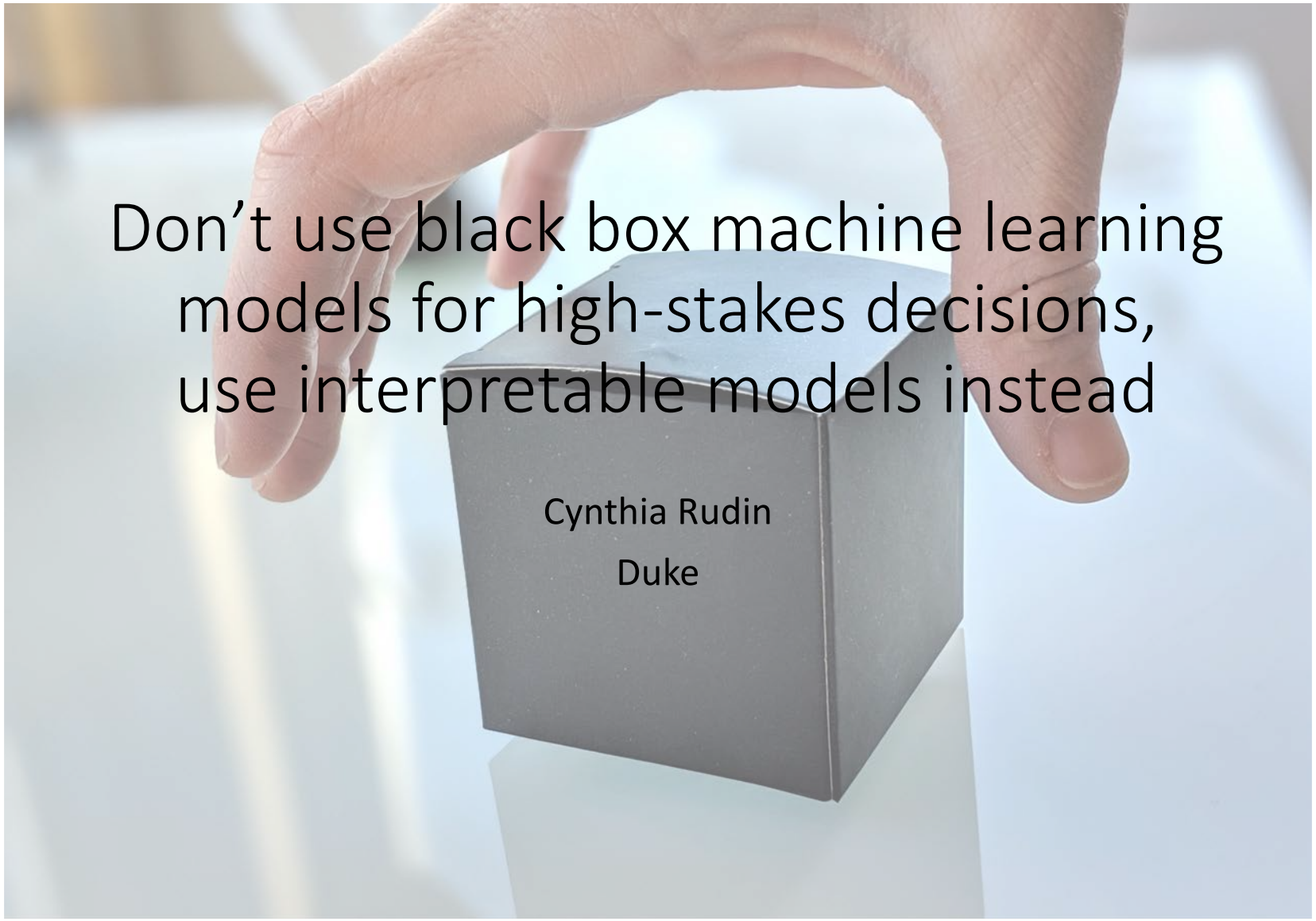


Don’t use black box machine learning models for high-stakes decisions, use interpretable models instead

Cynthia Rudin

Duke

Can a typographical error lead to years of extra prison time?

The New York Times

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

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

- A black box model is a formula that is either too complicated for any human to understand or is proprietary.
  - difficult to troubleshoot the overall model for bias
  - difficult to verify that an individual prediction is correct
  - it doesn't augment human decision makers, it just replaces them
- An interpretable machine learning model obeys a domain-specific set of constraints so that humans can better understand it.
- There's a spectrum.
- High-stakes decisions or troubleshooting
  - Criminal justice models, credit scoring, air pollution, airplane maintenance, many healthcare applications – anything high stakes

What happens when we use a black box?

