AI Explainability 360

Vijay Arya, Amit Dhurandhar, Dennis Wei

IBM Research AI

Jan 27th, 2020

 × 2019 IBM Corporation

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

 × 2019 IBM Corporation

AGENDA

- **Why Explainable AI?**
  - Types and Methods for Explainable AI
- AI Explainability 360 Toolkit
  - Taxonomy and Guidance

30

- Interactive Web Experience Demo

15

- Hands on session 1
  - Package Installation and Git walkthrough
  - Use case (Industry): Personal finance

45

Break 30

- Hands on session 2
  - Use case (Government): Health and nutrition

25

- Hands on session 3
  - Use case (Medicine): Clinical Medicine
  - Metrics

30

- Summary and future directions

30

 × 2019 IBM Corporation

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

Credit

Employment

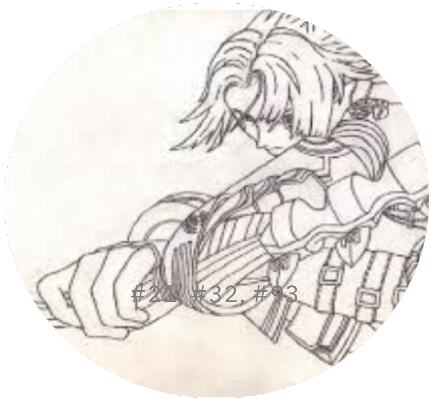
Admission

Sentencing

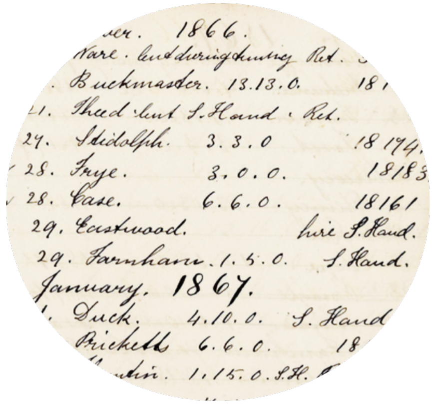
WHAT DOES IT TAKE TO TRUST A DECISION MADE BY A MACHINE (OTHER THAN THAT IT IS 99% ACCURATE)

Is it fair?

Is it easy to understand?



Did anyone tamper with it?



Is it accountable?

THE QUEST FOR "EXPLAINABLE AI"

CIO JOURNAL.

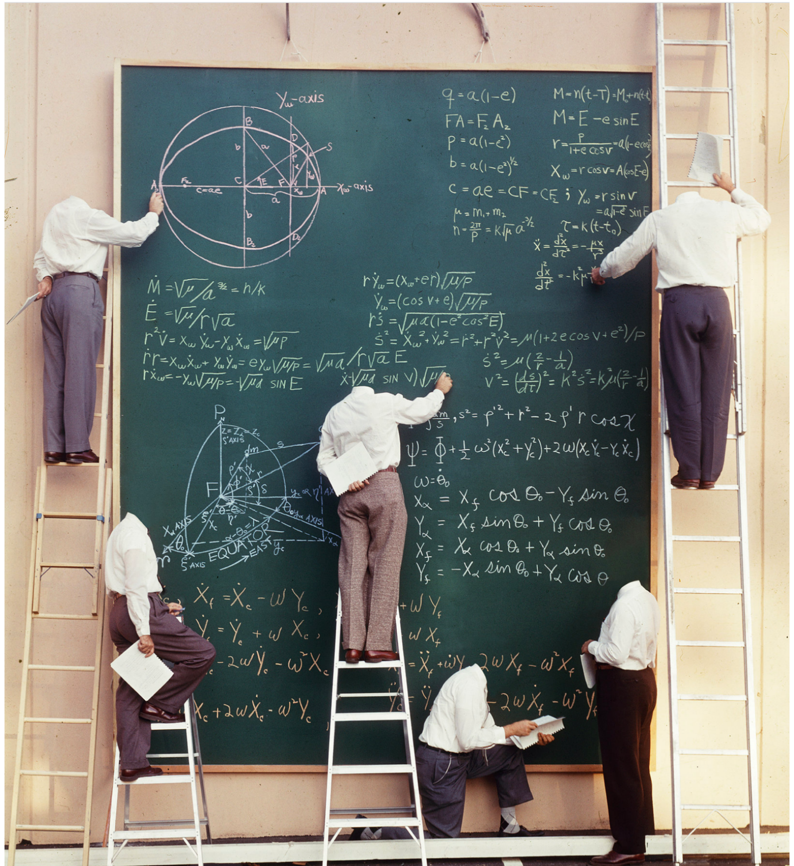
Companies Grapple With AI's Opaque Decision-Making Process

THE WALL STREET JOURNAL.

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

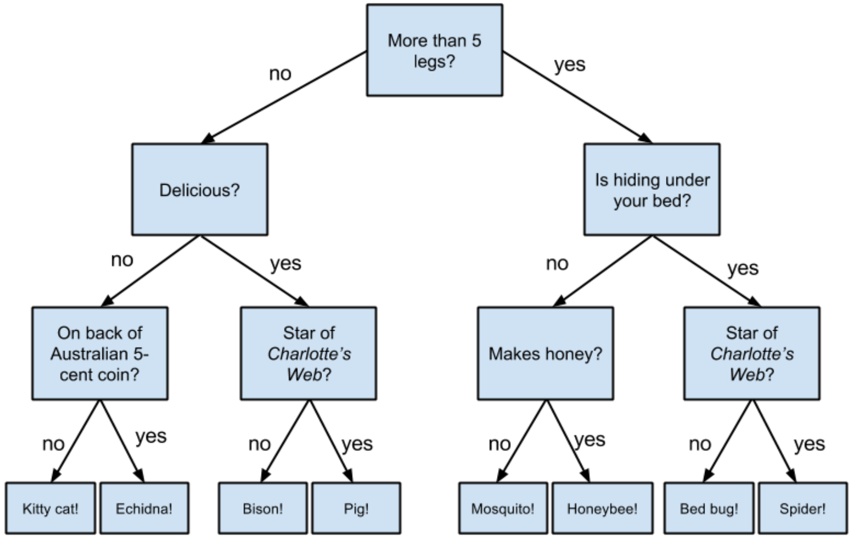
When a Computer Program Keeps You in Jail

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



 © 2019 IBM Corporation

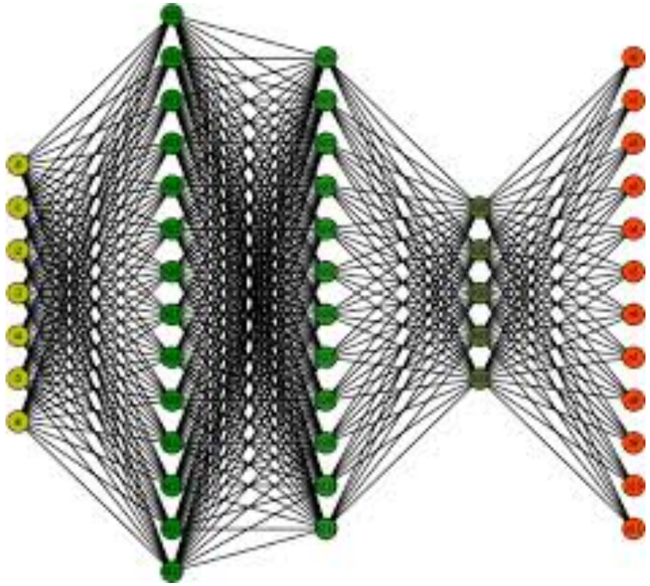
Decision Tree



Interpretable?

YES

Neural Network



Interpretable?

NO

 © 2019 IBM Corporation

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)

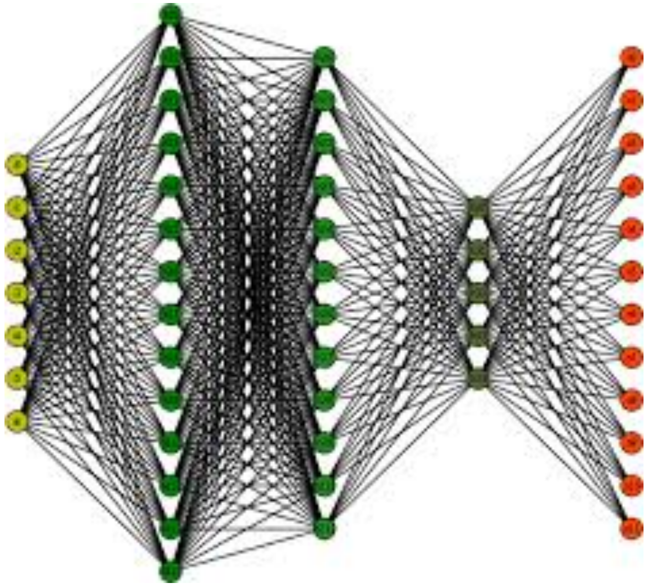
"meaningful" ???

 © 2019 IBM Corporation

Paul Nemitz, *Principal Advisor, European Commission*
Talk at IBM Research, Yorktown Heights, May, 4, 2018

Simplification

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



Insight

Check if code has comments

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???

