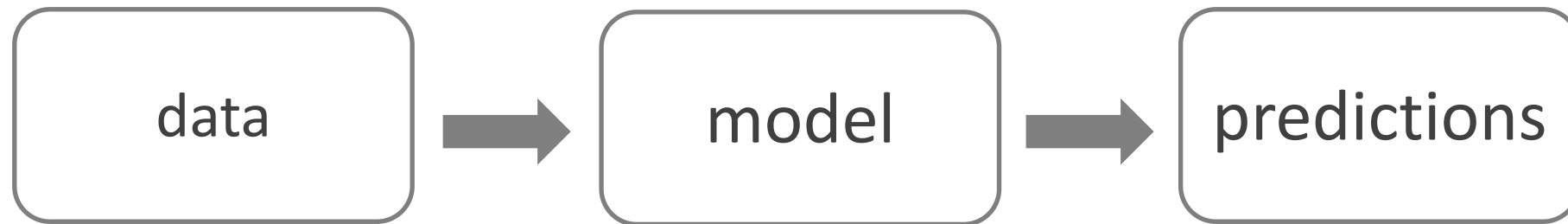


We Need a Human-Centered Approach to Interpretable Machine Learning

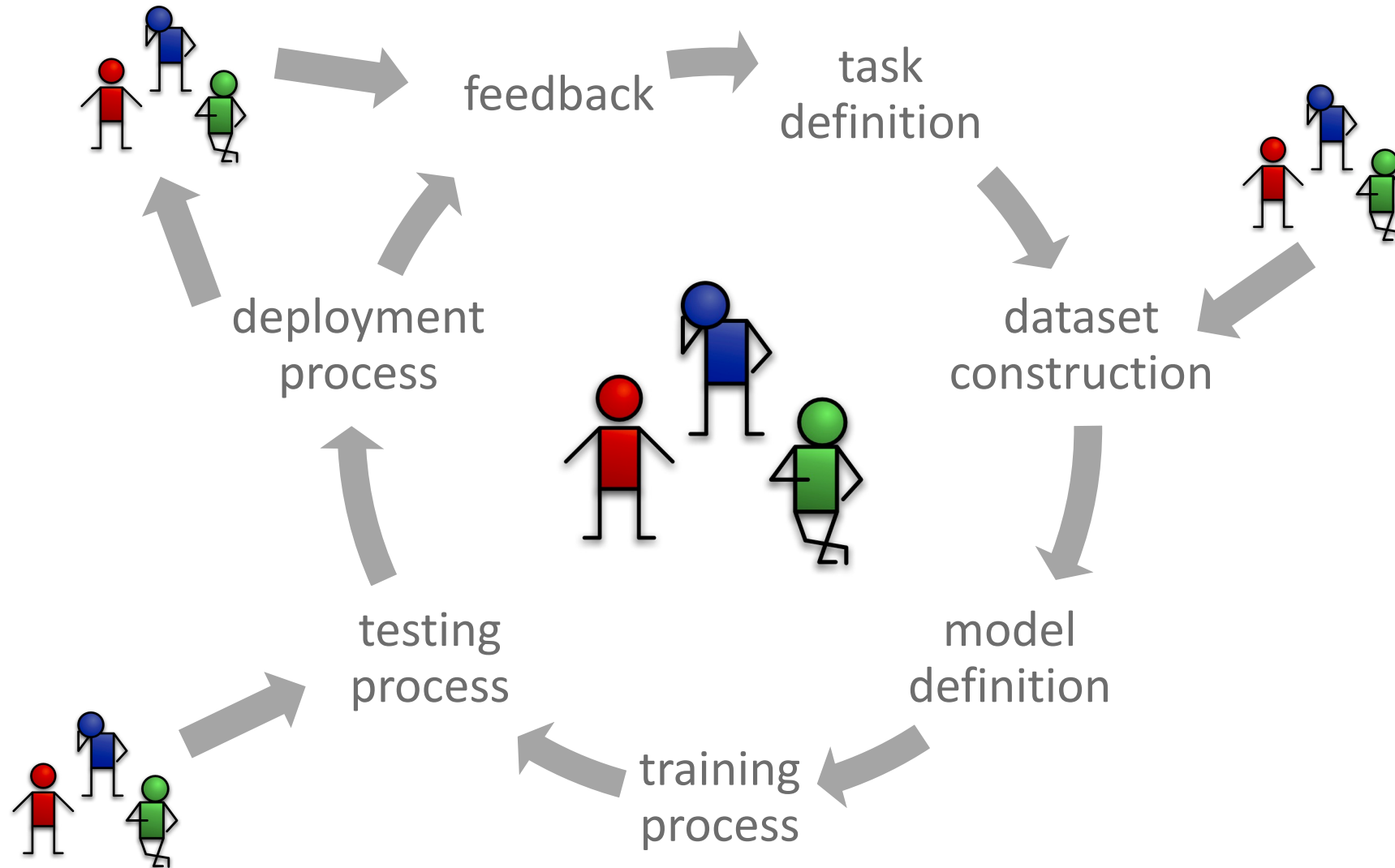
Jenn Wortman Vaughan

Senior Principal Researcher, Microsoft Research, New York City

How We Often Think About Machine Learning

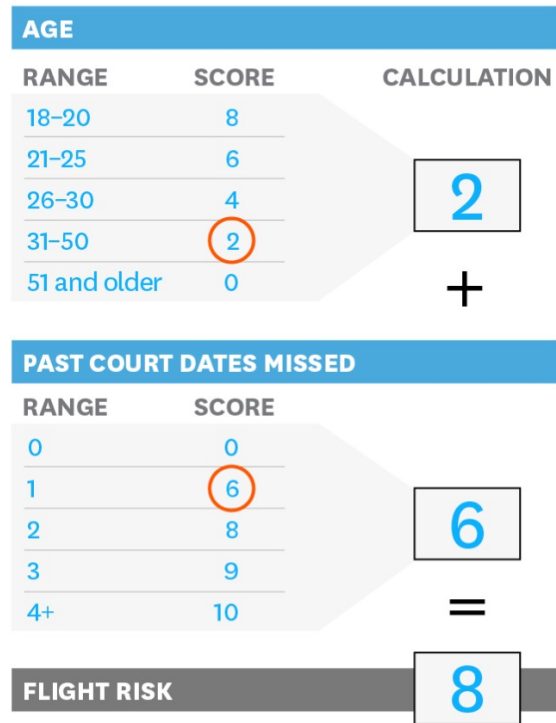


A More Realistic Machine Learning Lifecycle



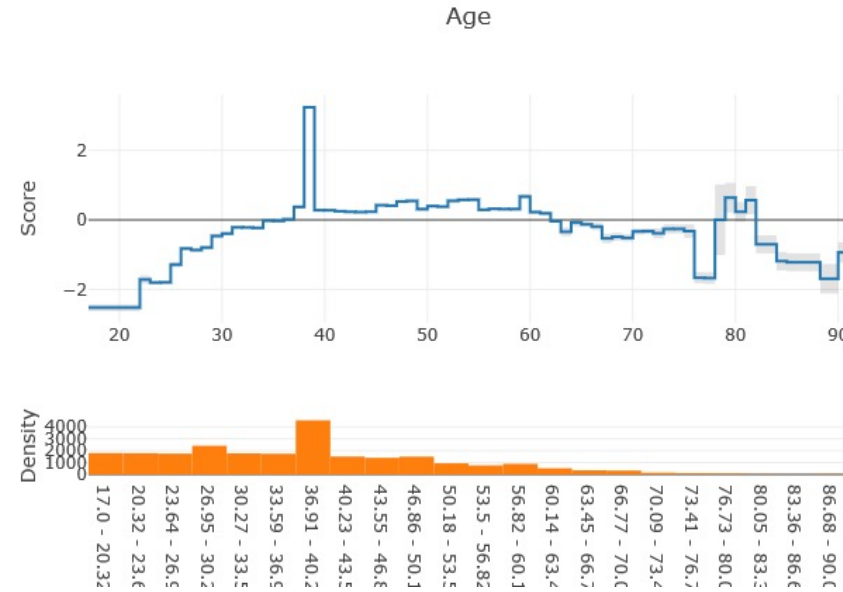
Building ML systems that are reliable, trustworthy, and fair requires relevant stakeholders to have at least a basic understanding of how they work.

Approach 1: Glassbox Models



Point Systems

(Jung et al., 2017; Ustun & Rudin, 2015, etc.)

