1.2 What and How of Modular Interpretability

A Case Study of Bespoke Surrogate Explainers for Tabular Data

- Decomposition of surrogates.
- Explaining text with surrogates.
- Explaining images with surrogates.
- Surrogates for tabular data.
  - Interpretable data representation.
  - Data sampling.
  - Explanation generation.
- Building surrogates from scratch.

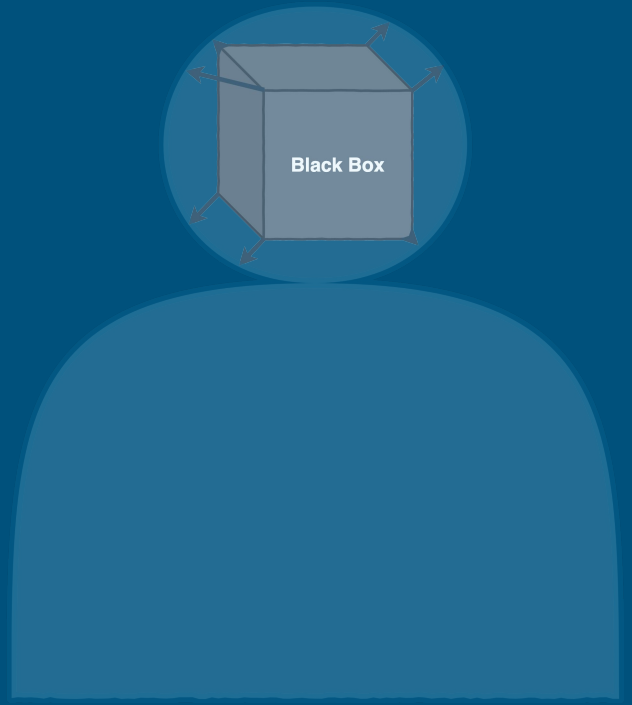
Kacper Sokol

What is Explainability?

No universally accepted definition.

- The Chinese Room Theorem
  - *Searle, 1980. Minds, brains, and programs.*
- Simulatability
  - *Lipton, 2018. The mythos of model interpretability.*
- Mental Models:

  Kulesza et al., 2013. Too much, too little, or just right? Ways explanations impact end users' mental models.
  - Functional -- operationalisation without understanding;
  - Structural -- appreciation of the underlying mechanism.



What is Explainability?

$$\texttt{Explainability} = \underbrace{\texttt{Reasoning}\,(\texttt{Transparency} \mid \texttt{Background Knowledge})}_{understanding}$$

- Transparency -- **insight** (of arbitrary complexity) into operation of a system.
- Background Knowledge -- implicit or explicit **exogenous** information.
- Reasoning -- **algorithmic** or **mental** processing of information.

Explainability → **explainee** walking away with **understanding**.

Explaining AI

Exemplars:

- [humidity=low, temperature=23, rain=no] → like
- [humidity=medium, temperature=23, rain=no] → dislike
- [humidity=high, temperature=23, rain=no] → dislike

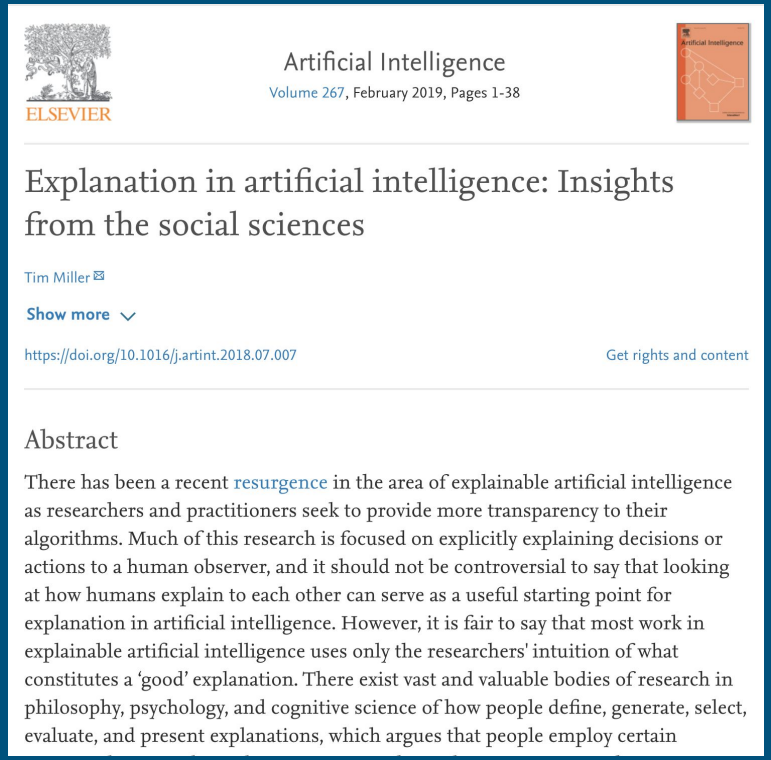
Logical rules:

- [humidity={low ∨ medium} ∧ 12<temperature≤22 ∧ rain=no] → like
- [humidity={medium ∨ high} ∧ 23<temperature≤27 ∧ rain=no] → dislike

With **23°C** and **clear sky**, the person *enjoys* the weather when the **humidity is low**.

With **clear sky** and **medium humidity**, the person *enjoys* the weather when the **temperature is between 13°C and 22°C**.

Human-centred Explainability



ELSEVIER

Artificial Intelligence

Volume 267, February 2019, Pages 1-38

Explanation in artificial intelligence: Insights from the social sciences

Tim Miller ✉

Show more

https://doi.org/10.1016/j.artint.2018.07.007

Get rights and content

Abstract

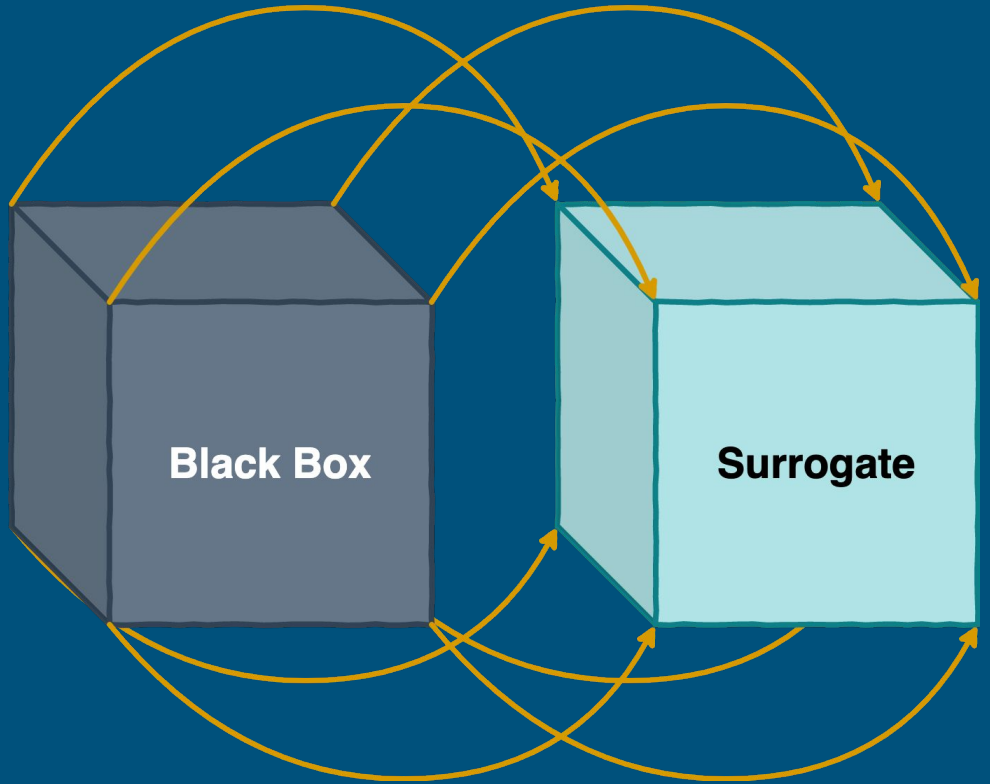
There has been a recent resurgence in the area of explainable artificial intelligence as researchers and practitioners seek to provide more transparency to their algorithms. Much of this research is focused on explicitly explaining decisions or actions to a human observer, and it should not be controversial to say that looking at how humans explain to each other can serve as a useful starting point for explanation in artificial intelligence. However, it is fair to say that most work in explainable artificial intelligence uses only the researchers' intuition of what constitutes a 'good' explanation. There exist vast and valuable bodies of research in philosophy, psychology, and cognitive science of how people define, generate, select, evaluate, and present explanations, which argues that people employ certain

