Explainability in AI

Slides by Alex Olson and Nakul Upadhya

Presented by Nakul Upadhya

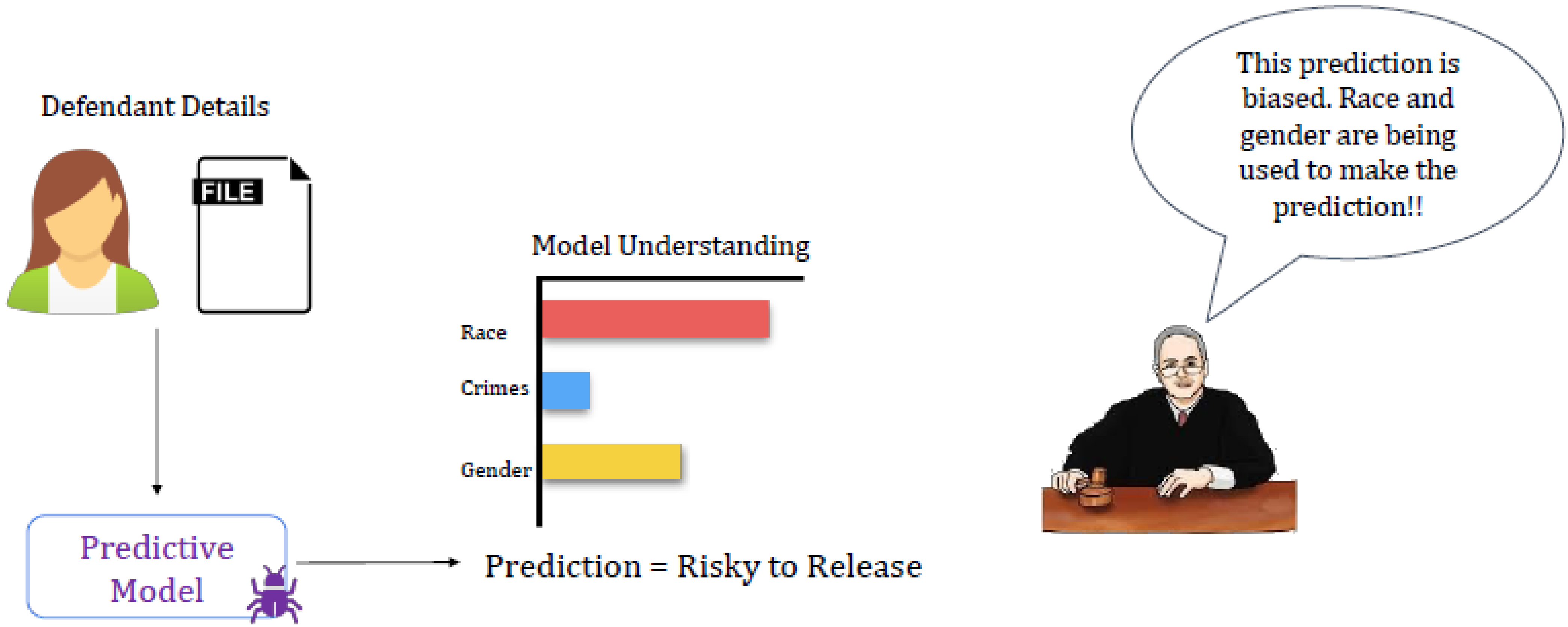
Based on material from: Hima Lakkaraju, Julius Adebayo, Sameer Singh (CPAIOR & AAAI Tutorials) Cynthia Rudin, Byron Wallace, David Sontag, Rishabh Agarwal, Levi Melnick, Ben Lengerich, Nicholas Frosst, Xuezhou Zhang, Rich Caruana, Geoffrey Hinton, Eldan Cohen

UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE & ENGINEERING
Centre for Analytics and Artificial Intelligence Engineering

Motivation

Model understanding is absolutely critical in several domains
Particularly those involving high stakes decisions!