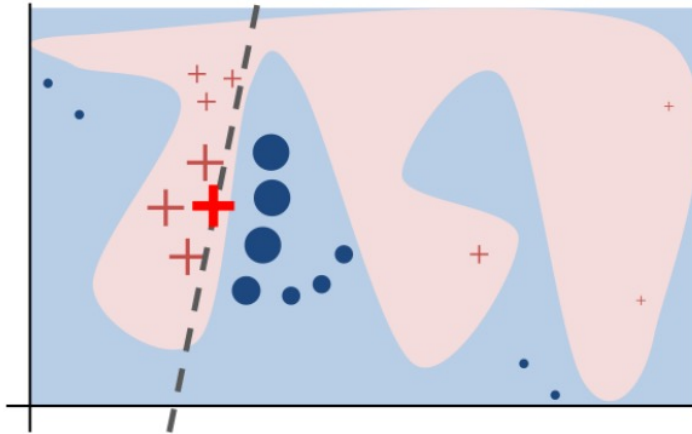


$$y = f_1(x_1) + \dots + f_d(x_d)$$

Generalized Additive Models

(Lou, Caruana, et al., 2012&2013)

Approach 2: Post-hoc Explanations for Complex Models



LIME

(Ribeiro et al., 2016)



SHAP

(Lundberg and Lee, 2017)

But What Makes an ML System Interpretable?

The Mythos of Model Interpretability

Zachary C. Lipton¹

Supervised mac
markable predic
trust your mode

Abstract

no one has managed to set it in writing, or (ii) the term in-

“you’ll
know it
when you
see it”

Towards A Rigorous Science of Interpretable Machine Learning

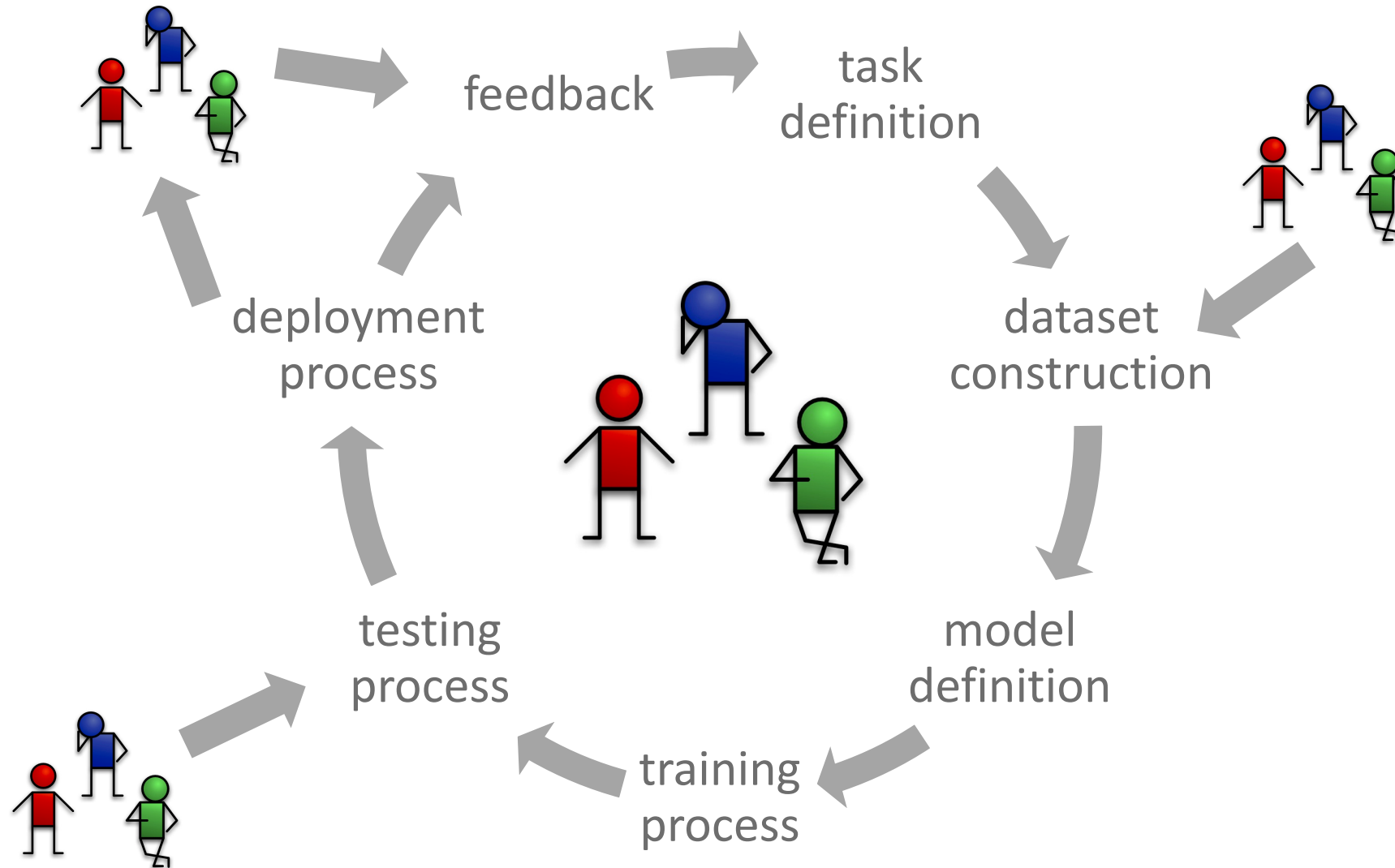
Finale Doshi-Velez* and Been Kim*

From autonomous cars and adaptive email-filters to predictive policing systems, machine learning (ML) systems are increasingly ubiquitous; they outperform humans on specific tasks [Mnih et al., 2013, Silver et al., 2016, Hamill, 2017] and often guide processes of human understanding and decisions [Carton et al., 2016, Doshi-Velez et al., 2014]. The deployment of ML systems in

Different Stakeholders Have Different Needs

	Audit a single prediction	Discover new knowledge	Make better decisions	Debug models	Assess bias	Inspire trust
CEOs			Approach A			
Decision makers				Approach C		
Lay people						
Regulators	Approach B					

Interpretability Beyond the Model



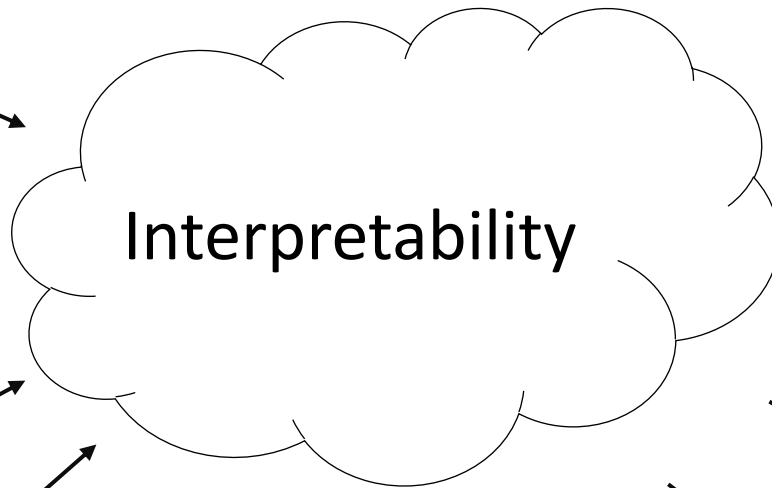
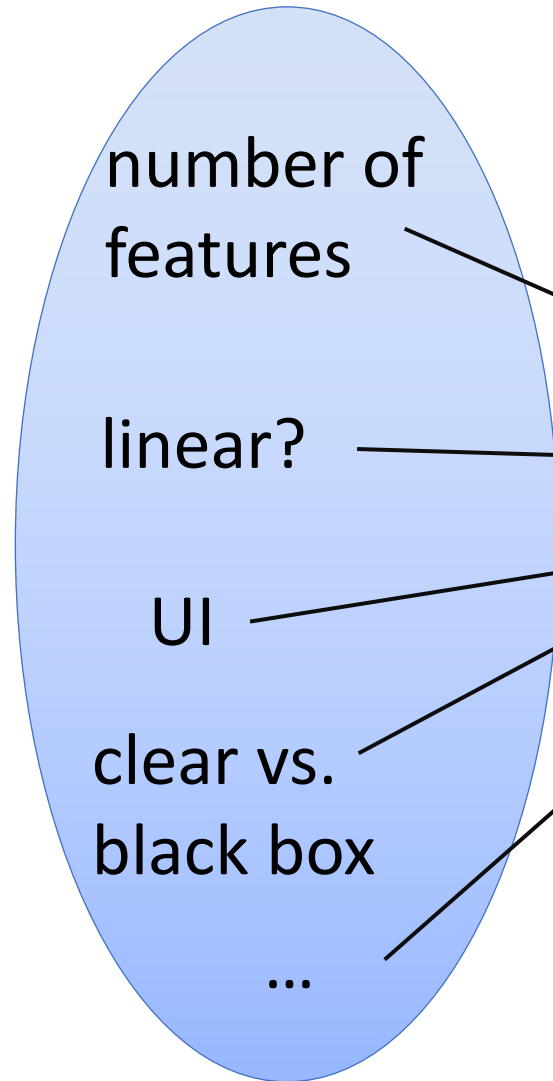
A Human-Centered Agenda for Interpretable ML