Human-centred explanations:

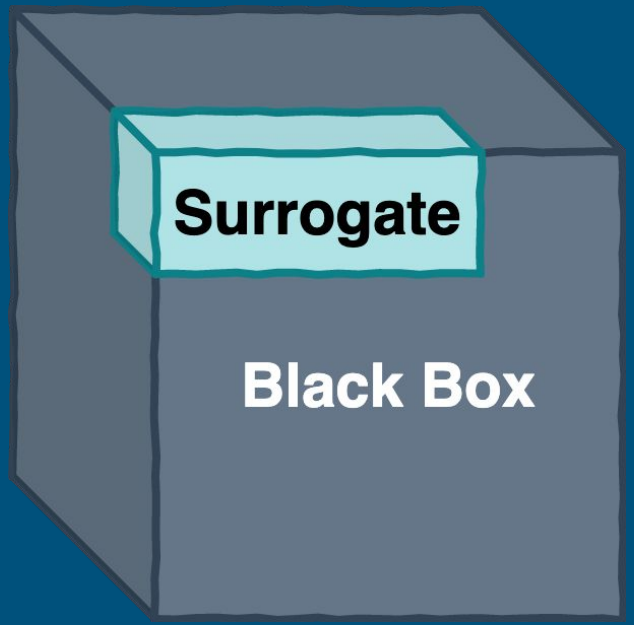
- interactive dialogue (bi-directional);
- contrastive statements.

Modular Surrogate Explainability

Scope: Global Surrogates

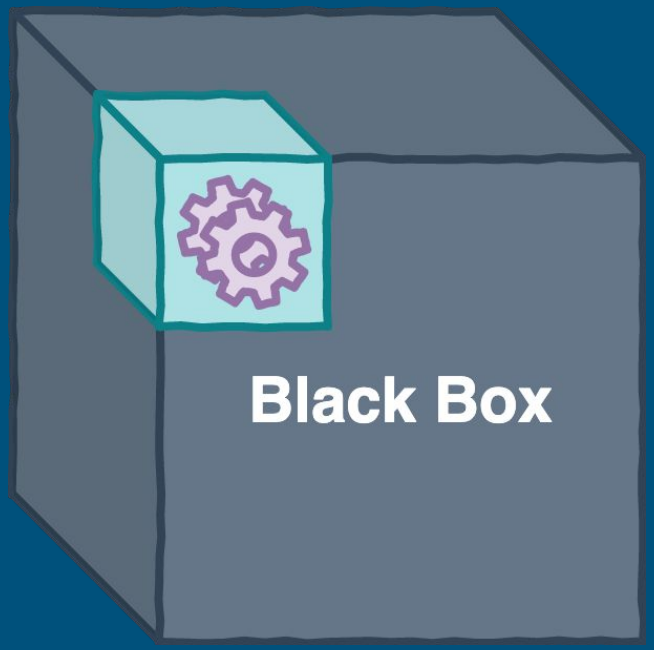


Scope: Local Surrogates



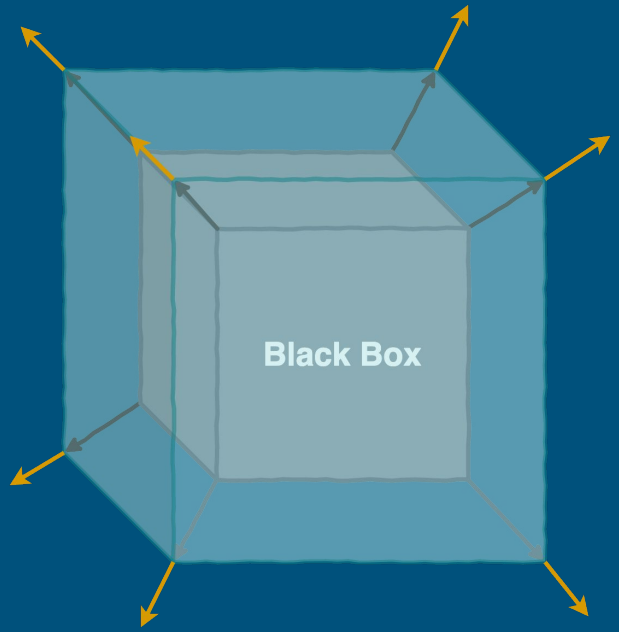
Mimicry Source: Structural Surrogate

(model)

