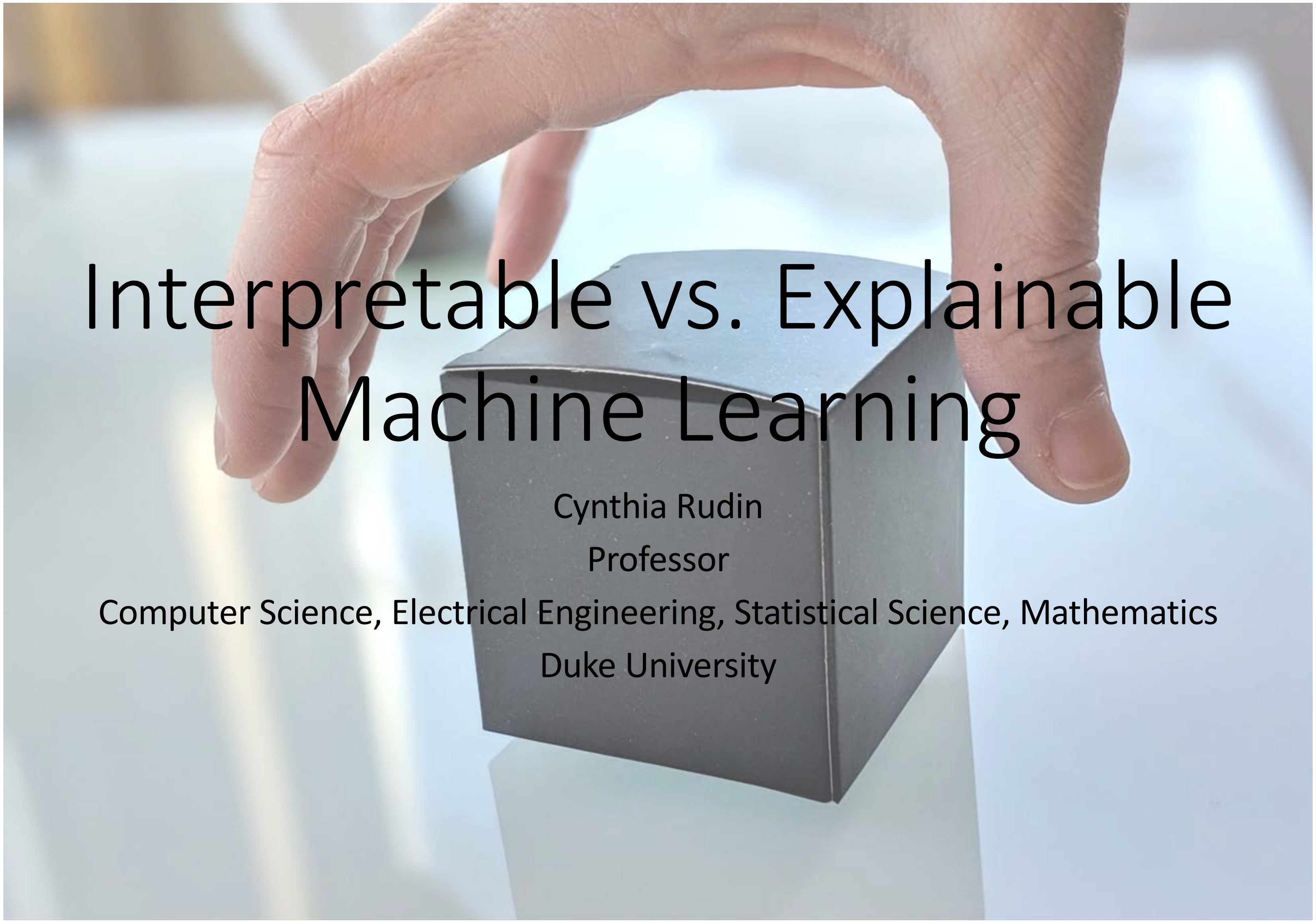


# Interpretable vs. Explainable Machine Learning

Cynthia Rudin
Professor
Computer Science, Electrical Engineering, Statistical Science, Mathematics
Duke University

A black box predictive model is a formula that is either too complicated to understand or proprietary.

What happens when you use a black box?

The New York Times

OP-ED CONTRIBUTOR

# When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

Glenn Rodriguez

# How bad is Sacramento's air, exactly? Google results appear at odds with reality, some say

BY MICHAEL MCGOUGH
AUGUST 07, 2018 09:26 AM, UPDATED AUGUST 07, 2018 09:26 AM


Smoke is affecting air quality all over California. Here's what it looks like at the Carr Fire, north of Redding, on July 31, 2018.
BY PAUL KITAGAKI JR.

Where did Breezometer go wrong?

THE WALL STREET JOURNAL.
English Edition ▼ | October 27, 2019 | Print Edition | Video

BUSINESS | HEALTH CARE | HEALTH

# Researchers Find Racial Bias in Hospital Algorithm

Healthier white patients were ranked the same as sicker black patients, according to study published in the journal Science

*By Melanie Evans and Anna Wilde Mathews*

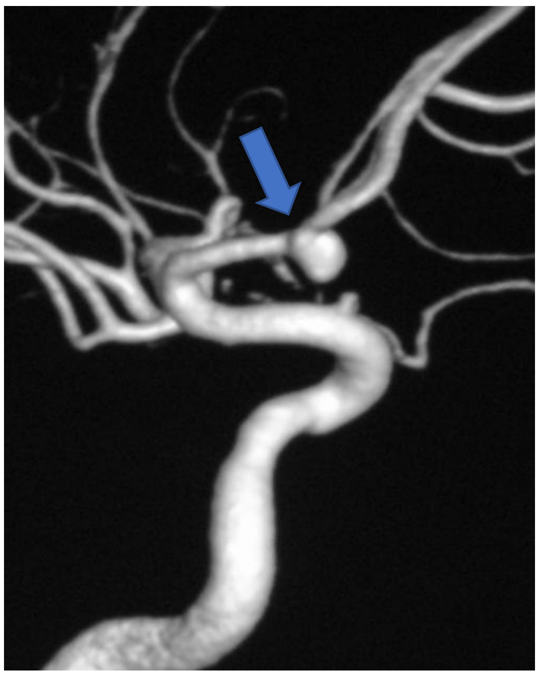
Updated Oct. 25, 2019 8:39 am ET

Black patients were less likely than white patients to get extra medical help, despite being sicker, when an algorithm used by a large hospital chose who got the additional attention, according to a new study underscoring the risks as technology gains a foothold in medicine.

And this is the tip of the iceberg…

- An interpretable machine learning model obeys a domain-specific set of constraints.
- My technical definition: An interpretable machine learning model is constrained in model form so that it is either useful to someone, or obeys structural knowledge of the domain, such as monotonicity, causality, structural (generative) constraints, additivity, or physical constraints that come from domain knowledge.
- There's a spectrum.

# Preventing Brain Damage in Critically Ill Patients



CT-angiography, Anterior Communicating Saccular Aneurysm

Head CT without contrast showing Subarachnoid Hemorrhage

- Seizure are common (20%)
- Seizure → Brain Damage
- Need EEG to detect seizures

Need to use EEG data to predict seizures to determine EEG duration

EEG is expensive and limited: 24hrs of monitoring is $1600-$4000

- 2HELPS2B was not created by doctors
- It is a ML model
- It is just as accurate as black box models.
- Doctors can decide themselves whether to trust it
- Doctors can calibrate the score with information not in the database
- Score can be explained to non-physicians

## 2HELPS2B

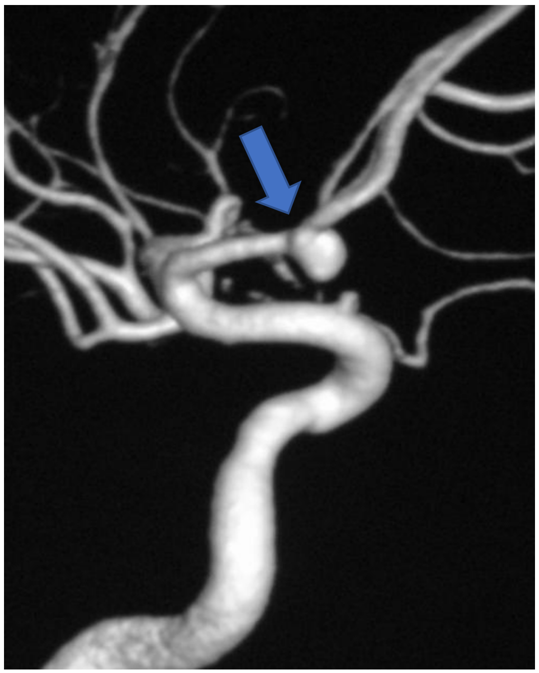
| | | | |
|---|---|---|---|
| 1. | Any cEEG Pattern with Frequency **2 H**z | 1 point | ··· |
| 2. | **E**pileptiform Discharges | 1 point | + ··· |
| 3. | Patterns include [**L**PD, LRDA, BIPD] | 1 point | + ··· |
| 4. | **P**atterns Superimposed with Fast or Sharp Activity | 1 point | + ··· |
| 5. | Prior **S**eizure | 1 point | + ··· |
| 6. | **B**rief Rhythmic Discharges | **2** points | + ··· |
| | | **SCORE** | = ··· |

| **SCORE** | 0 | 1 | 2 | 3 | 4 | 5 | 6+ |
|---|---|---|---|---|---|---|---|
| **RISK** | <5% | 11.9% | 26.9% | 50.0% | 73.1% | 88.1% | 95.3% |

| Variable |
| --- |
| PDR |
| BRDs |
| Unreactive background |
| Prior Sz |
| GRDA |
| LRDA |
| GPDs |
| LPDs |
| BIPDs |
| Infection |
| Inflammation |
| Neoplasm |
| ICH |
| Metabolic encephalopathy |
| Stroke |
| SAH |
| SDH |
| TBI |
| Hypoxic/ischemic |
| IVH |
| Hydrocephalus |
| Discharges |
| Frequency (>2Hz)[c] |

There are many variables to choose from.

# Preventing Brain Damage in Critically Ill Patients



CT-angiography, Anterior Communicating Saccular Aneurysm

Head CT without contrast showing Subarachnoid Hemorrhage

1-Hour Screening EEG

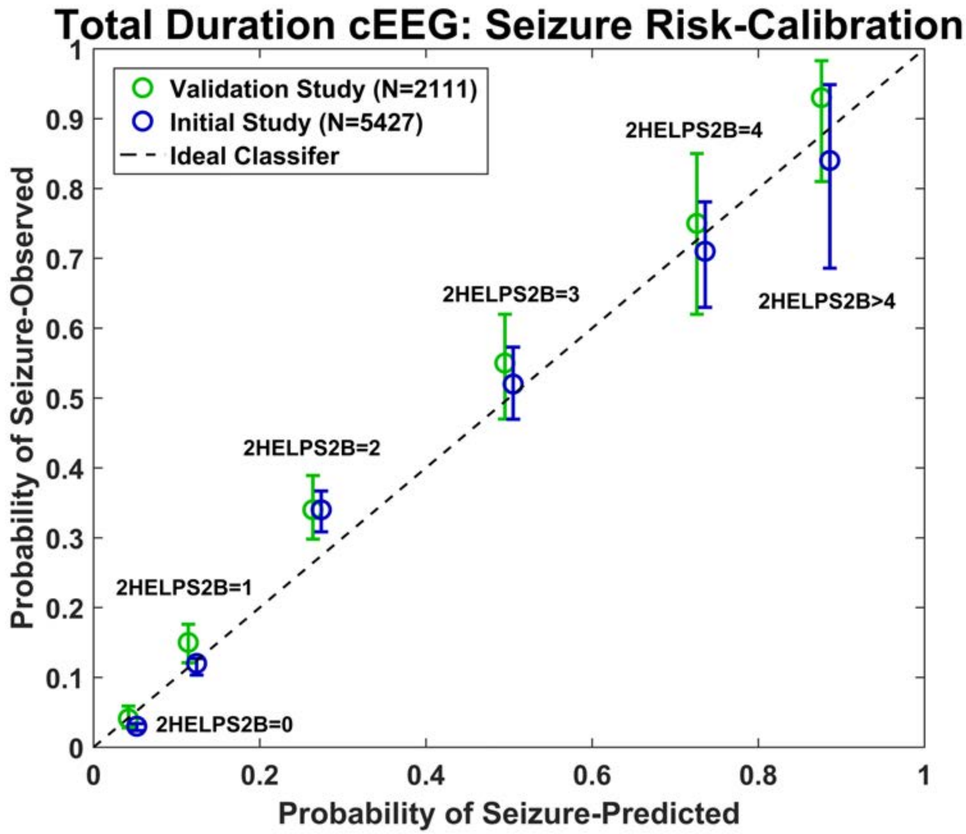
2HELPS2B=3 (high-risk)

- Placed on Continuous EEG for >72H
- Start on preventative medications

# So far…

- 2HELPS2B validated on independent multicenter cohort (N=2111)
- Implemented: University of Wisconsin, Massachusetts General Hospital/Harvard Medical School
- Ongoing implementation: Emory University, Duke University, Medical University of South Carolina, Free University of Brussels (Belgium)

- Resulted in **63.6%** reduction in duration of EEG monitoring per patient
  - $1,134.831 saving per patient[1]
- **2.82 X** More Patients Monitored
- **$6.1M** estimated savings in FY 2018 at MGH,UW




[1]2016 Medicare Reimbursement Most Common Professional Code

- So that’s how interpretable models are supposed to work…
  but don’t they lose accuracy?

