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

Global Taxonomy of Interpretable AI
AI4Media Workshop
Apr 29th, 2021

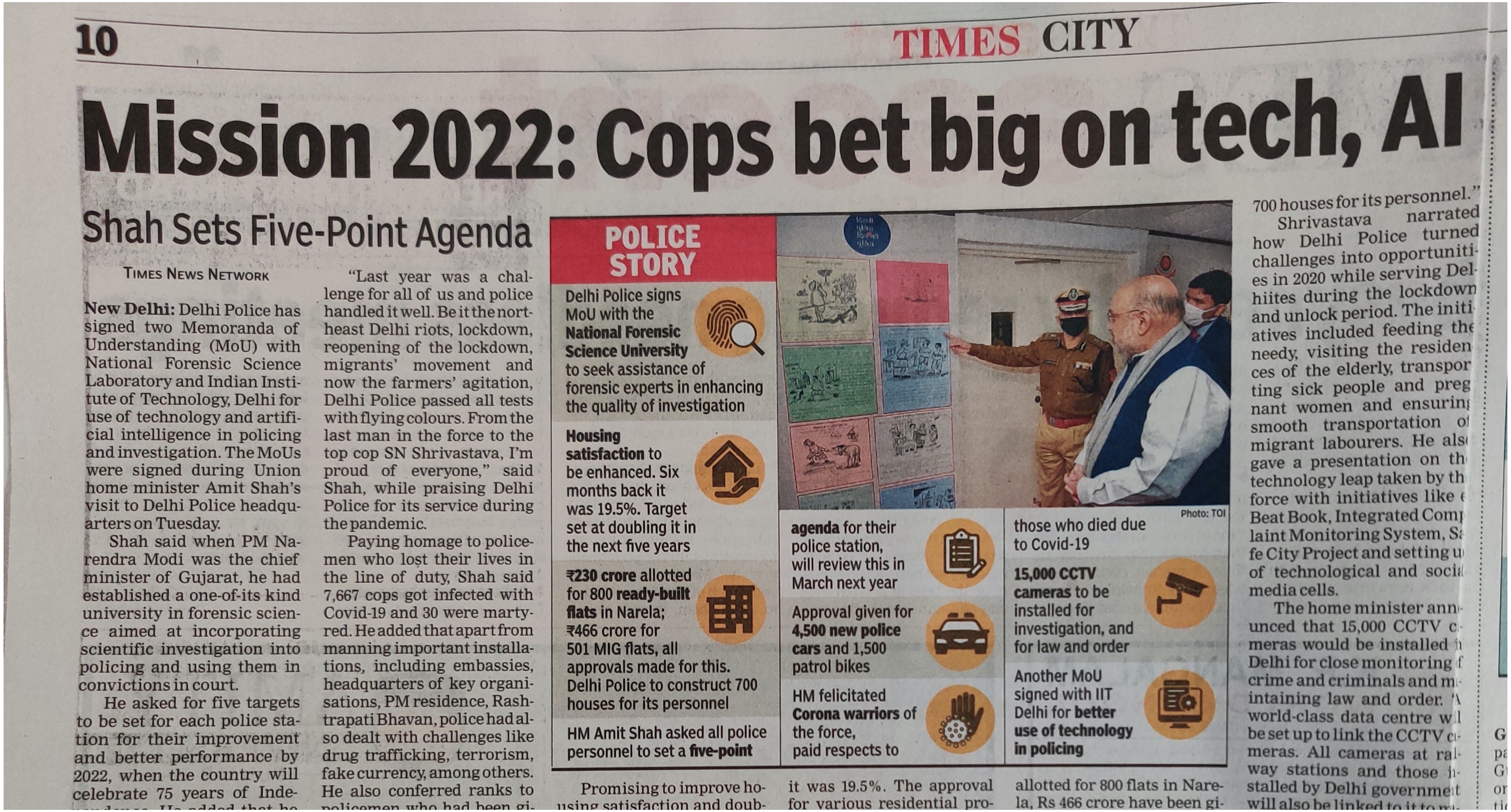
Vijay Arya
IBM Research
vijay.arya@in.ibm.com

Acknowledgments to colleagues at IBM Research and open-source contributors

AGENDA

- Introduction to Explainable AI
- AIX360 Toolkit
- Demo

NEWS THIS MORNING



Mission 2022: Cops bet big on tech, AI

Shah Sets Five-Point Agenda

TIMES NEWS NETWORK

New Delhi: Delhi Police has signed two Memoranda of Understanding (MoU) with National Forensic Science Laboratory and Indian Institute of Technology, Delhi for use of technology and artificial intelligence in policing and investigation. The MoUs were signed during Union home minister Amit Shah's visit to Delhi Police headquarters on Tuesday.

Shah said when PM Narendra Modi was the chief minister of Gujarat, he had established a one-of-its kind university in forensic science aimed at incorporating scientific investigation into policing and using them in convictions in court.

He asked for five targets to be set for each police station for their improvement and better performance by 2022, when the country will celebrate 75 years of Independence. He added that he

"Last year was a challenge for all of us and police handled it well. Be it the northeast Delhi riots, lockdown, reopening of the lockdown, migrants' movement and now the farmers' agitation, Delhi Police passed all tests with flying colours. From the last man in the force to the top cop SN Shrivastava, I'm proud of everyone," said Shah, while praising Delhi Police for its service during the pandemic.

Paying homage to policemen who lost their lives in the line of duty, Shah said 7,667 cops got infected with Covid-19 and 30 were martyred. He added that apart from manning important installations, including embassies, headquarters of key organisations, PM residence, Rashtrapati Bhavan, police had also dealt with challenges like drug trafficking, terrorism, fake currency, among others. He also conferred ranks to policemen who had been gi-

POLICE STORY

Delhi Police signs MoU with the **National Forensic Science University** to seek assistance of forensic experts in enhancing the quality of investigation

Housing satisfaction to be enhanced. Six months back it was 19.5%. Target set at doubling it in the next five years

₹230 crore allotted for 800 **ready-built flats** in Narela; ₹466 crore for 501 MIG flats, all approvals made for this. Delhi Police to construct 700 houses for its personnel

HM Amit Shah asked all police personnel to set a **five-point agenda** for their police station, will review this in March next year

Approval given for **4,500 new police cars** and 1,500 patrol bikes

HM felicitated **Corona warriors** of the force, paid respects to those who died due to Covid-19

15,000 CCTV cameras to be installed for investigation, and for law and order

Another MoU signed with IIT Delhi for **better use of technology in policing**

Photo: TOI

Promising to improve housing satisfaction and doub-

it was 19.5%. The approval for various residential pro-

allotted for 800 flats in Narela, Rs 466 crore have been gi-

700 houses for its personnel."