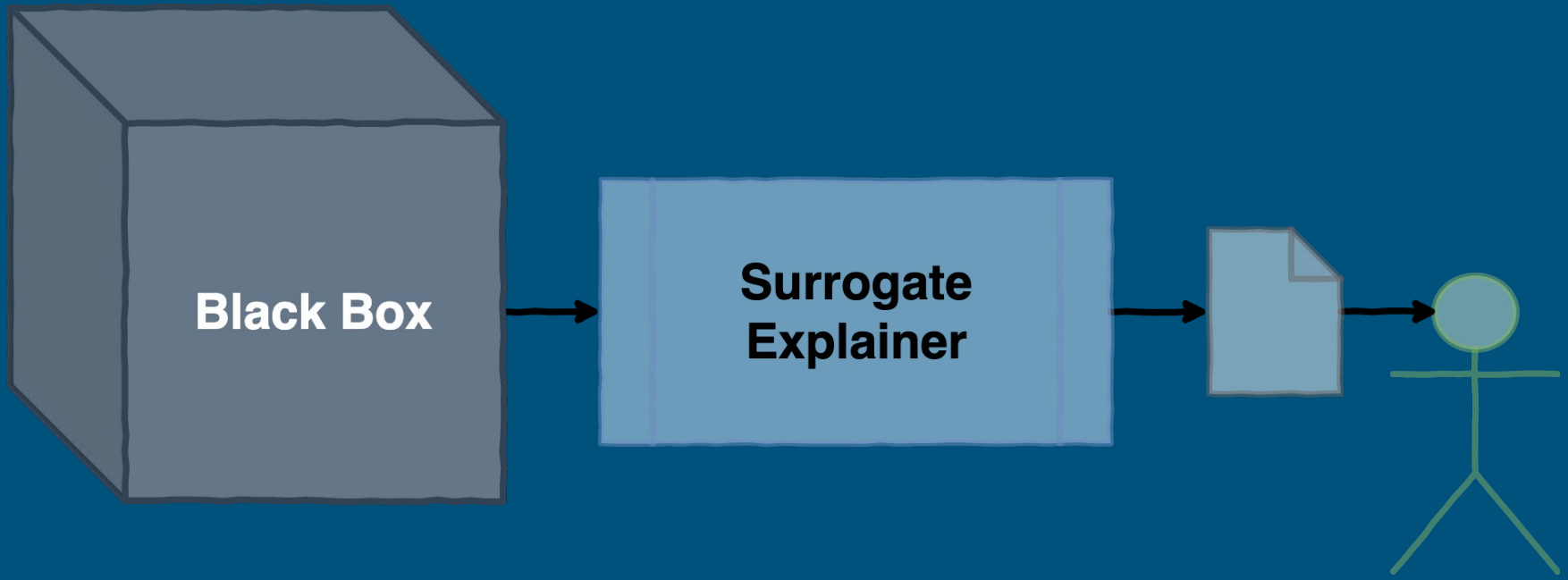


Mimicry Source: Behavioural Surrogate

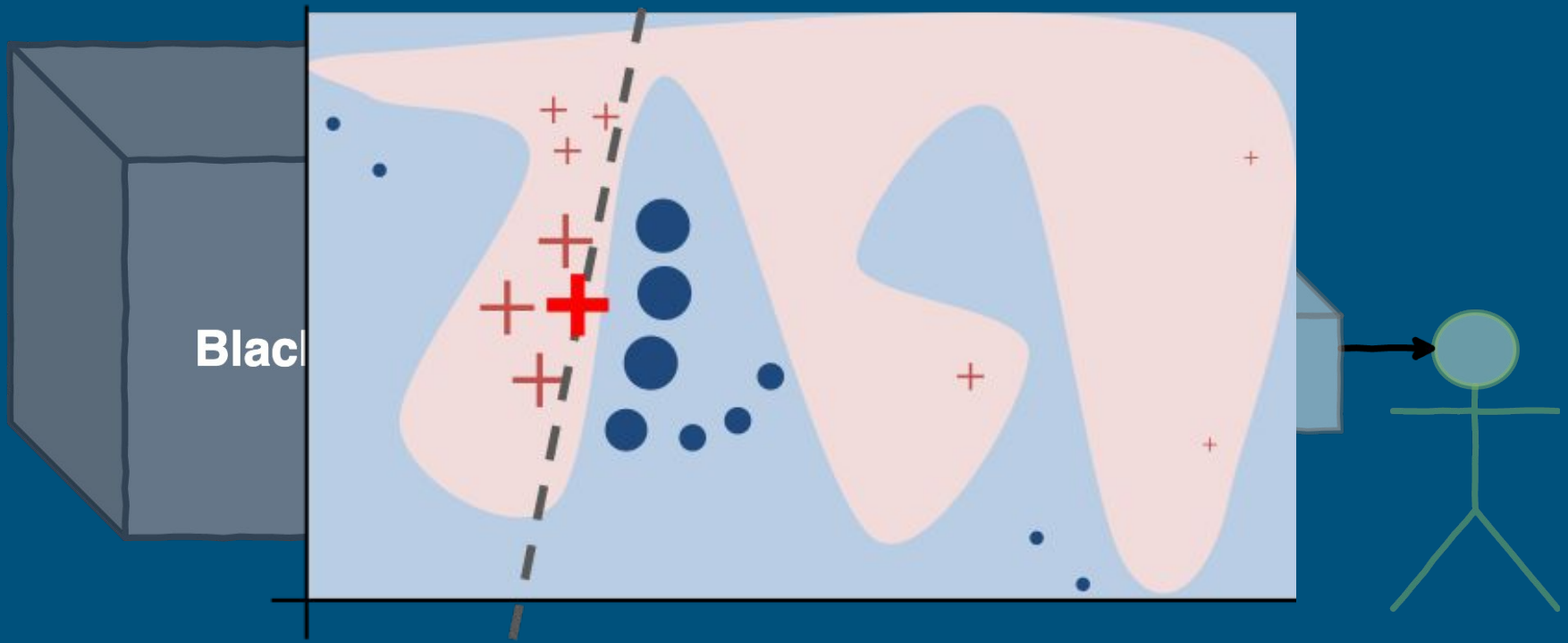
(predictions)



Surrogate Explainability in AI and ML



Surrogate Explainability in AI and ML



Ribeiro et al., 2016. "Why should I trust you?" Explaining the predictions of any classifier.

The Benefits of Surrogates

- **Model-agnostic** -- work with *any* black box.
- **Post-hoc** -- can be retrofitted into pre-existing predictors.
- **Data-universal** -- work with image, tabular and text data because of *interpretable data representations*.

Actual Surrogate Explainers

- Local Interpretable Model-agnostic Explanations (LIME)
  - *Ribeiro et al., 2016. "Why should I trust you?" Explaining the predictions of any classifier.*
- Anchor
  - *Ribeiro et al., 2018. Anchors: High-Precision Model-Agnostic Explanations.*
- SHapley Additive exPlanations (SHAP)
  - *Lundberg and Lee, 2017. A unified approach to interpreting model predictions.*
- RuleFit
  - *Friedman and Popescu, 2008. Predictive learning via rule ensembles.*

Universally Applicable Explainers

Caveat: No Free Lunch Theorem

Post-hoc Explainers have Poor Fidelity

nature machine intelligence

Perspective | Published: 13 May 2019

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

Nature Machine Intelligence **1**, 206–215(2019) | Cite this article

23k Accesses | **151** Citations | **233** Altmetric | Metrics

Abstract

Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that

Post-hoc Explainers have Poor Fidelity

nature machine intelligence

Explore our content ⌄ Journal information ⌄

nature › nature machine intelligence › perspectives › article

Perspective | Published: 13 May 2019

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin ✉

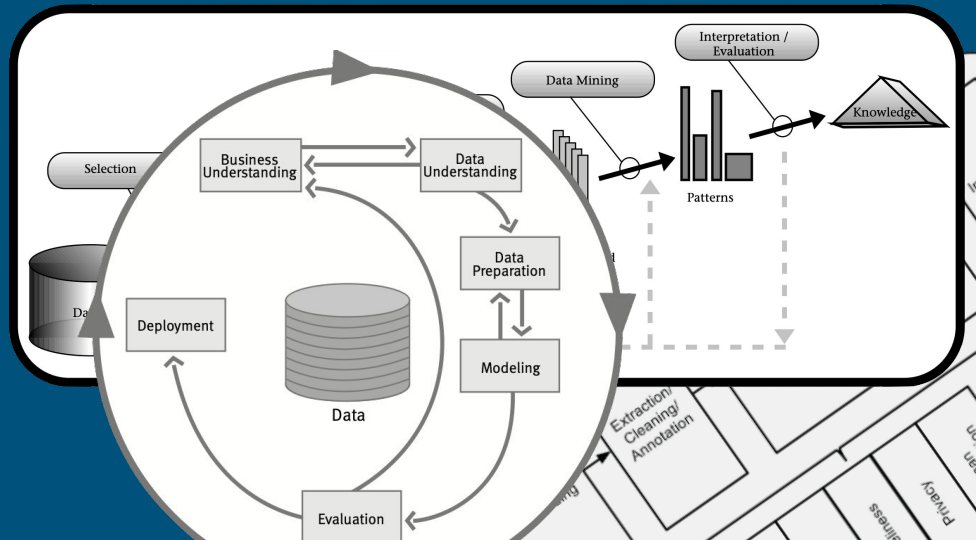
Nature Machine Intelligence **1**, 206–215(2019) | Cite this article

23k Accesses | **151** Citations | **233** Altmetric | Metrics

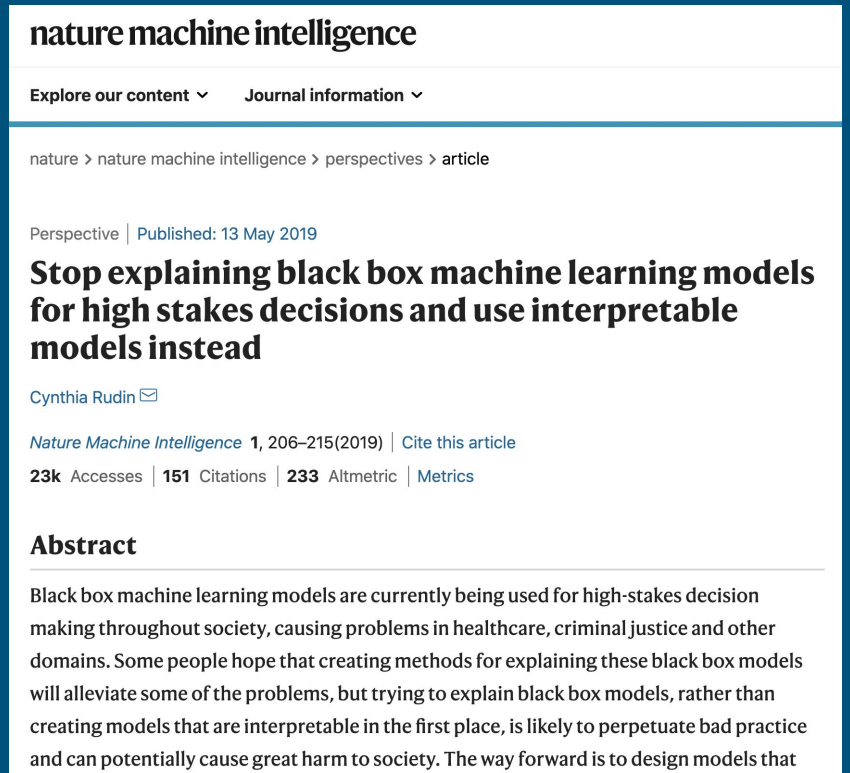
Abstract

Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that

- Explainability needs a process similar to: **KDD**, **CRISP-DM** or **BigData**.



Post-hoc Explainers have Poor Fidelity



nature machine intelligence

Explore our content ⌄ Journal information ⌄

nature › nature machine intelligence › perspectives › article

Perspective | Published: 13 May 2019

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin ✉

Nature Machine Intelligence **1**, 206–215(2019) | Cite this article

23k Accesses | **151** Citations | **233** Altmetric | Metrics

Abstract

Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that

- Explainability needs a process similar to: KDD, CRISP-DM or BigData.
- Engineers should spend time creating **informative features** and building **inherently transparent models**.

Bottom Line: It requires **effort**.

