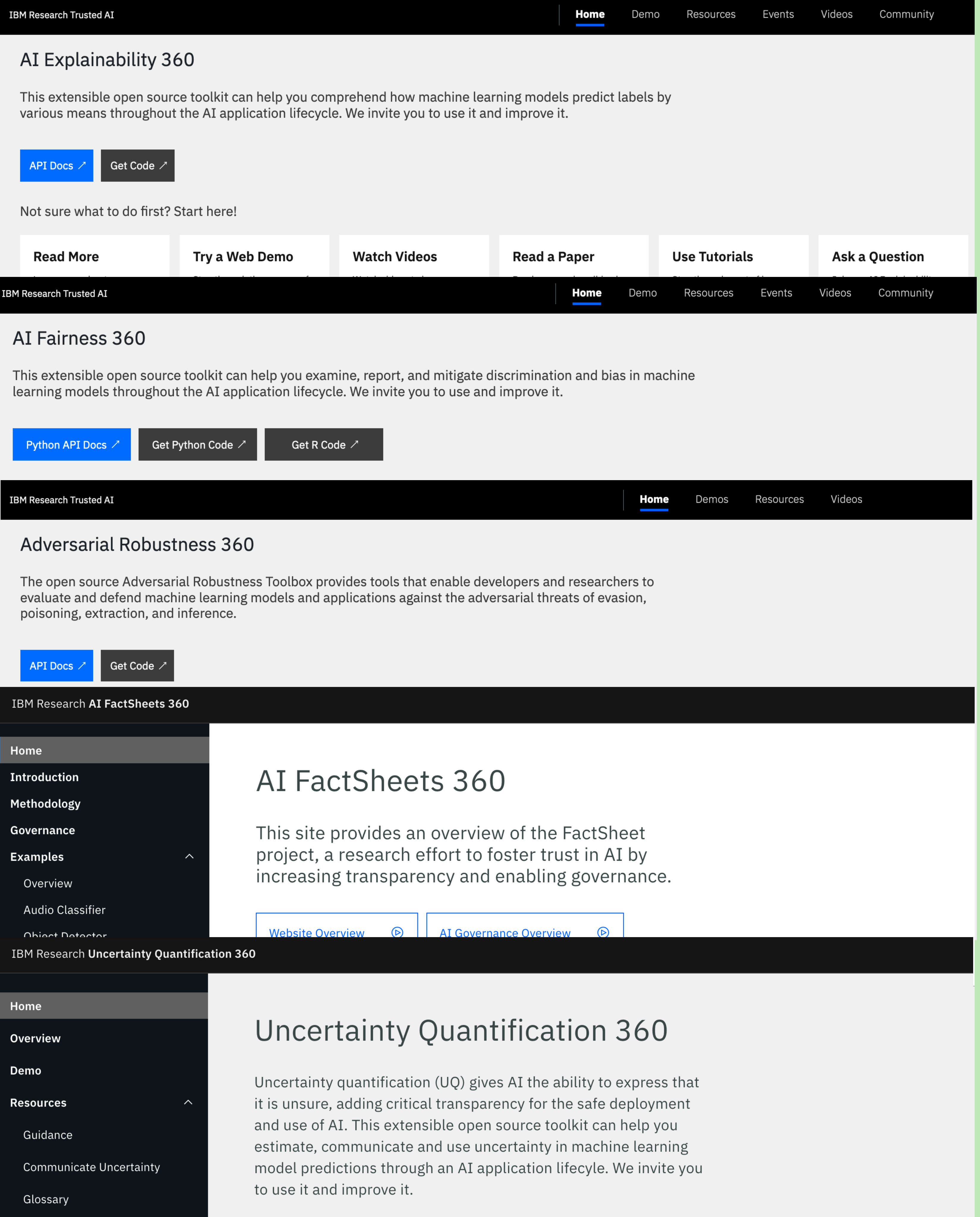
Questioning the AI: Towards Human Centered Explainable AI (XAI)

Research work 2018-2021

Q. Vera Liao
IBM Research

HCXAI
HUMAN-CENTERED EXPLAINABLE AI

HCXAI logo made by Upol Ehsan



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

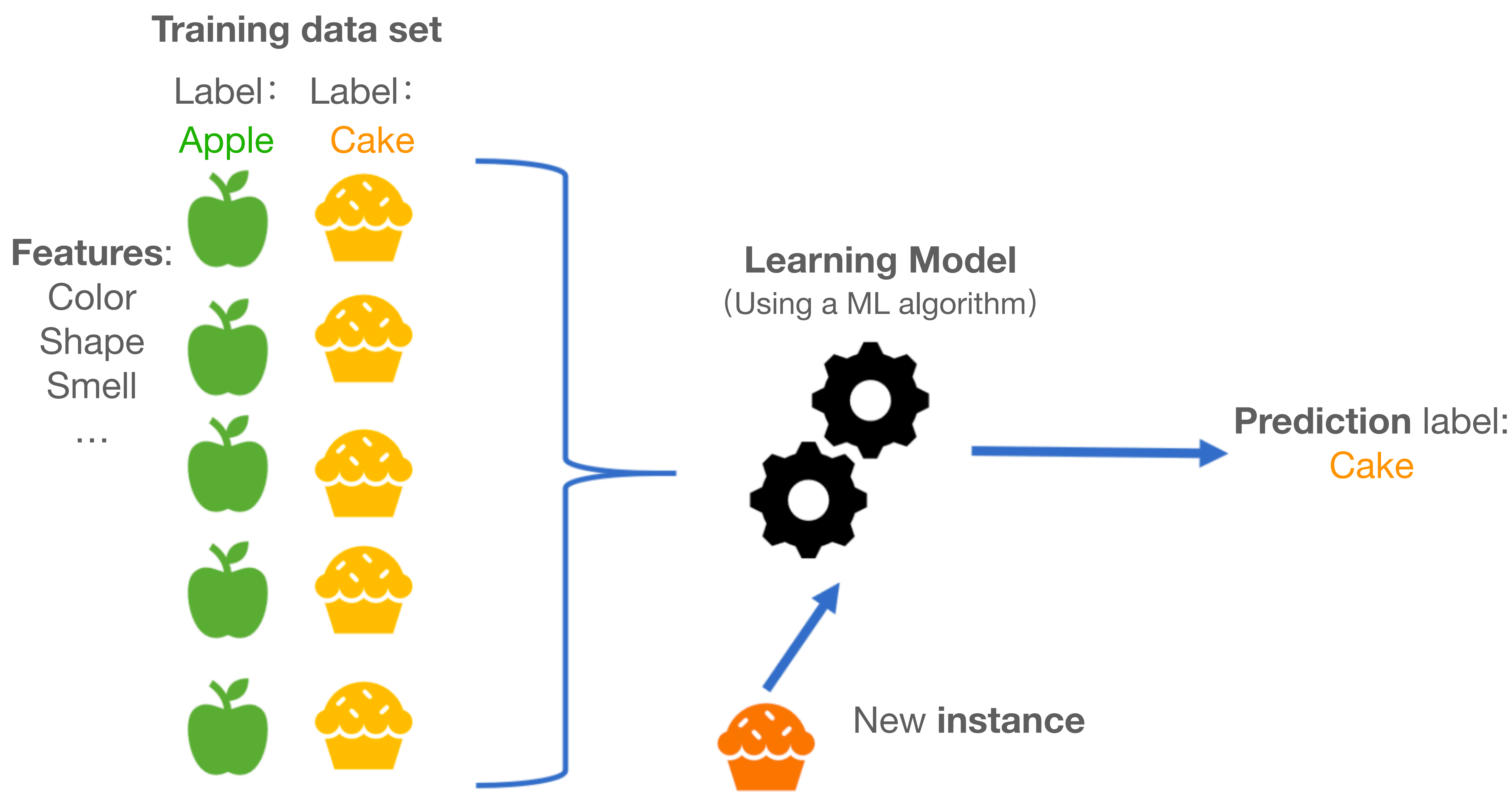
Broader definition:

(comprehensible/intelligible AI)

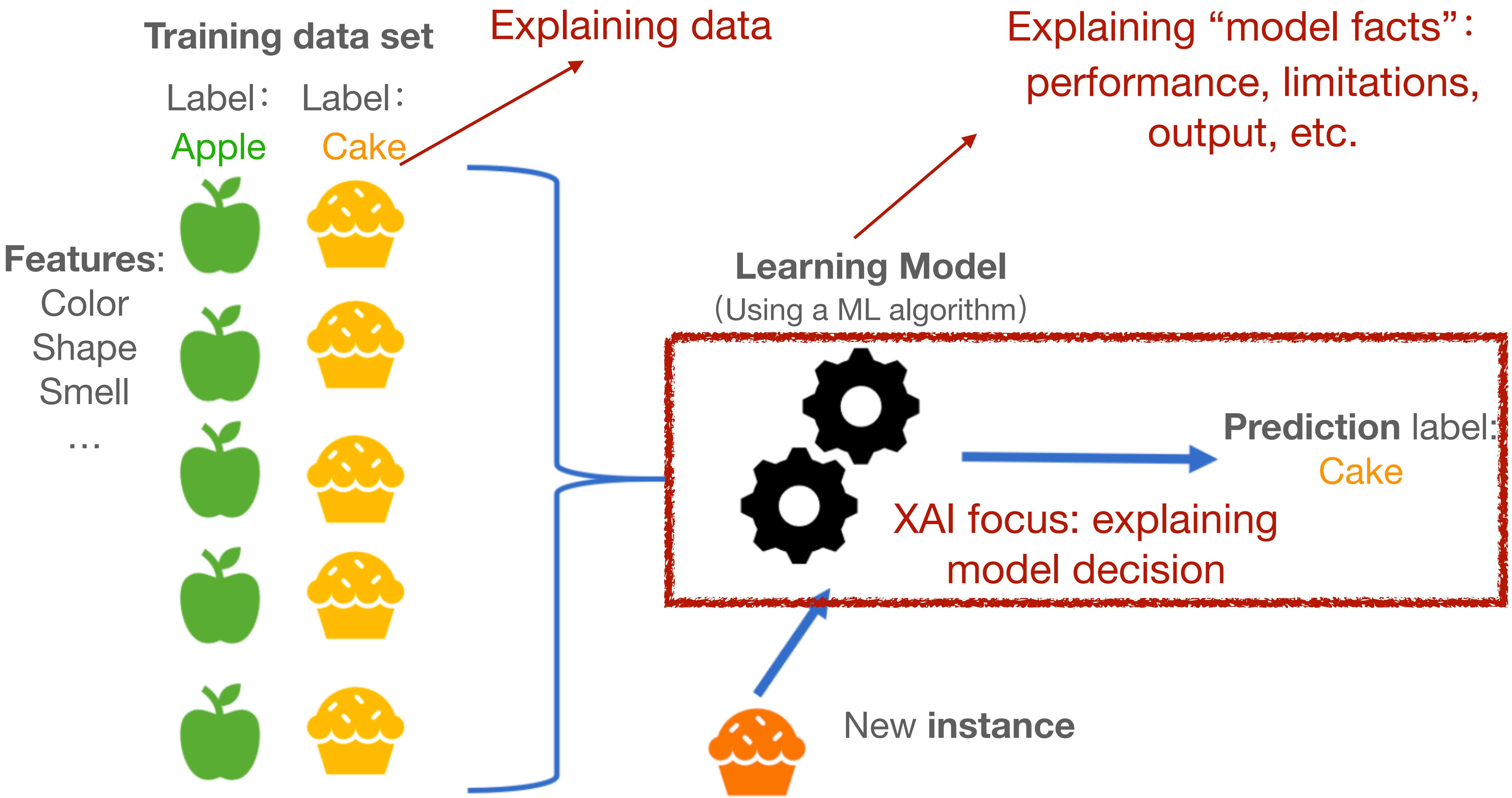
Everything that makes AI understandable (e.g., also including data, functions, performance, etc.)

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

Supervised Machine Learning



Supervised Machine Learning



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

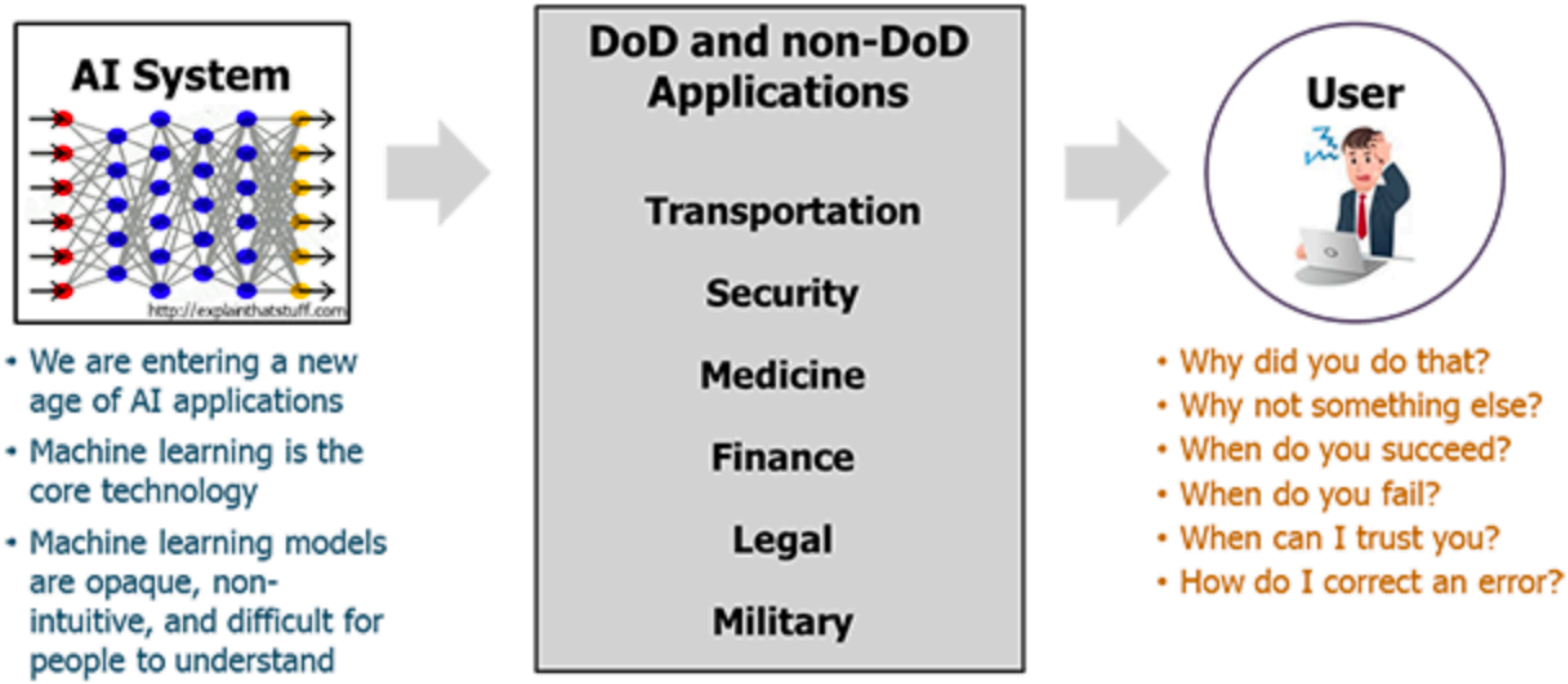
XAI in regulation: “rights to explanation”

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)