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?
We'll never know…

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…

The New York Times

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

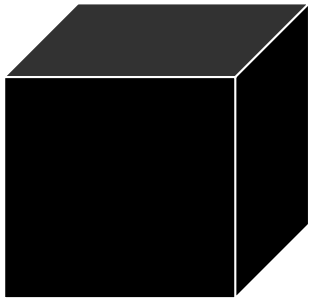
232

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)

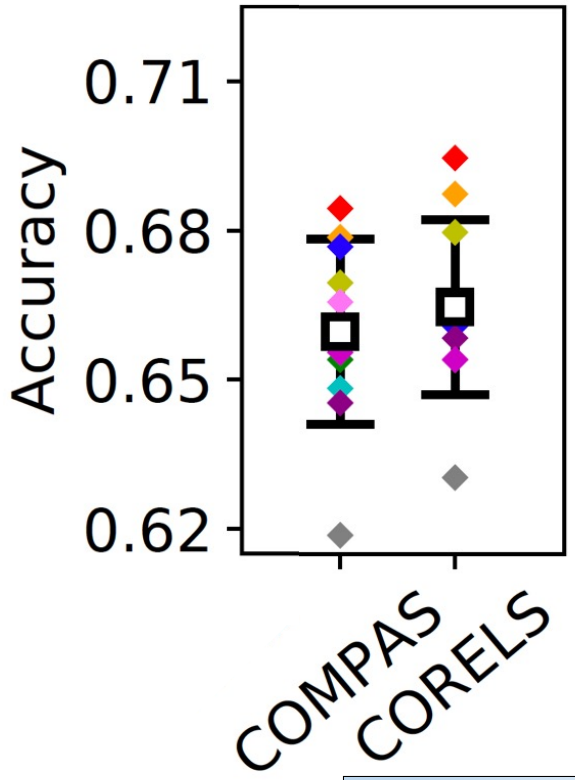


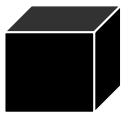
CORELS: (Certifiably Optimal RulE ListS, with Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, and Margo Seltzer, KDD 2017 & JMLR 2018)