XAI Process

- A **generic** eXplainable Artificial Intelligence Process is beyond our reach at the moment.
- Attempts:
  - XAI Taxonomy spanning social and technical desiderata.
    - *Sokol and Flach, 2020. Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches.*
  - Generic framework for black-box explainers.
    - *Henin and Le Métayer, 2019. Towards a generic framework for black-box explanations of algorithmic decision systems.*

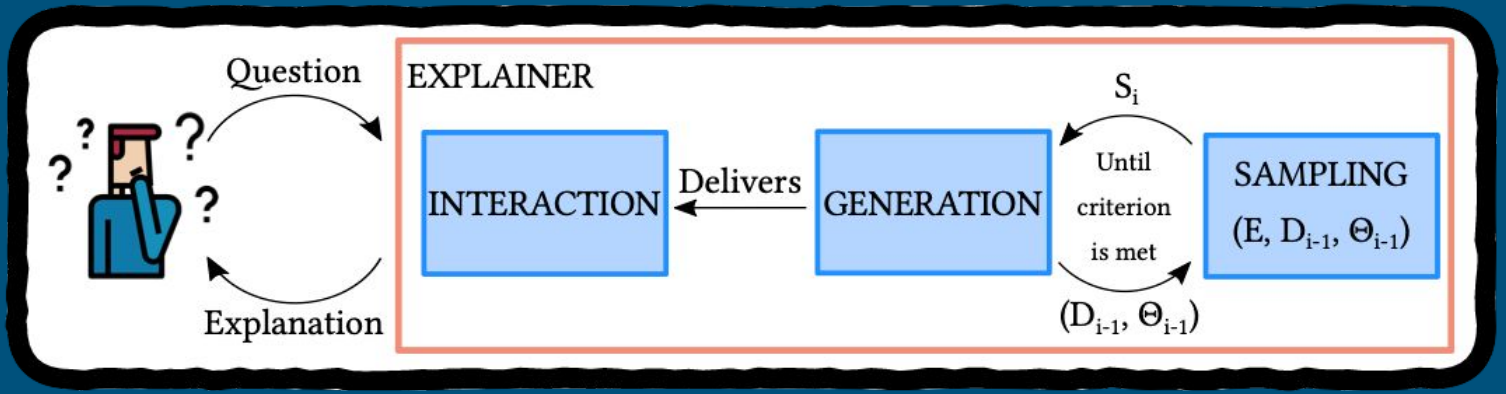
But we can design one for surrogate explainers.

XAI Process

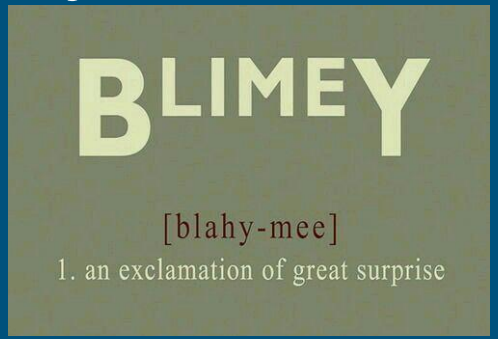
XAI Taxonomy spanning **social** and **technical** desiderata.

Functional	Operational	Usability	Safety	Validation
...	...	...	...	

Generic framework for black-box explainers.



bLIMEy, there has to be a better way...



bLIMEy → build LIME yourself

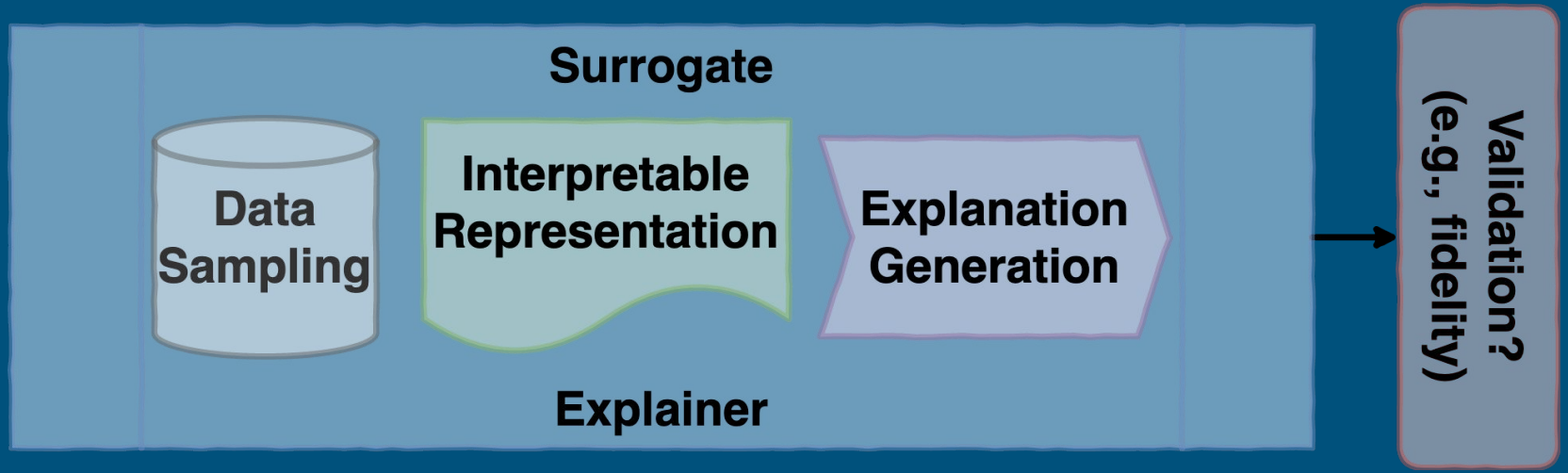
- *Framework* for building surrogate explainers.
- *Meta-algorithm* for operationalising them.
- Accompanied by analysis of surrogate building blocks (akin to a user guide).
- Practical recommendations (**this tutorial**, especially the hands-on part).

Good news: A means to build flexible, faithful, interactive, ... surrogates.

Not so good news: It requires **effort**.

Sokol et al., 2019. bLIMEy: Surrogate prediction explanations beyond LIME

Building Blocks of Surrogate Explainers

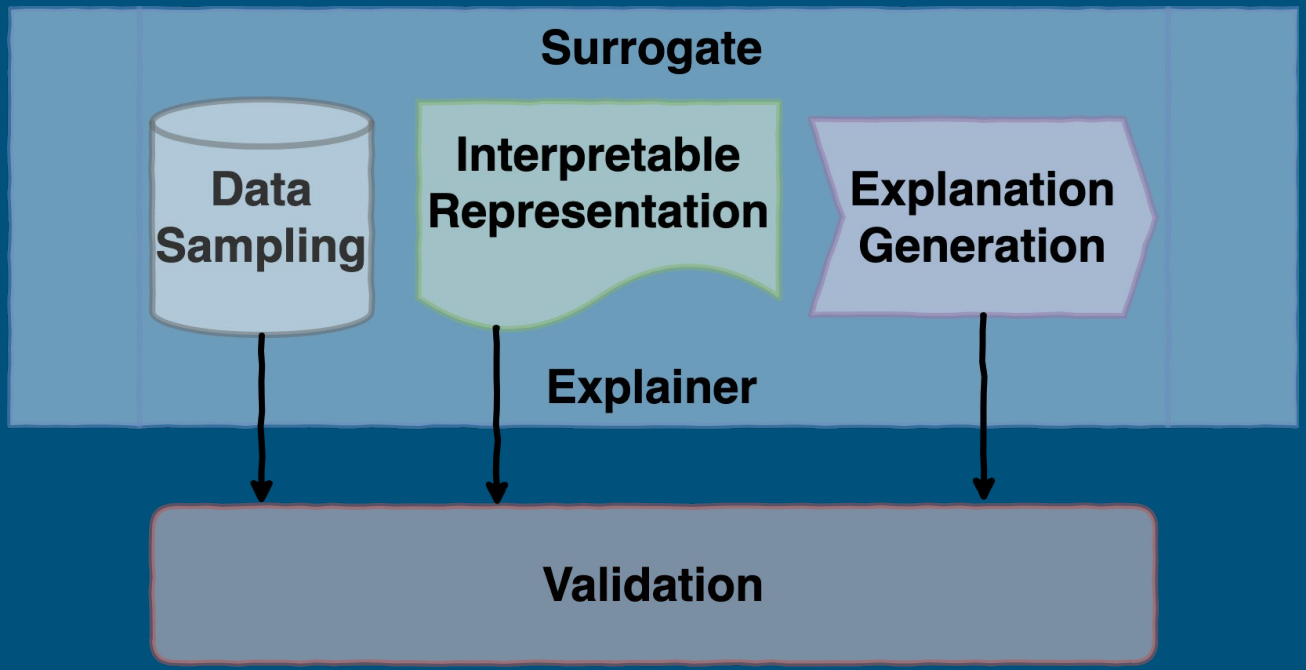


bLIMEy Framework

Tabular

Image

Text



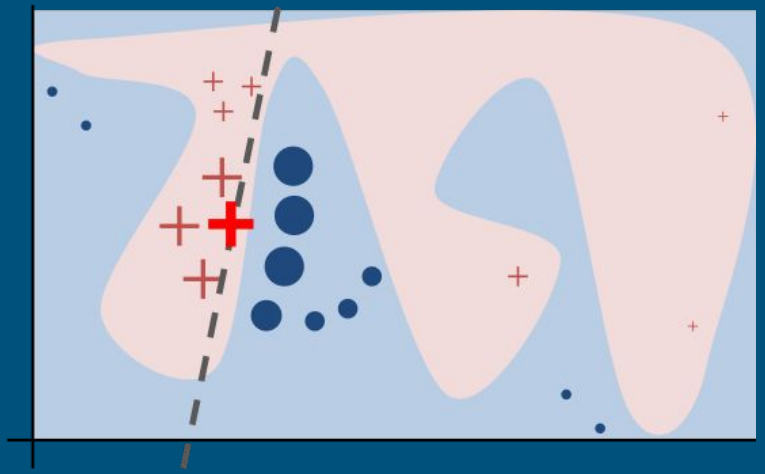
Learning Outcomes

Theory:

- Understand where surrogates come from.
- Have the right intuition about them.
- Know how to (correctly) interpret their explanations.

