Trusted AI - Explainability

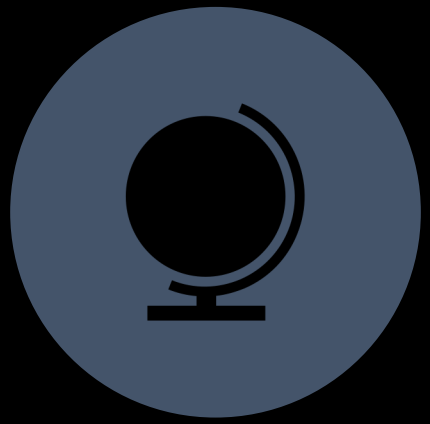
Guest Lecture, Oct 19th
Diptikalyan Saha, Vijay Arya
IBM Research

Plan

- 19th
  - Part 1
    - Overview of Three types of Explainability
    - Local Explainability – LIME, ANCHORS
    - Global Explainability – TREPAN
    - Counterfactual Explainability – Wachter
  - Part 2
    - Introduction to AIX360
      - Usage, Taxonomy
- 21st
  - Notebook walkthrough of Algorithms in AIX360

Explainability

LOCAL



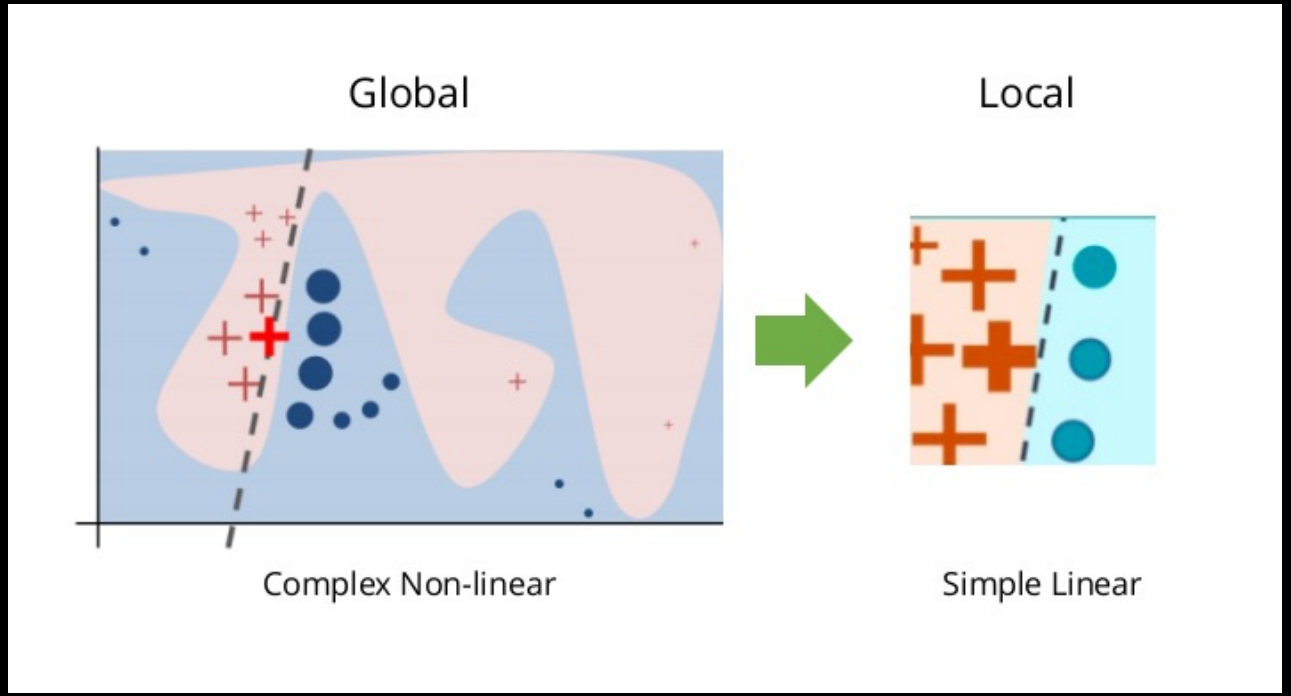
GLOBAL

COUNTERFACTUAL

https://github.com/marcotcr/lime

LIME

- Permute data
- Distance between permutations and original observation
- Make predictions on new data using original model
- Pick small number of best features
- Fit a simple linear model
- Feature weights from the simpler model