Shrivastava narrated how Delhi Police turned challenges into opportunities in 2020 while serving Delhiites during the lockdown and unlock period. The initiatives included feeding the needy, visiting the residences of the elderly, transporting sick people and pregnant women and ensuring smooth transportation of migrant labourers. He also gave a presentation on the technology leap taken by the force with initiatives like e-Beat Book, Integrated Complaint Monitoring System, Safe City Project and setting up of technological and social media cells.

The home minister announced that 15,000 CCTV cameras would be installed in Delhi for close monitoring of crime and criminals and maintaining law and order. 'A world-class data centre will be set up to link the CCTV cameras. All cameras at railway stations and those installed by Delhi government will also be linked to it to m-

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

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

Credit

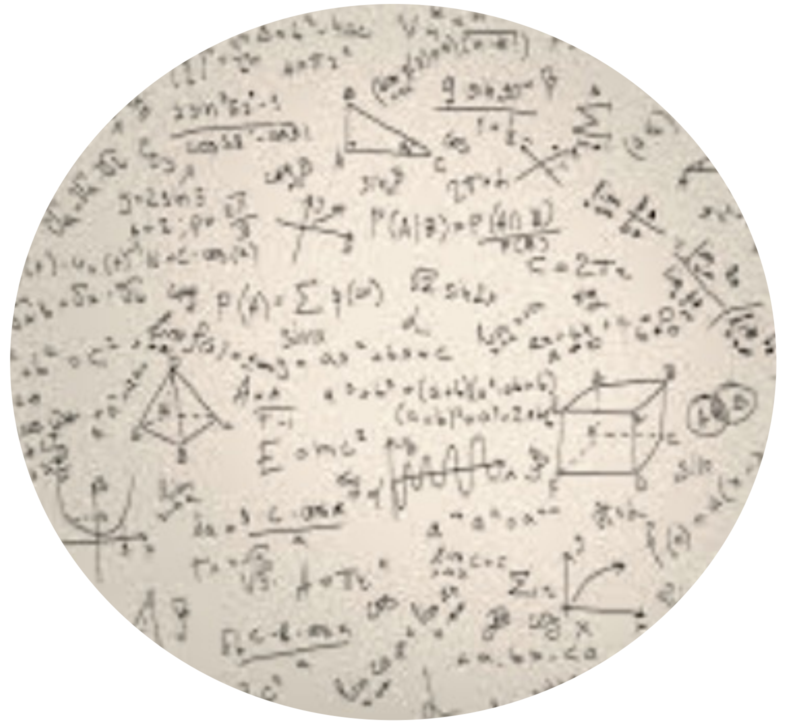
Employment

Admission

Sentencing

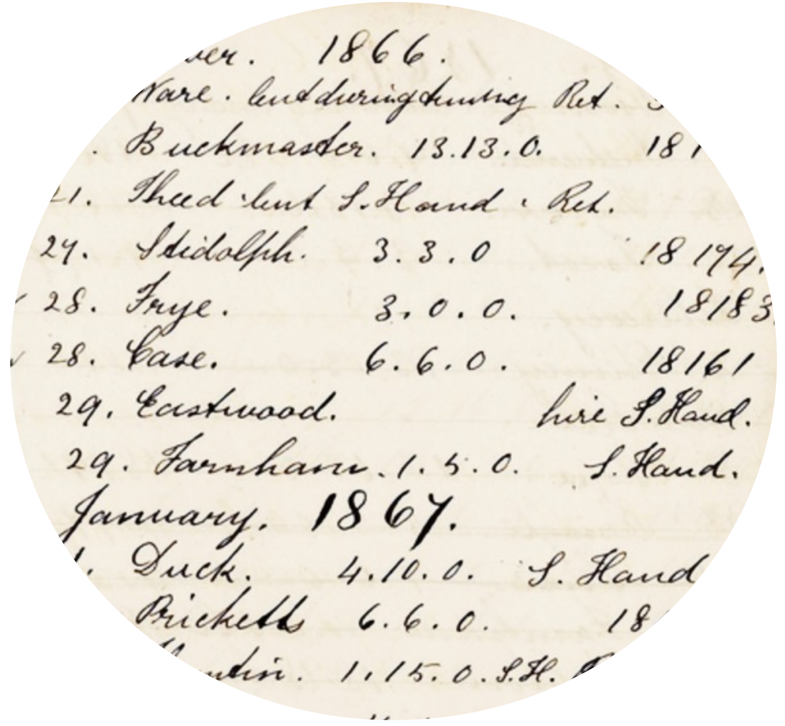
Healthcare

Is it fair?



Is it easy to understand?