Practice (hands-on part):

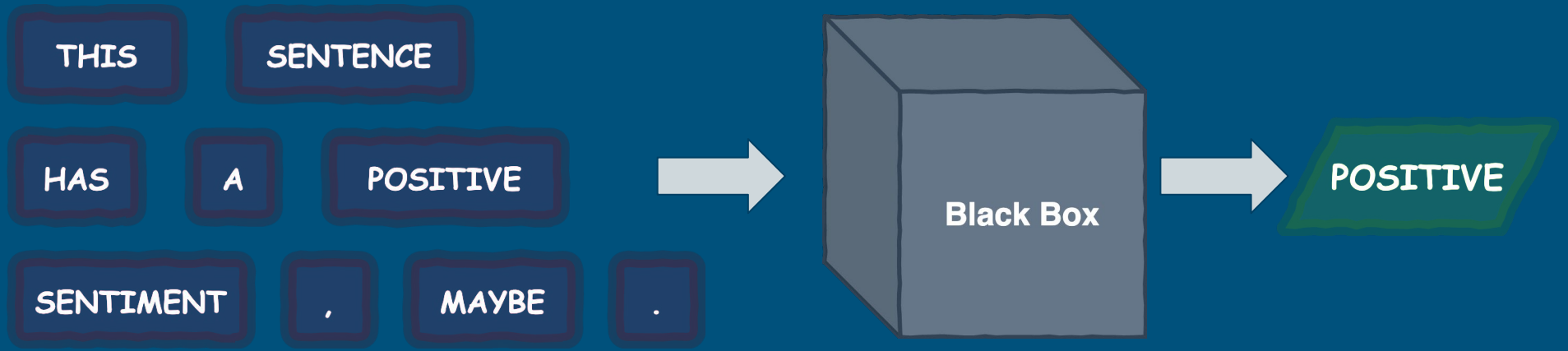
- Find out what components are needed and how to choose them.
- Learn how to evaluate them.
- Know how to use them to build explainers.



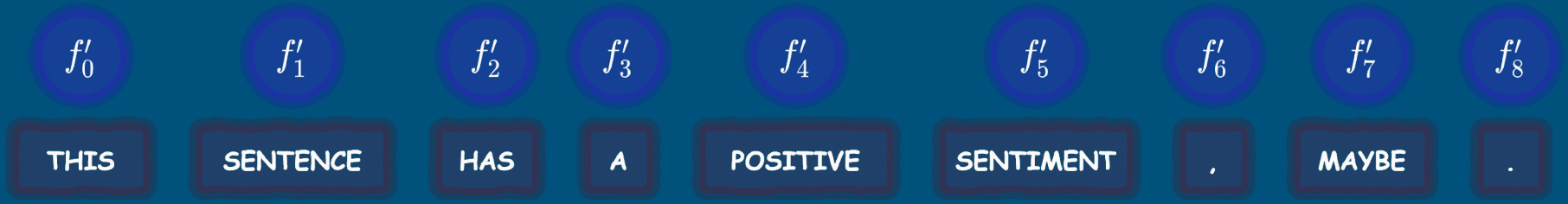
Ribeiro et al., 2016. "Why should I trust you?" Explaining the predictions of any classifier.

Surrogates for Text Data

Black-box Prediction

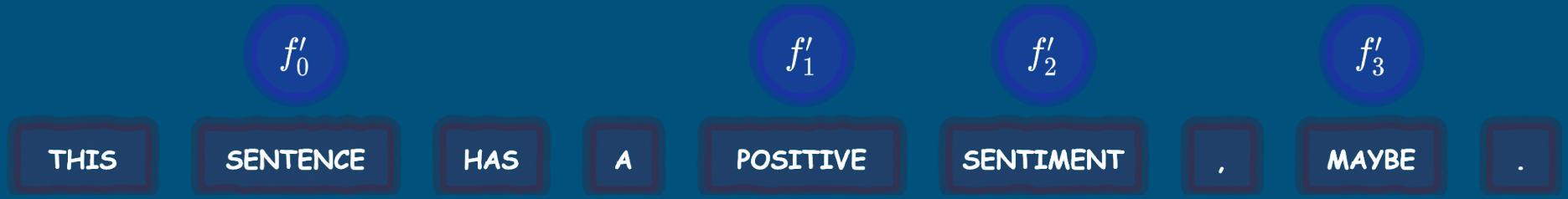


Interpretable Data Representation



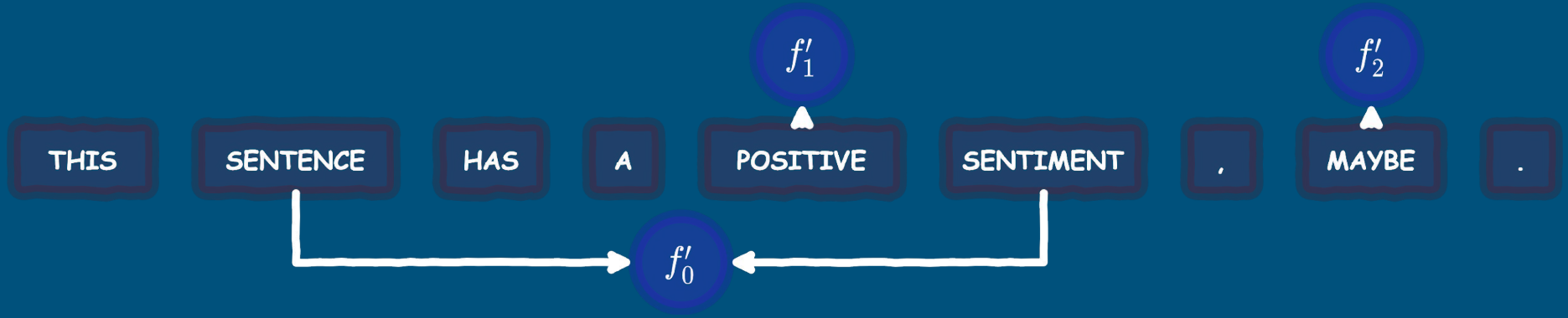
$$\mathring{x}' = [1, 1, 1, 1, 1, 1, 1, 1, 1]$$

Interpretable Data Representation ★



$$\mathring{x}' = [1, 1, 1, 1]$$

Interpretable Data Representation ★



$$\mathring{x}' = [1, 1, 1]$$

Data Sampling

- Sampling from the original data domain is ill-defined.
- We have to sample from the interpretable representation.

```
randint(low=0, high=1, shape=(3, 4))
```

