



AI Explainability 360

Vijay Arya, Amit Dhurandhar, Dennis Wei

IBM Research AI

Jan 27th, 2020



CONTRIBUTORS

This toolkit is the joint effort of many people:

Vijay Arya

Rachel K. E. Bellamy

Pin-Yu Chen

Amit Dhurandhar

Michael Hind

Samuel C. Hoffman

Stephanie Houde

Q. Vera Liao

Ronny Luss

Aleksandra Mojsilović

Sami Mourad

Pablo Pedemonte

Ramya Raghavendra

John Richards

Prasanna Sattigeri

Karthikeyan Shanmugam

Moninder Singh

Kush R. Varshney

Yunfeng Zhang



AGENDA



- **Why Explainable AI?**
 - Types and Methods for Explainable AI
- AI Explainability 360 Toolkit
 - Taxonomy and Guidance
- Interactive Web Experience Demo
- Hands on session 1
 - Package Installation and Git walkthrough
 - Use case (Industry): Personal finance
- Hands on session 2
 - Use case (Government): Health and nutrition
- Hands on session 3
 - Use case (Medicine): Clinical Medicine
 - Metrics
- Summary and future directions

30

15

45

Break 30

25

30

30



AI IS NOW USED IN MANY HIGH-STAKES DECISION MAKING APPLICATIONS



Credit



Employment



Admission



Sentencing

WHAT DOES IT TAKE TO TRUST A DECISION MADE BY A MACHINE (OTHER THAN THAT IT IS 99% ACCURATE)



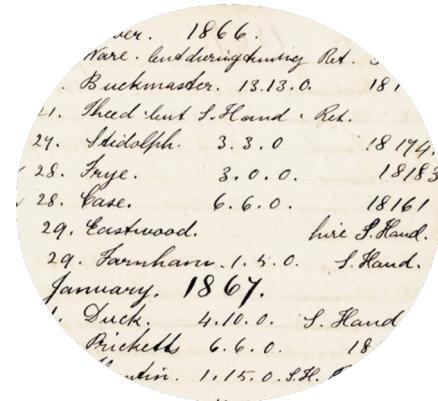
Is it fair?



Is it easy to understand?



Did anyone tamper with it?



Is it accountable?



THE QUEST FOR "EXPLAINABLE AI"

CIO JOURNAL

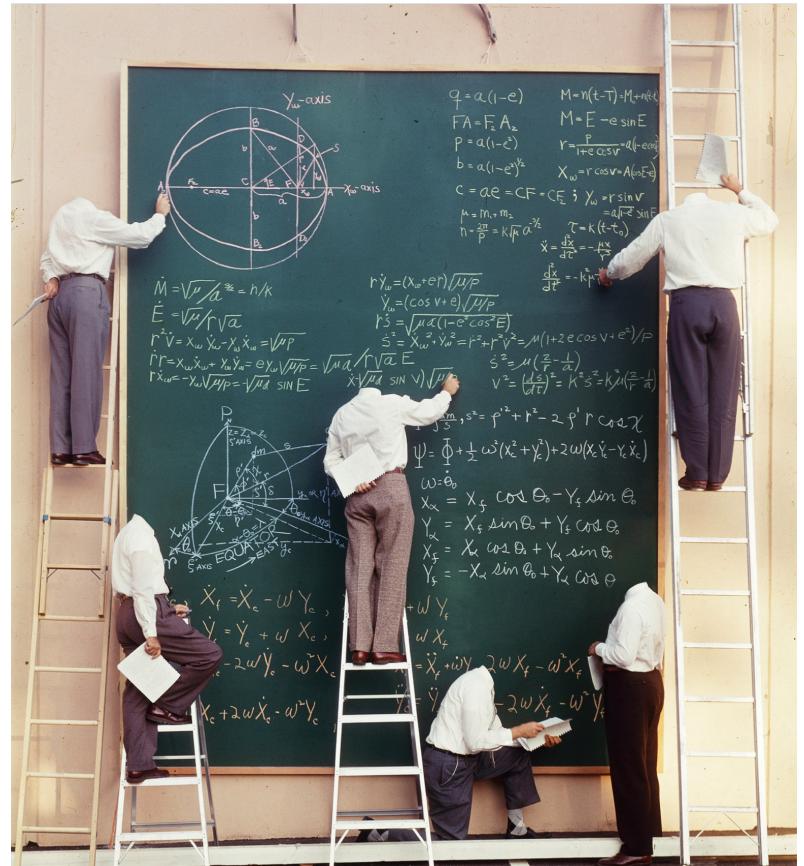
Companies Grapple With AI's Opaque Decision-Making Process
THE WALL STREET JOURNAL

Why Explainable AI Will Be the Next Big Disruptive Trend in Business 

When a Computer Program Keeps You in Jail

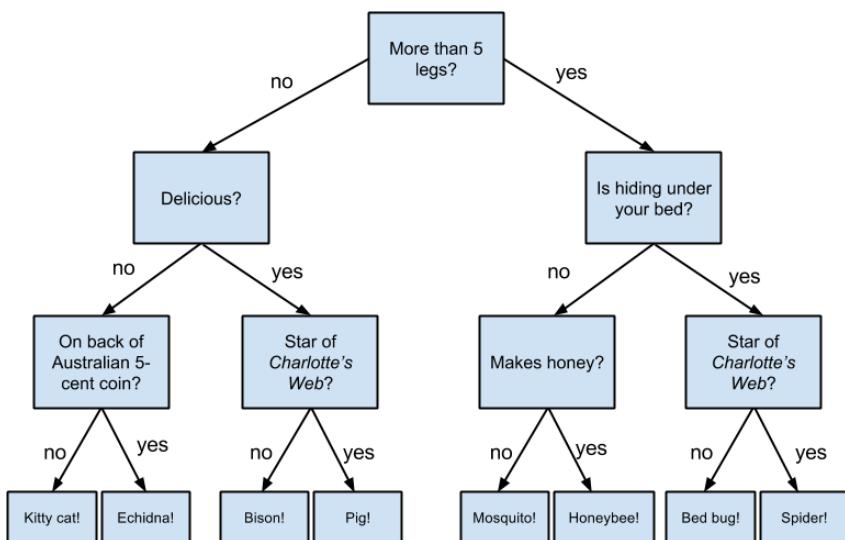
Don't Trust Artificial Intelligence? Time To Open The AI 'Black Box'

6 x 2019 IBM Corporation



WHY EXPLAINABLE AI?

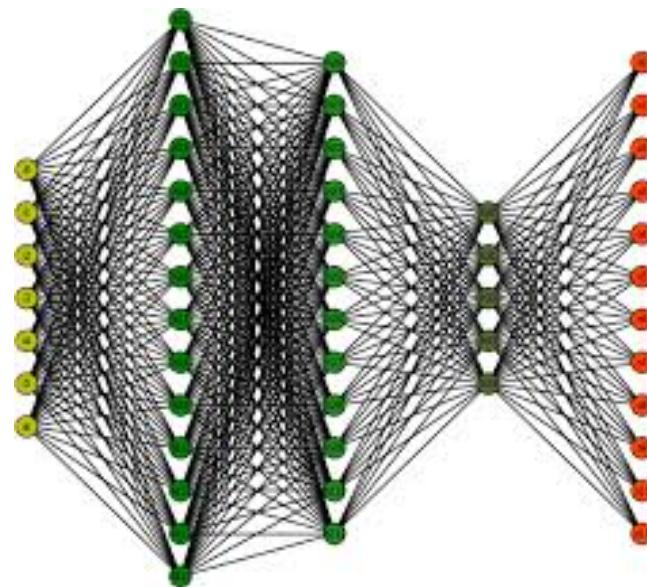
Decision Tree



Interpretable?

YES

Neural Network



Interpretable?

NO



BUT WHAT ARE WE ASKING FOR?

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)

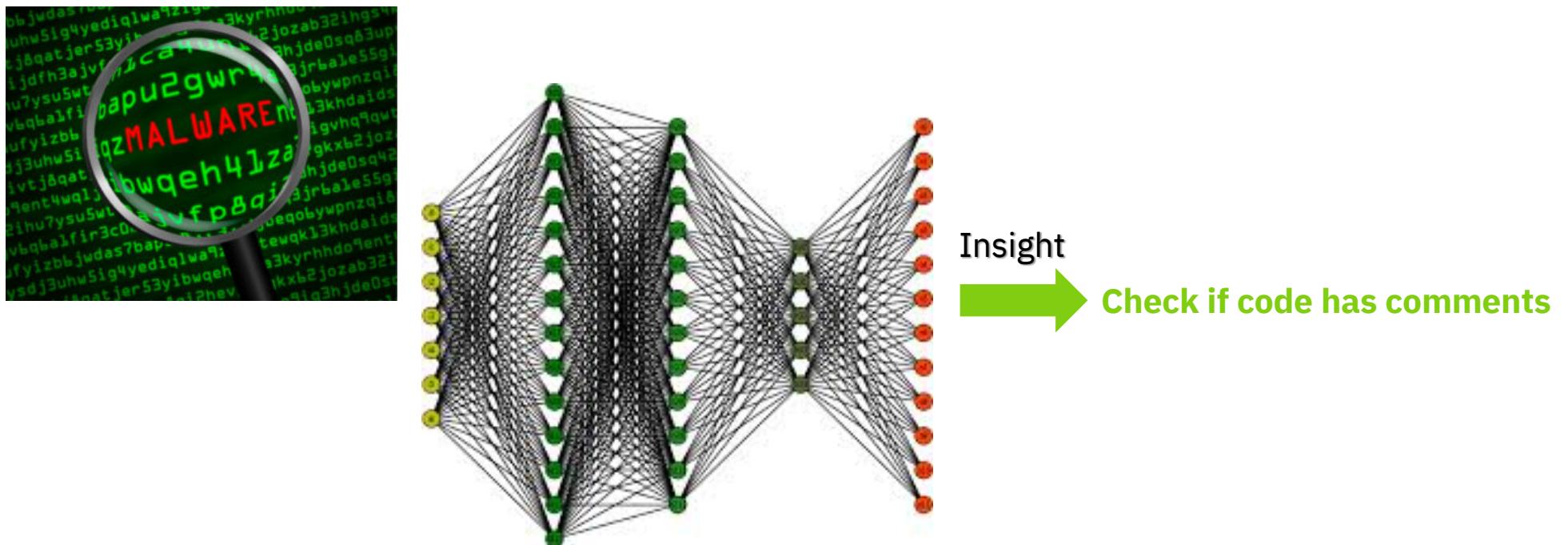
“**meaningful**” ???



WHY EXPLAINABLE AI?

