IBM Research AI Explainability 360 Toolkit

Vijay Arya, Rachel Bellamy, Pin-Yu Chen,
Payel Das, Amit Dhurandhar, MaryJo Fitzgerald,
Michael Hind, Samuel Hoffman,
Stephanie Houde, Vera Liao, Ronny Luss,
Sameep Mehta, Saska Mojsilovic, Sami Mourad,
Pablo Pedemonte, John Richards,
Prasanna Sattigeri, Moninder Singh,
Karthikeyan Shanmugam, Kush Varshney,
Dennis Wei, Yunfeng Zhang, Ramya Raghavendra

WHAT DOES IT TAKE TO TRUST AI DECISIONS? (BEYOND ACCURACY)

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

Credit

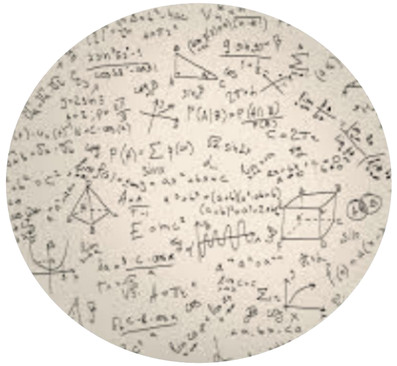
Employment

Admission

Sentencing

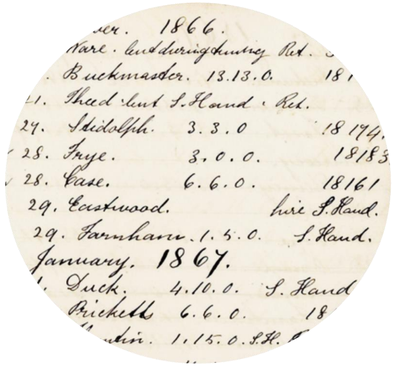
Healthcare

Is it fair?



Is it easy to understand?

Did anyone tamper with it?



Is it accountable?

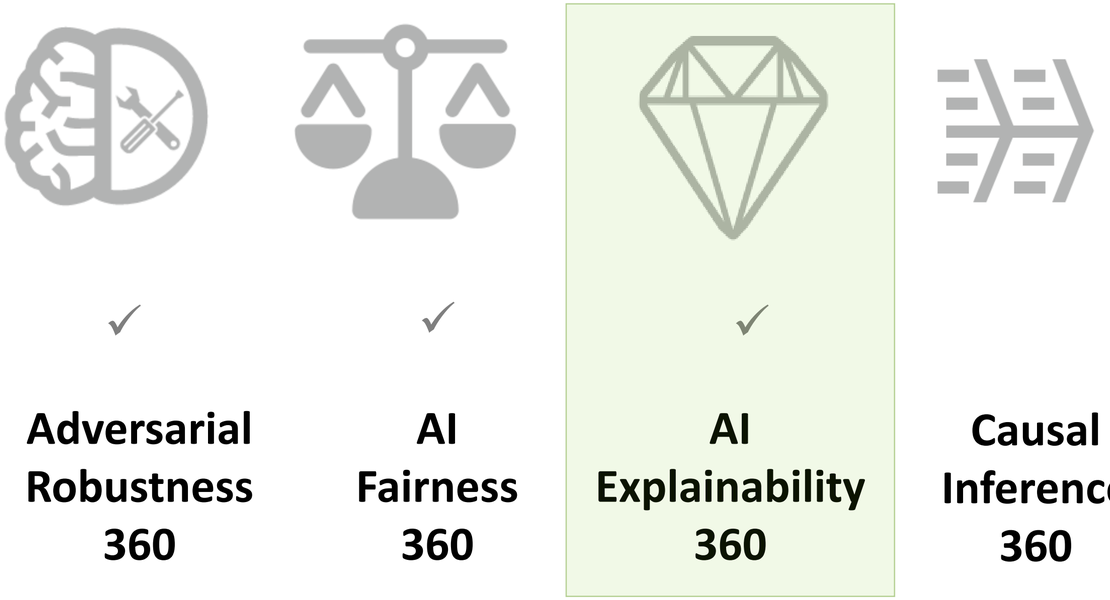
AIX360: IBM RESEARCH AI EXPLAINABILITY 360 TOOLKIT

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	8 innovations to explain data and AI models
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.