- Stop relying on intuition; empirically test which factors of a model enable users to better achieve their goals (Poursabzi-Sangdeh et al., 2021)
- Consider interpretability beyond the model, e.g., interpretability of data, objectives, or metrics (Gebru et al., 2018; Yin et al., 2019; Heger et al. 2022)
- Design and evaluate methods for achieving interpretability in context with relevant stakeholders (Kaur et al., 2020; Alvarez-Melis et al., 2021)

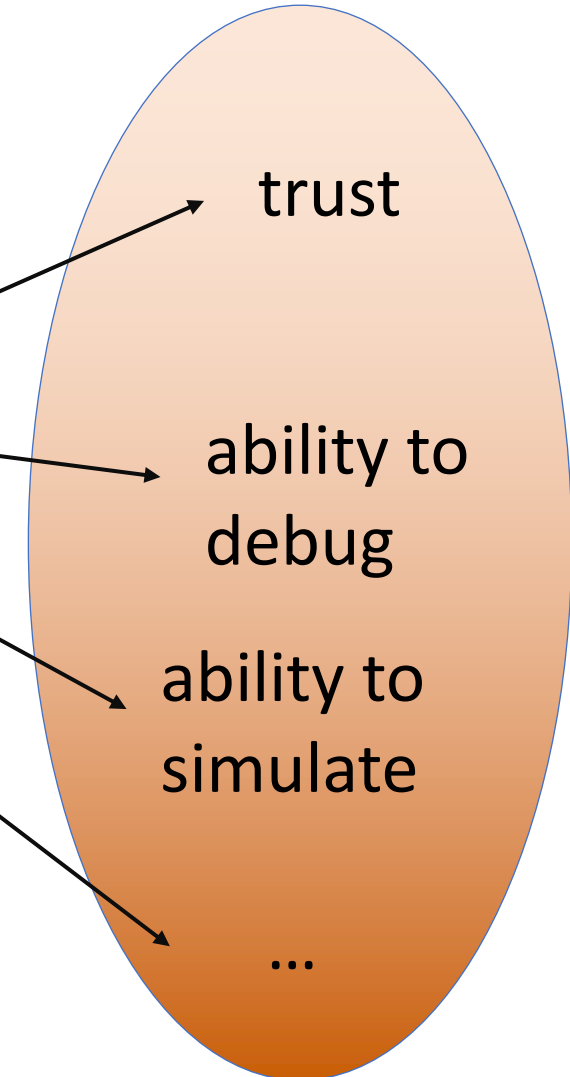
A Human-Centered Agenda for Interpretable ML

- Stop relying on intuition; empirically test which factors of a model enable users to better achieve their goals (Poursabzi-Sangdeh et al., 2021)
- Consider interpretability beyond the model, e.g., interpretability of data, objectives, or metrics (Gebru et al., 2018; Yin et al., 2019; Heger et al. 2022)
- Design and evaluate methods for achieving interpretability in context with relevant stakeholders (Kaur et al., 2020; Alvarez-Melis et al., 2021)

Properties of the system design

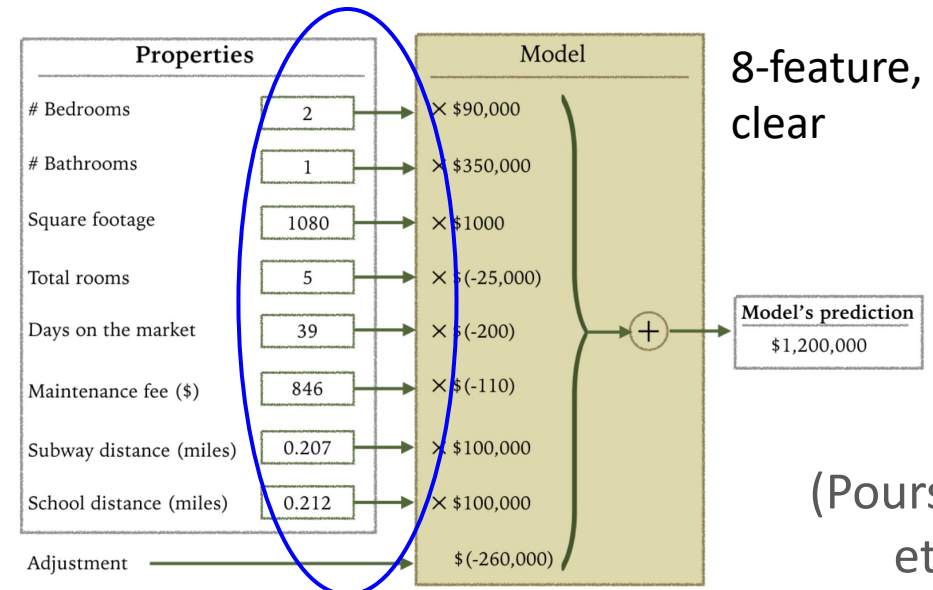
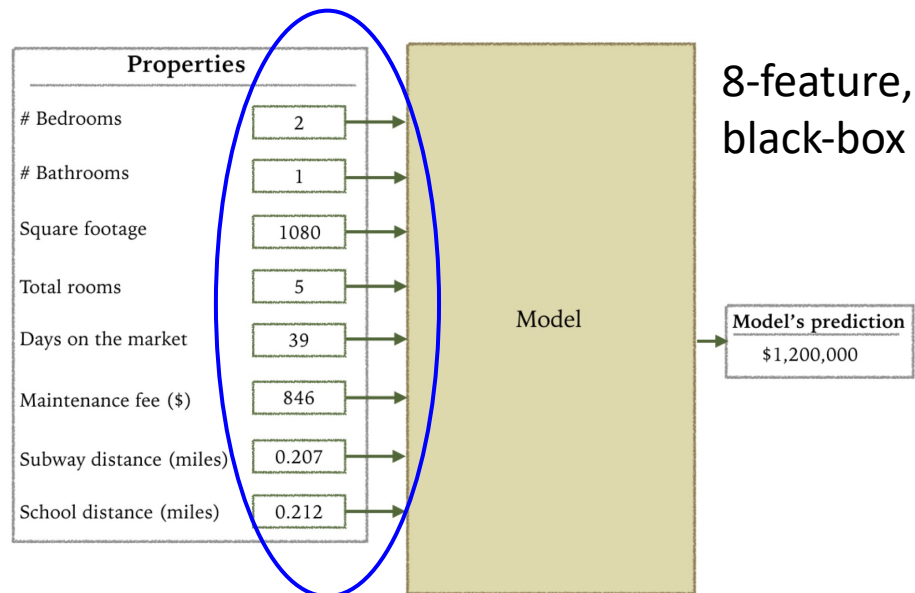
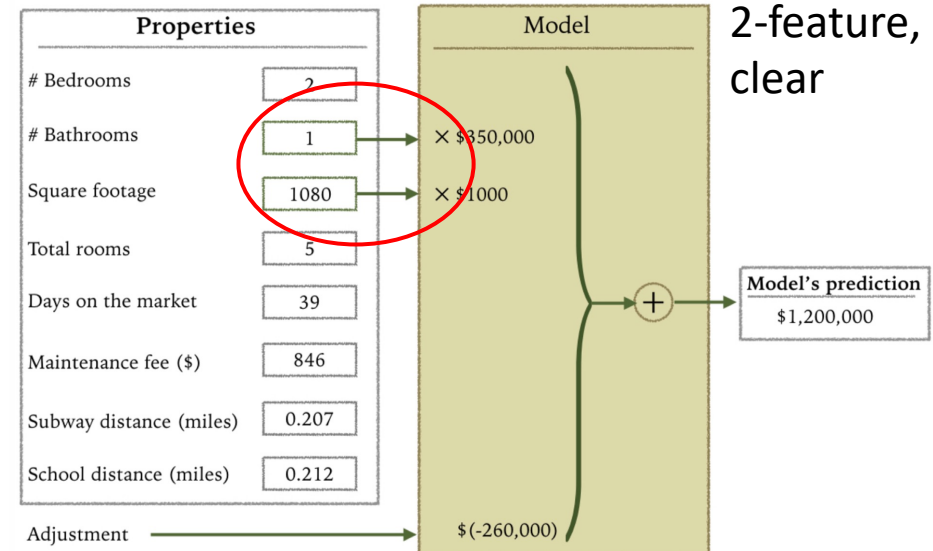
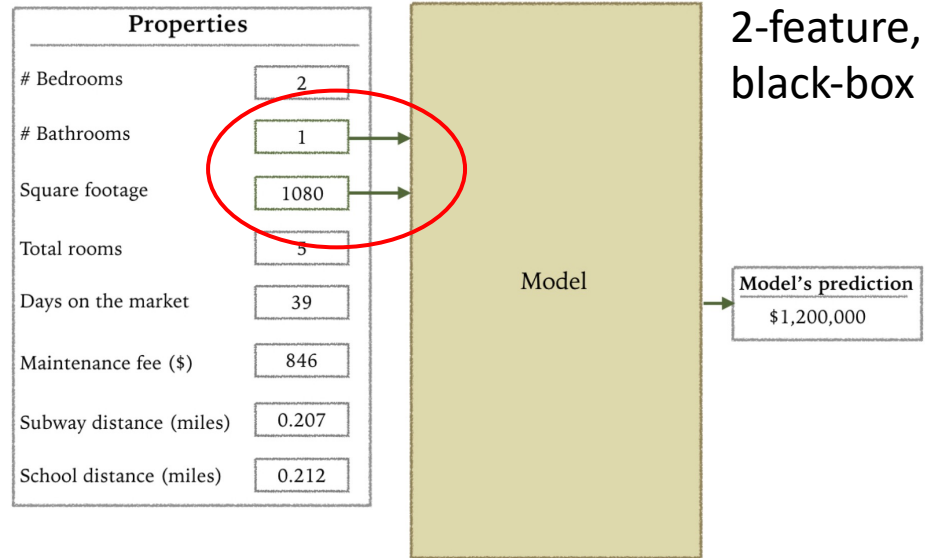