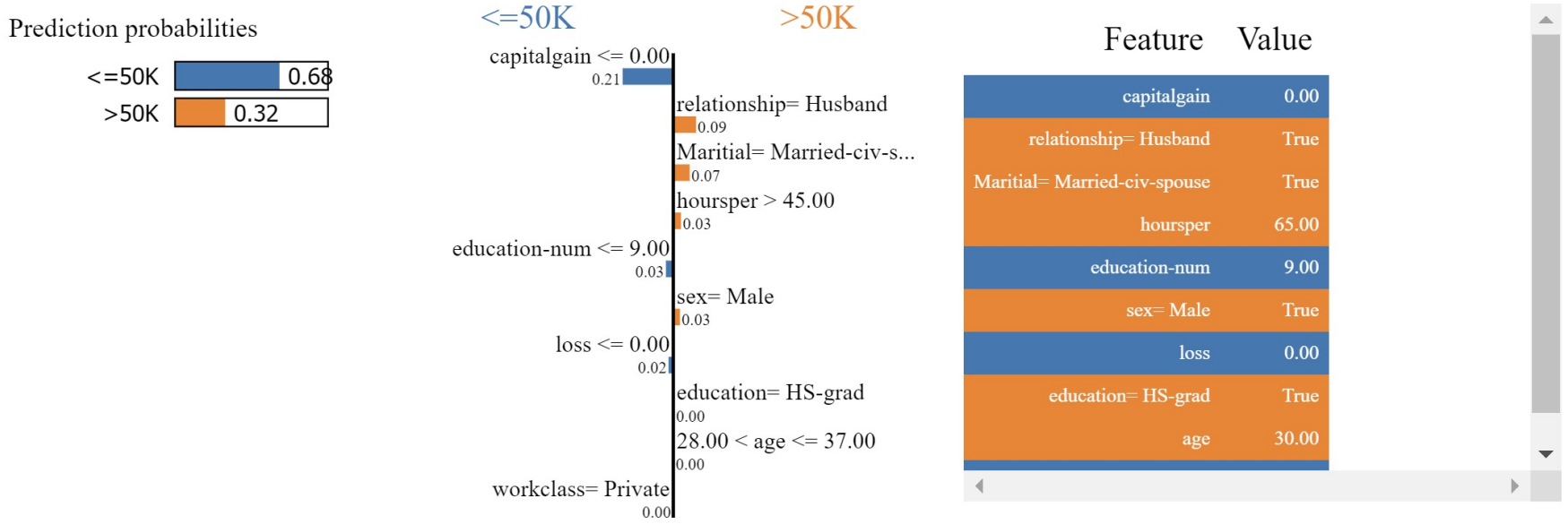


Demo

In [57]:
```
sample = X_test[0]
print(model.test1([sample]))
exp = explainer.explain_instance(sample, predict_fn)
print('Sample',sample)
print('Lime Explanation',exp.as_list())
```

```
[0]
Sample [3.0000e+01 2.0000e+00 2.0245e+05 1.1000e+01 9.0000e+00 2.0000e+00
 1.3000e+01 0.0000e+00 4.0000e+00 1.0000e+00 0.0000e+00 0.0000e+00
 6.5000e+01 3.8000e+01]
Lime Explanation [('capitalgain <= 0.00', -0.2050326069047271), ('relationship= Husband', 0.0920737113788543), ('Maritial= Married-civ-spouse', 0.06949948999332721), ('sex= Male', 0.029527907120445854), ('education-num <= 9.00', -0.029197828082035183), ('hoursper > 45.00', 0.0278694315190951), ('loss <= 0.00', -0.023773821411809164), ('education= HS-grad', 0.006204495072599 59), ('occupation= Transport-moving', 0.0030610736036260555), ('country= United-States', 0.0020635778905474665)]
```

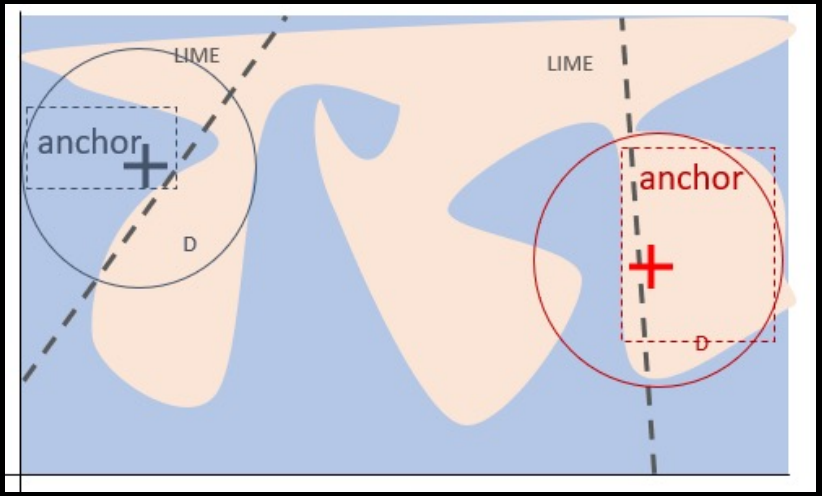
In [56]:
```
exp.show_in_notebook(show_table=True, show_all=False)
```



ANCHOR

https://github.com/marcotcr/anchor

- ANCHOR computes the region where the same prediction holds with high probability
- Easier to comprehend
- Unlike LIME it provides how faithful the explanation is
- Beam search technique in the space of candidate rules till it satisfies precision and coverage criteria




```python
from anchor import anchor_tabular

anchor_explainer = anchor_tabular.AnchorTabularExplainer(
                d.class_names,
                d.feature_names,
                 X_train,
                 d.categorical_names)
aexp = anchor_explainer.explain_instance(sample, model.test1, threshold=0.95)
print('Anchor: %s' % (' AND '.join(aexp.names())))
print('Precision: %.2f' % aexp.precision())
print('Coverage: %.2f' % aexp.coverage())
```

```
Anchor: capitalgain <= 0.00
Precision: 1.00
Coverage: 0.91
```

$$\max_{A \text{ s.t. } P(\text{prec}(A) \geq \tau) \geq 1-\delta} \text{cov}(A).$$

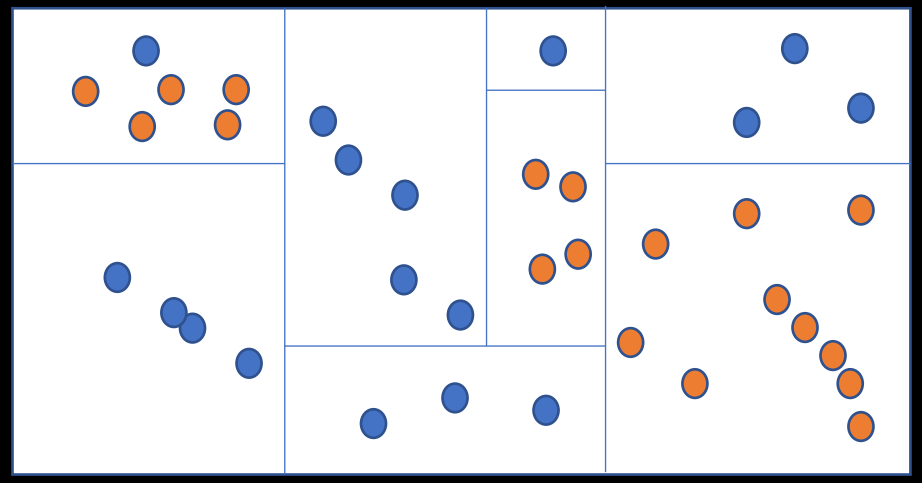
Cov(A) = probability A applies to samples from D
D = perturbation distribution around sample x
τ= desired level of precision
Prec(A) = anchor precision

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations." *AAAI* 2018.

Image Source https://drive.google.com/file/d/1QdBRUCn_yYtUlJaTyWghcp9XJZ4K2kY9/view

TrePan

- Approximate black-box model's decision boundary by decision tree
- Model Agnostic path coverage TrePan
  - Modified decision tree algorithm
    - Uses target model's decision (instead of training data's decision)
  - Decision tree algorithm forms boundary based on fewer number of samples near the leaf levels
    - TrePan generates more synthetic data to increase density to create more accurate boundary



```
from trepan import Trepan
from decisiontreehelper import convert_path
from helper import generate_feature_distributions

t = Trepan(X_train, model,d.feature_names,label, d.encode)
paths = t.build()
print('+++ Trepan #nodes = {}, #leaves = {}'.format(t.nodes, len(paths)))
paths1 = []
for p in paths:
    p1 = convert_path(p,d.feature_names, generate_feature_distributions(X))
    paths1.append(p1)

fidelity = t.fidelity(X_train)
print('train fidelity',fidelity)
fidelity = t.fidelity(X_test)
print('test fidelity',fidelity)

paths1[0]
```