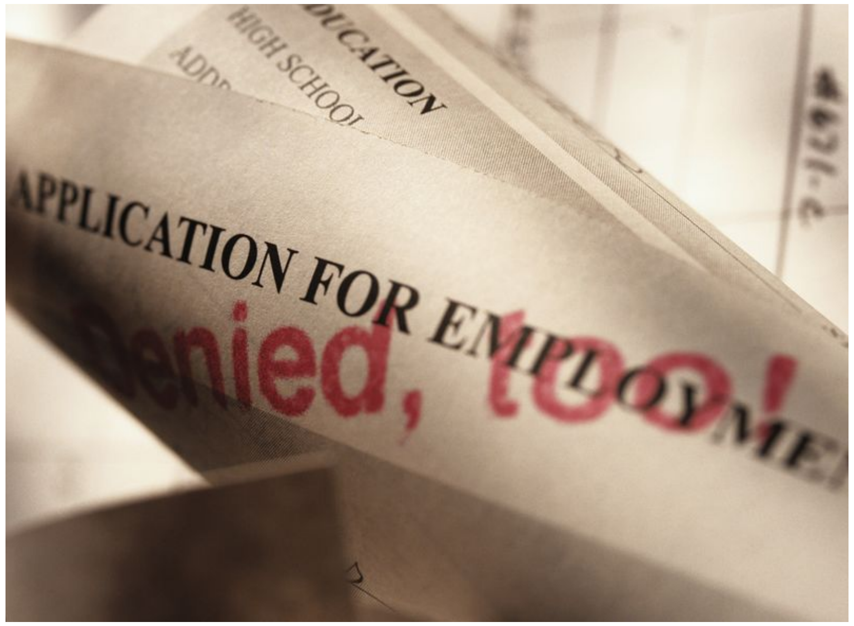
GDPR, 2016

XAI in research funding



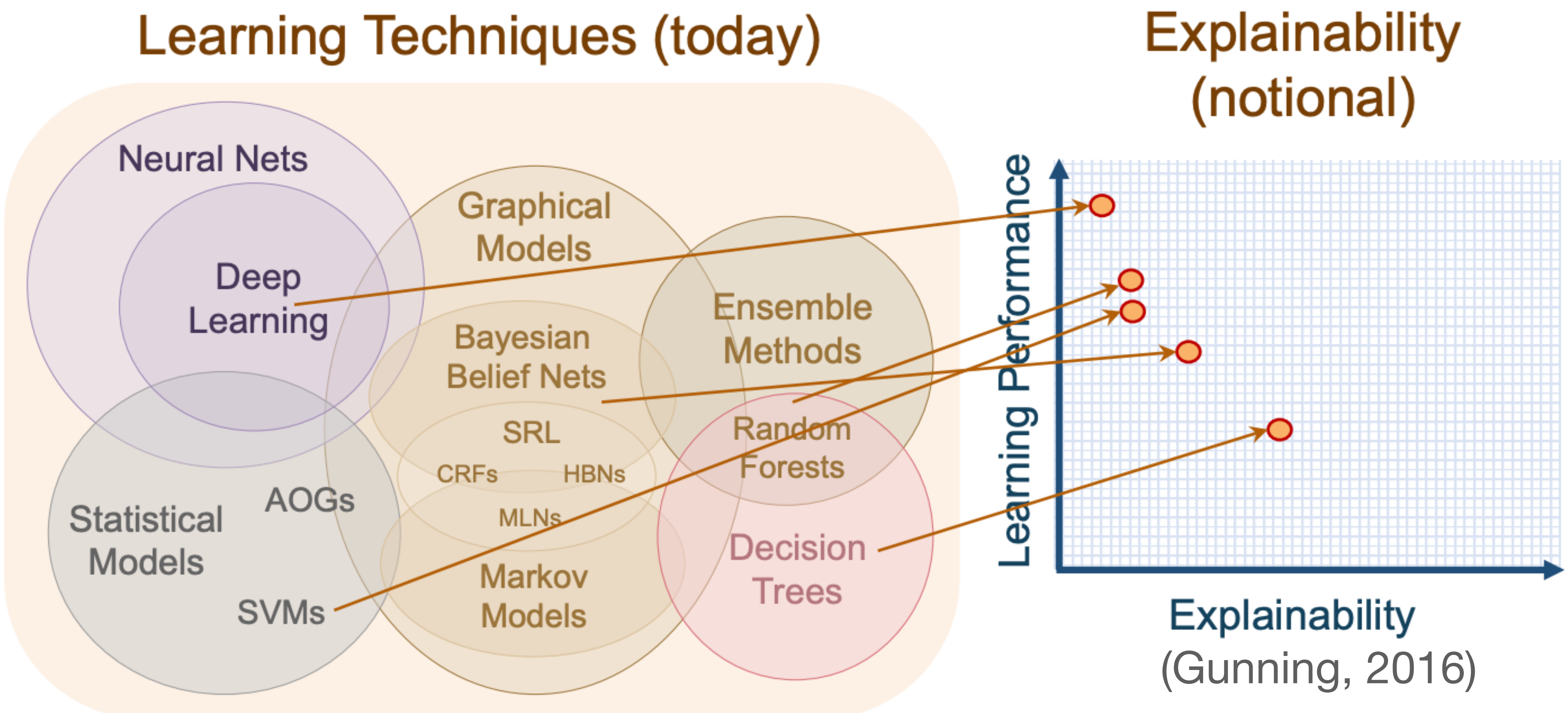
DARPA, 2016

AI is increasingly used in many high-stakes tasks



Performance-Explainability trade-off

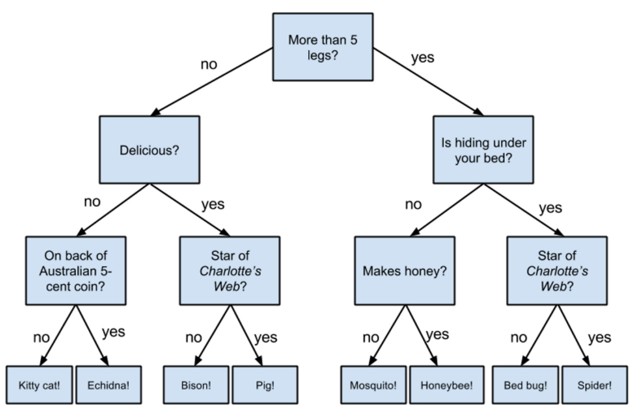
In ***average*** settings



(Gunning, 2016)

XAI

Directly explainable model

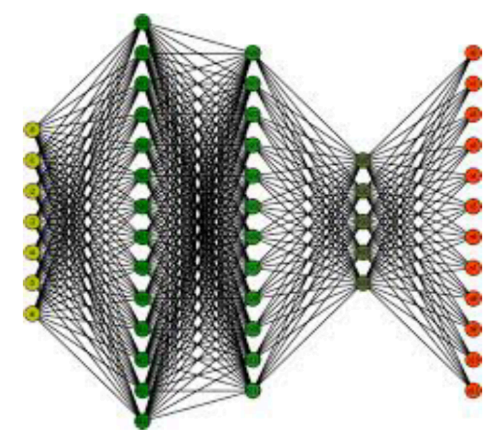


- Linear model
- Decision tree
- Rule-based model

Breaking the trade-off

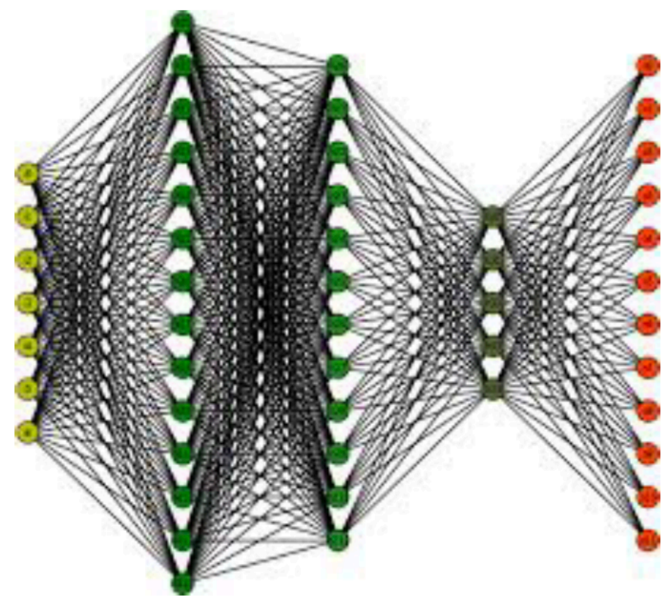
- Generalized linear rule model
- Generalized additive models
- ...

Post-hoc explainability

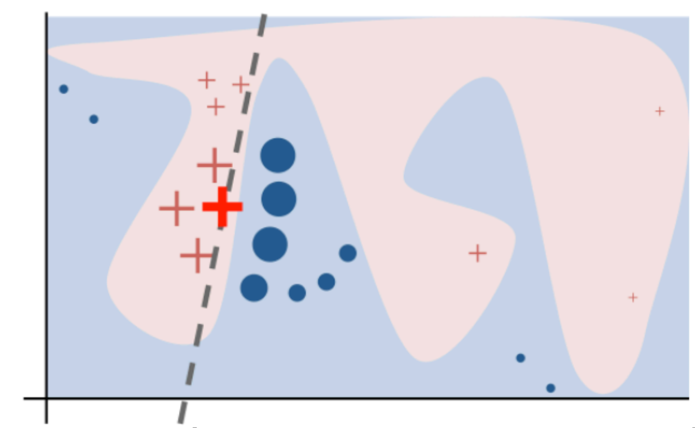


- Deep neural networks
- Ensemble models

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

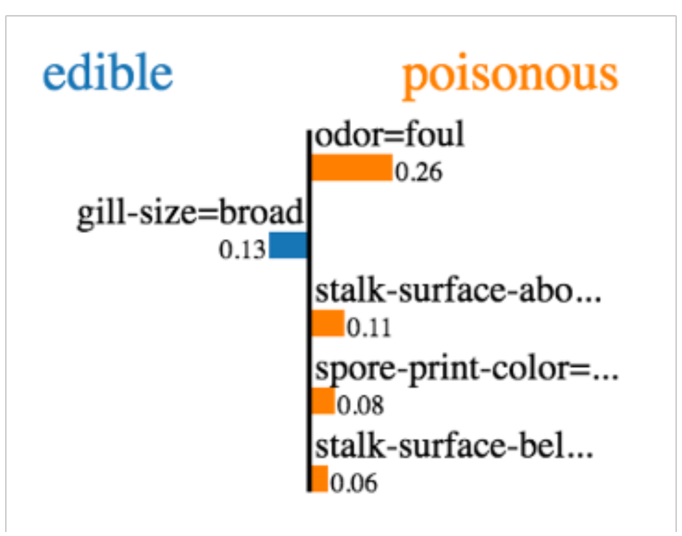


Neural network, not directly explainable

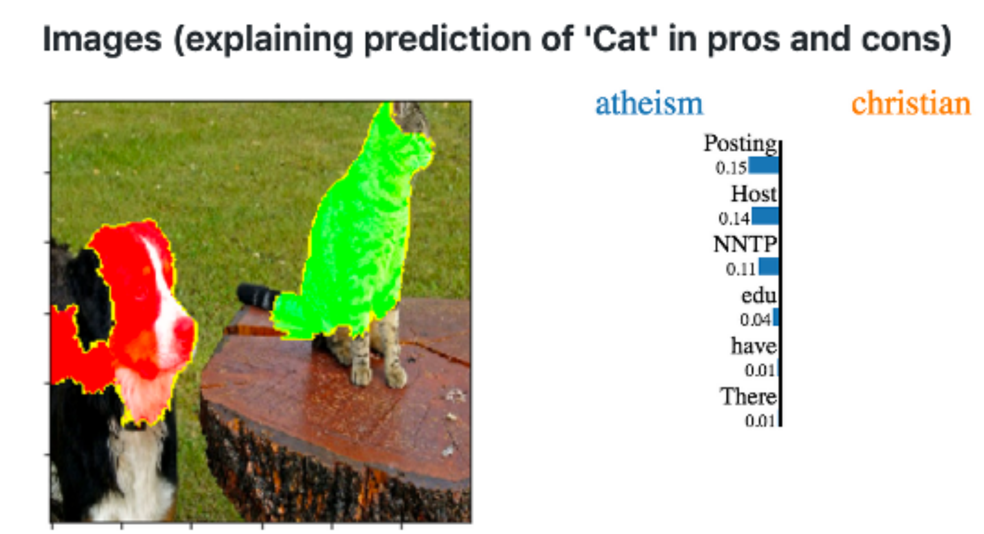


LIME (Ribeiro et al. 2016)

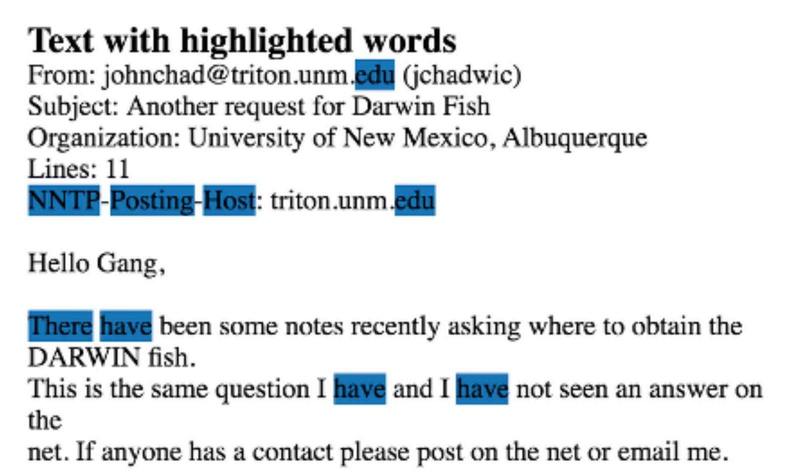
Use a *post-hoc* XAI technique



Tabuler data



Image



Texts

XAI

Transparent model

- Generalized linear rule model
- Generalized additive models
- …

Post-hoc explainability

Explaining the model (global)

- Feature importance
- Rule approximation
- Decision tree approximation

Explaining a decision (local)

- Local contribution
- Local rules
- Similar instances

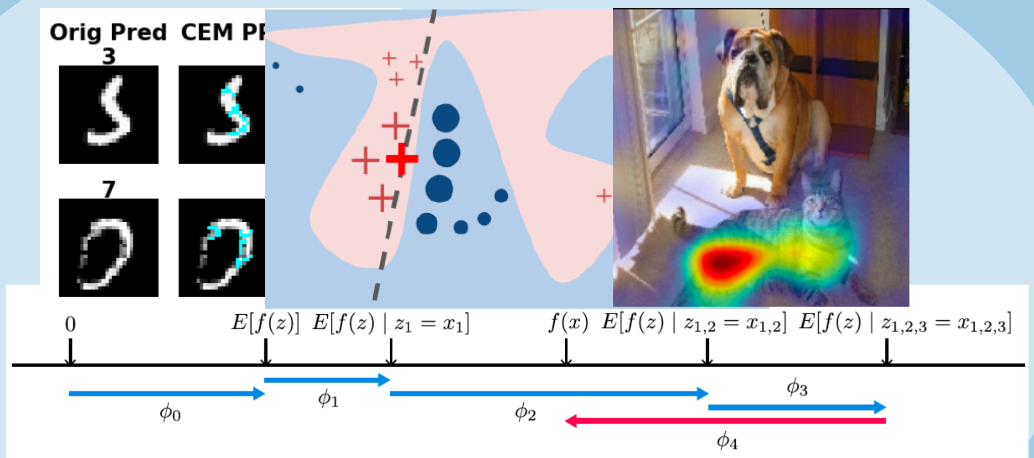
Inspecting counterfactual

- Feature influence
- Contrastive features
- Counterfactual instances

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

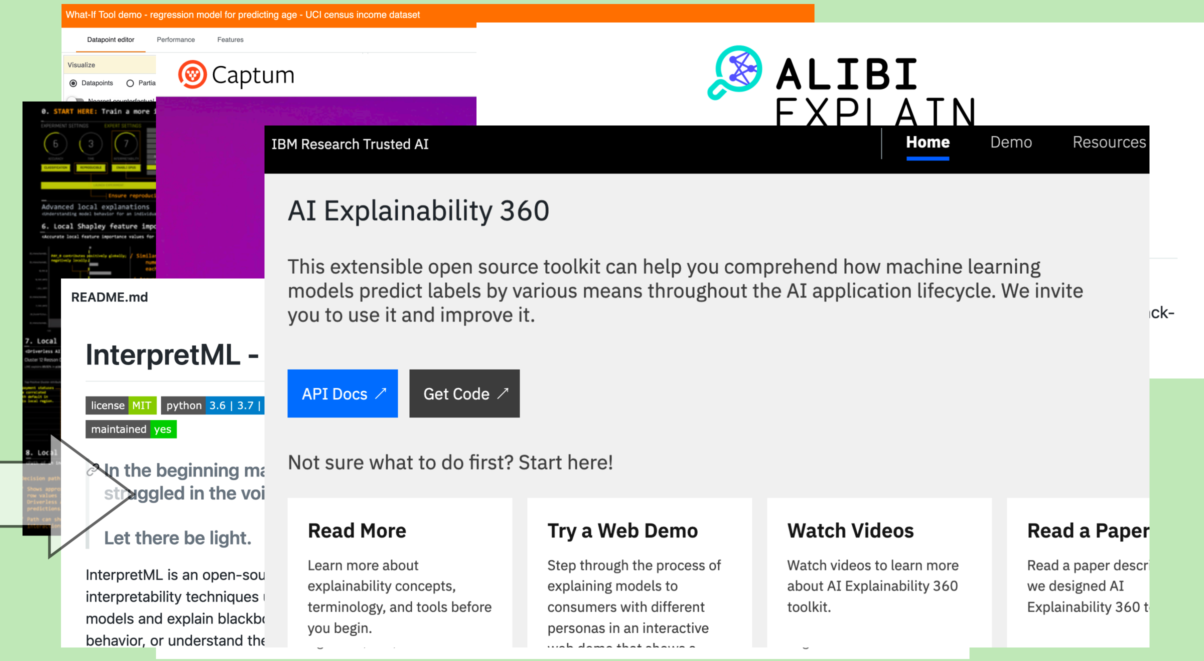
A growing collection of XAI techniques

XAI in Academia



An abundance of XAI algorithms

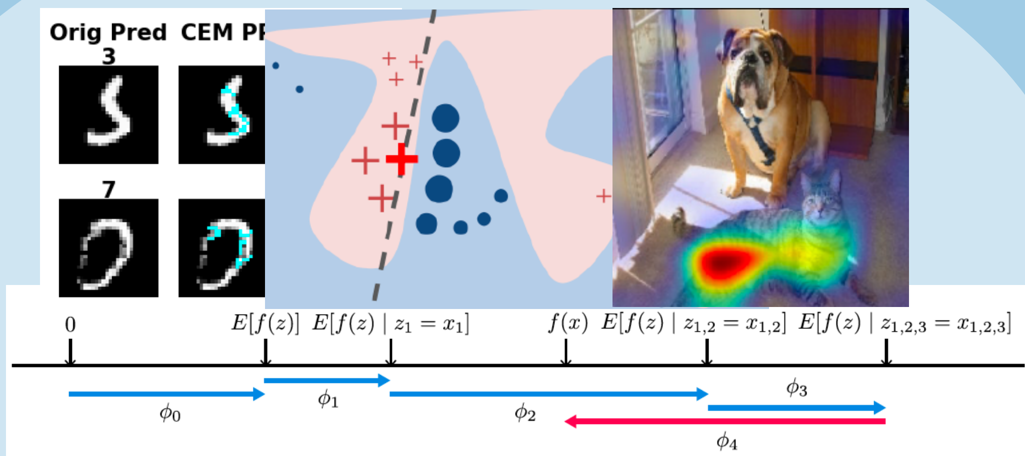
XAI in Practice



Toolbox of XAI techniques

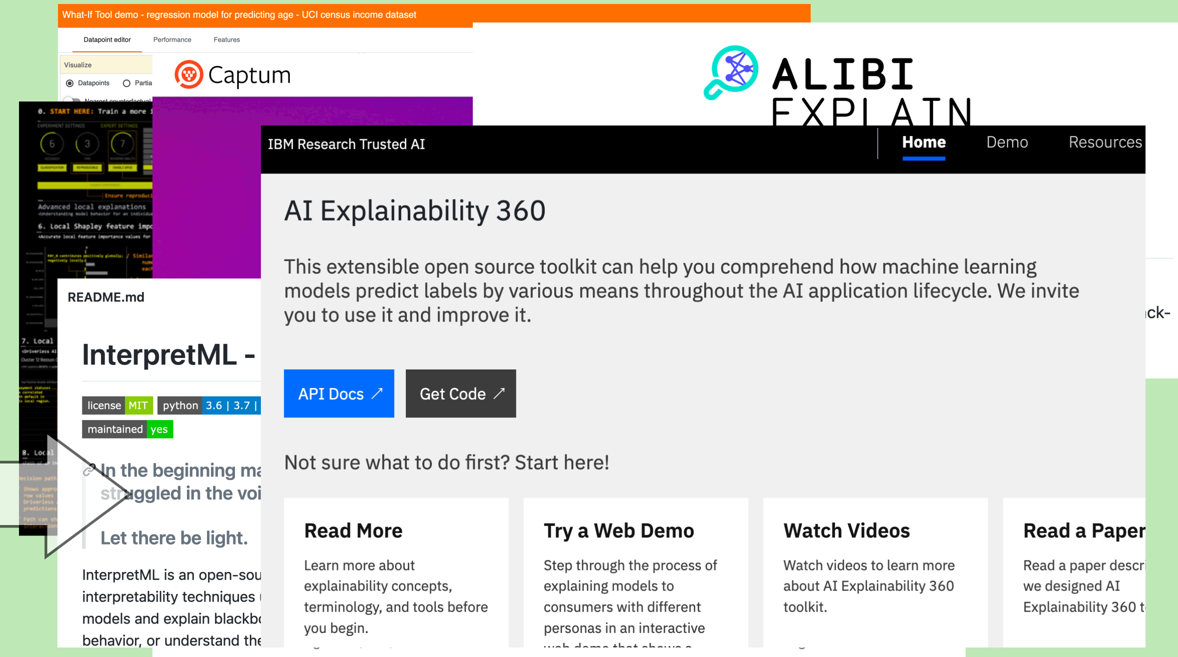
From academic research into a practitioners' toolbox

