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

Q. Vera Liao

IBM Research

HCXAI

HUMAN-CENTERED EXPLAINABLE AI

HCXAI logo made by Upol Ehsan

IBM Research Trusted AI

Home Demo Resources Events Videos Community

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 More **Try a Web Demo** **Watch Videos** **Read a Paper** **Use Tutorials** **Ask a Question**

IBM Research Trusted AI

Home Demo Resources Events Videos Community

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 ↗

Not sure what to do first? Start here!

IBM Research Trusted AI

Home Demos Resources Videos

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

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.

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 lifecyle. We invite you to use it and improve it.

Human-computer interaction (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)

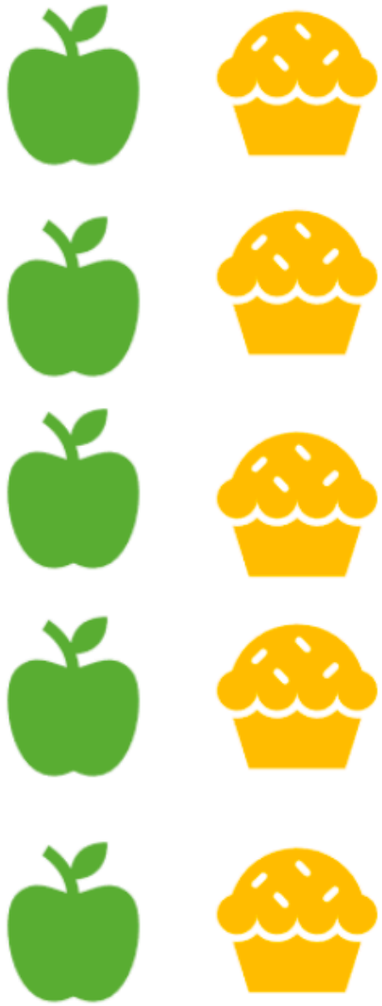
Everything that makes AI *understandable* (e.g., also including data, functions, performance, etc.)

Explaining Supervised Machine Learning

Training data set

Label: Apple

Label: Cake



Features:
Color
Shape
Smell
…

Learning Model
(Using a ML algorithm)

Prediction label:
Cake

New **instance**

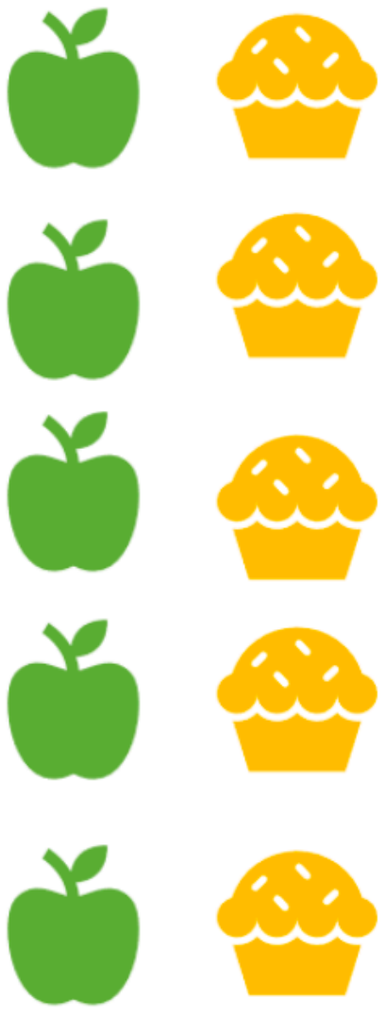
Explaining Supervised Machine Learning

Training data set

Label: Apple

Label: Cake

Features:
Color
Shape
Smell
…



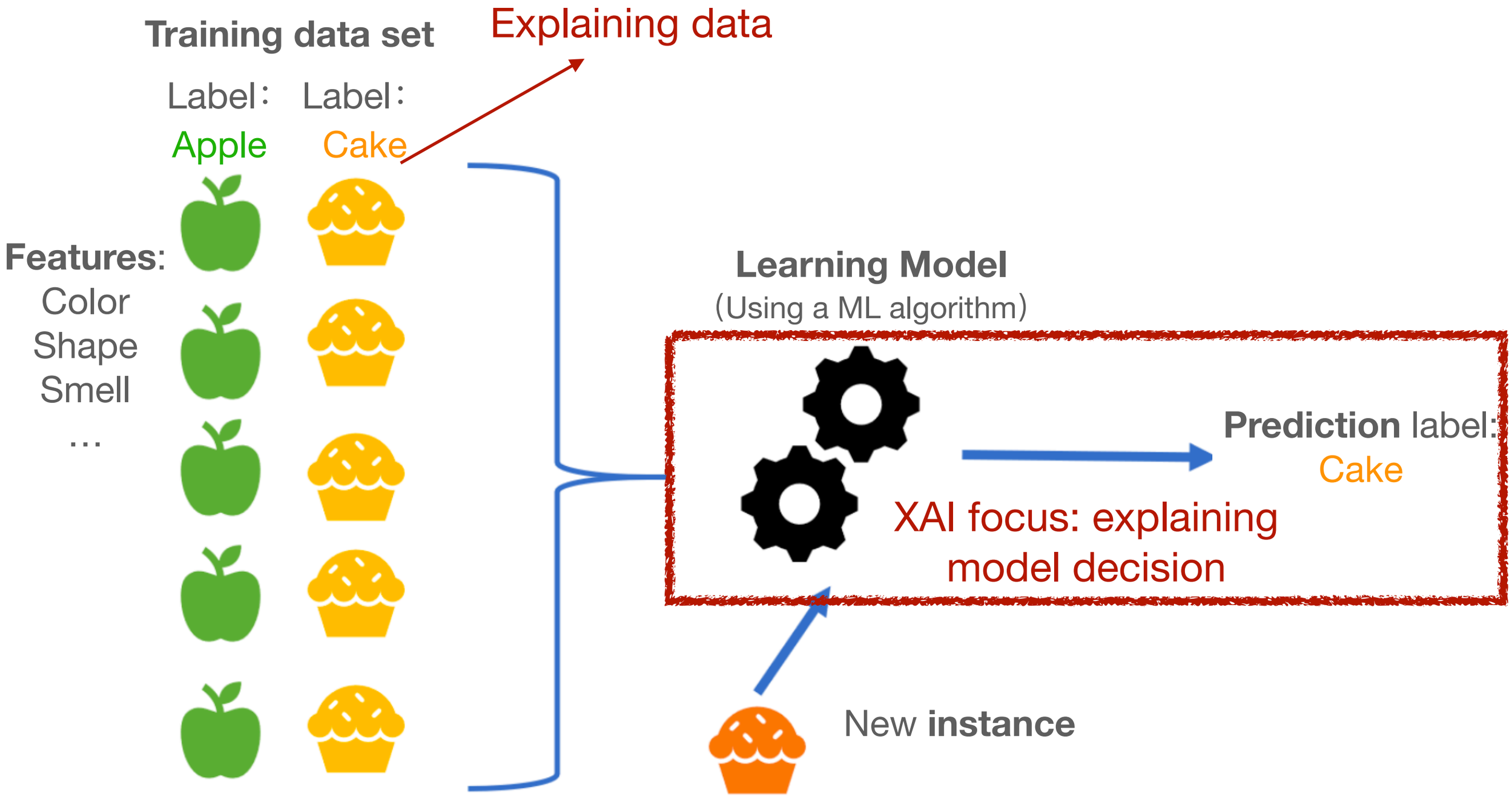
Learning Model
(Using a ML algorithm)

