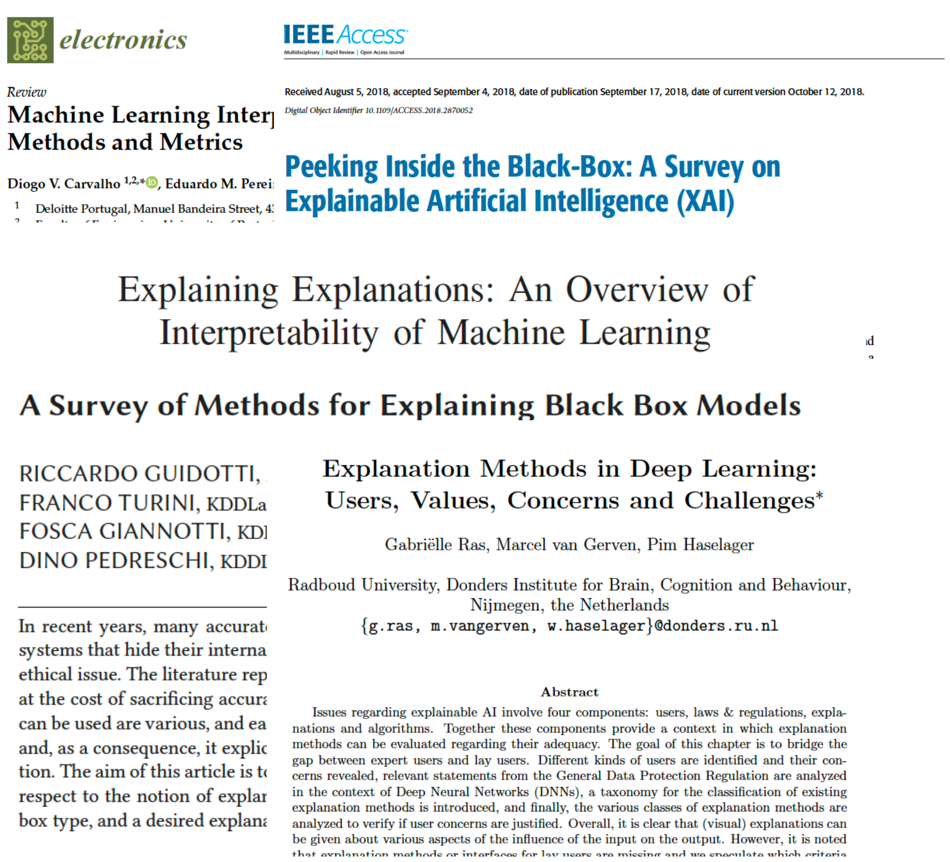
Questioning the AI: Towards User Centered Explainable AI (XAI)

Research work 2018-2021

Q. Vera Liao
IBM Research

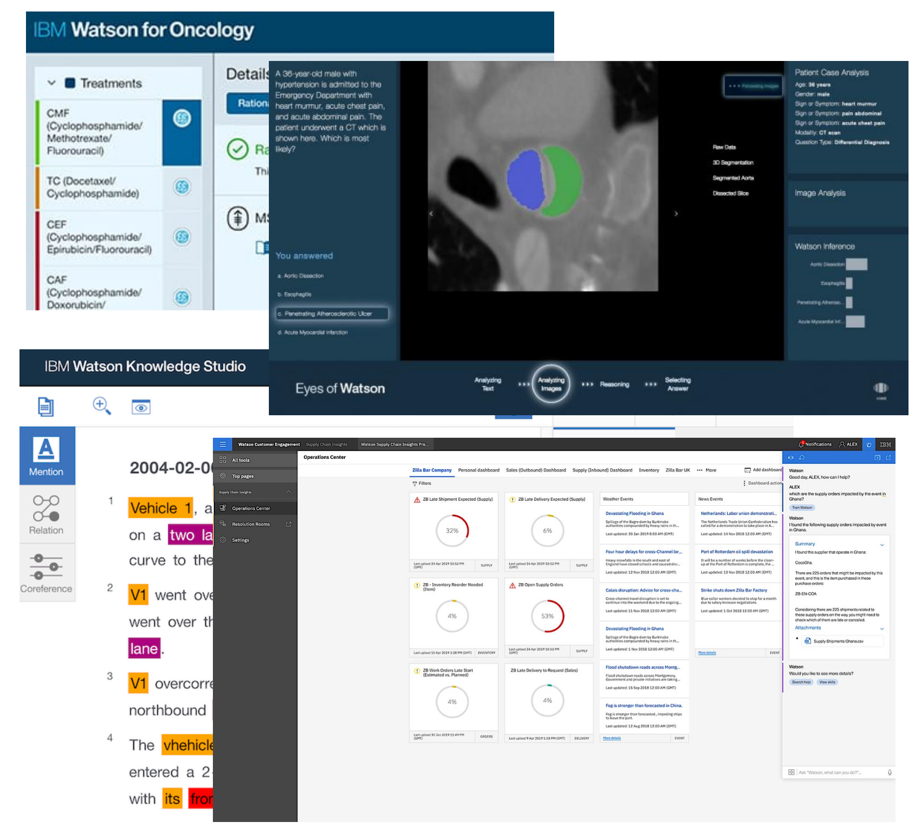
Our HCI research: **Bridging** work

Transfer emerging AI technologies by creating tangible **tools, guidelines, and** design **methods** that support practitioners to navigate the technical and design space



Inform usage

Identify gaps



Explainable AI (**XAI**): Definition

Narrow definition:

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

Broader definition:

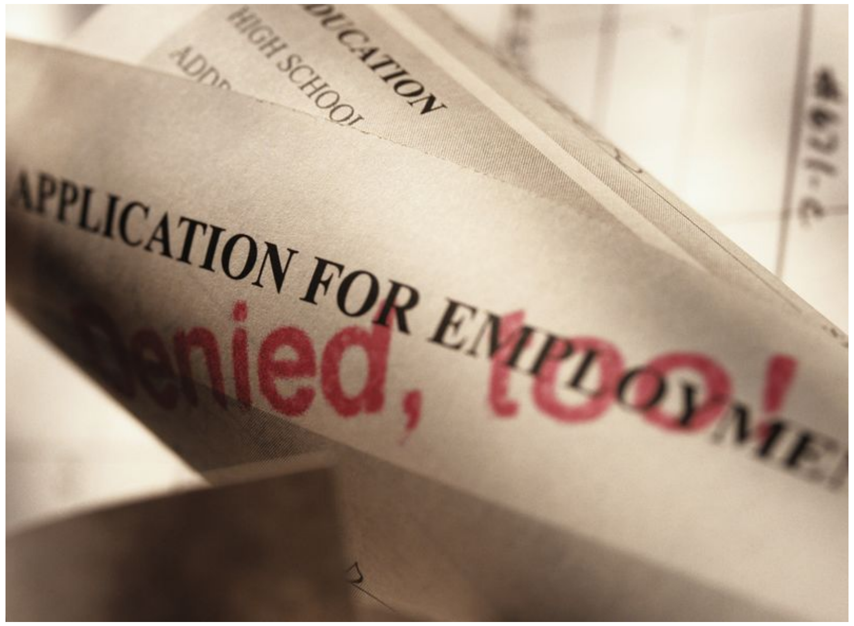
(comprehensible/intelligible AI)

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

XAI is not just ML (also explainable robotics, planning, etc.), but our current work focuses on **explaining supervised ML**

AI is increasingly used in many high-stakes tasks







The quest for explainable AI (XAI)

Companies Grapple With AI's Opaque Decision-Making Process

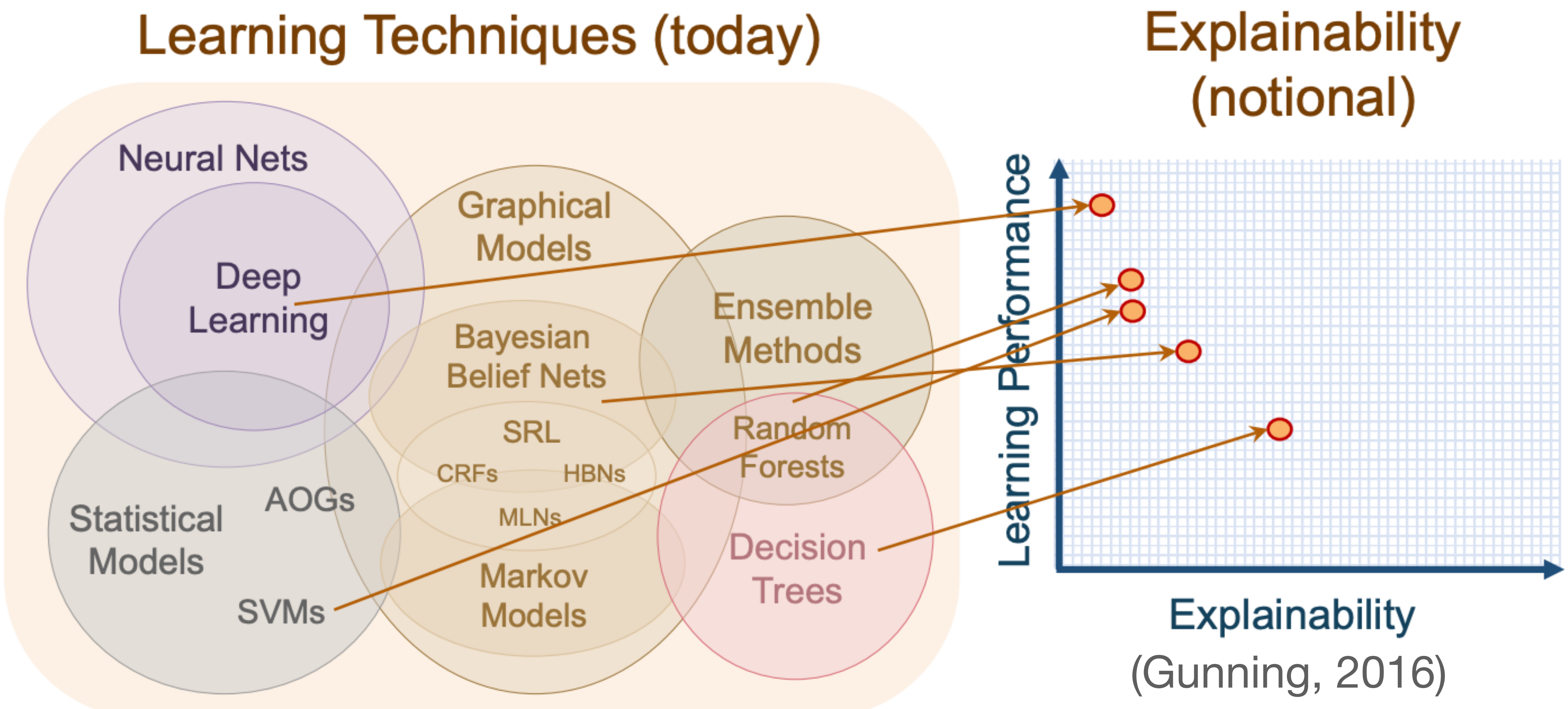
We Need AI That Is Explainable, Auditable, and Transparent

Why "Explainability" Is A Big Deal In AI

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

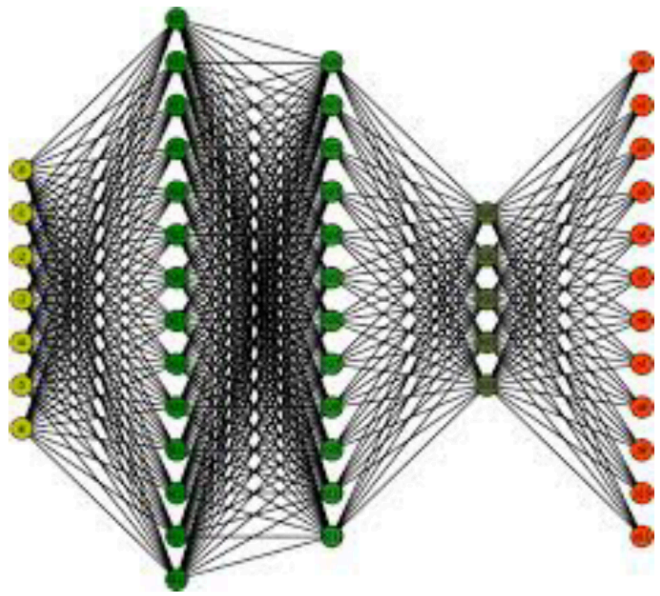
How Explainable AI Is Helping Algorithms Avoid Bias

The needs for XAI algorithms

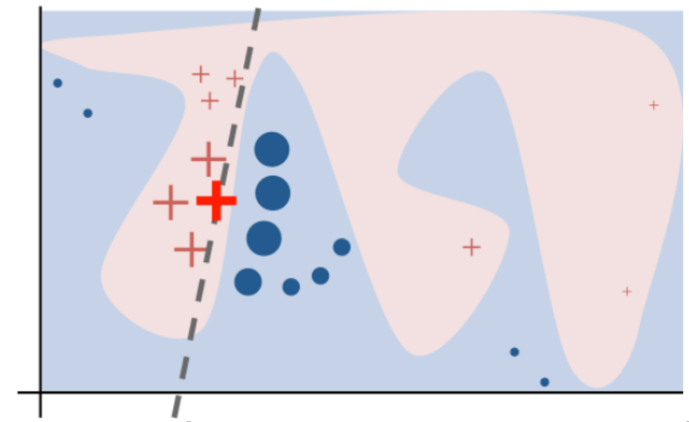


(Gunning, 2016)