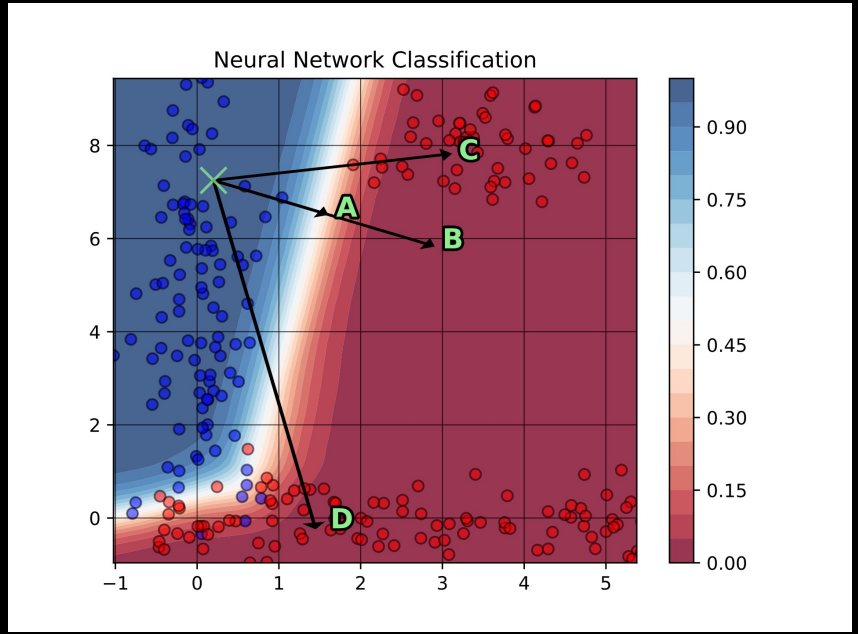
```
+++ Trepan #nodes = 385, #leaves = 172
train fidelity 1.0
test fidelity 0.9991712249295541
```

```
6]: {'age': [17, 90],
 'workclass': [0, 6],
 'fnlwgt': [13769, 1484705],
 'education': [0, 15],
 'education-num': [1, 16],
 'Maritial': [0, 6],
 'occupation': [0, 13],
 'relationship': [0, 5],
 'race': [0, 4],
 'sex': [0, 1],
 'capitalgain': [0, 5119],
 'loss': [0, 4356],
 'hoursper': [1, 99],
 'country': [0, 40]}
```

Craven, Mark, and Jude W. Shavlik. "Extracting tree-structured representations of trained networks." *Advances in neural information processing systems*. 1996.

Counterfactual Explainability

- A counterfactual explanation of a prediction describes the smallest change to the feature values that changes the prediction to a predefined output



Wachter et al. suggest minimizing the following loss:

$$L(x, x', y', \lambda) = \lambda \cdot (\hat{f}(x') - y')^2 + d(x, x')$$

https://github.com/BadrinathS/Counterfactual-generation-using-Wachter-paper

https://github.com/interpretml/DiCE https://github.com/carla-recourse/CARLA

```
# provide the trained ML model to DiCE's model object
backend = 'sklearn'
m = dice_ml.Model(model=model, backend=backend)
```

Generate diverse counterfactuals

```
# initiate DiCE
exp_random = dice_ml.Dice(d, m, method="random")
```

```
query_instances = x_train[4:6]
```

```
# generate counterfactuals
dice_exp_random = exp_random.generate_counterfactuals(query_instances, total_CFs=2, desired_class="opposite", verbose=False)
```

```
100%|██████████| 2/2 [00:00<00:00,  3.75it/s]
```

```
dice_exp_random.visualize_as_dataframe(show_only_changes=True)
```

```
Query instance (original outcome : 0)
```

	age	workclass	education	marital_status	occupation	race	gender	hours_per_week	income
0	27	Private	School	Single	Blue-Collar	White	Male	40	0

```
Diverse Counterfactual set (new outcome: 1.0)
```

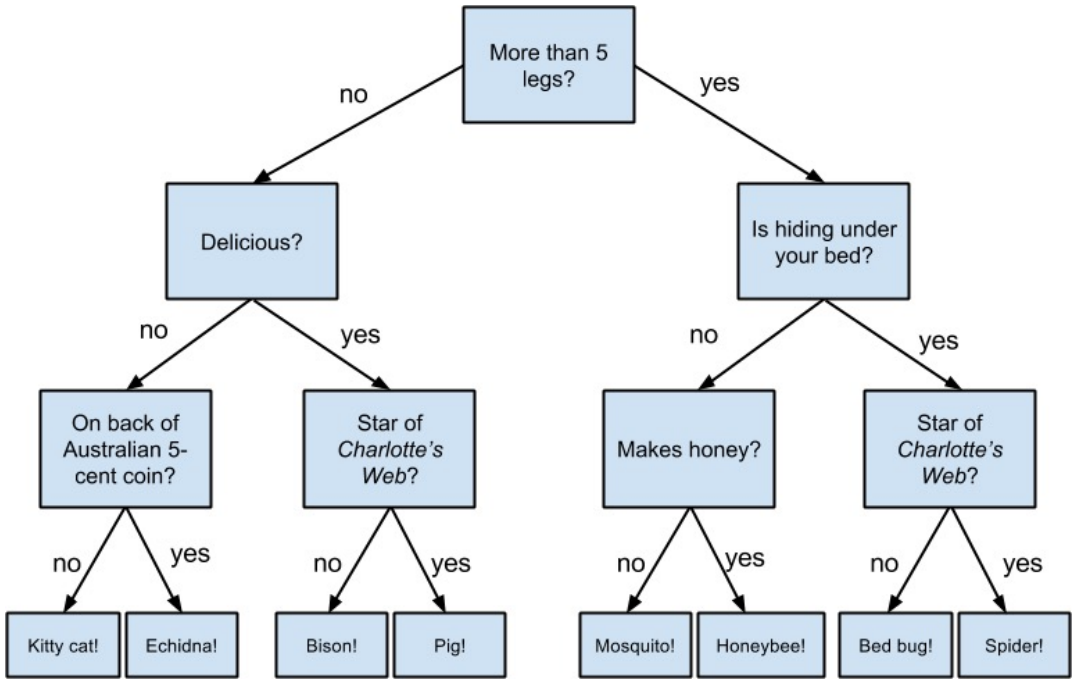
	age	workclass	education	marital_status	occupation	race	gender	hours_per_week	income
0	35.0	-	Prof-school	-	Sales	-	-	-	1
1	65.0	-	-	Married	-	-	-	68.0	1

