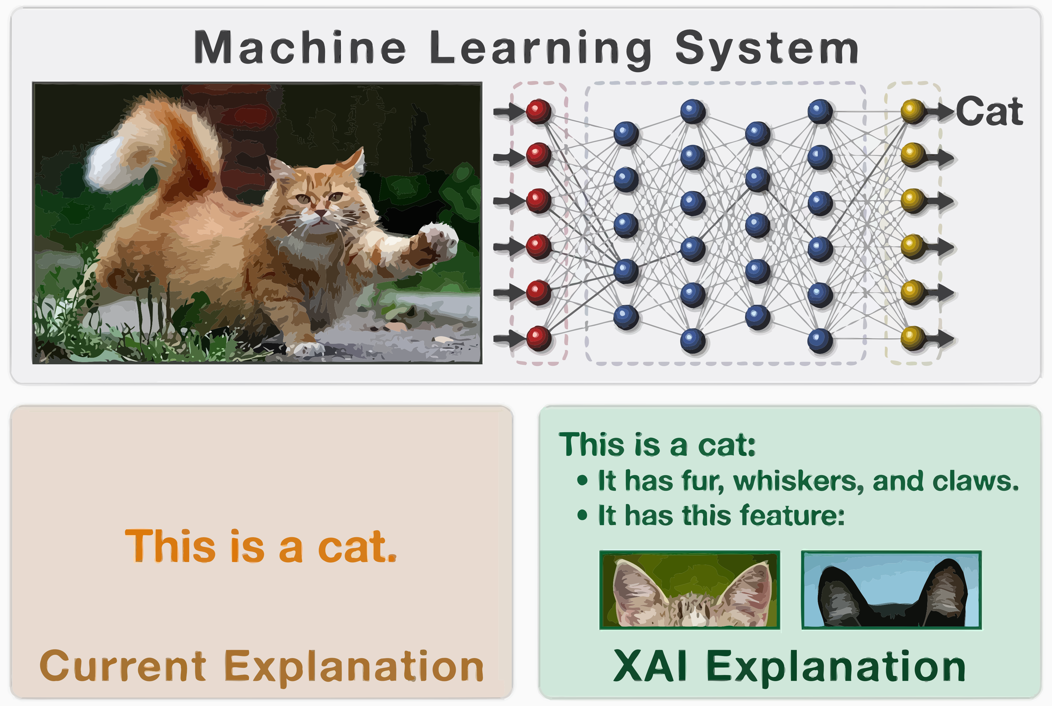
From Contrastive to Abductive Explanations and Back Again

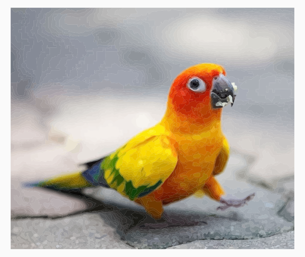
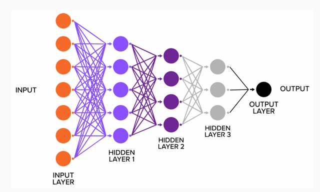
Alexey Ignatiev[1], Nina Narodytska[2], Nicholas Asher[3], and Joao Marques-Silva[3]

[1]Monash University, Melbourne, Australia [2]VMware Research, CA, USA [3]IRIT, CNRS, Toulouse, France

