Introduction to eXplainable AI (XAI)

Q. Vera Liao, Moninder Singh,
Yunfeng Zhang, Rachel Bellamy

IBM Research

CHI 2021

Latest slides available: https://hcixaitutorial.github.io

Who we are

- Researchers @ IBM Research
- Part of the team developed [IBM AI Explainability 360]
- Human-centered XAI

HCXAI
HUMAN-CENTERED EXPLAINABLE AI

HCXAI logo made by Upol Ehsan

Ask questions in Zoom Chat

Follow-up after the course: vera.liao@ibm.com
@QVeraLiao, www.qveraliao.com

Links

- Course website: https://hcixaitutorial.github.io/
- Course slides: http://qveraliao.com/xai_tutorial.pdf
- Pre-course notes:http://qveraliao.com/chi_course_notes.pdf
- AIX360: http://aix360.mybluemix.net/
- Install AIX360: https://github.com/Trusted-AI/AIX360
- Code demo:https://nbviewer.jupyter.org/github/IBM/AIX360/blob/master/examples/tutorials/HELOC.ipynb

Agenda

- Part 1: Overview presentation
  - What is explainable AI (XAI)?
  - How to explain? *With a use case*
  - Why is XAI important (*as the foundation for responsible AI*)?
  - How to design XAI?
- Part 2: Code demonstration with AIX360
  - Course notes: https://hcixaitutorial.github.io

Explainable AI (**XAI**): Definition

Narrow definition:

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

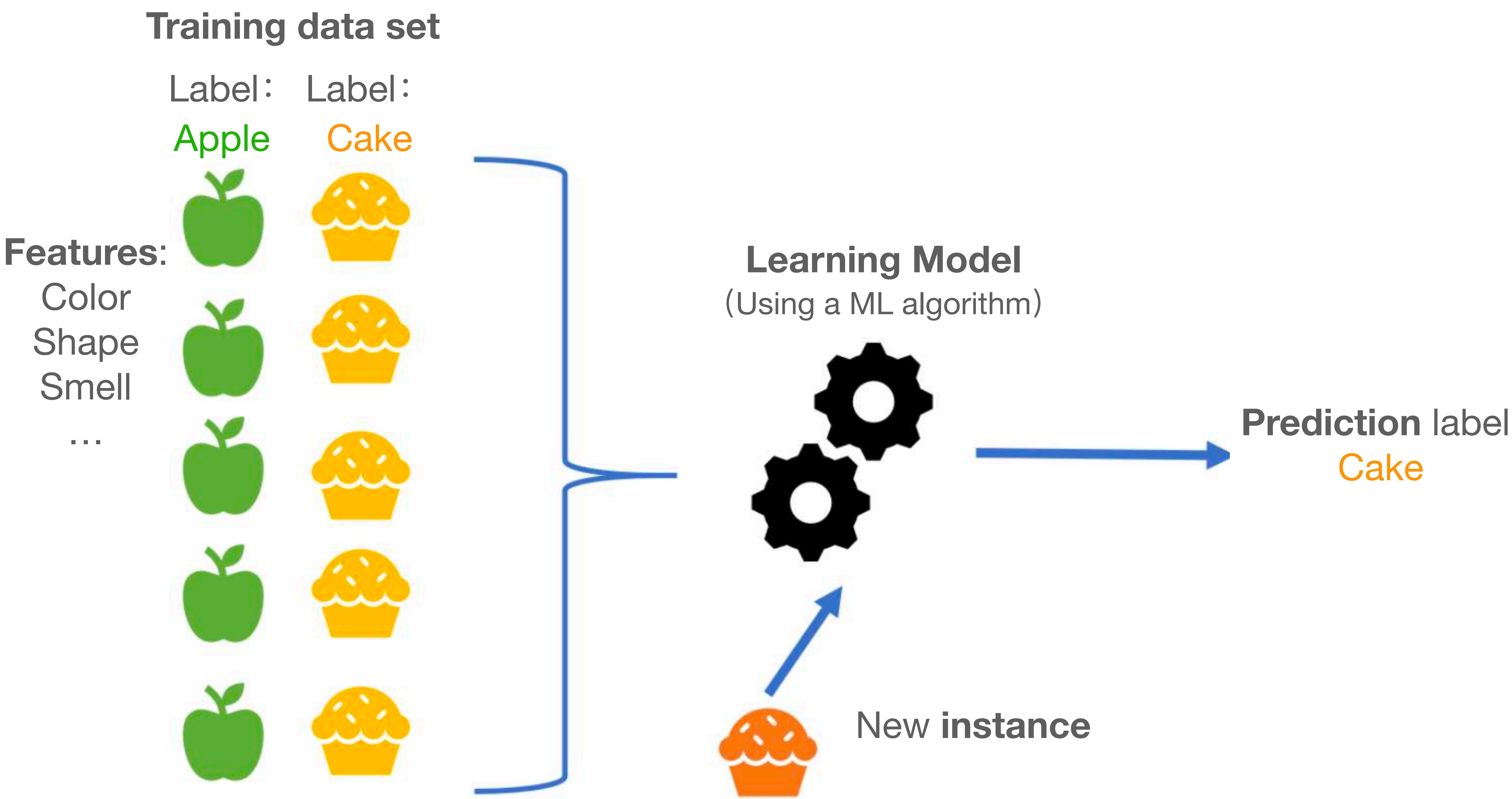
Broader definition:

(comprehensible/intelligible AI)

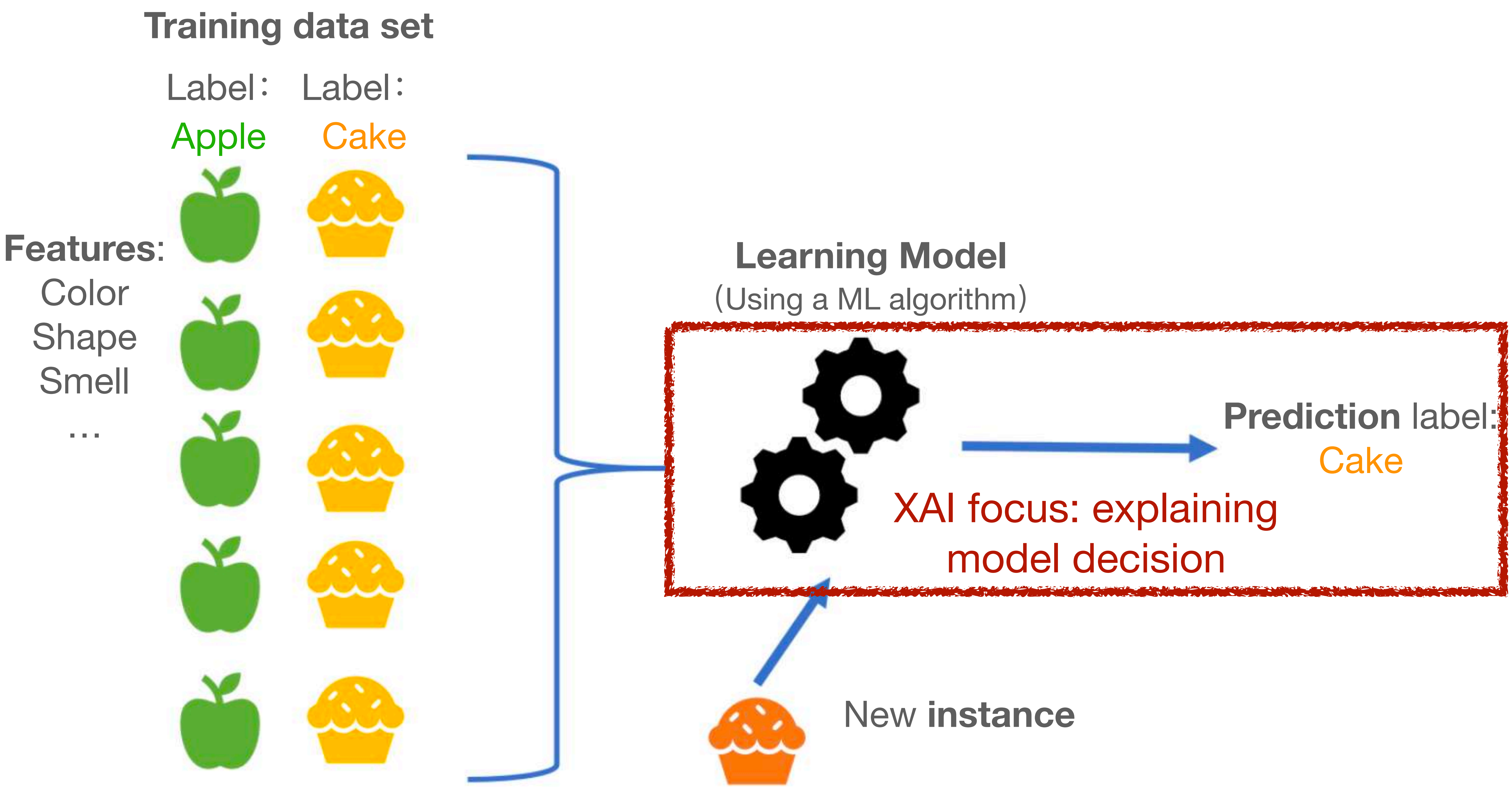
Everything that makes AI understandable (e.g., also including data, functions performance, etc.)

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

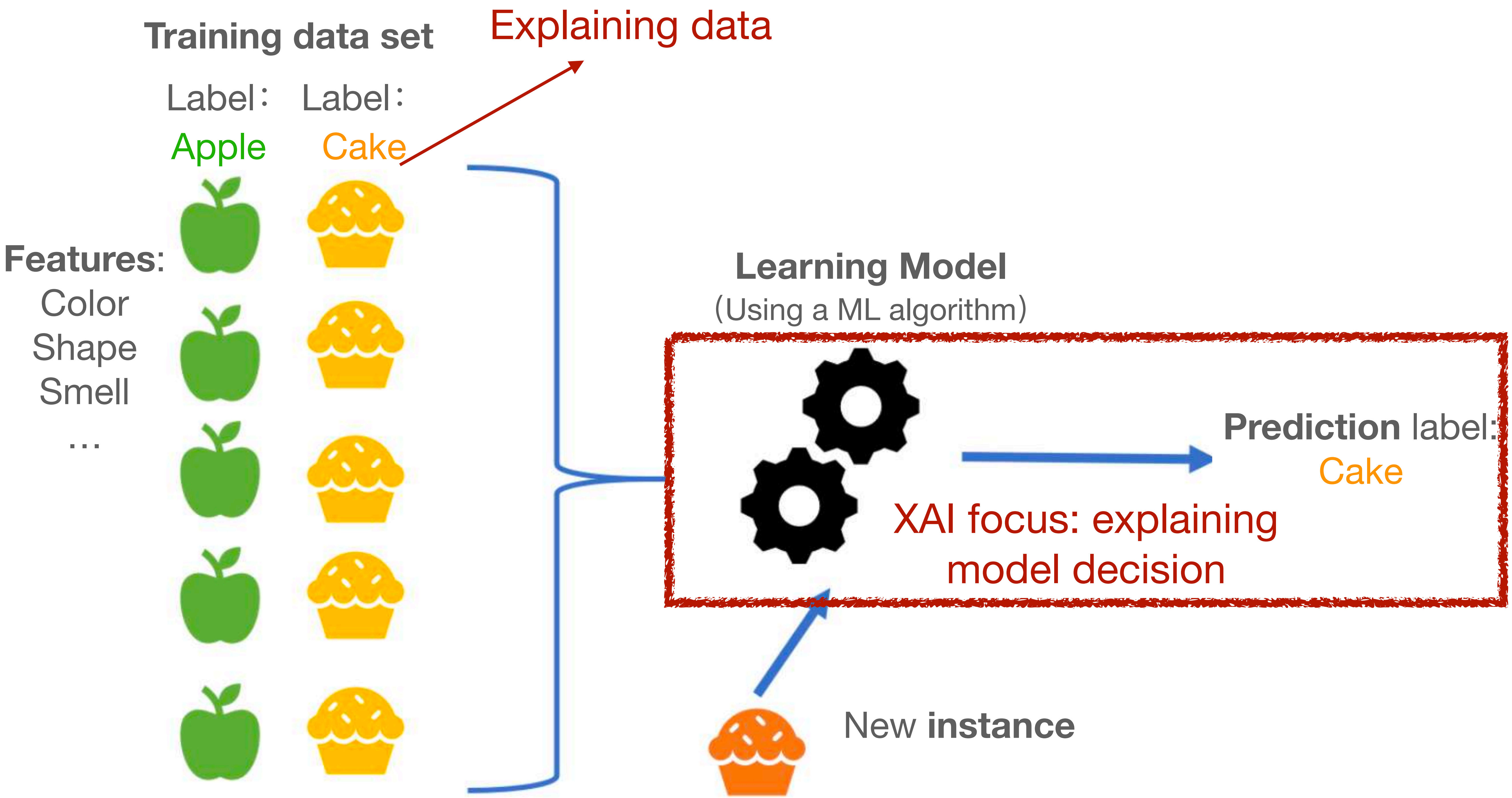
Supervised Machine Learning



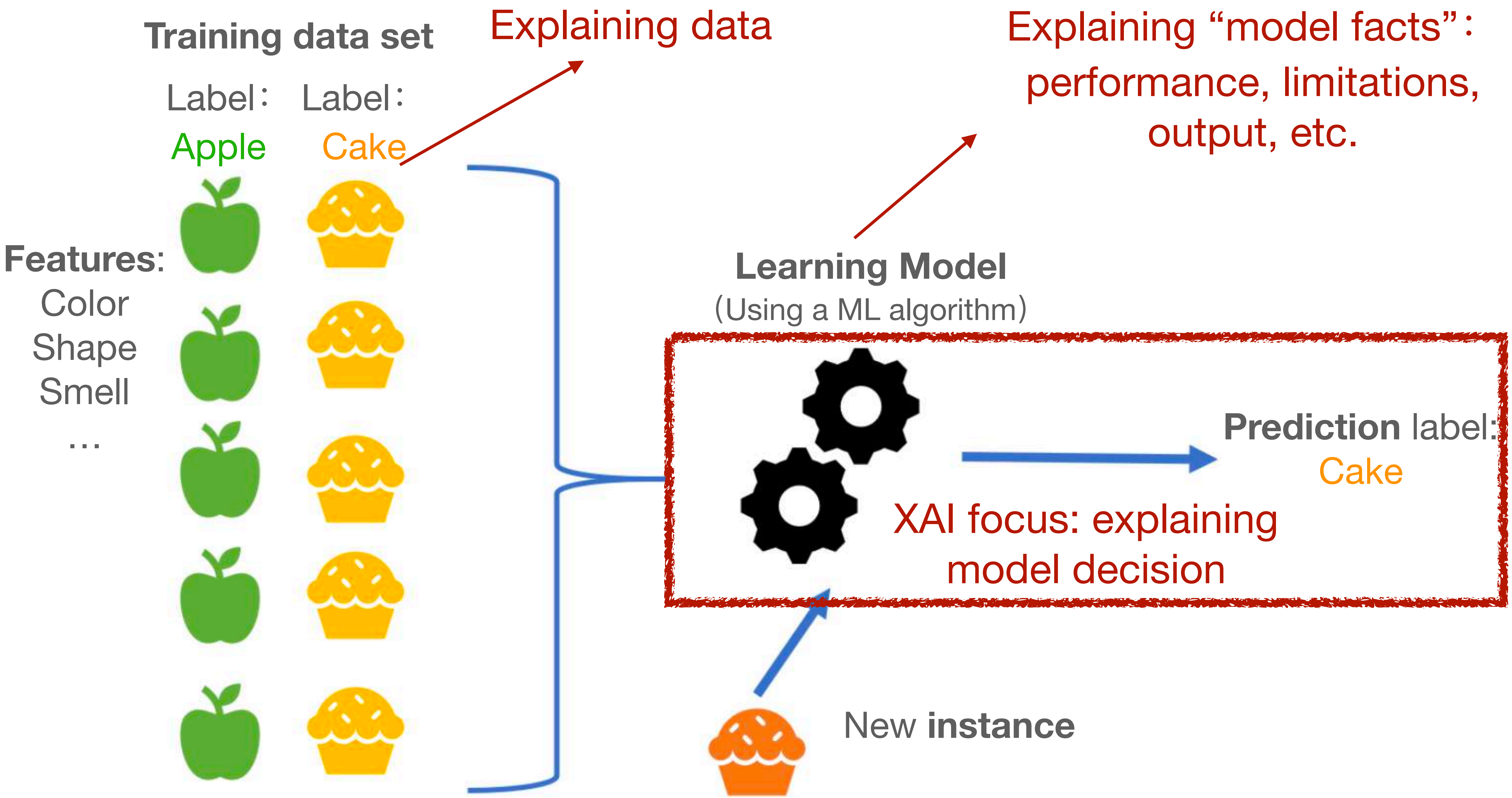
Supervised Machine Learning



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

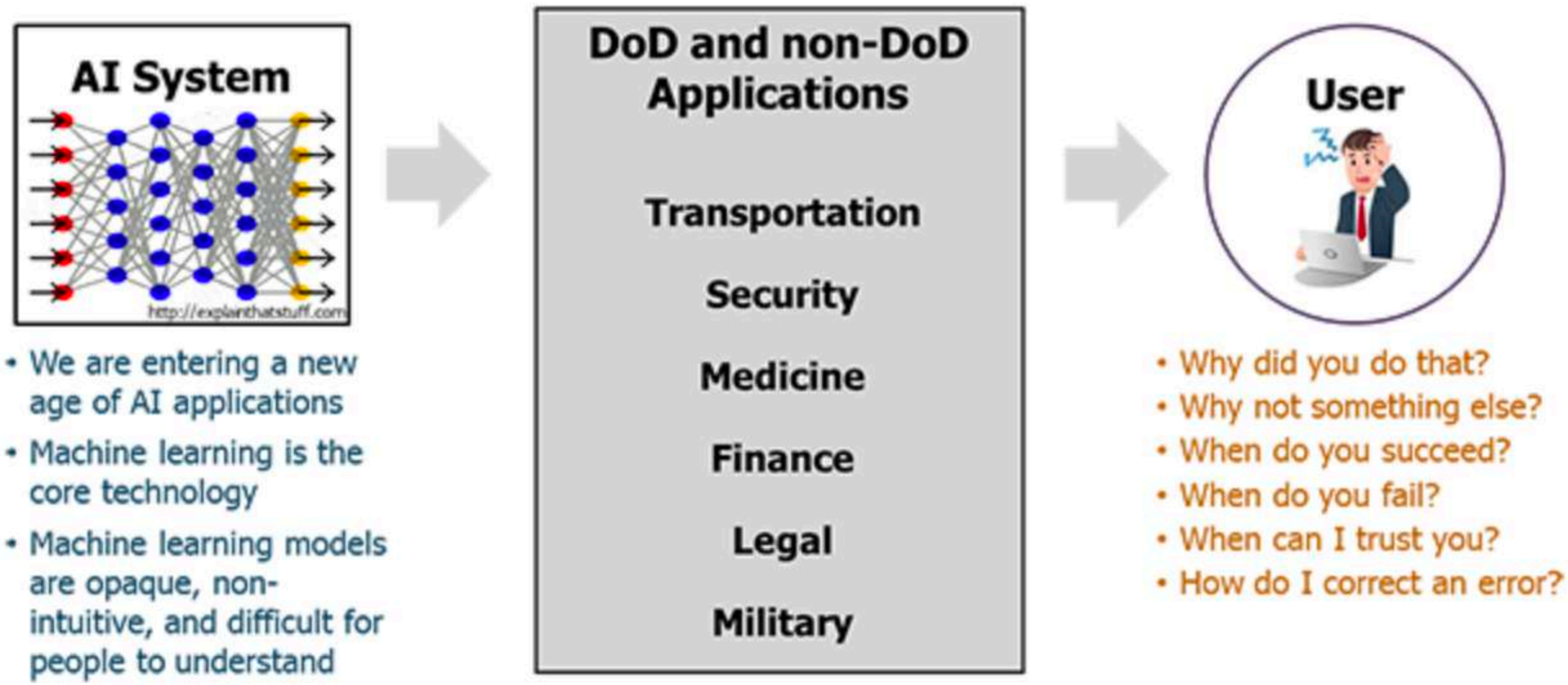
XAI in regulation: “rights to explanation”

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)

