A Brief Overview to Interpretable Machine Learning

Isabel Valera
July 27, 2019

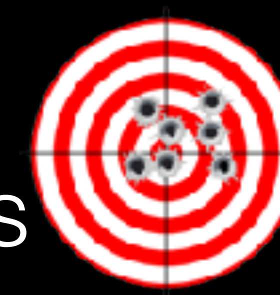
Slides by Amir Karimi

University admissions

Risk stratification for patients

Insurance policy assignment

Autonomous military systems



 Applying for a loan.

Autonomous vehicles

Recidivism prediction

Predictive policing

Targeted Political Ads

Job hiring

Interpret: *to explain or to present in understandable terms* *[Merriam-Webster]*

Interpretability

Interpretability (ML): *ability to explain or to present in understandable terms to a human [Doshi-Velez & Kim, 2017]*

Explanations: the currency in which we exchange beliefs *[Lombrozo, 2006]*

University admissions

Risk stratification for patients

Applying for a loan

Autonomous vehicles

Job hiring

Mechanisms?

Definitions?

Interpretability (ML): *ability to explain or to present in understandable terms to a human [Doshi-Velez & Kim, 2017]*

Measures?

Stakeholders?

Insurance policy assignment

Targeted Political Ads

Predictive policing

Autonomous military systems

Recidivism prediction

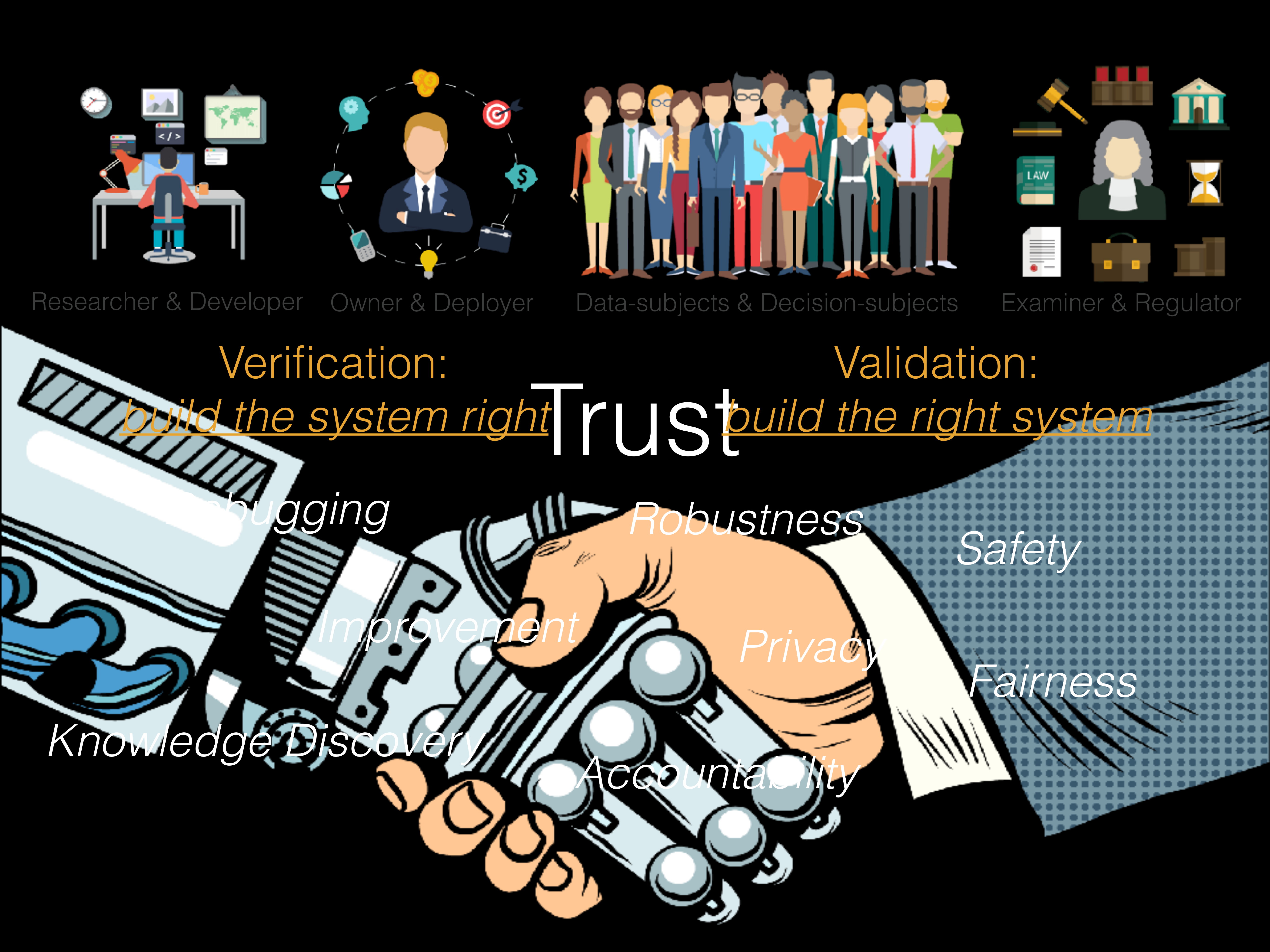
Researcher & Developer

Owner & Deployer

Stakeholders

Examiner & Regulator

Data-subjects & Decision-subjects


Researcher & Developer
Owner & Deployer
Data-subjects & Decision-subjects
Examiner & Regulator
LAW
Verification:
build the system right
Trust
Validation:
build the right system
Debugging
Robustness
Safety
Improvement
Privacy
Fairness
Knowledge Discovery
Accountability

Definitions

Simulatability

Interestingness

Decomposability

Explainability

Interpretability

Transparency

Informativeness

Justifiability

Understandability

Comprehensibility

Visibility

Intelligibility

Legibility

Scrutability

Usabillity

Definitions

Interpretability

Transparency

(model interpretability)

(inherent interpretability)

the level to which a system provides information about its internal workings or structure, and the data it has been trained with

Explainability

(decision interpretability)

(post-hoc interpretability)

the level to which a system can provide clarification (explanations) for the cause of its decisions/outputs.

Zachary Lipton 2016

Been Kim, Finale Doshi-Velez 2017

Leilani H. Gilpin et al. 2018

Richard J. Tomsett et al. 2018

"The truth,
the whole truth,
and nothing but the truth"

Contrastive

Selective

Social

Measures

Functionally-grounded

Human-grounded

Application-grounded

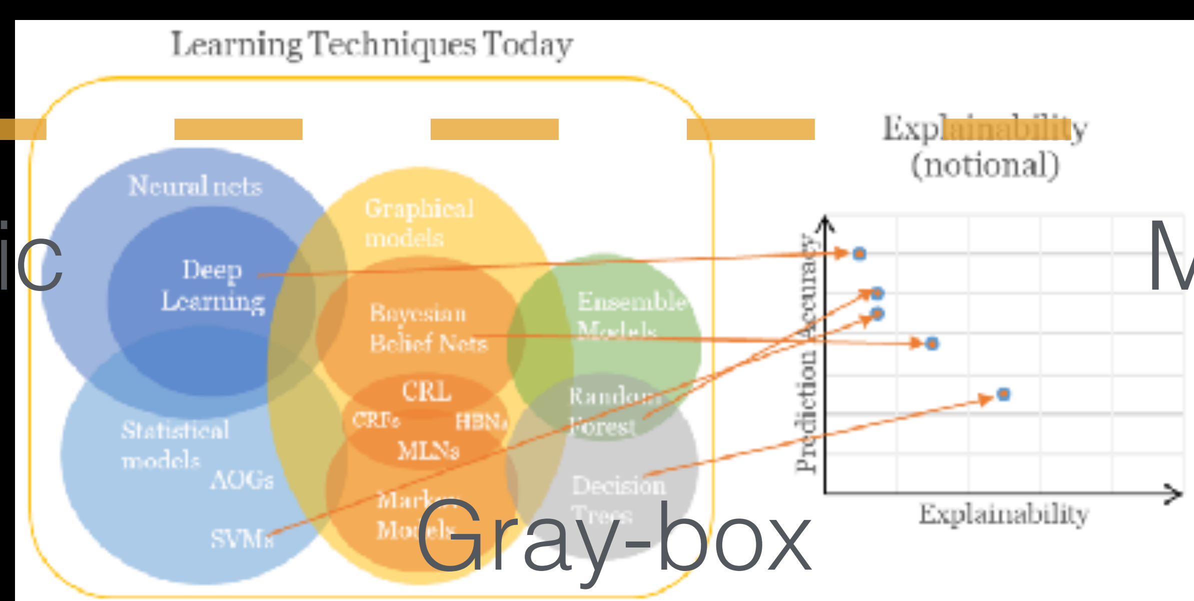
Adrian Weller 2017

Finale Doshi-Velez, Been Kim 2017

Tim Miller 2018

Mechanisms

Transparency (inherent/model interpretability)	Explainability (post-hoc/decision interpretability)
Simulatability	Feature-based (attribution)
Decomposability	Instance-based
Algorithmic transparency	Surrogate Models



