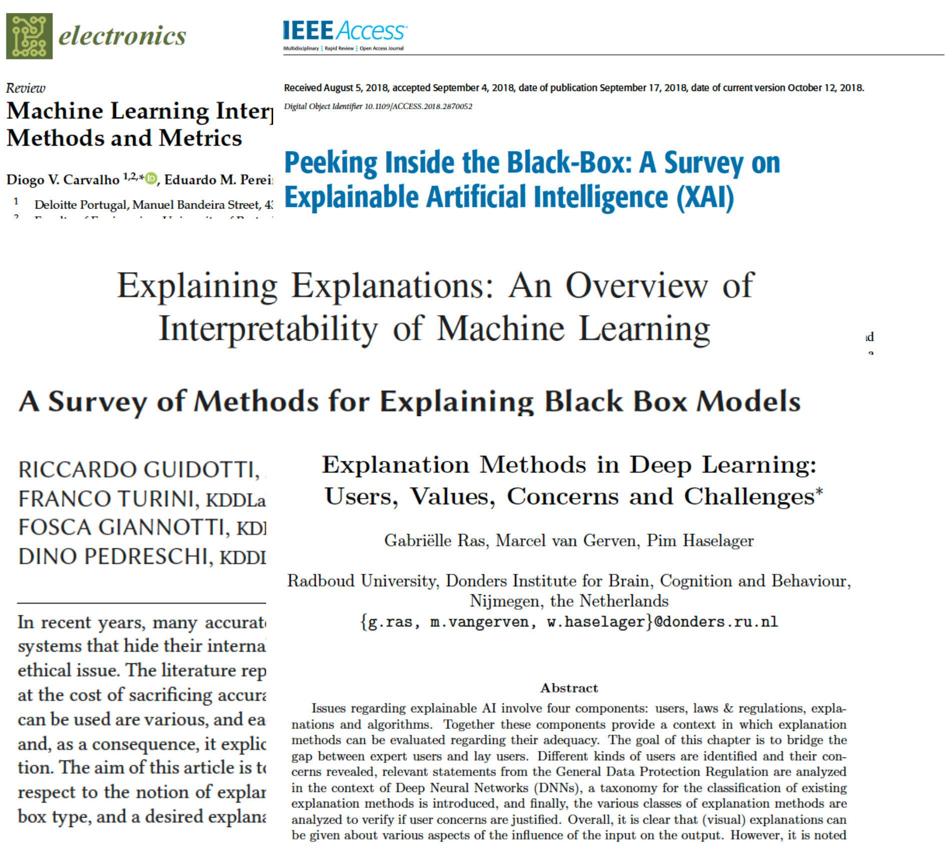
Questioning the AI: Toward Human Centered Interpretable Machine Learning

Research work 2018-2020

Q. Vera Liao
IBM Research

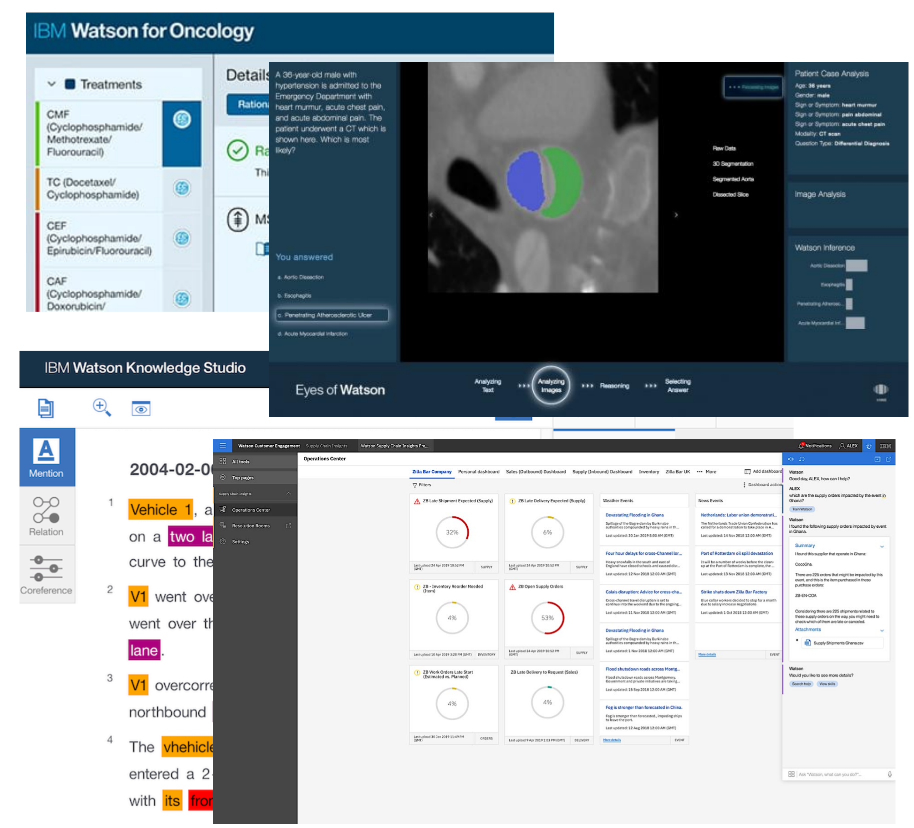
HCI research: **Bridging** work

Transfer emerging research or technologies into tangible ***tools*** and ***guidelines*** that help product teams navigate the design space



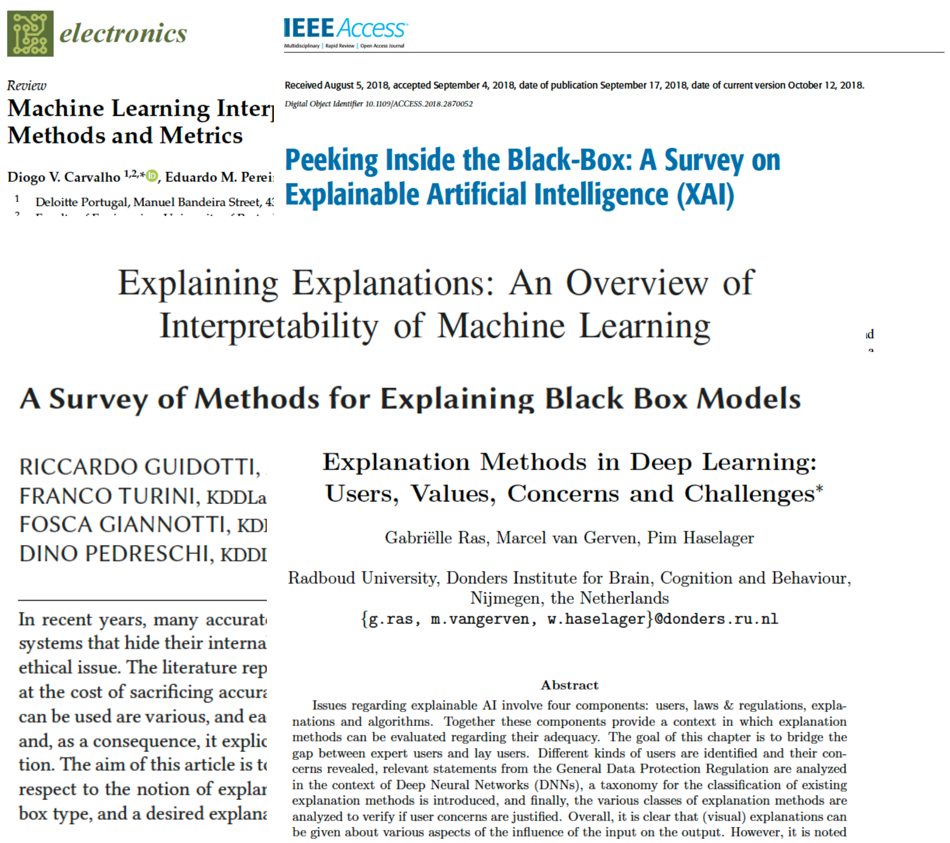
Inform usage

Identify gaps



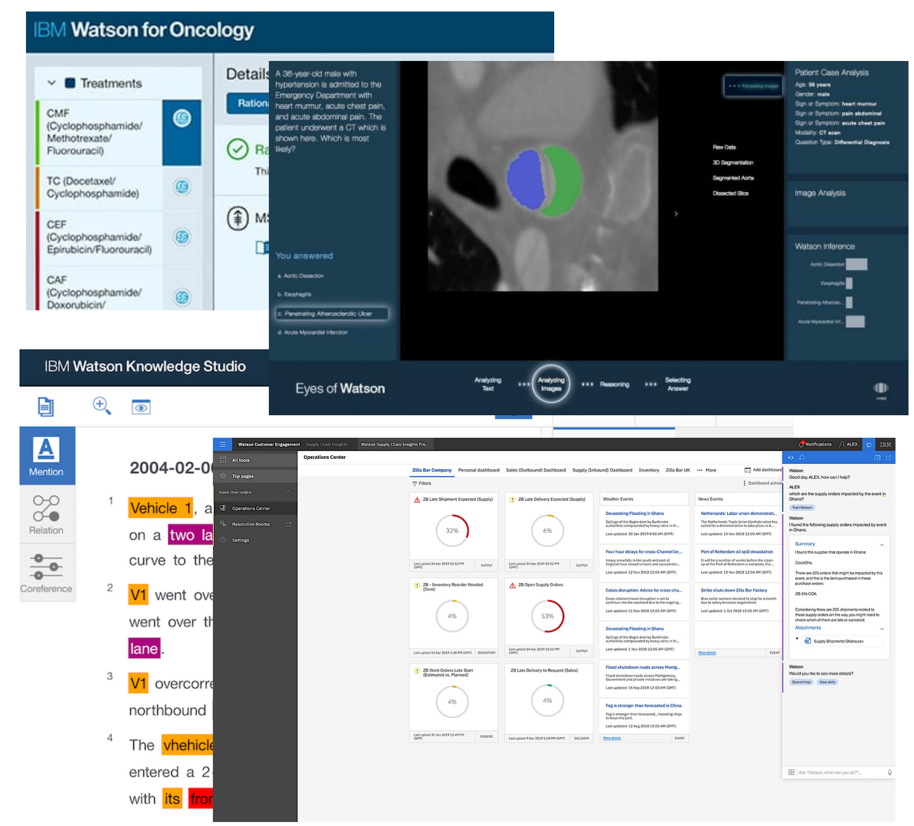
HCI research: **Bridging** work

Transfer emerging research or technologies into tangible ***tools*** and ***guidelines*** that help product teams navigate the design space



Inform usage

Identify gaps



Which explanation technique to use?

How to design XAI user experiences?

Terminologies and definitions

Interpretable ML $\approx$ **Explainable AI (XAI)**

Narrow definition:

Techniques and methods that make a ML model's decisions understandable by people

Broader (practitioners') definition:

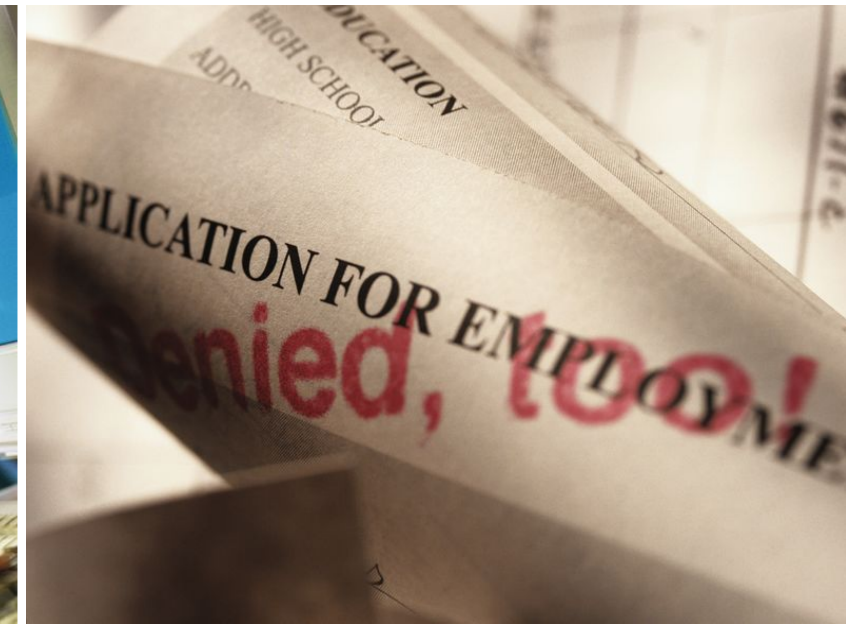
Everything that makes AI more understandable (e.g., also including data, functions, performance)

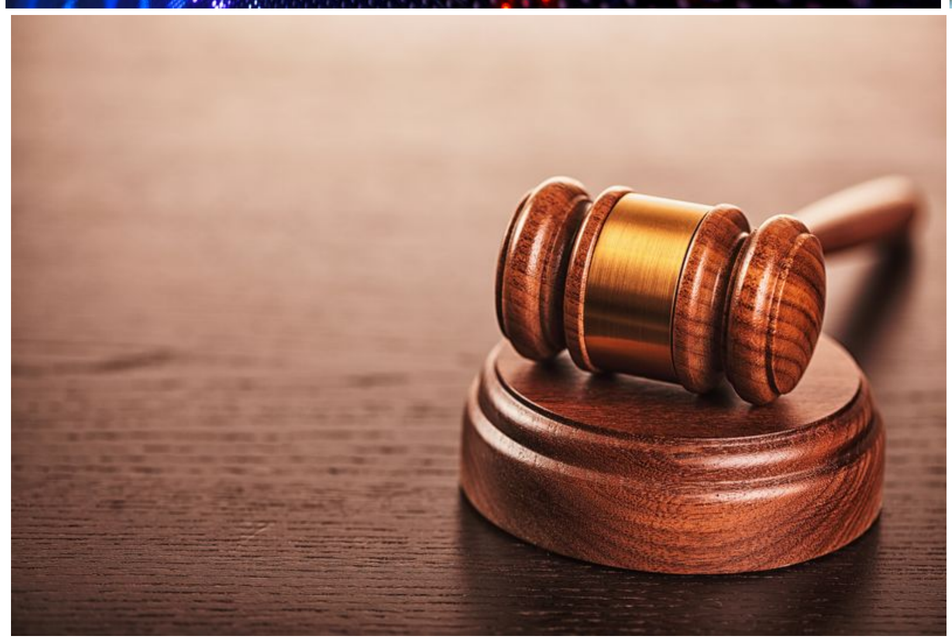
XAI is not just ML (also explainable robotics, planning, etc.), but I will focus on **explaining supervised ML**

Towards human centered XAI: Agenda

- **Background and motivation for HCXAI**
- **Research into design**
  - **Question-driven explainable AI (🎖 CHI 2020)**
  - Designing social transparency in AI systems (CHI 2021)
- **Research through design and case studies**
  - **XAI for fair ML (🎖 IUI 2019)**
  - XAI for AI decision support (FAccT* 2020)
  - XAI for active learning (CSCW 2020)
  - XAI for autoAI (IUI 2021)