XAI “post-hoc” algorithm example: LIME

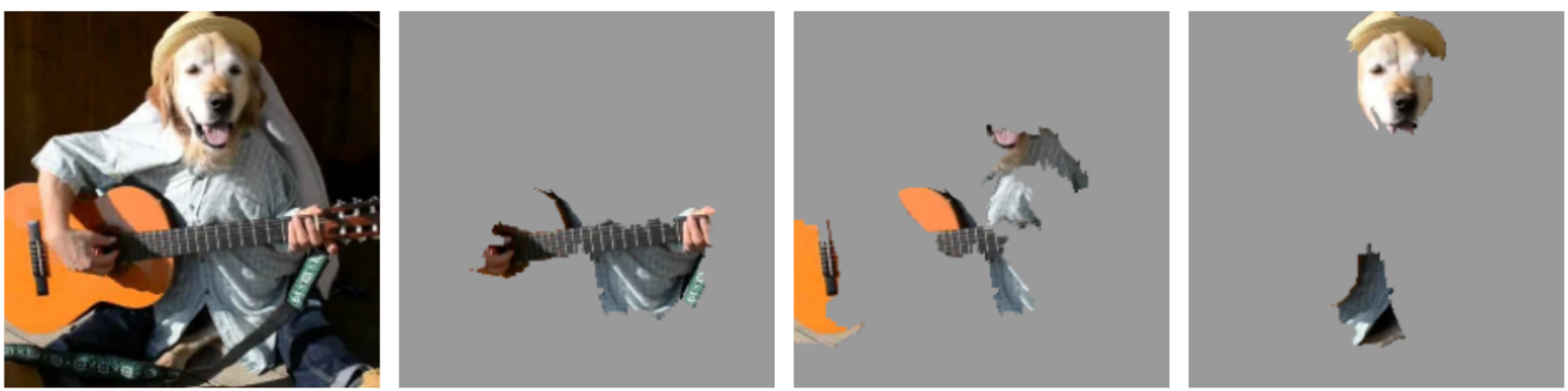


Neural network, not directly explainable

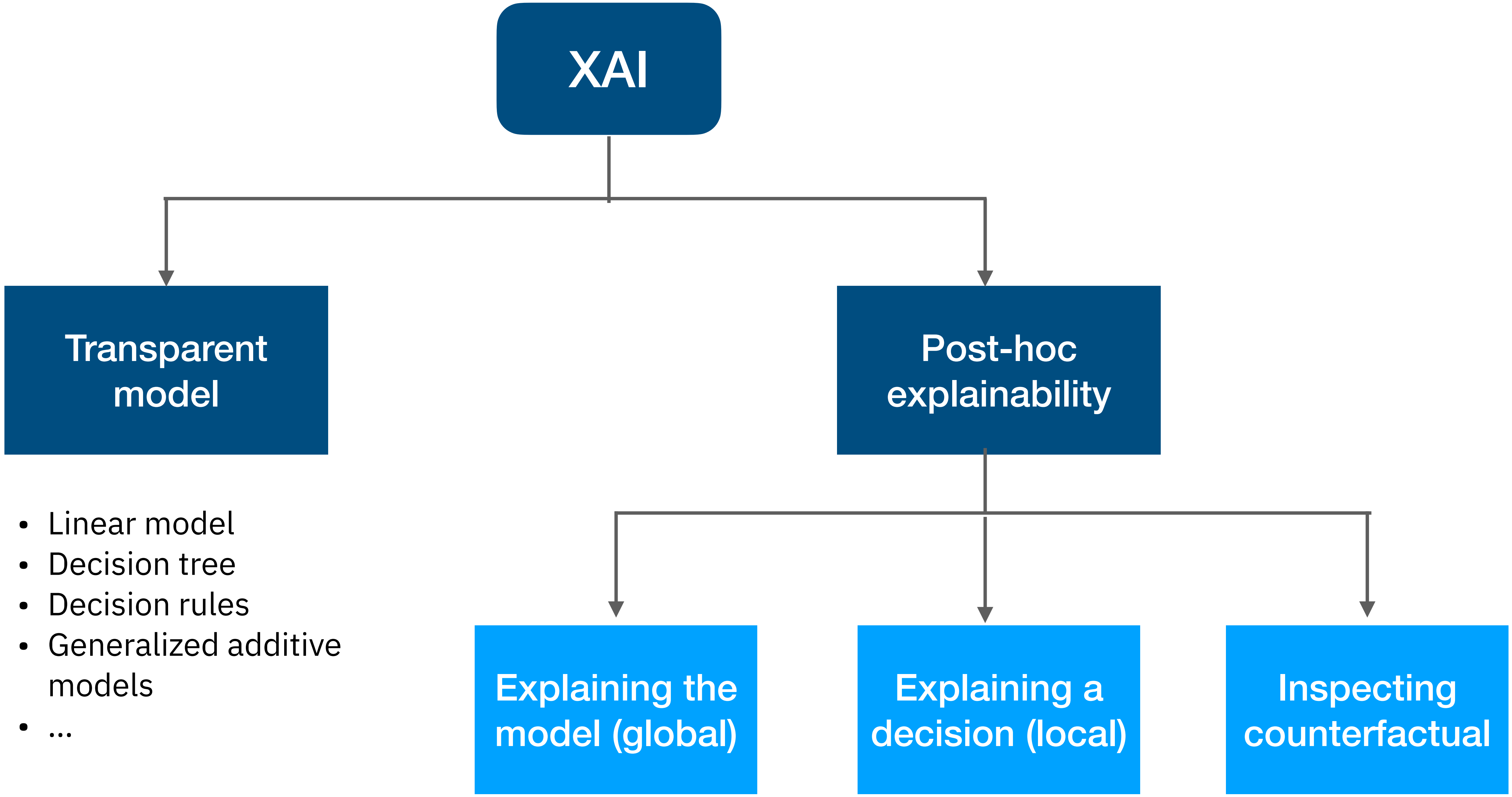


LIME (Ribeiro et al. 2016)

Use a *post-hoc* XAI technique



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



Transparent model

- Linear model
- Decision tree
- Decision rules
- Generalized additive models
- ...

Explaining the model (global)

- Feature importance
- Rule approximation
- Decision tree approximation
- Example based: prototypes

Explaining a decision (local)

- Feature contribution
- Local rules
- Example-based: similar instances

Inspecting counterfactual

- Feature influence
- Contrastive features
- Example-based: counterfactual instances

We will be teaching a virtual tutorial on this at CHI 2021! https://hcixaitutorial.github.io/

A growing collection of XAI techniques

electronics

Review

Machine Learning Interpretability: A Survey on Methods and Metrics

Diogo V. Carvalho [1,2,*], Eduardo M. Pereira [1] and
1 Deloitte Portugal, Manuel Bandeira Street, 43, 4150-47
2 Faculty of Engineering, University of Porto, Dr. Rober
3 INESC TEC, Dr. Roberto Frias Street, 4200-465 Porto, I
* Correspondence: diocarvalho@deloitte.pt

Received: 21 June 2019; Accepted: 24 July 2019; Published

Abstract: Machine learning systems are becoming in
has been expanding, accelerating the shift towar
algorithmically informed decisions have greater p
most of these accurate decision support systems rem
logic and inner workings are hidden to the user a
ratic
mad
ques
The

Explaining Explanations: An Overview of Interpretability of Machine Learning

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter and Lalana Kagal
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139
{lgilpin, davidbau, bzy, abajwa, specter, lkagal}@ mit.edu

Abstract—There has recently been a surge of work in explanatory artificial intelligence (XAI). This research area tackles the important problem that complex machines and algorithms

As a first step towards creating explanation mechanisms there is a new line of research in interpretability, loosel defined as the science of comprehending what a model did (o
le models and learning method
les include visual cues to fin
networks in image recognition

IEEE Access

Received August 5, 2018, accepted September 4, 2018, date of publication September 17, 2018, date of current version October 12, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2870052

Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)

AMINA ADADI AND MOHAMMED BERRADA
Computer and Interdisciplinary Physics Laboratory, Sidi Mohammed Ben Abdellah University, Fez 30050, Morocco
Corresponding author: Amina Adadi (amina.adadi@gmail.com)

ABSTRACT At the dawn of the fourth industrial revolution, we are witnessing a fast and widespread adoption of artificial intelligence (AI) in our daily life, which contributes to accelerating the shift towards a more algorithmic society. However, even with such unprecedented advancements, a key impediment to the use of AI-based systems is that they often lack transparency. Indeed, the black-box nature of these systems allows powerful predictions, but it cannot be directly explained. This issue has triggered a new debate on explainable AI (XAI). A research field holds substantial promise for improving trust and transparency of

A Multidisciplinary Survey and Framework for Design and Evalua

SINA M
ERIC D.

The need f
intelligenc
reasoning
to define, c
on differen
challenges for identifying appropriate design and evaluation methodology and consolidating knowledge from across efforts. To this end, this paper presents a survey and framework intended to share knowledge and experiences of XAI design and evaluation methods across multiple disciplines. Aiming to support diverse design goals and evaluation method in XAI research, after a thorough review of XAI related papers in the fields of machine learning, visualization, and human-compute

omies,

netot[b,e,f],
olina[g],
[b]ENSTA, Institute Polytechnique Paris and INRIA Flowers Team, Palaiseau, France
[c]University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain
[d]Basque Center for Applied Mathematics (BCAM), 48009 Bilbao, Bizkaia, Spain
[e]Segula Technologies, Parc d'activité de Pissaloup, Trappes, France
[f]Institut des Systèmes Intelligents et de Robotique, Sorbonne Universitè, France
mputational Intelligence, University of Granada, 18071 Granada, Spain
nica, 28050 Madrid, Spain

A Survey of Methods for Explaining

RICCARDO GUIDOTTI, ANNA MONREALE, SALV
FRANCO TURINI, KDDLab, University of Pisa, Italy
FOSCA GIANNOTTI, KDDLab, ISTI-CNR, Italy
DINO PEDRESCHI, KDDLab, University of Pisa, Italy

In recent years, many accurate decision support systems have
systems that hide their internal logic to the user. This lack of ex
ethical issue. The literature reports many approaches aimed at o
at the cost of sacrificing accuracy for interpretability. The appli
can be used are various, and each approach is typically develope
and, as a consequence, it explicitly or implicitly delineates its ov
tion. The aim of this article is to provide a classification of the m
respect to the notion of explanation and the type of black box
box type, and a desired explanation, this survey should help the

Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges*

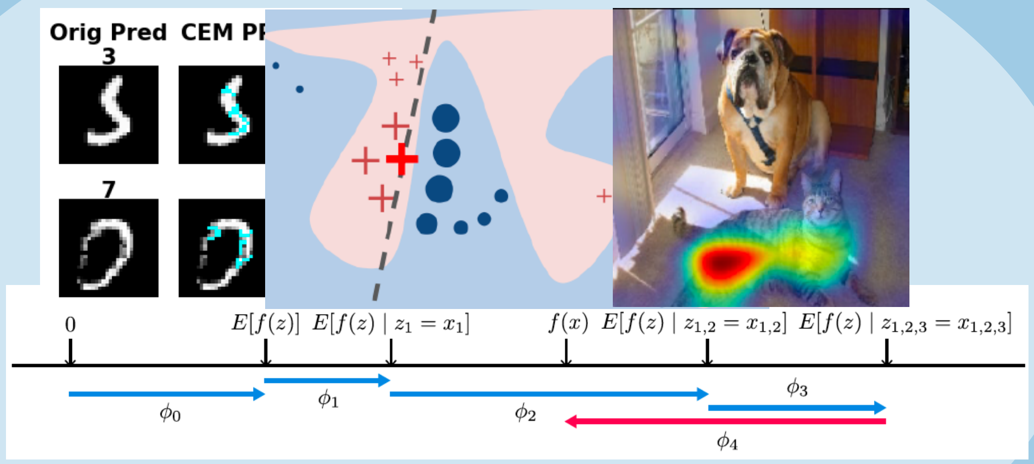
Gabriëlle Ras, Marcel van Gerven, Pim Haselager
Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, the Netherlands
{g.ras, m.vangerven, w.haselager}@donders.ru.nl

Abstract

Issues regarding explainable AI involve four components: users, laws & regulations, explanations and algorithms. Together these components provide a context in which explanation methods can be evaluated regarding their adequacy. The goal of this chapter is to bridge the gap between expert users and lay users. Different kinds of users are identified and their concerns revealed, relevant statements from the General Data Protection Regulation are analyzed in the context of Deep Neural Networks (DNNs), a taxonomy for the classification of existing explanation methods is introduced, and finally, the various classes of explanation methods are analyzed to verify if user concerns are justified. Overall, it is clear that (visual) explanations can be given about various aspects of the influence of the input on the output. However, it is noted