Model-specific ⟷ Model-agnostic
Global — Local
White-box — Gray-box — Black-box

Transparency (Simulatability) Example: Bayesian Rule Sets

Objective
Build classifiers that are comprised of a small number of short rules. Rules are restricted to disjunctive normal form (DNF), e.g., if X satisfies (condition A AND condition B) OR (condition C) OR · · ·, then Y = 1

Related work
Greedy methods where rules are added to the model one by one, do not generally produce high-quality sparse models.

Example (rule selection)
Predicting if a customer will accept a coupon for a nearby coffee house, where the coupon is presented by their car's mobile recommendation device

if a customer (goes to coffee houses $\geq$ once per month AND destination = no urgent place AND passenger $\neq$ kids)

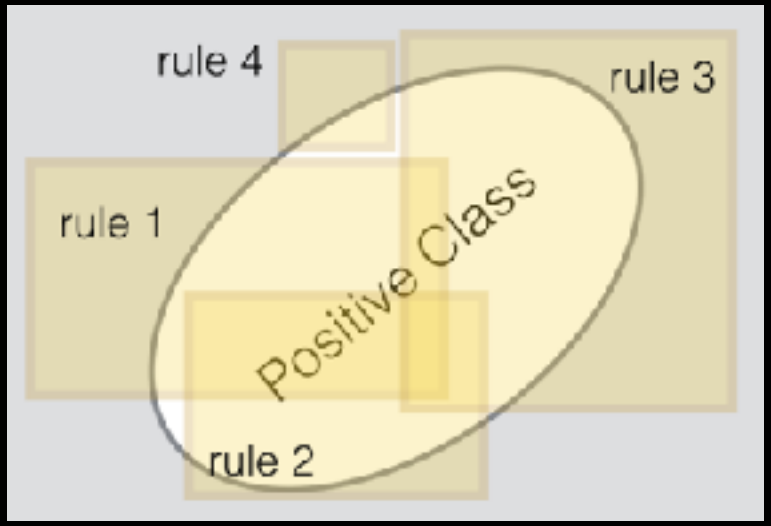
OR (goes to coffee houses $\geq$ once per month AND the time until coupon expires = one day)
then
predict the customer will accept the coupon for a coffee house.

Wang et al. 2016

Transparency (Simulatability) Example: Bayesian Rule Sets

Set of rules:

$$\mathcal{A} = \cup_{l=1}^{L} \mathcal{A}_l.$$



Beta-Binomial Prior (rule selection)

$$\mathcal{A}_l \sim \text{Bernoulli}(p_l) \qquad p_l \sim \text{Beta}(\alpha_l, \beta_l)$$

$$p(A; \{\alpha_l, \beta_l\}_l) = \prod_{l=1}^{L} \int \frac{B(M_l + \alpha_l, |\mathcal{A}_l| - M_l + \beta_l)}{B(\alpha_l, \beta_l)}$$

Poisson Prior (rule generation)

$$M \sim \text{Poisson}(\lambda) \qquad L_m \sim \text{Truncated} - \text{Poisson}(\eta)$$

$$p(A; \lambda, \eta) = \frac{1}{w(\lambda, \eta)} \text{Poisson}(M; \lambda) \prod_{m=1}^{M} \text{Poisson}(L_m; \eta) \frac{1}{\binom{J}{L_m}} \prod_{k=1}^{L_m} \frac{1}{K_{v_m,k}}$$

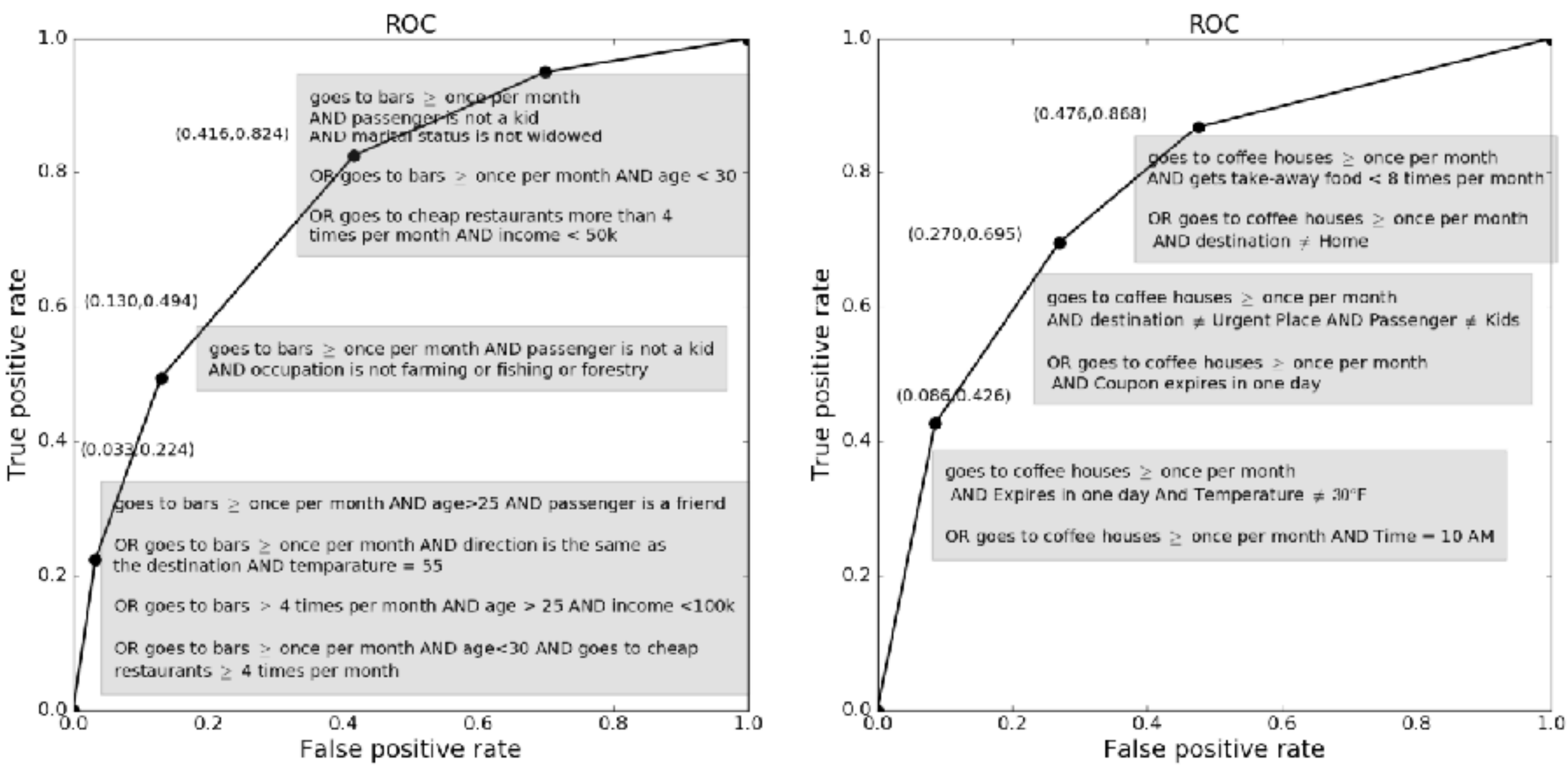
Likelihood

$$p(S|A, \alpha_+, \beta_+, \alpha_-, \beta_-) = \frac{B(\text{TP} + \alpha_+, \text{FP} + \beta_+)}{B(\alpha_+, \beta_+)} \frac{B(\text{TN} + \alpha_-, \text{FN} + \beta_-)}{B(\alpha_-, \beta_-)}$$

Inference An approximate inference method using association rule mining and a randomized search algorithm is used to find optimal BRS MAP models.

