

# Machine Learning Bias

---

AbdulMajedRaja RS

# Outline

- Recognizing the Problem
- What's Machine Learning Bias?
- Definition of “Fairness”
- Interpretable Machine Learning

Thoughts?

What if I told you Computers can lie?

Would you believe me?

# *Biased*-Google Translation at Work

The screenshot displays the Google Translate web interface. At the top, the 'Google Translate' logo is visible. Below it are buttons for 'Text' and 'Documents'. The interface is divided into two main sections, each showing a translation pair.

**Top Section:**

- Left Panel (English - DETECTED):** The text 'She is the doctor who treated her Nurse' is entered. Below it is a character count '40/5000' and a copy icon.
- Right Panel (FINNISH):** The translated text is 'Hän on lääkäri, joka kohteli sairaanhoitajaansa'. Below it is a character count '40/5000' and icons for copy, edit, and share.

**Bottom Section:**

- Left Panel (FINNISH):** The text 'Hän on lääkäri, joka kohteli sairaanhoitajaansa' is entered. Below it is a character count '47/5000' and a copy icon. A suggestion is shown: 'Did you mean: Hän on lääkäri, joka kohteli *sairanhoitaja ansa*'.
- Right Panel (ENGLISH):** The translated text is 'He is a doctor who treated his nurse'. Below it is a character count '47/5000' and icons for copy, edit, and share.

# The Problem - Samples

—

But Wait, Why is this concerning?

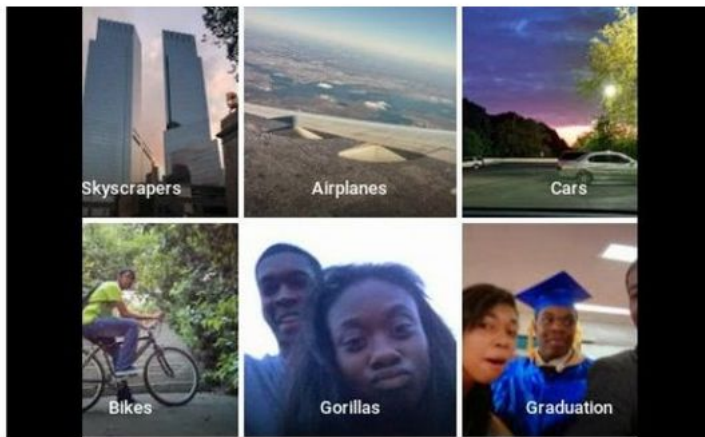
After all, This is just Google Translate

# *Biased*-Google Photos App at Work

## Google apologises for Photos app's racist blunder

🕒 1 July 2015

f 🗨️ 🐦 ✉️ Share



diri noir avec banan @jackyalcine · Jun 29  
Google Photos, y'all [redacted] My friend's not a gorilla.



813



394



\*\*\*

TWITTER

Mr Alcine tweeted Google about the fact its app had misclassified his photo

Perhaps, That's just Google.

Two instances can account for the entire industry, Huh?



# Microsoft's super-cool Teen Tweeting Bot Tay



# Much more!

## 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

Two Petty Theft Arrests

	
VERNON PRATER	BRISHA BORDEN
LOW RISK 3	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.

Oops, Got it!

There, definitely, is Bias!  
What's next?

# ML Bias - What

—

# What's Machine Learning Bias?

A Machine Learning Algorithm being  
“**unfair**” with its Predictions



A Machine Learning Algorithm missing  
“**Fairness**”

# ML Bias - (un)Fairness

---

# Disclaimer

No Common Consensus / Standard  
definition of Fairness

# ML Bias - un(Fairness)

- Group Fairness
- Individual Fairness



# ML Bias - Causes

—

# ML Bias - Causes

- Skewed sample
- Tainted examples
- Limited features
- Sample size disparity
- Proxies