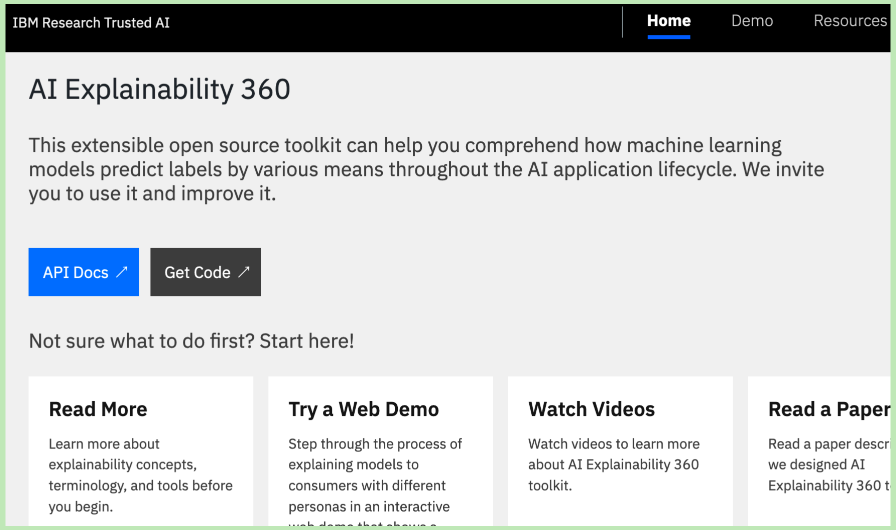
(AI) has achieved a notable momentum that, if harnessed
ations over many application sectors across the field. For this
ire community stands in front of the barrier of explainability,
brought by sub-symbolism (e.g. ensembles or Deep Neural
ype of AI (namely, expert systems and rule based models).
in the so-called *eXplainable* AI (XAI) field, which is widely
ctical deployment of AI models. The overview presented in
d contributions already done in the field of XAI, including a
r this purpose we summarize previous efforts made to define
ning a novel definition of explainable Machine Learning that
th a major focus on the audience for which the explainability
propose and discuss about a taxonomy of recent contributions

XAI in Academia



An abundance of XAI algorithms

XAI in Practice

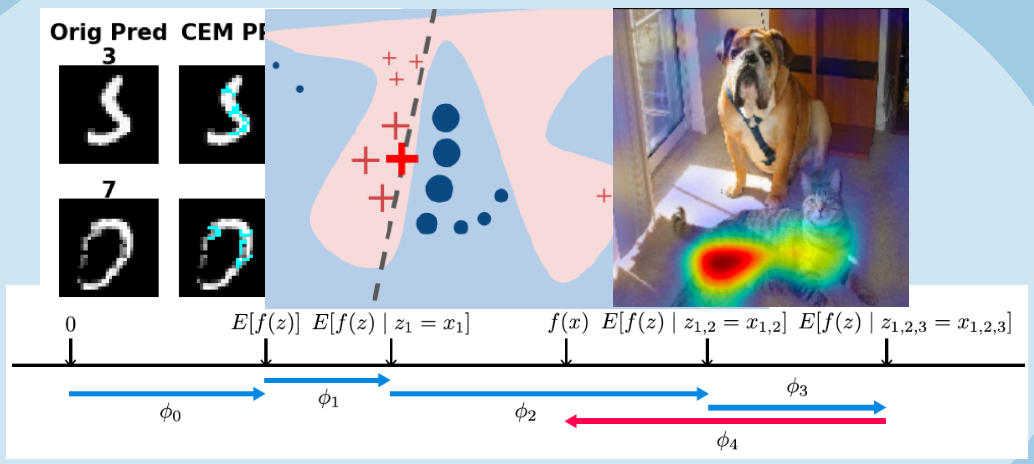


Toolbox of XAI techniques

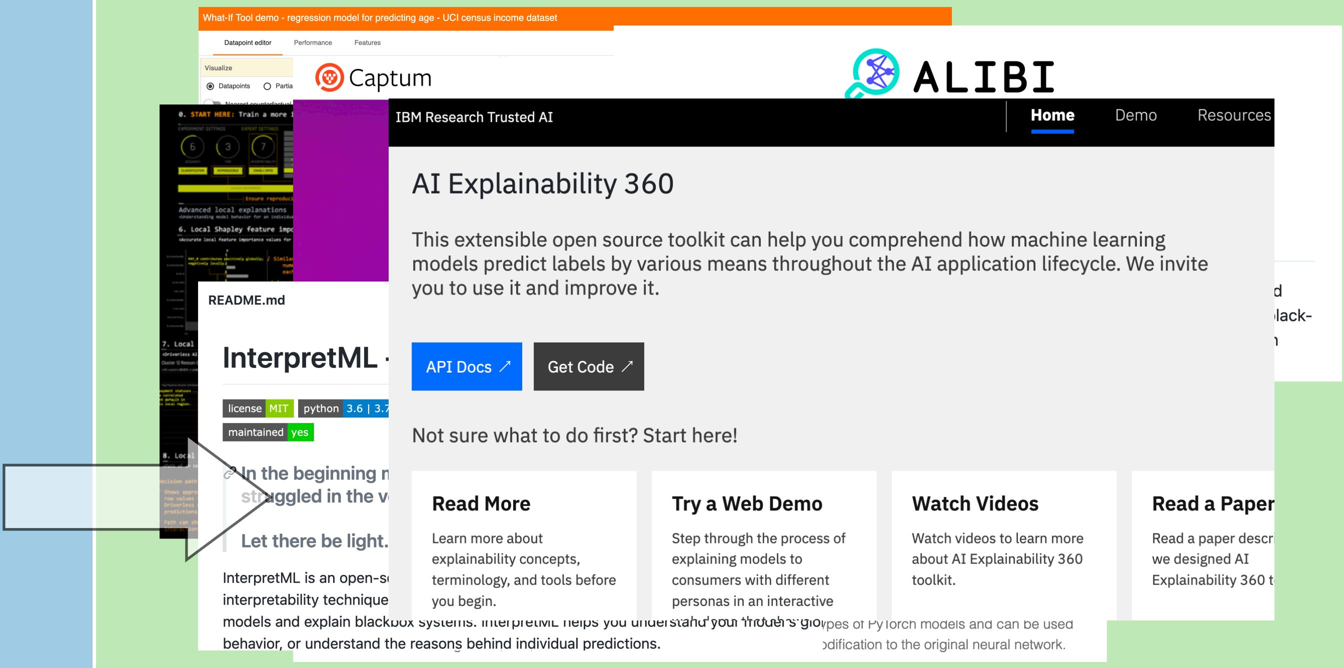
From academic research into a practitioners' toolbox

XAI in Academia

XAI in Practice



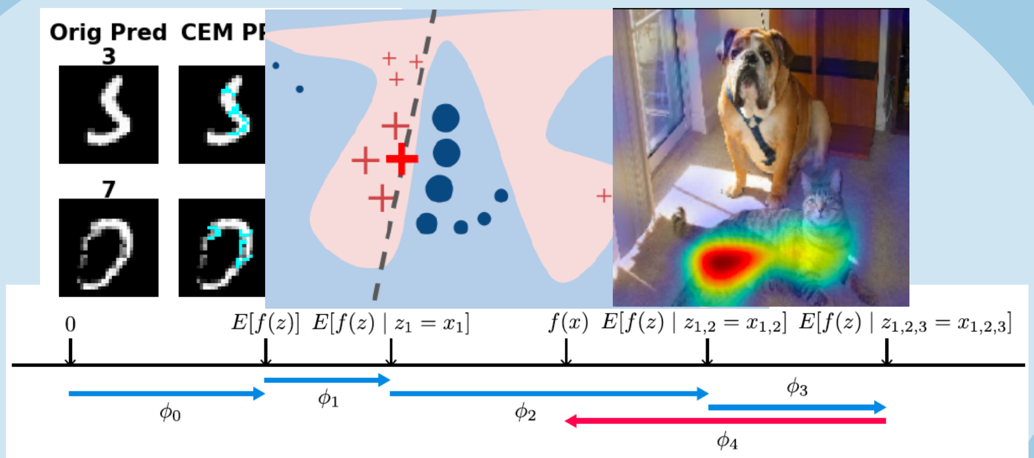
An abundance of XAI algorithms



Toolbox of XAI techniques

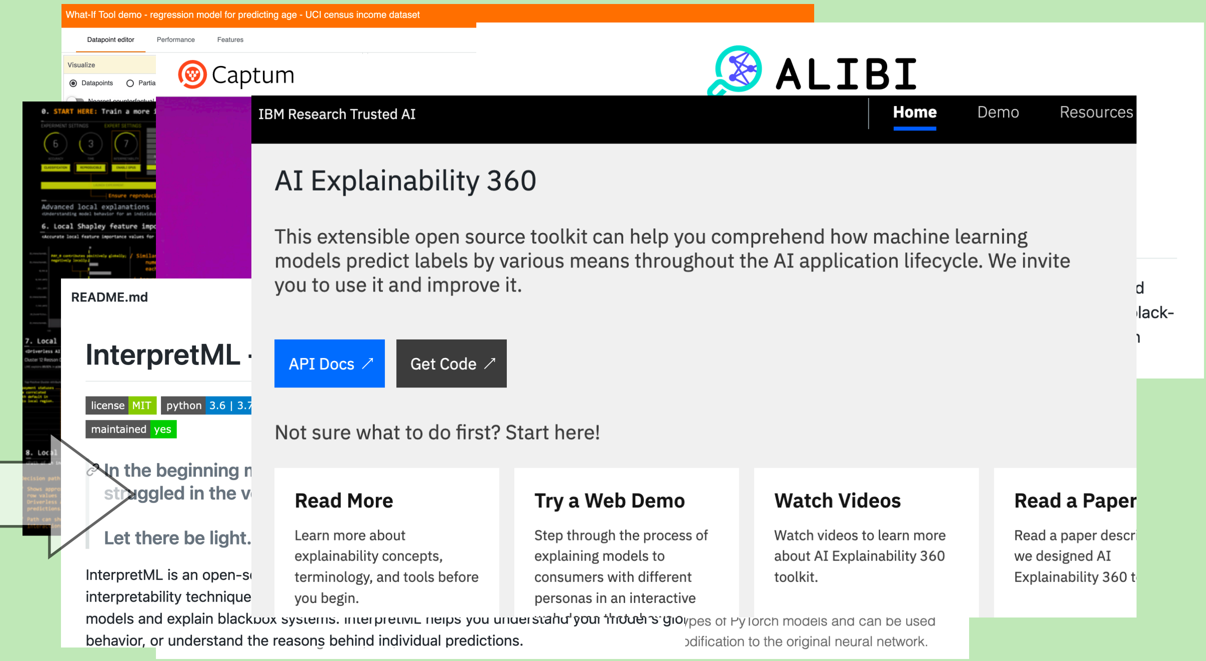
From academic research into a practitioners' toolbox