Wang et al. 2016

Transparency (Simulatability) Example: Bayesian Rule Sets



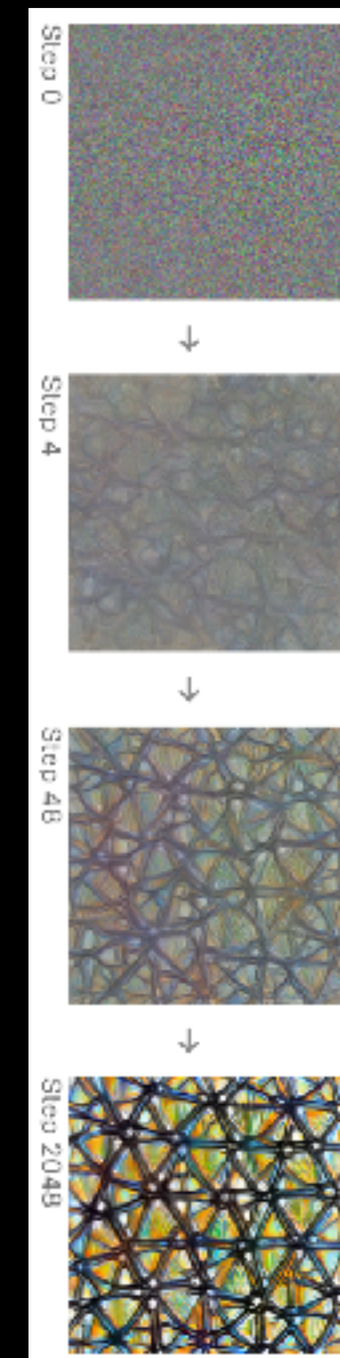
(a) Coupons for bars

(b) Coupons for coffee houses

Figure 11: ROC for data set of coupons for bars and coffee houses.

Wang et al. 2016

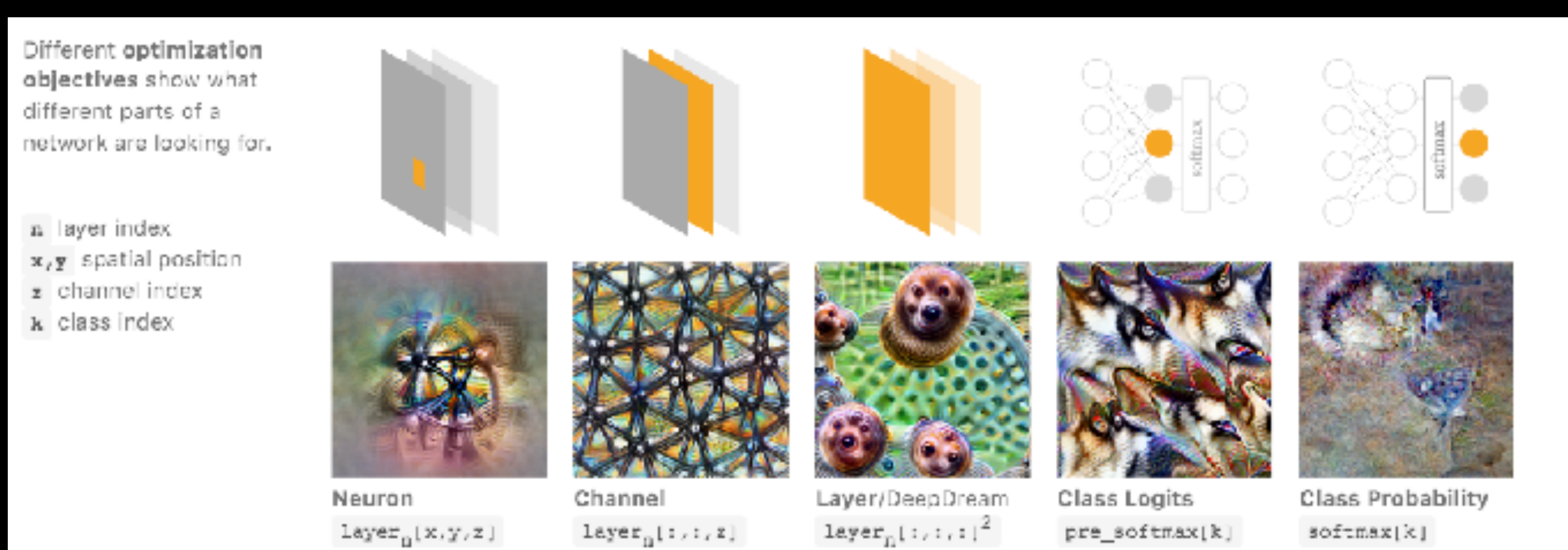
Transparency (Decomposability) Example: Feature Visualization



Definition What is a unit looking for?

Visualization by optimization For a unit of a neural network, find the input that maximizes the activation of that unit.

$$I^* = \arg\max_{I} \sum \hat{f}_{n,x,y,z}(I)$$



Erhan et al. 2009
Springenberg et al. 2014
Olah et al. 2017
Nguyen et al. 2017
Molnar 2019

Transparency (Decomposability) Example: Feature Visualization

Limitations

- Many visualization images are not interpretable and lack human concepts
- Fails to describe complex inter-unit interactions
- There are too many units to consider
- Limited to CNNs for image recognition
- Lacking human semantical concepts.

Erhan et al. 2009
Springenberg et al. 2014
Olah et al. 2017
Nguyen et al. 2017
Molnar 2019

Explainability (Feature-based) Example: Attribution (Saliency Maps)

Definition How does the input affect the output?

Objective Identify exactly which regions of an image are being used for discrimination.

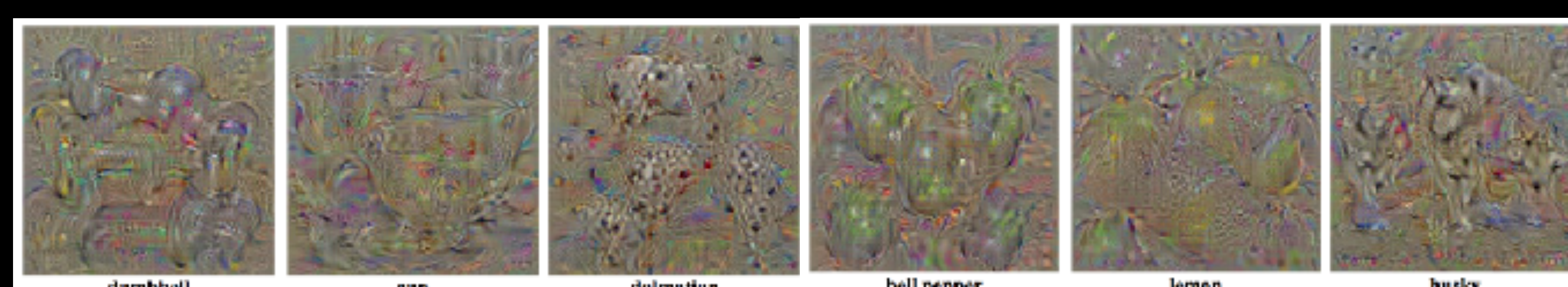
Linear score model for class c

$$S_c(\mathrm{I}) = w_c^T \mathrm{I} + b_c$$

Influence: magnitude of w_i correspond to importance of pixel i for classifying class c

