

# Questioning the AI: Towards **Human** **Centered Explainable AI (XAI)**

Research work 2018-2021

Q. Vera Liao  
IBM **Research**



IBM Research Trusted AI

HomeDemoResourcesEventsVideosCommunity

AI Explainability 360

This extensible open source toolkit can help you comprehend how machine learning models predict labels by various means throughout the AI application lifecycle. We invite you to use it and improve it.

API Docs ↗Get Code ↗

Not sure what to do first? Start here!

Read MoreTry a Web DemoWatch VideosRead a PaperUse TutorialsAsk a Question

IBM Research Trusted AI

HomeDemoResourcesEventsVideosCommunity

AI Fairness 360

This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. We invite you to use and improve it.

Python API Docs ↗Get Python Code ↗Get R Code ↗

IBM Research Trusted AI

HomeDemosResourcesVideos

Adversarial Robustness 360

The open source Adversarial Robustness Toolbox provides tools that enable developers and researchers to evaluate and defend machine learning models and applications against the adversarial threats of evasion, poisoning, extraction, and inference.

API Docs ↗Get Code ↗

IBM Research AI FactSheets 360

Home

Introduction

Methodology

Governance

Examples ^

Overview

Audio Classifier

Object Detector

AI FactSheets 360

This site provides an overview of the FactSheet project, a research effort to foster trust in AI by increasing transparency and enabling governance.

Website Overview ⓘAI Governance Overview ⓘ

IBM Research Uncertainty Quantification 360

Home

Overview

Demo

Resources ^

Guidance

Communicate Uncertainty

Glossary

Uncertainty Quantification 360

Uncertainty quantification (UQ) gives AI the ability to express that it is unsure, adding critical transparency for the safe deployment and use of AI. This extensible open source toolkit can help you estimate, communicate and use uncertainty in machine learning model predictions through an AI application lifecycle. We invite you to use it and improve it.

HCI research as **bridging work**: From toolboxes of AI algorithms to toolboxes of design materials



# 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**

# Supervised Machine Learning

## Training data set

Label:    Label:

Apple

Cake



**Features:**

Color

Shape

Smell

...

## Learning Model

(Using a ML algorithm)



New **instance**

**Prediction label:**

Cake



# Supervised Machine Learning

**Training data set**

Label:    Label:

Apple

Cake



**Features:**

Color

Shape

Smell

...

**Explaining data**

**Explaining “model facts”:**  
performance, limitations,  
output, etc.

**Learning Model**

(Using a ML algorithm)



**XAI focus: explaining  
model decision**

**Prediction label:**

Cake

**New instance**



# 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**



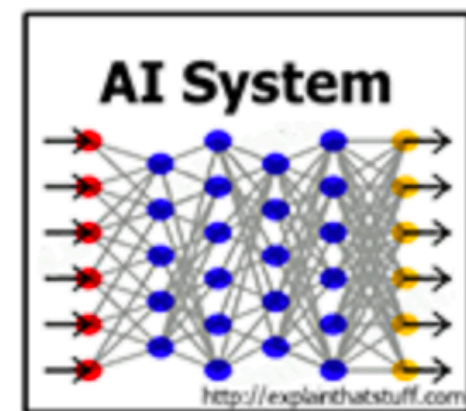
# XAI in regulation: “rights to explanation”

## 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)

GDPR, 2016

# XAI in research funding



- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

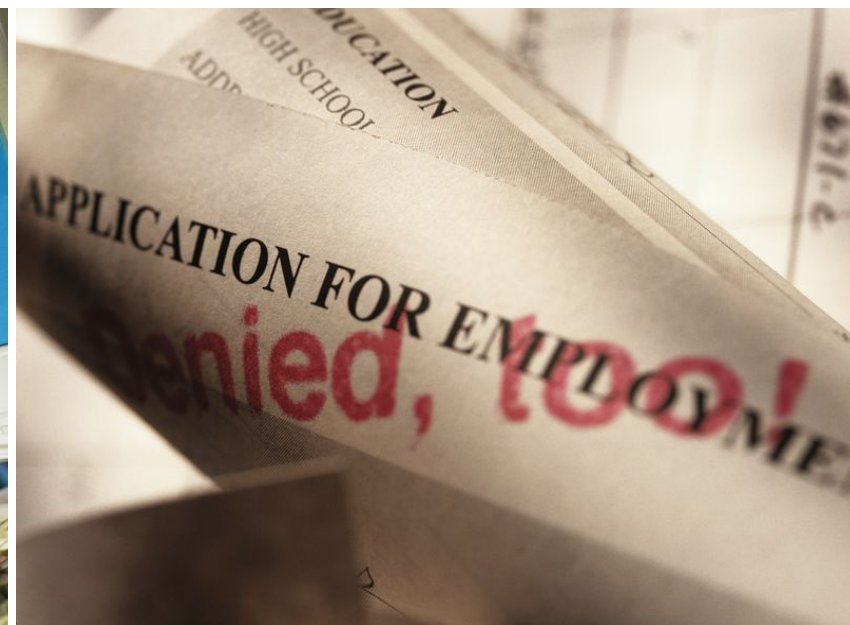


- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

**DARPA, 2016**

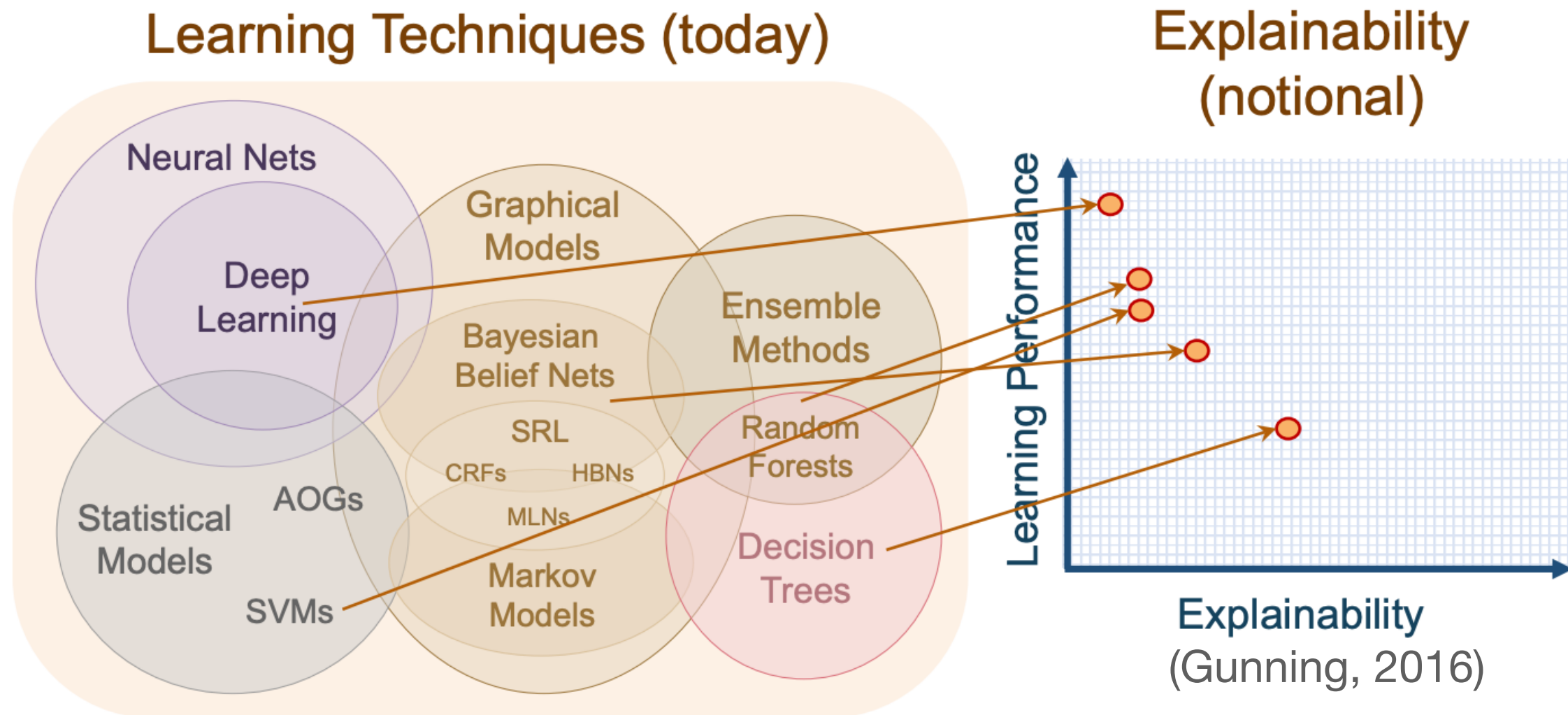


AI is increasingly used in many high-stakes tasks



# Performance-Explainability trade-off

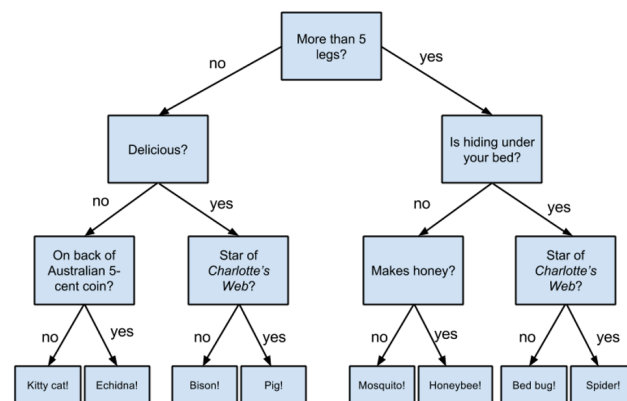
In **average** settings





# XAI

## Directly explainable model



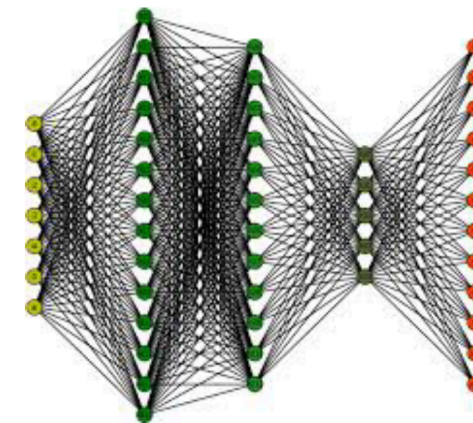
- Linear model
- Decision tree
- Rule-based model



Breaking the trade-off

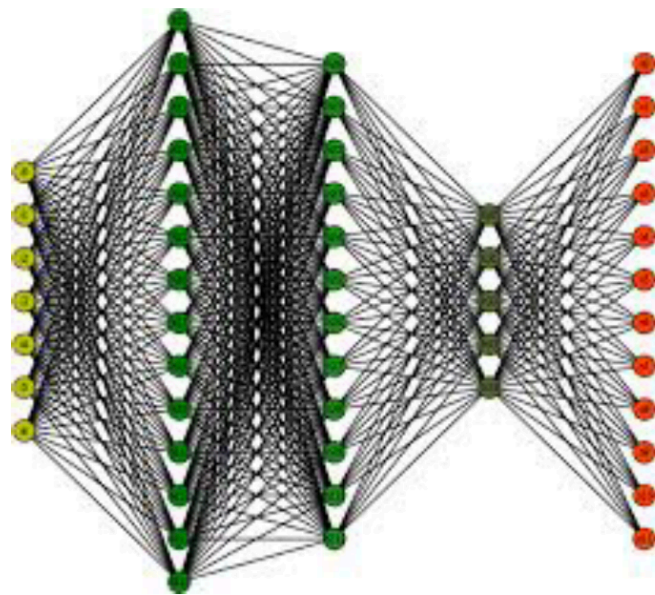
- Generalized linear rule model
- Generalized additive models
- ...

## Post-hoc explainability

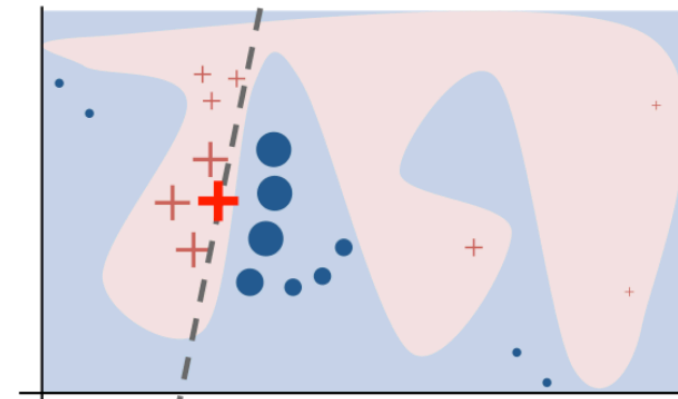
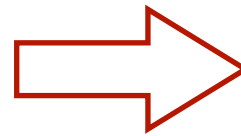


- Deep neural networks
- Ensemble models

# XAI “post-hoc”/reconstructive algorithm example: LIME

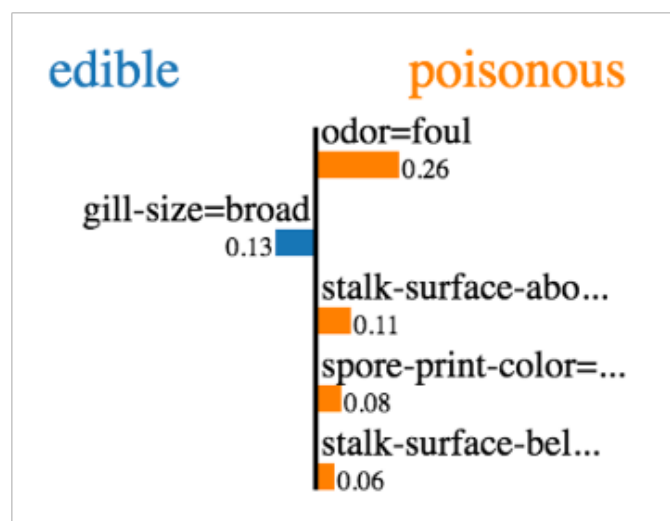
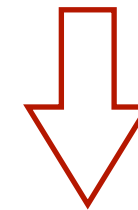


Neural network, not directly explainable



LIME (Ribeiro et al. 2016)

Use a *post-hoc* XAI technique



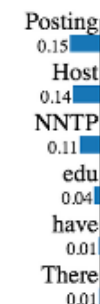
Tabular data

Images (explaining prediction of 'Cat' in pros and cons)



Image

atheism christian



Text with highlighted words

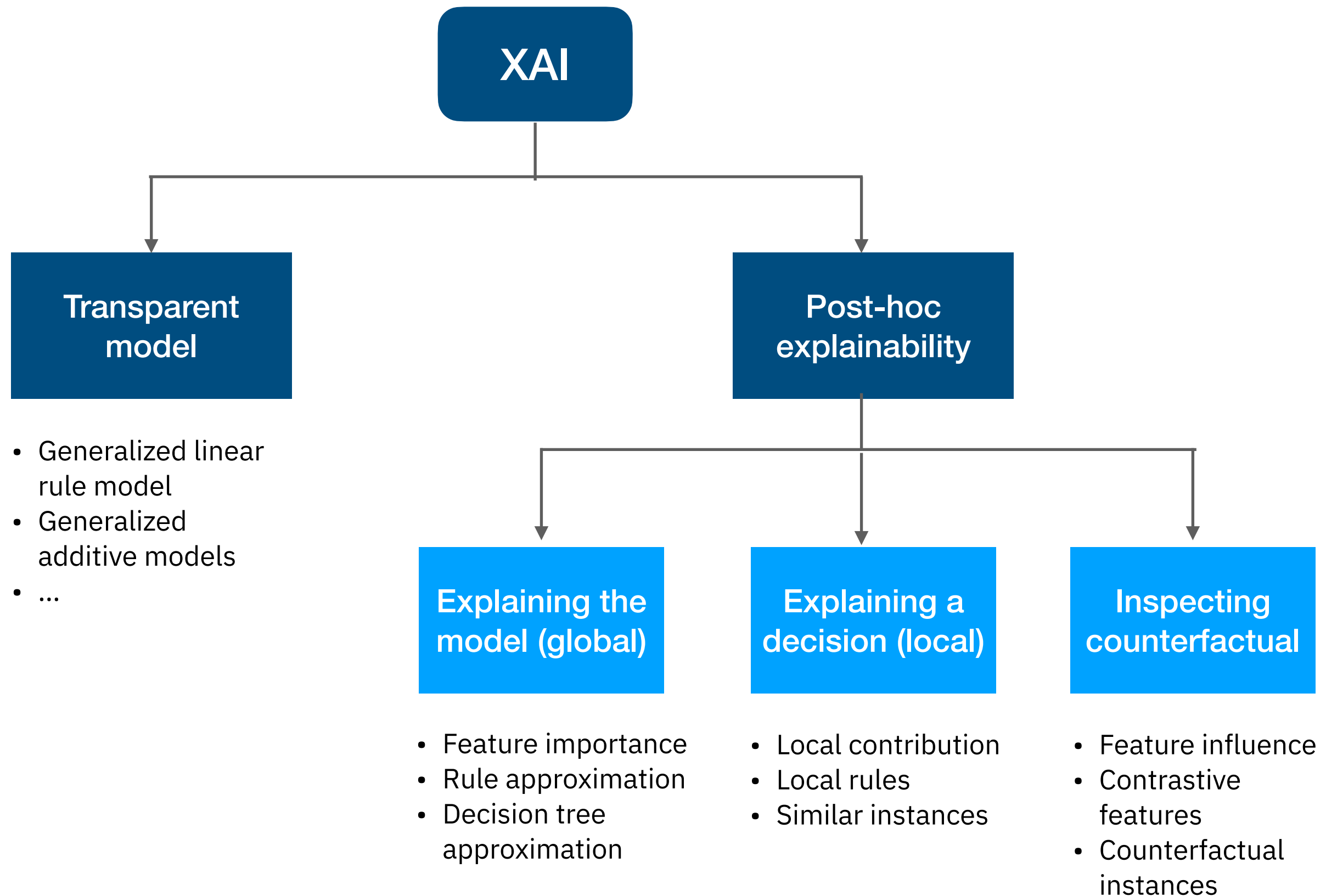
From: johnchad@triton.unm.edu (jchadwic)  
Subject: Another request for Darwin Fish  
Organization: University of New Mexico, Albuquerque  
Lines: 11  
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.  
This is the same question I have and I have not seen an answer on the net. If anyone has a contact please post on the net or email me.