Simplification

Understanding what's truly happening can help build simpler systems.



WHY EXPLAINABLE AI? (CONTINUED)

Debugging

Can help to understand what is wrong with a system.



Self driving car slowed down but
wouldn't stop at red light???



WHY EXPLAINABLE AI? (CONTINUED)

Existence of Confounders

Can help to identify spurious correlations.

Pneumonia



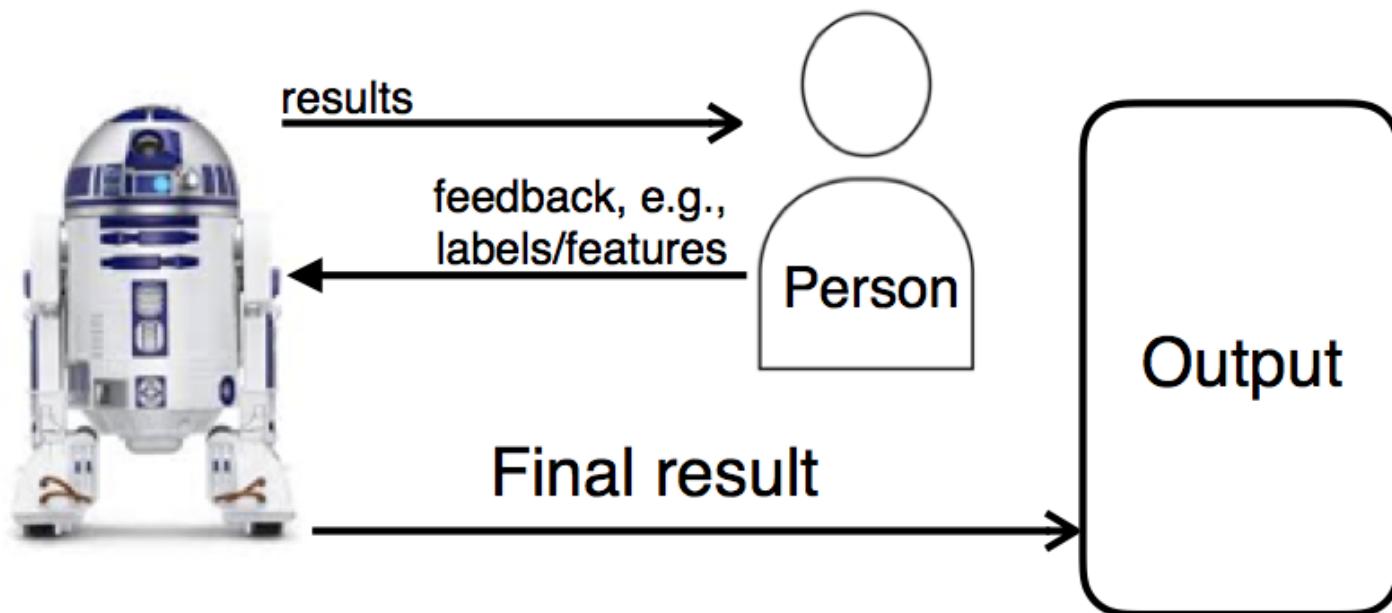
Diabetes



WHY EXPLAINABLE AI? (CONTINUED)

Enhance Performance

Humans in combination with a system can be much more effective than just a more accurate system.



WHY EXPLAINABLE AI? (CONTINUED)

Fairness

Is the decision making system fair?



Robustness and Generalizability

Is the system basing decisions on the correct features?



Wide Spread Adoption

Interesting article

Geoff Hinton Dismissed The Need For Explainable AI: 8 Experts Explain Why He's Wrong

Hinton: "I'm an expert on trying to get the technology to work, not an expert on social policy. One place where I do have technical expertise that's relevant is [whether] regulators should insist that you can explain how your AI system works. I think that would be a complete disaster."

[Geoff Hinton Dismissed - The Need For Explainable AI: 8 Experts Explain Why He's Wrong](#)



THREE DIMENSIONS OF EXPLAINABILITY

One explanation does not fit all: There are many ways to explain things.

directly interpretable

The oldest AI formats, such as decision rule sets, decision trees, and decision tables are simple enough for people to understand. Supervised learning of these models is directly interpretable.

global (model-level)

Shows the entire predictive model to the user to help them understand it (e.g. a small decision tree, whether obtained directly or in a post hoc manner).

static

The interpretation is simply presented to the user.

vs.

post hoc interpretation

Start with a black box model and probe into it with a companion model to create interpretations. The black box model continues to provide the actual prediction while the interpretation improves human interactions.

vs.

local (instance-level)

Only show the explanations associated with individual predictions (i.e. what was it about this particular person that resulted in her loan being denied).

vs.

interactive (visual analytics)

The user can interact with interpretation.

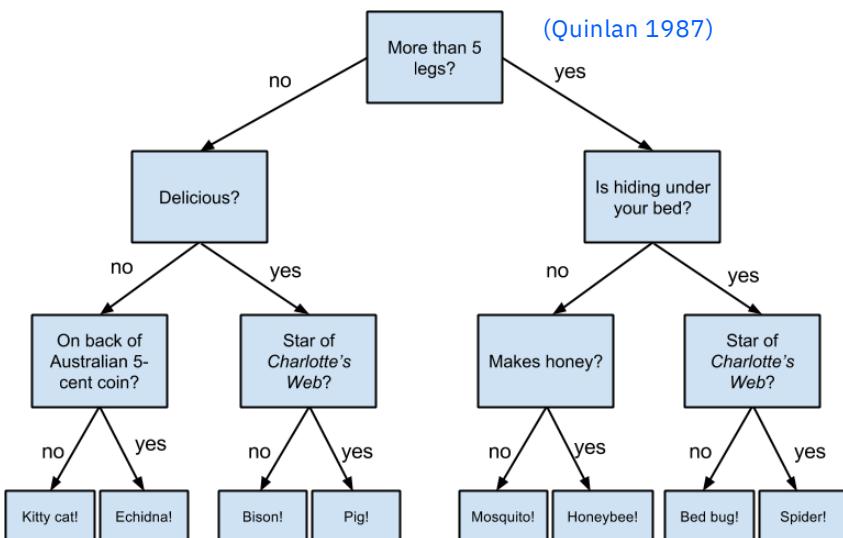


EXPLANATION METHOD TYPES

Directly interpretable

The oldest AI formats, such as decision rule sets, decision trees, and decision tables are simple enough for people to understand. Supervised learning of these models is directly interpretable.

Decision Tree



Rule List

(Wang and Rudin 2016)

```

if capital-gain>$7298.00
else if Young,Never-married,
else if Grad-school,Married,
else if Young,capital-loss=0,
else if Own-child,Never-married,
else if Bachelors,Married,
else if Bachelors,Over-time,
else if Exec-managerial,Married,
else if Married,HS-grad,
else if Grad-school,
else if Some-college,Married,
else if Prof-specialty,Married,
else if Assoc-degree,Married,
else if Part-time,
else if Husband,
else if Prof-specialty,
else if Exec-managerial,Male,
else if Full-time,Private,
else (default rule)
then probability to make over 50K = 0.986
then probability to make over 50K = 0.003
then probability to make over 50K = 0.748
then probability to make over 50K = 0.072
then probability to make over 50K = 0.015
then probability to make over 50K = 0.655
then probability to make over 50K = 0.255
then probability to make over 50K = 0.531
then probability to make over 50K = 0.300
then probability to make over 50K = 0.266
then probability to make over 50K = 0.410
then probability to make over 50K = 0.713
then probability to make over 50K = 0.420
then probability to make over 50K = 0.013
then probability to make over 50K = 0.126
then probability to make over 50K = 0.148
then probability to make over 50K = 0.193
then probability to make over 50K = 0.026
then probability to make over 50K = 0.066.
  
```

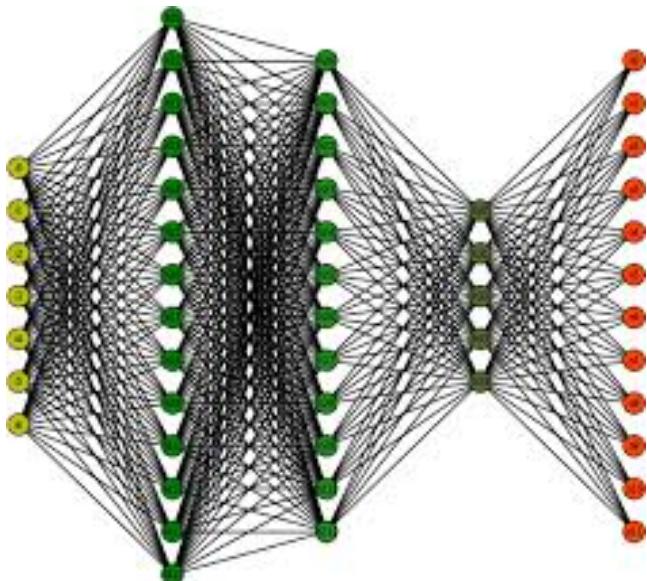


EXPLANATION METHOD TYPES (CONTINUED)

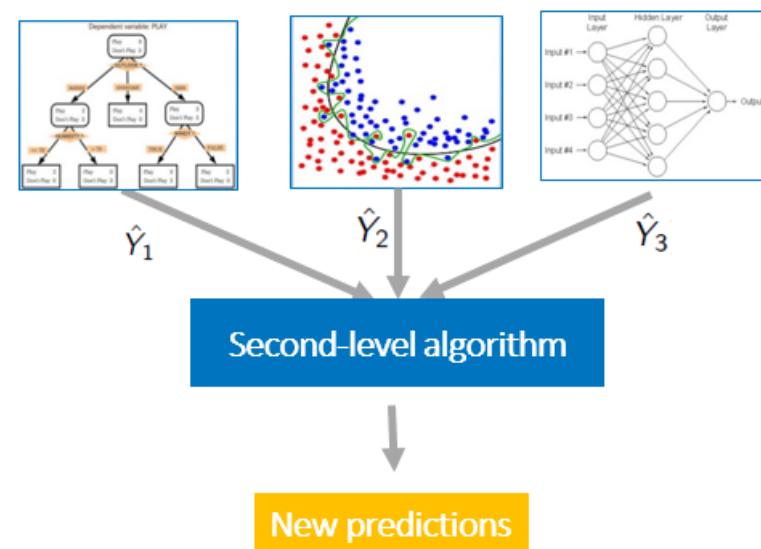
Post hoc interpretation

Start with a black box model and probe into it with a companion model to create interpretations. The black box model continues to provide the actual prediction while interpretation improve human interactions.

