Explainable ML Models

~ Krishna Balaga
krbalaga@in.ibm.com
Deep Learning Developer Advocate

94%

of companies believe AI is key to competitive advantage .

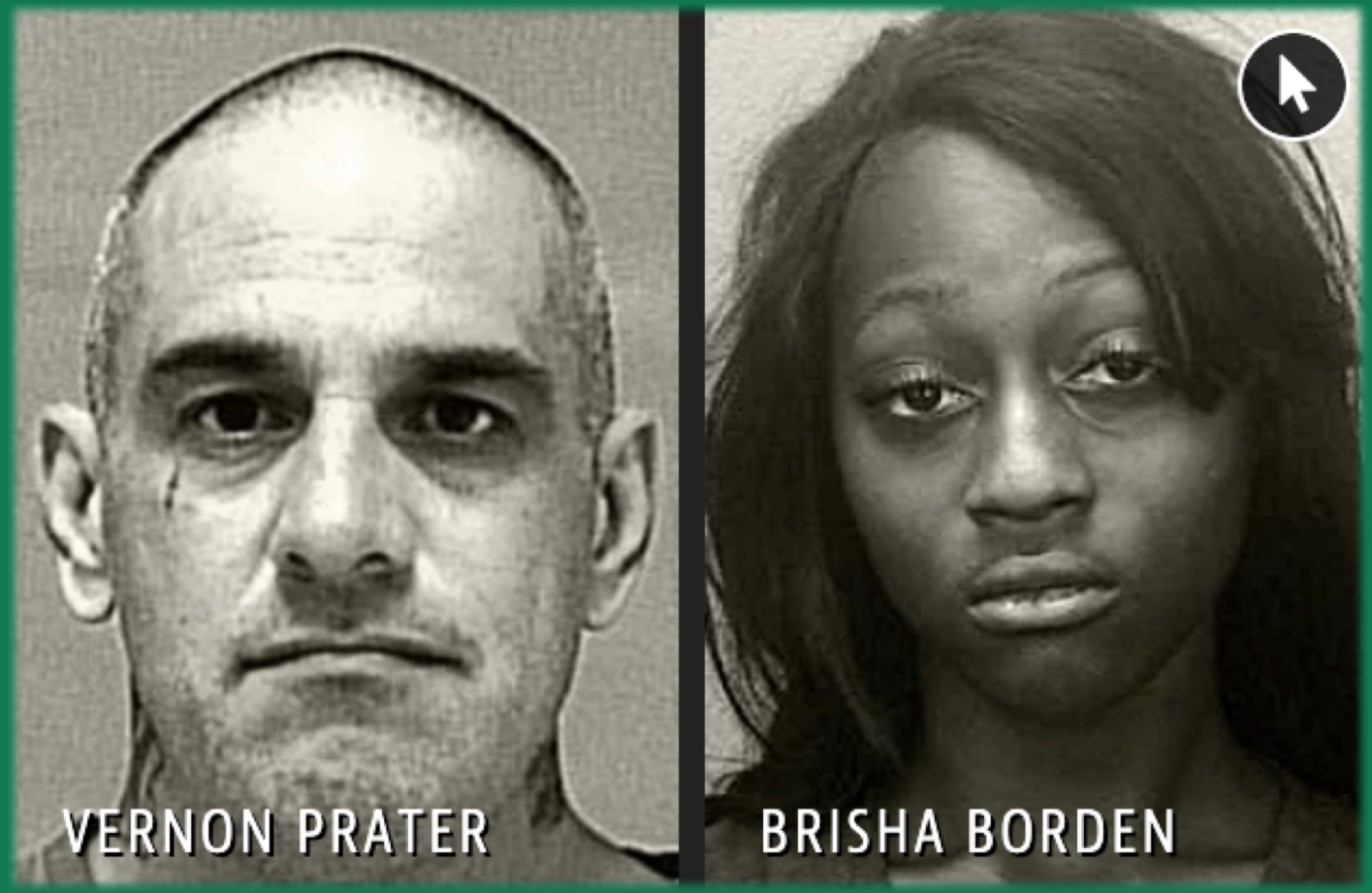
– IDC

1 in 20

companies have extensively incorporated AI in offerings or processes.

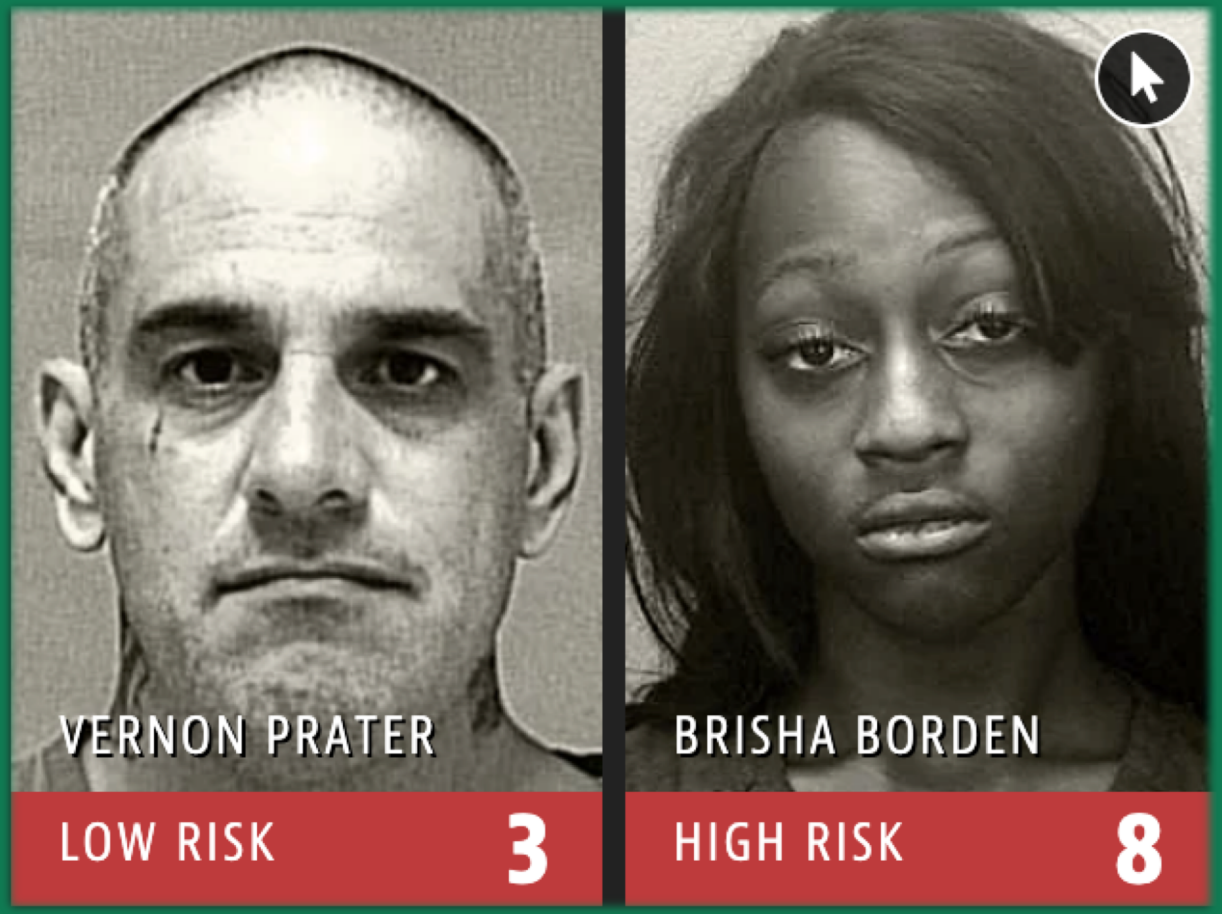
– MIT Sloan Management Review

What Happened So Far?



What Happened So Far?