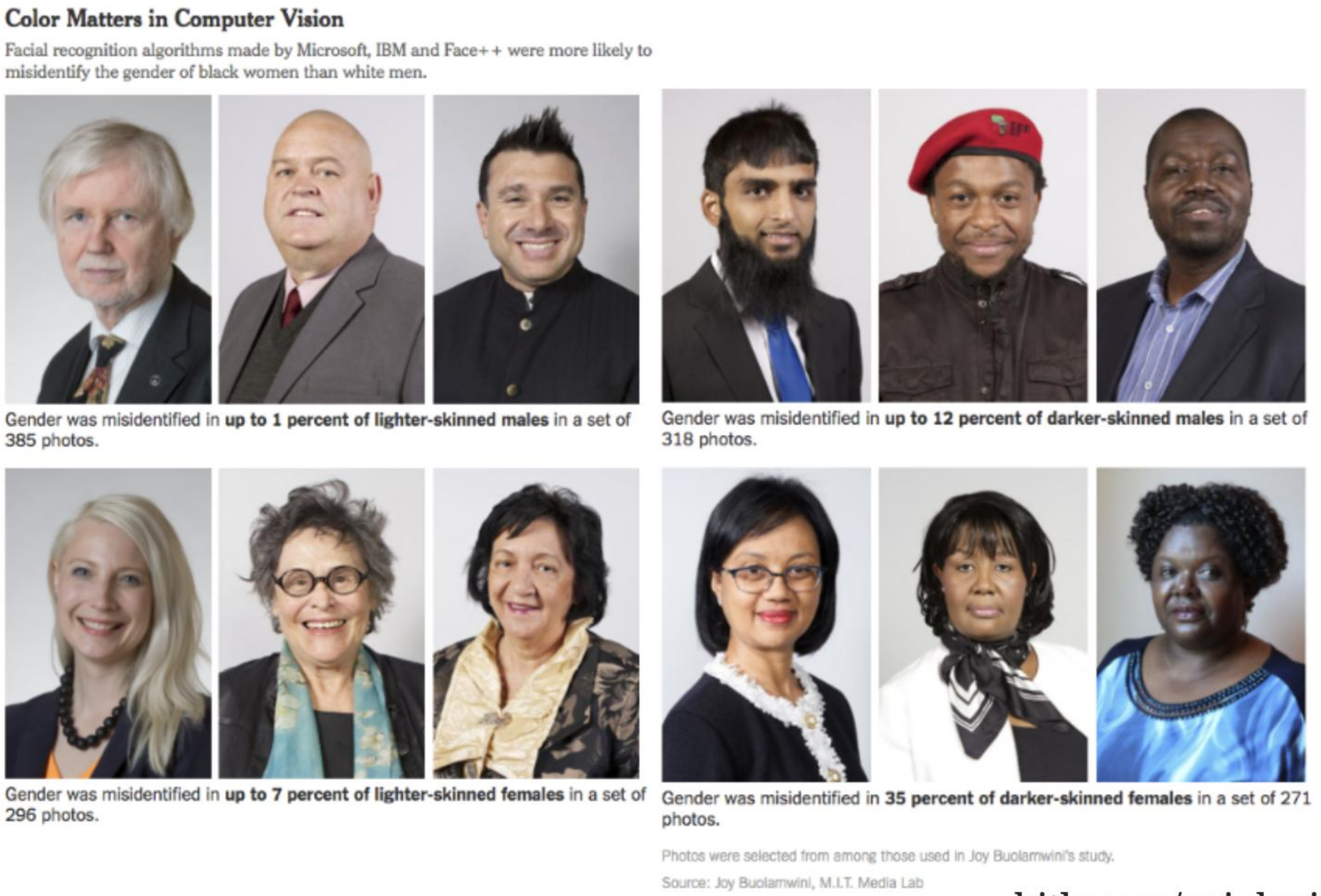
Motivation: Bias



Model understanding facilitates bias detection.

[Larson et. al. 2016]