Prediction label:
Cake

XAI focus: explaining model decision

New **instance**

Explaining Supervised Machine Learning



Explaining Supervised Machine Learning

Training data set

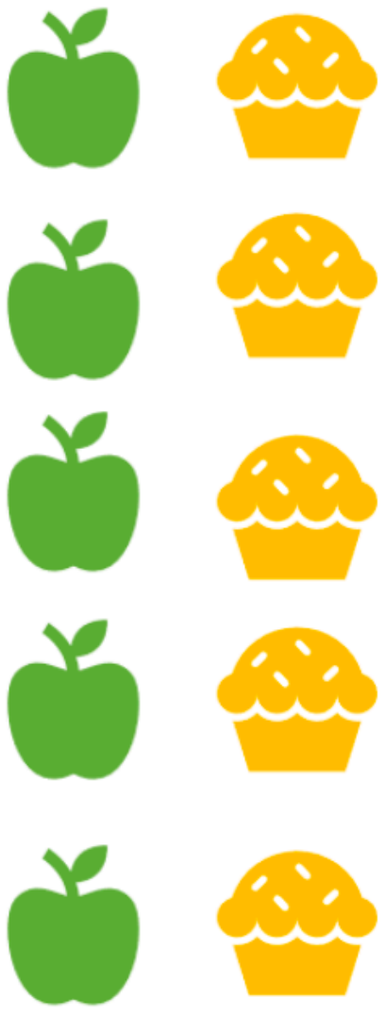
Explaining data

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

Label: Apple

Label: Cake

Features:
Color
Shape
Smell
…



Learning Model
(Using a ML algorithm)

Prediction label:
Cake

XAI focus: explaining model decision

New **instance**

The quest for explainable AI (XAI)

Companies Grapple With AI's Opaque Decision-Making Process

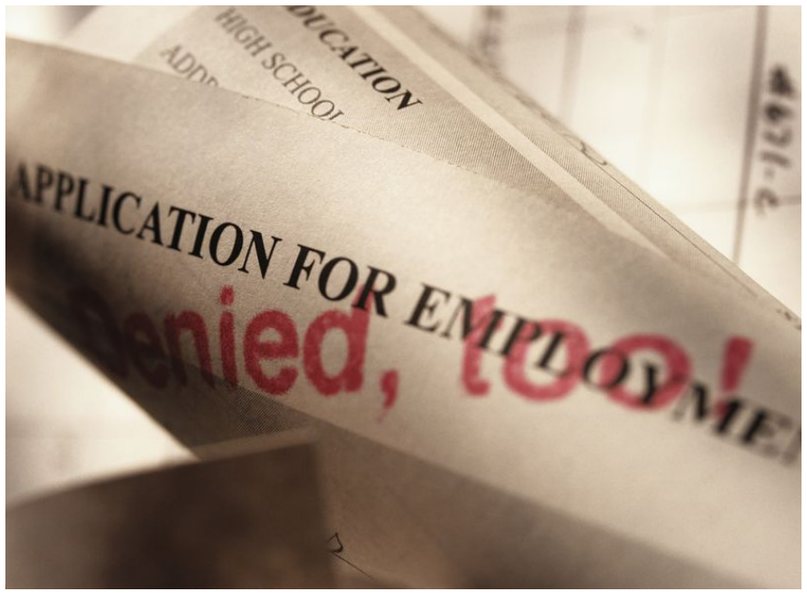
We Need AI That Is Explainable, Auditable, and Transparent

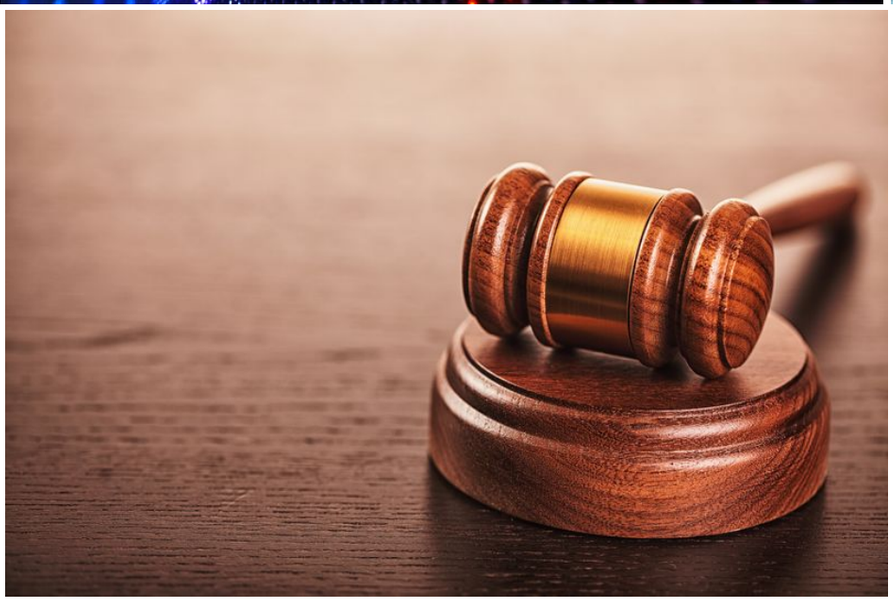
Why "Explainability" Is A Big Deal In AI