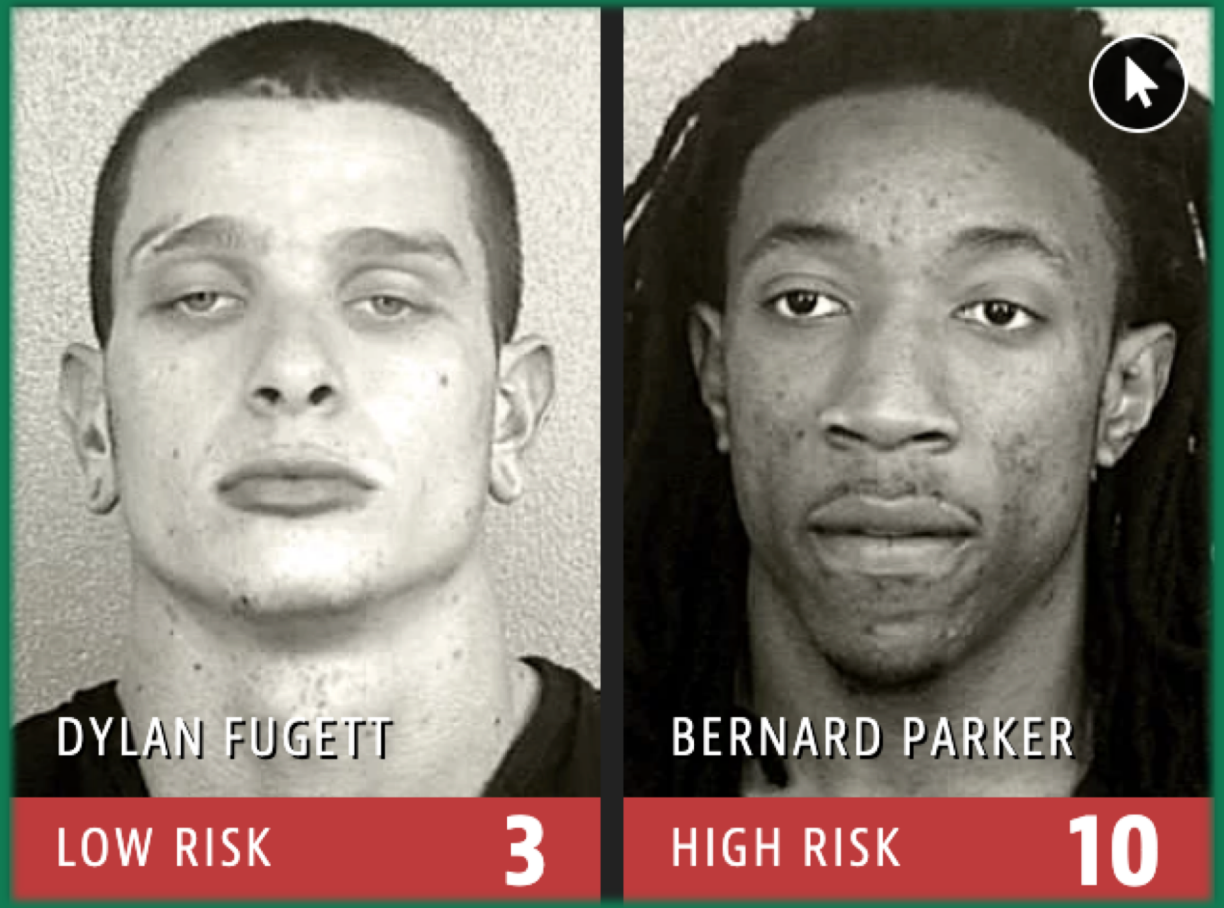
What Happened So Far?



COMPAS

(Correctional Offender Management Profiling for Alternative Sanctions)

What Happened So Far?



What Happened So Far?

Why Am I getting this Decision?

BlackBox AI

Business Owner

Can I Trust our AI Decisions?

Customer Support

How do I answer this complaint?

IT & Operations

How do I monitor the model?

Data Scientist

Is this the Best Model?

Internal Audit

Are these decisions Fair?

The Mighty Karna of Enterprise Deployments

ADVERSARY

/ˈadvəs(ə)ri/

(Noun)

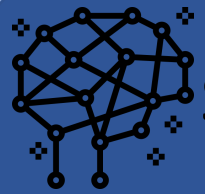
One's opponent in a contest, conflict, or dispute.

The Third Wave of Intelligence

Symbolic AI

Logic and Rules represent Knowledge

No Learning capability and poor handling of uncertainty



Statistical AI

Statistical models for Domain Training

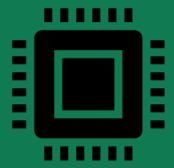
No Contextual capability and absolutely minimal explainability

Explainable AI

Systems build exp0lanatory models

Systems dynamically learn and reason with new tasks

Factors driving rapid advancements of AI



GPUs,On-Chip Neural Nets

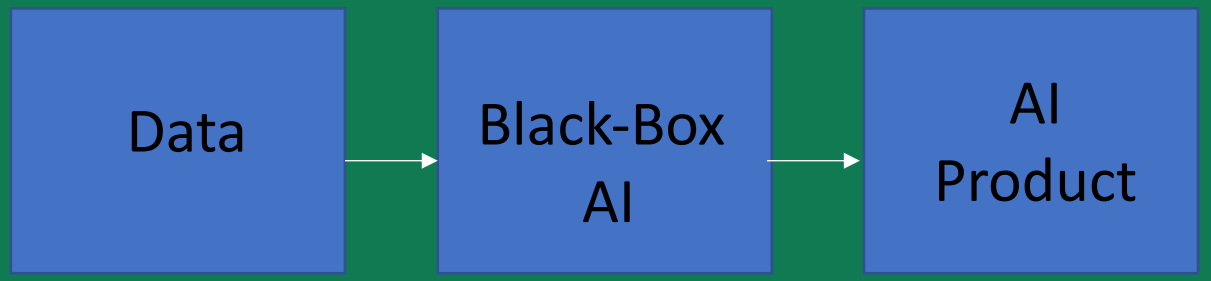
Big-Data Availability

Cloud compute Infrastructure

New Algorithms

What is Explainable AI?

Black Box AI

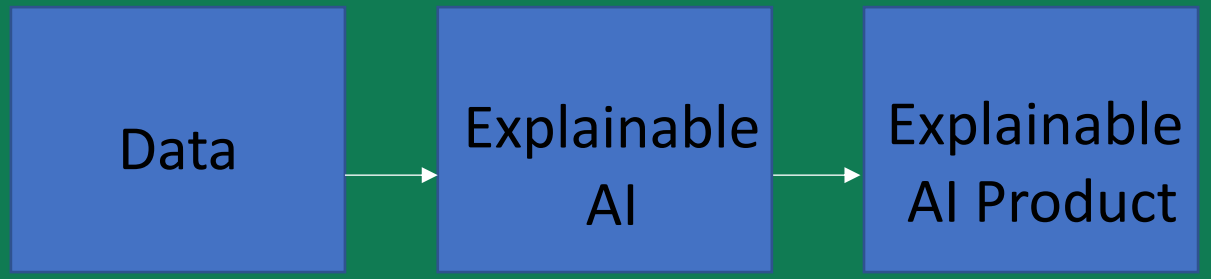


Decision, Recommendation

Confusion with today's AI Black Box

- Why did you do that?
- Why did it succeed or fail?
- How do I correct this error?

Explainable AI



Decision

Explanation

Advantages of Explainable AI

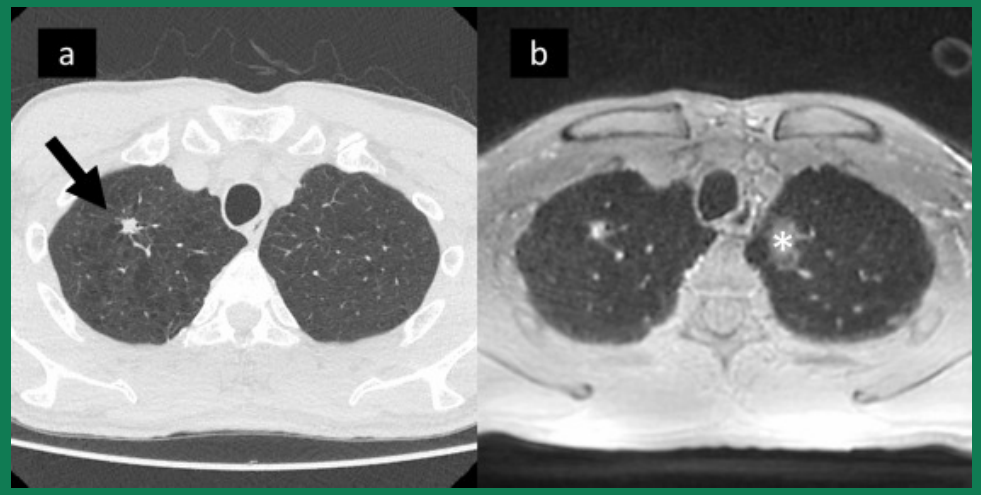
- I Understand Why
- I Know why it succeeds or fails
- I Understand, so I can **Trust**

