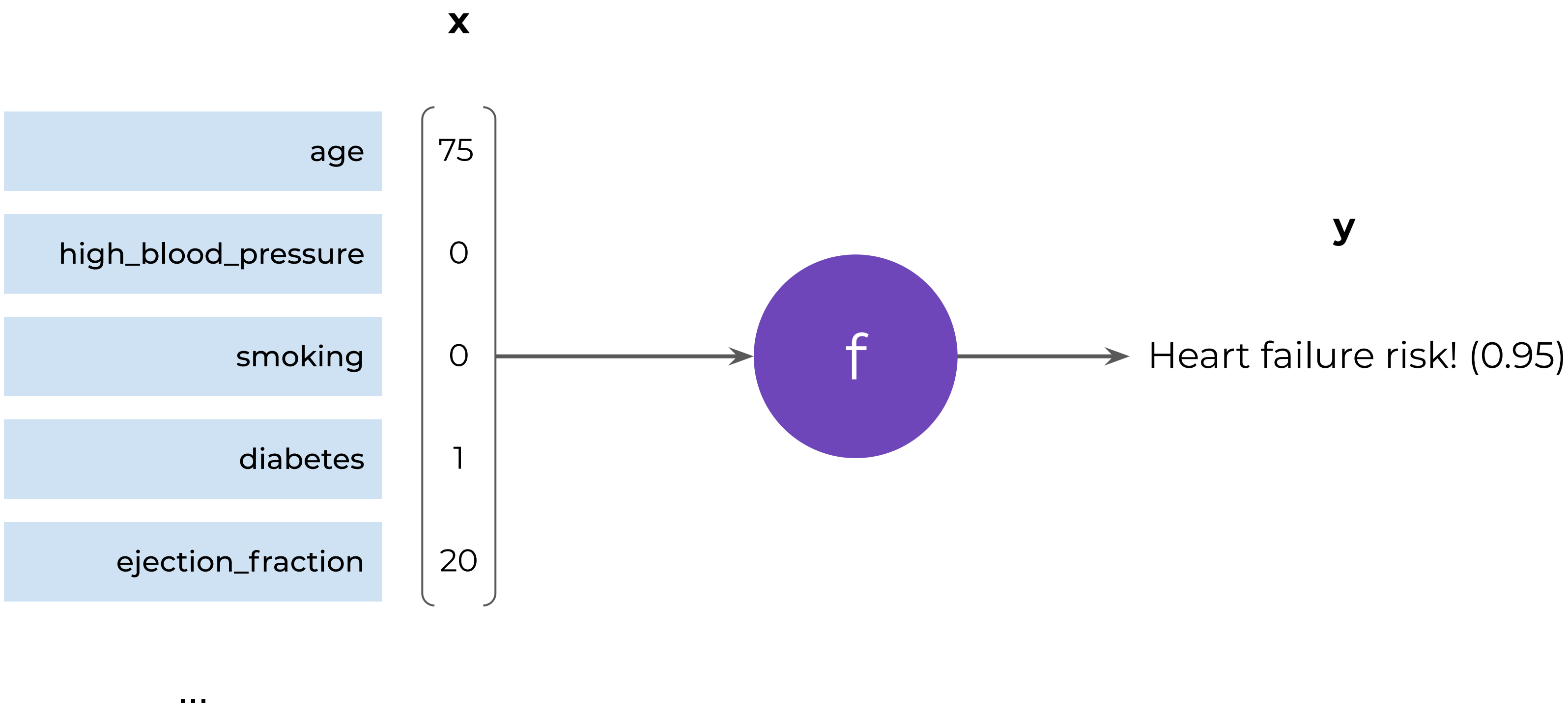
Towards Rigorous Interpretation: a Formalisation of Feature Attribution

Darius Afchar[1,2], Romain Hennequin[1], Vincent Guigue[2]

