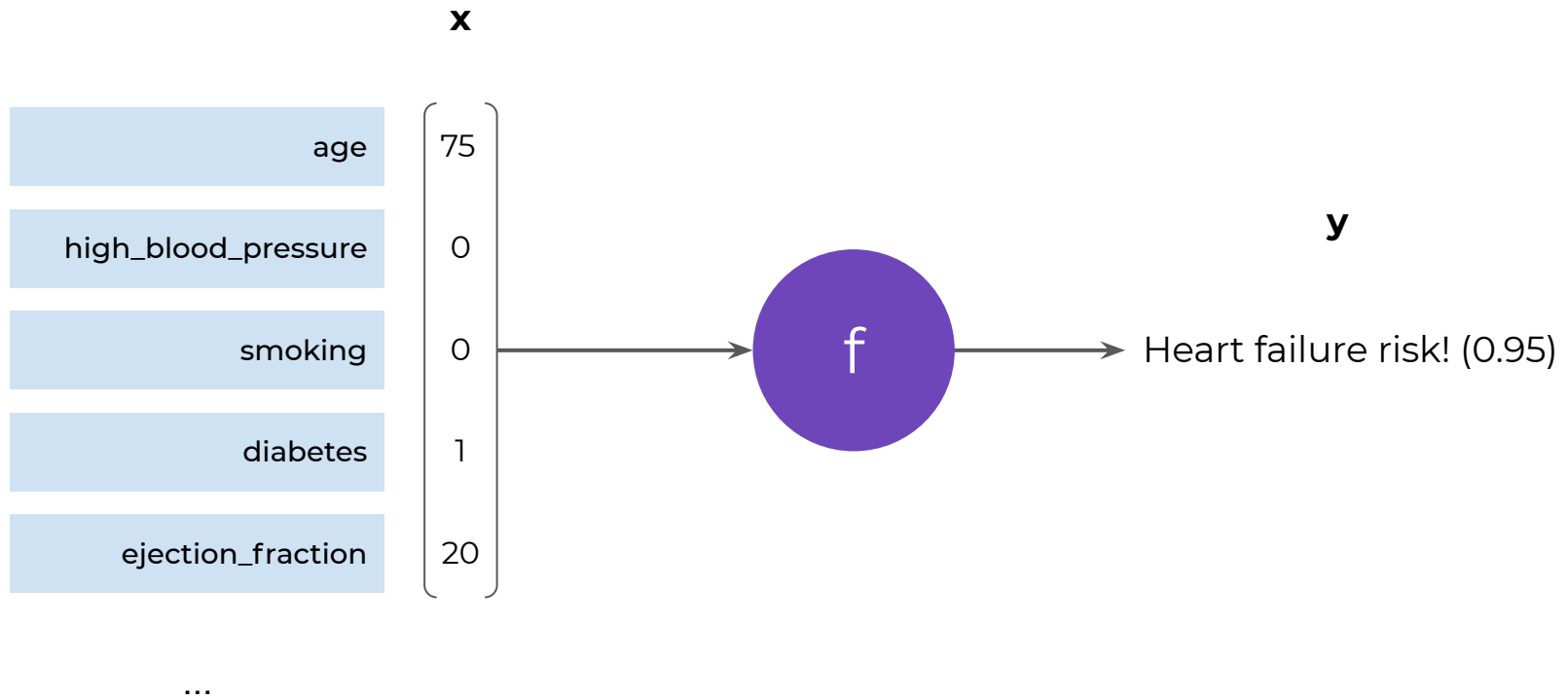


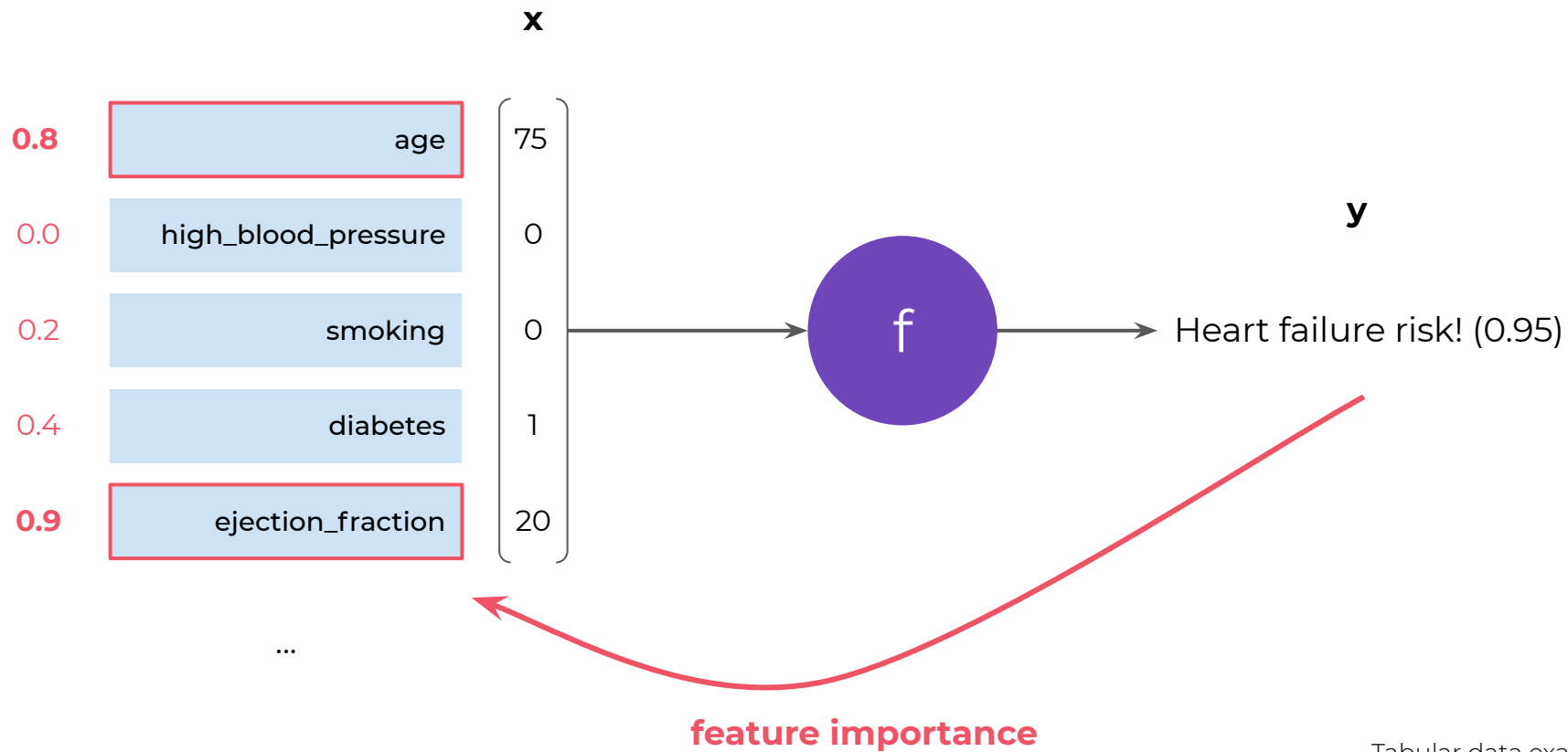
Towards Rigorous Interpretation: a Formalisation of Feature Attribution

Darius Afchar^{1,2}, Romain Hennequin¹, Vincent Guigue²

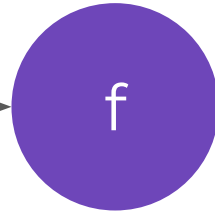
Feature Attribution?