Why Explainability? :Verify the ML Model

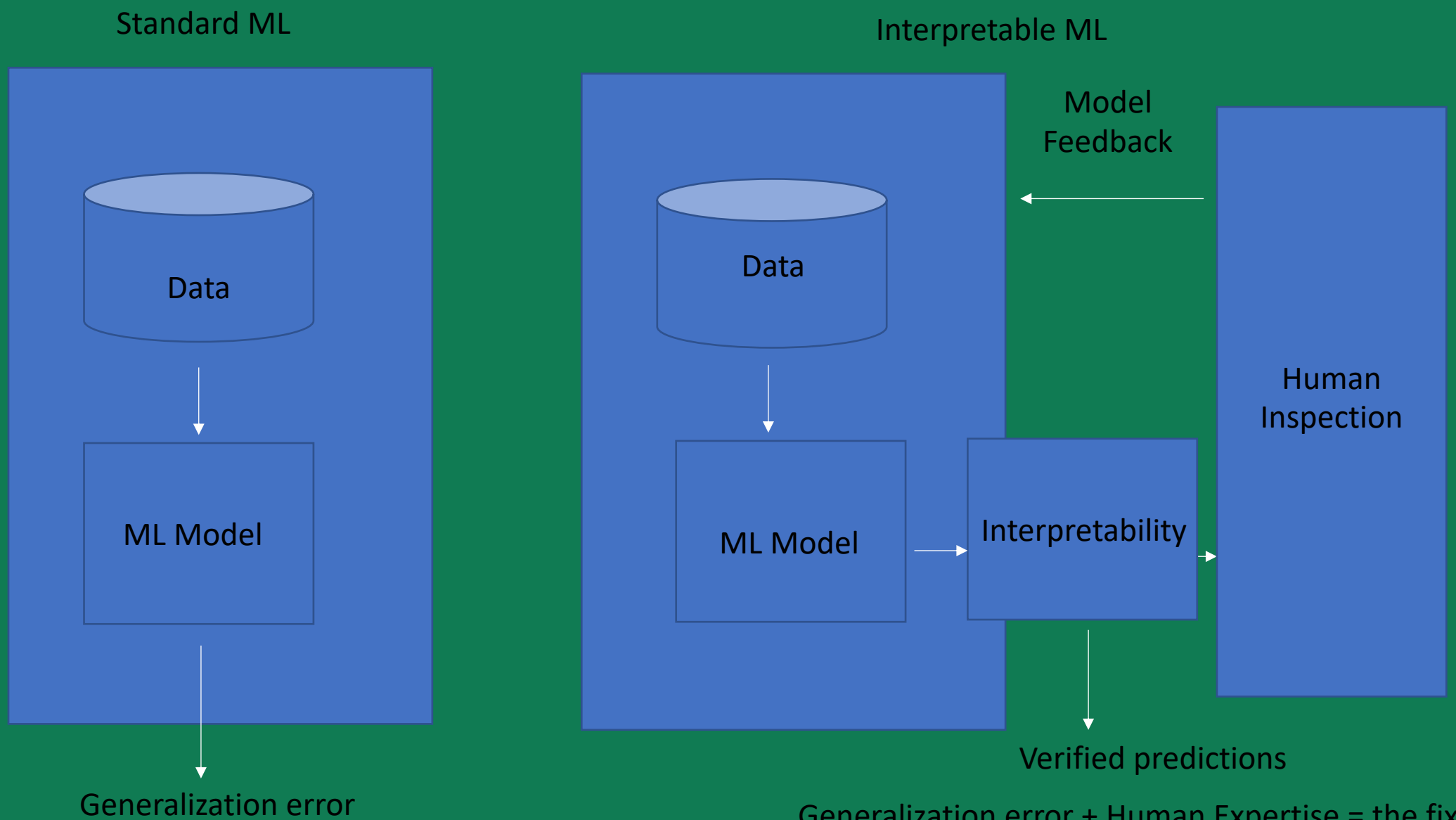
Wrong Decisions can be costly!

Autonomous car crashes,
Because it wrongly recognizes…

AI System mis-classifies patient's disease



Why Explainability? : Improve the ML Model



Why Explainability? : Laws against Discrimination

Citizenship

Immigration and control Act

Sex

Equal Pay Act of 1963;
Civil Rights Act of 1964

Age

Age Discrimination in Employment Act of 1967

Race

Civil Rights Act of 1964

Disability Status

Rehabilitation Act f 1973;
Americans with Disabilities Act of 1990

And More..

Fairness

Privacy

SR 11-7: Guidance on Model Risk Management

BOARD OF GOVERNORS
OF THE FEDERAL RESERVE SYSTEM
WASHINGTON, D.C. 20551

GDPR

Transparency

Explainability

CALIFORNIA
CONSUMER
PRIVACY
ACT OF 2018

GDPR Concerns Around Lack of Explainability in AI

“

Companies should commit to ensuring systems that could fall under GDPR, including AI, will be compliant. The threat of sizeable fines of €20 million or 4% of global turnover provides a sharp incentive.”

- European Commission

SR 11-7 and OCC regulations for Financial Institutions

SR 11-7: Guidance on Model Risk Management

BOARD OF GOVERNORS
OF THE FEDERAL RESERVE SYSTEM
WASHINGTON, D.C. 20551

What's driving Stress Testing and Model Risk Management efforts?

Regulatory efforts

SR 11-7 says "Banks benefit from **conducting model stress testing** to check performance over a wide range of inputs and parameter values, including extreme values, **to verify that the model is robust**"

In fact, **SR14-03** explicitly calls for **all models used for Dodd-Frank Act Company-Run Stress Tests must fall under the purview of Model Risk Management.**

In addition **SR12-07** calls for **incorporating validation or other type of independent review of the stress testing framework to ensure the integrity of stress testing processes and results.**

JOHN HILL

GLOBAL HEAD OF MODEL RISK GOVERNANCE, CREDIT SUISSE

"In the current regulatory environment, model validation policies must be fully compliant with the requirements of SR11-7. While SR11-7 officially applies to US conforming bank and non-US banks doing business in the US, many European financial firms have adopted SR11-7 as their standard as well."

Credit Lending in a Black-box ML world

Credit Line Increase

Query AI System

Credit Lending Model

Credit Lending Score = 0.3

Request Denied

Why? Why not? How?

Foundations and Techniques

Achieving Explainable AI

Approach 1: Post-hoc explain a given AI model

- Individual prediction explanations in terms of input features, influential examples, concepts, local decision rules
- Global prediction explanations in terms of entire model in terms of partial dependence plots, global feature importance, global decision rules

Approach 2: Build an interpretable model

- Logistic regression, Decision trees, Decision lists and sets, Generalized Additive Models (GAMs)

Achieving Explainable AI

Approach 1: Post-hoc explain a given AI model

- Individual prediction explanations in terms of input features, influential examples, concepts, local decision rules
- Global prediction explanations in terms of entire model in terms of partial dependence plots, global feature importance, global decision rules

Top label: “fireboat”

Why did the network label this image as “fireboat”?

The Attribution Problem

Attribute a model's prediction on an input to features of the input

Examples:

- Attribute an object recognition network's prediction to its pixels
- Attribute a text sentiment network's prediction to individual words
- Attribute a lending model's prediction to its features

A reductive formulation of "why this prediction" but surprisingly useful :-)

Application of Attributions

- **Debugging model predictions**
  E.g., Attribution an image misclassification to the pixels responsible for it
- **Generating an explanation for the end-user**
  E.g., Expose attributions for a lending prediction to the end-user
- **Analyzing model robustness**
  E.g., Craft adversarial examples using weaknesses surfaced by attributions
- **Extract rules from the model**
  E.g., Combine attribution to craft rules (pharmacophores) capturing prediction logic of a drug screening network

We will cover the following attribution methods**

- Ablations
- Gradient based methods
- Score Backpropagation based methods
- Shapley Value based methods

Ablations

Drop each feature and attribute the change in prediction to that feature

Useful tool but not a perfect attribution method. Why?

- Unrealistic inputs
- Improper accounting of interactive features
- Computationally expensive

Feature Gradient

Attribution to a feature is feature value times gradient, i.e., xi* $\partial y/\partial$

- Gradient captures sensitivity of output w.r.t. feature

Gradients in the vicinity of the input seem like noise

$$IG(input, base) ::= (input - base) * \int_{0}^{1} \nabla F(\alpha * input + (1-\alpha) * base)\, d\alpha$$

Original Image

Integrated Gradient Baselines

What is a baseline?

Ideally, the baseline is an informationless input for the model
E.g., Black image for image models
E.g., Empty text or zero embedding vector for text models
Integrated Gradients explains F(input) - F(baseline) in terms of input features

Why is this image labelled as “clog”?

Original image

Integrated Gradients(for label “clog”)
“Clog”

“Clog”

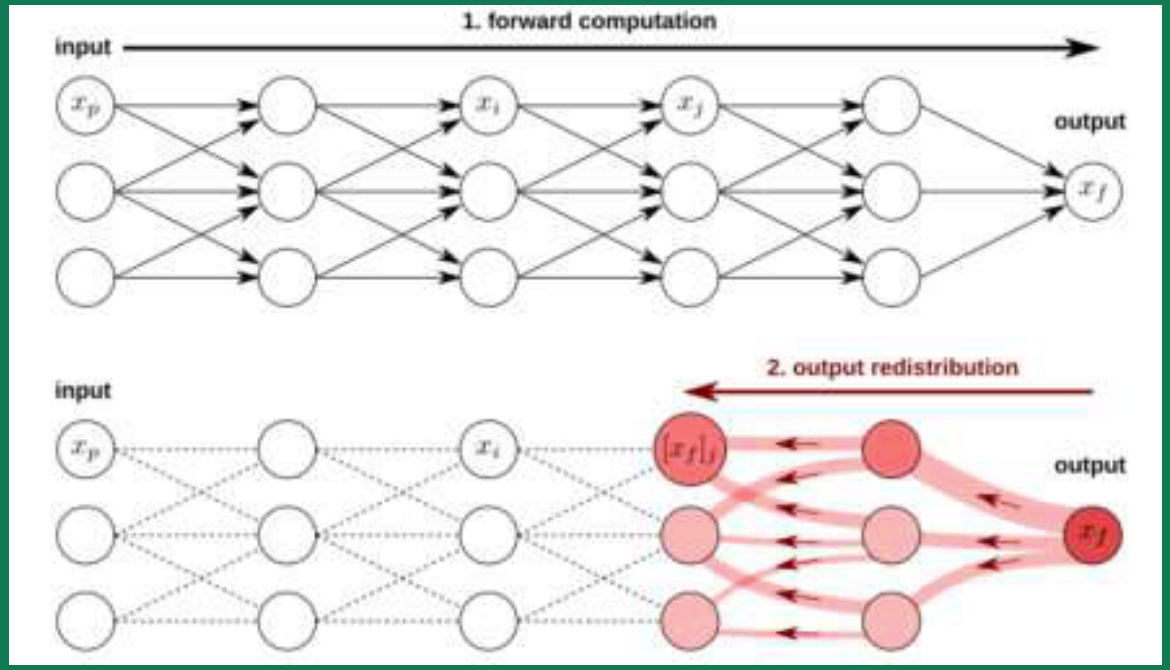
Score Back-Propagation based Methods

Re-distribute the prediction score through the neurons in the network

- LRP [JMLR 2017], DeepLift [ICML 2017], Guided BackProp [ICLR 2014]

Easy case: Output of a neuron is a linear function
of previous neurons (i.e., ni = $\sum$ wij * nj)
e.g., the logit neuron

- Re-distribute the contribution in proportion to
the coefficients wij



The Shapley Value

Given a player i, and a set $S \subseteq N$, the marginal contribution of i to S is

$$m_i(S) = v(S \cup \{i\}) - v(S)$$

How much does i contribute by joining S?