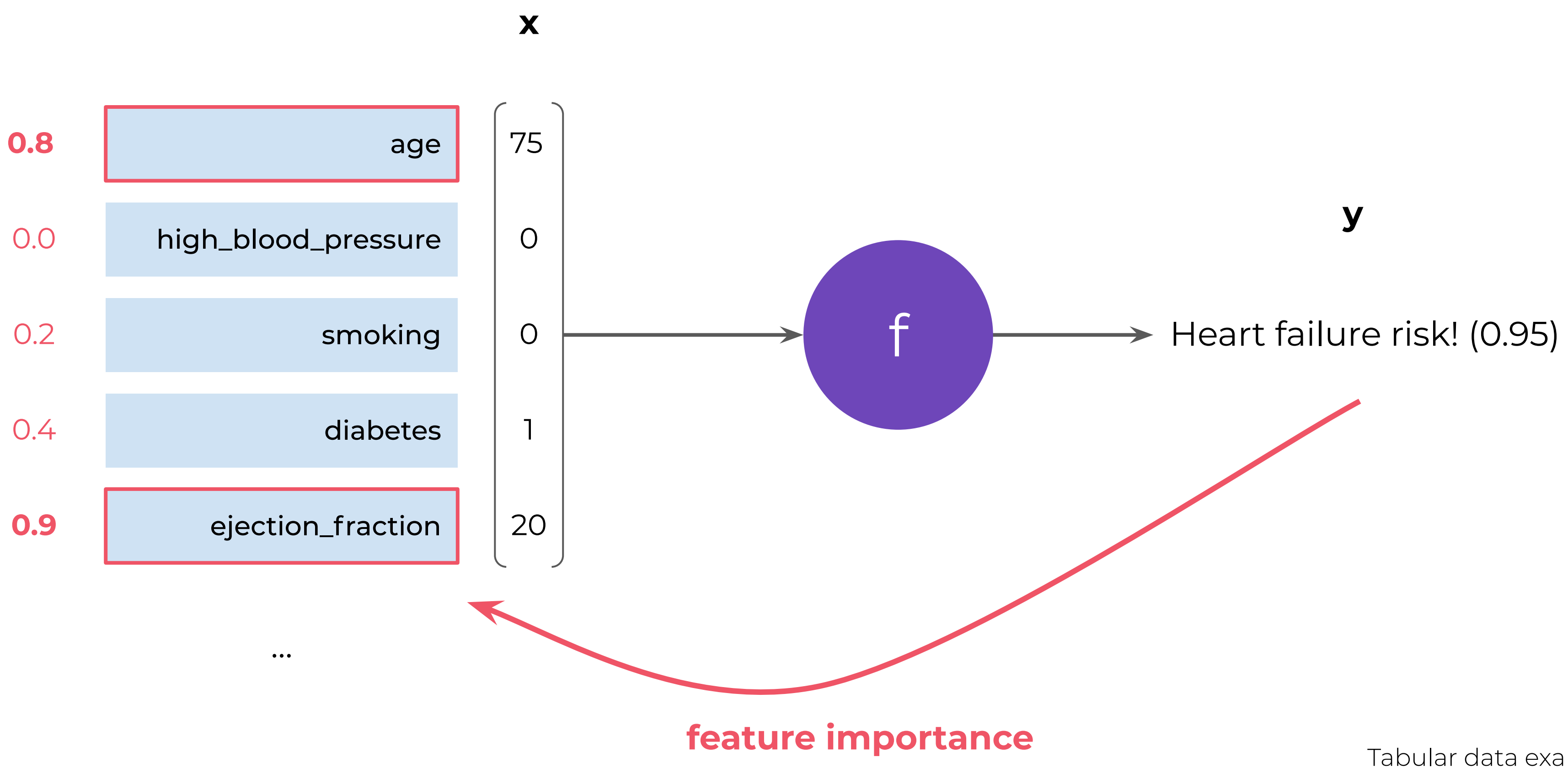
[1] deezer [2] MLIA Machine Learning & Deep Learning for Information Access

Feature Attribution?

- Feature-based interpretation method
- *"What input is most responsible for a given prediction?"*



Tabular data example



Tabular data example

x

f

y

Cow (0.9)

Cat (0.3)

Image data example

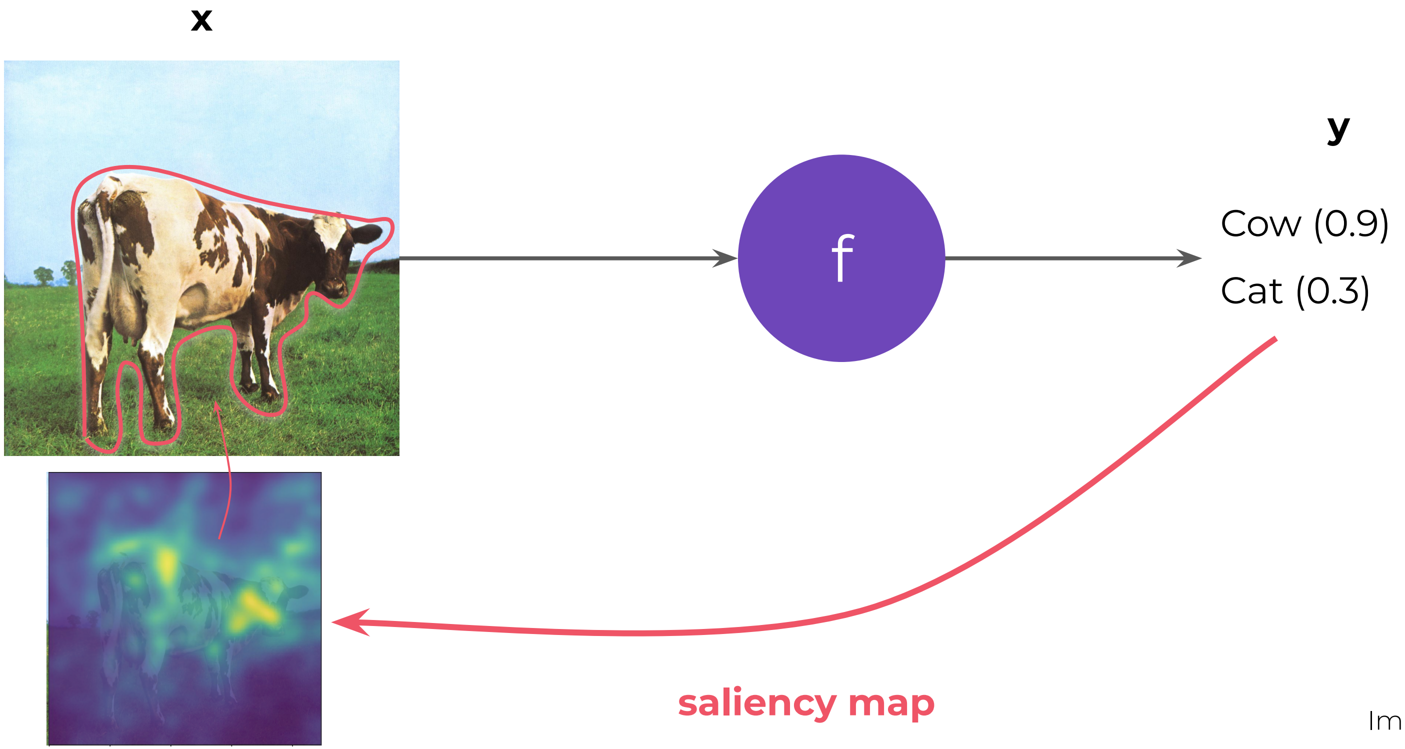
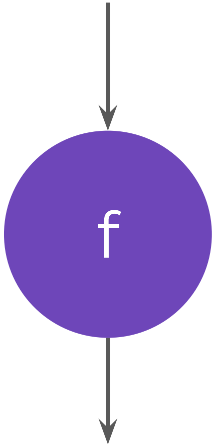