Texts





Check out our **CHI2021 Course** materials, with links to AIX360 code libraries:  
<https://hcixaitutorial.github.io/>

Review

# Machine Learning Interpretability: A Survey on Methods and Metrics

Diogo V. Carvalho<sup>1,2,\*</sup>, Eduardo M. Pereira<sup>1</sup> and

<sup>1</sup> Deloitte Portugal, Manuel Bandeira Street, 43, 4150-47  
<sup>2</sup> Faculty of Engineering, University of Porto, Dr. Rober  
<sup>3</sup> 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 toward algorithmically informed decisions have greater power. Most of these accurate decision support systems remain logic and inner workings are hidden to the user. 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-computer interaction, we present a categorization of

## 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, loosely defined as the science of comprehending what a model did (could). Models and learning methods include visual cues to find patterns in image recognition

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

AMINA ADADI<sup>1</sup> 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 growing collection of XAI techniques

## A Survey of Methods for Explaining

RICCARDO GUIDOTTI, ANNA MONREALE, SALVATORE  
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 explainability is an ethical issue. The literature reports many approaches aimed at explaining the internal logic of the model, but at the cost of sacrificing accuracy for interpretability. The approaches that can be used are various, and each approach is typically developed for a specific purpose. As a consequence, it explicitly or implicitly delineates its own scope of application. The aim of this article is to provide a classification of the methods in the respect to the notion of explanation and the type of black 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

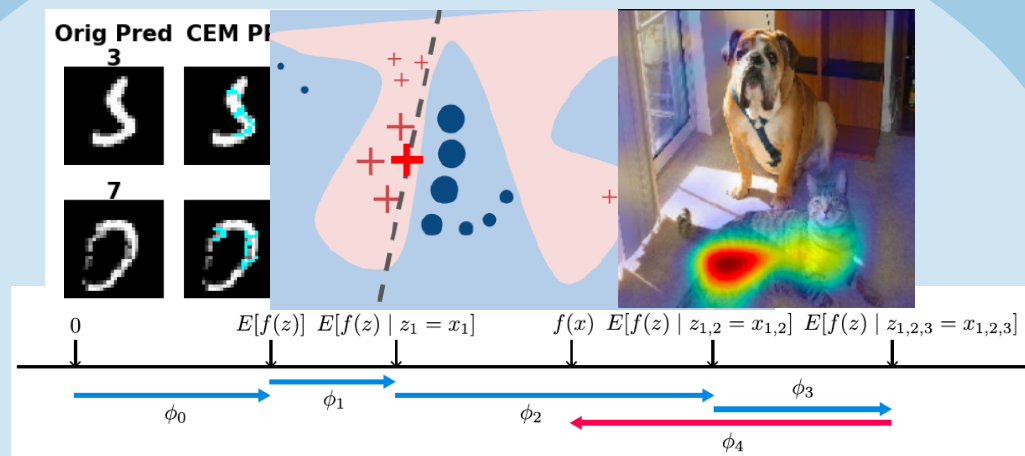
Radboud University, Donders Institute for Brain, Cognition and Behaviour,  
Nijmegen, the Netherlands  
{g.ras, m.vangerven, w.haselager}@donders.ru.nl

### 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 that explanation methods or interfaces for lay users are missing and we speculate which criteria

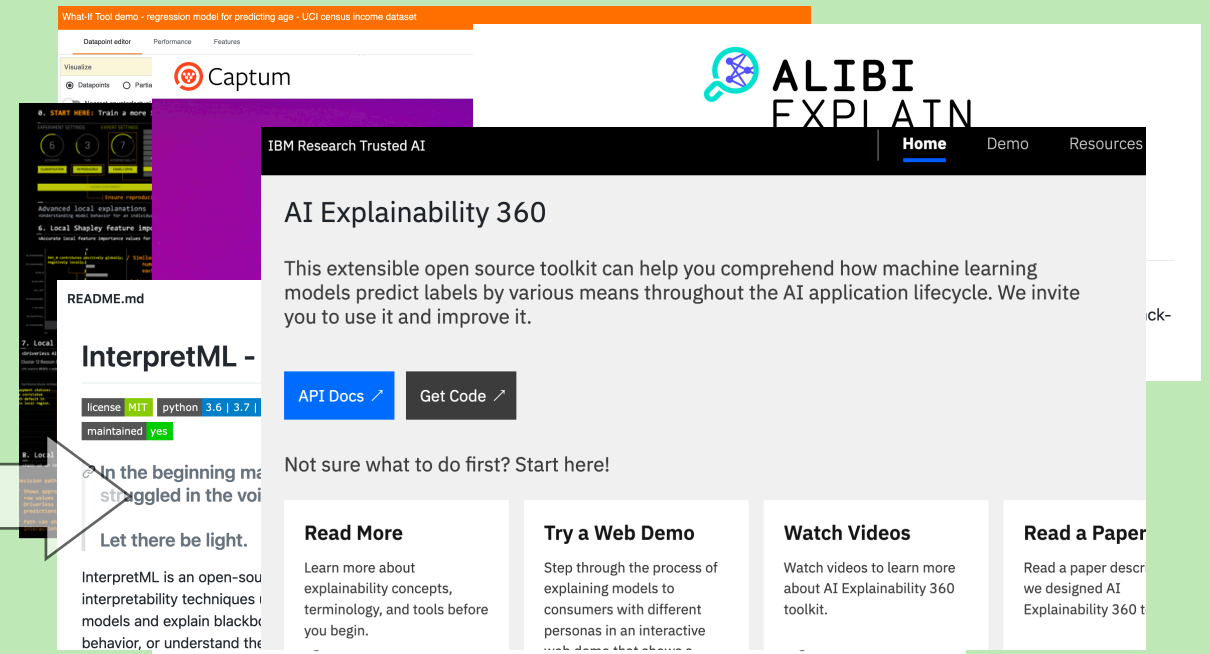
(AI) has achieved a notable momentum that, if harnessed properly, can lead to significant improvements over many application sectors across the field. For this reason, the research community stands in front of the barrier of explainability, brought by sub-symbolism (e.g. ensembles or Deep Neural Networks). The type of AI (namely, expert systems and rule based models). In the so-called *eXplainable* AI (XAI) field, which is widely used for the practical deployment of AI models. The overview presented in this paper summarizes contributions already done in the field of XAI, including a taxonomy of XAI methods. For this purpose we summarize previous efforts made to define explainability and propose a novel definition of explainable Machine Learning that takes into account a major focus on the audience for which the explainability is required. We propose and discuss about a taxonomy of recent contributions

# XAI in Academia



**An abundance of XAI algorithms**

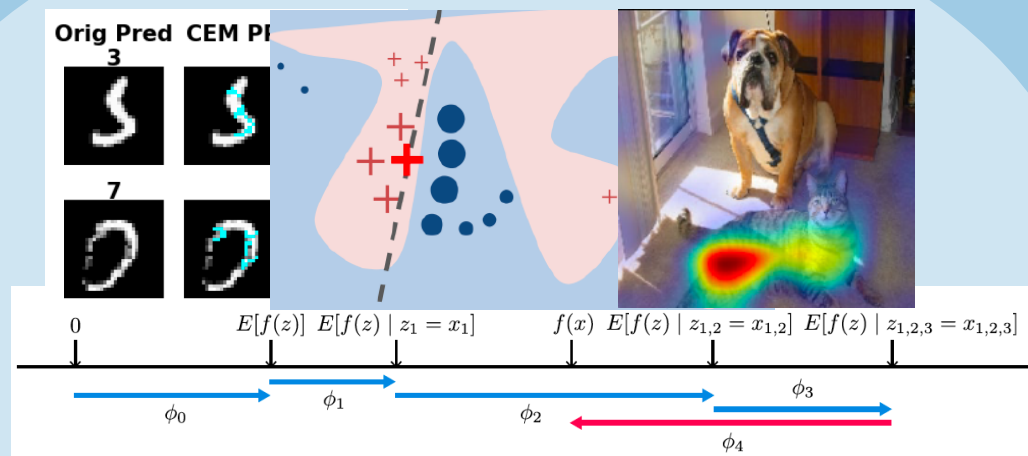
# XAI in Practice



**Toolbox of XAI techniques**

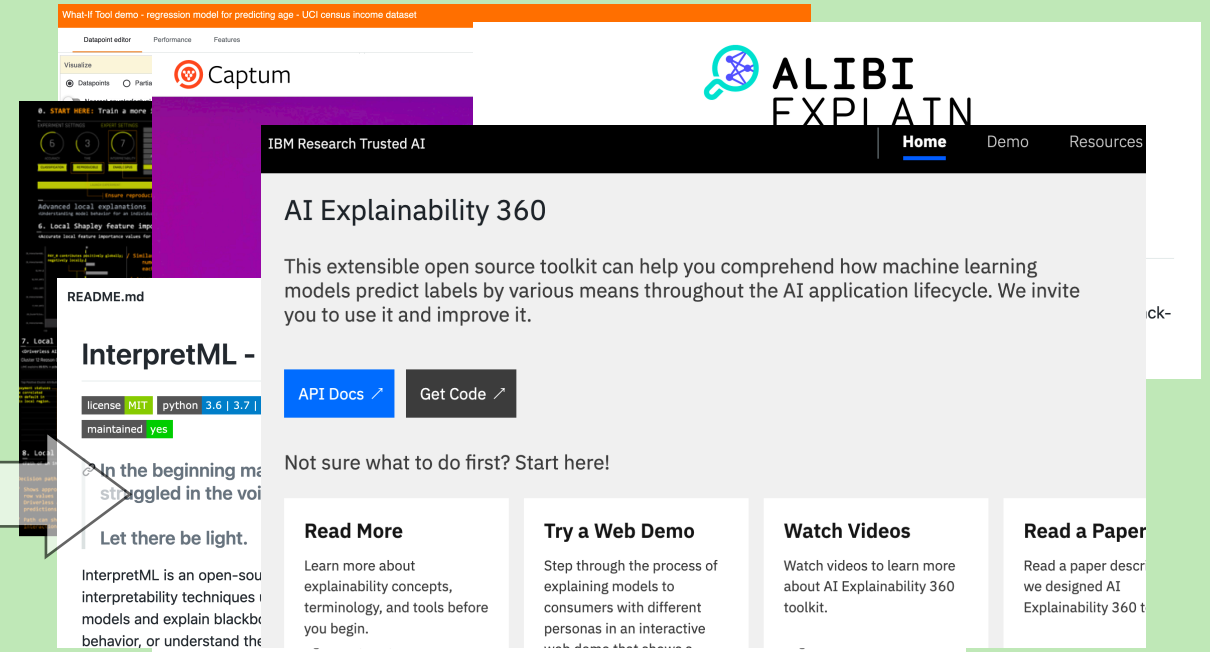
**From academic research into a practitioners' toolbox**

# XAI in Academia



**An abundance of XAI algorithms**

# XAI in Practice



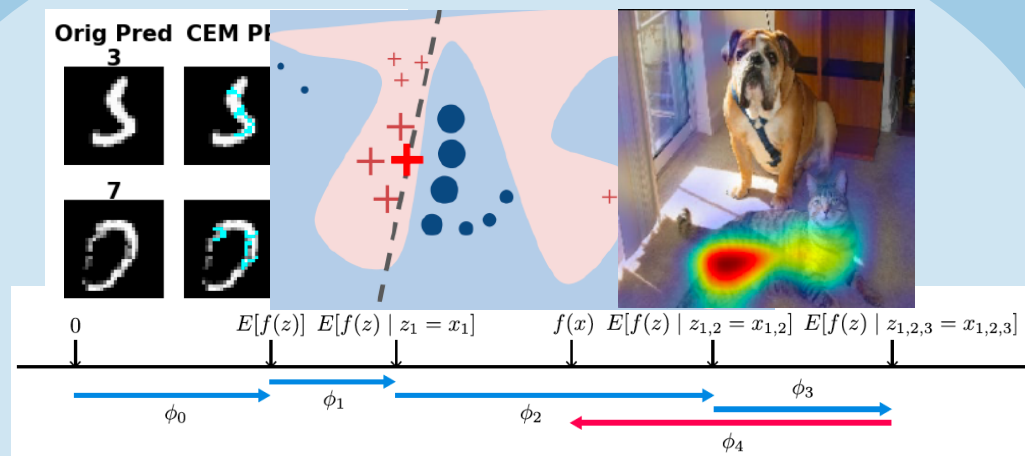
## Toolbox of XAI techniques



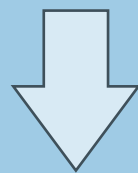
**Real-world XAI systems? Serving many domains and user groups**



# XAI in Academia



**An abundance of XAI algorithms**



Cognitive science

HCI

Social sciences

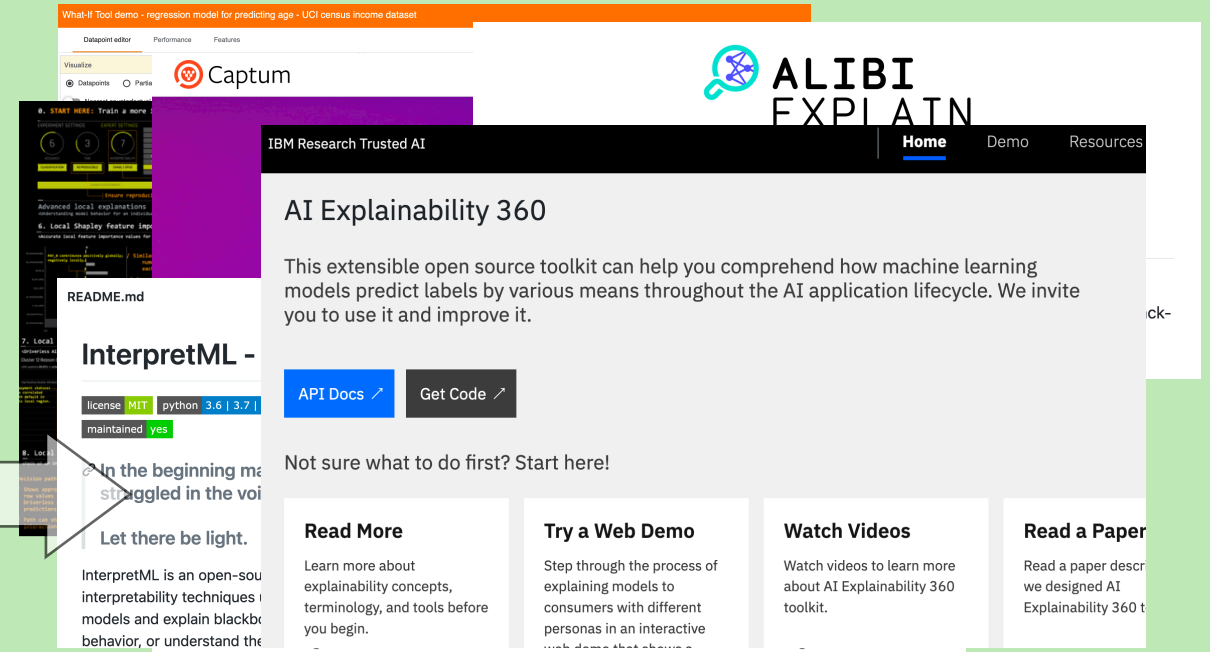
Philosophy

Law

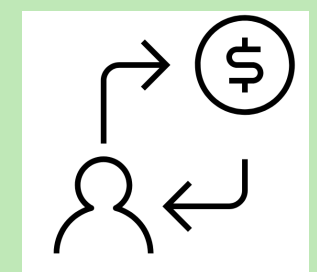
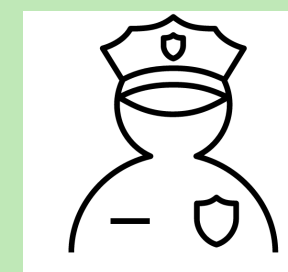
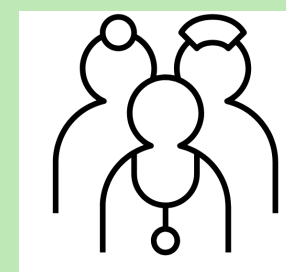
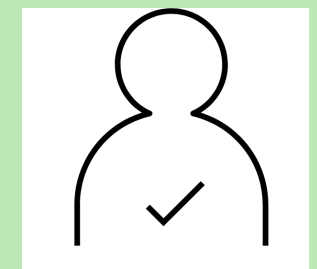
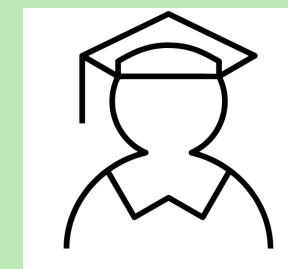
**Inter-disciplinary perspectives**