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

How Explainable AI Is Helping Algorithms Avoid Bias

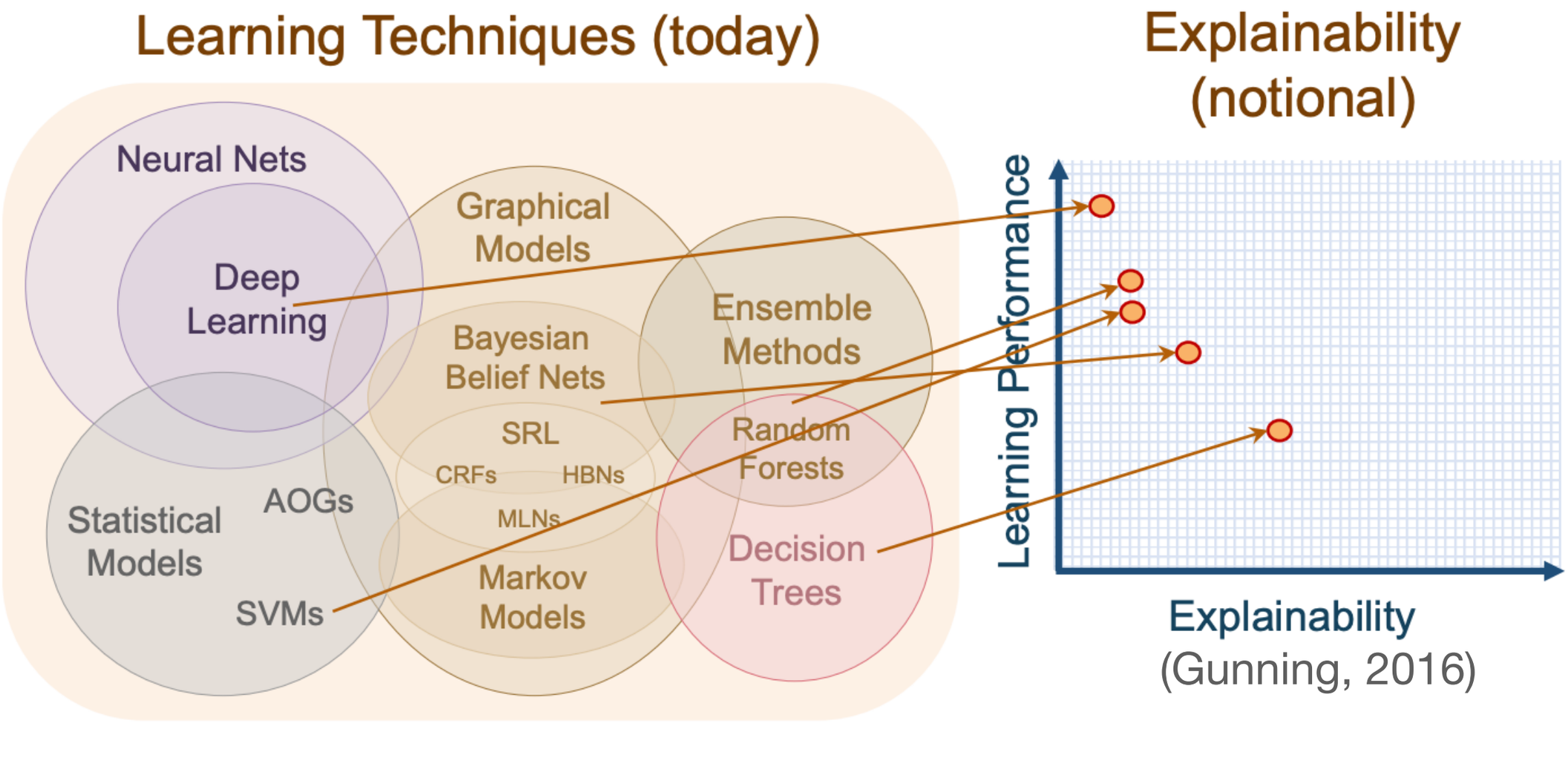
AI is increasingly used in many high-stakes tasks