Properties of human behavior



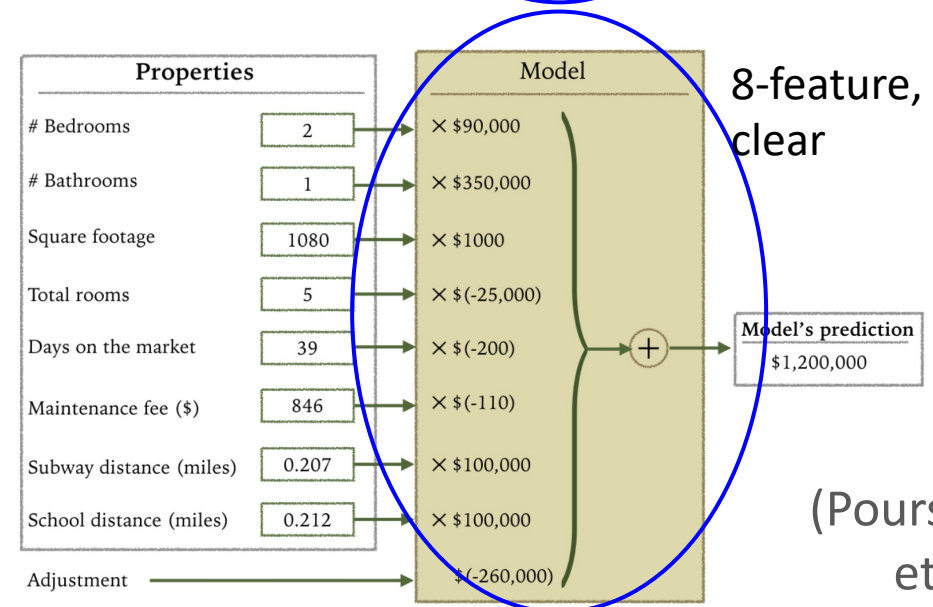
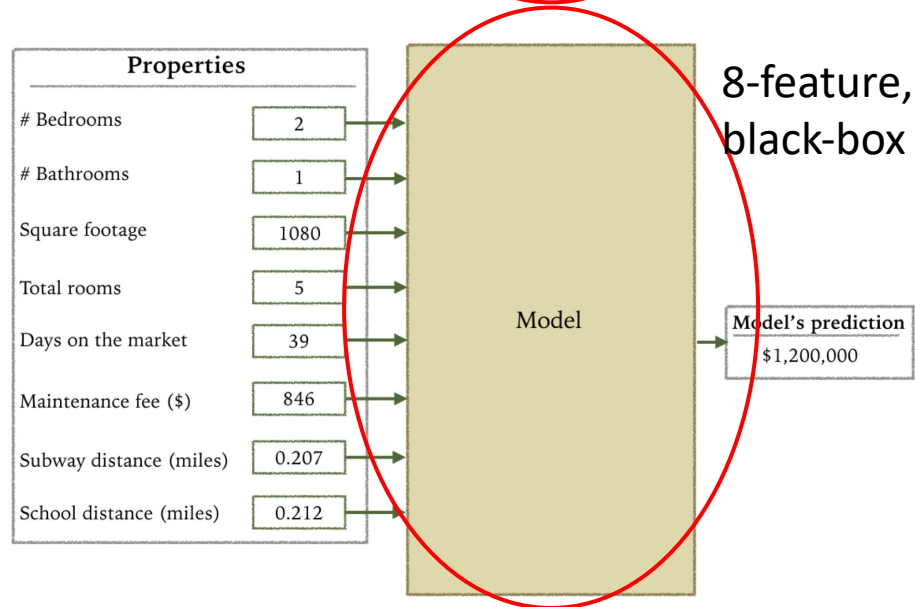
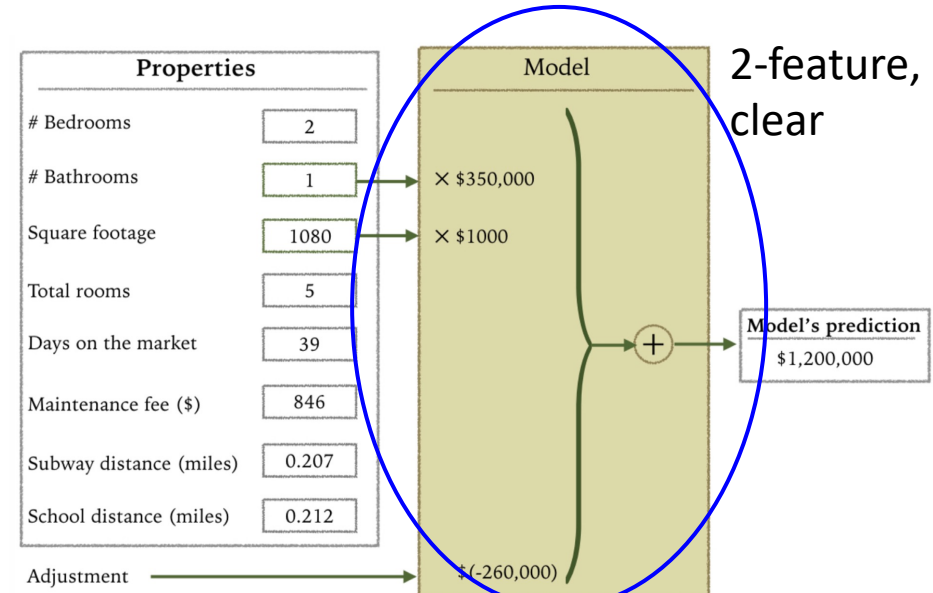
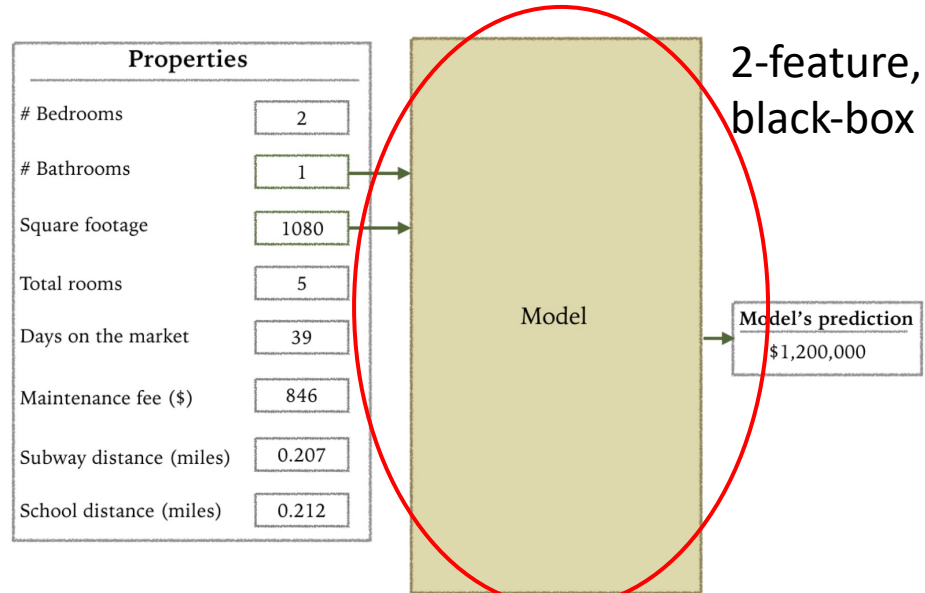
(Poursabzi-Sangdeh et al., 2021)