Existence of Confounders

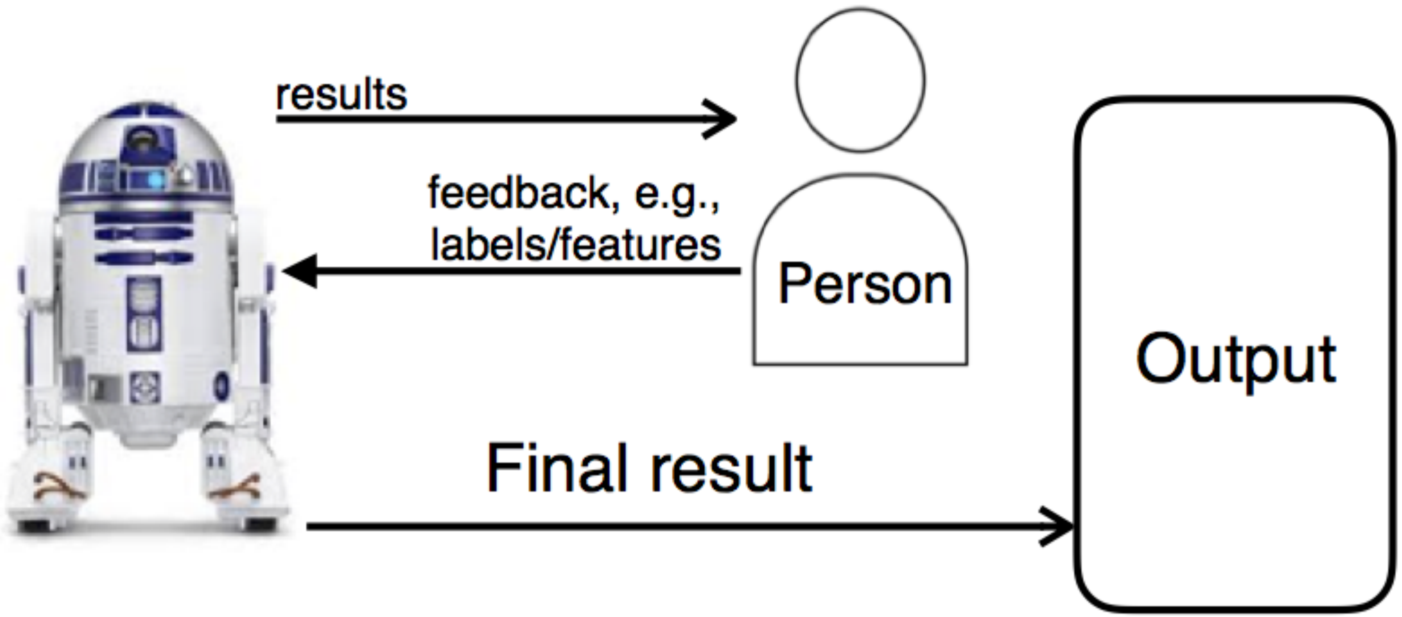
Can help to identify spurious correlations.

Pneumonia

~~Diabetes~~

Enhance Performance

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



 × 2019 IBM Corporat:

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]

 × 2019 IBM Corporation

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

directly interpretable vs. **post hoc interpretation**

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.

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.

global (model-level) vs. **local (instance-level)**

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).

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).

static vs. **interactive (visual analytics)**

static
The interpretation is simply presented to the user.

interactive (visual analytics)
The user can interact with interpretation.

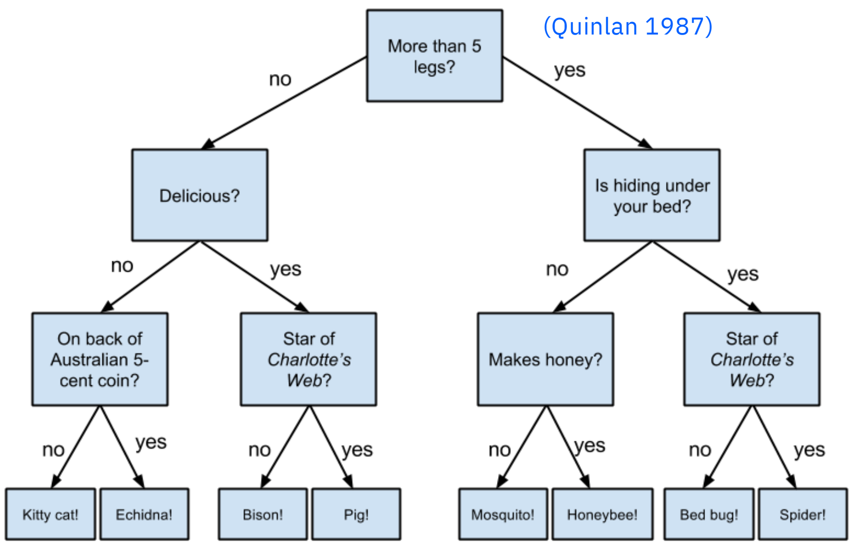
 × 2019 IBM Corporation

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

(Quinlan 1987)



Rule List

(Wang and Rudin 2016)

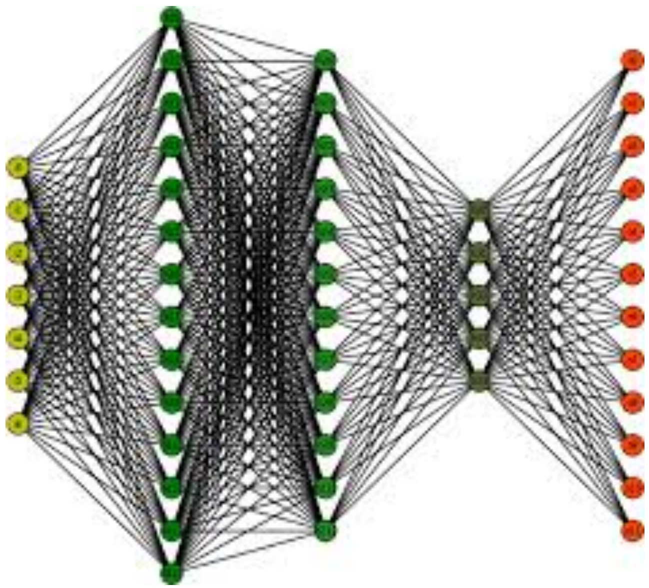
if	capital-gain>$7298.00	then probability to make over 50K	= 0.986
else if	Young,Never-married,	then probability to make over 50K	= 0.003
else if	Grad-school,Married,	then probability to make over 50K	= 0.748
else if	Young,capital-loss=0,	then probability to make over 50K	= 0.072
else if	Own-child,Never-married,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

 © 2019 IBM Corporation

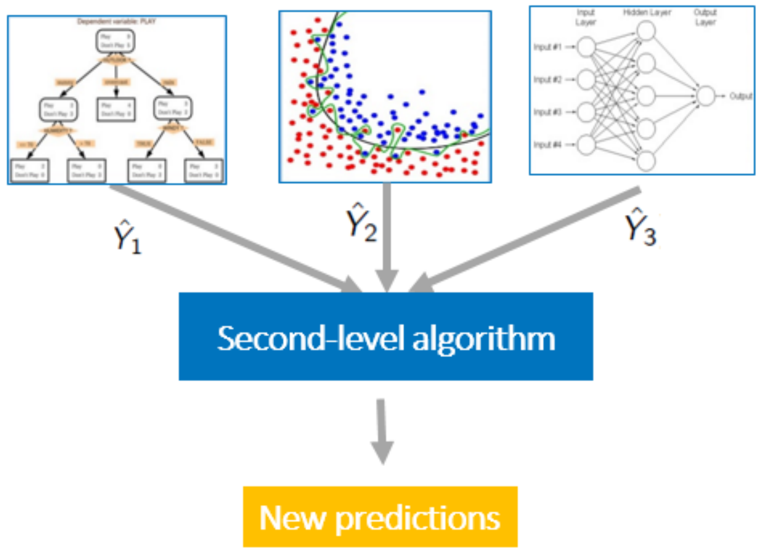
Post hoc interpretation

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



Post hoc (local) interpretation

Locally Interpretable Model Agnostic Explanations (LIME)

(Ribeiro et. al. 2016)

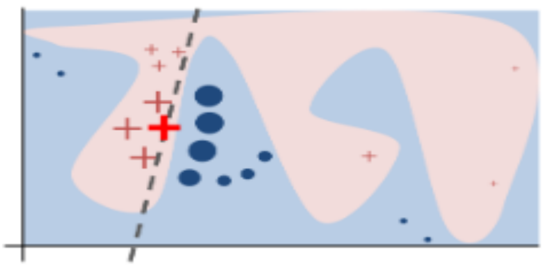


Figure 1. Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background. The bright bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity 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 π_x, Length of explanation K
  $\mathcal{Z} \leftarrow \{\}$
  for $i \in \{1, 2, 3, ..., N\}$ **do**
    $z_i' \leftarrow sample\_around(x')$
    $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z_i', f(z_i), \pi_x(z_i) \rangle$
  end for
  $w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ $\triangleright$ with z_i' as features, $f(z)$ as target
  return w

 × 2019 IBM Corporation

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

$$J_b(\mathsf{S}) = \frac{1}{n^2} \sum_{i,j=1}^{n} k(x_i, x_j) - \mathrm{MMD}^2(\mathcal{F}, X, X_{\mathsf{S}})$$
$$= \frac{2}{n|\mathsf{S}|} \sum_{i \in [n], j \in \mathsf{S}} k(x_i, y_j) - \frac{1}{|\mathsf{S}|^2} \sum_{i,j \in \mathsf{S}} k(y_i, x_j).$$

 © 2019 IBM Corporation

Post hoc (local) interpretation

Saliency Maps

(Sinmoyan et. al. 2013)

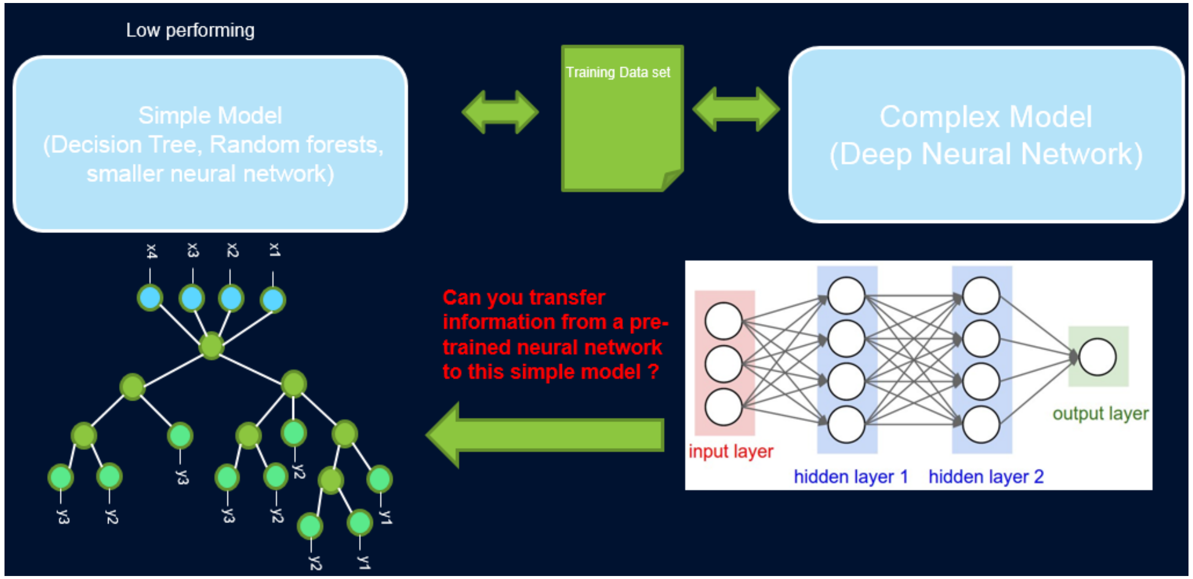
$$w = \left.\frac{\partial S_c}{\partial I}\right|_{I_0}.$$

