Practitioner Interpretability Needs

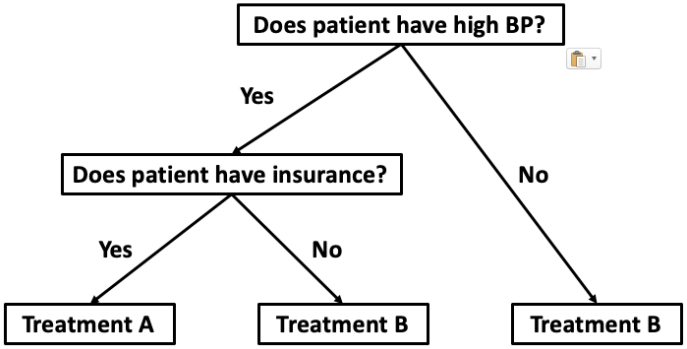
CS 282 BR Topics in Machine Learning:
Interpretability and Explainability

Ike Lage

01/30/2023

Overview

- Interpretable ML focus:
  - Developing new interpretable models and explanation methods



Overview

- Interpretable ML focus:
  - Developing new interpretable models and explanation methods
- Much less explored:
  - How useful are these tools actually to users?

Overview

- Interpretable ML focus:
  - Developing new interpretable models and explanation methods
- Much less explored:
  - How useful are these tools actually to users?
- These papers:
  - **How do ML practitioners use interpretability tools, and what are their unmet needs?**

Outline

- **Research paper:**
  - “Human Factors in Model Interpretability” by Hong et al.
- **Research paper:**
  - “Interpreting Interpretability: Understanding Data Scientists’ Use of Interpretability Tools for Machine Learning” by Kaur et al.
- **Discussion**

Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs

SUNGSOO RAY HONG, New York University, USA
JESSICA HULLMAN, Northwestern University, USA
ENRICO BERTINI, New York University, USA

Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs

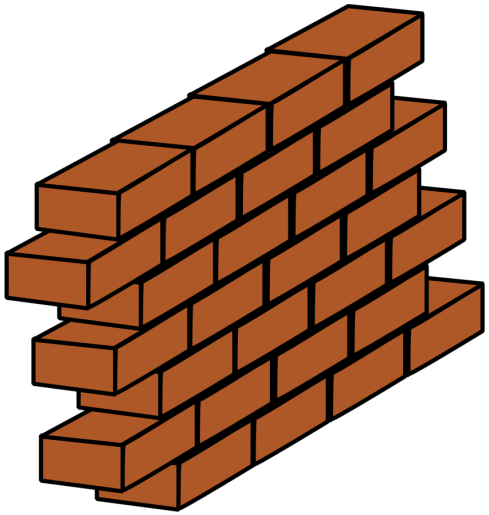
SUNGSOO RAY HONG, New York University, USA
JESSICA HULLMAN, Northwestern University, USA
ENRICO BERTINI, New York University, USA

- Contributions:
  - Conducts an interview study to understand industry practitioners' existing needs and uses for interpretability
  - Presents findings on roles, stages and goals related to interpretability
  - Identifies aspects of interpretability under-supported by existing technical solutions

Research Question

Fancy models
(not always
practical)