Experimental Conditions



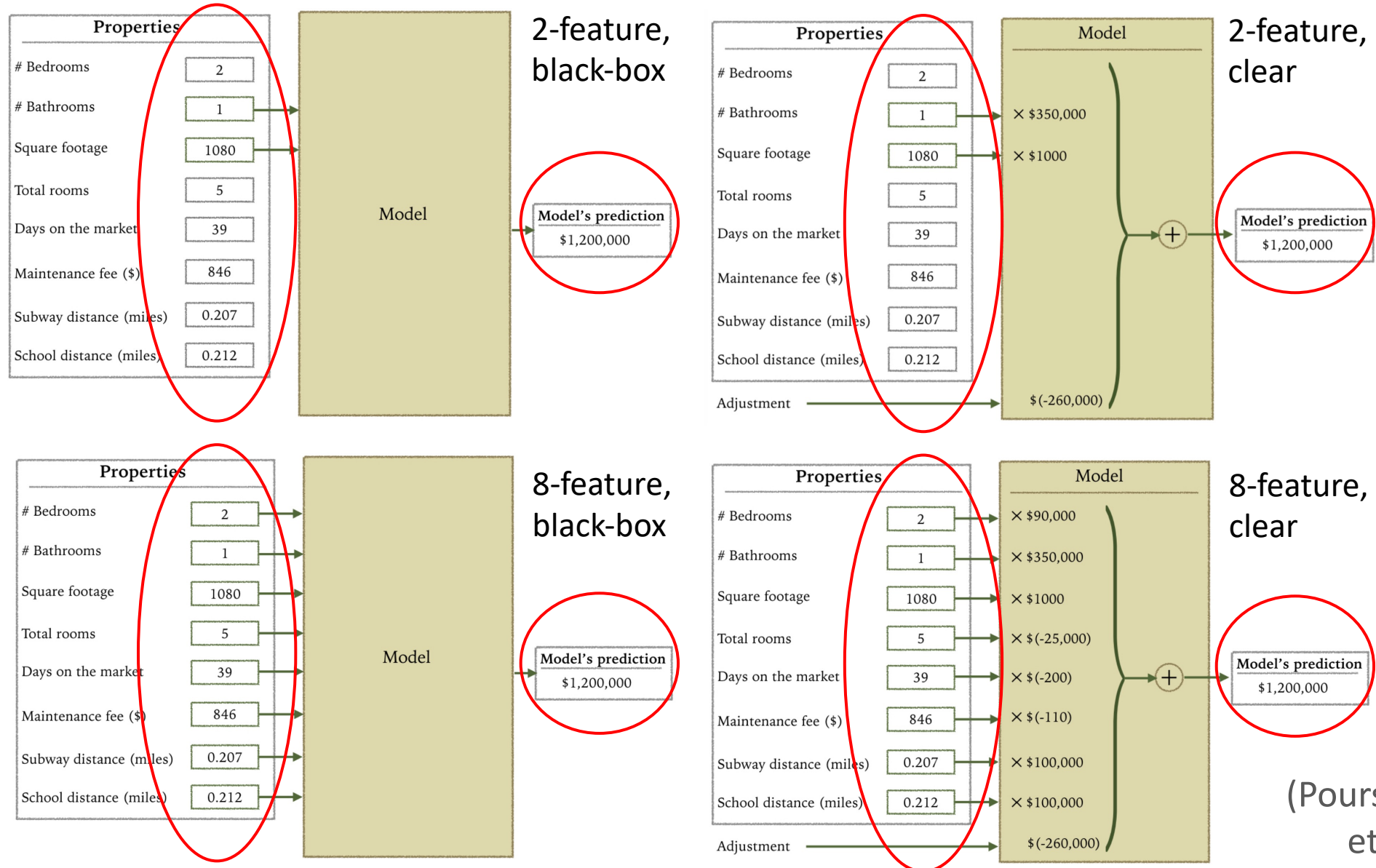
(Poursabzi-Sangdeh et al., 2021)

Experimental Conditions



(Poursabzi-Sangdeh et al., 2021)

Experimental Conditions



(Poursabzi-Sangdeh et al., 2021)

Results and Implications

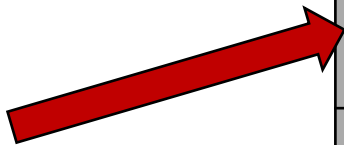
- Participants who were shown a clear model with a small number of features were best able to simulate the model's predictions
- However, we found no improvements in the degree to which participants followed the model's predictions when it was beneficial to do so
- Transparency reduced people's ability to detect when the model made a sizable mistake and correct it, seemingly due to information overload
- Generally, researchers should rely on rigorous experimentation over intuition when designing and evaluating interpretable models

A Human-Centered Agenda for Interpretable ML

- Stop relying on intuition; empirically test which factors of a model enable users to better achieve their goals (Poursabzi-Sangdeh et al., 2019)
- Consider interpretability beyond the model, e.g., interpretability of data, objectives, or metrics (Gebru et al., 2018; Yin et al., 2019; Heger et al. 2022)
- Design and evaluate methods for achieving interpretability in context with relevant stakeholders (Kaur et al., 2020; Alvarez-Melis et al., 2021)

Zoom in on data scientists

	Audit a single prediction	Discover new knowledge	Make better decisions	Debug models	Assess bias	Inspire trust
Decision makers						
Data scientists						
Lay people						
Regulators						



How do data scientists perceive and use interpretability tools?

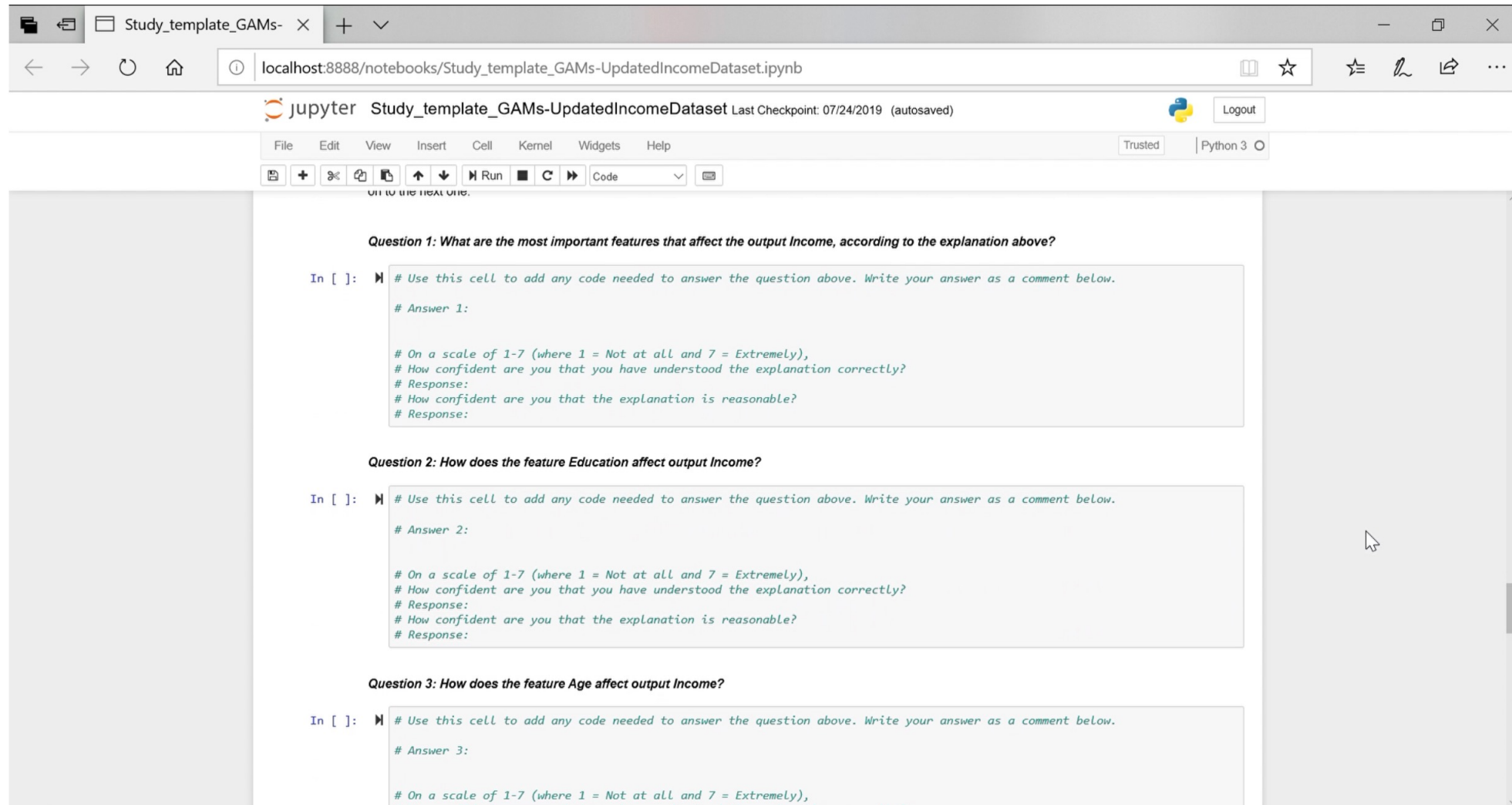
What are key challenges towards their use of these tools?

