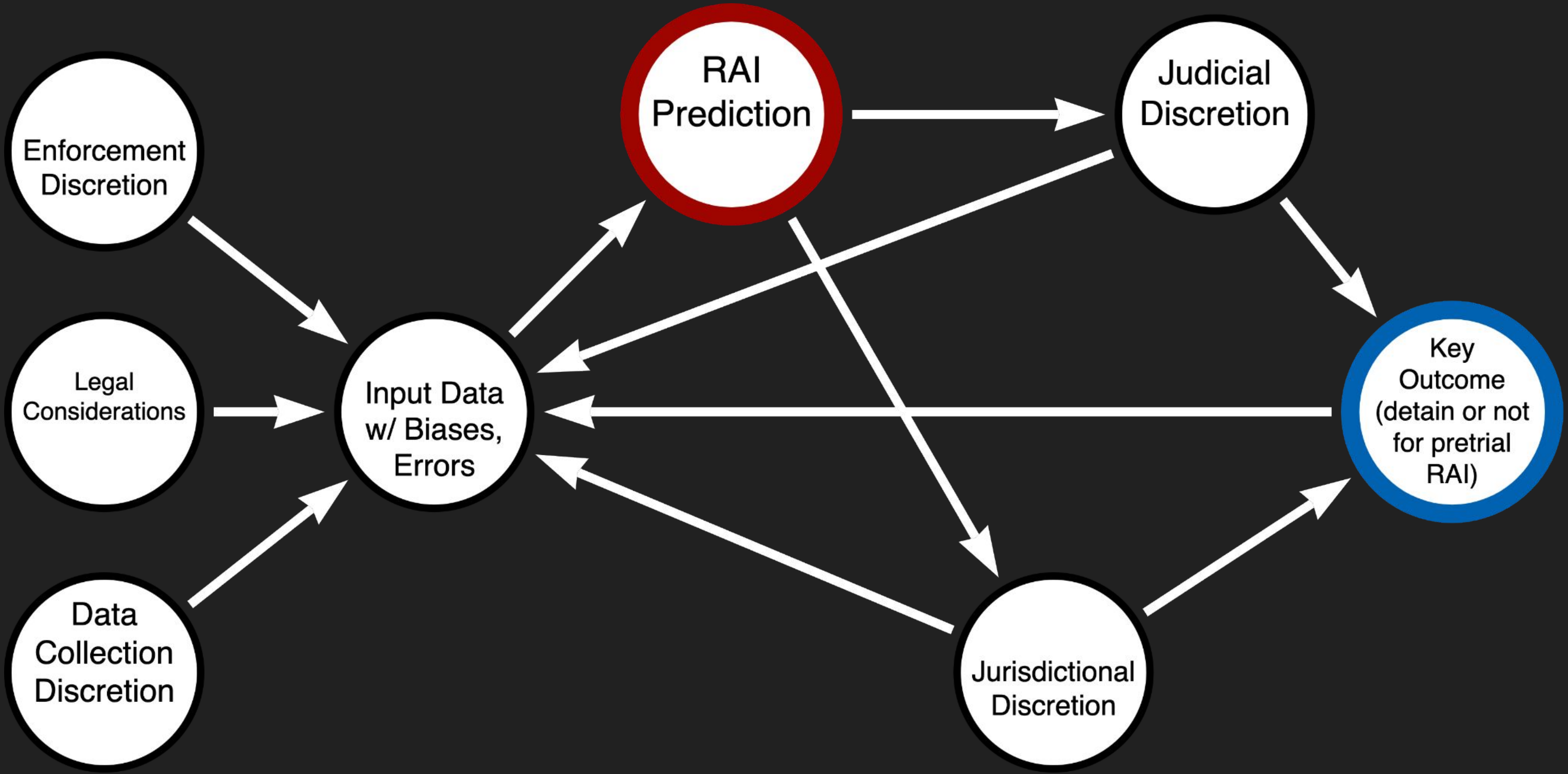
Normative Values Embedded in Use of RAI Datasets

- Research on RAIs has broad ethical implications

Established Norms in Other Fields

- Forensic Psychology & Criminology
  - Produced many existing RAIs
  - Train people who use RAIs
- Data collection practices
- Standards for assessing RAIs
- Guidelines for RAI use in practice

RAI Outcome Network


RAI Prediction
Judicial Discretion
Enforcement Discretion
Legal Considerations
Input Data w/ Biases, Errors
Key Outcome (detain or not for pretrial RAI)
Data Collection Discretion
Jurisdictional Discretion

Mismatch Between AI Fairness Practices and Criminal Justice Research

- Focus on SOTA optimization of *methods* in benchmarking experiments decontexualizes the data
  - "Granted bail" as a positive label

- Conference publication workflow does not incentivize meaningful engagement
  - Norms around performing experiments on benchmark datasets to demonstrate real-world applicability

Call To Arms

When using CJ-related datasets:

Things to avoid…

- Generically illustrating or benchmarking a new fairness algorithm/measure
- Making broad conclusions about practices in CJ solely from the use of CJ datasets

Call To Arms

When using CJ-related datasets:

Things to be careful about…

- Using the datasets thoughtfully, ideally with a deep understanding of the details and in collaboration with CJ researchers
- Accurately describing the limitations, implications, or potential uses of the work

Call To Arms

When using CJ-related datasets:

Things that might be helpful:

- Building datasheets and model cards for CJ datasets/RAIs
- Identifying implicit assumptions in fairness algorithms
- Investigating where problems in data and interpretation could make algorithms fail
- Investigating the use of different metrics and their problems

Thank you!