AI is increasingly used in many high-stakes tasks





The quest for explainable AI (XAI)

Companies Grapple With AI's Opaque Decision-Making Process

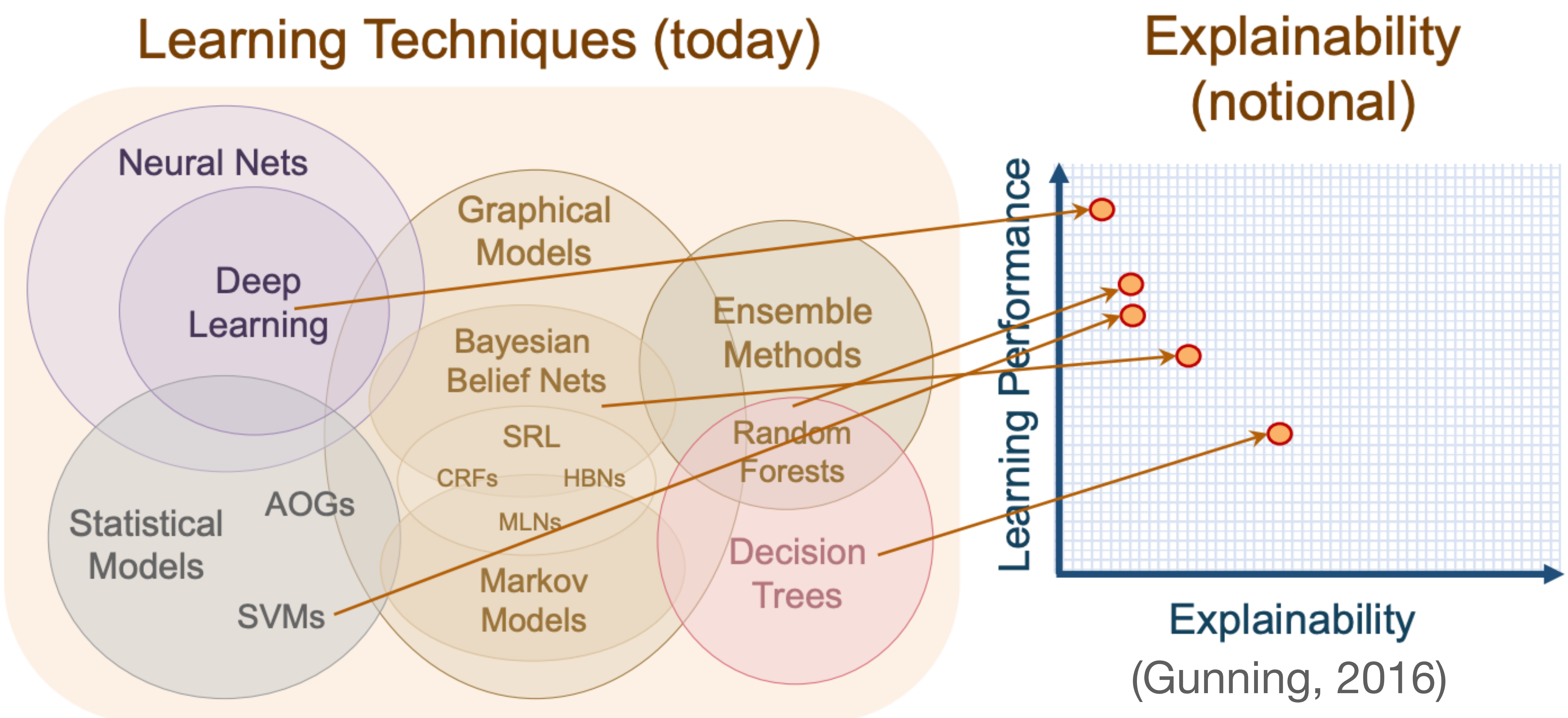
We Need AI That Is Explainable, Auditable, and Transparent

Why "Explainability" Is A Big Deal In AI

From black box to white box: Reclaiming human power in AI

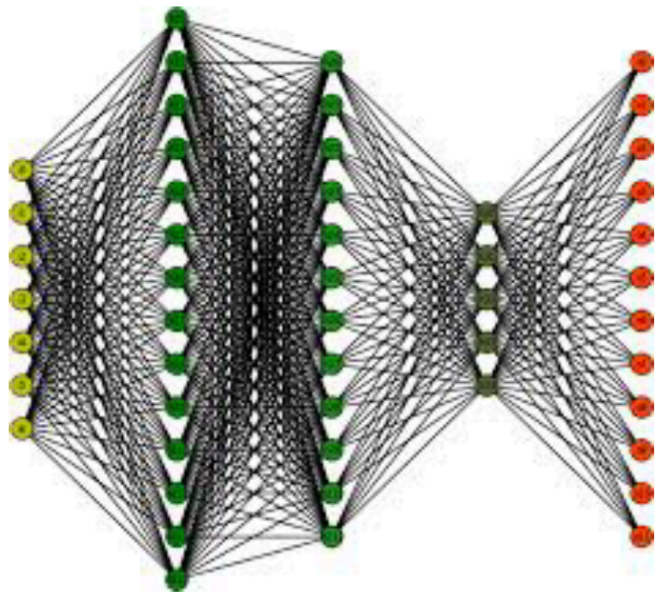
How Explainable AI Is Helping Algorithms Avoid Bias

XAI is hard: it is technical

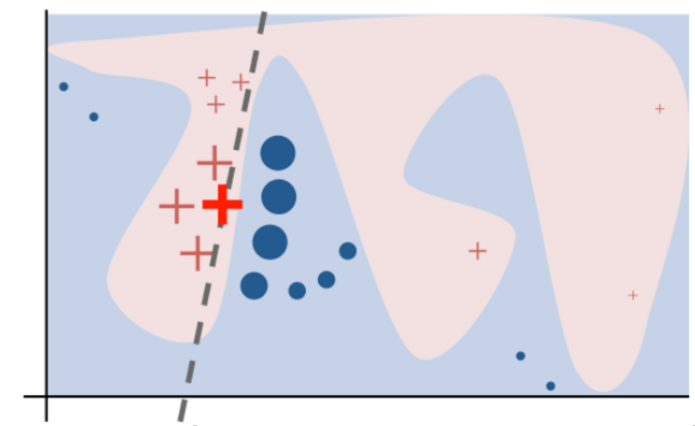


(Gunning, 2016)

XAI is hard: it is technical

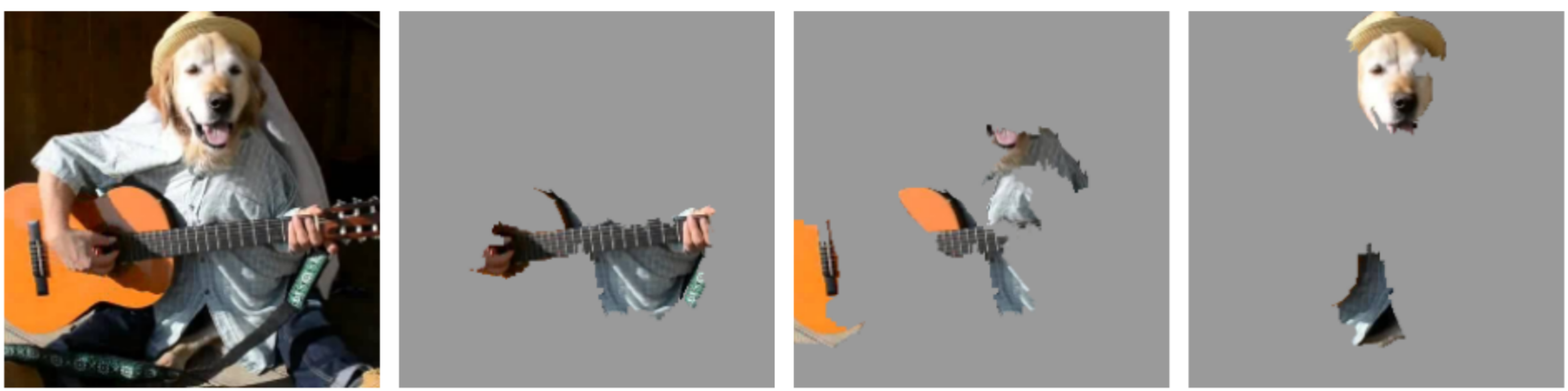


Neural network, not directly explainable



LIME (Ribeiro et al. 2016)

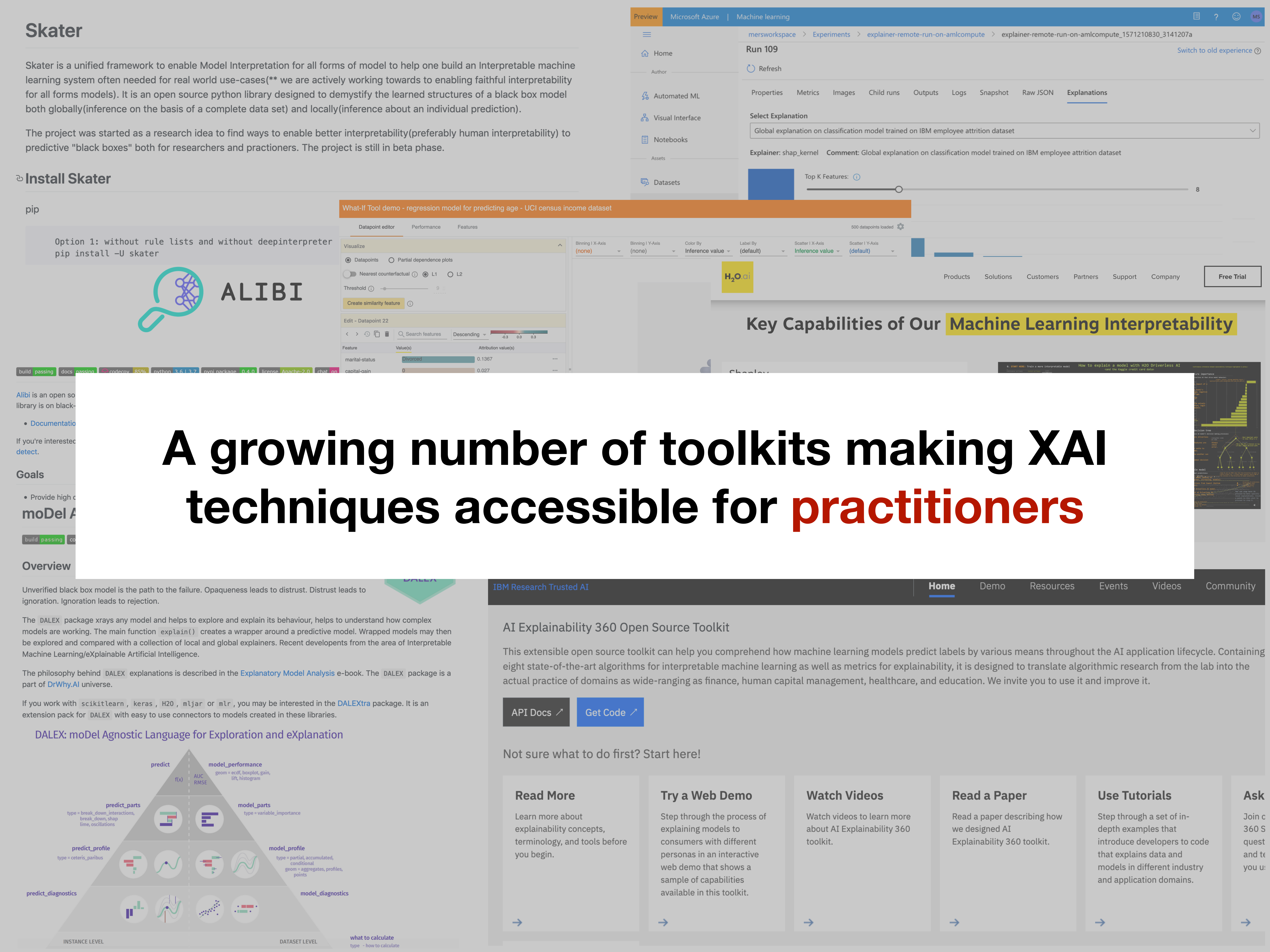
Use a *post-hoc* XAI technique



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

A growing collection of XAI techniques

A growing number of toolkits making XAI techniques accessible for practitioners



AI Explainability 360 Open Source Toolkit

This extensible open source toolkit can help you comprehend how machine learning models predict labels by various means throughout the AI application lifecycle. Containing eight state-of-the-art algorithms for interpretable machine learning as well as metrics for explainability, it is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, human capital management, healthcare, and education. We invite you to use it and improve it.

API Docs ↗ Get Code ↗

Not sure what to do first? Start here!

Read More

Learn more about explainability concepts, terminology, and tools before you begin.

→

Try a Web Demo

Step through the process of explaining models to consumers with different personas in an interactive web demo that shows a sample of capabilities available in this toolkit.

→

Watch Videos

Watch videos to learn more about AI Explainability 360 toolkit.

→

Read a Paper

Read a paper describing how we designed AI Explainability 360 toolkit.

→

Use Tutorials

Step through a set of in-depth examples that introduce developers to code that explains data and models in different industry and application domains.

→

Ask a Question

Join our AI Explainability 360 Slack Channel to ask questions, make comments, and tell stories about how you use the toolkit.

→

View Notebooks

Open a directory of Jupyter notebooks in GitHub that provide working examples of explainability in sample datasets. Then share your own notebooks!

→

Contribute

You can add new algorithms and metrics in GitHub. Share Jupyter notebooks showcasing how you have enabled explanations in your machine learning application.

→

Learn how to put this toolkit to work for your application or industry problem. Try these tutorials.