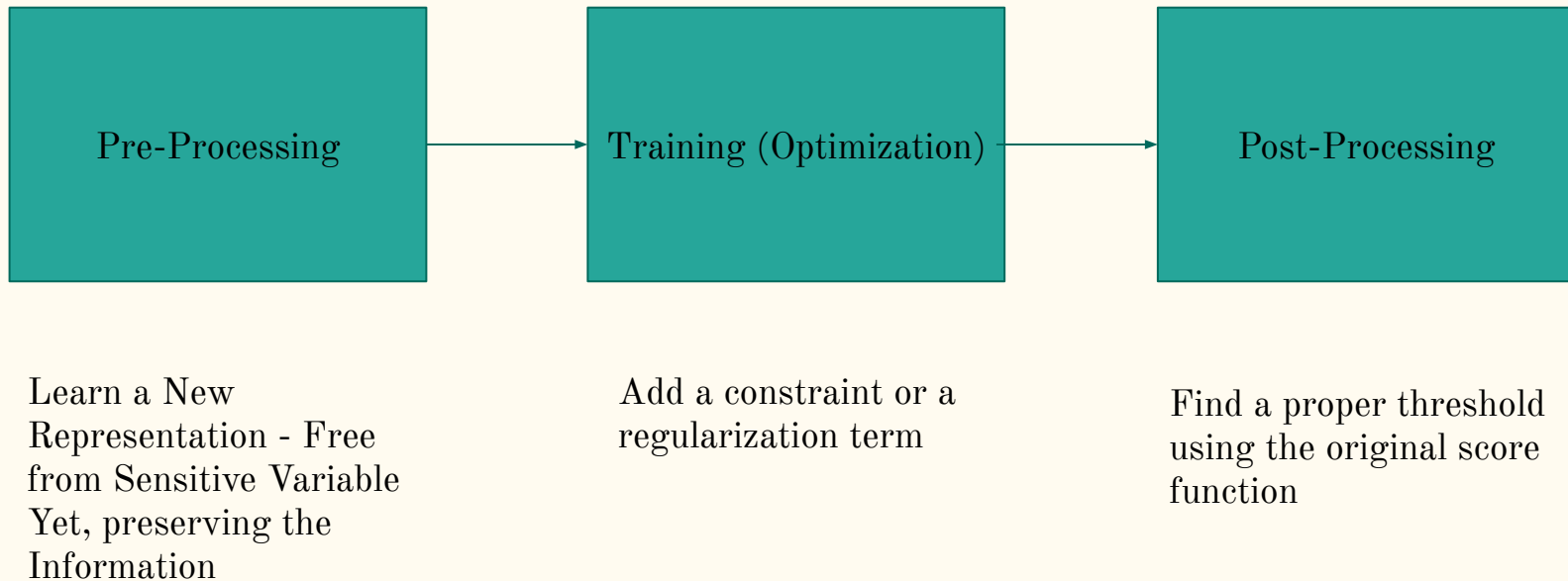
# ML Bias - Mitigate

—

Mitigation

Also means, Improving Fairness

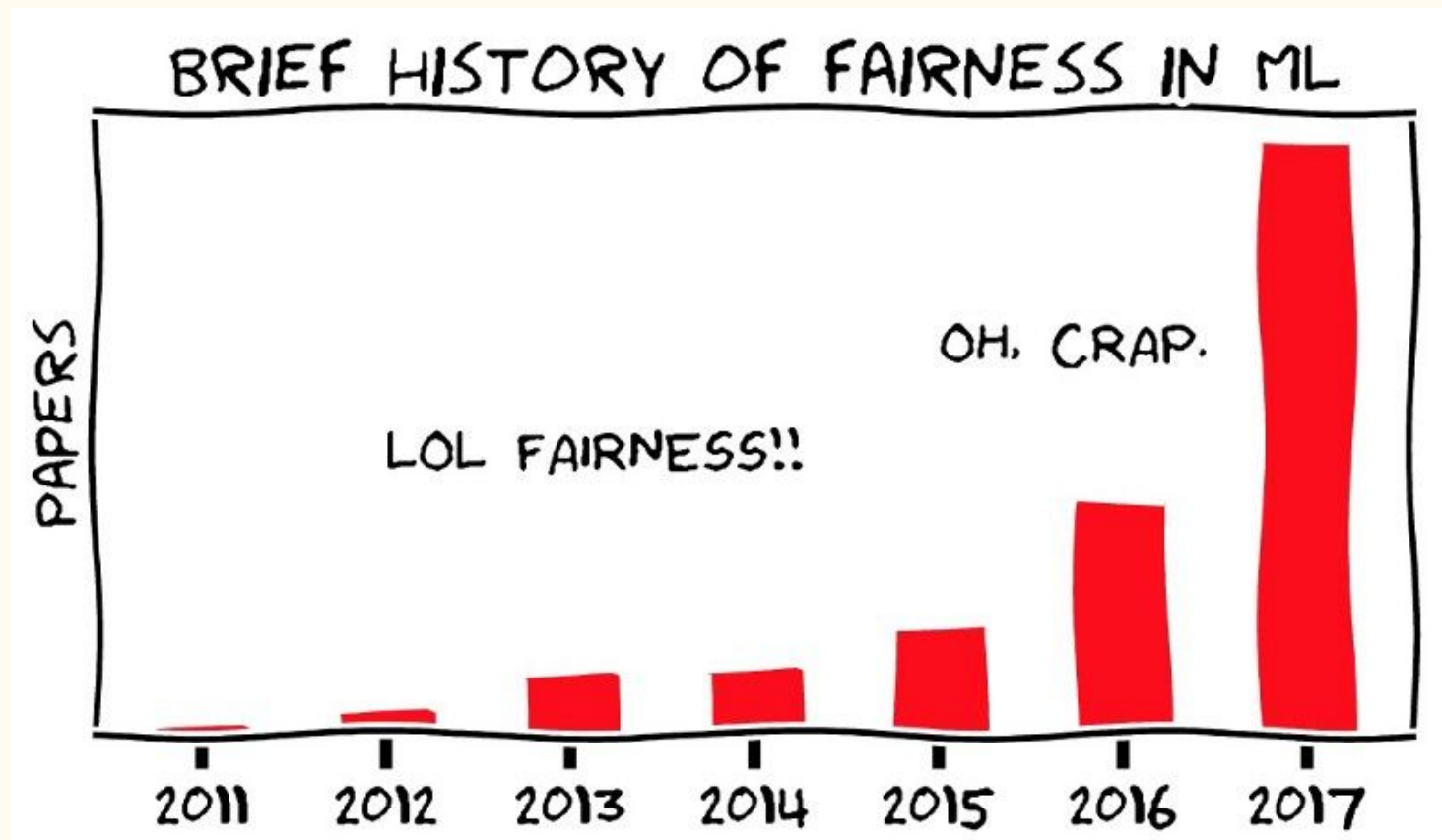
# ML Bias - Improving Fairness



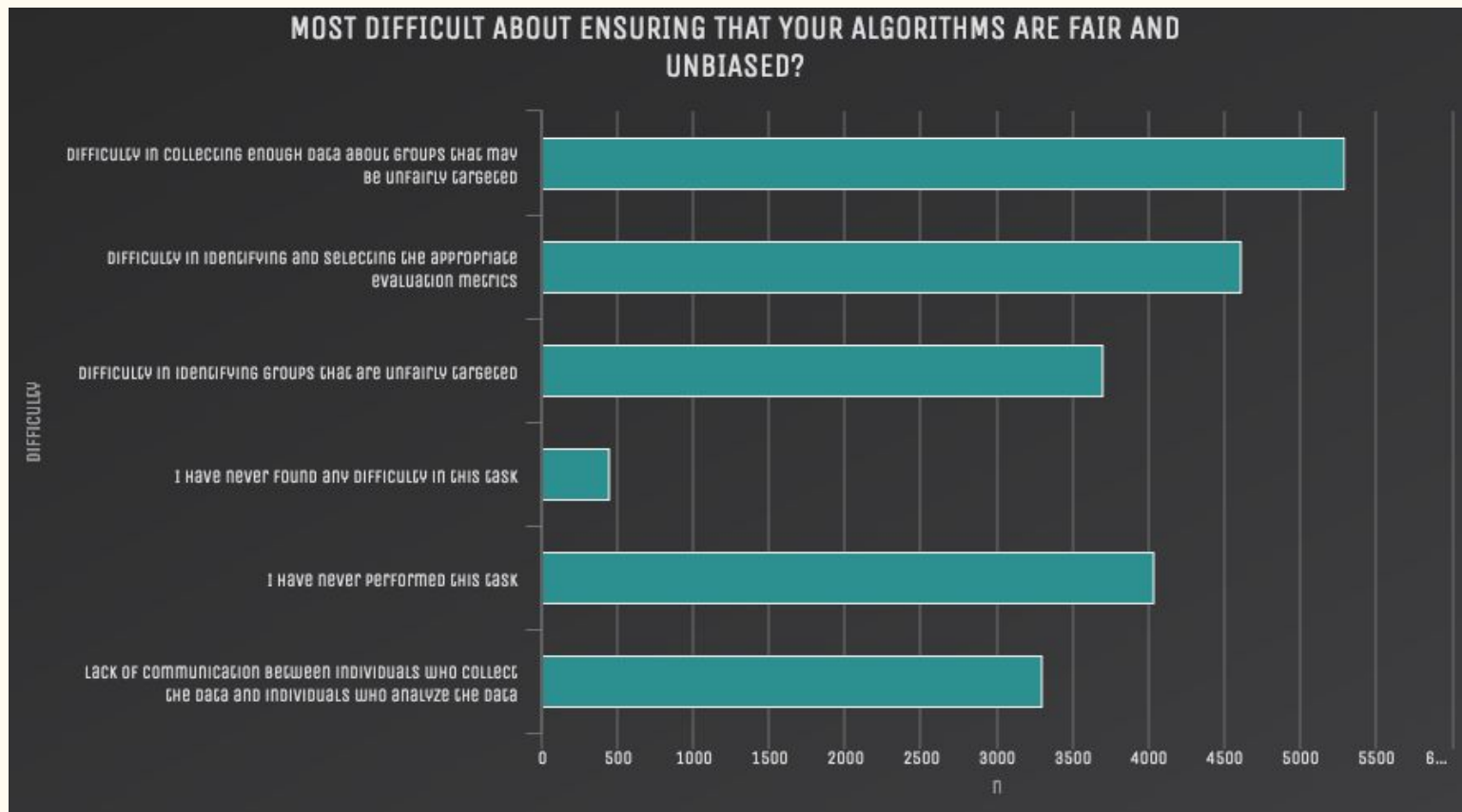
# ML Bias - Happening

—

# Mention of ML Fairness in Research Papers



# Difficulties in ensuring ML Algorithm is unbiased

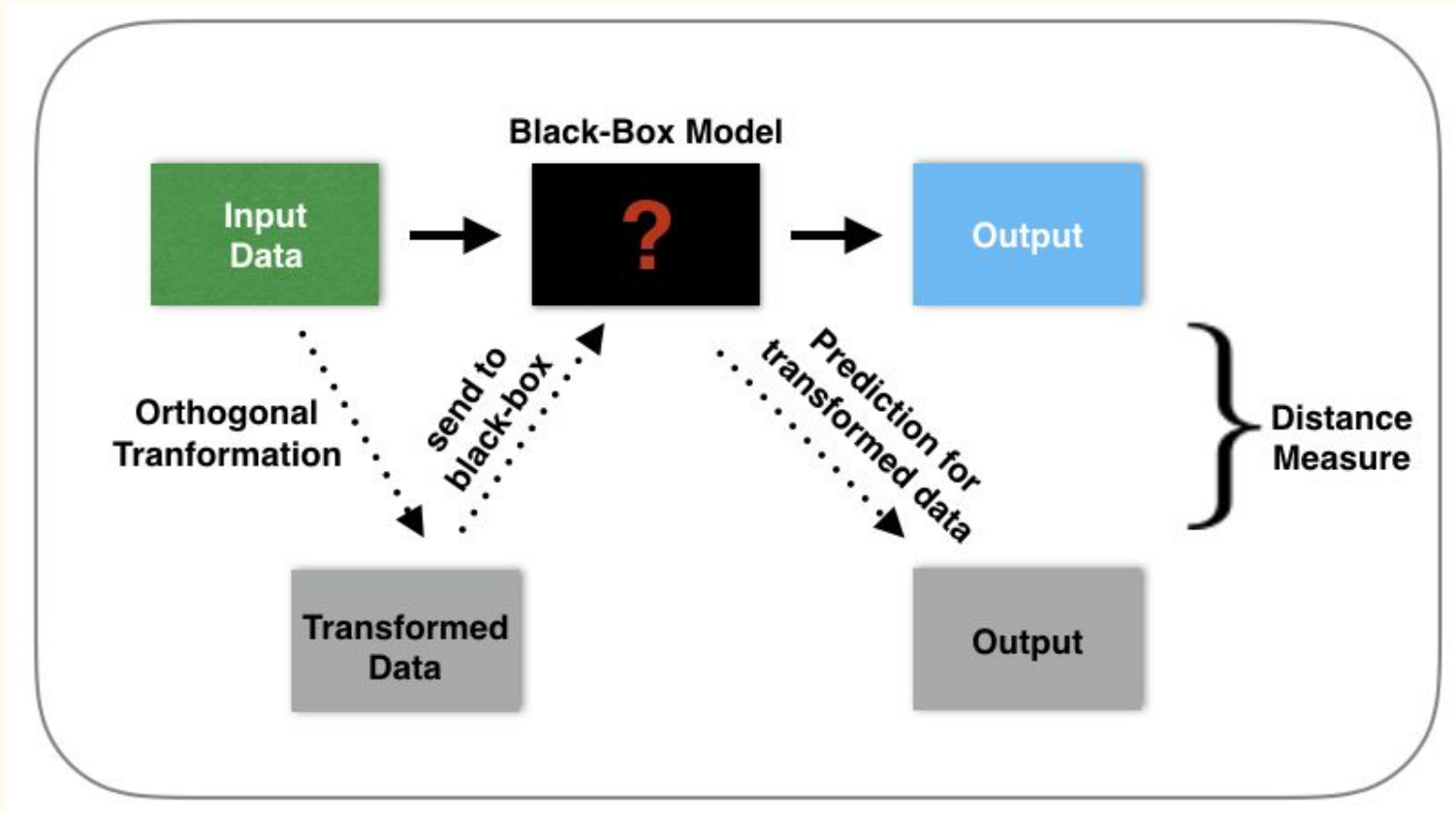