Image data example

x

`I am so glad to do an ICML presentation I could do this all day! Mom and Dad will be proud!`

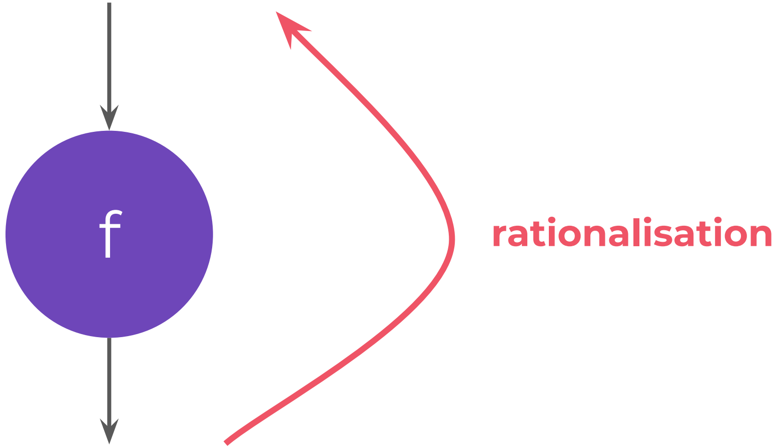


y Overly enthusiastic (0.98)

NLP example

x

`I am so` **`glad`** `to do an ICML presentation I could do this all day! Mom and Dad will be` **`proud`**`!`



y Overly enthusiastic (0.98)

NLP example

Feature Attribution?

- Feature-based interpretation
- *"What input is most responsible for a given prediction?"*
- Many terms designate the same thing

Feature Attribution?

- Feature-based interpretation
- *"What input is most responsible for a given prediction?"*
- Many terms designate the same thing

feature contribution

feature importance

responsibility

prime implicant

input relevance

variable selection

saliency map

minimal sufficient subset

feature ranking

a posteriori explanation

sparse approximation

Interpretability is in question

- The issue is not just philosophical!
- Many works rely on **intuitive notions** of interpretability / **heuristics**
- ... and are **ill-evaluated**

Interpretability is in question

The (Un)reliability of Saliency Methods

Pieter-Jan Kind
Google Brain*
{pikinder,

Maximilian Al
TU-Berlin

Hooker† Julius Adebayo

Exploratory Not Explanatory: Counterfactual Analysis of Saliency Maps for Deep Reinforcement Learning

ksha Atrey, Kaleigh Clary & Dav
rsity of Massachusetts

On the Robustness of Interpretability Methods

David Alvarez-Melis [1] Tommi S. Jaakkola [1]

the insight gained from a single attributi
ight be too brittle, and lead to a false se
address this limitation
behavior

When Explanations Lie: Why Many Modified BP Attributions Fail

Sixt [1] Maximilian Granz [1] Tim Landgraf [1]