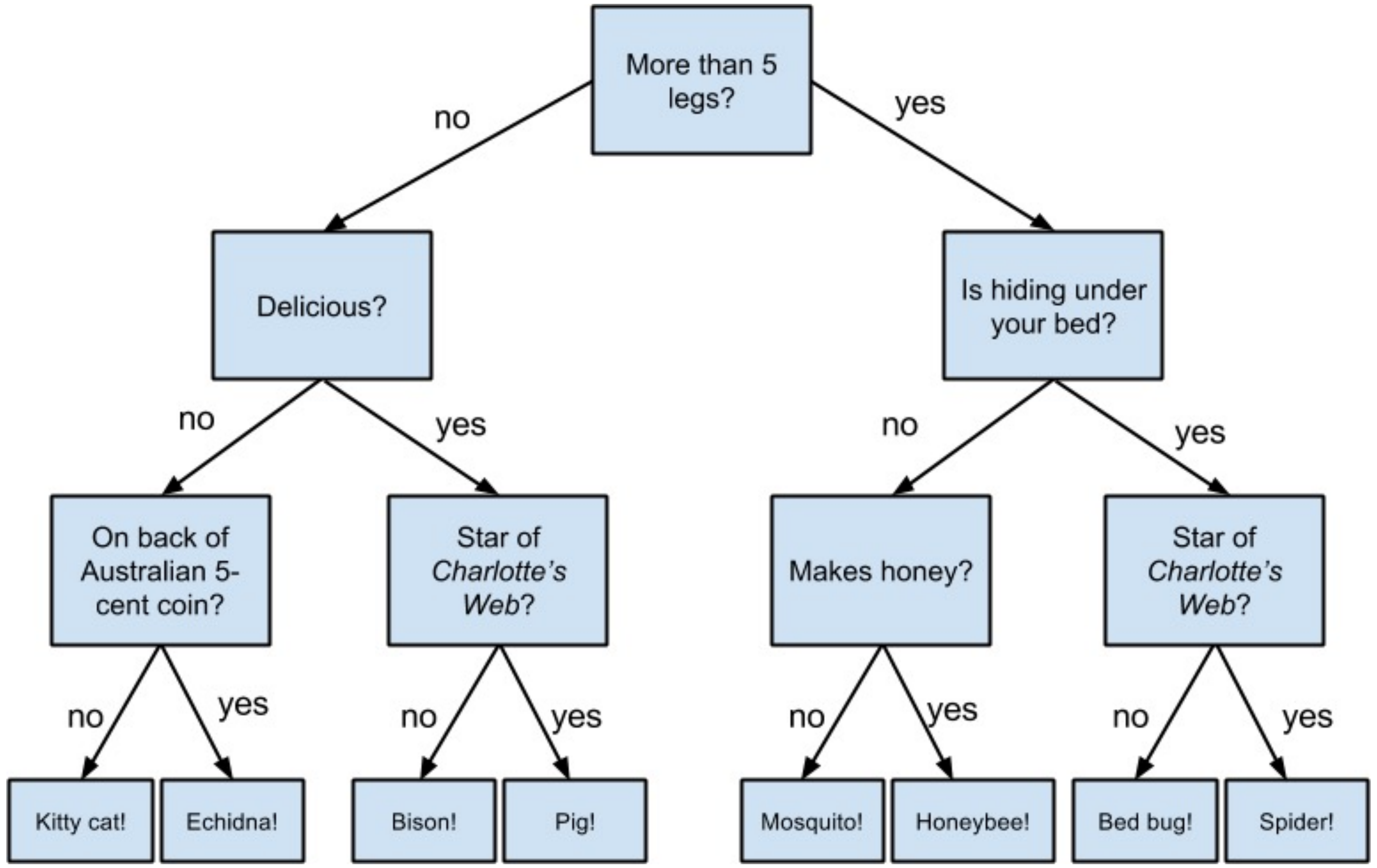
Did anyone tamper with it?



Is it accountable?

WHY EXPLAINABLE AI?

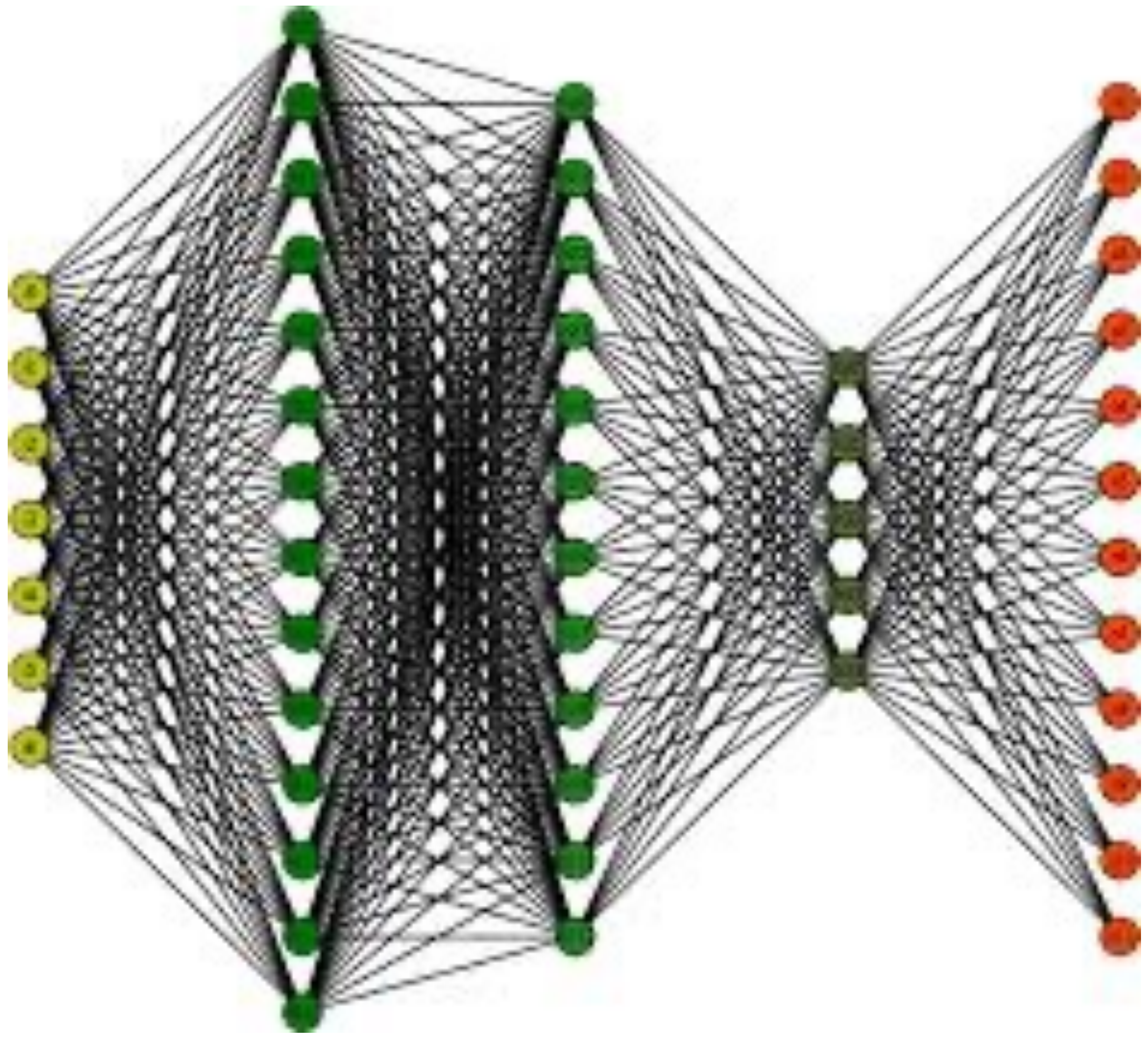
Decision Tree



Interpretable?