(Deep) Neural Network



Ensembles



EXPLANATION METHOD TYPES (CONTINUED)

Post hoc (local) interpretation

Locally Interpretable Model Agnostic Explanations (LIME)

(Ribeiro et. al. 2016)

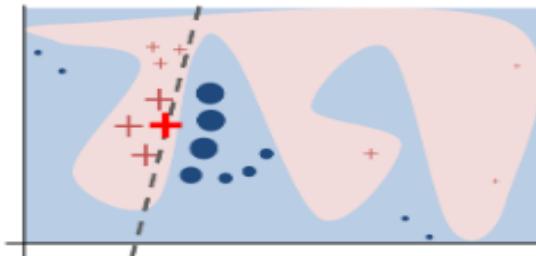


Figure 1. Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background. The bright bold red cross is the instance being explained. LIME samples instances, gets predictions using f , and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the explanation that is locally (but not globally) faithful.



Algorithm 1 Sparse Linear Explanations using LIME

```
Require: Classifier  $f$ , Number of samples  $N$ 
Require: Instance  $x$ , and its interpretable version  $x'$ 
Require: Similarity kernel  $\pi_x$ , Length of explanation  $K$ 
 $\mathcal{Z} \leftarrow \{\}$ 
for  $i \in \{1, 2, 3, \dots, N\}$  do
     $z'_i \leftarrow \text{sample\_around}(x')$ 
     $\mathcal{Z} \leftarrow \mathcal{Z} \cup (z'_i, f(z_i), \pi_x(z_i))$ 
end for
 $w \leftarrow \text{K-Lasso}(\mathcal{Z}, K)$   $\triangleright$  with  $z'_i$  as features,  $f(z_i)$  as target
return  $w$ 
```



EXPLANATION METHOD TYPES (CONTINUED)

Post hoc (local) interpretation

Maximum Mean Discrepancy Critic

(Kim et. al. 2016)

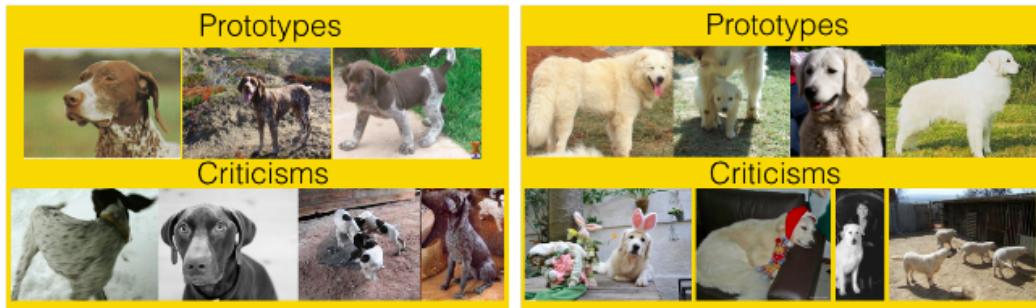


Figure 2: Learned prototypes and criticisms from Imagenet dataset (two types of dog breeds)

Health care



Prototypes

$$f(x) = \frac{1}{n} \sum_{i \in [n]} k(x, x_i) - \frac{1}{m} \sum_{j \in [m]} k(x, z_j).$$

Criticisms

$$\begin{aligned} J_b(S) &= \frac{1}{n^2} \sum_{i,j=1}^n k(x_i, x_j) - \text{MMD}^2(\mathcal{F}, X, X_S) \\ &= \frac{2}{n|S|} \sum_{i \in [n], j \in S} k(x_i, y_j) - \frac{1}{|S|^2} \sum_{i,j \in S} k(y_i, x_j). \end{aligned}$$



EXPLANATION METHOD TYPES (CONTINUED)

Post hoc (local) interpretation

Saliency Maps

(Simonyan et. al. 2013)



$$w = \frac{\partial S_c}{\partial I} \Big|_{I_0} .$$

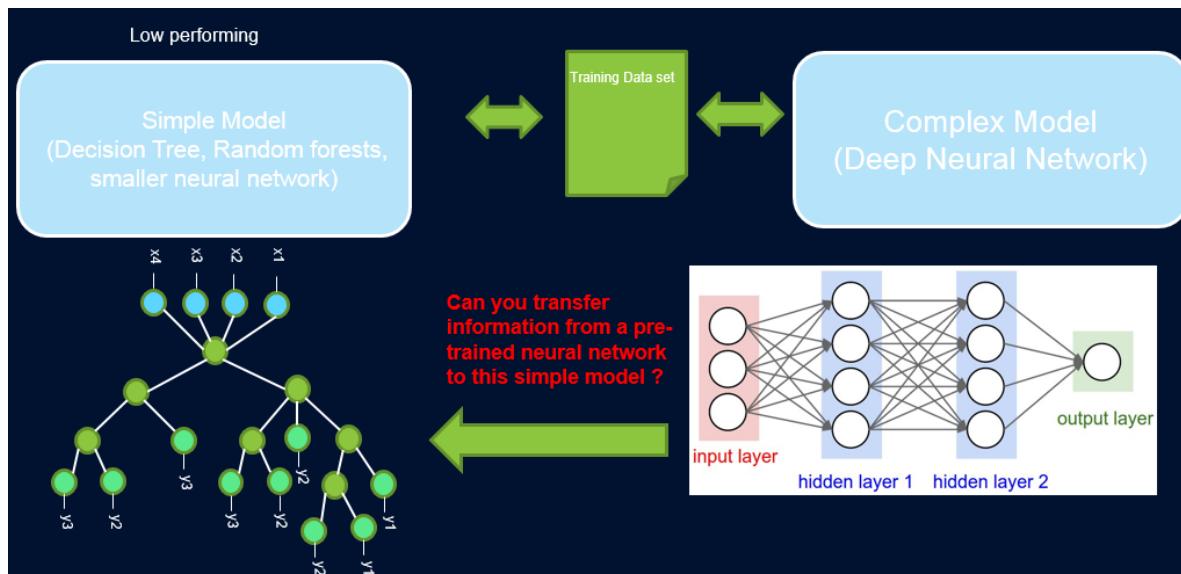


EXPLANATION METHOD TYPES (CONTINUED)

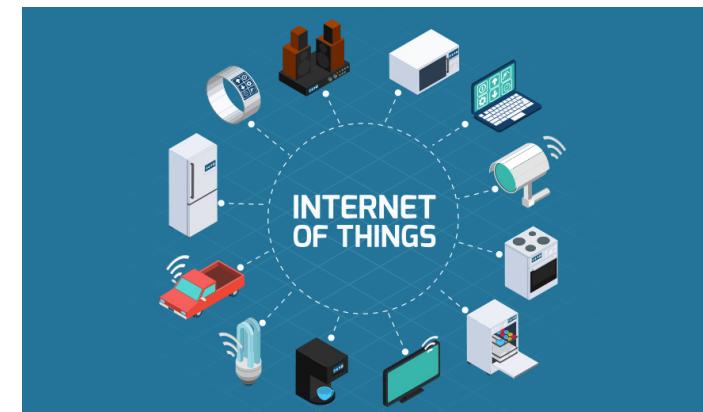
Post hoc (global) interpretation

Knowledge Distillation

(Hinton et. al. 2015)



Complex Systems



$$\frac{\partial C}{\partial z_i} = \frac{1}{T} (q_i - p_i) = \frac{1}{T} \left(\frac{e^{z_i/T}}{\sum_j e^{z_j/T}} - \frac{e^{v_i/T}}{\sum_j e^{v_j/T}} \right)$$



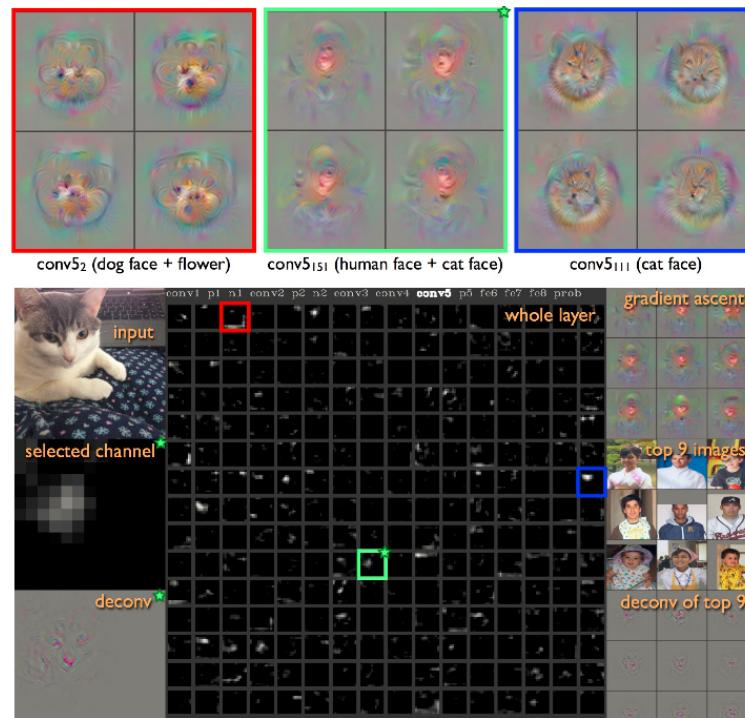
EXPLANATION METHOD TYPES (CONTINUED)

Static/Interactive (visual) interpretation

Start with a black box model and probe into it with a companion model to create interpretations. The black box model continues to provide the actual prediction while the interpretation improves human interactions.

Deep Visualization

(Yosinski et. al. 2015)



ONE EXPLANATION DOES NOT FIT ALL

Different stakeholders require explanations for different purposes and with different objectives. Explanations will have to be tailored to their needs.

End users

“Why did you recommend this treatment?”

Who: Physicians, judges, loan officers, teacher evaluators

Why: trust/confidence, insights(?)

Affected users

“Why was my loan denied? How can I be approved?”

Who: Patients, accused, loan applicants, teachers

Why: understanding of factors

Regulatory bodies

“Prove that your system didn't discriminate.”

Who: EU (GDPR), NYC Council, US Gov't, etc.