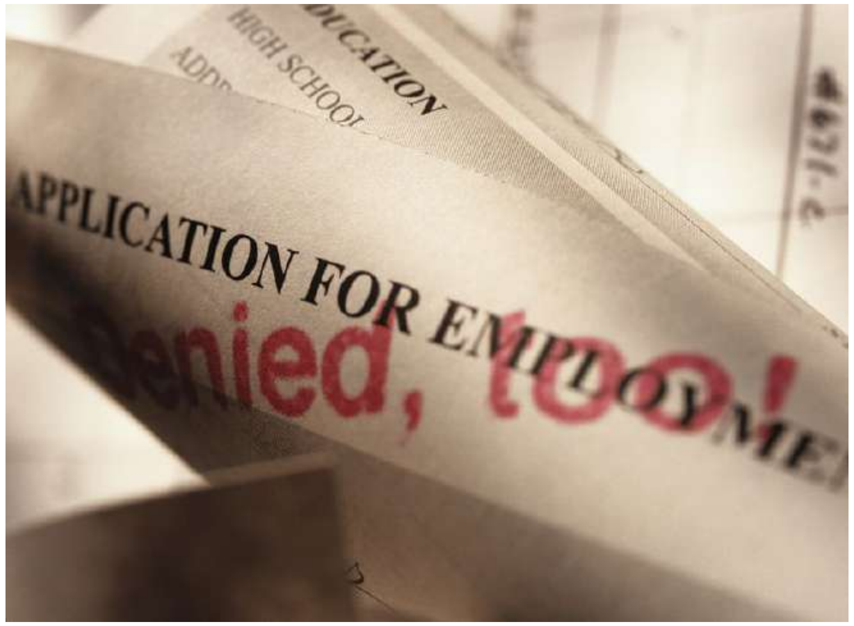
GDPR, 2016

XAI in research funding



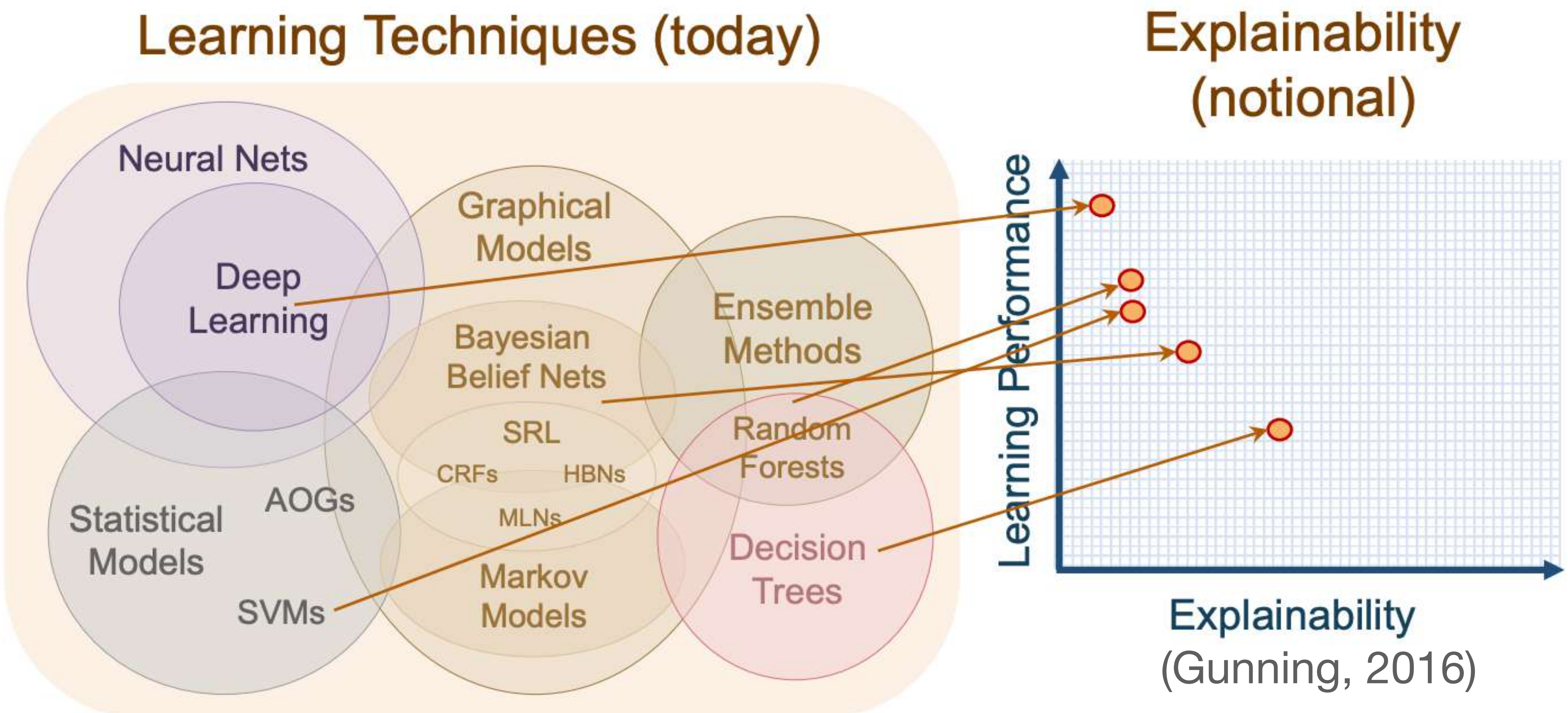
DARPA, 2016

AI is increasingly used in many high-stakes tasks



Performance-Explainability trade-off

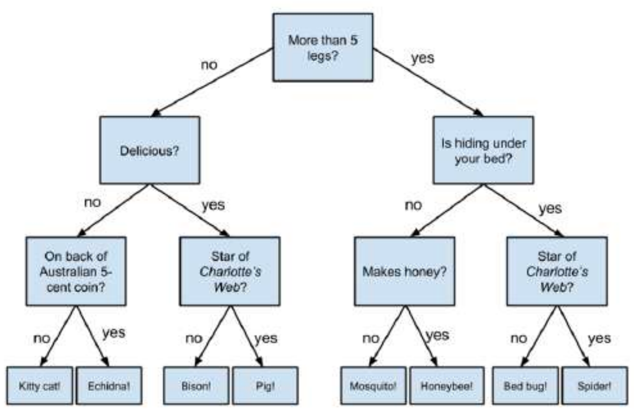
In ***average*** settings



(Gunning, 2016)

XAI

Directly explainable model

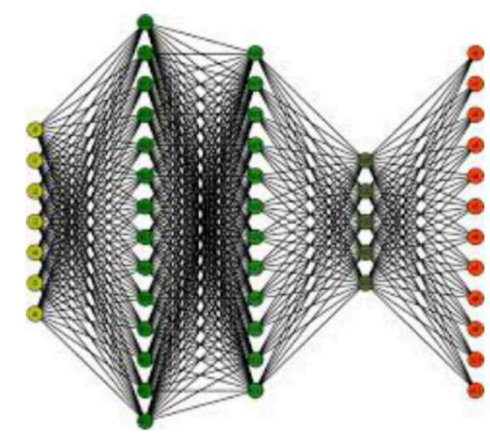


- Linear model
- Decision tree
- Rule-based model

Breaking the trade-off

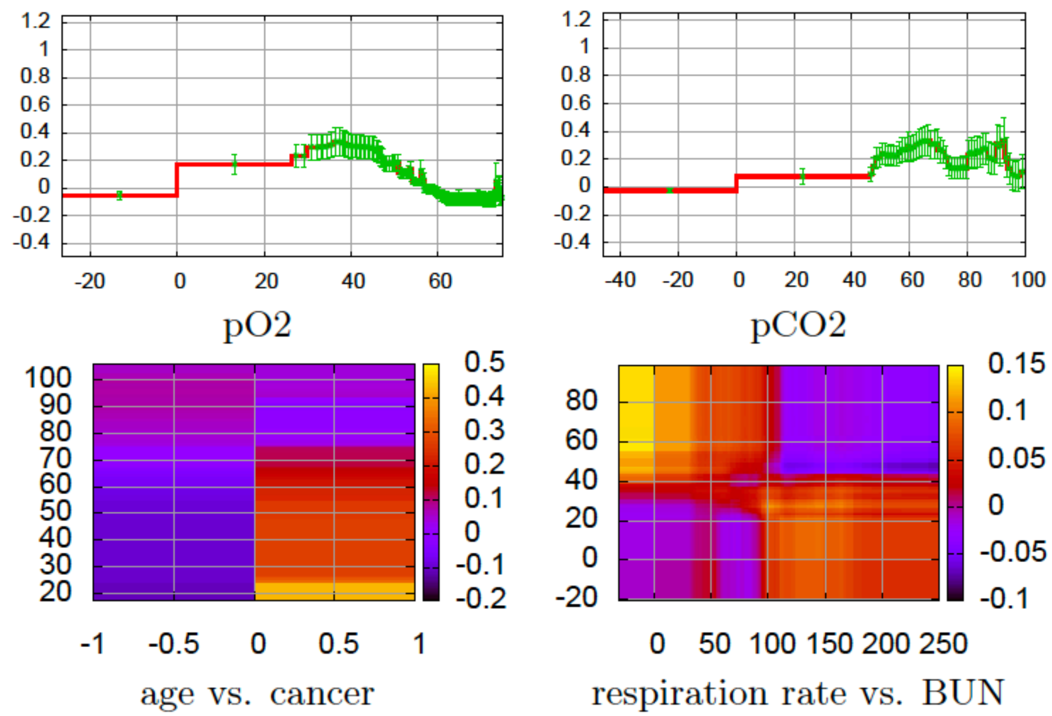
- Generalized linear rule model
- Generalized additive models
- ...

Post-hoc explainability



- Deep neural networks
- Ensemble models

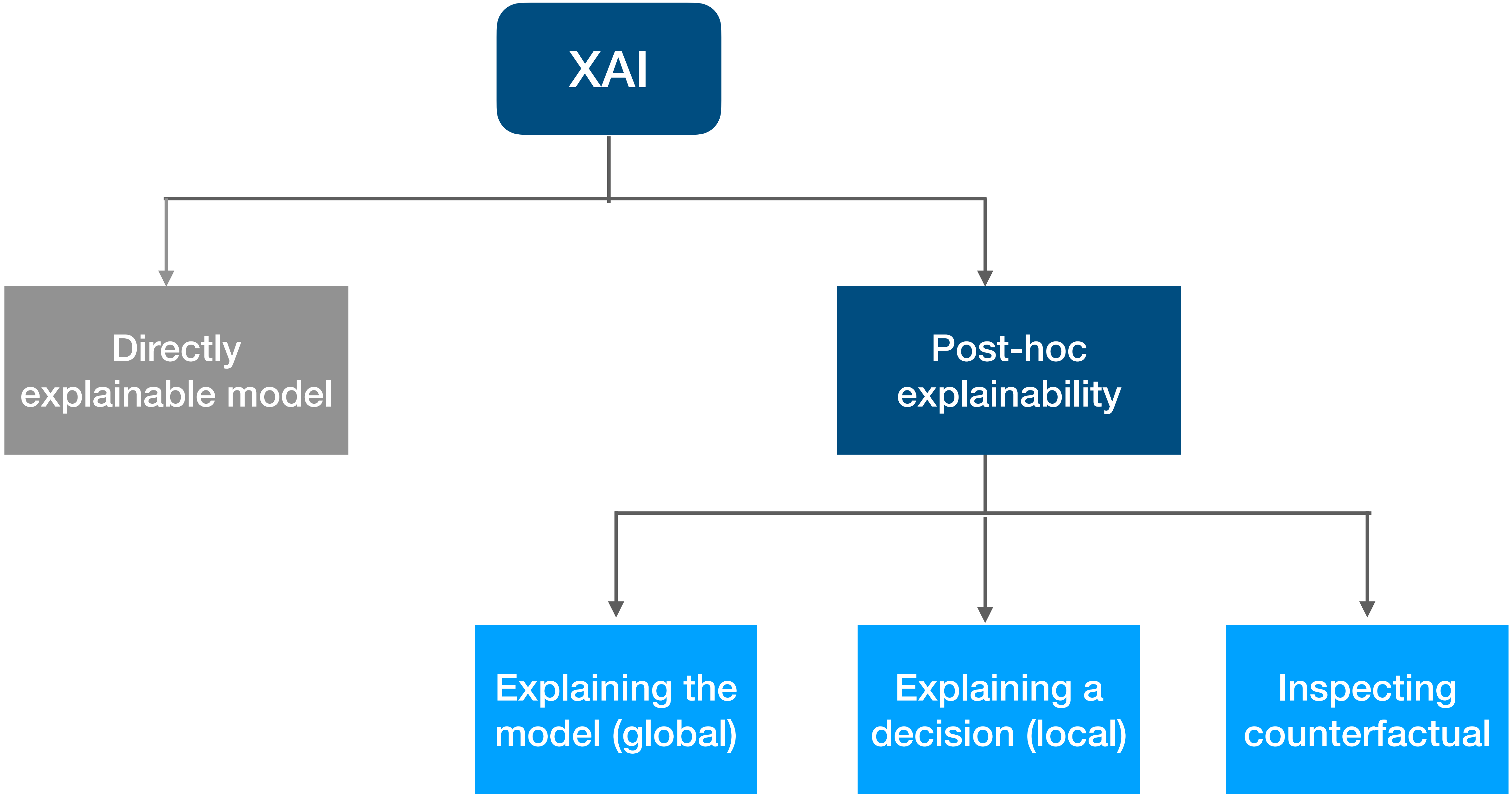
Examples of high-performing directly explainable models



Generalized additive model with pairwise interaction (GA^2M) (Caruana et al., 2015)

Generalized Linear Rule Model (GLRM) (Wei et al., 2019)

Wei et al. Generalized Linear Rule Models. ICML 2019 (**GLRM** for regression: https://github.com/IBM/AIX360/blob/master/aix360/algorithms/rbm/GLRM.py)
Dash et al. Boolean Decision Rules via Column Generation, NeurIPS 2018 (**BRCG** for classification: https://github.com/IBM/AIX360/blob/master/aix360/algorithms/rbm/BRCG.py)
Wang & Rudin (2015). Falling rule lists. In *Artificial Intelligence and Statistics*



Guidotti et al. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* .

XAI

I will:

- Use a **fictional use case** and show fictional explanations
- Focus on **methods**, not algorithmic details
- Provide references to example algorithms at the bottom, and links to code if available in AIX360

A use case： A decision-support ML system for loan application approval

Customer： Jason

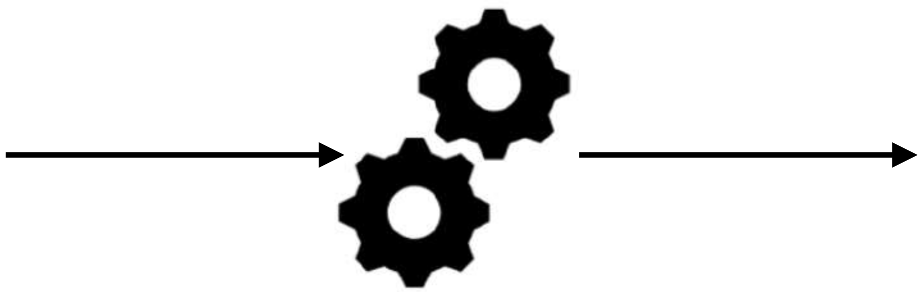
Assets score： **88**

No. Of satisfactory trades: **0**

Mo. since account open： **3**

Number of inquiries： **1**

Debt percentage： **10%**



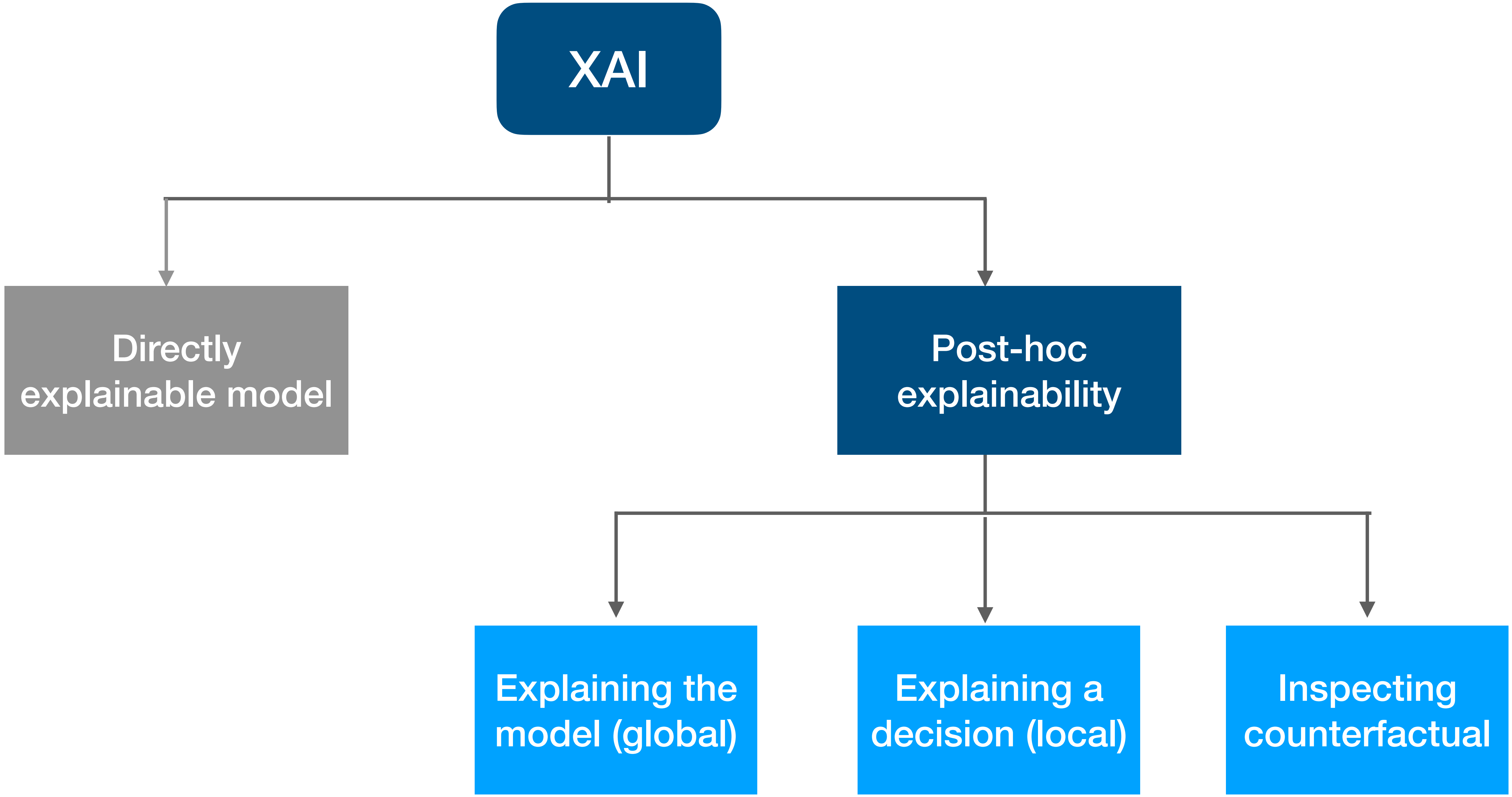
Risk of failing to repay: low

Data scientist
Must ensure the model works appropriately before deployment

Loan officer
Needs to assess the model’s prediction and make the final judgment

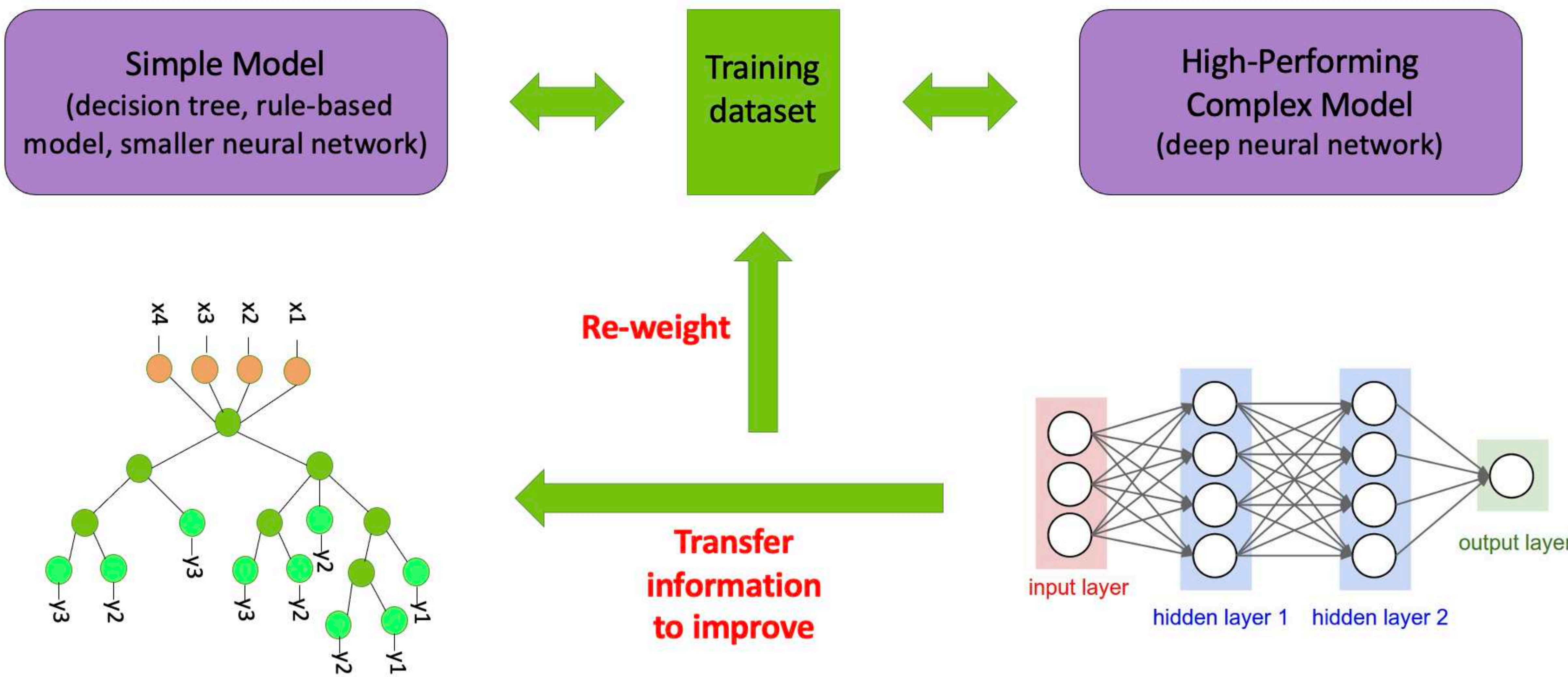
Bank customer
Wants to understand the reason for the application result