 × 2019 IBM Corporation

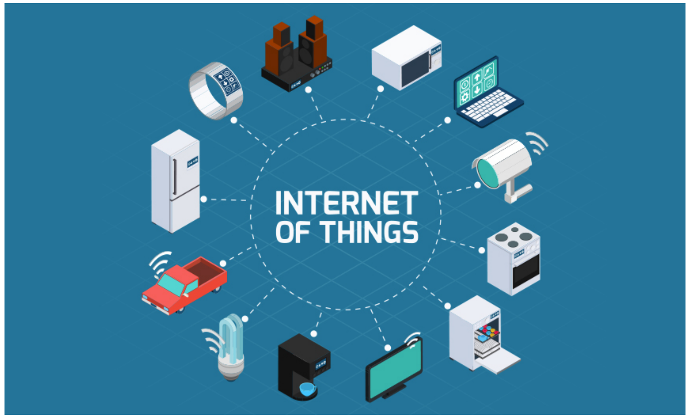
Post hoc (global) interpretation

Knowledge Distillation

(Hinton et. al. 2015)



Complex Systems



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

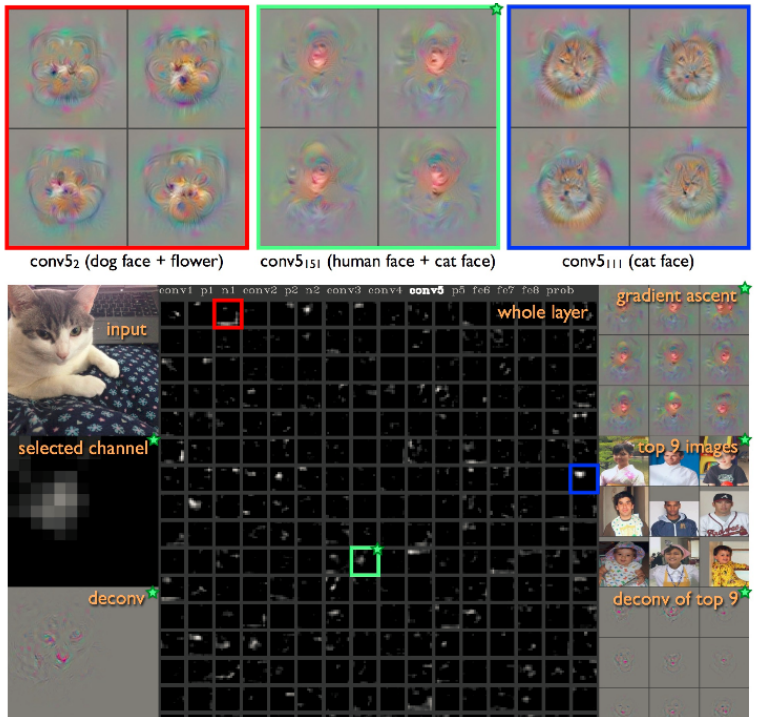
 × 2019 IBM Corporation

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)



 © 2019 IBM Corporation

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.

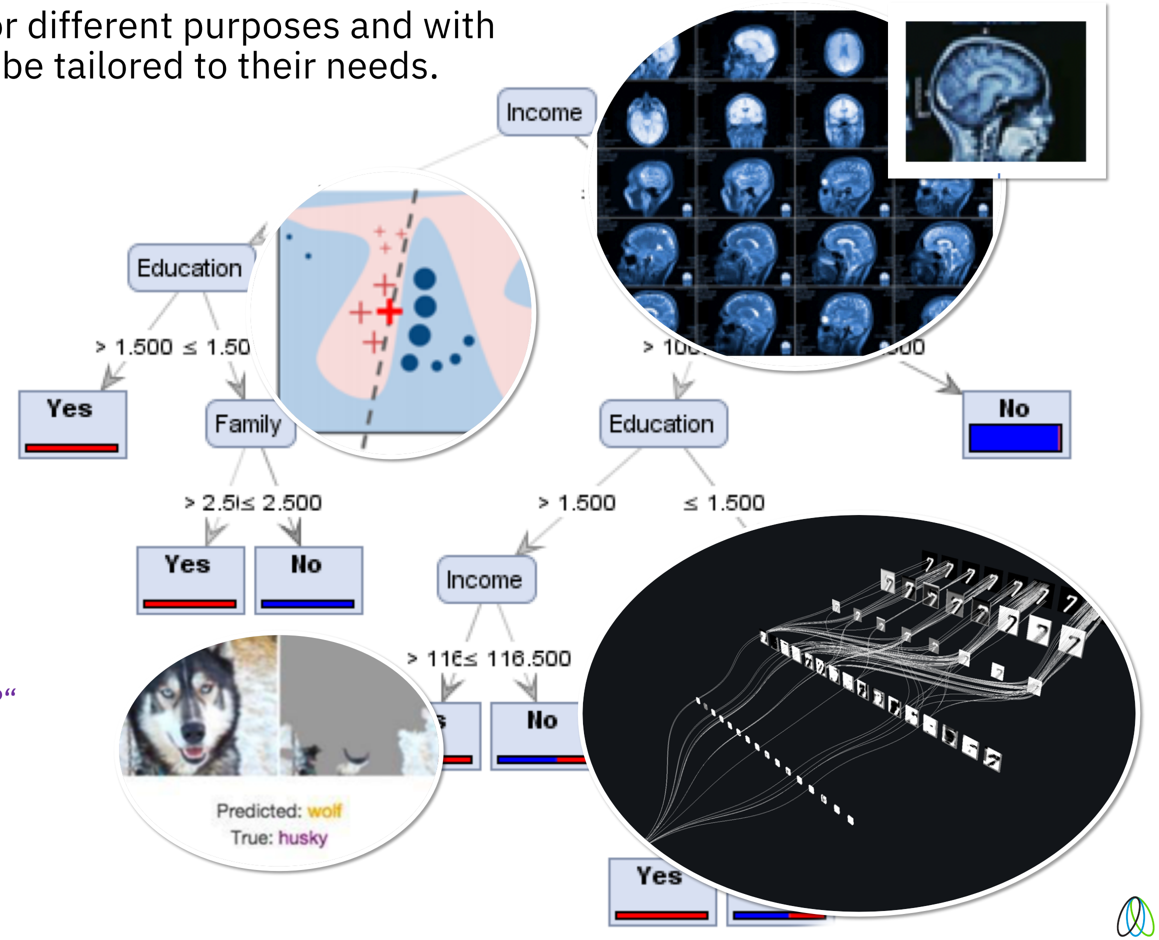
Why: ensure fairness for constituents

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

 × 2019 IBM Corporation

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 AlleyWatch

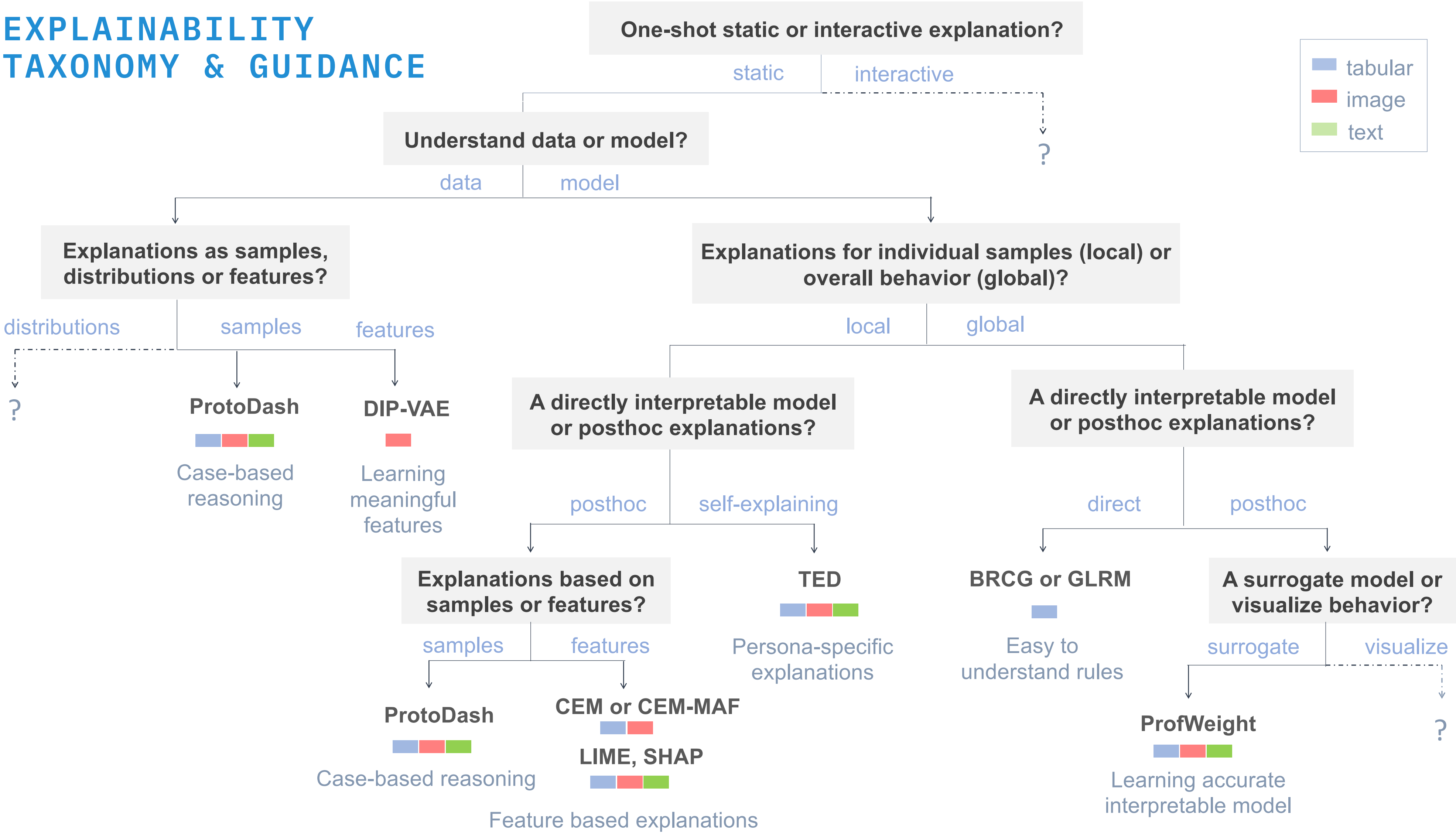
Don't Trust Artificial Intelligence? Time To Open The AI 'Black Box'