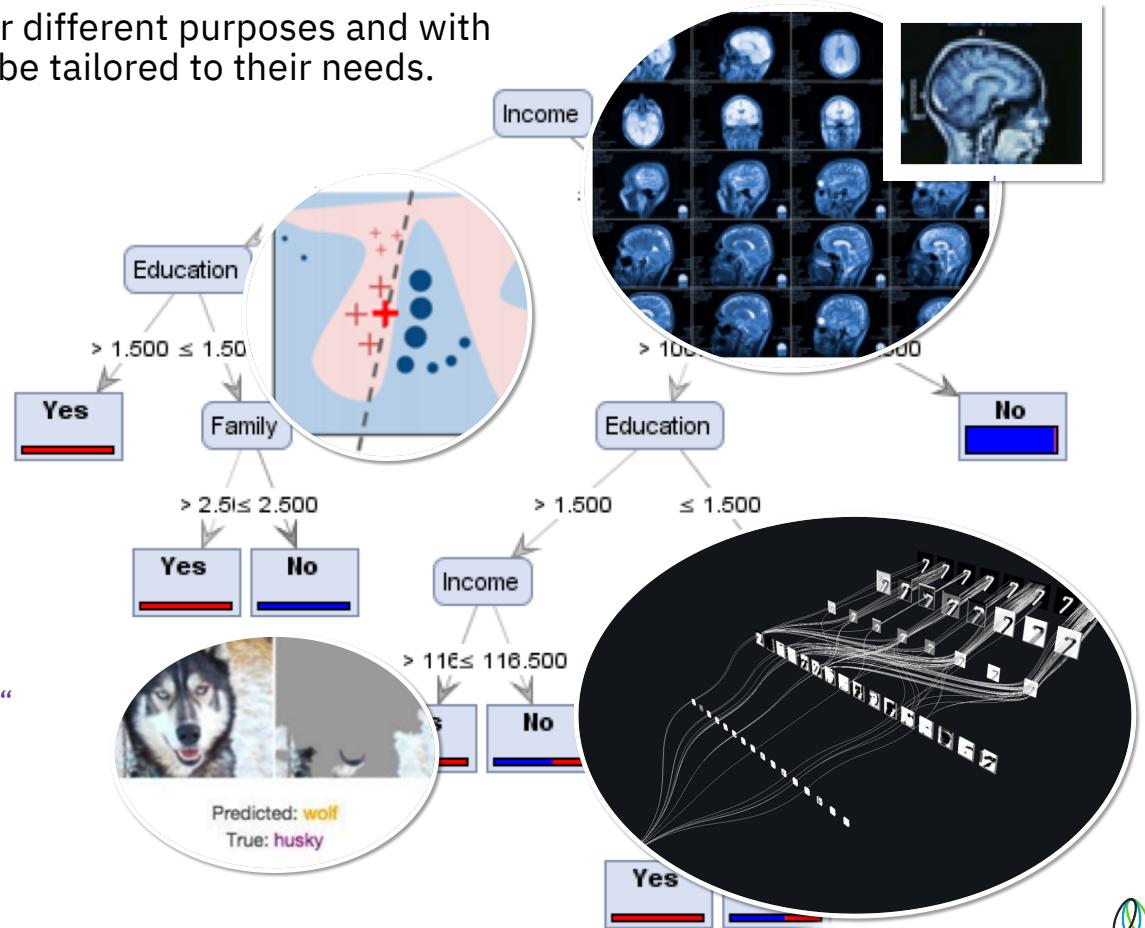
Why: ensure fairness for constituents

AI system builders/stakeholders

“Is the system performing well? How can it be improved?”

Who: EU (GDPR), NYC Council, US Gov't, etc.

Why: ensure or improve performance



AGENDA



- Why Explainable AI?
 - Types and Methods for Explainable AI
- **AI Explainability 360 Toolkit**
 - Taxonomy and Guidance
- Interactive Web Experience Demo
- Hands on session 1
 - Package Installation and Git walkthrough
 - Use case (Industry): Personal finance
- Hands on session 2
 - Use case (Government): Health and nutrition
- Hands on session 3
 - Use case (Medicine): Clinical Medicine
 - Metrics
- Summary and future directions

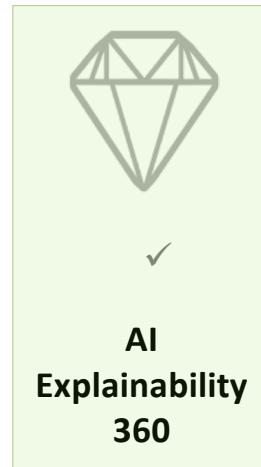
AIX360: IBM RESEARCH AI EXPLAINABILITY 360 TOOLKIT

Goals

- Support a community of users and contributors who will together help make models and their predictions more transparent.
- Support and advance research efforts in explainability.
- Contribute efforts to engender trust in AI.

IBM Research AIX360	
Explainability Algorithms	10 algorithms to explain data and AI models + 2 metrics
Repositories	github.ibm.com/AIX360 github.com/IBM/AIX360
Interactive Experience	aix360.mybluemix.net
API	aix360.readthedocs.io
Tutorials	13 notebooks (finance, healthcare, lifestyle, Attrition, etc.)
Developers	> 15 Researchers + Software engineers across YKT, India, Argentina

Trusted AI Toolkits



Adversarial
Robustness
360

AI
Fairness
360

AI
Explainability
360

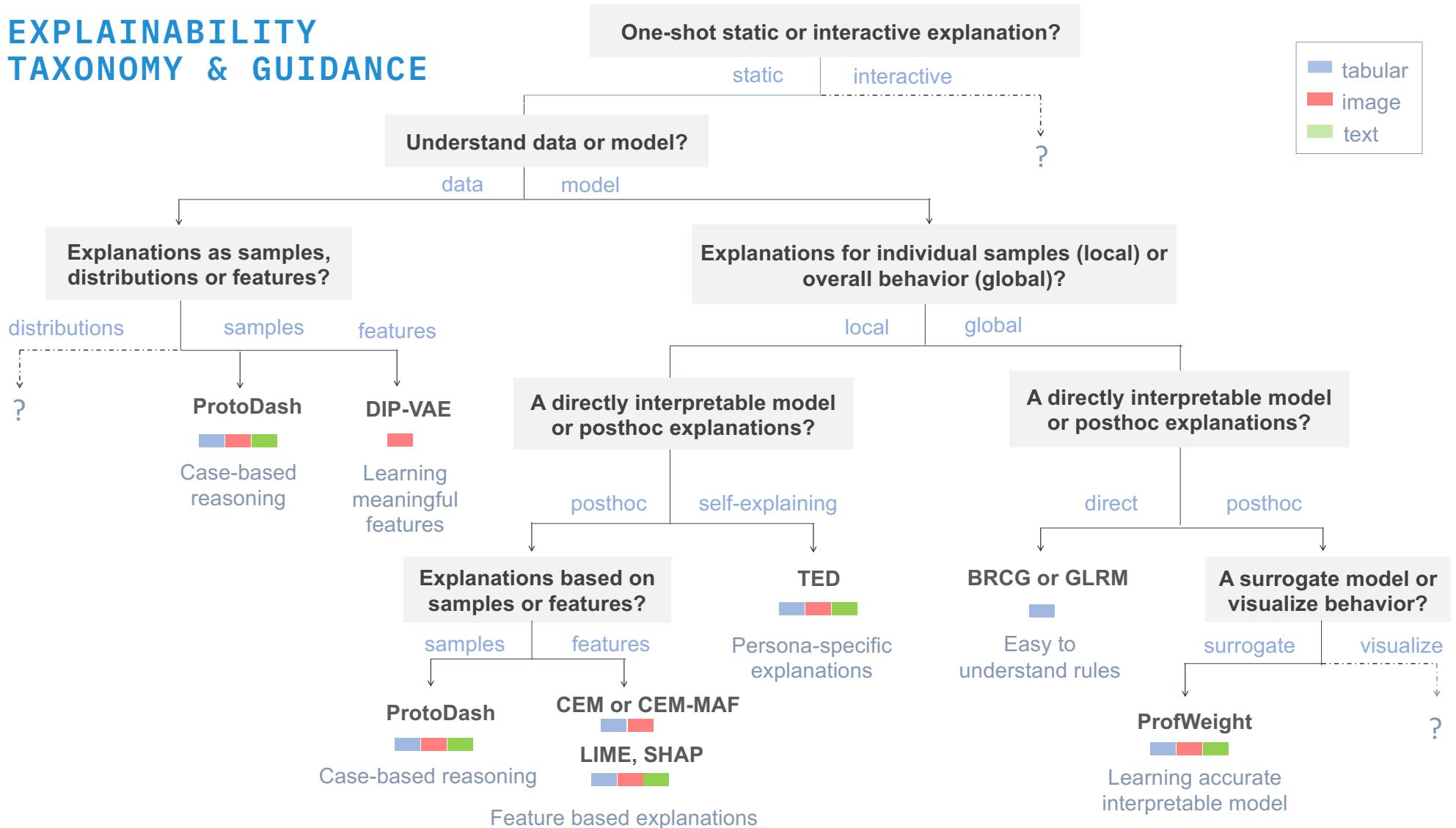
Causal
Inference
360

Why Explainable AI Will Be the Next Big Disruptive Trend in Business 

Don't Trust Artificial
Intelligence? Time To Open The
AI 'Black Box'

CIO JOURNAL
Companies Grapple With AI's Opaque Decision-Making Process
THE WALL STREET JOURNAL

EXPLAINABILITY TAXONOMY & GUIDANCE



AIX360: AI EXPLAINABILITY OPENSOURCE LANDSCAPE

Toolkit	Data Explanations	Directly Interpretable	Local Post-hoc	Global Post-hoc	Custom Explanation	Metrics
IBM AIX360	2	2	5	1	1	2
Seldon Alibi			✓	✓		
Oracle Skater		✓	✓	✓		
H2o		✓	✓	✓		
Microsoft Interpret		✓	✓	✓		
Ethical ML				✓		
DrWhyDalEx				✓		

All algorithms of AIX360 are developed by IBM Research

AIX360 also provides demos, tutorials, and guidance on explanations for different use cases.

