COLUMBIA UNIVERSITY
Data Science Institute

Trustworthy AI:
New Properties for New Complex Systems

Jeannette M. Wing
Avanessians Director of Data Science and Professor of Computer Science, Columbia University
Adjunct Professor of Computer Science, Carnegie Mellon University

Formal Methods at Scale
Washington, DC
September 25, 2019

Trustworthy Computing

- Trustworthy =
  - + Reliability
    - Does it do the right thing?
  - + Security
    - How vulnerable is it to attack?
  - + Privacy
    - Does it protect a person's identity and data?
  - + Usability
    - Can a human use it easily?

- Computing = hardware + software + people

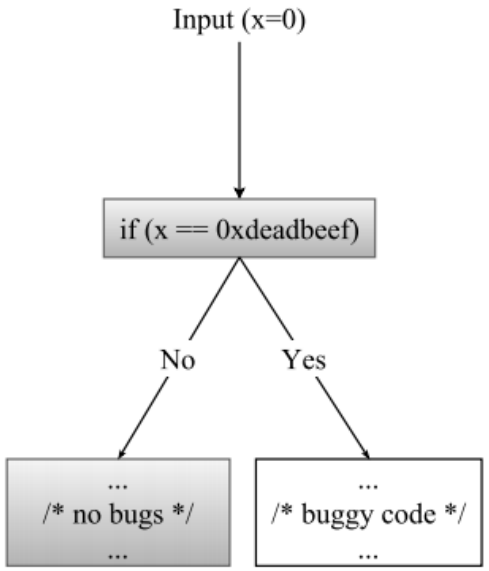
Trustworthy AI

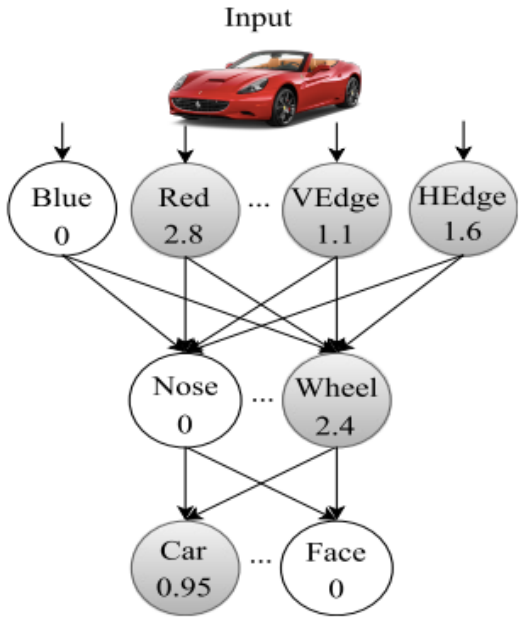
- Trustworthy =
  - + Reliability/Safety
    - Does it do the right thing?
  - + Security
    - How vulnerable is it to attack?
  - + Privacy
    - Does it protect a person's identity and data?
  - + Usability
    - Can a human use it easily?
- AI = data + algorithm + context

FAT(E) → FASTER

- + Fairness
  - Are the model outcomes unbiased?
- + Accountable
  - Who or what is responsible for the outcome?
- + Transparent (Explainable)
  - How was the outcome produced?
- + Ethical
  - Was the data collected in an ethical manner?
  - Will the outcome be used in an ethical manner?
- + Robustness
  - How sensitive is the outcome to a change in the input?

DeepXplore: Testing Deep Learning Systems





Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana, "Deep Xplore: Automated Whitebox Testing of Deep Learning Systems, *Proceedings of the 26th ACM Symposium on Operating Systems Principles*, October 2017, Best Paper Award.