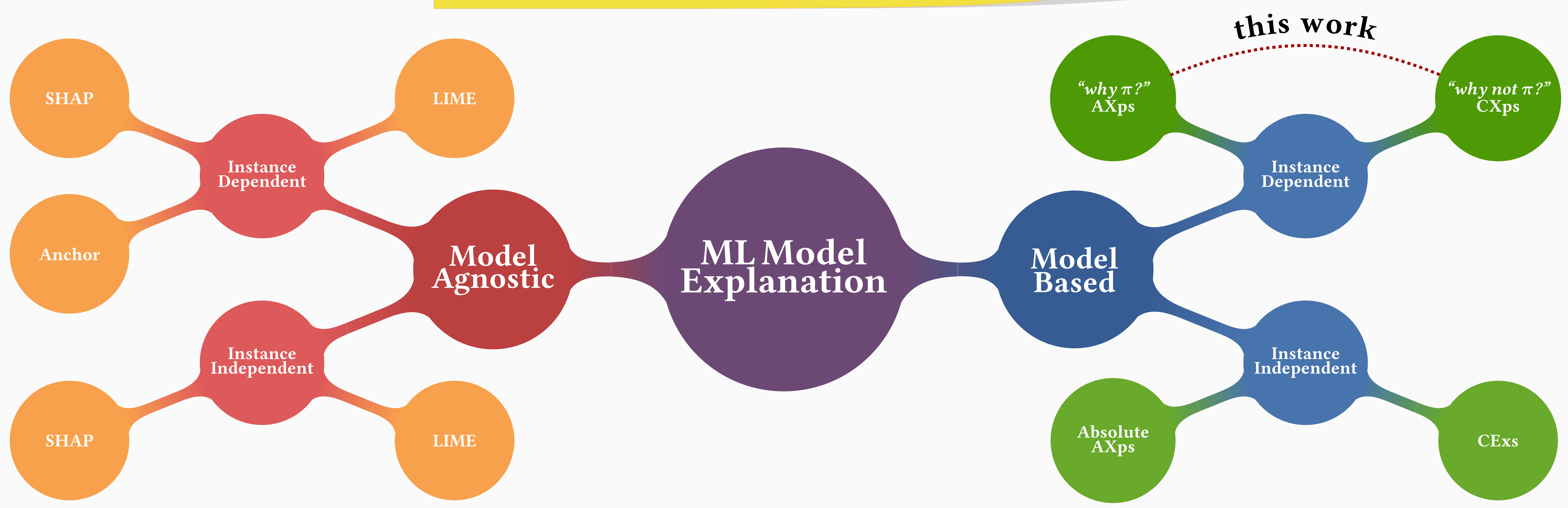
eXplainable AI



Why? Status Quo

	A parrot	Machine learning algorithm
		
Learns random phrases	✓	✓
Doesn't understand s**t about what it learns	✓	✓
Occasionally speaks nonsense	✓	✓

Taxonomy of ML Model Explanations



Formal Explanations

classifier $\tau : \mathbb{F} \to \mathcal{K}$, **instance** $\mathbf{v}$ **s.t.** $\tau(\mathbf{v}) = c$

abductive explanation $\mathcal{X}$

$$\forall (\mathbf{x} \in \mathbb{F}) \,.\, \bigwedge_{j \in \mathcal{X}} (x_j = v_j) \rightarrow (\tau(\mathbf{x}) = c)$$

contrastive explanation $\mathcal{Y}$

$$\exists (\mathbf{x} \in \mathbb{F}) \,.\, \bigwedge_{j \notin \mathcal{Y}} (x_j = v_j) \wedge (\tau(\mathbf{x}) \neq c)$$

Explanation Examples

$\mathbb{F} = \{0, 1, 2\}^5$ $\quad \mathcal{K} = \{\ominus, \oplus\}$

R_0:	IF	$x_1 = 1 \wedge x_2 = 1$	THEN $\ominus$
R_1:	ELSE IF	$x_3 \neq 1$	THEN $\oplus$
R_{DEF}:	ELSE		THEN $\ominus$

