Human-Centered Computing
Research Group

Freie Universität Berlin

«Human-Centered Data Science»

Post-hoc Interpretability: Understanding User Needs

Prof. Dr. Claudia Müller-Birn

Human-Centered Computing, Institute of Computer Science

Freie Universität Berlin

June 23, 2022

Lecture Overview

Recap

Intrinsic vs. Post-hoc Interpretability

A Human-Centered Perspective (Mental Model on Explanations, Experts Mental Models = Explanation Methods, LIME)

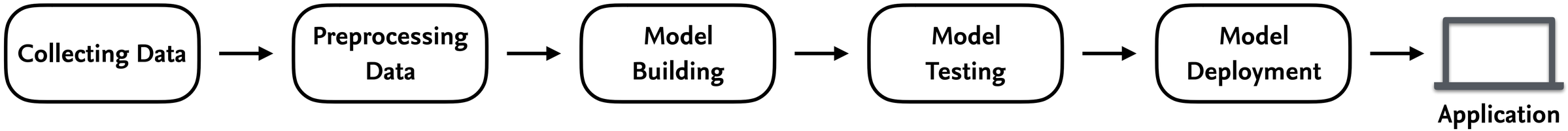
Break

A Human-Centered Perspective (Perspectives on Explanations, Methods for Understanding Explanation Needs, Challenge of Interpretability)

Course Evaluation

Recap

Scope of Human-Centered Data Science



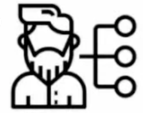
Human and the machine are equally important actors in carrying out data science.

Motivating the Need for Transparency

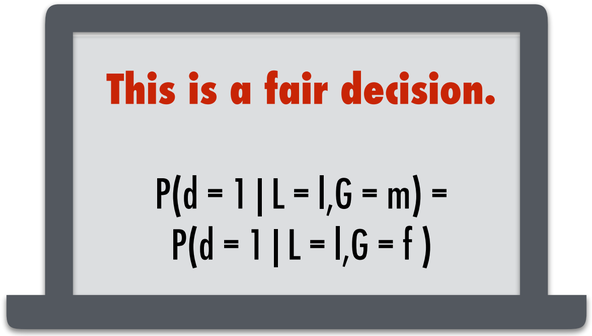
User *Why is this fair?*

Engineer *What do I have to do to make it a fair solution?*

Why is this fair but not the other approach?



Developer



Researcher *What are the assumptions?*

How can I correct an error?

Engineer

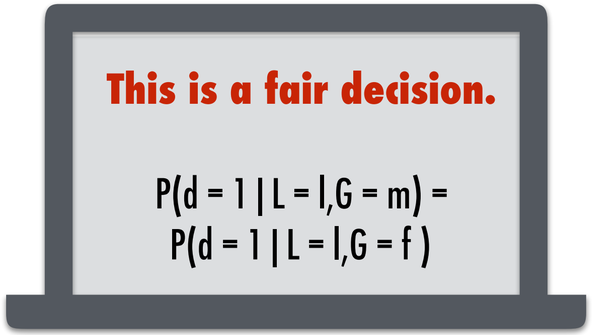
When can I trust this decision?

User

Questions adapted from D. Gunning, Explainable artificial intelligence (xAI), Tech. rep., Defense Advanced Research Projects Agency (DARPA) (2017).
https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

Motivating the Need for Transparency

Use models that are intrinsically interpretable and known to be easy for humans to understand.



Questions adapted from D. Gunning, Explainable artificial intelligence (xAI), Tech. rep., Defense Advanced Research Projects Agency (DARPA) (2017).
https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

Overview on Transparency: Openness of Data/Code

Transparency

Openness of Data/Code

(Intrinsic) Interpretability

Algorithmic Transparency

Level of the whole pipeline

- Model (code) and training/test data are publicly inspectable.
- Individual decisions are reproducible, and changes are logged and version controlled.

Lipton, Z. C. (2018). The mythos of model interpretability. Commun. ACM, 61(10), 36–43. http://doi.org/10.1145/3233231

Overview on Transparency: (Intrinsic) Interpretability

Transparency

Openness of Data/Code

(Intrinsic) Interpretability

Algorithmic Transparency

Decomposability

Level of individual components

- Each part of the model admits an intuitive explanation.
- Example: Each node in a decision tree might correspond to a plain text description.

Simulatability

Level of the entire model

- A model might be called *transparent* if a person can contemplate the entire model at once.
- Example: Lasso regression is are more interpretable than dense linear models.

Lipton, Z. C. (2018). The mythos of model interpretability. Commun. ACM, 61(10), 36–43. http://doi.org/10.1145/3233231

Overview on Transparency: Algorithmic Transparency

Transparency

Openness of Data/Code

(Intrinsic) Interpretability

Algorithmic Transparency

Level of the learning algorithm

» Transparency applies at the level of the learning algorithm itself.

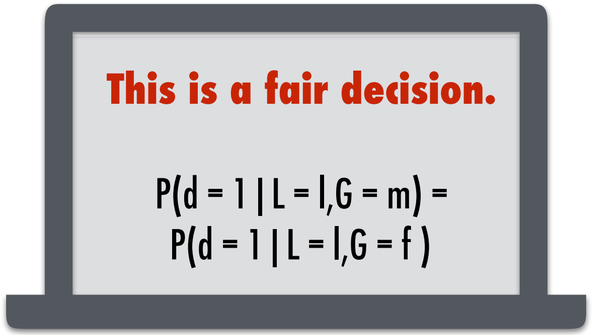
» Example: Understand the error metrics in linear regression.

Lipton, Z. C. (2018). The mythos of model interpretability. Commun. ACM, 61(10), 36–43. http://doi.org/10.1145/3233231

Intrinsic vs. Post-hoc Interpretability Techniques

Use models that are intrinsically interpretable and known to be easy for humans to understand.

Intrinsic Interpretability Techniques



Train a black box model and apply post-hoc interpretability techniques to provide explanations.

Post-hoc Interpretability Techniques

Questions adapted from D. Gunning, Explainable artificial intelligence (xAI), Tech. rep., Defense Advanced Research Projects Agency (DARPA) (2017).
https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

Intrinsic vs. Post-hoc Interpretability Techniques (*cont.*)

Intrinsic Interpretability Techniques

Your intended audience can

» Understand how the model works (input data, features, basic mechanics), and how specific predictions/classifications/etc. were made by showing the inner working.

» Most methods are *model-specific* and have primarily a *global scope*.

» It refers to models that can be interpreted by themselves. They are *transparent*.

Post-hoc Interpretability Techniques

Your intended audience can

» Understand how the model works by applying methods that analyze the model after training. These methods are decoupled from the main model.

» Most methods are *model-agnostic* and have primarily a *local scope*.

» It refers to *explanation methods* that are used to create explanations.

Carvalho, Diogo V., Eduardo M. Pereira, and Jaime S. Cardoso. "Machine learning interpretability: A survey on methods and metrics." Electronics 8.8 (2019): 832.

A Human-Centered Perspective on Post-hoc Interpretability Techniques

A Participatory Process for Interpretability Techniques

WHAT to explain?

Expert Mental Model

What happens to the best of our knowledge?

What can be explained?

What does an expert mental model of the system look like?

User Mental Model

How do users currently make sense of the system?

What is the user mental model of the system based on its current UI?

How does it differ from the expert mental model?

Target Mental Model

Which key components of the algorithm do users want to be made transparent in the UI?

To what extent are users actually interested in the rationale behind the algorithms?

HOW to explain?

User Mental Model ⟷ Target Mental Model

How can the target mental model be reached through UI design?

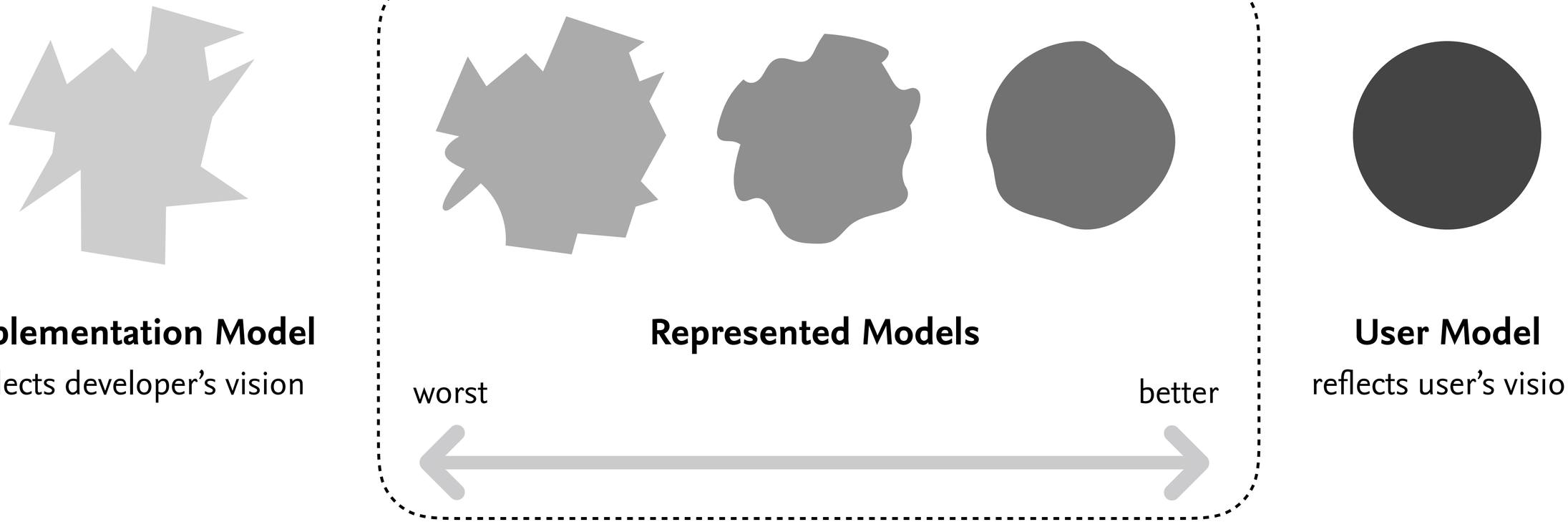
How and where can transparency be integrated into the UI of the system?

How has the users mental model developed?

Has the target mental model been reached?

Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In Conf. on Intelligent User Interfaces, 211–223.

Types of Mental Models



How can we understand this spectrum?