AGENDA: AI EXPLAINABILITY IN PRACTICE

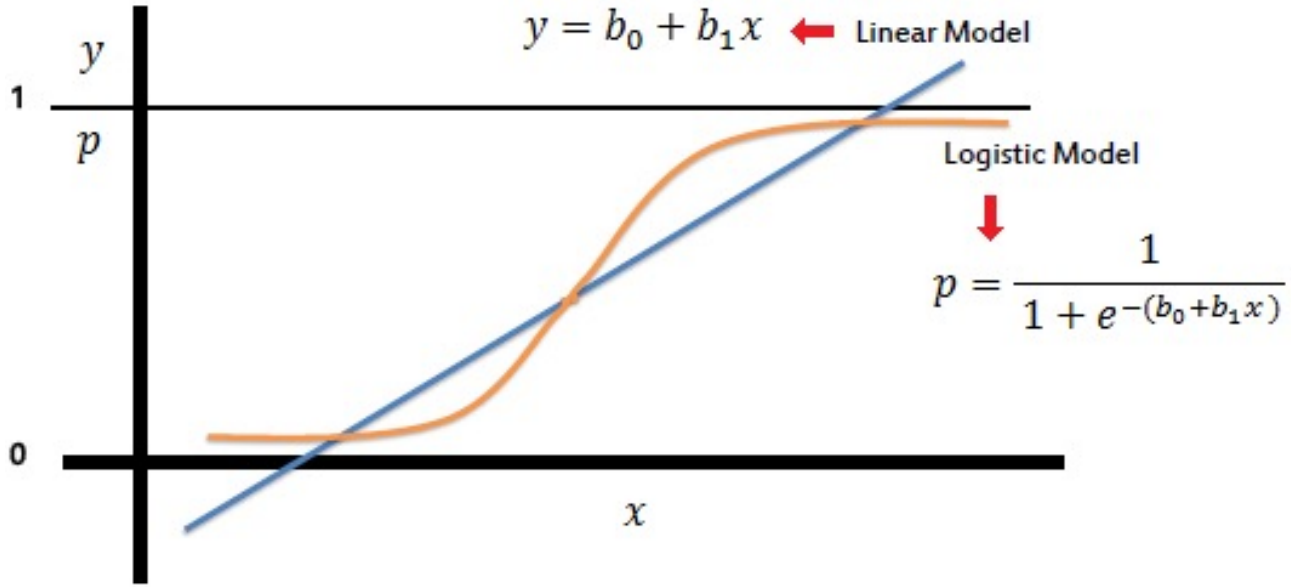
- Types and Methods for Explainable AI
- One explanation does not Fit All:
  - AI Explainability 360 (AIX360) toolkit
  - Taxonomy
  - Github & other resources
- Preview of Oct 21st Hands-on Class.
  - Installation instructions
  - Demo

EXAMPLES OF INTERPRETABLE AND NON-INTERPRETABLE MODELS

Decision Tree



Linear & logistic regression Models

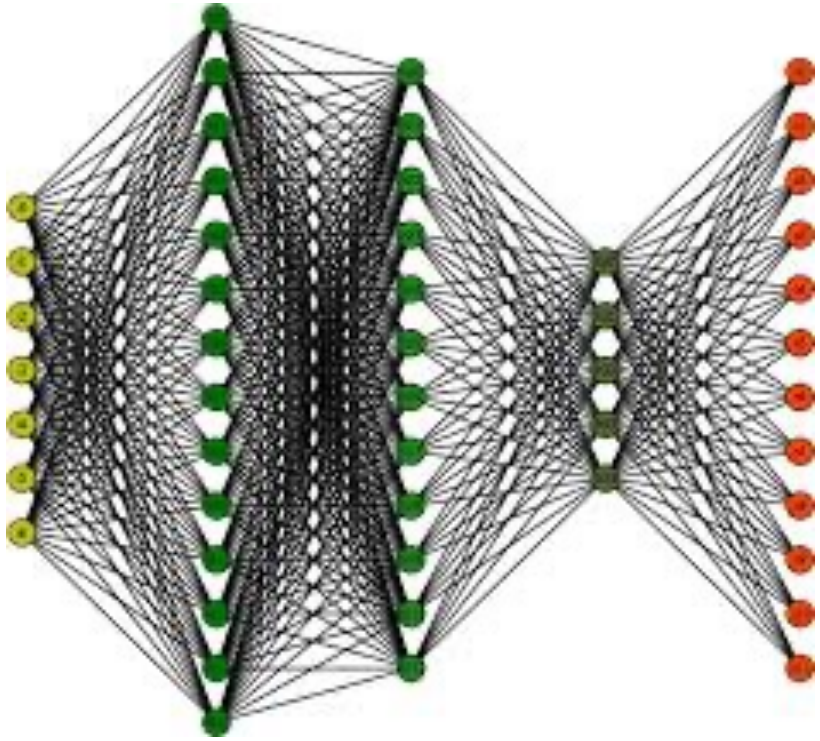


Interpretable?

YES

EXAMPLES OF INTERPRETABLE AND NON-INTERPRETABLE MODELS

Neural Network



XGBoost

Interpretable?

NO

ONE EXPLANATION DOES NOT FIT ALL

Three dimensions of explainability: many ways to explain.

directly interpretable **vs.** **post hoc interpretation**

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.

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)**

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).

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)**

The interpretation is simply presented to the user.

The user can interact with interpretation.

Post hoc (local) interpretation

Locally Interpretable Model Agnostic Explanations (LIME)