# Interpretable Machine Learning



# Today - Modelling Architecture



# IML - Definition

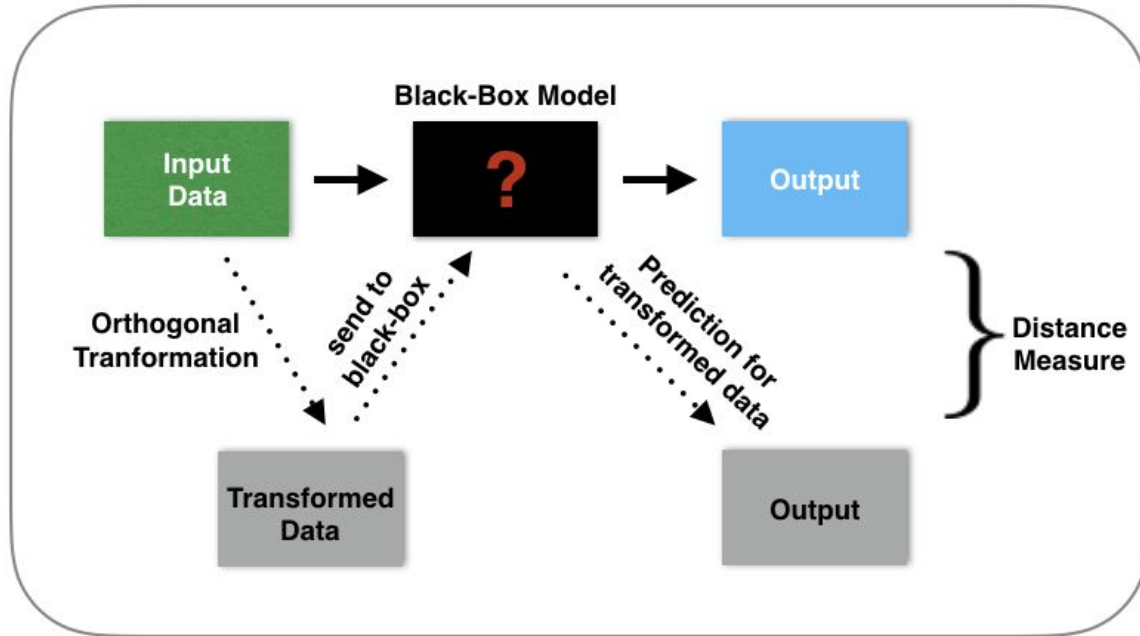
Interpretable Machine Learning refers to methods and models that make the behavior and **predictions of machine learning systems understandable to humans.**

# IML - Benefits

- **Fairness:** Ensuring that predictions are unbiased and do not implicitly or explicitly discriminate against protected groups. An interpretable model can tell you why it has decided that a certain person should not get a loan, and it becomes easier for a human to judge whether the decision is based on a learned demographic (e.g. racial) bias.
- **Privacy:** Ensuring that sensitive information in the data is protected.
- **Reliability or Robustness:** Ensuring that small changes in the input do not lead to large changes in the prediction.
- **Causality:** Check that only causal relationships are picked up.
- **Trust:** It is easier for humans to trust a system that explains its decisions compared to a black box.