Paper: One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques:

<https://arxiv.org/abs/1909.03012v1>



AGENDA



- Why Explainable AI?
 - Types and Methods for Explainable AI
- AI Explainability 360 Toolkit
 - Taxonomy and Guidance
- **Interactive Web Experience Demo**
- Hands on session 1
 - Package Installation and Git walkthrough
 - Use case (Industry): Personal finance
- Hands on session 2
 - Use case (Government): Health and nutrition
- Hands on session 3
 - Use case (Medicine): Clinical Medicine
 - Metrics
- Summary and future directions

The screenshot shows the homepage of the AI Explainability 360 Open Source Toolkit. At the top, there's a navigation bar with links for Home, Demo, Resources, Events, Videos, and Community. Below the navigation, a banner for "IBM Research Trusted AI" features the toolkit's name and a brief description: "This extensible open source toolkit can help you comprehend how machine learning models predict labels by various means throughout the AI application lifecycle. Containing eight state-of-the-art algorithms for interpretable machine learning as well as metrics for explainability, it is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, human capital management, healthcare, and education. We invite you to use it and improve it." Two buttons, "API Docs" and "Get Code", are located below the banner. A main heading "Not sure what to do first? Start here!" is followed by a grid of seven cards, each representing a different toolkit component:

- Boolean Decision Rules via Column Generation (Light Edition)**: Directly learn accurate and interpretable 'or'-of-'and' logical classification rules.
- Generalized Linear Rule Models**: Directly learn accurate and interpretable weighted combinations of 'and' rules for classification or regression.
- ProfWeight**: Improve the accuracy of a directly interpretable model such as a decision tree using the confidence profile of a neural network.
- Teaching AI to Explain its Decisions**: Predict both labels and explanations with a model whose training set contains features, labels, and explanations.
- Contrastive Explanations Method**: Generate justifications for neural network classifications by highlighting minimally sufficient features, and minimally and critically absent features.
- Contrastive Explanations Method with Monotonic Attribute Functions**: Contrastive explanations for colored images or images with rich structure.
- Disentangled Inferred Prior VAE**: Learn disentangled representations for interpreting unlabeled data.
- ProtoDash**: Select prototypical examples from a dataset.

Each card has a small "→" icon at the bottom right corner.

<http://aix360.mybluemix.net/>

AGENDA



- Why Explainable AI?
 - Types and Methods for Explainable AI
- AI Explainability 360 Toolkit
 - Taxonomy and Guidance
- Interactive Web Experience Demo
- **Hands on session 1**
 - Package Installation and Git walkthrough
 - Use case (Industry): Personal finance
- Hands on session 2
 - Use case (Government): Health and nutrition
- Hands on session 3
 - Use case (Medicine): Clinical Medicine
 - Metrics
- Summary and future directions

The image shows two overlapping web pages. The left page is a GitHub repository for 'IBM / AIX360'. It features a banner for 'Join GitHub today' and a summary of repository statistics: 197 commits, 7 branches, 0 packages, 0 releases, and 11 contributors. The right page is a Jupyter nbviewer notebook titled 'Credit Approval Tutorial'. The notebook content discusses the use of various methods in the AI Explainability 360 Toolkit for credit approval models, mentioning the FICO Explainable Machine Learning Challenge, Boolean Rule Column Generation, Logistic Rule Regression, and CEM (Contrastive Explanations Method). It also covers the FICO HELOC Dataset and its use for explaining predictions based on HELOC Dataset. Both pages have a dark theme.

<http://github.com/IBM/AIX360>

<https://github.com/IBM/AIX360/tree/master/examples>

AGENDA



- Why Explainable AI?
 - Types and Methods for Explainable AI
- AI Explainability 360 Toolkit
 - Taxonomy and Guidance
- Interactive Web Experience Demo
- Hands on session 1
 - Package Installation and Git walkthrough
 - Use case (Industry): Personal finance
- **Hands on session 2**
 - Use case (Government): Health and nutrition
- Hands on session 3
 - Use case (Medicine): Clinical Medicine
 - Metrics
- Summary and future directions

The screenshot shows a Jupyter Notebook interface with the following details:

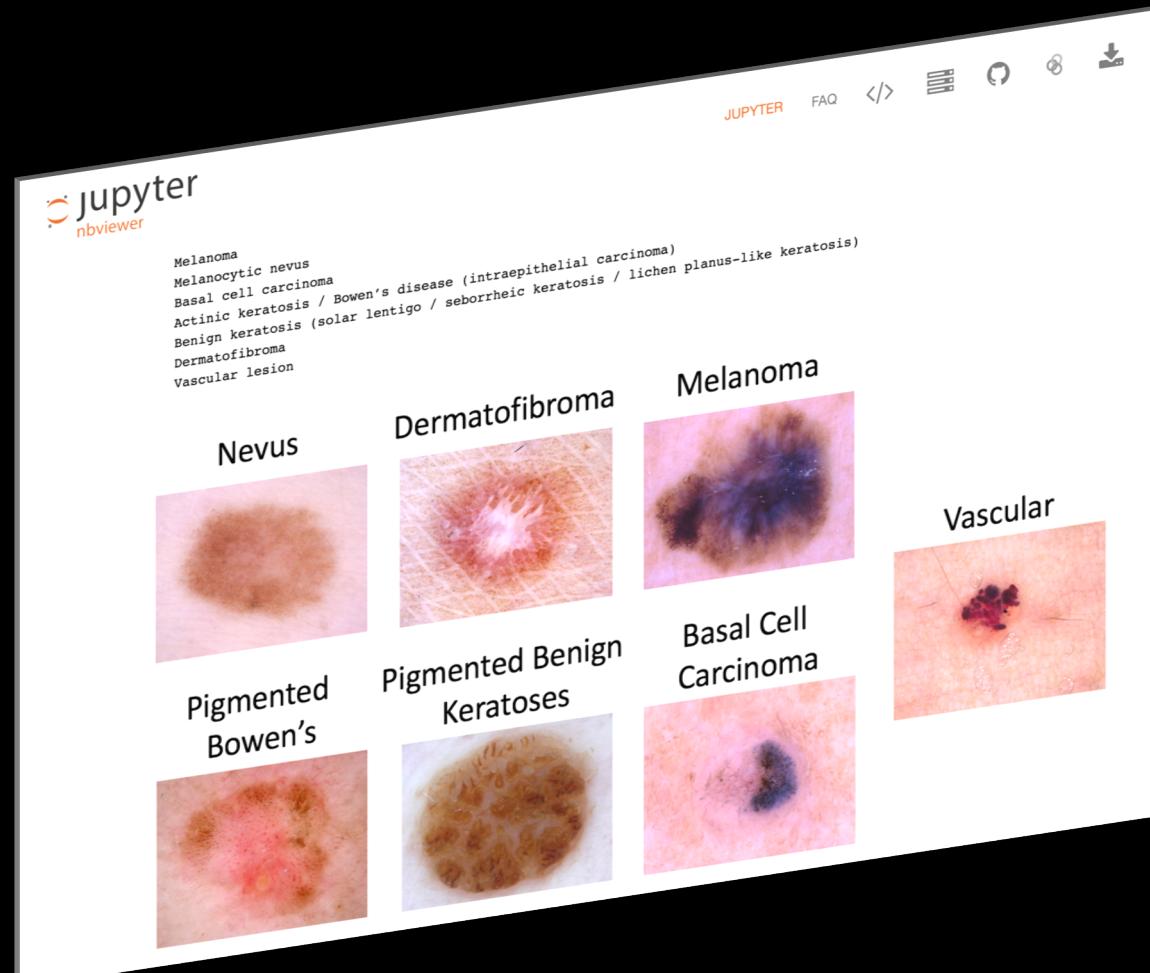
- Title:** Health and Lifestyle Survey Questions Tutorial
- Content Summary:** In this tutorial, we showcase how the ProtoDash explainer algorithm from AI Explainability 360 Toolkit implemented through the `ProtoDashExplainer` class could be used to summarize the National Health and Nutrition Examination Survey (NHANES) datasets ([Study 1](#)) available through the Center for Disease Control and Prevention (CDC). Moreover, we also show how the algorithm could be used to distill interesting relationships between different facets of life (i.e. early childhood and income), which were found by scientists ([Study 2](#)) through decades of rigorous experimentation. This study shows that in using ProtoDash, one can potentially uncover such insights cheaply, which could then be reaffirmed through rigorous experimentation.
- Text Block:** Data from this survey is typically used in epidemiological studies and health science research, which helps develop public health policy, direct and design health programs and services, and expand health knowledge. Thus, the impact of understanding these datasets and the relationships that may exist between them are far reaching for a social scientist.
- Section Header:** Introduction to Center for Disease Control and Prevention (CDC) datasets
- Text Block:** The [NHANES CDC questionnaire datasets](#) are surveys conducted by the organization involving thousands of civilians about various facets of their daily lives. There are 44 questionnaires that collect data about income, occupation, health, early childhood and many other behavioral and lifestyle aspects of individuals living in the US. These questionnaires are thus a rich source of information indicative of the quality of life of many civilians.
- Text Block:** This tutorial presents two studies. We first see how a CDC questionnaire answered by thousands of individuals could be summarized by looking at answers given by a few prototypical users. Next, an interesting endeavor is to uncover relationships between different aspects of life by analyzing data across the different CDC questionnaires. In the second study, we do exactly that with the help of the ProtoDash explainer algorithm. We show how the algorithm is able to uncover an interesting [insight](#) known only through decades of experimentation, solely from the questionnaire datasets. This by no means suggests the method as a substitute for rigorous experimentation, but showcases it as an avenue for obtaining interesting insights at low cost, which could inspire further indepth studies. The manner in which this is accomplished is by finding prototypical individuals for each of the questionnaires and then evaluating how well they represent the income questionnaire (w.r.t. the method's objective function). The more representative these prototypes are, the more that questionnaire is indicative/representative of income.
- Text Block:** For this use case, we are selecting prototypes from specific questionnaires. Hence, the group we want to explain is the dataset itself, which — in this case — are the questionnaires. We are not training an AI model. Rather, we are trying to summarize each questionnaire, which was filled by thousands of people, by selecting a few representative individuals for each of them.