AIX360: DIFFERENT WAYS TO EXPLAIN

One explanation does not fit all

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

End users/customers (trust)

Doctors: *Why did you recommend this treatment?*

Customers: *Why was my loan denied?*

Teachers: *Why was my teaching evaluated in this way?*

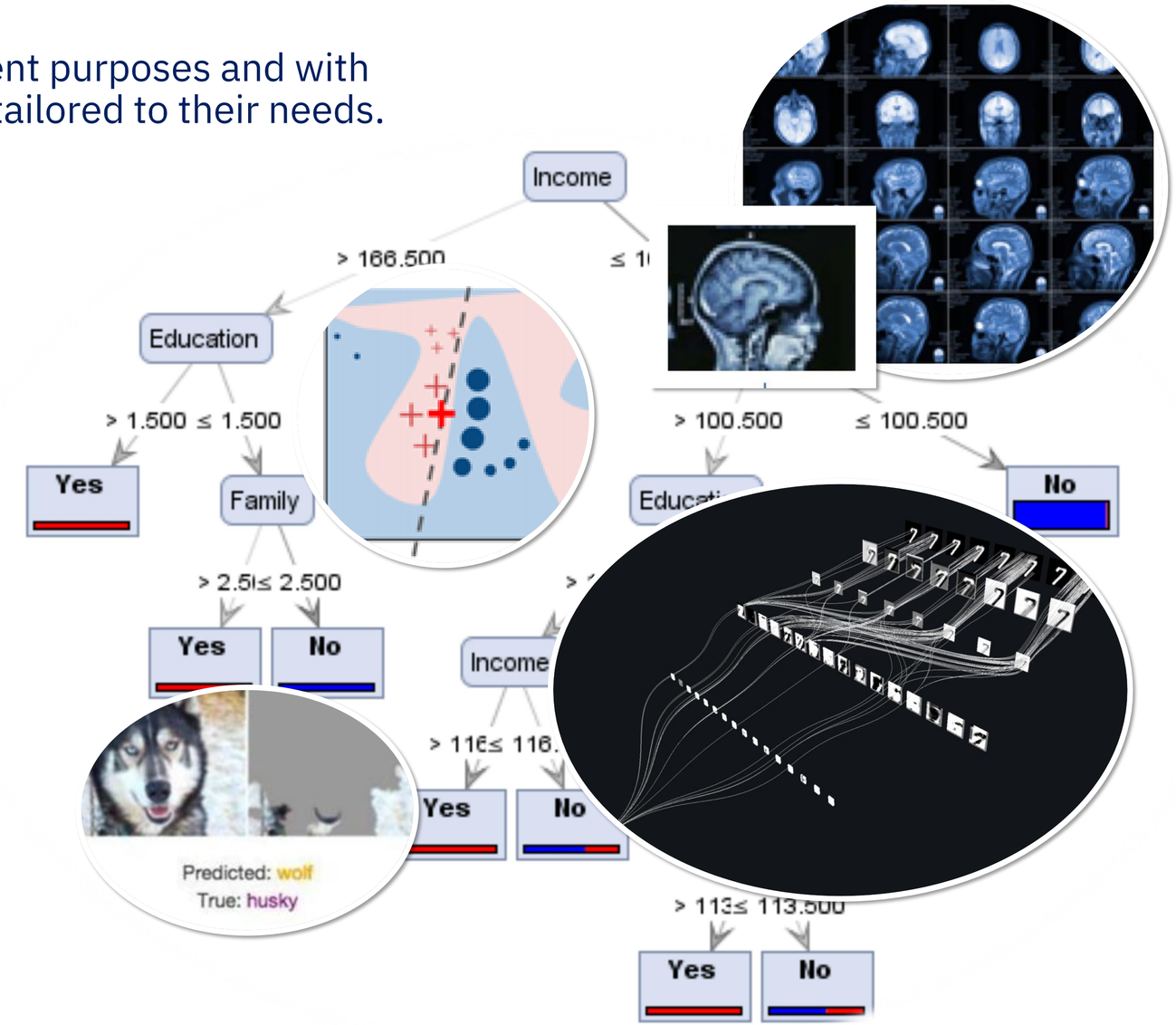
Gov't/regulators (compliance, safety)