Given a permutation $\sigma \in \Pi(N)$ of players, let the predecessors of i in σ be

$$P_i(\sigma) = \{j \in N \mid \sigma(j) < \sigma(i)\}$$

We write $m_i(\sigma) = m_i(P_i(\sigma))$

Evaluating Attribution Methods

Human Review

Have humans review attributions and/or compare them to (human provided)
groundtruth on “feature importance”

Pros:
- Helps assess if attributions are human-intelligible
- Helps increase trust in the attribution method

Cons:
- Attributions may appear incorrect because model reasons differently
- Confirmation bias

Axiomatic Justification

Inspired by how Shapley Values are justified

- List desirable criteria (axioms) for an attribution method
- Establish a uniqueness result: X is the only method that satisfies these criteria

Integrated Gradients, SHAP, QII, Strumbelj & Konenko are justified in this manner

Theorem [Integrated Gradients, ICML 2017]: Integrated Gradients is the
unique path-integral method satisfying: Sensitivity, Insensitivity, Linearity
preservation, Implementation invariance, Completeness, and Symmetry

Some limitations and caveats

Attributions are pretty shallow

Attributions do not explain:

- Feature interactions
- What training examples influenced the prediction
- Global properties of the model

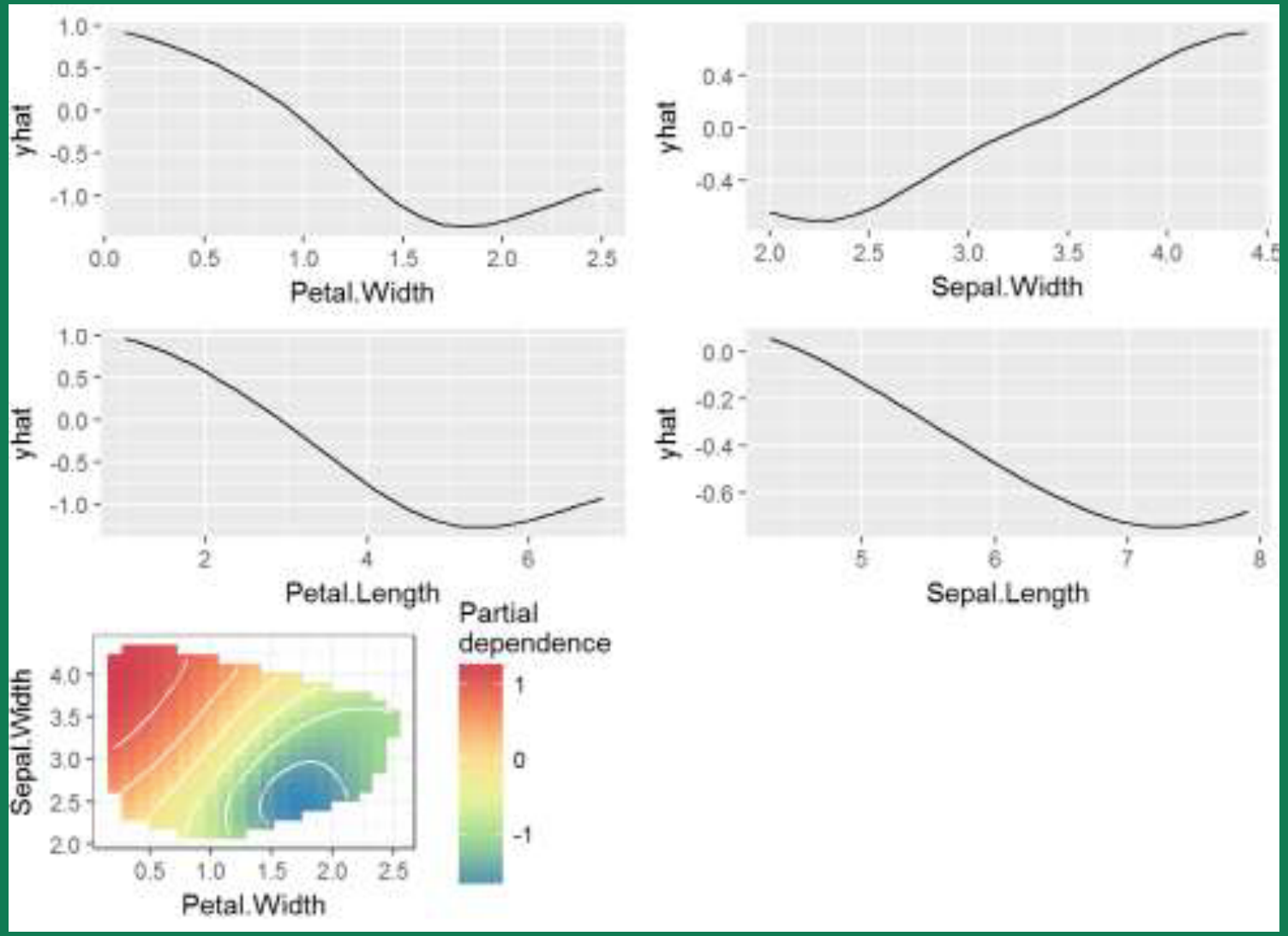
An instance where attributions are useless:

- A model that predicts TRUE when there are even number of black pixels and FALSE otherwise

Other types of Post-hoc Explanations

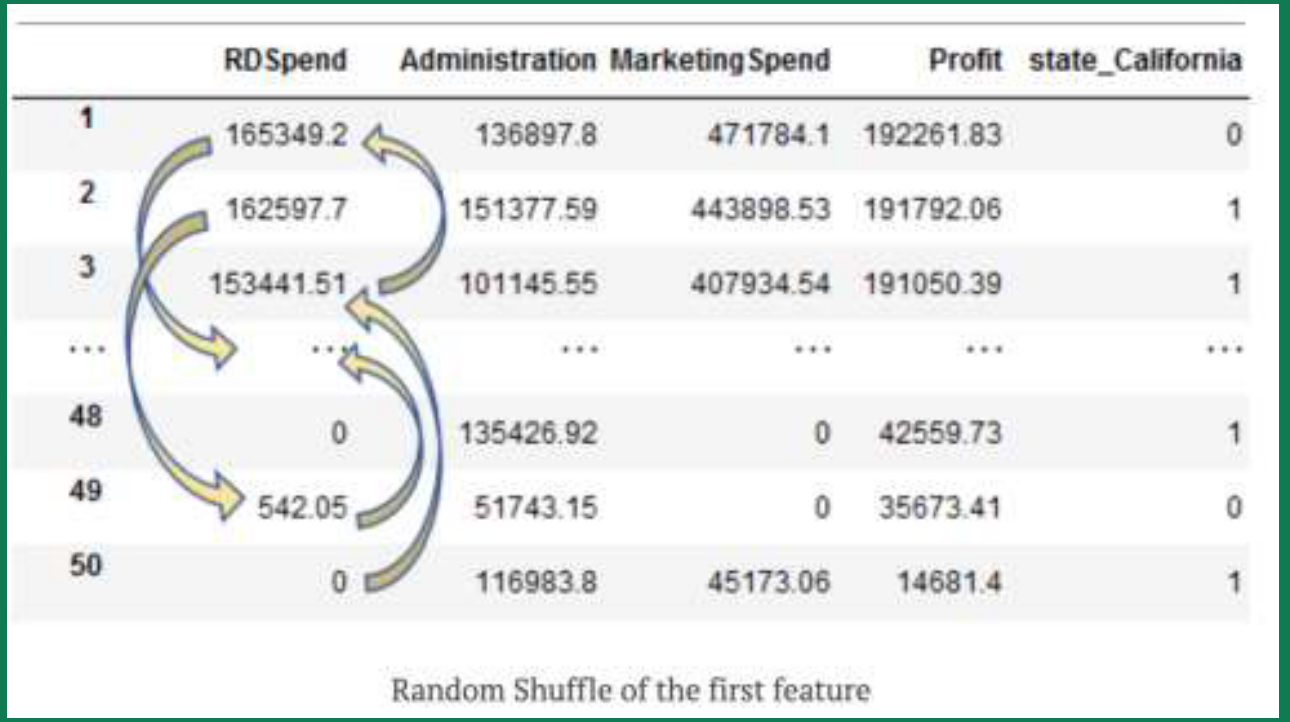
Global Explanations Methods

- Partial Dependence Plot: Shows the marginal effect one or two features have on the predicted outcome of a machine learning model



Global Explanations Methods

- Permutations: The importance of a feature is the increase in the prediction error of the model after we permuted the feature's values, which breaks the relationship between the feature and the true outcome.