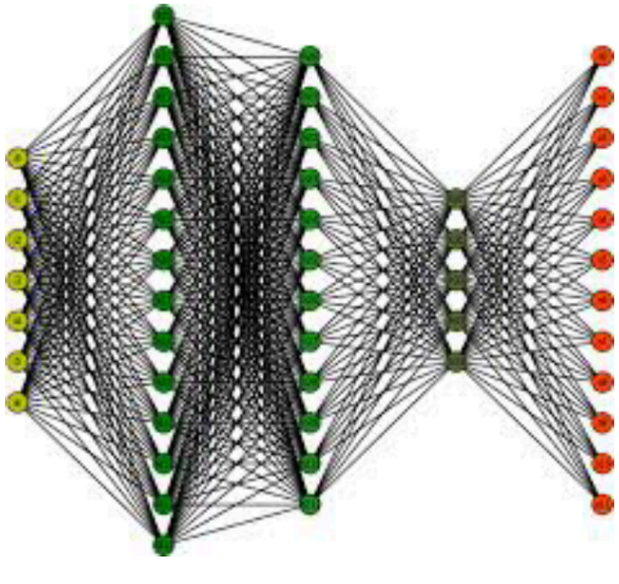


Performance-Explainability trade-off

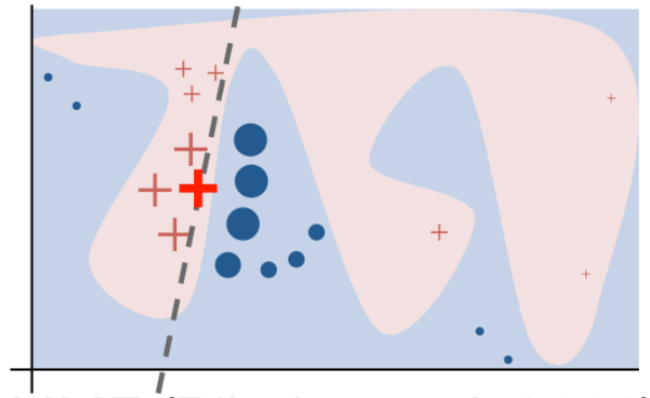
In ***average*** settings



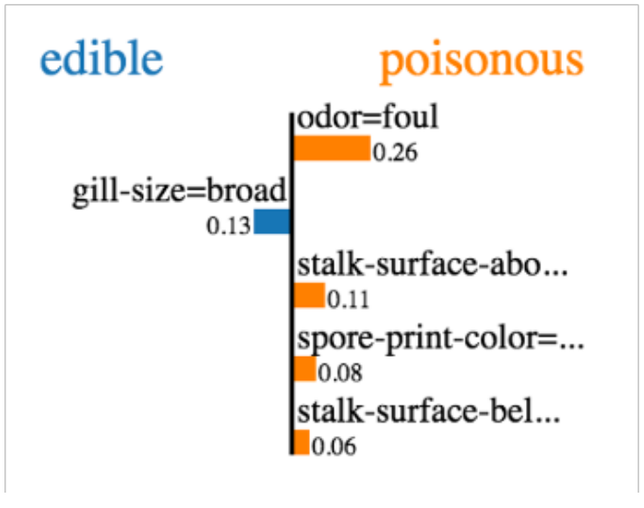
XAI “post-hoc” algorithm example: LIME



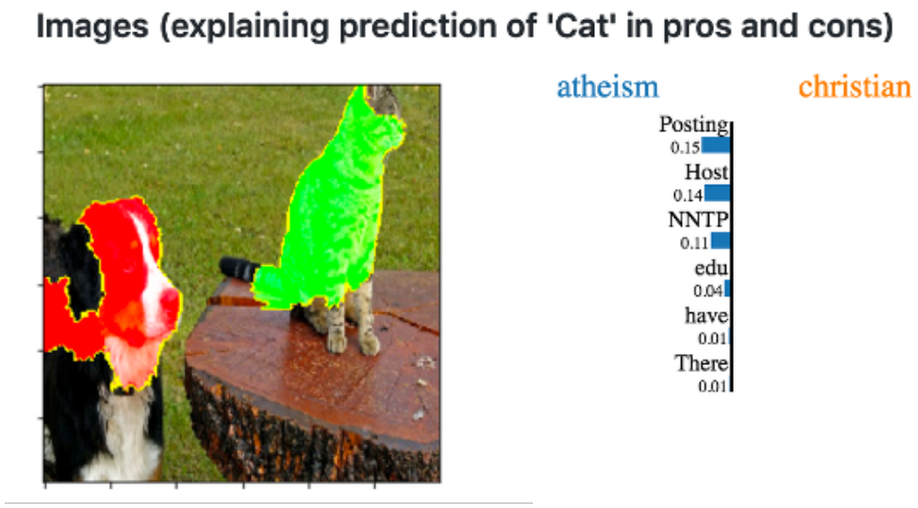
Neural network, not directly explainable