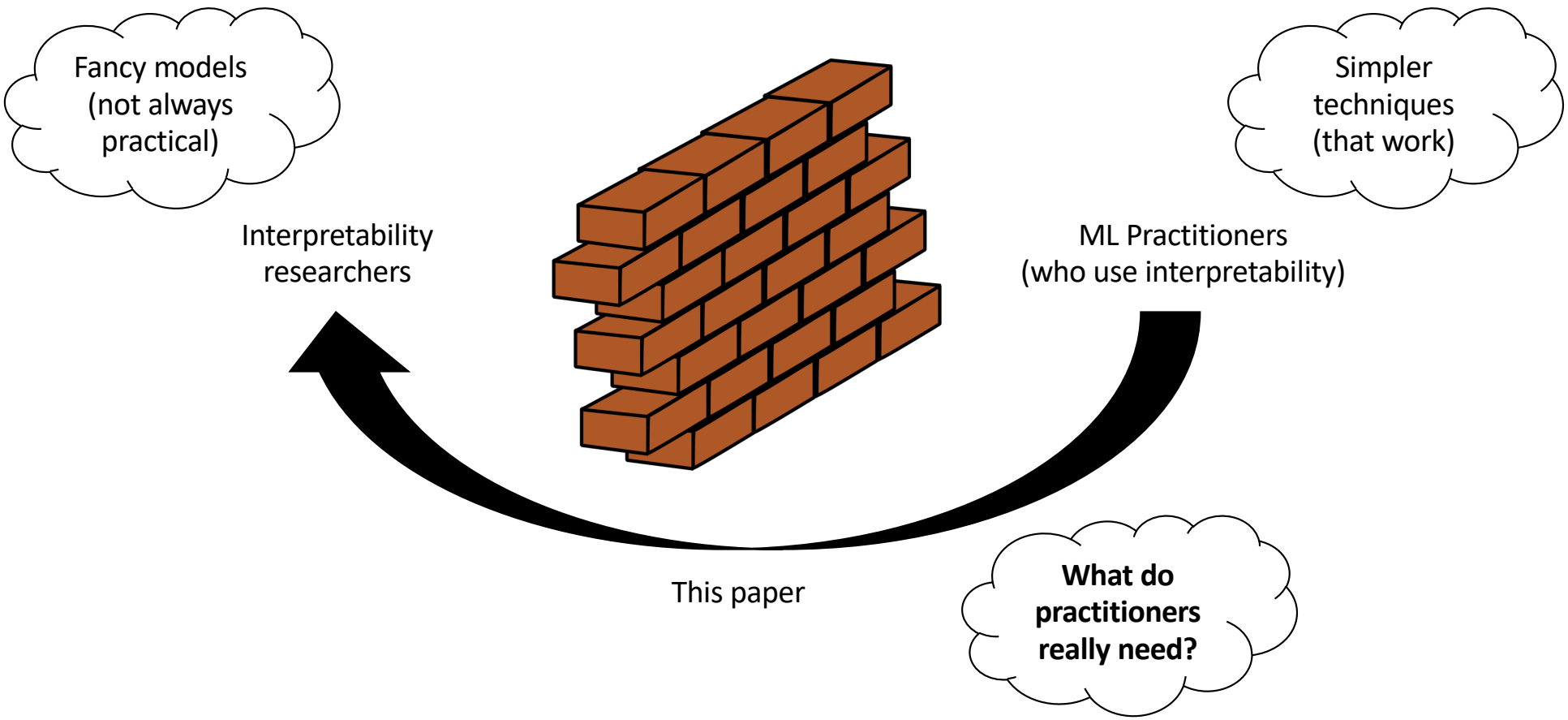
Interpretability
researchers



Simpler
techniques
(that work)

ML Practitioners
(who use interpretability)

Research Question



Methodology: Qualitative Studies

- Useful for exploratory research
- Can generate hypotheses to test quantitively

Methodology

- Study type:
  - Semi-structured interviews
- Recruitment:
  - 22 participants from convenience and snowball sampling
- Data analysis:
  - Qualitative coding

Methodology

- Study type:
  - Semi-structured interviews
- Recruitment:
  - 22 participants from convenience and snowball sampling
- Data analysis:
  - Qualitative coding

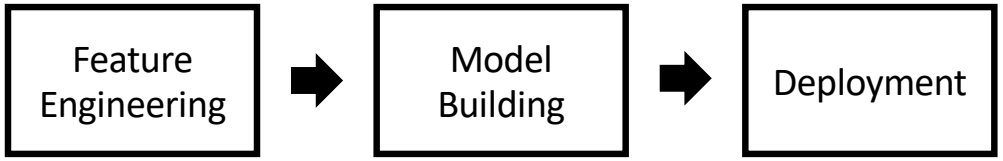
Iteratively build up a set of codes by looking at data and comparing notes with other annotators