Based on FICO XAI Challenge

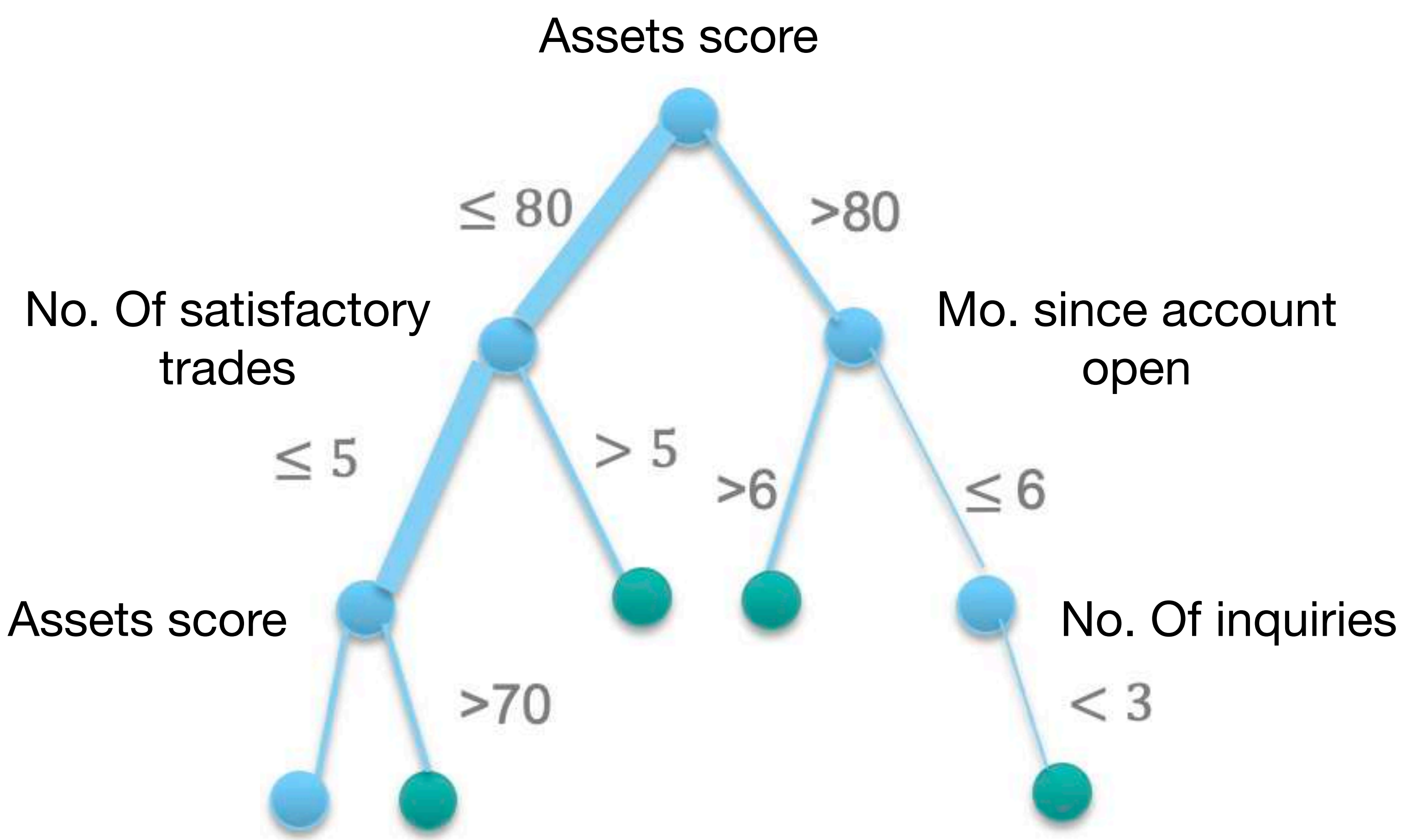


- Feature importance
- Rule approximation
- Decision tree approximation

Post-hoc global explanation: knowledge distillation (approximation)



Explaining the model： decision-tree approximation



How does the model make decision? Is the logic reasonable?

Data Scientist

Dhurandhar et al. Improving Simple Models with Confidence Profiles. NeurIPS 2018 (**ProfWeight**: https://github.com/Trusted-AI/AIX360/blob/master/aix360/algorithms/profwt/profwt.py)

Explaining the model: rule approximation

- If {**assets score> 90, Mo. since account opening>6**}:**Low** risk
- Else if {**Debt percentage< 15**}:**Low** risk

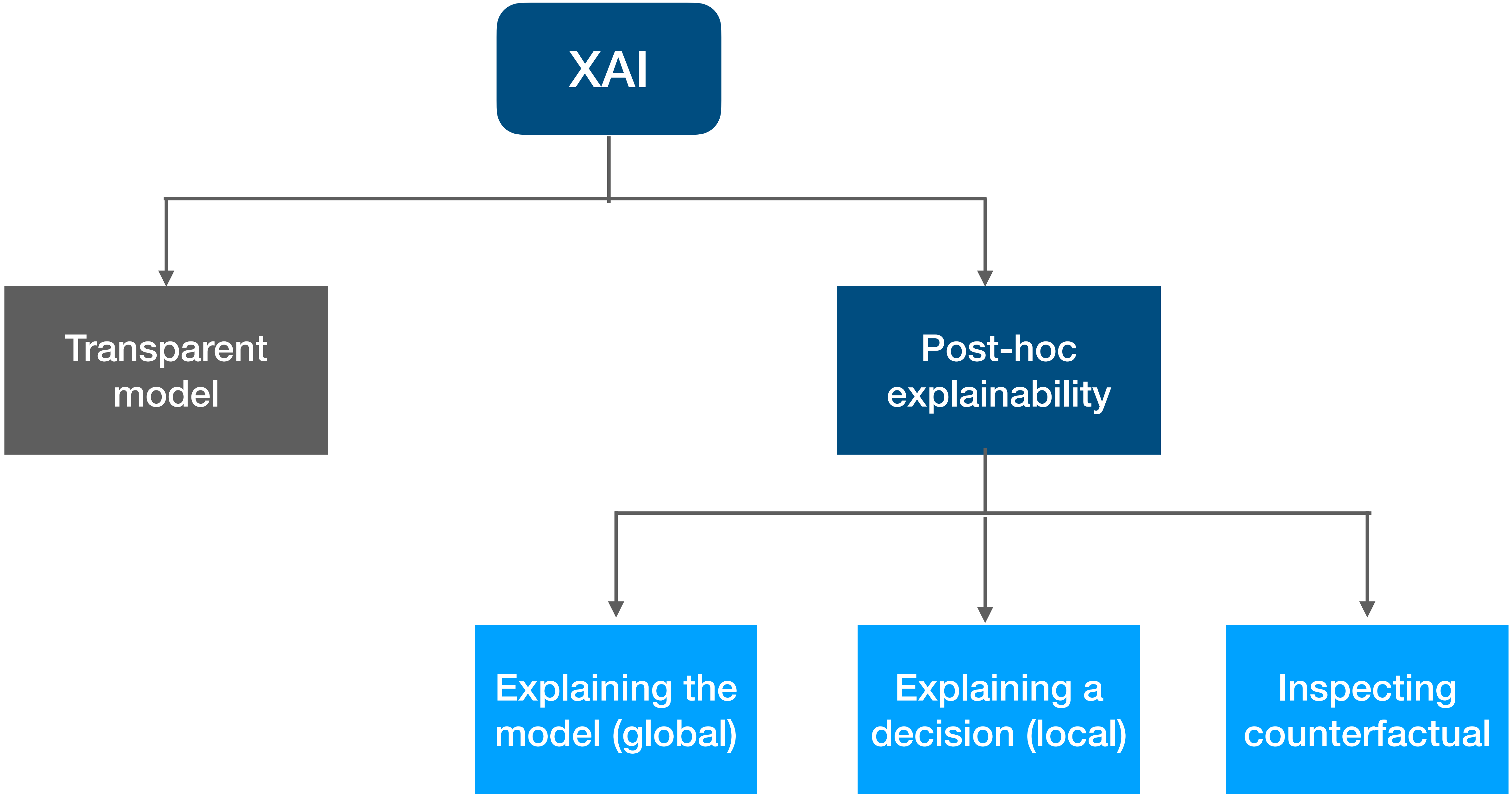
How does the model make decision? Is the logic reasonable?

Data scientist

What kind of customers does the model consider as low risk?

Loan officer

Lakkaraju et al., 2019. Faithful and customizable explanations of black box models. AIES 2019



- Feature importance
- Rule approximation
- Decision tree approximation

- Local feature contribution
- Local rules
- Prototypical examples

Explaining a prediction: local feature contribution

Customer： Jason

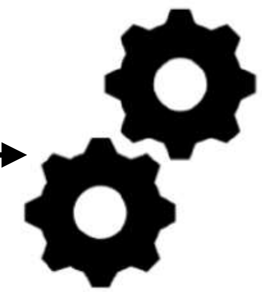
Assets score： **88**

No. Of satisfactory trades: **0**

Mo. since account open： **3**

No. of inquiries： **1**

Debt percentage： **10%**



Risk of failing to repay: low

Loan officer

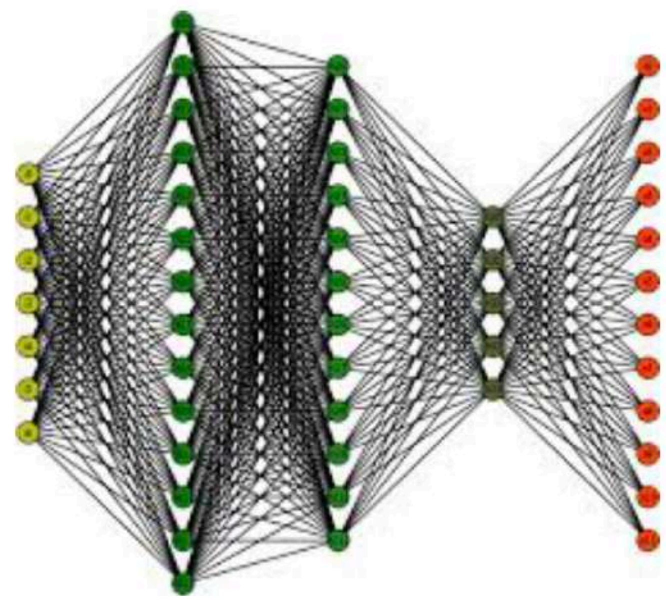
Why is Jason predicted of low risk?

Can I trust this prediction?

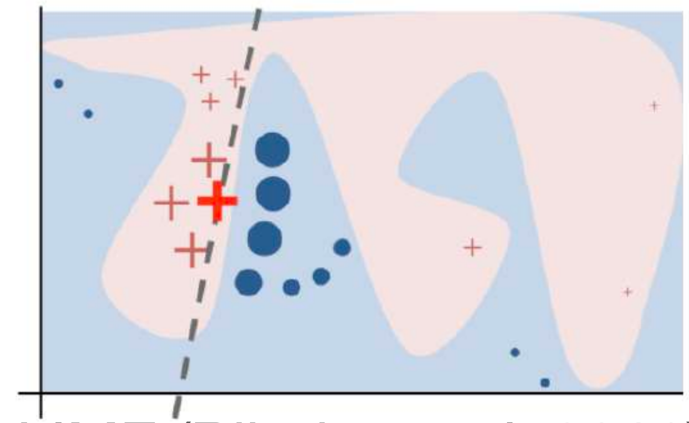
Ribeiro, et al. Why should i trust you? Explaining the predictions of any classifier. KDD 2016 (**LIME:** https://github.com/Trusted-AI/AIX360/blob/master/aix360/algorithms/lime/lime_wrapper.py)

Lundberg and Lee. A Unified Approach to Interpreting Model Predictions. NeurIPS 2016 (**SHAP**:https://github.com/Trusted-AI/AIX360/blob/master/aix360/algorithms/shap/shap_wrapper.py)

XAI “post-hoc” algorithm example: LIME

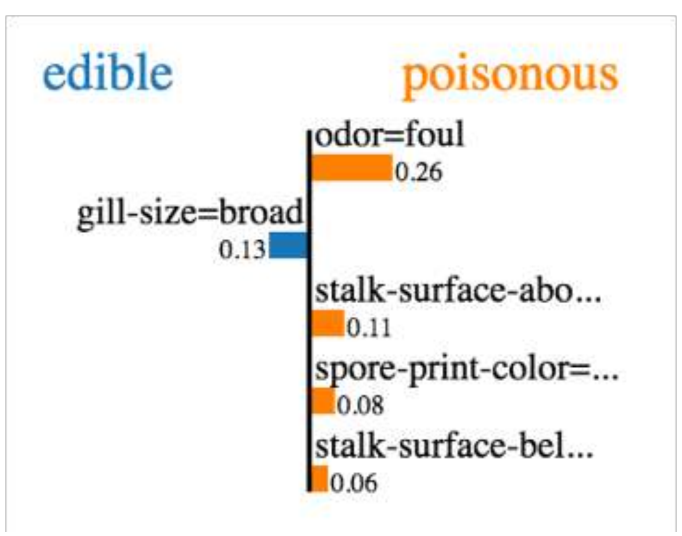


Neural network, not directly explainable

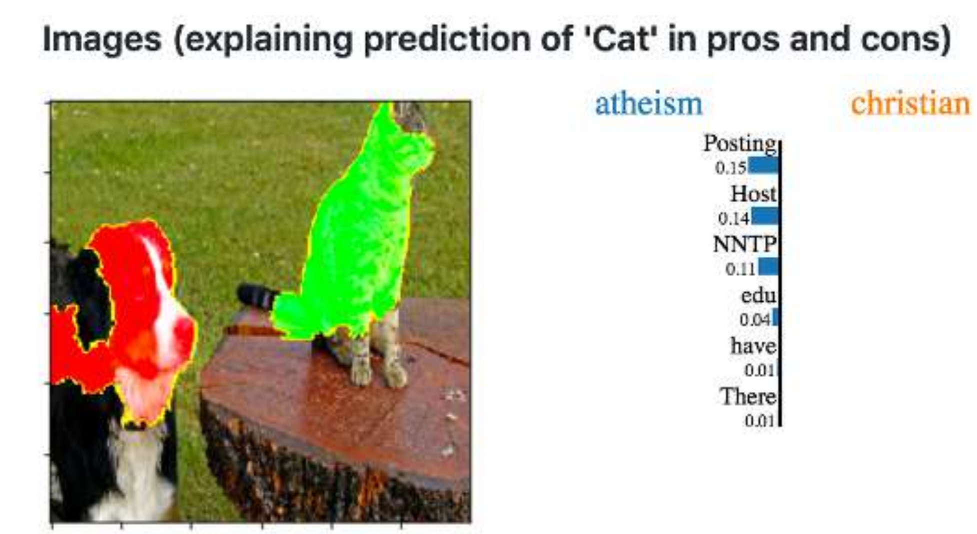


LIME (Ribeiro et al. 2016)

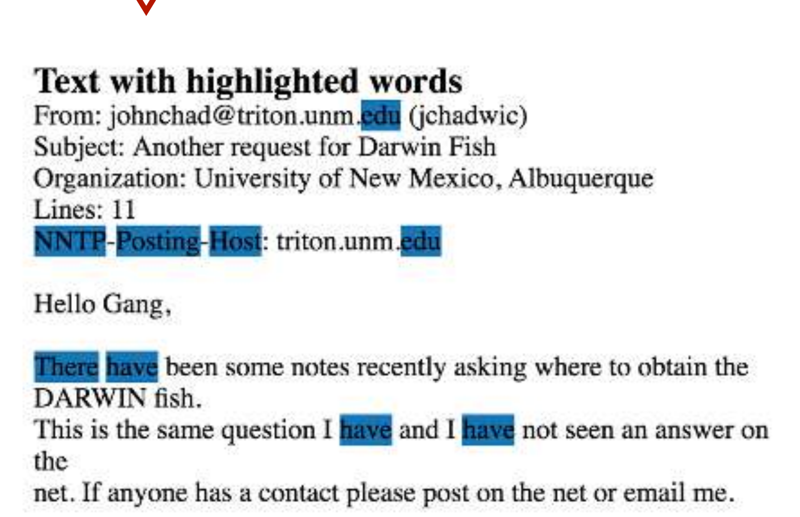
Use a *post-hoc* XAI technique



Tabuler data



Image



Texts

Explaining a prediction: prototypical/similar examples

Customer： Jason
Assets score： **88**
No. Of satisfactory trades: **0**
Mo. since account open： **3**
No. of inquiries： **1**
Debt percentage： **10%**

Risk of failing to repay: low

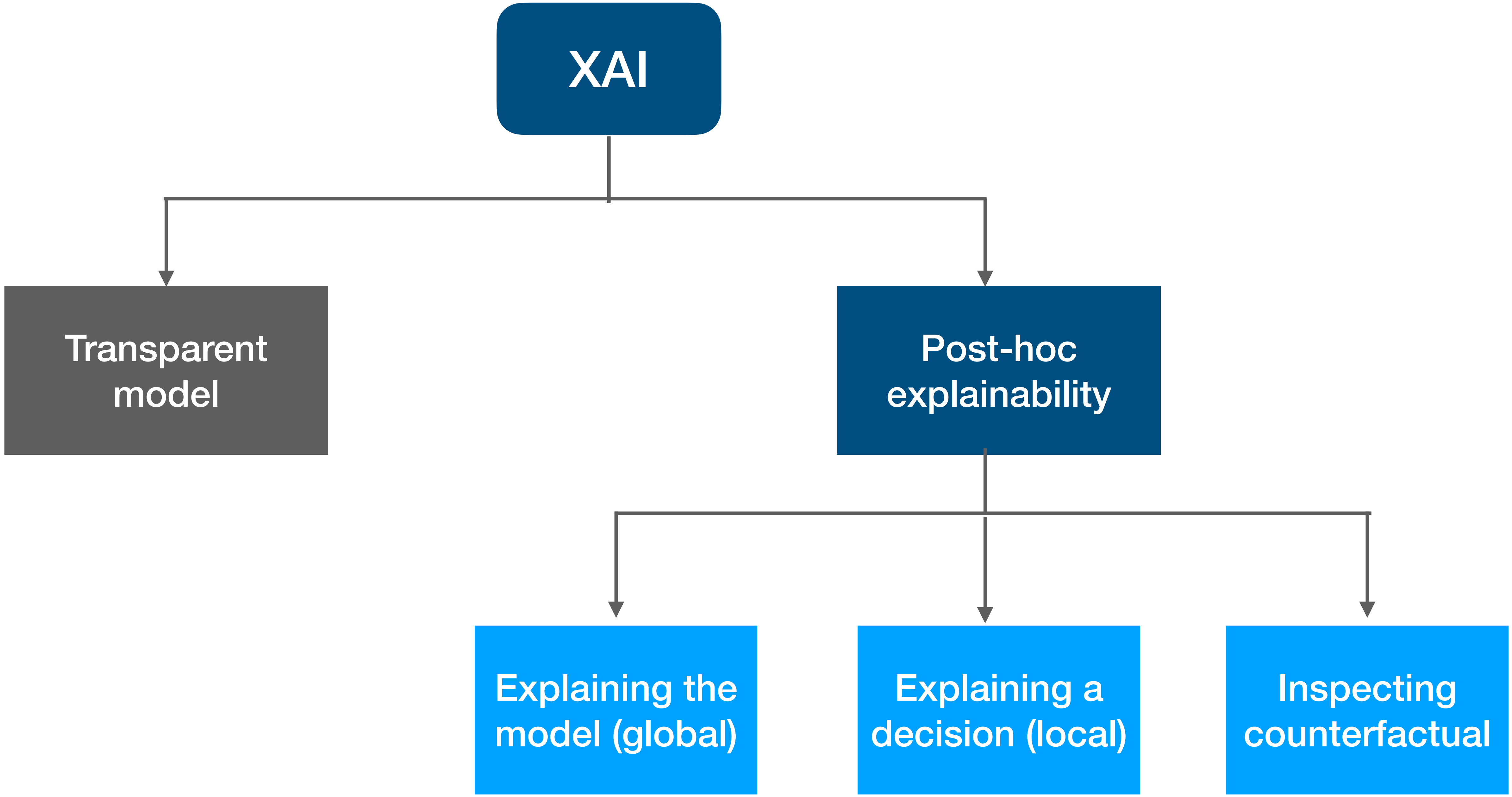
James
Assets score： **86**
No. Of satisfactory trades: **0**
Mo. since account open： **4**
No. of inquiries： **1**
Debt percentage： **7%**
Repaid on time

Danielle
Assets score： **89**
No. Of satisfactory trades: **0**
Mo. since account open： **3**
No. of inquiries： **1**
Debt percentage： **9%**
Repaid on time

Loan officer

Why is Jason predicted of low risk?
Can I trust this prediction?