$$x'_0 = [1, 0, 0, 1] \qquad x'_2 = [0, 0, 0, 1]$$

$$x'_1 = [1, 0, 1, 0]$$

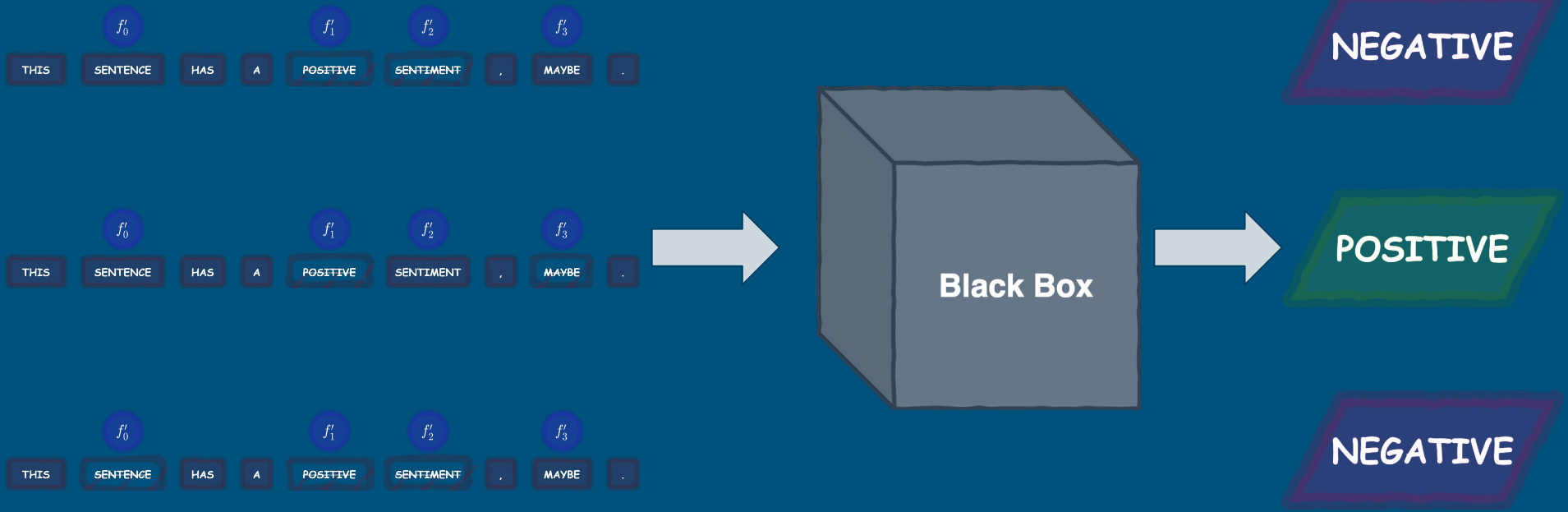
Explanation Generation -- Restoring Sample

$$x'_0 = [1, 0, 0, 1]$$

$$x'_1 = [1, 0, 1, 0]$$

$$x'_2 = [0, 0, 0, 1]$$

Explanation Generation -- Predicting Sample



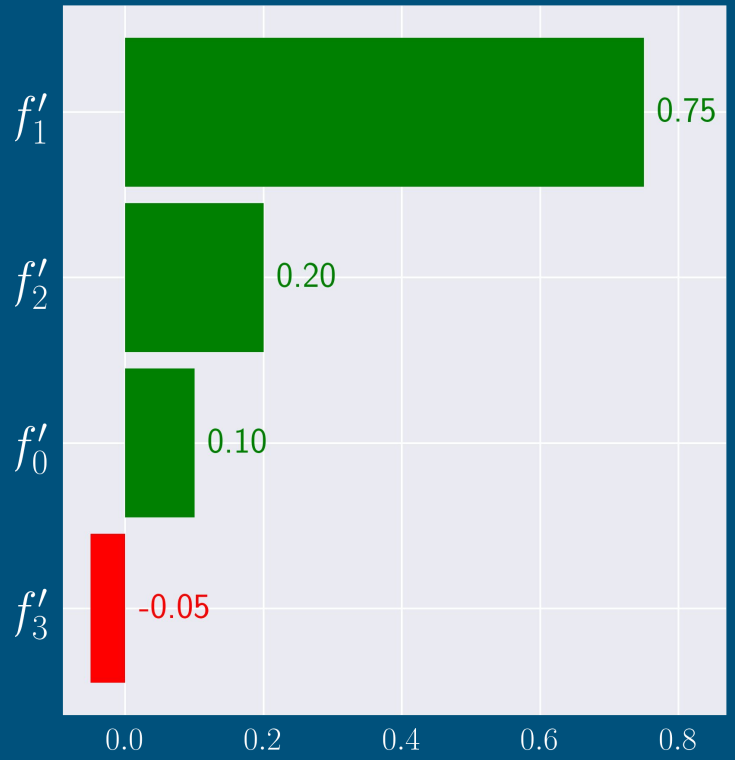
Explanation Generation -- Weighting Sample

$\mathring{x}' = [1, 1, 1, 1]$

	Distance	Similarity
$x'_0 = [1, 0, 0, 1]$	d=2	s=0.62
$x'_1 = [1, 0, 1, 0]$	d=2	s=0.62
$x'_2 = [0, 0, 0, 1]$	d=3	s=0.23

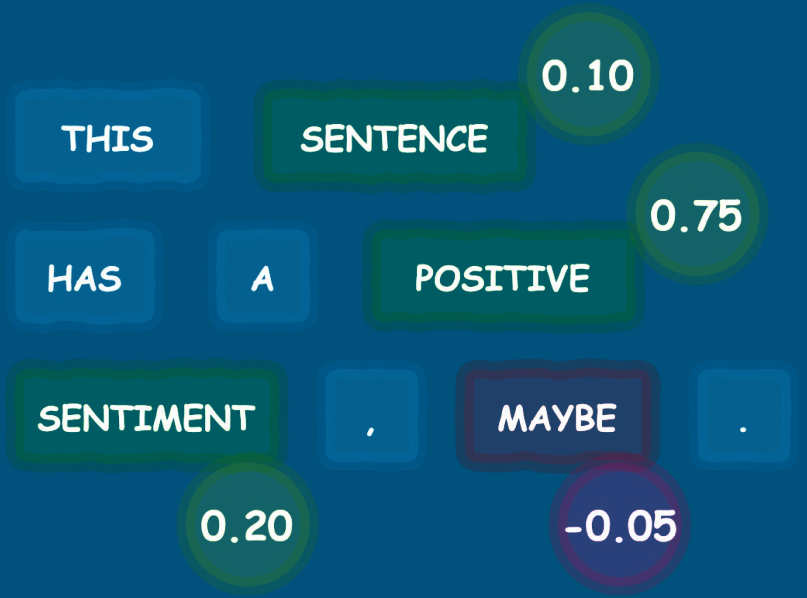
Explanation Generation -- Fitting Surrogate

- Fit a sparse linear classifier.
- Use the binary vectors as data.
- Use the black box-predictions for the sampled and reversed data as labels:
  - 0 -- negative sentiment;
  - 1 -- positive sentiment.
- Weight instances according to the similarity scores.
- Extract model coefficients as importance scores (explanation).



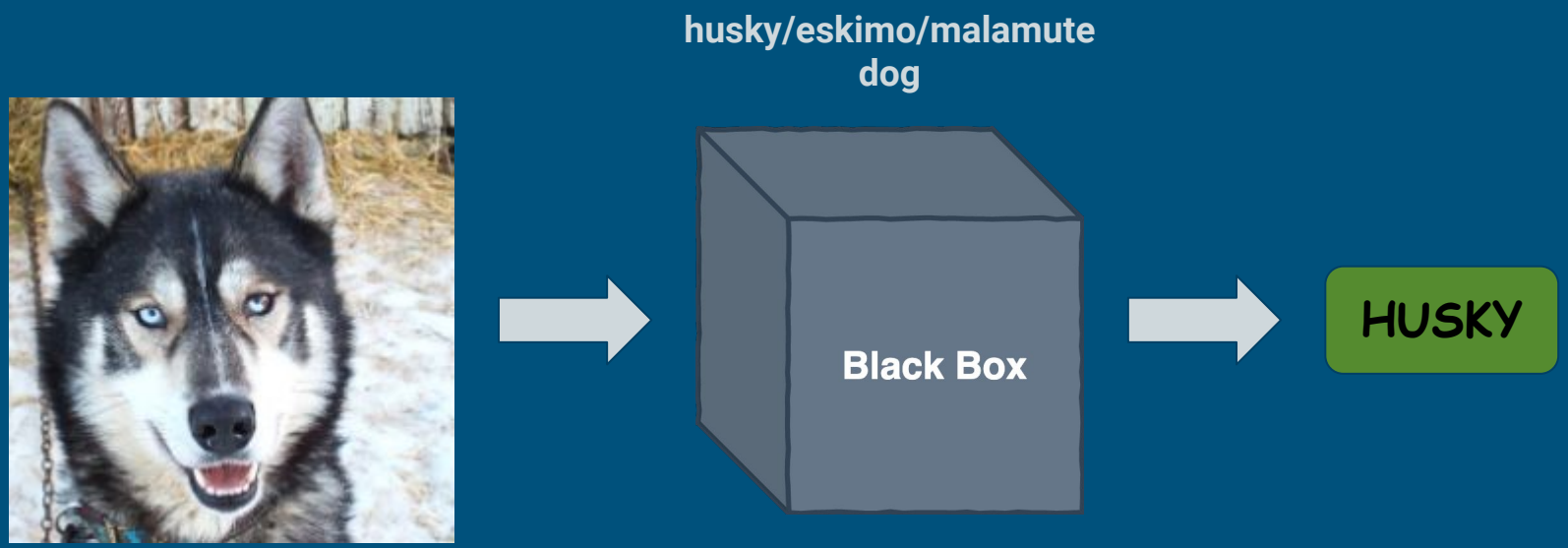
Explanation Generation -- Fitting Surrogate

- Fit a sparse linear classifier.
- Use the binary vectors as data.
- Use the black box-predictions for the sampled and reversed data as labels:
  - 0 -- negative sentiment;
  - 1 -- positive sentiment.
- Weight instances according to the similarity scores.
- Extract model coefficients as importance scores (explanation).