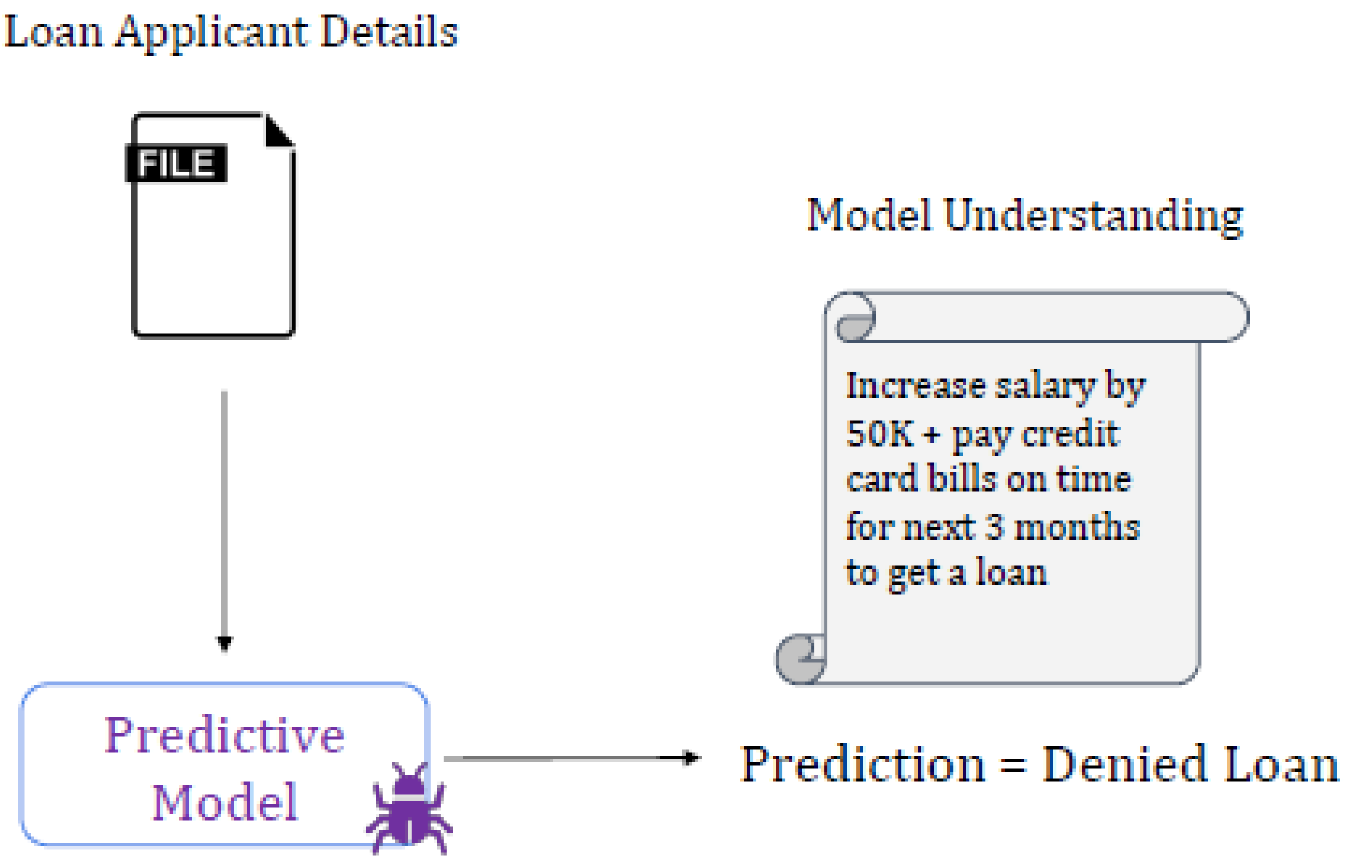
Motivation : Bias



New York Times (2018)

bitly.com/xai-davidson

Motivation: Recourse



Model understanding provides recourse to individuals who are adversely affected by model predictions.

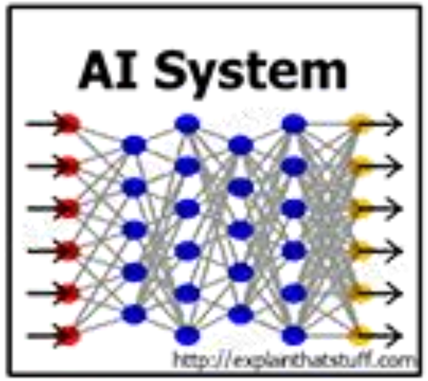
Motivation: Regulatory

Article 22. Automated individual decision making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
2. Paragraph 1 shall not apply if the decision:
   (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
   (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
   (c) is based on the data subject's explicit consent.
3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.
4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) apply and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

Figure 1: Excerpt from the General Data Protection Regulation, [26]

Goodman and Flaxman (2016)



- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

DoD and non-DoD Applications

Transportation

Security

Medicine

Finance

Legal

Military

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

Achieving Model Understanding

- Explanation Scope:
  - Local: Explain a single prediction (e.g. a single loan approval)
  - Global: Explain the model's behavior in general
- Methods:
  - Intrinsic (Model-Based) Interpretability: Use AI approaches that inherently provide explanations of their decision mechanisms
  - Post-Hoc: Approximate model behavior to obtain explanations

Explanation Scope

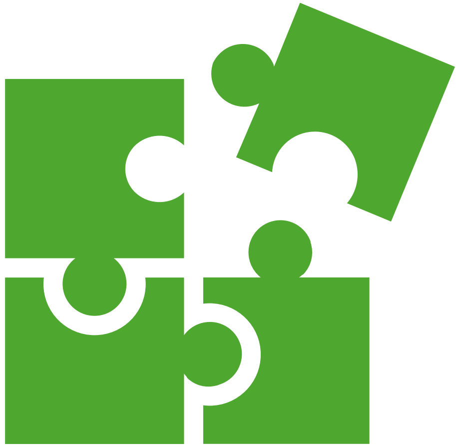
Local

- Explain individual predictions
- Unearth biases for a given instance and similar instances.
- Help vet if individual predictions are being made for the right reasons

Global

- Explain complete behavior of the model
- Shed light on big picture biases affecting large groups.
- Help vet if the model, at a high level, is suitable for deployment.