Gurumoorthy et al. Efficient Data Representation by Selecting Prototypes with Importance Weights", ICDM 2019 (**ProtoDash**: https://github.com/Trusted-AI/AIX360/blob/master/aix360/algorithms/protodash/PDASH.py)

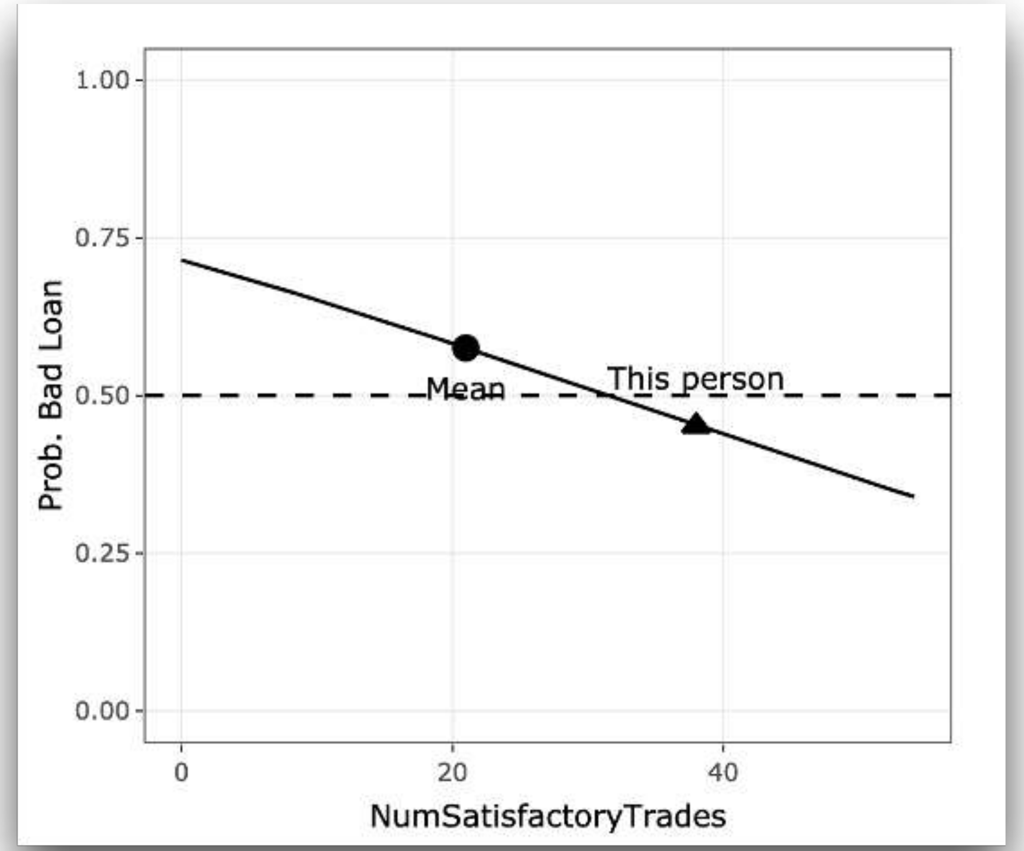


- Feature importance
- Rule approximation
- Decision tree approximation

- Local feature contribution
- Local rules
- Prototypical examples

- Feature influence
- Contrastive features
- Counterfactual examples

Inspecting counterfactual of instance: feature influence



Loan officer

What if Jason fails more trades?

Example techniques: Partial Dependence Plot (PDP), Individual Conditional Expectation (ICE), Accumulated Local Effects (ALE) plot (read in an e-book: https://christophm.github.io/interpretable-ml-book/)

Inspecting counterfactual of prediction: contrastive feature

Customer： Ana

Assets score： **65**

No. Of satisfactory trades: **1**

Mo. since account open： **12**

No. of inquiries： **4**

Debt percentage： **50%**

Risk of failing to repay: high

- If **{debt percentage under 30%},** you will no longer be predicted of high risk

Bank customer

Why was my loan application rejected?
How can I improve in the future?

Dhurandhar, et al. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. NeurIPS 2018
(**CEM**:https://github.com/Trusted-AI/AIX360/blob/master/aix360/algorithms/contrastive/CEM.py)

Inspecting counterfactual of prediction: counterfactual example

Customer： Ana

Assets score： **65**

No. Of satisfactory trades: **1**

Mo. since account open： **12**

No. of inquiries： **4**

Debt percentage： **50%**

Risk of failing to repay: high

Sue

Assets score： **66**

No. Of satisfactory trades: **1**

Mo. since account open： **12**

No. of inquiries： **3**

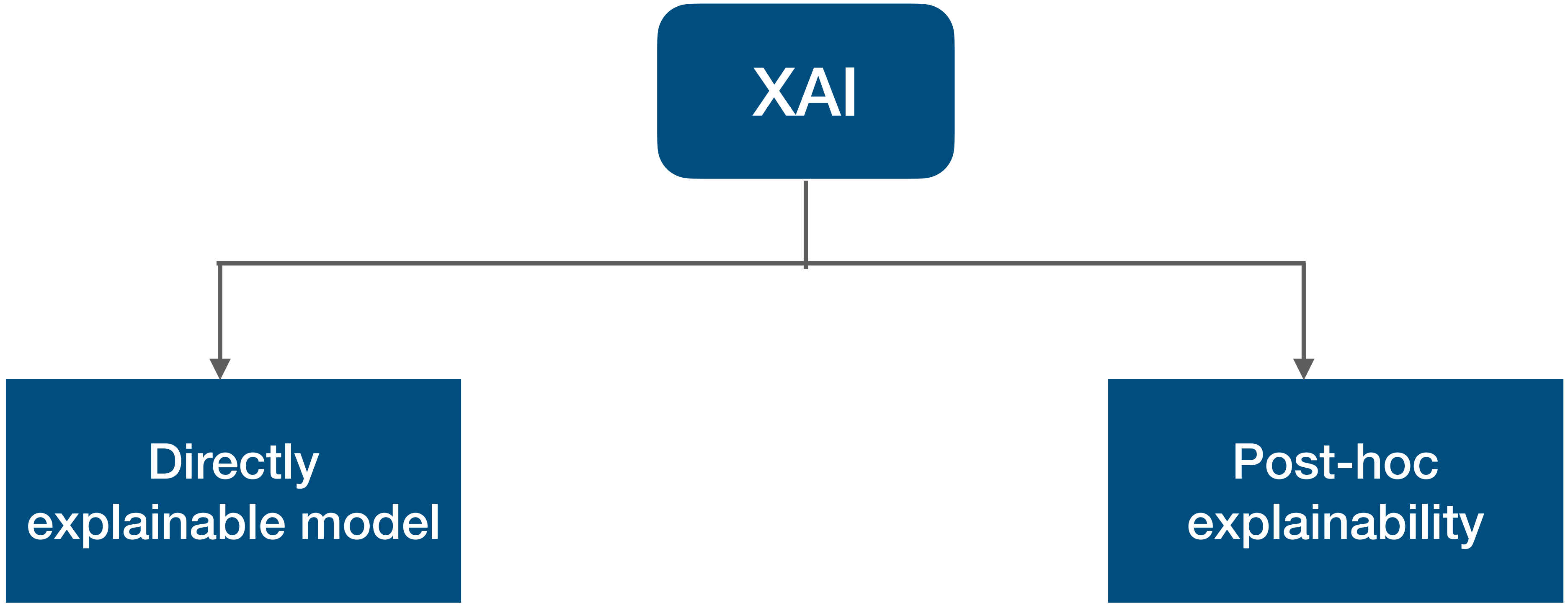
Debt percentage： 28%

Repaid on time

Bank customer

Why was my loan application rejected?
How can I improve in the future?

Mothilal et al. Explaining machine learning classifiers through diverse counterfactual explanations. FAT* 2020



- Not always perform well
- Sometimes take more human effort to train
- Sometimes impossible to train (e.g., using pre-trained or proprietary models)

- Can be applied to any model
- But usually an approximation, not always faithful, much debated topic, see:

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*

Briefly on XAI evaluation

Inherent "goodness" metrics

- **Fidelity/faithfulness**
- Stability
- Compactness
- …

Faithfulness

Correlation between the feature importance assigned by the interpretability algorithm and the effect of features on model accuracy.

User-dependent measures

- Comprehensibility
- Explanation satisfaction
- …

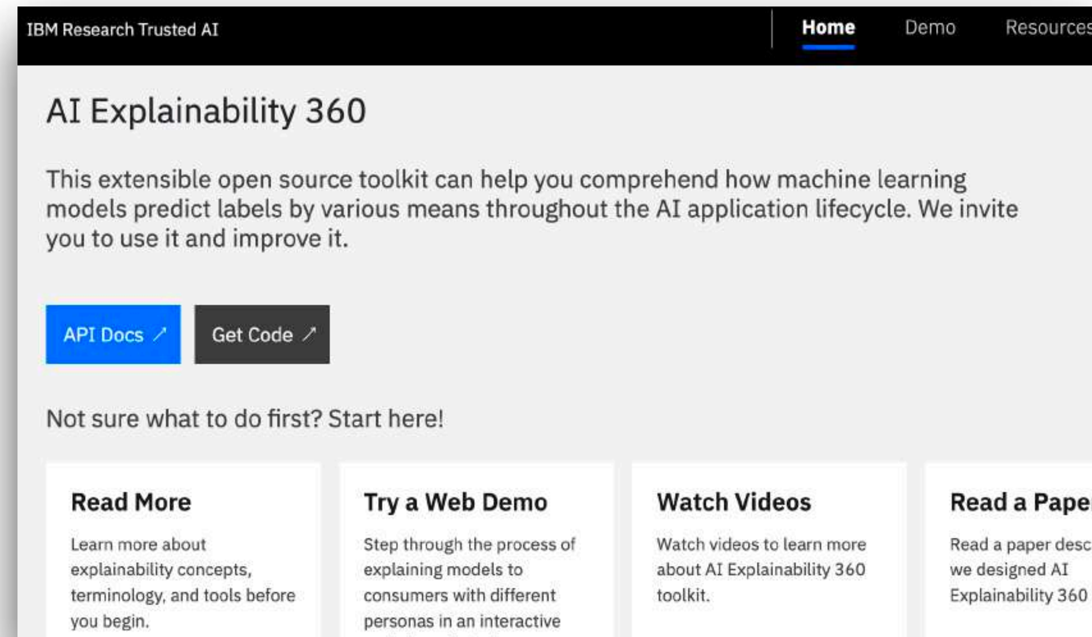
Task oriented measures

- Task performance
- Impact on AI interaction
  - Trust (calibration) in model
- Task or AI system satisfaction

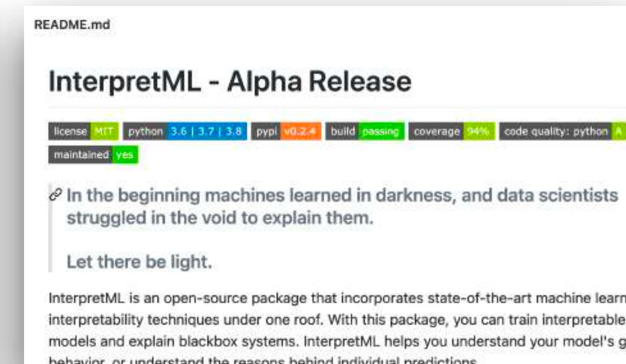
In later slides:user-centered design by identifying **"user requirements"** to satisfy

Carvalho et al. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*
Hoffman et al. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv*
Sokol., & Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. *FAT* 2020*
Doshi-Velez & Kim, (2017). Towards a rigorous science of interpretable machine learning. *arXiv*

XAI open-source toolkits

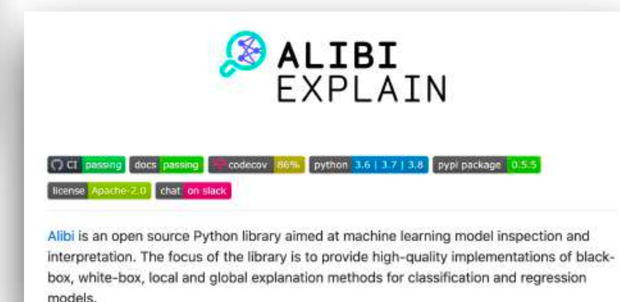


AIX 360
http://aix360.mybluemix.net/

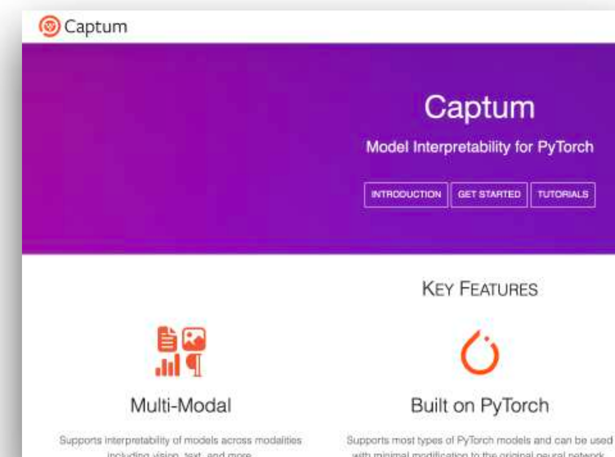


Microsoft Interpret

H2o



Sheldon Alibi



PyTorch Captum

Oracle Skater

Arya, et al. (2019). [One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques](). *arXiv*

Why is XAI important?

Why is XAI the foundation for responsible AI?

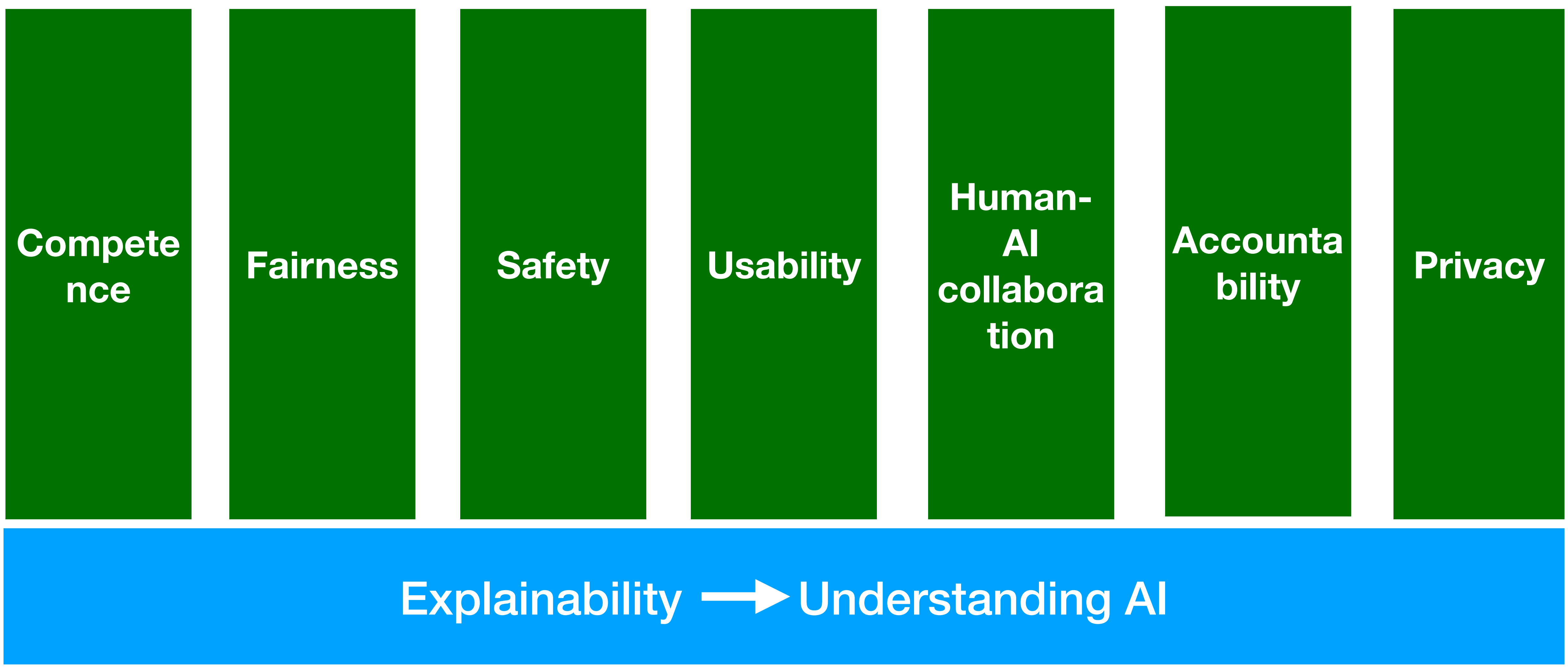
Responsible/ethical/trustworthy AI

	Berkman Klein Center	IEEE Ethically Aligned Design
Close Match	Accountability	Accountability
	Transparency & explainability	Transparency
	Promotion of human values	Human rights
	Safety & security	Well-being
Similar	Human control of technology	Effectiveness
	Fairness & non-discrimination	Awareness of misuse
	Professional responsibility	Competence
	Privacy	Data agency