# Modelling Architecture - with IML

## Overview of the attribute significance procedure

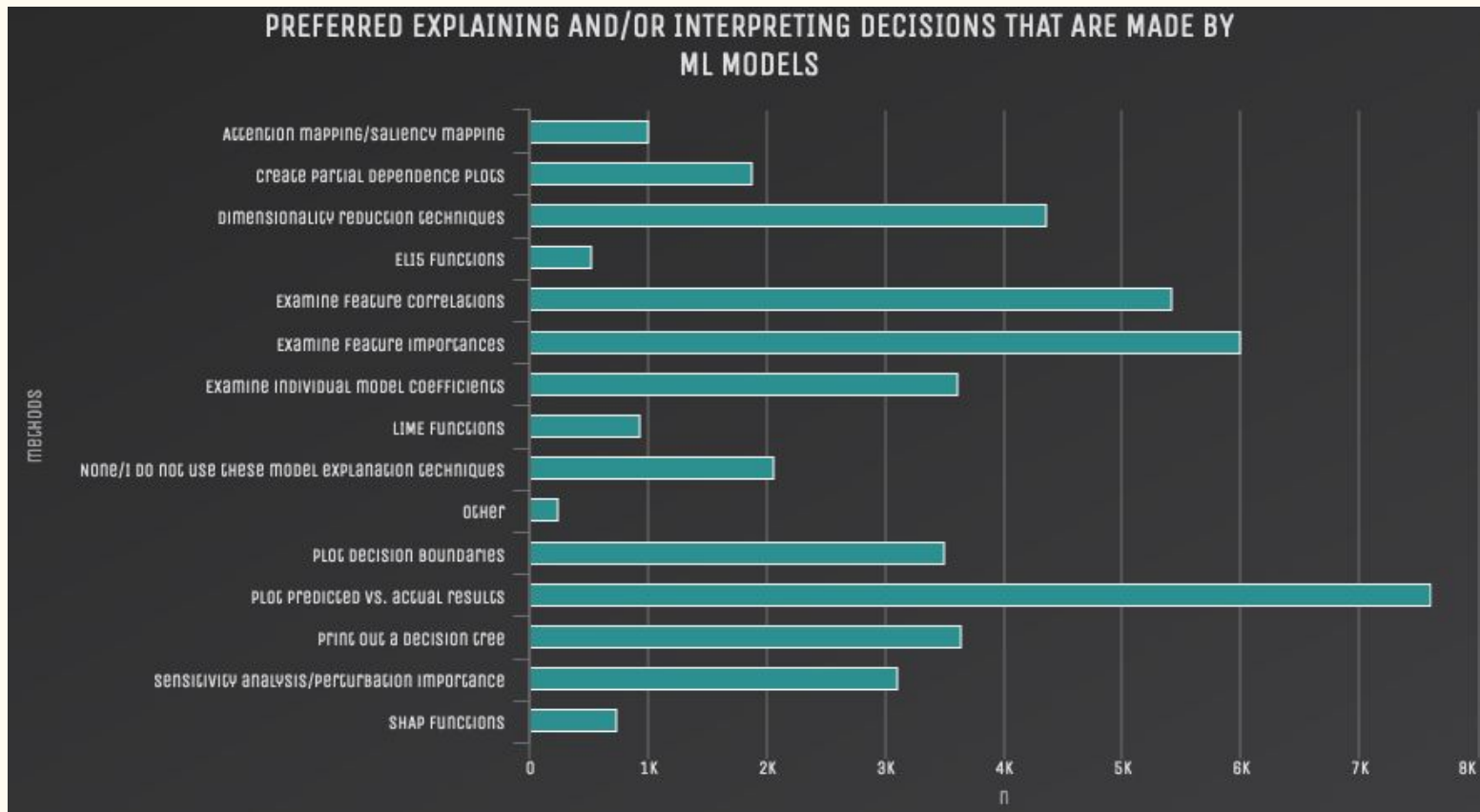


### Attribute Variable Significance



Figure S-4: Figure shows the input attribute ranking across all the ranking methods in FairML for the Bank's probability-of-default model audit.

# Preferred Explaining - Model Interpretation



# References

- <https://developers.google.com/machine-learning/fairness-overview/>
- <https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb>
- [https://www.youtube.com/watch?v=fMym\\_BKWQzk](https://www.youtube.com/watch?v=fMym_BKWQzk)
- <https://www.kaggle.com/nulldata/ml-bias-impl-perspective-recommendation#media-coverage-about-bias-in-ml>
- Doshi-Velez, Finale, and Been Kim. “Towards a rigorous science of interpretable machine learning,” no. ML: 1–13. <http://arxiv.org/abs/1702.08608> ( 2017)
- <https://christophm.github.io/interpretable-ml-book/>
- <https://github.com/adebayoj/fairml/>

Thank you!