Problems with Shapley-value-based explanations
measures

Elizabeth Kumar [1] Suresh

Abstract

eoretic formulations of featu
e become popular as a way to
earning models. These

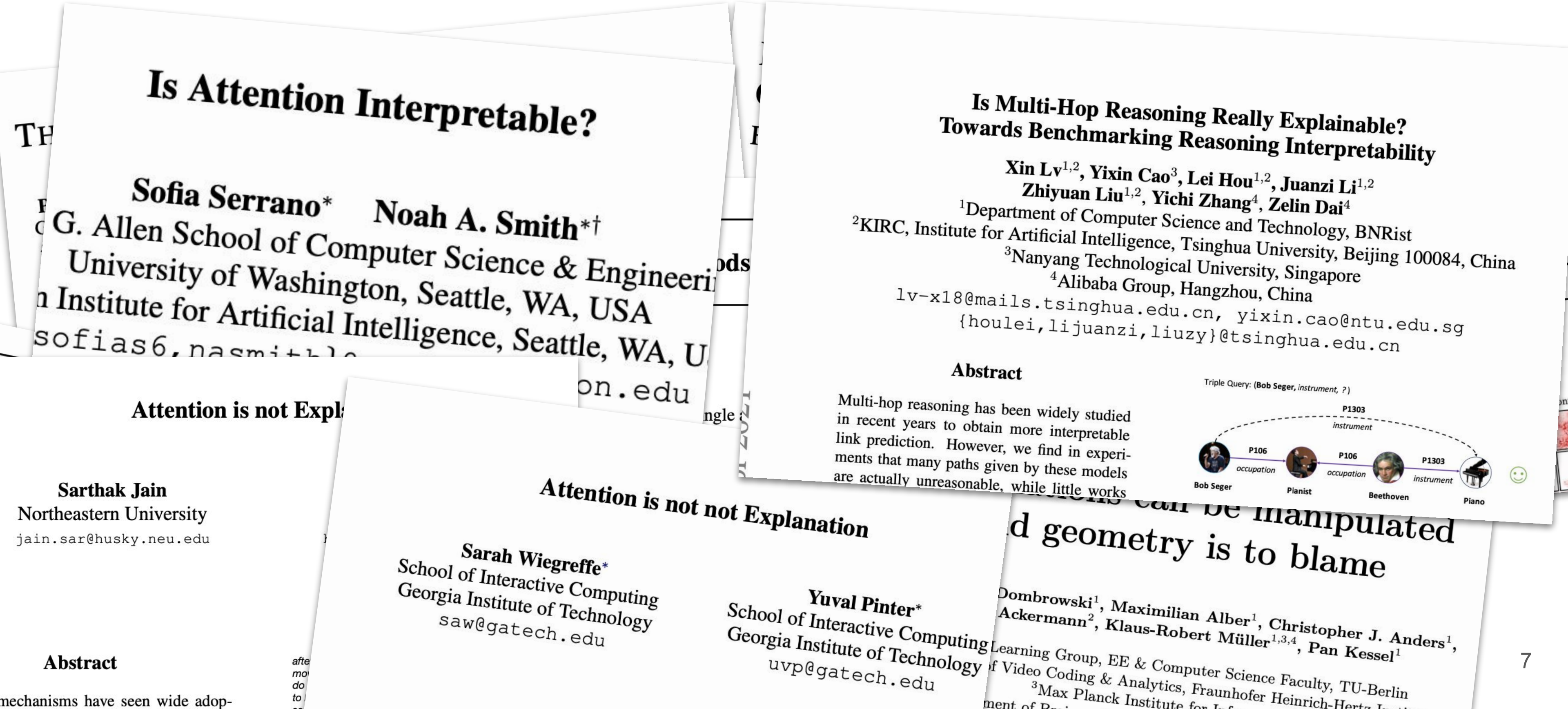
Challenging common interpretability assumptions in feature attribution explanations

J. Zico Kolter

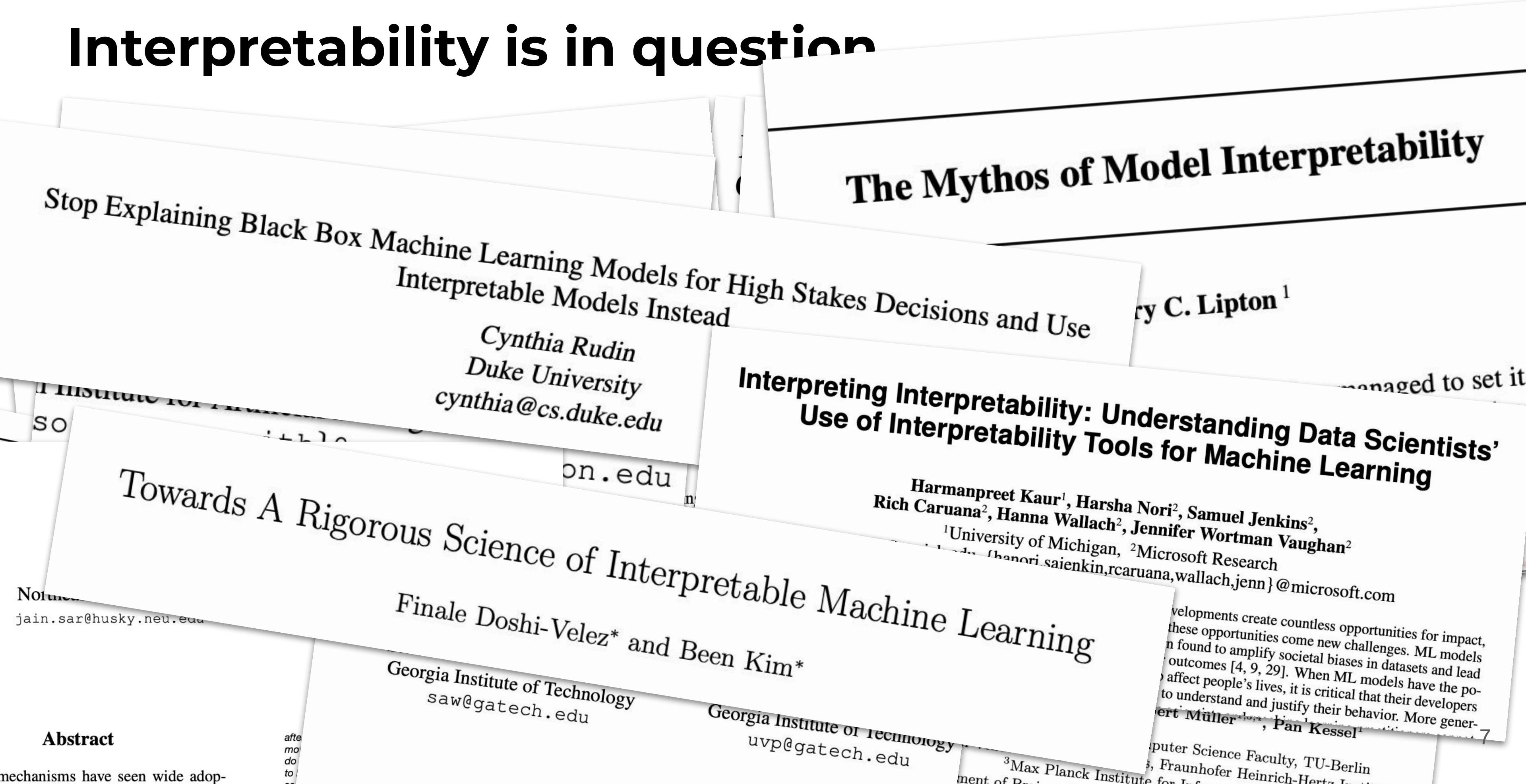
Explanations can be manipulated and geometry is to blame