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

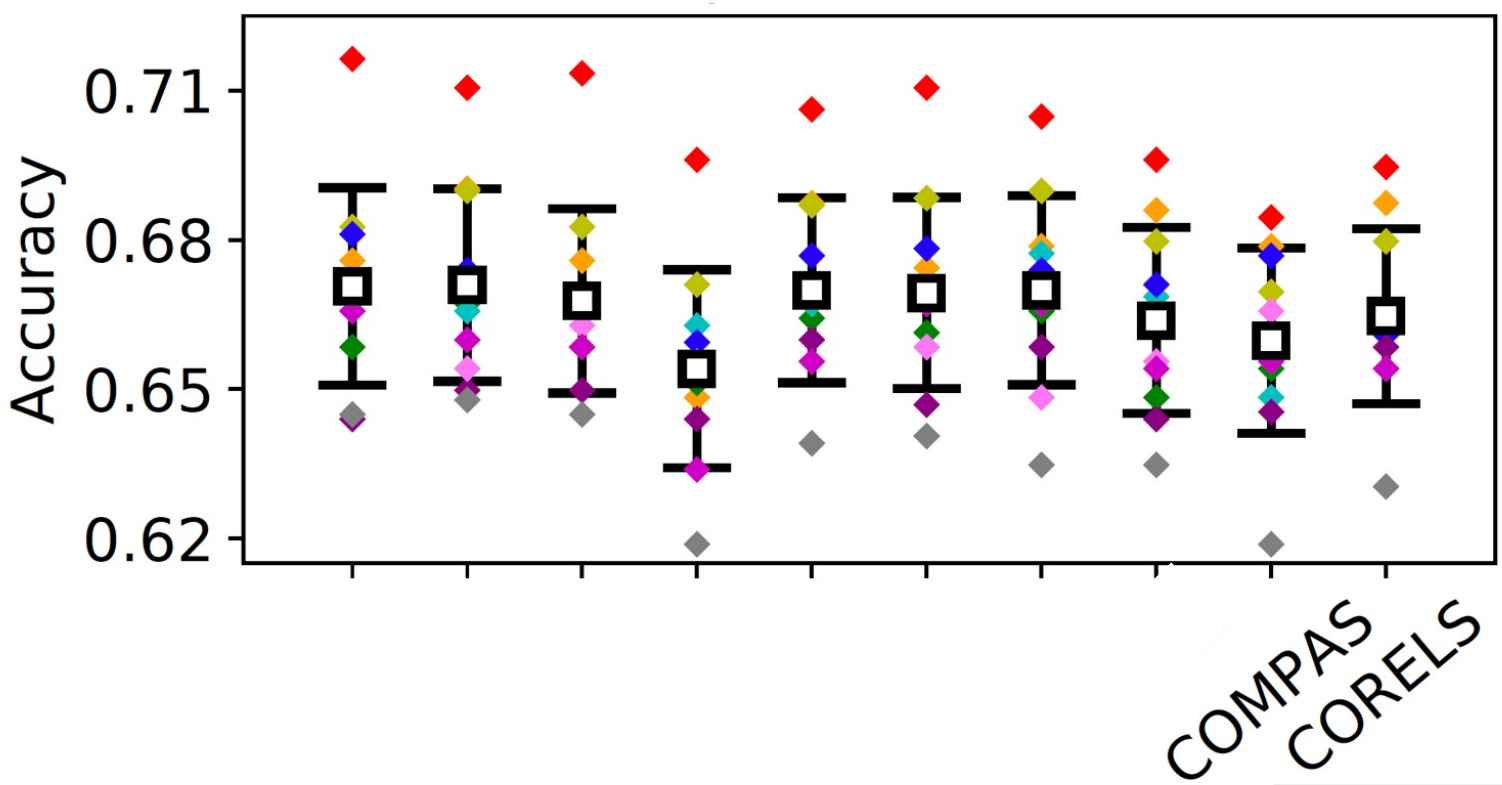
Prediction of re-arrest within 2 years

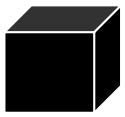




If age=19-20 and sex=male, then predict arrest
else if age=21-22 and priors=2-3 then predict arrest
else if priors >3 then predict arrest
else predict no arrest

Prediction of re-arrest within 2 years





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

Perhaps we are using complicated models when we don't need them

There’s no benefit from complicated models for re-arrest prediction in criminal justice.

Interpretable Classification Models for Recidivism Prediction. Zeng et al., Journal of the Royal Statistical Society, 2016.
Learning Certifiably Optimal Rule Lists for Categorical Data. Angelino et al., Journal of Machine Learning Research, 2018.

There’s no benefit from complicated models for lots of problems.

(Holte, Very Simple Classification Rules Perform Well on Most Commonly Used Datasets, 1993).

Sleep apnea screening

Energy grid reliability (underground power events)

Adult ADHD screening

Seizure prediction in ICU patients

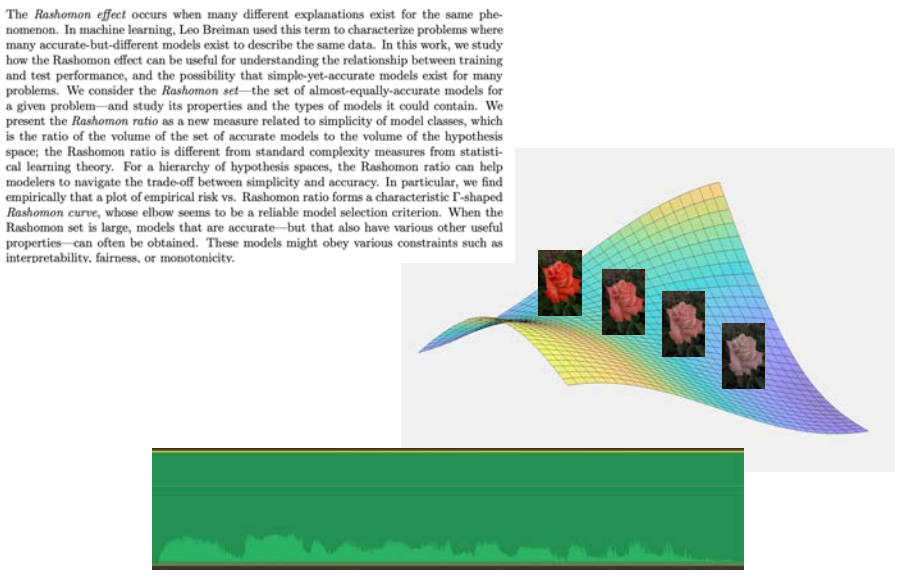
Financial risk assessment

Crime series detection

Depends on data representation.

Problem spectrum

age 45
congestive heart failure? yes
takes aspirin
smoking? no
gender M
exercise? yes
allergies? no
number of past strokes 2
diabetes? yes



Tabular: All features are interpretable

- many problems in criminal justice, healthcare, social sciences, equipment reliability & maintenance, etc.
- features include counts, categorical data

Raw: Features are individually uninterpretable

- pixels/voxels, words, a bit of a sound wave

Problem spectrum

Very interpretable models (if you can optimize)

With minor pre-processing, all methods have similar performance

Neural networks

Tabular: All features are interpretable

- many problems in criminal justice, healthcare, social sciences, equipment reliability & maintenance, etc.
- features include counts, categorical data

Raw: Features are individually uninterpretable

- pixels/voxels, words, a bit of a sound wave

Where are black boxes more accurate?