YES

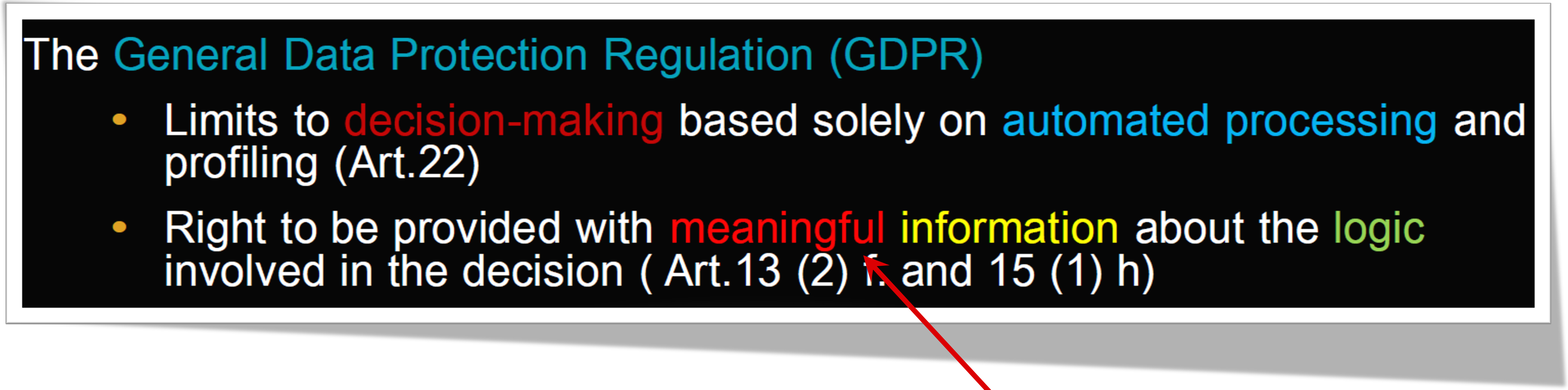
Neural Network



Interpretable?

NO

Regulations



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” ???

Related applications:

- Auditing models in regulated industries for bias, compliance, risk, etc.

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

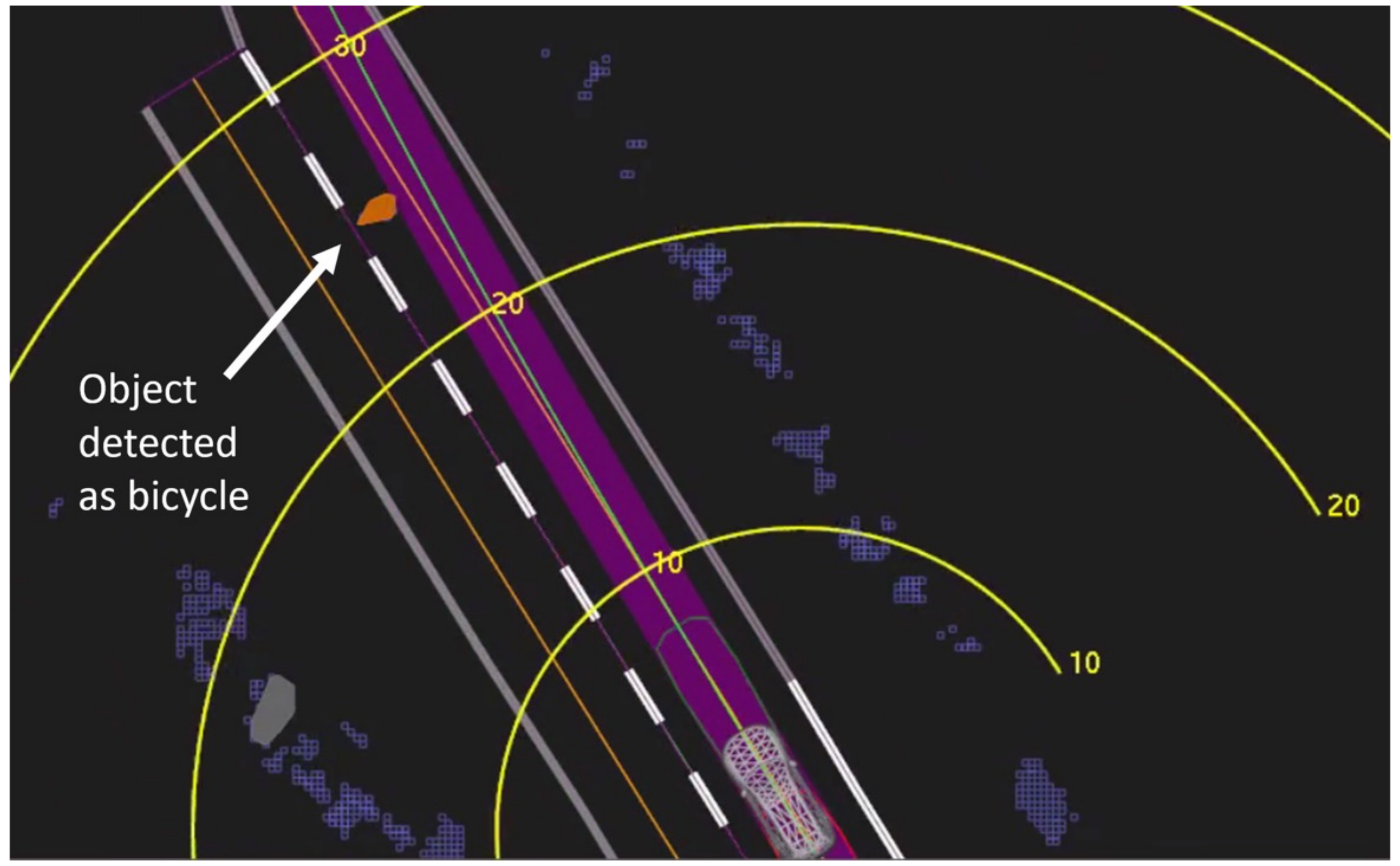
WHY EXPLAINABLE AI? (CONTINUED)

Debugging

Can help to understand what is wrong with a system / improve its performance.

Accident involving a pedestrian cyclist

Self driving car didn’t slow down even though the sensors detected the cyclist. why?



https://en.wikipedia.org/wiki/Death_of_Elaine_Herzberg

Existence of Confounders

Can help to identify spurious correlations.