The New York Times

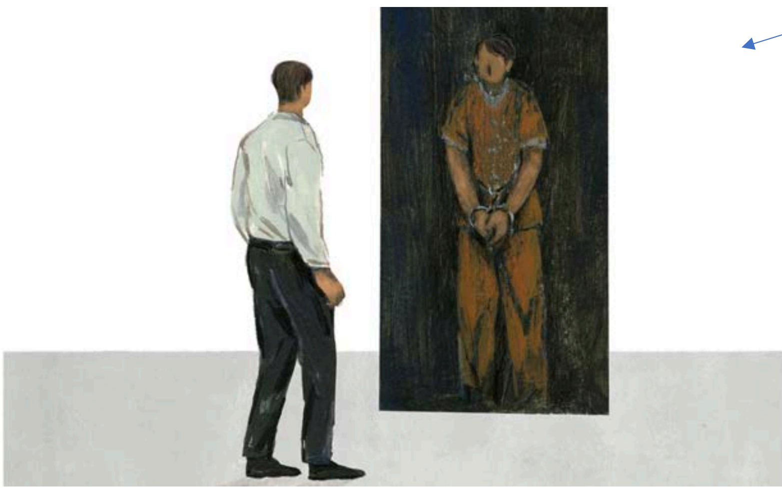
OP-ED CONTRIBUTOR

# When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

232

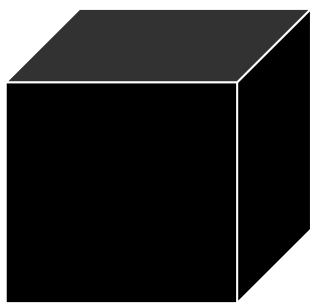


Glenn Rodriguez was denied parole because of a miscalculated "COMPAS" score.

How accurate is COMPAS?
Data from Florida can tell us…

# COMPAS vs. CORELS

COMPAS: (Correctional Offender Management Profiling for Alternative Sanctions)

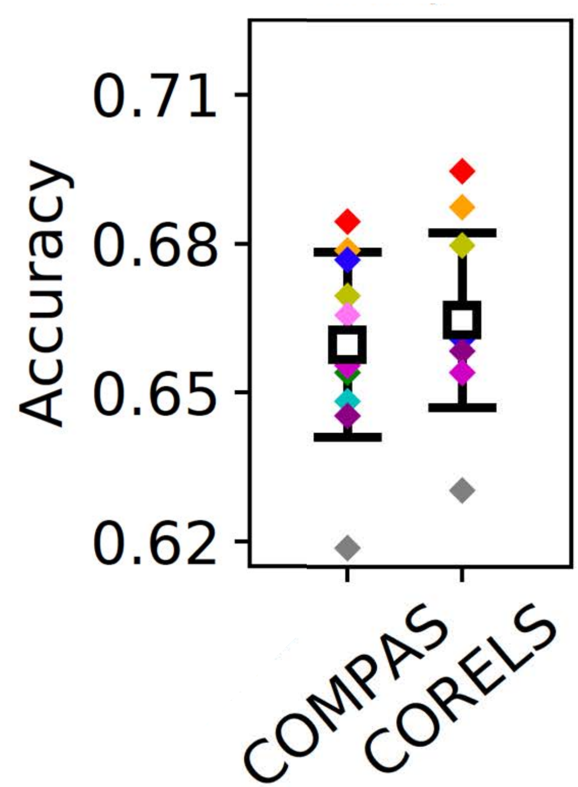


CORELS: (Certifiably Optimal RulE ListS, with Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, and Margo Seltzer, KDD 2017 & JMLR 2018)

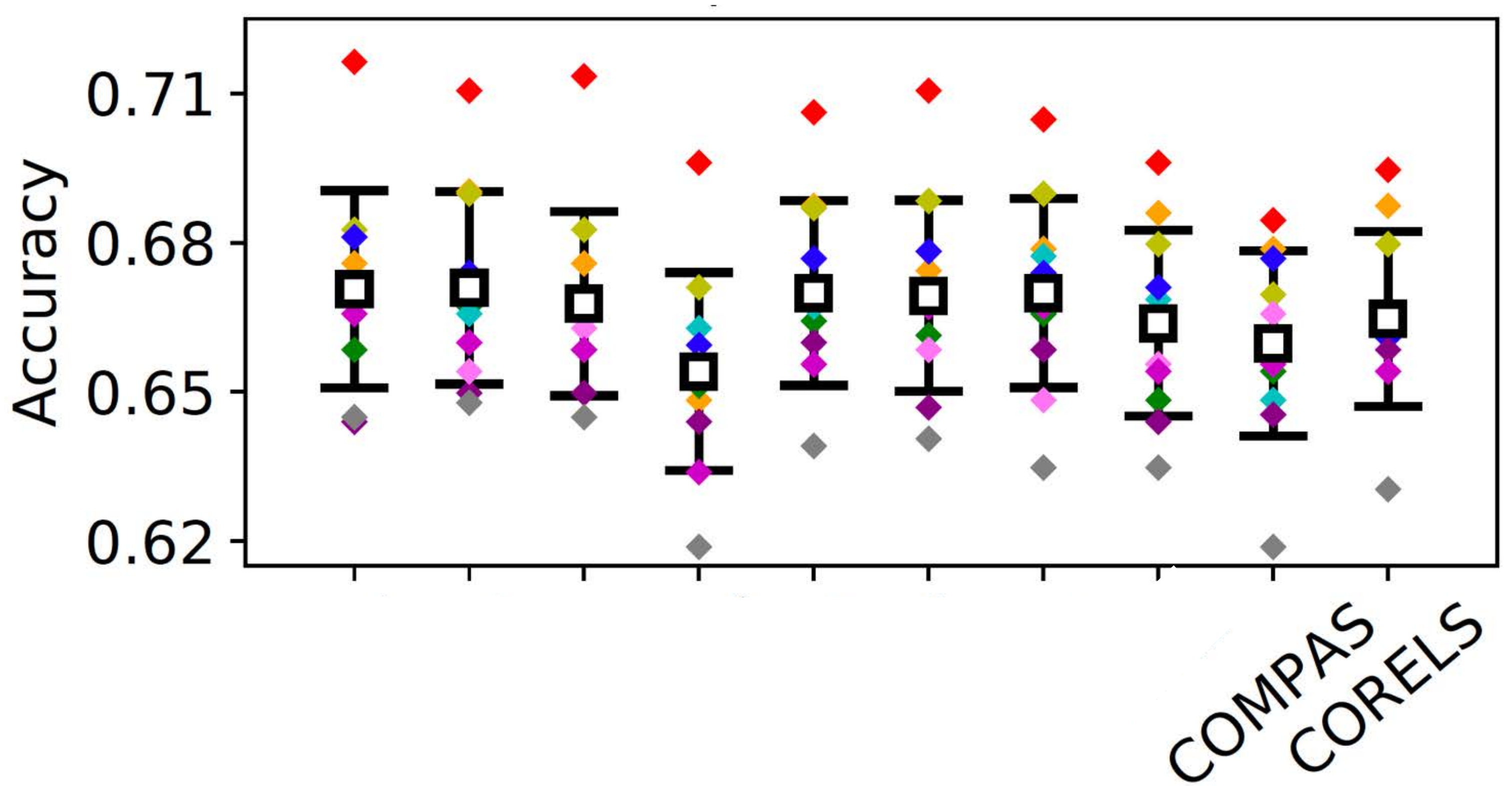
Here is the machine learning model:

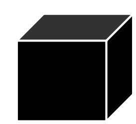
If age=19-20 and sex=male, then predict arrest
else if age=21-22 and priors=2-3 then predict arrest
else if priors >3 then predict arrest
else predict no arrest

# Prediction of re-arrest within 2 years



# Prediction of re-arrest within 2 years



If age=19-20 and sex=male, then predict arrest
else if age=21-22 and priors=2-3 then predict arrest
else if priors >3 then predict arrest
else predict no arrest

- Interpretable ML – When you use a model that is not black box.

- Explainable ML – When you use a black box and explain it afterwards (posthoc)
  - Start with a black box.
    - Create another model that approximates it.
    - Compute derivatives of it.
    - Visualize what part of the input the model is paying attention to.

    :

# Interpretable Models $\neq$ Explanations of Black Box Models

- Trusting a black box means you trust the database it was built from

- Double Trouble: Forces you to rely on two models instead of one. Those models necessarily disagree with each other
  - An explanation that is right 90% of the time is wrong 10% of the time.

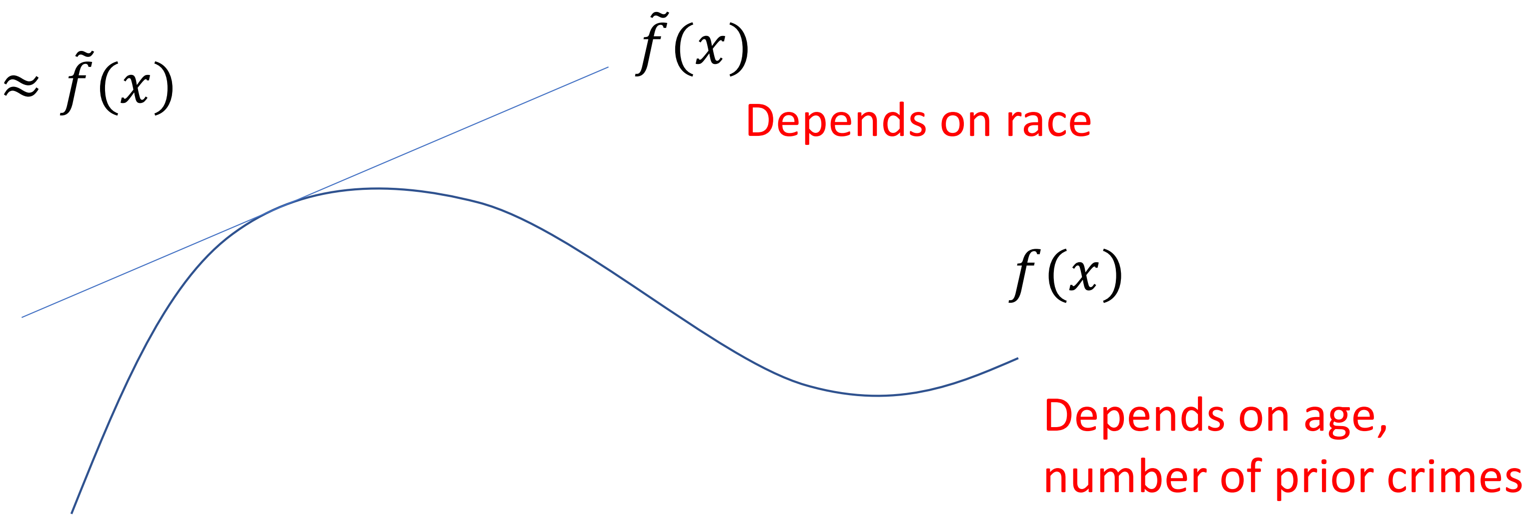
- Typos are a problem when inputting data into black box models.