Credit Approval

See how to explain credit approval models using the

Medical Expenditure

See how to create

Dermoscopy

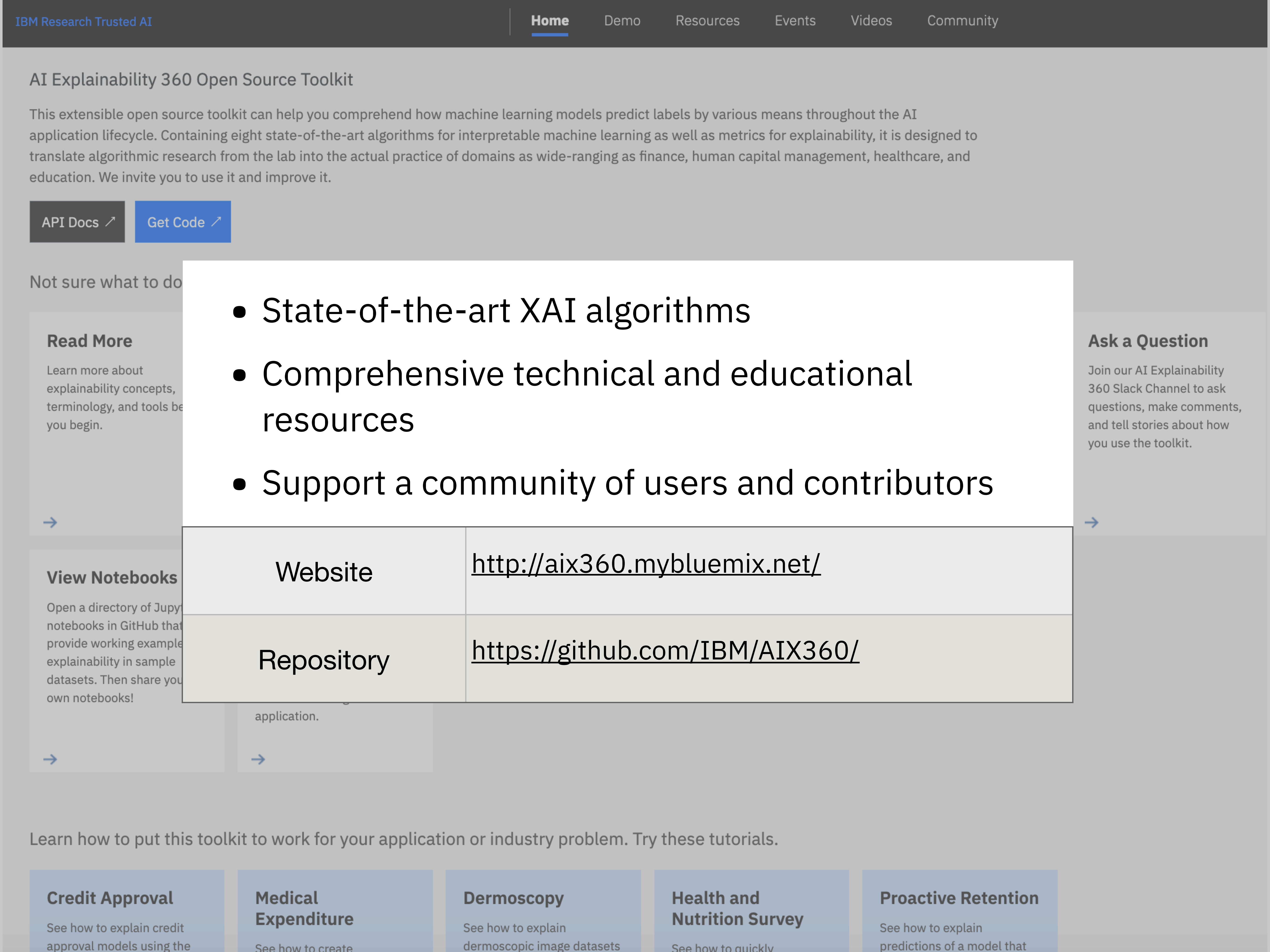
See how to explain dermoscopic image datasets

Health and Nutrition Survey

See how to quickly

Proactive Retention

See how to explain predictions of a model that



- State-of-the-art XAI algorithms
- Comprehensive technical and educational resources
- Support a community of users and contributors

Website	http://aix360.mybluemix.net/
Repository	https://github.com/IBM/AIX360/

XAI is hard: it has to be user-centered

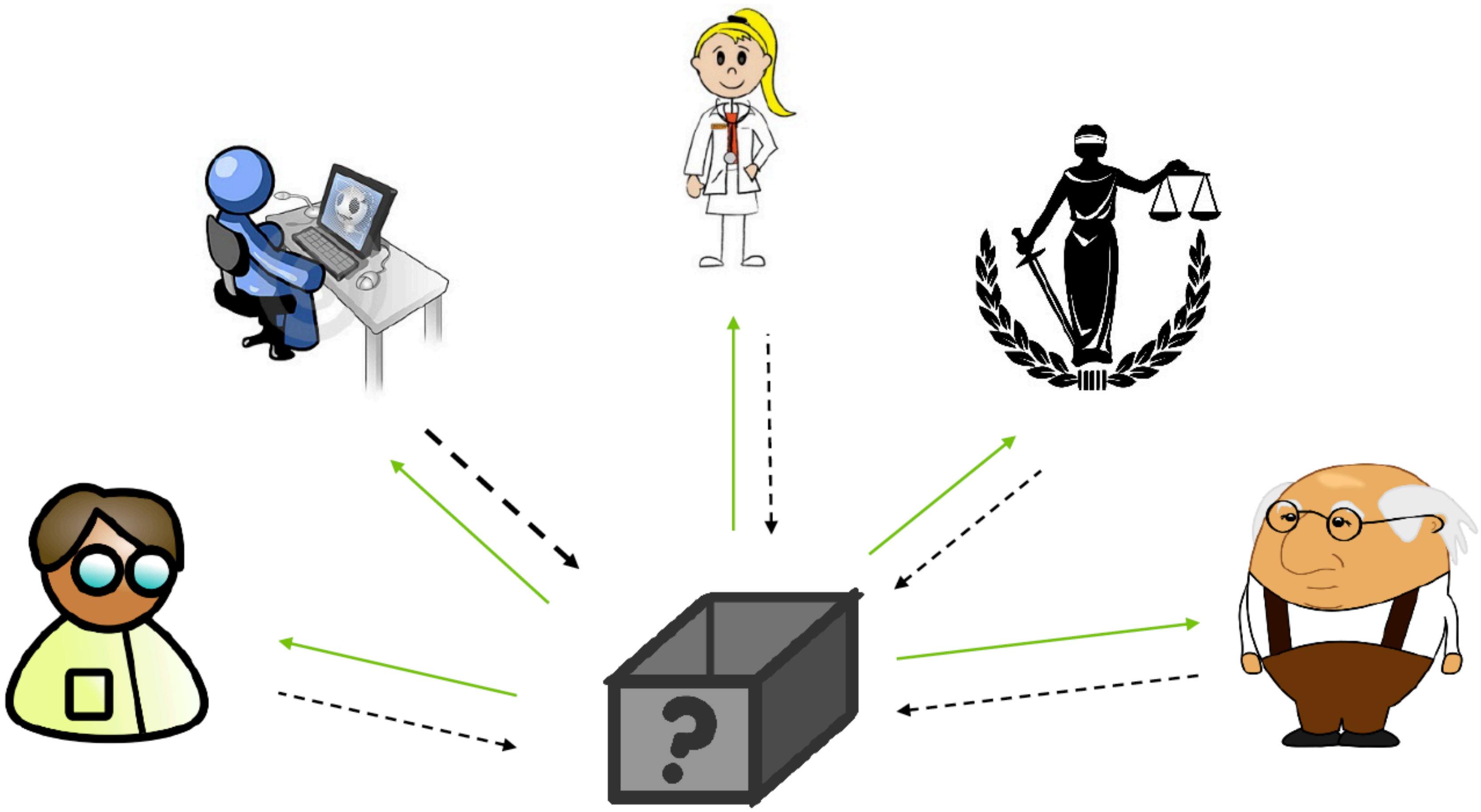
The General Data Protection Regulation (GDPR)

- Limits to decision-making based solely on automated processing and profiling (Art.22)
- Right to be provided with meaningful information about the logic involved in the decision (Art.13 (2) f. and 15 (1) h)

"meaningful" ???

(Nemitz, 2018)

XAI is hard: it has to be user-centered



(Hind et al., 2019)

Which explanation technique to use?

How to design XAI user experiences?

Motivation: Research into **XAI Design Practices**

Why AI design practitioners?

- Bridging roles connecting user needs and XAI techniques

→ Develop design methods to support creating HCXAI

- Understanding real-world user needs for XAI

→ Inform future directions of XAI

Liao et al. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. CHI 2020

A technical space people are not quite in there yet… how to talk about it?

Study probe: algorithm informed **XAI Questions**

Category of Methods	Explanation Method	Definition	Algorithm Examples	Question Type
Explain the model	Global feature importance	Describe the weights of features used by the model (including visualization that shows the weights of features)	[41, 60, 69, 90]	**How**
(**Global**)	Decision tree approximation	Approximate the model to an interpretable decision-tree	[11, 47, 52]	**How**, Why, Why not, What if
	Rule extraction	Approximate the model to a set of rules, e.g., if-then rules	[26, 93, 102]	**How**, Why, Why not, What if
Explain a prediction	Local feature importance and saliency method	Show how features of the instance contribute to the model's prediction (including causes in parts of an image or text)	[61, 74, 83, 85, 101]	**Why**
(**Local**)	Local rules or trees	Describe the rules or a decision-tree path that the instance fits to guarantee the prediction	[39, 75, 99]	**Why**, **How to still be this**
Inspect counterfactual	Feature influence or relevance method	Show how the prediction changes corresponding to changes of a feature (often in a visualization format)	[8, 33, 36, 51]	**What if**, How to be that, How to still be this
	Contrastive or counterfactual features	Describe the feature(s) that will change the prediction if perturbed, absent or present	[27, 91, 100]	**Why**, **Why not**, **How to be that**
Example based	Prototypical or representative examples	Provide example(s) similar to the instance and with the same record as the prediction	[13, 48, 50]	**Why**, How to still be this
	Counterfactual example	Provide example(s) with small differences from the instance but with a different record from the prediction	[37, 55, 66]	**Why**, **Why not**, How to be that

- User needs for XAI are represented as **prototypical questions**
- A **question** can be answered by one or multiple **XAI methods**
- An **XAI method** can be implemented by one or multiple **XAI algorithms**

“ *An explanation is an answer to a question* (Wellman, 2011; Miller 2018)
The effectiveness of an explanation depends on the question asked (Bromberger, 1992)

Question: **Why** is this husky classified as wolf?

XAI method: **local feature** (pixels) **contribution**

XAI algorithms:

- LIME (Ribeiro et al. 2016)
- SHAP (Lundberg and Lee 2017)
- …

Study probe: algorithm informed **XAI Questions**

Category of Methods	Explanation Method	Definition	Algorithm Examples	Question Type
Explain the model (**Global**)	Global feature importance	Describe the weights of features used by the model (including visualization that shows the weights of features)	[41, 60, 69, 90]	**How**
	Decision tree approximation	Approximate the model to an interpretable decision-tree	[11, 47, 52]	**How**, Why, Why not, What if
	Rule extraction	Approximate the model to a set of rules, e.g., if-then rules	[26, 93, 102]	**How**, Why, Why not, What if
Explain a prediction (**Local**)	Local feature importance and saliency method	Show how features of the instance contribute to the model's prediction (including causes in parts of an image or text)	[61, 74, 83, 85, 101]	**Why**
	Local rules or trees	Describe the rules or a decision-tree path that the instance fits to guarantee the prediction	[39, 75, 99]	**Why**, **How to still be this**
Inspect counterfactual	Feature influence or relevance method	Show how the prediction changes corresponding to changes of a feature (often in a visualization format)	[8, 33, 36, 51]	**What if**, How to be that, How to still be this
	Contrastive or counterfactual features	Describe the feature(s) that will change the prediction if perturbed, absent or present	[27, 91, 100]	**Why**, **Why not**, **How to be that**
Example based	Prototypical or representative examples	Provide example(s) similar to the instance and with the same record as the prediction	[13, 48, 50]	**Why**, How to still be this
	Counterfactual example	Provide example(s) with small differences from the instance but with a different record from the prediction	[37, 55, 66]	**Why**, **Why not**, How to be that

+

Input (data)**, output, performance**

(Lim et al., 2009)

Methodology

- Interviewed **20 participants**
- **16 AI products** in IBM

1. Walk through the AI system
2. Common questions users might ask
3. Discuss each question card
4. General challenges to create XAI products

Understanding input (training data): **What kind of data does the system learn from?**
- What is the *source* of the data?
- How are the *labels/ground-truth* produced?

Inspecting what if changing a case/counterfactual questions: what if, how to be that, how to still be this
- What would the system predict if the case changes to...?
- How should this case change to get a different prediction?
- What are the scope of changes permitted for this case to still get the same prediction?
- What kind of cases get a different/same prediction?

Understanding the model globally: How does the system make predictions (overall logic)?
- What algorithm is used?
- What *rules* does the system use to make predictions?
- *What features* does the model consider or not consider?
- How does the model *weigh/reason with these features*?

Understanding output: **What kind of output/predictions does the system give?**
- What does the system output *mean*?
- How can I use the output of the system?

Other category (add your own question)

Understanding prediction for a particular case: Why this? Why not that?
- Why is this case given this prediction? Why is it NOT predicted that?
- What *feature(s)* of this case lead to the model's prediction for it?
- *What kind of cases* are predicted this?
- Why are [cases A and B] given *the same prediction*?
- Why are [cases A and B] given *different predictions*?

Understanding model performance and certainty: How accurate/reliable are the system's predictions?
- *How often* does the system make mistakes?
- *When/under what situation* is the system likely to be correct/wrong?

Methodology

- Interviewed **20 participants**
- **16 AI products** in IBM

1. Walk through the AI system
2. Common questions users might ask
3. Discuss each question card
4. General challenges to create XAI products

Understanding input (training data): What kind of data does the system learn from?
- What is the *source* of the data?
- How are the *labels/ground-truth* produced?

Inspecting what if changing a case/counterfactual questions: what if, how to be that, how to still be this
- What would the system predict if the case changes to...?
- How should this case change to get a different prediction?
- What are the scope of changes permitted for this case to still get the same prediction?
- What kind of cases get a different/same prediction?

Understanding the model globally: How does the system make predictions (overall logic)?
- What algorithm is used?
- What *rules* does the system use to make predictions?
- *What features* does the model consider or not consider?
- How does the model *weigh/reason with these features*?

Understanding output: What kind of output/predictions does the system give?
- What does the system output *mean*?
- How can I use the output of the system?

Other category (add your own question)

Understanding prediction for a particular case: Why this? Why not that?
- Why is this case given this prediction? Why is it NOT predicted that?
- What *feature(s)* of this case lead to the model's prediction for it?
- *What kind of cases* are predicted this?
- Why are [cases A and B] given *the same prediction*?
- Why are [cases A and B] given *different predictions*?

Understanding model performance and certainty: How accurate/reliable are the system's predictions?
- *How often* does the system make mistakes?
- *When/under what situation* is the system likely to be correct/wrong?

XAI question bank

Data

- **What kind of data does the system learn from?**
- What is the source of the data?
- How were the labels/ground-truth produced?
- * What is the sample size?
- * What data is the system NOT using?
- * What are the limitations/biases of the data?
- * How much data [like this] is the system trained on?

Output

- **What kind of output does the system give?**
- What does the system output mean?
- How can I best utilize the output of the system ?
- * What is the scope of the system's capability? Can it do…?
- * How is the output used for other system component(s) ?