Pneumonia + Asthma

Low-Risk (Treat as outpatient)
(NN with 86% accuracy)

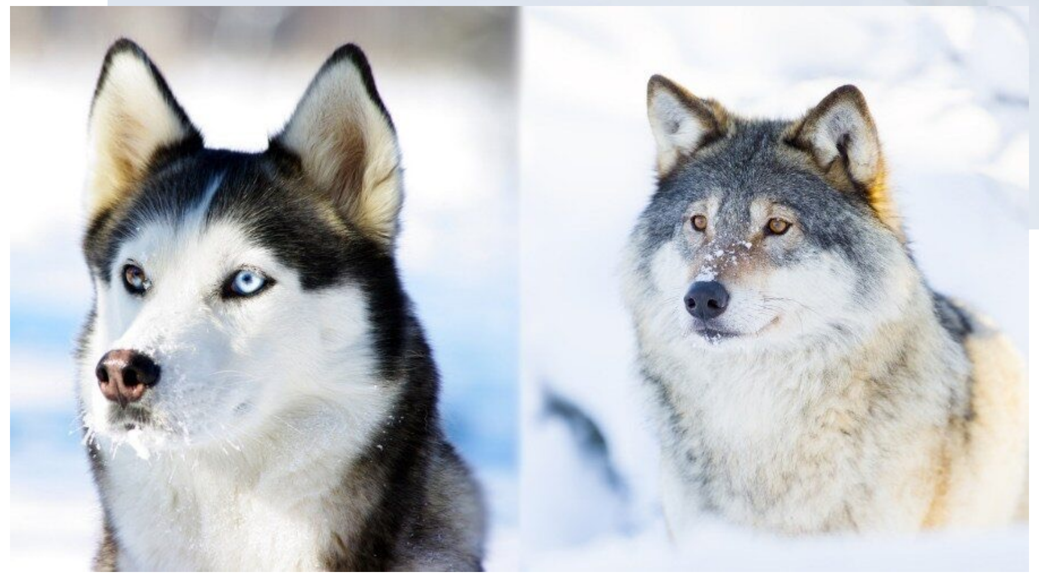
i.e., patients with pneumonia and history of asthma have lower risk of dying from pneumonia than the general population.

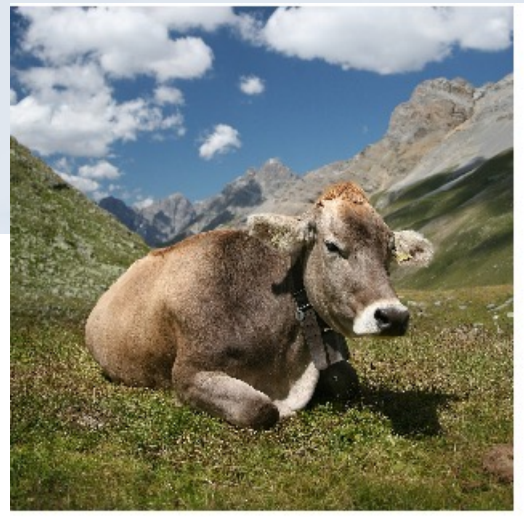
(Rich Caruana, et al)

WHY EXPLAINABLE AI? (CONTINUED)

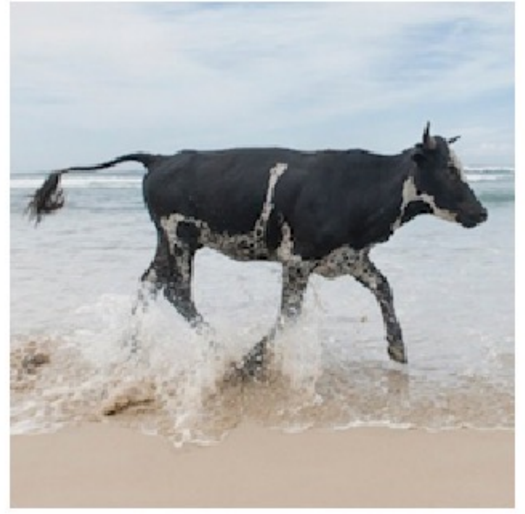
Robustness and Generalizability

Is the system basing decisions on the correct features?





(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98



(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(Berry, et al Caltech, ClarifAI.com)

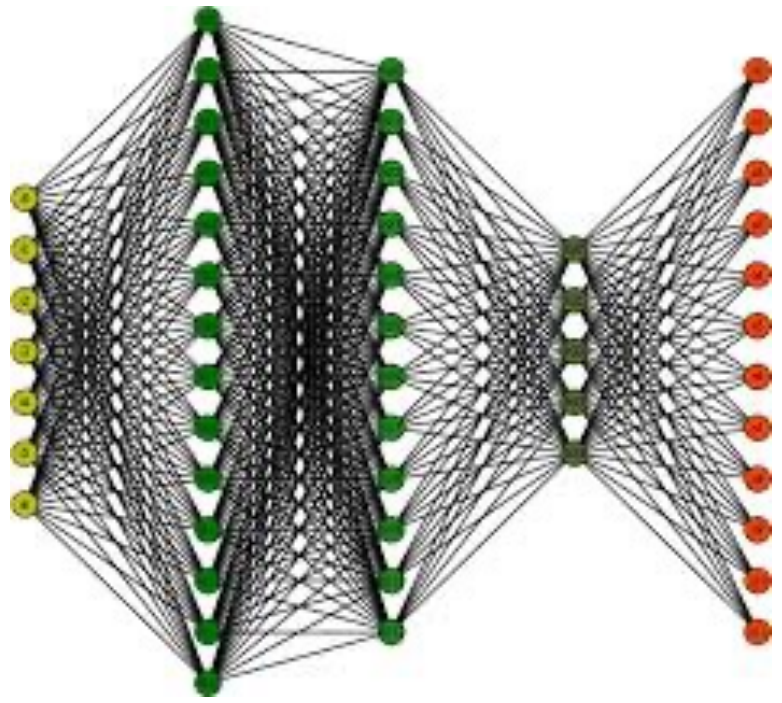
Is the decision-making system fair?

Widespread Adoption

WHY EXPLAINABLE AI? (CONTINUED)