	RDSpend	Administration	MarketingSpend	Profit	state_California
1	165349.2	136897.8	471784.1	192261.83	0
2	162597.7	151377.59	443898.53	191792.06	1
3	153441.51	101145.55	407934.54	191050.39	1
...	...	...	...	...	...
48	0	135426.92	0	42559.73	1
49	542.05	51743.15	0	35673.41	0
50	0	116983.8	45173.06	14681.4	1

Random Shuffle of the first feature

Approach 2: Build an interpretable model

- Logistic regression, Decision trees, Decision lists and sets, Generalized Additive Models (GAMs)

Decision Trees

Is the person fit?

Age < 30 ? Yes

Eats a lot of pizzas? No

Exercises in the morning? Yes

Unfit. Fit

No yes

Decision List

If Past-Respiratory-Illness =Yes **and** Smoker =Yes **and** Age $\geq$ 50, **then** Lung Cancer

Else if Allergies =Yes **and** Past-Respiratory-Illness =Yes, **then** Asthma

Else if Family-Risk-Respiratory =Yes, **then** Asthma

Else if Family-Risk-Depression =Yes, **then** Depression

Else if Gender =Female **and** Short-Breath-Symptoms =Yes, **then** Asthma

Else if BMI $\geq$ 0.2 **and** Age$\geq$ 60, **then** Diabetes

Else if Frequent-Headaches =Yes **and** Dizziness =Yes, **then** Depression

Else if Frequency-Doctor-Visits $\geq$ 0.3, **then** Diabetes

Else if Disposition-Tiredness =Yes, **then** Depression

Else if Chest-Pain =Yes **and** Nausea **and** Yes, **then** Diabetes

Else Diabetes

Decision Set

If Allergies =Yes **and** Smoker =Yes **and** Irregular-Heartbeat =Yes, **then** Asthma

If Allergies =Yes **and** Past-Respiratory-Illness =Yes **and** Avg-Body-Temperature $\geq$ 0.1, **then** Asthma

If Smoker =Yes **and** BMI $\geq$ 0.2 **and** Age $\geq$ 60, **then** Diabetes

If Family-Risk-Diabetes =Yes **and** BMI $\geq$ 0.4 =Frequency-Infections $\geq$ 0.2, **then** Diabetes

If Frequency-Doctor-Visits $\geq$ 0.4 **and** Childhood-Obesity =Yes **and** Past-Respiratory-Illness =Yes, **then** Diabetes

If Family-Risk-Depression =Yes **and** Past-Depression =Yes **and** Gender =Female, **then** Depression

If BMI $\geq$ 0.3 **and** Insurance-Coverage =None **and** Avg-Blood-Pressure $\geq$ 0.2, **then** Depression

If Past-Respiratory-Illness =Yes **and** Age $\geq$ 50 **and** Smoker =Yes, **then** Lung Cancer

If Family-Risk-LungCancer =Yes **and** Allergies =Yes **and** Avg-Blood-Pressure $\geq$ 0.3, **then** Lung Cancer

If Disposition-Tiredness =Yes **and** Past-Anemia =Yes **and** BMI $\geq$ 0.3 **and** Rapid-Weight-Loss =Yes, **then** Leukemia

If Family-Risk-Leukemia =Yes **and** Past-Blood-Clotting =Yes **and** Frequency-Doctor-Visits $\geq$ 0.3, **then** Leukemia

If Disposition-Tiredness =Yes **and** Irregular-Heartbeat =Yes **and** Short-Breath-Symptoms =Yes **and** Abdomen-Pains =Yes, **then** Myelofibrosis

GLMs and GAMs

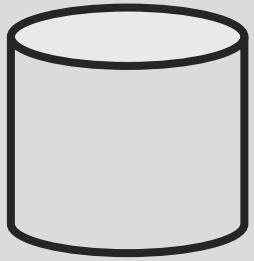
Model	Form	Intelligibility	Accuracy
Linear Model	$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$	+++	+
Generalized Linear Model	$g(y) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$	+++	+
Additive Model	$y = f_1(x_1) + ... + f_n(x_n)$	++	++
Generalized Additive Model	$g(y) = f_1(x_1) + ... + f_n(x_n)$	++	++
Full Complexity Model	$y = f(x_1, ..., x_n)$	+	+++

AI Fairness 360

A rich open source library to allow model builders to investigate and fix bias in models

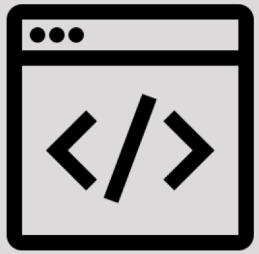
Build / Test Time

Data Selection



Pre-Processing

Algorithm Design



In-Processing

Test Deployment

Post Processing

AI Fairness 360

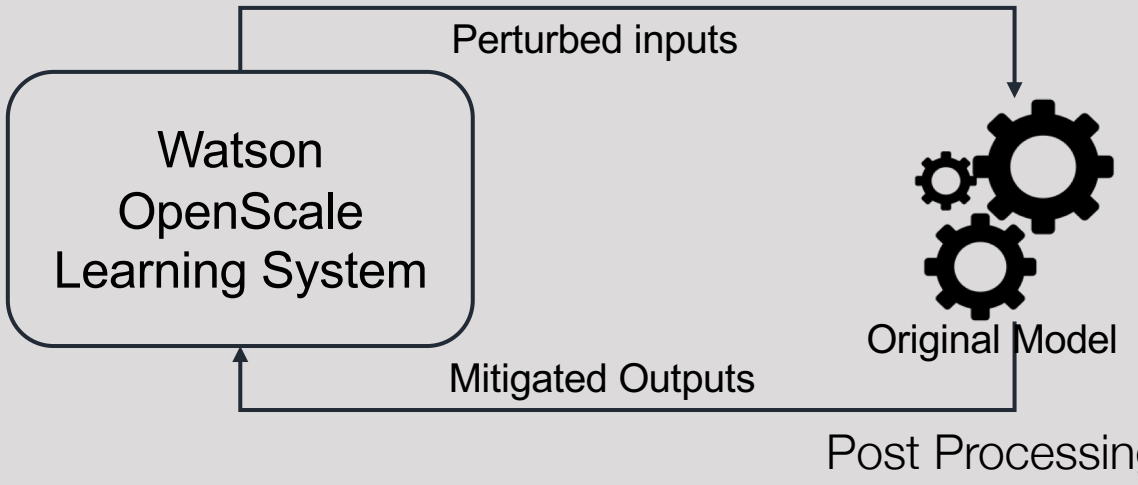
- Data scientist / Developer driven
- Suite of several algorithms applicable at different points of lifecycle
- Python based integration
- Backed by leading fairness researchers at IBM

Watson OpenScale

Extends state of art with novel data perturbation techniques with use of online scoring that make it most effective for production scenarios

Production runtime

Production Deployment



- Introduces data perturbation with active scoring to verify bias actually exists in applications
- Zero code Integration
- Final mitigated output comes from original model
- Allows individual bias
- OOTB Dashboards and custom reporting via open data mart
- Designed by leading IBM researchers

Watson OpenScale

business owner

I won't build an AI System that I can't understand. What can I do?

IBM Watson OpenScale

100% understandable

nothing

0% of potential AI Systems allowed to proceed

IT and dev ops

I don't want to waste time. Are there out-of-the-box tools and frameworks to deploy and instrument my AI System?

yes

3-5x reduction in time-to-market

data science team

If my AI System isn't instrumented then I can't understand and improve it. Can I get all system sessions monitored and traced down to individual predictions?