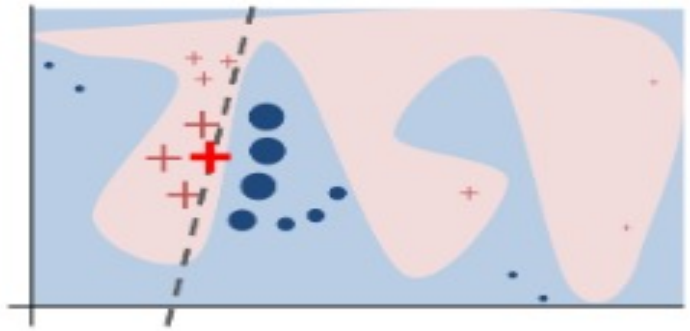
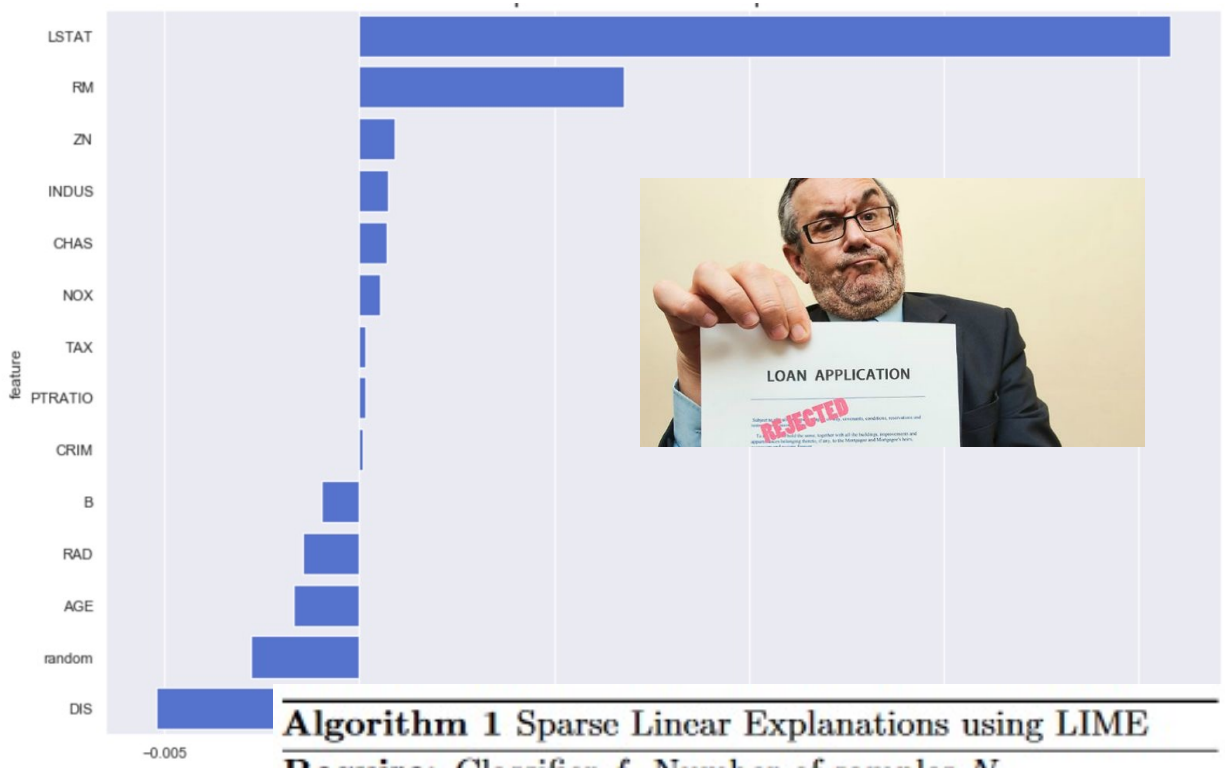


Figure 1. Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background. The bright bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity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

(Ribeiro et. al. 2016)



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$(\mathcal{Z}, K)$ ▷ with z'_i as features, $f(z)$ as target
return w

Post hoc (local) interpretation

SHAP (Lundberg and Lee, Strumbelj and Kononenko)

Definition 3 *A coalitional form game is a tuple* $\langle N, v \rangle$, *where* $N = \{1, 2, \ldots, n\}$ *is a finite set of n players, and* $v : 2^N \to \Re$ *is a characteristic function such that* $v(\varnothing) = 0$.

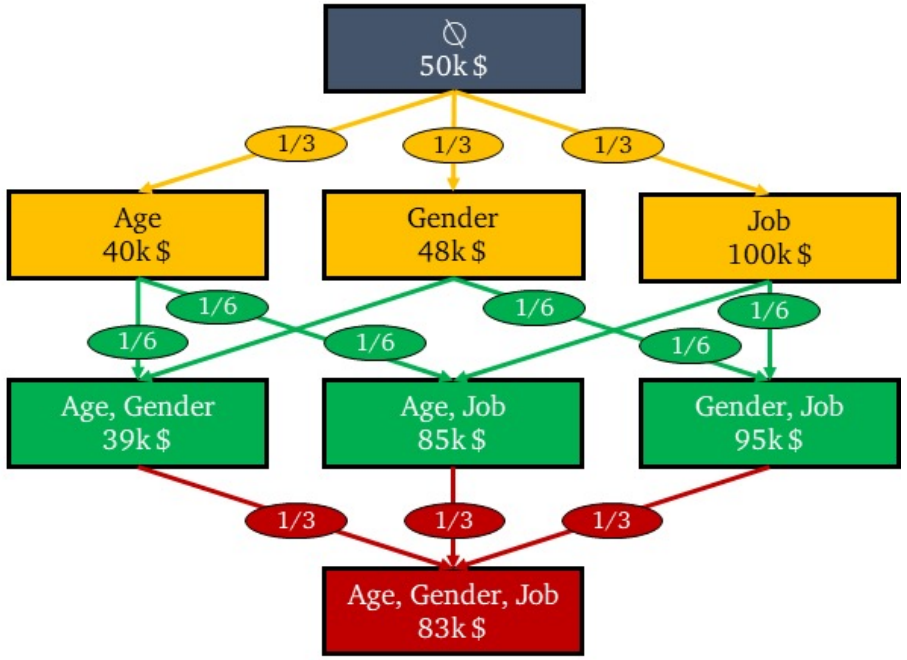
$$shapley(i) = \sum_{S \subseteq N \setminus \{i\}, s=|S|} \frac{(n-s-1)!s!}{n!}(v(S \cup \{i\}) - v(S)), \quad i = 1, \ldots, n.$$

(Lloyd Shapley, 1953)

$$\begin{aligned} SHAP_{Age}(x_0) = &[(1 \times \binom{3}{1}]^{-1} \times MC_{Age,\{Age\}}(x_0) + \\ &[(2 \times \binom{3}{2}]^{-1} \times MC_{Age,\{Age,Gender\}}(x_0) + \\ &[(2 \times \binom{3}{2}]^{-1} \times MC_{Age,\{Age,Job\}}(x_0) + \\ &[(3 \times \binom{3}{3}]^{-1} \times MC_{Age,\{Age,Gender,Job\}}(x_0) + \\ = &\frac{1}{3} \times (-10k\$) + \frac{1}{6} \times (-9k\$) + \frac{1}{6} \times (-15k\$) + \frac{1}{3} \times (-12k\$) \\ = &-11.33k\$ \end{aligned}$$

Example

Prediction(sample)
= Average +
shap(age) + shap(gender) + shap(job)
(commonly viewed as *force plots*)