<https://github.com/IBM/AIX360/tree/master/examples>

AGENDA



- Why Explainable AI?
 - Types and Methods for Explainable AI
- AI Explainability 360 Toolkit
 - Taxonomy and Guidance
- Interactive Web Experience Demo
- Hands on session 1
 - Package Installation and Git walkthrough
 - Use case (Industry): Personal finance
- Hands on session 2
 - Use case (Government): Health and nutrition
- **Hands on session 3**
 - Use case (Medicine): Clinical Medicine
 - Metrics
- Summary and future directions



<https://github.com/IBM/AIX360/tree/master/examples>

AGENDA



- Why Explainable AI?
 - Types and Methods for Explainable AI
- AI Explainability 360 Toolkit
 - Taxonomy and Guidance
- Interactive Web Experience Demo
- Hands on session 1
 - Package Installation and Git walkthrough
 - Use case (Industry): Personal finance
- Hands on session 2
 - Use case (Government): Health and nutrition
- Hands on session 3
 - Use case (Medicine): Clinical Medicine
 - Metrics
- **Summary and future directions**

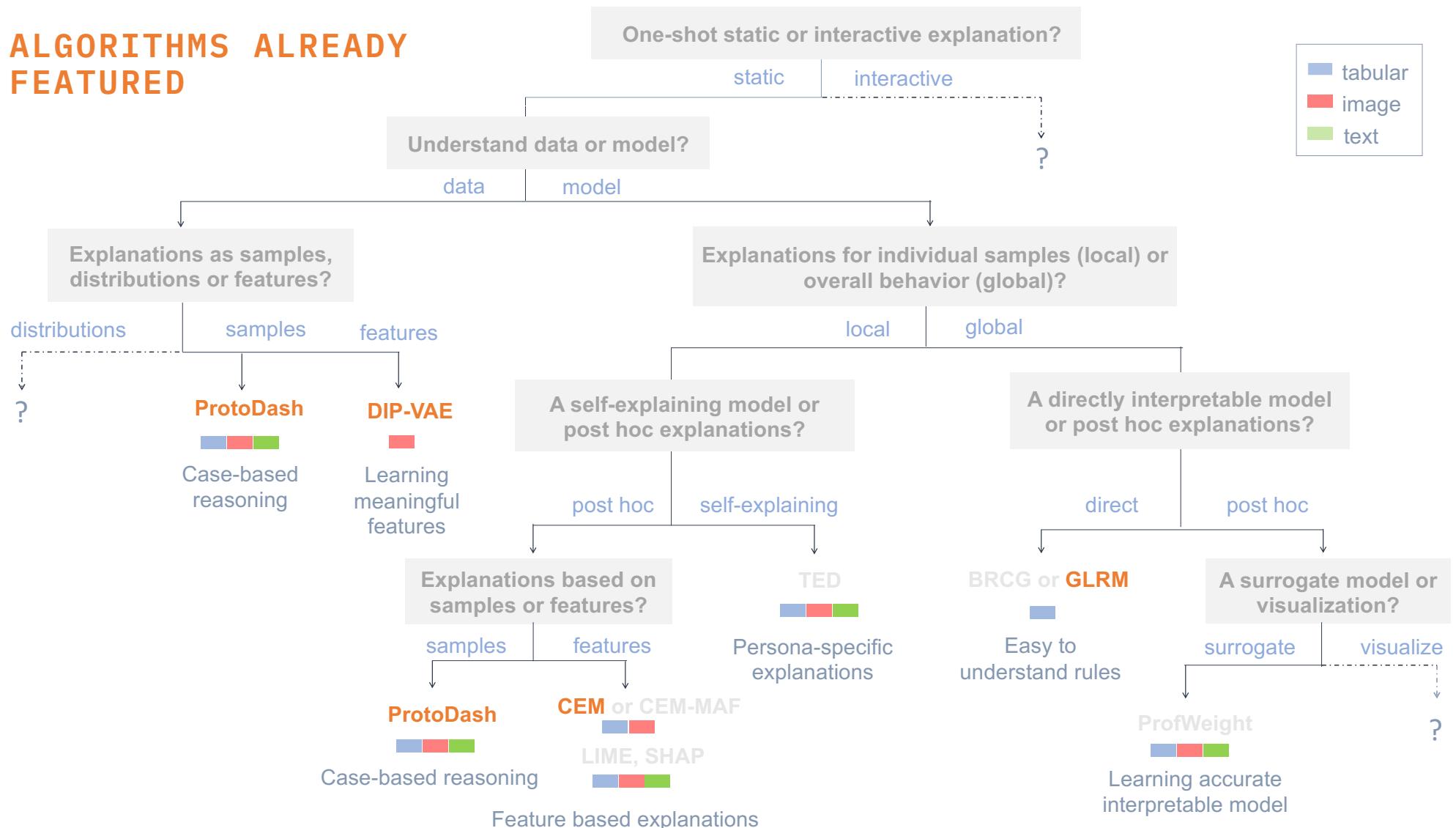


Summary and Future Directions

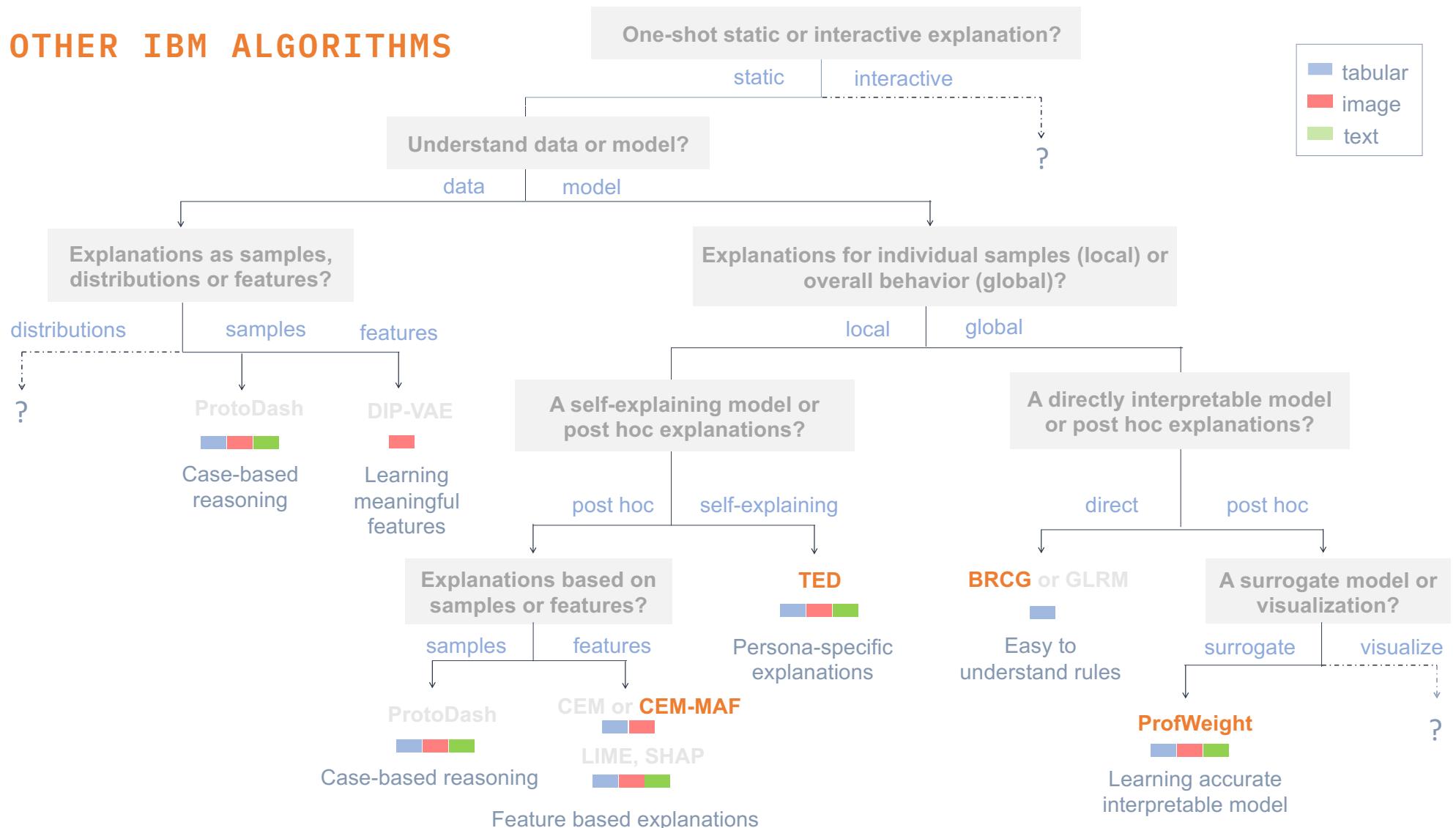
- **Algorithm Summary**
- AIX360 for Developers
- Future Directions in Explainability
- Future Directions for AIX360



ALGORITHMS ALREADY FEATURED



OTHER IBM ALGORITHMS



CEM-MAF: CONTRASTIVE EXPLANATIONS FOR COMPLEX IMAGES

MODEL - LOCAL - POST HOC

CEM produces

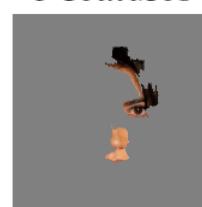
- Pertinent positives (PP): Present, minimally sufficient to yield classification
- Pertinent negatives (PN): Absent but (minimal) **addition** would change classification

Define **addition** in terms of higher-level concepts
e.g. high cheekbones, hair color, hair length

Represent concepts using *monotonic attribute functions* (MAF)

Advantages:

- More realistic output images
- Interpretable additions (PN)

INPUT	INPUT + PN	PP
old, male, not smiling	old, male, smiling	20 features 
		+ cheekbones
young, female, not smiling	young, male, not smiling	5 features 
		+ single hair color, - bangs



BRCG: BOOLEAN RULES VIA COLUMN GENERATION MODEL - GLOBAL - DIRECTLY INTERPRETABLE

Learns Boolean rules for binary classification

- Disjunctive normal form (DNF, OR of ANDs)
- Conjunctive normal form (CNF, AND of ORs)



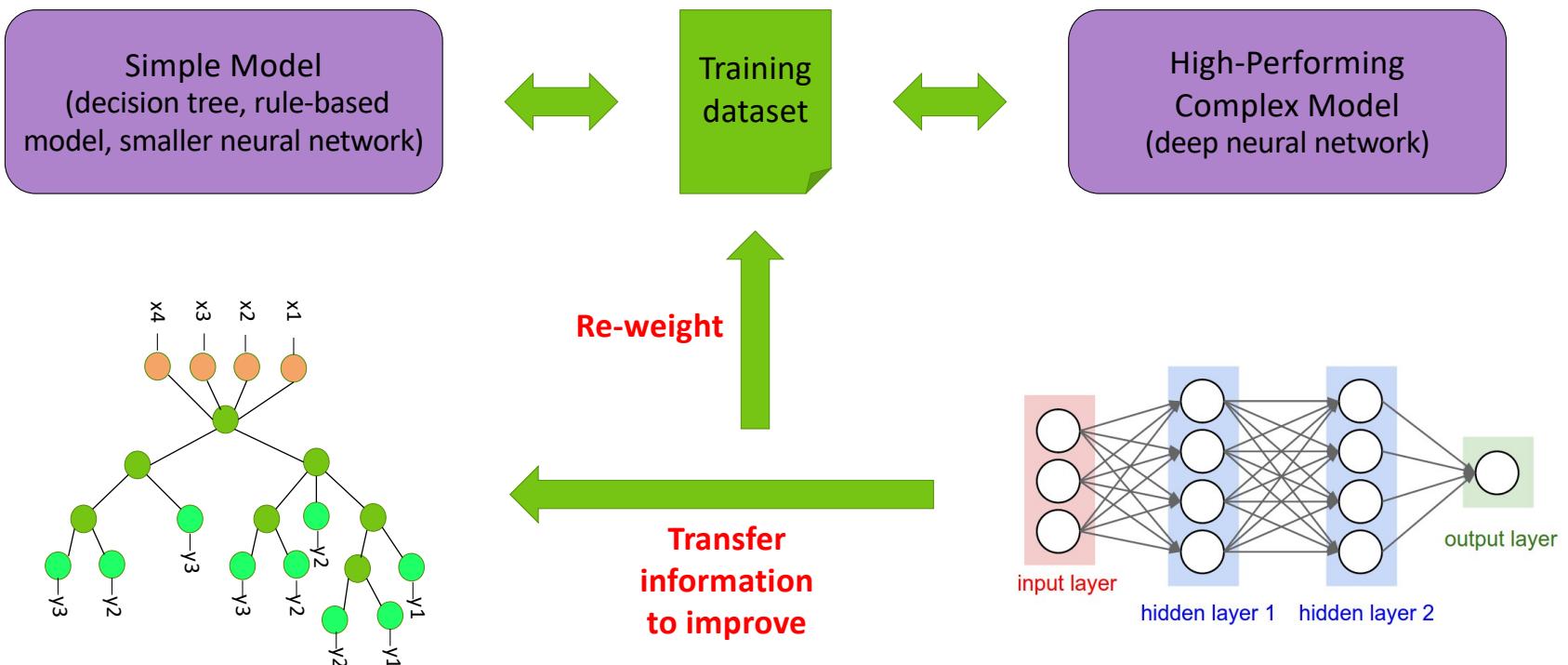
BRCG and GLRM are complementary rule-based methods