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				✓		

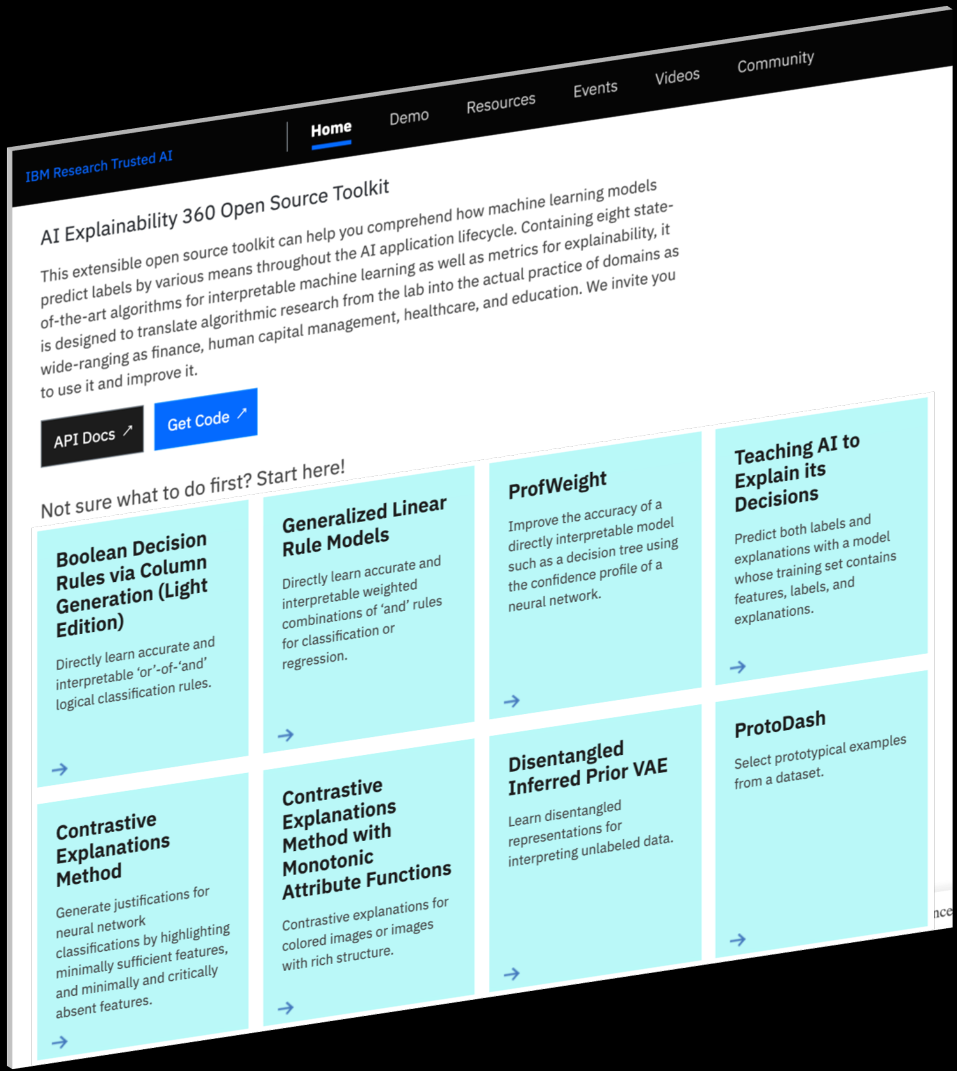
All algorithms of AIX360 are developed by IBM Research
AIX360 also provides demos, tutorials, and guidance on explanations for different use cases.
Paper: One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques: https://arxiv.org/abs/1909.03012v1