Results

- Interpretability Roles

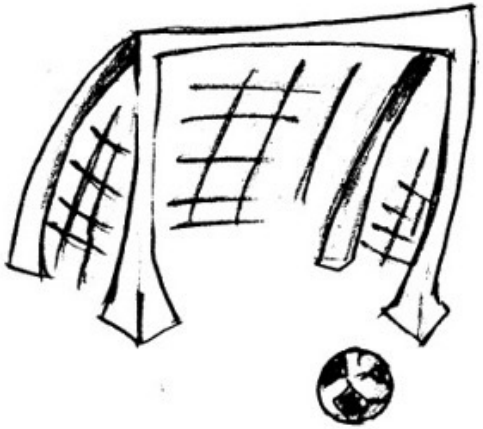
Results

- Interpretability Roles
- Interpretability Stages



Results

- Interpretability Roles
- Interpretability Stages
- Interpretability Goals



Results: Interpretability Roles

- Model builders
- Model breakers
- Model consumers

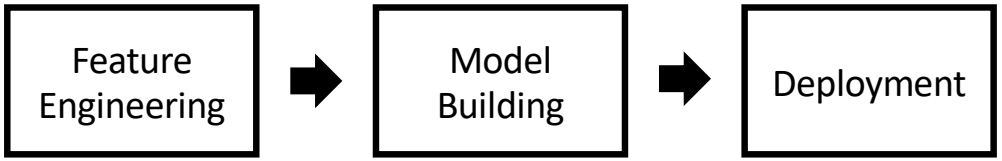
Results: Interpretability Roles

- Model builders
- Model breakers
- Model consumers

What methods are designed for different roles?

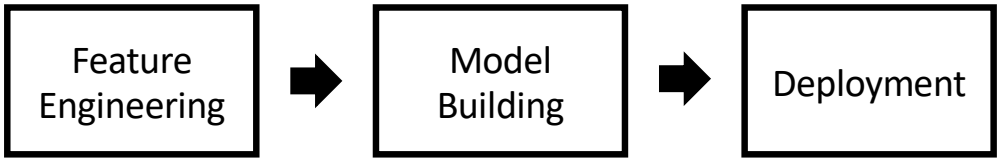
Results: Interpretability Stages

- Ideation and conceptualization stage
- Building and validation stage
- Deployment, maintenance and use stage



Results: Interpretability Stages

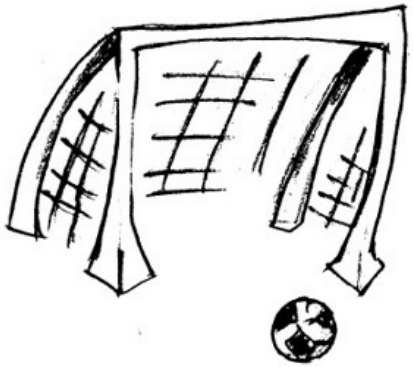
- Ideation and conceptualization stage
- Building and validation stage
- Deployment, maintenance and use stage



What methods are designed for different stages?

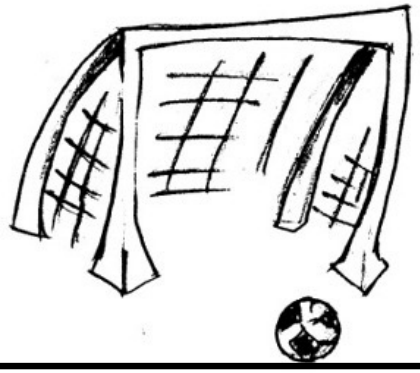
Results: Interpretability Goals

- Interpretability for model validation and improvement
- Interpretability for decision making and knowledge discovery
- Interpretability to gain confidence and obtain trust