Image from Cooper, A., Reimann, R., & Cronin, D. (2007). About face 3: the essentials of interaction design. John Wiley & Sons.

A Participatory Process for Transparency Design

WHAT to explain?

Expert Mental Model

What happens to the best of our knowledge?

What can be explained?

What does an expert mental model of the system look like?

User Mental Model

How do users currently make sense of the system?

What is the user mental model of the system based on its current UI?

How does it differ from the expert mental model?

Target Mental Model

Which key components of the algorithm do users want to be made transparent in the UI?

To what extent are users actually interested in the rationale behind the algorithms?

HOW to explain?

User Mental Model ⟷ Target Mental Model

How can the target mental model be reached through UI design?

How and where can transparency be integrated into the UI of the system?

How has the users mental model developed?

Has the target mental model been reached?

Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In Conf. on Intelligent User Interfaces, 211–223.

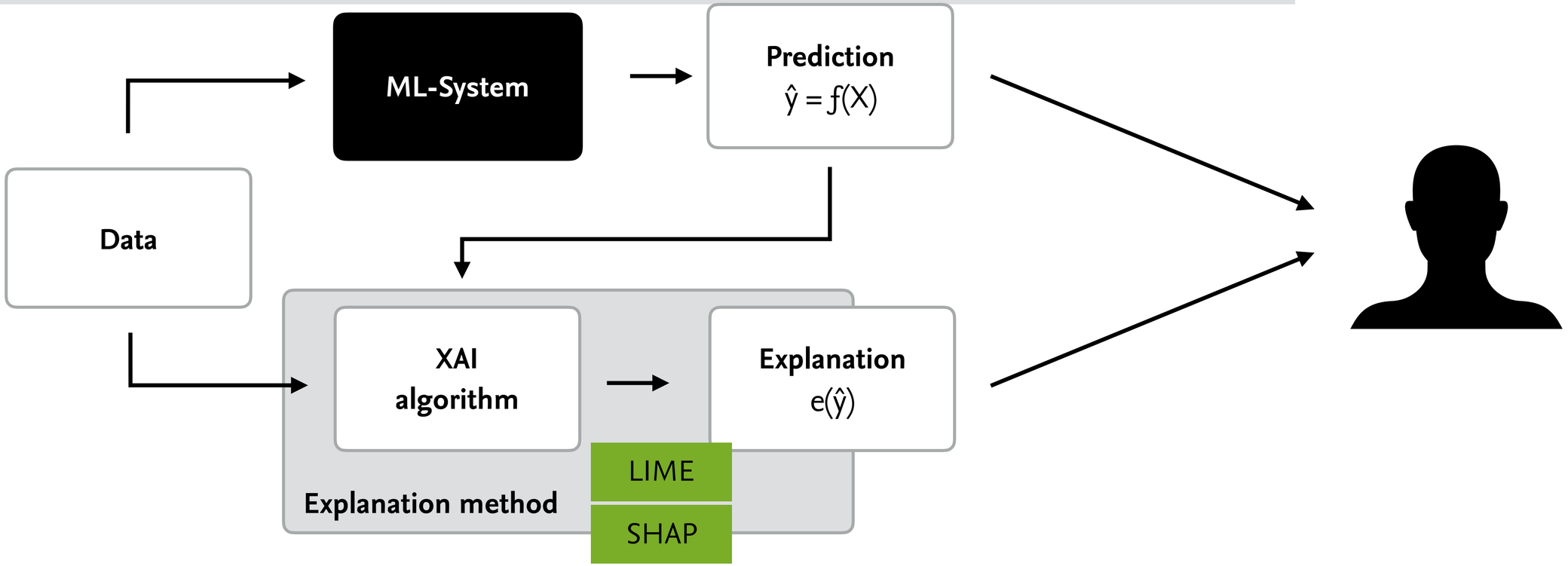
Overview on Categories of Explanation Methods

Category of Methods	Explanation Method	Definition
Explain a prediction (**Local**)	Local rules or trees	»Describe the rules or a decision-tree path that the instance fits to guarantee the prediction
Explain the model (**Global**)	Global feature importance	»Describe the weights of features used by the model (including visualization that shows the weights of features)
	Decision tree approximation	»Approximate the model to an interpretable decision-tree
	Rule extraction	»Approximate the model to a set of rules, e.g., if-then rules
Inspect counterfactual	Feature influence or relevance method	»Show how the prediction changes corresponding to changes of a feature (often in a visualization format)
	Contrastive or counterfactual features	»Describe the feature(s) that will change the prediction if perturbed, absent or present
Example based	Prototypical or representative examples	»Provide example(s) similar to the instance and with the same record as the prediction
	Counterfactual example	»Provide example(s) with small differences from the instance but with a different record from the prediction

Liao, Q. V., Gruen, D., & Miller, S. (2020, January 8). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. http://doi.org/10.1145/3313831.3376590

Explanation Method

An explanation method is a pattern or a mechanisms that generates explanations to establish post-hoc interpretability. An explanation method is based on an explainable AI (XAI) algorithm.



Liao, Q. V., Gruen, D., & Miller, S. (2020, January 8). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. http://doi.org/10.1145/3313831.3376590
Image from Carvalho, Diogo V., Eduardo M. Pereira, and Jaime S. Cardoso. "Machine learning interpretability: A survey on methods and metrics." Electronics 8.8 (2019): 832.

Local Interpretable Model-Agnostic Explanations (LIME)

Introduced by Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin in 2016.

LIME aims to help users to build trust in a prediction by explaining individual predictions.

The explanation method works on images as well as textual data.

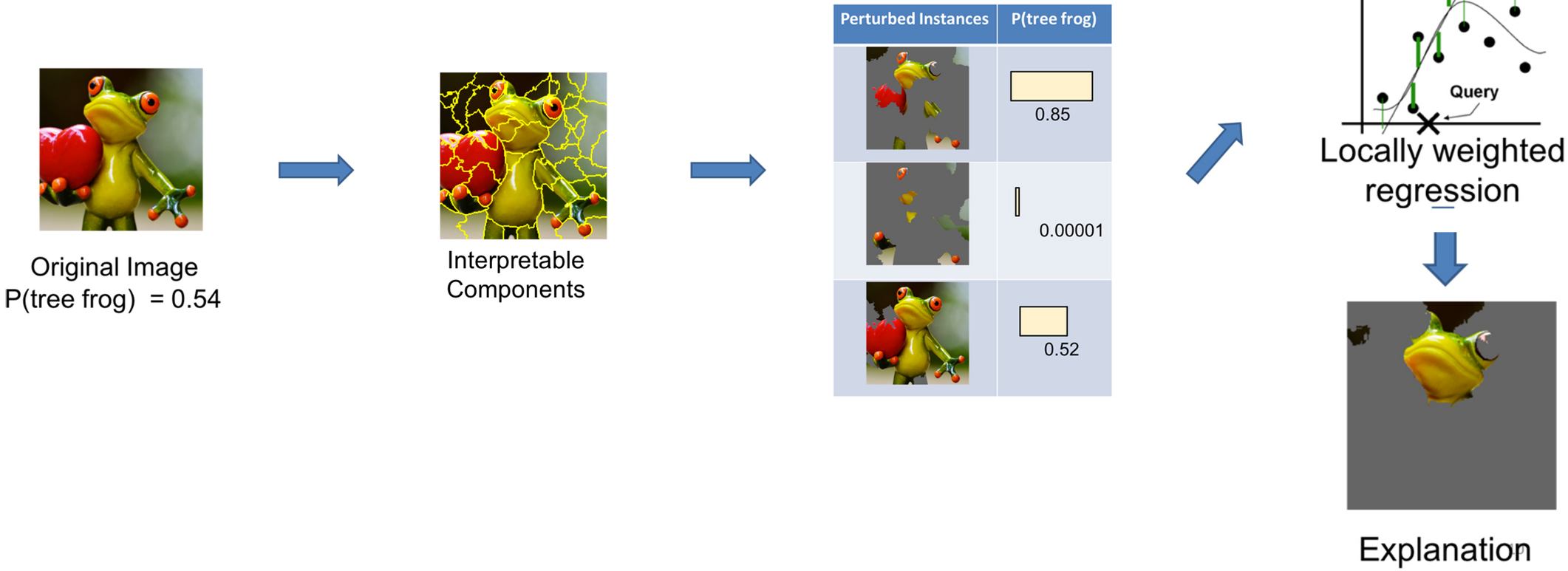
Fast computation but based on heuristics.

Key Ideas:

- Pick a model class interpretable by humans
- Locally approximate global (blackbox) model

Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD. 2016; pp. 1135–1144.
Github: https://github.com/marcotcr/lime