 × 2019 IBM Corporation

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

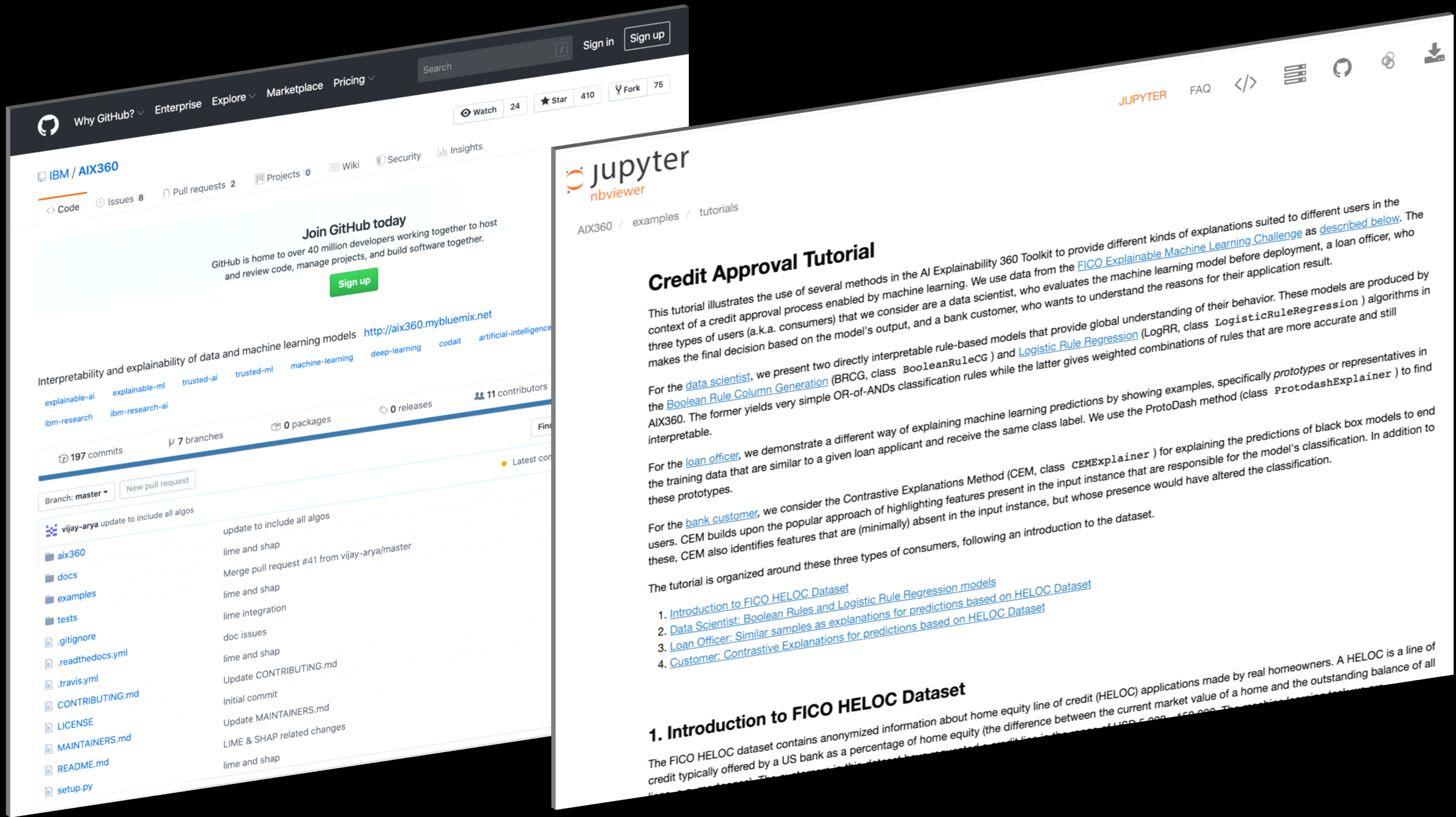
 × 2019 IBM Corporation



http://aix360.mybluemix.net/

AGENDA

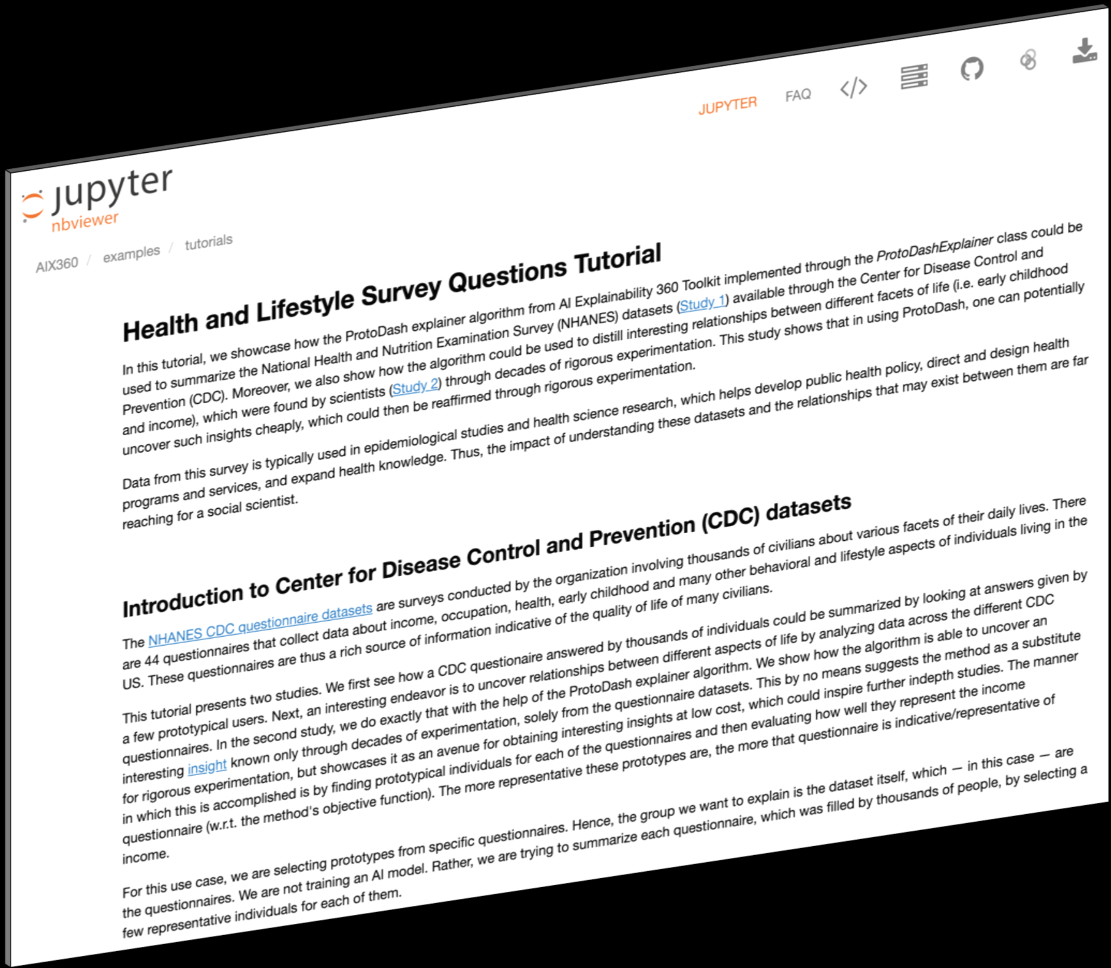
- Why Explainable AI?
  - Types and Methods for Explainable AI
- AI Explainability 360 Toolkit
  - Taxonomy and Guidance
- Interactive Web Experience Demo
- **Hands on session 1**
  - Package Installation and Git walkthrough
  - Use case (Industry): Personal finance
- Hands on session 2
  - Use case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



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


jupyter nbviewer

JUPYTER FAQ

AIX360 / examples / tutorials

Health and Lifestyle Survey Questions Tutorial

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.

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.

Introduction to Center for Disease Control and Prevention (CDC) datasets

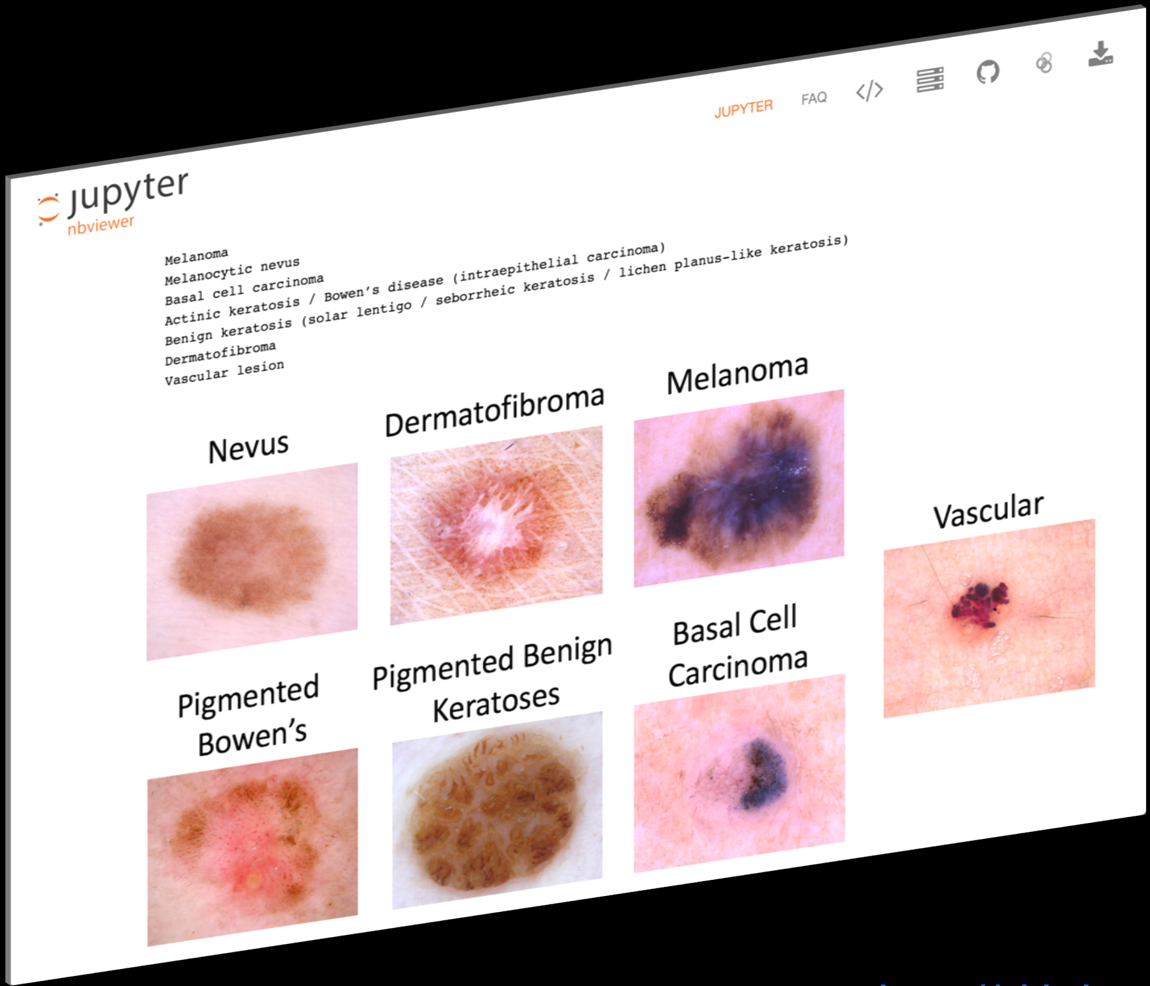
The NHANES CDC questionnaire datasets are surveys conducted by the organization involving thousands of civilians about various facets of their daily lives. There are 44 questionnaires that collect data about income, occupation, health, early childhood and many other behavioral and lifestyle aspects of individuals living in the US. These questionnaires are thus a rich source of information indicative of the quality of life of many civilians.

This tutorial presents two studies. We first see how a CDC questionaire 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.