- If you can produce an interpretable model, why explain a black box? (e.g., COMPAS vs CORELS)

# Interpretable Models ≠ Explanations of Black Box Models

- “Explanations” are not actually explanations of what the model is doing. **Approximations are not explanations**! Gets variable importance wrong.

$$f(x) \approx \tilde{f}(x)$$



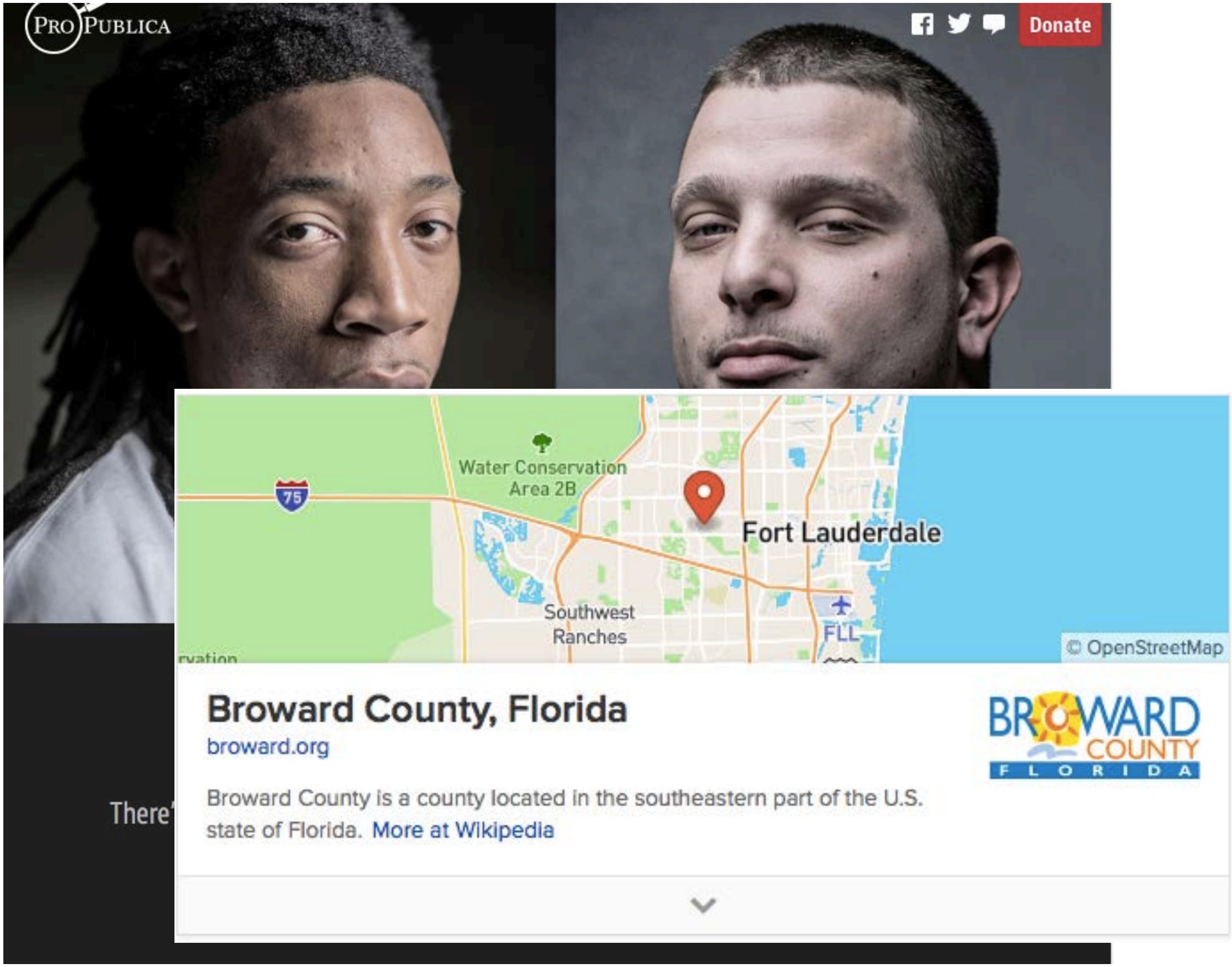

*Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)*

# 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


ProPublica
Donate
Water Conservation Area 2B
75
Fort Lauderdale
Southwest Ranches
FLL
© OpenStreetMap
Broward County, Florida
broward.org
Broward County is a county located in the southeastern part of the U.S. state of Florida. More at Wikipedia
BROWARD COUNTY FLORIDA

# What ProPublica Did

- They showed that FPR and FNR varied by race.

- They suggested maybe this might not be a good comparison, we should condition on age and number of priors and reexamine.

- After conditioning on age and number of priors, still found a linear model with a low pvalue for the race covariate.

- Concluded that COMPAS depends on race.

# What ProPublica Did

- They showed that FPR and FNR varied by race.
  - This is a property of the data, not necessarily the model. In Broward County, the blacks in the database are younger and have more priors.
- They suggested maybe this might not be a good comparison, we should condition on age and number of priors and reexamine.
  - Good idea
- After conditioning on age and number of priors, still found a linear model with a low pvalue for the race covariate.
  - We don't think COMPAS is linear in their covariates
- Concluded that COMPAS depends on race.
  - Bad idea

# A peek inside COMPAS?

COMPAS - Correctional Offender Management Profiling for Alternative Sanctions. By Northpointe, Inc.

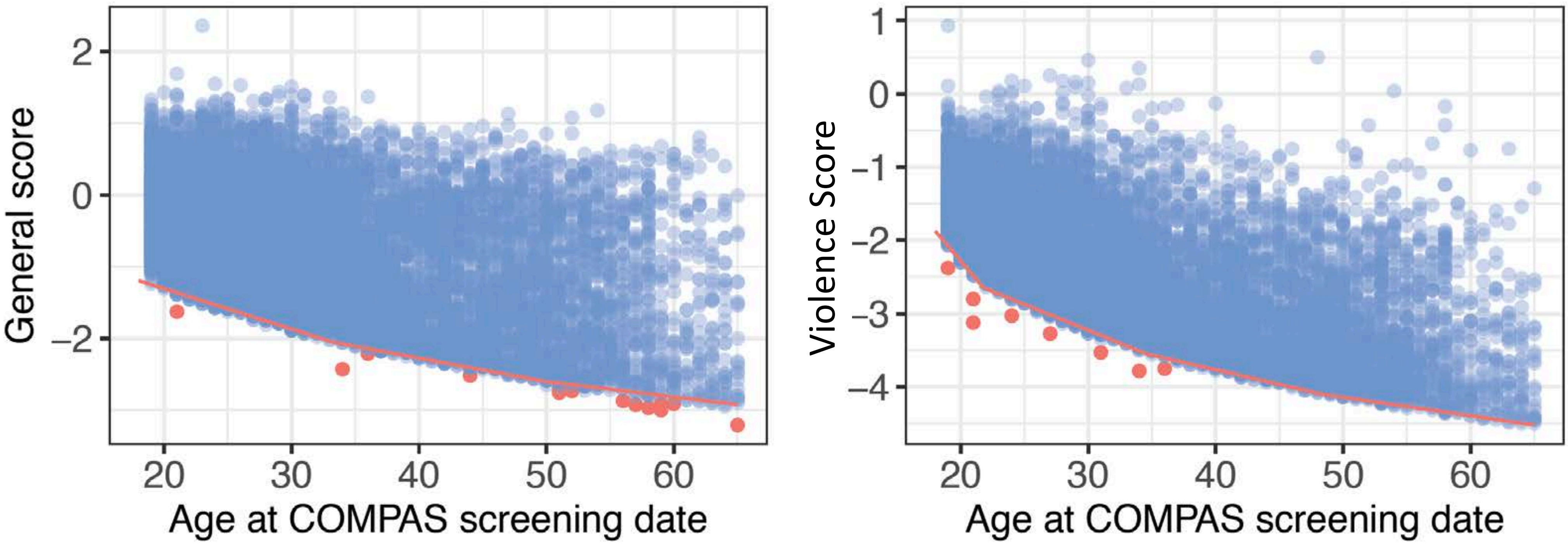
**Conjecture:** *The COMPAS general recidivism model is a nonlinear additive model. Its dependence on age in Broward County is approximately a linear spline, defined as follows:*

$$\begin{aligned}\text{for ages } &\leq 33.26,\ f_{\text{age}}(\text{age}) = -0.056 \times \text{age} - 0.179\\ \text{for ages between } 33.26 \text{ and } 50.02,\ &f_{\text{age}}(\text{age}) = -0.032 \times \text{age} - 0.963\\ \text{for ages } &\geq 50.02,\ f_{\text{age}}(\text{age}) = -0.021 \times \text{age} - 1.541.\end{aligned}$$

*Similarly, the COMPAS violence recidivism model is a nonlinear additive model, with a dependence on age that is approximately a linear spline, defined by:*

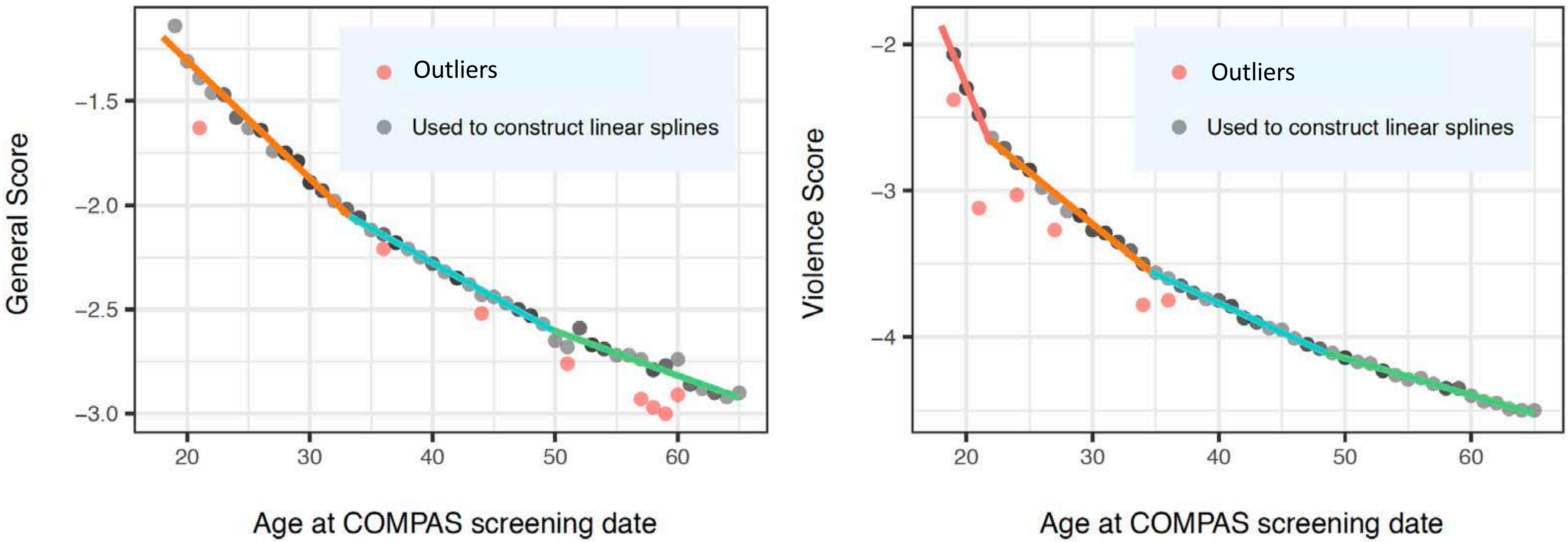
$$\begin{aligned}\text{for ages } &\leq 21.77,\ f_{\text{viol age}}(\text{age}) = -0.205 \times \text{age} + 1.815\\ \text{for ages between } 21.77 \text{ and } 34.58,\ &f_{\text{viol age}}(\text{age}) = -0.070 \times \text{age} - 1.113\\ \text{for ages between } 34.58 \text{ and } 48.36,\ &f_{\text{viol age}}(\text{age}) = -0.040 \times \text{age} - 2.166\\ \text{for ages } &\geq 48.36,\ f_{\text{viol age}}(\text{age}) = -0.025 \times \text{age} - 2.882.\end{aligned}$$

Rudin, Wang, and Coker. The *Age* of Secrecy and Unfairness in Recidivism Prediction. Harvard Data Science Review (Accepted)




Scatter plot of COMPAS scores vs age for all individuals in Broward County FL.