XAI in Academia



An abundance of XAI algorithms

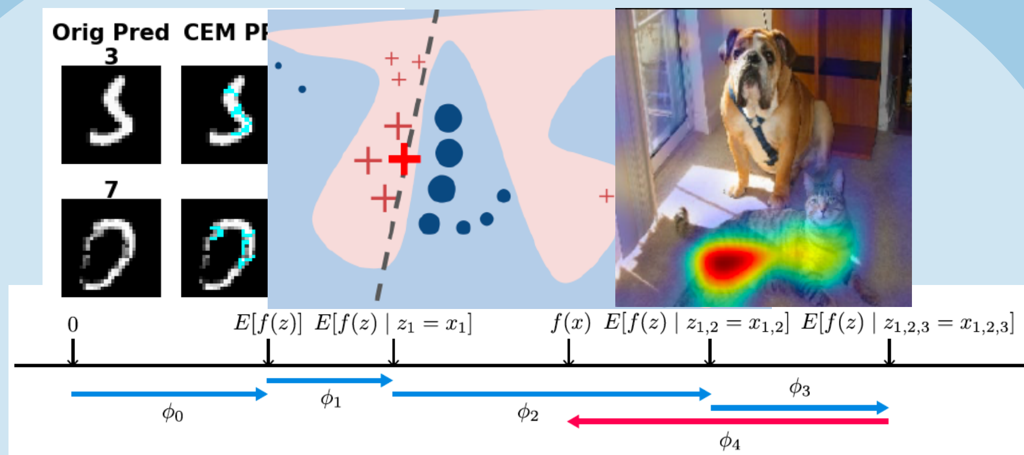
XAI in Practice



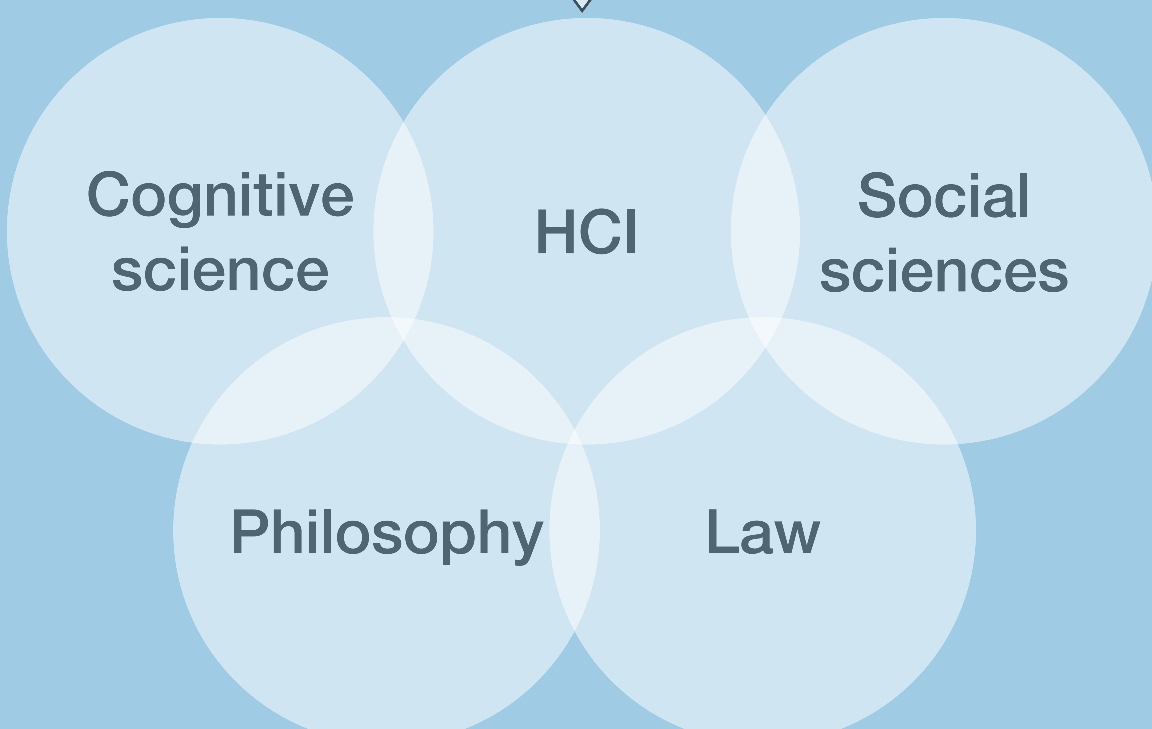
Toolbox of XAI techniques

Towards real-world XAI: serving many domains and user groups

XAI in Academia

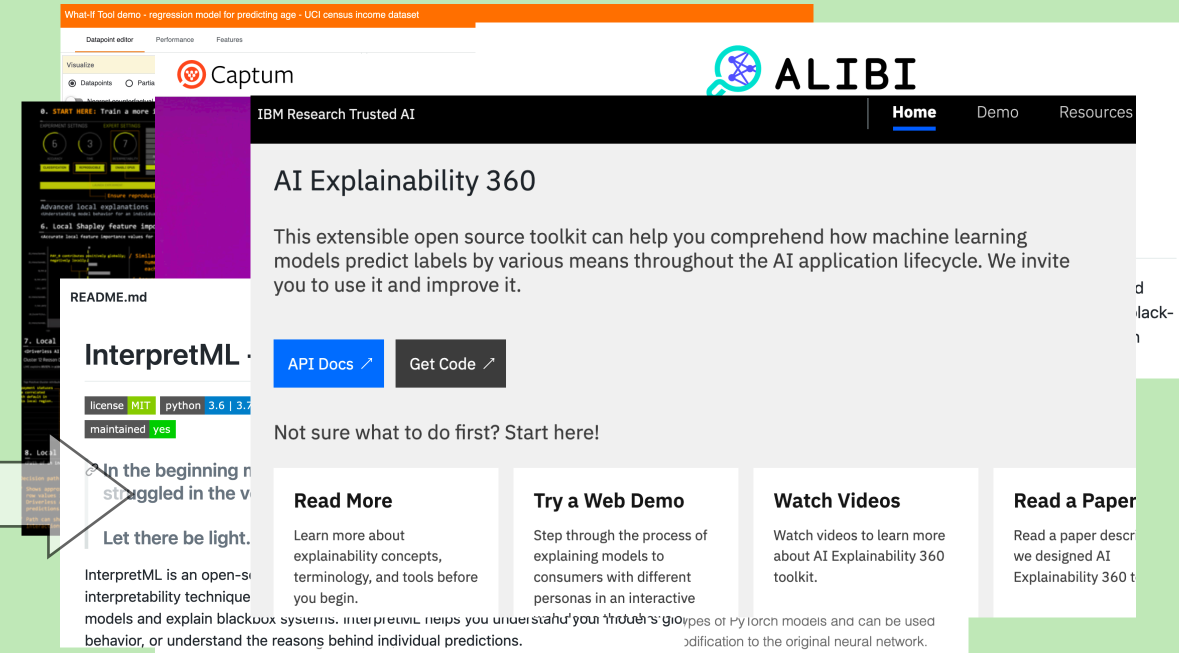


An abundance of XAI algorithms



Inter-disciplinary perspectives

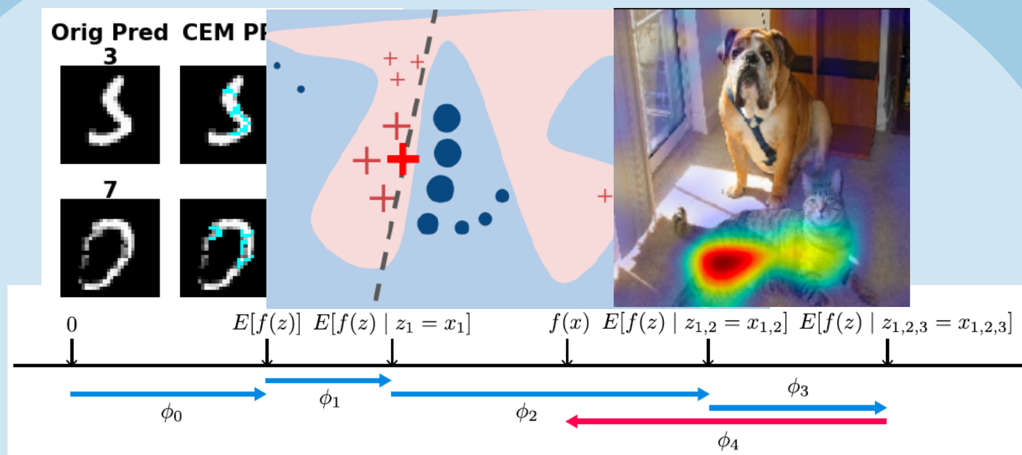
XAI in Practice



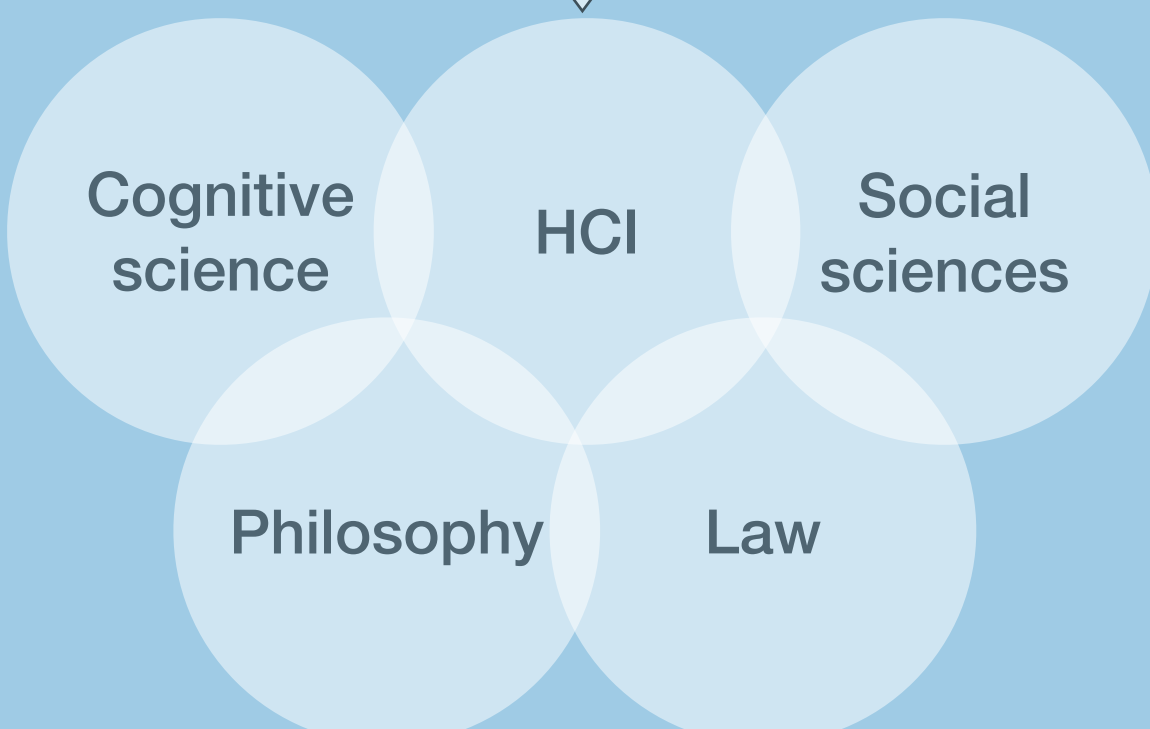
Toolbox of XAI techniques

Towards real-world XAI: serving many domains and user groups

XAI in Academia



An abundance of XAI algorithms



Inter-disciplinary perspectives

Inter-disciplinary perspectives

- **Plurality of motivation** for explanation: diagnosis, predicting the future, sense-making, justification, reconciling dissonance, etc. (Kiel 2006; Lombrozo, 2006)
- Explanatory power is **recipient dependent,** including the question asked (**explanatory relevance**) (Hilton, 1990; Walton, 2004)
- More complexities:
  - The **plurality of cognitive processes** (Petty and Cacioppo, 1986; Horne et al, 2013)
  - **Socio-technical systems** (Ehsan et al., 2021)

From XAI algorithms to XAI UX

With a toolbox:

How to **select**?

How to **translate**?

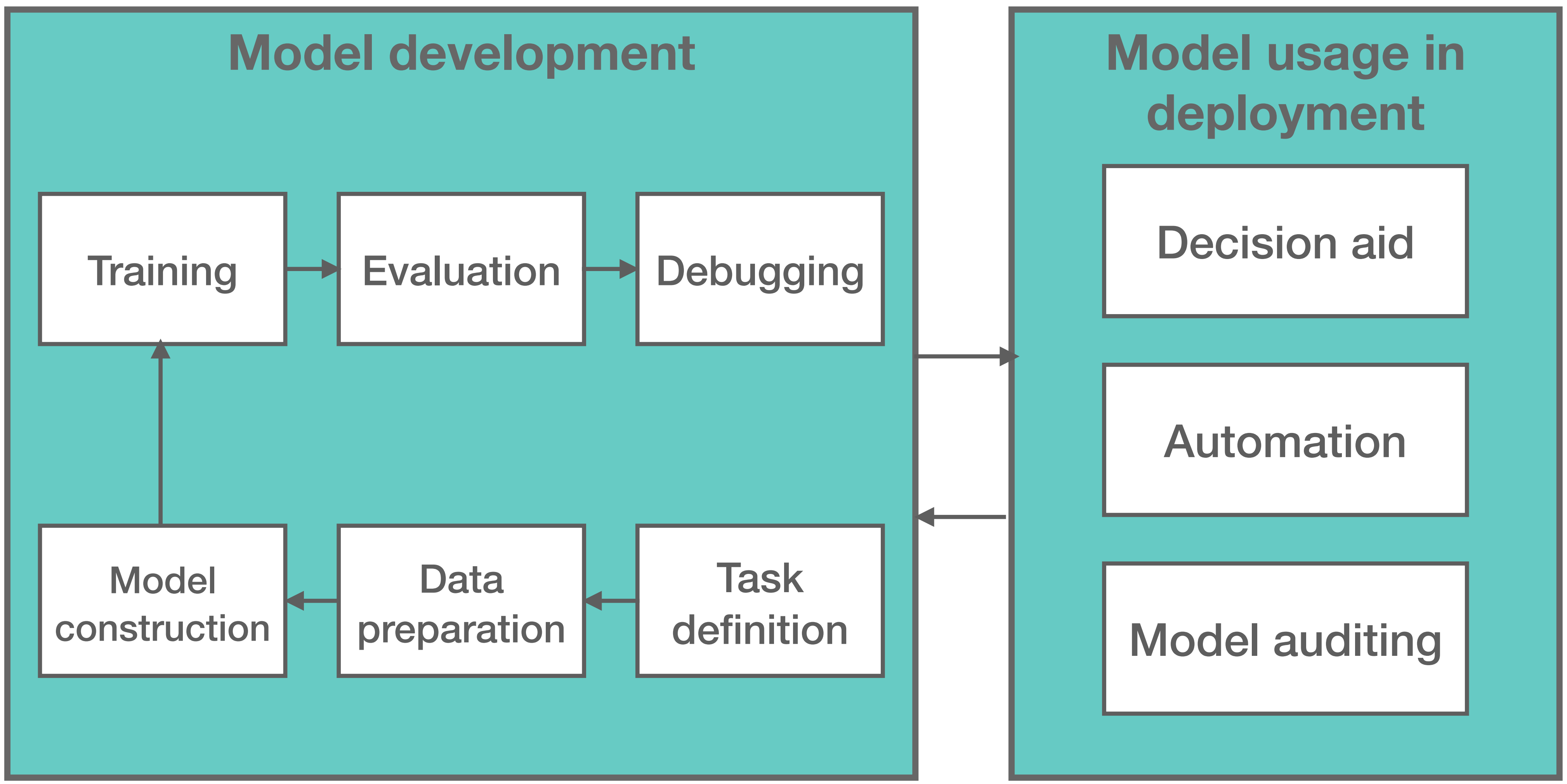
Our paths:

- Develop context-specific design guidelines: **HCI research with XAI use cases**
- Tackle the design process: **User centered design of XAI**

XAI use cases in AI lifecycle

Model evaluation and selection (IUI2021)
XAI consumer: Data scientist

Trust calibration and decision support (FAT* 2020, CHI 2021 🎖)
XAI consumer: Decision-maker



Delegation support (ongoing)
XAI consumer: Domain expert

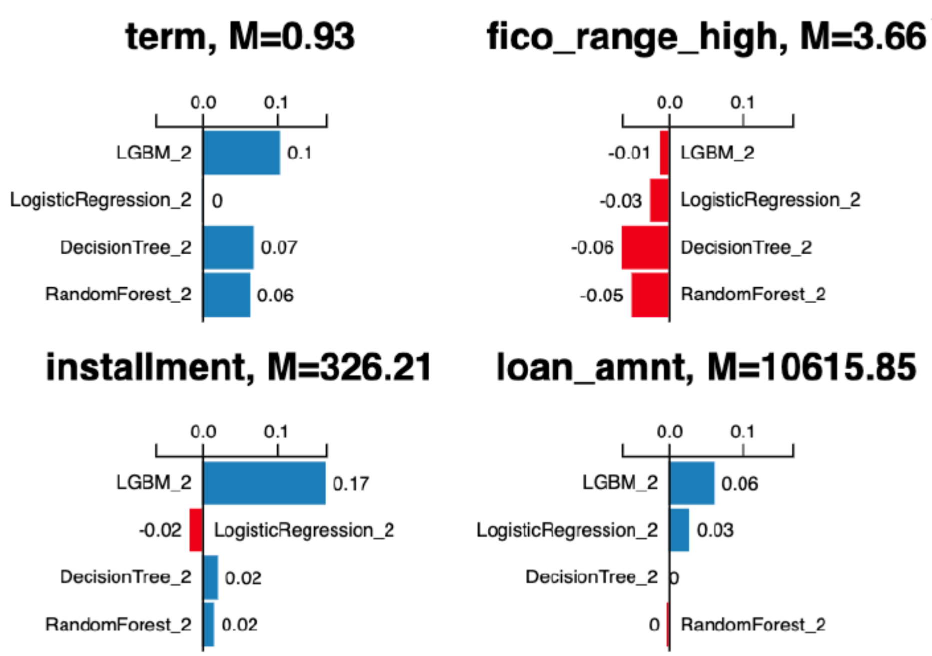
Explainable active learning (CSCW 2020)
XAI consumer: Annotator (domain expert)

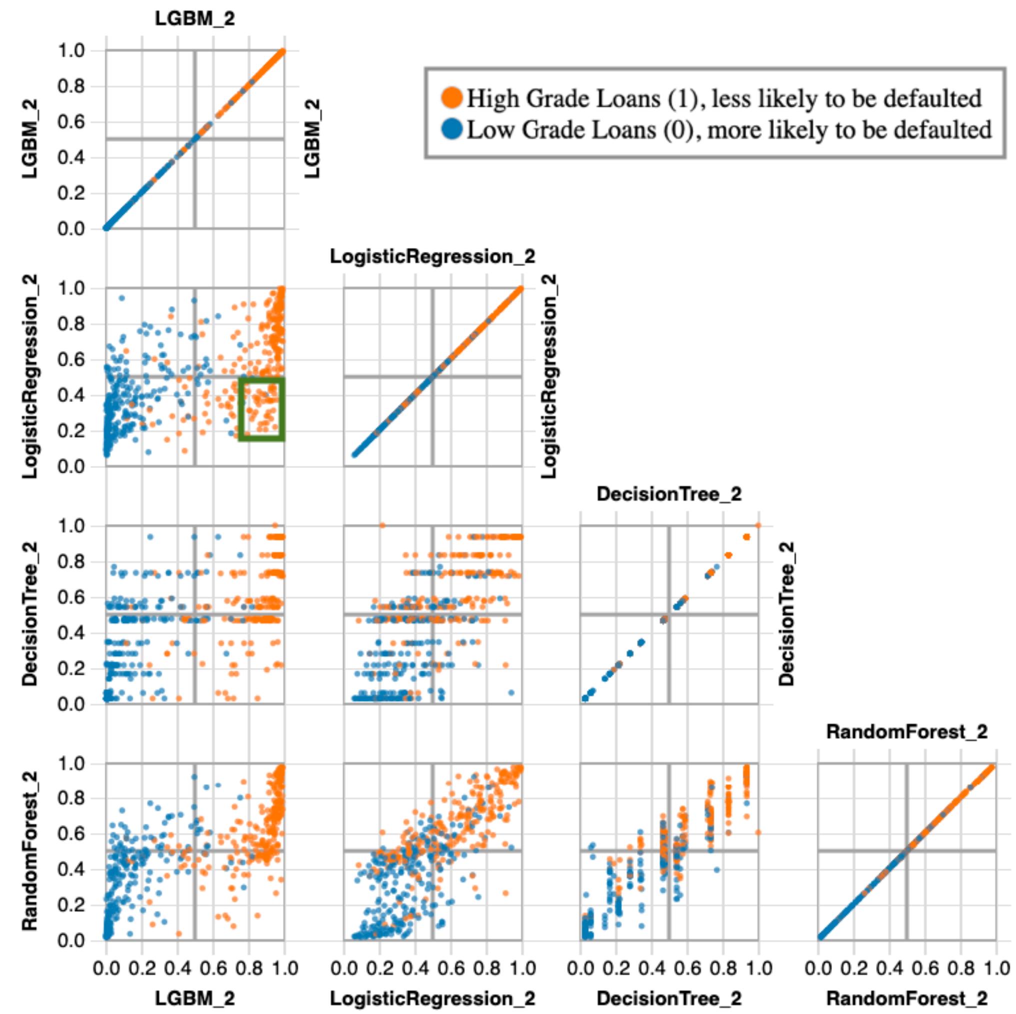
Fairness assessment (IUI 2019 🎖)
XAI consumer: Regulator, impacted groups

XAI for **model evaluation and selection**

	f1	accuracy	roc_auc	precision	recall	neg_log_loss
LGBM_2	0.922	0.923	0.923	0.926	0.918	-2.66
LogisticRegression_2	0.699	0.712	0.712	0.725	0.675	-9.95
DecisionTree_2	0.694	0.707	0.706	0.719	0.67	-10.1
RandomForest_2	0.752	0.755	0.755	0.756	0.747	-8.46

(a) Screenshot of the Metrics Table showing metrics for four selected models.



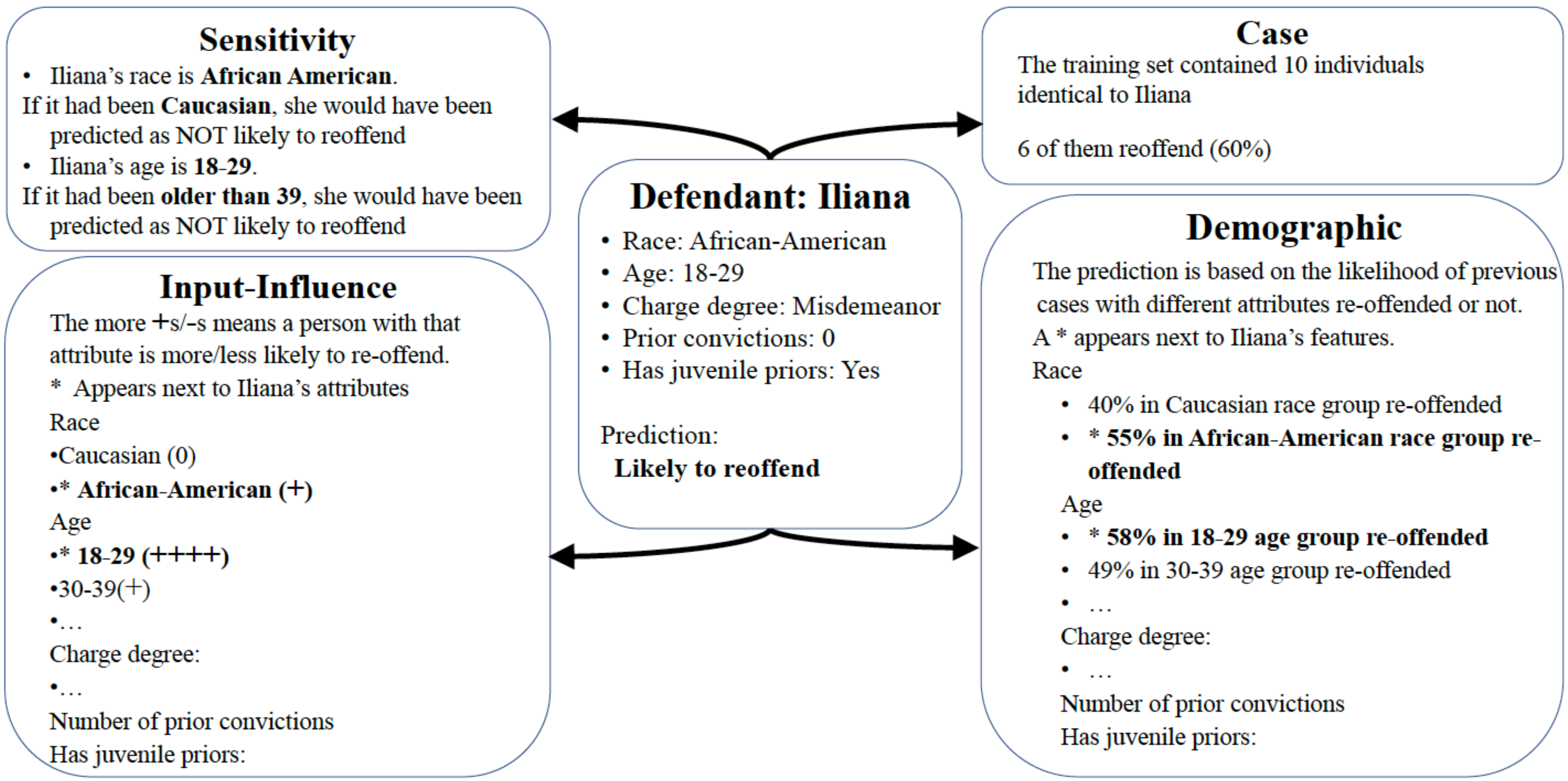


Data scientist

How does each model make predictions?
Why are these instances predicted differently by these models?
Why is this model making a wrong prediction?

Narkar et al. Model LineUpper: Supporting Interactive Model Comparison at Multiple Levels for AutoML. IUI 2021