XAI in Academia



An abundance of XAI algorithms

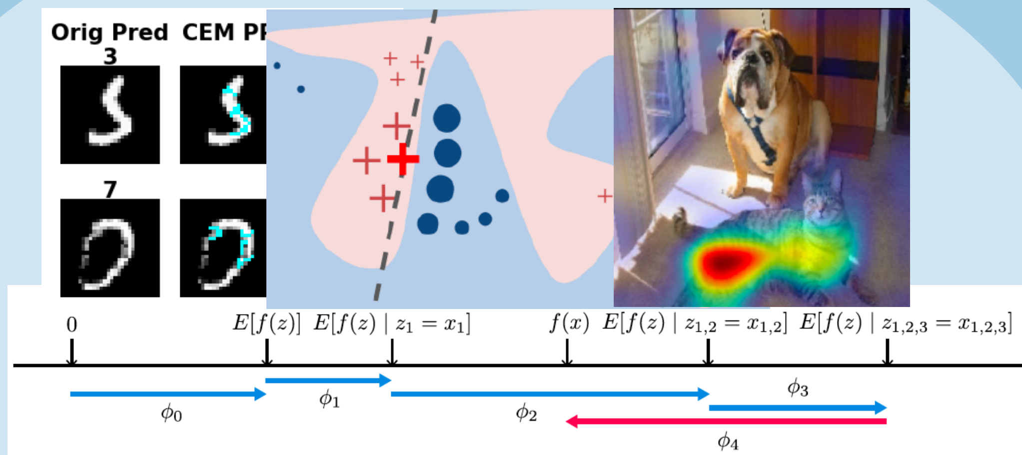
XAI in Practice



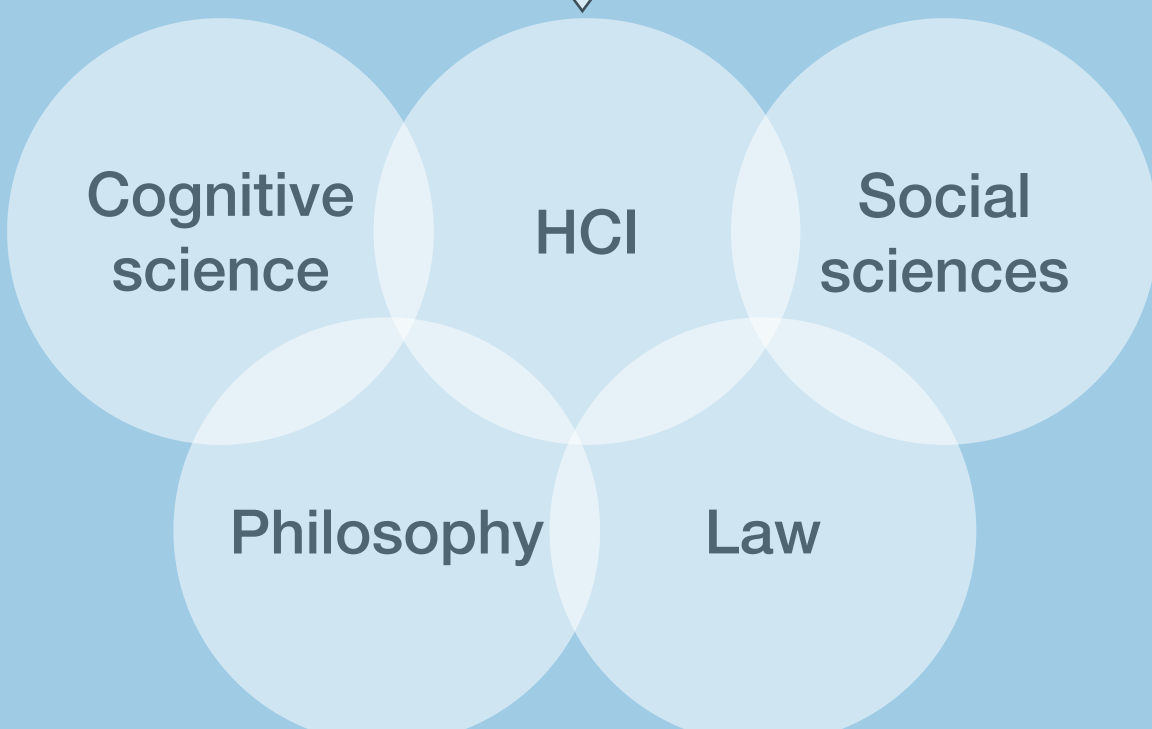
Toolbox of XAI techniques

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

XAI in Academia

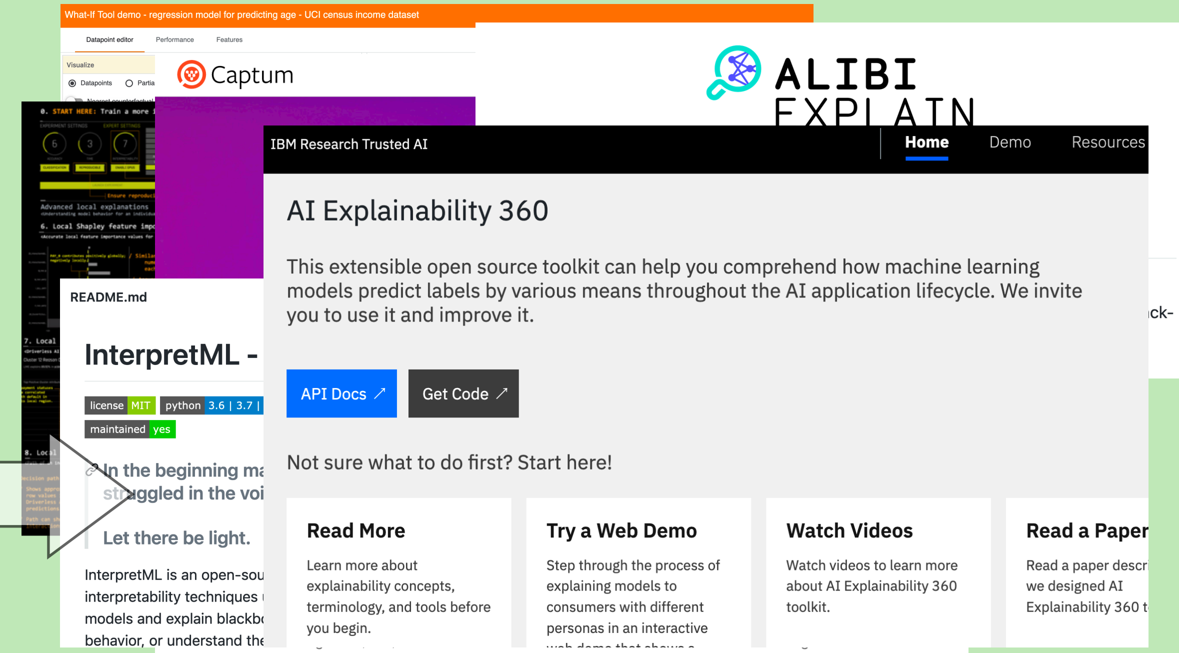


An abundance of XAI algorithms



Inter-disciplinary perspectives

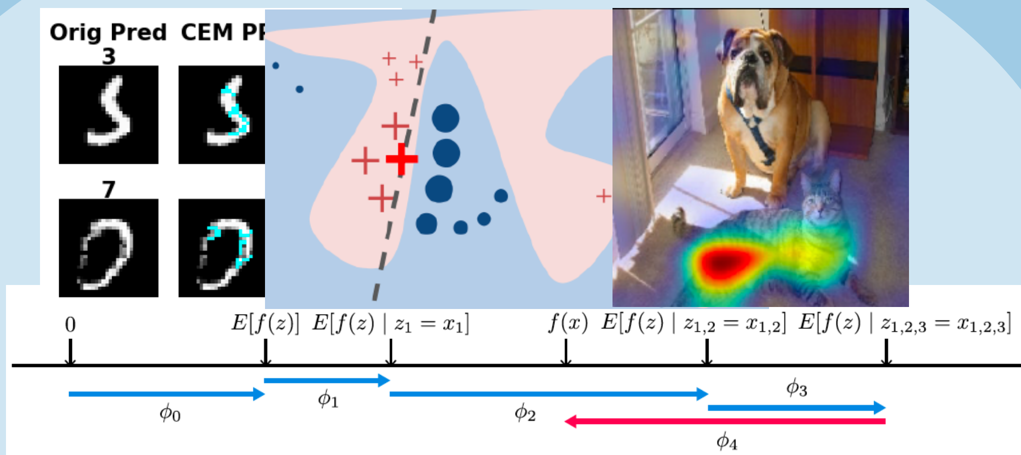
XAI in Practice



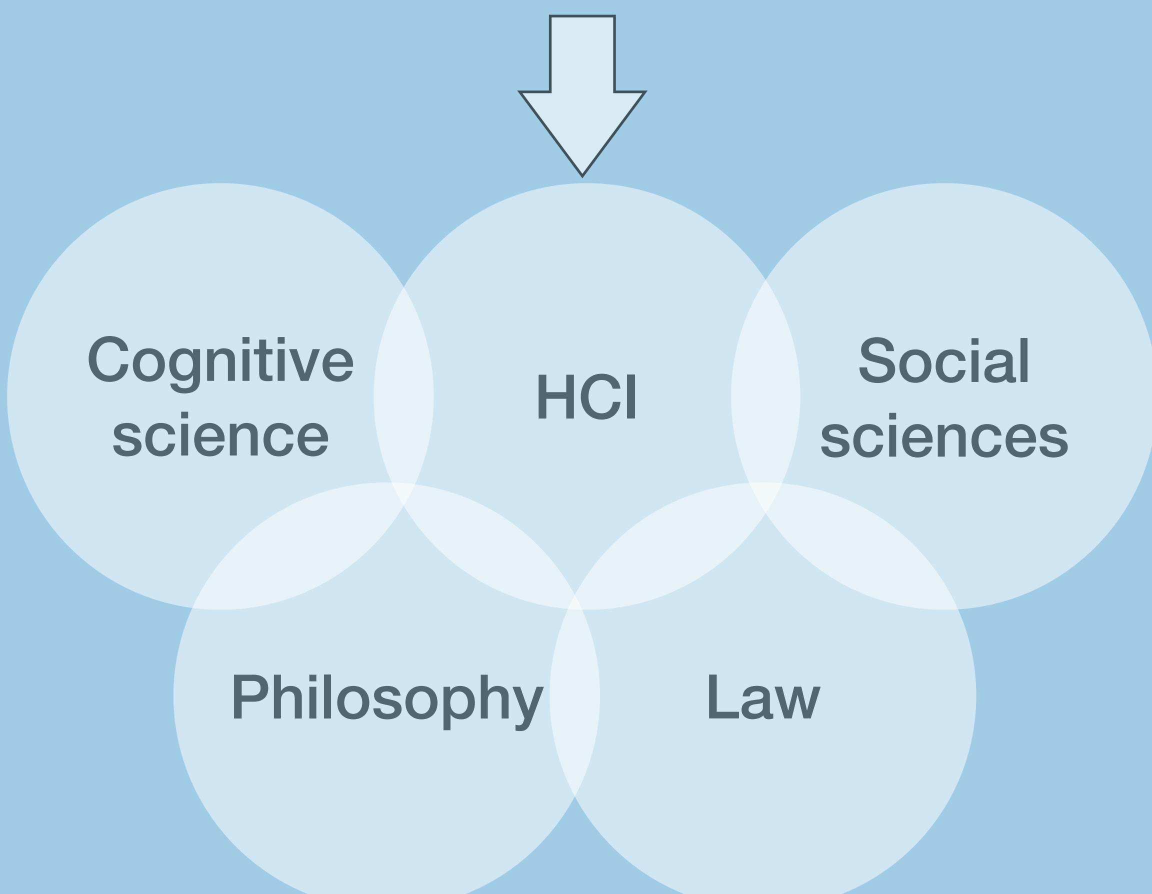
Toolbox of XAI techniques

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

XAI in Academia



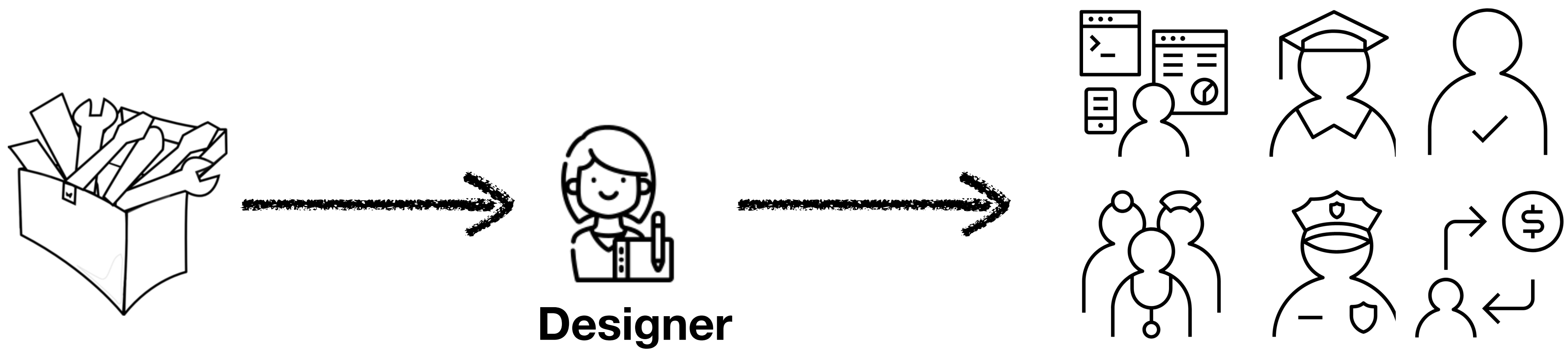
An abundance of XAI algorithms

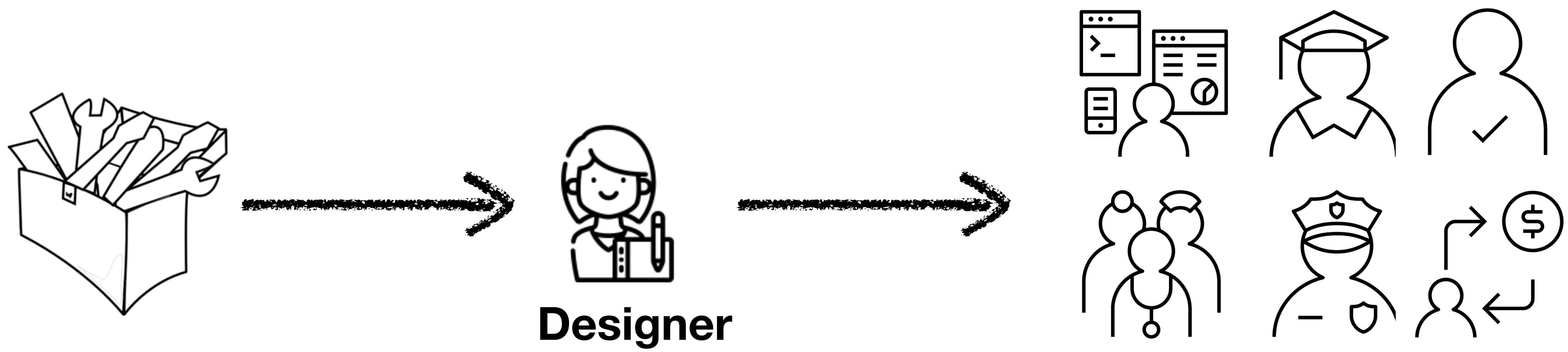


Inter-disciplinary perspectives

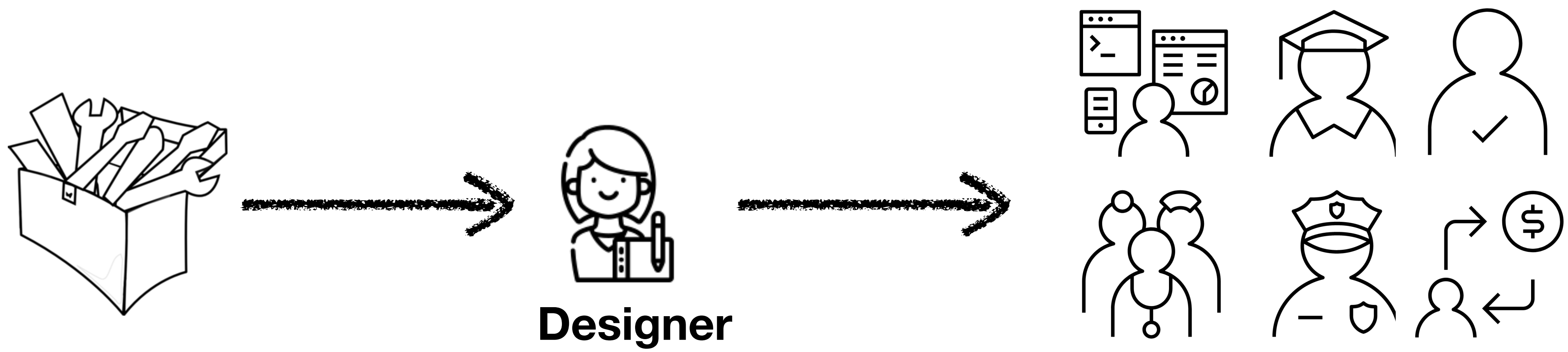
Inter-disciplinary perspectives

- **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)