Nonlinear score model for class c, near image I_0

$$S_c(\mathrm{I}) \approx w_c^T \mathrm{I} + b_c \qquad w = \left.\frac{\partial S_c}{\partial I}\right|_{I_0}$$



Simonyan et al. 2013 Kindermans et al. 2017
Fong & Vedaldi 2017 Sundararajan et al. 2017

Explainability (Feature-based) Example:
Attribution (Saliency Maps)

Class-Activation Maps (CAM)

$f_k(x, y)$ activation of unit *k* in the last conv layer at spatial *(x, y)*

$F_k = \Sigma_{x,y} f_k(x, y)$ *k*-th unit in the fully connected layer

$S_c = \Sigma_k w_k^c F_k = \Sigma_{x,y} \Sigma_k w_k^c f_k(x, y)$ score for class c

$M_c(x, y) = \Sigma_k w_k^c f_k(x, y)$ class activation map of class c

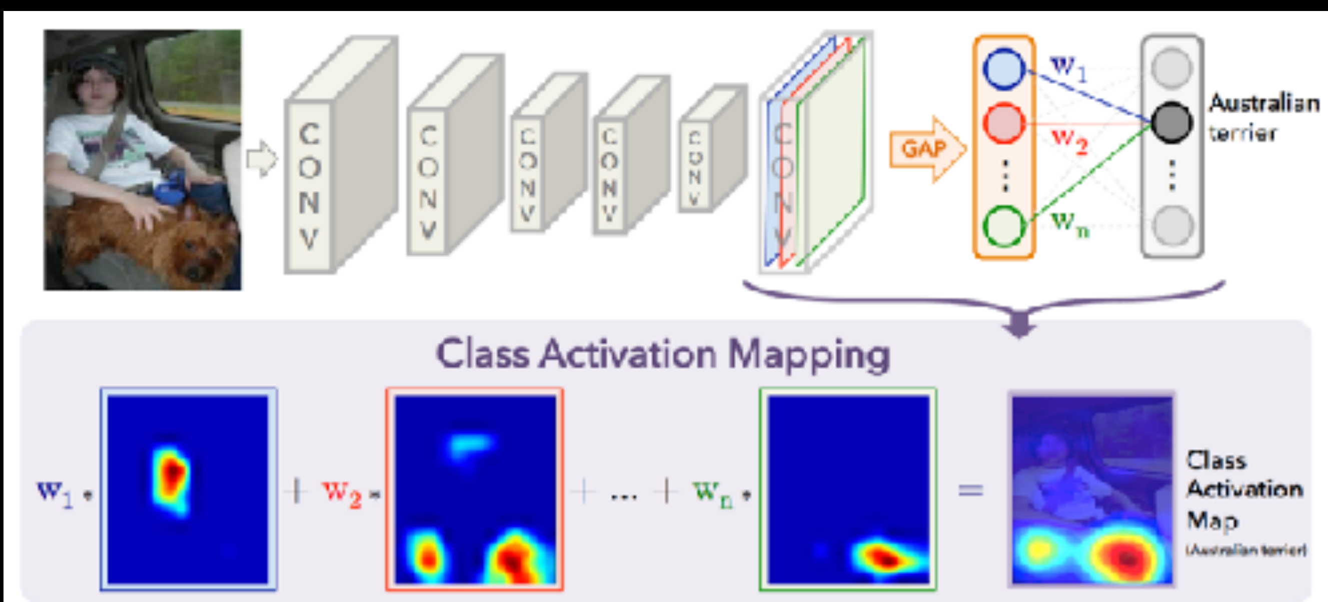
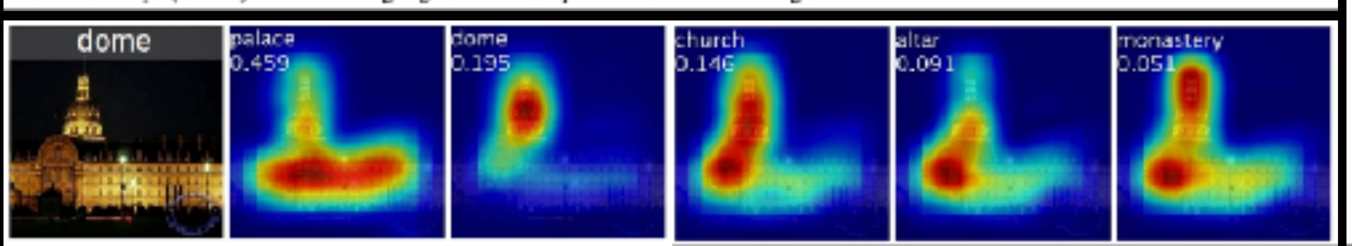


Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.



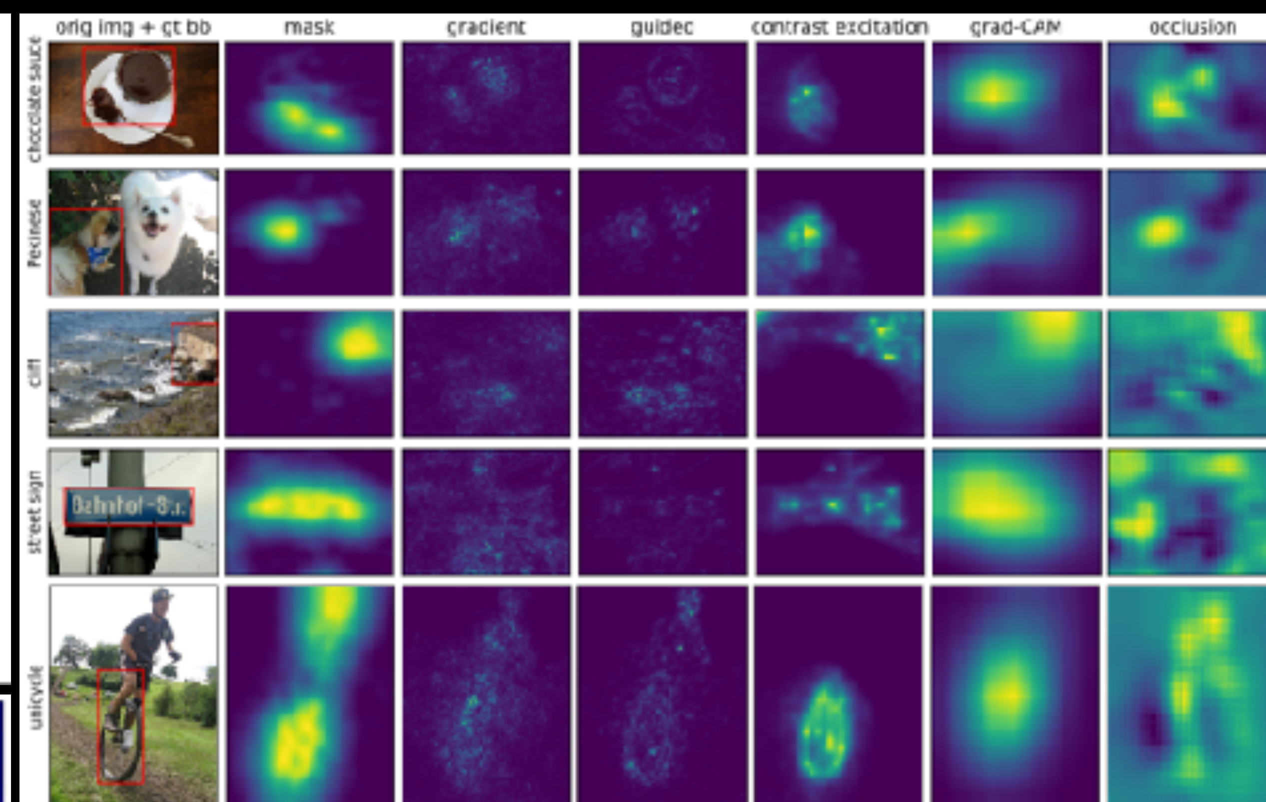


Figure 2. Comparison with other saliency methods. From left to right: original image with ground truth bounding box, learned mask subtracted from 1 (our method), gradient-based saliency [15], guided backprop [16, 8], contrastive excitation backprop [20], Grad-CAM [14], and occlusion [19].