Simplification

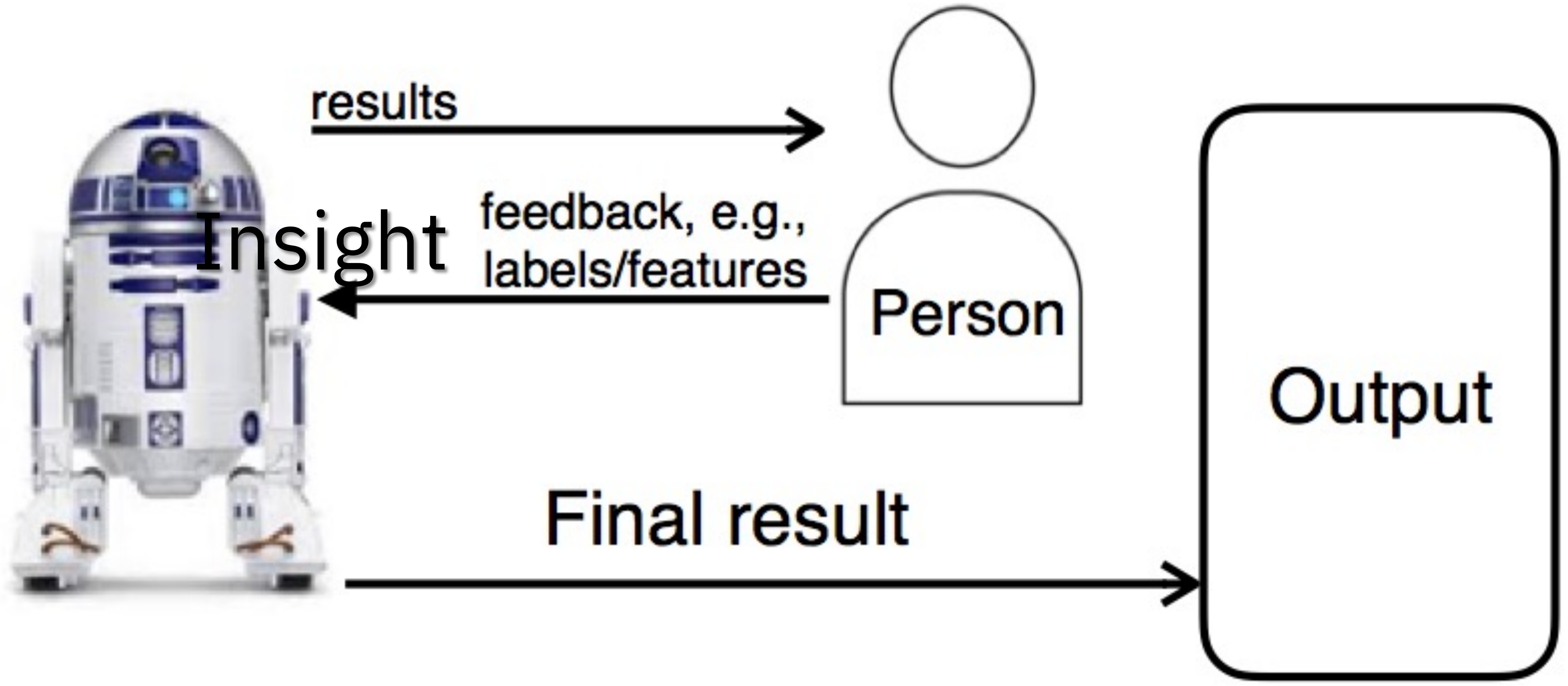
Understanding what’s truly happening can help build simpler systems.



Check if code has comments

Enhance Performance

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



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.

Affected users

"Why was my loan denied? How can I be approved?"

Who: Patients, accused, loan applicants, teachers

Why: understanding of factors

End Domain users

"Why did you recommend this treatment?"

Who: Physicians, judges, loan officers, teacher evaluators

Why: trust/confidence, insights(?)

Regulatory bodies

"Prove that your system didn't discriminate."

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

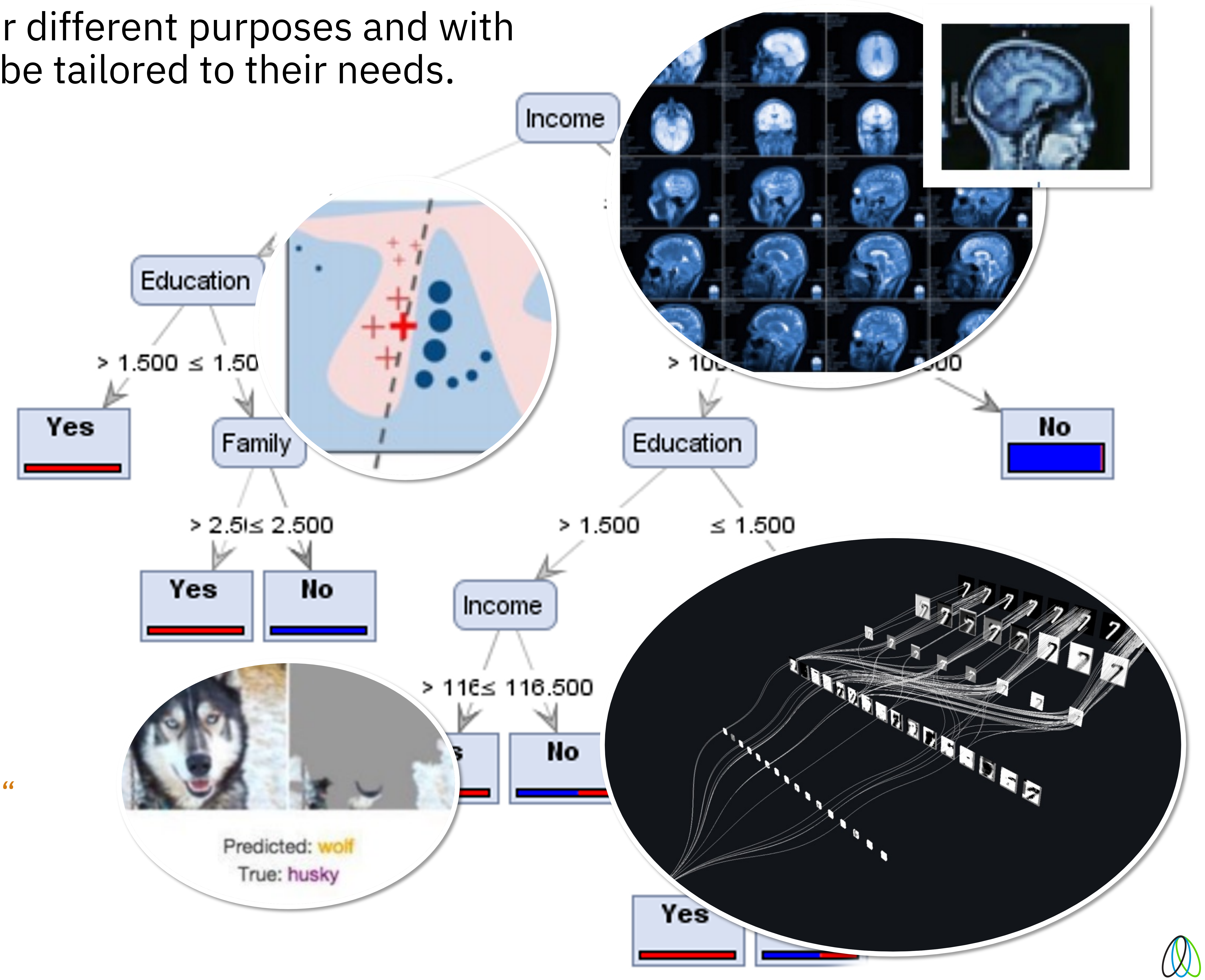
Why: ensure fairness for constituents

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



AIX360: AI EXPLAINABILITY 360 TOOLKIT

(LINUX FOUNDATION AI, JMLR)

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.