# XAI in Practice

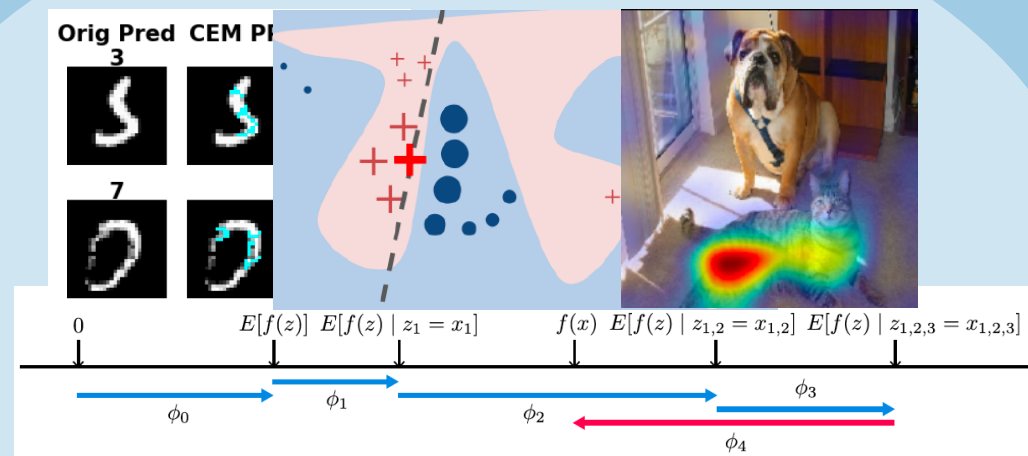


**Toolbox of XAI techniques**

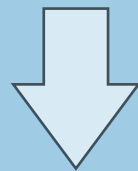


**Real-world XAI systems? Serving many domains and user groups**

# XAI in Academia



## An abundance of XAI algorithms



Cognitive  
science

HCI

Social  
sciences

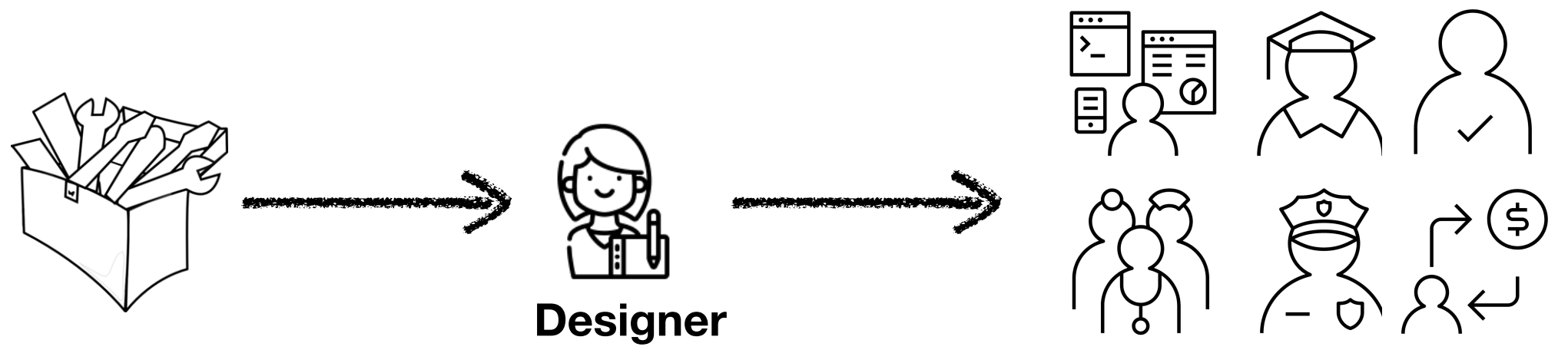
Philosophy

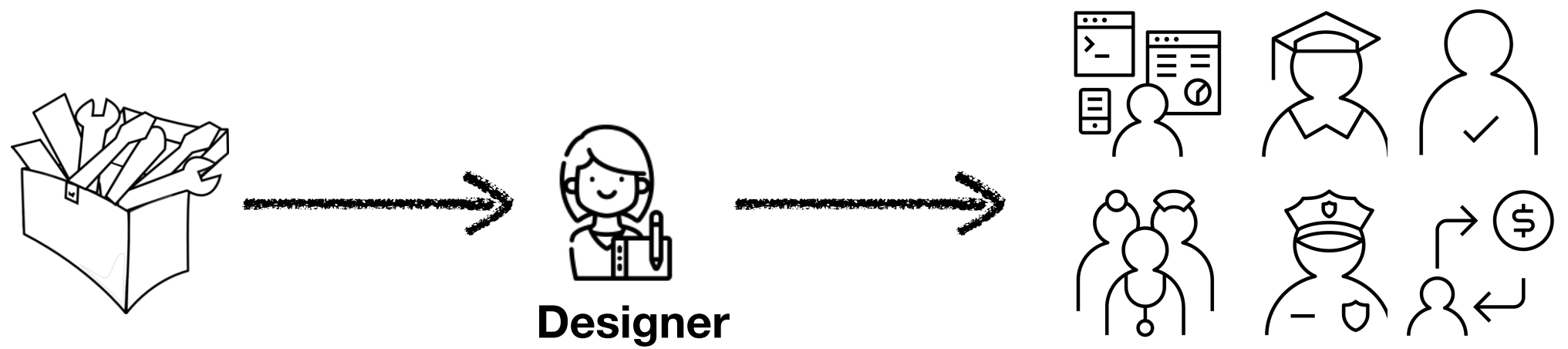
Law

## 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 psychological processes** (Petty and Cacioppo, 1986; Horne et al, 2013)
  - **Socio-technical systems** (Ehsan et al., 2021)

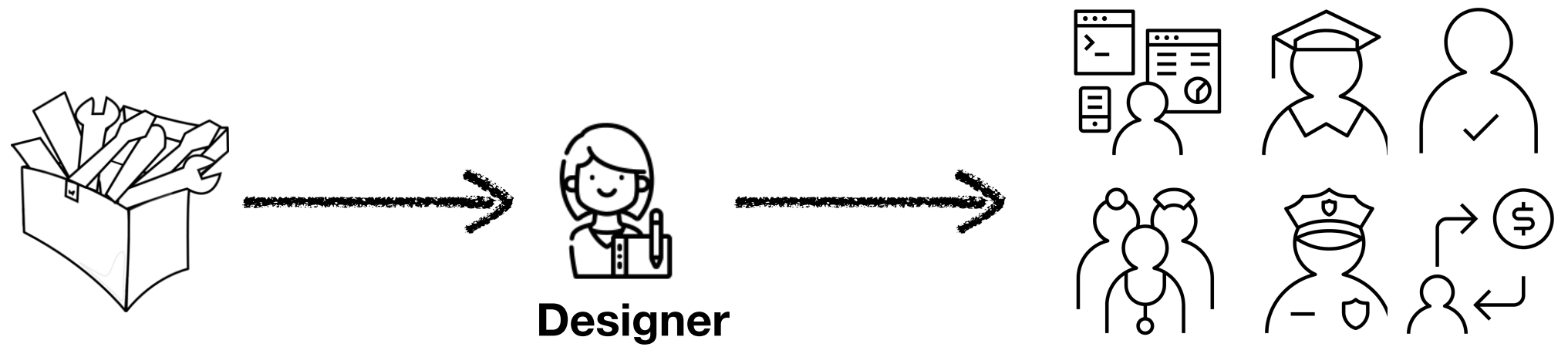




How to **select**?

How to **translate**?



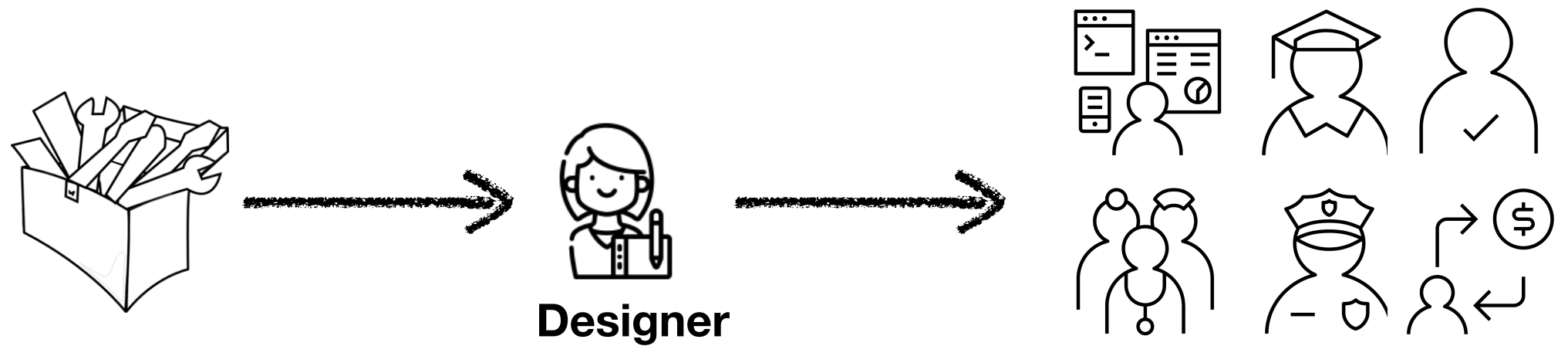


How to **select**?      How to **translate**?

Thread 1: Study and support design practices for XAI UX

---

Thread 2: HCI research with XAI use cases



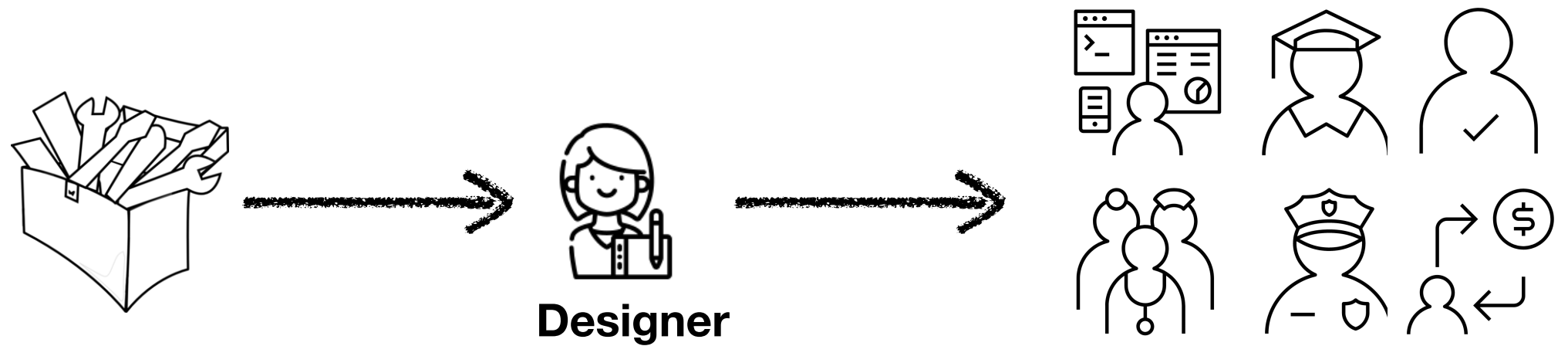
How to **select**?      How to **translate**?

Thread 1: Study and support design practices for XAI UX

---

Thread 2: HCI research with XAI use cases

**Suitability** for different  
usage contexts (*What  
contexts?*)



How to **select**?      How to **translate**?

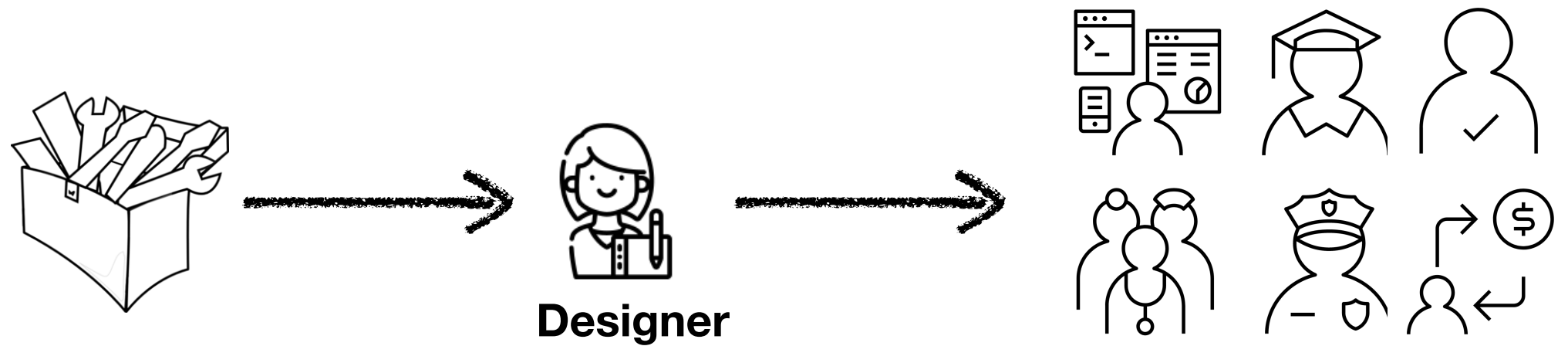
Thread 1: Study and support design practices for XAI UX

---

Thread 2: HCI research with XAI use cases

**Suitability** for different  
usage contexts (*What  
contexts?*)

Where are the  
limitations and  
**breakdowns**?



How to **select**?

How to **translate**?

Thread 1: Study and support design practices for XAI UX

---

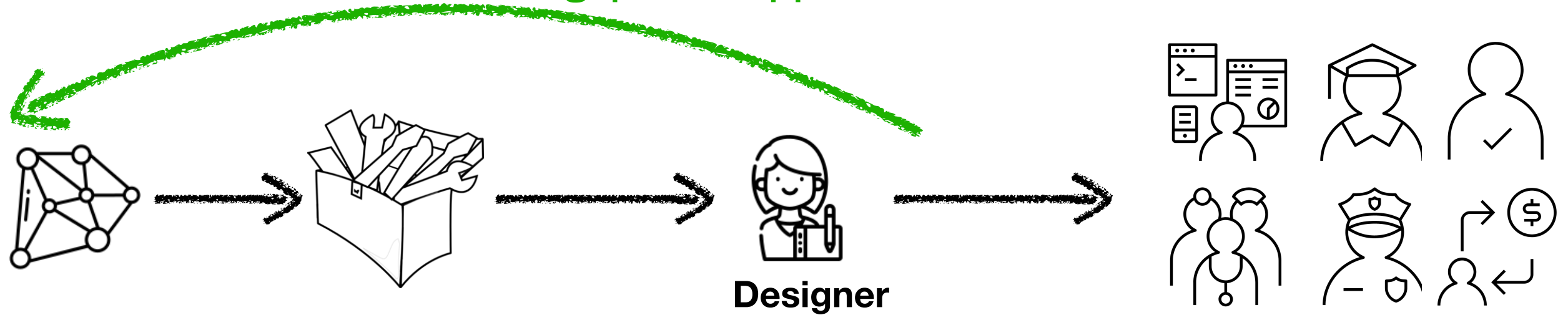
Thread 2: HCI research with XAI use cases

**Suitability** for different  
usage contexts (*What  
contexts?*)

Where are the  
limitations and  
**breakdowns**?

What's **beyond the  
toolbox** to achieve  
understanding?

## Contextualize XAI algorithms Inform gaps and opportunities



How to **select**?

How to **translate**?

Thread 1: Study and support design practices for XAI UX

---

Thread 2: HCI research with XAI use cases

**Suitability** for different  
usage contexts (*What  
contexts?*)

Where are the  
limitations and  
**breakdowns**?

What's **beyond the  
toolbox** to achieve  
understanding?