Example 1: Image Classification



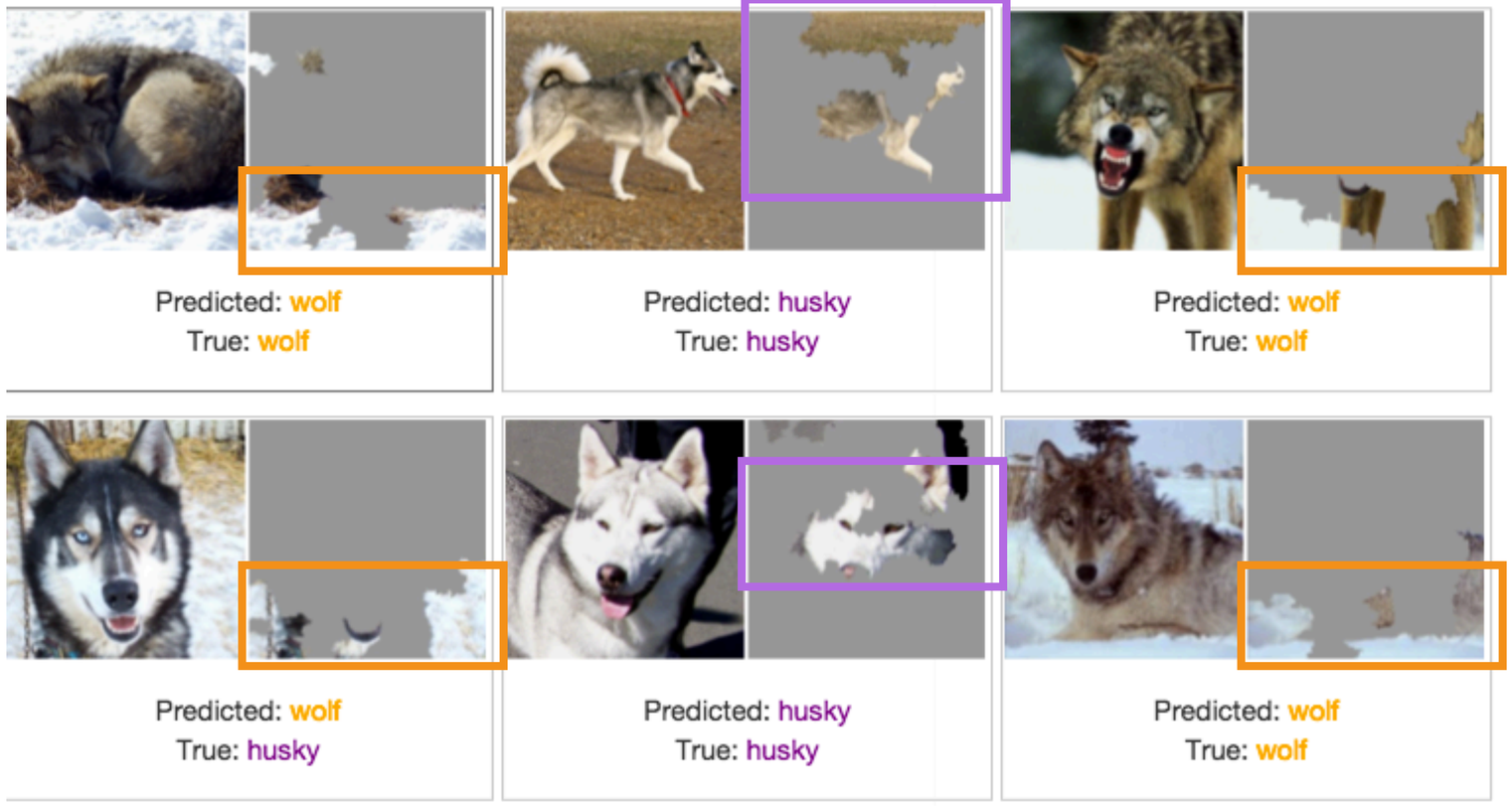
Ribeiro, M.T.; Singh, S.; Guestrin, C. "Introduction to Local Interpretable Model-Agnostic Explanations (LIME)" https://www.kdnuggets.com/2016/08/introduction-local-interpretable-model-agnostic-explanations-lime.html

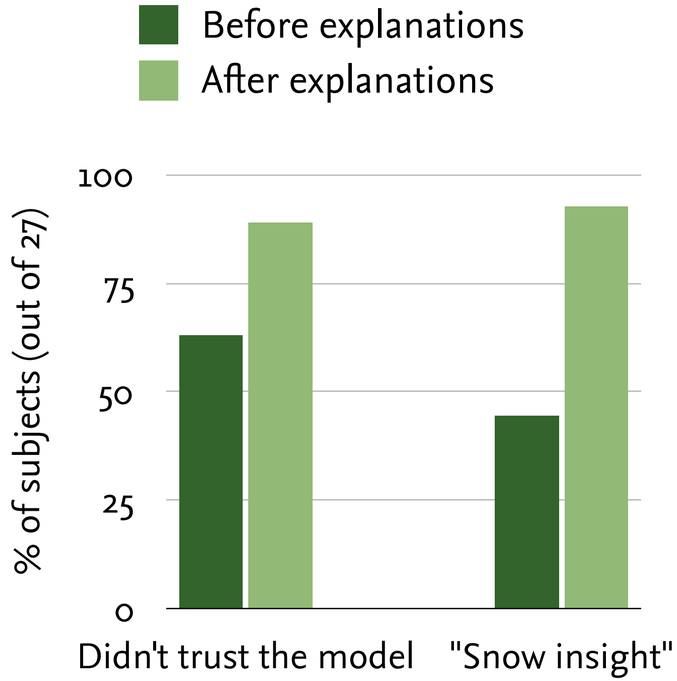
Example 2: Understanding Model Predictions

Only one mistake - do you trust this model?

Slides adapted from Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

Example 2: Understanding Model Predictions (*cont.*)

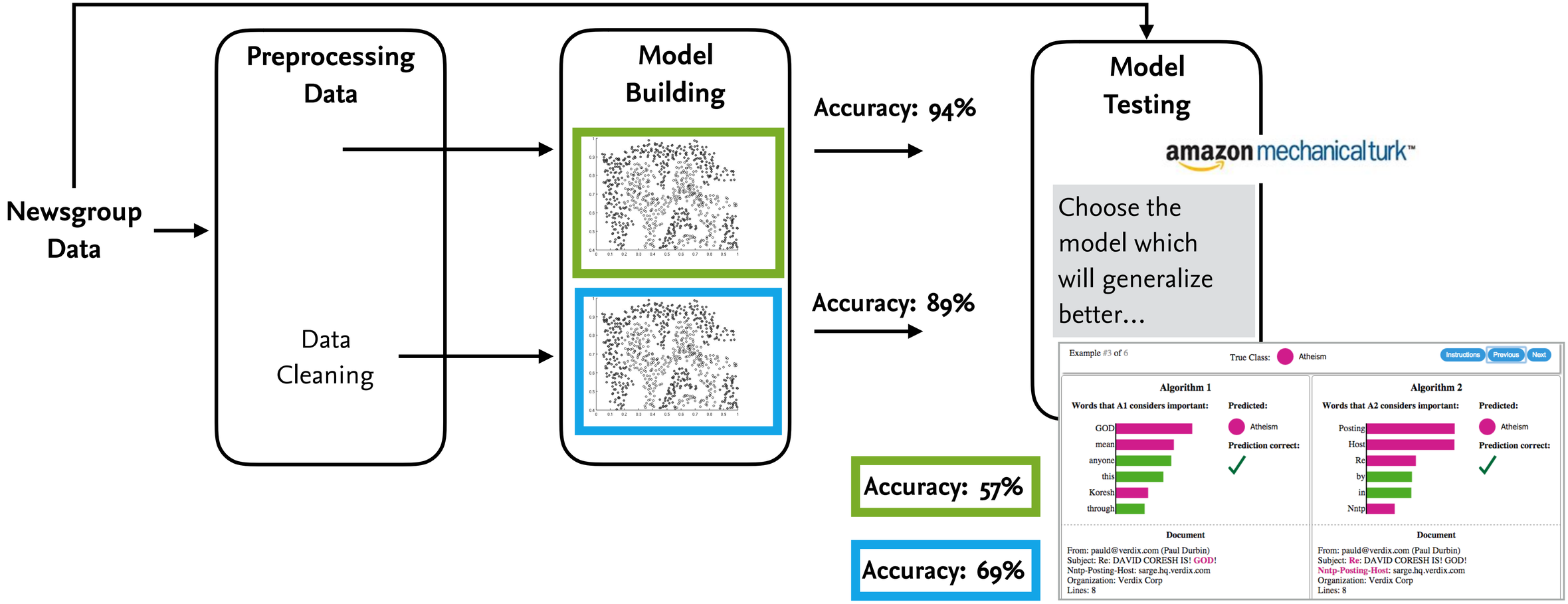




It is a snow detector.

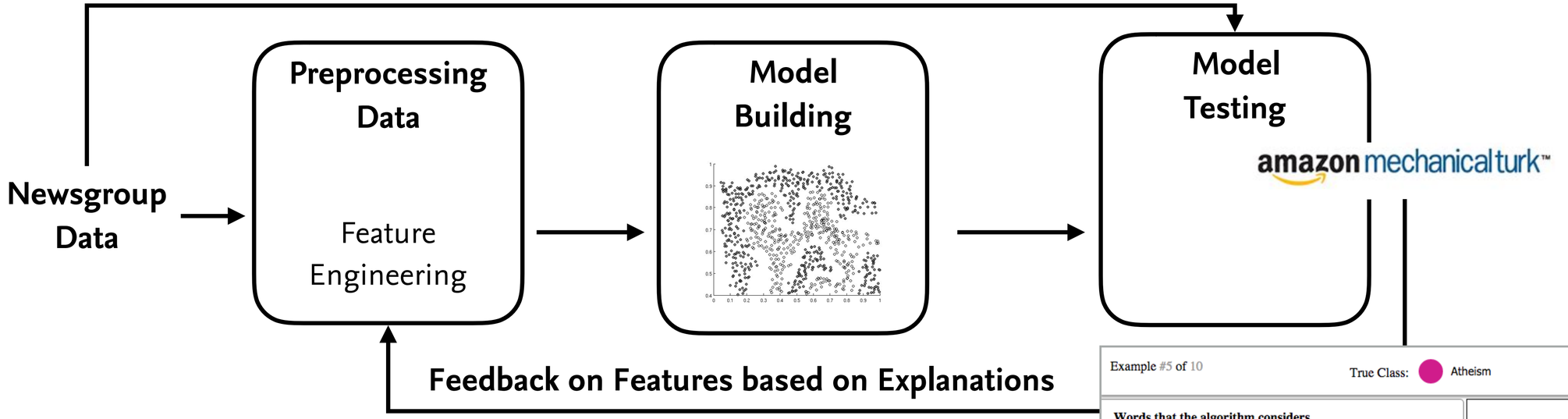
Slides adapted from Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