Kathrin Dombrowski[1], Maximilian Alber[1], Christopher J. Anders[1],
Marcel Ackermann[2], Klaus-Robert Müller[1,3,4], Pan Kessel[1]
Machine Learning Group, EE & Computer Science Faculty, TU-Berlin
rtment of Video Coding & Analytics, Fraunhofer Heinrich-Hertz
[3]Max Planck Institute for

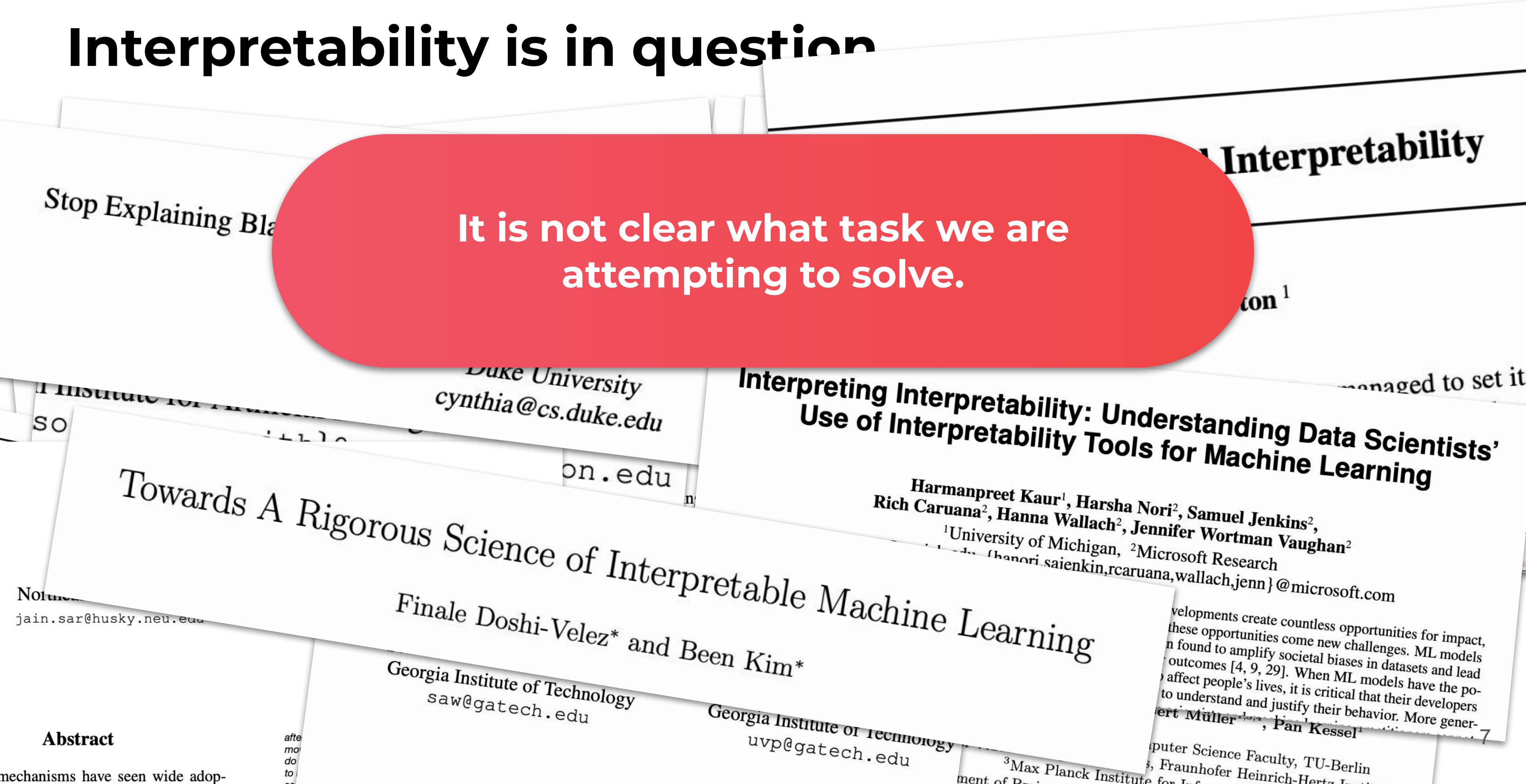