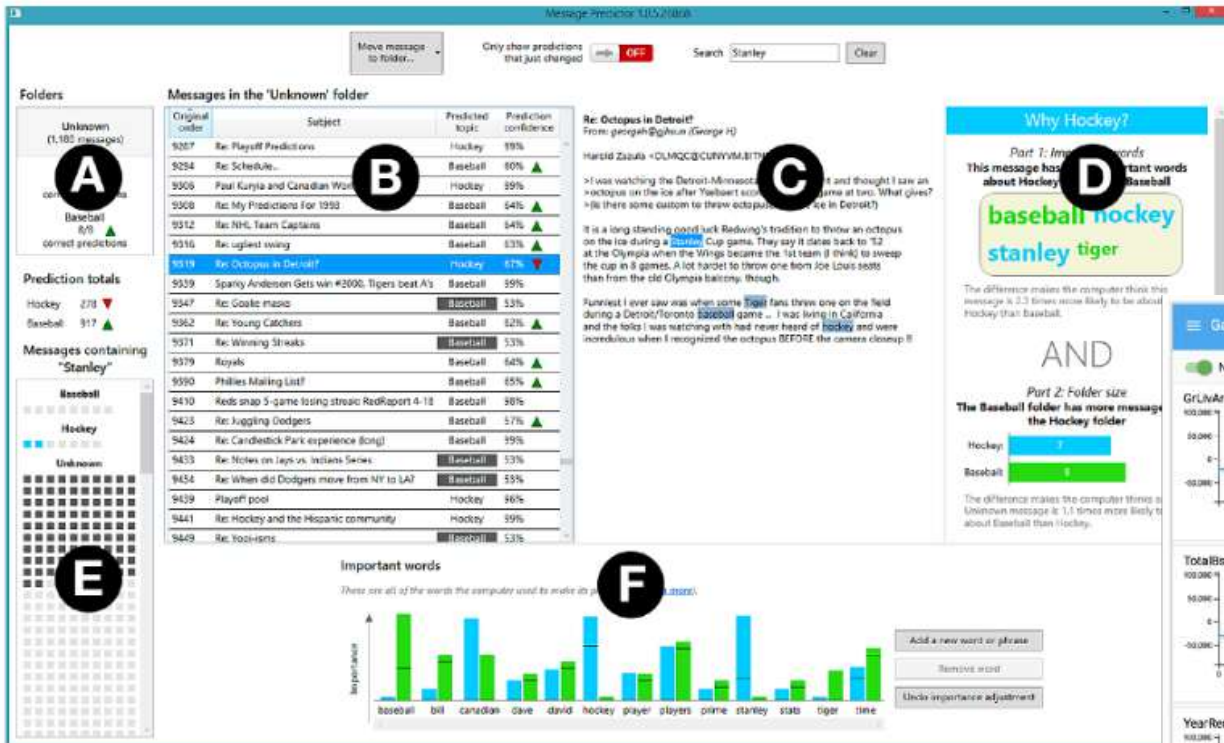
https://cyber.harvard.edu/publication/2020/principled-ai
https://ethicsinaction.ieee.org/

(Shneiderman, 2021)

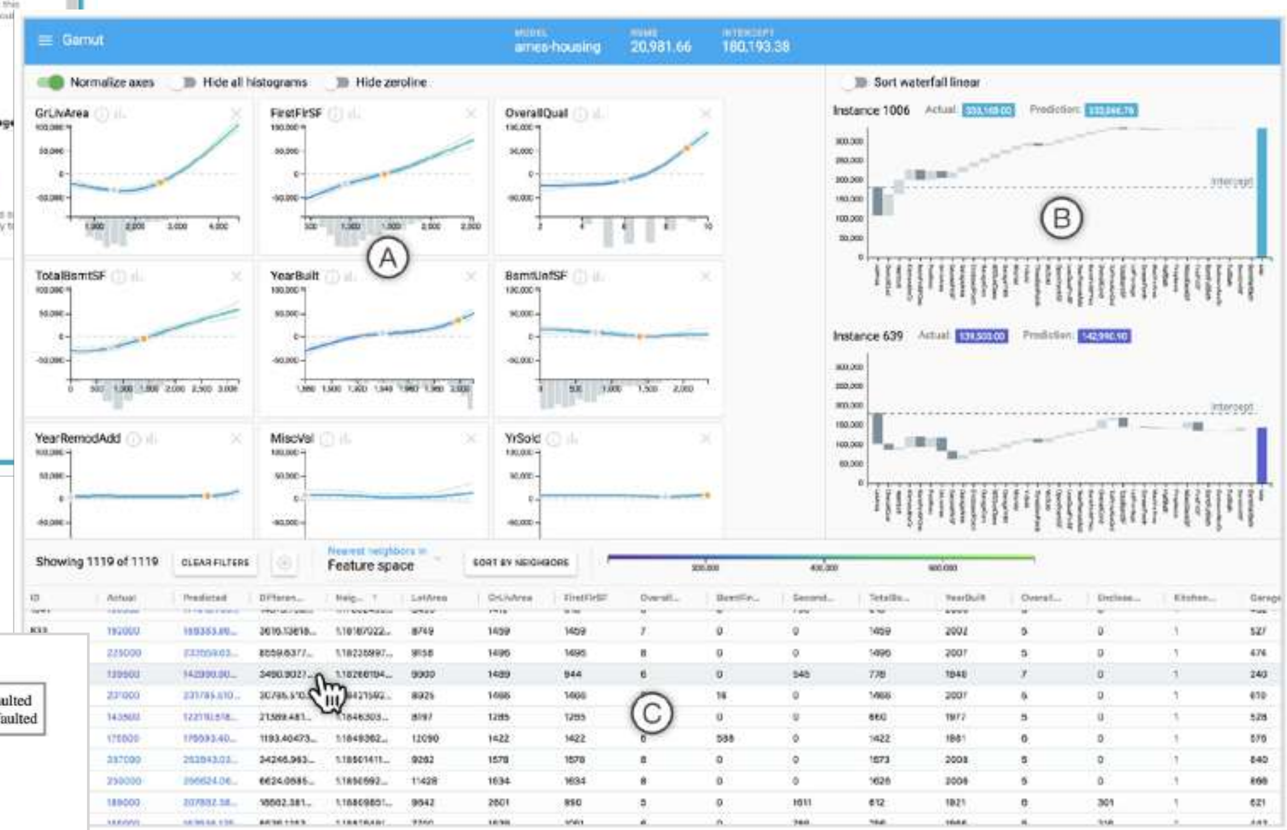
Explainability as the foundation for responsible AI



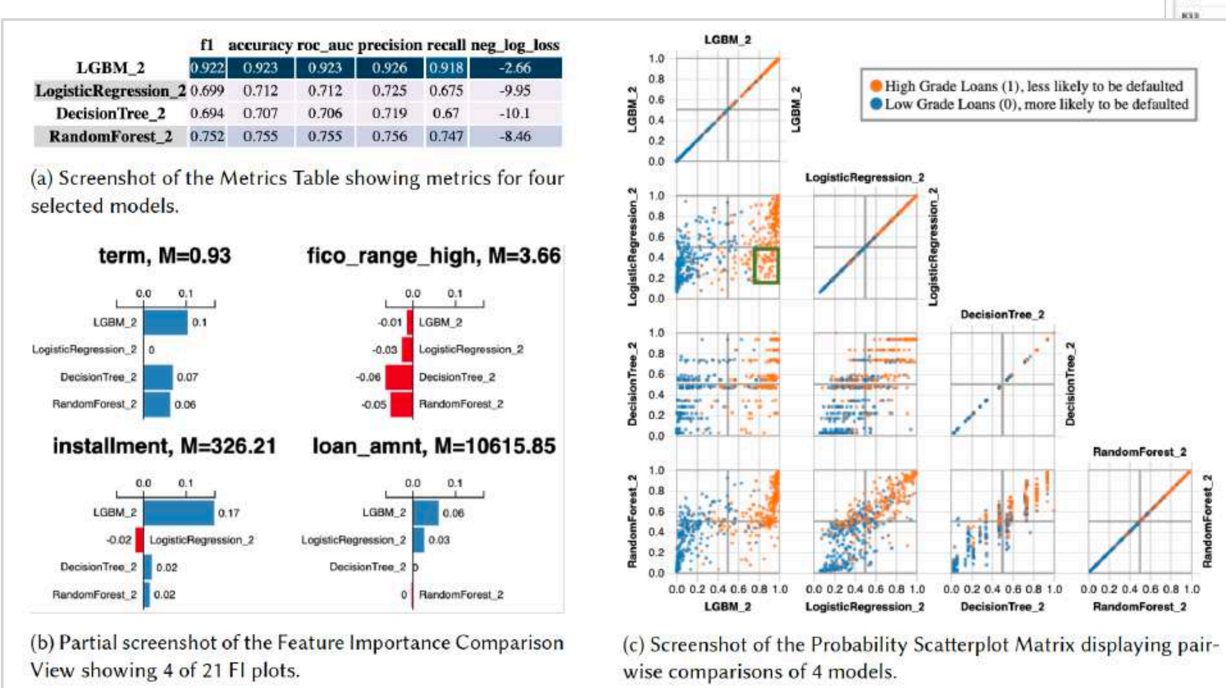
XAI for improving model (competence)



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

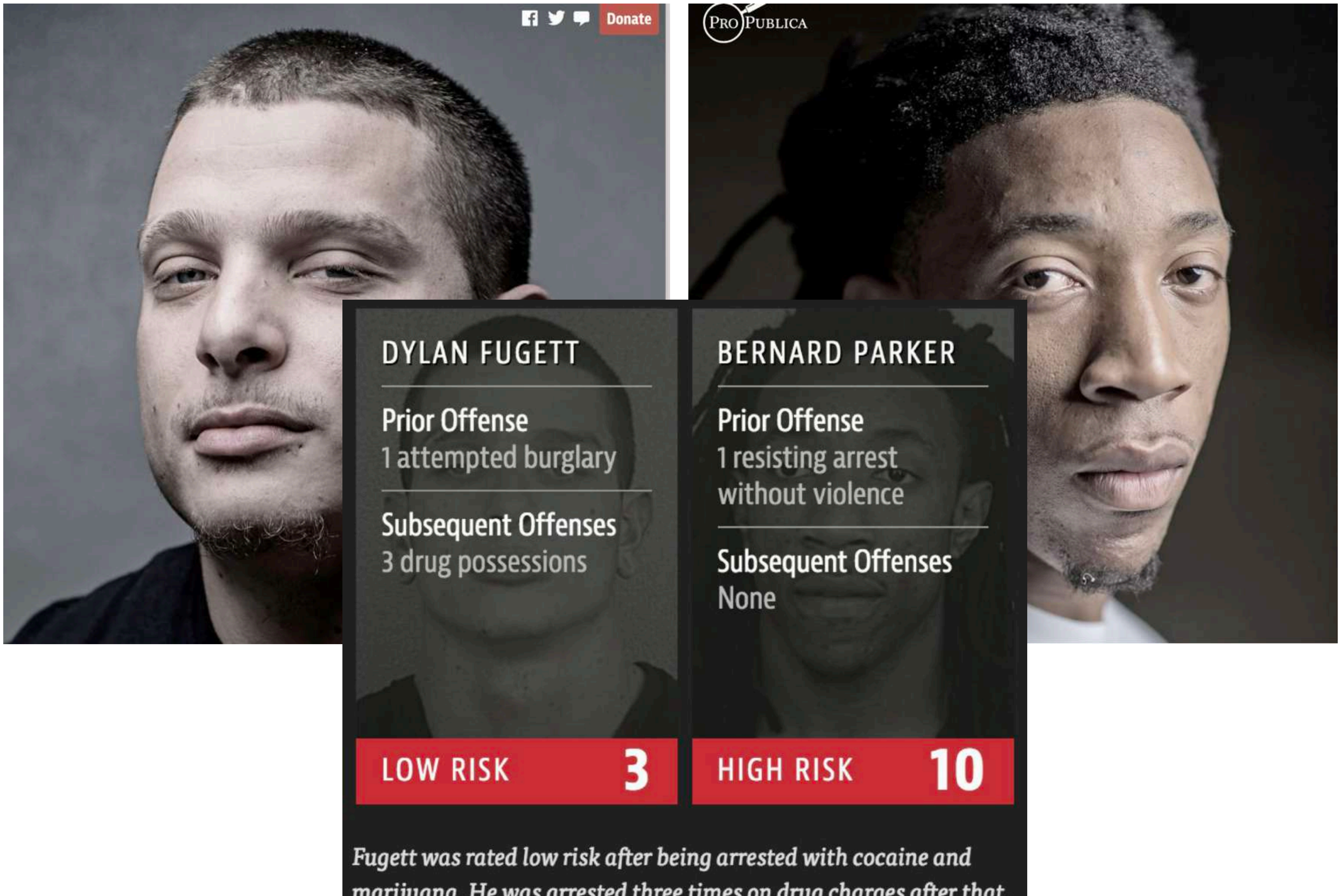
Fair ML: What is unwanted bias?

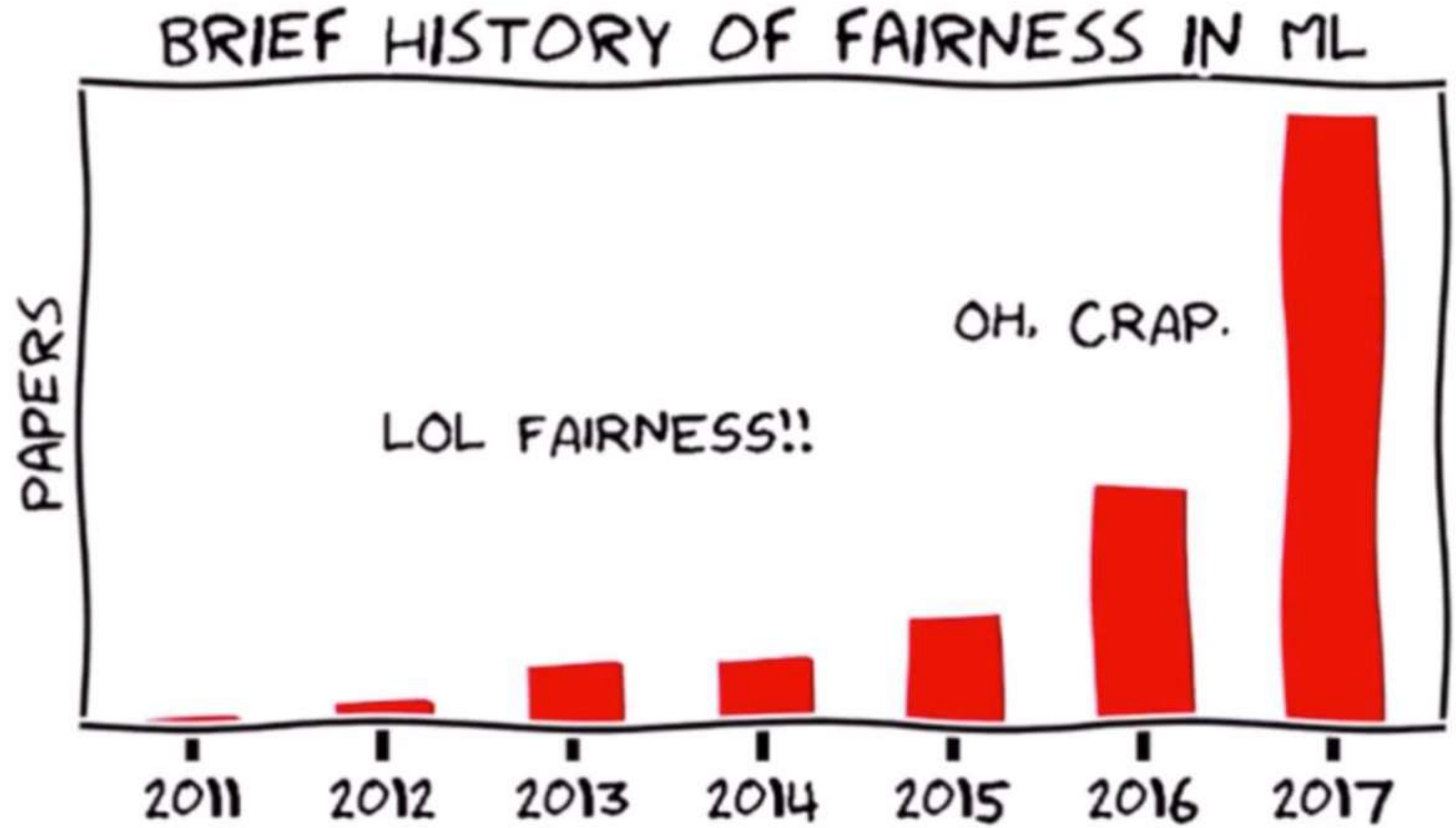
Discrimination becomes objectionable when it places certain **unprivileged** groups at a systematic disadvantage

Illegal in certain contexts

(Barocas and Selbst, 2017)

Discrimination in COMPAS




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

(Hardt, 2017)

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

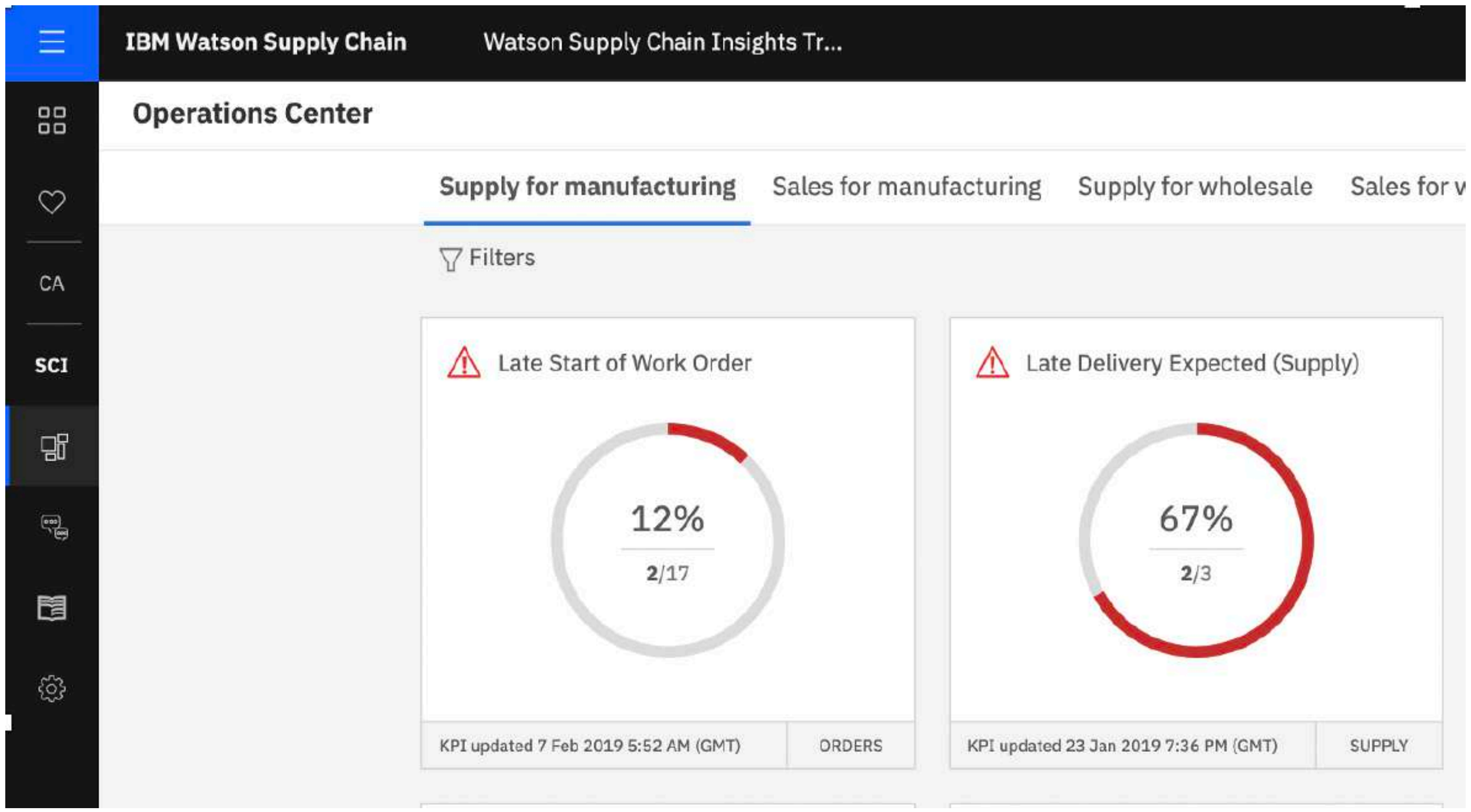
Has juvenile priors:

Explain a prediction:
Individual fairness

Explain the model:
Group fairness

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

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)