# Thread: HCI Research with XAI Use Cases

I will discuss

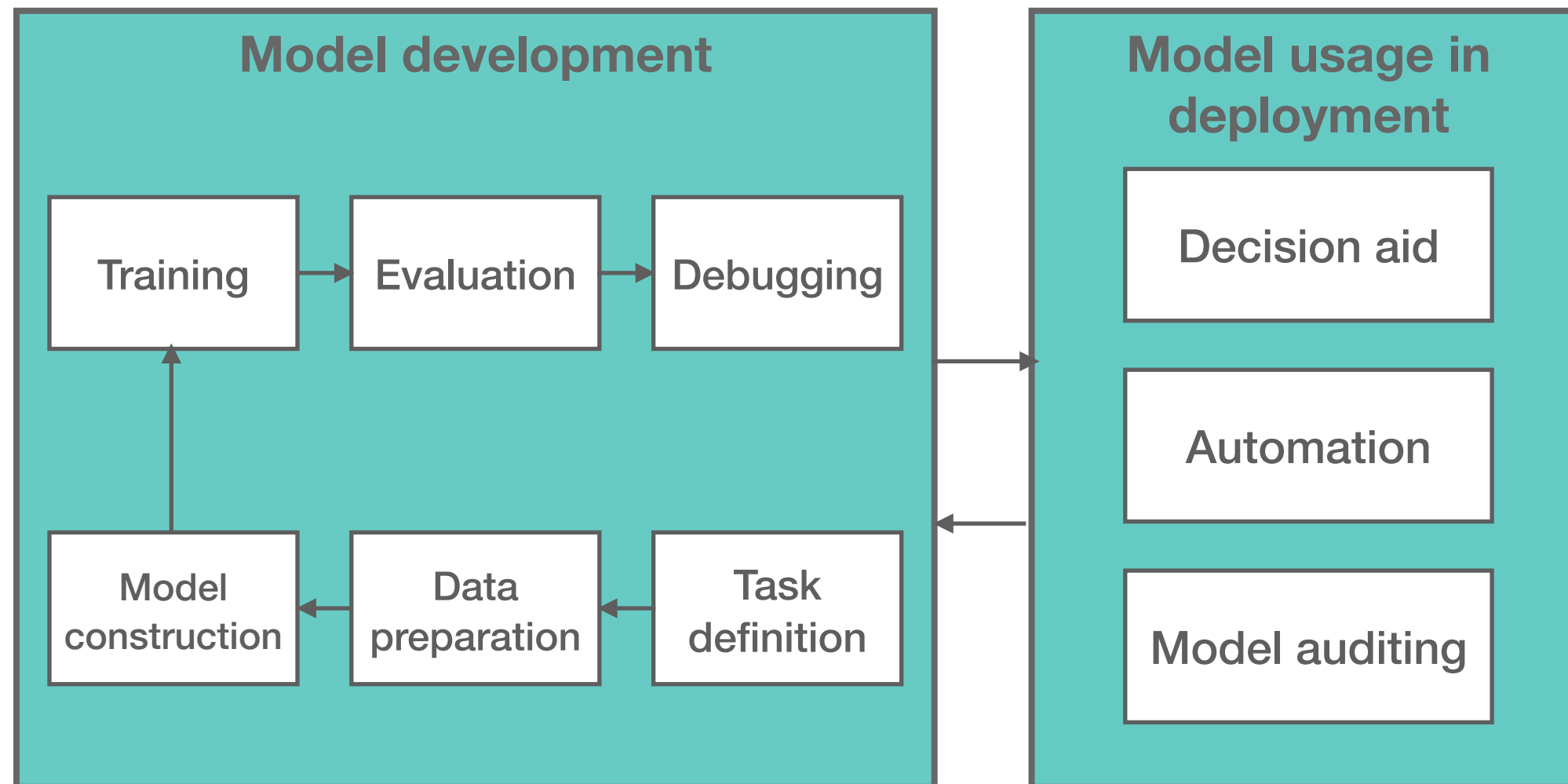
- What use cases
- Why these use cases
- What I have learned

I might not delve into:

- Explanation details
- Research design and results

But please interrupt if you are curious!

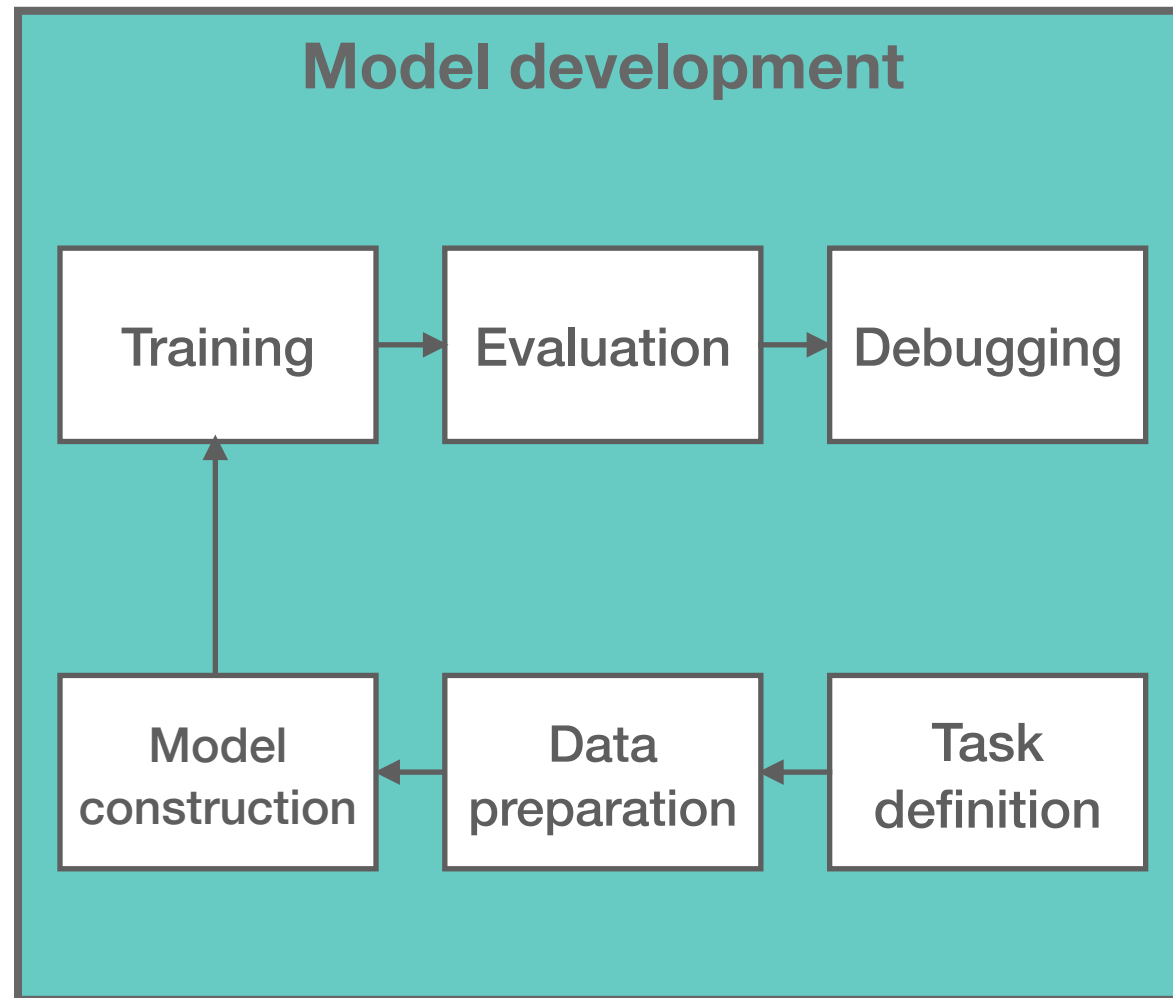
# XAI use cases in AI lifecycle



# XAI use cases in AI lifecycle

**Model debugging or selection** (IUI2021)

XAI user: **Data scientist**



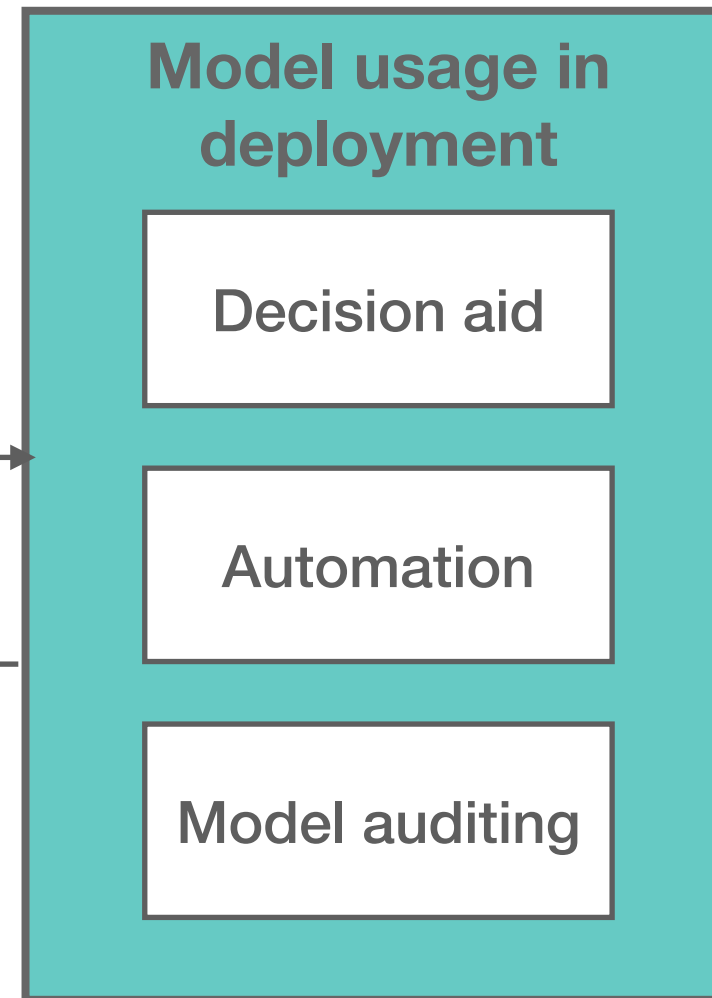
**Explainable active learning** (CSCW 2020)

XAI user: **Annotator (domain expert)**

**Trust calibration and decision**

**support** (FAT\* 2020, CHI 2021 🏆)

XAI user: **Decision-maker**



**Delegation support**

(ongoing)

XAI consumer:

**Domain expert**

**Fairness assessment** (IUI 2019 🏆)

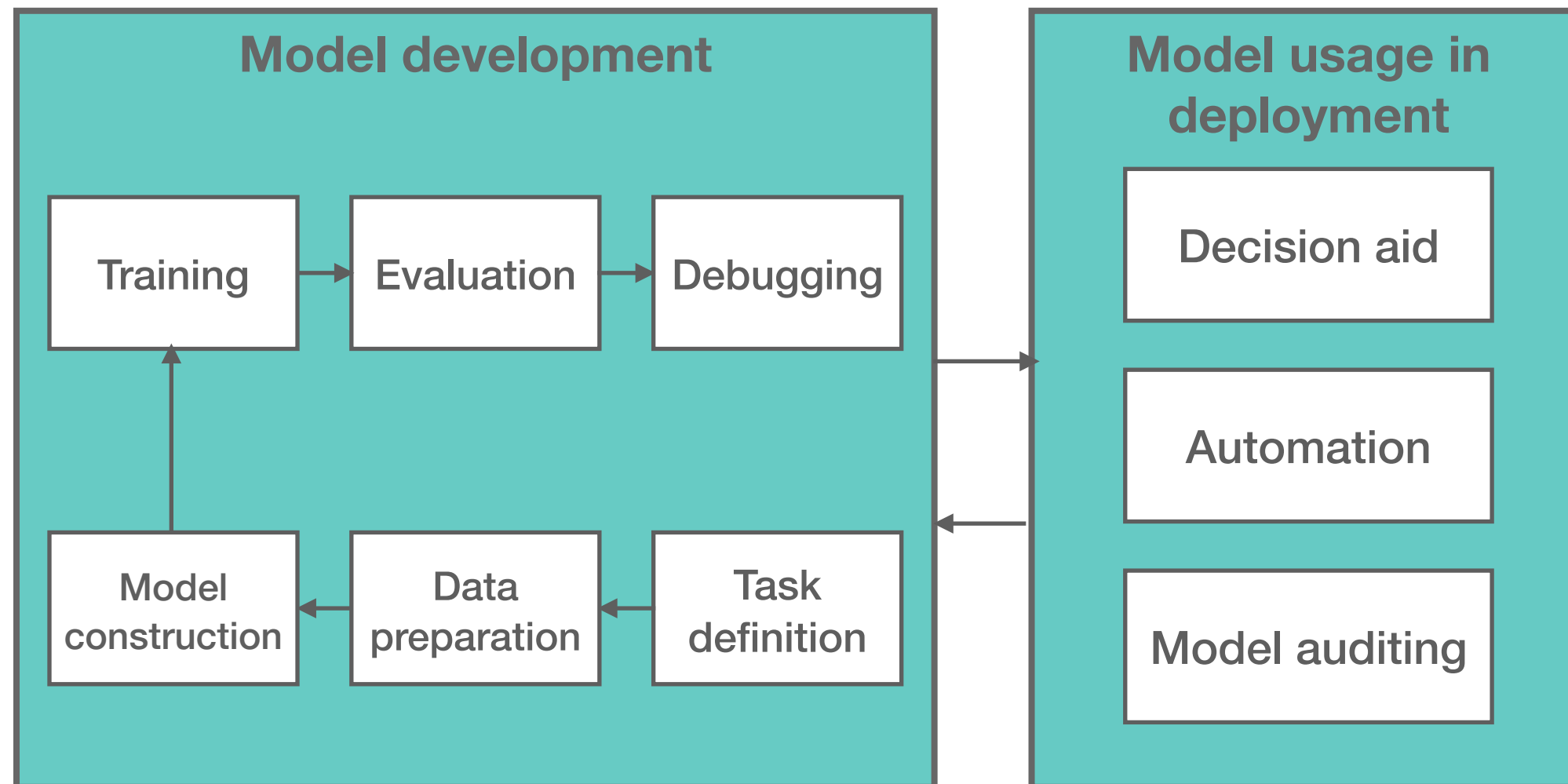
XAI user: **Regulator, impacted groups**



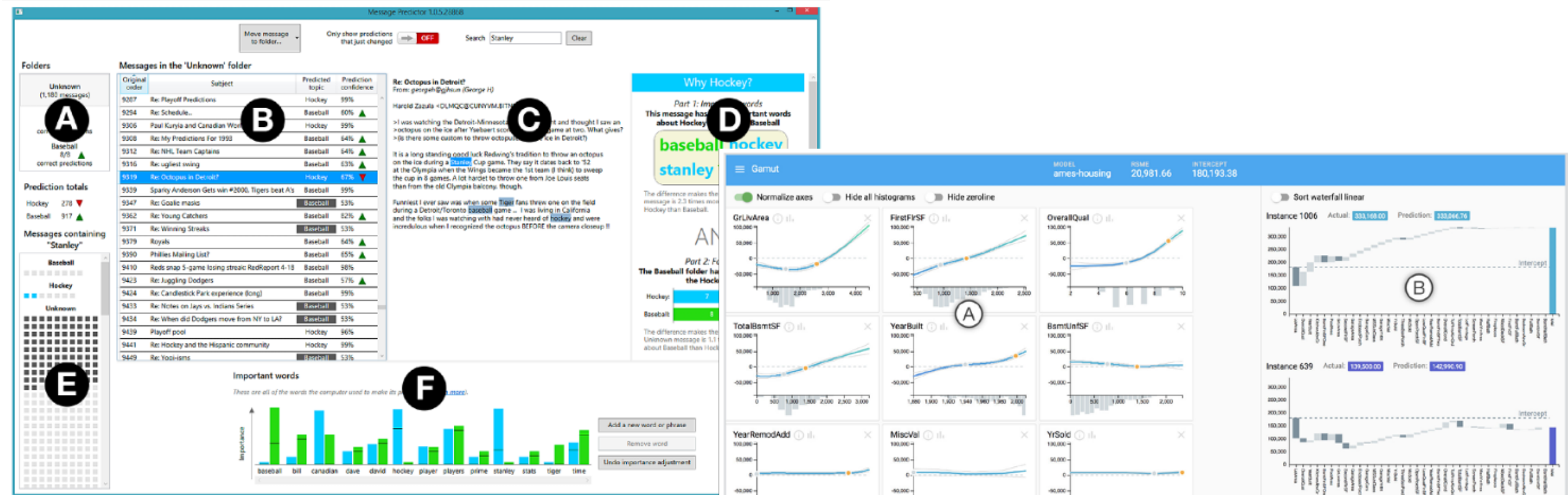
# XAI use cases in AI lifecycle

**Model debugging or selection** (IUI2021)

XAI user: **Data scientist**



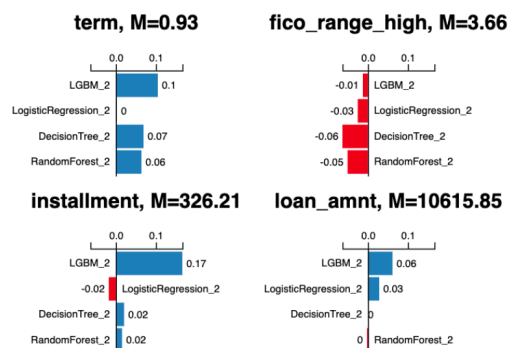
# XAI for model debugging and selection



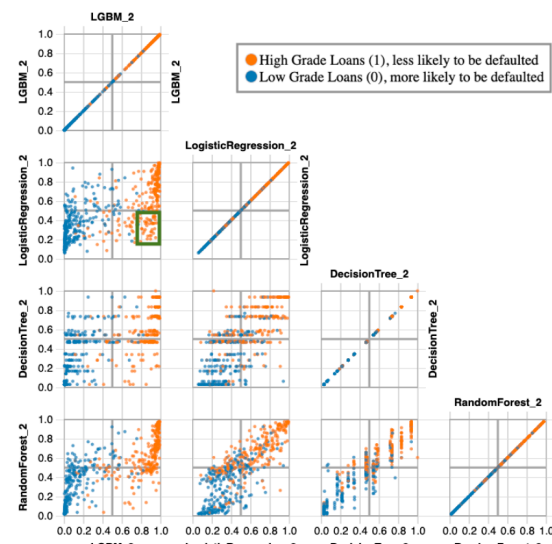
Explanatory debugging  
(Kulesza et al, 2015)

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

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



(b) Partial screenshot of the Feature Importance Comparison View showing 4 of 21 FI plots.