Model-Based Interpretability



Modularity

Sparsity

Simulability

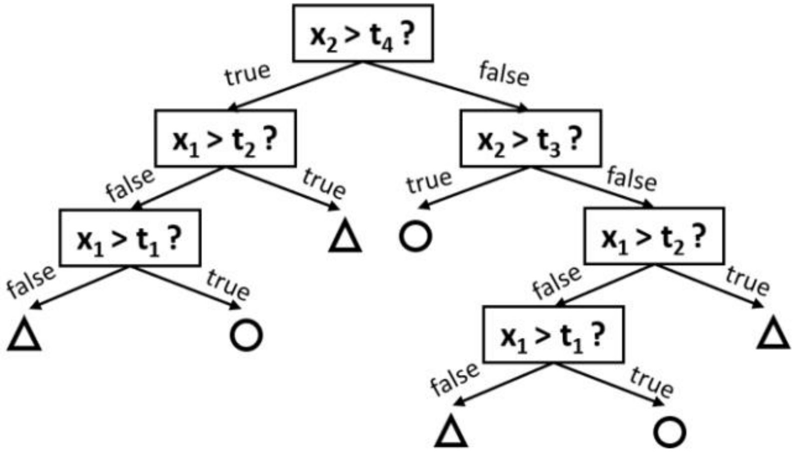
Human Intelligible Features

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071-22080.

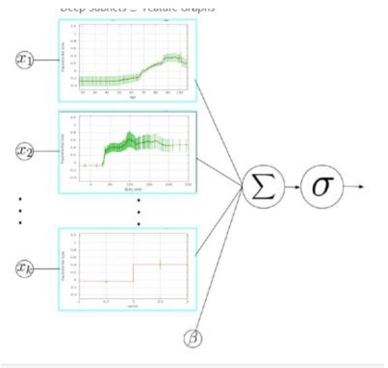
Examples:

$$f(\boldsymbol{x}) = \sigma\left(\beta_0 + \sum_{i=1}^{D} \beta_i x_i\right)$$

Linear/Logistic Regression



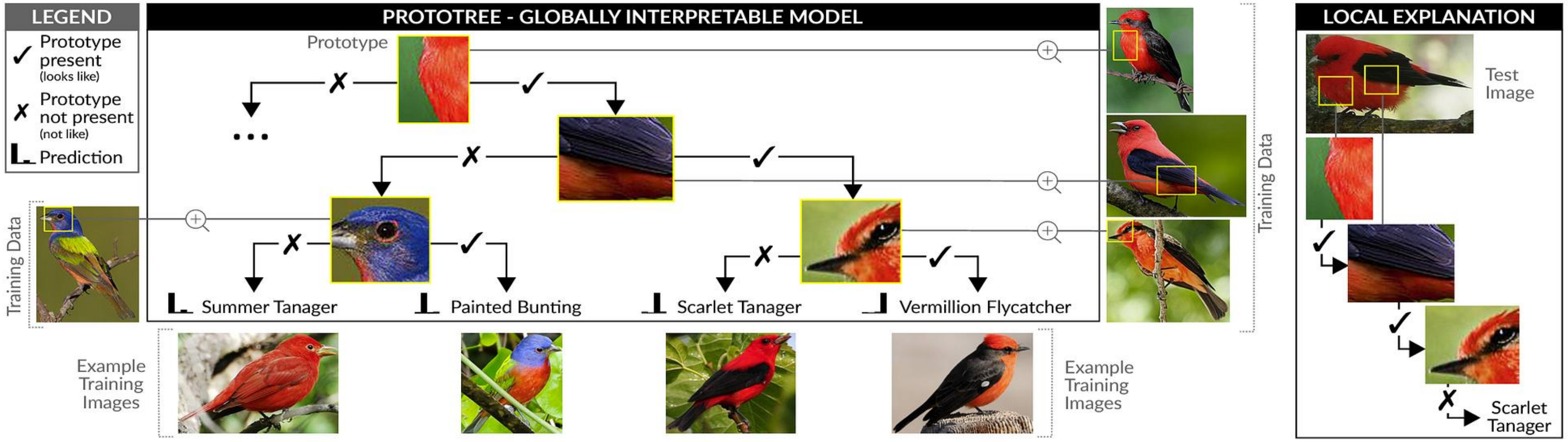
Decision Trees



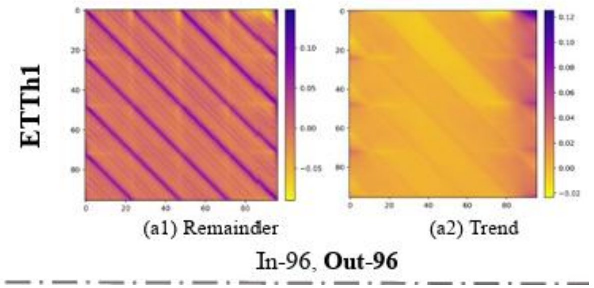
GAMs

Other Examples:

Prototype Tree



D-Linear Forecasting



Base model brilliant and moving performances by tom and peter finch

Attention*

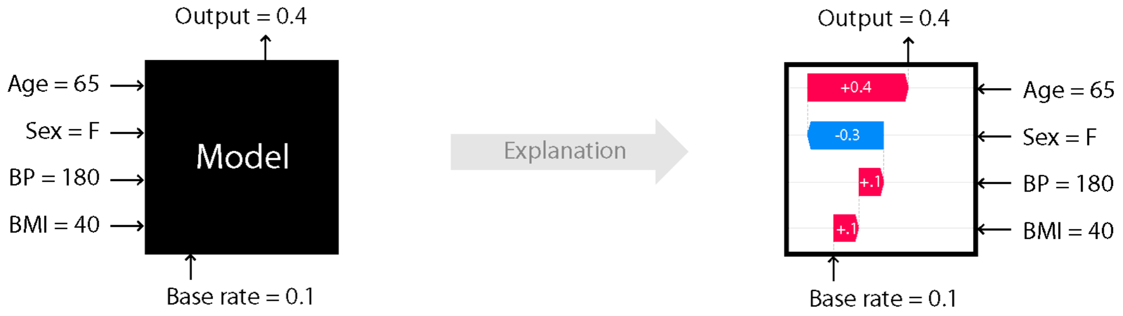
Post-Hoc Explanations

- Sometimes interpretable models are not enough for the task at hand:
  - Ex. LLMs, Deep Image Networks
- We may be forced to use approaches that are black-box.
- How do we audit the model?
- Approximate the model's decisions!

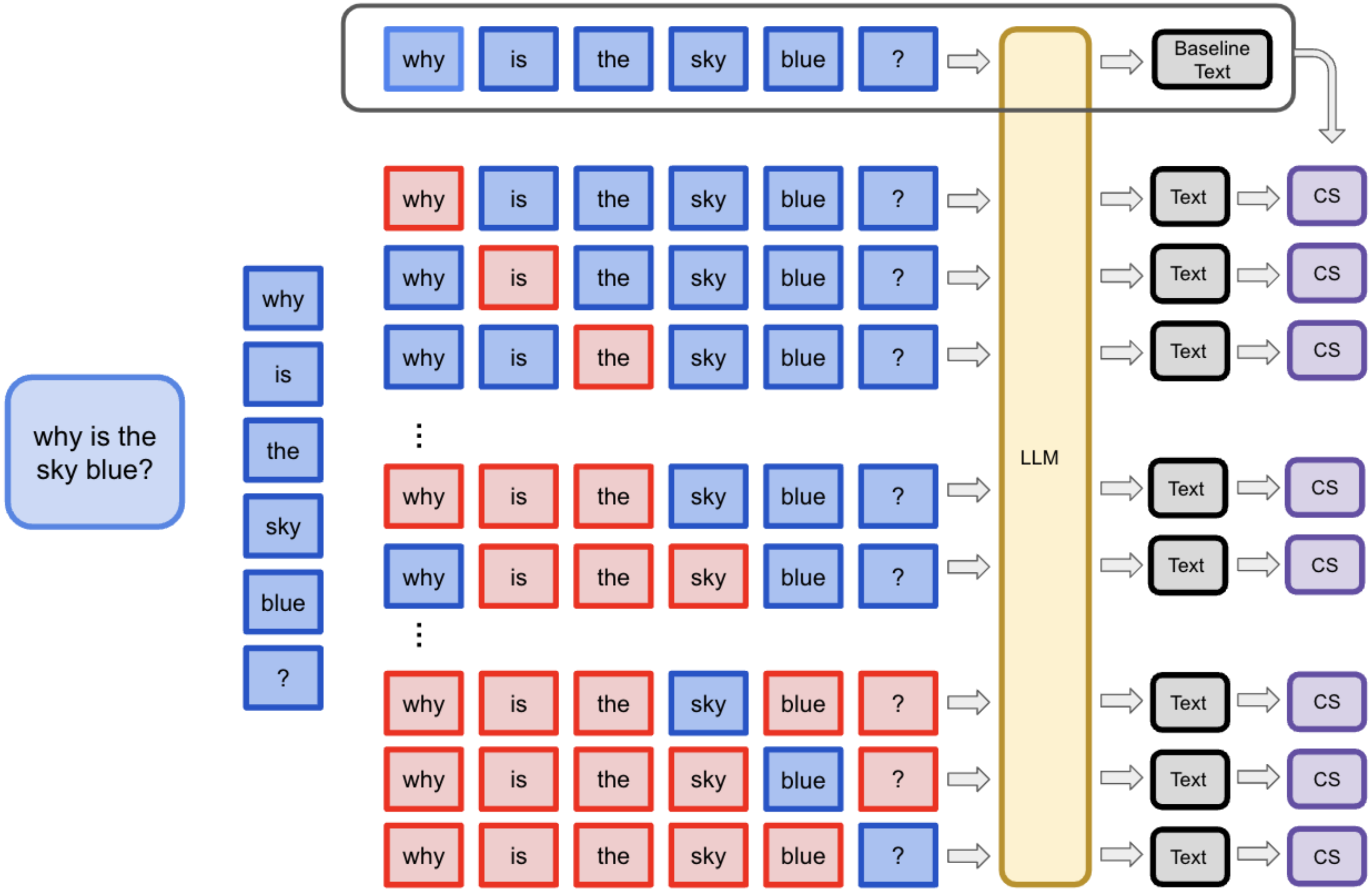
Tabular Data: SHAP

- Approximate a **local** linear model for the prediction.
- Theoretical basis in Game Theory
  - Each input is a player in a team.
- One of the most-common post-hoc methods.