Results: Interpretability Goals

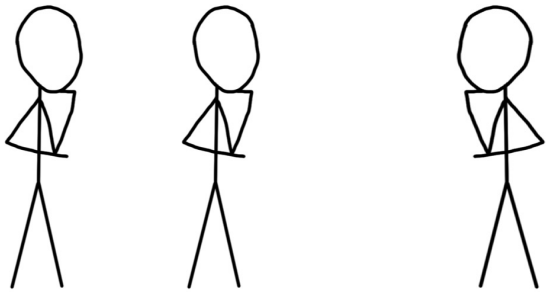
- Interpretability for model validation and improvement
- Interpretability for decision making and knowledge discovery
- Interpretability to gain confidence and obtain trust



What methods are designed for different stages?

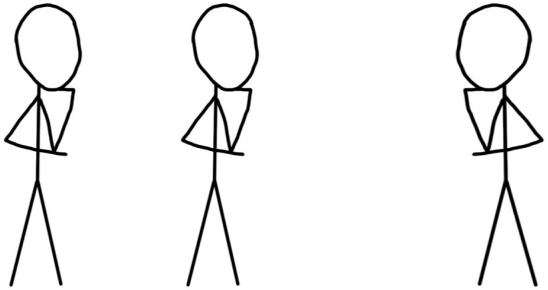
Themes: Interpretability is Cooperative

- Important for communicating with domain experts and stakeholders
- Facilitate trust, sometimes just by virtue of including an explanation



Themes: Interpretability is Cooperative

- Important for communicating with domain experts and stakeholders
- Facilitate trust, sometimes just by virtue of including an explanation



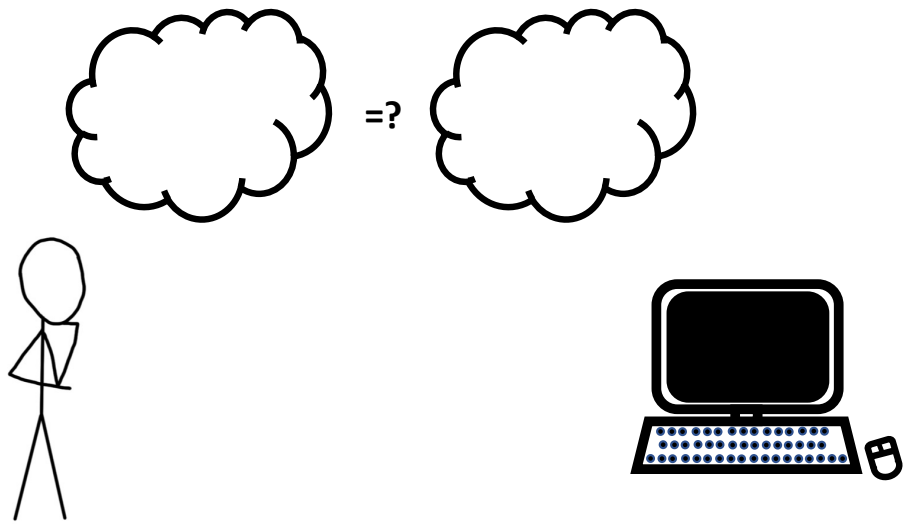
Better tools vs. better data science training for communication?

Themes: Interpretability is a Process

- Important across many different stages
- Dialogue with the model for continued use

Themes: Interpretability is Mental Model Comparison

- Understanding what end-users need is important
- Translating human hypotheses into ML models



Themes: Interpretability is Context-Dependent

- Good explanations depend on the user
- How detailed should it be? What skepticism will they bring to it?