DeepXplore

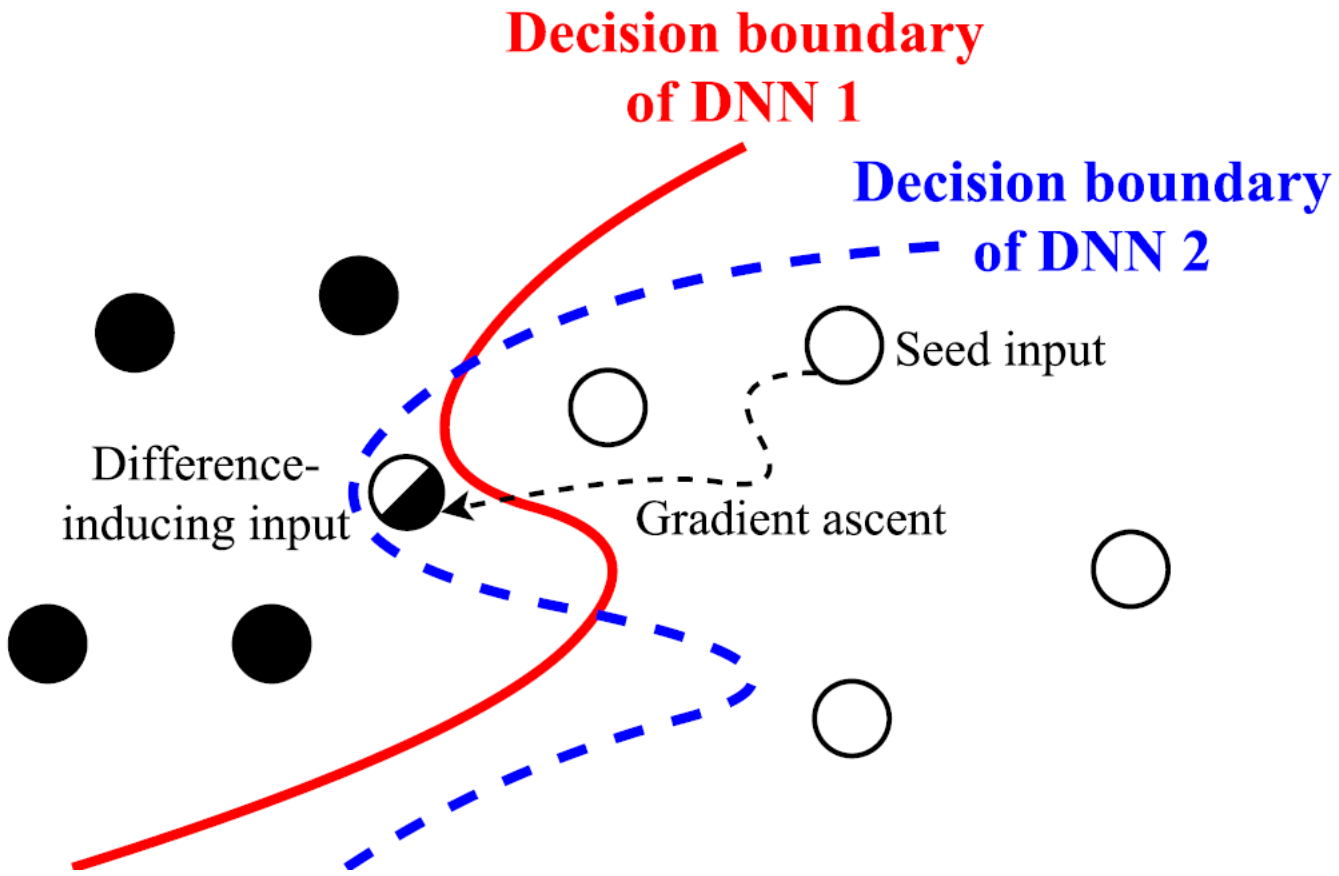
Seed,
No accident

Darker,
Accident

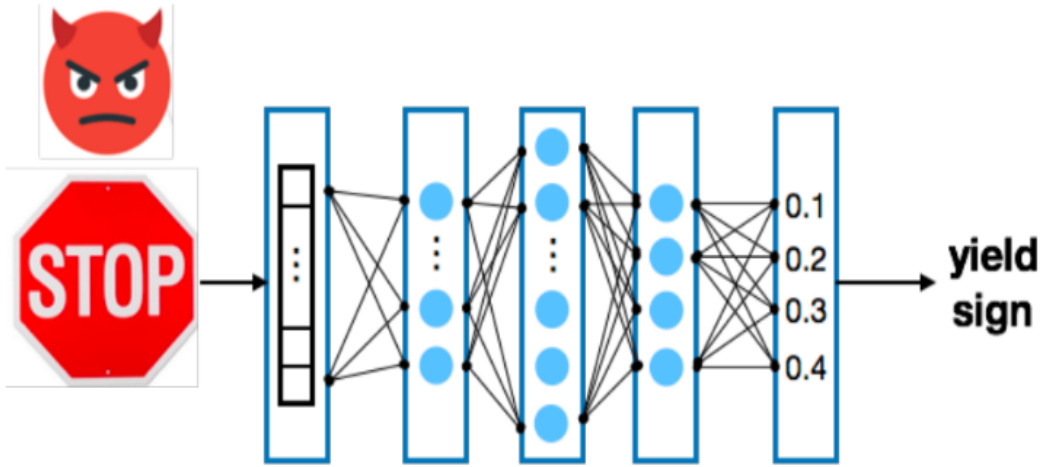
- Efficiently and systematically tests DNNS of hundreds of thousands of neurons without labeled data (only needs unlabeled seeds)
- Key ideas: neuron coverage (akin to code coverage), differential testing, and domain-specific constraints for focusing on realistic inputs
- Testing as a joint optimization problem (maximize both number of differences and neuron coverage)
- Found 1000s of fatal errors in 15 state-of-the-art DNNs for ImageNet, self-driving cars, and PDF/Android malware