XAI for **fairness assessment**



Auditor

Impacted groups

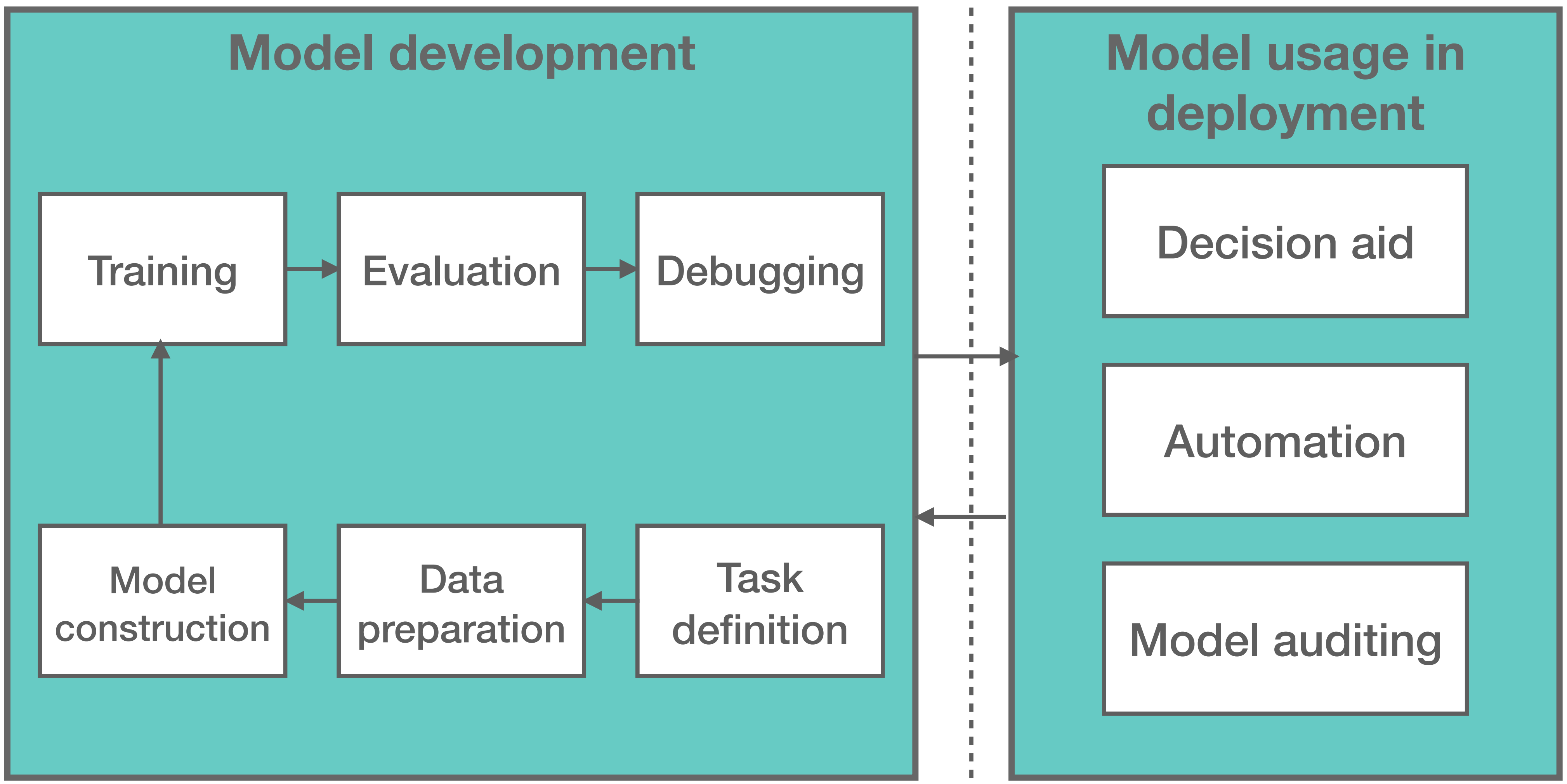
Is the way the model makes risk predictions fair?
Why is this person predicted of high risk?
Is he/she treated fairly?

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

Lessons learned: From XAI algorithms to XAI UX

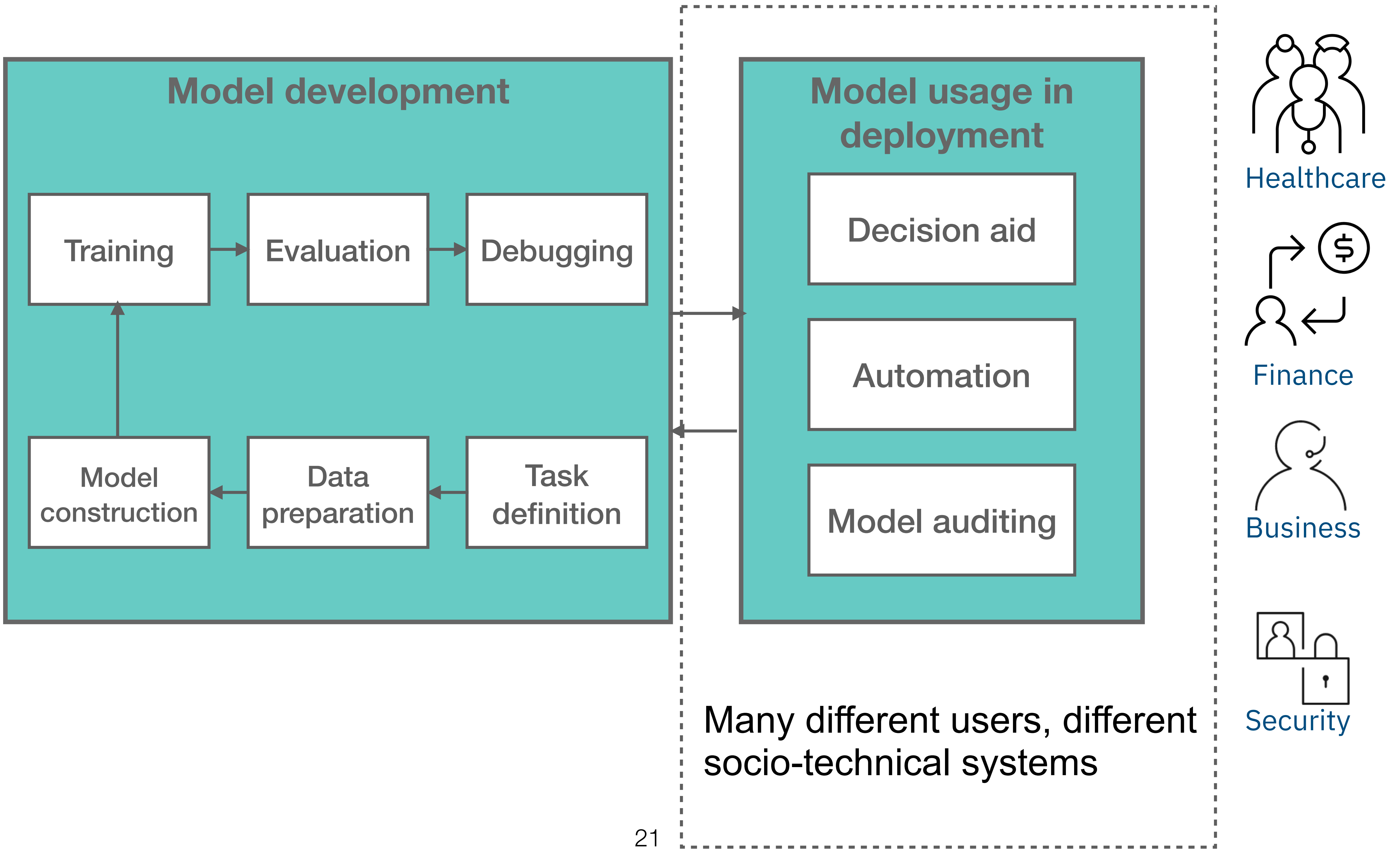
- No one-fits-all solutions
- XAI UX often needs multiple types of explanations
  - Anticipate *when* and *where* users want *what* explanations
- Beware of the potential harm of XAI
  - Unwarranted trust and confidence
  - Distraction and cognitive workload
  - Can unequalized or marginalize certain groups
- Under-developed "translation" design space
  - Algorithmic output needs communication, elaboration, constraints, integration, etc.
  - Drive adapting or developing new XAI algorithms
- Breakdowns more often, translation design more necessary, on the model usage side

Why break-downs in model usage?



"*Inmates running the asylum*": XAI algorithms developed by AI researchers with their own intuitions, values and epistemic background

Why break-downs in model usage?



From XAI algorithms to XAI UX

With a toolbox:

How to **select**?

How to **translate**?

How to **expand?**

Our paths:

- Develop context-specific design guidelines: **HCI research with XAI use cases**
- Tackle the design process: **User centered design of XAI**
- Socially situated explainability by making visible the AI contexts

Towards "social transparency" in AI systems

Customer: Scout Inc. **Product:** Access Management (SaaS) **Product ID (PID):** 43523X

Recommendation: Sell at $100 per account per month

Justification: the AI system considered the following components

[❍] *Quota goals* [❍] *Comparative pricing: what similar customers pay* [❍] *Cost: $55 /account/month*

1

For this customer, 3 members of your team received pricing recommendations in past sales. However, 1 out 3 have sold at the recommended price. Click to see more details.

2

Nadia M.
Sales Assoc. (AB34)

Action: Reject Recommendation ⇔ **Outcome:** No Sale
Comment: Long-term profitable customer; main revenue from a different vertical ; selling at cost price to maintain relationship
Oct 2, 2019

3

Eric C.
Sales Manager (XZ89)

Action: Accept Recommendation ⇔ **Outcome:** Sale
Comment: Recommended price aligned with profit margins; customer felt the price was fair
Dec 14, 2019

4

4W

What
Who
Why
When

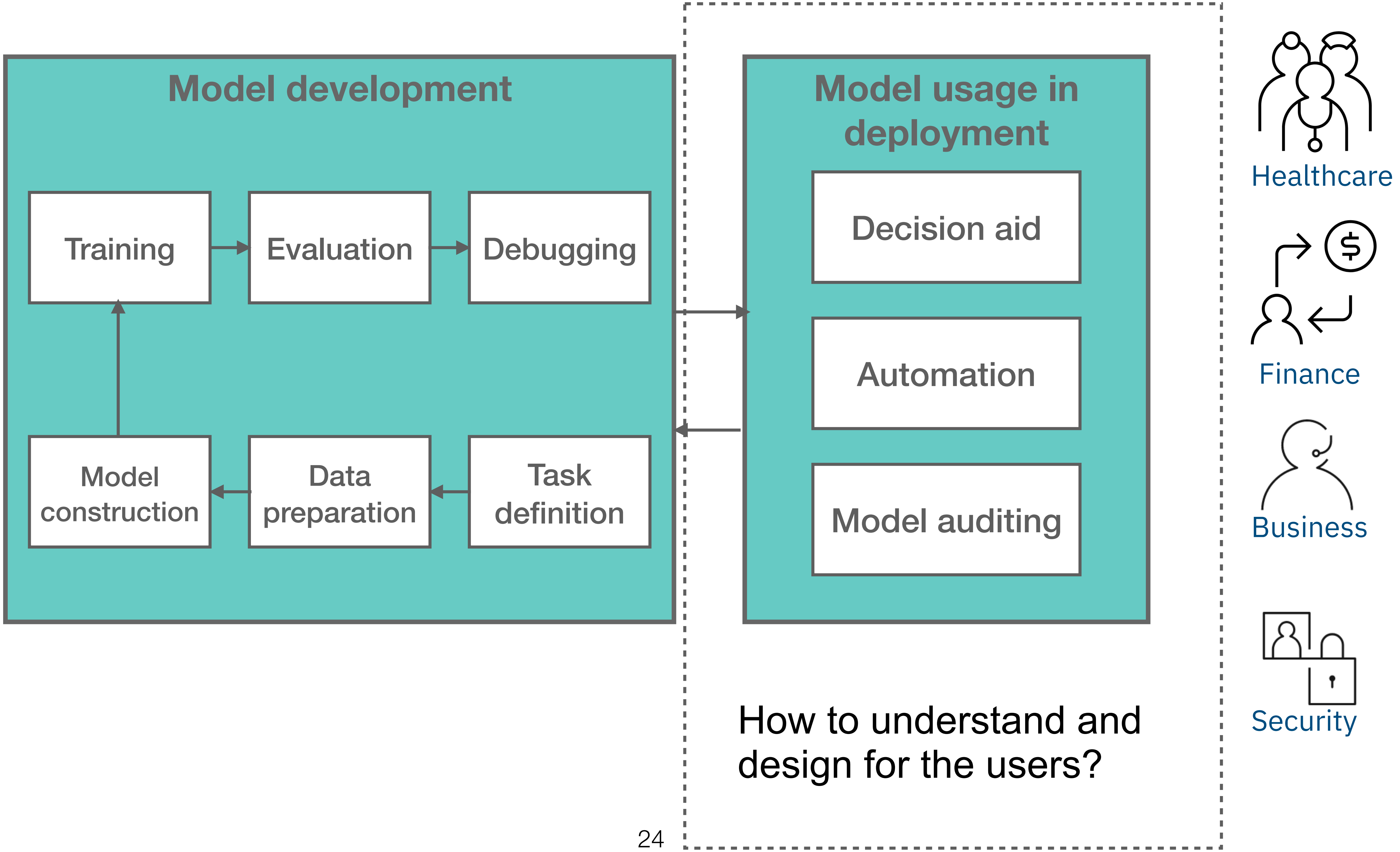
Jess W.
Sales Director (RE43)

Action: Reject Recommendation ⇔ **Outcome:** Sale
Comment: Covid-19 pandemic mode; cannot lose long-term profitable customer; offered 10% below cost price
May 6, 2020

5

Ehsan et al. Expanding Explainability: Towards Social Transparency in AI systems. To appear in CHI 2021

Why break-downs in model usage?



From XAI algorithms to XAI UX

With a toolbox:

How to **select**?

How to **translate**?

How to **expand**?

Our paths:

- Develop context-specific design guidelines: **HCI research with XAI use cases**
- Tackle the design process: **User centered design of XAI**
- Socially situated explainability by making visible the social contexts

Where we started: Research into **XAI Design Practices**

Why AI design practitioners?

- Bridging roles connecting user needs and XAI techniques

Research questions:

- What is the design space of XAI UX?
- What are the design challenges?

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

+

Model facts: **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 was the system trained on?**
- What is the source of the training data?
- How were the labels/ground-truth produced?
- What is the sample size of the training data?
- What dataset(s) is the system NOT using?
- What are the limitations/biases of the data?
- What is the size, proportion, or distribution of the training data with given feature(s)/feature-value(s)?

Output

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

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 kind of rules does it follow?
  - How does [feature X] impact its predictions?
  - What are the top rules/features that determine its predictions?
- What kind of algorithm is used?
  - How were the parameters set?