What opportunities do we, as researchers, have to make them better?

Interdisciplinary Approach

- Recruited a team of ML and HCI researchers plus data scientists with experience building and using interpretability tools
- Put the data scientists we studied in context
- Analyzed data through a mix of qualitative and quantitative methods
 1. Pilot interviews ($N = 6$) to identify challenges faced by data scientists in their day-to-day work
 2. Interview study ($N = 11$) to observe data scientists' ability to use interpretability tools when faced with these challenges
 3. Large-scale survey ($N = 197$) to scale up these results

Interview Study Setup



The screenshot displays a Jupyter Notebook interface in a web browser. The browser's address bar shows the URL `localhost:8888/notebooks/Study_template_GAMs-UpdatedIncomeDataset.ipynb`. The Jupyter interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help), a toolbar with icons for file operations and execution, and a status bar indicating the environment is 'Trusted' and using 'Python 3'. The notebook content consists of three sequential code cells, each preceded by a question:

Question 1: What are the most important features that affect the output Income, according to the explanation above?

```
In [ ]: # Use this cell to add any code needed to answer the question above. Write your answer as a comment below.  
  
# Answer 1:  
  
# On a scale of 1-7 (where 1 = Not at all and 7 = Extremely),  
# How confident are you that you have understood the explanation correctly?  
# Response:  
# How confident are you that the explanation is reasonable?  
# Response:
```

Question 2: How does the feature Education affect output Income?

```
In [ ]: # Use this cell to add any code needed to answer the question above. Write your answer as a comment below.  
  
# Answer 2:  
  
# On a scale of 1-7 (where 1 = Not at all and 7 = Extremely),  
# How confident are you that you have understood the explanation correctly?  
# Response:  
# How confident are you that the explanation is reasonable?  
# Response:
```

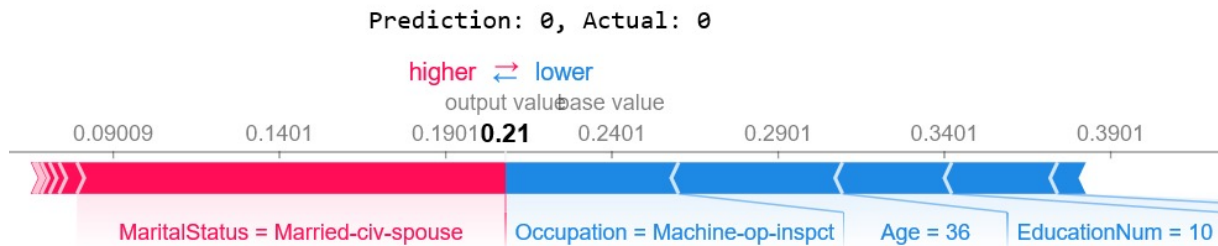
Question 3: How does the feature Age affect output Income?

```
In [ ]: # Use this cell to add any code needed to answer the question above. Write your answer as a comment below.  
  
# Answer 3:  
  
# On a scale of 1-7 (where 1 = Not at all and 7 = Extremely),
```

(Kaur et al., 2020)

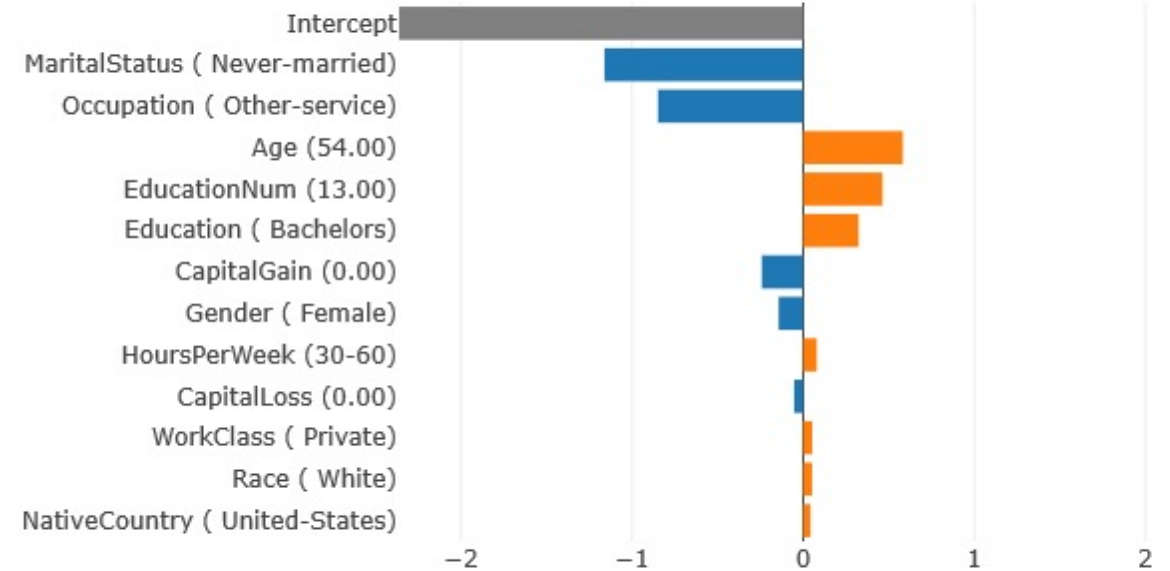
Explanation Types: Local Feature Importance