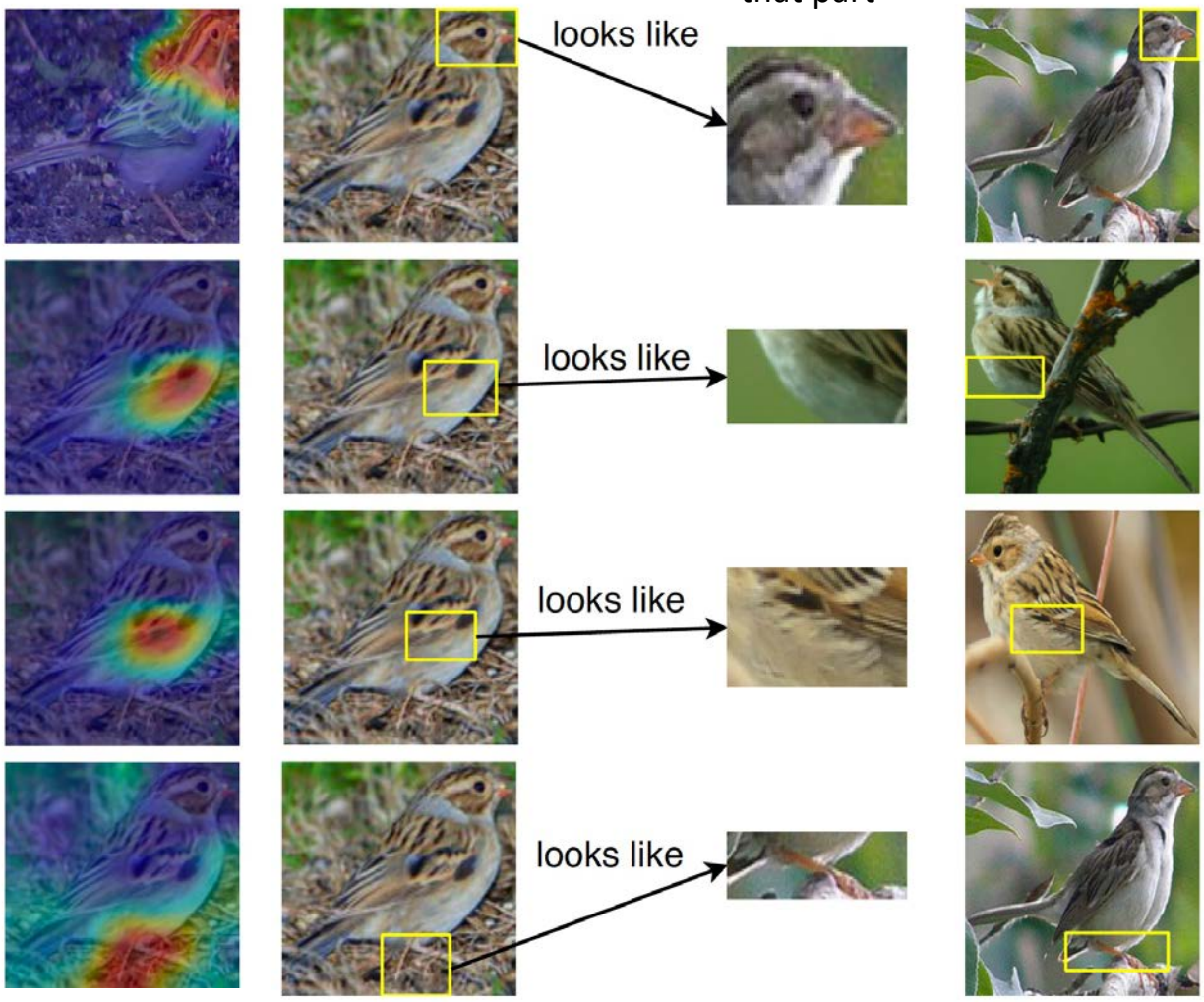
- Challenge 1: Interpretable neural networks for computer vision

Why is this bird classified as a clay-colored sparrow?