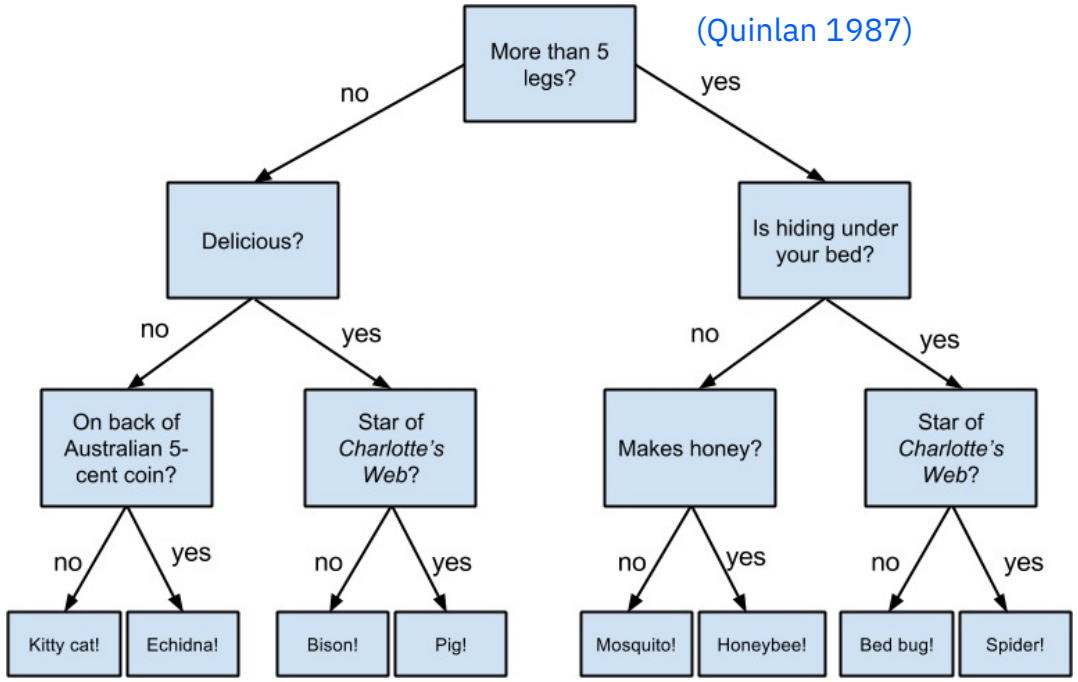
(Samuele Mazzanti, Medium article)

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

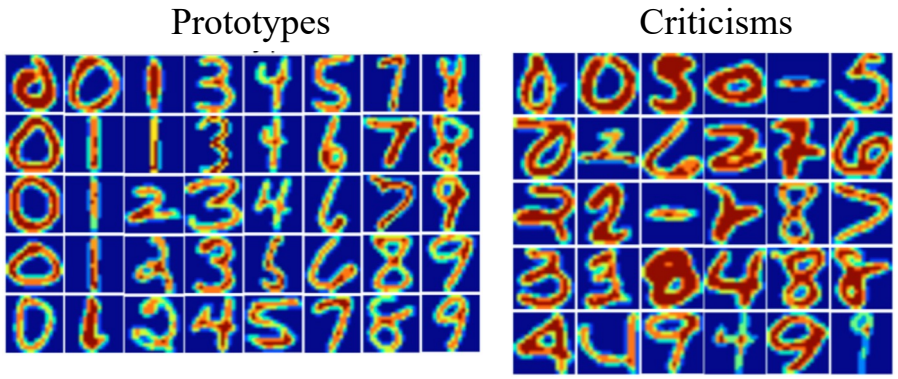
(Wang and Rudin 2016)

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.

Post hoc (local) interpretation

MMD-Critic (Kim et. al. 2016)

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



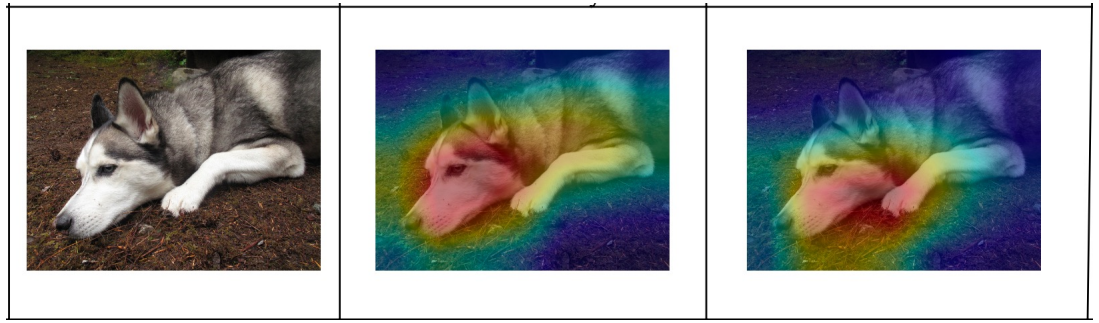
Prototypes and criticisms for handwritten digits dataset

Health care applications

Post hoc (local) interpretation

Saliency Maps

(Sinmoyan et. al. 2013)



(Use carefully, Cynthia Rudin, 2019)

Other methods
- Grad-cam (Selvaraju, et al)
- LRP (Heatmapping.org)

Other ‘similar’ Concepts
- Permutation feature importance
- Influence functions

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.