Performance

- **How accurate/precise/reliable are the predictions?**
- How often does the system make mistakes?
- In what situations is the system likely to be correct/incorrect?
- * What are the limitations of the system?
- * What kind of mistakes is the system likely to make?
- * Is the system's performance good enough for…

How (global)

- **How does the system make predictions?**
- What features does the system consider?
  - * Is [feature X] used or not used for the predictions?
- What is the system's overall logic?
  - How does it weigh different features?
  - What rules does it use?
  - How does [feature X] impact its predictions?
  - * What are the top rules/features it uses?
- * What kind of algorithm is used?
  - * How are the parameters set?

Why

- **Why/how is this instance given this prediction?**
- What feature(s) of this instance leads to the system's prediction?
- Why are [instance A and B] given the same prediction?

Why not

- **Why/how is this instance NOT predicted…?**
- Why is this instance predicted P instead of Q?
- Why are [instance A and B] given different predictions?

What If

- **What would the system predict if this instance changes to…?**
- What would the system predict if this feature of the instance changes to…?
- What would the system predict for [a different instance]?

How to be that

- **How should this instance change to get a different prediction?**
- How should this feature change for this instance to get a different prediction?
- What kind of instance gets a different prediction?

How to still be this

- **What is the scope of change permitted to still get the same prediction?**
- What is the [highest/lowest/…] feature(s) one can have to still get the same prediction?
- What is the necessary feature(s) present or absent to guarantee this prediction?
- What kind of instance gets this prediction?

Others

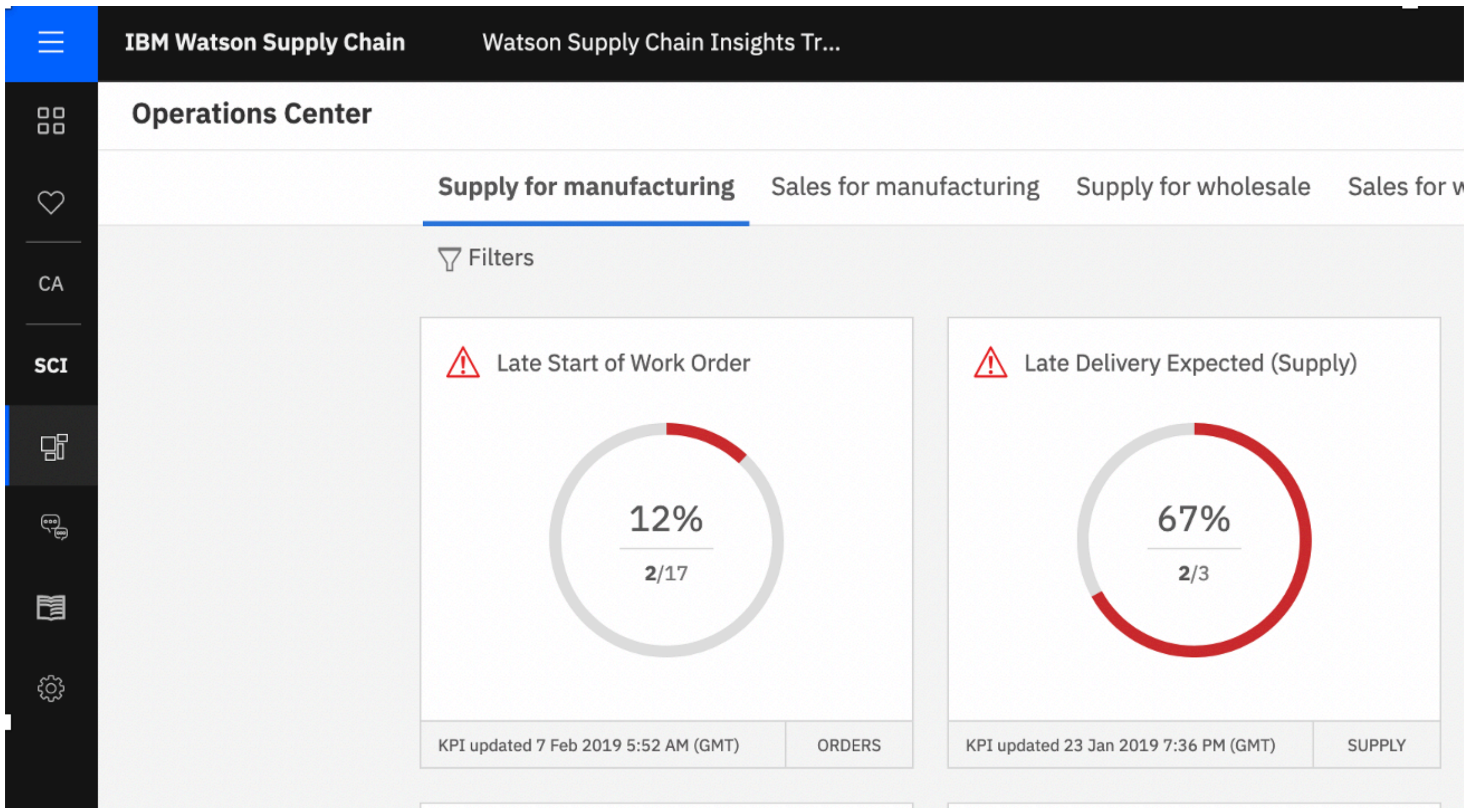
- * How/what/why will the system change/adapt/improve/drift over time? (change)
- * How to improve the system? (change)
- * Why using or not using this feature/rule/data? (follow-up)
- * What does [ML terminology] mean? (terminological)
- * What are the results of other people using the system? (social)

XAI design challenge 1: Variability of XAI needs

Diverse end goals for explainability

- To gain further insights for the decision
- To appropriately evaluate AI’s capability
- To adapt usage or interaction
- To improve AI performance
- Ethical responsibilities of AI products

To gain further insights for the decision



Why
How to be that

> *Users need to know why the system is saying this will be late because the reason is going to determine what their next action is...If it's because of a weather event, so no matter what you do you're not going to improve this number, versus something small, if you just make a quick call, you can get that number down* (I-5)

To appropriately evaluate AI's capability

Performance
How

“ *There is a calibration of trust, whether people will use it over time. But also saying hey, we know this fails in this way* (I-6)

XAI design challenge 1: Variability of XAI needs

Diverse end goals for explainability

- To gain further insights for the decision
- To appropriately evaluate AI's capability
- To adapt usage or interaction
- To improve AI performance
- Ethical responsibilities of AI products

Also varying XAI needs: **User group**, **usage point**, **algorithm and data type**, **decision context**

XAI design challenge 2: Gaps between algorithmic output and human explanations

Human explanations are

- **Selective**
- **Contrastive**
- **Interactive**
- **Tailored for recipients**

Design attempt to mimic how people, especially domain experts, explain

XAI design challenge 3: “in the dark” design process

- Challenge **navigating the technical capabilities**
- **Communication barriers** between designers, data scientists and other stakeholders
- **Cost of time and resource** impeding buy-in

> *It remains in this weird limbo where people know it's important. People see it happen. They don't know how to make it happen. And everybody's feeling their way in the dark with no lights.* (I-8)

XAI Question Bank

Data

- **What kind of data does the system learn from?**
- What is the source of the data?
- How were the labels/ground-truth produced?
- * What is the sample size?
- * What data is the system NOT using?
- * What are the limitations/biases of the data?
- * How much data [like this] is the system trained on?

Output

- **What kind of output does the system give?**
- What does the system output mean?
- How can I best utilize the output of the system ?
- * What is the scope of the system's capability? Can it do…?
- * How is the output used for other system component(s) ?

Performance

- **How accurate/precise/reliable are the predictions?**
- How often does the system make mistakes?
- In what situations is the system likely to be correct/incorrect?
- * What are the limitations of the system?
- * What kind of mistakes is the system likely to make?
- * Is the system's performance good enough for…

How (global)

- **How does the system make predictions?**
- What features does the system consider?
  - * Is [feature X] used or not used for the predictions?
- What is the system's overall logic?
  - How does it weigh different features?
  - What rules does it use?
  - How does [feature X] impact its predictions?
  - * What are the top rules/features it uses?
- * What kind of algorithm is used?
  - * How are the parameters set?

Why

- **Why/how is this instance given this prediction?**
- What feature(s) of this instance leads to the system's prediction?
- Why are [instance A and B] given the same prediction?

Why not

- **Why/how is this instance NOT predicted…?**
- Why is this instance predicted P instead of Q?
- Why are [instance A and B] given different predictions?

What If

- **What would the system predict if this instance changes to…?**
- What would the system predict if this feature of the instance changes to…?
- What would the system predict for [a different instance]?

How to be that

- **How should this instance change to get a different prediction?**
- How should this feature change for this instance to get a different prediction?
- What kind of instance gets a different prediction?

How to still be this

- **What is the scope of change permitted to still get the same prediction?**
- What is the [highest/lowest/…] feature(s) one can have to still get the same prediction?
- What is the necessary feature(s) present or absent to guarantee this prediction?
- What kind of instance gets this prediction?

Others

- * How/what/why will the system change/adapt/improve/drift over time? (change)
- * How to improve the system? (change)
- * Why using or not using this feature/rule/data? (follow-up)
- * What does [ML terminology] mean? (terminological)
- * What are the results of other people using the system? (social)

- A checklist representing the space of user needs for XAI
- Understand real-world user questions to derive design guidelines
- New questions (with *) inform gaps in XAI technical work

XAI Question Bank

Data

- **What kind of data does the system learn from?**
- What is the source of the data?
- How were the labels/ground-truth produced?
- * What is the sample size?
- * What data is the system NOT using?
- * What are the limitations/biases of the data?
- * How much data [like this] is the system trained on?

Output

- **What kind of output does the system give?**
- What does the system output mean?
- How can I best utilize the output of the system ?
- * What is the scope of the system's capability? Can it do…?
- * How is the output used for other system component(s) ?

Performance

- **How accurate/precise/reliable are the predictions?**
- How often does the system make mistakes?
- In what situations is the system likely to be correct/incorrect?
- * What are the limitations of the system?
- * What kind of mistakes is the system likely to make?
- * Is the system's performance good enough for…

How (global)

- **How does the system make predictions?**
- What features does the system consider?
  - * Is [feature X] used or not used for the predictions?
- What is the system's overall logic?
  - How does it weigh different features?
  - What rules does it use?
  - How does [feature X] impact its predictions?
  - * What are the top rules/features it uses?
- * What kind of algorithm is used?
  - * How are the parameters set?

Why

- **Why/how is this instance given this prediction?**
- What feature(s) of this instance leads to the system's prediction?
- Why are [instance A and B] given the same prediction?

Why not

- **Why/how is this instance NOT predicted…?**
- Why is this instance predicted P instead of Q?
- Why are [instance A and B] given different predictions?

What If

- **What would the system predict if this instance changes to…?**
- What would the system predict if this feature of the instance changes to…?
- What would the system predict for [a different instance]?

How to be that

- **How should this instance change to get a different prediction?**
- How should this feature change for this instance to get a different prediction?
- What kind of instance gets a different prediction?

How to still be this

- **What is the scope of change permitted to still get the same prediction?**
- What is the [highest/lowest/…] feature(s) one can have to still get the same prediction?
- What is the necessary feature(s) present or absent to guarantee this prediction?
- What kind of instance gets this prediction?

Others

- * How/what/why will the system change/adapt/improve/drift over time? (change)
- * How to improve the system? (change)
- * Why using or not using this feature/rule/data? (follow-up)
- * What does [ML terminology] mean? (terminological)
- * What are the results of other people using the system? (social)

- A checklist representing the space of user needs for XAI
- **Understand real-world user questions and how to address them**
- New questions (with *) inform gaps in XAI technical work

Understand XAI questions and desired solutions

Input: Provide comprehensive transparency of training data, especially the limitations

Output: Contextualize the system's output in downstream tasks and the users' overall workflow

Performance: Help users understand the limitation of the AI and make it actionable

Global model: Choose appropriate level of details to explain the model

Local decision: Provide resources for "why not"

Counterfactual: Consider opportunities as utility features for analytics or system exploration

XAI Question Bank

Data

- **What kind of data does the system learn from?**
- What is the source of the data?
- How were the labels/ground-truth produced?
- * What is the sample size?
- * What data is the system NOT using?
- * What are the limitations/biases of the data?
- * How much data [like this] is the system trained on?

Output

- **What kind of output does the system give?**
- What does the system output mean?
- How can I best utilize the output of the system ?
- * What is the scope of the system's capability? Can it do…?
- * How is the output used for other system component(s) ?

Performance

- **How accurate/precise/reliable are the predictions?**
- How often does the system make mistakes?
- In what situations is the system likely to be correct/incorrect?
- * What are the limitations of the system?
- * What kind of mistakes is the system likely to make?
- * Is the system's performance good enough for…

How (global)

- **How does the system make predictions?**
- What features does the system consider?
  - * Is [feature X] used or not used for the predictions?
- What is the system's overall logic?
  - How does it weigh different features?
  - What rules does it use?
  - How does [feature X] impact its predictions?
  - * What are the top rules/features it uses?
- * What kind of algorithm is used?
  - * How are the parameters set?

Why

- **Why/how is this instance given this prediction?**
- What feature(s) of this instance leads to the system's prediction?
- Why are [instance A and B] given the same prediction?

Why not

- **Why/how is this instance NOT predicted…?**
- Why is this instance predicted P instead of Q?
- Why are [instance A and B] given different predictions?

What If

- **What would the system predict if this instance changes to…?**
- What would the system predict if this feature of the instance changes to…?
- What would the system predict for [a different instance]?

How to be that

- **How should this instance change to get a different prediction?**
- How should this feature change for this instance to get a different prediction?
- What kind of instance gets a different prediction?

How to still be this

- **What is the scope of change permitted to still get the same prediction?**
- What is the [highest/lowest/…] feature(s) one can have to still get the same prediction?
- What is the necessary feature(s) present or absent to guarantee this prediction?
- What kind of instance gets this prediction?