Limitations There's reason to think that all our present answers aren't quite right

The Unreliability of Saliency Methods (Kindermans et al. 2017)
Interpretation of Neural Networks is Fragile (Ghorbani et al. 2017)

Zhou et al. 2015
Selvaraju et al. 2016

Explainability (Instance-based) Example: Counterfactual Explanations

Counterfactual explanation: "You were denied a loan because your annual income was £30,000. If your income had been £45,000, you would have been offered a loan."

Nearest counterfactual explanation: The set of features resulting in the desired prediction while remaining at min distance from the original set of features for the individual.

Additional considerations: Plausibility, Diversity

$$x^* \in \underset{x^{\mathrm{CF}}}{\operatorname{argmin}} \quad d(x^{\mathrm{F}}, x^{\mathrm{CF}})$$

$$s.t. \quad f(x^{\mathrm{F}}) \neq f(x^{\mathrm{CF}})$$

$$x^{\mathrm{CF}} \in \mathcal{P}\mathrm{lausible}$$

Sandra Wachter et al. 2017

Methods for Generating Counterfactual Explanations

Counterfactual Explanations without Opening the Black Box - Wachter et al. 2017

Interpretable Predictions of Tree-based Ensembles via Feature Tweaking - Tolomei et al. 2017

Actionable Recourse in Linear Classification - Ustan et al. 2018

Minimum Observable Counterfactuals - Google PAIR team 2019

Limitations of current methods:

- Lacking closeness guarantees
- Linear / convex models
- Homogenous data spaces
- Limited coverage
- Differentiable distance metrics

Model Agnostic Counterfactual Explanations for Consequential Decisions - Karimi et al. 2019

Counterfactual Explanations without Opening the Black Box

Problem setup $\arg\min_{x^{\mathrm{CF}}} \max_{\lambda} d(x^{\mathrm{F}}, x^{\mathrm{CF}}) + \lambda(f(x^{\mathrm{CF}}) - c)$

Optimization Maximization over λ is done by iteratively solving for x^{CF} (using ADAM) and increasing λ until a sufficiently close solution is found.

Distance function $d(x^{\mathrm{F}}, x^{\mathrm{CF}}) = \sum_{d \in D} \frac{x_d^{\mathrm{F}} - x_d^{\mathrm{CF}}}{\mathrm{MAD}_d}$

$$\mathrm{MAD}_d = \mathrm{median}_{n \in P}(|X_{n,d} - \mathrm{median}_{m \in P}(X_{m,d})|)$$

- Captures intrinsic volatility
- Robustness to outliers
- L_1 norm induces sparsity

Example Counterfactuals Predict whether women of Pima heritage are at risk of diabetes.
Fully-connected neural network (8-20-20-1) —> risk score ∈ [0,1]
"If your Plasma glucose concentration was 158.3 & your 2-Hour serum insulin level was 160.5, your risk score would have been 0.5"

Limitations

- Restricted to differentiable functions f() and distances d()
- Cannot accommodate heterogeneous data
- Lacking closeness guarantee

Sandra Wachter et al. 2017

Interpretable Predictions of Tree-based Ensembles via Feature Tweaking

Ensemble formulation $\hat{f} = \phi(\hat{h}_1, \cdots, \hat{h}_K) \quad \hat{h}_k : \mathcal{X} \to \mathcal{Y}$

Majority voting ϕ $\hat{f}(x) = -1 \iff \left(\sum_{k=1}^{K} \hat{h}_k(x)\right) \leq 0$

Tree-based classifiers $\{\hat{h}_k\}_{i=1}^{K} \equiv \mathcal{T} = \{T_k\}_{i=1}^{K}$

Objective tweak the original input feature vector x so as to adjust the prediction made by the ensemble from -1 to +1

Gabriele Tolomei et al. 2017

Interpretable Predictions of Tree-based Ensembles via Feature Tweaking

Positive and Negative Paths

$$p_{k,j} = \{(x_1 \gtreqless \theta_1), \cdots, (x_n \gtreqless \theta_d)\} \; j^{th} \text{ path of } k^{th} \text{ tree}$$

$$P_k^+ = \bigcup_{j \in T_k} p_{j,k}^+, \quad P_k^- = \bigcup_{j \in T_k} p_{j,k}^-, \quad P_k = P_k^+ \bigcup P_k^-$$

Identify instances classified as +1

For each $p_{k,j}^+ \in P_k^+$, associate instance

$x_j^+ \in \mathcal{X}$ that satisfies the path.

Restrict instances to ε-satisfactory (for tree *k*)

$$x_{j(\epsilon)}^+[i] = \begin{cases} \theta_i - \epsilon & \text{if the i-th condition is } (x_i \leq \theta_i) \\ \theta_i + \epsilon & \text{if the i-th condition is } (x_i > \theta_i) \end{cases}$$

Finding nearest counterfactual (for ensemble)

$$x^{\mathrm{CF}} = \underset{\hat{f}(x_{j(\epsilon)}^+) = +1}{\arg\min} \; d(x^{\mathrm{F}}, x_{j(\epsilon)}^+)$$

Limitations

- Restricted to counterfactuals of the ε-satisfactory form
- Only applies to ensembles of binary tree base classifiers
- Lacking existence & plausibility guarantees
- $O(2^d)$ complexity (d: number of features)

Actionable Recourse in Linear Classification

Individual $\textit{features}\ x = [1, x_1, \cdots, x_d] \subseteq \mathcal{X}_0 \cup \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_d$

and a binary $\textit{label}\ y = \{-1, +1\}$

Linear Classifier $f(x) = \operatorname{sign}(\langle w, x \rangle) \quad w = [w_0, w_1, \cdots, w_d] \subseteq \mathbb{R}^{d+1}$

Problem Formulation

$$\begin{aligned} \min \quad & \operatorname{cost}(a; x) \\ \text{s.t.} \quad & f(x + a) = 1 \\ & a \in A(x) \end{aligned}$$

$$a_j \in A_j(x_j) \subseteq \{a_j \in \mathbb{R} | a_j + x_j \in \mathcal{X}_j\}$$

$A_j(x) = \{0\}$ if feature *j* is immutable

$\operatorname{cost}(\cdot; x) : A(x) \rightarrow \mathbb{R}_+$ is a user-specified cost

Berk Ustun et al. 2018

Actionable Recourse in Linear Classification

Alternative Formulation (Integer Linear Programming)

$$
\begin{aligned}
\min \quad & \text{cost} \\
\text{s.t.} \quad & \text{cost} = \sum_{j \in J_A} \sum_{k=1}^{m_j} c_{jk} v_{jk} && && \text{(2a)} \\
& \sum_{j \in J_A} w_j a_j \geq \sum_{j=0}^{d} w_j x_j && && \text{(2b)} \\
& a_j = \sum_{k=1}^{m_j} a_{jk} v_{jk} && j \in J_A && \text{(2c)} \\
& 1 = u_j + \sum_{k=1}^{m_j} v_{jk} && j \in J_A && \text{(2d)} \\
& a_j \in \mathbb{R} && j \in J_A && \\
& u_j \in \{0, 1\} && j \in J_A && \\
& v_{jk} \in \{0, 1\} && k = 1 \ldots m_j,\ j \in J_A &&
\end{aligned}
$$

Constraints

(2a) precomputed cost
$c_{jk} = \text{cost}(x_j + a_{jk}; x_j)$

(2b) Counterfactuals must flip the prediction

(2c, 2d) Restrict the actions a_j to a grid of $m_j + 1$ feasible values $a_j \in \{0, a_{j1}, \ldots, a_{jm_j}\}$