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





x

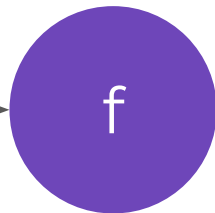
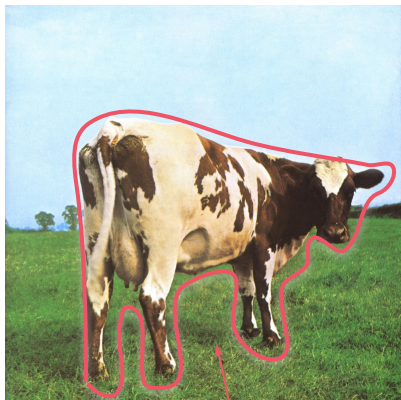


y

Cow (0.9)

Cat (0.3)

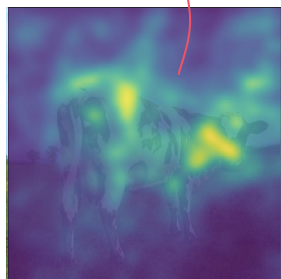
x



y

Cow (0.9)

Cat (0.3)

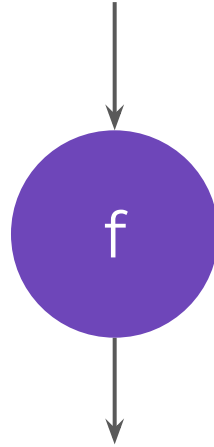


saliency map

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!

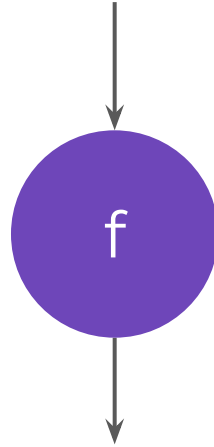


y

Overly enthusiastic (0.98)

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)



rationalisation

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

variable
selection

input relevance

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 Kingma¹, Julius Adebayo¹
Google Brain*
{pikinder, adebayo}@google.com

Maximilian Alber¹
TU-Berlin

EXPLORATORY NOT EXPLANATORY: COUNTERFACTUAL ANALYSIS OF SALIENCY MAPS FOR DEEP REINFORCEMENT LEARNING

Shikha Atrey, Kaleigh Clary & David
University of Massachusetts Lowell
atrey, kclary, kolter@uml.edu

On the Robustness of Interpretability Methods

David Alvarez-Melis¹ Tommi S. Jaakkola¹

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

When Explanations Lie: Why Many Modified BP Attributions Fail

Sixt¹ Maximilian Granz¹ Tim Landgraf¹

Problems with Shapley-value-based explanations measures

Elizabeth Kumar¹ Suresh

Abstract

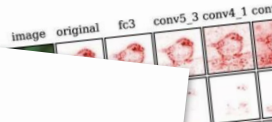
theoretic formulations of feature
become popular as a way to
learning models. These
game theory

Challenging common interpretability assumptions in feature attribution explanations

I. Zico Kolter

Explanations can be manipulated and geometry is to blame

Kathrin Dombrowski¹, Maximilian Alber¹, Christopher J. Anders¹,
Marcel Ackermann², Klaus-Robert Müller^{1,3,4}, Pan Kessel¹
Machine Learning Group, EE & Computer Science Faculty, TU-Berlin
Department of Video Coding & Analytics, Fraunhofer Heinrich-Hertz Inst
³Max Planck Institute for Inf
Department of Ps



Interpretability is in question

Is Attention Interpretable?