https://github.com/peikexin9/deepxplore

DNNs are differentiable. Using gradient ascent to solve the optimization problem. Here eventually from the seed input we find input values that will cause the two DNNs to differ in their output.

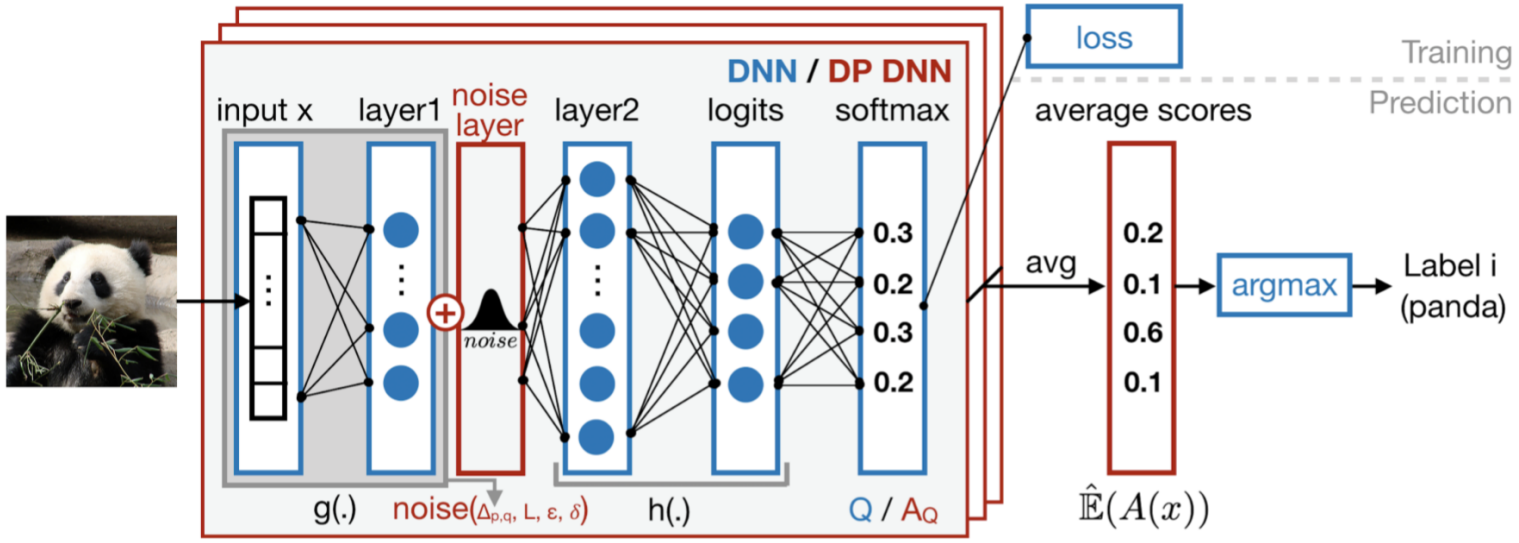


DP and Machine Learning: PixelDP

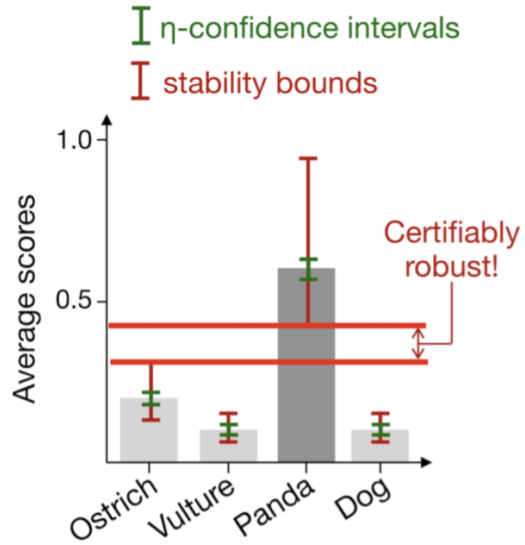


Mathias Lecuyer, Baggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana, “Certified Robustness to Adversarial Examples with Differential Privacy, arXiv:1802.03471v2, IEEE Security and Privacy (‘‘Oakland’’) 2019.

1. Add a noise layer a la Differential Privacy



(a) PixelDP DNN Architecture



(b) Robustness Test Example

2. Provable guarantee from DP says classifier is robust to some degree of input perturbations.

Trustworthy Computing and Formal Methods

$E, M \models P$

M: program (code), protocol, model of concurrent system, distributed system, hardware
$\models$: logics and tools, e.g., model checking, theorem proving ($\vdash$), SMT solvers
P: correctness properties (safety and liveness)

E: environment is often implicit

Trustworthy AI and Formal Methods

$$E, M \models P$$

M: machine-learned model, …, program (code)

$\models$: probabilistic logics and tools

P : probabilistic, stochastic

E: stochastic process or distribution that generates the inputs on which M's outputs need to be verified;