Solution

Standard Integer Linear Program solvers, e.g., CPLEX

Limitations

- Restricted to linear models
- *Cannot handle heterogeneous data*

MACE: Model-Agnostic Counterfactual Explanations

$$x^* \in \underset{x^{\mathrm{CF}}}{\operatorname{argmin}} \quad d(x^{\mathrm{F}}, x^{\mathrm{CF}})$$
$$s.t. \quad f(x^{\mathrm{F}}) \neq f(x^{\mathrm{CF}})$$
$$x^{\mathrm{CF}} \in \mathcal{P}\text{lausible}$$

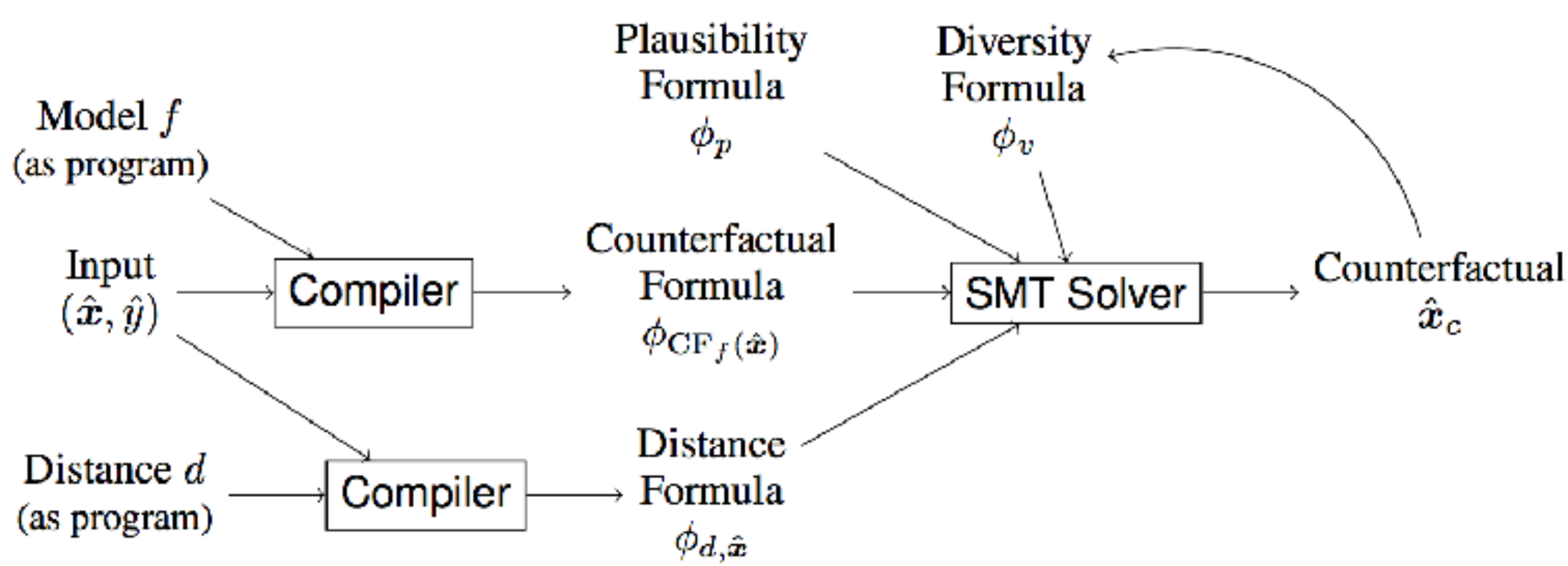


Figure 1: Architecture Overview for Model-Agnostic Counterfactual Explanations (MACE)

First-Order Predicate Logic

Function Symbols
(e.g., addition $+$, multiplication $\times$)

↓

Expressions
(e.g., $(-x + 2) \times (y + 3)$)

Predicate Symbols
(e.g., equality $=$, lesser than $<$)

↓

Atomic Formulae
(e.g., $e < e'$, $e \leq e'$, $e = e'$)

(Quantifier-free) Formula: Boolean combinations ($\wedge$, $\vee$, $\neg$) of atomic formulae or clauses, e.g., $[\,(x + 2) \times (y + 3) \leq x \times y + 16\,] \wedge [1 \leq x]$

A formula is satisfiable if $\exists$ a solution satisfying the all atomic formulae, e.g., $x \rightarrow 2$, $y \rightarrow 1$ assigns true because $[\,16 \leq 18\,] \wedge [1 \leq 2]$

Standard SMT (Satisfiability Modulo Theories) solvers can verify the satisfiability of a formula, e.g., Z3, CVC3, pySMT

Programs, Static Single Assignment Form (SSA), Path Formulae, and the Characteristic Formula ϕ

Model: $f : \mathcal{X} \to \{0, 1\}$

Program: a collection of variables, constants, function symbols, assignment commands, if-else conditionals, for-loops, and return statements

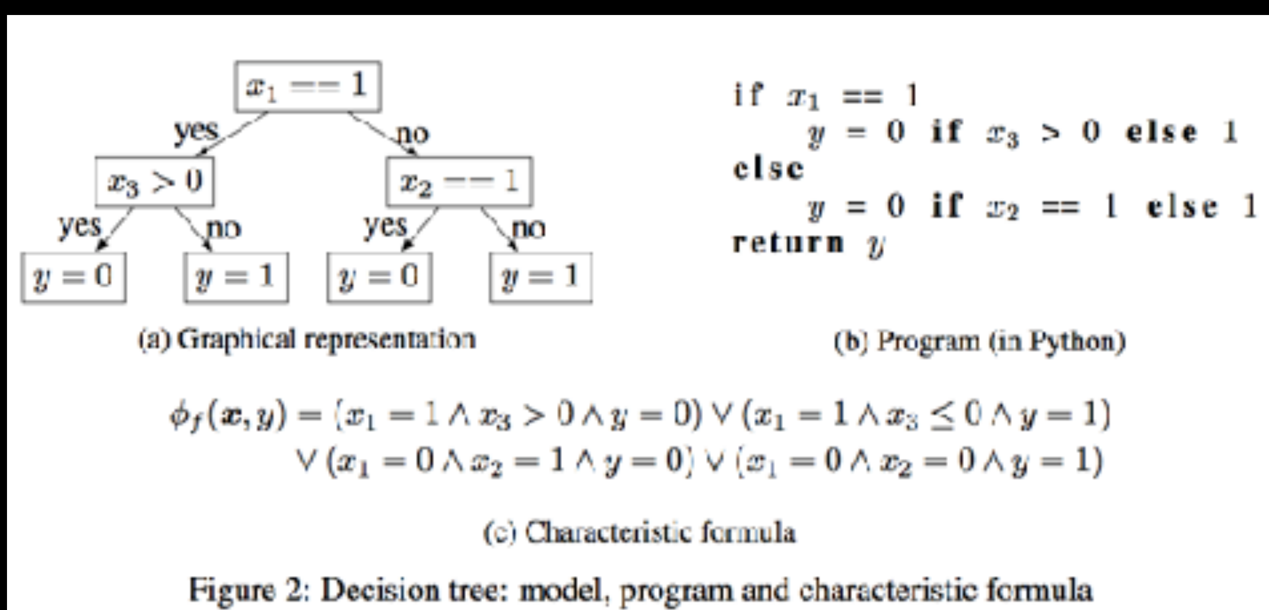
SSA Form: every non-input variable is defined before being used, and assigned at most once during execution

Path formula: a possible execution of the program yielding y

Characteristic formula ϕ: disjunction (or) of all path formulae in the program. $\phi_f(x, y) \text{ is valid} \iff f(x) = y$

Examples

$$f : \{0,1\}^2 \times \mathbb{R} \to \{0,1\}$$



(a) Graphical representation

```
if x1 == 1
    y = 0 if x3 > 0 else 1
else
    y = 0 if x2 == 1 else 1
return y
```