Sofia Serrano* Noah A. Smith*†

G. Allen School of Computer Science & Engineering
University of Washington, Seattle, WA, USA
Institute for Artificial Intelligence, Seattle, WA, USA
sofias6, nasmit@cs.washington.edu

Attention is not Expl

Sarthak Jain

Northeastern University
jain.sar@husky.neu.edu

Abstract

mechanisms have seen wide adop-

after
mo
do
to

Attention is not not Explanation

Sarah Wiegrefe*
School of Interactive Computing
Georgia Institute of Technology
saw@gatech.edu

Yuval Pinter*
School of Interactive Computing
Georgia Institute of Technology
uvp@gatech.edu

Is Multi-Hop Reasoning Really Explainable? Towards Benchmarking Reasoning Interpretability

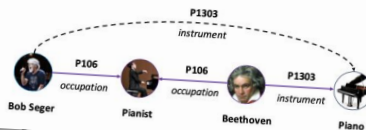
Xin Lv^{1,2}, Yixin Cao³, Lei Hou^{1,2}, Juanzi Li^{1,2},
Zhiyuan Liu^{1,2}, Yichi Zhang⁴, Zelin Dai⁴

¹Department of Computer Science and Technology, BNRist
²KIRC, Institute for Artificial Intelligence, Tsinghua University, Beijing 100084, China
³Nanyang Technological University, Singapore
⁴Alibaba Group, Hangzhou, China
lv-x18@mails.tsinghua.edu.cn, yixin.cao@ntu.edu.sg
{houlei, lijuanzi, liuzy}@tsinghua.edu.cn

Abstract

Multi-hop reasoning has been widely studied in recent years to obtain more interpretable link prediction. However, we find in experiments that many paths given by these models are actually unreasonable, while little works

Triple Query: (Bob Seger, instrument, ?)



... can be manipulated d geometry is to blame

Dombrowski¹, Maximilian Alber¹, Christopher J. Anders¹,
Ackermann², Klaus-Robert Müller^{1,3,4}, Pan Kessel¹

¹Learning Group, EE & Computer Science Faculty, TU-Berlin
²Video Coding & Analytics, Fraunhofer Heinrich-Hertz
³Max Planck Institute for Informatics
⁴Department of Brain and Cognitive Sciences, MIT

Interpretability is in question

The Mythos of Model Interpretability

Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead

Cynthia Rudin
Duke University
cynthia@cs.duke.edu

John C. Lipton¹

Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning

Harmanpreet Kaur¹, Harsha Nori², Samuel Jenkins²,
Rich Caruana², Hanna Wallach², Jennifer Wortman Vaughan²
¹University of Michigan, ²Microsoft Research
{hankaur, harsha.nori, saienkin, rcaruana, wallach, jenn}@microsoft.com

Towards A Rigorous Science of Interpretable Machine Learning

Finale Doshi-Velez* and Been Kim*

Georgia Institute of Technology
saw@gatech.edu

Georgia Institute of Technology
uvp@gatech.edu

Developments create countless opportunities for impact, these opportunities come new challenges. ML models are found to amplify societal biases in datasets and lead to outcomes [4, 9, 29]. When ML models have the potential to affect people's lives, it is critical that their developers understand and justify their behavior. More generalist Müller, Pan Kessel

Computer Science Faculty, TU-Berlin
Fraunhofer Heinrich-Hertz Institute

³Max Planck Institute for Informatics

Abstract

mechanisms have seen wide adop-

Interpretability is in question

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

Stop Explaining Bla

Interpretability

con¹

managed to set it

Duke University
cynthia@cs.duke.edu

Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning

Harmanpreet Kaur¹, Harsha Nori², Samuel Jenkins², Rich Caruana², Hanna Wallach², Jennifer Wortman Vaughan²
¹University of Michigan, ²Microsoft Research
{hankaur, harsha.nori, saienkin, rcaruana, wallach, jenn}@microsoft.com

Towards A Rigorous Science of Interpretable Machine Learning

Finale Doshi-Velez* and Been Kim*

Georgia Institute of Technology
saw@gatech.edu

Georgia Institute of Technology
uvp@gatech.edu

developments create countless opportunities for impact, these opportunities come new challenges. ML models n found to amplify societal biases in datasets and lead outcomes [4, 9, 29]. When ML models have the po affect people's lives, it is critical that their developers to understand and justify their behavior. More gener- ert Müller¹, Pan Kessel¹

Computer Science Faculty, TU-Berlin
Fraunhofer Heinrich-Hertz Institute

³Max Planck Institute for In-

Abstract

mechanisms have seen wide adop-

after
mo
do
to

Interpretability is in question

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

... can we mathematically write it down?