(c) Screenshot of the Probability Scatterplot Matrix displaying pair-wise comparisons of 4 models.

GAMUT  
(Hohman et al, 2019)

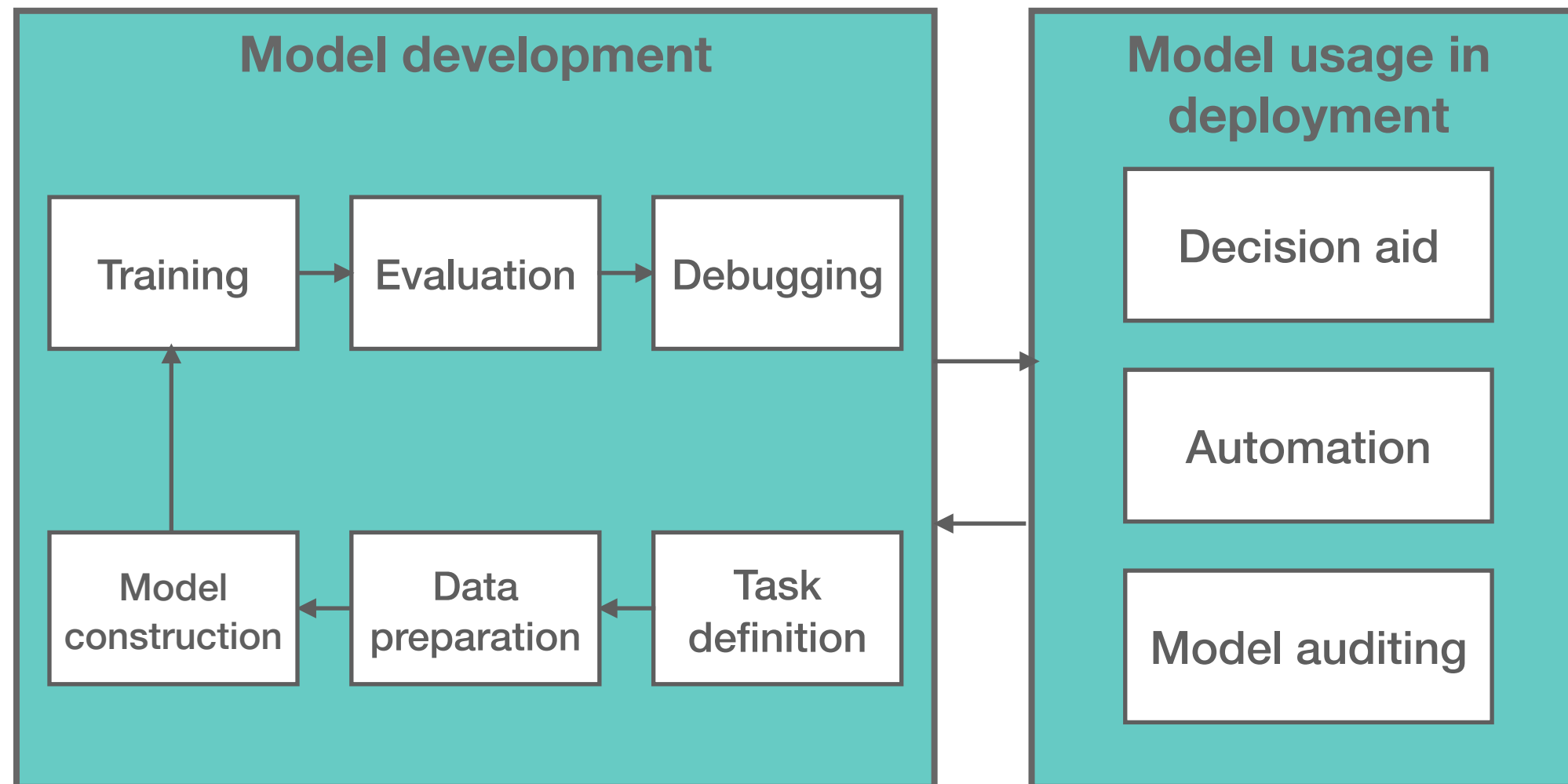
# XAI use cases in AI lifecycle

**Model debugging or selection** (IUI2021)

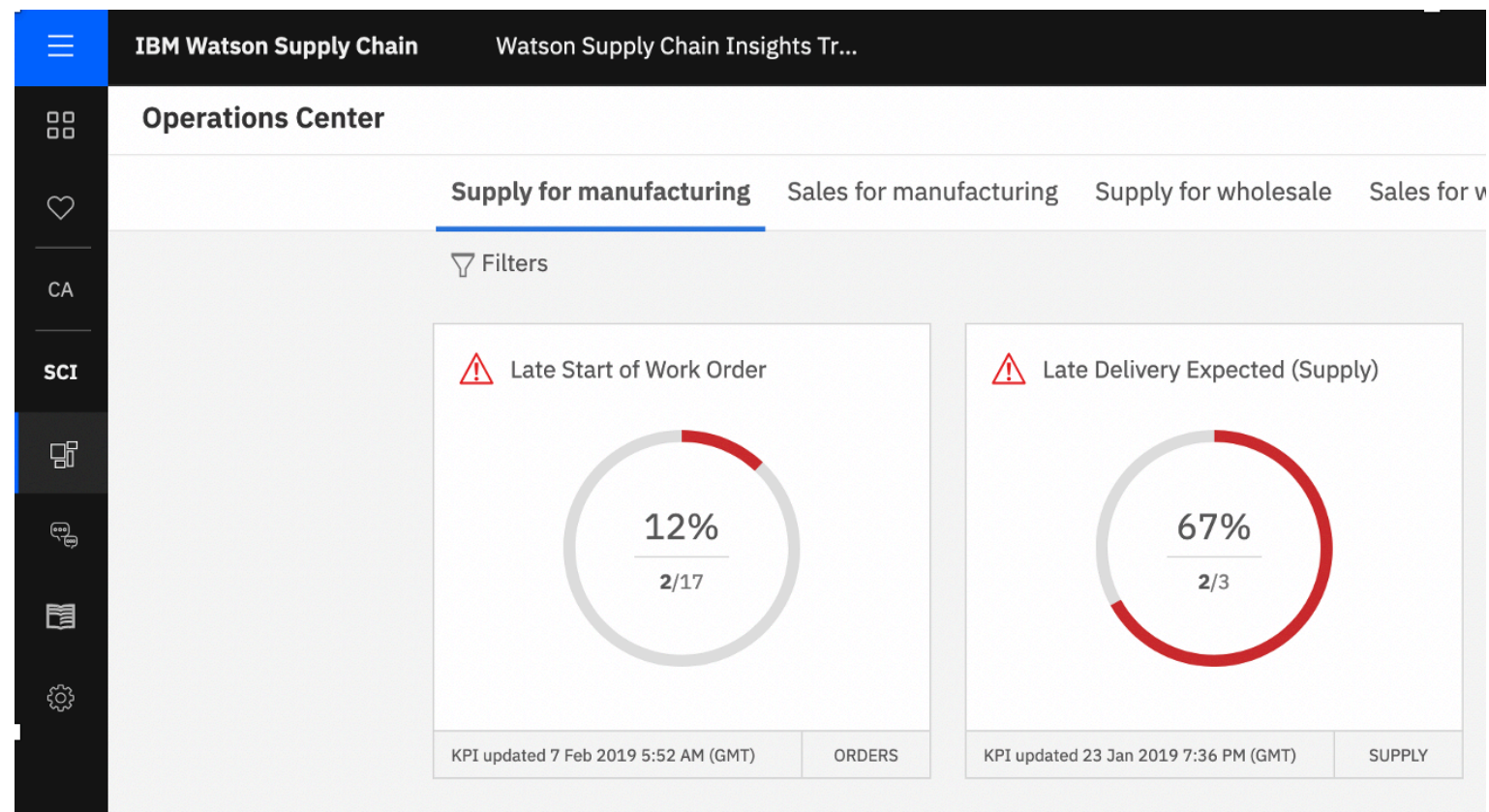
XAI user: **Data scientist**


**Trust calibration and decision support** (FAT\* 2020, CHI 2021 )

XAI user: **Decision-maker**



# XAI for actionable decision-making



 *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 (1-5)*

# XAI for human-AI collaboration and **trust calibration**

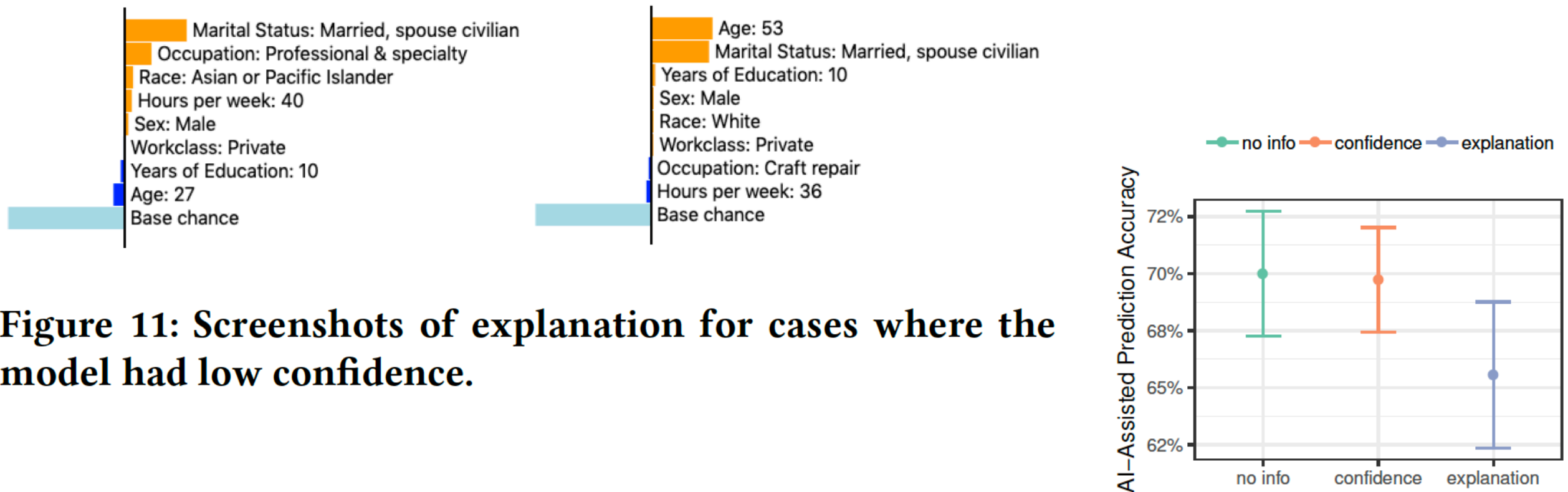


“ There is a calibration of trust, whether people will use it over time. But also saying hey, we know this fails in this way (I-6)



# XAI for **trust calibration** in decision-making

Caveat: Explanation can lead to unwarranted trust!



**Figure 11: Screenshots of explanation for cases where the model had low confidence.**



**Decision-makers**

Can I trust this prediction? ❌

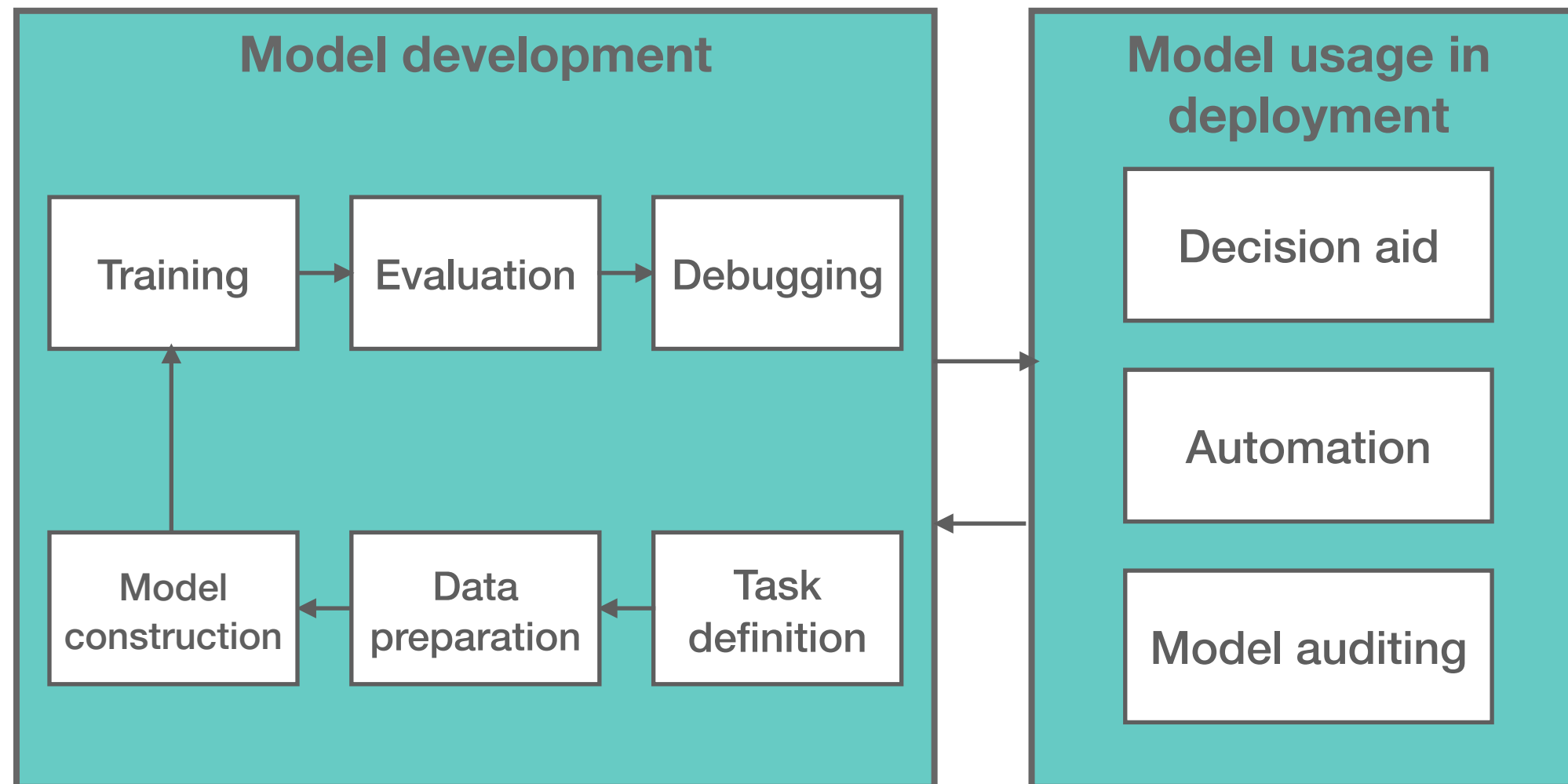
# XAI use cases in AI lifecycle

**Model debugging or selection** (IUI2021)

XAI user **Data scientist**

**Trust calibration and decision support** (FAT\* 2020, CHI 2021 🏆 )

XAI user **Decision-maker**



**Fairness assessment** (IUI 2019 🏆 )

XAI user: **Regulator, impacted groups**

# Fair ML: What is unwanted bias?



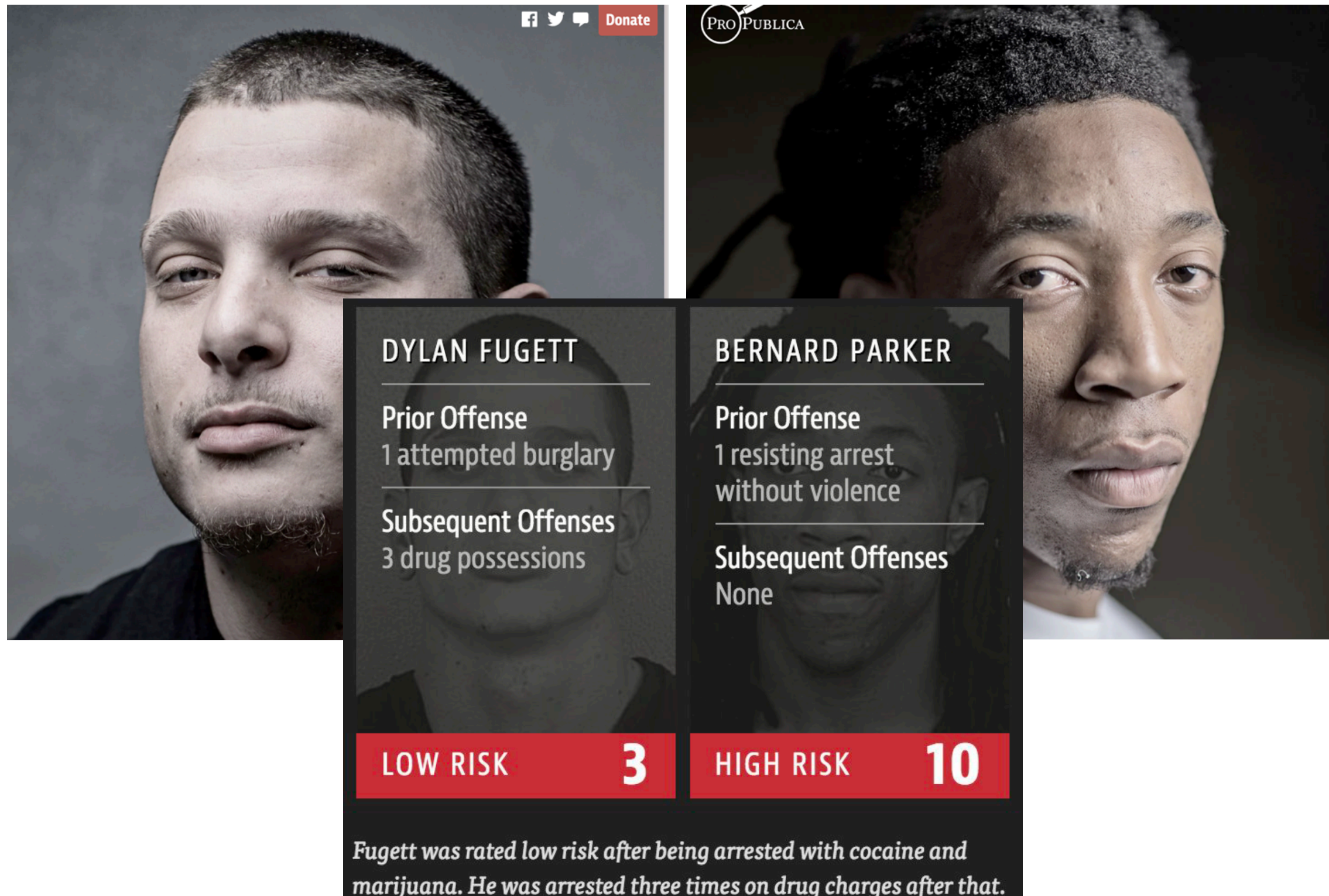
Discrimination becomes objectionable when it places certain **unprivileged** groups at a systematic disadvantage