Design Opportunities

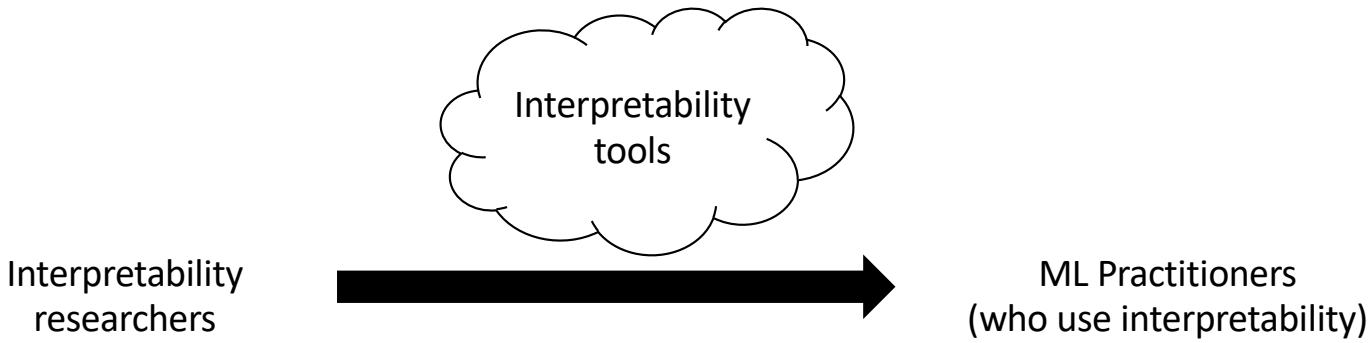
- Integrating human expectations
- Communicating and summarizing behavior
- Scalable and integratable tools
- Post-deployment support

Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning

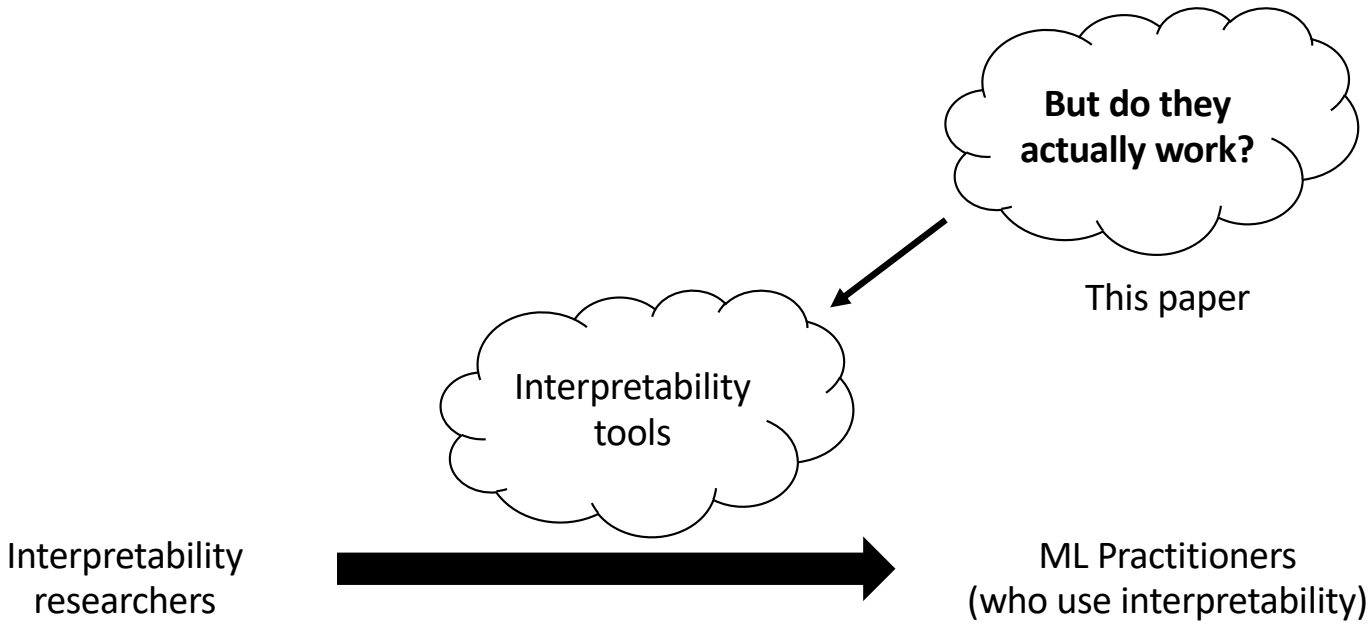
Harmanpreet Kaur[1], Harsha Nori[2], Samuel Jenkins[2], Rich Caruana[2], Hanna Wallach[2], Jennifer Wortman Vaughan[2]
[1]University of Michigan, [2]Microsoft Research
harmank@umich.edu, {hanori,sajenkin,rcaruana,wallach,jenn}@microsoft.com