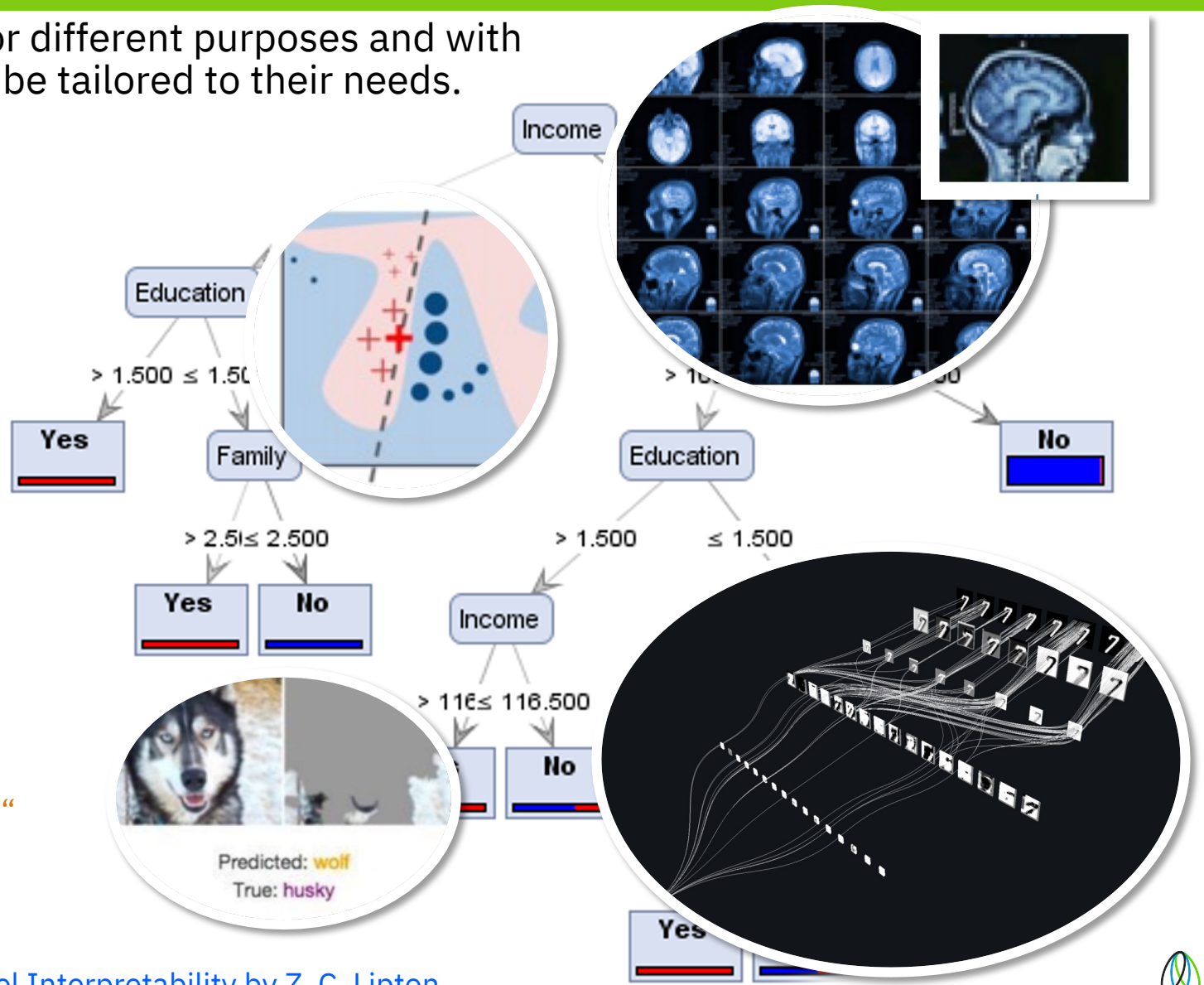
Why: ensure fairness for constituents

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



Recommend: The Mythos of Model Interpretability by Z. C. Lipton

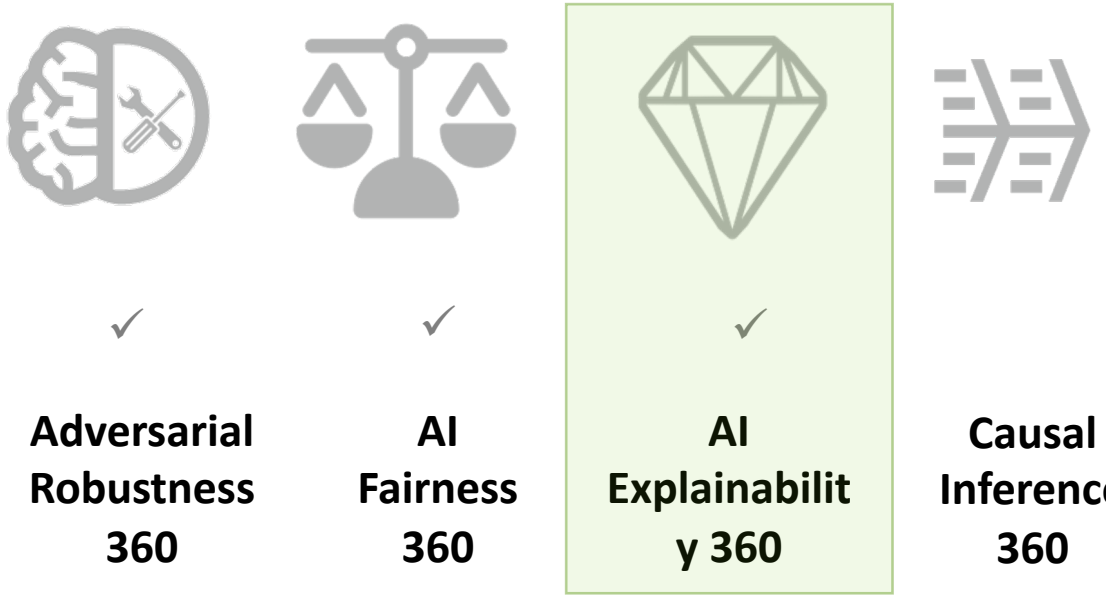
AIX360: AI EXPLAINABILITY 360 TOOLKIT

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.

AI Explainability 360 (Linux Foundation AI)	
Explainability Algorithms	10 ways to explain data and AI models
Repositories	github.com/Trusted-AI/AIX360
Interactive Experience	aix360.mybluemix.net
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



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 OPENSOURCE LANDSCAPE

Toolkit	Data Explanations	Directly Interpretable	Local Post-Hoc	Global Post-Hoc	Self Explaining	Metrics
AIX360	✓	✓	✓	✓	✓	✓
Alibi			✓			
Skater		✓	✓	✓		
H2O		✓	✓	✓		
InterpretML		✓	✓	✓		
EthicalML-XAI				✓		
DALEX			✓	✓		
tf-explain			✓	✓		
iNNvestigate			✓			
modelStudio	✓	✓	✓	✓		
ELI5		✓	✓	✓		
Iml		✓	✓	✓		
Captum			✓			
WIT	✓		✓	✓		