Others

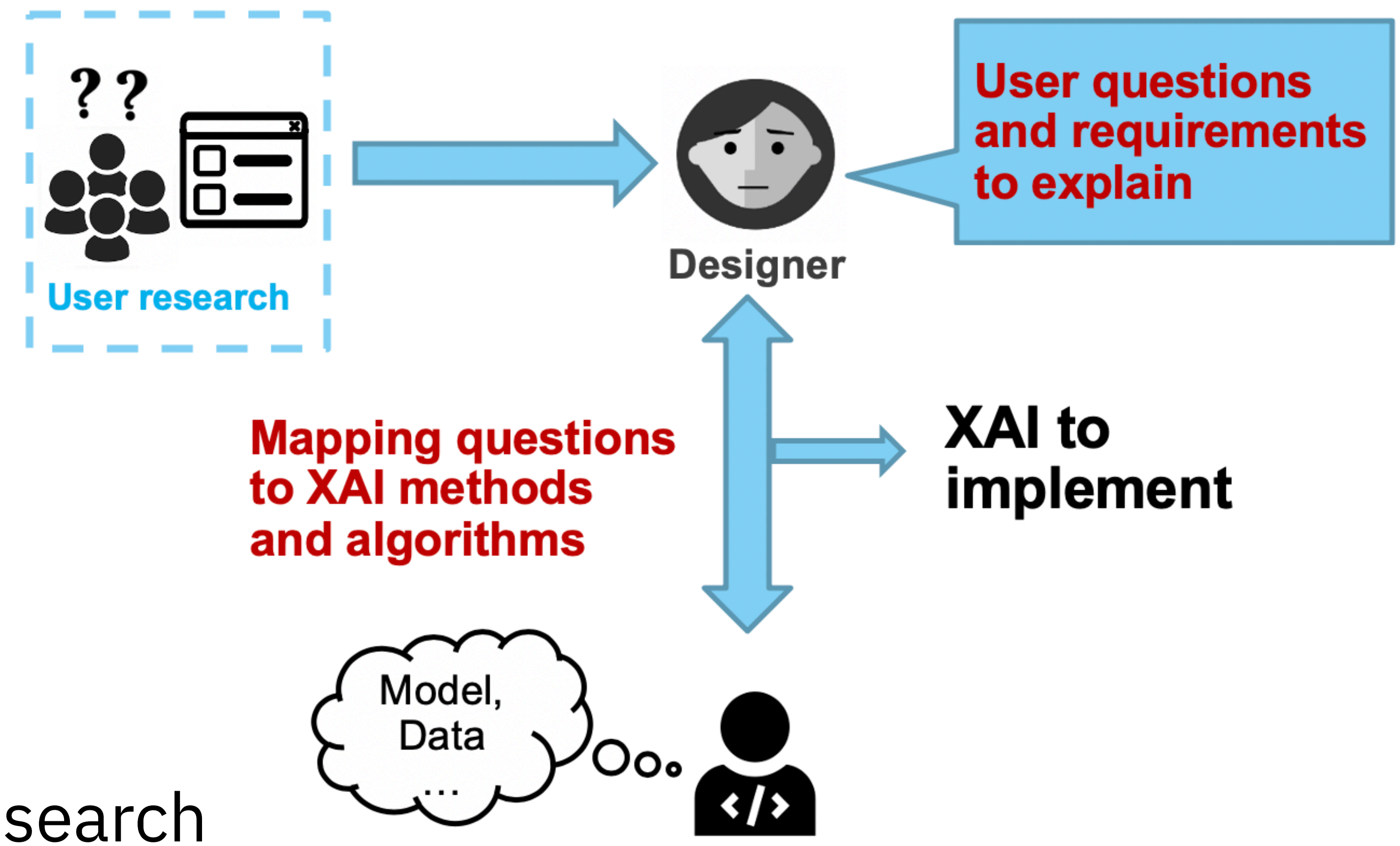
- * How/what/why will the system change/adapt/improve/drift over time? (change)
- * How to improve the system? (change)
- * Why using or not using this feature/rule/data? (follow-up)
- * What does [ML terminology] mean? (terminological)
- * What are the results of other people using the system? (social)

- A checklist representing the space of user needs for XAI
- Understand real-world user questions and how to address them
- **New questions (with *) inform gaps in XAI technical work**

Opportunities for future technical XAI work

- Explain data bias and generalizability
- Explain output of multiple models
- Explain system changes
- Multi-level global explanations
- Interactive counterfactual explanations
- Social explanations
- Personalized and adaptive explanations

Supporting the process: **question-driven XAI design**

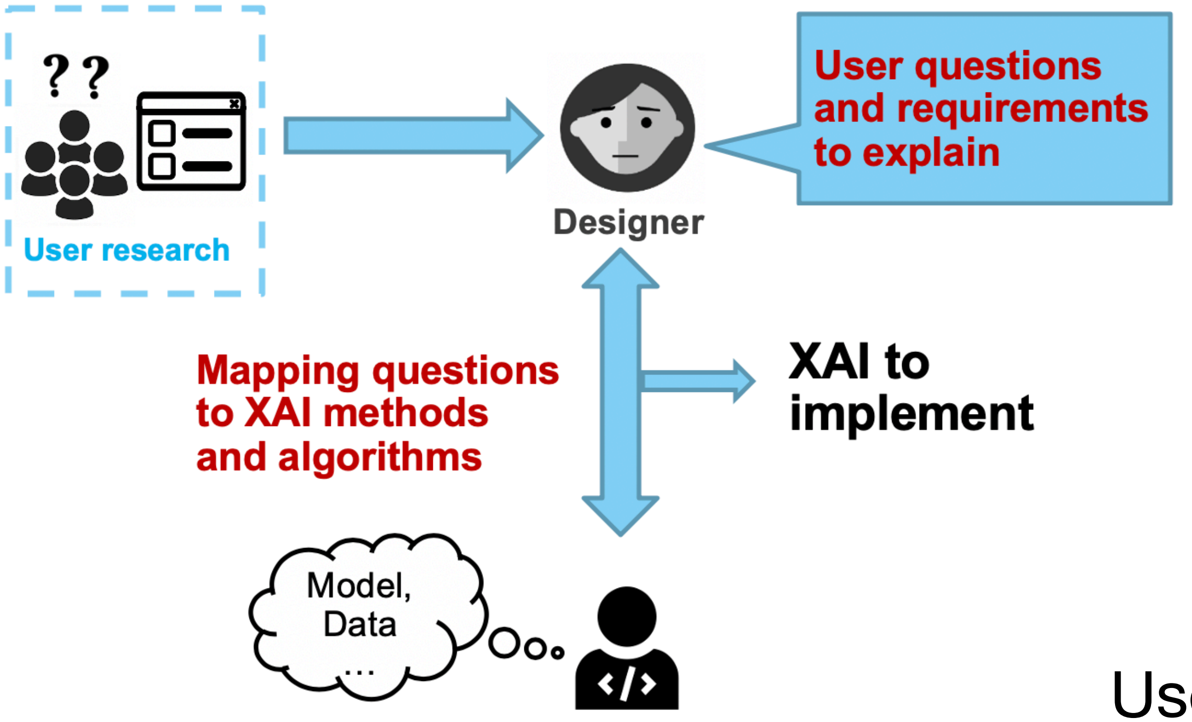


Through user research

- **Questions** elicitation
- Identify user **requirements** to address the *questions*

Working with data scientists and the team

- Map the *questions* to **XAI technique(s)**
- **Iteratively** evaluate by the user requirements and fill the gaps



XAI Question Bank

Data
- **What kind of data does the system learn from?**
- What is the source of the data?
- How were the labels/ground-truth produced?
- * What is the sample size?
- * What data is the system NOT using?
- * What are the limitations/biases of the data?
- * How much data [like this] is the system trained on?

Output
- **What kind of output does the system give?**
- What does the system output mean?
- How can I best utilize the output of the system ?
- * What is the scope of the system's capability? Can it do…?
- * How is the output used for other system component(s) ?

Performance
- **How accurate/precise/reliable are the predictions?**
- How often does the system make mistakes?
- In what situations is the system likely to be correct/incorrect?
- * What are the limitations of the system?
- * What kind of mistakes is the system likely to make?
- * Is the system's performance good enough for…

How (global)
- **How does the system make predictions?**
- What features does the system consider?
  - * Is [feature X] used or not used for the predictions?
- What is the system's overall logic?
  - How does it weigh different features?
  - What rules does it use?
  - How does [feature X] impact its predictions?
  - * What are the top rules/features it uses?
- * What kind of algorithm is used?
  - * How are the parameters set?

Why
- **Why/how is this instance given this prediction?**
- What feature(s) of this instance leads to the system's prediction?
- Why are [instance A and B] given the same prediction?

Why not
- **Why/how is this instance NOT predicted…?**
- Why is this instance predicted P instead of Q?
- Why are [instance A and B] given different predictions?

What If
- **What would the system predict if this instance changes to…?**
- What would the system predict if this feature of the instance changes to…?
- What would the system predict for [a different instance]?

How to be that
- **How should this instance change to get a different prediction?**
- How should this feature change for this instance to get a different prediction?
- What kind of instance gets a different prediction?

How to still be this
- **What is the scope of change permitted to still get the same prediction?**
- What is the [highest/lowest/…] feature(s) one can have to still get the same prediction?
- What is the necessary feature(s) present or absent to guarantee this prediction?
- What kind of instance gets this prediction?

Others
- * How/what/why will the system change/adapt/improve/drift over time? (change)
- * How to improve the system? (change)
- * Why using or not using this feature/rule/data? (follow-up)
- * What does [ML terminology] mean? (terminological)
- * What are the results of other people using the system? (social)

Use XAI question bank to guide question elicitation

Category of Methods	Explanation Method	Definition	Algorithm Examples	Question Type
Explain the model (**Global**)	Global feature importance	Describe the weights of features used by the model (including visualization that shows the weights of features)	[41, 60, 69, 90]	**How**
	Decision tree approximation	Approximate the model to an interpretable decision-tree	[11, 47, 52]	**How**, Why, Why not, What if
	Rule extraction	Approximate the model to a set of rules, e.g., if-then rules	[26, 93, 102]	**How**, Why, Why not, What if
Explain a prediction (**Local**)	Local feature importance and saliency method	Show how features of the instance contribute to the model's prediction (including causes in parts of an image or text)	[61, 74, 83, 85, 101]	**Why**
	Local rules or trees	Describe the rules or a decision-tree path that the instance fits to guarantee the prediction	[39, 75, 99]	**Why**, **How to still be this**
Inspect counterfactual	Feature influence or relevance method	Show how the prediction changes corresponding to changes of a feature (often in a visualization format)	[8, 33, 36, 51]	**What if**, How to be that, How to still be this
	Contrastive or counterfactual features	Describe the feature(s) that will change the prediction if perturbed, absent or present	[27, 91, 100]	**Why**, **Why not**, **How to be that**
Example based	Prototypical or representative examples	Provide example(s) similar to the instance and with the same record as the prediction	[13, 48, 50]	**Why**, How to still be this
	Counterfactual example	Provide example(s) with small differences from the instance but with a different record from the prediction	[37, 55, 66]	**Why**, **Why not**, How to be that

A guide to mapping questions to XAI techniques for supervised ML

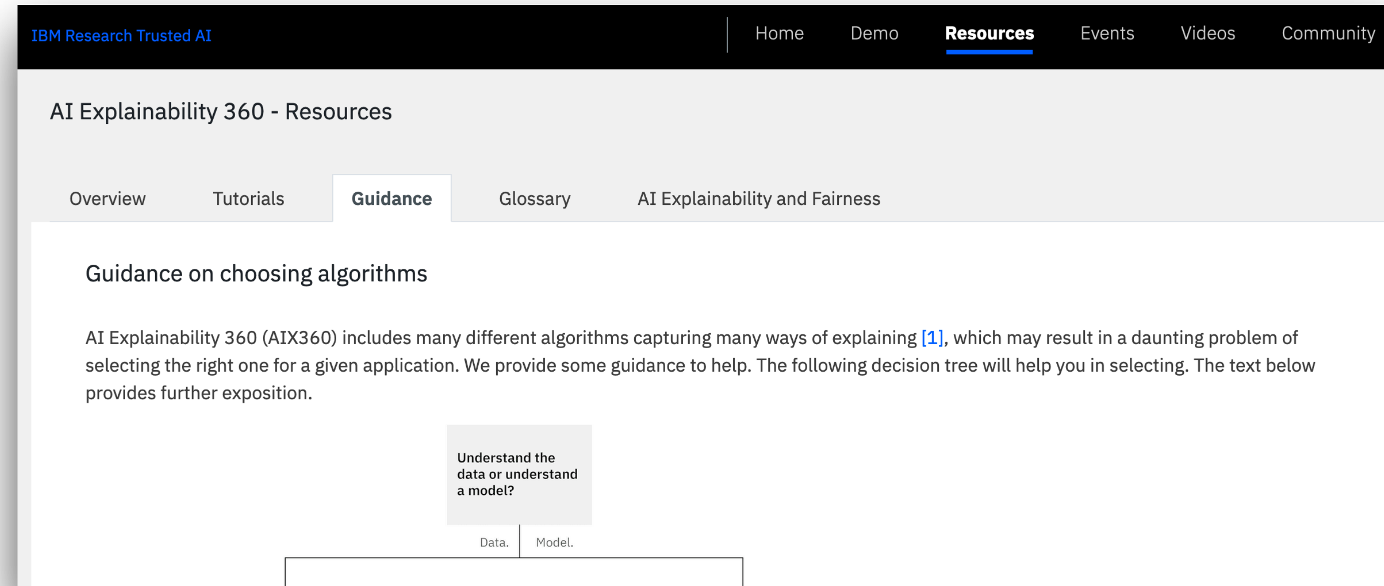
Summary, and putting research into practices…

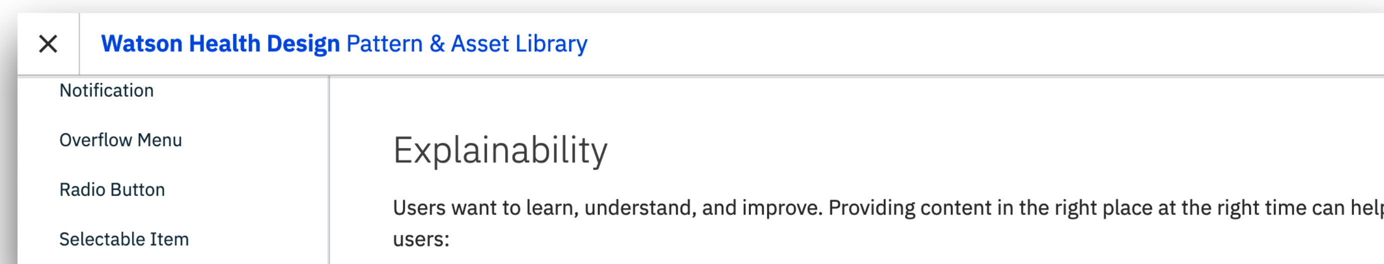
Real-world user needs for **nine categories of AI explainability**

- Guidelines to address them
- Opportunities for future algorithmic work

Challenges faced by design practitioners

- **XAI Question Bank**
- **Question-driven XAI design process**





Liao et al. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. CHI 2020

Research through design and case studies

Use case	Fair ML	AI-assisted decision	Active learning	Auto AI/ML
User	Regulator, impacted group	Decision-maker	Annotator	Model builder
Key RQ	How do different styles of explanation impact **fairness judgment**?	Can local explanation improve **decision outcomes**?	Can local explanation improve **model training** and **annotator experience**?	Can interactive explanation support **model selection**?