yes

Improved performance of AI system from 50% to 300%

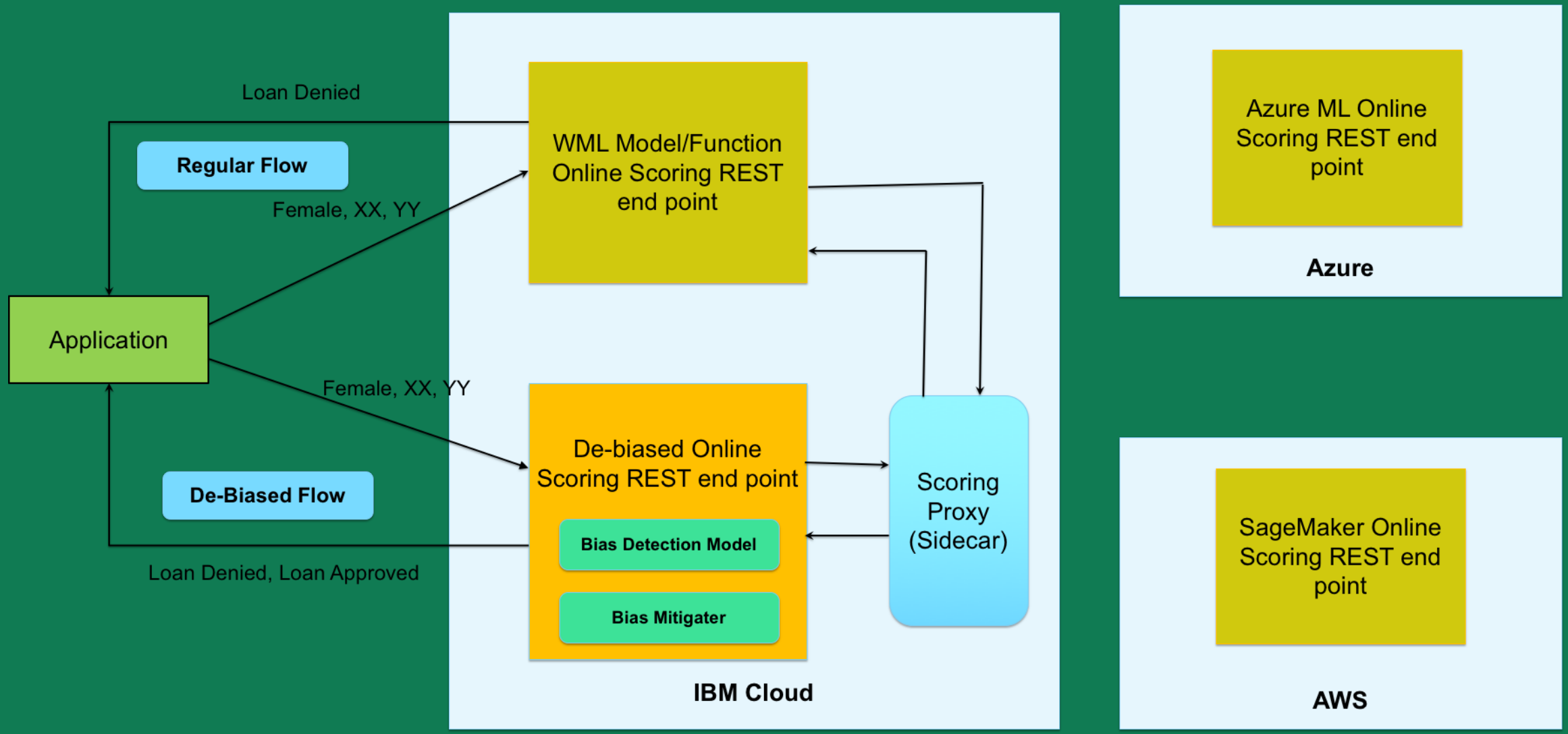
Deployed AI System

Satisfied customer?

yes

priceless

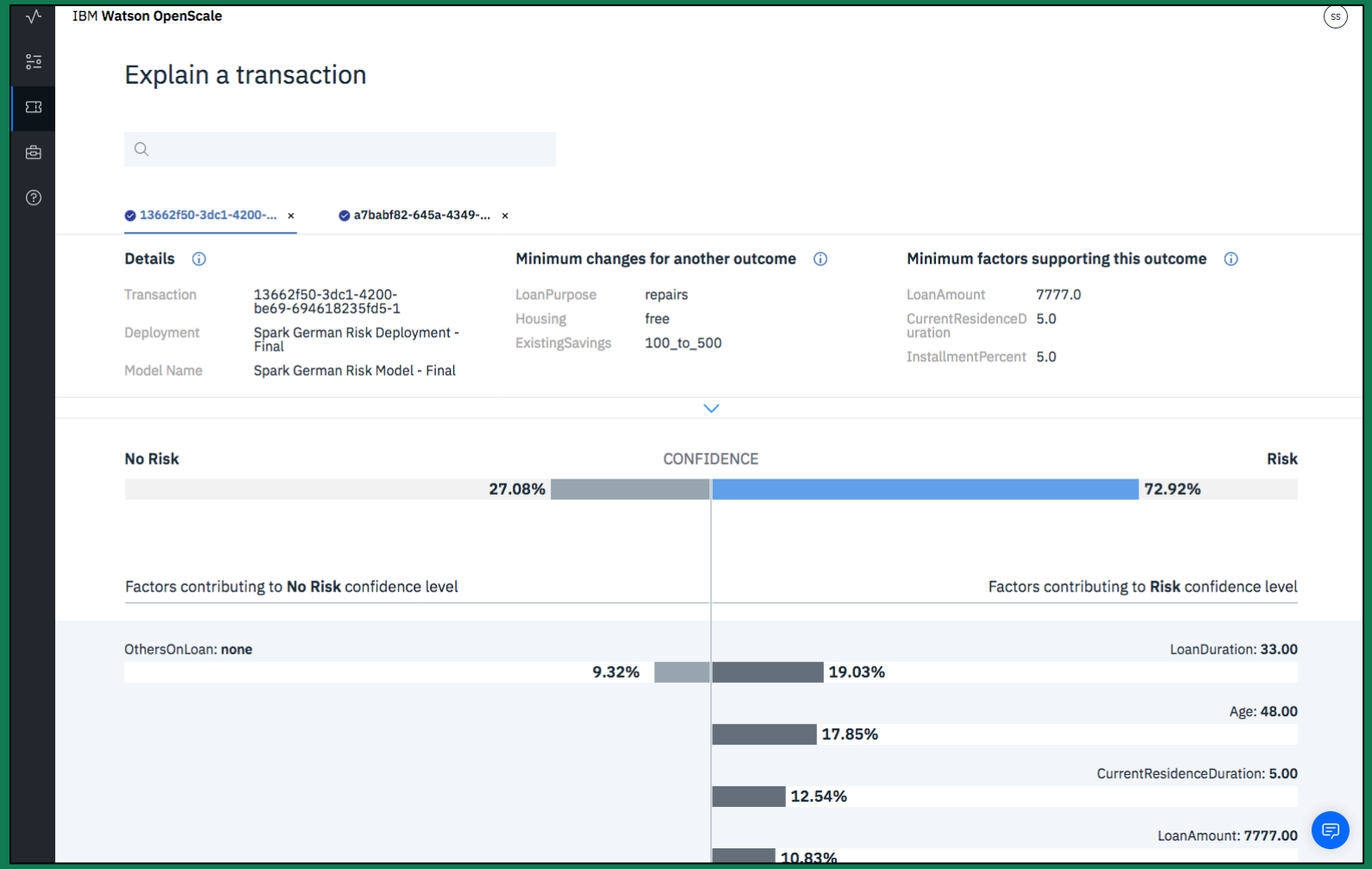
- We continually monitor how AI makes predictions to individual transactions to ensure it produced fair outcomes.
- Input
  - Model deployment, fairness attribute, favored/protected population, class label, favorable/unfavorable outcome.
- Output:
  - Is the model biased on some attribute for a specific value (Gender=Female)
  - What is the source of the bias? (Model is biased for loan amount > $2M)
  - Data which can be sent for manual labelling which will help mitigate bias.
  - Natural language explanation of how bias was computed and detected


Loan Denied
Regular Flow
Female, XX, YY
Application
Female, XX, YY
De-Biased Flow
Loan Denied, Loan Approved
WML Model/Function Online Scoring REST end point
De-biased Online Scoring REST end point
Bias Detection Model
Bias Mitigater
Scoring Proxy (Sidecar)
IBM Cloud
Azure ML Online Scoring REST end point
Azure
SageMaker Online Scoring REST end point
AWS

Explainability

Explain how AI arrived at a prediction

- Uses contrastive techniques to explain model behavior in the vicinity of the target data point.
- Identifies feature weighting of most and least important features
- Displays factors that influence a prediction in simple terms.
- Explanation in terms of the top-K features which played a key role in the prediction. E.g., The loan was rejected because: (1) Credit score=average, (2) Loan Amount>$2M and (3) Area=Downtown.



Bias Mitigation – Guardrails against AI Bias

Watson OpenScale

Online Post Processing
(Intercept & Fix Features)

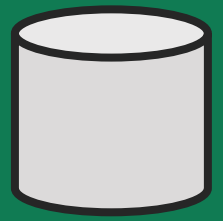
AI Fairness 360

In-Processing
(Intercept & Fix Algorithm)

AI Fairness 360

Pre-Processing
(Intercept & Fix Training Data)