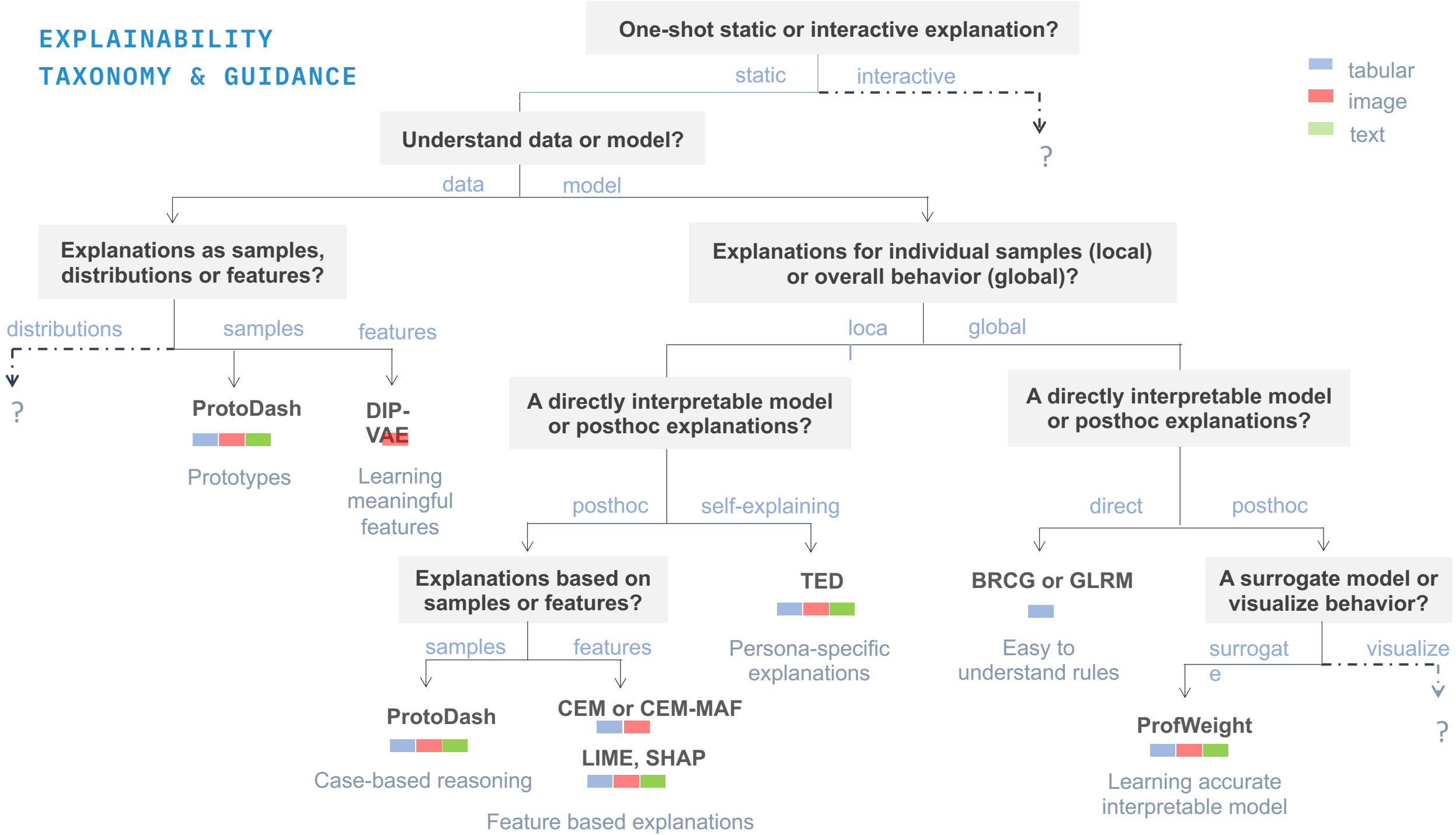
AIX360 also provides demos, tutorials, and guidance on explanations for different use cases.

See Also:

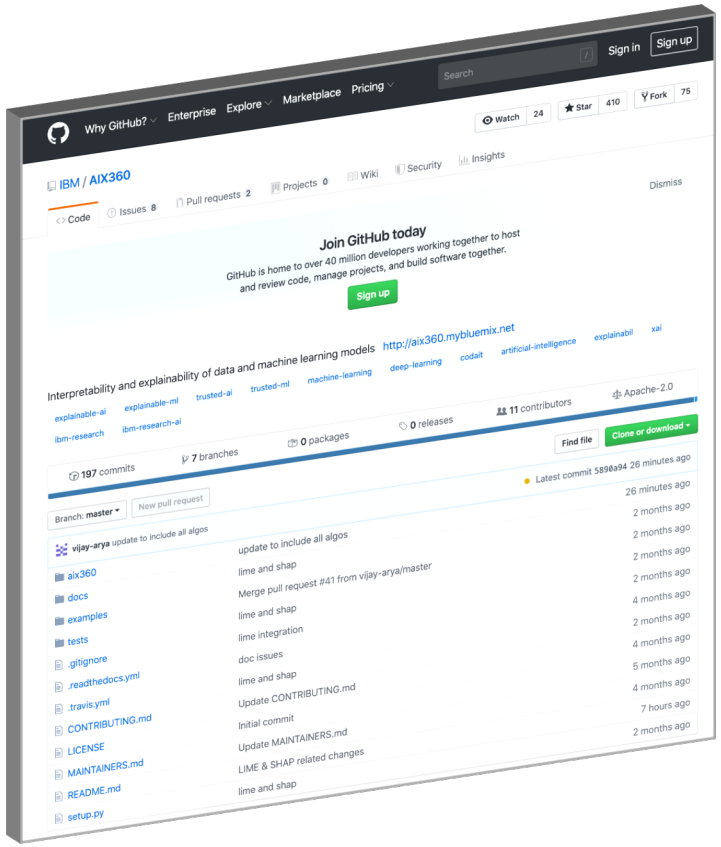
“One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques.”: https://arxiv.org/abs/1909.03012v1

“AI Explainability 360: Impact and Design”: https://arxiv.org/abs/2109.12151

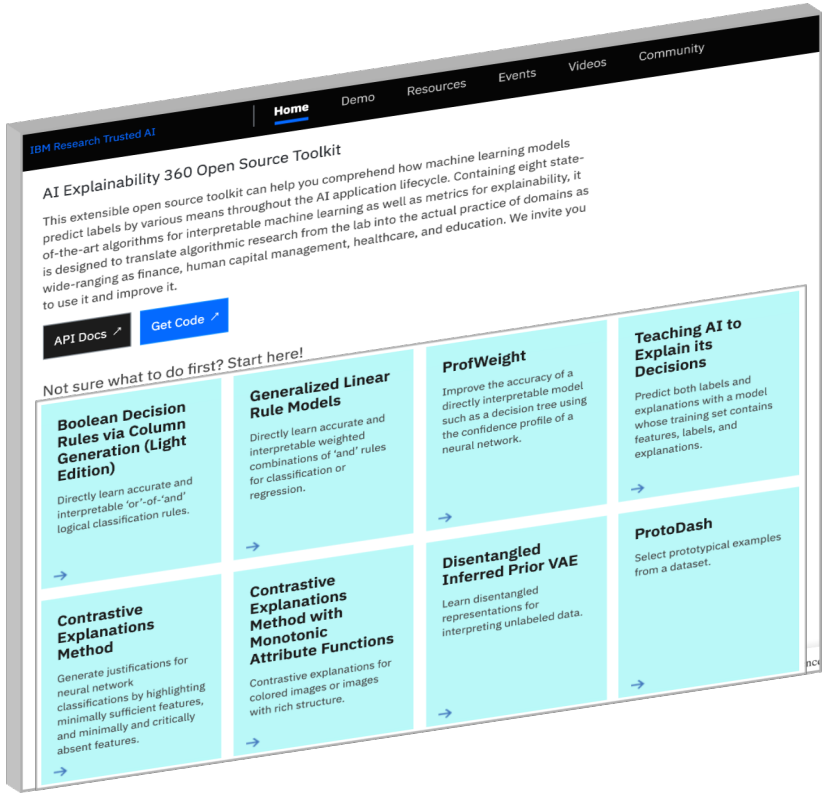
EXPLAINABILITY TAXONOMY & GUIDANCE



HANDS-ON SESSION: AI EXPLAINABILITY 360 (OCT 21)



https://github.com/Trusted-AI/AIX360



http://aix360.mybluemix.net/data

- FICO Tutorial notebook: Directly Interpretable, Prototypes, & Counterfactual explanations
- Health and Nutrition survey tutorial notebook: Prototype explanations /Or a SHAP example