Surrogates for Image Data

Black-box Prediction



Interpretable Data Representation

$$\mathring{x}' = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]$$

Interpretable Data Representation ★

$$\mathring{x}' = [1, 1, 1, 1, 1, 1, 1]$$

Data Sampling

- Sampling from the original data domain is ill-defined.
- We have to sample from the interpretable representation.

```
randint(low=0, high=1, shape=(3, 7))
```

$$x'_0 = [0, 0, 0, 1, 1, 1, 1] \quad x'_1 = [0, 1, 1, 1, 0, 1, 0]$$

$$x'_2 = [1, 1, 1, 0, 1, 1, 1]$$

Explanation Generation -- Restoring Sample

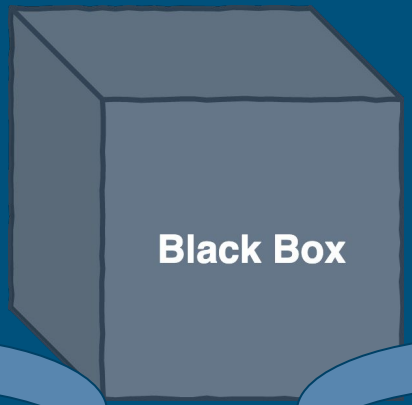
How to remove super-pixels?

Using occlusion.

$$x'_0 = [0, 0, 0, 1, 1, 1, 1] \qquad x'_1 = [0, 1, 1, 1, 0, 1, 0]$$

Explanation Generation -- Restoring Sample

What occlusion colour to use?

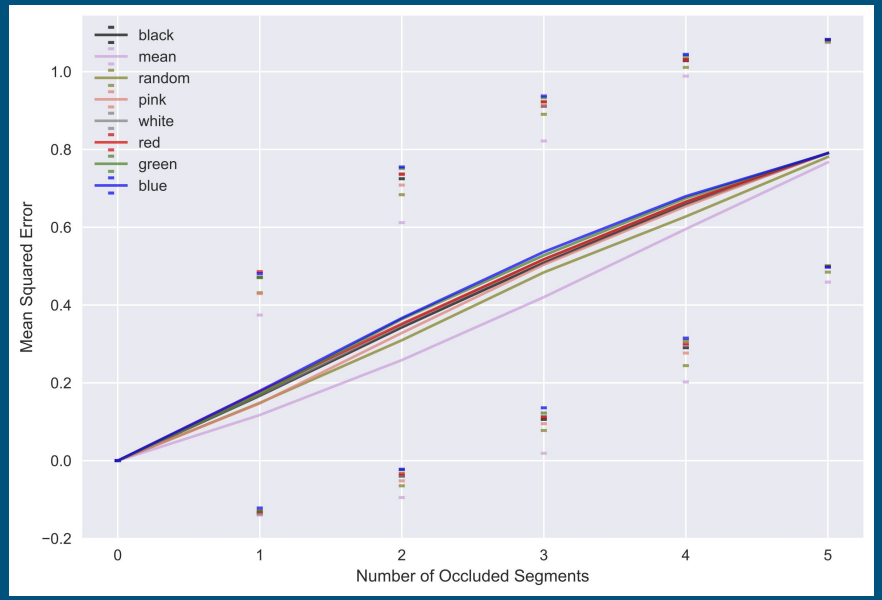


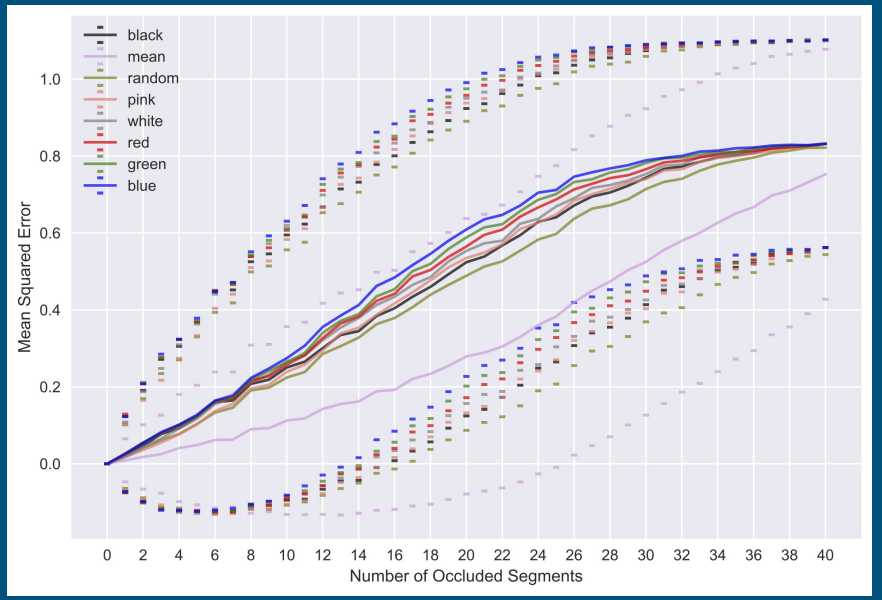
MALAMUTE

HUSKY

Explanation Generation -- Restoring Sample

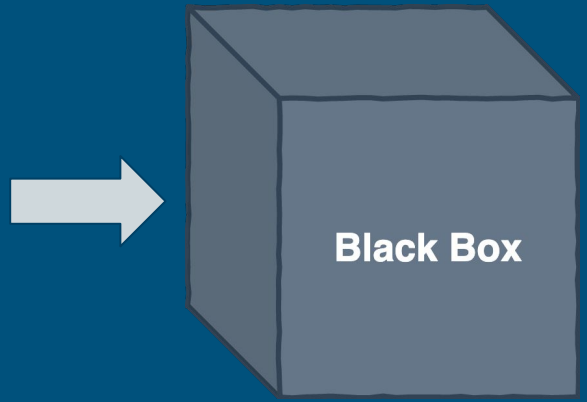
What occlusion colour and segmentation granularity to use?





Sokol and Flach, 2020.Towards Faithful and Meaningful Interpretable Representations.

Explanation Generation -- Predicting Sample



MALAMUTE

HUSKY

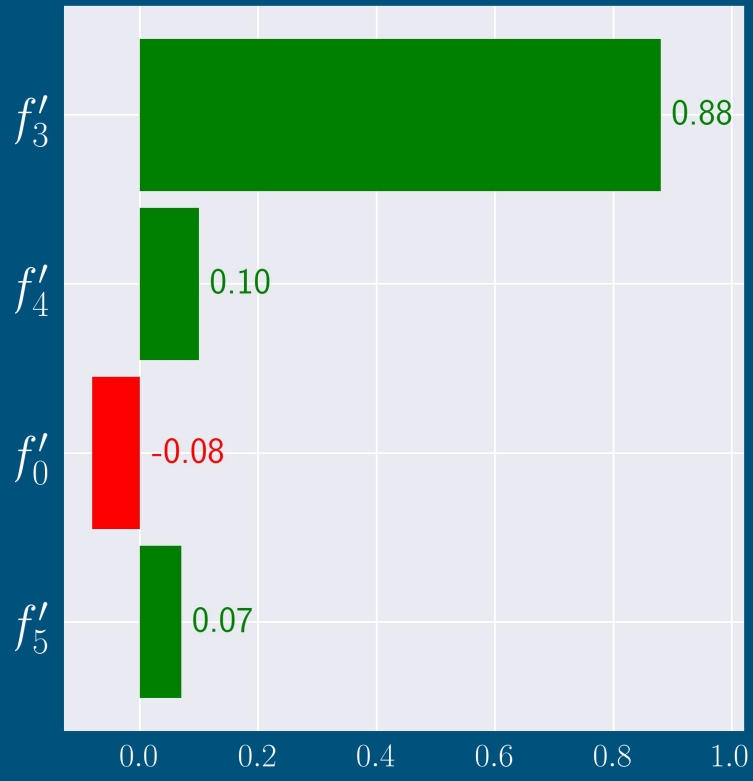
ESKIMO

Explanation Generation -- Weighting Sample

		Distance	Similarity
	$x'_0 = [0, 0, 0, 1, 1, 1, 1]$	d=3	s=0.57
$\mathring{x}' = [1, 1, 1, 1, 1, 1, 1]$	$x'_1 = [0, 1, 1, 1, 0, 1, 0]$	d=3	s=0.57
	$x'_2 = [1, 1, 1, 0, 1, 1, 1]$	d=1	s=0.09