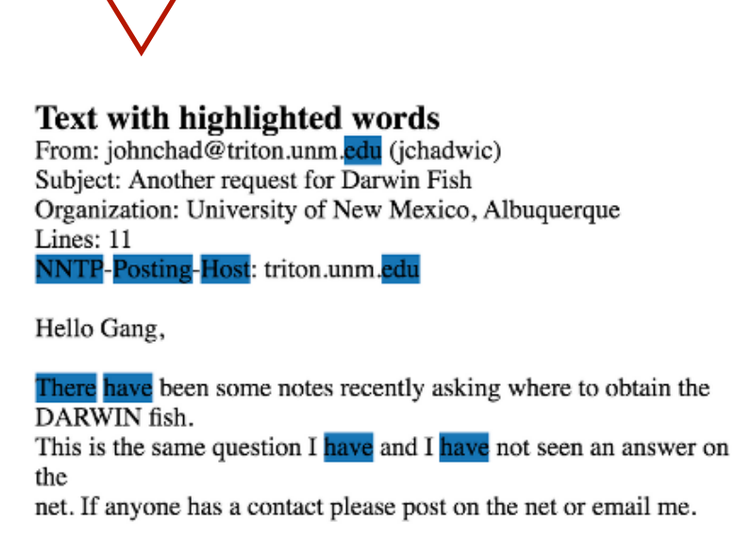
Use a *post-hoc* XAI technique



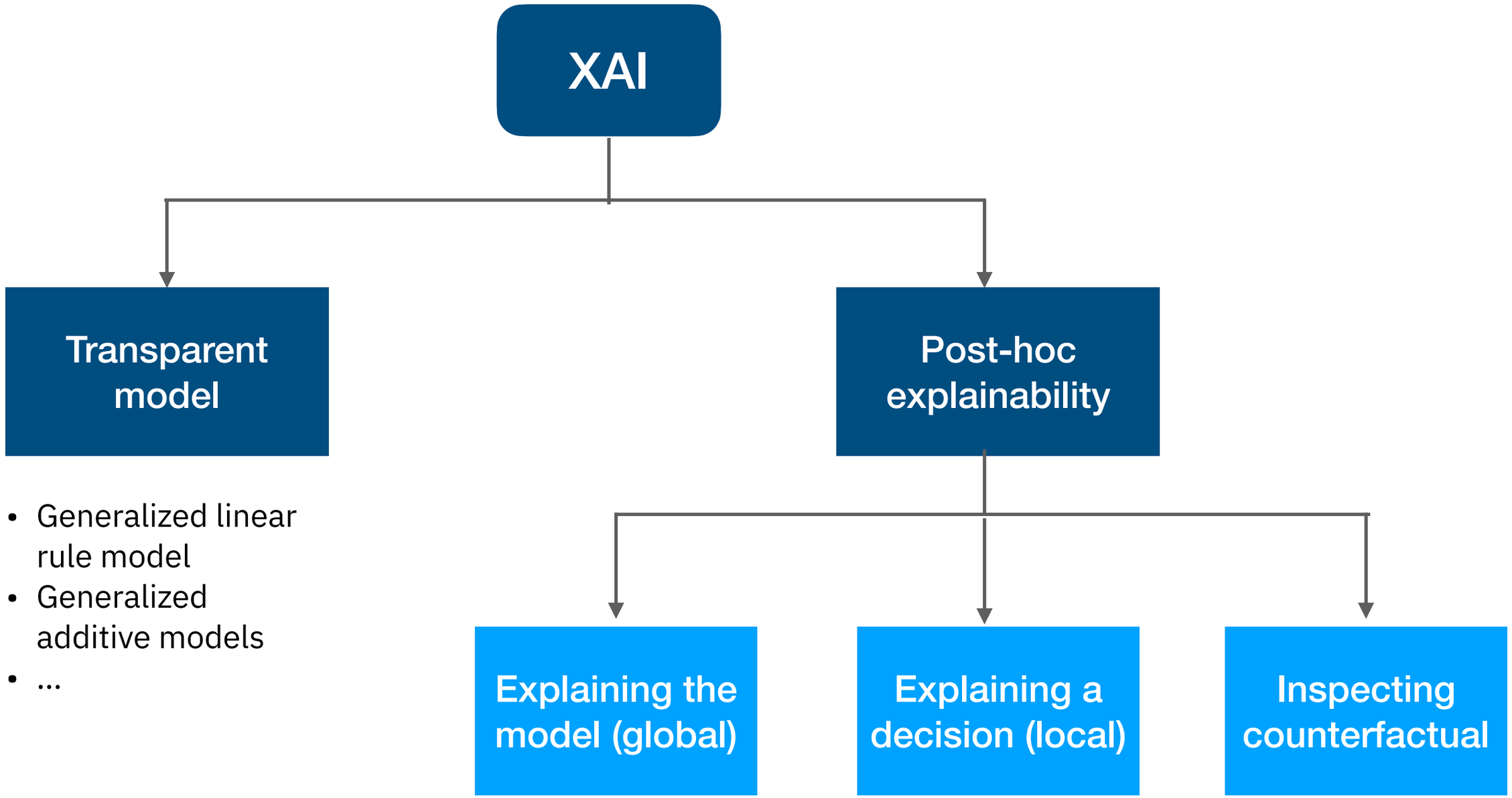
Tabuler data



Image



Texts



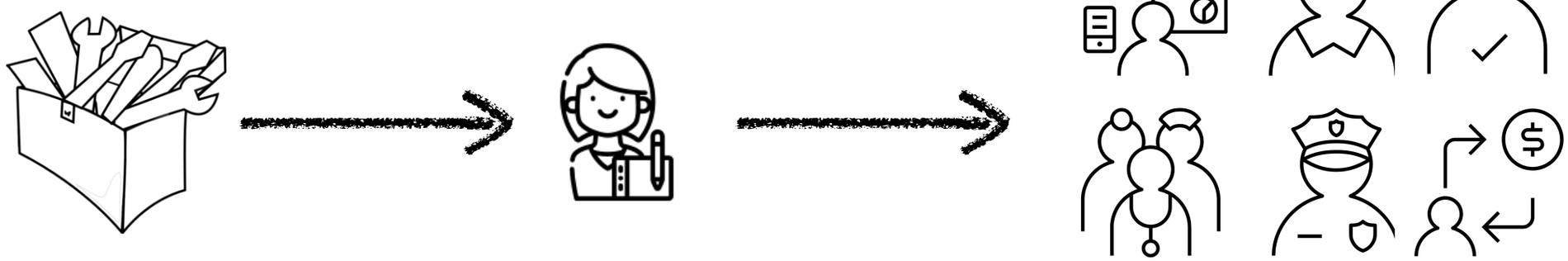
- Generalized linear rule model
- Generalized additive models
- …

- Feature importance
- Rule approximation
- Decision tree approximation

- Local contribution
- Local rules
- Similar instances

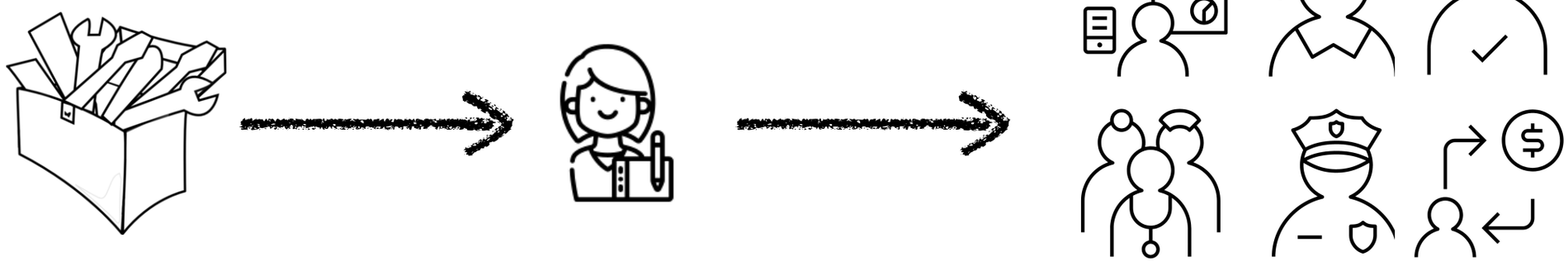
- Feature influence
- Contrastive features
- Counterfactual instances

AIX360: http://aix360.mybluemix.net/
Also check out our **CHI2021 Course materials**: https://hcixaitutorial.github.io/



Toolbox of XAI techniques

Real-world XAI applications



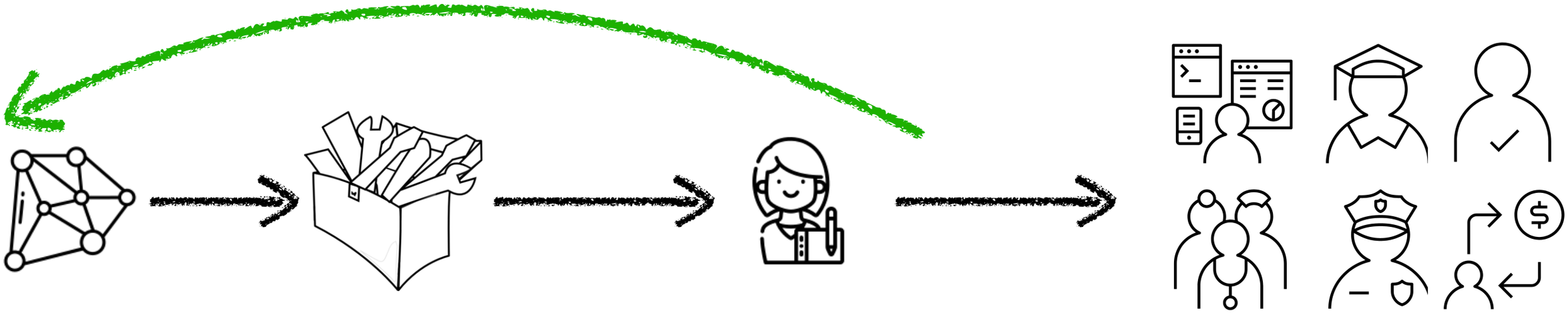
Toolbox of XAI techniques

Real-world XAI applications

How to **select**?

How to **translate**?

Inform **gaps** and **opportunities**



Toolbox of XAI techniques

Real-world XAI applications

How to **select**?

How to **translate**?