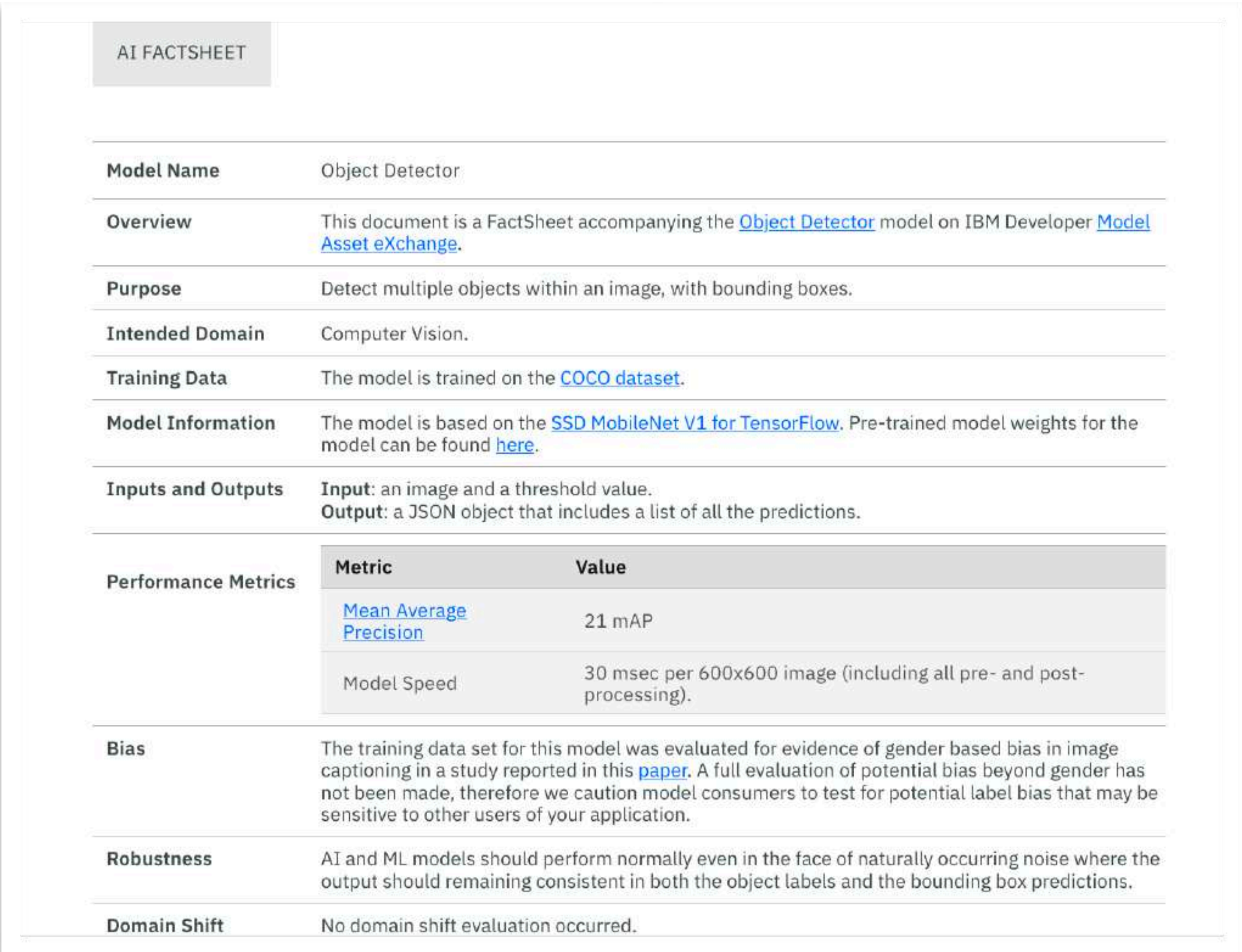
Liao et al. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. CHI 2020

XAI for better control and human-AI collaboration

“ *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)

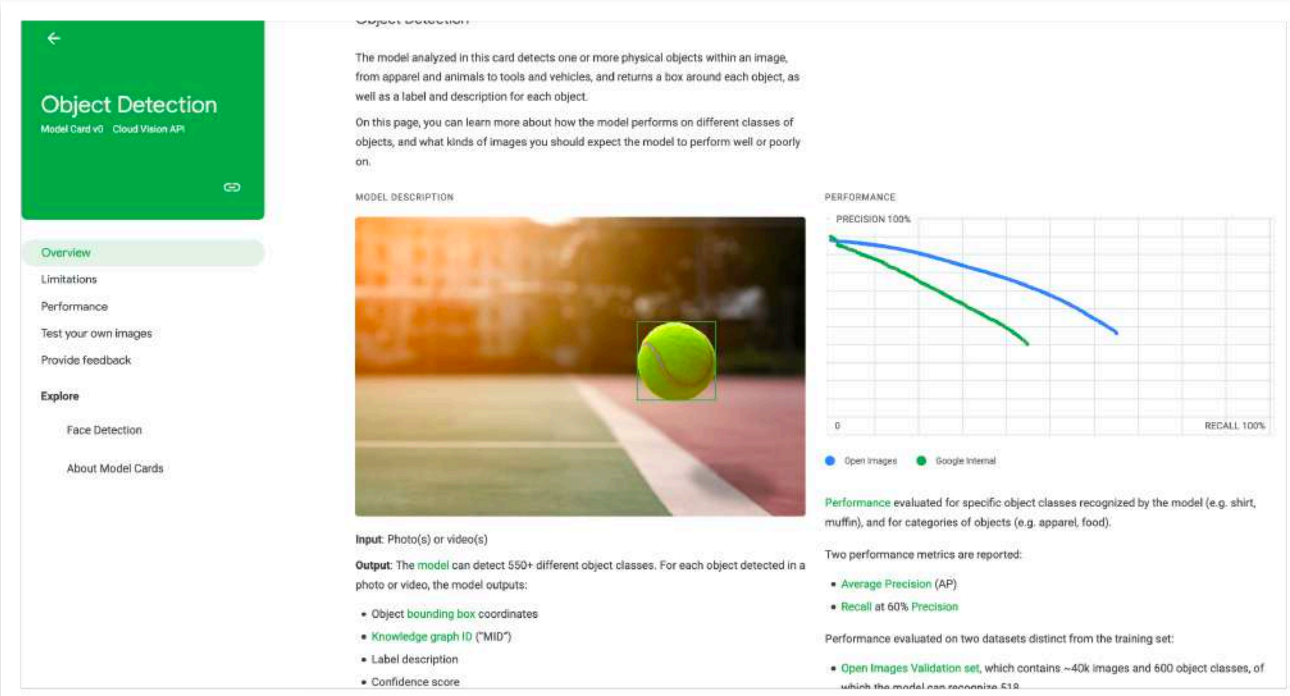
Liao et al. [Questioning the AI: Informing Design Practices for Explainable AI User Experiences](). CHI 2020

Trends: AI documentation and governance (accountability)



AI FACTSHEET

Model Name	Object Detector
Overview	This document is a FactSheet accompanying the Object Detector model on IBM Developer Model Asset eXchange.
Purpose	Detect multiple objects within an image, with bounding boxes.
Intended Domain	Computer Vision.
Training Data	The model is trained on the COCO dataset.
Model Information	The model is based on the SSD MobileNet V1 for TensorFlow. Pre-trained model weights for the model can be found here.
Inputs and Outputs	**Input**: an image and a threshold value. **Output**: a JSON object that includes a list of all the predictions.
Performance Metrics	**Metric** / **Value** Mean Average Precision: 21 mAP Model Speed: 30 msec per 600x600 image (including all pre- and post-processing).
Bias	The training data set for this model was evaluated for evidence of gender based bias in image captioning in a study reported in this paper. A full evaluation of potential bias beyond gender has not been made, therefore we caution model consumers to test for potential label bias that may be sensitive to other users of your application.
Robustness	AI and ML models should perform normally even in the face of naturally occurring noise where the output should remaining consistent in both the object labels and the bounding box predictions.
Domain Shift	No domain shift evaluation occurred.



IBM FactSheets

Google Model Cards

How to design XAI UX?

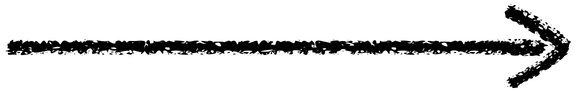
~~How to design XAI UX?~~

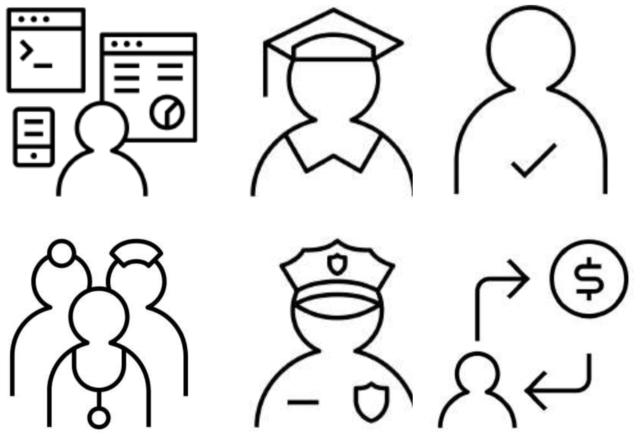
What are the design challenges?

What are some solutions explored?

XAI design as activities from XAI algorithms to XAI UX

A toolbox of XAI techniques





XAI UX

How to **select**?

How to **translate**?

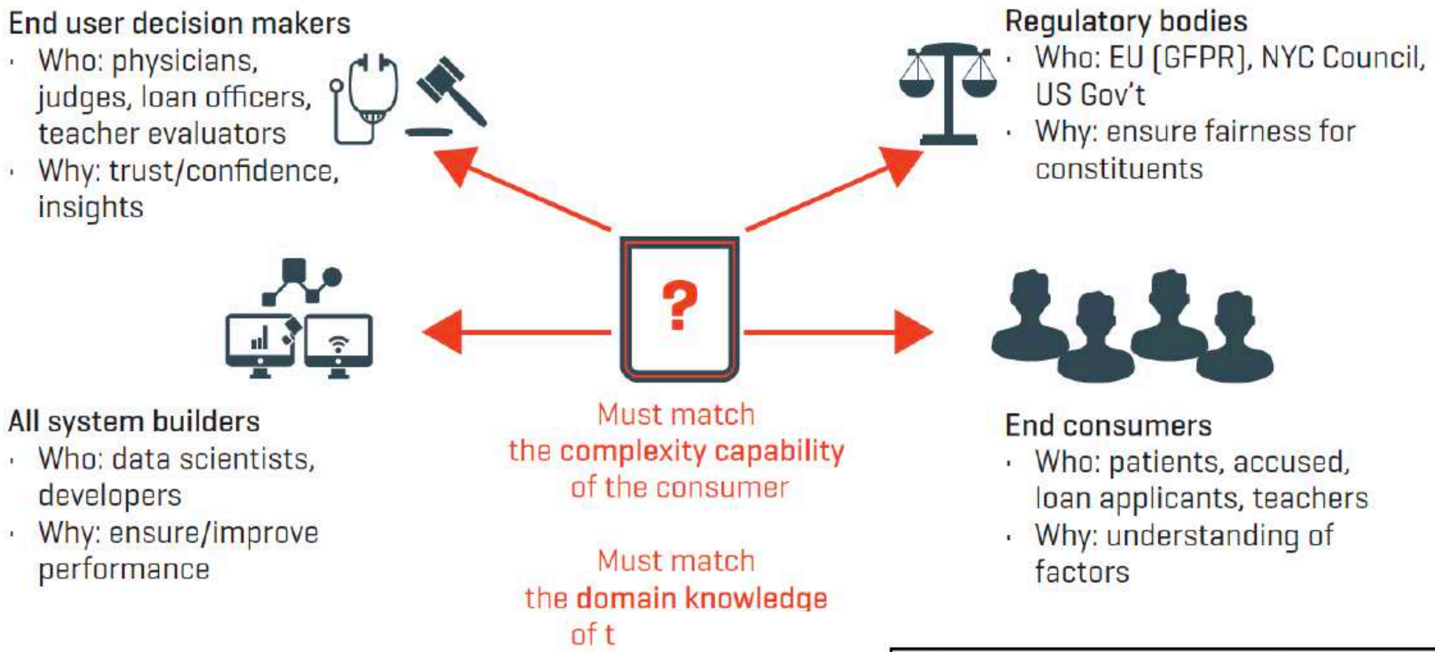
Design Challenge 1: No one-fits-all solutions

Many objectives

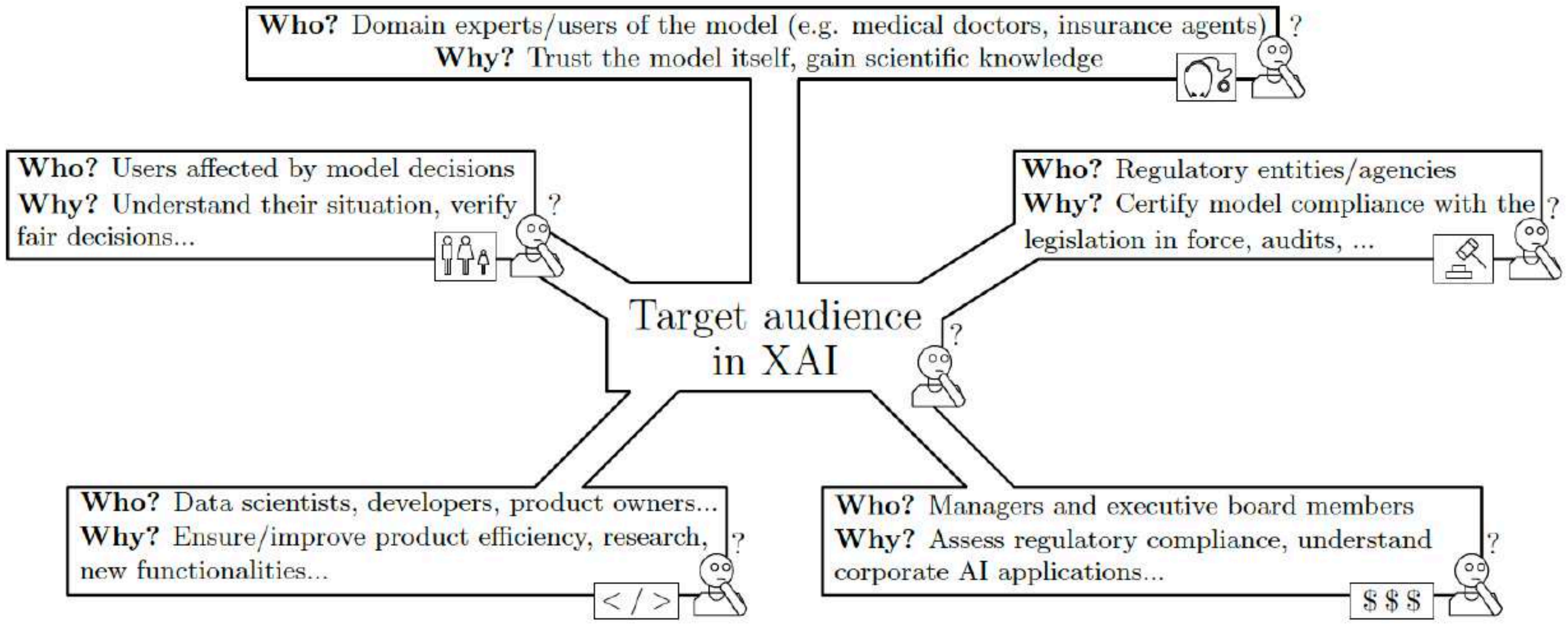
Compete nce	Fairness	Safety	Usability	Human- AI collabora tion	Accounta bility	Privacy

Explainability → Understanding AI

Many user groups



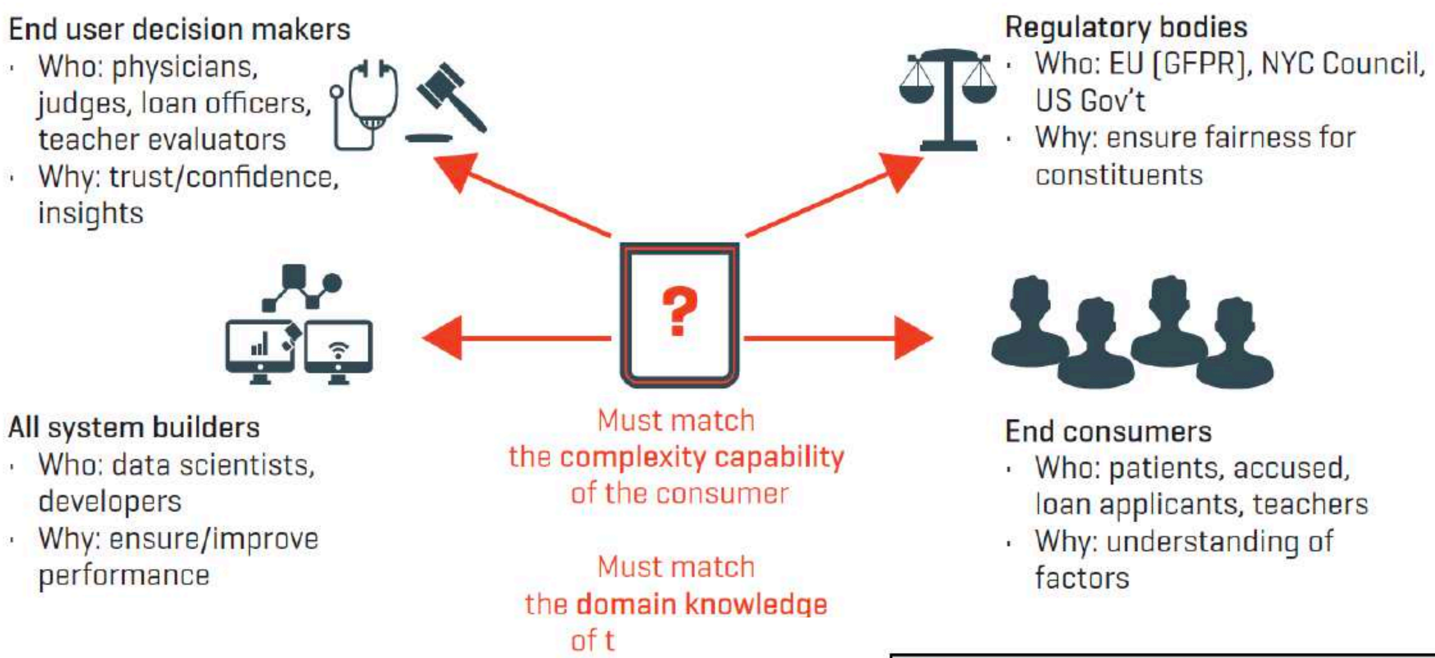
(Hind et al., 2019)



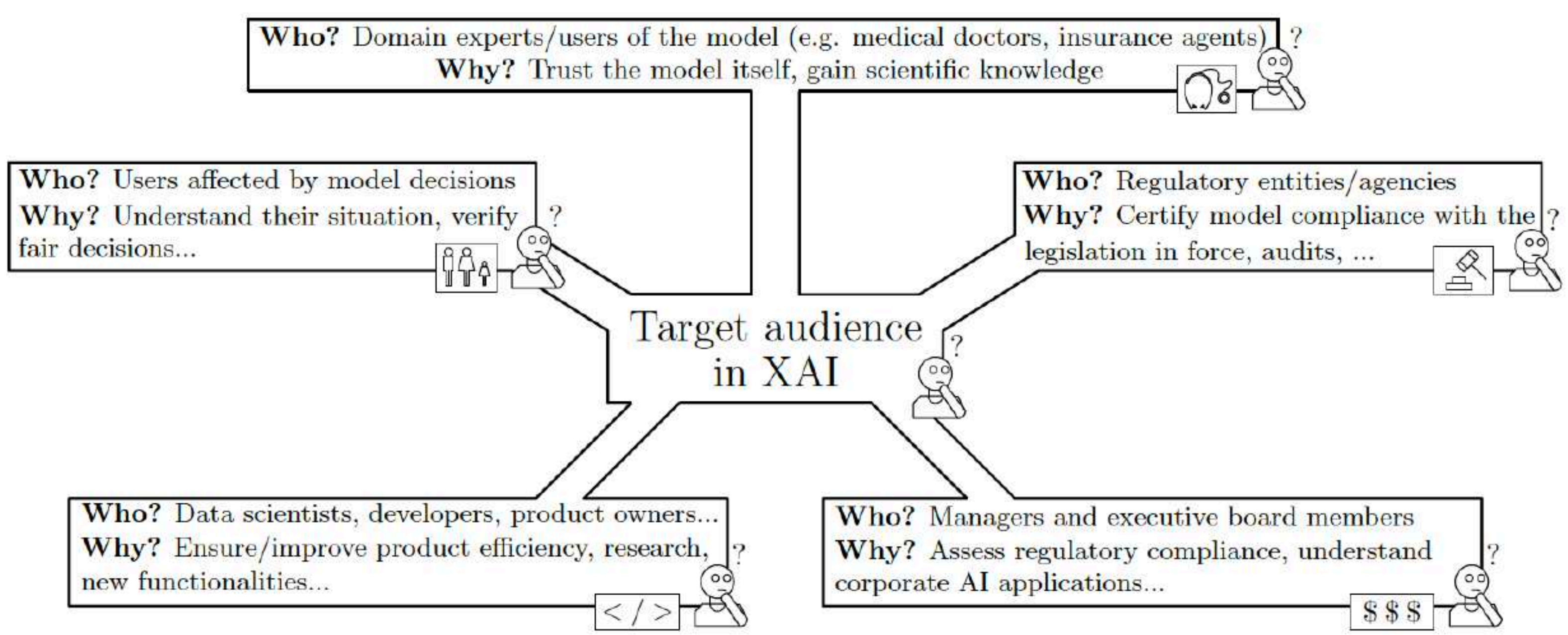
- Model developers
- Domain experts
- Regulators
- Business owners
- Decision-makers
- Impacted groups