Because this part of the bird

that part

of a prototypical clay-colored sparrow



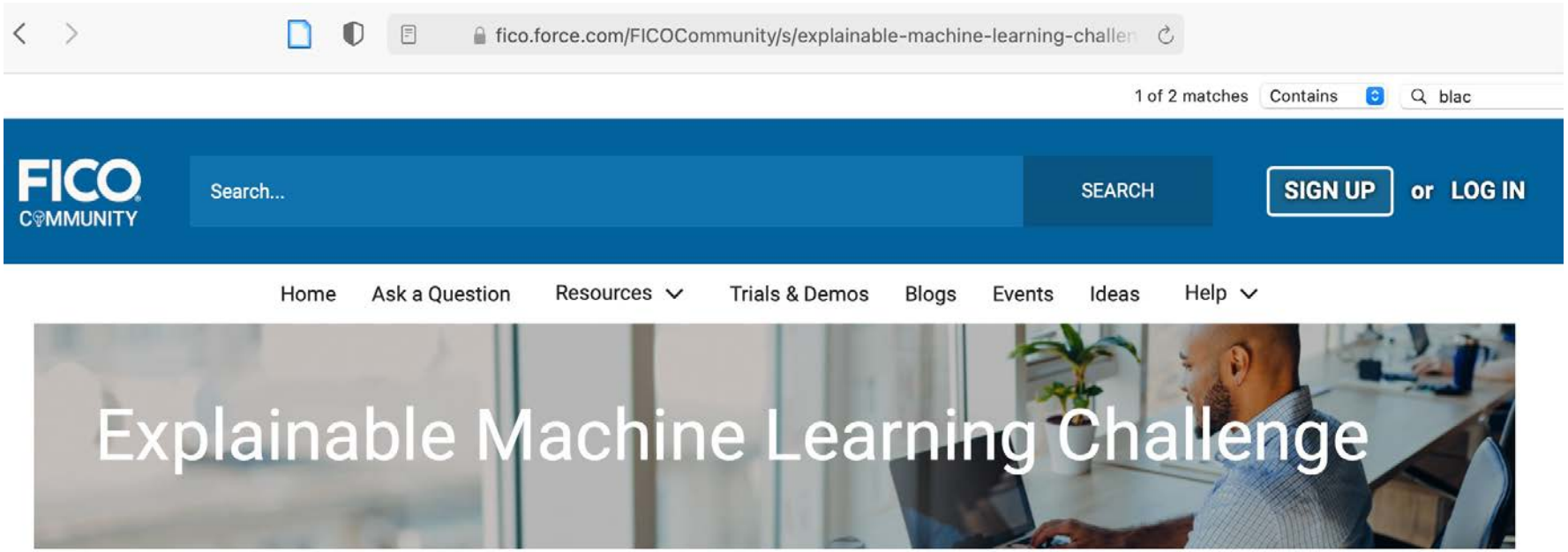
ProtoPNet, Chen et al. 2019

Where are black boxes more accurate?

- Challenge 1: Interpretable neural networks for computer vision

  Interpretable neural network accuracy = black box accuracy

- Challenge 2: Really hard benchmark datasets



Home Equity Line of Credit (HELOC) Dataset

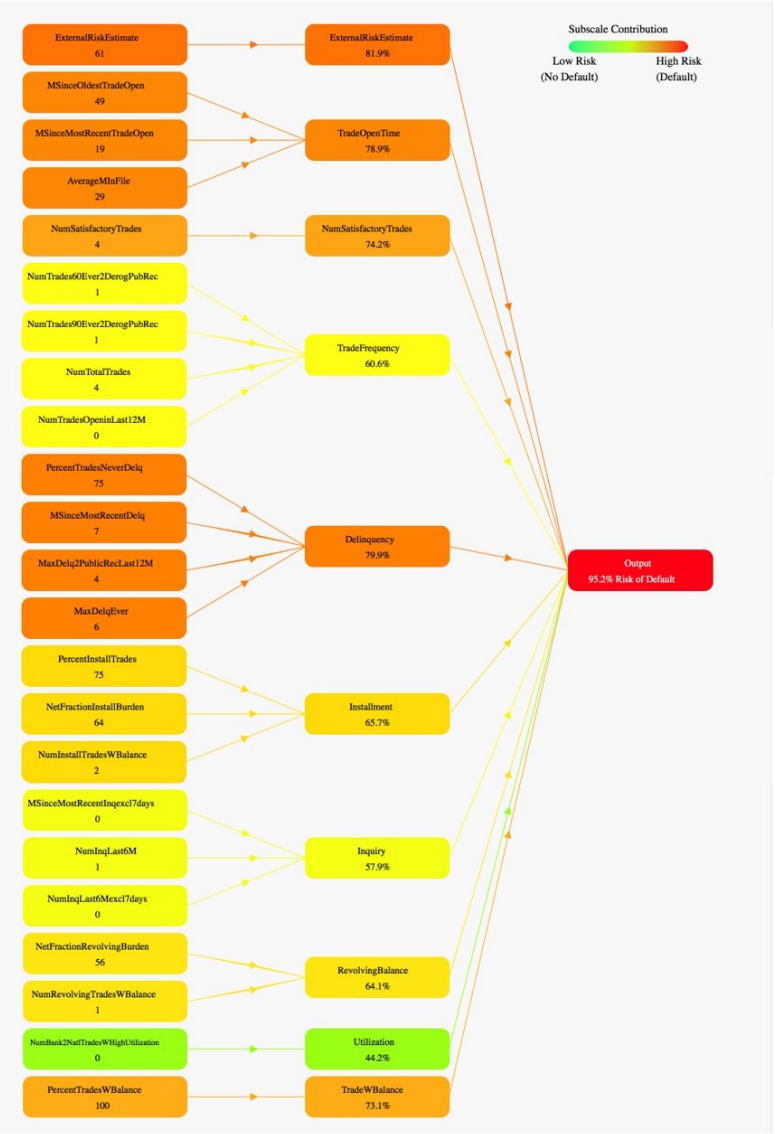
This competition focuses on an anonymized dataset of Home Equity Line of Credit (HELOC) applications made by real homeowners. A HELOC is a line of credit typically offered by a bank as a percentage of home equity (the difference between the current market value of a home and its purchase price). The customers in this dataset have requested a credit line in the range of $5,000 - $150,000. The fundamental task is to use the information about the applicant in their credit report to predict whether they will repay their HELOC account within 2 years. This prediction is then used to decide whether the homeowner qualifies for a line of credit and, if so, how much credit should be extended.