Why

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

Why not

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

How to be that

- **How should this instance change to get a different prediction Q?**
- What is the minimum change required for this instance to get a different prediction Q?
- How should a given feature change for this instance to get a different prediction Q?
- What kind of instance is predicted of [a different outcome Q]?

How to still be this

- **What is the scope of change permitted for this instance to still get the same prediction?**
- What is the range of value permitted for a given feature for this prediction to stay the same?
- What is the necessary feature(s)/feature-value(s) present or absent to guarantee this prediction?
- What kind of instance gets the same prediction?

What If

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

Others

- How/why will the system change/adapt/improve/drift over time? (change)
- Can I, and if so, how do I, improve the system? (change)
- Why is the system using or not using a given feature/rule/data? (follow-up)
- What does [a machine learning terminology] mean? (terminological)
- What are the results of other people using the system? (social)

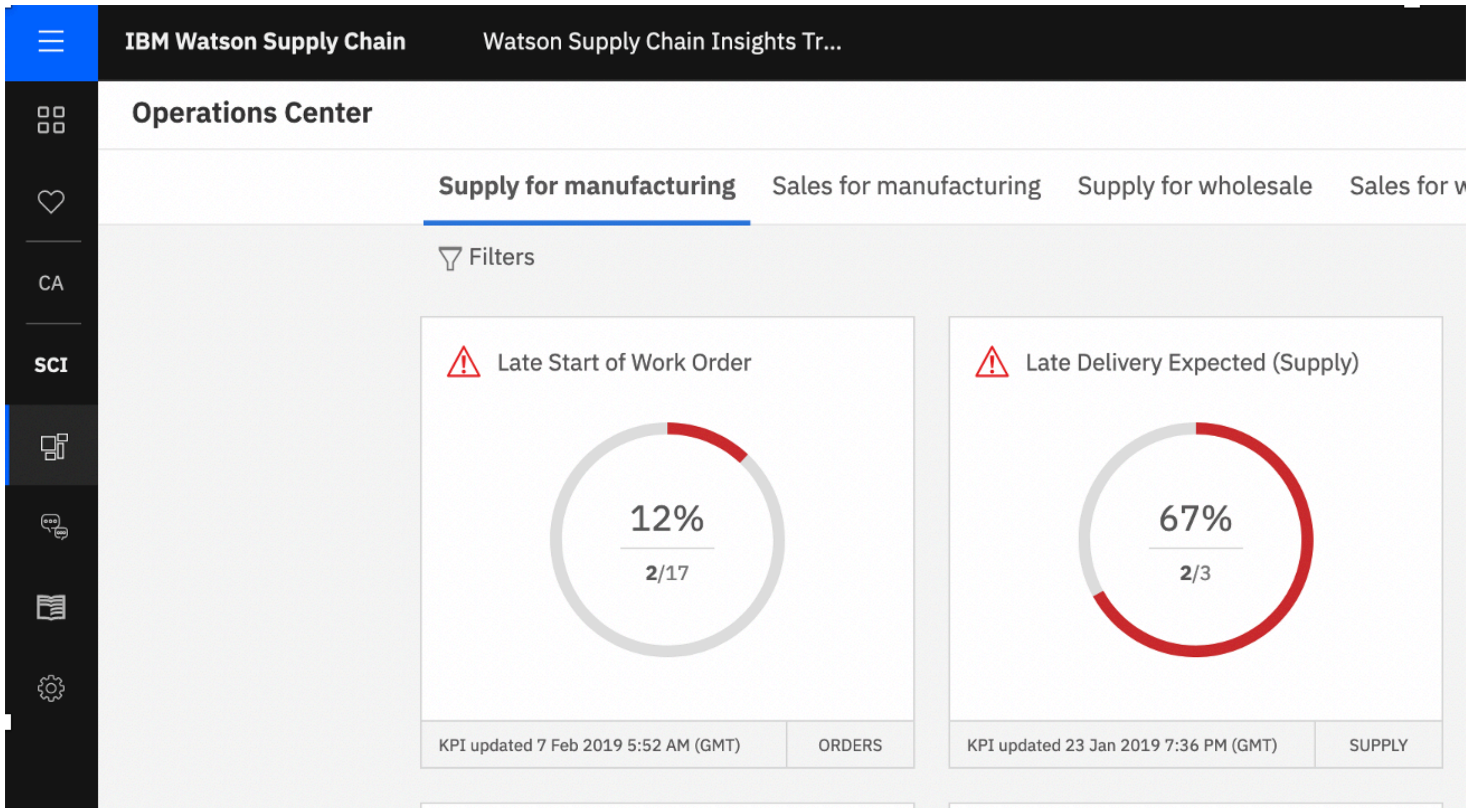
XAI design challenge 1: Variability of XAI needs

Diverse objectives of explainability

- To gain further insights for the decision
- To appropriately evaluate AI's capability
- To adapt usage or control
- To learn about a domain
- Legal or ethical requirement

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

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 control
- 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-desired explanations

Human explanations are

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

“Translation” design attempt to mimic how people, especially domain experts, explain

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

- Challenge **navigating the technical capabilities**

“ *finding the right pairing to put the ideas of what’s right for the user together with what’s doable given the tools or the algorithms*

- **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.*

XAI in Academia

Inform usage →

← Identify gaps

XAI in Practice

Opportunities for technical XAI work

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

Guidelines to address XAI user needs

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 limitations of the AI and make it actionable

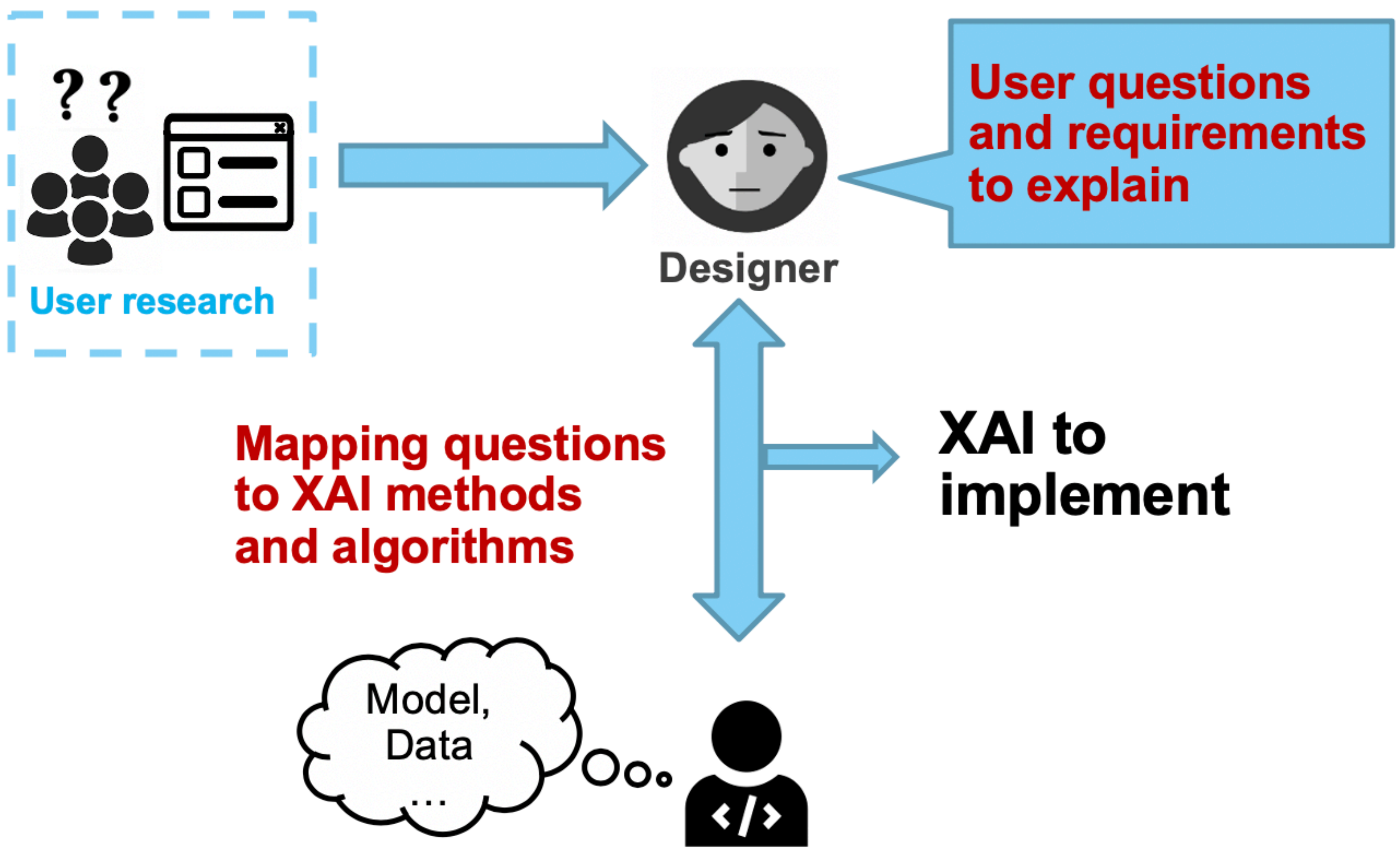
Global model: Choose appropriate level of details to explain the model

Local decision: Provide resources for "why not"

Counterfactual: Consider opportunities as utility features for analytics or exploration

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

Supporting the process: Question-driven XAI design



Design pain points to address:

- Identify application, user and interaction specific XAI needs
- Enable a "designedly" understanding of XAI by reframing the technical space
- Support designer-AI-engineer communication and collaboration

Question-Driven XAI Design

Step 1

Identify user questions

Elicit user needs for XAI as questions

Also gather user intentions and expectations for asking the questions

Designers, users

Step 2

Analyze questions

Cluster questions into categories and prioritize categories for the XAI UX to focus on

Summarize user intentions and expectations to identify key user requirements

Designers, product team

Step 3

Map questions to modeling solutions

Map prioritized question categories to candidate XAI techniques as a set of functional elements that the design should cover

A mapping guide for supervised ML is provided for reference

Designers, data scientists

Step 4

Iteratively design and evaluate

Create a design including the candidate elements identified in step 3

Iteratively valuate the design with the user requirements identified in step 2 and fill the gaps

Designers, data scientists, users

Liao et al. Question-Driven Design Process for Explainable AI User Experiences. (Under review)

XAI Question Bank

Data

- **What kind of data was the system trained on?**
- What is the source of the training data?
- How were the labels/ground-truth produced?
- What is the sample size of the training data?
- What dataset(s) is the system NOT using?
- What are the limitations/biases of the data?
- What is the size, proportion, or distribution of the training data with given feature(s)/feature-value(s)?

Output

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

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 kind of rules does it follow?
    - How does [feature X] impact its predictions?
    - What are the top rules/features that determine its predictions?
- What kind of algorithm is used?
    - How were the parameters set?

Why

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

Why not

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

How to be that

- **How should this instance change to get a different prediction Q?**
- What is the minimum change required for this instance to get a different prediction Q?
- How should a given feature change for this instance to get a different prediction Q?
- What kind of instance is predicted of [a different outcome Q]?

How to still be this

- **What is the scope of change permitted for this instance to still get the same prediction?**
- What is the range of value permitted for a given feature for this prediction to stay the same?
- What is the necessary feature(s)/feature-value(s) present or absent to guarantee this prediction?
- What kind of instance gets the same prediction?