- Contributions:
  - Evaluates whether interpretability tools help ML practitioners understand models
  - Contextual inquiry and survey of how practitioners use ML tools
  - Find that data scientists over trust and misuse interpretability tools

Research Question



Research Question



Methodology: Overview

Pilot interviews

- N = 6
- Identified issues to test in contextual inquiry

Contextual Inquiry

- N = 11
- Can users find issues identified in pilot when given standard tools?

Survey

- N = 197
- Validate and quantify findings in a large sample

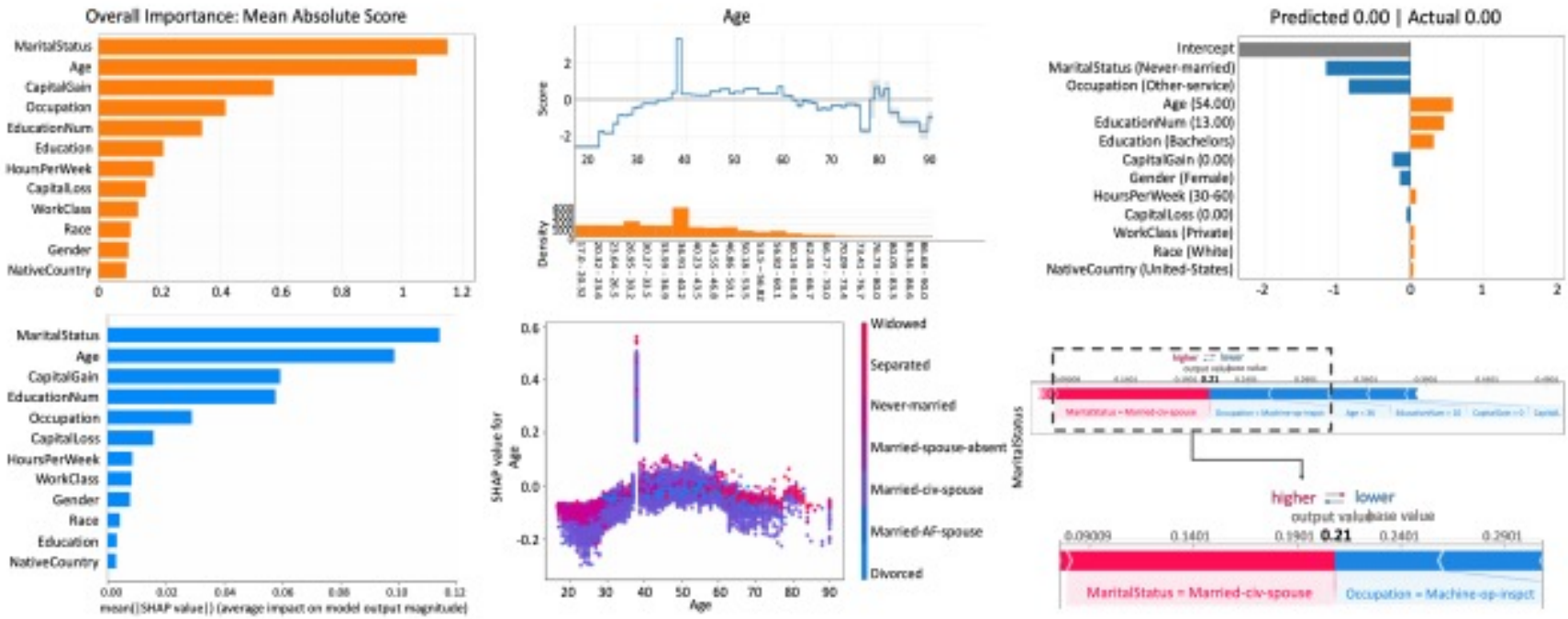
Pilot Study: Common Issues for Data Scientists