About the data

- ~10K loan applicants
- Factors:
  - External Risk Estimate
  - Months Since Oldest Trade Open
  - Months Since Most Recent Trade Open
  - Average Months In File
  - Number of Satisfactory Trades
  - Number Trades 60+ Ever
  - Number Trades 90+ Ever
  - Number of Total Trades
  - Number Trades Open In Last 12 Months
  - Percent Trades Never Delinquent
  - Months Since Most Recent Delinquency
  - Max Delinquency / Public Records Last 12 Months
  - Max Delinquency Ever
  - Percent Installment Trades
  - Net Fraction of Installment Burden
  - Number of Installment Trades with Balance
  - Months Since Most Recent Inquiry excluding 7 days
  - Number of Inquiries in Last 6 Months
  - Number of Inquiries in Last 6 Months excluding 7 days.
  - Net Fraction Revolving Burden. (Revolving balance divided by credit limit.)
  - Number Revolving Trades with Balance
  - Number Bank/Natl Trades with high utilization ratio
  - Percent of Trades with Balance

Best black box accuracy
(boosted decision trees) 73%

Best black box AUC
(2-layer neural network) .80



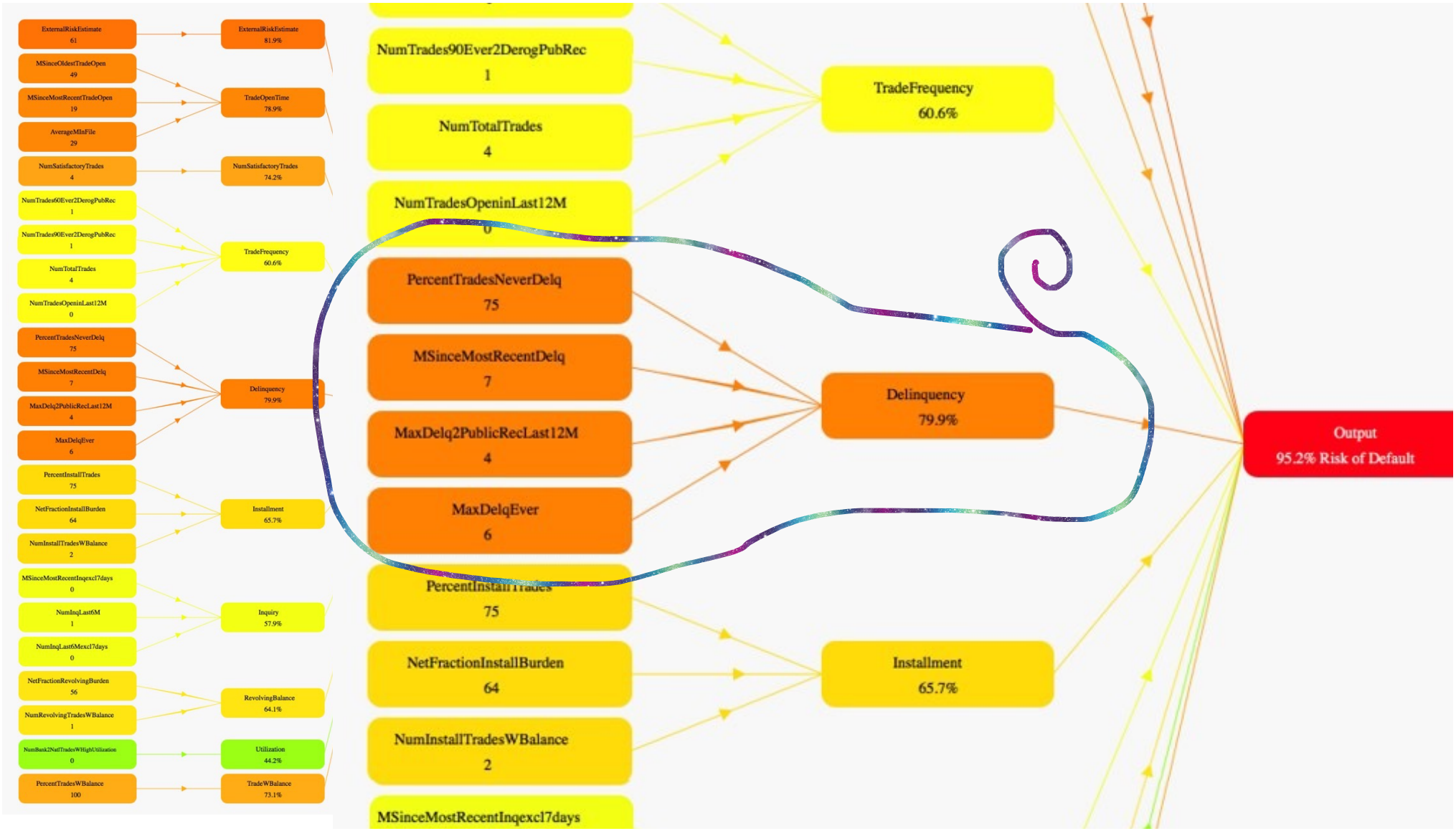
Best black box accuracy
(boosted decision trees) 73%