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

- 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**

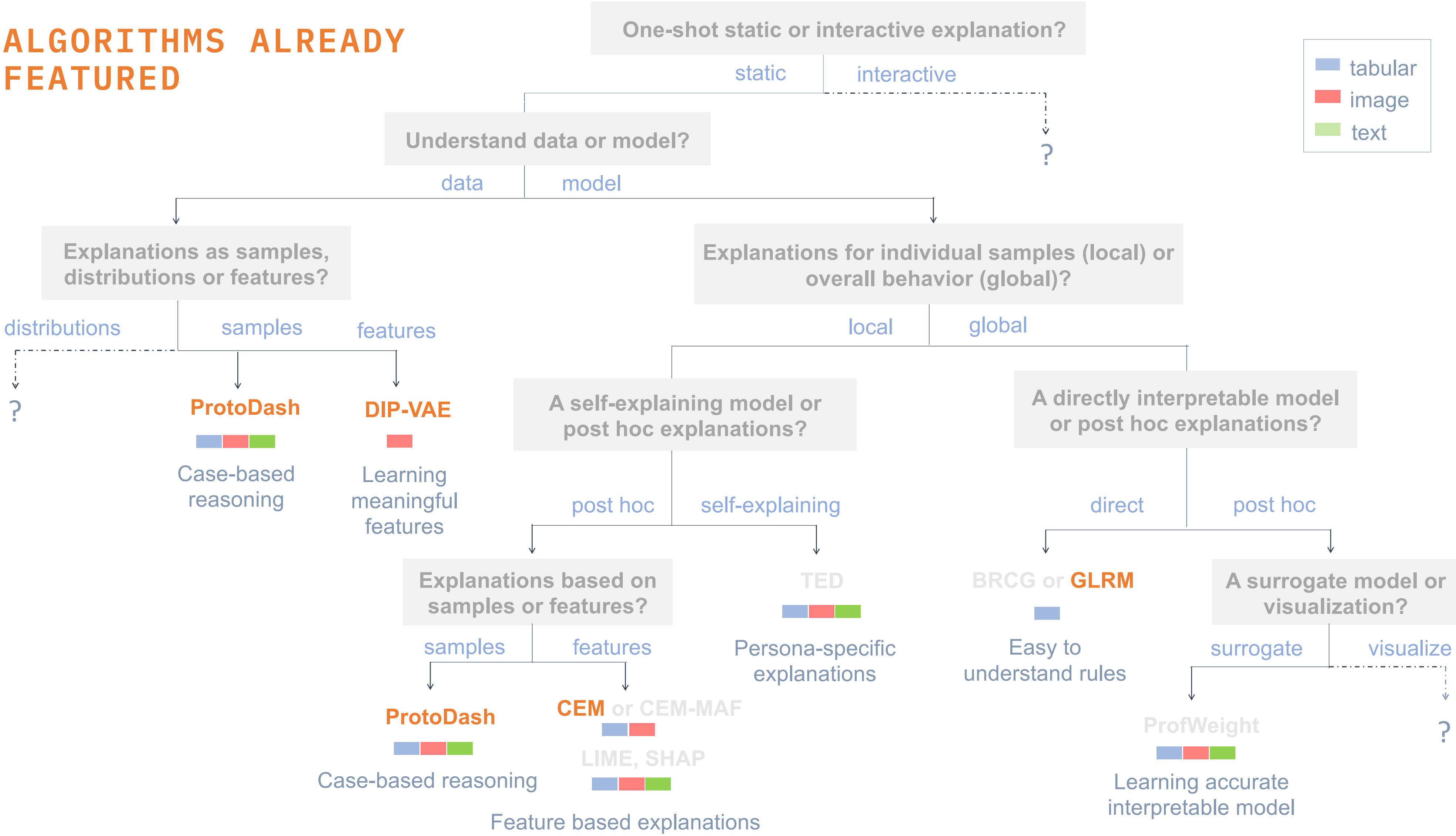
 © 2019 IBM Corporation

Summary and Future Directions

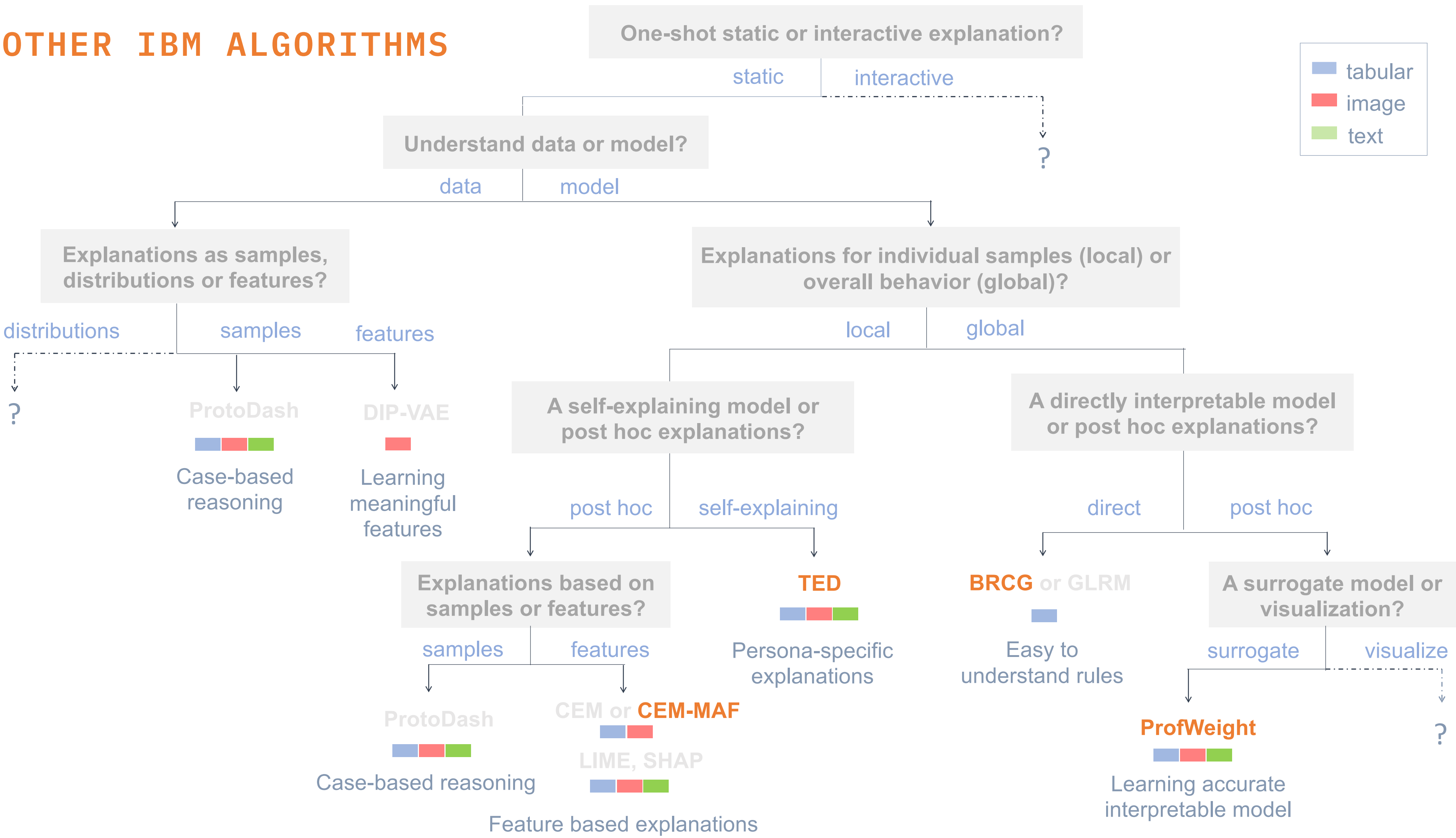
- **Algorithm Summary**
- AIX360 for Developers
- Future Directions in Explainability
- Future Directions for AIX360

 × 2019 IBM Corporation

ALGORITHMS ALREADY FEATURED



OTHER IBM ALGORITHMS



CEM-MAF: CONTRASTIVE EXPLANATIONS FOR COMPLEX IMAGES

MODEL – LOCAL – POST HOC

CEM produces

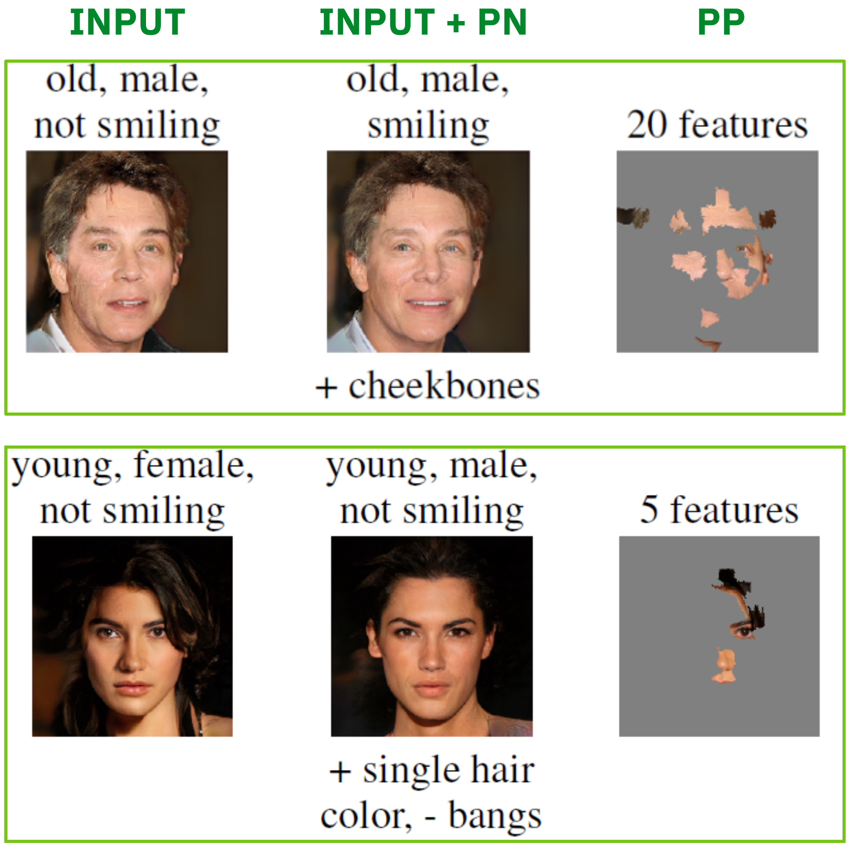
- Pertinent positives (PP): Present, minimally sufficient to yield classification
- Pertinent negatives (PN): Absent but (minimal) **addition** would change classification

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)



 × 2019 IBM Corporation

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)

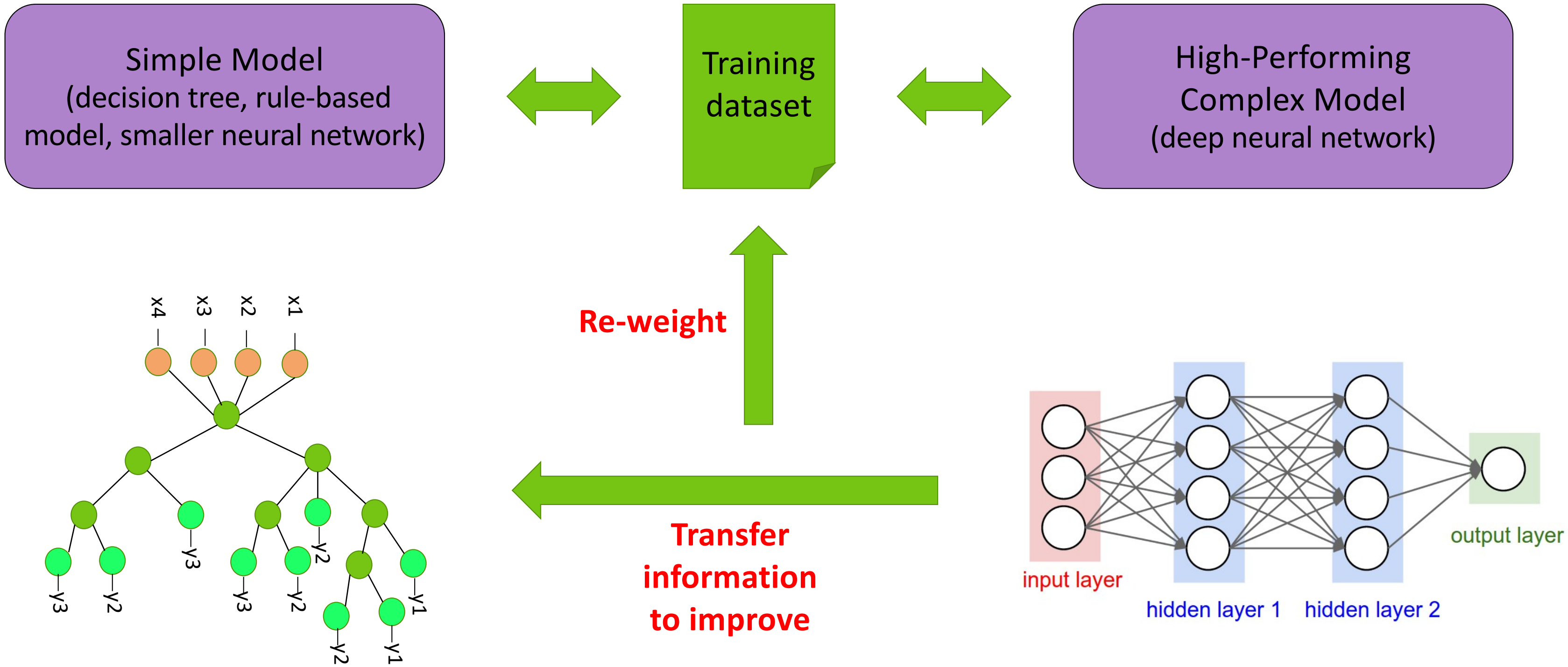
accounts < 5 OR # accounts ≥ 5 AND Debt > $1000 → Credit risk = high