Rudin, Wang, and Coker. The *Age* of Secrecy and Unfairness in Recidivism Prediction. Harvard Data Science Review (Accepted)




Rudin, Wang, and Coker. The *Age* of Secrecy and Unfairness in Recidivism Prediction. Harvard Data Science Review (Accepted)

# A peek inside COMPAS?

- Take COMPAS remainder:

  COMPAS – $f_{age}$

and examine whether it depends on race…

it doesn't seem to.

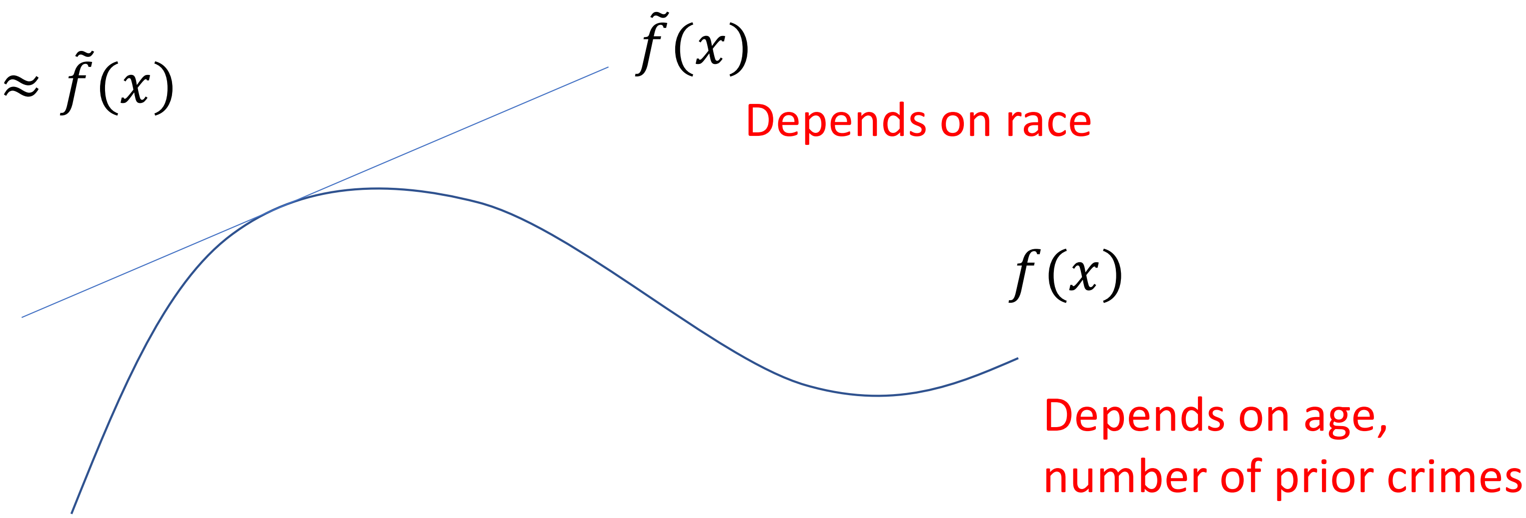
(We ran machine learning methods *with and without race* to see if they need race to predict COMPAS well. They performed similarly.)

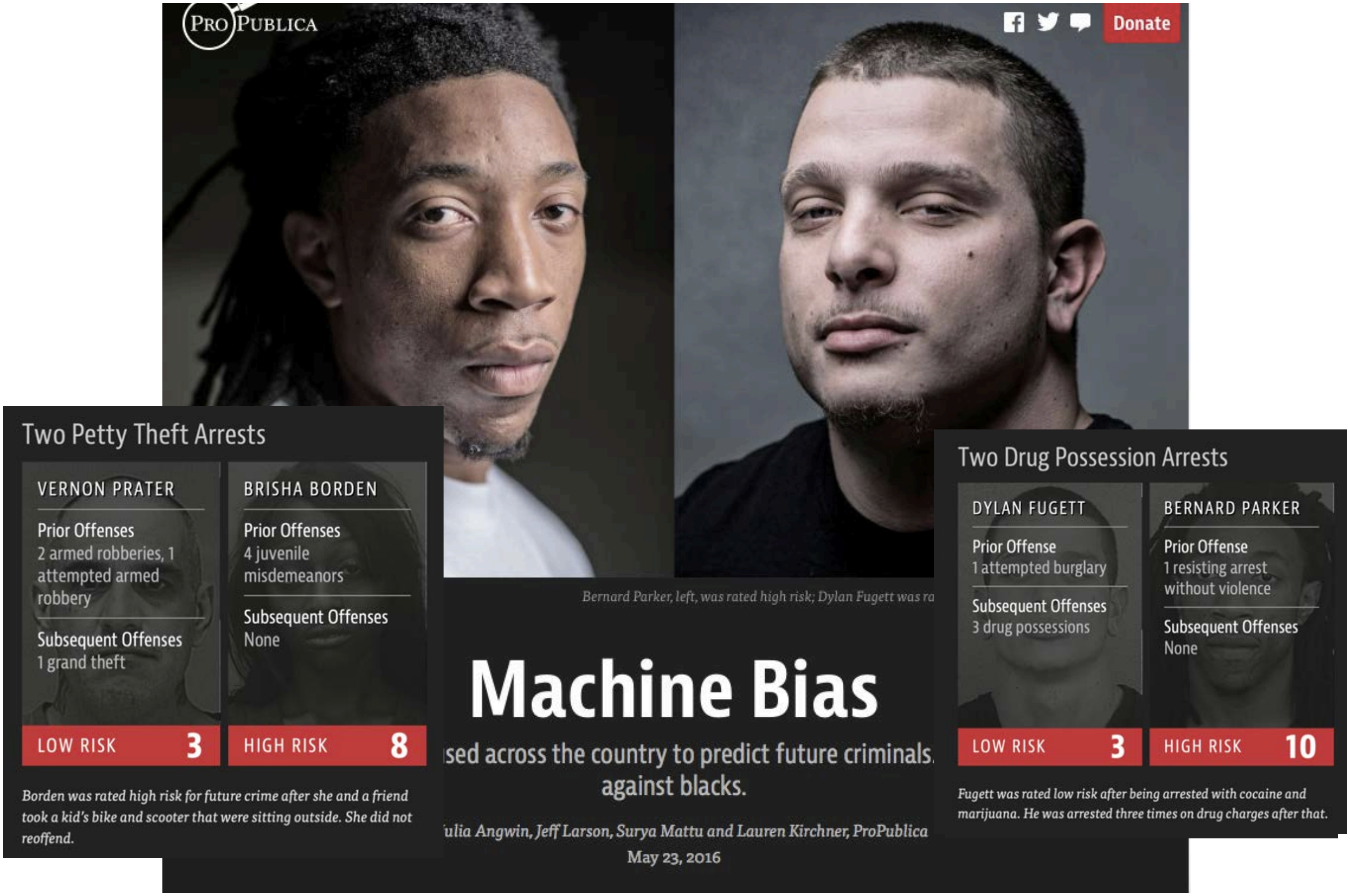
Rudin, Wang, and Coker. The *Age* of Secrecy and Unfairness in Recidivism Prediction. Harvard Data Science Review (Accepted)

# Interpretable Models ≠ Explanations of Black Box Models

- "Explanations" are not actually explanations of what the model is doing. **Approximations are not explanations**! Gets variable importance wrong.

$$f(x) \approx \tilde{f}(x)$$




ProPublica
Donate
Bernard Parker, left, was rated high risk; Dylan Fugett was ra
Machine Bias
sed across the country to predict future criminals.
against blacks.
ulia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016
Two Petty Theft Arrests
VERNON PRATER
Prior Offenses
2 armed robberies, 1 attempted armed robbery
Subsequent Offenses
1 grand theft
LOW RISK 3
BRISHA BORDEN
Prior Offenses
4 juvenile misdemeanors
Subsequent Offenses
None
HIGH RISK 8
Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.
Two Drug Possession Arrests
DYLAN FUGETT
Prior Offense
1 attempted burglary
Subsequent Offenses
3 drug possessions
LOW RISK 3
BERNARD PARKER
Prior Offense
1 resisting arrest without violence
Subsequent Offenses
None
HIGH RISK 10
Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

137 factors entered by hand for each survey

1% error rate → 75% chance of at least one typo on a survey

This is a serious disadvantage to complicated or proprietary models.

In Florida….?

The New York Times

OP-ED CONTRIBUTOR

## When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

232

| Name | COMPAS<br>Violence Decile | #<br>Priors | Selected<br>Prior Charges | Selected<br>Subsequent Charges |
|---|---|---|---|---|
| Vilma<br>Dieppa | 1 | 4 | Aggravated Battery (F,1),<br>Child Abuse (F,1),<br>Resist Officer w/Violence (F,1) | |
| David<br>Selzer | 1 | 14 | Battery on Law Enforc Officer (F,3),<br>Aggravated Assault W/Dead Weap (F,1),<br>Aggravated Battery (F,1),<br>Resist/obstruct Officer W/viol (F,1) | |
| Berry<br>Sanders | 1 | 15 | Attempted Murder 1st Degree (F,1),<br>Resist/obstruct Officer W/viol (F,1),<br>Agg Battery Grt/Bod/Harm (F,1),<br>Carrying Concealed Firearm (F,1) | Armed Sex Batt/vict<br>12 Yrs + (F,2), Aggravated<br>Assault W/dead Weap (F,3),<br>Kidnapping (F,1) |
| Fernando<br>Walker | 1 | 22 | Aggrav Battery w/Deadly Weapon (F,1),<br>Driving Under The Influence (M,2),<br>Carrying Concealed Firearm (F,1) | |
| Steven<br>Glover | 1 | 28 | Robbery / Deadly Weapon (F,11),<br>Poss Firearm Commission Felony (F,7) | |
| Rufus<br>Jackson | 1 | 40 | Resist/obstruct Officer W/viol (F,3),<br>Battery on Law Enforc Officer (F,2),<br>Attempted Robbery Deadly Weapo (F,1),<br>Robbery 1 / Deadly Weapon (F,1) | |
| Miguel<br>Gonzalez | 2 | 6 | Murder in the First Degree (F,1),<br>Aggrav Battery w/Deadly Weapon (F,1),<br>Carrying Concealed Firearm (F,1) | |