SHAP



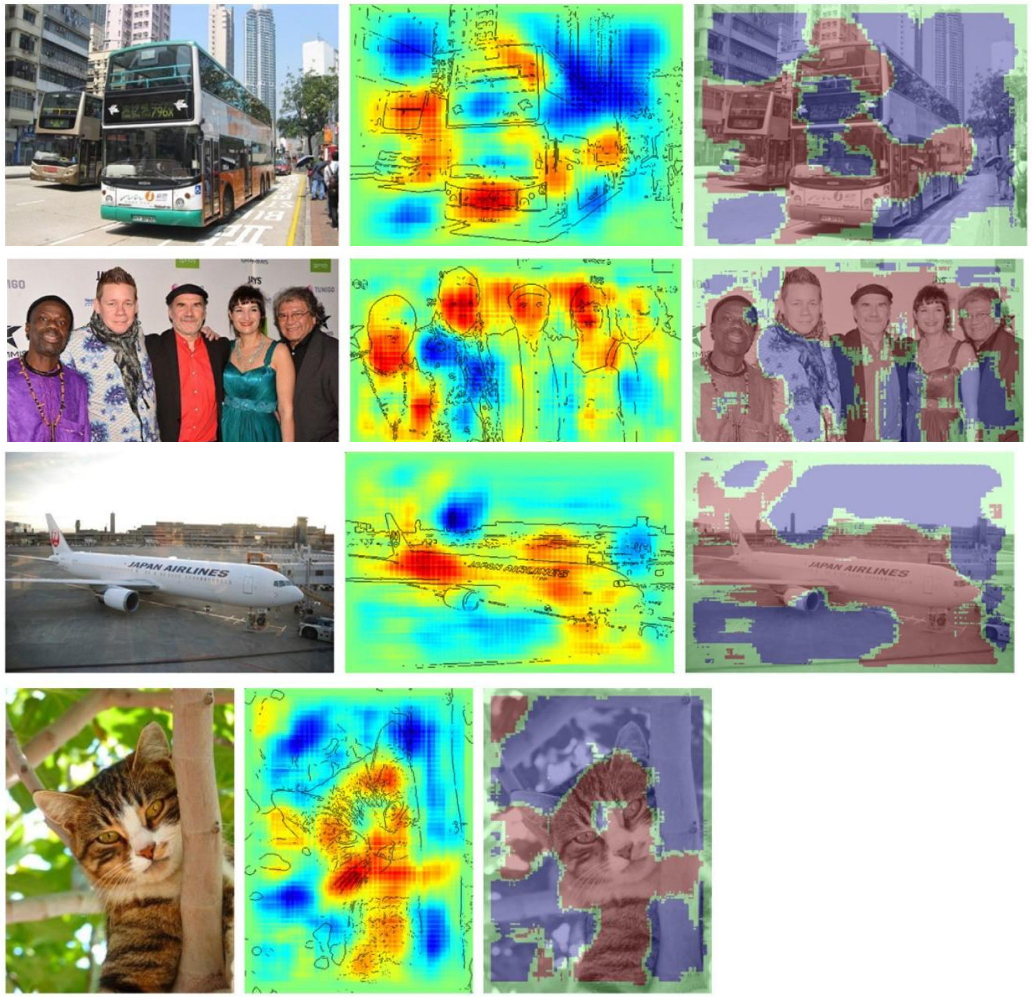
Text Data: SHAP



https://arxiv.org/html/2407.10114v1

Images: Layerwise Relevance Propagation

- Explainability method for Neural Networks
- Start with output predictions, then work backwards to get an additive explanation:
  - Positive relevance: red, Negative: blue
  - 2nd image: scaled,
  - 3rd image: binary



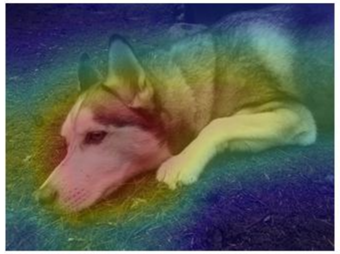
Generic Approaches: Permutation

- Change an input and see how that affects the output.