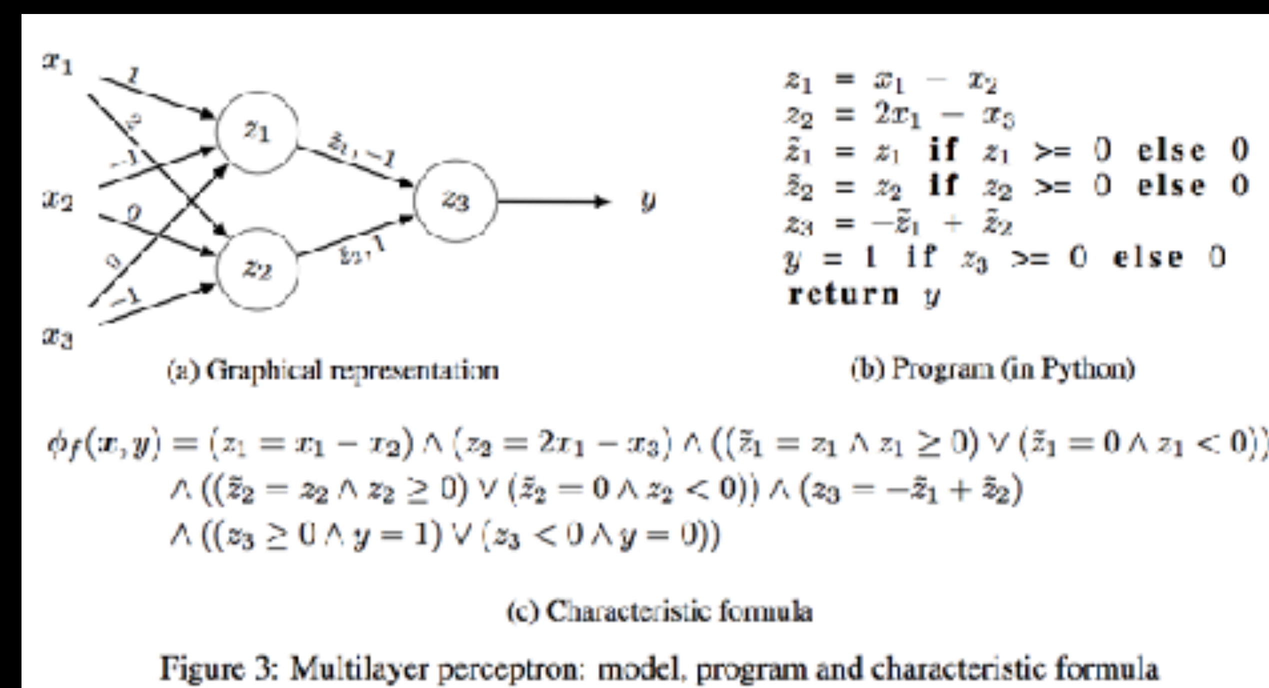
(b) Program (in Python)

$$\phi_f(\boldsymbol{x}, y) = (x_1 = 1 \wedge x_3 > 0 \wedge y = 0) \vee (x_1 = 1 \wedge x_3 \leq 0 \wedge y = 1) \\ \vee (x_1 = 0 \wedge x_2 = 1 \wedge y = 0) \vee (x_1 = 0 \wedge x_2 = 0 \wedge y = 1)$$

(c) Characteristic formula

Figure 2: Decision tree: model, program and characteristic formula

$$f : \mathbb{R}^3 \to \{0,1\}$$



(a) Graphical representation

```
z1 = x1 - x2
z2 = 2x1 - x3
z̃1 = z1 if z1 >= 0 else 0
z̃2 = z2 if z2 >= 0 else 0
z3 = -z̃1 + z̃2
y = 1 if z3 >= 0 else 0
return y
```

(b) Program (in Python)

$$\phi_f(\boldsymbol{x}, y) = (z_1 = x_1 - x_2) \wedge (z_2 = 2x_1 - x_3) \wedge ((\tilde{z}_1 = z_1 \wedge z_1 \geq 0) \vee (\tilde{z}_1 = 0 \wedge z_1 < 0)) \\ \wedge ((\tilde{z}_2 = z_2 \wedge z_2 \geq 0) \vee (\tilde{z}_2 = 0 \wedge z_2 < 0)) \wedge (z_3 = -\tilde{z}_1 + \tilde{z}_2) \\ \wedge ((z_3 \geq 0 \wedge y = 1) \vee (z_3 < 0 \wedge y = 0))$$

(c) Characteristic formula

Figure 3: Multilayer perceptron: model, program and characteristic formula

$$\phi_f(x, y) \text{ is valid} \iff f(x) = y$$

Counterfactual + Distance Formulae

Characteristic formula:

$$\phi_f(x, y) \text{ is valid}$$
$$\iff f(x) = y$$

Counterfactuals:
(given factual input f(x̂) = ŷ, and ϕ_f)

$$\mathrm{CF}_f(\hat{x}) = \{x \in \mathcal{X} | f(x) \neq f(\hat{x})\}$$

Counterfactual formula:
(given factual input f(x̂) = ŷ, and ϕ_f)

$$\phi_{\mathrm{CF}_f(\hat{x})}(x) = \phi_f(x, 1 - \hat{y}) \text{ is valid}$$
$$\iff x \in \mathrm{CF}_f(\hat{x})$$

Distance formula:

$$\phi_{d,\hat{x}}(x, \delta) \text{ is valid}$$
$$\iff d(x, \hat{x}) \leq \delta; \delta \in [0, 1]$$

Restricted CF formula:
(given factual input f(x̂) = ŷ, and ϕ_f)

$$\phi_{\hat{x},\delta}(x) = \phi_{\mathrm{CF}_f \hat{x}}(x) \wedge \phi_{d,\hat{x}}(x, \delta) \text{ is valid}$$
$$\iff x \in \mathrm{CF}_f(\hat{x}) \wedge d(x, \hat{x}) \leq \delta$$

Algorithm

$$x^* \in \operatorname*{argmin}_{x^{\mathrm{CF}}} \quad d(x^{\mathrm{F}}, x^{\mathrm{CF}})$$
$$s.t. \quad f(x^{\mathrm{F}}) \neq f(x^{\mathrm{CF}})$$
$$x^{\mathrm{CF}} \in \mathcal{P}\text{lausible}$$

$$\longrightarrow \quad x^* \leftarrow \mathsf{SAT}(\phi_{\mathrm{CF}_f(\hat{x})}(x) \wedge \phi_{d,\hat{x}}(x,\delta) \wedge \phi_{g,\hat{x}})$$

Algorithm 1: Binary Search for Nearest Counterfactuals with Satisfiability Oracle

Input: Factual $\hat{\boldsymbol{x}}$, counterfactual formula $\phi_{\mathrm{CF}_f(\hat{\boldsymbol{x}})}$, distance formula $\phi_{d,\hat{\boldsymbol{x}}}$, constraints formula $\phi_{g,\hat{\boldsymbol{x}}}$, accuracy ϵ

Output: Counterfactual $\hat{\boldsymbol{x}}_\epsilon$, distance $\delta_{\max} = d(\hat{\boldsymbol{x}}_\epsilon, \hat{\boldsymbol{x}})$, lower bound $\delta_{\min}$ on (2)

Let $\delta_{\min} \leftarrow 0$ and $\delta_{\max} \leftarrow 1$

while $\delta_{\max} - \delta_{\min} > \epsilon$ **do**

- Let $\delta \leftarrow \frac{\delta_{\min} + \delta_{\max}}{2}$
- Let $\phi_{\hat{\boldsymbol{x}},\delta}(\boldsymbol{x}) \leftarrow \phi_{\mathrm{CF}_f(\hat{\boldsymbol{x}})}(\boldsymbol{x}) \wedge \phi_{d,\hat{\boldsymbol{x}}}(\boldsymbol{x},\delta) \wedge \phi_{g,\hat{\boldsymbol{x}}}$
- Let $\boldsymbol{x} \leftarrow \mathsf{SAT}(\phi_{\hat{\boldsymbol{x}},\delta})$
- **if** $\boldsymbol{x}$ *is "unsatisfiable"* **then**
  - Let $\delta_{\min} \leftarrow \delta$
- **else**
  - Let $\hat{\boldsymbol{x}}_\epsilon \leftarrow \boldsymbol{x}$ and $\delta_{\max} \leftarrow \delta$

return $\hat{\boldsymbol{x}}_\epsilon$, $\delta_{\min}$, $\delta_{\max}$

Experiments

Table 1: Comparison of approaches for generating counterfactual explanations, based on the supported model types, data types, distance types, and plausibility constraints (actionability, data type & range).