SHAP



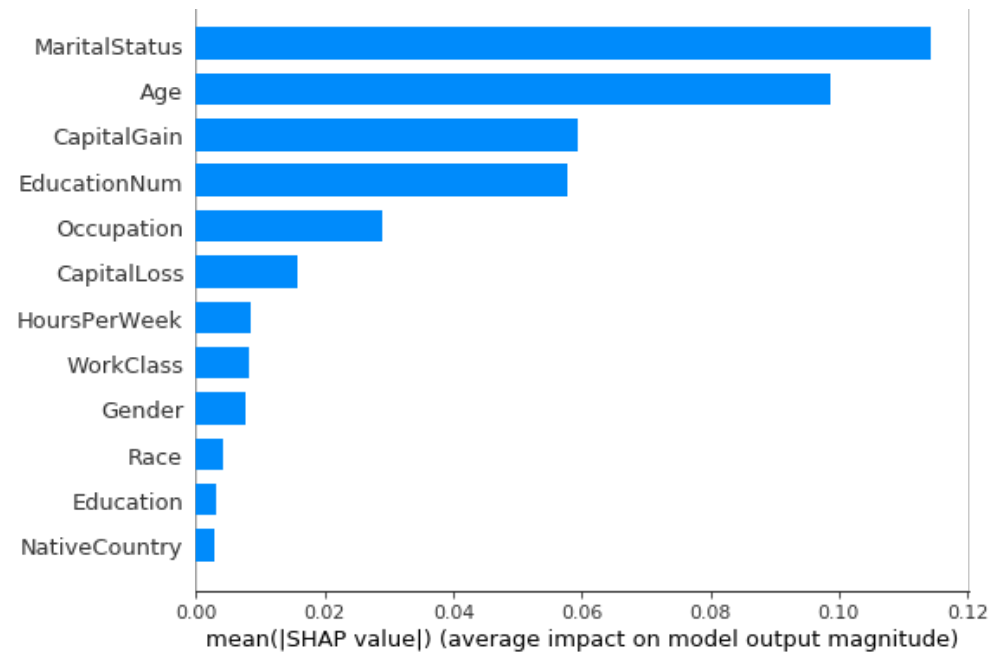
GAM

Predicted 0.00 | Actual 0.00

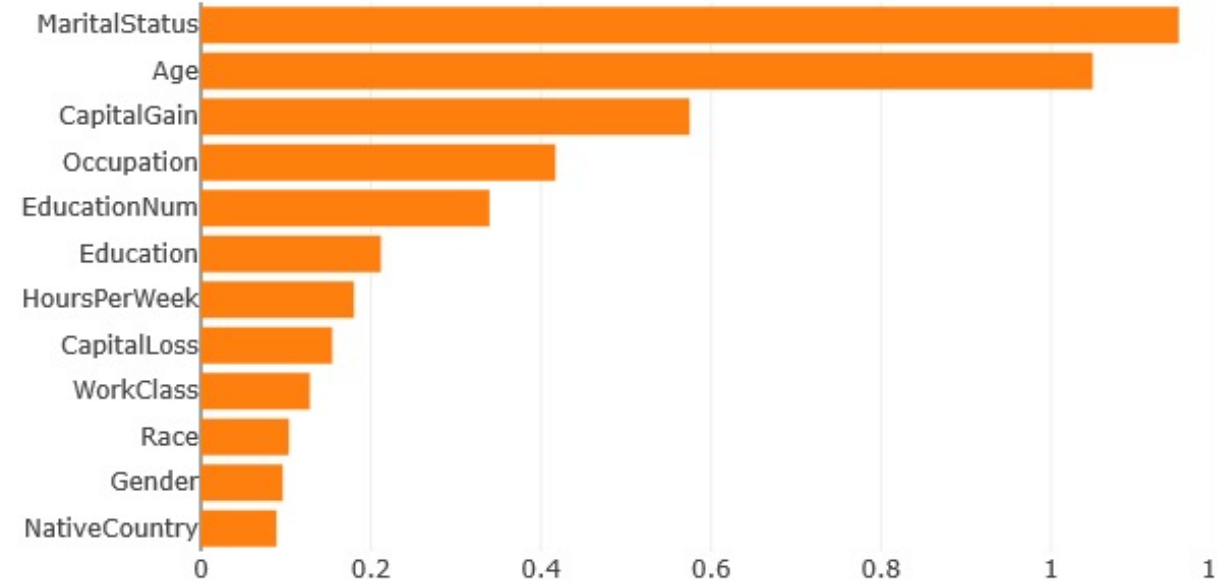


Explanation Types: Global Feature Importance

SHAP

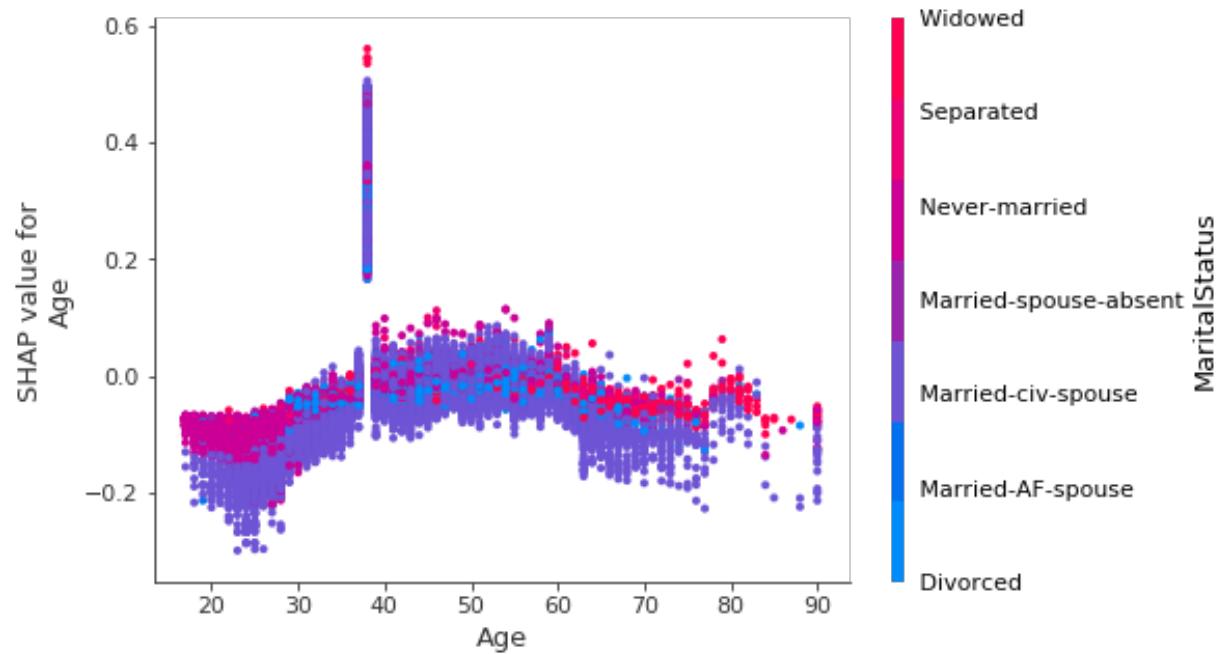


GAM

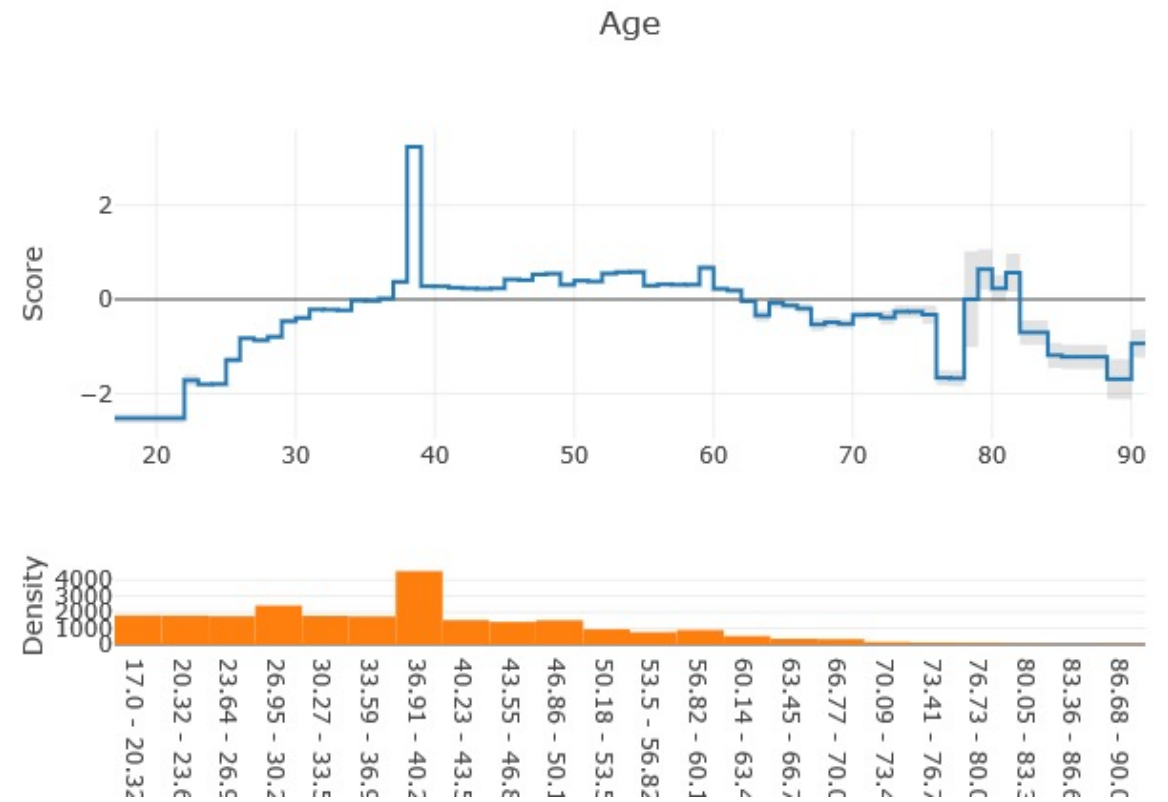


Explanation Types: Feature Impact on Prediction

SHAP

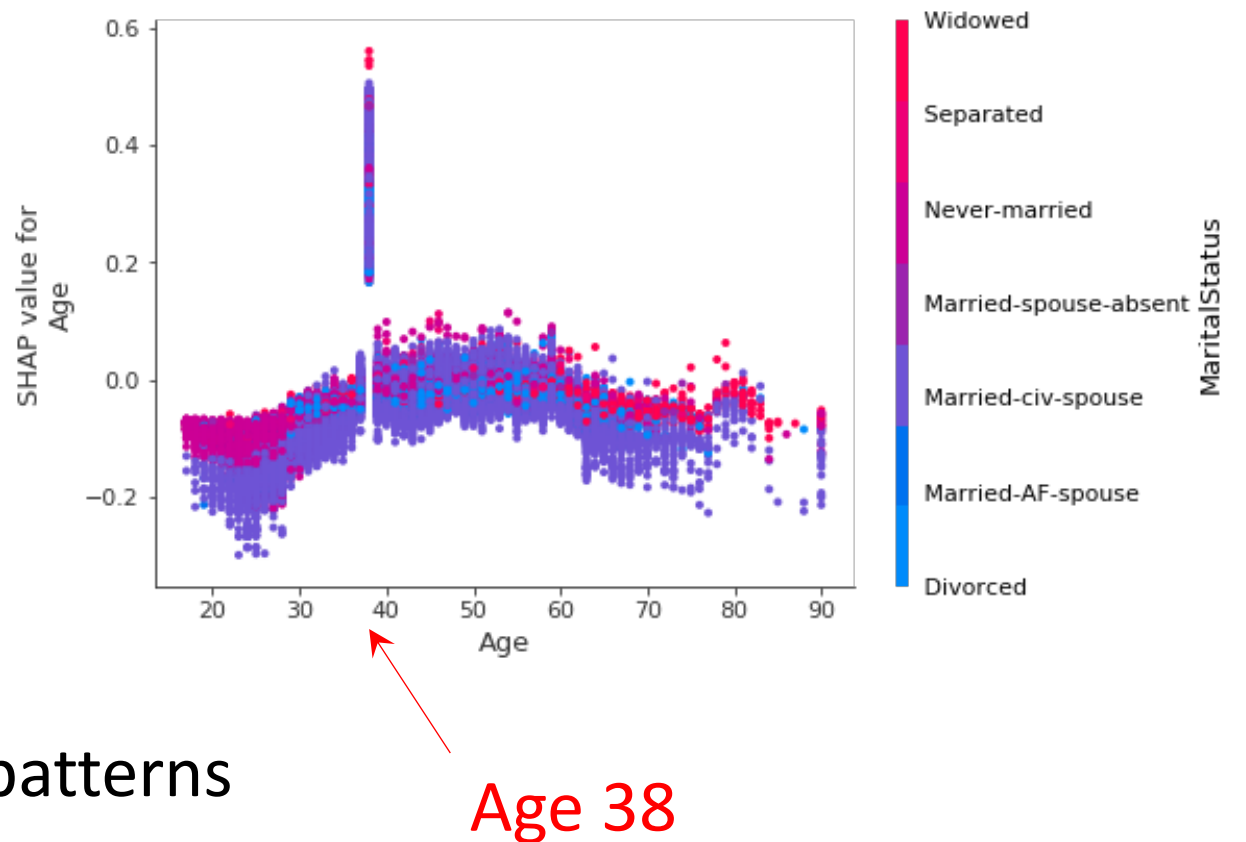


GAM



Challenges in the Data Science Pipeline

1. Missing values
2. Redundant features
3. Duplicate data masked as unique
4. Temporal changes in the data
5. Ad hoc categorization
6. Challenges recognizing debugging patterns



Results: Overuse of Tools/Overly Trusting Models

*“Age 38 seems to have the highest positive influence on income based on the plot. **Not sure why, but I guess if that’s what’s shown... makes sense.**” (P9, GAMs)*

*“Test of means says the same thing as SHAP about age. All’s good!”
(P8, SHAP)*

Results: Underuse of Tools

“[The tool] assigns a quantity that is important to know, but it’s showing that in a way that makes you misinterpret that value. Now I want to go back and check all my answers” ...
*“Okay, so, it’s not showing me a whole lot more than what I can infer on my own. Now I’m thinking... **is this an interpretability tool?**” (P4, SHAP)*

Results: Social Context is Important

“I guess this is a publicly available tool... must be doing something right. I think it makes sense.” (P8, SHAP)