Theme	Description	Incorporation into Contextual Inquiry	Num.
Missing values	Many methods for dealing with missing values (e.g., coding as a unique value or imputing with the mean) can cause biases or leakage in ML models.	Replaced the value for the "Age" feature with 38 (the mean) for 10% of the data points with an income of >$50k, causing predictions to spike at 38. Asked about the relationship between "Age" and "Income."	4 of 11
Changes in data	Data can change over time (e.g., new categories for an existing feature).	Asked whether the model (trained on 1994 data) would work well on current data after adjusting for inflation.	10 of 11
Duplicate data	Unclear or undefined naming conventions can lead to accidental duplication of data.	Modified the "WorkClass" feature to have duplicate values: "Federal Employee," "Federal Worker," "Federal Govt." Asked about the relationship between "WorkClass" and "Income."	1 of 11
Redundant features	Including the same feature in several ways can distribute importance across all of them, making each appear to be less important.	Included two features, "Education" and "EducationNum," that represent the same information. Asked about the relationships between each of these and "Income."	3 of 11
Ad-hoc categorization	Category bins can be chosen arbitrarily when converting a continuous feature to a categorical feature.	Converted "HoursPerWeek" into a categorical feature, binning arbitrarily at 0–30, 30–60, 60–90, and 90+ hours. Asked about the relationship between "HoursPerWeek" and "Income."	3 of 11
Debugging difficulties	Identifying potential model improvements based on only a small number of data points is difficult.	Asked people to identify ways to improve accuracy based on local explanations for 20 misclassified data points.	8 of 11

Contextual Inquiry: Tools

GAM

SHAP



Contextual Inquiry: Results

- Misuse and disuse
- Social context is important
- Visualizations can be misleading

Methodology: Large Scale Survey

- Study type:
  - Survey based on example queries from previous tools
- Recruitment:
  - 197 participants from the mailing list of a large tech company
- Data analysis:
  - Coded open ended responses
  - Ran statistical tests to compare outcomes by condition

Large Scale Survey: Conditions

- Explanation type
  - GAM
  - SHAP
- Visualization type
  - normal
  - manipulated ← Do people trust obviously wrong explanations less?

Results: Performance with explanations

- GAM >> SHAP
- Better results with good explanations than manipulated

Results: Factors that affect willingness to deploy

- Deployment decisions made on intuition
- Explanations used to superficially justify deployment
- Some participants suspicious of model and used tool as intended

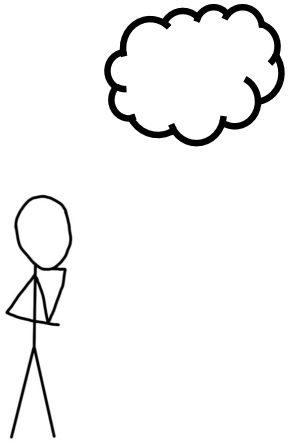
Results: Factors that affect willingness to deploy

- Deployment decisions made on intuition
- Explanations used to superficially justify deployment
- Some participants suspicious of model and used tool as intended

How to push people towards deliberative reasoning?

Results: Mental models of interpretability tools

- Participants largely did not understand tools well
- Despite that, they believed tools effective for many uses



Is it bad for explanations to persuade people without understanding?

Results: Tension between cognitive and social factors

- Participants with more ML background understood explanations better
- More ML experience -> less confidence in explanations -> lower deployment

How do we make ML explanations more accessible?