Approach	Models	Data-types	Distances	Plausibility
Proposed (MACE)	tree, forest, lr, mlp	heterogeneous	$\ell_p \; \forall \, p$	✓
Minimum Observable (MO)[3]	-	heterogeneous	$\ell_p \; \forall \, p$	✓
Feature Tweaking (FT) [28]	tree, forest	heterogeneous	$\ell_p \; \forall \, p$	x
Actionable Recourse (AR) [29]	lr	numeric, binary	ℓ_1, ℓ_∞	x

Datasets:
Adult
Credit
COMPAS

Software:
pySMT + Z3
Scikit-learn

Metrics:
distance $\delta = d(x, \hat{x})$
coverage Ω: $\mathbb{1}[\hat{x} \in \mathcal{P}]$

Experiments:
40,000+

Quantitative Analysis - Ω

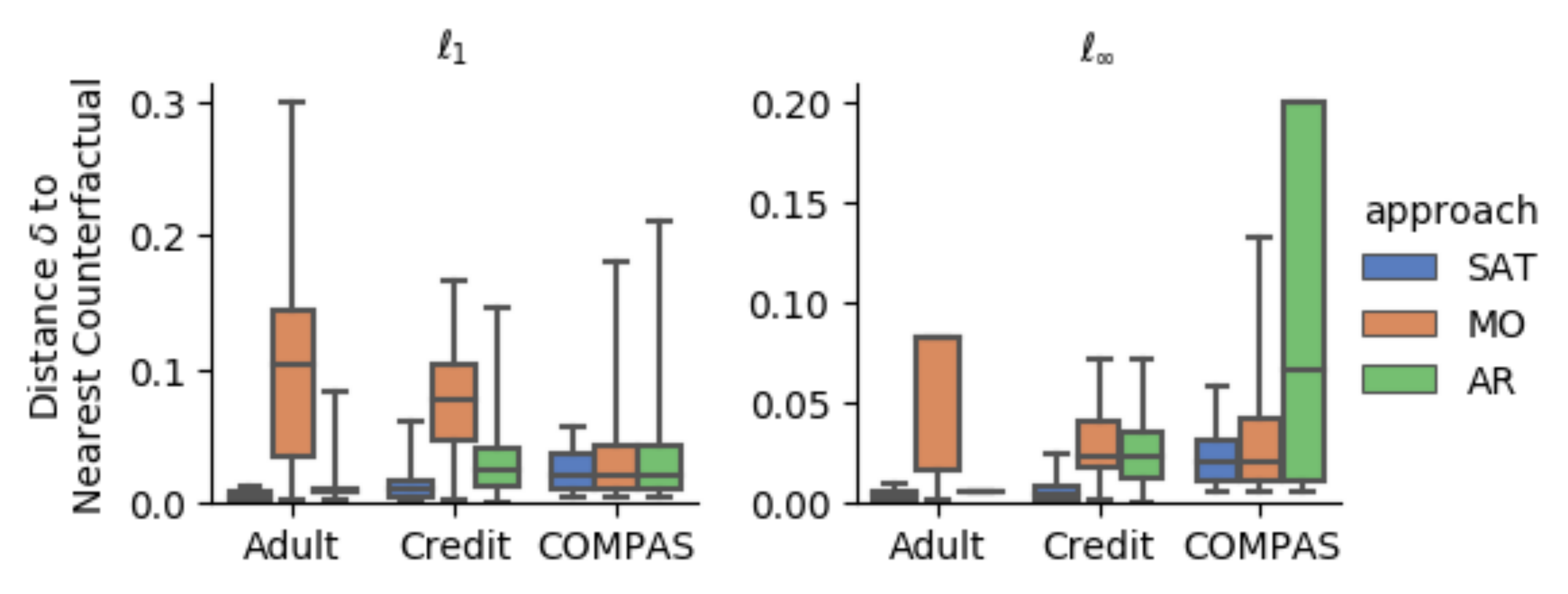
Table 2: Coverage Ω computed on $N = 500$ factual samples. For comparison, MO and MACE always have 100% coverage by definition and by design, respectively. Cells are shaded when tests are not supported. The higher the %, the higher the coverage (better performance).

		Adult			Credit			COMPAS		
		ℓ_0	ℓ_1	ℓ_∞	ℓ_0	ℓ_1	ℓ_∞	ℓ_0	ℓ_1	ℓ_∞
tree	PFT	0%	0%	0%	68%	68%	68%	74%	74%	74%
forest	PFT	0%	0%	0%	99%	99%	99%	100%	100%	100%
lr	AR		18%	0.4%		100%	100%		100%	100%

Quantitative Analysis - δ

Table 3: Percentage of improvement in distances, computed as $100 * \mathbb{E}[1 - \boldsymbol{\delta}_{\text{MACE}}/\boldsymbol{\delta}_{\text{Other}}]$. $N = \Omega_{\text{MACE}} \cap \Omega_{\text{Other}}$ factual samples. Cells are shaded when tests are not supported. The higher the % the better the improvement.

		Adult			Credit			COMPAS		
		ℓ_0	ℓ_1	ℓ_∞	ℓ_0	ℓ_1	ℓ_∞	ℓ_0	ℓ_1	ℓ_∞
tree	MACE vs MO	47%	81%	72%	67%	97%	94%	1%	5%	5%
	MACE vs PFT				53%	97%	96%	15%	56%	54%
forest	MACE vs MO	51%	82%	71%	68%	97%	96%	1%	6%	6%
	MACE vs PFT				53%	96%	96%	4%	28%	27%
lr	MACE vs MO	62%	93%	88%	80%	82%	81%	3%	7%	6%
	MACE vs AR		5%	91%		41%	71%		10%	38%
mlp	MACE vs MO	85%	99%	98%	89%	99%	99%	58%	92%	88%



Qualitative Analysis

$$x^* \leftarrow \mathsf{SAT}\big(\phi_{\mathrm{CF}_{f}(\hat{x})}(x) \wedge \phi_{d,\hat{x}}(x, \delta) \wedge \phi_{g,\hat{x}}\big)$$

Plausibility Formula ϕ_g: $(\hat{x}_{age} \leq x_{age})$

Table 4: Percentage of factual samples for which the nearest counterfactual sample requires a change in age for a random forest trained on the Adult dataset, and the corresponding increase in distance to nearest counterfactual when restricting the approaches not to change age: $100 \times \mathbb{E}[\delta_{\text{restr.}}/\delta_{\text{unrestr.}} - 1]$. The higher %, the greater the increase in distance.

	ℓ_0		ℓ_1		ℓ_∞	
	% age-change	rel. dist. increase	% age-change	rel. dist. increase	% age-change	rel. dist. increase
MACE	13.2%	9.0%	20.4%	100.3%	84.4%	32.8%
MO	78.8%	50.9%	92.0%	245.7%	95.6%	193.3%

Still a lot to do…

- Observed **features are indirect, noisy and potentially biased** measurements of the "state of the world"
- How these features have been **measured** to design a proper similarity metric for each of them
- **Dependencies** between features (correlations, cofounders, causal graphs, etc.)
- **Measurement and data collection processes** should be not ignored when studying fairness and interpretability (ethics) in ML

The most influential of our four stake-holder communities is the users — the one that's barely represented in the literature — because, as in the 1980s, failure to satisfy users of AI technology in the long run will be the most likely cause of another 'AI Winter'. Unfulfilled expectations and/or a smaller-than-hoped-for market will lead to investment drying up.

– Alun Preece et al.

Thank you!