Stop Explaining Bl

Interpretability

Duke University
cynthia@cs.duke.edu

Interpreting Interpretability: Understanding Data Scientists' Machine Learning

Towards A

Jenkins²,
an Vaughan²
arch
enn}@microsoft.com

Georgia Institute of Technology
saw@gatech.edu

Georgia Institute of Technology
uvp@gatech.edu

ents create countless opportunities for impact,
opportunities come new challenges. ML models
found to amplify societal biases in datasets and lead
outcomes [4, 9, 29]. When ML models have the po-
affect people's lives, it is critical that their developers
to understand and justify their behavior. More gener-
ert Müller¹, Pan Kessel¹

Abstract

mechanisms have seen wide adop-

³Max Planck Institute
Fraunhofer Heinrich-Hertz
computer Science Faculty, TU-Berlin

Yes.

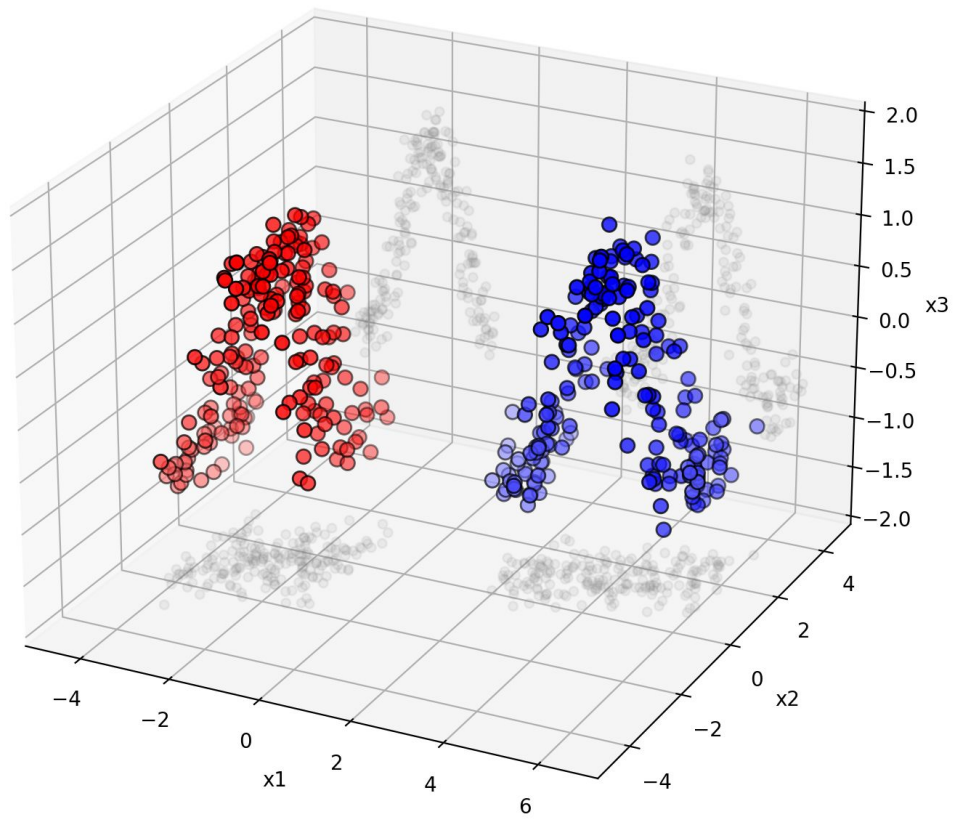
- Leveraging simple considerations on functionality
- No evasive / intuitive concepts
- Allows task-specificity