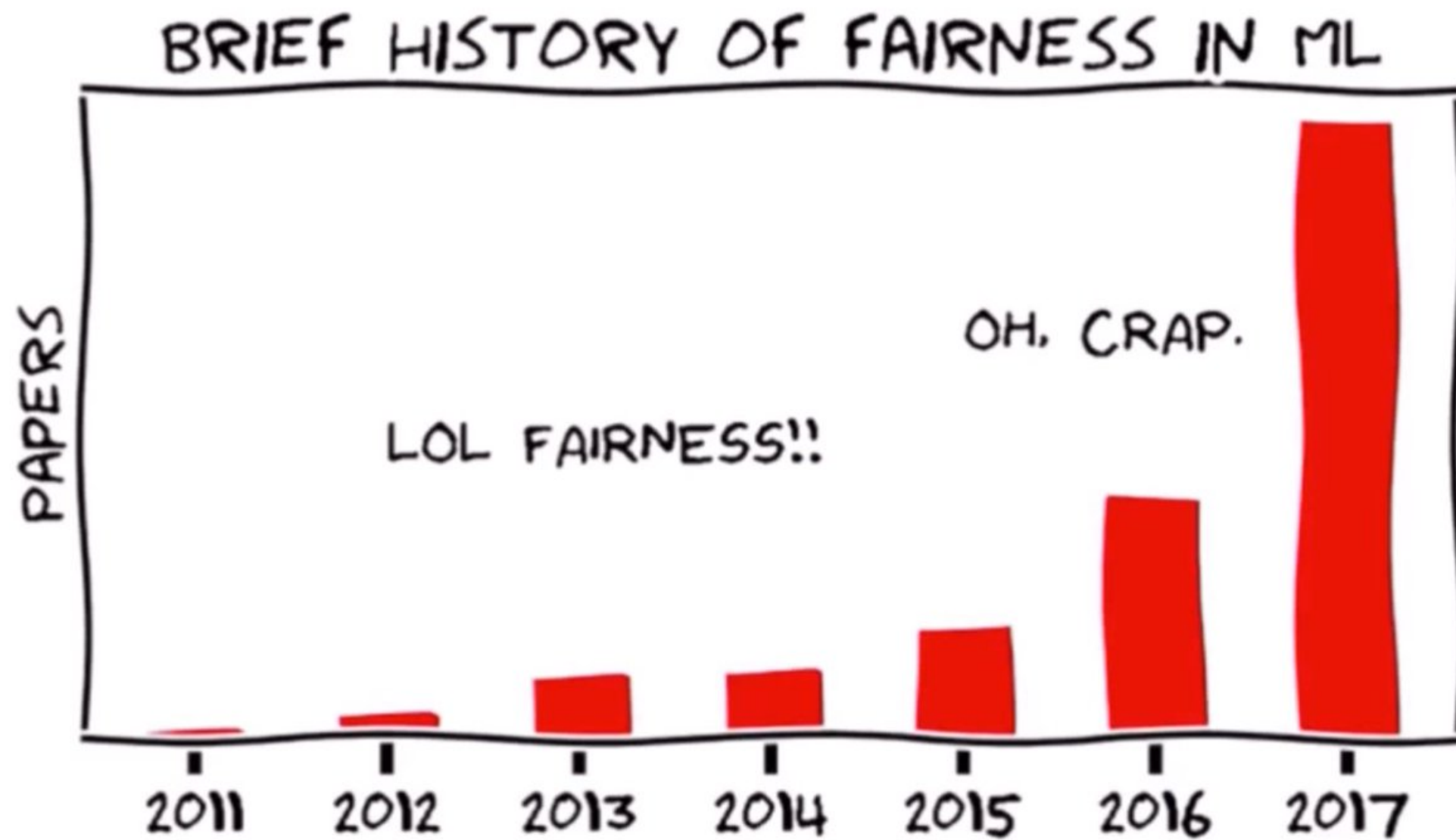
Illegal in certain contexts

(Barocas and Selbst, 2017)



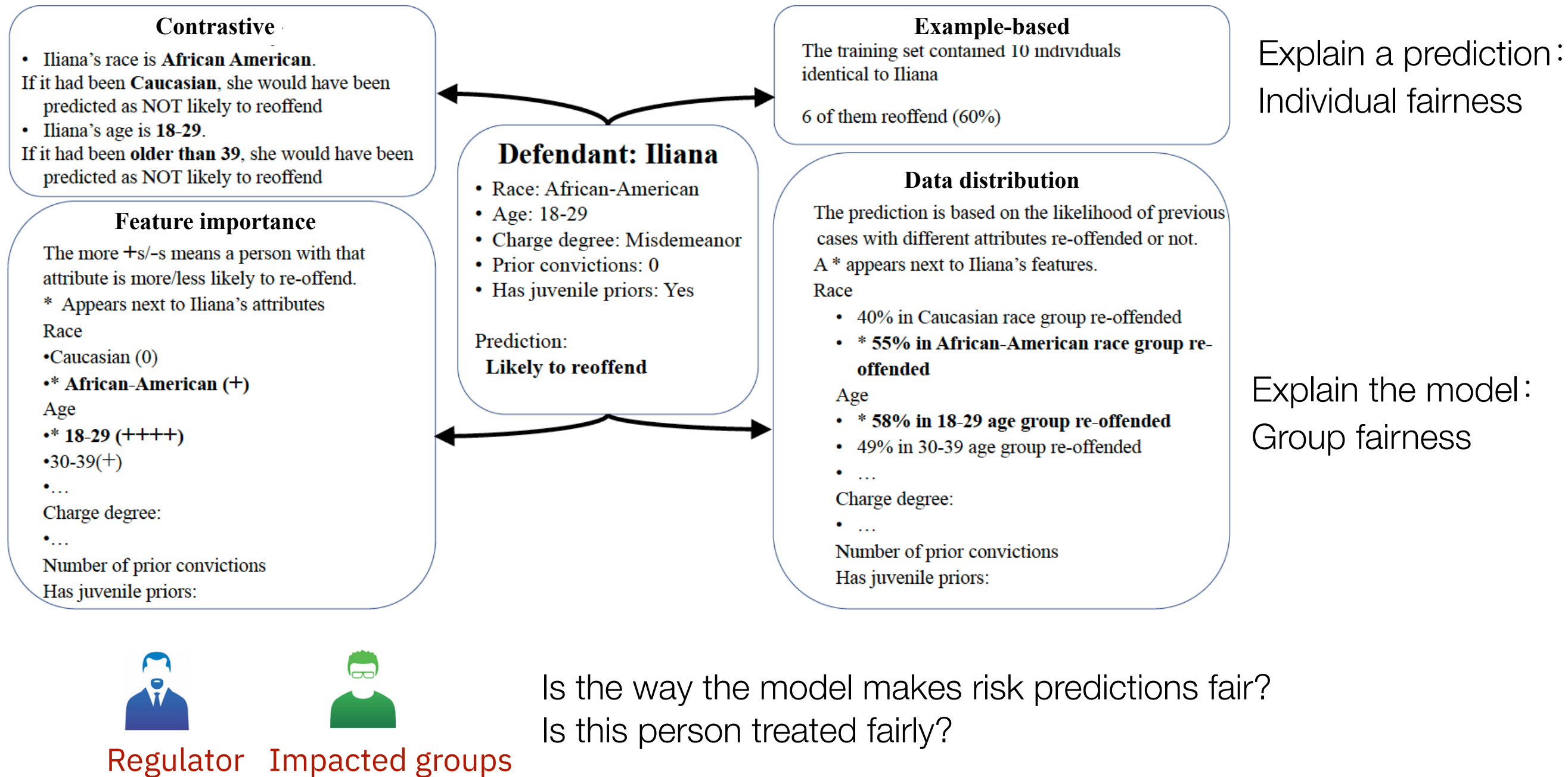
# Discrimination in COMPAS





(Hardt, 2017)

# XAI as interfaces for scrutinizing discrimination





# Lessons learned: From XAI algorithms to XAI UX

- No one-fits-all solutions
- XAI UX often needs multiple types of explanation/transparency information
  - Anticipate *when* and *where* users want *what* explanations
- Beware of the potential risk of XAI
  - Unwarranted trust and confidence
  - Distraction and information workload
  - Disparate effect: disadvantage people with “non-ideal” ability and motivation to process XAI
- Under-developed “translation” design space
- Algorithmic explanations may not satisfy all users’ information needs to achieve understanding of AI



HCAI: “understanding” lies in the recipient

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) i. and 15 (1) h)



**“meaningful” ???**

(Nemitz, 2018)

“Understanding” lies in the recipient:  
beyond the toolbox



**XAI techniques**



Information needs to achieve  
understanding of AI:

- General AI knowledge gaps
- Domain knowledge gaps

“Understanding” lies in the recipient:  
beyond the toolbox



**XAI techniques**



**XAI UX**

“Sense-making is not just about opening the closed box of AI, but also about who is around the box, and the socio-technical factors that govern the use of the AI system and the decision. Thus the 'ability' in explainability does not lie exclusively in the guts of the AI system

**Information needs to achieve understanding of AI:**

- General AI knowledge gaps
- Domain knowledge gaps
- “Socially situated understanding”

# 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

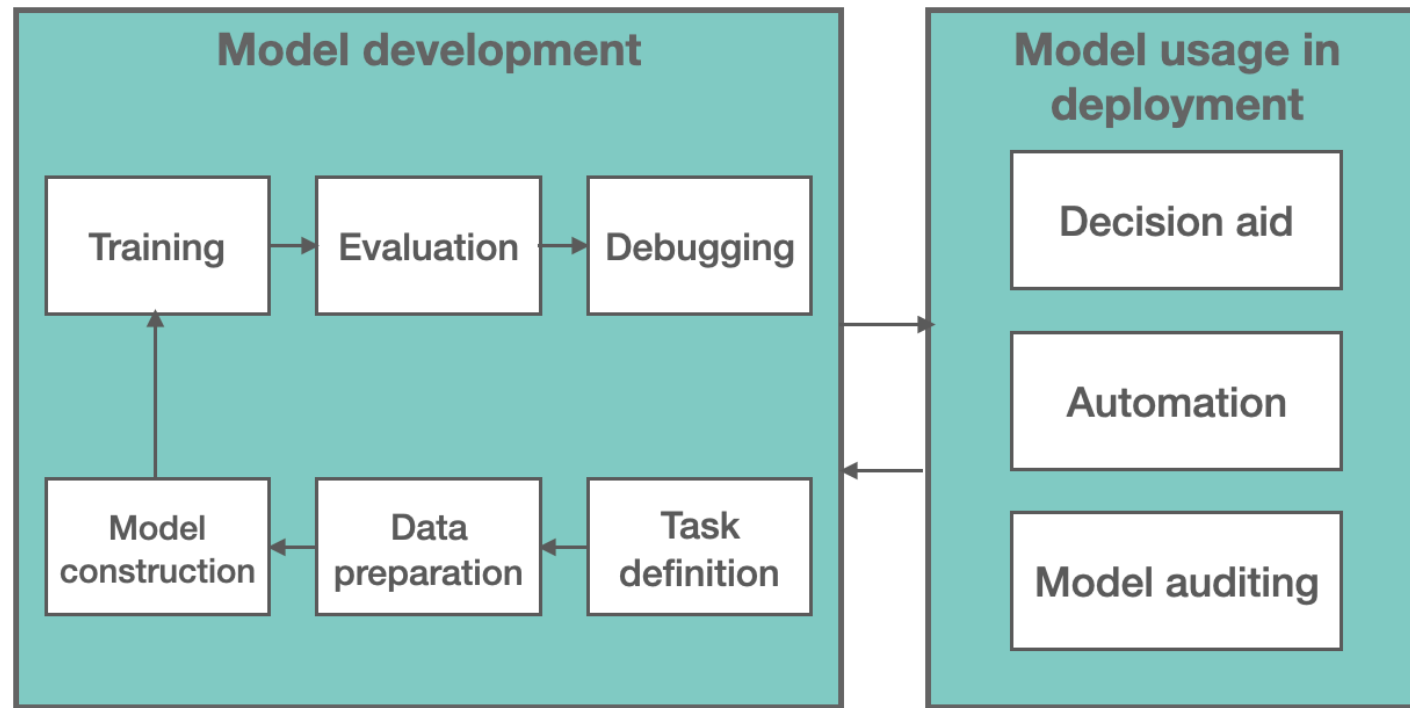
May 6, 2020

5

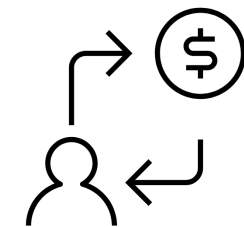




# Many user objectives + user groups + domains + social contexts



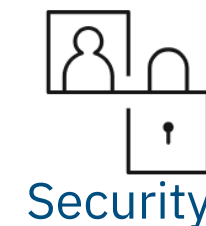
Healthcare



Finance



Business



Security

## End user decision makers

- Who: physicians, judges, loan officers, teacher evaluators
- Why: trust/confidence, insights

## Regulatory bodies

- Who: EU [GDPR], NYC Council, US Gov't
- Why: ensure fairness for constituents



## All system builders

- Who: data scientists, developers
- Why: ensure/improve performance



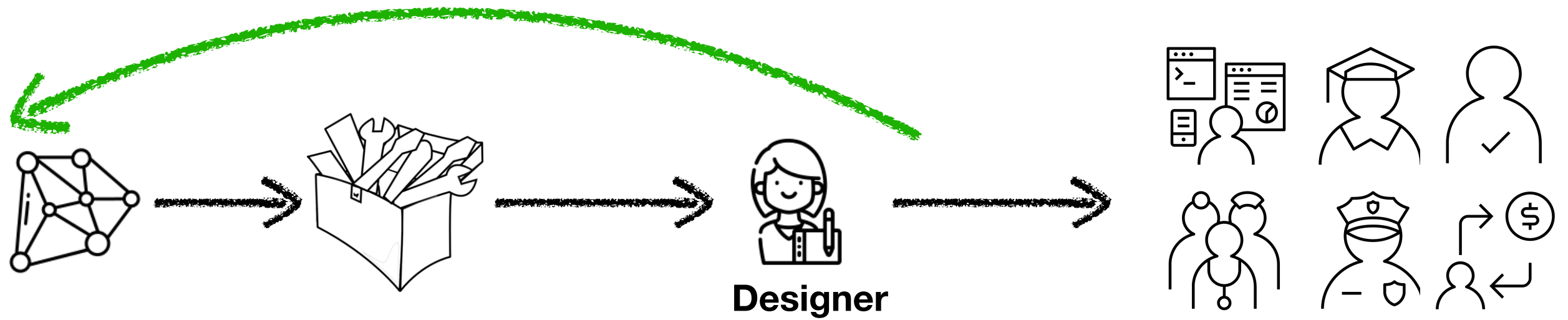
## End consumers

- Who: patients, accused, loan applicants, teachers
- Why: understanding of factors

Must match  
the **complexity capability**  
of the consumer

Must match  
the **domain knowledge**  
of the consumer

(Hind et al., 2019)



How to **select**?      How to **translate**?

## Thread 1: Study and support design practices for XAI UX

---

Thread 2: HCI research with XAI use cases

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

## Research questions:

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



Review

# Machine Learning Interpretability: A Survey on Methods and Metrics

Diogo V. Carvalho<sup>1,2,\*</sup>, Eduardo M. Pereira<sup>1</sup> and

<sup>1</sup> Deloitte Portugal, Manuel Bandeira Street, 43, 4150-47  
<sup>2</sup> Faculty of Engineering, University of Porto, Dr. Rober  
<sup>3</sup> 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 toward  
algorithmically informed decisions have greater po  
most of these accurate decision support systems rem  
logic and inner workings are hidden to the user  
ratic  
mach  
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 ex  
planatory 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, loosely  
defined as the science of comprehending what a model did (c  
le models and learning method  
les include visual cues to fin  
networks in image recognition

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

AMINA ADADI<sup>1</sup> 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 technical space people are not quite in there yet... how to talk about it?