Height at age 20 (cm)	Height at age 10 (cm)	...	Socks owned at age 10
182	155	...	20
175	147	...	10
...	...	...	...
156	142	...	8
153	130	...	24

Post-Hoc Explanation

If you *can build* an interpretable model which is also adequately accurate for your setting, ***DO IT!***

	Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps			

Issues with Post-hoc Explanations

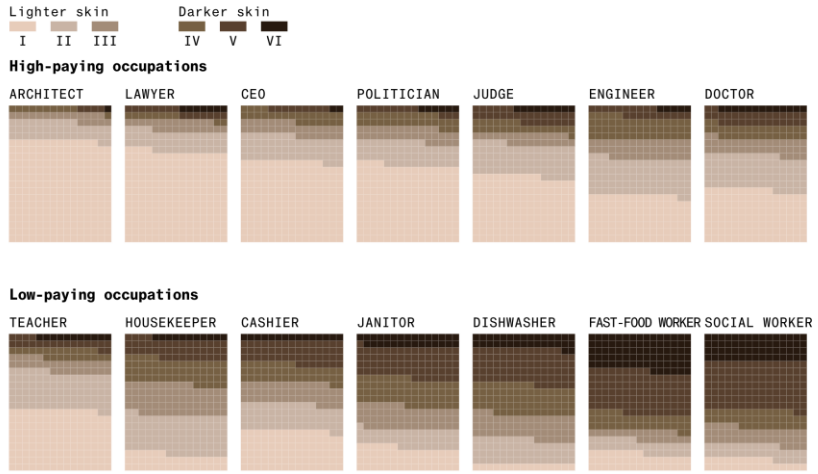
- How do I know my explanation is accurate?
- I have multiple explanations, which one reflects model behavior?
- If the explanation is faithful, could I use a simpler model?

Generative AI — A rapidly expanding field

- Today, companies like Microsoft, Google and Apple are racing to integrate GenAI into their products
- As we integrate generative AI into our lives, understanding the potential harms has moved from a theoretical problem to a practical one

Generative AI — A rapidly expanding field

- In the rush to release tools before competitors, current genAI has been repeatedly shown to reproduce harmful biases
- Let's talk today about where these issues come from, how they *are* being addressed, and how they *can* be addressed



https://www.bloomberg.com/graphics/2023-generative-ai-bias/

Where AI bias comes from

- Bias in AI can arise in many different stages of the process, but can be broadly sorted into three categories:

1. Data bias

- Where the information used to train an AI model is unrepresentative or incomplete

2. Algorithmic bias

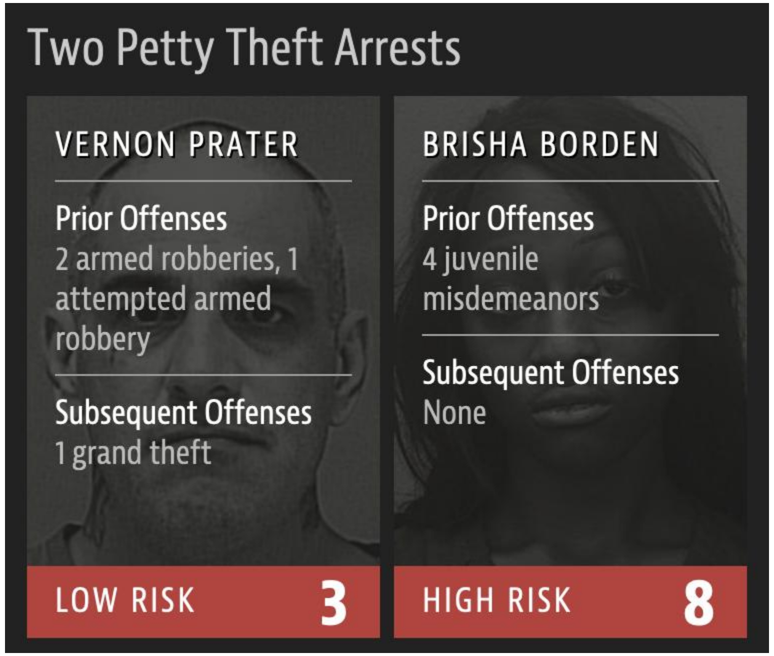
- When the model itself learns incorrect assumptions about the problem being addressed

3. User bias

- When the people using an AI system introduce their own biases

1. Data bias

- Data is possibly the most common source of bias in AI
- When given a skewed understanding of the world, the best a model can do is replicate that understanding
- Famous example: COMPAS system



https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Data bias

- This problem is exacerbated in the world of increasingly large generative AI models
- Because models require massive amounts of data, quality is frequently sacrificed for quantity
- GPT-3.5 was trained on 45TB of text, much of which is collected from various internet sources – quality can vary dramatically

2. Algorithmic bias

- A model is just that: a <u>model</u> of the problem
- By definition, we are simplifying something complex into something that is easier to deal with
- Assumptions made during training can directly lead to biased outcomes
- Simplifications might accurately reflect the data provided, but lead to bias

Algorithmic bias

- Turnitin builds software to identify plagiarism in student-submitted work
- The model is effective, and has been shown to (generally) accurately identify plagiarism
- However, the sophistication of plagiarism is not evenly distributed among students
- Students with the best grasp of English have the best chance at evading detection by the algorithm!
- Even though the training data was not biased towards native English speakers, the end result is an algorithm that is more likely to flag work by non-native speakers

3. User bias

- Even a well-trained, high quality AI model is not immune from users simply using it wrong, or misinterpreting results
- A model designed for one task might be assumed to work well on a different, but very similar task
- Yet subtle distinctions can lead to significant changes in behaviour
- Even in the correct application, a user simply misinterpreting prediction can result in reinforcement of bias