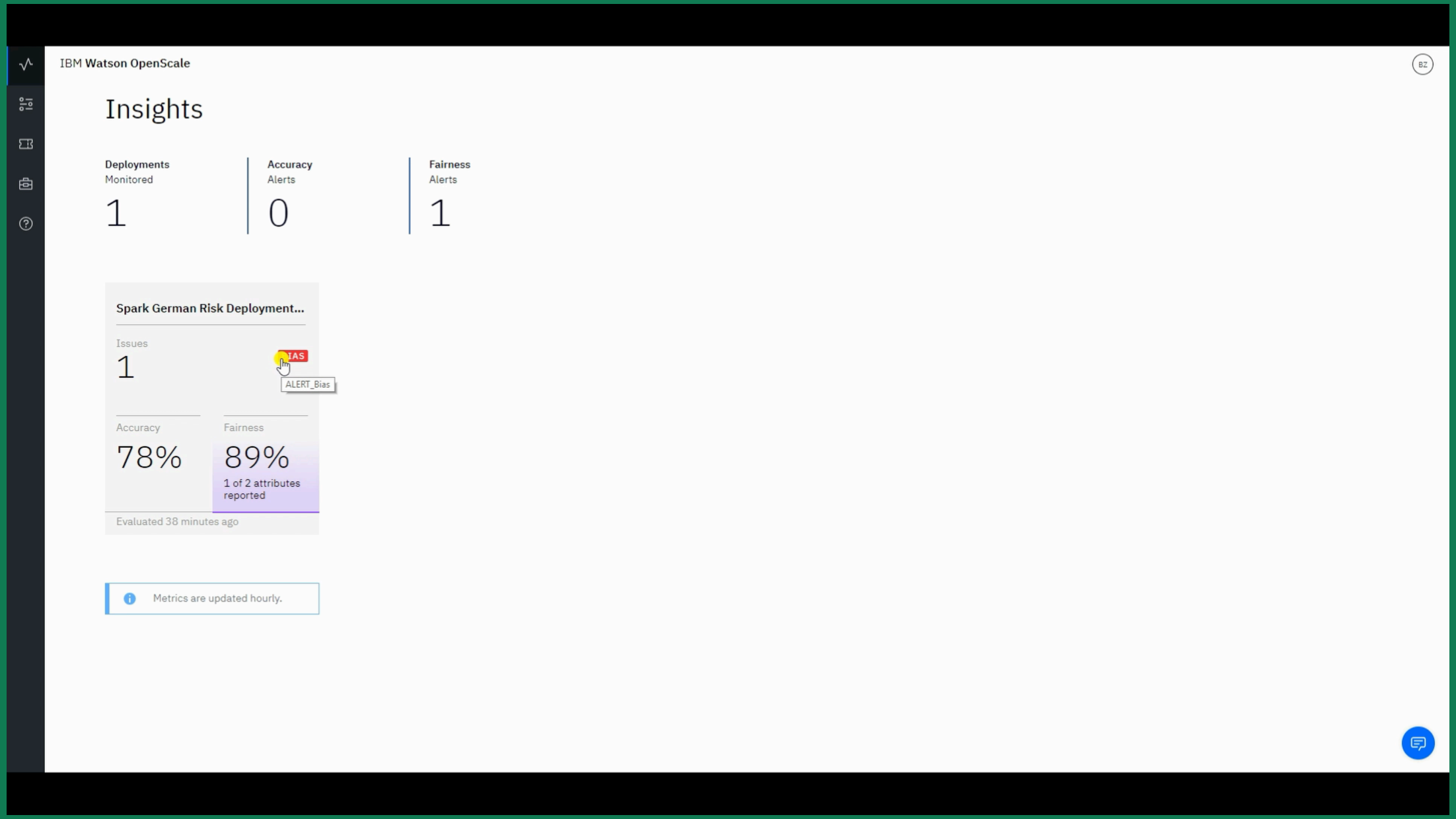
Data Selection



Algorithm

Runtime

Bias Mitigation – Guardrails against AI Bias

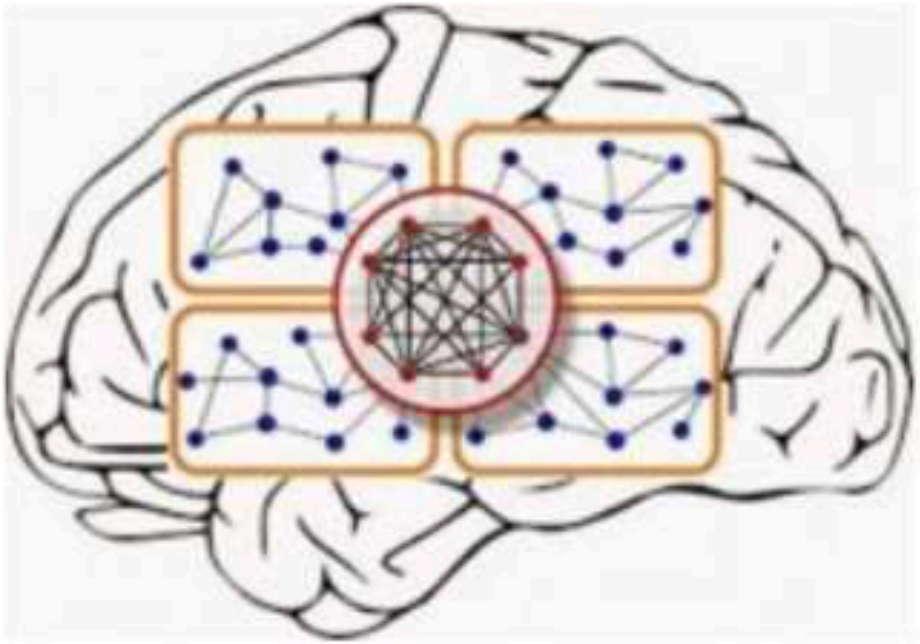


Learning New Insights

"It's not a human move. I've never seen a human play this move." (Fan Hui)

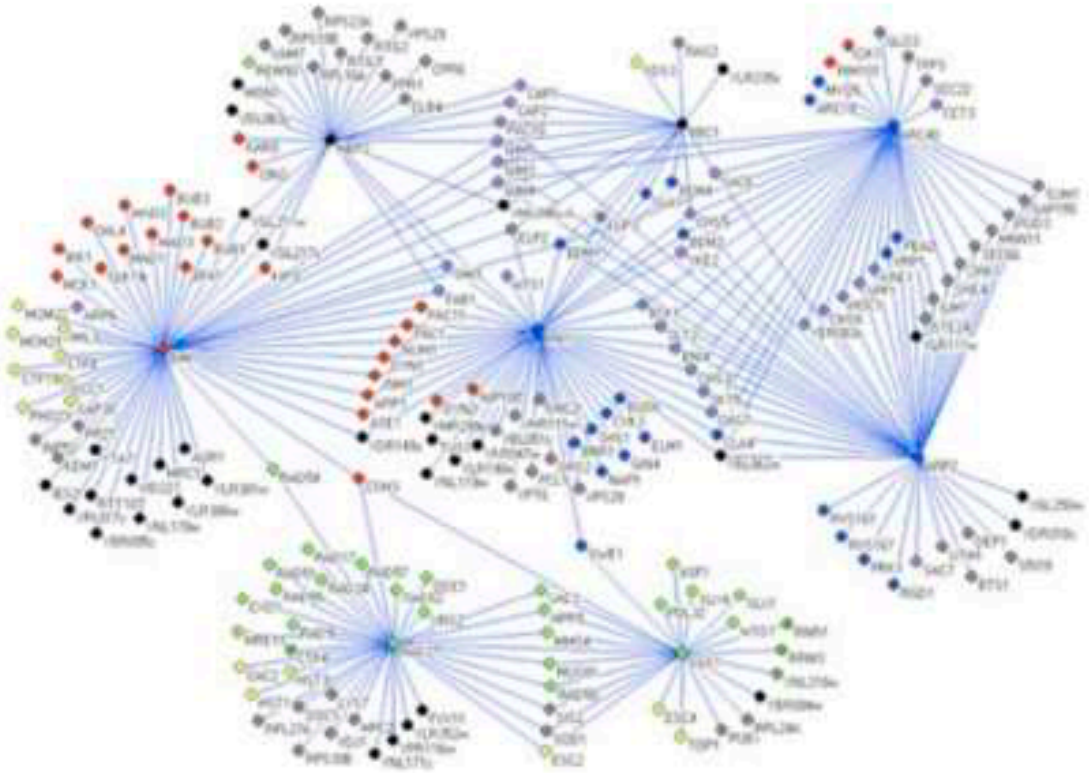
Old promise:

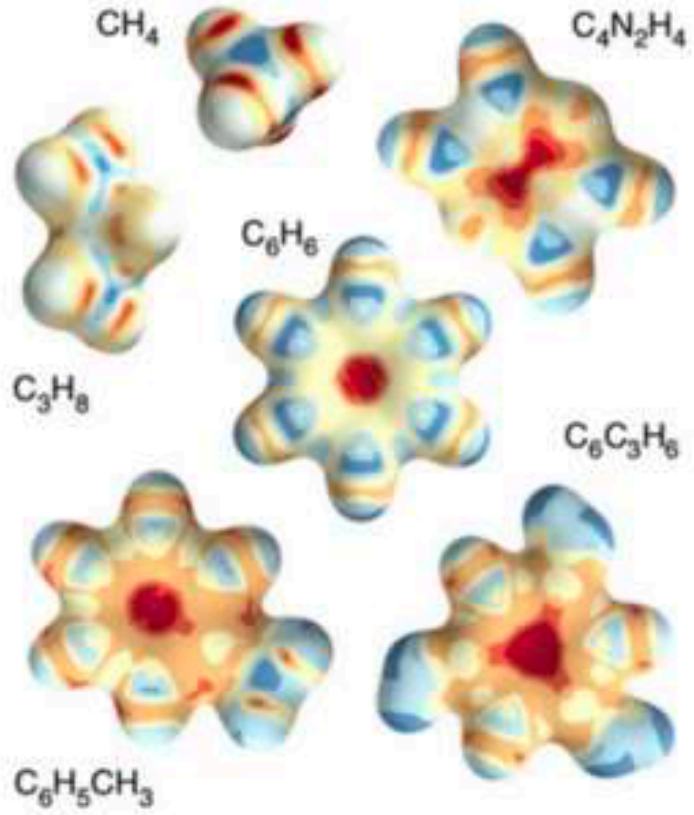
"Learn about the human brain."