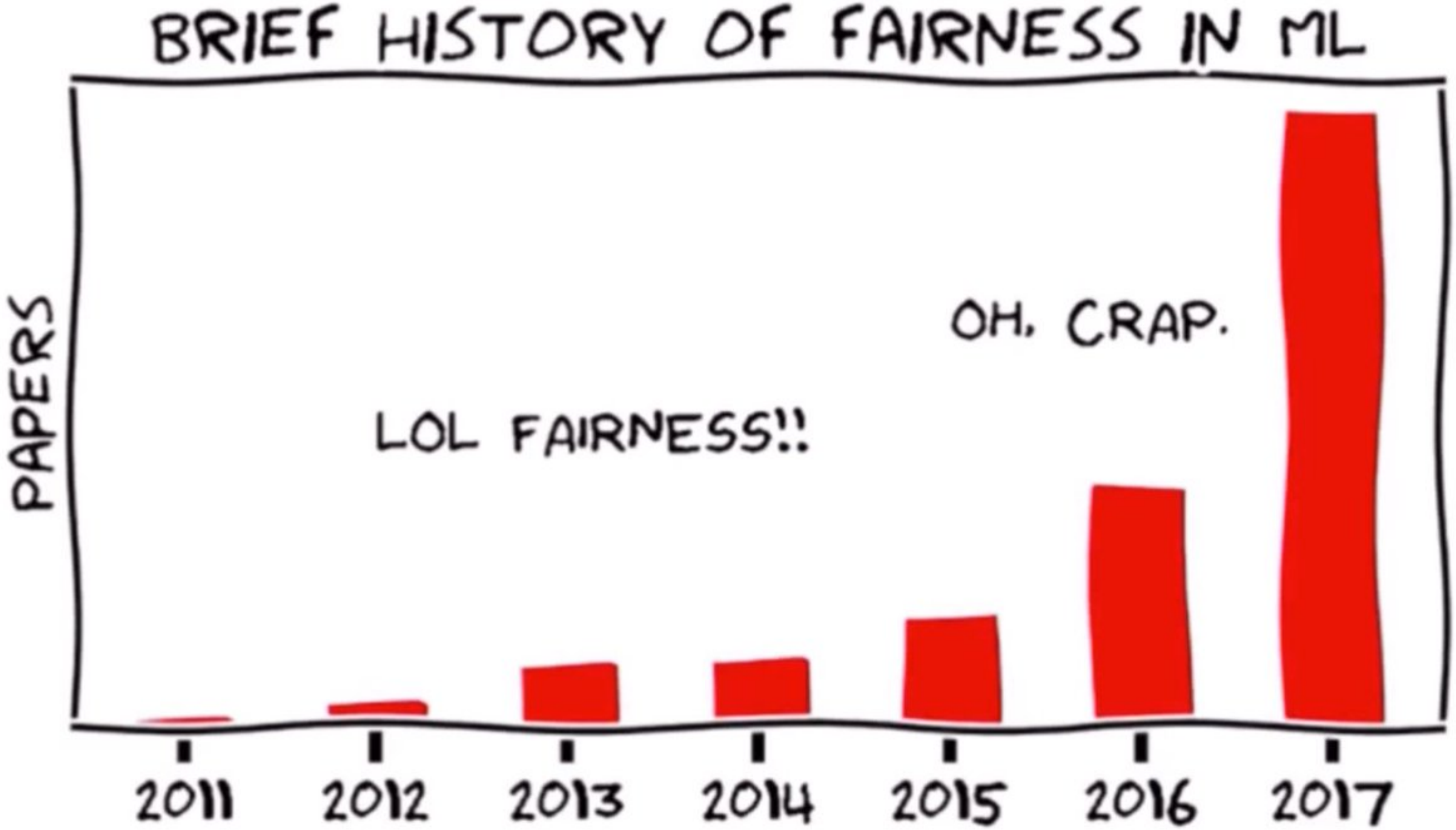
Dodge et al. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. IUI 2019

Fair ML: What is unwanted bias?

Discrimination becomes objectionable when it places certain **unprivileged** groups at a systematic disadvantage

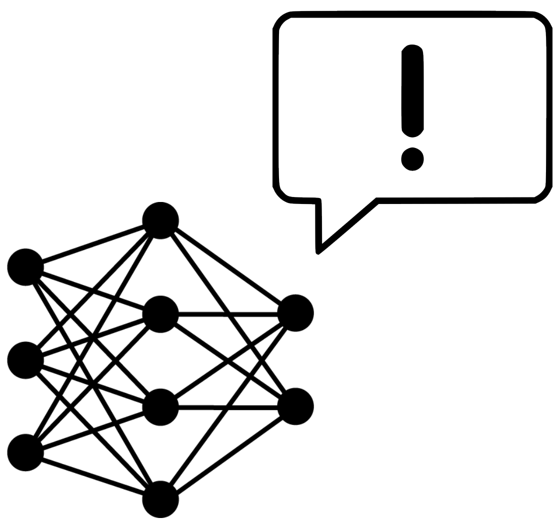
Illegal in certain contexts

(Barocas and Selbst, 2017)

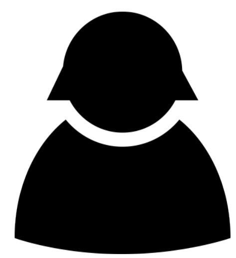

BRIEF HISTORY OF FAIRNESS IN ML
PAPERS
OH. CRAP.
LOL FAIRNESS!!
2011
2012
2013
2014
2015
2016
2017

(Hardt, 2017)

XAI as interfaces for scrutinizing model biases



Developer



Regulator

Impacted group

End goal of XAI:
Fairness **calibration**

If the model is biased

→ Help identify the bias

If the model is unbiased

→ Help foster trust and confidence

Prototype and use case: explaining COMPAS

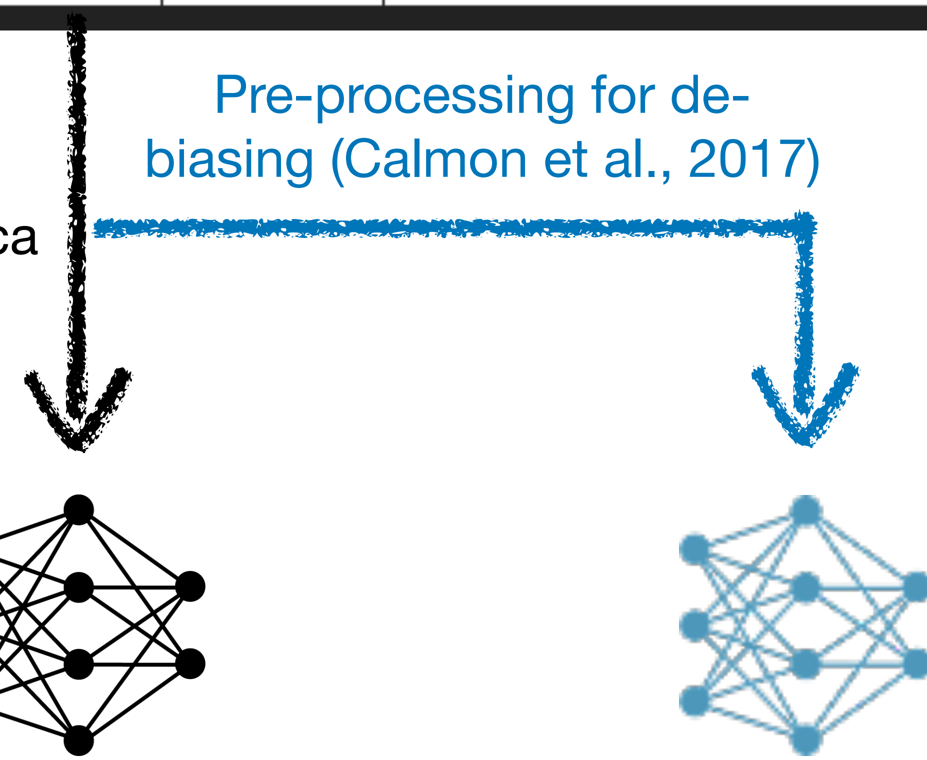
COMPAS is a software used to assess the recidivism risk of a defendant who posts a bail. Widely criticized as biased.

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

(Larson et al. ProPublica, 2016)

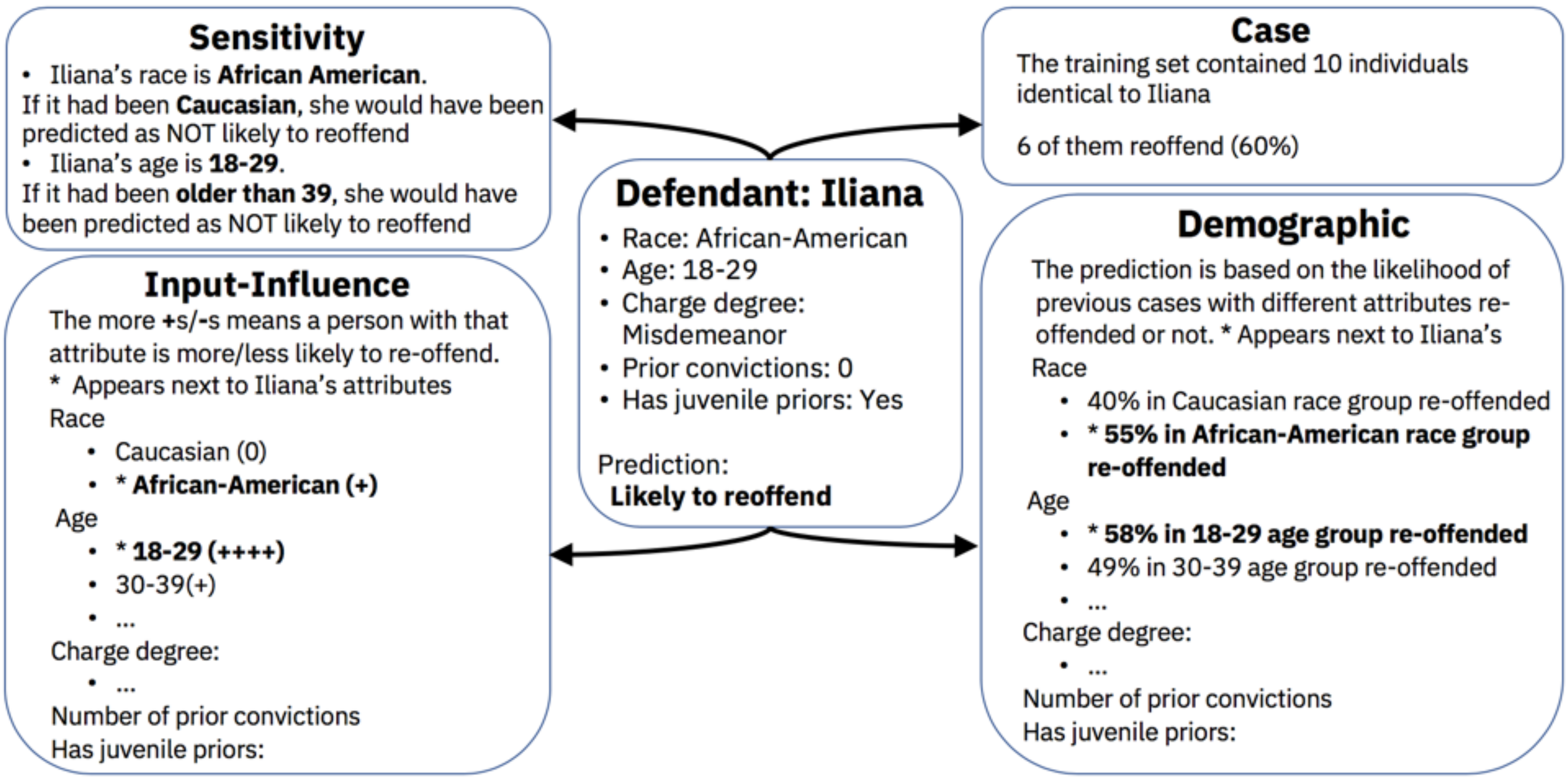
Pre-processing for de-biasing (Calmon et al., 2017)

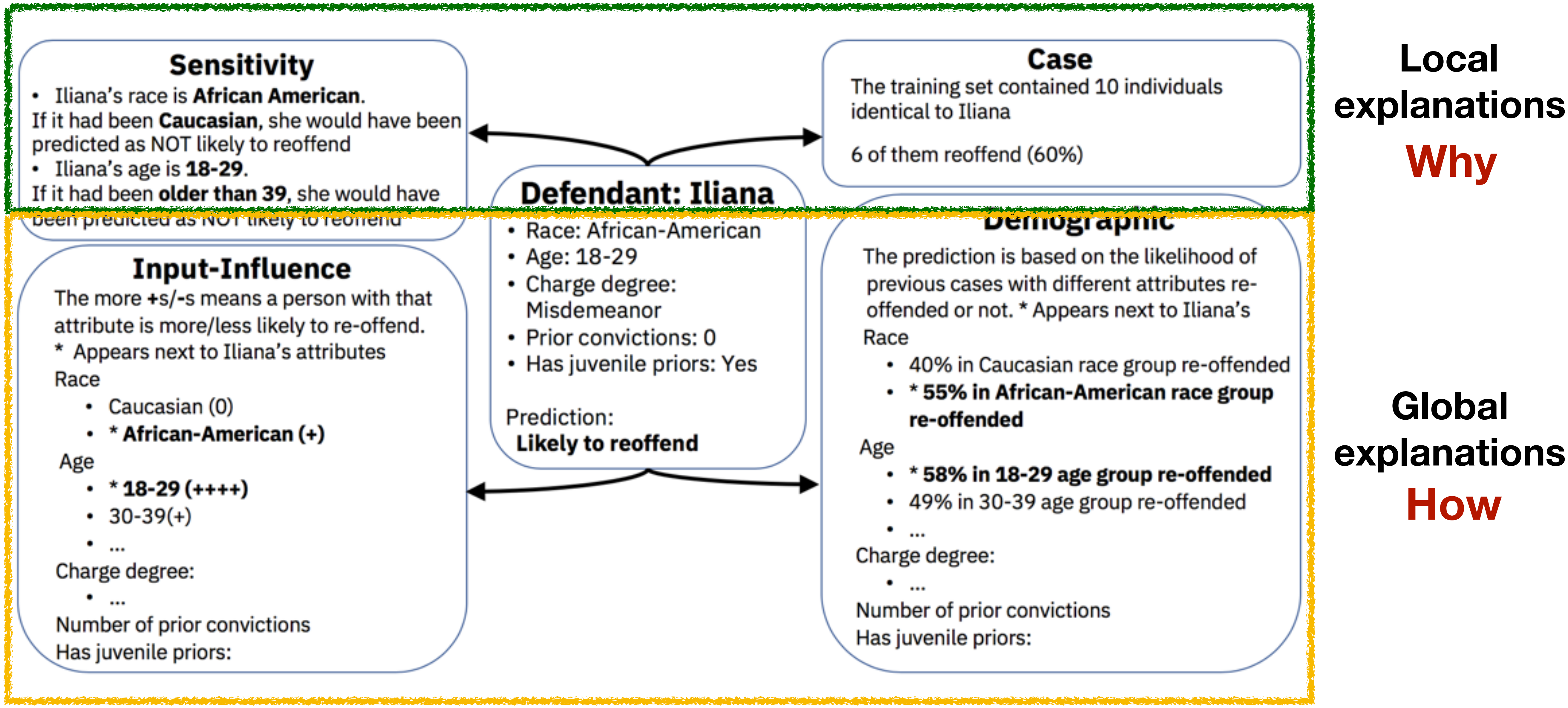
Trained with ProPublica data 10K records