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	

 × 2019 IBM Corporation

PROFWEIGHT: IMPROVING INTERPRETABLE SURROGATES
MODEL – GLOBAL – POST HOC

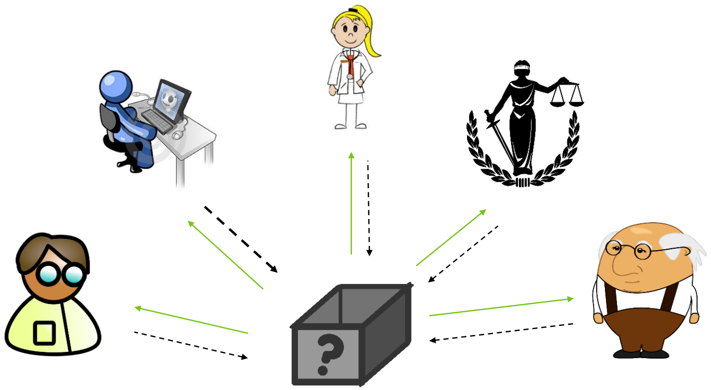


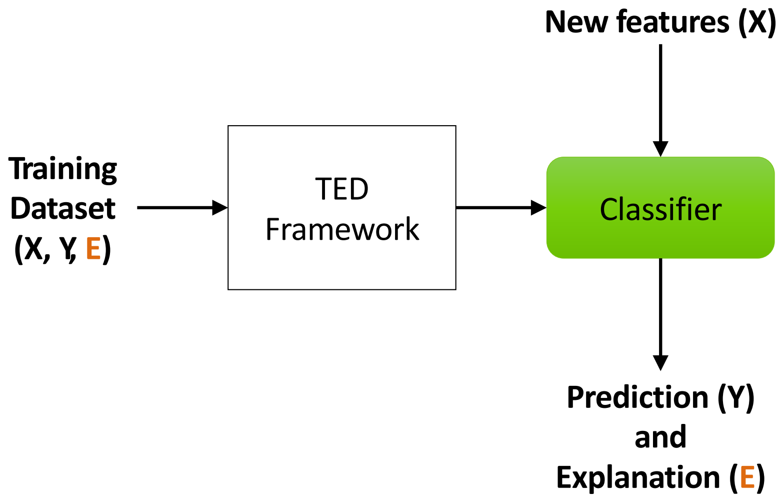
 × 2019 IBM Corporation

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

 × 2019 IBM Corporation

Summary and Future Directions

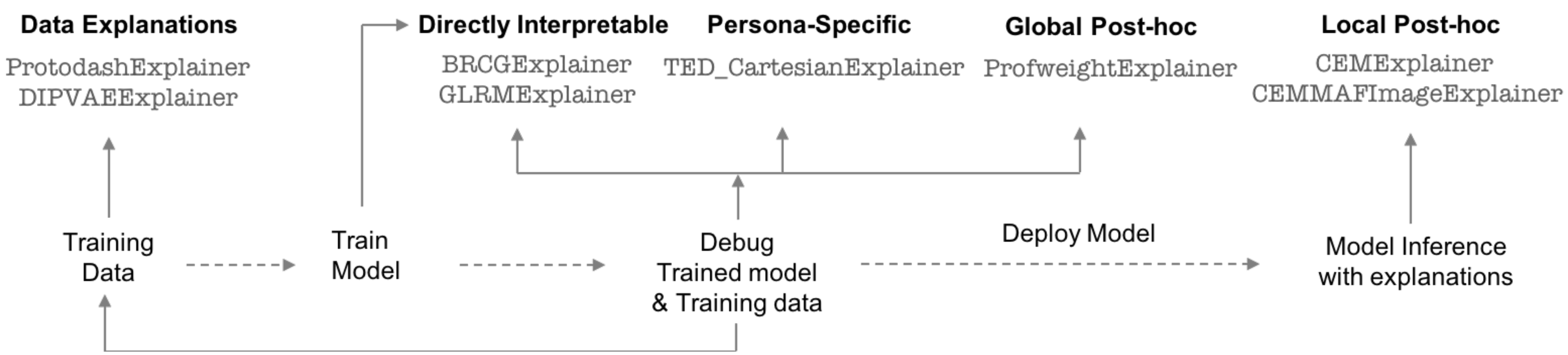
- Algorithm Summary
- **AIX360 for Developers**
- Future Directions in Explainability
- Future Directions for AIX360

AIX360 CLASS HIERARCHY

- DIExplainer (Directly Interpretable unsupervised)
  - ProtodashExplainer
  - DIPVAEExplainer
- DISExplainer (Directly Interpretable Supervised)
  - BRCGExplainer
  - GLRMExplainer
  - 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

 © 2019 IBM Corporation

CLASSES IN ML PIPELINE



 × 2019 IBM Corporation

Summary and Future Directions

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

 × 2019 IBM Corporation

Local-to-Global Interpretation

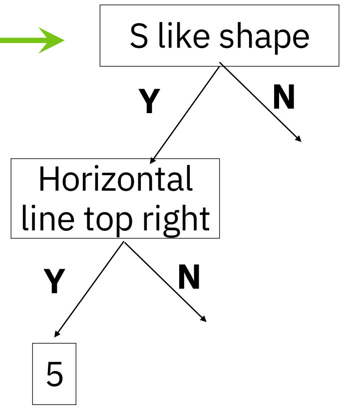
Local explanation methods could

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

Boolean Formula

$$(x_1 \vee x_2 \vee x_3) \wedge (\neg x_1 \vee \neg x_2 \vee \neg x_4)$$

Globally Transparent Model

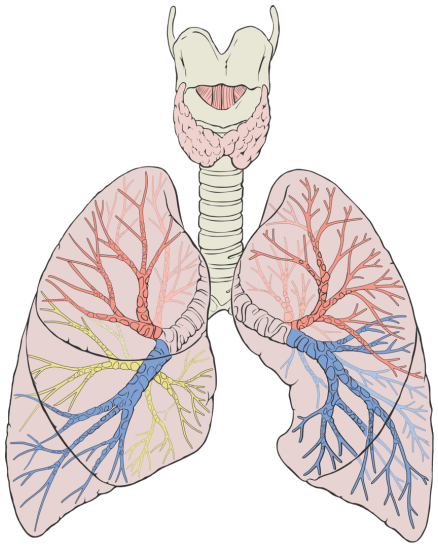


 × 2019 IBM Corporation

Causality

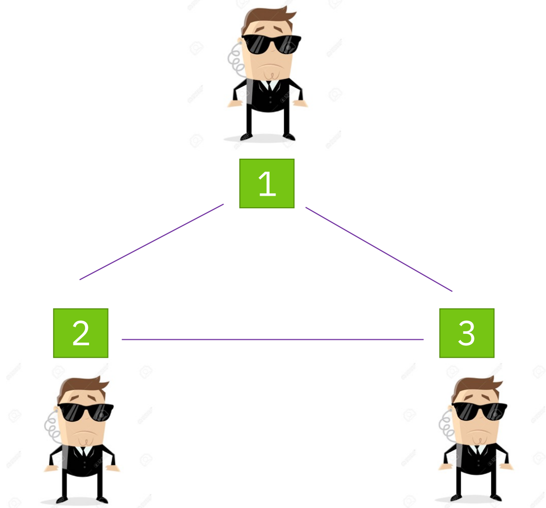
What is the true cause for an event? Interpretability methods can be used to identify where to look (reduce search space) before causal methods are applied.

?



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.



 © 2019 IBM Corporation

Summary and Future Directions

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

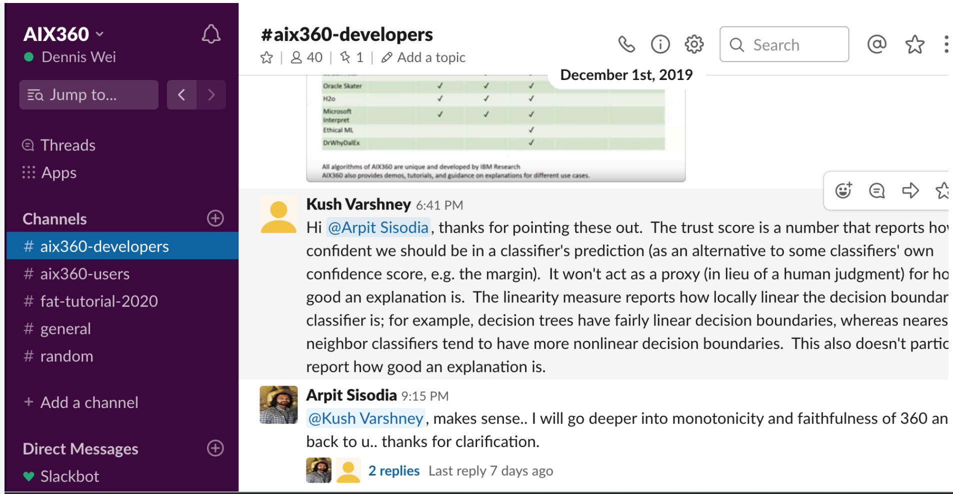
The future of AIX360 is people like you!