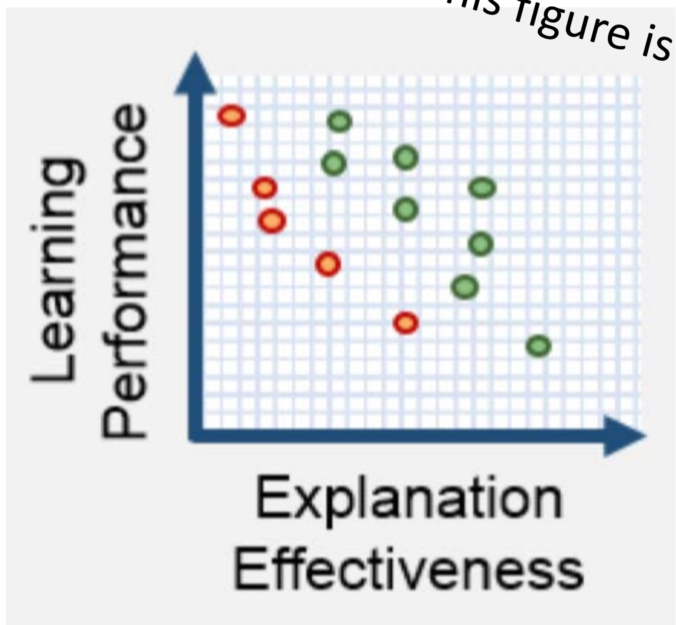
| Name | COMPAS Violence Decile | # Priors | Selected Prior Charges | Selected Subsequent Charges |
|---|---|---|---|---|
| Vilma Dieppa | 1 | 4 | Aggravated Battery (F,1), Child Abuse (F,1), Resist Officer w/Violence (F,1) | |
| David Selzer | 1 | 14 | Battery on Law Enforc Officer (F,3), Aggravated Assault W/Dead Weap (F,1), Aggravated Battery (F,1), Resist/obstruct Officer W/viol (F,1) | |
| Berry Sanders | 1 | 15 | Attempted Murder 1st Degree (F,1), Resist/obstruct Officer W/viol (F,1), Agg Battery Grt/Bod/Harm (F,1), Carrying Concealed Firearm (F,1) | Armed Sex Batt/vict 12 Yrs + (F,2), Aggravated Assault W/dead Weap (F,3), Kidnapping (F,1) |
| Fernando Walker | 1 | 22 | Aggrav Battery w/Deadly Weapon (F,1), Driving Under The Influence (M,2), Carrying Concealed Firearm (F,1) | |
| Steven Glover | 1 | 28 | Robbery / Deadly Weapon (F,11), Poss Firearm Commission Felony (F,7) | |
| Rufus Jackson | 1 | 40 | Resist/obstruct Officer W/viol (F,3), Battery on Law Enforc Officer (F,2), Attempted Robbery Deadly Weapo (F,1), Robbery 1 / Deadly Weapon (F,1) | |
| Miguel Gonzalez | 2 | 6 | Murder in the First Degree (F,1), Aggrav Battery w/Deadly Weapon (F,1), Carrying Concealed Firearm (F,1) | |
| William Kelly | 2 | 17 | Aggravated Assault (F,5), Aggravated Assault W/dead Weap (F,2), Shoot/throw Into Vehicle (F,2), Battery Upon Detainee (F,1) | |
| Richard Campbell | 2 | 21 | Armed Trafficking In Cocaine (F,1), Poss Weapon Commission Felony (F,1), Carrying Concealed Firearm (F,1) | |
| John Coleman | 2 | 25 | Attempt Murder in the First Degree (F,1), Carrying Concealed Firearm (F,1), Felon in Pos of Firearm or Amm (F,1) | |
| Oscar Pope | 2 | 38 | Aggravated Battery (F,3), Robbery / Deadly Weapon (F,3), Kidnapping (F,1), Carrying Concealed Firearm (F,2) | Grand Theft in the 3rd Degree (F,3) |
| Travis Spencer | 3 | 16 | Aggravated Assault W/dead Weap (F,1), Burglary Damage Property>$1000 (F,1), Burglary Unoccupied Dwelling (F,1) | |
| Michael Avila | 3 | 17 | Aggravated Assault W/dead Weap (F,2), Aggravated Assault w/Firearm (F,2), Discharge Firearm From Vehicle (F,1), Home Invasion Robbery (F,1) | Fail Register Vehicle (M,2) |

| | | | | |
|---|---|---|---|---|
| Richard Campbell | 2 | 21 | Armed Trafficking In Cocaine (F,1), Poss Weapon Commission Felony (F,1), Carrying Concealed Firearm (F,1) | |
| John Coleman | 2 | 25 | Attempt Murder in the First Degree (F,1), Carrying Concealed Firearm (F,1), Felon in Pos of Firearm or Amm (F,1) | |
| Oscar Pope | 2 | 38 | Aggravated Battery (F,3), Robbery / Deadly Weapon (F,3), Kidnapping (F,1), Carrying Concealed Firearm (F,2) | Grand Theft in the 3rd Degree (F,3) |
| Travis Spencer | 3 | 16 | Aggravated Assault W/dead Weap (F,1), Burglary Damage Property>$1000 (F,1), Burglary Unoccupied Dwelling (F,1) | |
| Michael Avila | 3 | 17 | Aggravated Assault W/dead Weap (F,2), Aggravated Assault w/Firearm (F,2), Discharge Firearm From Vehicle (F,1), Home Invasion Robbery (F,1) | Fail Register Vehicle (M,2) |
| Terrance Murphy | 3 | 20 | Solicit to Commit Armed Robbery (F,1), Armed False Imprisonment (F,1), Home Invasion Robbery (F,1) | Driving While License Revoked (F,3) |
| Anthony Hawthorne | 3 | 25 | Attempt Sexual Batt / Vict 12+ (F,1), Resist/obstruct Officer W/viol (F,1), Poss Firearm W/alter/remov Id# (F,1) | |
| Stephen Brown | 3 | 36 | Carrying Concealed Firearm (F,2), Battery On Law Enforce Officer (F,1), Kidnapping (F,1), Aggravated Battery (F,1) | Driving While License Revoked (F,3) |
| Samuel Walker | 3 | 36 | Murder in the First Degree (F,1), Poss Firearm Commission Felony (F,1), Solicit to Commit Armed Robbery (F,1) | Petit Theft 100−300 (M,1) |
| Jesse Bernstein | 4 | 10 | Aggravated Battery / Pregnant (F,1), Sex Battery Vict Mental Defect (F,1), Shoot/throw In Occupied Dwell (F,1) | Tresspass in Struct/Convey Occupy (M,1) |
| Shandedra Hardy | 4 | 16 | Aggrav Battery w/Deadly Weapon (F,1), Felon in Pos of Firearm or Amm (F,4) | Resist/Obstruct W/O Violence (M,1), Possess Drug Paraphernalia (M,1) |

# Back to Interpretable vs Explainable…

This figure is phony baloney

The tradeoff doesn't happen like this




Static dataset?

Are they talking about explaining black boxes?

From the DARPA XAI BAA, 2016

PERSPECTIVE
https://doi.org/10.1038/s42256-019-0048-x

nature machine intelligence

# Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

**Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that are inherently interpretable. This Perspective clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare and computer vision.**

There has been an increasing trend in healthcare and criminal justice to leverage machine learning (ML) for high-stakes prediction applications that deeply impact human lives. Many of not. There is a spectrum between fully transparent models (where we understand how all the variables are jointly related to each other) and models that are lightly constrained in model form (such as models

- Typos (e.g., Glenn Rodriguez's COMPAS calculation)
- Black box models *still* force you to trust the dataset.
- Double trouble: Forces you to rely on two models instead of one.

  Those models necessarily disagree with each other
  - An explanation that is right 90% of the time is wrong 10% of the time.
- The explanations are not really explanations, they don't use the same variables.

(Propublica scandal: They said COMPAS depends on age, criminal history, and *race*. But their analysis is wrong - it probably *only* depends on race through age and criminal history.)

- If you can produce an interpretable model, why explain black boxes? Do you really want to extend the authority of the black box?

# Some current projects

- Almost-matching-exactly for matching treatment and control units
- Optimal sparse decision trees, and optimal sparse decision lists
- Scoring systems (sparse linear models with integer coefficients)
- Interpretable neural networks

# Almost Matching Exactly

Cynthia Rudin

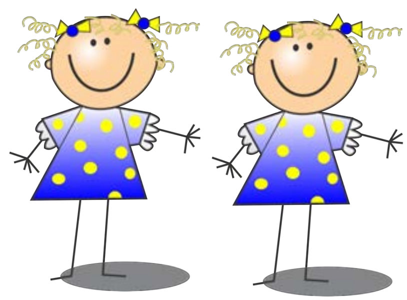
Professor of Computer Science, Electrical and Computer Engineering and Statistical Science, Duke University

Joint work with Alex Volfovsky, Sudeepa Roy, Tianyu Wang, Marco Morucci, Usaid Awan, Vittorio Orlandi, Harsh Parikh and Yameng Liu

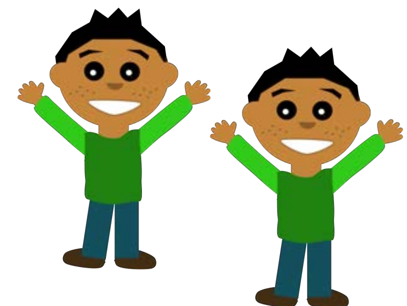
# In Observational Data

Ideally…

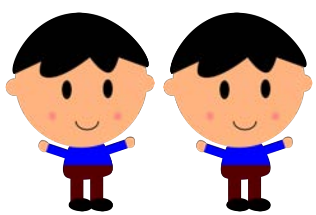
control treated



control treated



control treated



control treated

# In Observational Data

$\underset{n \times p}{X}, \quad \underset{n \times 1}{Y}, \quad \underset{\substack{n \times 1 \\ \{0,1\}}}{T}$ observational data, SUTVA, strong ignorability

Stroke Sumatripan (for migraines)

Matching is useful because it is ***interpretable.***

Most matching methods don't try to match exactly.

Most matching methods used a fixed distance metric between units.

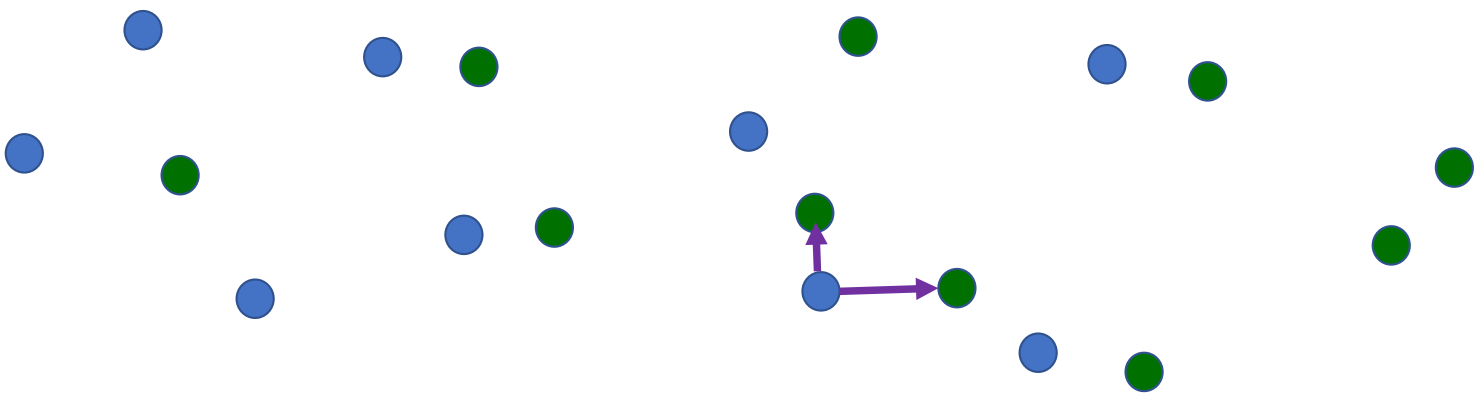
covariates: age, gender, heart conditions, blood pressure, toenail length, eyeball width, etc

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| treated patient | | | | | | | | |
| Marietta | [ 50 | F | 1 0 1 1 | 68 | 1.5cm | 2cm | 1 0 3 0 ..... ] | |
| control patient 1 | | | | | | | | |
| Lee Ann | [ 50 | F | 1 0 1 1 | 68 | 14cm | 1cm | 4 1 5 6 ..... ] | |

# Almost Matching Exactly

- Goal: Match treatment and control units using *important* covariates.

*Learn* the distance metric between units on a holdout training set



# Almost Matching Exactly

- Goal: Match treatment and control units using *important* covariates.