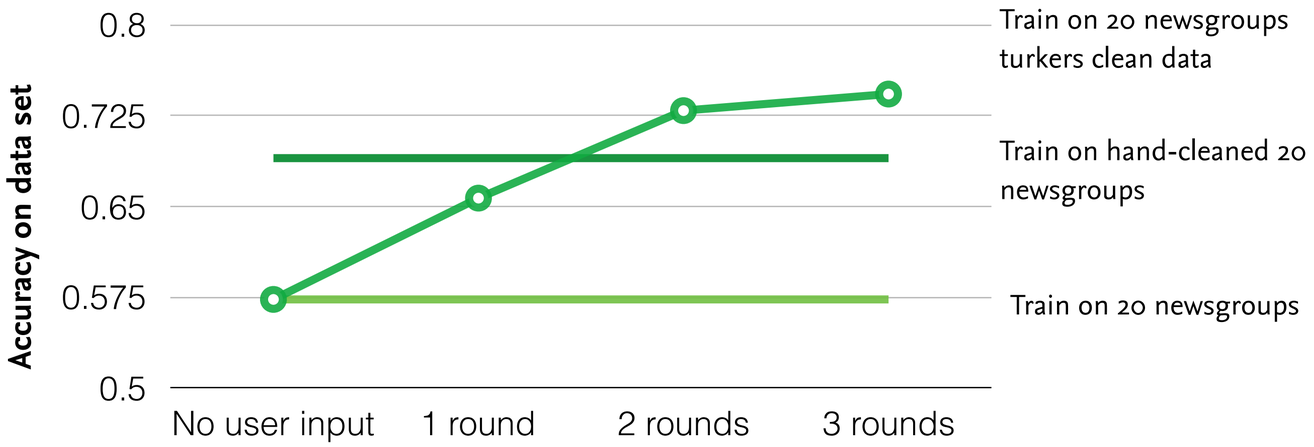
Example 3: Model Selection

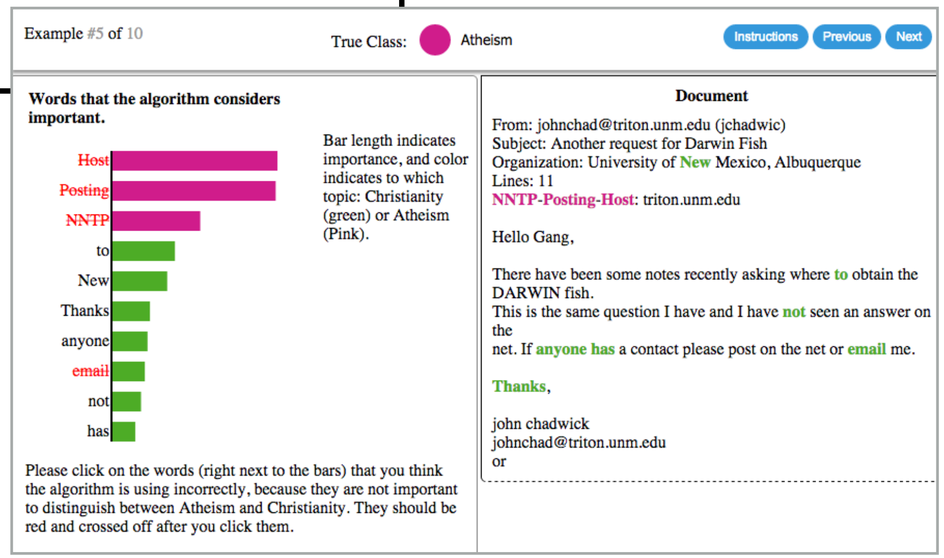


Slides adapted from Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.
Data set at http://qwone.com/~jason/20Newsgroups/

Example 4: Model Improvement







Slides adapted from Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

Microsoft's InterpretML

InterpretML is an open-source package that incorporates many state-of-the-art machine learning interpretability techniques, i.e. explanation methods (e.g., LIME, SHAP).

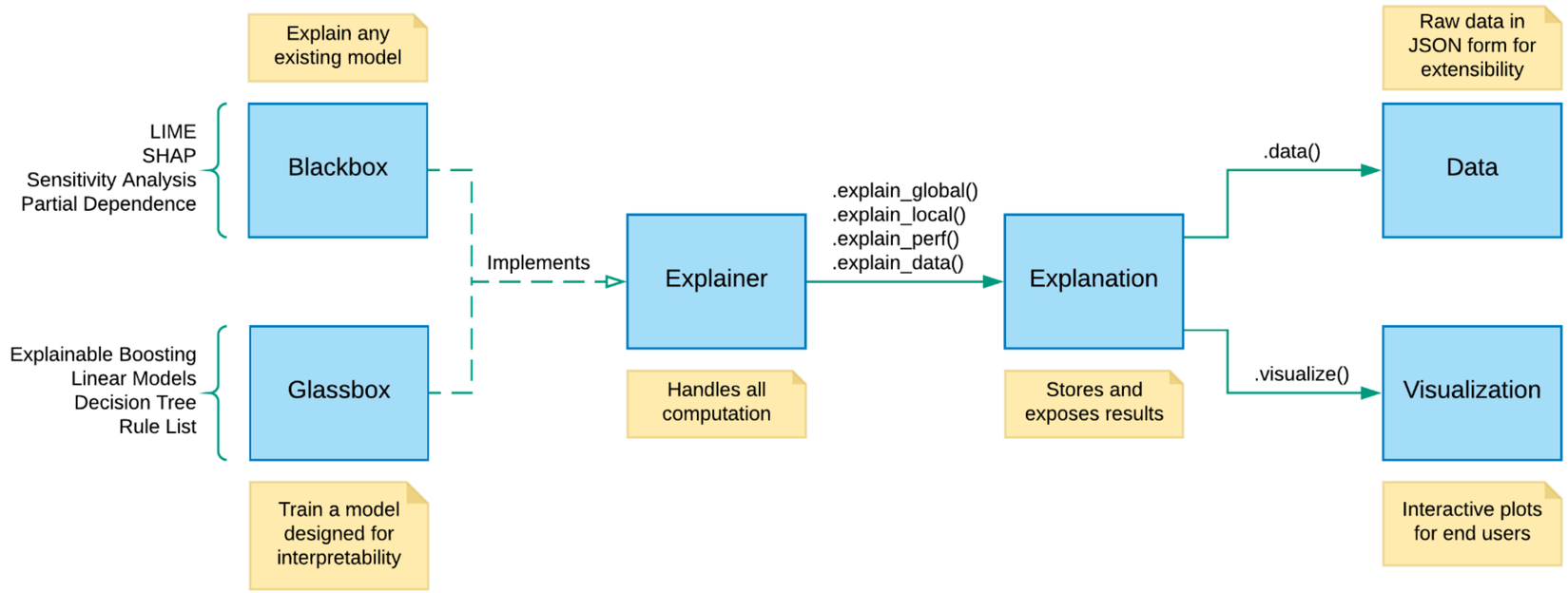
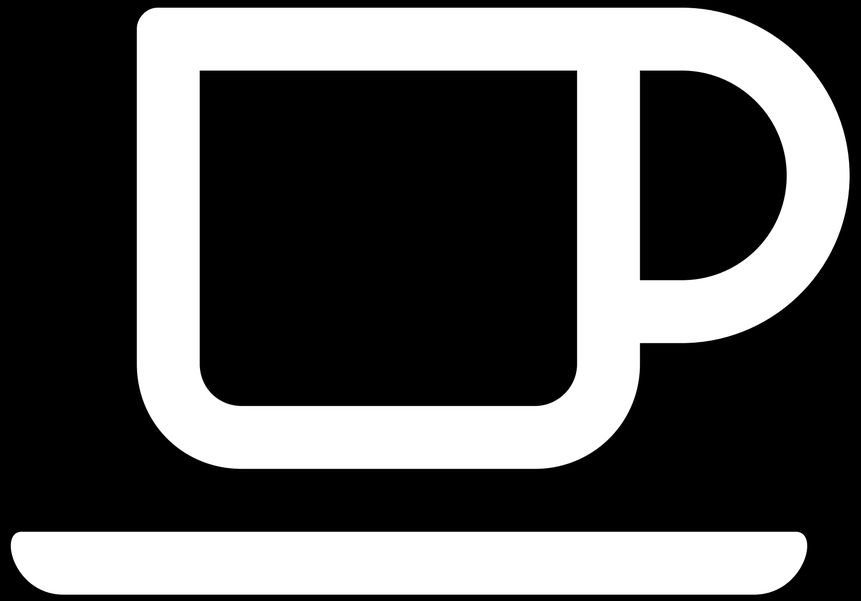


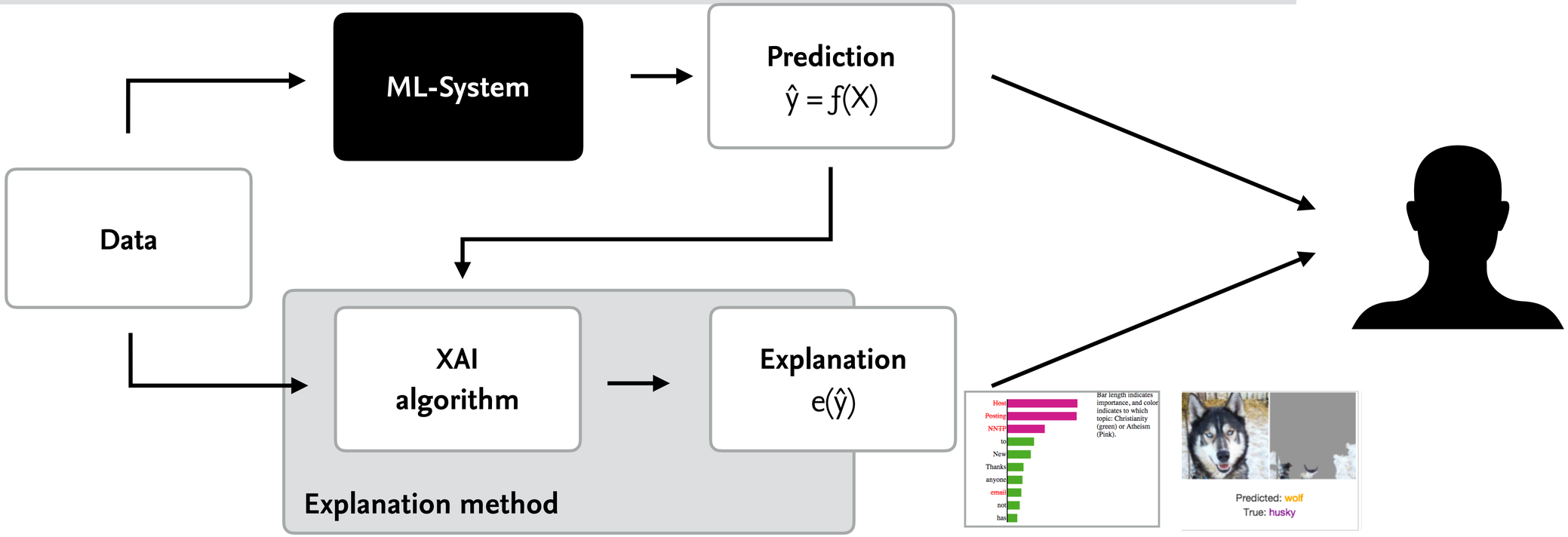
Image taken from Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). Interpretml: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223.
Github Repo: https://github.com/interpretml/interpret



5 minutes break

Explanation Method

An explanation method is a pattern or a mechanisms that generates explanations to establish post-hoc interpretability. An explanation method is based on an explainable AI (XAI) algorithm.