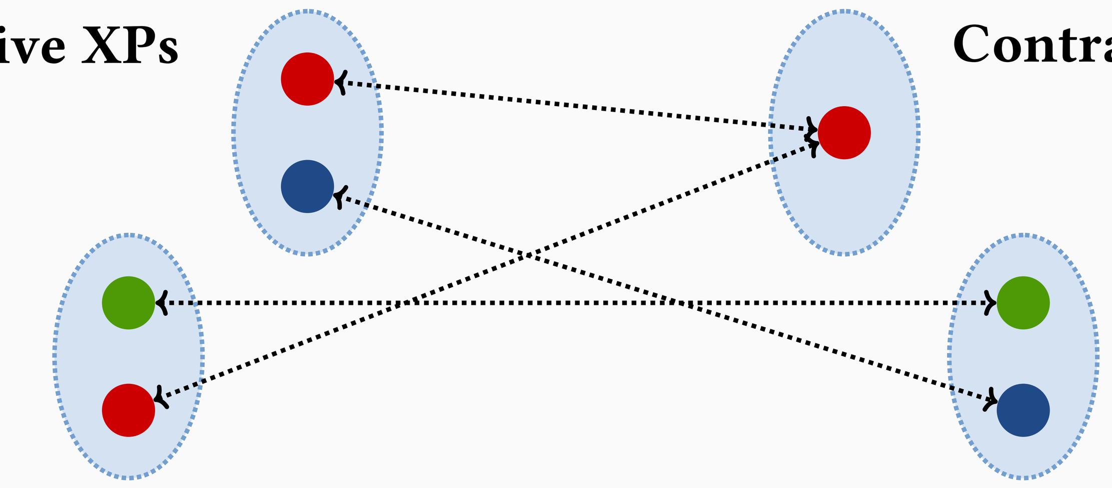
observe $\tau(1, 1, 1, 1, 1) = \ominus$

AXps $\mathbb{X} = \{\{1, 2\}, \{3\}\}$

CXps $\mathbb{Y} = \{\{1, 3\}, \{2, 3\}\}$

Minimal Hitting Set Duality

Abductive XPs **Contrastive XPs**



AXps are minimal hitting sets of CXps, and vice versa.

Enumerating CXps

Function CXpEnum(τ, $\mathbf{v}$, c)
Input: τ: ML model, $\mathbf{v}$: Input instance, $c = \tau(\mathbf{v})$: Prediction

```
𝓘 ← ∅                              // Block CXps
while true:
    μ ← ExtractCXp(τ, v, c, 𝓘)
    if μ = ∅: break
    ReportCXp(μ)
    𝓘 ← 𝓘 ∪ ⋁_{j∈μ} (x_j = v_j)
```

Function ExtractCXp(τ, $\mathbf{v}$, c, $\mathcal{I}$)
Input: τ: classifier, $\mathbf{v}$: Input instance, $c = \tau(\mathbf{v})$: Prediction, $\mathcal{I}$: Blocked CXps
Output: $\mathcal{S}$: Minimal set

```
𝓢 ← [|v|]
foreach j ∈ 𝓢:
    if SAT(⋀_{i∉𝓢\{j}} (x_i = v_i) ∧ 𝓘 ∧ τ(x) ≠ c):
        𝓢 ← 𝓢 \ {j}
return 𝓢                           // 𝓢 is CXp
```