*Learn* the distance metric between units on a holdout training set

FLAME – Fast Large-scale Almost Matching Exactly

– for categorical data

– distance metric is a weighted Hamming distance

DAME – Dynamic Almost Matching Exactly

– sister algorithm to FLAME

MALTS – Matching After Learning to Stretch

– for continuous data

– distance metric is a Mahalanobis distance (a stretch matrix)

# FLAME - Fast Large-Scale Almost Matching Exactly
# DAME - Dynamic Almost Matching Exactly

- Alternates between:
  - ML step: choose which covariates to match on.
  - Matching step: find matched groups using either an efficient SQL query or a bit-vector computation

- FLAME-DAME hybrid
  - Run FLAME using backwards elimination until the number of covariates is small enough to run DAME

Say only the first 10 out of 40 covariates are relevant.
Eliminate covariate subsets in this order:

t=1 40

t=2 40,39

t=3 40,39,38

t=4 40,39,38,37

t=5 40,39,38,37,36

:

FLAME iterations

t=31 40,39,…,13,12,11,10

t=32 40,39,…,13,12,11,9

t=33 40,39,…,13,12,11,10,9

t=34 40,39,…,13,12,11,8

t=35 40,39,…,13,12,11,10,8

t=36 40,39,…,13,12,11,9,8

t=37 40,39,…,13,12,11,10,9,8

DAME iterations

t=large 40,39,…,13,12,11,10,9,8,6,3

Stop iterating here – if I eliminate anything else, I can't predict the outcome.

# FLAME - Fast Large-Scale Almost Matching Exactly
# DAME - Dynamic Almost Matching Exactly

- Produces high quality matched groups
- Covariates used for matching can (together) predict the outcome well.

# MALTS – Matching After Learning to Stretch

- ML step: learn how much to stretch each covariate
- Matching step: find matched groups as k-nearest neighbors

# An Example from the LaLonde Dataset

| Age | Education | Black | Hispanic | Married | Nodegree | Income 1975 (re75) |
|---|---|---|---|---|---|---|
| 2.745 | 1.61 | 0.331 | 0.389 | 0.206 | 0.434 | 0.164 |

← Stretch matrix
(for normalized features)

A matched group

| Age | Education | Black | Hispanic | Married | No degree | Income in 1975 (re75) | Income in 1978 (re78) | T |
|---|---|---|---|---|---|---|---|---|
| 23 | 12 | 1 | 0 | 0 | 0 | 0 | 4728.73 | 0 |
| 22 | 12 | 1 | 0 | 1 | 0 | 0 | 664.98 | 0 |
| 22 | 12 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 24 | 12 | 1 | 0 | 0 | 0 | 0 | 10344.09 | 0 |
| 25 | 12 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 27 | 12 | 1 | 0 | 1 | 0 | 0 | 11821.81 | 0 |
| 23 | 12 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 23 | 12 | 1 | 0 | 0 | 0 | 0 | 4843.18 | 1 |
| 22 | 12 | 1 | 0 | 0 | 0 | 0 | 18678.08 | 1 |
| 25 | 12 | 1 | 0 | 0 | 0 | 0 | 2348.97 | 1 |
| 25 | 12 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |

# FLAME/DAME/MALTS

- Python Code: https://github.com/almost-matching-exactly/
- R FLAME Code: https://github.com/JerryChiaRuiChang/FLAME
- Papers on my website: https://users.cs.duke.edu/~cynthia/papers.html

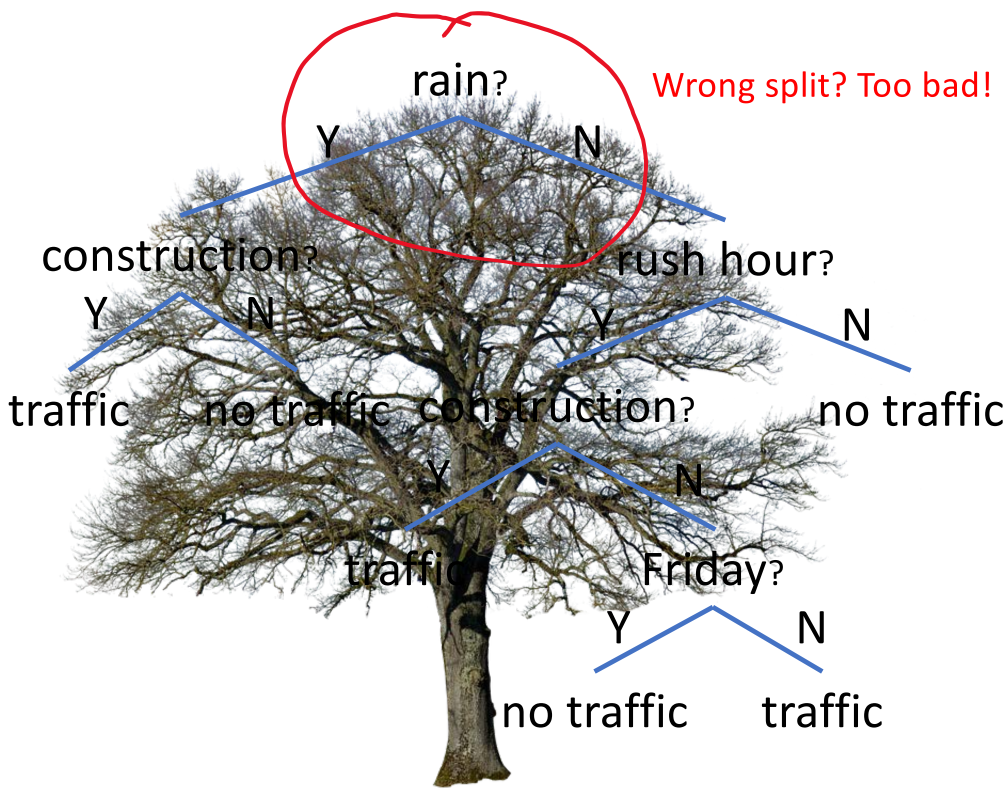
# Some current projects

- Almost-matching-exactly for matching treatment and control units
- Optimal sparse decision trees, and optimal sparse decision lists
- Scoring systems (sparse linear models with integer coefficients)
- Interpretable neural networks