## 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 c  
at the cost of sacrificing accuracy for interpretability. The appli  
can be used are various, and each approach is typically develop  
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, expla  
nations 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 con  
cerns 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  
that explanation methods or interfaces for lay users are missing and we speculate which criteria

computational Intelligence, University of Granada, 18071 Granada, Spain  
nica, 28050 Madrid, Spain

(AI) has achieved a notable momentum that, if harnessed  
tions 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  
type of AI (namely, expert systems and rule based models).  
in the so-called *eXplainable* AI (XAI) field, which is widely  
actical deployment of AI models. The overview presented in  
id contributions already done in the field of XAI, including a  
r this purpose we summarize previous efforts made to define  
ing 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



# Study probe: algorithm informed **XAI Questions**

Category of Methods	Explanation Method	Definition	Algorithm Examples	Question Type
Explain the model ( <b>Global</b> )	Global feature importance	Describe the weights of features used by the model (including visualization that shows the weights of features)	[41, 60, 69, 90]	<b>How</b>
	Decision tree approximation	Approximate the model to an interpretable decision-tree	[11, 47, 52]	<b>How, Why, Why not, What if</b>
	Rule extraction	Approximate the model to a set of rules, e.g., if-then rules	[26, 93, 102]	<b>How, Why, Why not, What if</b>
Explain a prediction ( <b>Local</b> )	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]	<b>Why</b>
	Local rules or trees	Describe the rules or a decision-tree path that the instance fits to guarantee the prediction	[39, 75, 99]	<b>Why, How to still be this</b>
<b>Inspect counterfactual</b>	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]	<b>What if, How to be that, How to still be this</b>
	Contrastive or counterfactual features	Describe the feature(s) that will change the prediction if perturbed, absent or present	[27, 91, 100]	<b>Why, Why not, How to be that</b>
<b>Example based</b>	Prototypical or representative examples	Provide example(s) similar to the instance and with the same record as the prediction	[13, 48, 50]	<b>Why, How to still be this</b>
	Counterfactual example	Provide example(s) with small differences from the instance but with a different record from the prediction	[37, 55, 66]	<b>Why, Why not, How to be that</b>

- 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**

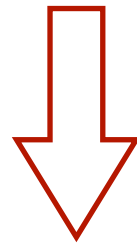


*An explanation is an answer to a question (Wellman, 2011; Miller 2018)*

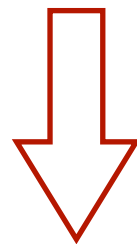
*The effectiveness of an explanation depends on the question asked (Bromberger, 1992)*



**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 ( <b>Global</b> )	Global feature importance	Describe the weights of features used by the model (including visualization that shows the weights of features)	[41, 60, 69, 90]	<b>How</b>
	Decision tree approximation	Approximate the model to an interpretable decision-tree	[11, 47, 52]	<b>How, Why, Why not, What if</b>
	Rule extraction	Approximate the model to a set of rules, e.g., if-then rules	[26, 93, 102]	<b>How, Why, Why not, What if</b>
Explain a prediction ( <b>Local</b> )	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]	<b>Why</b>
	Local rules or trees	Describe the rules or a decision-tree path that the instance fits to guarantee the prediction	[39, 75, 99]	<b>Why, How to still be this</b>
<b>Inspect counterfactual</b>	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]	<b>What if, How to be that, How to still be this</b>
	Contrastive or counterfactual features	Describe the feature(s) that will change the prediction if perturbed, absent or present	[27, 91, 100]	<b>Why, Why not, How to be that</b>
<b>Example based</b>	Prototypical or representative examples	Provide example(s) similar to the instance and with the same record as the prediction	[13, 48, 50]	<b>Why, How to still be this</b>
	Counterfactual example	Provide example(s) with small differences from the instance but with a different record from the prediction	[37, 55, 66]	<b>Why, Why not, How to be that</b>

+

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 potential limitations/biases of the data?
- What is the size, proportion, or distribution of the training data with given feature(s)/feature-value(s)?

## 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 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?

## 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 should I best utilize the output of the system?
- How should the output fit in my workflow?

## Why not

## How to be that (a different prediction)

- **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]?

## 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 to still be this (the current prediction)

- **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?

## How

(global model-wide explanation)

- **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?

## 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? (improvement)
- Why is the system using or not using a given algorithm/feature/rule/dataset? (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 for 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: fairness, privacy, etc.

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:** mimic how 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 and implementation cost**  
impeding buy-in from data scientists and the team

“*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

### 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

## XAI in Practice

### 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

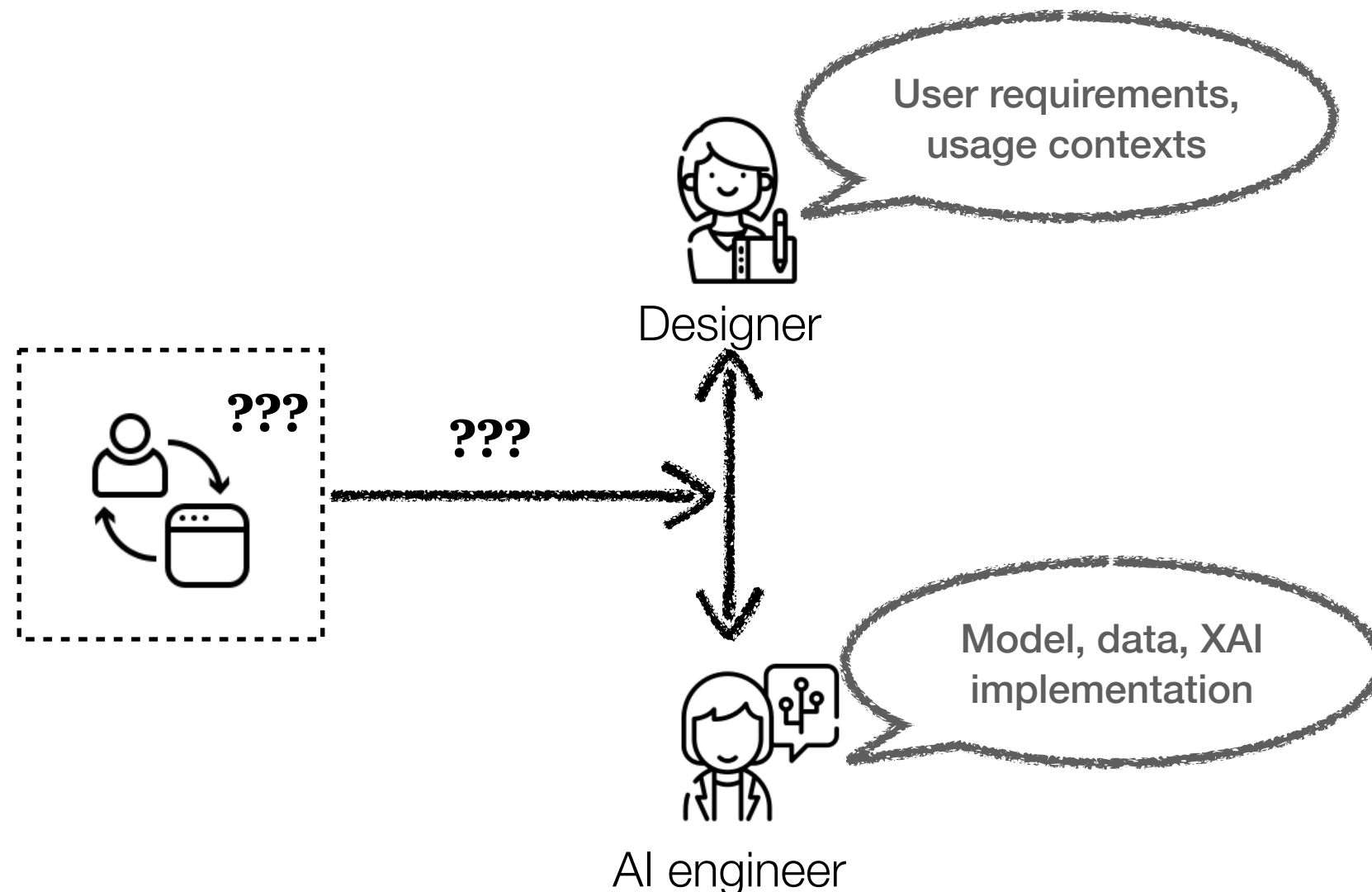
**Performance:** Help users understand the limitations of the AI and make it actionable

**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

# User-centered design process: **Question-driven XAI design**



Pain points to address:

- Thoroughly identify interaction specific XAI user needs
- Enable a “designedly” understanding of XAI techniques to find the right pairing
- Support designer-engineer collaboration

# 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 potential limitations/biases of the data?
- What is the size, proportion, or distribution of the training data with given feature(s)/feature-value(s)?

## 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 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?

## 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 should I best utilize the output of the system?
- How should the output fit in my workflow?

## How to be that (a different prediction)

- **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]?

## 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 to still be this (the current prediction)

- **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

- **How does the system make predictions?**
- What features does the system consider?
  - Is [feature X] used or not used for the predictions?

- **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]?

## How

(global model-wide explanation)

- 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?

## Others

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

Question	Explanations	Example XAI techniques
<b>Global how</b>	<ul style="list-style-type: none"> <li>Describe what algorithm is used and what features are considered, if a user is only interested in a high-level view</li> <li>Describe the general model logic as feature impact<sup>+</sup>, rules<sup>+</sup> or decision-trees<sup>•</sup> (sometimes need to explain with a surrogate simple model)</li> </ul>	<a href="#">ProfWeight</a> <sup>++•</sup> , <a href="#">Feature Importance</a> <sup>+</sup> , <a href="#">PDP</a> <sup>+</sup> , <a href="#">BRCG</a> <sup>+</sup> , <a href="#">GLRM</a> <sup>+</sup> , <a href="#">Rule List</a> <sup>+</sup> , <a href="#">DT Surrogate</a> <sup>•</sup>
<b>Why</b>	<ul style="list-style-type: none"> <li>Describe what key features of the particular instance determine the model's prediction of it<sup>+</sup></li> <li>Describe rules<sup>+</sup> that the instance fits to guarantee the prediction</li> <li>Show similar examples<sup>•</sup> with the same predicted outcome to justify the model's prediction</li> </ul>	<a href="#">LIME</a> <sup>+</sup> , <a href="#">SHAP</a> <sup>+</sup> , <a href="#">LOCO</a> <sup>+</sup> , <a href="#">Anchors</a> <sup>+</sup> , <a href="#">ProtoDash</a> <sup>•</sup>
<b>Why not</b>	<ul style="list-style-type: none"> <li>Describe what changes are required for the instance to get the alternative prediction and/or what features of the instance guarantee the current prediction<sup>+</sup></li> <li>Show prototypical examples<sup>+</sup> that had the alternative outcome</li> </ul>	<a href="#">CEM</a> <sup>+</sup> , <a href="#">Prototype counterfactual</a> <sup>+</sup> , <a href="#">ProtoDash</a> <sup>+</sup> (on alternative class)
<b>How to be that</b>	<ul style="list-style-type: none"> <li>Highlight features that if changed (increased, decreased, absent, or present) could alter the prediction<sup>+</sup></li> <li>Show examples with small differences but had a different outcome than the prediction<sup>+</sup></li> </ul>	<a href="#">CEM</a> <sup>+</sup> , <a href="#">Counterfactuals</a> <sup>+</sup> , <a href="#">DiCE</a> <sup>+</sup>
<b>What if</b>	<ul style="list-style-type: none"> <li>Show how the prediction changes corresponding to the inquired change</li> </ul>	<a href="#">PDP</a> , <a href="#">ALE</a> , <a href="#">What-if Tool</a>
<b>How to still be this</b>	<ul style="list-style-type: none"> <li>Describe feature ranges<sup>+</sup> or rules<sup>+</sup> that could guarantee the same prediction</li> <li>Show examples that are different from the particular instance but still had the same outcome</li> </ul>	<a href="#">CEM</a> <sup>+</sup> , <a href="#">Anchors</a> <sup>+</sup>
<b>Performance</b>	<ul style="list-style-type: none"> <li>Provide performance metrics of the model</li> <li>Show confidence or uncertainty information for each prediction</li> <li>Describe potential strengths and limitations of the model</li> </ul>	Precision, Recall, Accuracy, F1, AUC Confidence <a href="#">FactSheets</a> , <a href="#">Model Cards</a>
<b>Data</b>	<ul style="list-style-type: none"> <li>Document comprehensive information about the training data, including the source, provenance, type, size, coverage of population, potential biases, etc.</li> </ul>	<a href="#">FactSheets</a> , <a href="#">DataSheets</a>
<b>Output</b>	<ul style="list-style-type: none"> <li>Describe the scope of output or system functions</li> <li>Suggest how the output should be used for downstream tasks or user workflow</li> </ul>	<a href="#">FactSheets</a> , <a href="#">Model Cards</a>

Questions as *re-framing* the technical space of XAI

Questions as "*boundary objects*" supporting designer-engineer collaboration



# Question-Driven XAI Design

Step 1

Identify user questions

Step 2

Analyze questions

Step 3

Map questions to modeling solutions

Step 4

Iteratively design and evaluate

---

Elicit user needs for XAI as questions

Also gather user intentions and expectations for asking the 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

---

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

---

Create a design including the candidate elements identified in step 3

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

**Designers, users**

**Designers, product team**

**Designers, data scientists**

**Designers, data scientists, users**



A running example

## Adverse Event Prediction for Healthcare

HealthMind is developing an AI based dashboard system to help clinicians assess patients' readmission risks at discharge time.

By simply providing a risk score, the system is of limited use for clinicians. **Clinicians need to understand how the system arrives at a risk score for a patient in order to feel confident in the judgment and identify effective interventions to improve the patient's health outcomes.**

The team needs to develop an explainable AI system but is not sure where to start.



HealthMind's AI based dashboard

# 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**

# Identify relevant questions

Elicit user questions to identify what types of explanation are needed

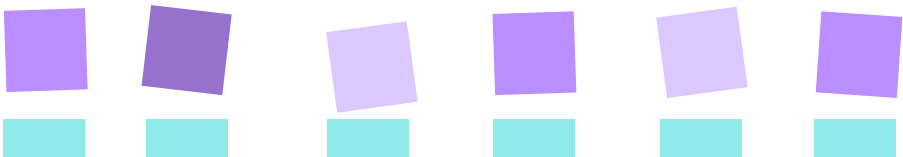
Also collect the **intention and expectation** behind these user questions

Task description

An AI based dashboard presents patients' readmission risk scores to help clinicians to identify high-risk

User Journey (optional)

Questions from User 1



Questions from User 2



# Identify relevant questions

---

Elicit user questions to identify what types of explanation are needed

Also collect the **intention and expectation** behind these user questions

What are the main risk factors for this person?

*“Help me better understand the patient, discover otherwise non-obvious factors, e.g. social status or community factors”*

What is the population of the training data?