Liao, Q. V., Gruen, D., & Miller, S. (2020, January 8). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. http://doi.org/10.1145/3313831.3376590
Image from Carvalho, Diogo V., Eduardo M. Pereira, and Jaime S. Cardoso. "Machine learning interpretability: A survey on methods and metrics." Electronics 8.8 (2019): 832.

Explanation as a Product

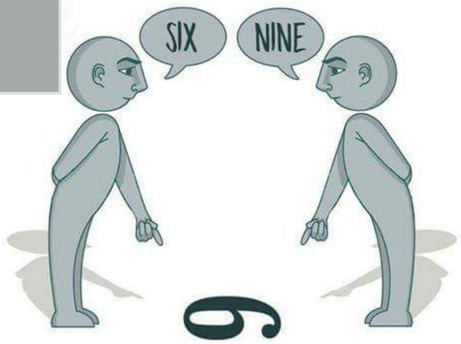
Non-pragmatic Theory of Explanation

The explanation should be the correct answer to the why-question.

Pragmatic Theory of Explanation

The explanation should be a good answer for an explainer to give when answering the why-question to an audience.

“Rashomon Effect”



Kim, B.; Khanna, R.; Koyejo, O.O. Examples are not enough, learn to criticize! Criticism for interpretability. In Advances in Neural Information Processing Systems; MIT Press, 2016; pp. 2280–2288.

Explanation as a Process

Cognitive Dimension

» Relates to knowledge acquisition.

» An explanation is derived by abductive inference, i.e., the causes of an event are identified and, then, a subset of these causes are selected as the explanation.

Social Dimension

» Relates to the social interaction.

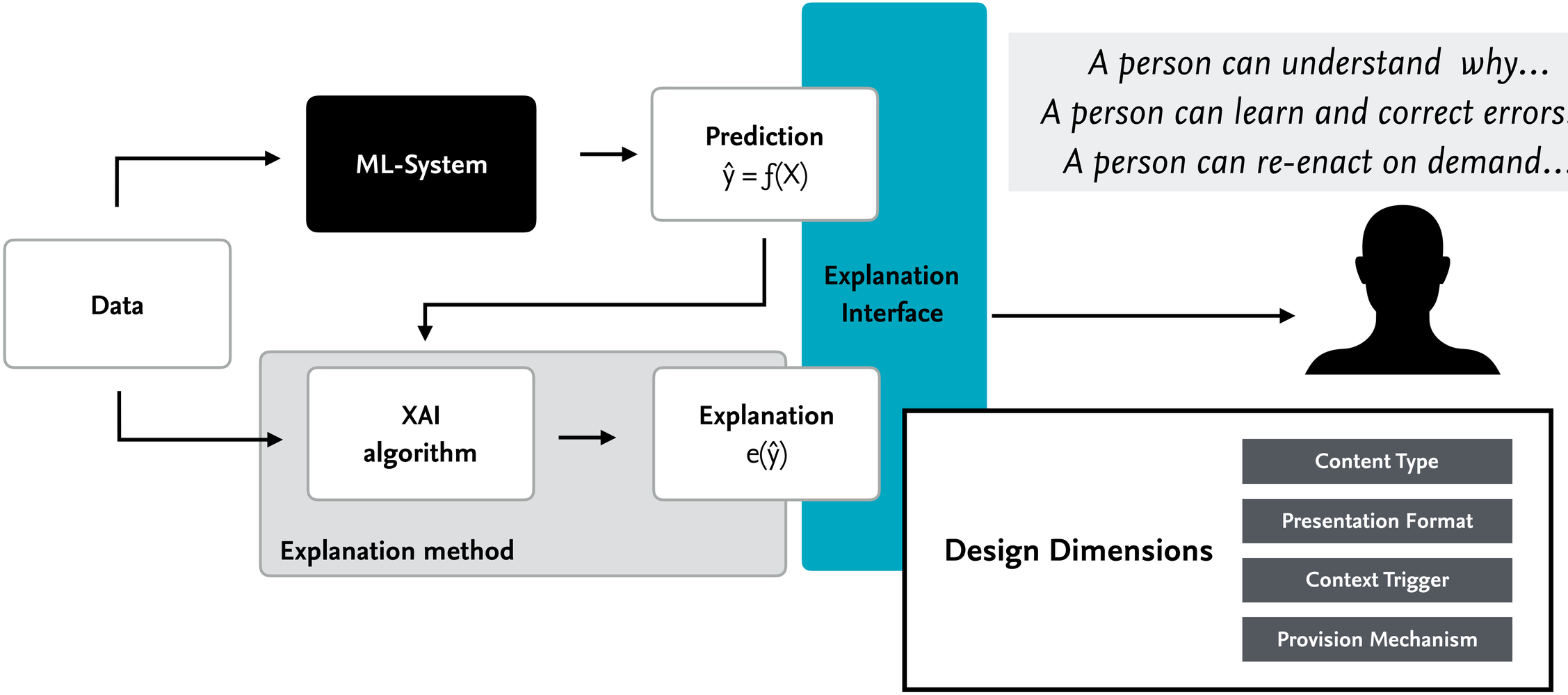
» Knowledge is transferred from the explainer to the explainee.

» The explainee need to receive enough information to understand the causes of an event or a decision.

These dimensions emphasize the subjectivity of explanations, highlighting the need to adapt the explanation to the audience.

Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." Artificial Intelligence 267 (2019): 1-38..

Explanantions to Explanation User Interfaces



A Participatory Process for Transparency Design

WHAT to explain?

Expert Mental Model

What happens to the best of our knowledge?

What can be explained?

What does an expert mental model of the system look like?

User Mental Model

How do users currently make sense of the system?

What is the user mental model of the system based on its current UI?

How does it differ from the expert mental model?

Target Mental Model

Which key components of the algorithm do users want to be made transparent in the UI?

To what extent are users actually interested in the rationale behind the algorithms?

HOW to explain?

User Mental Model ⟷ Target Mental Model

How can the target mental model be reached through UI design?

How and where can transparency be integrated into the UI of the system?

How has the users mental model developed?

Has the target mental model been reached?

Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In Conf. on Intelligent User Interfaces, 211–223.

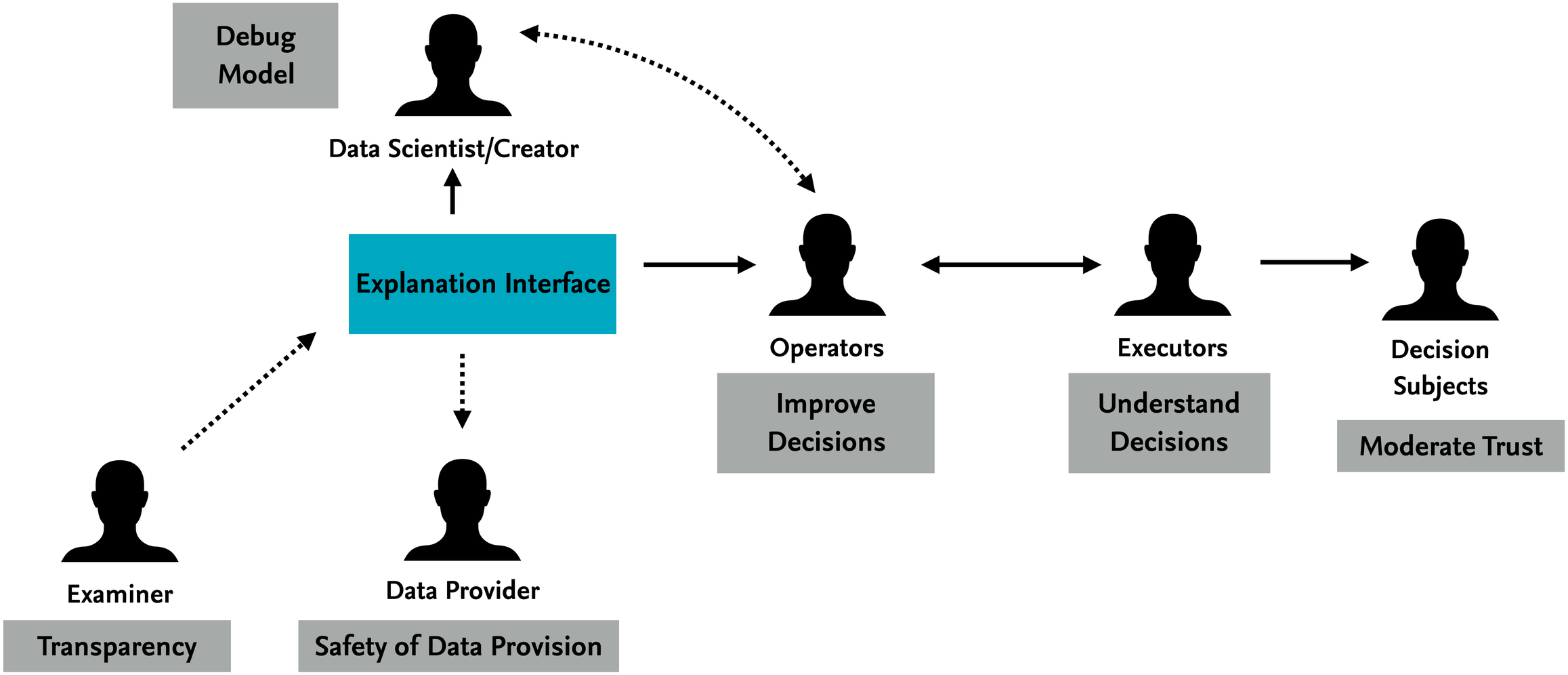
Explanations from a Human-Centered Design Perspective

Audience: For whom are the explanation made?