AI Explainability 360	
Explainability Algorithms	10 ways to explain data and AI models
Repositories	github.com/Trusted-AI/AIX360
Interactive Experience	aix360.mybluemix.net/data
API	aix360.readthedocs.io
Tutorials	> 13 tutorial notebooks (finance, healthcare, lifestyle, Attrition, etc.)
Developers	> 15 Researchers, Software engineers across US, 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'

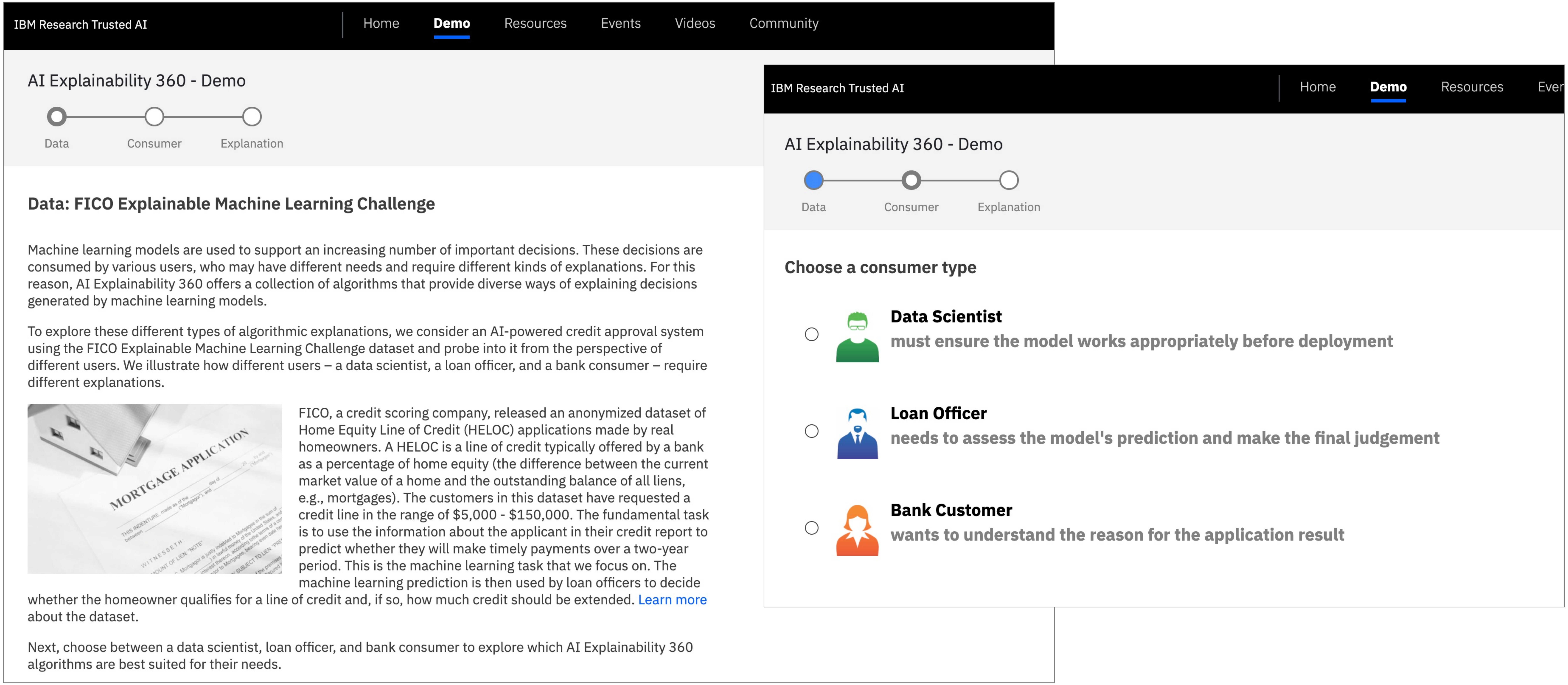
CIO JOURNAL.

Companies Grapple With AI's Opaque Decision-Making Process

THE WALL STREET JOURNAL.

DEMO

http://aix360.mybluemix.net/data



AI Explainability 360 - Demo

Data: FICO Explainable Machine Learning Challenge

Machine learning models are used to support an increasing number of important decisions. These decisions are consumed by various users, who may have different needs and require different kinds of explanations. For this reason, AI Explainability 360 offers a collection of algorithms that provide diverse ways of explaining decisions generated by machine learning models.

To explore these different types of algorithmic explanations, we consider an AI-powered credit approval system using the FICO Explainable Machine Learning Challenge dataset and probe into it from the perspective of different users. We illustrate how different users – a data scientist, a loan officer, and a bank consumer – require different explanations.

FICO, a credit scoring company, released an anonymized dataset of Home Equity Line of Credit (HELOC) applications made by real homeowners. A HELOC is a line of credit typically offered by a bank as a percentage of home equity (the difference between the current market value of a home and the outstanding balance of all liens, e.g., mortgages). The customers in this dataset have requested a credit line in the range of $5,000 - $150,000. The fundamental task is to use the information about the applicant in their credit report to predict whether they will make timely payments over a two-year period. This is the machine learning task that we focus on. The machine learning prediction is then used by loan officers to decide whether the homeowner qualifies for a line of credit and, if so, how much credit should be extended. Learn more about the dataset.

Next, choose between a data scientist, loan officer, and bank consumer to explore which AI Explainability 360 algorithms are best suited for their needs.