Interpretability is in question



Interpretability is in question

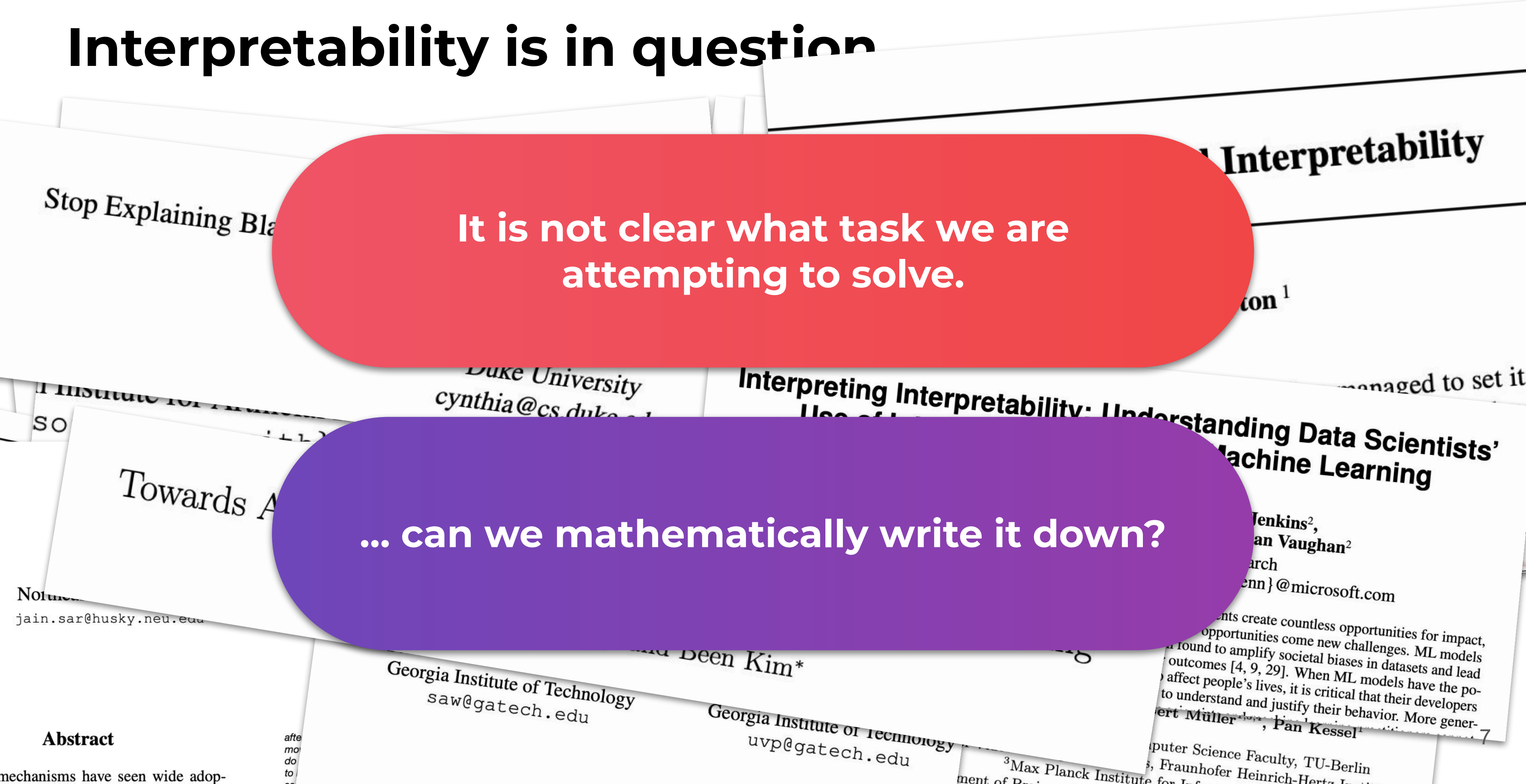


Interpretability is in question



It is not clear what task we are attempting to solve.

Interpretability is in question



It is not clear what task we are attempting to solve.

… can we mathematically write it down?

Yes.