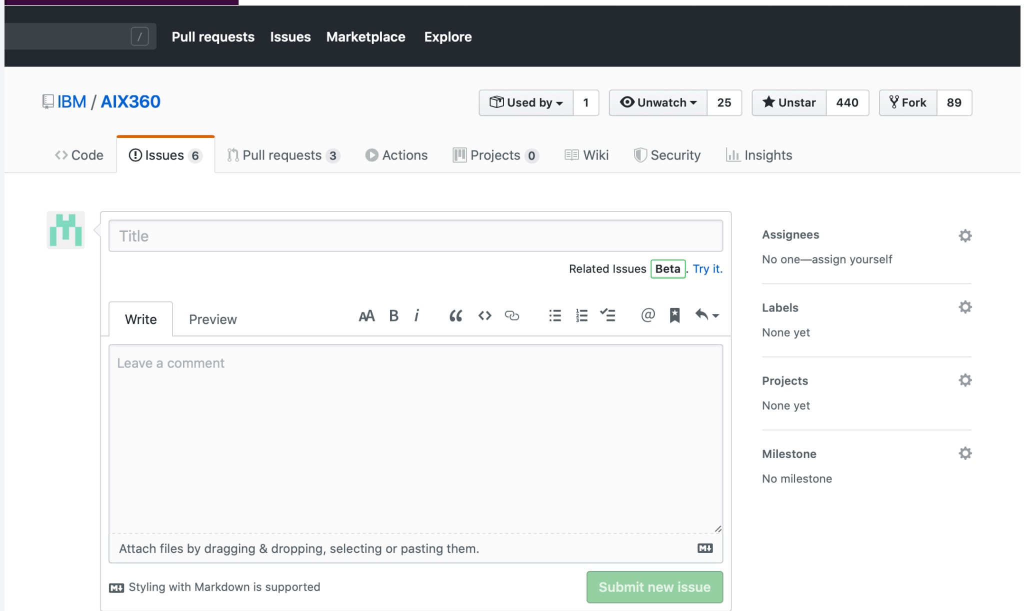
 × 2019 IBM Corporation

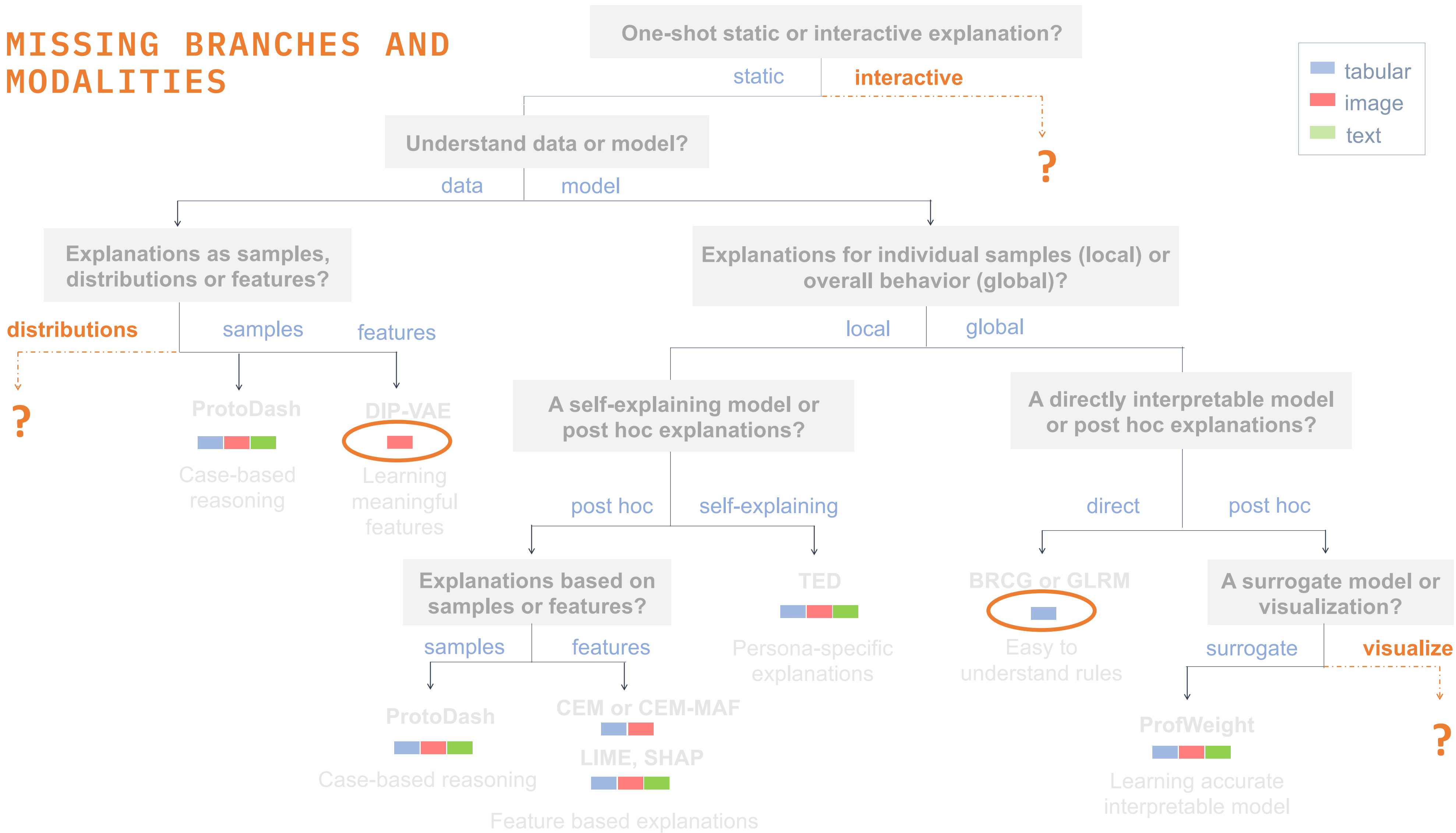
CONTRIBUTING TO AIX360

Want to contribute?

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



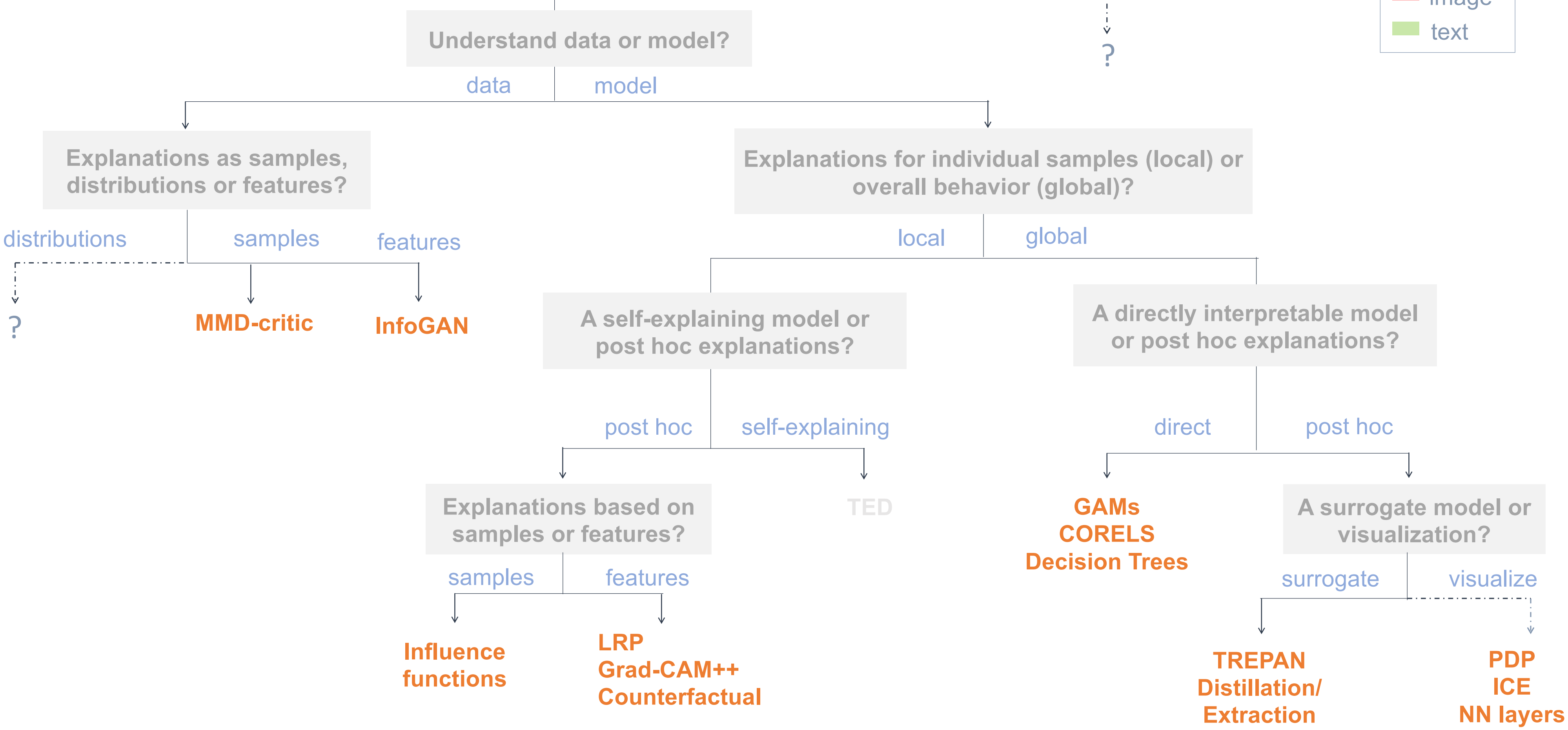




MISSING BRANCHES AND MODALITIES
One-shot static or interactive explanation?
static
interactive
?
tabular
image
text
Understand data or model?
data
model
Explanations as samples, distributions or features?
Explanations for individual samples (local) or overall behavior (global)?
distributions
samples
features
local
global
?
ProtoDash
Case-based reasoning
DIP-VAE
Learning meaningful features
A self-explaining model or post hoc explanations?
A directly interpretable model or post hoc explanations?
post hoc
self-explaining
direct
post hoc
Explanations based on samples or features?
TED
Persona-specific explanations
BRCG or GLRM
Easy to understand rules
A surrogate model or visualization?
samples
features
surrogate
visualize
ProtoDash
Case-based reasoning
CEM or CEM-MAF
LIME, SHAP
Feature based explanations
ProfWeight
Learning accurate interpretable model
?

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

 × 2019 IBM Corporation