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

electronics

MDPI

Review

Machine Learning Interpretability: A Survey on Methods and Metrics

Diogo V. Carvalho [1,2,*], Eduardo M. Pereira [1] and

1 Deloitte Portugal, Manuel Bandeira Street, 43, 4150-47
2 Faculty of Engineering, University of Porto, Dr. Rober
3 INESC TEC, Dr. Roberto Frias Street, 4200-465 Porto,
* Correspondence: diocarvalho@deloitte.pt

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

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

A Multidisciplinary Survey and Framework for Design and Evaluat

SINA MOH
ERIC D. RA

The need for
intelligence a
reasoning be
to define, des
on different
challenges fo
across effort
experiences
design goals
fields of machine learning, visualization, and human-compute

Explaining Explanations: An Overview of Interpretability of Machine Learning

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

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

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

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

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

Digital Object Identifier 10.1109/ACCESS.2018.2870052

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

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

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

es,

e,f,

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
at the cost of sacrificing accuracy for interpretability. The appli
can be used are various, and each approach is typically develope
and, as a consequence, it explicitly or implicitly delineates its ov
tion. The aim of this article is to provide a classification of the m
respect to the notion of explanation and the type of black box
box type, and a desired explanation, this survey should help the

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

Gabriëlle Ras, Marcel van Gerven, Pim Haselager

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

Abstract

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

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

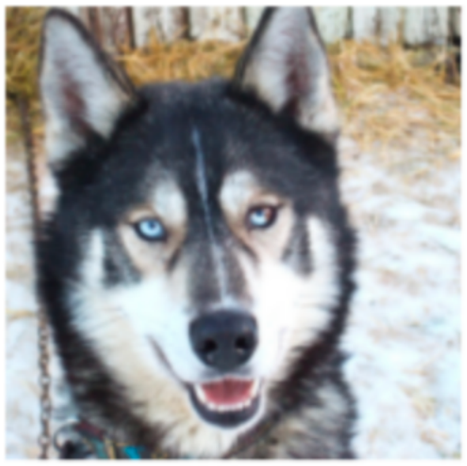
(AI) has achieved a notable momentum that, if harnessed
tions over many application sectors across the field. For this
ire community stands in front of the barrier of explainability,
brought by sub-symbolism (e.g. ensembles or Deep Neural
ype of AI (namely, expert systems and rule based models).
in the so-called *eXplainable* AI (XAI) field, which is widely
ctical deployment of AI models. The overview presented in
d contributions already done in the field of XAI, including a
r this purpose we summarize previous efforts made to define
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 (**Global**)	Global feature importance	Describe the weights of features used by the model (including visualization that shows the weights of features)	[41, 60, 69, 90]	**How**
	Decision tree approximation	Approximate the model to an interpretable decision-tree	[11, 47, 52]	**How**, Why, Why not, What if
	Rule extraction	Approximate the model to a set of rules, e.g., if-then rules	[26, 93, 102]	**How**, Why, Why not, What if
Explain a prediction (**Local**)	Local feature importance and saliency method	Show how features of the instance contribute to the model's prediction (including causes in parts of an image or text)	[61, 74, 83, 85, 101]	**Why**
	Local rules or trees	Describe the rules or a decision-tree path that the instance fits to guarantee the prediction	[39, 75, 99]	**Why**, **How to still be this**
Inspect counterfactual	Feature influence or relevance method	Show how the prediction changes corresponding to changes of a feature (often in a visualization format)	[8, 33, 36, 51]	**What if**, How to be that, How to still be this
	Contrastive or counterfactual features	Describe the feature(s) that will change the prediction if perturbed, absent or present	[27, 91, 100]	**Why**, **Why not**, **How to be that**
Example based	Prototypical or representative examples	Provide example(s) similar to the instance and with the same record as the prediction	[13, 48, 50]	**Why**, How to still be this
	Counterfactual example	Provide example(s) with small differences from the instance but with a different record from the prediction	[37, 55, 66]	**Why**, **Why not**, How to be that

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

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



Question: Why is this husky classified as wolf?

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

XAI algorithms:

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

Study probe: algorithm informed **XAI Questions**

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

+

Model facts: **data, output, performance**

(Lim et al., 2009)

Methodology

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

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

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

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

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

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

Other category (add your own question)

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

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

Methodology

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

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

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

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

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

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

Other category (add your own question)

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

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

XAI Question Bank

Data

- **What kind of data was the system trained on?**
- What is the source of the training data?
- How were the labels/ground-truth produced?
- What is the sample size of the training data?
- What dataset(s) is the system NOT using?
- What are the 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 different 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

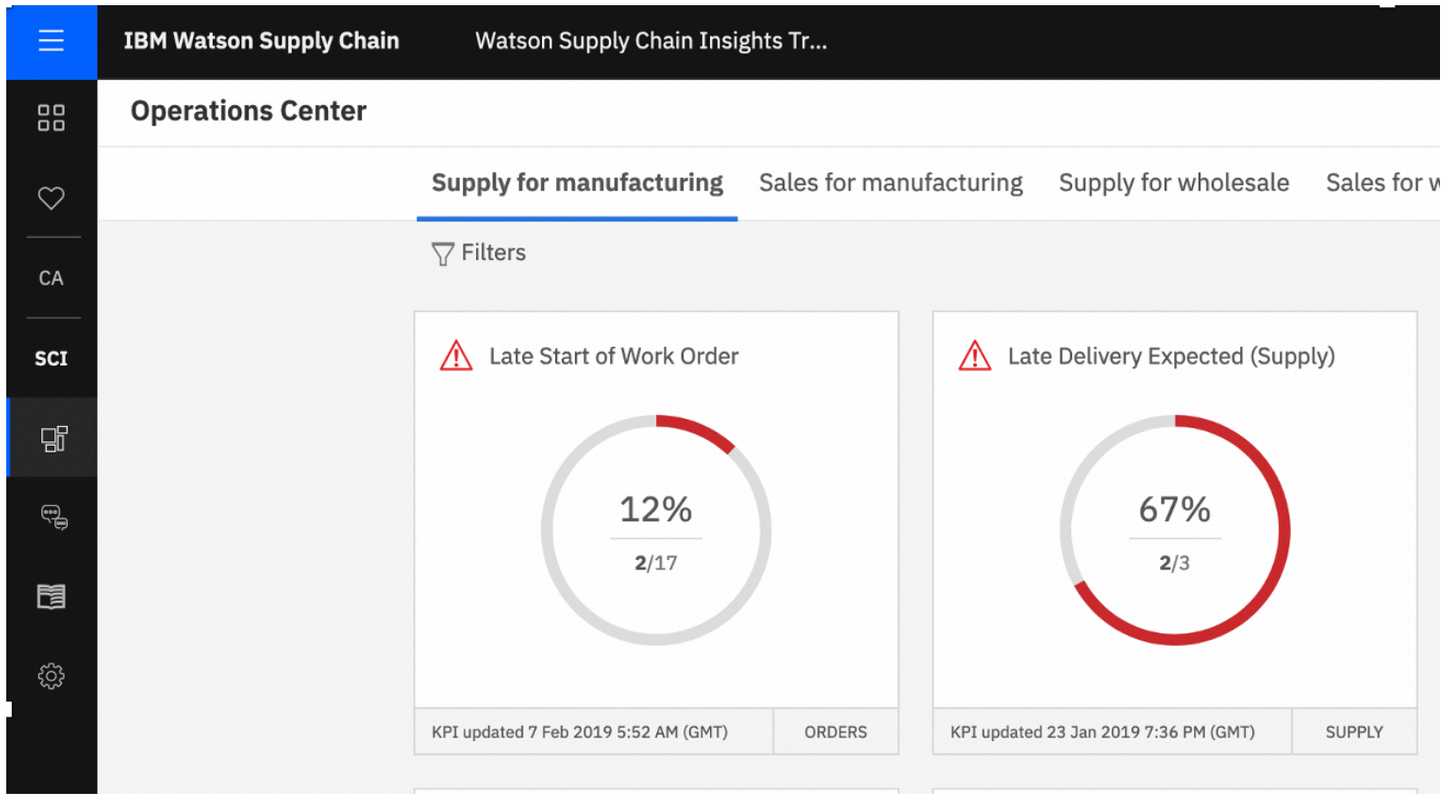
- How/why will the system change/adapt/improve/drift over time? (change)
- Can I, and if so, how do I, improve the system? (improvement)
- Why is the system using or not using a given algorithm/feature/rule/ dataset? (follow-up)
- What does [a machine learning terminology] mean? (terminological)
- What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