Statistically unfair model

Statistically fair model


Sensitivity
• Iliana's race is African American.
If it had been Caucasian, she would have been predicted as NOT likely to reoffend
• Iliana's age is 18-29.
If it had been older than 39, she would have been predicted as NOT likely to reoffend
Input-Influence
The more +s/-s means a person with that attribute is more/less likely to re-offend.
* Appears next to Iliana's attributes
Race
• Caucasian (0)
• * African-American (+)
Age
• * 18-29 (++++)
• 30-39(+)
• ...
Charge degree:
• ...
Number of prior convictions
Has juvenile priors:
Defendant: Iliana
• Race: African-American
• Age: 18-29
• Charge degree: Misdemeanor
• Prior convictions: 0
• Has juvenile priors: Yes
Prediction:
Likely to reoffend
Case
The training set contained 10 individuals identical to Iliana
6 of them reoffend (60%)
Demographic
The prediction is based on the likelihood of previous cases with different attributes re-offended or not. * Appears next to Iliana's
Race
• 40% in Caucasian race group re-offended
• * 55% in African-American race group re-offended
Age
• * 58% in 18-29 age group re-offended
• 49% in 30-39 age group re-offended
• ...
Charge degree:
• ...
Number of prior convictions
Has juvenile priors:


Sensitivity
• Iliana's race is African American.
If it had been Caucasian, she would have been predicted as NOT likely to reoffend
• Iliana's age is 18-29.
If it had been older than 39, she would have been predicted as NOT likely to reoffend
Case
The training set contained 10 individuals identical to Iliana
6 of them reoffend (60%)
Defendant: Iliana
• Race: African-American
• Age: 18-29
• Charge degree: Misdemeanor
• Prior convictions: 0
• Has juvenile priors: Yes
Prediction:
Likely to reoffend
Input-Influence
The more +s/-s means a person with that attribute is more/less likely to re-offend.
* Appears next to Iliana's attributes
Race
• Caucasian (0)
• * African-American (+)
Age
• * 18-29 (++++)
• 30-39(+)
• ...
Charge degree:
• ...
Number of prior convictions
Has juvenile priors:
Demographic
The prediction is based on the likelihood of previous cases with different attributes re-offended or not. * Appears next to Iliana's
Race
• 40% in Caucasian race group re-offended
• * 55% in African-American race group re-offended
Age
• * 58% in 18-29 age group re-offended
• 49% in 30-39 age group re-offended
• ...
Charge degree:
• ...
Number of prior convictions
Has juvenile priors:
Local explanations
Why
Global explanations
How

Research questions

Fairness **calibration**

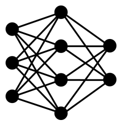
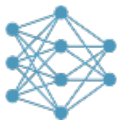
How do different styles of explanation impact fairness judgment?

- **Fairness calibration**?
- Surfacing **individual fairness** issue—similar individuals receiving different treatment?
- Perceived **inherently less fair**?

How do individual factors mediate the impact?

Experimental design

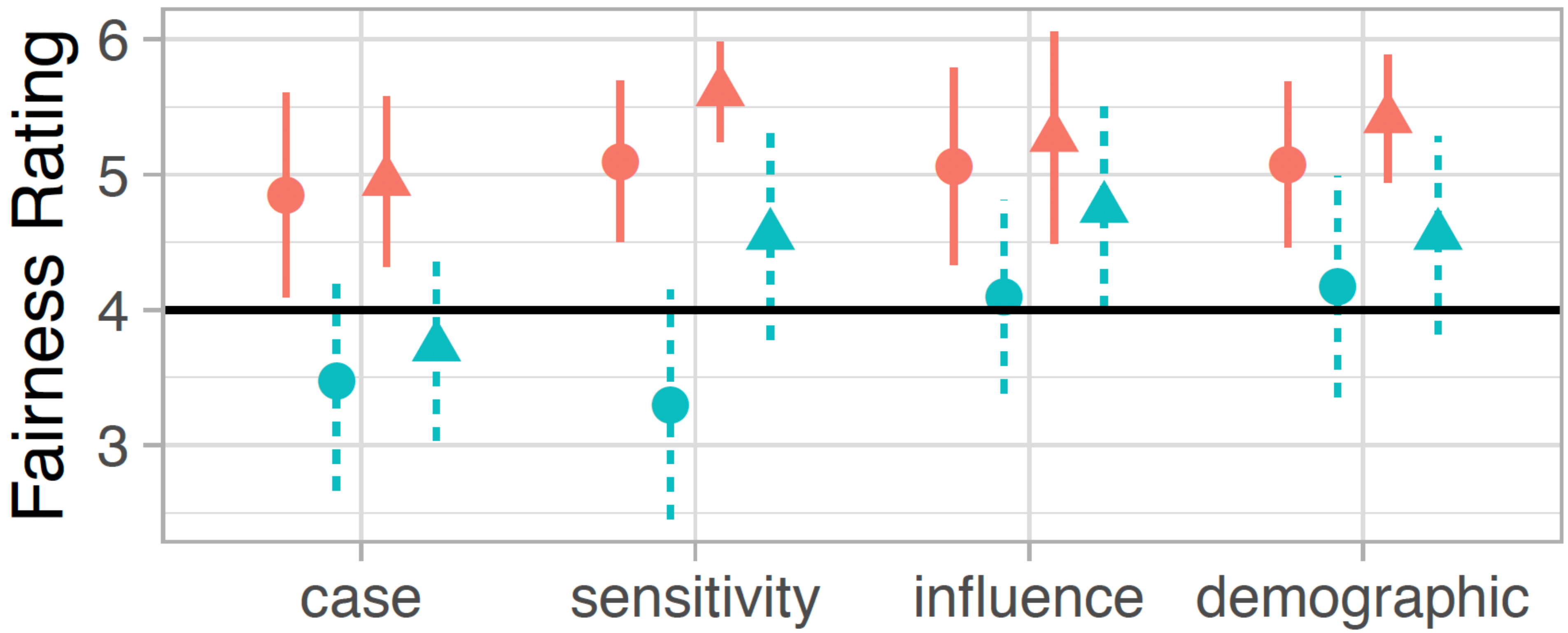
160 MTurk participants

	Input influence	Data demographic	Sensitivity	Case based
	20	20	20	20
	20	20	20	20

Sampled 6 instances from test data, oversampled 1/3 disparately impacted individuals (*individual unfairness*)

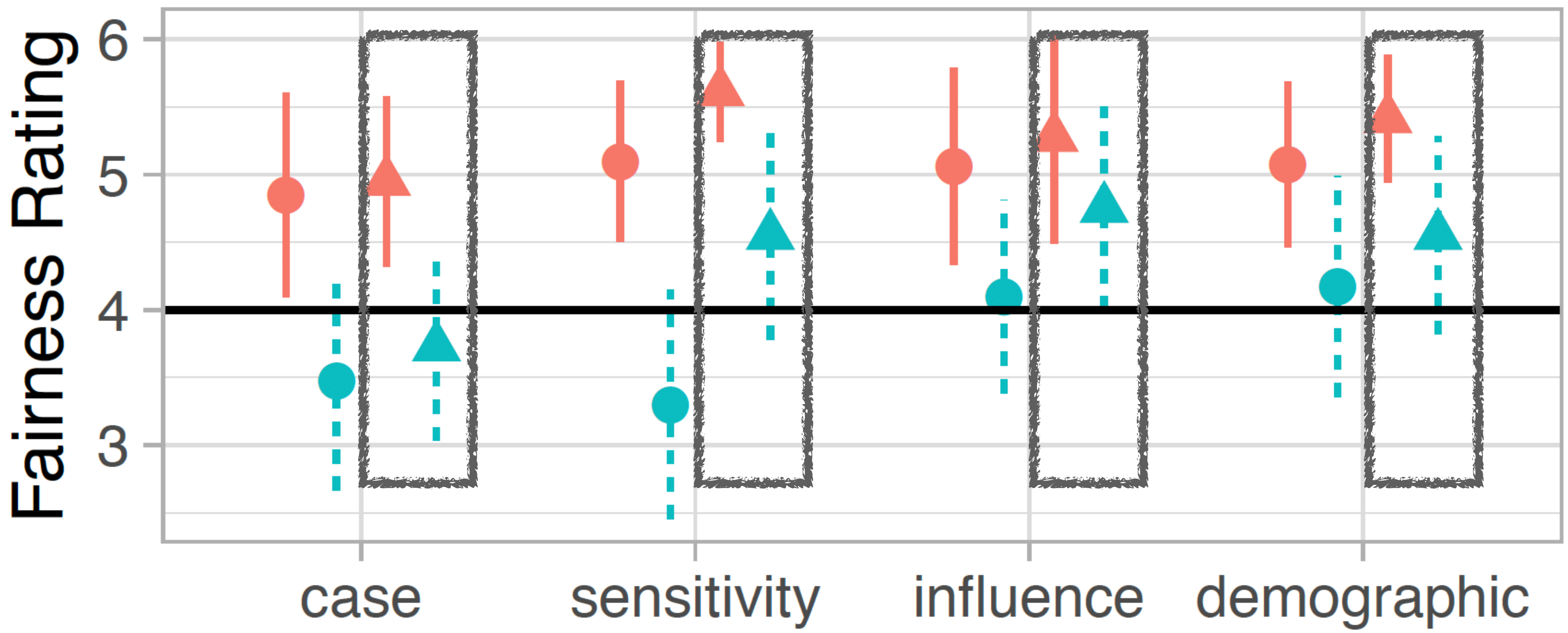
Present model prediction and explanation
Rate "**how the software made the prediction was fair**" and explain the rating

Survey: demographic, prior position on general ML fairness, fairness of race feature, cognitive styles



Legend: data process *(raw=●, processed=▲)*, and sample group *(impacted=blue dashed lines, non-impacted=red solid lines)*

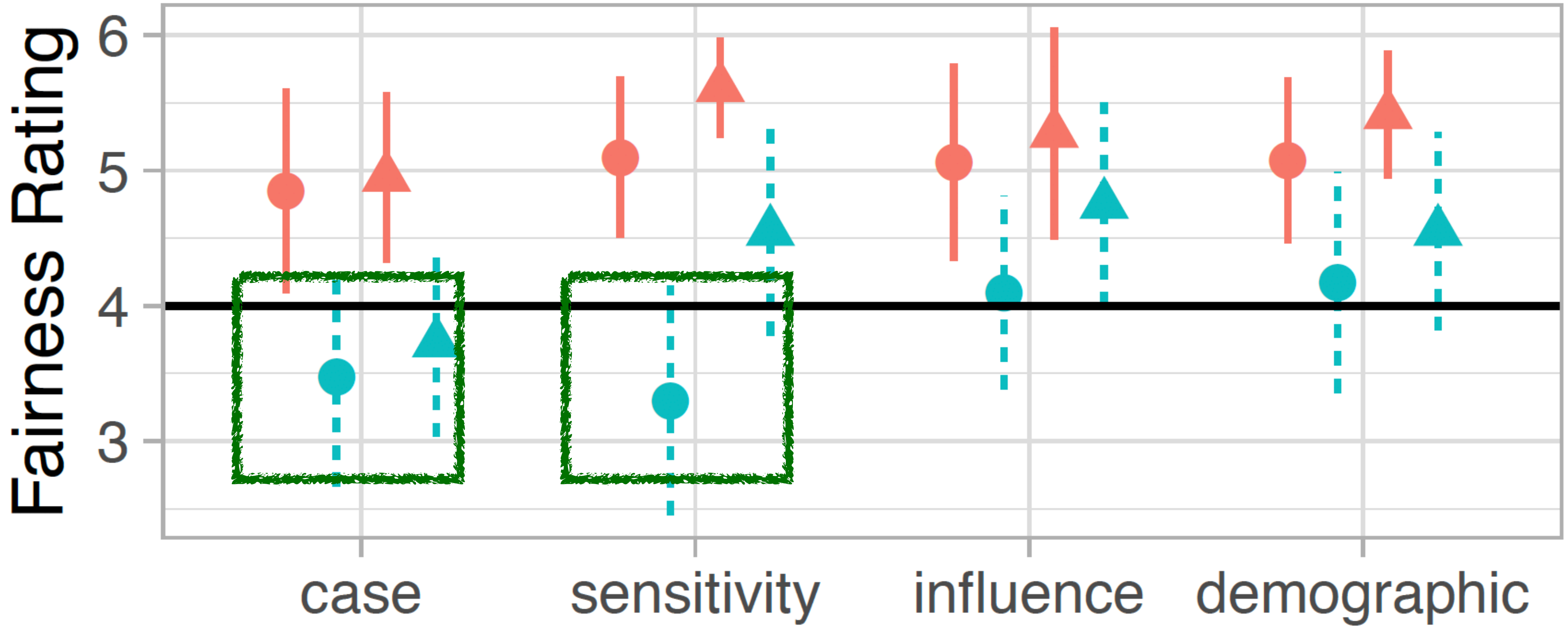
Fairness calibration?