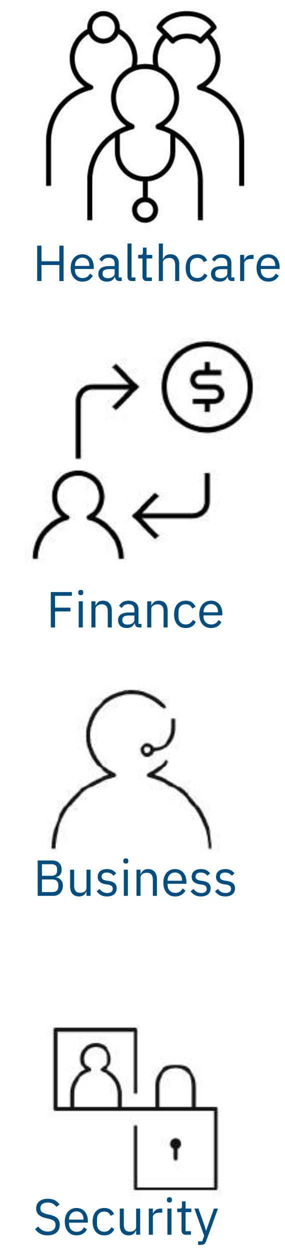
(Arrieta et al, 2019)

Many user groups+many domains+social contexts



(Hind et al., 2019)





- Model developers
- Domain experts
- Regulators
- Business owners
- Decision-makers
- Impacted groups

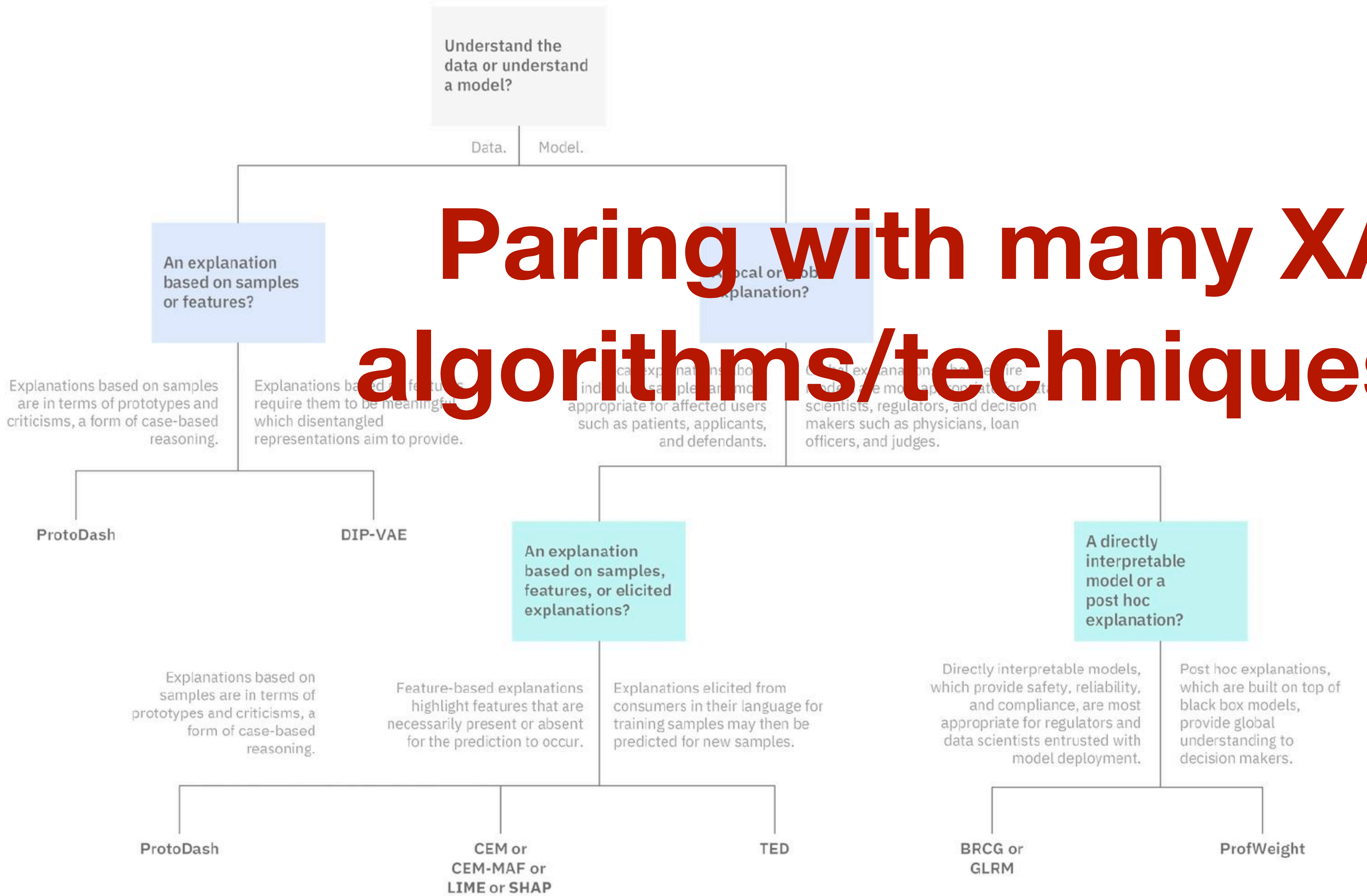
(Arrieta et al, 2019)

AI Explainability 360 - Resources

Guidance on choosing algorithms

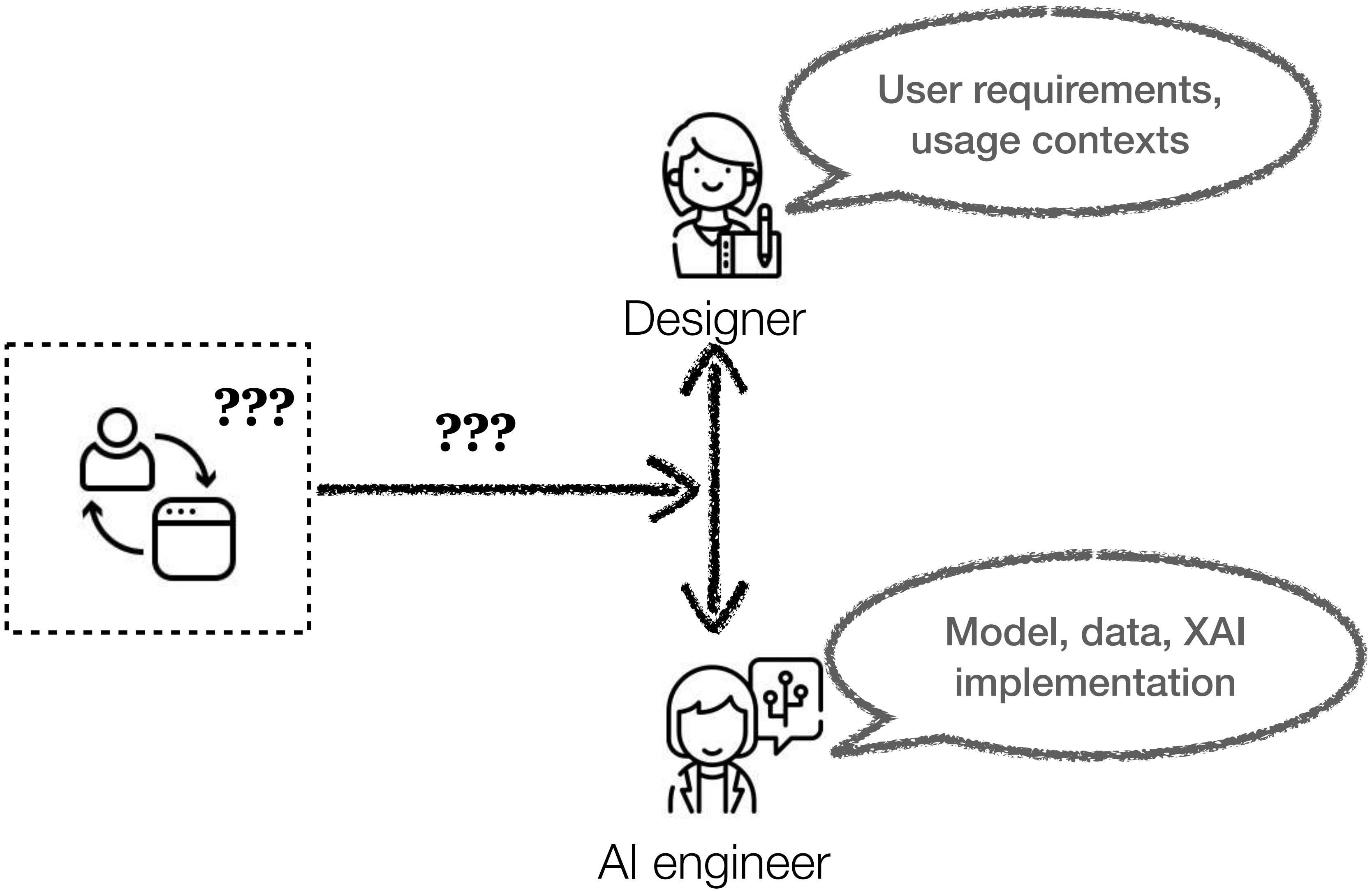
AI Explainability 360 (AIX360) includes many different algorithms capturing many ways of explaining [1], which may result in a daunting problem of selecting the right one for a given application. We provide some guidance to help. The following decision tree will help you in selecting. The text below provides further exposition.

Paring with many XAI algorithms/techniques??



source: IBM Research AI Explainability 360

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)

User needs for explainability = Questions

> *An explanation is an answer to a question* (Wellman, 2011; Miller 2018)
>
> *Explanatory relevance and effectiveness depends on the question asked* (Bromberger, 1992; Hilton, 1990; Walton, 2004)
>
> "*Intelligibility types*": why, how-to, why not, what if… (Lim and Dei, 2019)

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)

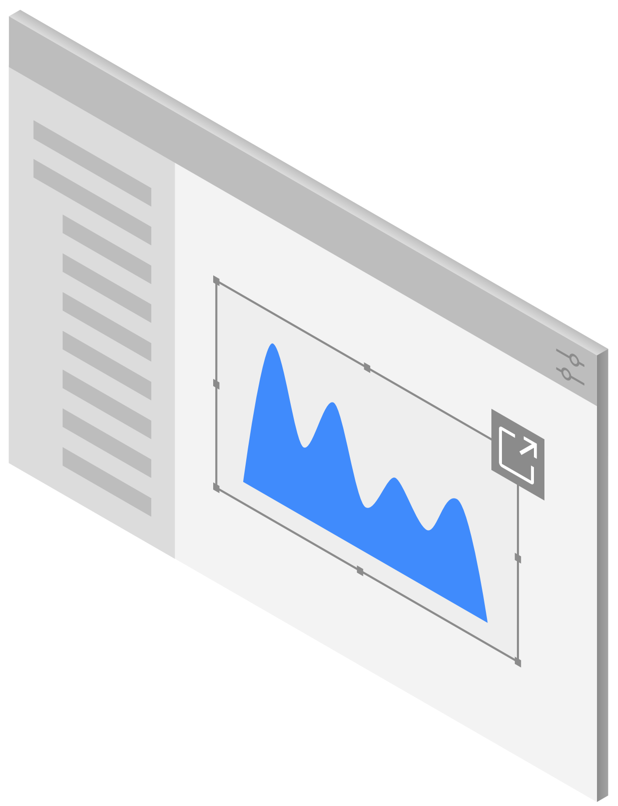
A running example

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

Task description	An AI based dashboard presents patients' readmission risk scores to help clinicians to identify high-risk
User Journey (optional)	
Questions from User 1	
Questions from User 2	

Identify relevant questions

Elicit user questions to identify what types of explanation are needed

Also collect the intention and expectation behind these user questions

What are the main risk factors for this person?

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

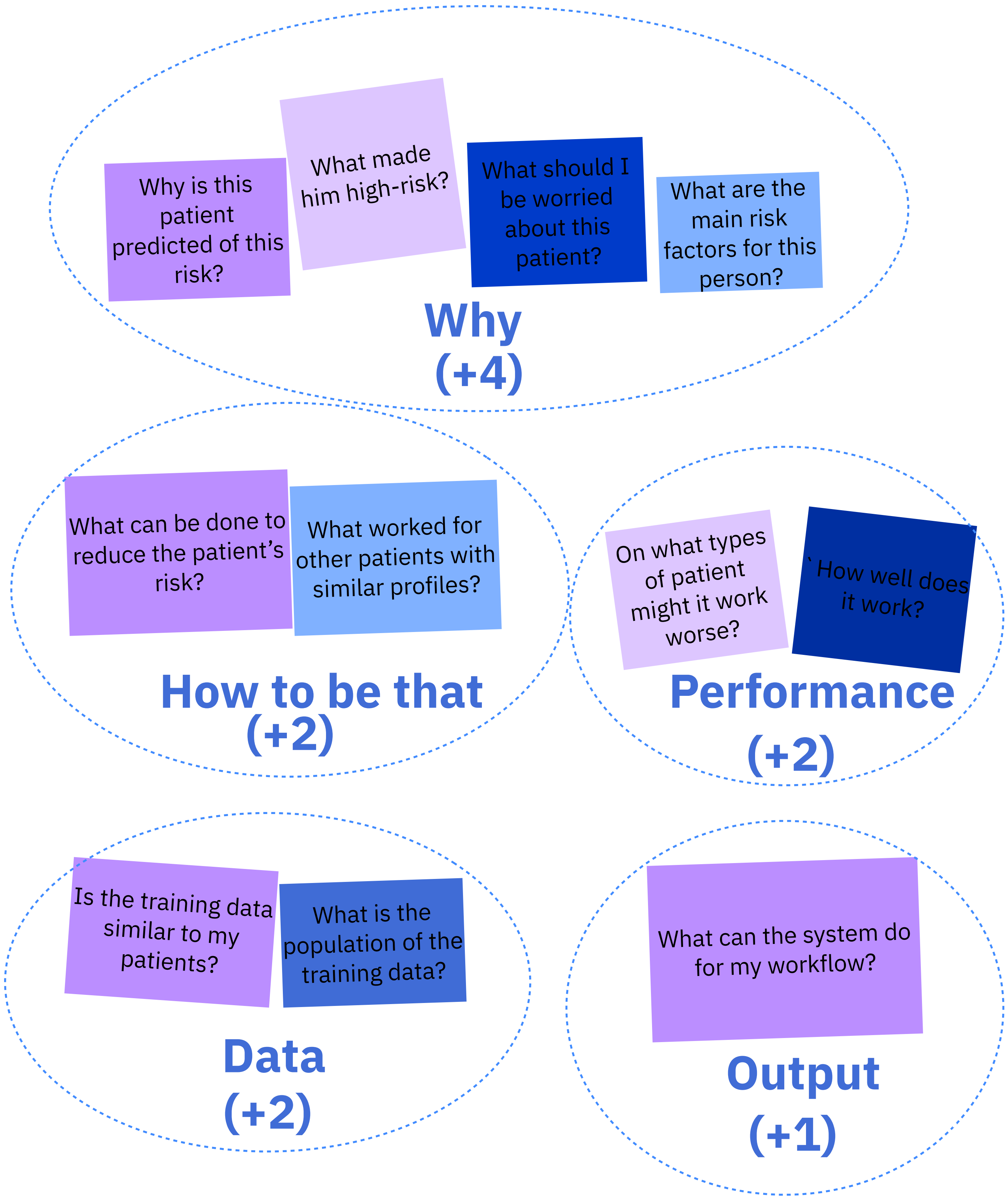
Liao et al. [Question-Driven Design Process for Explainable AI User Experiences](). (Working paper)