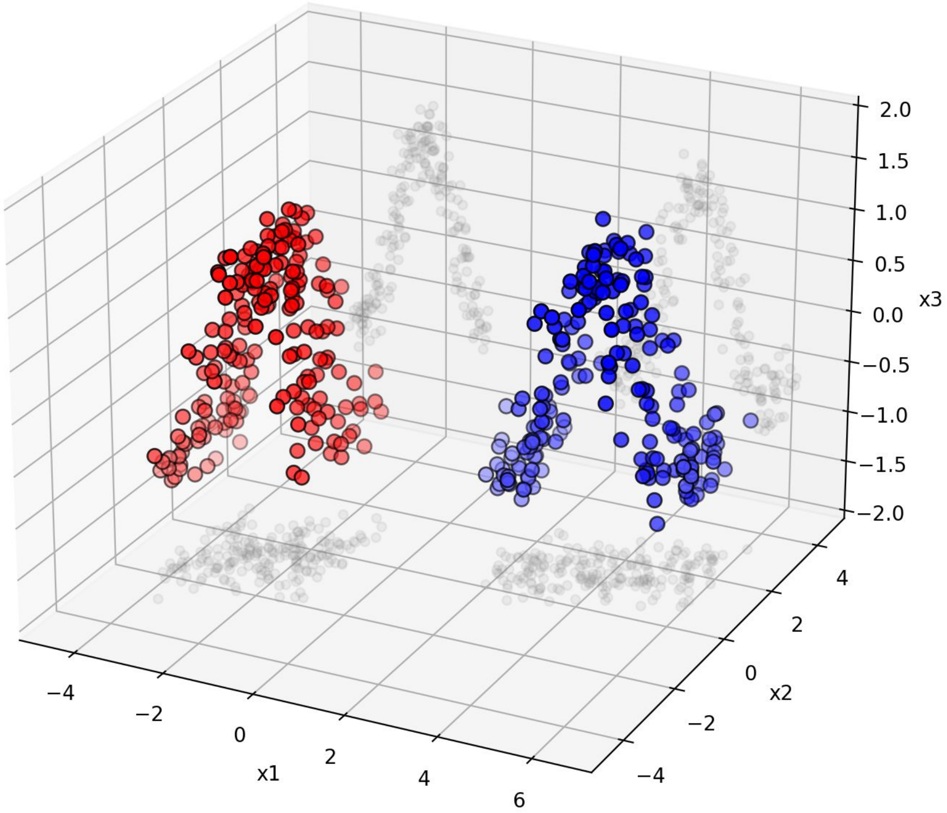
- Leveraging simple considerations on functionality
- No evasive / intuitive concepts
- Allows task-specificity

Problem 1 (Global subset selection)**.** *Given a relation R, find a subset of indices $I^* \subset [n]$ that minimises*

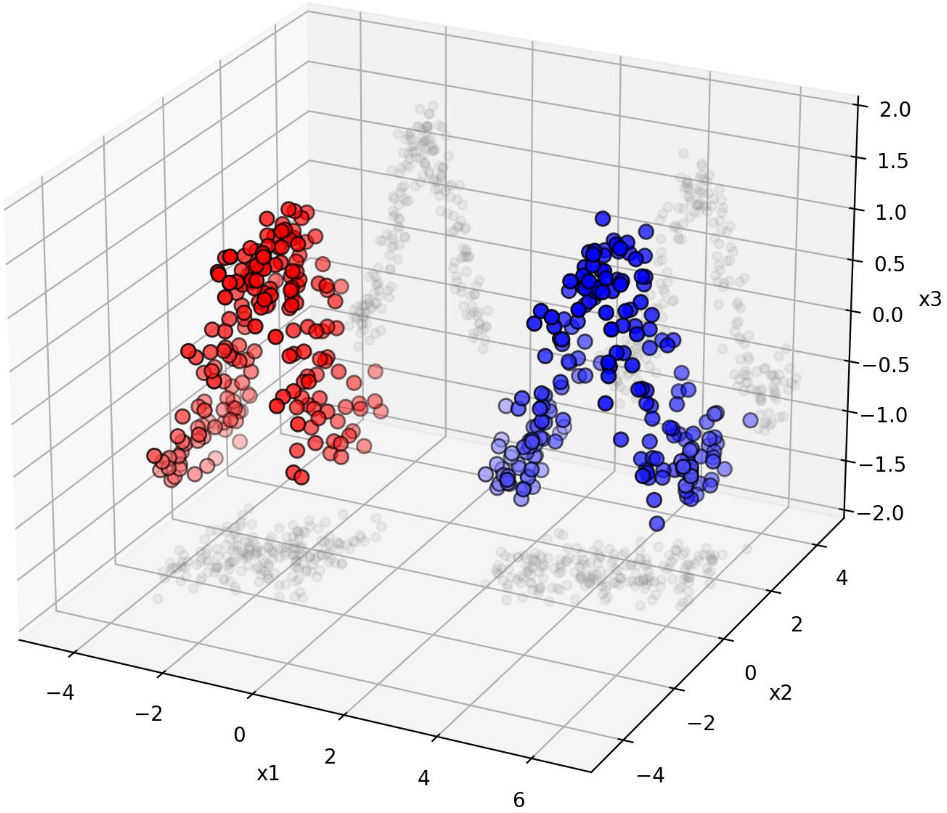
$$\min_{J \subset [n]} \mathit{Card}(J) \quad \text{s.t.} \quad \forall x,\ x \in A_J(R)$$

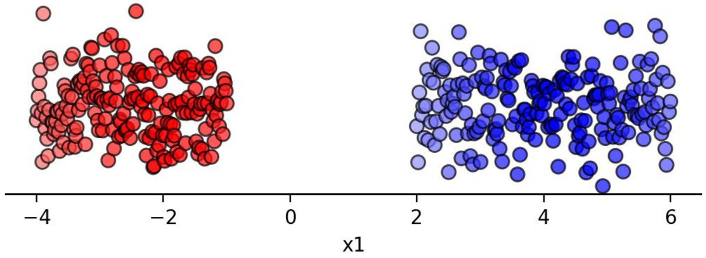
Problem 2 (Instance-wise subset selection)**.** *Given a relation R, for all $x \in \mathcal{X}$, find a local subset of indices $I^*(x) \subset [n]$ that minimises*

$$\min_{J \subset [n]} \mathit{Card}(J) \quad \text{s.t.} \quad x \in A_J(R)$$



x1
x2
x3
−4
−2
0
2
4
6
4
2
0
−2
−4
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0