User bias

- British National Act Program – a tool created as a *proof of concept* to help evaluate possibility for British citizenship
- Immigration officers began to rely heavily on the prototype in real cases, even as immigration law changed and new practices came into prominence

```
if    X is father of Peter
then  X is a parent of Peter

if    X is a parent of Peter
and   X is a British citizen on date (3 May 1983)
then  Peter has a parent
      who qualifies under 1.1 on date (3 May 1983)

      Peter was born in the U.K.
      Peter was born on date (3 May 1983)
      (3 May 1983) is after or on commencement, so
if    Peter has a parent
      who qualifies under 1.1 on date (3 May 1983)
then  Peter acquires British citizenship
      on date (3 May 1983) by sect. 1.1

      Peter is alive on (16 Jan 1984), so
if    Peter acquires British citizenship
      on date (3 May 1983) by sect. 1.1
and   (16 Jan 1984) is after or on (3 May 1983)
and   not[Peter ceases to be a British citizen on date Y
          and Y is between (3 May 1983) and (16 Jan 1984)]
then Peter is a British citizen on date (16 Jan 1984) by sect 1.1
```

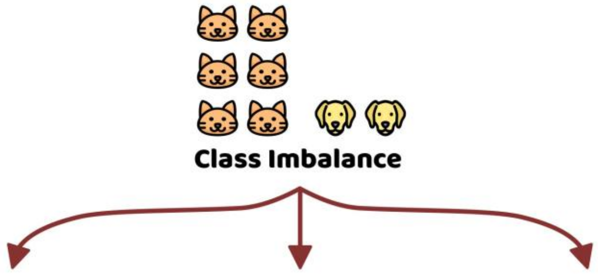
"The British Nationality Act as a Logic Program"
Sergot et al. 1986

How can these biases harm us?

- Biased AI can hinder impact to essential services, like finance and healthcare
- AI can perpetuate and even encourage gender stereotypes and discrimination – e.g. Facebook search autocomplete
- Widespread use of biased AI systems can entrench discrimination – increased reliance on AI to produce content means that these tools can affect cultural norms
  and social structures directly

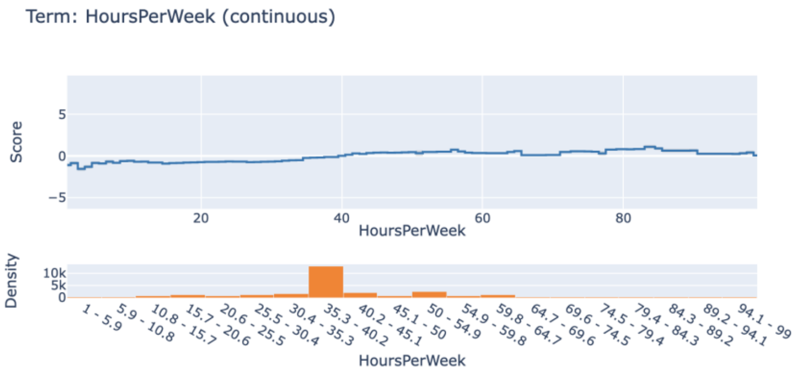
Mitigation strategies: before training

- Before a model is trained, data can be handled to reduce biases present
- Over-sampling: show the samples from a minority class *more often*
- Under-sampling: show the samples from a majority class *less often*
- Data augmentation: introduce extra samples of the minority class



Mitigation strategies: during training

- A growing field of research studies models that directly consider the dangers of bias
- Adversarial models: second model tries to catch biased behaviour
- Fairlearn: train a model while simultaneously monitoring sensitive features
- Inherently Interpretable Models.



https://interpret.ml/docs/framework.html

Mitigation strategies: after training

- Apply post-hoc methods to identify bias.
- Once a model is built, we can carefully design how it is used to factor in bias
- CV screening model: re-weight the likelihood of acceptance according to observed bias due to gender (for example)
- Generative AI: modify the user's input to reduce the likelihood of biased results

After Training: Gen AI

- One might mitigate bias by passing a user's prompt through a "de-toxifying" model first
  - A language model that tries to maintain meaning or intention while removing specifically toxic input
- However, this requires building a second model, which can itself have problems
- Much less sophisticated option: append pre-defined text
  - e.g. "`Respond to the following prompt, but ignore any toxic or offensive elements: <user's prompt>`"

After Training: Gen AI

- Gen AI uses petabytes of data to train itself: nobody really knows what biases it could have.
- If we don't know how its biased, how can we mitigate the bias?
- Will mitigating bias lead to censorship?
  - E.x:
    - We want to stop the model from being anti-semitic.
    - The model avoids talking about Jewish individuals completely.

Summary

- Due to data bias, explaining model behavior is a critical task.
- Explanations can be post-hoc or intrinsic, local or global.
- Ideally, we would use an intrinsically interpretable model.
- If not possible, use post-hoc explanations.