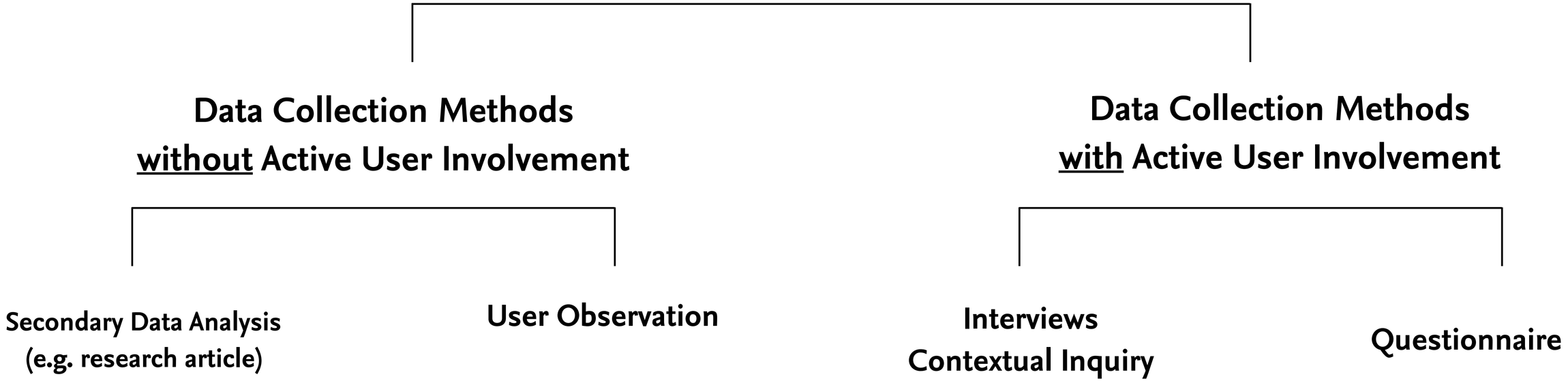
Purpose: How will your explanation being used?

Context: What factors will impact whether/how they use it?

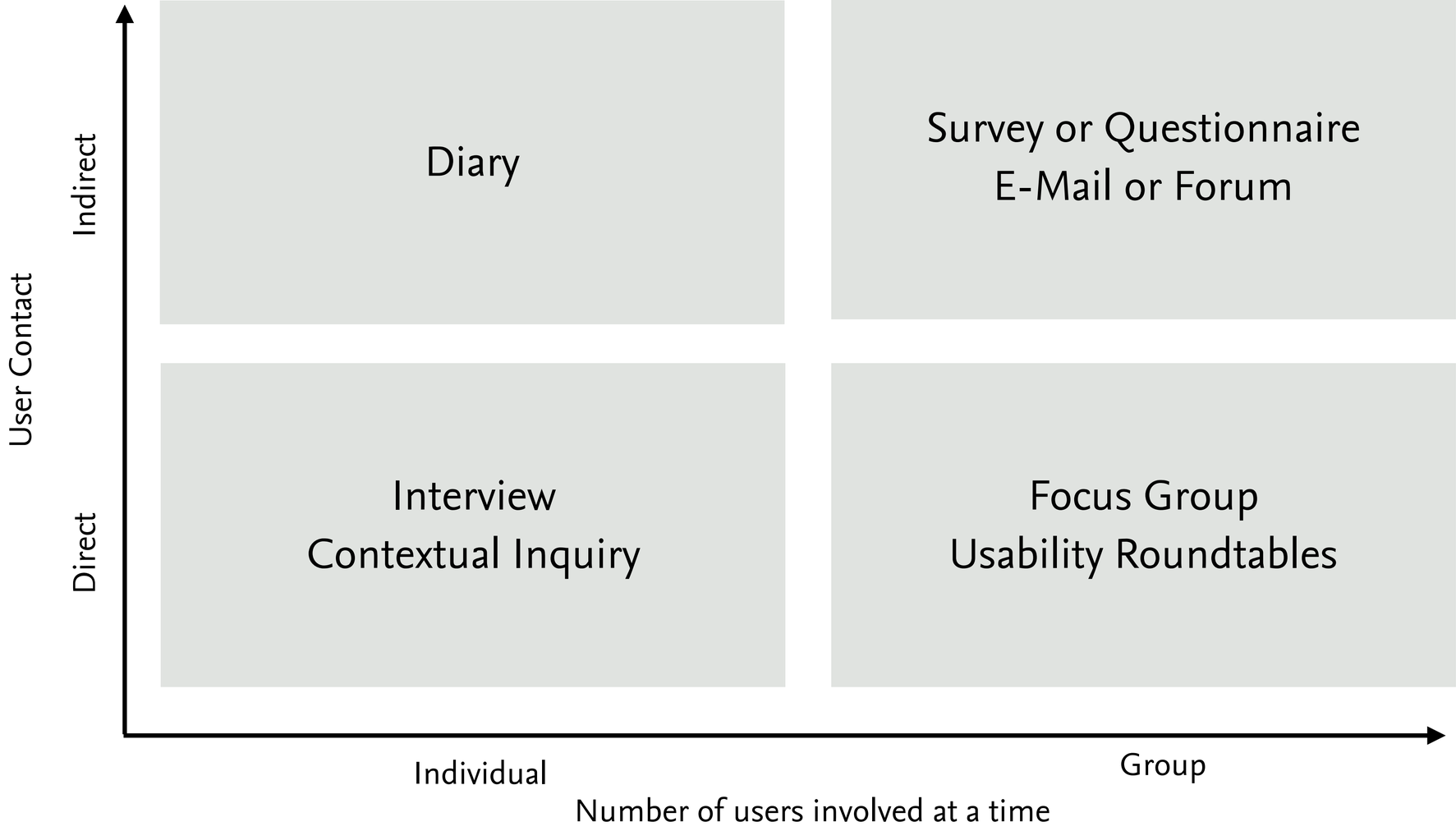
Different Explanation Needs



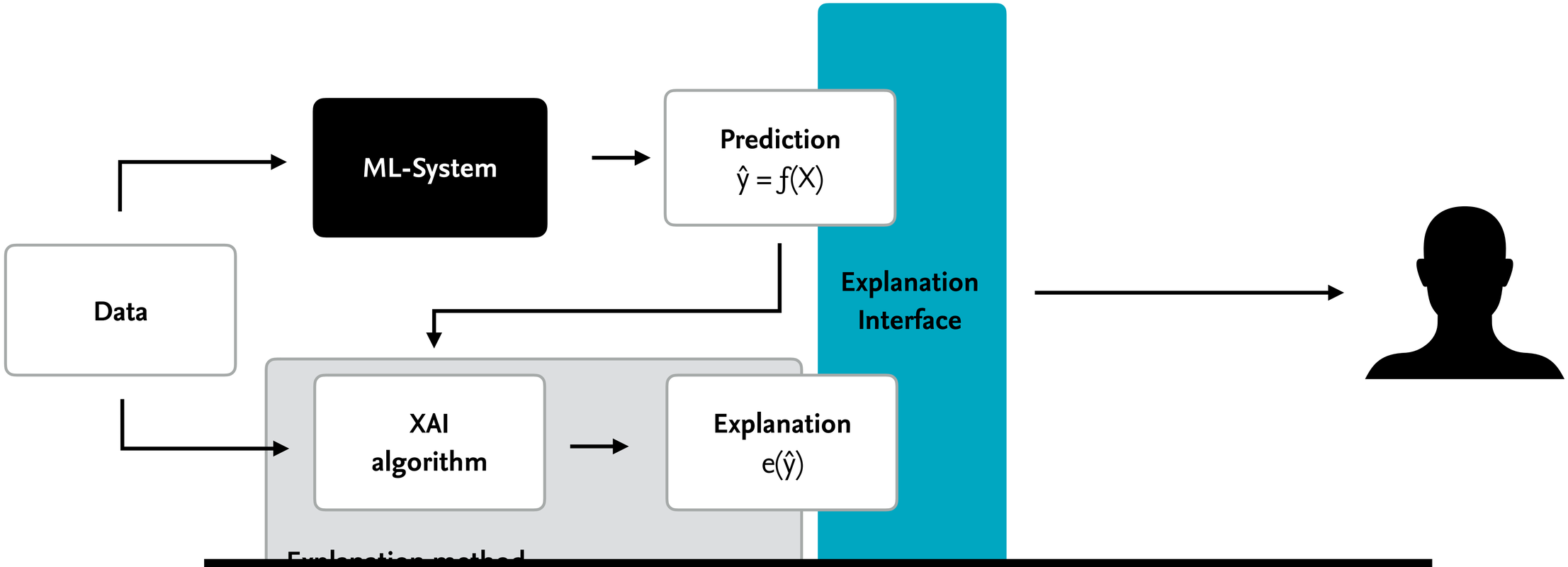
Eliciting Explanation Needs



Dimensions of Methods with User Involvement



Explanation User Interface



How do data scientists perceive and use interpretability tools?
What are key challenges towards their use of these tools?

Understanding Data Scientists' Use of Interpretability Tools

Context: Data scientist and ML practitioners at a large technology company in the US

Used Data Set: Adult Data Set

Model: Light Gradient Boosting Machine (LightGBM)

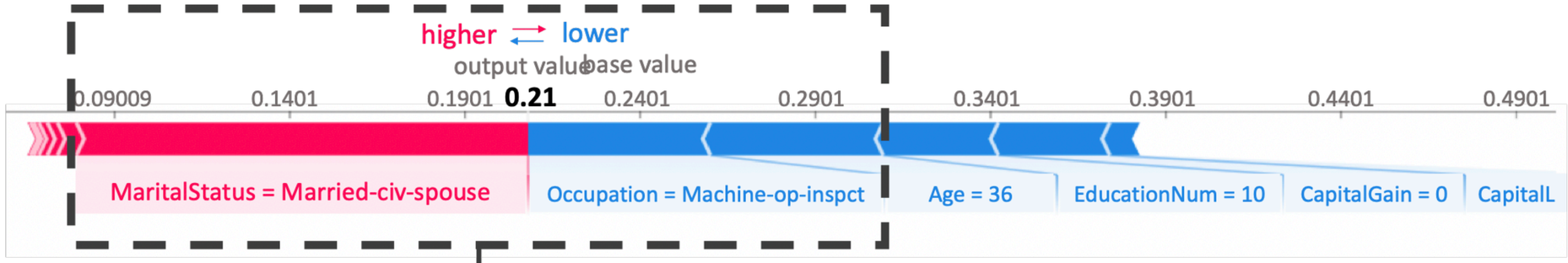
Used Explanation Method, i.e., interpretability tool: SHAP Python Package consisting of