Designer



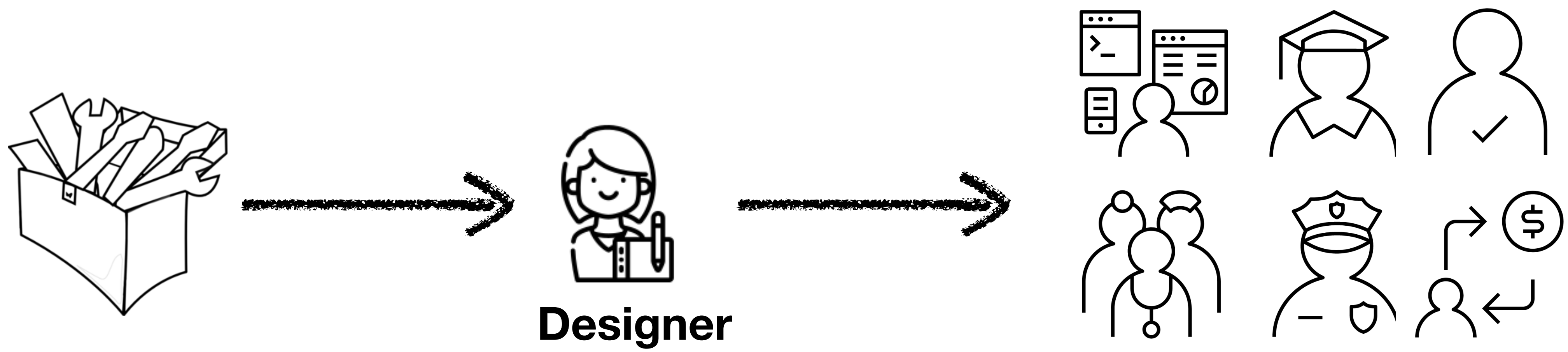
How to **select**?   How to **translate**?



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

Thread 1: Study and support design practices for XAI UX

Thread 2: HCI research with XAI use cases

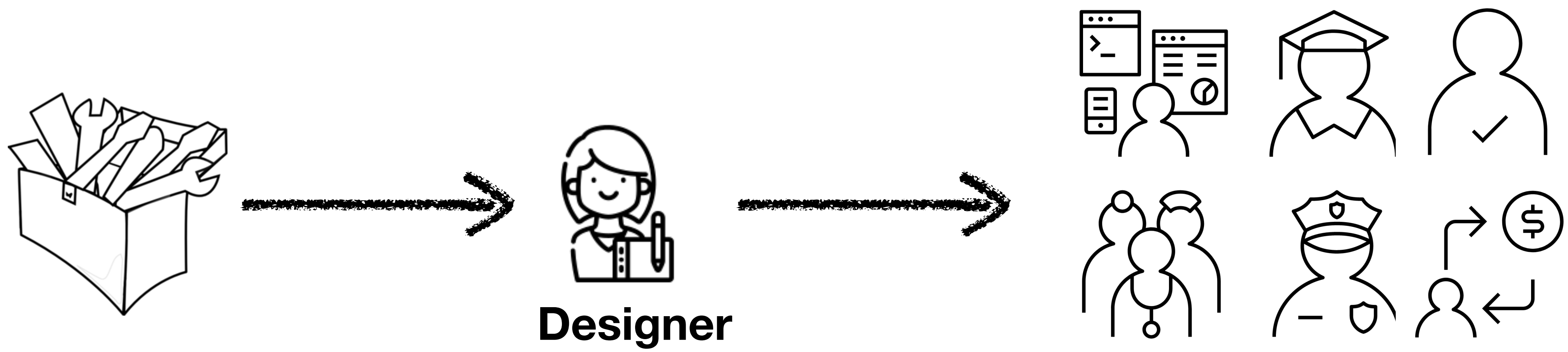


How to **select**? How to **translate**?

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*?)



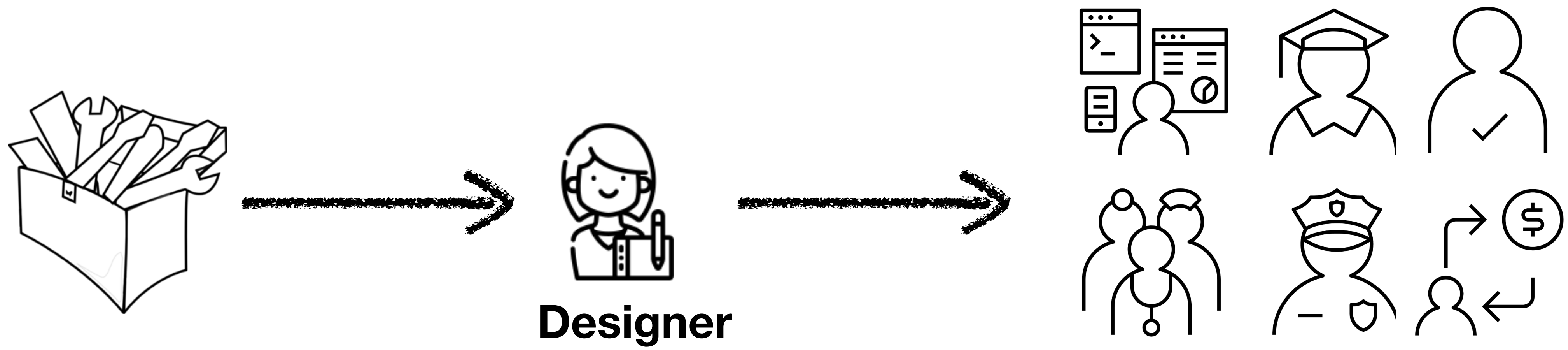
How to **select**?   How to **translate**?

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

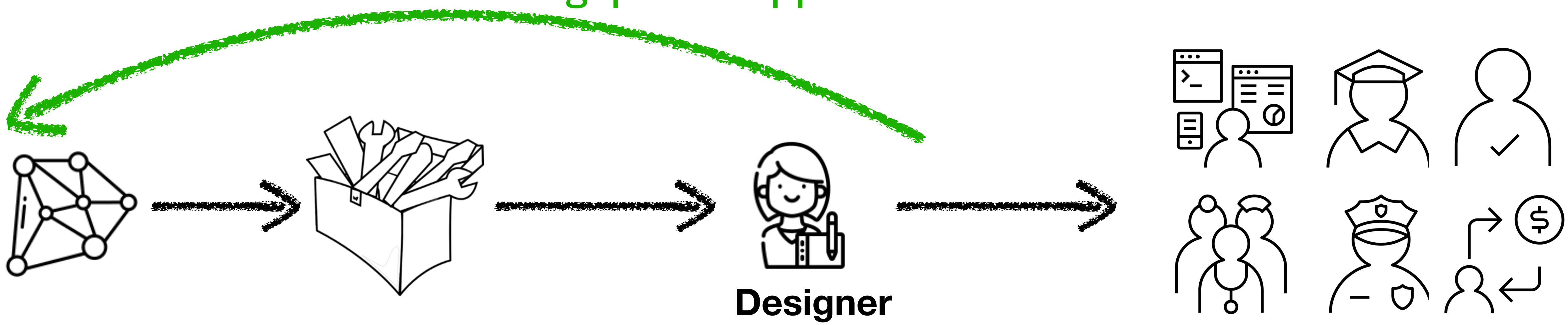
Suitability for different usage contexts (*What contexts*?)

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?

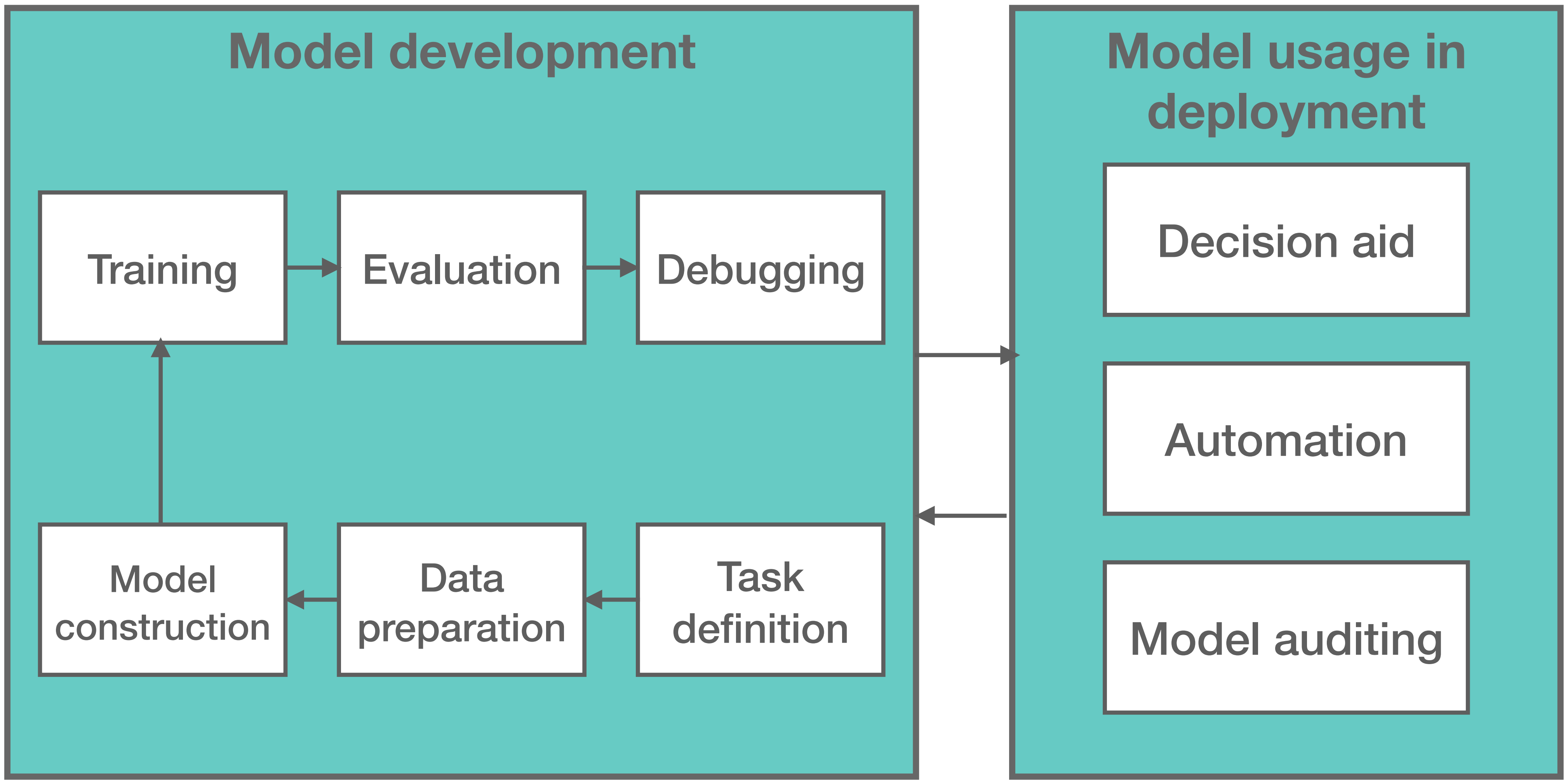
Thread: HCI Research with XAI Use Cases

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!

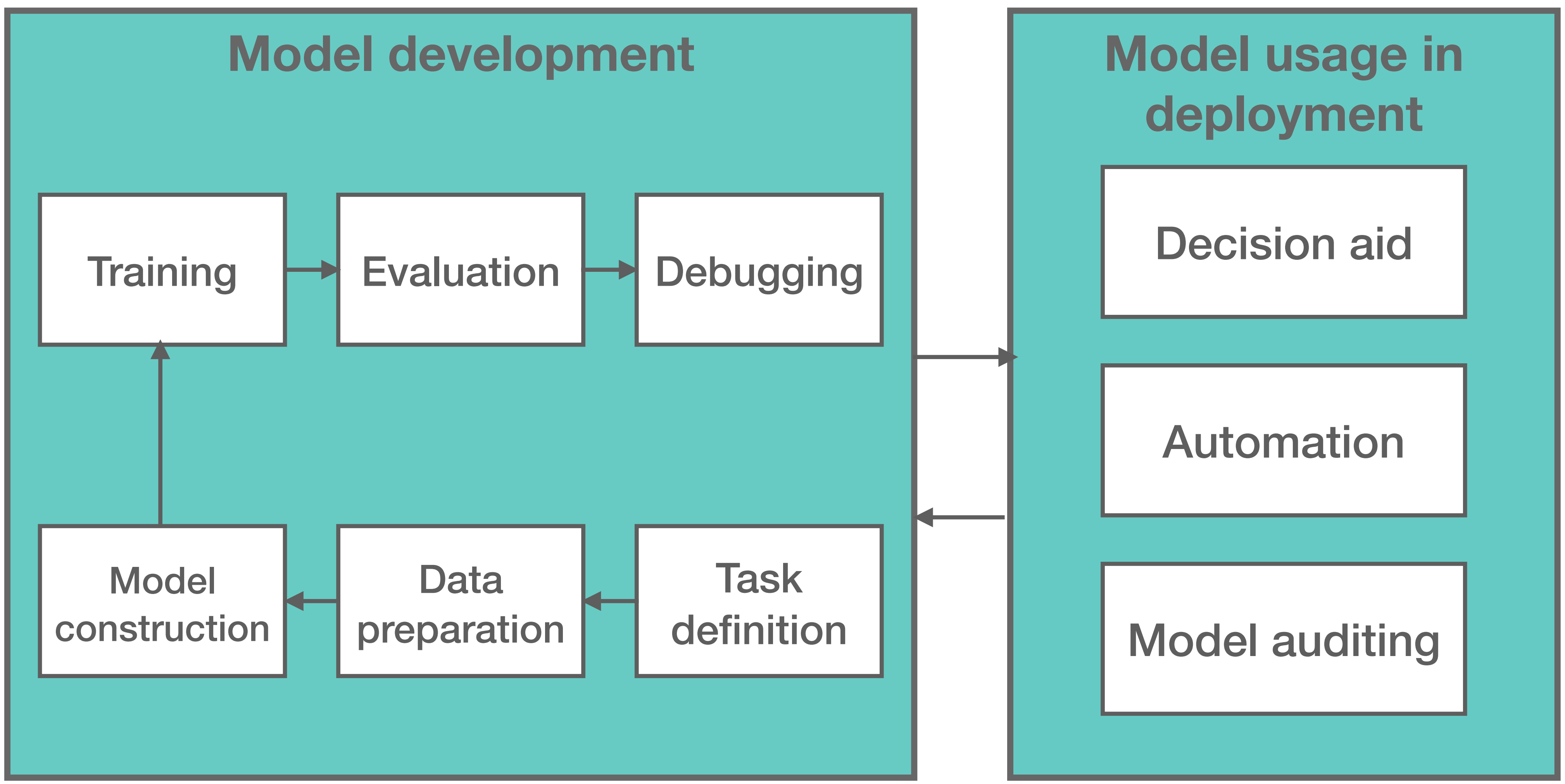
XAI use cases in AI lifecycle



XAI use cases in AI lifecycle

Model debugging or selection (IUI2021)
XAI user: Data scientist

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



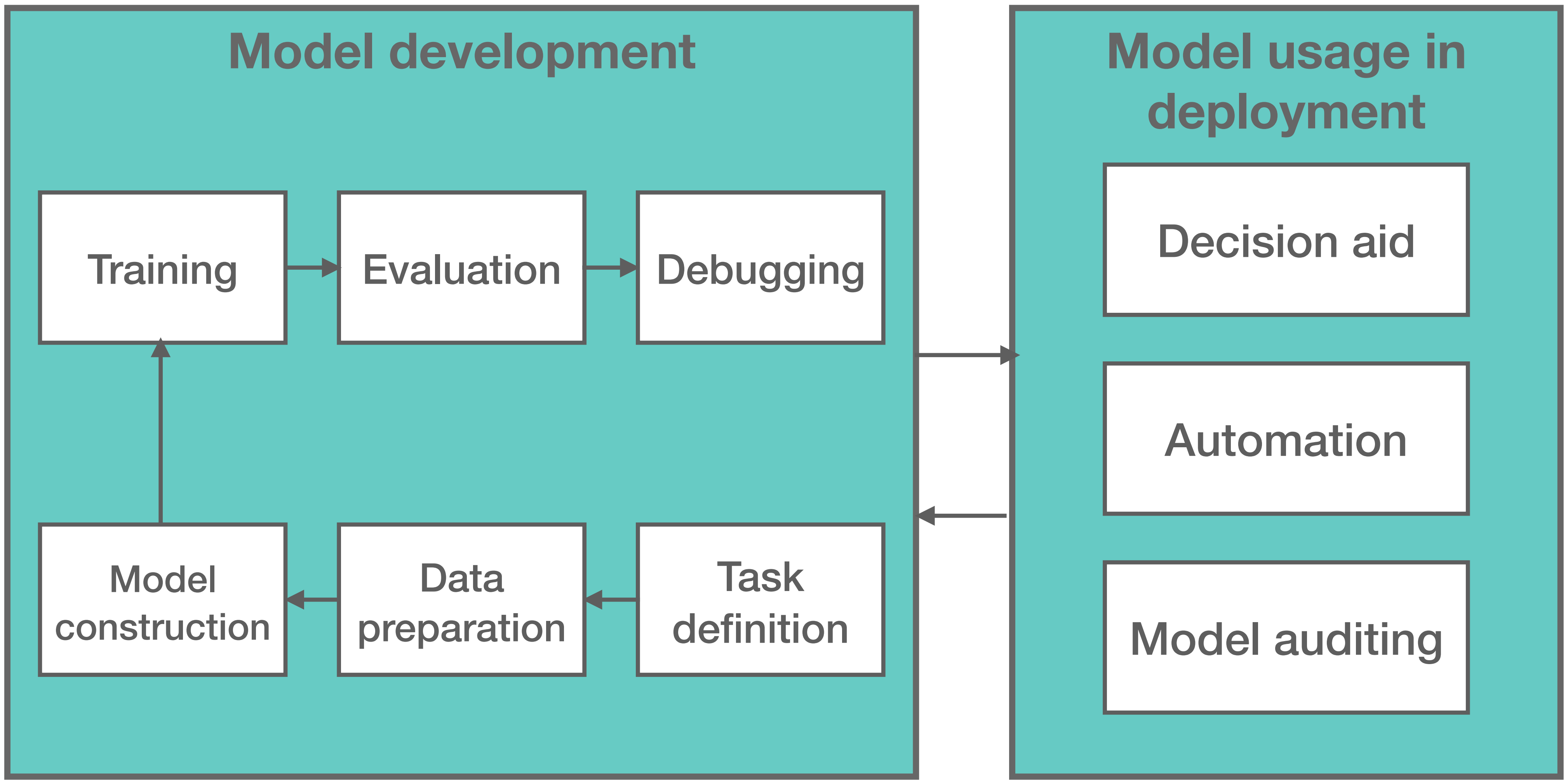
Delegation support (ongoing)
XAI consumer: Domain expert