Concretely, think of E as datasets/data distributions used for building M, but both **uncertainty** and **bias** are typically inherent to such datasets/data distributions

Technical Challenges and Opportunities

$$E, M \models P$$

- M and P are inherently—structurally and semantically—different from a progam and a correctness property about program behavior
- As much work goes into modeling the environment E (e.g., input distribution, probabilistic graphical model, or stochastic process) as constructing M itself
- “Correctness” properties P, such as fairness and robustness, have inherent ambiguity or stochasticity germane to the intended uses of M
- Statistical nature of M, P, and E means that AI verification is more amenable to methods over continuous domains (e.g., interval analysis) rather than exact solver based approaches (e.g., SAT)

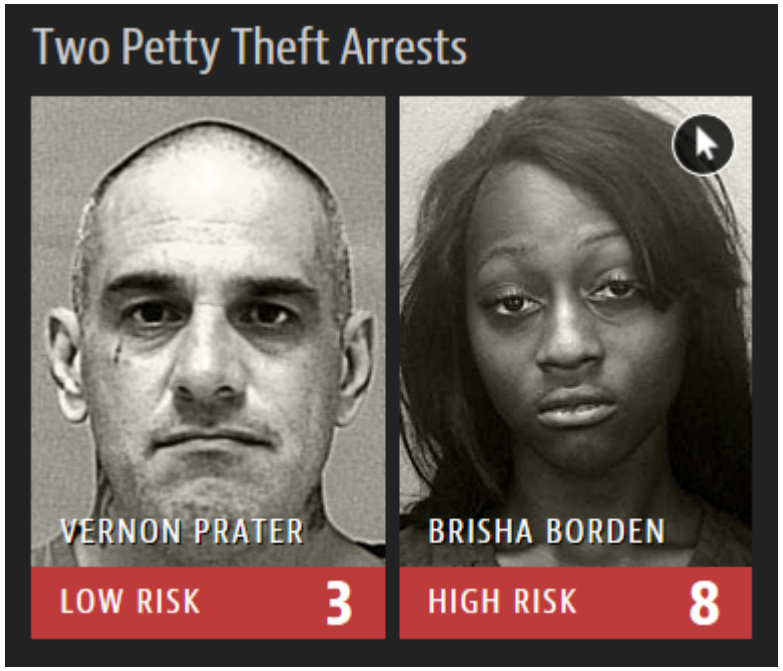
Formal Methods Needs and Opportunities

$$E, M \models P$$

- Need new specification languages (logics and models) for E, M, and P
- Need new verification techniques for $\models$
- Need to interpret verification failures ("counterexamples") for fixing E, M, and/or P
- Need to fit specification and verification steps in AI/ML workflow
- Need to relate back to application and end user, i.e., what M is being used for:
  Can one better trust M, if it passes a verification check? Why?
  What more needs to be done?