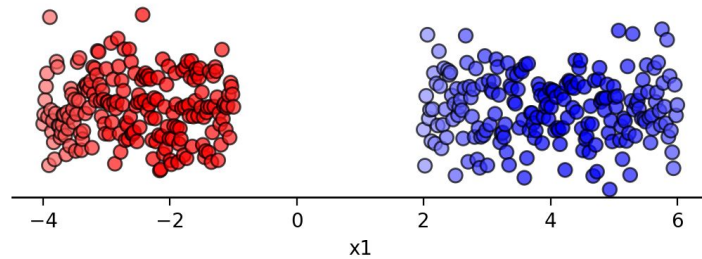
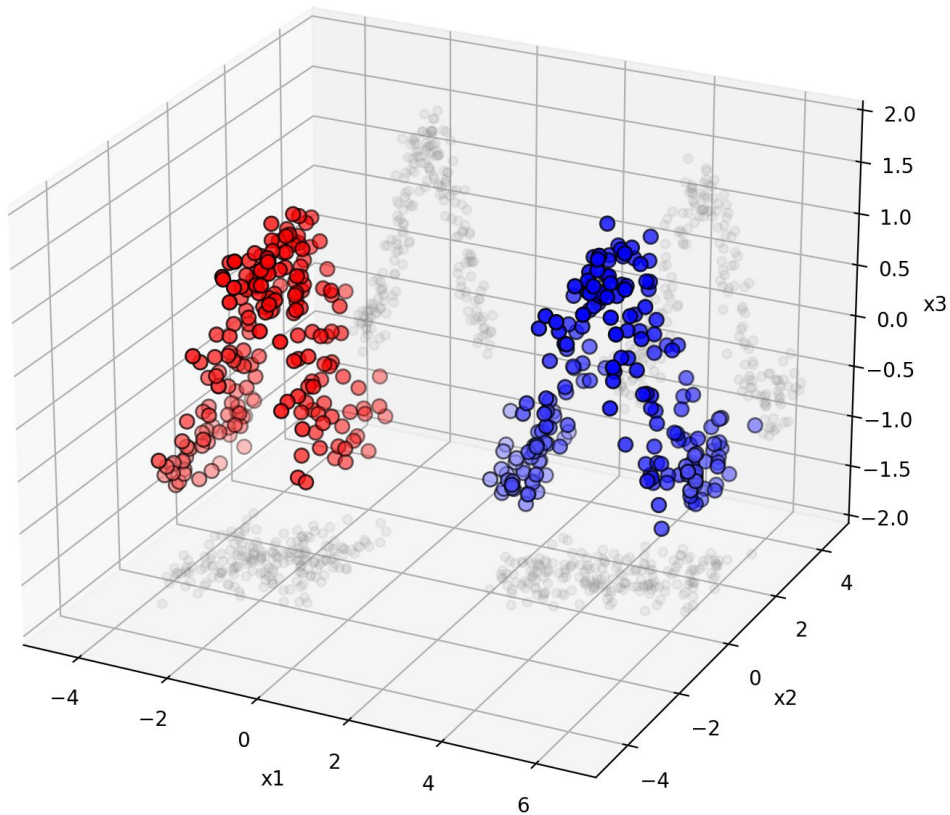
Problem 1 (Global subset selection). *Given a relation R , find a subset of indices $I^* \subset [n]$ that minimises*

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

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*

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



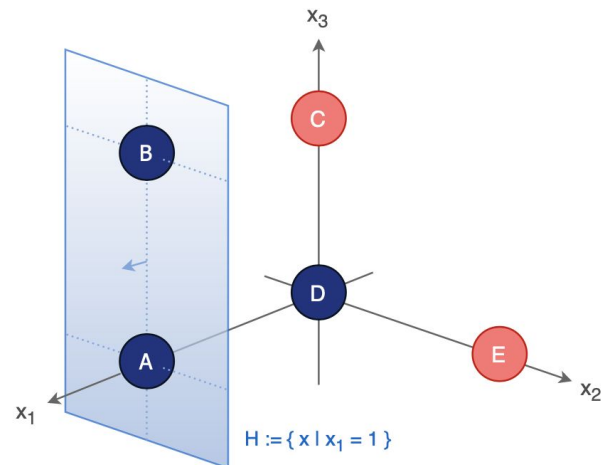


$$f_1 := \begin{cases} x_1 > 0 \mapsto \text{blue} \\ x_1 \leq 0 \mapsto \text{red} \end{cases}$$