What If

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

Others

- How/why will the system change/adapt/improve/drift over time? (change)
- Can I, and if so, how do I, improve the system? (change)
- Why is the system using or not using a given feature/rule/data? (follow-up)
- What does [a machine learning terminology] mean? (terminological)
- What are the results of other people using the system? (social)

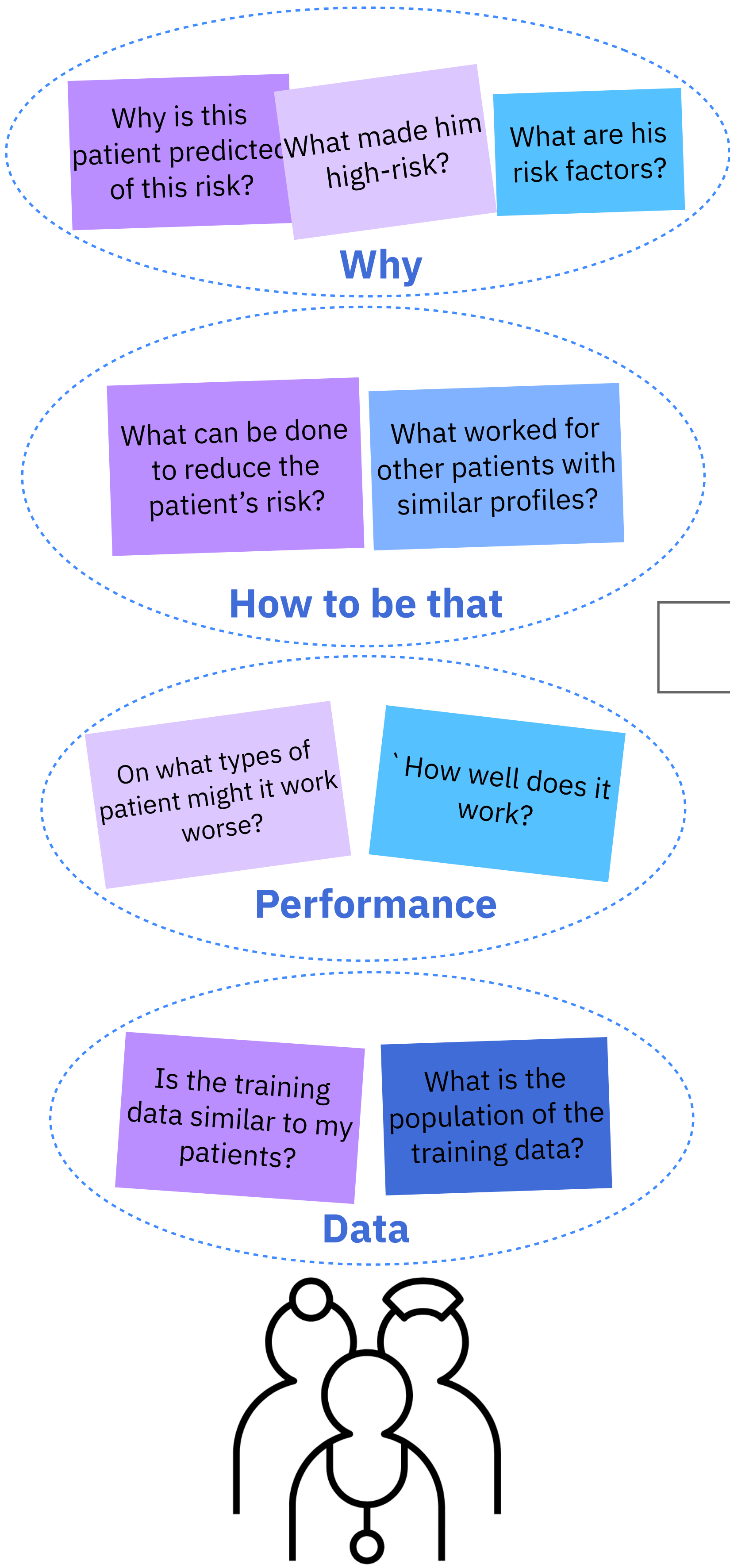
How to select: identify user needs for XAI as ***questions***

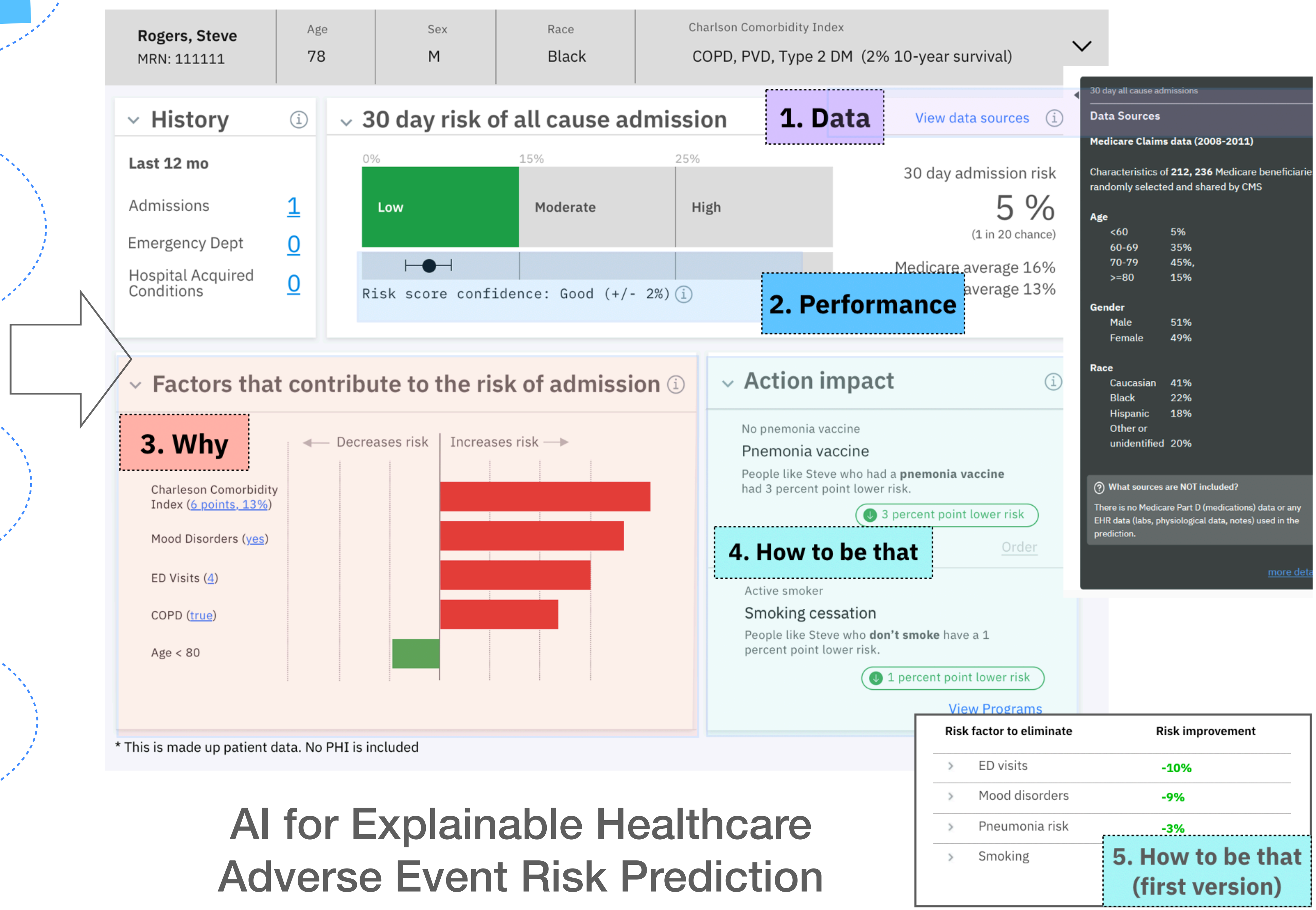
Liao et al. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. CHI 2020 🎖

Question	Explanations	Example XAI techniques
Global how	• Describe what algorithm is used and what features are considered, if a user is only interested in a high-level view • Describe the general model logic as feature impact*, rules✢ or decision-trees• (sometimes need to explain with a surrogate simple model)	ProfWeight*, Feature Importance*, PDP*, BRCG✢ , GLRM✢ , Rule List✢ , DT Surrogate•
Why	• Describe what key features of the particular instance determine the model's prediction of it* • Describe rules✢ that the instance fits to guarantee the prediction • Show similar examples• with the same predicted outcome to justify the model's prediction	LIME*, SHAP*, LOCO*, Anchors✢, ProtoDash•
Why not	• Describe what changes are required for the instance to get the alternative prediction and/or what features of the instance guarantee the current prediction* • Show prototypical examples✢ that had the alternative outcome	CEM* , Prototype counterfactual✢ , ProtoDash✢ (on alternative class)
How to be that	• Highlight features that if changed (increased, decreased, absent, or present) could alter the prediction* • Show examples with small differences but had a different outcome than the prediction✢	CEM*, Counterfactuals*, DiCE✢
What if	• Show how the prediction changes corresponding to the inquired change	PDP, ALE, What-if Tool
How to still be this	• Describe feature ranges* or rules✢ that could guarantee the same prediction • Show examples that are different from the particular instance but still had the same outcome	CEM*, Anchors✢
Performance	• Provide performance metrics of the model • Show confidence information for each prediction • Describe potential strengths and limitations of the model	Precision, Recall, Accuracy, F1, AUC Confidence FactSheets, Model Cards
Data	• Document comprehensive information about the training data, including the source, provenance, type, size, coverage of population, potential biases, etc.	FactSheets, DataSheets
Output	• Describe the scope of output or system functions • Suggest how the output should be used for downstream tasks or user workflow	FactSheets, Model Cards

How to translate: support collaborative problem-solving between data scientists and designers with "***boundary objects***"

Liao et al. Question-Driven Design Process for Explainable AI User Experiences. (Under review)





AI for Explainable Healthcare Adverse Event Risk Prediction

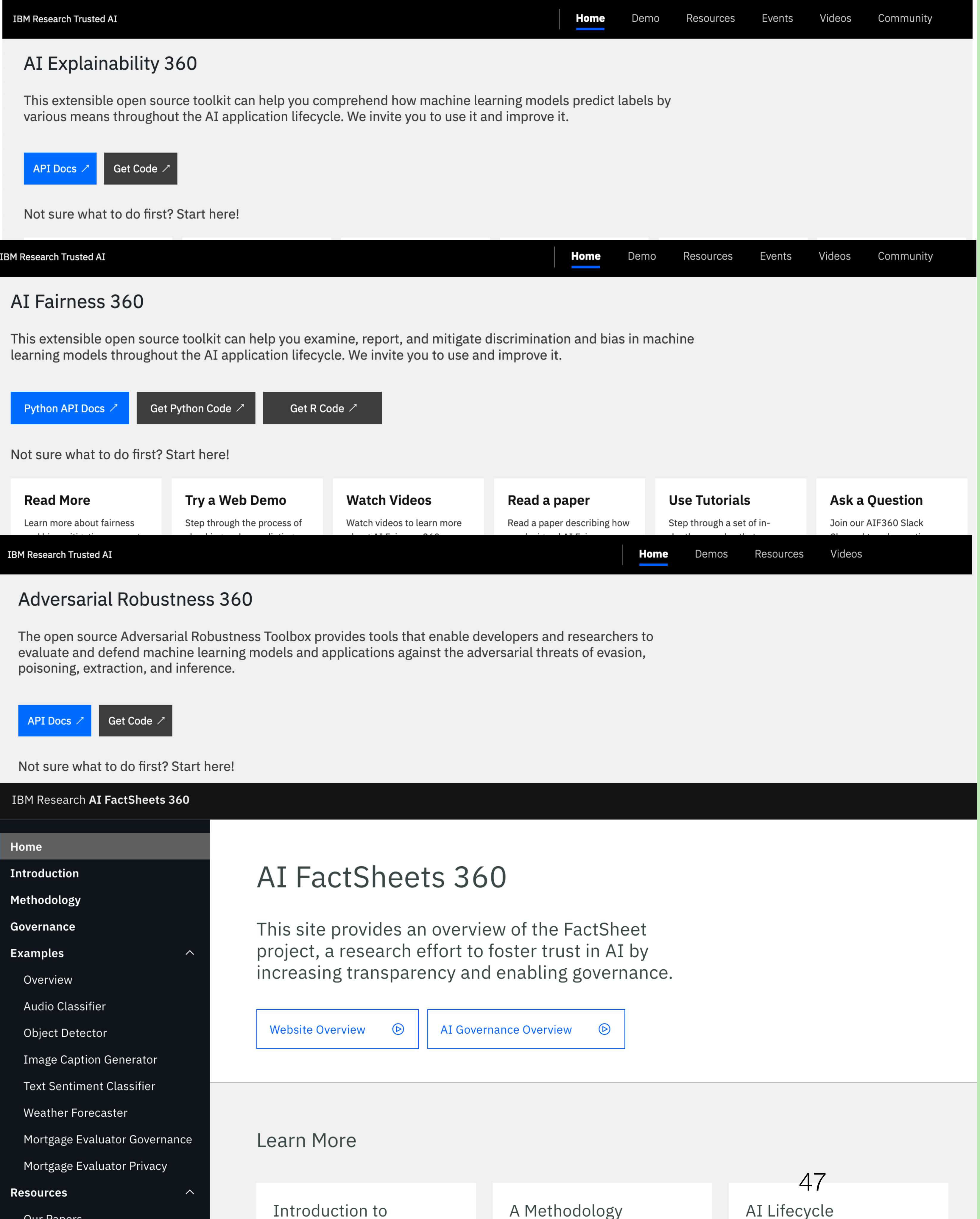
Liao et al. Question-Driven Design Process for Explainable AI User Experiences. (Under review)

Conclusions: **Bridging** work

- **Human-centered** re-framing of technical spaces
  - Contextualize the tools by the human needs, values, and conditions they serve
  - Thinking “outside the toolbox”
- **Responsible** understanding and and use of the toolbox
  - Examine breakdowns, limitations and potential harm
  - User-centered design vision drives technical development
- **Actionable** frameworks, design assets and methods that practitioners can readily use

From a toolbox of **AI algorithms** to a toolbox of **design materials**

From a toolbox of **AI algorithms** to a toolbox of **design materials**



Thank **YOU!**

…and thanks to

Rachel Bellamy, Amit Dhurandhar, Jonathan Dodge, Casey Dugan, Upol Ehsan, Bhavya Ghai, Werner Geyer, Daniel Gruen, Jaesik Han, Michael Hind, Stephanie Houde, David Millen, David Piorkowski, Aleksandra Mojsilović, Sarah Miller, Klaus Mueller, Michael Muller, Shweta Narkar, Milena Pribić, John Richards, Mark Riedl, Daby Sow, Chenhao Tan, Richard Tomsett, Kush Varshney, Dakuo Wang, Justin Weisz, Yunfeng Zhang

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