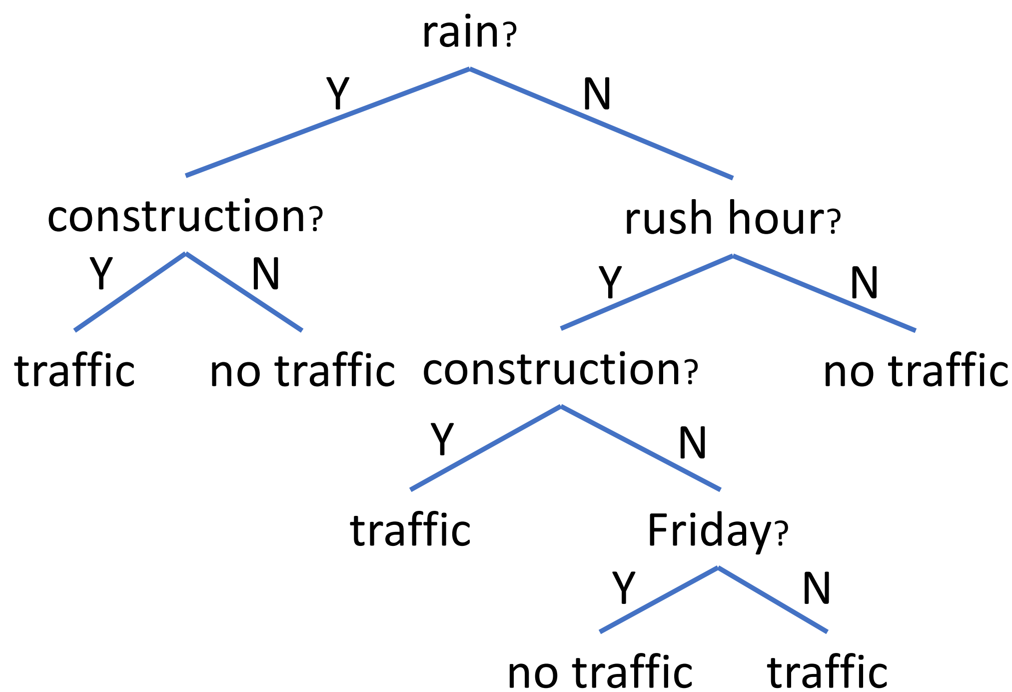
# Optimal Decision Trees

- With Margo Seltzer, Xiyang Hu, Chudi Zhong, Jimmy Lin

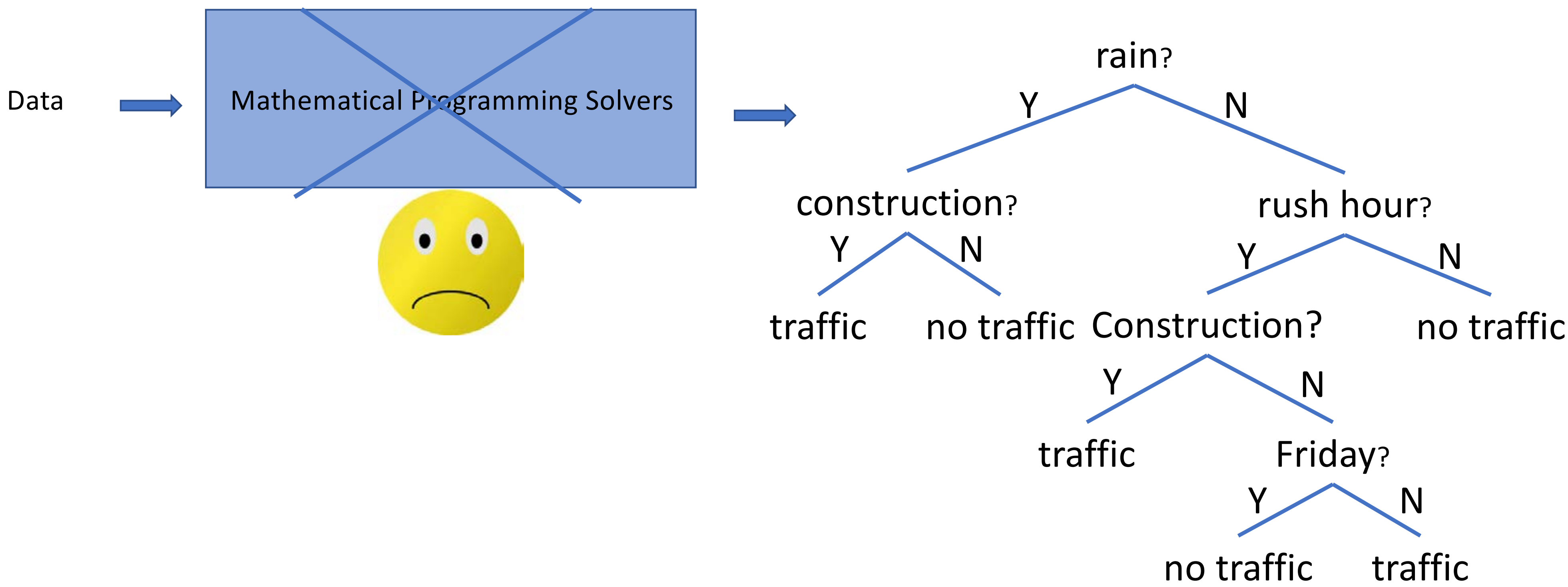
# Optimal Sparse Decision Trees



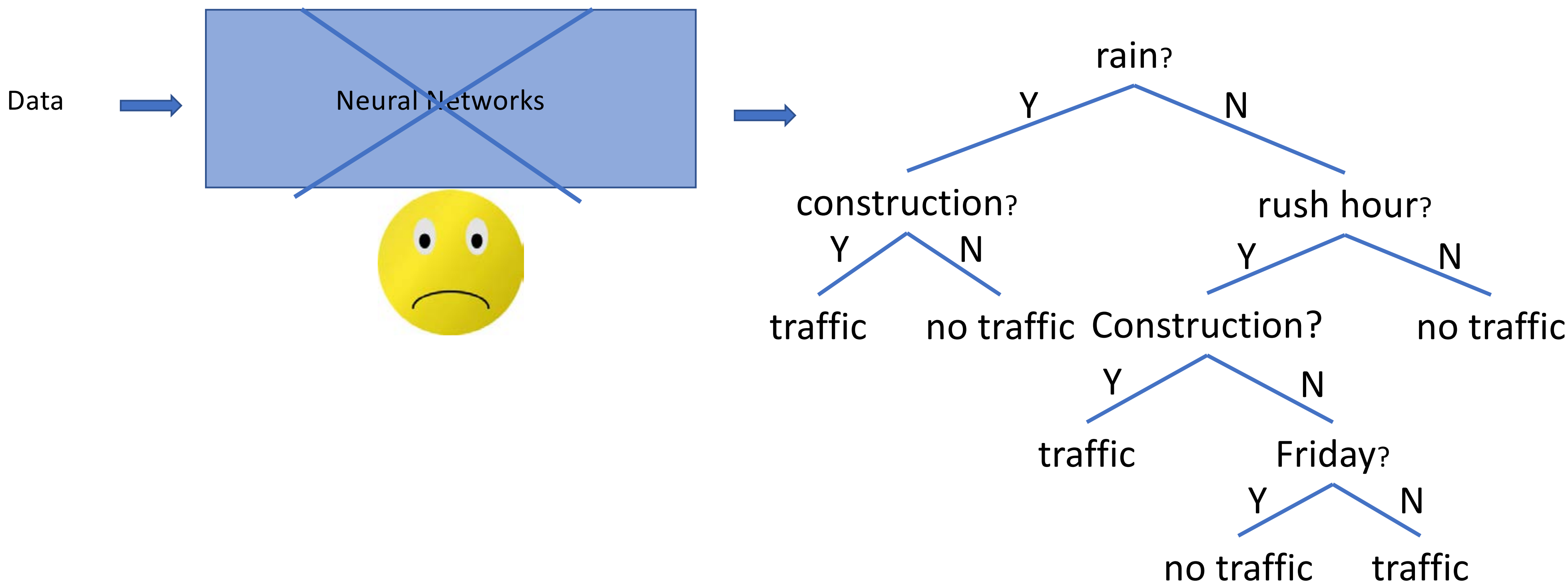
# Optimal Sparse Decision Trees



# Optimal Sparse Decision Trees



# Optimal Sparse Decision Trees



# Optimal Sparse Decision Trees

$$\min_{\text{tree}} \hat{L}(\text{tree}, \{(x_i, y_i)\}_i) \text{ where}$$

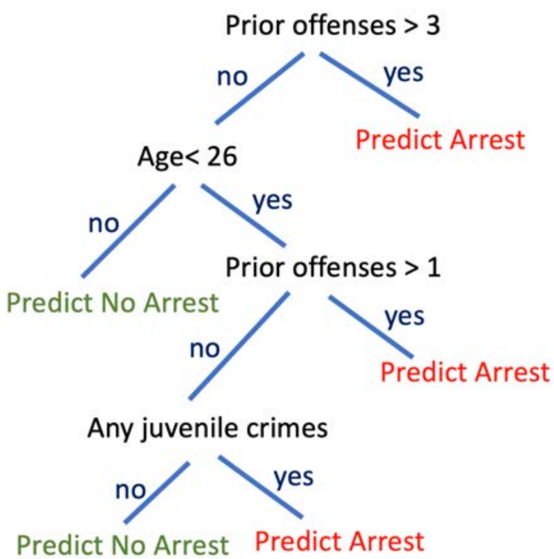
$$\hat{L}(\text{tree}, \{(x_i, y_i)\}_i) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} 1_{[\text{tree}(x_i) \neq y_i]}}_{\text{Misclassification error}} + \underbrace{C(\#\,\text{leaves in tree})}_{\text{Sparsity}}$$

We solve this to optimality.
No greedy splitting and pruning like C4.5 and CART
The key: very efficient branch & bound combined with computer systems.

# Optimal Sparse Decision Trees

$$\min_{\text{tree}} \hat{L}(\text{tree}, \{(x_i, y_i)\}_i) \text{ where}$$

$$\hat{L}(\text{tree}, \{(x_i, y_i)\}_i) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} 1_{[\text{tree}(x_i) \neq y_i]}}_{\text{Misclassification error}} + \underbrace{C(\#\,\text{leaves in tree})}_{\text{Sparsity}}$$

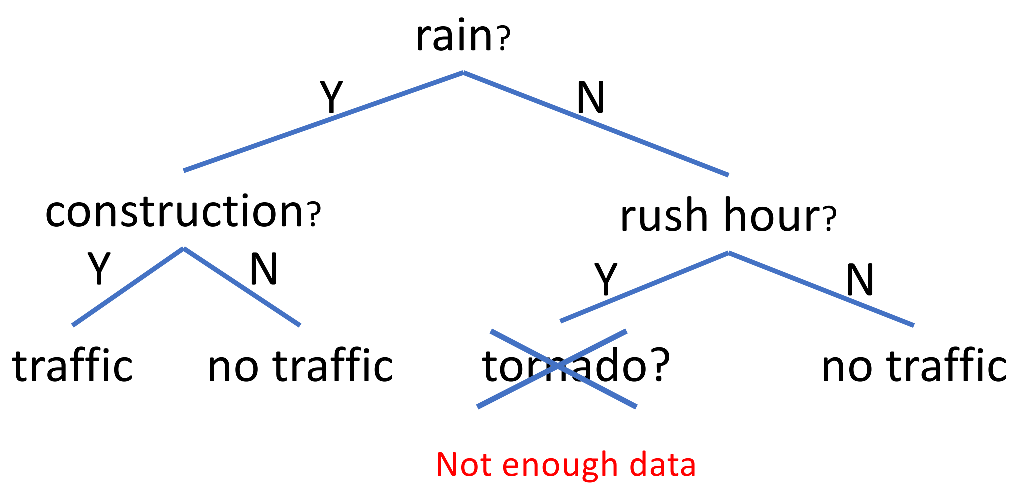



An example of an optimal tree on the Broward County Florida re-arrest data

# Optimal Sparse Decision Trees

## Analytical Bounds Reduce the Search Space

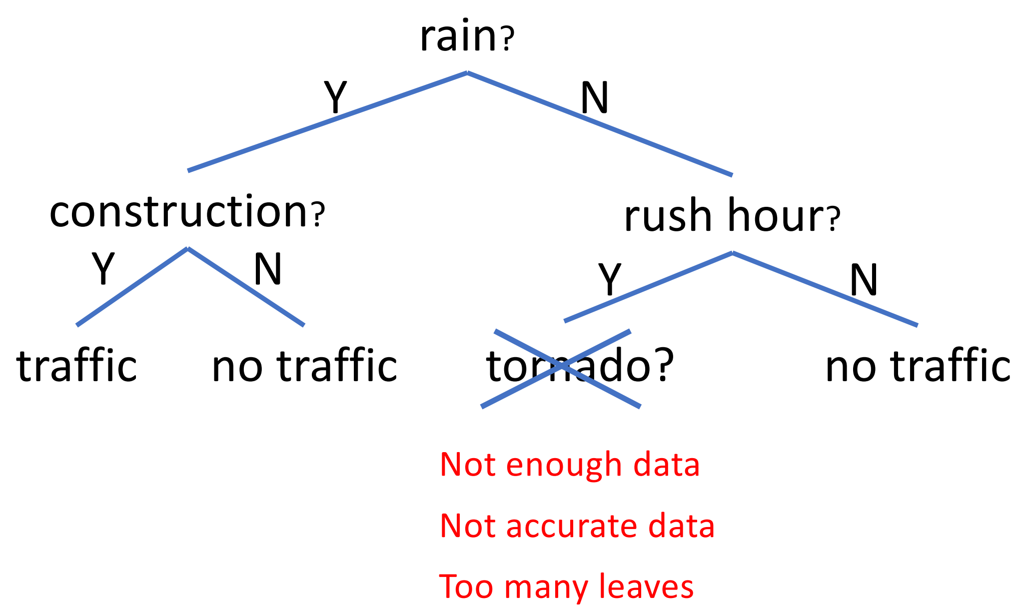
This collection of theorems show that some partial trees can never be extended to form optimal trees.



# Optimal Sparse Decision Trees

## Analytical Bounds Reduce the Search Space

This collection of  theorems show that some partial trees can never be extended to form optimal trees.