XAI design challenge 1: Variability of XAI needs

Diverse objectives for explainability

- To gain insights for the decision or action
- To appropriately evaluate AI's capability
- To adapt usage or control
- To learn about a domain
- To improve the model
- Legal or ethical compliance: auditing for fairness, privacy, security, etc.

To gain further insights for the decision



Why
How to be that

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

To appropriately evaluate AI's capability

Performance
How

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

XAI design challenge 2: Gaps between algorithmic output and human 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

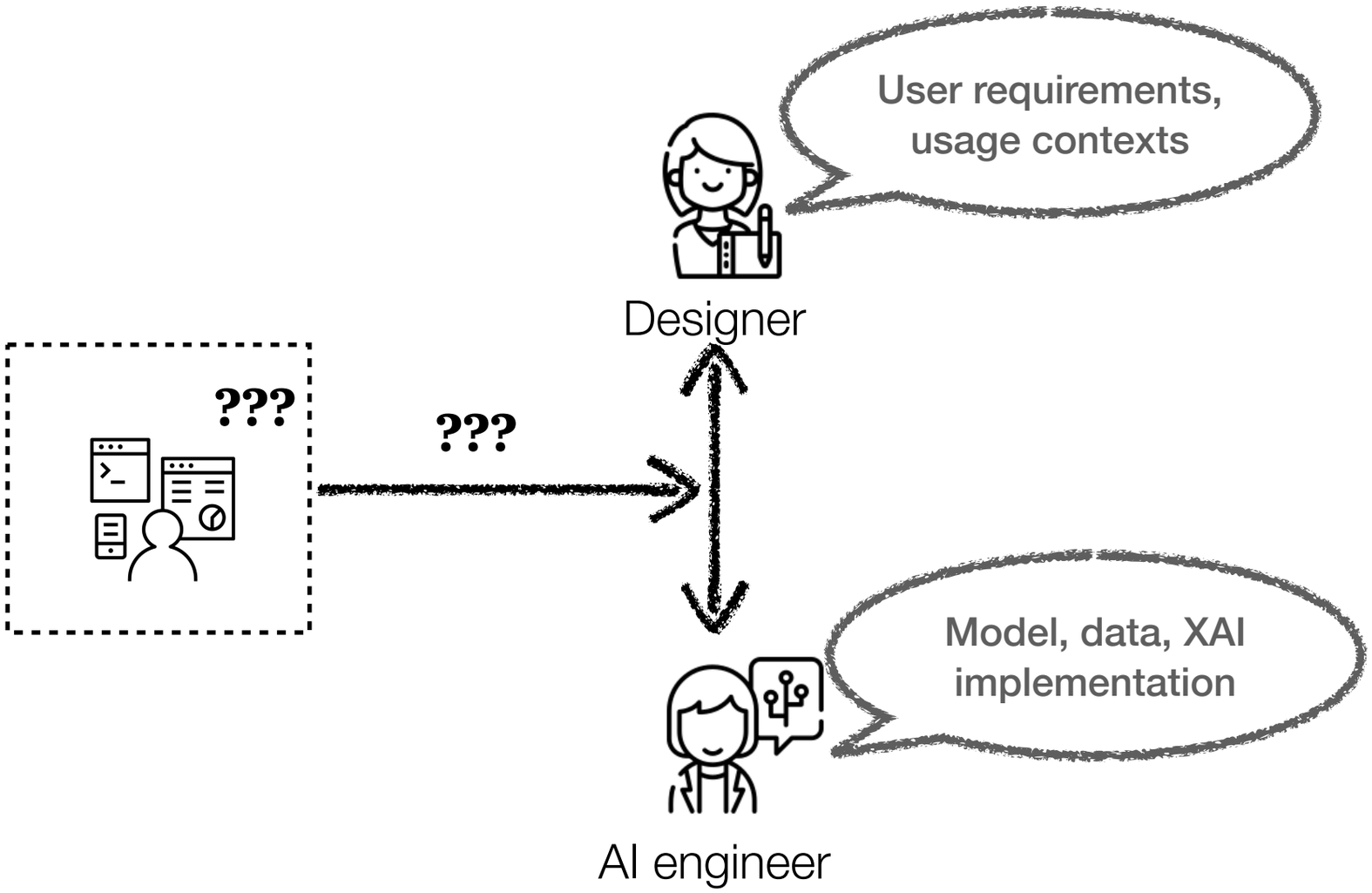
- Challenge **navigating the technical capabilities**

> *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*

- **Implementation cost** (technical debt) 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.*

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 different 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)

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 type of patient might it work worse?

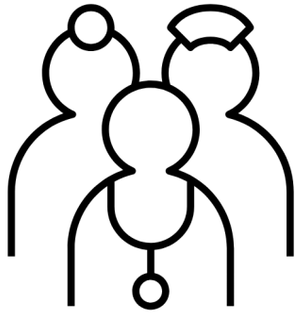
How well does it work?

Performance

Is the training data similar to my patients?

What is the population of the training data?