*“I didn’t fully grasp what SHAP values were. This is a pretty popular tool and I get the log-odds concept in general. I figure they were showing SHAP values for a reason. Maybe it’s easier to judge relationships using log-odds instead of predicted value. **Anyway, so it made sense I suppose.**” (P6, SHAP)*

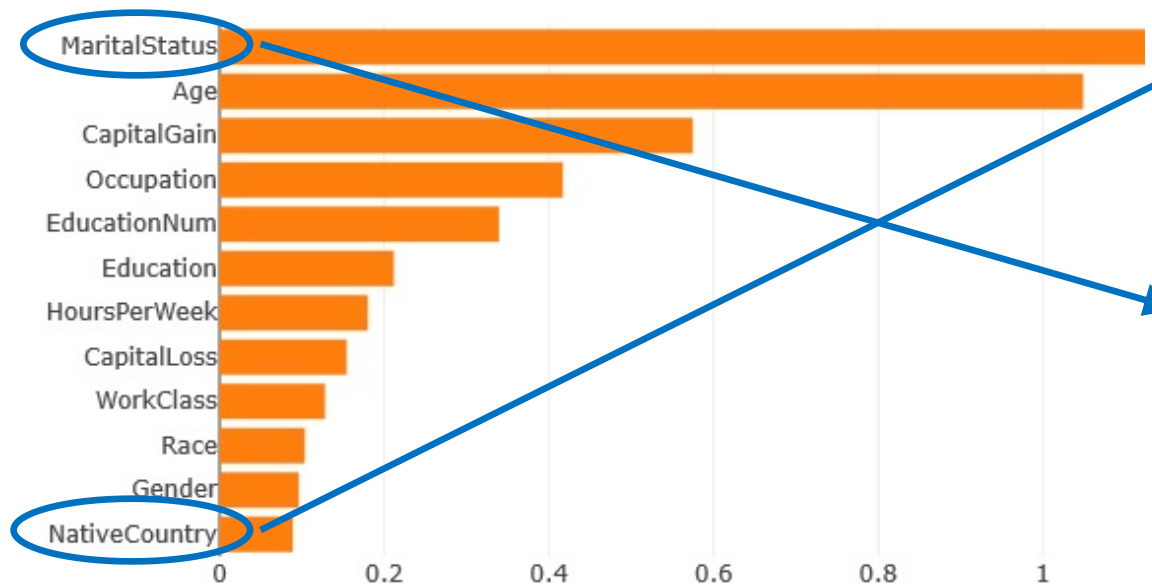
Survey Setup

- Demographic/experience questions
- Simulated exploration of the dataset, model, and interpretability tool
- Four blocks of questions about the dataset and model
- Follow-up questions about the interpretability tool and model

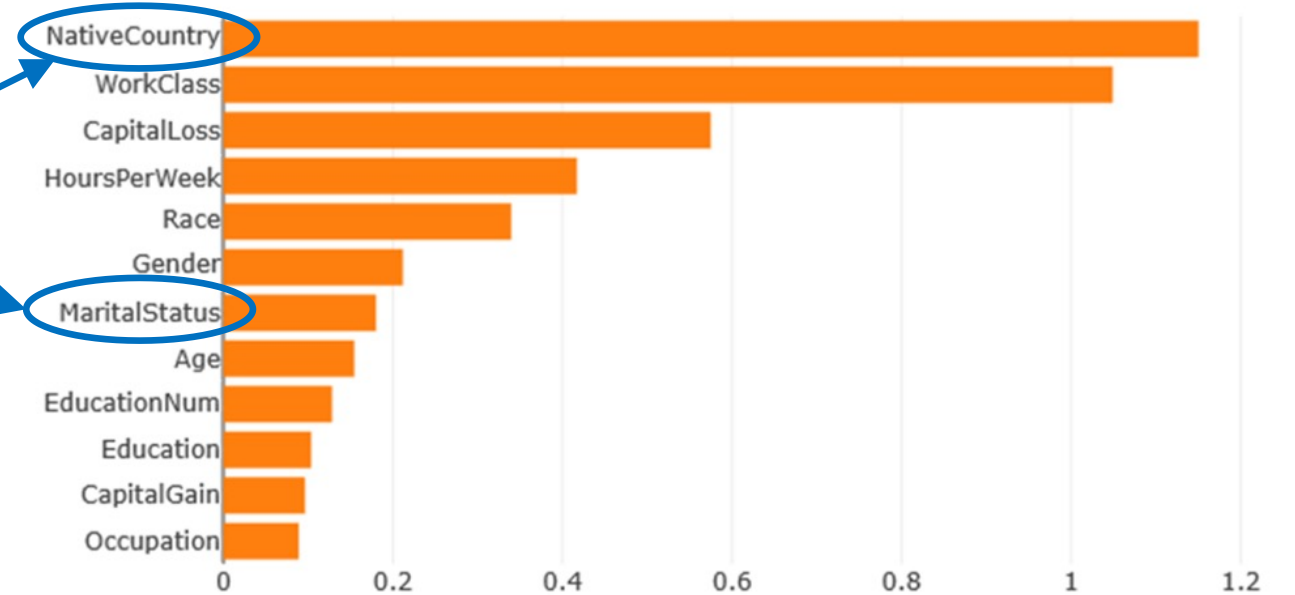
Survey Setup

- Controlled experiments, 2-by-2 design
 - GAMs or SHAP
 - Normal or manipulated global feature importance values

Overall Importance:
Mean Absolute Score



Overall Importance:
Mean Absolute Score



Sample of Quantitative Results

- Participants had higher accuracy on multiple choice questions about the visualizations using GAMs compared with SHAP
- Participants who used GAMs were more confident compared with those who used SHAP
- Manipulating the feature importance values reduced participants' confidence that the explanations were reasonable
- ... but didn't lead to increased suspicion about the model or interpretability tool

Takeaways

- People are central to machine learning systems and stakeholders need a basic understanding of how they work
- “Simple” doesn’t necessarily imply interpretable
- We need to...
 - Stop relying on our intuition about interpretability
 - Design and evaluate methods for achieving interpretability in context with relevant stakeholders
 - Consider interpretability beyond the model