# Optimal Sparse Decision Trees

## Represent a tree by its leaves

rain & construction & traffic

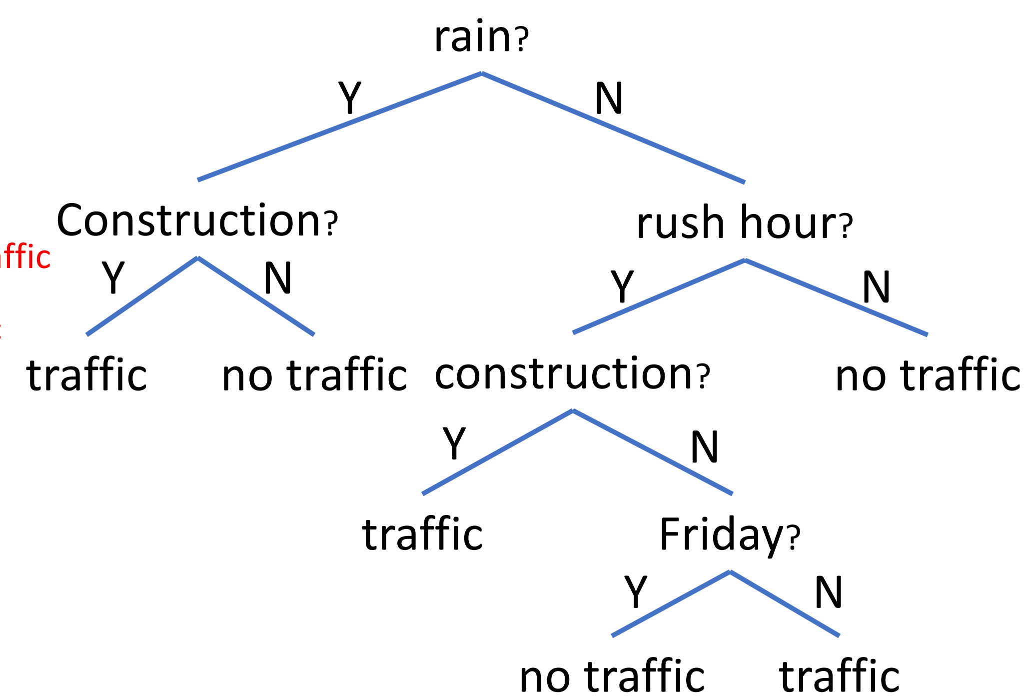
rain & no construction & no traffic

no rain & rush hour & construction & traffic

no rain & rush hour & no construction & Friday and no traffic

no rain & rush hour & no construction & Friday and traffic

no rain & no rush hour & no traffic



# Optimal Sparse Decision Trees

## Permutation map: Discover identical trees already evaluated

rain & construction & traffic

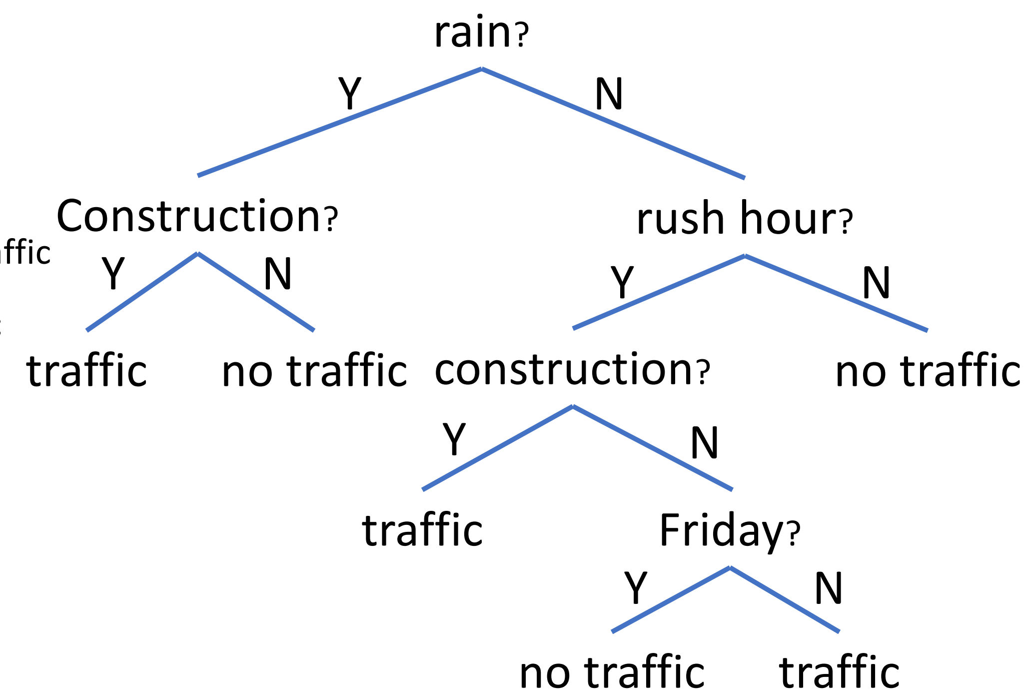
rain & no construction & no traffic

no rain & rush hour & construction & traffic

no rain & rush hour & no construction & Friday and no traffic

no rain & rush hour & no construction & Friday and traffic

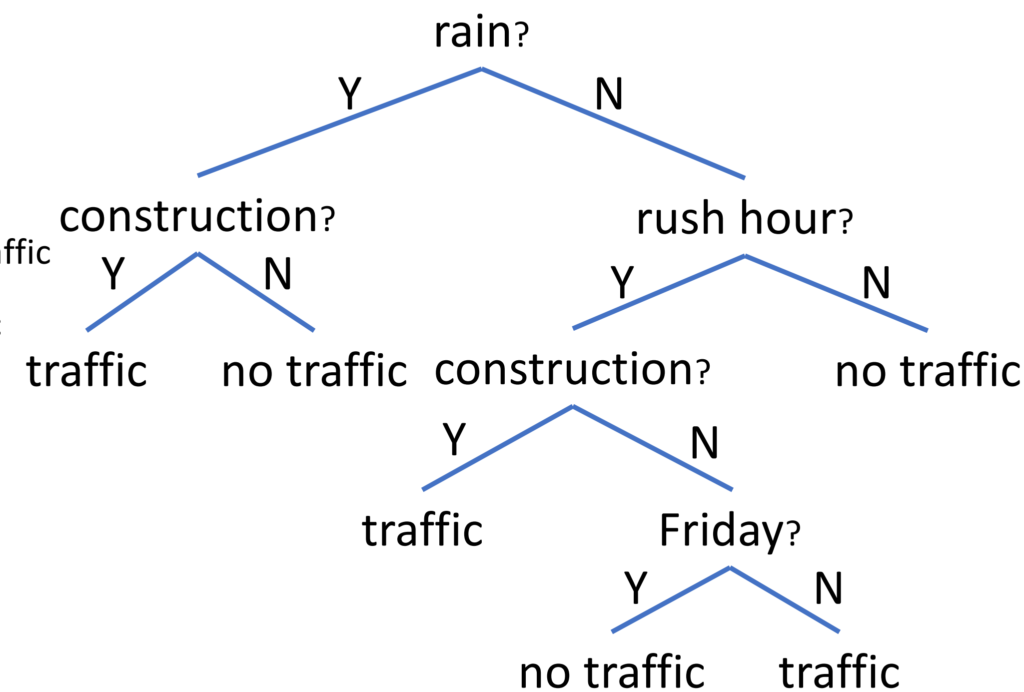
no rain & no rush hour & no traffic



# Optimal Sparse Decision Trees

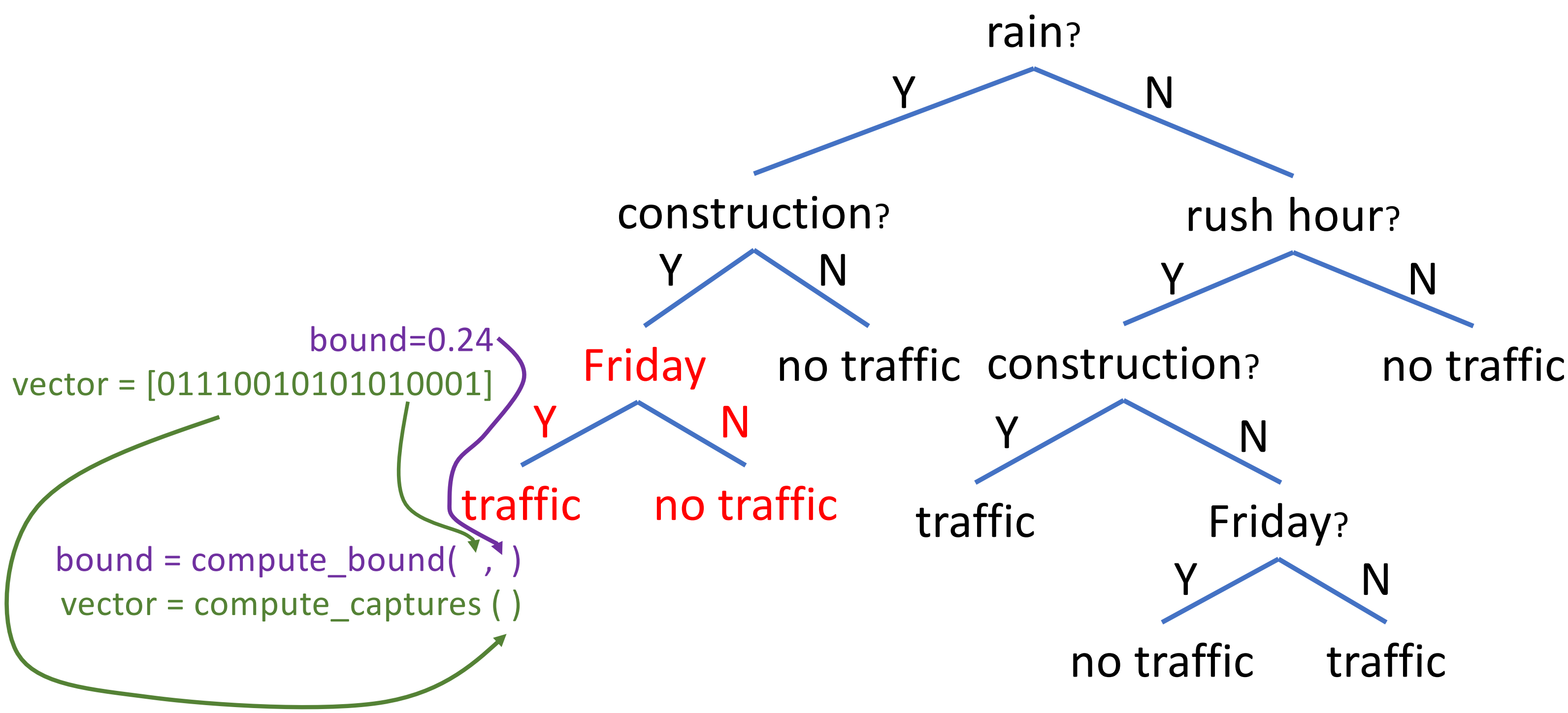
## Bit-vectors describe data represented by each leaf

rain & construction & traffic
[1000010001001110000..........................0]
rain & no construction & no traffic
[0110001000000000110..........................1]
no rain & rush hour & construction & traffic
[0001000100000001000..........................0]
no rain & rush hour & no construction & Friday and no traffic
[0000100000000000001..........................0]
no rain & rush hour & no construction & Friday and traffic
[0000000010000000000..........................0]
no rain & no rush hour & no traffic
[0000000000011000000..........................0]



# Optimal Sparse Decision Trees

## Incremental computation of objective and bounds



# Optimal Sparse Decision Trees

Strong analytical bounds

+

Leaf-based representation

+

Permutation map

+

Caching of intermediate results

+

Incremental computation

=

Fast Implementation

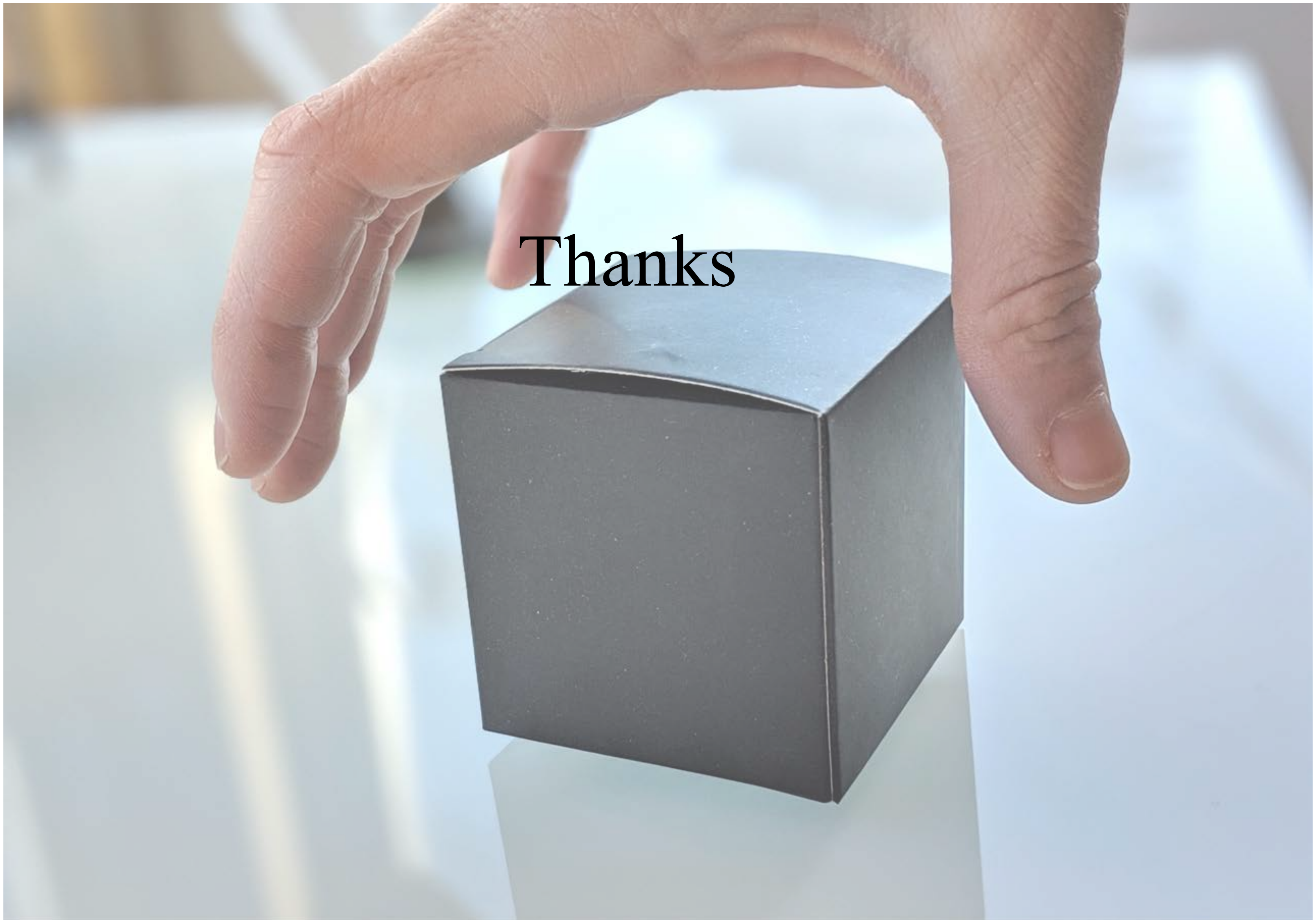
# Optimal Sparse Decision Trees

NeurIPS 2019 (spotlight)

Xiyang Hu, Cynthia Rudin, Margo Seltzer

Code: https://github.com/xiyanghu/OSDT

Paper: https://arxiv.org/abs/1904.12847


Thanks