COMPAS Data

Proprietary algorithms widely used by judges to help determine risk of re-offense are almost twice as likely to mistakenly flag black defendants than white defendants.



Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, "Machine Bias," *ProPublica*, May 23, 2016.

Impossibility result

A risk score could either be equally predictive or equally wrong for all races—but not both.

Northepointe ("statistical parity")

- Calibration

ProPublica ("equal odds")

- Balance for the positive class
- Balance for the negative class

Average score for whites who reoffend =
Fraction of whites who reoffend

and similarly for blacks

Average score assigned to whites who reoffend =
Average score assigned to blacks who reoffend

and similarly for those who do not reoffend

Jon Kleinberg, Sendhil Mullainathan, Manish Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores,," Proceedings of Innovations in Theoretical Computer Science (ITCS), 2017.

Formalizing P (for two different notions of fairness)

P: probability distribution (population)
f: real-valued function
We assume we can measure a "protected" attribute A in $\mathscr{A}$, a qualification attribute Y in $\mathscr{Y}$, and other attributes $\mathbf{X}$ in $\mathscr{X}$ that are inputs to f.

- Statistical parity

$$|\mathbb{E}_P(f(X) \mid A = a) - \mathbb{E}_P(f(X) \mid A = a')| \leq \epsilon$$

for all a and a'.

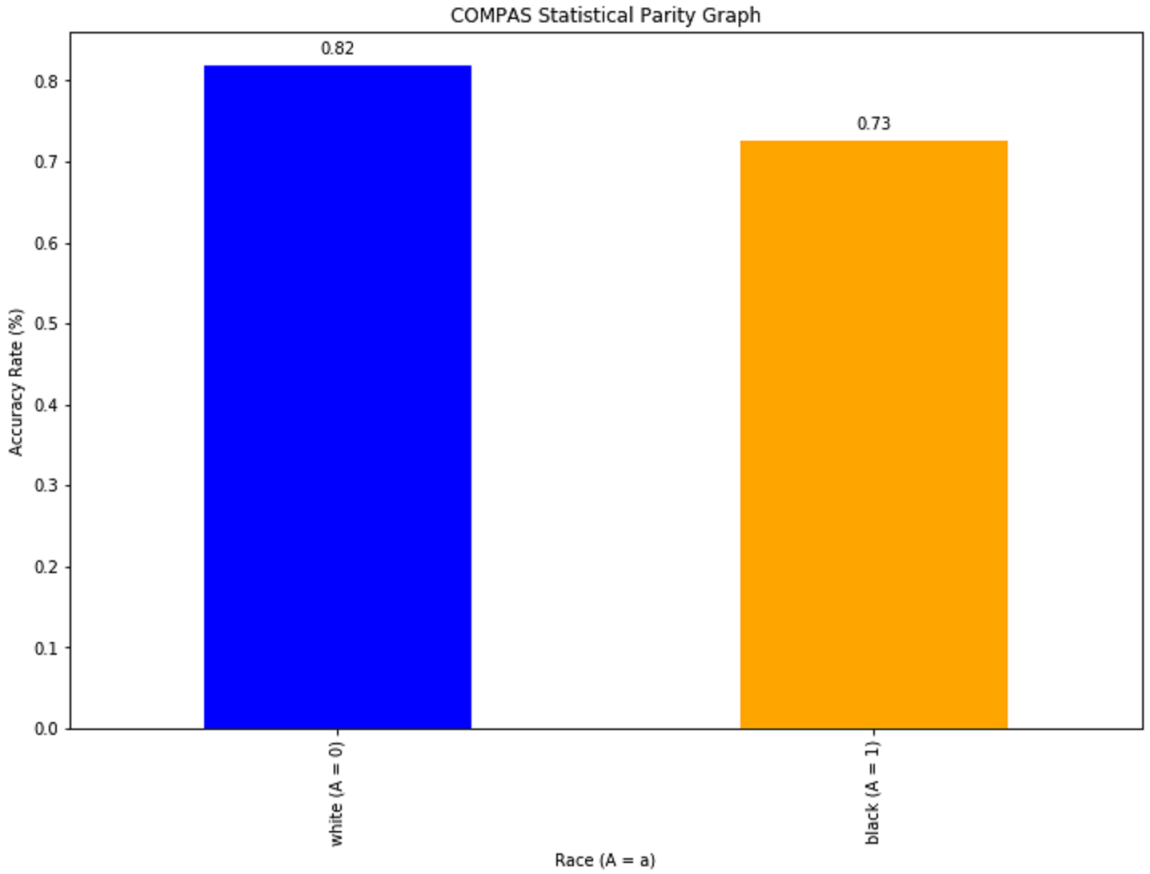
- Equal odds

$$|\mathbb{E}_P(f(X) \mid Y = y, A = a) - \mathbb{E}_P(f(X) \mid Y = y, A = a')| \leq \epsilon$$

for all y, a, and a'.