	GLRM	BRCG
Model produced	Generalized linear model (e.g. linear/logistic regression)	Binary classifier
Rule combination method	Linear combination	Logical OR or AND
Directly interpretable?	Yes	Even more so
How interpretability achieved	Few rules, short rules	
Optimization technique	Column generation	



PROFWEIGHT: IMPROVING INTERPRETABLE SURROGATES

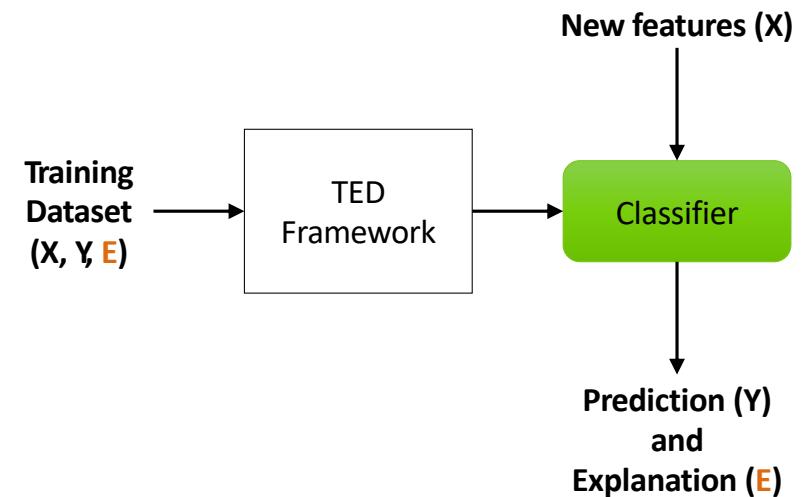
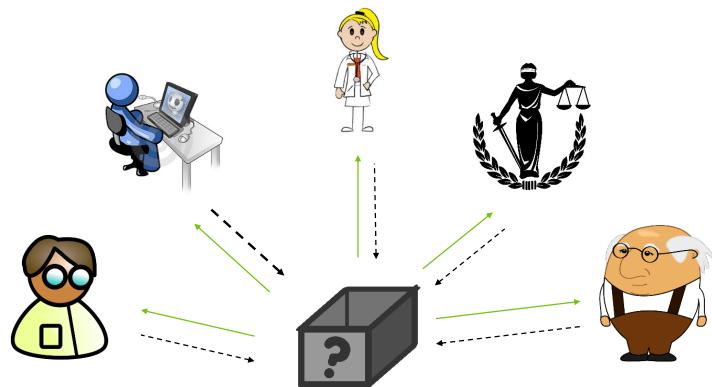
MODEL - GLOBAL - POST HOC



TED: TEACHING EXPLANATIONS FOR AI DECISIONS

MODEL - LOCAL - SELF-EXPLAINING

Different explanation consumers require different explanations



Consumer provides **training explanations** in addition to training labels

Learn to predict both label and explanation for unseen data point





Summary and Future Directions

- Algorithm Summary
- **AIX360 for Developers**
- Future Directions in Explainability
- Future Directions for AIX360

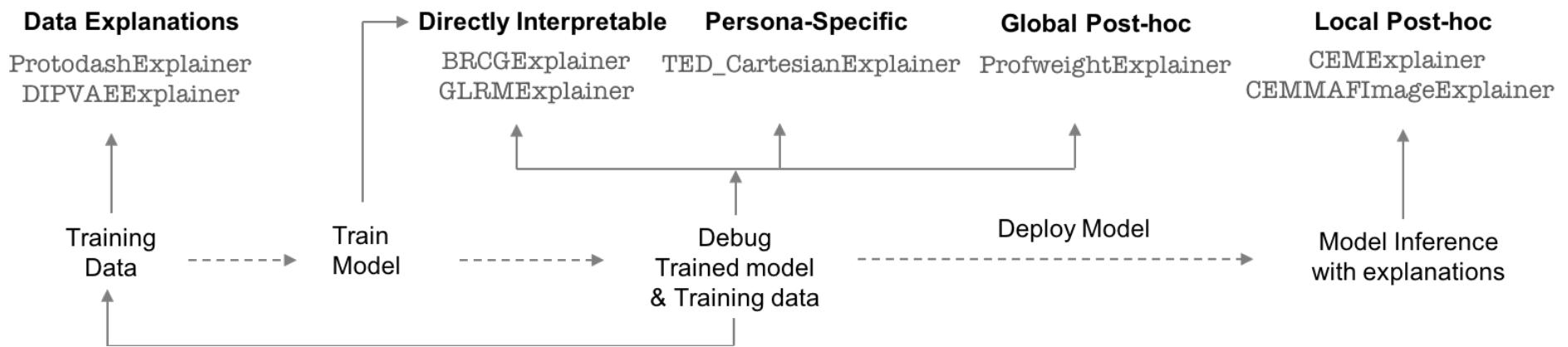


AIX360 CLASS HIERARCHY

- ❑ DIExplainer (Directly Interpretable unsupervised)
 - ProtodashExplainer
 - DIPVAEExplainer
- ❑ DISExplainer (Directly Interpretable Supervised)
 - BRCGExplainer
 - GLRMEExplainer
 - TED_CartesianExplainer
- ❑ LocalBBExplainer (Local Black-Box)
 - LIME Explainers
 - SHAP KernelExplainer
- ❑ LocalWBExplainer (Local White-Box)
 - CEMExplainer
 - CEM_MAFImageExplainer
 - SHAP Explainers
- ❑ GlobalBBExplainer (Global Black-Box)
- ❑ GlobalWBExplainer (Global White-Box)
 - ProfweightExplainer



CLASSES IN ML PIPELINE





Summary and Future Directions

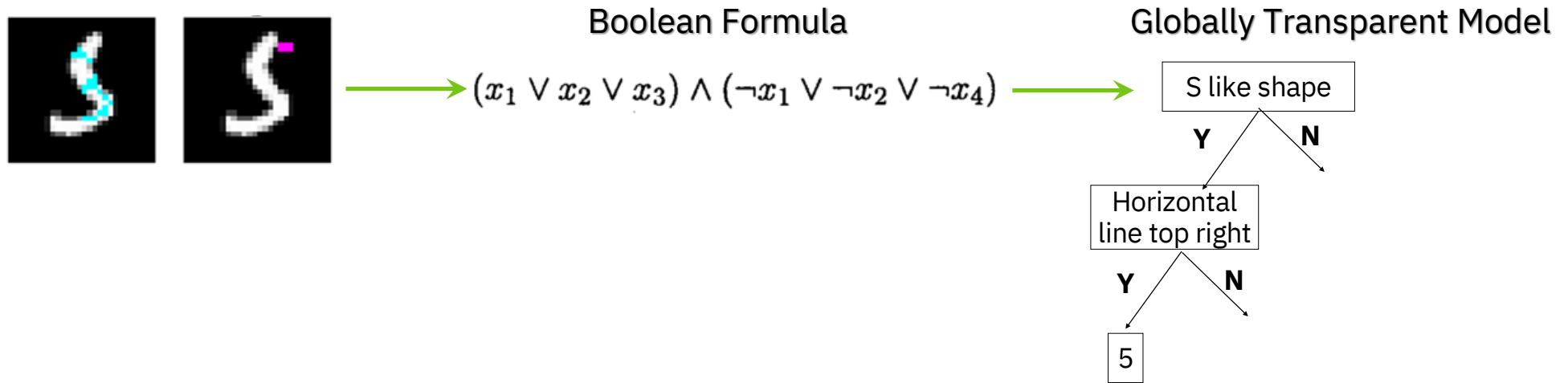
- Algorithm Summary
- AIX360 for Developers
- **Future Directions in Explainability**
- Future Directions for AIX360



Local-to-Global Interpretation

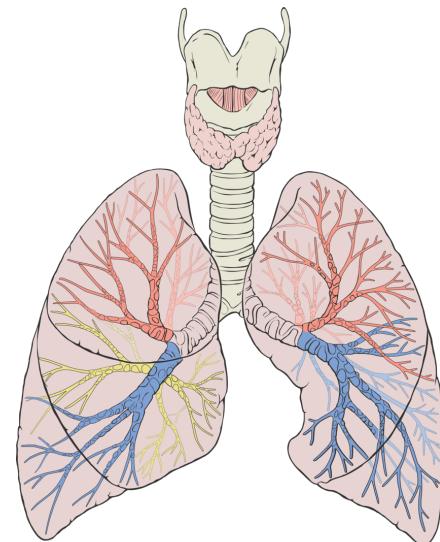
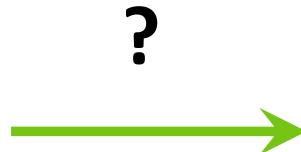
Local explanation methods could

- Extract useful features or a superset of rules to be passed to logic programs
- Be integrated into a coarse-to-fine hierarchy of explanations



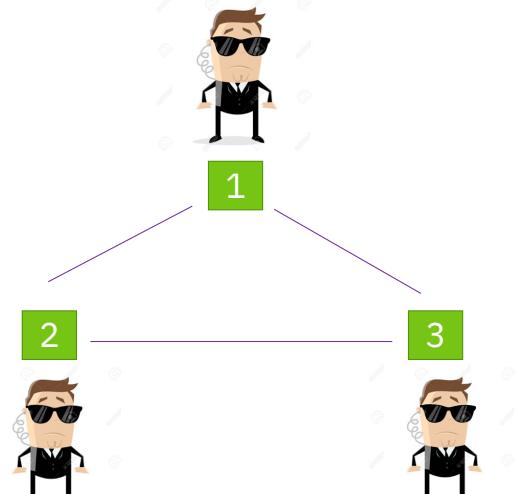
Causality

What is the true cause for an event? Interpretability methods can be used to identify where to look (reduce search space) before causal methods are applied.