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**

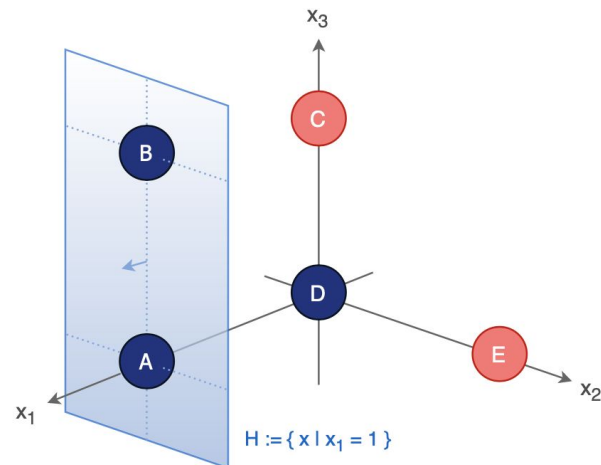


Formalising: what is at stake

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)
<i>LIME</i> (Cat.)	16.2 ± 1.3	0:05:54
<i>LIME</i> (Cont.)	27.4 ± 1.6	0:05:47
<i>attr-GAM</i>	24.5 ± 1.5	0:00:25
Shapley ($\mathbb{E}(f)$)	74.3 ± 1.1	0:16:29
<i>SHAP</i> ($f(\mathbb{E})$)	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
<i>Integrated Gradient</i>	18.5 ± 1.4	0:00:24
<i>Expected Gradient</i>	21.4 ± 1.4	0:03:42
<i>attr-GA</i>^{∞}M	81.7 ± 1.1	0:17:44*
<i>attr-GA</i> ² M	52.5 ± 1.8	$\ll *$
<i>attr-GA</i> ³ M	74.1 ± 1.3	$< *$
<i>attr-GA</i> ⁴ M	81.2 ± 1.1	$< *$
<i>InterpretableNN</i>	79.7 ± 1.2	$\simeq *$
<i>Archipelago</i>	70.2 ± 1.1	$\simeq *$
<i>L2X</i>	23.7 ± 1.6	32:53:16
<i>INVASE</i>	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 (%)
<i>LIME</i> (Cat.)	29.9 ± 1.7
<i>LIME</i> (Cont.)	46.6 ± 1.6
<i>attr-GAM</i>	61.5 ± 1.1
Shapley ($\mathbb{E}(f)$)	79.5 ± 1.1
<i>SHAP</i> ($f(\mathbb{E})$)	23.7 ± 1.5
Gradient	61.6 ± 1.3
Gradient × Input	54.5 ± 1.3
<i>Integrated Gradient</i>	39.7 ± 1.5
<i>Expected Gradient</i>	41.8 ± 1.5
<i>attr-GA</i>[∞]<i>M</i>	92.9 ± 0.6
<i>attr-GA</i> ² <i>M</i>	63.7 ± 1.4
<i>attr-GA</i> ³ <i>M</i>	81.2 ± 1.4
<i>attr-GA</i> ⁴ <i>M</i>	90.7 ± 1.1
<i>InterpretableNN</i>	86.9 ± 0.9
<i>Archipelago</i>	88.8 ± 0.7
<i>L2X</i>	37.5 ± 1.6
<i>INVASE</i>	61.3 ± 1.7

Thank you!

Towards Rigorous Interpretation: a Formalisation of
Feature Attribution

Darius Afchar^{1,2}, Romain Hennequin¹, Vincent Guigue²