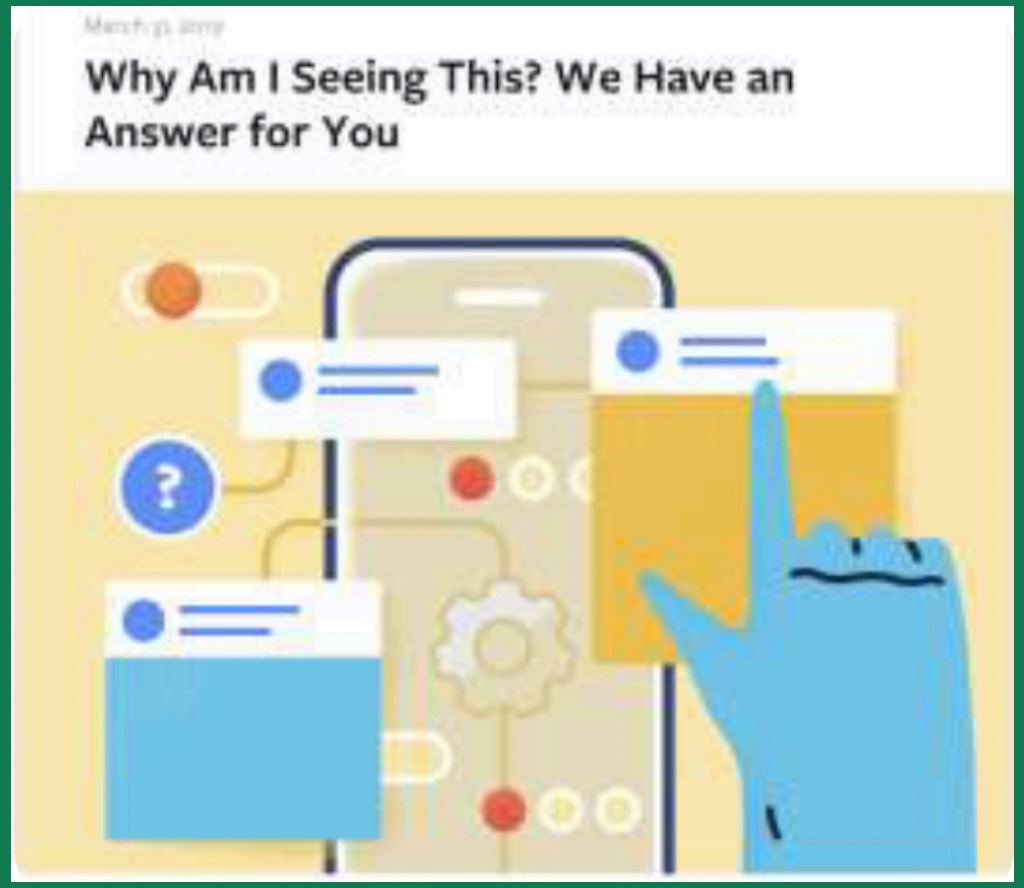
Learning New Insights

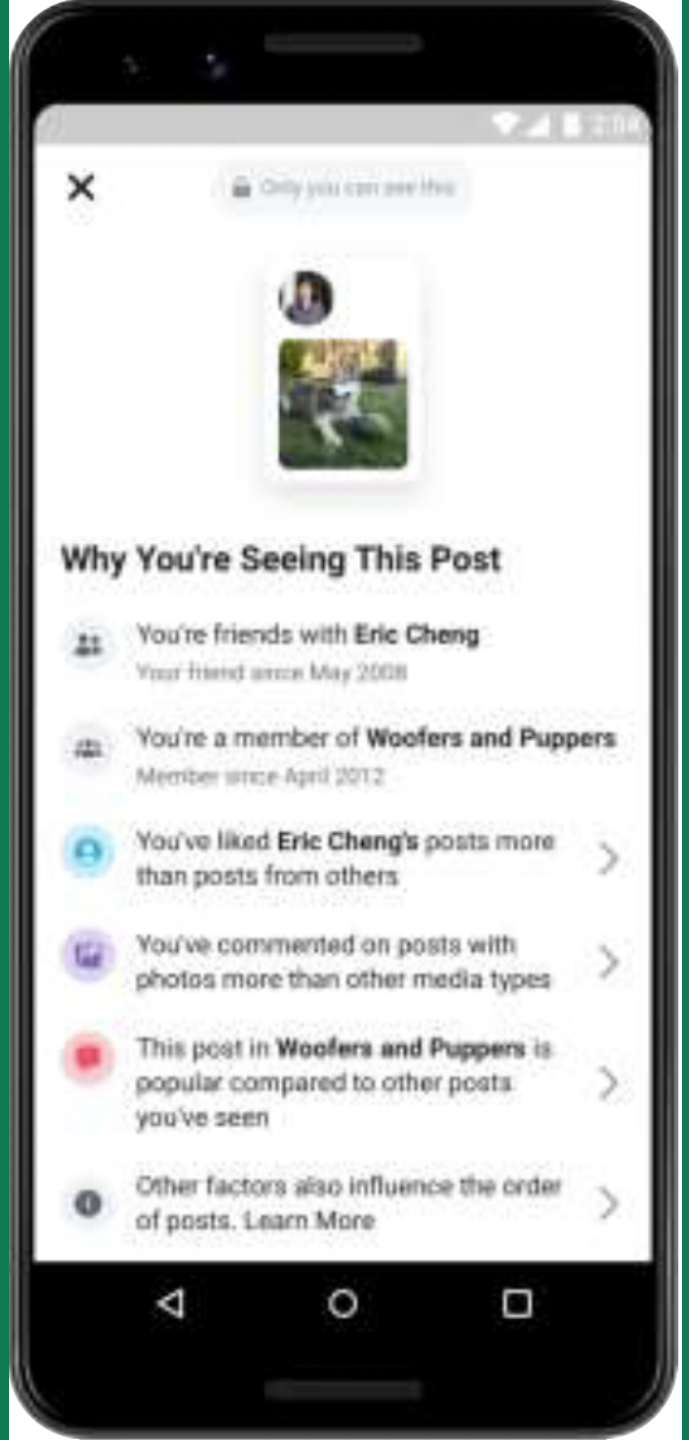
Learn about the physical / biological / chemical mechanisms.
(e.g. find genes linked to cancer, identify binding sites …)



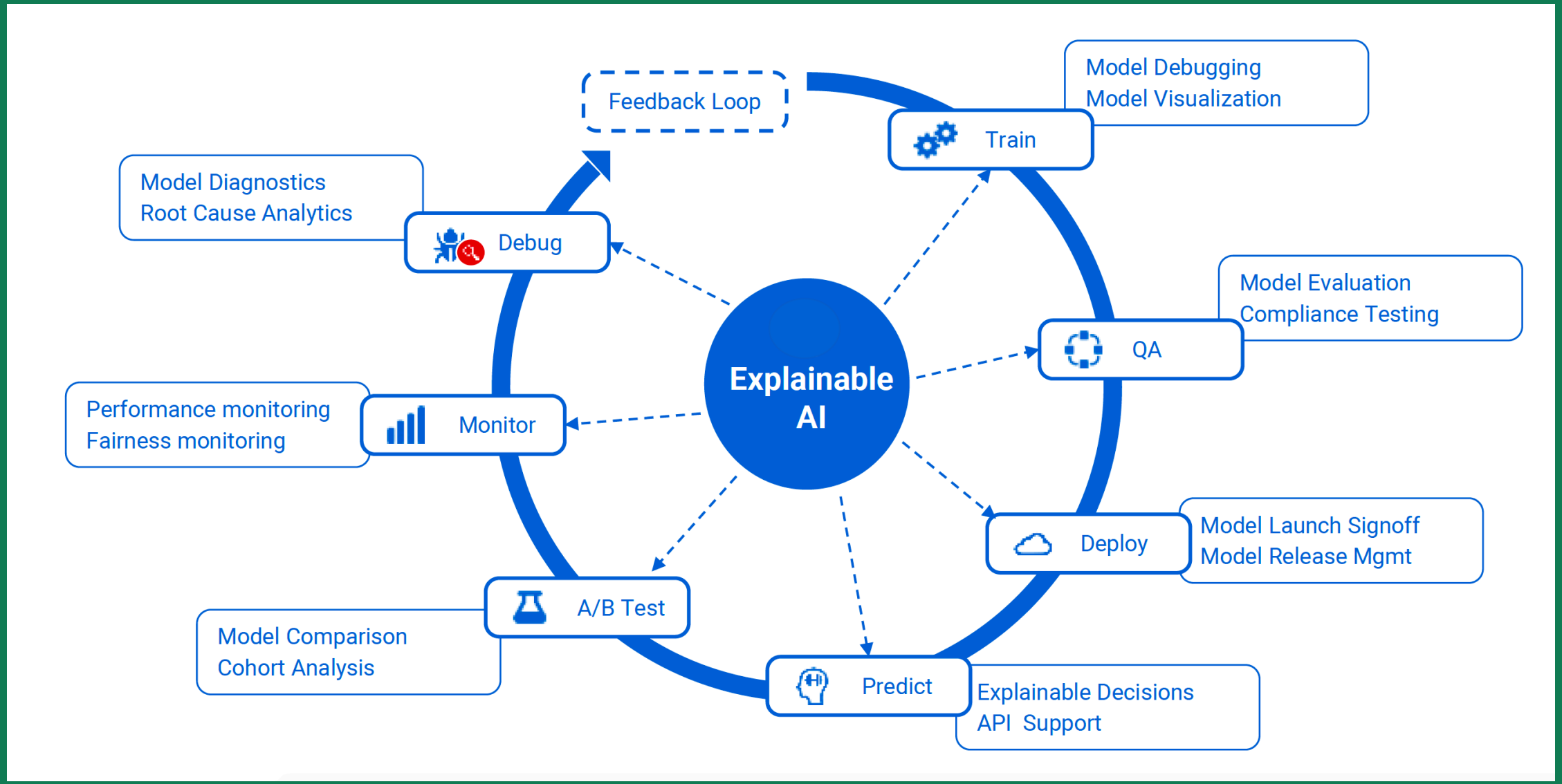


Example: Facebook adds Explainable AI to build Trust





AI Explainability by Design



Thank You