Best black box AUC
(2-layer neural network) .80

IBM model (First Prize): 6 questions
Accuracy = 71.8%
AUC = .62

Our entry (won FICO Recognition Prize):
Two-layer additive risk model
10 subscales + one final scoring model

Accuracy = 73.8%
AUC = .806


ExternalRiskEstimate
61
MSinceOldestTradeOpen
49
MSinceMostRecentTradeOpen
19
AverageMInFile
29
NumSatisfactoryTrades
4
NumTrades60Ever2DerogPubRec
1
NumTrades90Ever2DerogPubRec
1
NumTotalTrades
4
NumTradesOpeninLast12M
0
PercentTradesNeverDelq
75
MSinceMostRecentDelq
7
MaxDelq2PublicRecLast12M
4
MaxDelqEver
6
PercentInstallTrades
75
NetFractionInstallBurden
64
NumInstallTradesWBalance
2
MSinceMostRecentInqexcl7days
0
NumInqLast6M
1
NumInqLast6Mexcl7days
0
NetFractionRevolvingBurden
56
NumRevolvingTradesWBalance
1
NumBank2NatlTradesWHighUtilization
0
PercentTradesWBalance
100
ExternalRiskEstimate
81.9%
TradeOpenTime
78.9%
NumSatisfactoryTrades
74.2%
TradeFrequency
60.6%
Delinquency
79.9%
Installment
65.7%
Inquiry
57.9%
RevolvingBalance
64.1%
Utilization
44.2%
TradeWBalance
73.1%
NumTrades90Ever2DerogPubRec
1
NumTotalTrades
4
NumTradesOpeninLast12M
0
PercentTradesNeverDelq
75
MSinceMostRecentDelq
7
MaxDelq2PublicRecLast12M
4
MaxDelqEver
6
PercentInstallTrades
75
NetFractionInstallBurden
64
NumInstallTradesWBalance
2
MSinceMostRecentInqexcl7days
TradeFrequency
60.6%
Delinquency
79.9%
Installment
65.7%
Output
95.2% Risk of Default

Delinquency Subscores

Intervals	Points
0-59	+1.567
59-84	+1.012
84-89	+0.601
89-96	+0.366
96-Inf	-0.147
-7	0
-8	0
-9	0

PercentTradesNeverDelq

Intervals	Points
0-8	-0.058
9-17	-0.058
18-32	-0.22
33-47	-0.392
48-Inf	-0.482
-7	+0.198
-8	+0.137
-9	0

MSinceMostRecentDelq

Intervals	Points
0-3	+0.806
4-5	+0.806
6	+0.408
7-8	-0.147
9-Inf	-0.147
-7	0
-8	0
-9	0

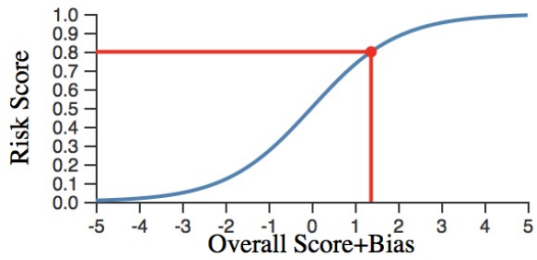
MaxDelq2PublicRecLast12M

Intervals	Points
0-2	-0.017
3	-0.147
4-5	-0.147
6	-0.147
7-Inf	-0.147
-7	0
-8	0
-9	0