Reinforcement Learning

Explanation methods are essentially communication methods that convey feature importances or representative examples. One could envision these methods being used in multiagent systems for teaching one another.





Summary and Future Directions

- Algorithm Summary
- AIX360 for Developers
- Future Directions in Explainability
- **Future Directions for AIX360**



The future of AIX360 is people like you!



CONTRIBUTING TO AIX360

Want to contribute?

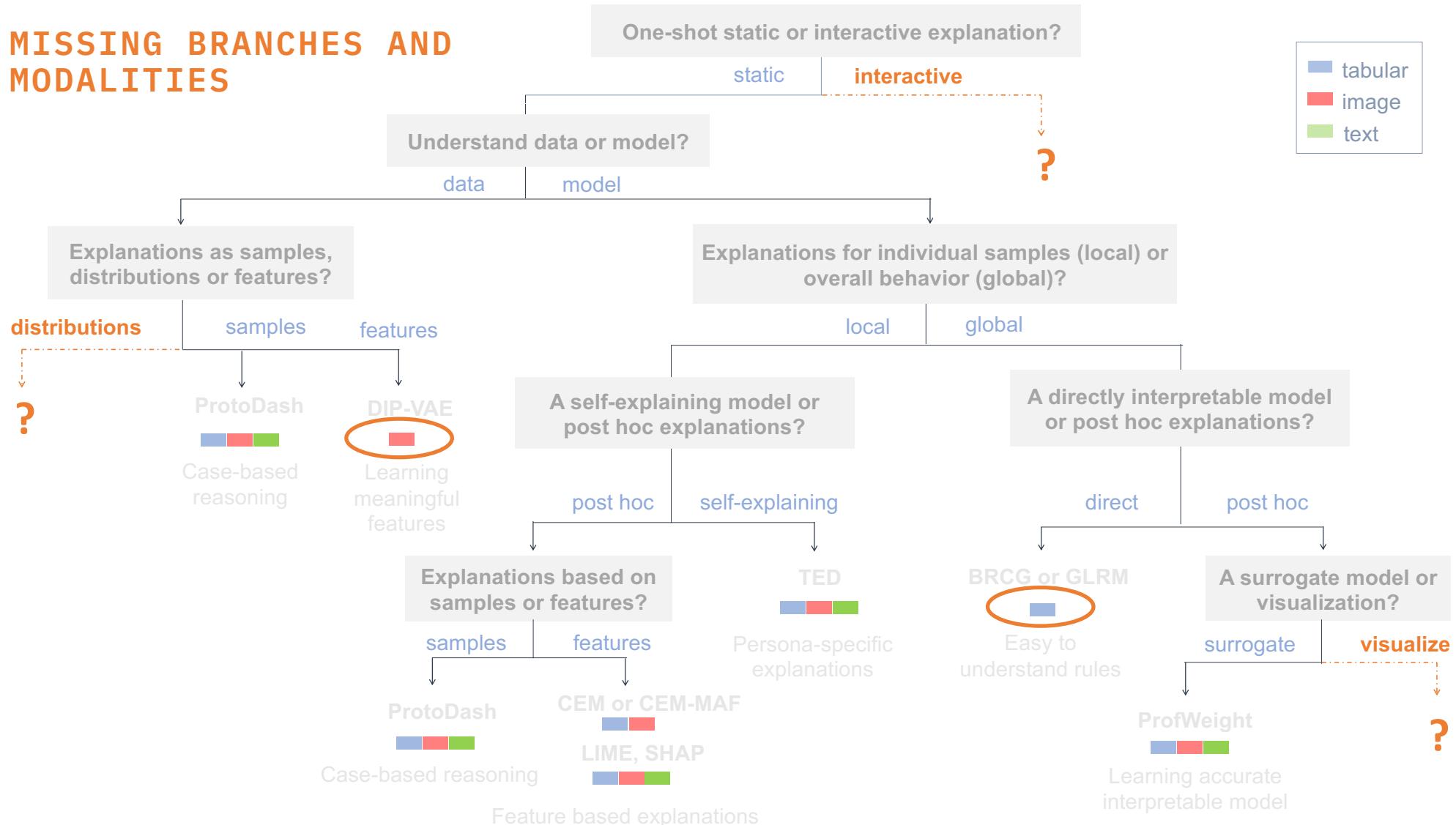
- Start a discussion in our Slack workspace
- Create a GitHub issue
- Get working!

The image shows two screenshots illustrating the contribution process for AIX360.

Slack Workspace: The left screenshot shows the AIX360 Slack workspace interface. The sidebar lists channels: # aix360-developers (selected), # aix360-users, # fat-tutorial-2020, # general, and # random. The main area shows a message from Kush Varshney (@ArpitSisodia) at 6:41 PM, responding to Arpit Sisodia (@KushVarshney) about trust scores and classifier explanations. Arpit replies at 9:15 PM asking for deeper explanation. The bottom navigation bar includes Pull requests, Issues, Marketplace, and Explore.

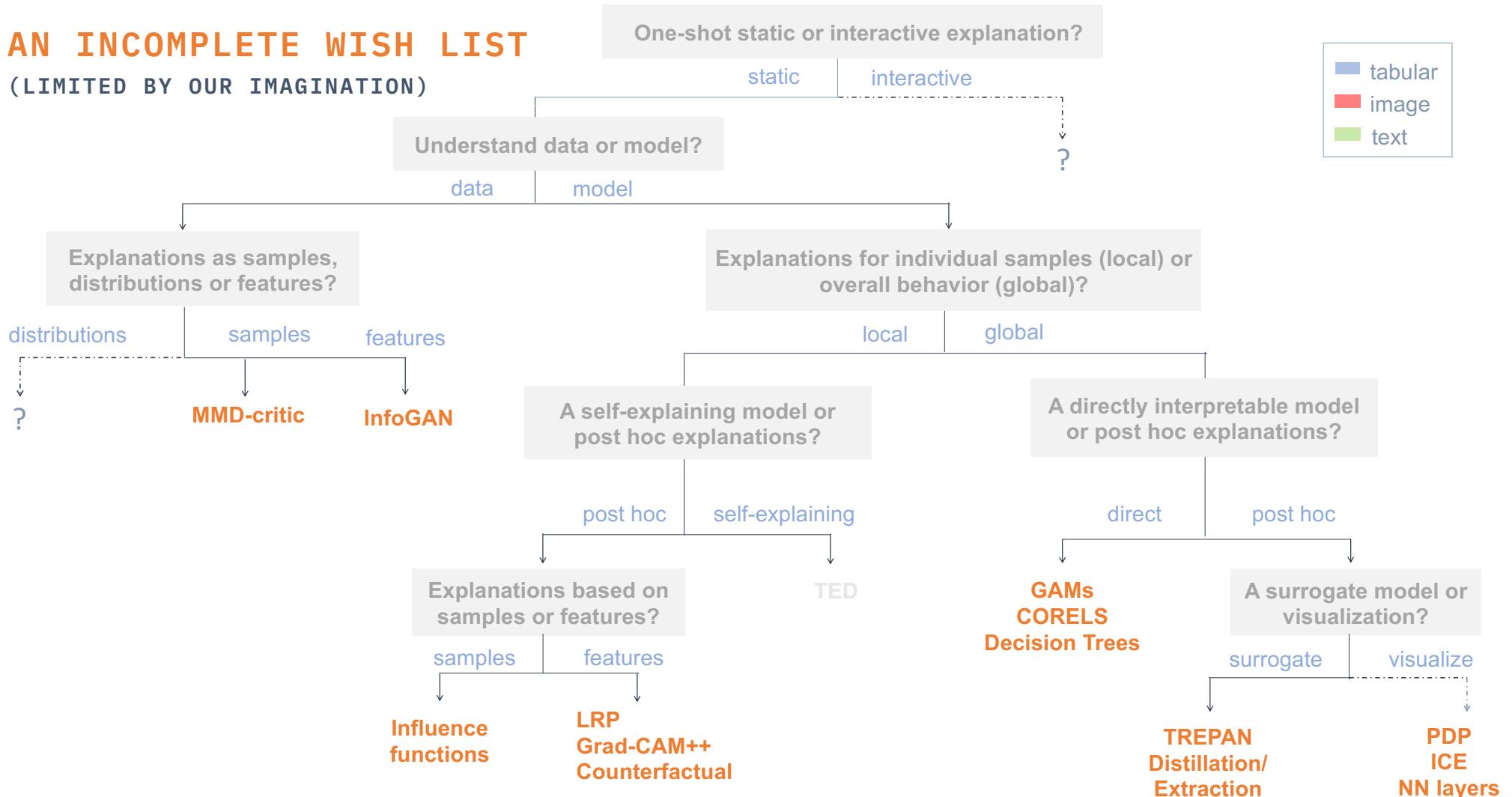
GitHub Repository: The right screenshot shows the GitHub repository page for IBM / AIX360. The repository has 440 stars and 89 forks. The Issues tab is selected, showing 6 open issues, 3 pull requests, and 0 projects. The right sidebar provides options for Assignees, Labels, Projects, and Milestone, all currently set to "None yet". A large central area is for creating a new issue, featuring fields for Title, Write, Preview, and a rich text editor. A note says "Styling with Markdown is supported". A "Submit new issue" button is at the bottom right.

MISSING BRANCHES AND MODALITIES



AN INCOMPLETE WISH LIST

(LIMITED BY OUR IMAGINATION)



SUMMARY

- Why Explainable AI?
 - **Trust**, societal calls, better systems, etc.
- AIX360 Toolkit
 - **Many ways to explain**
 - 10 algorithms and 2 metrics (currently)
 - Data vs. model, local vs. global, direct vs. post hoc
- Toward an Explainability Community
 - Users: web demo, 3 in-depth use cases
 - Developers: Solicit contributions to fill in gaps and expand scope