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

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

XAI use cases in AI lifecycle

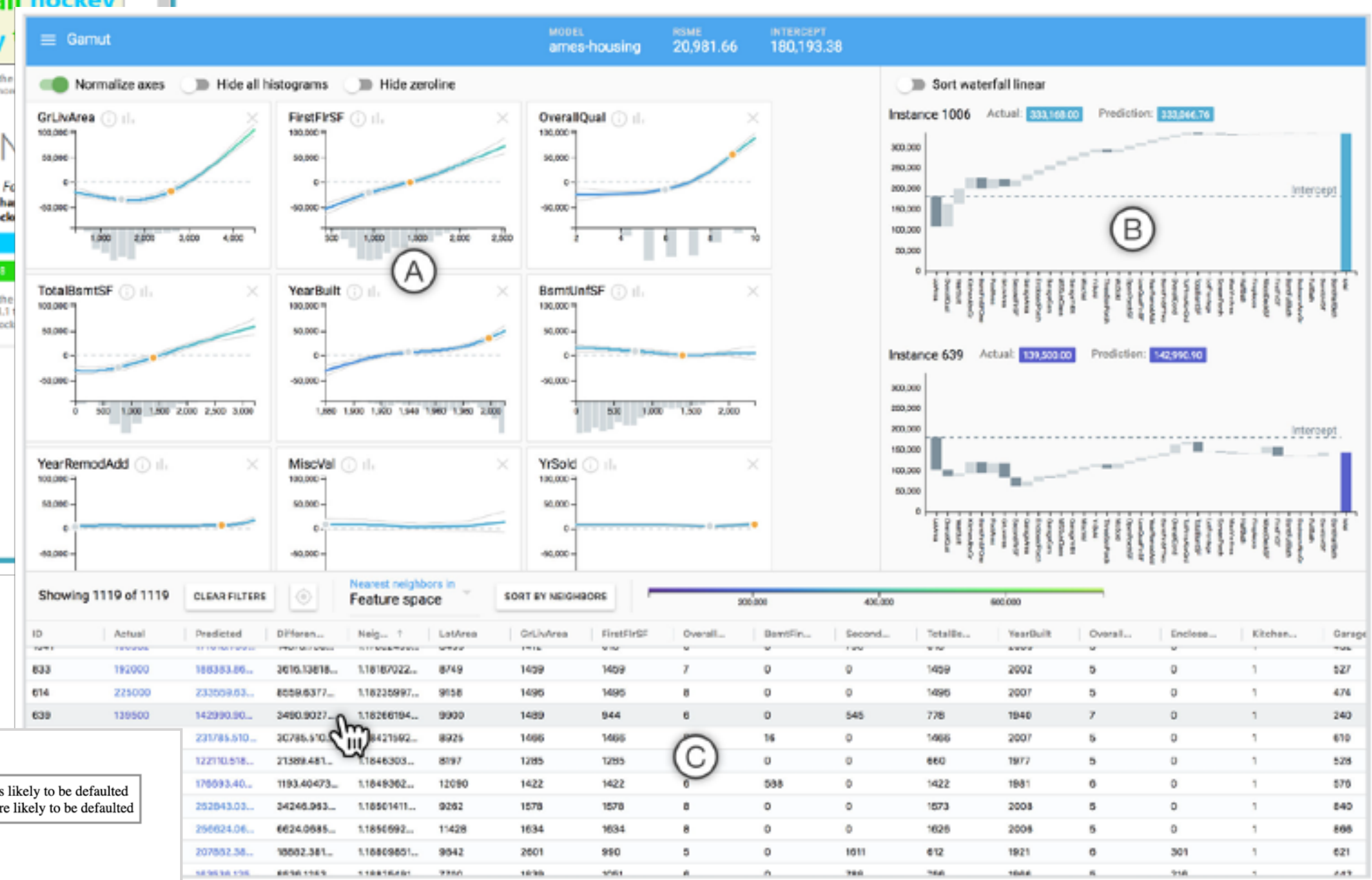
Model debugging or selection (IUI2021)
XAI user: Data scientist



XAI for model debugging and selection

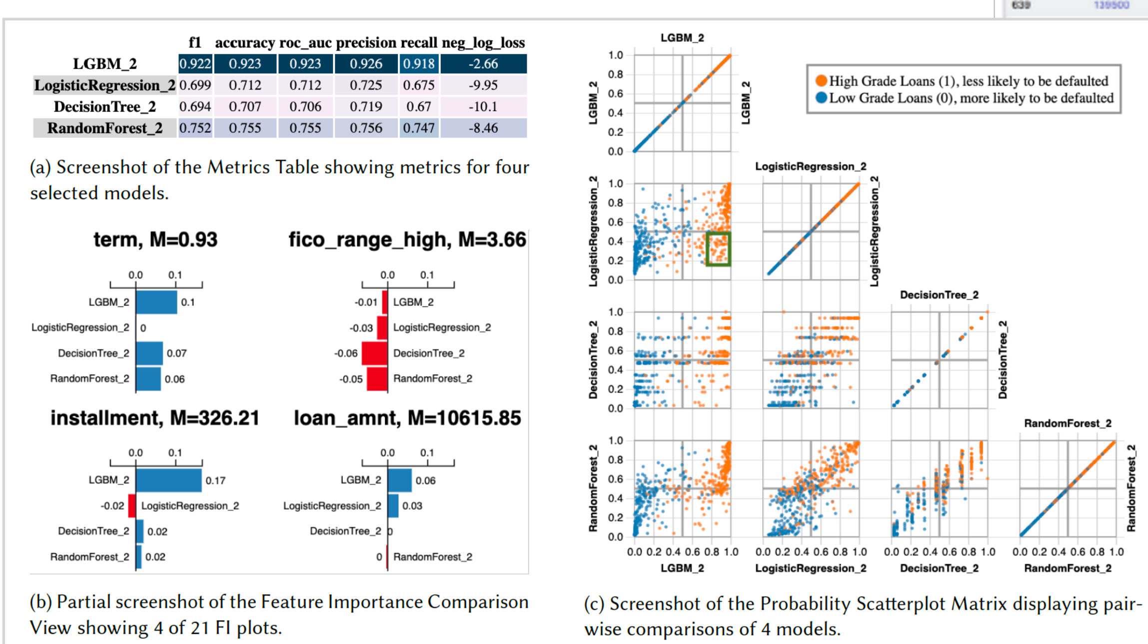
Explanatory debugging

(Kulesza et al, 2015)



GAMUT

(Hohman et al, 2019)



	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.

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

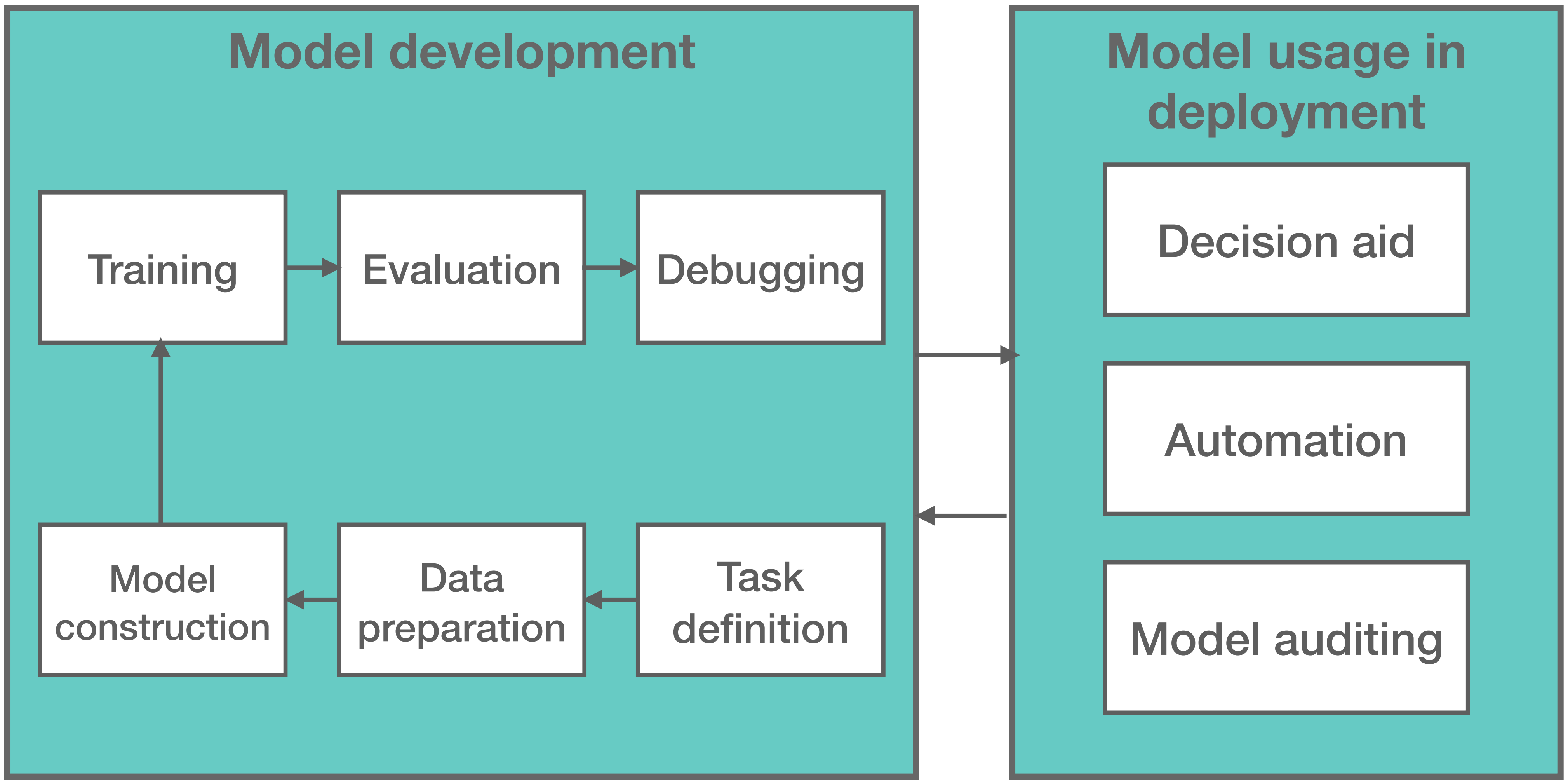
(c) Screenshot of the Probability Scatterplot Matrix displaying pair-wise comparisons of 4 models.

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

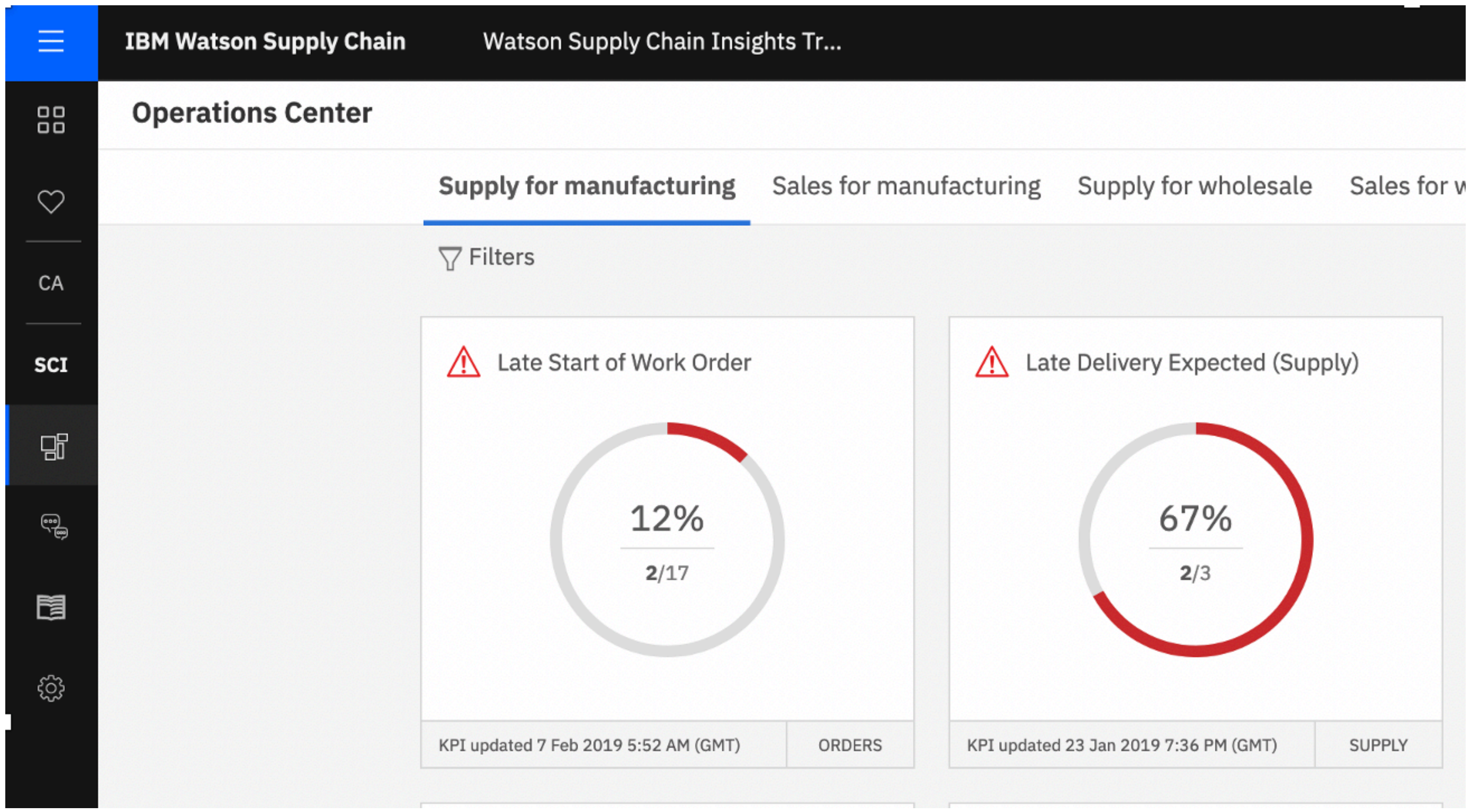
XAI use cases in AI lifecycle

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 (I-5)

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

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)

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

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

Caveat: Explanation can lead to unwarranted trust!