MaxDelqEver

Overall Score	1.613
Bias	-0.237
Associated Risk (for subscale Delinquency)	**79.8%**

Activation Function



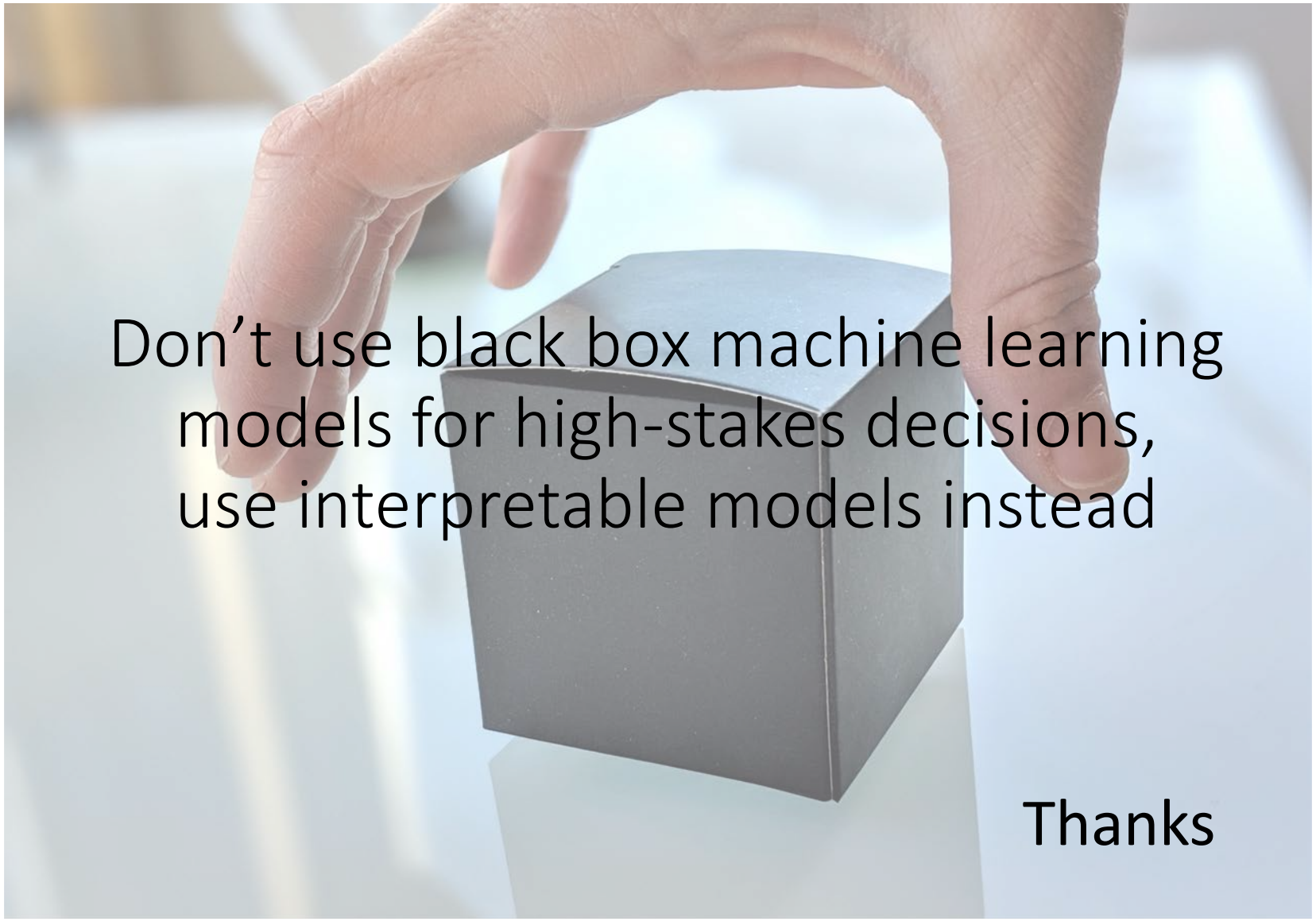
Where are black boxes more accurate?

- Challenge 1: Interpretable neural networks for computer vision

  Interpretable neural network accuracy = black box accuracy

- Challenge 2: Really hard benchmark datasets

  Interpretable model accuracy = black box accuracy


Don’t use black box machine learning models for high-stakes decisions, use interpretable models instead
Thanks