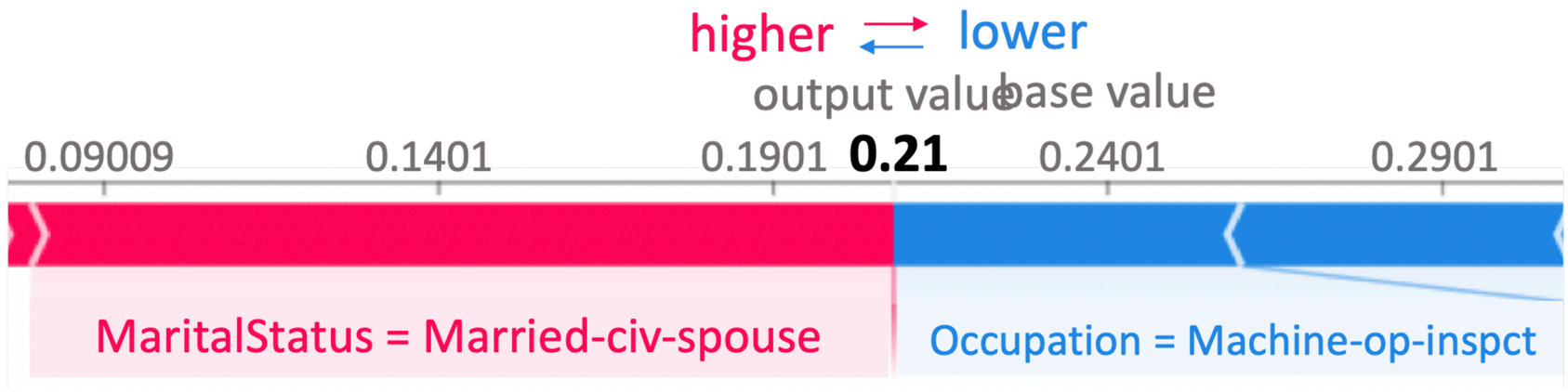
- Local explanations for individual predictions
- Global explanations (by aggregating the importance scores for many predictions)
- Dependence plots for single input features

https://archive.ics.uci.edu/ml/datasets/Adult

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, 1–14. DOI:https://doi.org/10.1145/3313831.3376219

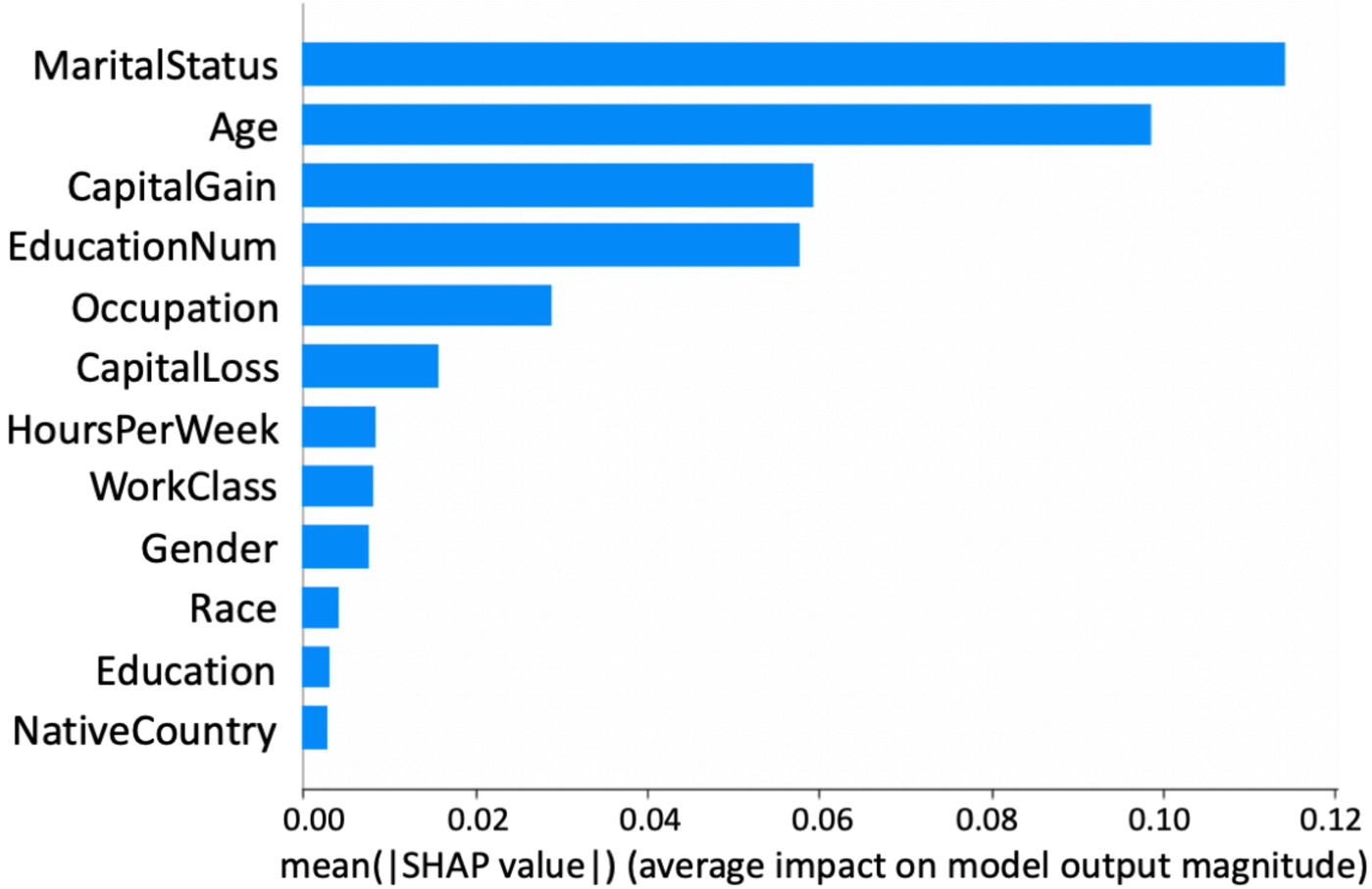
Local Explanations How an Individual Prediction was made





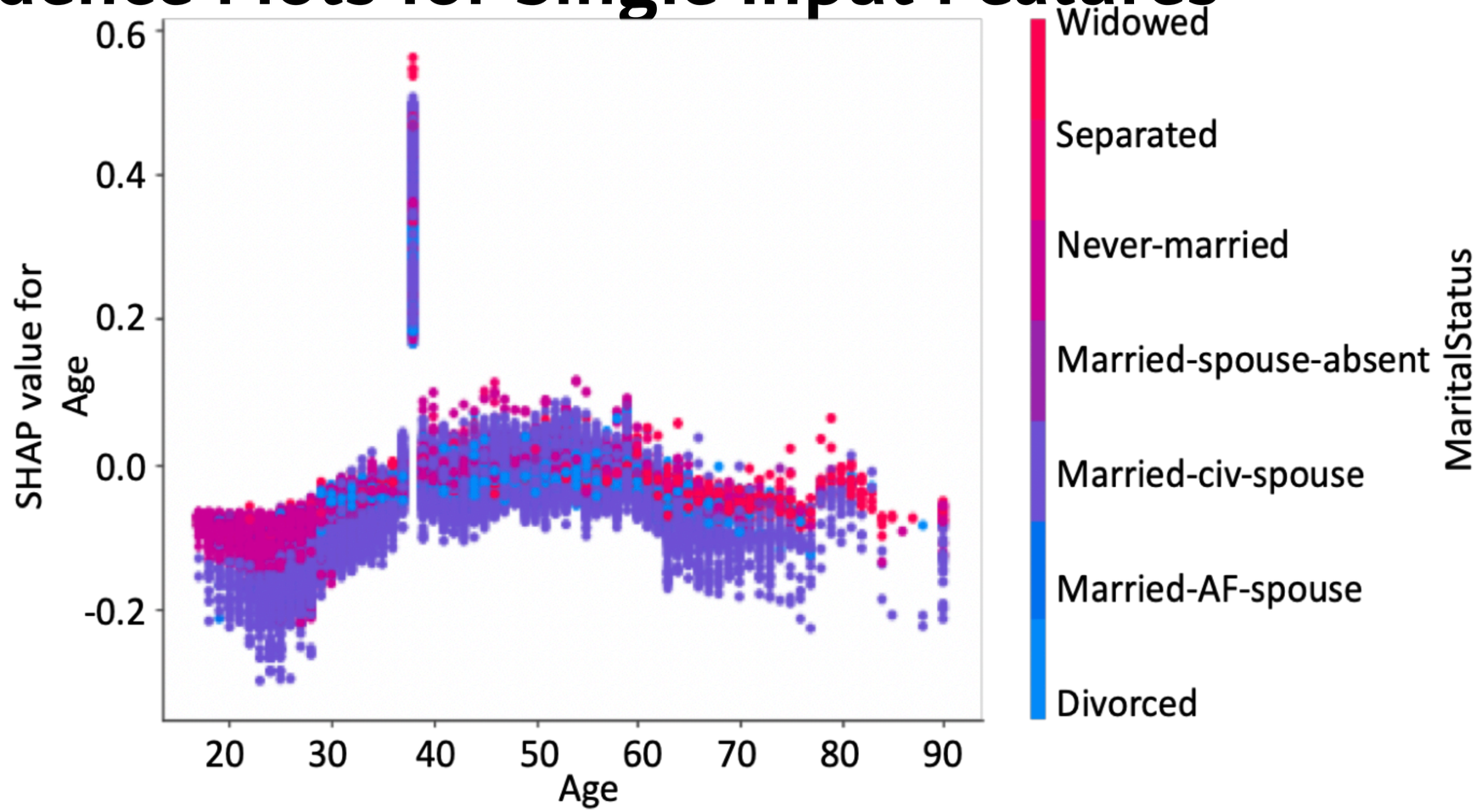
Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, 1–14. DOI:https://doi.org/10.1145/3313831.3376219

Global Explanations What the Model Learned Overall from Training Data



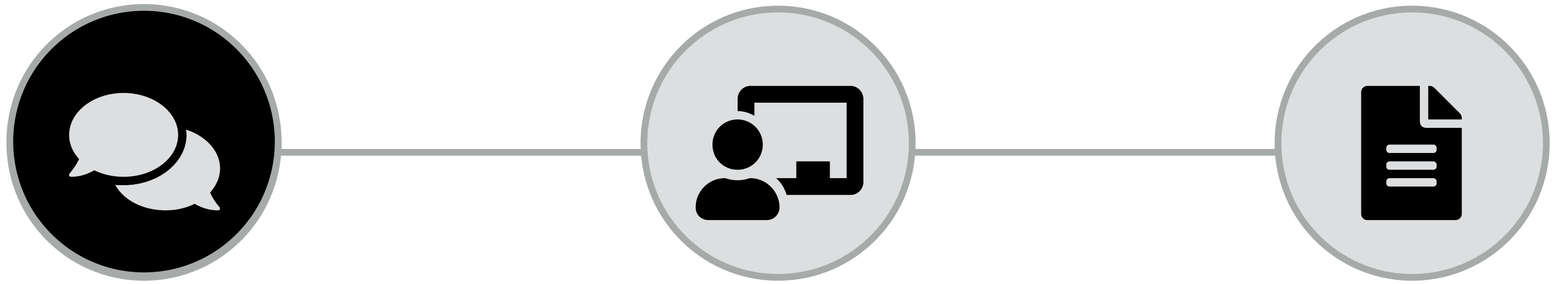
Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, 1–14. DOI:https://doi.org/10.1145/3313831.3376219

Dependence Plots for Single Input Features



Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, 1–14. DOI:https://doi.org/10.1145/3313831.3376219

Study Design



Pilot Interviews
(N=6)

Contextual Inquiry
(N=11)

Survey
(N=197)

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, 1–14. DOI:https://doi.org/10.1145/3313831.3376219

Types of Interviews

Unstructured: Similar to conversations that focus on one topic.