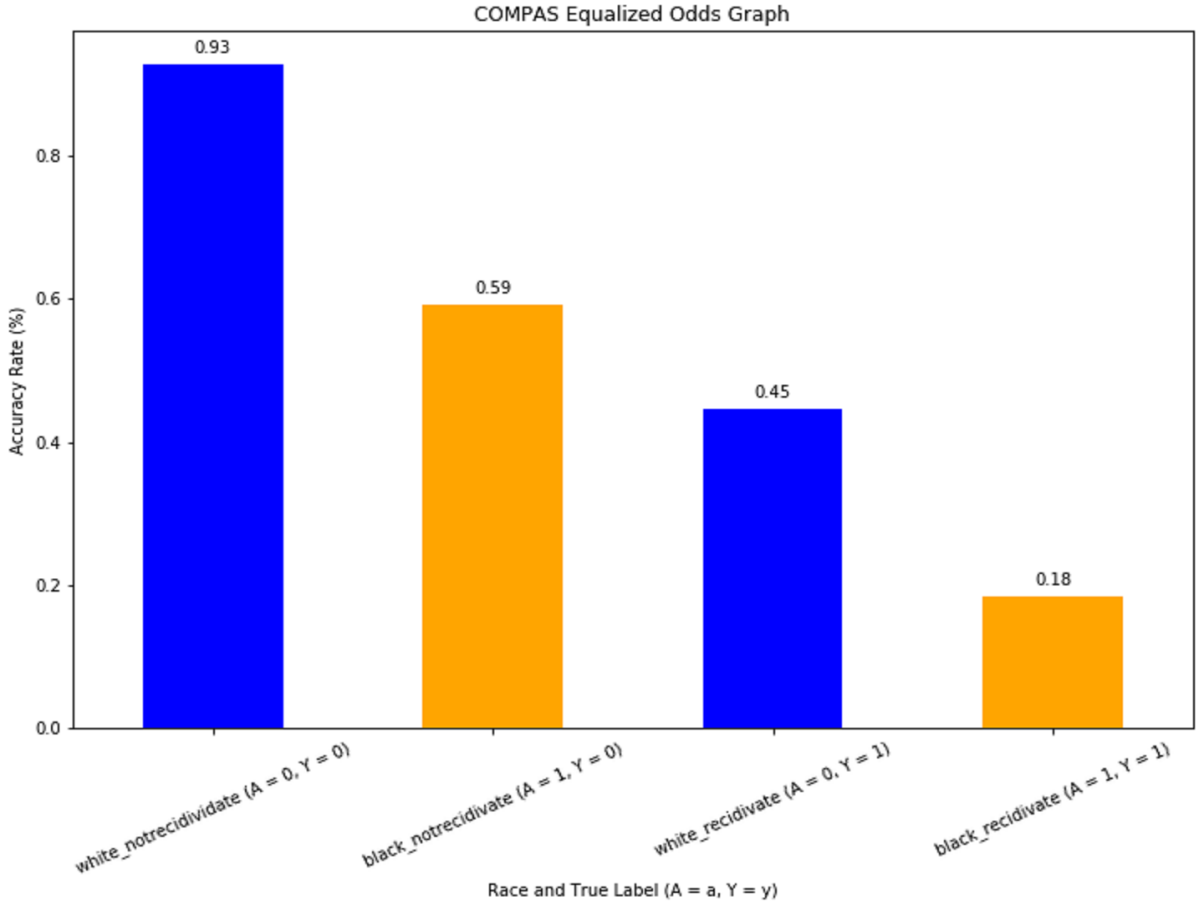
COMPAS Data

Northpointe: Statistical Parity



A = white A = black

ProPublica: Equalized Odds



A = white A = black A = white A = black

Y = did not recidivate Y = did recidivate

Trustworthy AI meets Formal Methods

$E, M \models P$

COLUMBIA UNIVERSITY
Data Science Institute

Thank You