Legend: data process *(raw=●, processed=▲)*, and sample group *(impacted=blue dashed lines, non-impacted=red solid lines)*

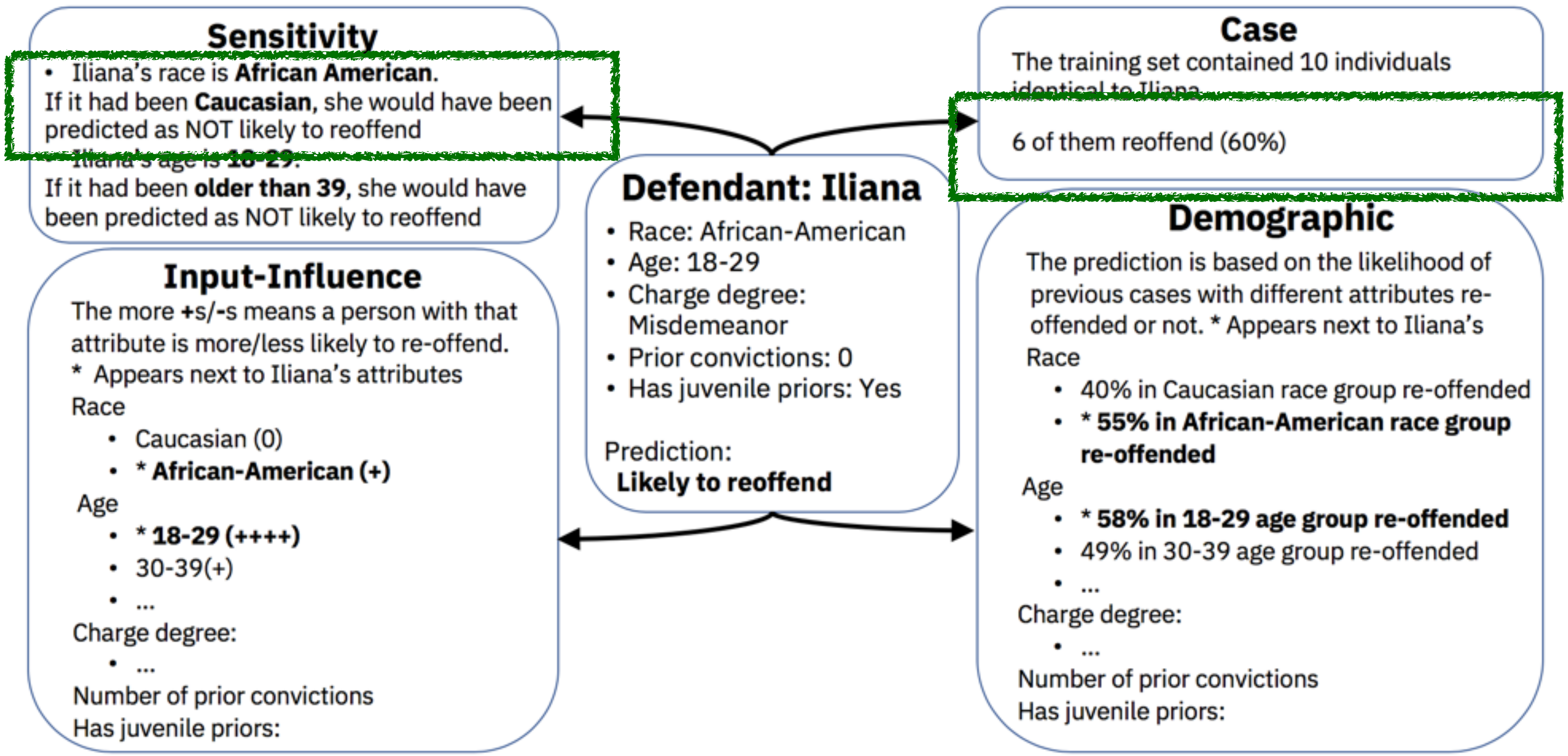
All styles of explanation supported fairness calibration

Surfacing individual fairness issue?

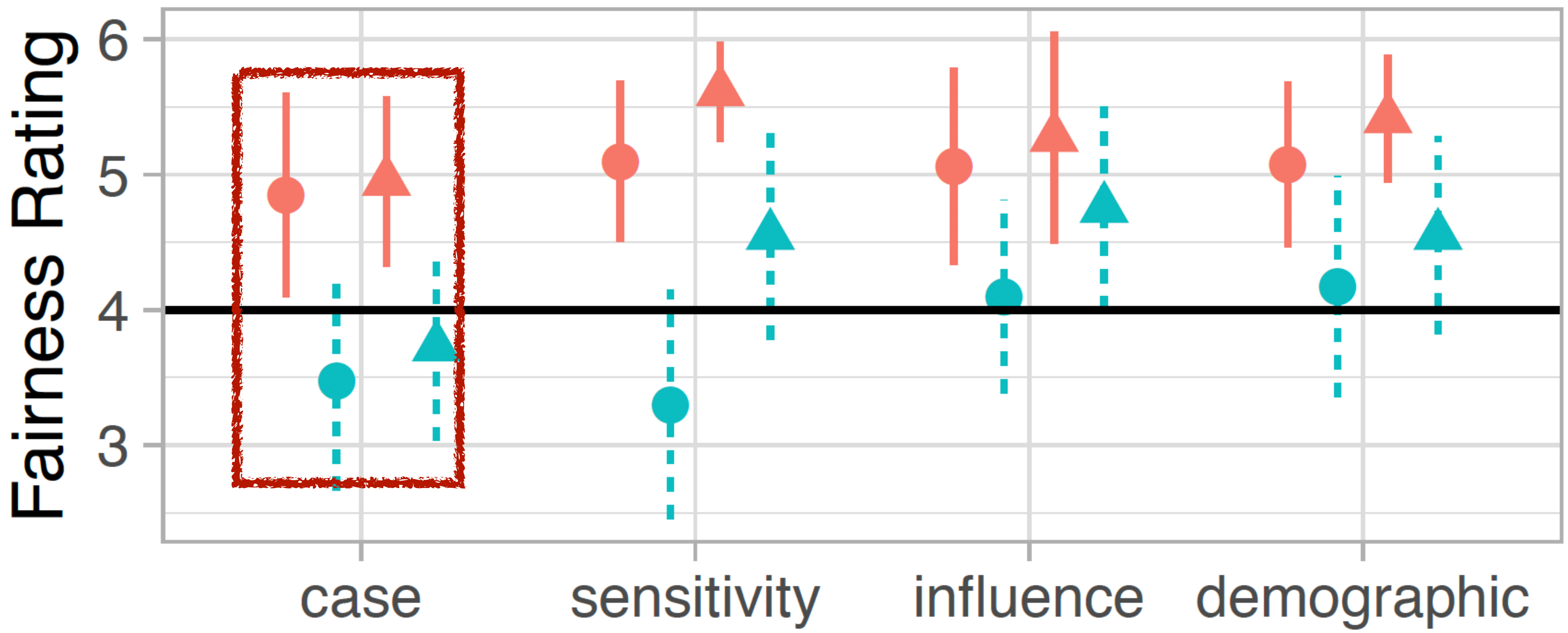


Legend: data process *(raw=●, processed=▲)*, and sample group *(impacted=blue dashed lines, non-impacted=red solid lines)*

Local (why) explanations are more effective in surfacing individual fairness issue


Sensitivity
• Iliana's race is African American.
If it had been Caucasian, she would have been predicted as NOT likely to reoffend
• Iliana's age is 18-29.
If it had been older than 39, she would have been predicted as NOT likely to reoffend
Case
The training set contained 10 individuals identical to Iliana.
6 of them reoffend (60%)
Defendant: Iliana
• Race: African-American
• Age: 18-29
• Charge degree: Misdemeanor
• Prior convictions: 0
• Has juvenile priors: Yes
Prediction:
Likely to reoffend
Input-Influence
The more +s/-s means a person with that attribute is more/less likely to re-offend.
* Appears next to Iliana's attributes
Race
• Caucasian (0)
• * African-American (+)
Age
• * 18-29 (++++)
• 30-39(+)
• ...
Charge degree:
• ...
Number of prior convictions
Has juvenile priors:
Demographic
The prediction is based on the likelihood of previous cases with different attributes re-offended or not. * Appears next to Iliana's
Race
• 40% in Caucasian race group re-offended
• * 55% in African-American race group re-offended
Age
• * 58% in 18-29 age group re-offended
• 49% in 30-39 age group re-offended
• ...
Charge degree:
• ...
Number of prior convictions
Has juvenile priors:

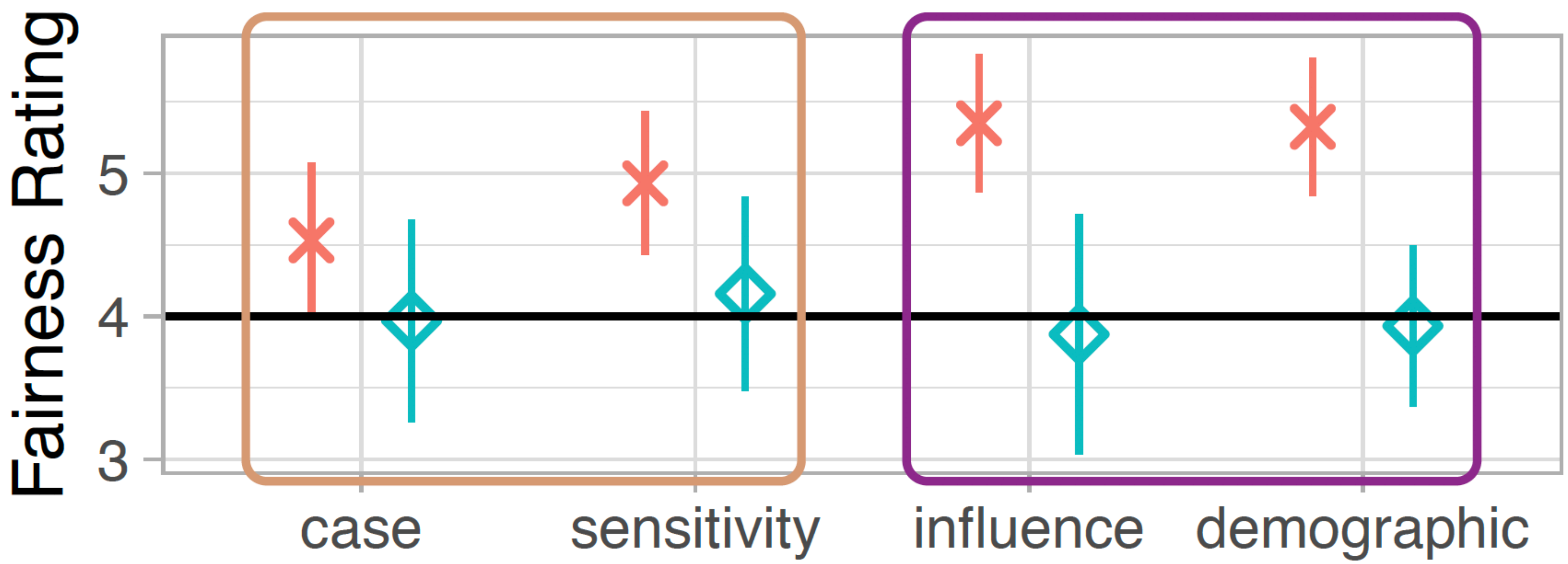
Inherently less fair?



Legend: data process *(raw=●, processed=▲)*, and sample group *(impacted=blue dashed lines, non-impacted=red solid lines)*

Case-based explanation is perceived to be inherently less fair

Individual differences: prior position on ML fairness



Participants who consider "*ML fair to use*" (×) rated the system to be fairer when presented with **global explanations**

Design guidelines: XAI supporting model scrutinization

Design and evaluate with the **goal of calibration**

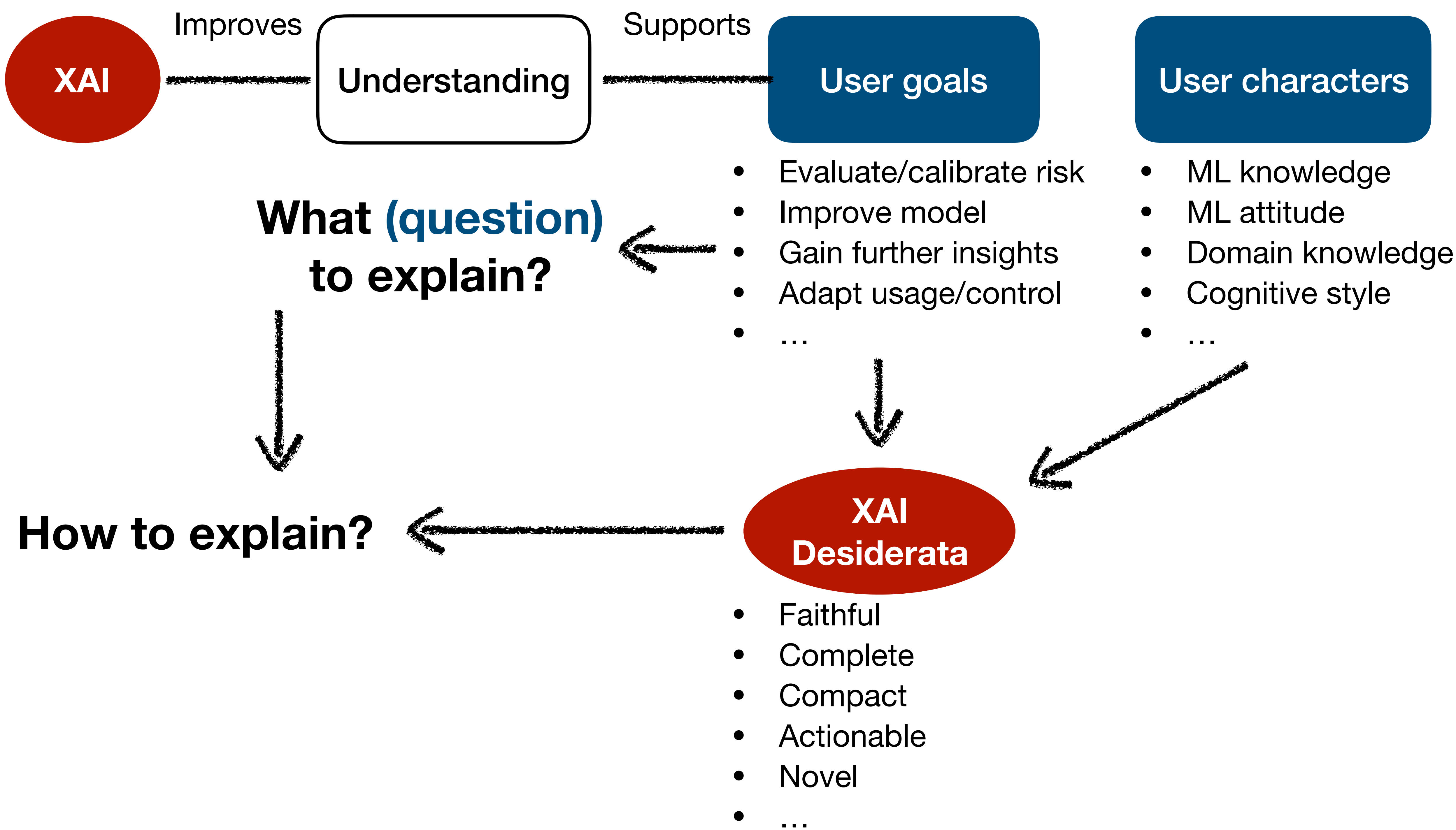
- Start with “ground truth” of model biases/problems

No one-size-fits-all

- Types of fairness problems
- Offsetting v.s. accommodating individual difference
- Fine-grained scrutinization: Data, feature fairness, feature importance, feature interaction, procedural fairness

Dodge et al. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. IUI 2019

Concluding remarks: toward **contextualized** and **actionable** human-centered XAI



Thank **YOU!**

…and thanks to

Rachel Bellamy, Amit Dhurandhar, Jonathan Dodge, Bhavya Ghai, Werner Geyer, Daniel Gruen, Michael Hind, Stephanie Houde, David Millen, Aleksandra Mojsilović, Sarah Miller, Michael Muller, John Richards, Chenhao Tan, Richard Tomsett, Kush Varshney, Justin Weisz, Yunfeng Zhang

Q. Vera Liao
vera.liao@ibm.com
www.qveraliao.com
@QVeraLiao