Enumerating AXps and CXps

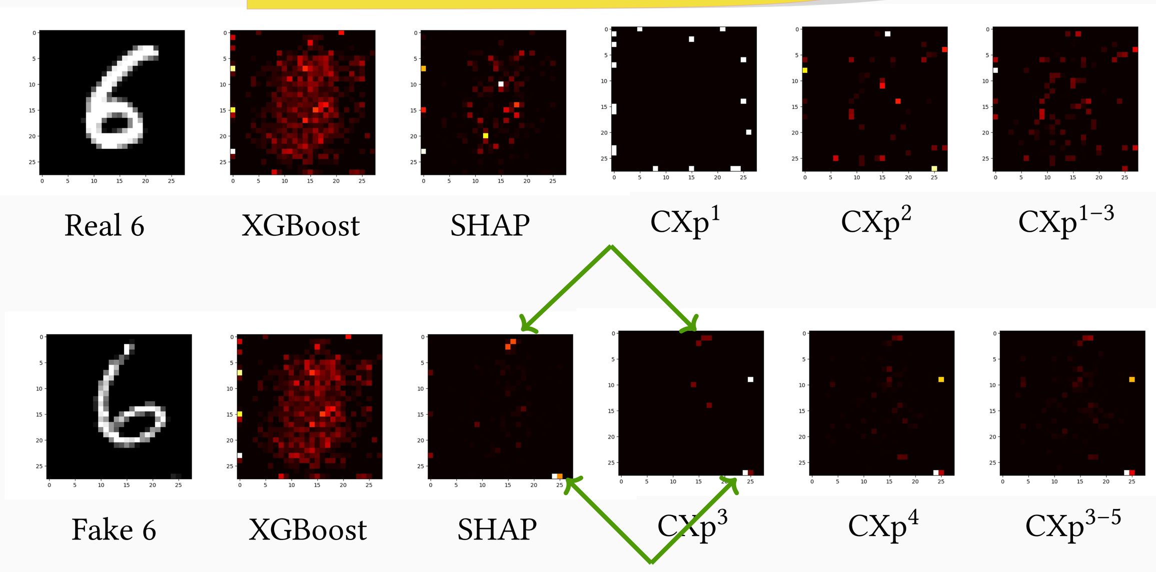
Function XpEnum(τ, $\mathbf{v}$, c)
Input: τ: ML model, $\mathbf{v}$: Input instance, $c = \tau(\mathbf{v})$: Prediction

```
𝒦 = (𝒩, 𝒫) ← (∅, ∅)                  // Block AXps & CXps
while true:
    (st_λ, λ) ← FindMHS(𝒫, 𝒩)         // MHS of 𝒫 s.t. 𝒩
    if ¬st_λ: break
    st_c' ← SAT(⋀_{j∈λ} (x_j = v_j) ∧ τ(x) ≠ c)
    if ¬st_c':                        // entailment holds
        ReportAXp(λ)
        𝒩 ← 𝒩 ∪ ⋁_{j∈λ} (x_j ≠ v_j)
    else:
        μ ← ExtractCXp(τ, v, c, 𝒫)
        ReportCXp(μ)
        𝒫 ← 𝒫 ∪ ⋁_{j∈μ} (x_j = v_j)
```

Explanation Enumeration Results

	Dataset					
	Adult	**Lending**	**Recidivism**	**Compas**	**German**	**Spambase**
# of instances	5579.0	4414.0	3696.0	778.0	1000.0	2344.0
total time (sec.)	7666.9	443.8	3688.0	78.4	16 943.2	6859.2
minimal time (sec.)	0.1	0.0	0.1	0.0	0.2	0.1
average time (sec.)	1.4	0.1	1.0	0.1	16.9	2.9
maximal time (sec.)	13.1	0.8	8.9	0.5	193.0	23.1
total oracle calls	492 990.0	69 653.0	581 716.0	21 227.0	748 164.0	176 354.0
minimal oracle calls	14.0	11.0	17.0	13.0	23.0	12.0
average oracle calls	88.4	15.8	157.4	27.3	748.2	75.2
maximal oracle calls	581.0	73.0	1426.0	134.0	7829.0	353.0
total # of AXps	52 137.0	8105.0	60 688.0	1931.0	59 222.0	18 876.0
average # of AXps	9.4	1.8	16.4	2.5	59.2	8.1
average AXp size	5.3	1.9	6.4	3.8	7.5	4.6
total # of CXps	66 219.0	8663.0	77 784.0	3558.0	66 781.0	24 774.0
average # of CXps	11.9	2.0	21.1	4.6	66.8	10.6
average CXp size	2.4	1.4	2.6	1.5	3.6	2.3

Debugging SHAP



Real 6 XGBoost SHAP CXp^1 CXp^2 CXp^{1-3}

Fake 6 XGBoost SHAP CXp^3 CXp^4 CXp^{3-5}

The **"real vs fake"** images. The first row shows results for the *real image 6*; the second – results for the *fake image 6*. The first column shows examples of inputs; the second – heatmaps of XGBoost's important features; the third – heatmaps of SHAP's explanation. Last three columns show heatmaps of CXp of different cardinality. The brighter pixels are more influential features.