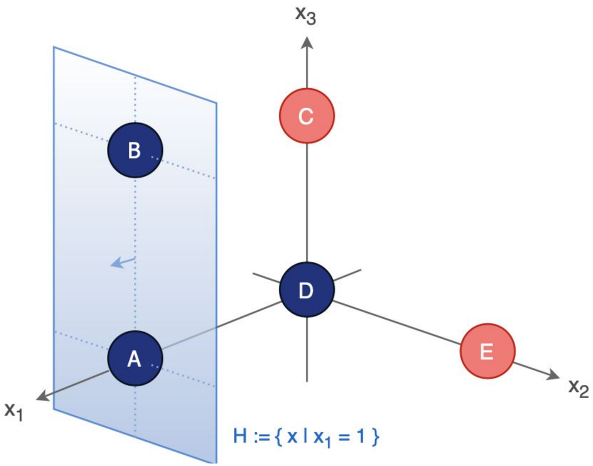


$$f_1 := \begin{cases} x_1 > 0 \mapsto \text{blue} \\ x_1 \leq 0 \mapsto \text{red} \end{cases}$$

Formalising: what is at stake

1. How to do that locally (*instance-wise* case)?

"Are we free to do whatever we want with the selected variables?" *NO*

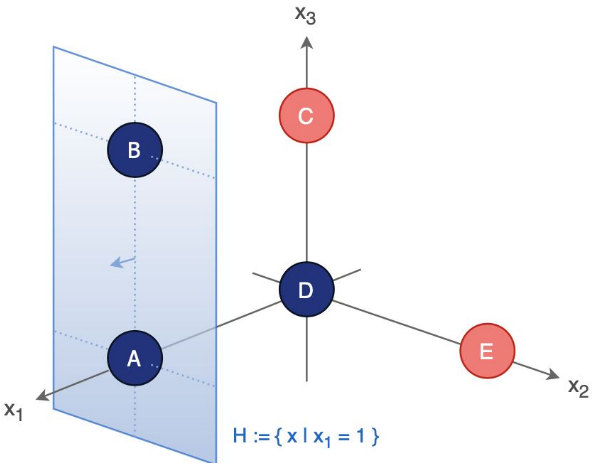


Formalising: what is at stake

1. How to do that locally (*instance-wise* case)?

   "Are we free to do whatever we want with the selected variables?" NO

2. How to do that probabilistically?



Contributions

Compare attribution methods on the same theoretical ground

Contributions

Compare attribution methods on the same theoretical ground

Rigorous evaluations without ground-truth

Method	Acc (%)	T (h:m:s)
LIME (Cat.)	16.2 ± 1.3	0:05:54
LIME (Cont.)	27.4 ± 1.6	0:05:47
attr-GAM	24.5 ± 1.5	0:00:25
Shapley ($\mathbb{E}(f)$)	74.3 ± 1.1	0:16:29
SHAP ($f(\mathbb{E})$)	15.7 ± 1.3	0:17:41
Gradient	26.5 ± 1.5	0:00:04
Gradient$\times$Input	22.6 ± 1.5	0:00:04
Integrated Gradient	18.5 ± 1.4	0:00:24
Expected Gradient	21.4 ± 1.4	0:03:42
attr-GA$^{\infty}$M	$\mathbf{81.7 \pm 1.1}$	0:17:44*
attr-GA^{2}M	52.5 ± 1.8	$\ll *$
attr-GA^{3}M	74.1 ± 1.3	$< *$
attr-GA^{4}M	81.2 ± 1.1	$< *$
InterpretableNN	79.7 ± 1.2	$\simeq *$
Archipelago	70.2 ± 1.1	$\simeq *$
L2X	23.7 ± 1.6	32:53:16
INVASE	7.4 ± 0.9	44:15:44

Contributions

Compare attribution methods on the same theoretical ground

Rigorous evaluations without ground-truth

Derive some properties, prove failure cases

Method	Property verification rate (%)
LIME (Cat.)	29.9 ± 1.7
LIME (Cont.)	46.6 ± 1.6
attr-GAM	61.5 ± 1.1
Shapley ($\mathbb{E}(f)$)	79.5 ± 1.1
SHAP ($f(\mathbb{E})$)	23.7 ± 1.5
Gradient	61.6 ± 1.3
Gradient $\times$ Input	54.5 ± 1.3
Integrated Gradient	39.7 ± 1.5
Expected Gradient	41.8 ± 1.5
attr-GA$^\infty$M	$\mathbf{92.9 \pm 0.6}$
attr-GA^{2}M	63.7 ± 1.4
attr-GA^{3}M	81.2 ± 1.4
attr-GA^{4}M	90.7 ± 1.1
InterpretableNN	86.9 ± 0.9
Archipelago	88.8 ± 0.7
L2X	37.5 ± 1.6
INVASE	61.3 ± 1.7

Thank you!

Towards Rigorous Interpretation: a Formalisation of Feature Attribution

Darius Afchar[1,2], Romain Hennequin[1], Vincent Guigue[2]