*“Without knowing if it applies to my patients I can’t trust it”*

# Question-Driven XAI Design

Step 1

Identify user  
questions



Step 2

Analyze  
questions

---

Elicit user needs for  
XAI as questions

Also gather user  
intentions and  
expectations for  
asking the questions

**Designers,  
users**

---

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

Summarize user intentions  
and expectations to  
identify key user  
requirements

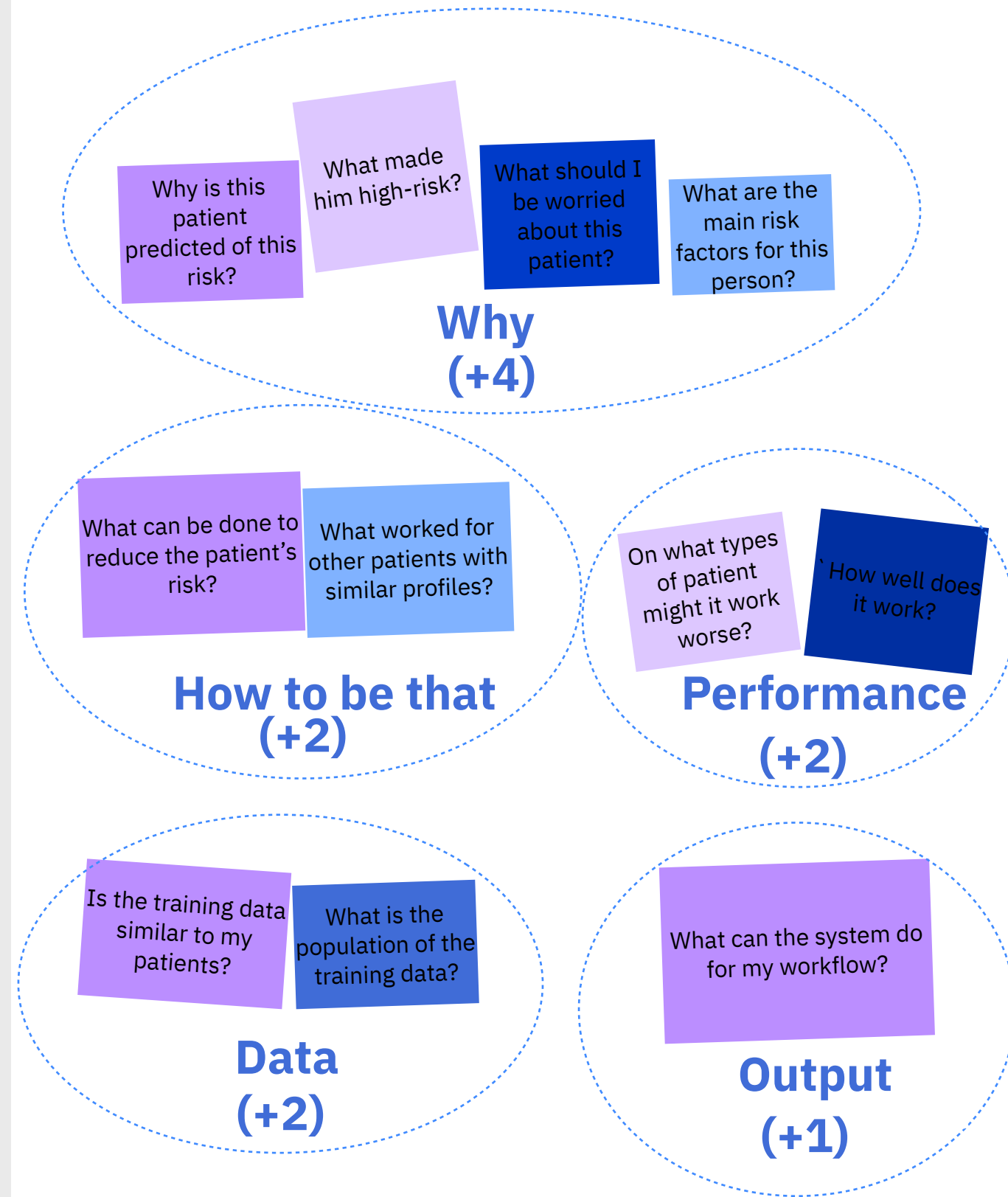
**Designers,  
product team**

Categorize and prioritize questions,  
identify key user requirements

Cluster similar questions across users  
into categories (use the Question Bank  
to guide labeling if needed)

Prioritize clusters with more questions

Summarize user intentions and  
expectations to identify key user  
requirements





Categorize and prioritize questions, identify key user requirements

Cluster similar questions across users into categories (use the Question Bank to guide labeling if needed)

Prioritize clusters with more questions

Summarize user intentions and expectations to identify key user requirements

User requirements			
UR1: Discover new information about the patient	<i>“Help me better understand the patient, discover</i>	<i>“Help me see the patient as a whole”</i>	<i>“I want to know what is unique about this patient”</i>
UR2: Determine effective next steps for the patient	<i>“Help me determine the right intervention”</i>	<i>“Help us decide where and how to focus our resources on”</i>	<i>“To know what actions we can take with this patient”</i>
UR3: Increase confidence to use the tool	<i>“I will be more comfortable using the tool”</i>	<i>“Without knowing if it applies to my patients I can’t trust it”</i>	
UR4: Appropriately evaluate the reliability of a prediction	<i>“So I know whether I should lean on my own experience”</i>		

# Question-Driven XAI Design

Step 1

Identify user questions

Step 2

Analyze questions

Step 3

Map questions to modeling solutions

---

Elicit user needs for XAI as questions

Also gather user intentions and expectations for asking the 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

---

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, users**

**Designers, product team**

**Designers, data scientists**

Question	Explanations	Example XAI techniques
<b>Global how</b>	<ul style="list-style-type: none"> <li>Describe what algorithm is used and what features are considered, if a user is only interested in a high-level view</li> <li>Describe the general model logic as feature impact<sup>+</sup>, rules<sup>+</sup> or decision-trees<sup>•</sup> (sometimes need to explain with a surrogate simple model)</li> </ul>	<a href="#">ProfWeight</a> <sup>++•</sup> , <a href="#">Feature Importance</a> <sup>+</sup> , <a href="#">PDP</a> <sup>+</sup> , <a href="#">BRCG</a> <sup>+</sup> , <a href="#">GLRM</a> <sup>+</sup> , <a href="#">Rule List</a> <sup>+</sup> , <a href="#">DT Surrogate</a> <sup>•</sup>
<b>Why</b>	<ul style="list-style-type: none"> <li>Describe what key features of the particular instance determine the model's prediction of it<sup>+</sup></li> <li>Describe rules<sup>+</sup> that the instance fits to guarantee the prediction</li> <li>Show similar examples<sup>•</sup> with the same predicted outcome to justify the model's prediction</li> </ul>	<a href="#">LIME</a> <sup>+</sup> , <a href="#">SHAP</a> <sup>+</sup> , <a href="#">LOCO</a> <sup>+</sup> , <a href="#">Anchors</a> <sup>+</sup> , <a href="#">ProtoDash</a> <sup>•</sup>
<b>Why not</b>	<ul style="list-style-type: none"> <li>Describe what changes are required for the instance to get the alternative prediction and/or what features of the instance guarantee the current prediction<sup>+</sup></li> <li>Show prototypical examples<sup>+</sup> that had the alternative outcome</li> </ul>	<a href="#">CEM</a> <sup>+</sup> , <a href="#">Prototype counterfactual</a> <sup>+</sup> , <a href="#">ProtoDash</a> <sup>+</sup> (on alternative class)
<b>How to be that</b>	<ul style="list-style-type: none"> <li>Highlight features that if changed (increased, decreased, absent, or present) could alter the prediction<sup>+</sup></li> <li>Show examples with small differences but had a different outcome than the prediction<sup>+</sup></li> </ul>	<a href="#">CEM</a> <sup>+</sup> , <a href="#">Counterfactuals</a> <sup>+</sup> , <a href="#">DiCE</a> <sup>+</sup>
<b>What if</b>	<ul style="list-style-type: none"> <li>Show how the prediction changes corresponding to the inquired change</li> </ul>	<a href="#">PDP</a> , <a href="#">ALE</a> , <a href="#">What-if Tool</a>
<b>How to still be this</b>	<ul style="list-style-type: none"> <li>Describe feature ranges<sup>+</sup> or rules<sup>+</sup> that could guarantee the same prediction</li> <li>Show examples that are different from the particular instance but still had the same outcome</li> </ul>	<a href="#">CEM</a> <sup>+</sup> , <a href="#">Anchors</a> <sup>+</sup>
<b>Performance</b>	<ul style="list-style-type: none"> <li>Provide performance metrics of the model</li> <li>Show confidence or uncertainty information for each prediction</li> <li>Describe potential strengths and limitations of the model</li> </ul>	Precision, Recall, Accuracy, F1, AUC Confidence <a href="#">FactSheets</a> , <a href="#">Model Cards</a>
<b>Data</b>	<ul style="list-style-type: none"> <li>Document comprehensive information about the training data, including the source, provenance, type, size, coverage of population, potential biases, etc.</li> </ul>	<a href="#">FactSheets</a> , <a href="#">DataSheets</a>
<b>Output</b>	<ul style="list-style-type: none"> <li>Describe the scope of output or system functions</li> <li>Suggest how the output should be used for downstream tasks or user workflow</li> </ul>	<a href="#">FactSheets</a> , <a href="#">Model Cards</a>

Questions as re-framing the technical space of XAI

Questions as "*boundary objects*" supporting designer-engineer 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 evaluate the design with the user requirements identified in step 2 and fill the gaps

**Designers, data scientists, users**

Why is this patient predicted of this risk? What made him high-risk? What are his risk factors?

**Why**

What can be done to reduce the patient's risk? What worked for other patients with similar profiles?

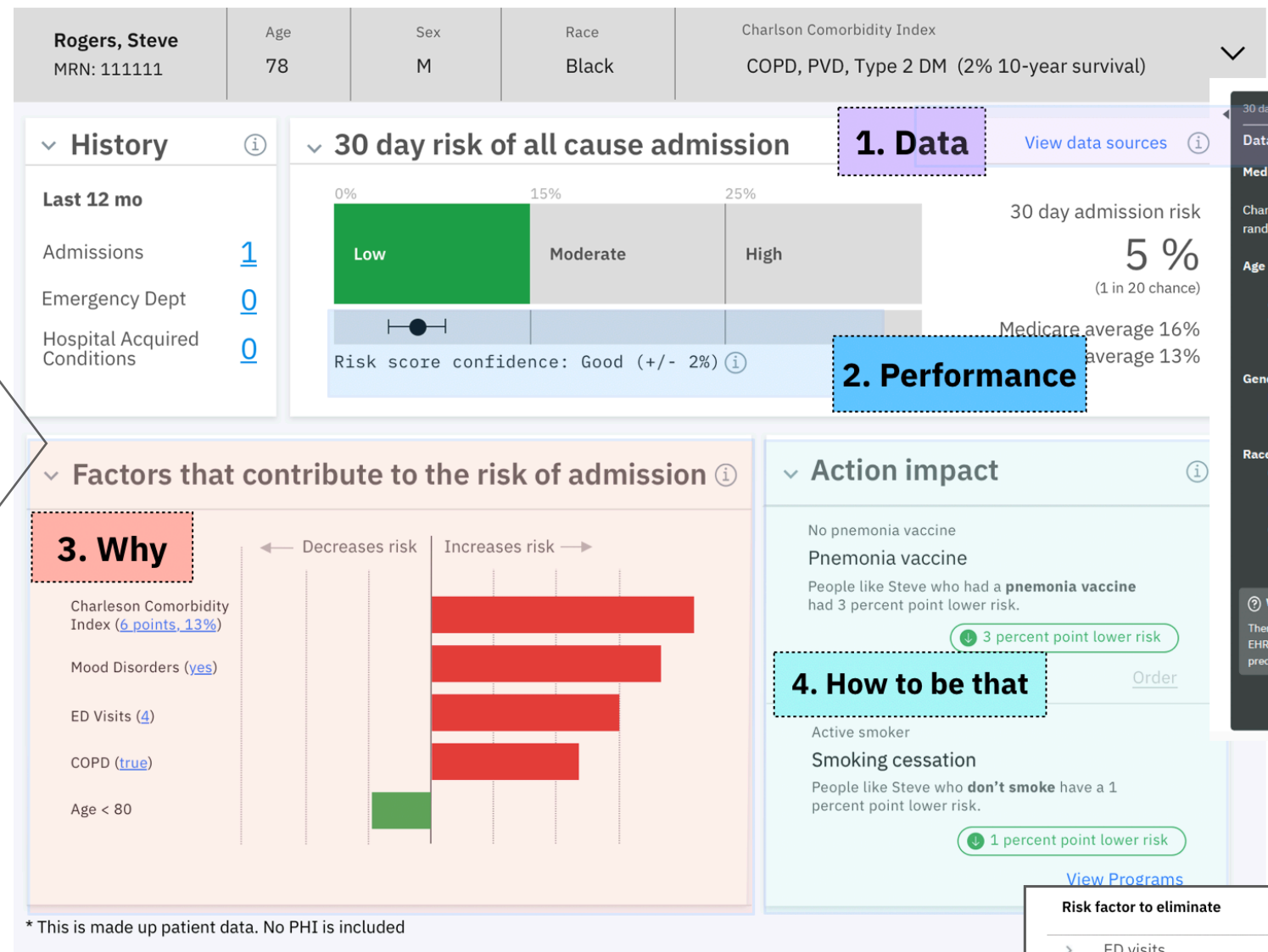
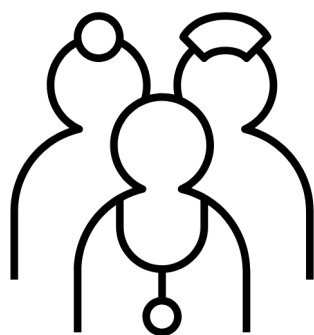
**How to be that**

On what types of patient might it work worse? How well does it work?

**Performance**

Is the training data similar to my patients? What is the population of the training data?

**Data**



## AI for Explainable Healthcare Adverse Event Risk Prediction

# Conclusions: **Bridging** work

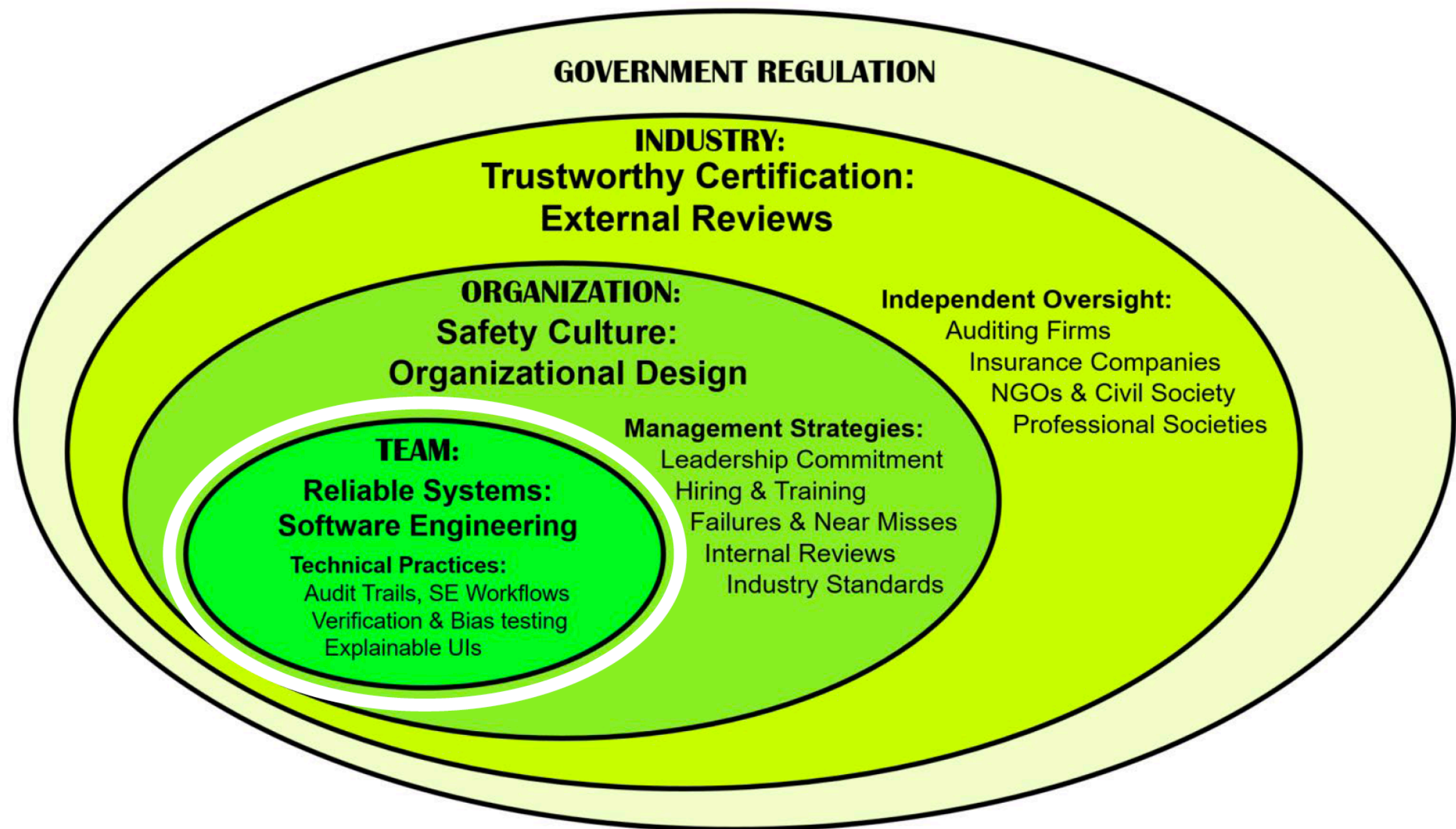
- **Human-centered** re-framing of technical spaces
  - Contextualize the tools by the human needs, values, and conditions they serve
  - Thinking “outside the toolbox” by centering on user needs and goals
- **Responsible** use of the toolbox
  - Examine breakdowns, limitations and potential harm
  - Not assuming “ideal users”
  - Enable user-centered design to drive technical development
- **Actionable** design assets and methods that practitioners can readily use

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





# Human-Centered AI: Beyond explainability



(Shneiderman, 2021)

# More resources for XAI

## Toolkits/Libraries

- [AIX 360](#)
- [Sheldon Alibi](#)
- [Oracle Skater](#)
- [H2o MLI](#)
- [Microsoft Interpret](#)
- [PyTorch Captum](#)

## Readings

- [Interpretable ML e-book](#)
- [A big list of resources](#)

## Design guidelines

- [Google PAIR: Explainability+Trust](#)
- [SAP Design Guidelines for Explainability](#)
- [IBM Design for AI: Explainability](#)
- [UXAI for Designers](#)
- [Lingua Franca: Transparency](#)

# Examples of **translation design** from XAI algorithms to XAI UX

An ***under-developed*** space

- Choose the right modality to communicate, e.g. visual or text-based
- Choose the right amount of information or level of granularity, e.g. how many features or examples
- Integrate XAI into the overall user workflow and experience. Sometimes it means to minimize distraction
- To achieve understanding, users may require additional information about the domain (e.g., what a feature means), AI (e.g., what a terminology means), socio-organizational contexts, etc.
- Sometimes need to link explanations to other evidence or guidelines (e.g., “how-to” for changing a feature) to support users’ objectives
- Sometimes need to put constraints or revise raw features due to security or privacy concerns

# 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](mailto:vera.liao@ibm.com)  
[www.qveraliao.com](http://www.qveraliao.com)  
@QVeraLiao