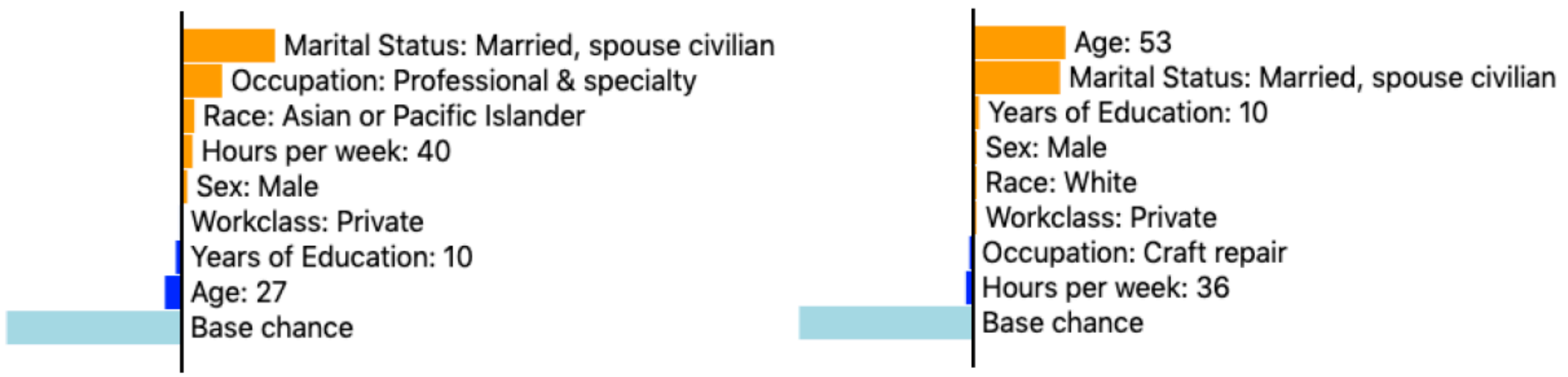
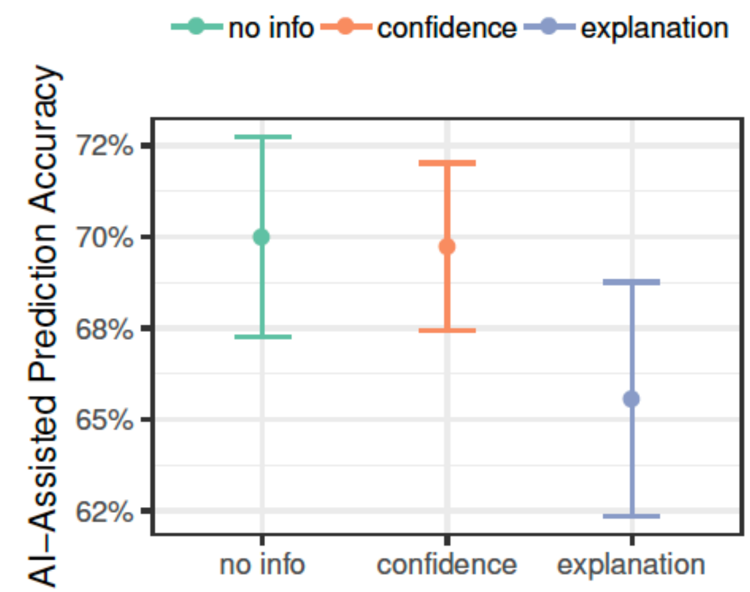


Figure 11: Screenshots of explanation for cases where the model had low confidence.



Can I trust this prediction?

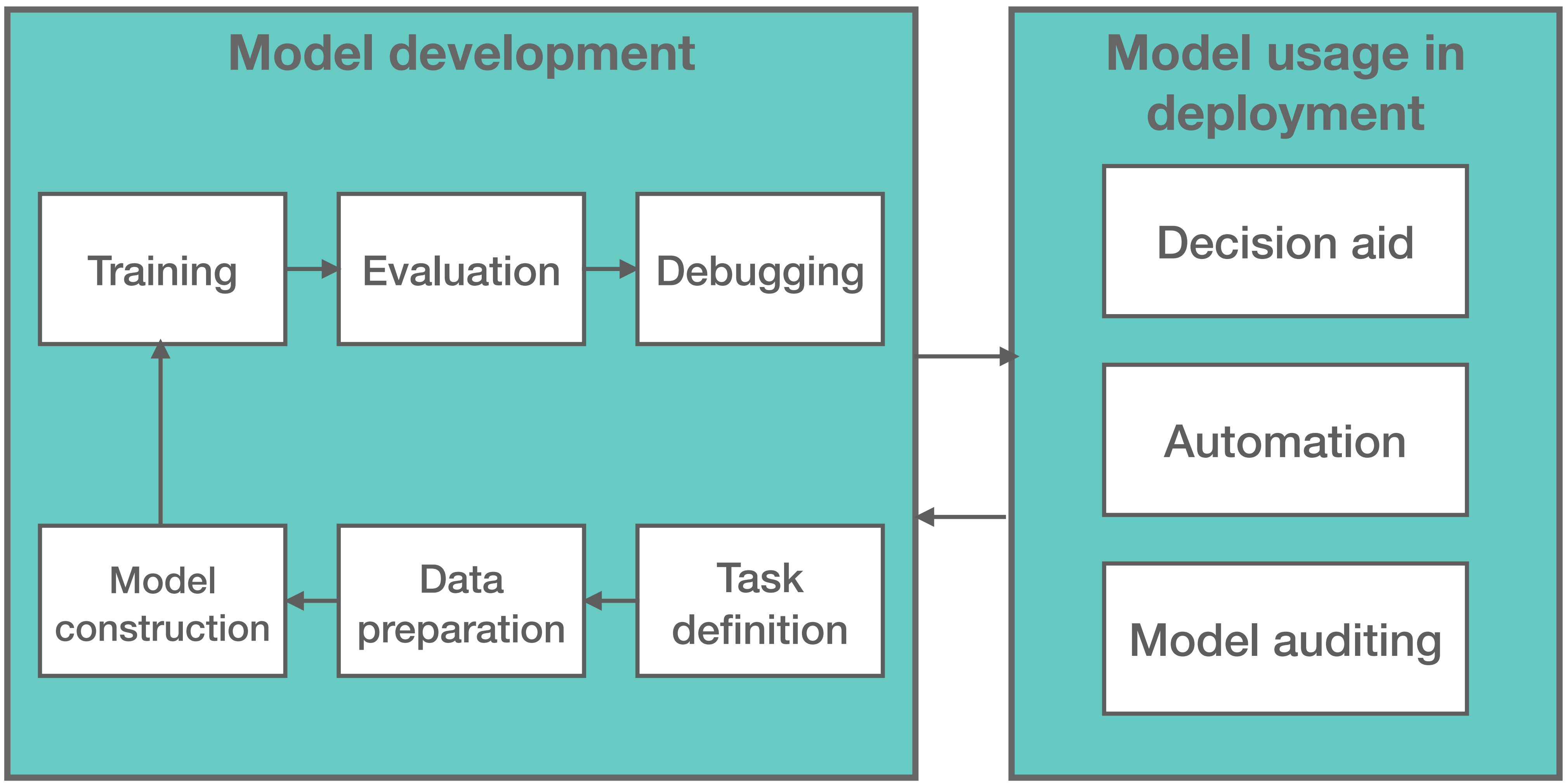
Decision-makers

Zhang et al. [Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making](). *FAT* 2020*
Poursabzi-Sangdeh,et al.. Manipulating and measuring model interpretability. *CHI 2021*
Bansal et al. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. *CHI 2021*

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

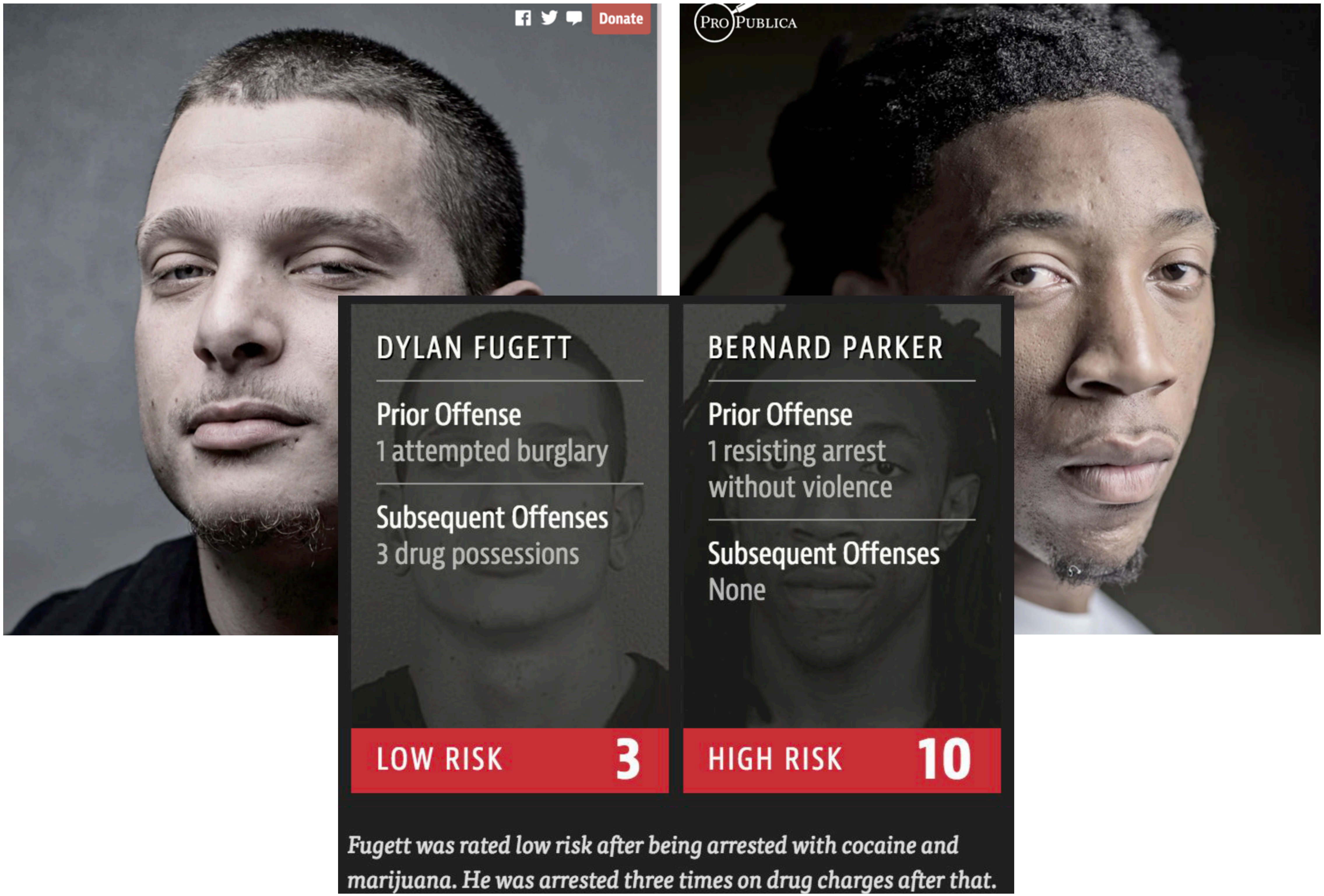
Fair ML: What is unwanted bias?

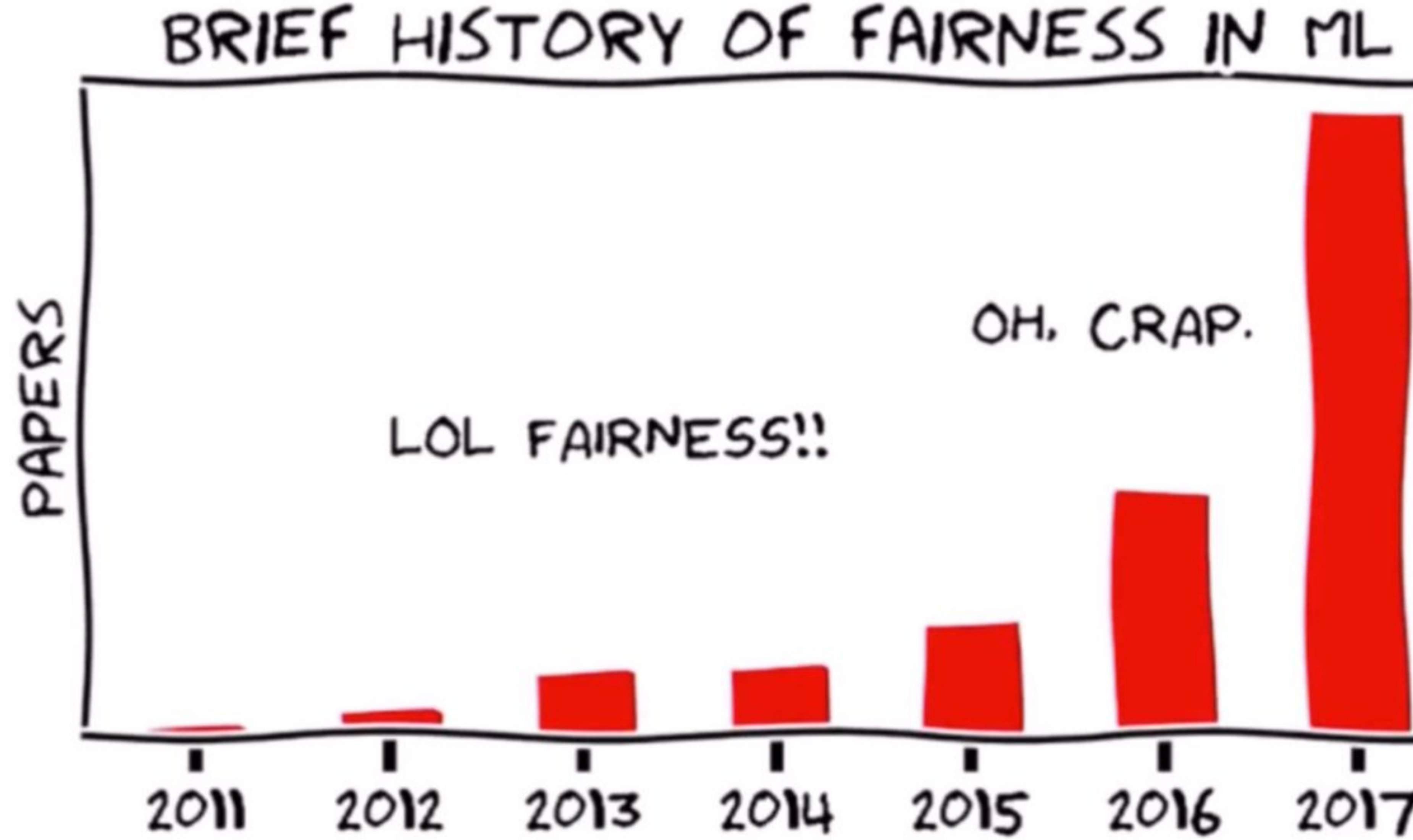
Discrimination becomes objectionable when it places certain **unprivileged** groups at a systematic disadvantage

Illegal in certain contexts

(Barocas and Selbst, 2017)

Discrimination in COMPAS




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

(Hardt, 2017)

Suitability

XAI as interfaces for scrutinizing discrimination

Contrastive

- Iliana's race is **African American**.

If it had been **Caucasian**, she would have been predicted as NOT likely to reoffend

- Iliana's age is **18-29**.

If it had been **older than 39**, she would have been predicted as NOT likely to reoffend

Feature importance

The more +s/-s means a person with that attribute is more/less likely to re-offend.

* Appears next to Iliana's attributes

Race

•Caucasian (0)

•* **African-American (+)**

Age

•* **18-29 (++++)**

•30-39(+)

•…

Charge degree:

•…

Number of prior convictions

Has juvenile priors:

Defendant: Iliana

- Race: African-American
- Age: 18-29
- Charge degree: Misdemeanor
- Prior convictions: 0
- Has juvenile priors: Yes

Prediction:

Likely to reoffend

Example-based

The training set contained 10 individuals identical to Iliana

6 of them reoffend (60%)

Data distribution

The prediction is based on the likelihood of previous cases with different attributes re-offended or not.

A * appears next to Iliana's features.

Race

- 40% in Caucasian race group re-offended
- *** 55% in African-American race group re-offended**

Age

- *** 58% in 18-29 age group re-offended**
- 49% in 30-39 age group re-offended
- …

Charge degree:

- …

Number of prior convictions

Has juvenile priors:

Explain a prediction: Individual fairness

Explain the model: Group fairness

Regulator

Impacted groups

Is the way the model makes risk predictions fair?

Is this person 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 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

HCXAI: "understanding" lies in the recipient

The General Data Protection Regulation (GDPR)

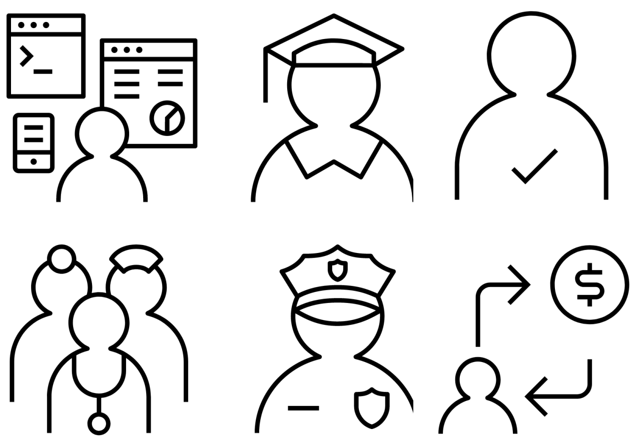
- Limits to decision-making based solely on automated processing and profiling (Art.22)
- Right to be provided with meaningful information about the logic involved in the decision (Art.13 (2) f. and 15 (1) h)

"meaningful" ???