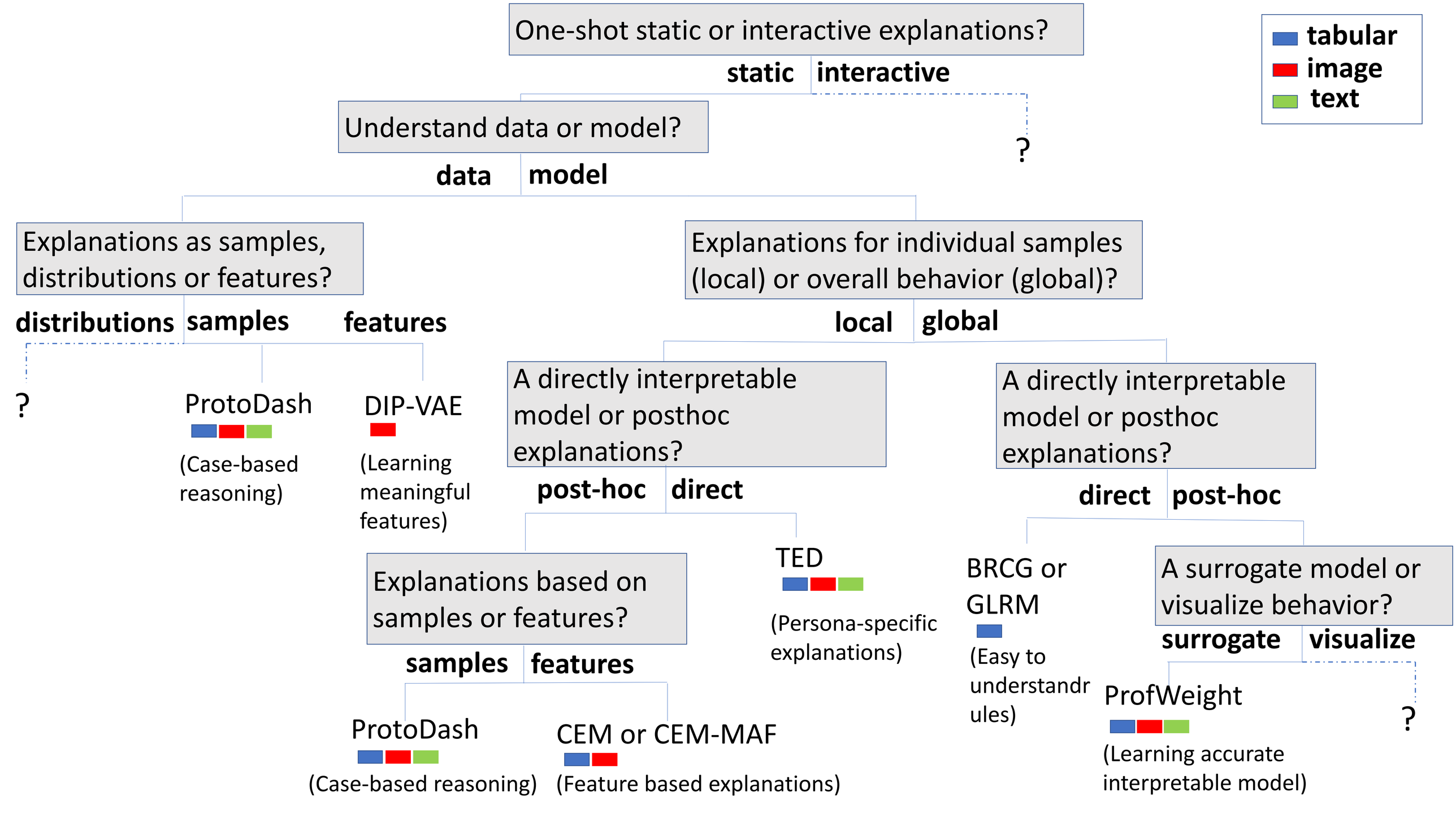
Prove to me that you didn't discriminate.

Developers (quality, "debuggability")

Is our system performing well?

How can we improve it?




One-shot static or interactive explanations?
static
interactive
?
tabular
image
text
Understand data or model?
data
model
Explanations as samples, distributions or features?
distributions
samples
features
?
ProtoDash
(Case-based reasoning)
DIP-VAE
(Learning meaningful features)
Explanations for individual samples (local) or overall behavior (global)?
local
global
A directly interpretable model or posthoc explanations?
post-hoc
direct
Explanations based on samples or features?
samples
features
ProtoDash
(Case-based reasoning)
CEM or CEM-MAF
(Feature based explanations)
TED
(Persona-specific explanations)
A directly interpretable model or posthoc explanations?
direct
post-hoc
BRCG or GLRM
(Easy to understandr ules)
A surrogate model or visualize behavior?
surrogate
visualize
ProfWeight
(Learning accurate interpretable model)
?

AIX360: COMPETITIVE LANDSCAPE

Toolkit	Data Explanations	Directly Interpretable	Local Post-hoc	Global Post-hoc	Custom Explanation	Metrics
IBM AIX360	2	2	3	1	1	2
Seldon Alibi			✓	✓		
Oracle Skater		✓	✓	✓		
H2o		✓	✓	✓		
Microsoft Interpret		✓	✓	✓		
Ethical ML				✓		
DrWhyDalEx				✓		

All algorithms of AIX360 are unique and developed by IBM Research
AIX360 also provides demos, tutorials, and guidance on explanations for different use cases.

OUTLINE

Demo

FICO demo: Showcase how the toolkit can serve different types of users for a credit approval application.

Tutorial Notebook

FICO Tutorial: Show the notebook that implements methods to create diverse explanations.

Questions

?