(+) Provides more detailed answers and allows more spontaneity

(-) More time consuming, possibility of losing control of the interview

(-) Generates a lot of information

Structured: Very well planned and controlled interviews.

(+) Efficient, can save time and cover lots of ground quickly

(+) Easy to compare interview results

(-) Can get boring for interviewee

(-) May miss details

Semi-structured: A focused interview design that allows a certain degree of flexibility.

Common Issues faced by Data Scientists

Theme	Description	Incorporation into Contextual Inquiry
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."
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.
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."
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."
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."
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.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, 1–14. DOI:https://doi.org/10.1145/3313831.3376219

Study Design

Pilot Interviews
(N=6)

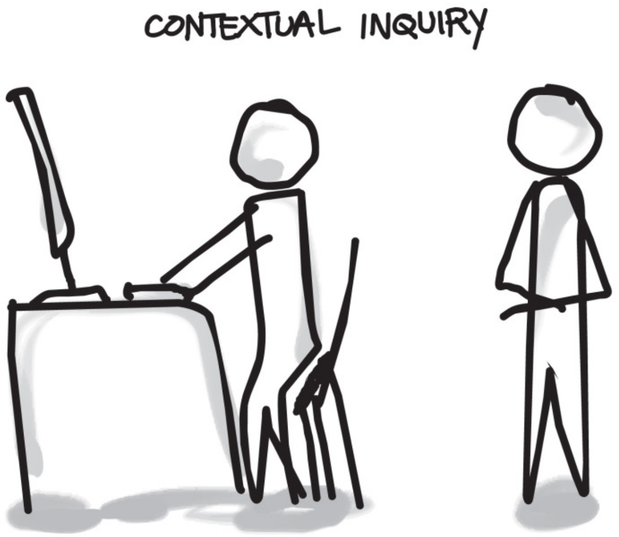
Contextual Inquiry
(N=11)

Survey
(N=197)

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, 1–14. DOI:https://doi.org/10.1145/3313831.3376219

Contextual Inquiry

Combine the strengths of both, interviews and user observation, to get findings that are more complete and convincing.



Master-Apprentice-Model of Learning by Beyer and Holzblatt

- »Observing and asking questions of the user as if she is the master craftsman
- »The interviewer the new apprentice

Discover the real requirements of the work.

Drives the creative process in original design or in considering new features or functionality.

Results of Contextual Inquiry

Misalignment between data scientists' understanding of interpretability tools and these tools' intended use.

Misuse resulted from over-trusting the tools because of their visualizations but also because of their availability as open source package.

Under-use of tools because of a limited understanding of the provided explanations and their meaning.

The shown visualizations are **misleading** because of their lack of explanations.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, 1–14. DOI:https://doi.org/10.1145/3313831.3376219

Study Design

Pilot Interviews
(N=6)

Contextual Inquiry
(N=11)

Survey
(N=197)

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, 1–14. DOI:https://doi.org/10.1145/3313831.3376219

Surveys

» In the process of conducting surveys, questionnaires are used.

» Questionnaires are a well-established technique for collecting demographic data and users' opinions.

» Questionnaires needs effort and skill are needed to ensure that questions are clearly worded and the data collected can be analyzed efficiently.

» It can be harder to develop good questions compared with structured interview questions.

- To what extent is Machine Learning a part of your daily job role?
  - Scale 1-7 (where 1 = Not at all, and 7 = Extremely)
- How long have you been practicing Machine Learning? Please enter time in months (e.g., 6 for 6 months, 12 for 1 year)
  - Textbox
- How familiar are you with interpretability tools for machine learning?
  - Scale 1-7
- How familiar are you with SHAP, a tool for explaining black box models?
  - Scale 1-7
- How many hours (estimate) have you spent using SHAP, a tool for explaining black box ML models?
  - I have not used SHAP before, Less than 10 hours, 10-20 hours, 20-50 hours, 50-100 hours, More than 100 hours
- How familiar are you with GAMs (Generalized Additive Models)?
  - Scale 1-7
- How many hours (estimate) have you spent building and using GAMs (Generalized Additive Models)?
  - I have not used GAMs before, Less than 10 hours, 10-20 hours, 20-50 hours, 50-100 hours, More than 100 hours
- Are there other interpretability tools you are familiar with? Please list these below
  - Textbox
- How many hours (estimate) have you spent using other interpretability tools?
  - I have not used any other interpretability tools, Less than 10 hours, 10-20 hours, 20-50 hours, 50-100 hours, More than 100 hours

Results of Survey

Factors that Affect Willingness to Deploy

» Most participants gave the models high deployment ratings based on intuition, instead of critically evaluating the explanations.

» Participants who gave the underlying models low deployment ratings used the interpretability tools in their intended ways.

Mental Models of Interpretability Tools

» Participants did not use the interpretability tools as intended. Qualitative analysis of participants' descriptions of the visualizations indicates that most participants did not have an accurate understanding of the visualizations.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, 1–14. DOI:https://doi.org/10.1145/3313831.3376219

Insights from the Study

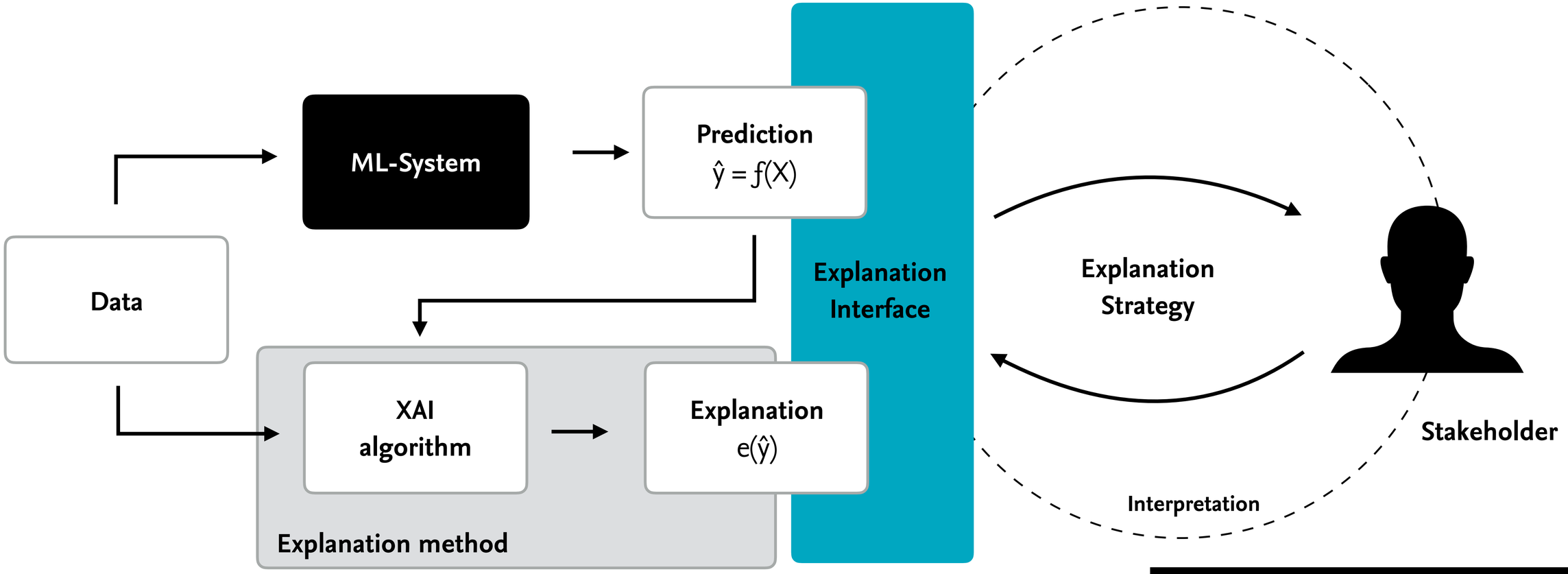
Participants often do not fully understand the underlying conceptional model of the tool.

The “ease of use” of interpretability tools reduces participant’s critically thinking.

The diagrams conveyed trust in the quality of the underlying explanation method.

So what?

Ensuring Interpretability by Explanation User Interfaces



Next week more :)

Jesse Josua Benjamin, Christoph Kinkeldey, Claudia Müller-Birn, Tim Korjakow, and Eva-Maria Herbst. 2022. Explanation Strategies as an Empirical-Analytical Lens for Socio-Technical Contextualization of Machine Learning Interpretability. Proc. ACM Hum.-Comput. Interact. 6, GROUP.

Course Evaluation

Check your Insights

What is the difference between intrinsic and post-hoc interpretability?

What is an explanation method? What is an explanation?

What are model-specific and model-agnostic explanation methods?

What are possible explanation needs that should be considered when designing explanations?

When looking at explanation as a process, what are the considered dimensions?

What are useful characteristics of an explanation user interface?

«Human-Centered Data Science»

Next week: Post-hoc Interpretability: Understanding Human Cognitive Abilities

Prof. Dr. Claudia Müller-Birn

Human-Centered Computing, Institute of Computer Science

Freie Universität Berlin

June 30, 2022