Data



Rogers, Steve MRN: 111111	Age 78	Sex M	Race Black	Charlson Comorbidity Index COPD, PVD, Type 2 DM (2% 10-year survival)

History

Last 12 mo

Admissions	1
Emergency Dept	0
Hospital Acquired Conditions	0

30 day risk of all cause admission

1. Data

View data sources

0% 15% 25%

Low | Moderate | High

Risk score confidence: Good (+/- 2%)

30 day admission risk

5 %

(1 in 20 chance)

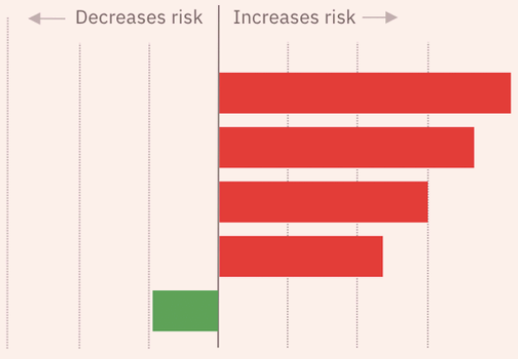
Medicare average 16%

average 13%

2. Performance

Factors that contribute to the risk of admission

3. Why



Action impact

No pnemonia vaccine

Pnemonia vaccine

People like Steve who had a **pnemonia vaccine** had 3 percent point lower risk.

3 percent point lower risk

4. How to be that

Order

Active smoker

Smoking cessation

People like Steve who **don't smoke** have a 1 percent point lower risk.

1 percent point lower risk

View Programs

30 day all cause admissions

Data Sources

Medicare Claims data (2008-2011)

Characteristics of **212, 236** Medicare beneficiarie randomly selected and shared by CMS

Age	
<60	5%
60-69	35%
70-79	45%,
>=80	15%
Gender	
Male	51%
Female	49%
Race	
Caucasian	41%
Black	22%
Hispanic	18%
Other or unidentified	20%

What sources are NOT included?

There is no Medicare Part D (medications) data or any EHR data (labs, physiological data, notes) used in the prediction.

more deta

* This is made up patient data. No PHI is included

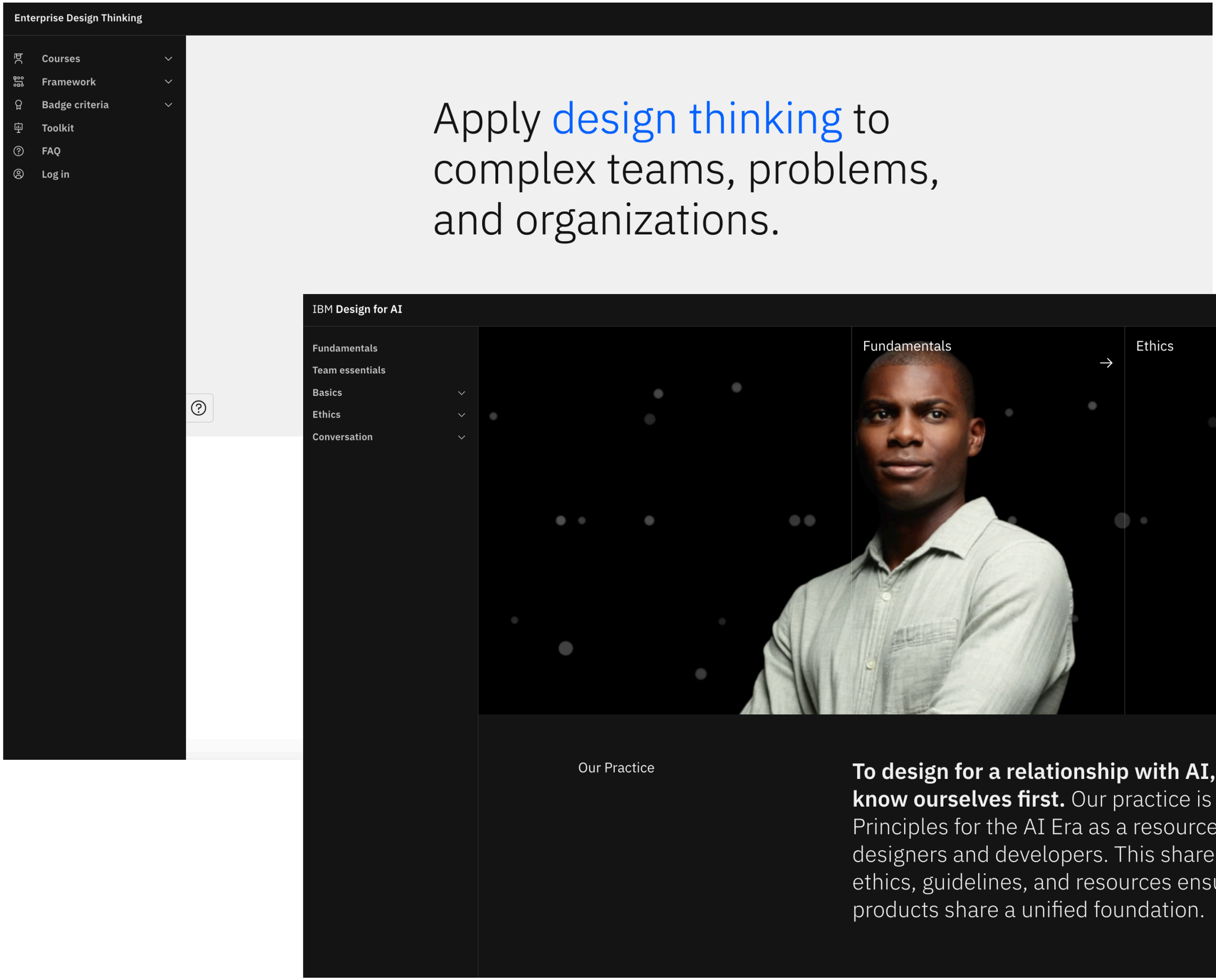
Risk factor to eliminate	Risk improvement
ED visits	-10%
Mood disorders	-9%
Pneumonia risk	-3%
Smoking	

5. How to be that (first version)

AI for Explainable Healthcare Adverse Event Risk Prediction

Liao et al. [Question-Driven Design Process for Explainable AI User Experiences](#).

https://www.ibm.com/design/thinking/



https://www.ibm.com/design/ai/

Thank **YOU!**

…and thanks to

Rachel Bellamy, Amit Dhurandhar, Jonathan Dodge, Upol Ehsan, Bhavya Ghai, 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, Justin Weisz, Yunfeng Zhang

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