AI Explainability 360 - Demo

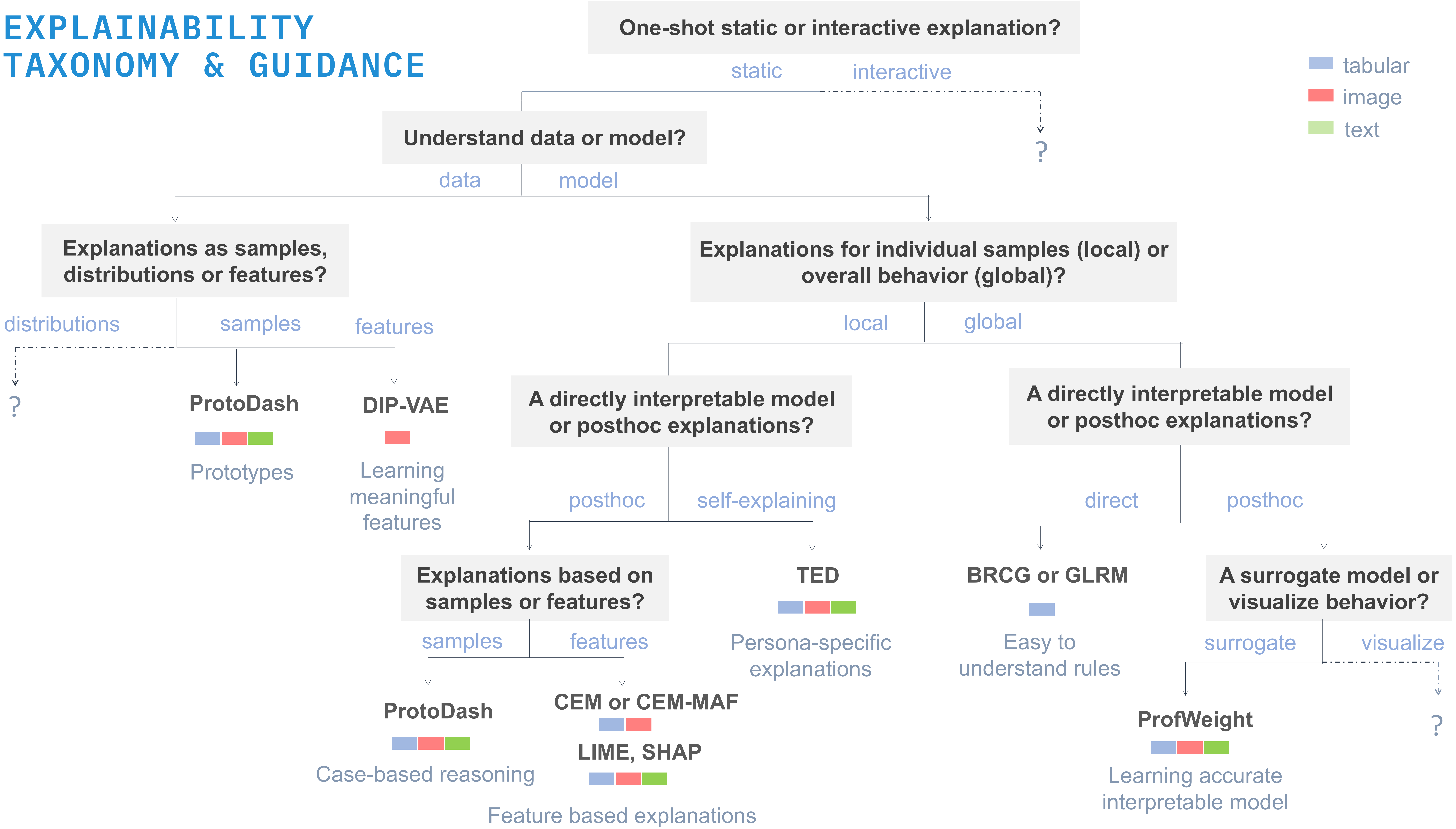
Choose a consumer type

○ **Data Scientist**
must ensure the model works appropriately before deployment

○ **Loan Officer**
needs to assess the model's prediction and make the final judgement

○ **Bank Customer**
wants to understand the reason for the application result

EXPLAINABILITY TAXONOMY & GUIDANCE



🟦 tabular
🟥 image
🟩 text

One-shot static or interactive explanation?

static

interactive

?

Understand data or model?

data

model

Explanations as samples, distributions or features?

distributions

?

samples

ProtoDash
🟦🟥🟩
Prototypes

features

DIP-VAE
🟥
Learning meaningful features

Explanations for individual samples (local) or overall behavior (global)?

local

A directly interpretable model or posthoc explanations?

posthoc

Explanations based on samples or features?

samples

ProtoDash
🟦🟥🟩
Case-based reasoning

features

CEM or CEM-MAF
🟦🟥

LIME, SHAP
🟦🟥🟩

Feature based explanations

self-explaining

TED
🟦🟥🟩
Persona-specific explanations

global

A directly interpretable model or posthoc explanations?

direct

BRCG or GLRM
🟦
Easy to understand rules

posthoc

A surrogate model or visualize behavior?

surrogate

ProfWeight
🟦🟥🟩
Learning accurate interpretable model

visualize

?

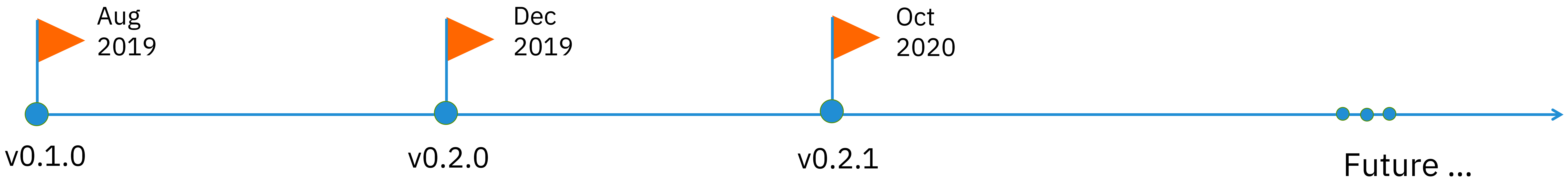
EXPLAINABILITY OPENSOURCE LANDSCAPE

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

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

AIX360: ROADMAP

THANK YOU



Aug 2019

v0.1.0

- 8 explainability algorithms
- 2 metrics
- Demos
- Tutorial notebooks
- Taxonomy
- PyPI, API (readthedocs)

Dec 2019

v0.2.0

- + LIME
- + SHAP
- (10 explainability algorithms)
- Dependency updates
- Bug fixes, closed PRs

Oct 2020

v0.2.1

- License updates
- Minor update to CEM
- FeatureBinarizerFromTrees for Directly interpretable explainers
- Minor updates to BRCG due to Pandas update
- Updates to Heloc tutorial
- Abstraction class for global black box
- Minor bug fixes, comment updates, closed issues/PRs, etc.
- PyPI package updated

Future ...

- Frameworks
- Algorithms
- Metrics
- Conda packaging
- Automated testing of notebooks
- Community contributions are welcome

Metric	Value
Forks	151
Stars	691
Github clones/day (last 14-day avg)	2.9
Github visits/day (last 14-day avg)	145
PyPI downloads/month	823
AIX360 Slack users	190
Closed PRs	67
Public presentations/tutorial views	4023

Current explainability algorithms in AIX360

Data Explanations	Directly Interpretable	Local Post-hoc	Global Post-hoc	Custom Explanation
2	2	5	1	1