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

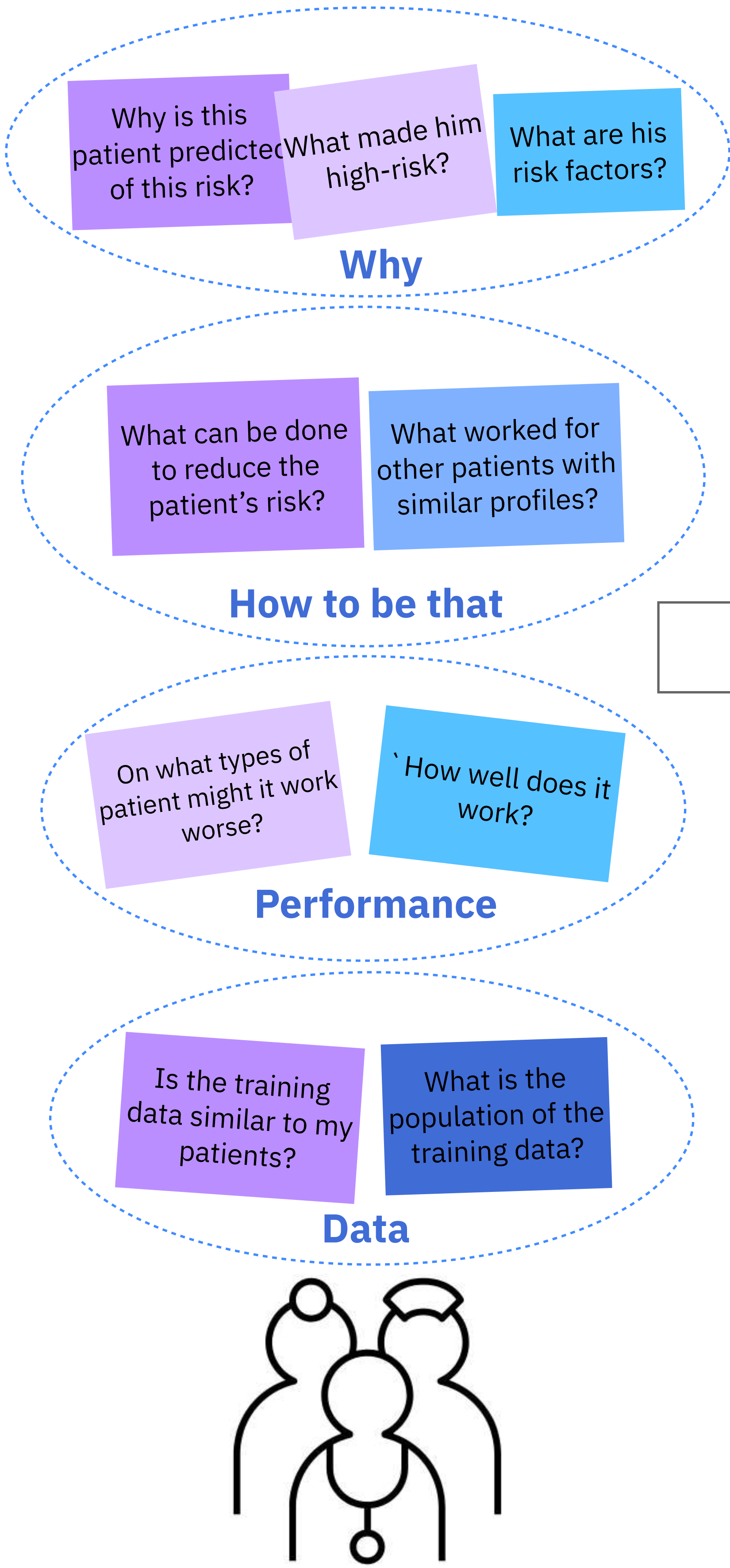
Iteratively design and evaluate

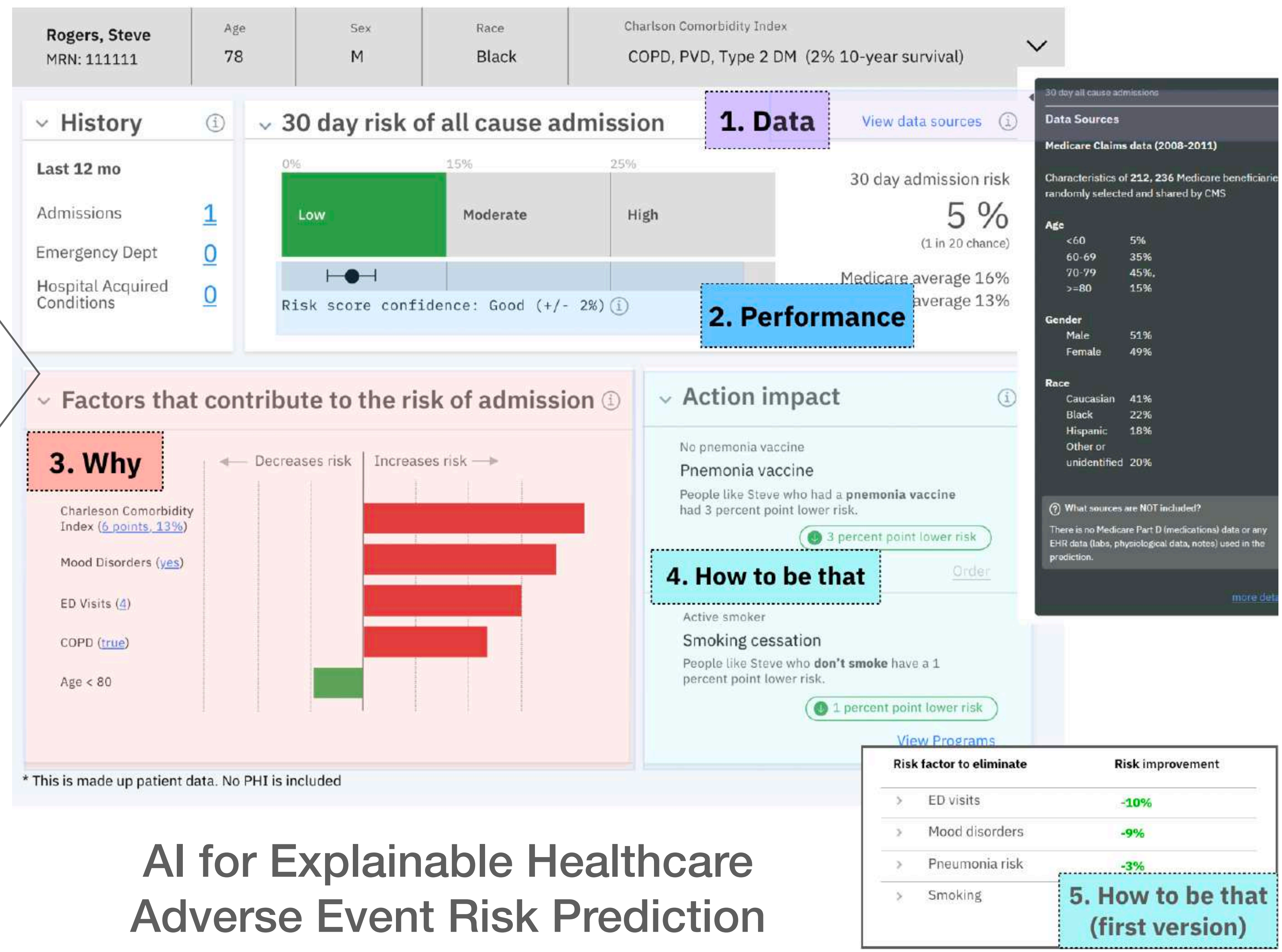
Create a design including the candidate elements identified in step 3

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)

Design Challenge 2: Gaps between XAI algorithmic output and human explanations

Human explanations are

- Contrastive
- Selective
- Interactive
- Tailored for recipients

Liao et al. [Questioning the AI: Informing Design Practices for Explainable AI User Experiences](). CHI 2020 🎖
Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*

Human explanations are

- Contrastive

You only bought fruits

Why didn't you buy me chocolate at Trader Joe's?

You went to Whole Foods

Inspecting counterfactual: contrastive feature

Customer： Ana

Assets score： **65**

No. Of satisfactory trades: **1**

Mo. since account open： **12**

No. of inquiries： **4**

Debt percentage： **50%**

Risk of failing to repay: high

- If **{debt percentage under 30%},** you will no longer be predicted of high risk

Bank customer

Why was my loan application rejected? How can I improve in the future?

Dhurandhar, et al. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. NeurIPS 2018 (**CEM**:https://github.com/Trusted-AI/AIX360/blob/master/aix360/algorithms/contrastive/CEM.py)

Human explanations are

- Contrastive
- Selective
- Interactive
- Tailored for recipients

“Translation” design: e.g. mimic how experts explain

Liao et al. [Questioning the AI: Informing Design Practices for Explainable AI User Experiences](). CHI 2020
Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*

Design Challenge 3: Limitations and Risks of XAI

Just to pick a few…

Explanation can lead to unwarranted trust in model

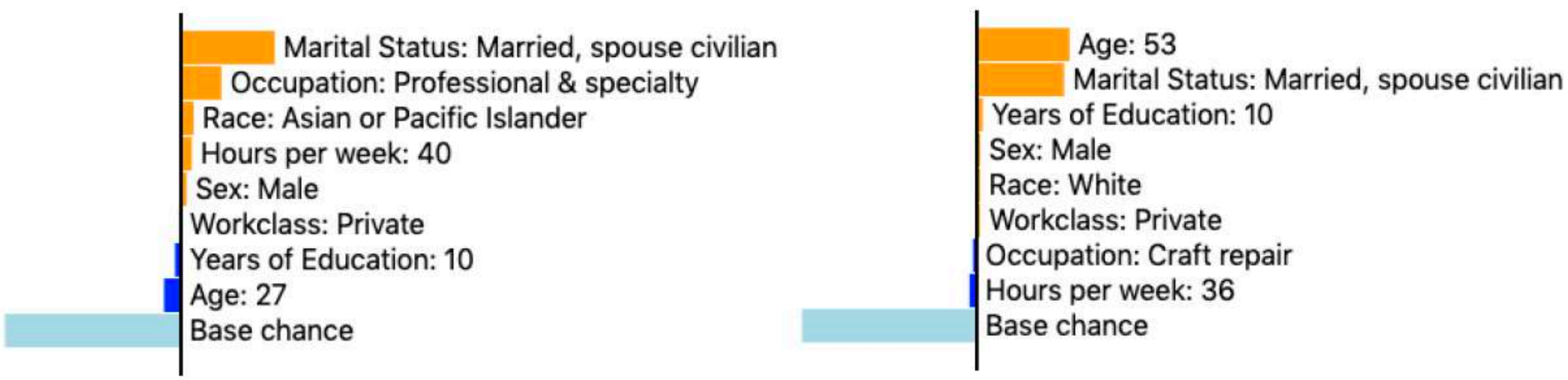


Figure 11: Screenshots of explanation for cases where the model had low confidence.

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*

“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

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)

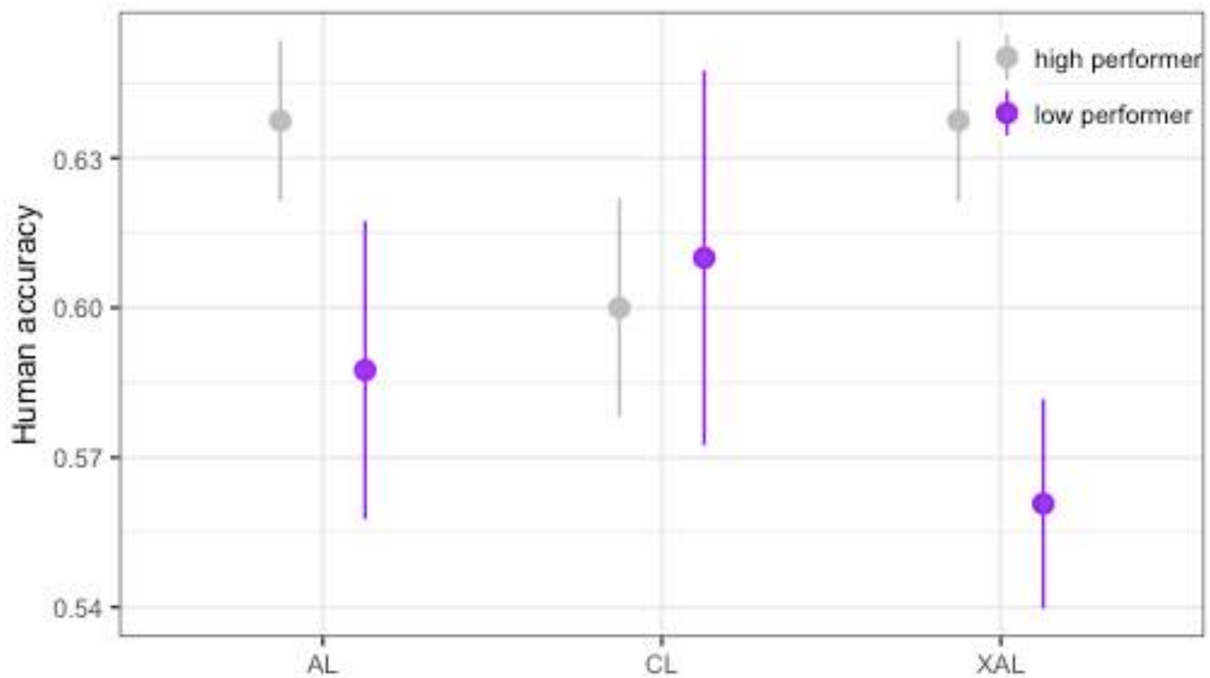
“meaningful” ???

(Nemitz, 2018)

Disparity of experience?

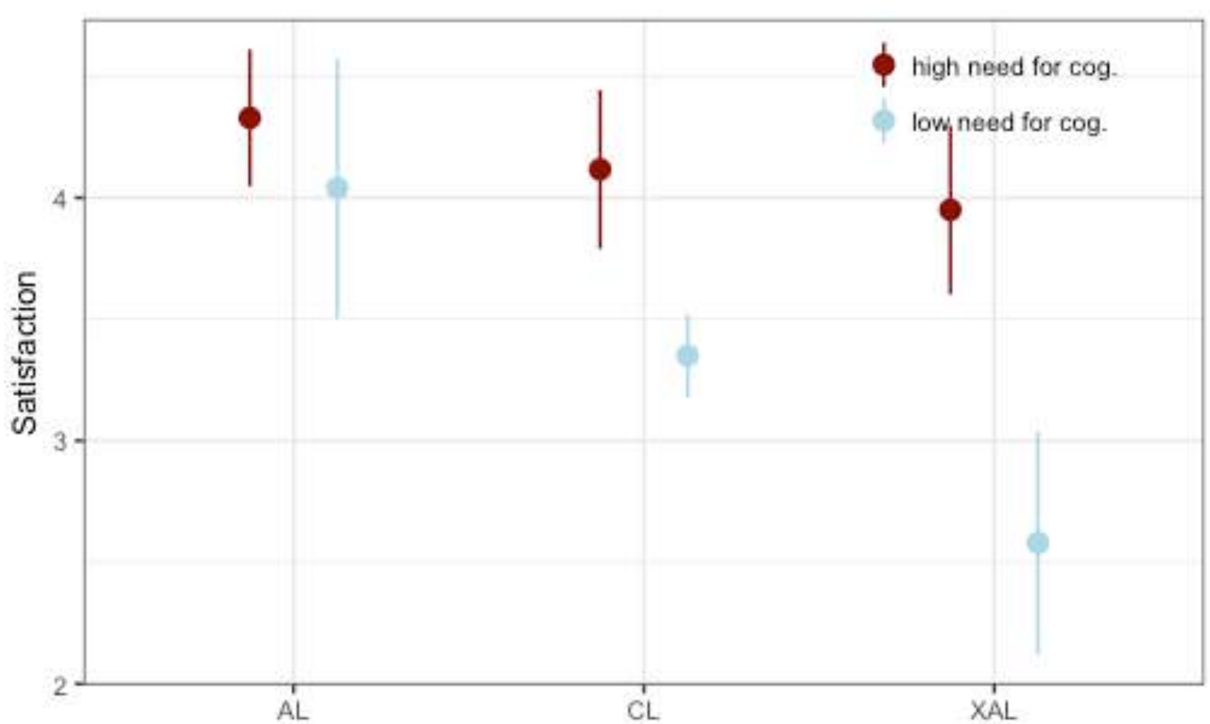
Ghai et al. [Explainable Active Learning (XAL): Toward AI Explanations as Interfaces for Machine Teachers](). CSCW 2021
Buçinca, at el. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. CSCW2021

Disparity of experience with XAI



Reduce human accuracy due to **unwarranted trust** in wrong predictions

But only for those **less familiar** with the domain



Reduce task satisfaction

But only for those with **low need for cognition** score

Ghai et al. [Explainable Active Learning (XAL): Toward AI Explanations as Interfaces for Machine Teachers](). CSCW 2021
Buçinca, at el. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. CSCW2021

“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

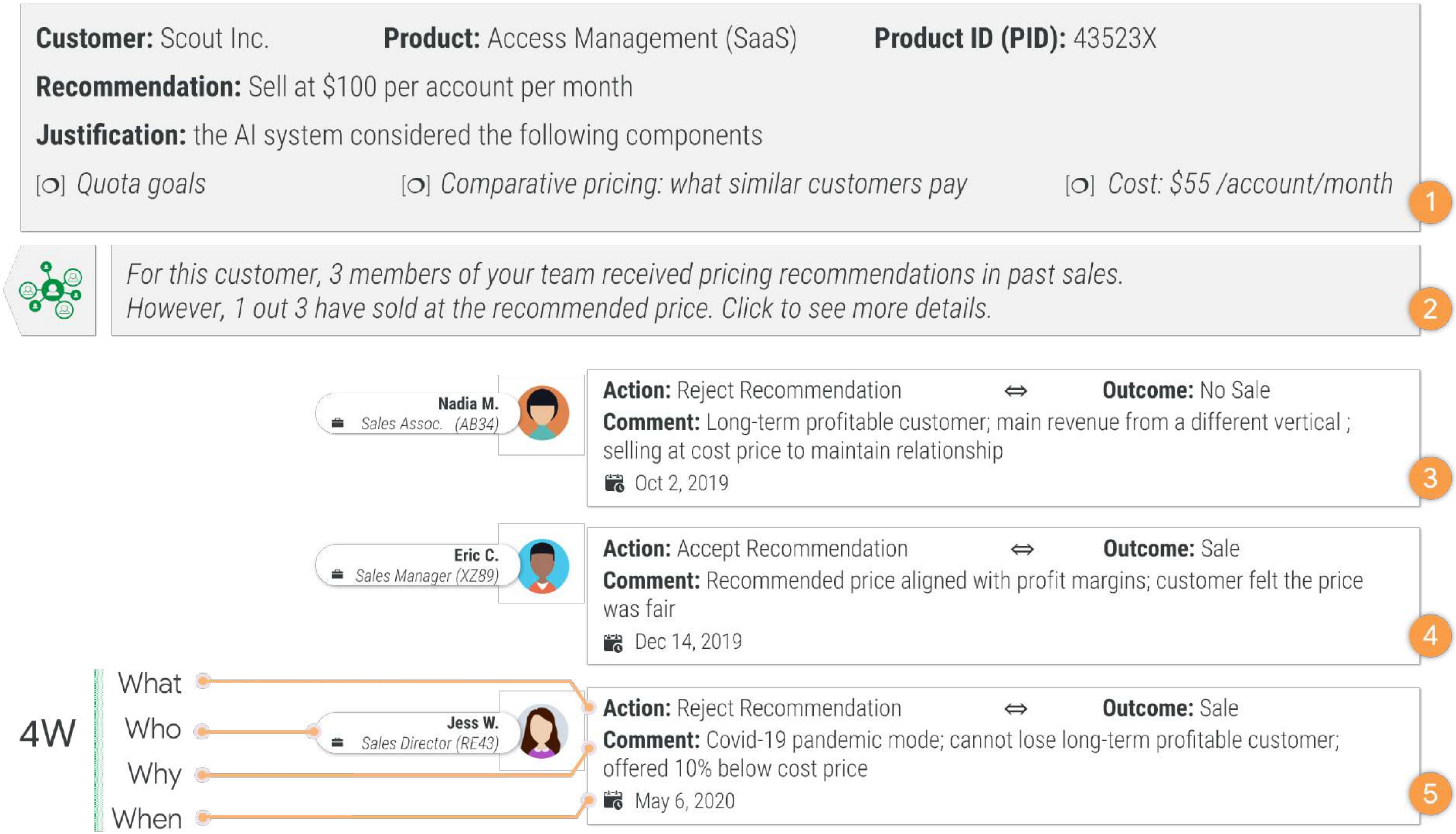
Information needs to achieve understanding of AI:

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

> *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***

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

Towards "social transparency" in AI systems



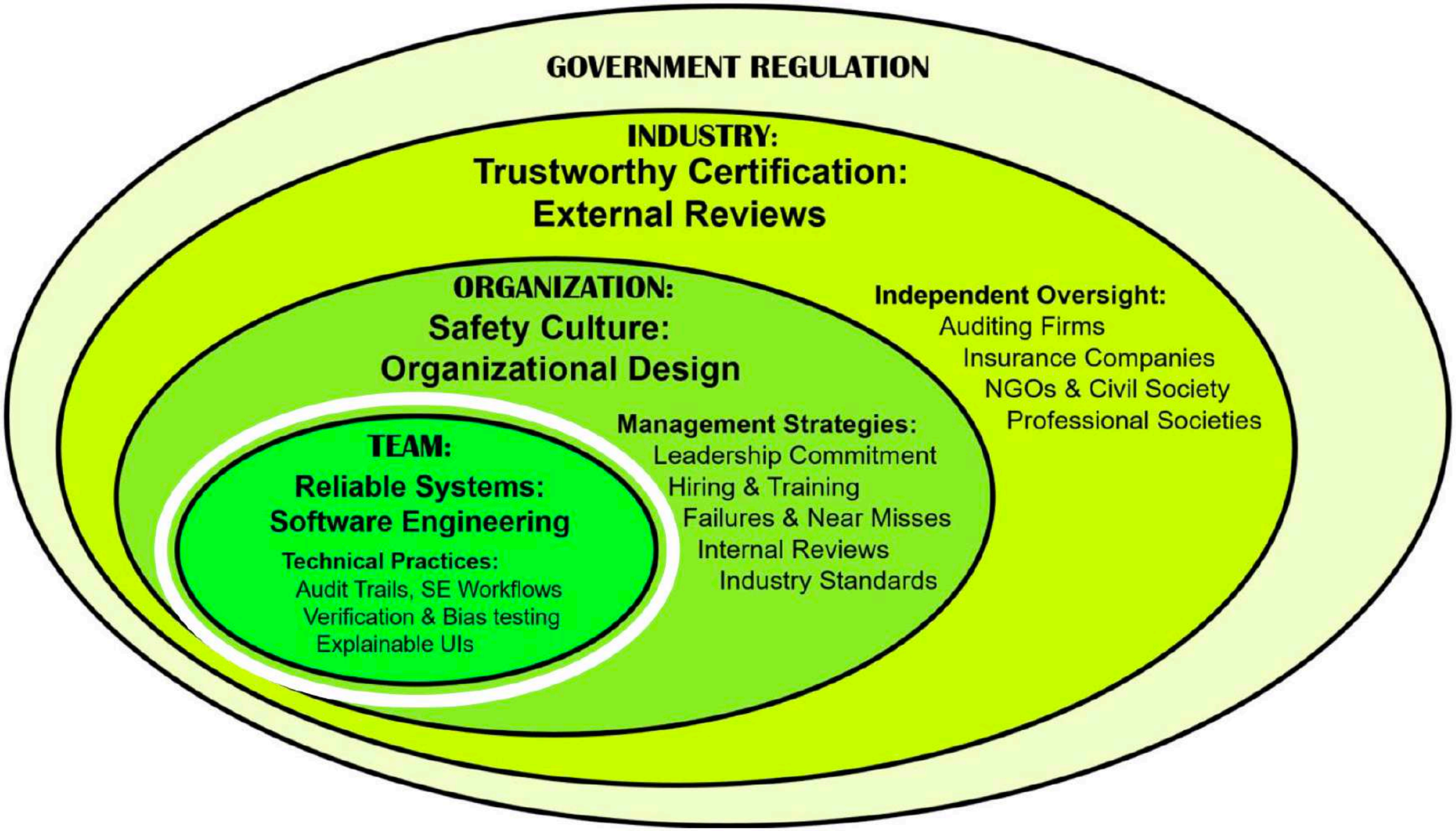
Ehsan et al. [Expanding Explainability: Towards Social Transparency in AI systems](). To appear in CHI 2021

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

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

Thank **YOU!**

Q. Vera Liao
www.qveraliao.com
@QVeraLiao