(Nemitz, 2018)

"Understanding" lies in the recipient: beyond the toolbox

XAI techniques



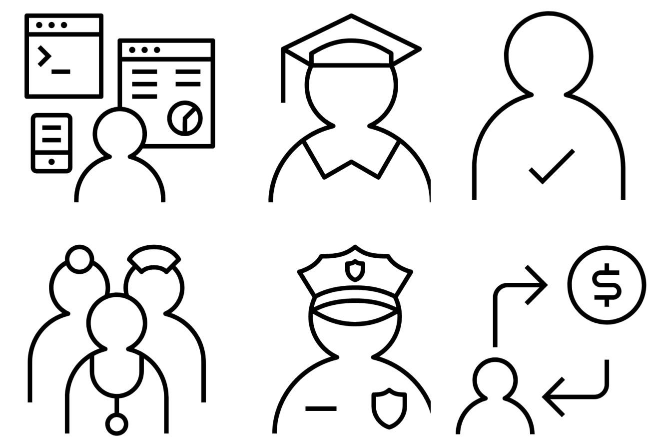
XAI UX

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

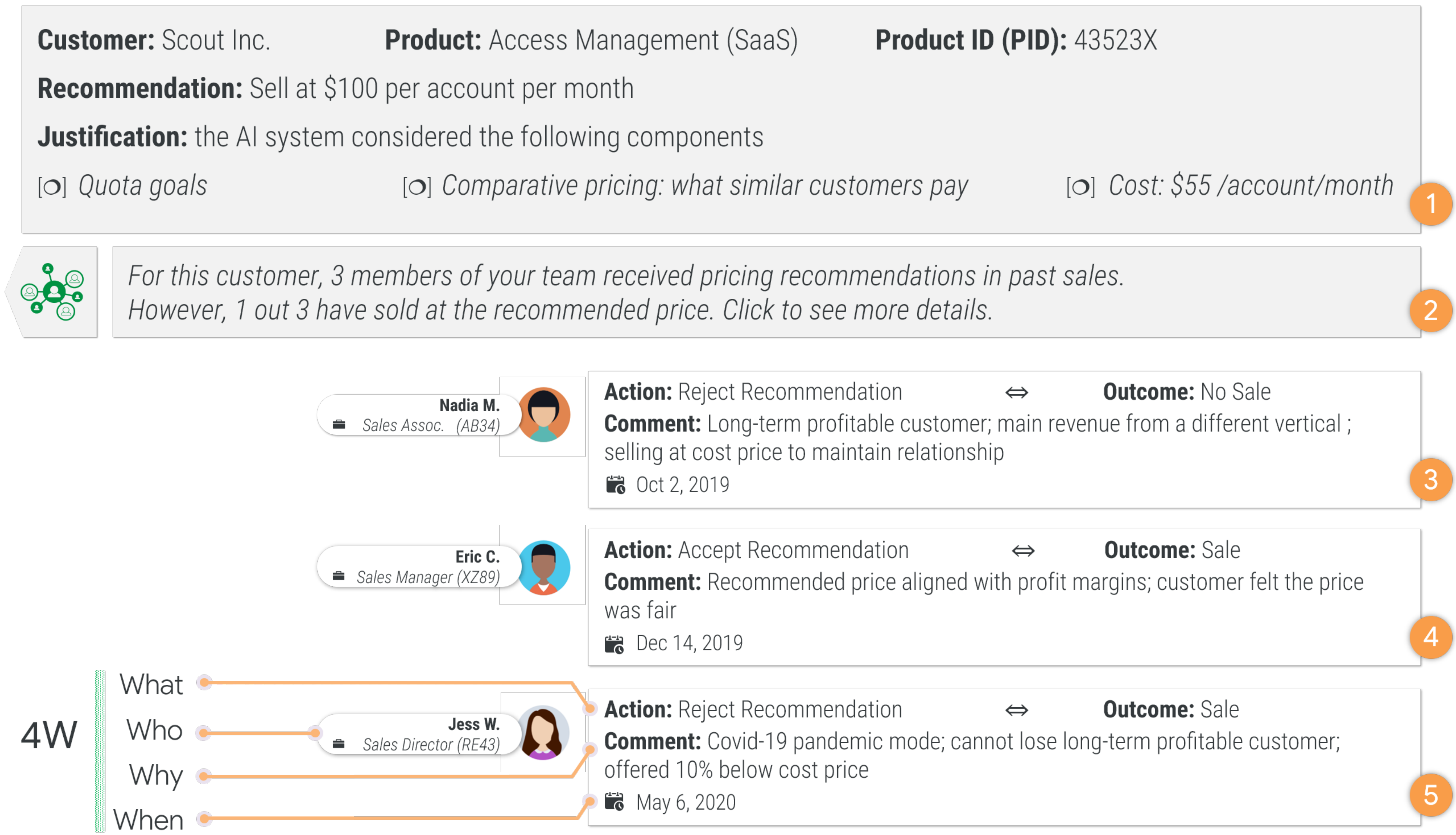
Sense-making is not just about opening the closed box of AI, but also about who is around the box, and the socio-technical factors that govern the use of the AI system and the decision. Thus the 'ability' in explainability does not lie exclusively in the guts of the AI system

Information needs to achieve understanding of AI:

- General AI knowledge gaps
- Domain knowledge gaps
- "Socially situated understanding"

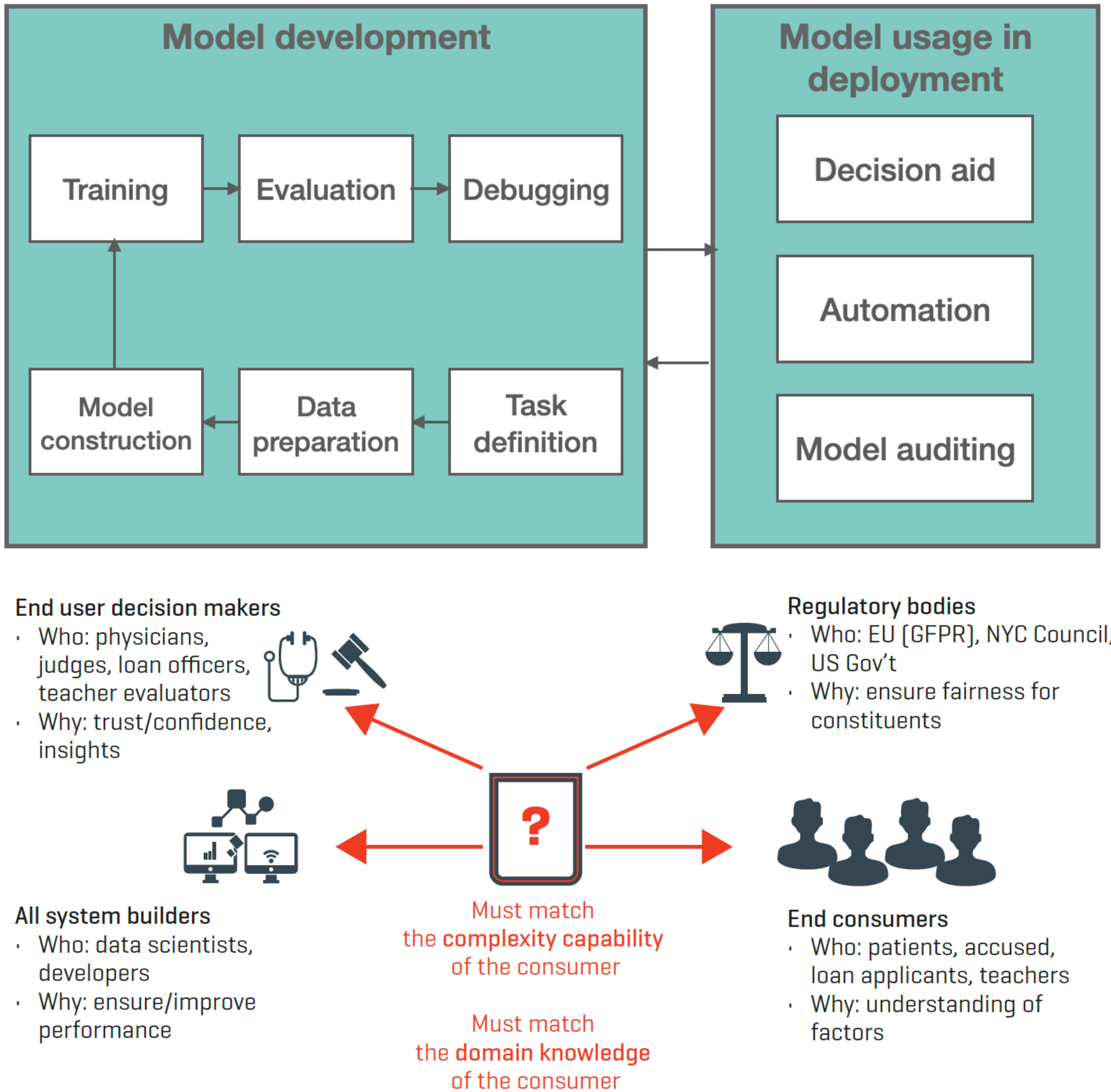
Ehsan et al. [Expanding Explainability: Towards Social Transparency in AI systems](). To appear in CHI 2021

Towards “social transparency” in AI systems

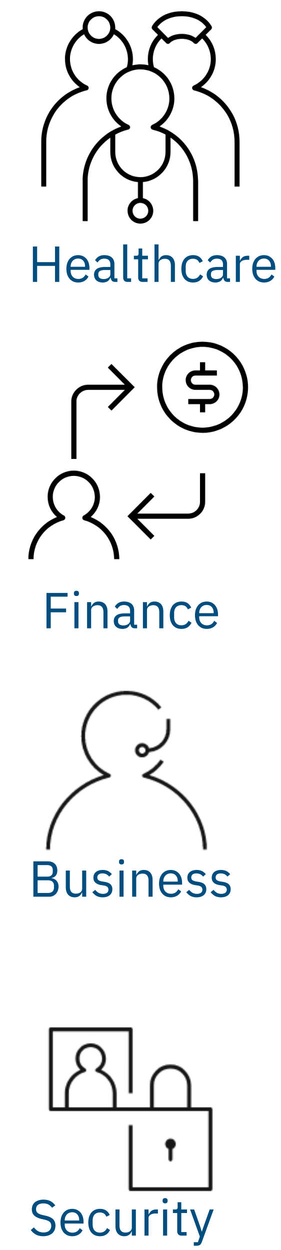


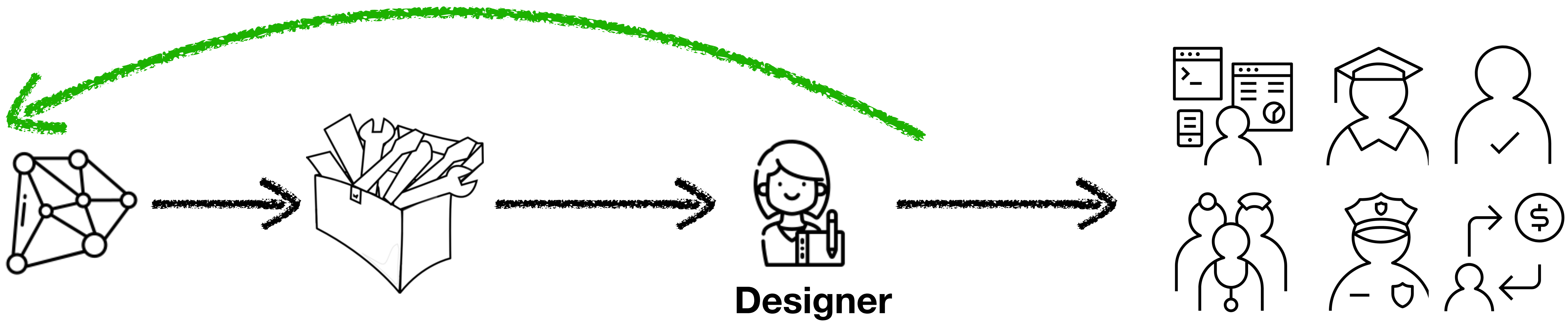
Ehsan et al. [Expanding Explainability: Towards Social Transparency in AI systems](). To appear in CHI 2021

Many user objectives + user groups + domains + social contexts



(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?

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 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

+

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)?

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?

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 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?

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?

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 differen prediction Q?
- What kind of instance is predicted of [a different outcome Q]?

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

- **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)

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

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

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*

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

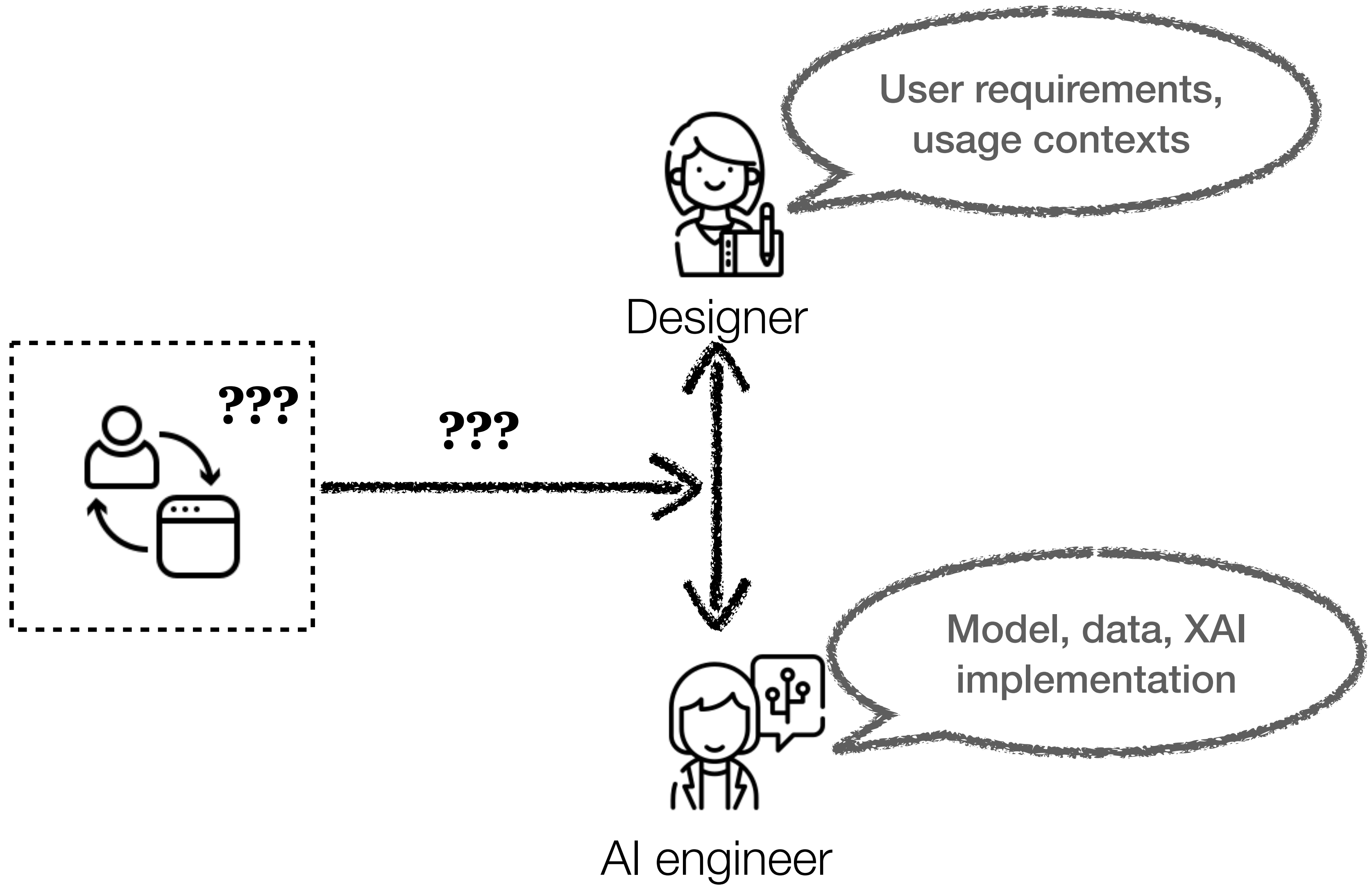
Global model: Choose appropriate level of details to explain the model

Local decision: Provide resources for "why not"

Counterfactual: Consider opportunities as utility features for analytics or exploration

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

User-centered design process: Question-driven XAI design



Pain points to address:

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

Liao et al. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. CHI 2020
Liao et al. Question-Driven Design Process for Explainable AI User Experiences. (Working paper)

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)?

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?

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 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?

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?

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 differen prediction Q?
- What kind of instance is predicted of [a different outcome Q]?

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

- **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)

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 or uncertainty 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

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

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

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

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](). (Working paper)

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

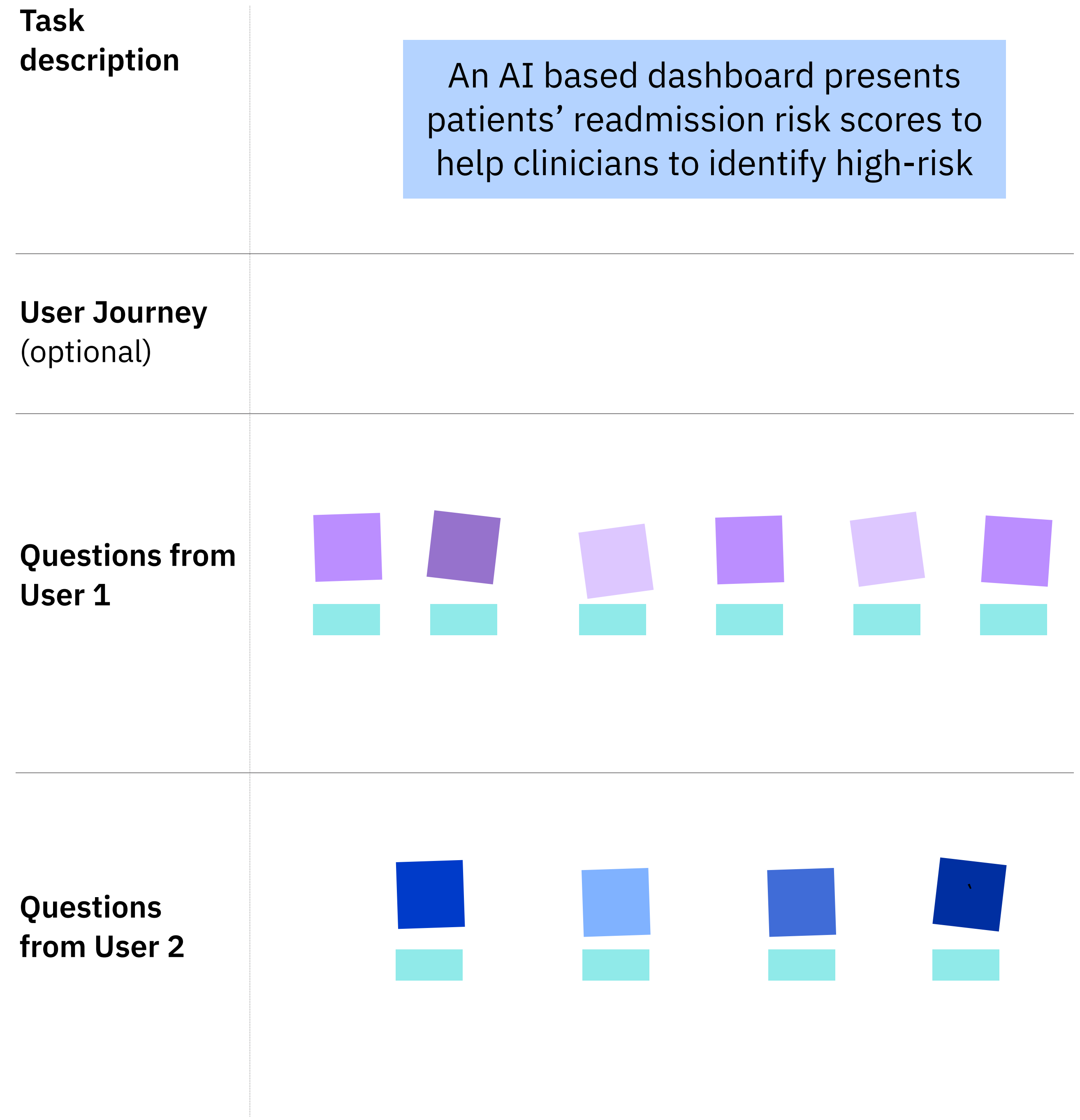
Designers, users

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

Identify relevant questions

Elicit user questions to identify what types of explanation are needed

Also collect the intention and expectation behind these user questions



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?

What is the population of the training data?

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

"Without knowing if it applies to my patients I can't trust it"

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

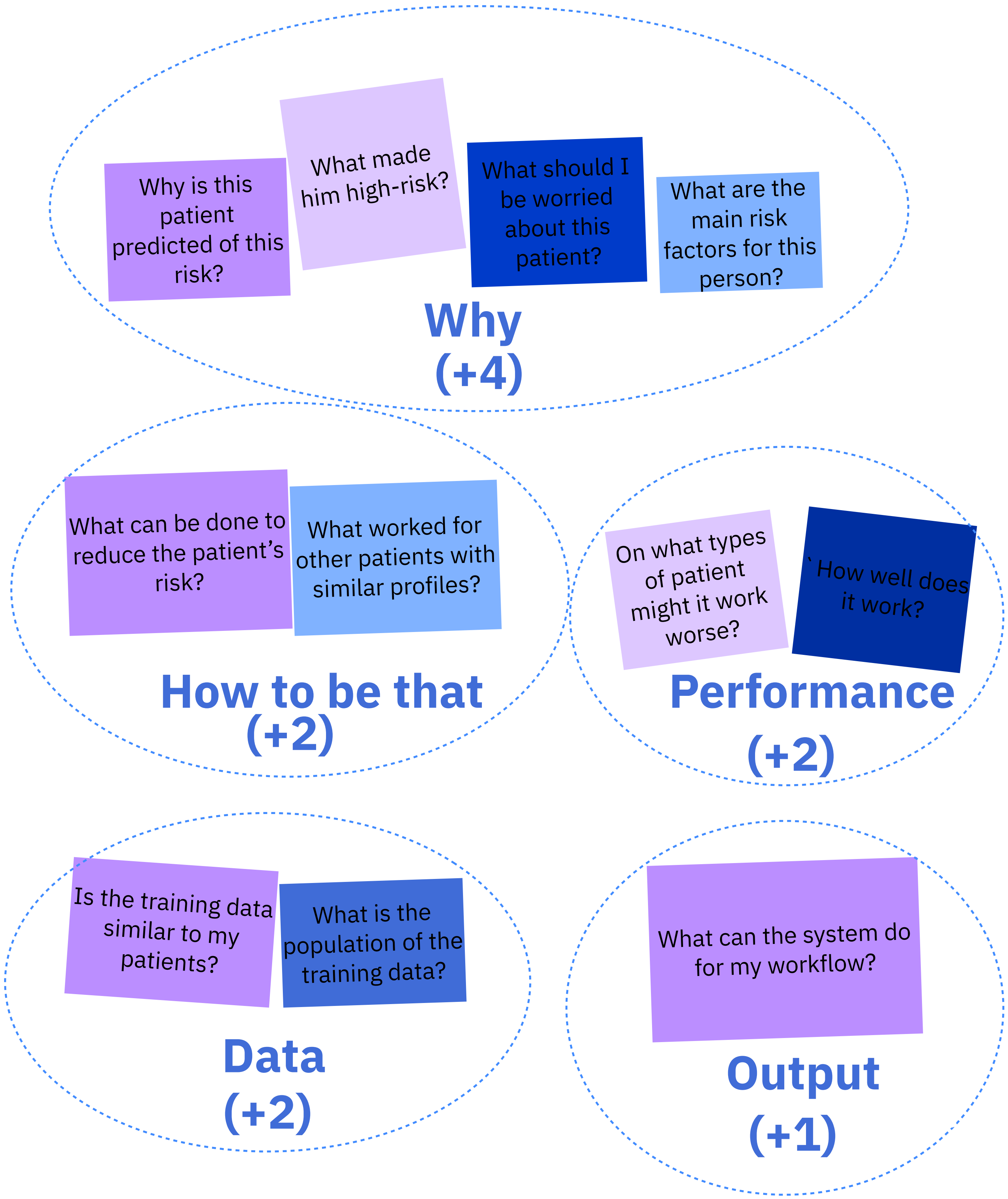
Liao et al. Question-Driven Design Process for Explainable AI User Experiences. (Working paper)

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

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

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

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 or uncertainty 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

Questions as re-framing the technical space of XAI

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

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

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

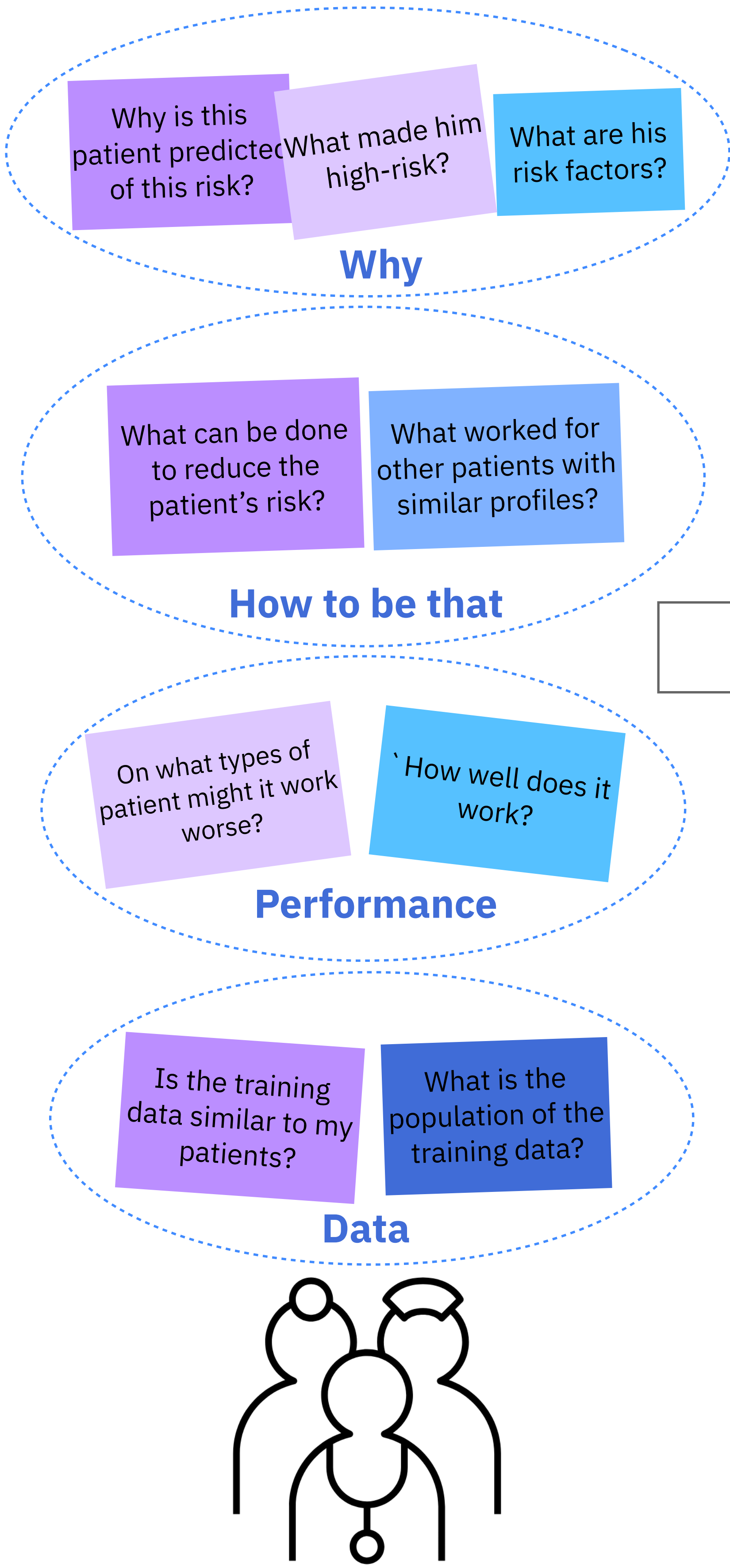
Iteratively design and evaluate

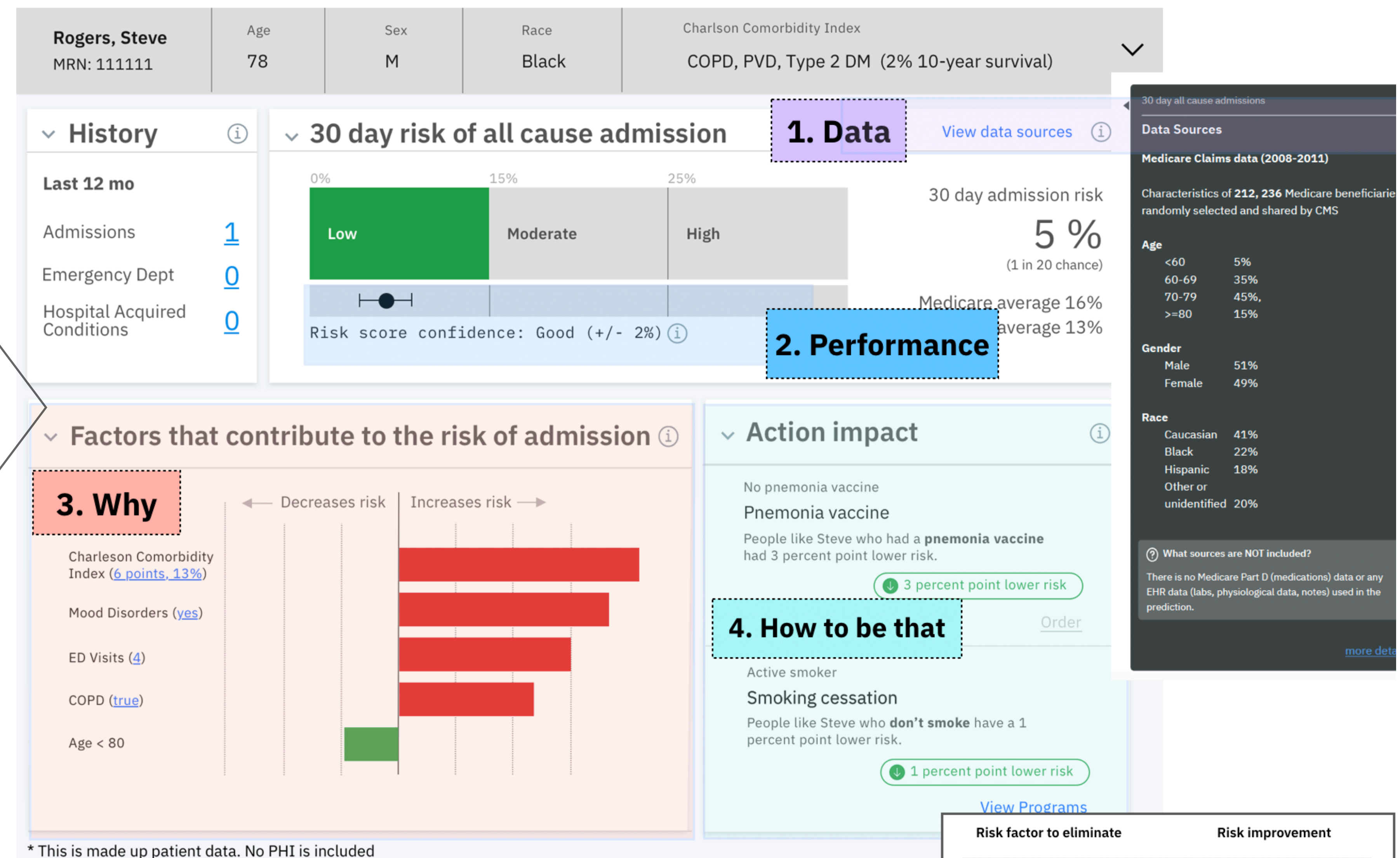
Create a design including the candidate elements identified in step 3

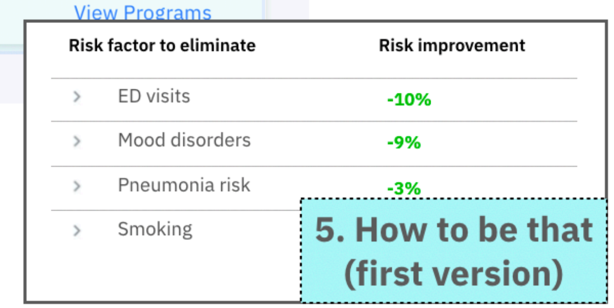
Iteratively valuate the design with the user requirements identified in step 2 and fill the gaps

Designers, data scientists, users

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







AI for Explainable Healthcare Adverse Event Risk Prediction

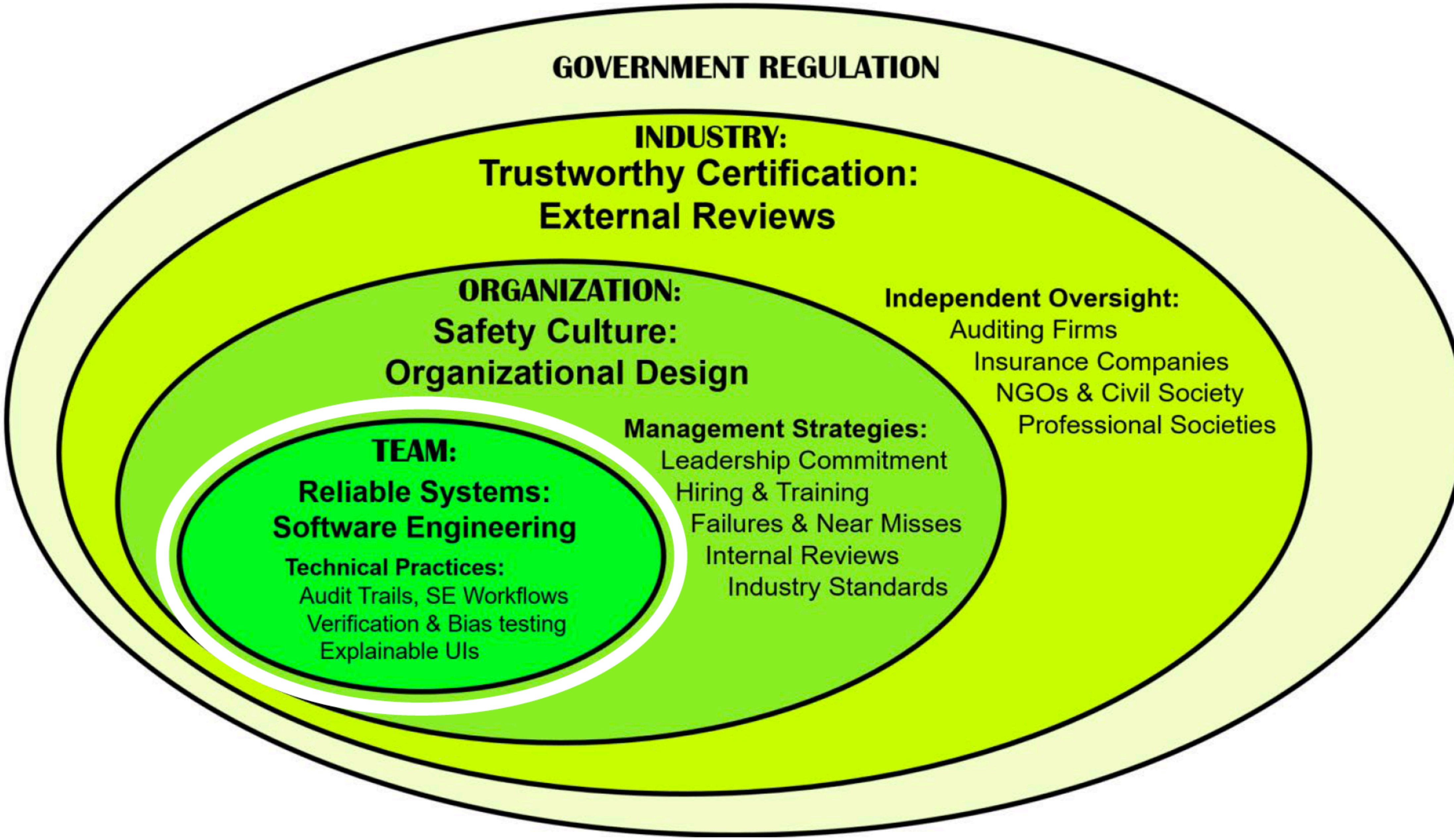
Liao et al. Question-Driven Design Process for Explainable AI User Experiences. (Working paper)

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
www.qveraliao.com
@QVeraLiao