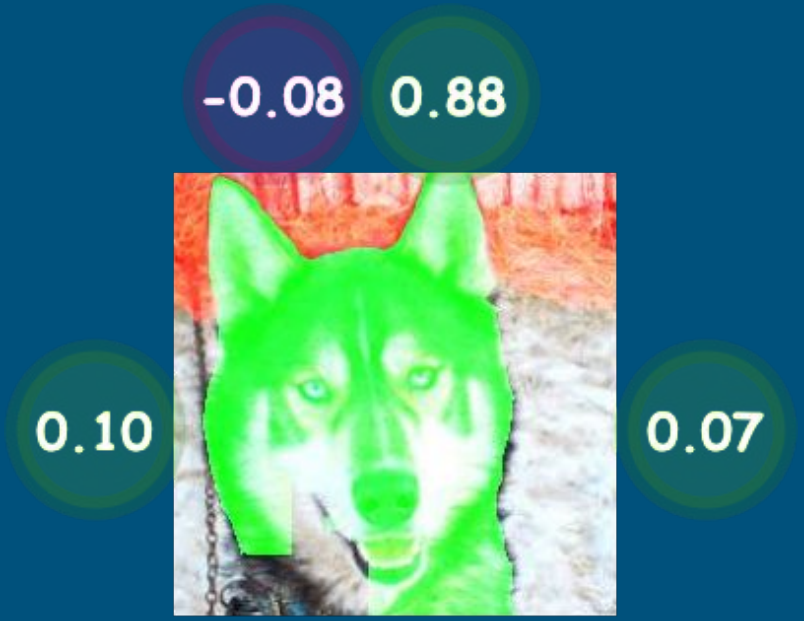
Explanation Generation -- Fitting Surrogate

- Fit a sparse linear classifier.
- Use the binary vectors as data.
- Use the black box-predictions for the sampled and reversed data as labels:
  - How for multi-class problems?
- Weight instances according to the similarity scores.
- Extract model coefficients as importance scores (explanation).

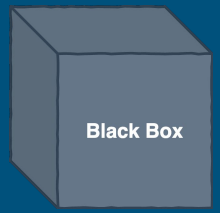


Explanation Generation -- Fitting Surrogate

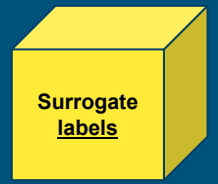
- Fit a sparse linear classifier.
- Use the binary vectors as data.
- Use the black box-predictions for the sampled and reversed data as labels:
  - How for multi-class problems?
- Weight instances according to the similarity scores.
- Extract model coefficients as importance scores (explanation).



Explanation Generation -- Fitting Surrogate



Black box is a multi-class classifier.
Surrogate is a (linear) binary classifier.



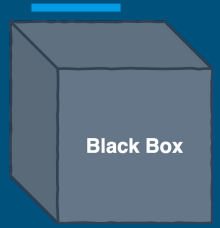
MALAMUTE
HUSKY
ESKIMO
ESKIMO
HUSKY

Explained Class:

HUSKY

0
1
0
0
1

Explanation Generation -- Fitting Surrogate



Black box is a probabilistic multi-class classifier.
Surrogate is a (linear) regressor.

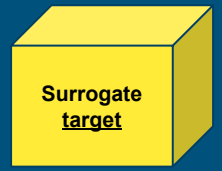
.97 MALAMUTE	.01 HUSKY	.02 ESKIMO	.01 HUSKY
.06 MALAMUTE	.88 HUSKY	.06 ESKIMO	.88 HUSKY
.13 MALAMUTE	.23 HUSKY	.64 ESKIMO	.23 HUSKY
.11 MALAMUTE	.10 HUSKY	.79 ESKIMO	.10 HUSKY
.42 MALAMUTE	.58 HUSKY	.00 ESKIMO	.58 HUSKY

Explained Class:

HUSKY

Explanation Generation -- Fitting Surrogate

Black box is a probabilistic multi-class classifier.
Surrogate is a multi-output regressor (tree).



.97 MALAMUTE	.01 HUSKY	.02 ESKIMO
.06 MALAMUTE	.88 HUSKY	.06 ESKIMO
.13 MALAMUTE	.23 HUSKY	.64 ESKIMO
.11 MALAMUTE	.10 HUSKY	.79 ESKIMO
.42 MALAMUTE	.58 HUSKY	.00 ESKIMO

Explained Classes:

HUSKY

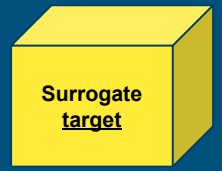
ESKIMO

.01 HUSKY	.02 ESKIMO
.88 HUSKY	.06 ESKIMO
.23 HUSKY	.64 ESKIMO
.10 HUSKY	.79 ESKIMO
.58 HUSKY	.00 ESKIMO

Sokol and Flach, 2020. LIMEtree: Interactively Customisable Explanations Based on Local Surrogate Multi-output Regression Trees.

Explanation Generation -- Fitting Surrogate

Black box is a probabilistic multi-class classifier.
Surrogate is a multi-output regressor (tree).



.97 MALAMUTE	.01 HUSKY	.02 ESKIMO		.01 HUSKY	.02 ESKIMO
.06 MALAMUTE	.88 HUSKY	.06 ESKIMO		.88 HUSKY	.06 ESKIMO
.13 MALAMUTE	.23 HUSKY	.64 ESKIMO		.23 HUSKY	.64 ESKIMO
.11 MALAMUTE	.10 HUSKY	.79 ESKIMO		.10 HUSKY	.79 ESKIMO
.42 MALAMUTE	.58 HUSKY	.00 ESKIMO		.58 HUSKY	.00 ESKIMO

Explained Classes:

HUSKY

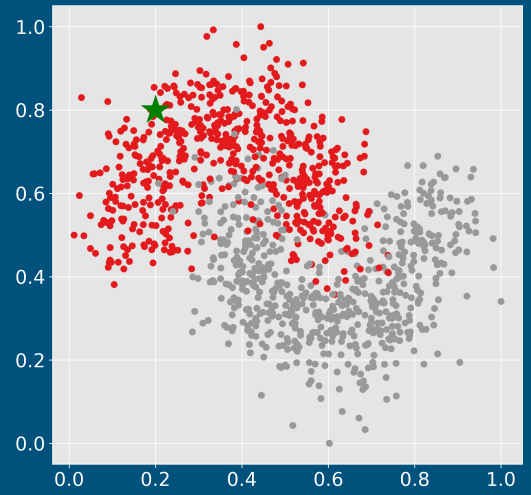
ESKIMO

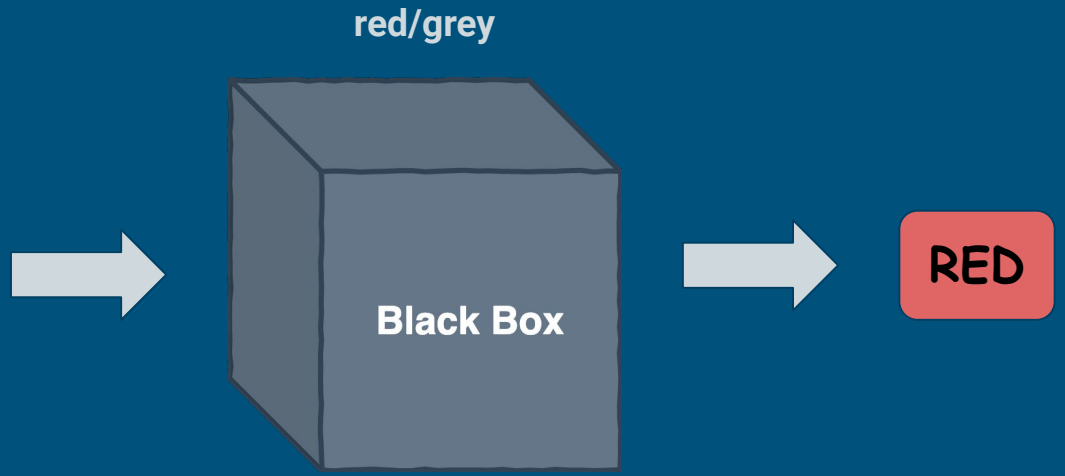
Sokol and Flach, 2020. LIMEtree: Interactively Customisable Explanations Based on Local Surrogate Multi-output Regression Trees.

Surrogates for Tabular Data

Black-box Prediction

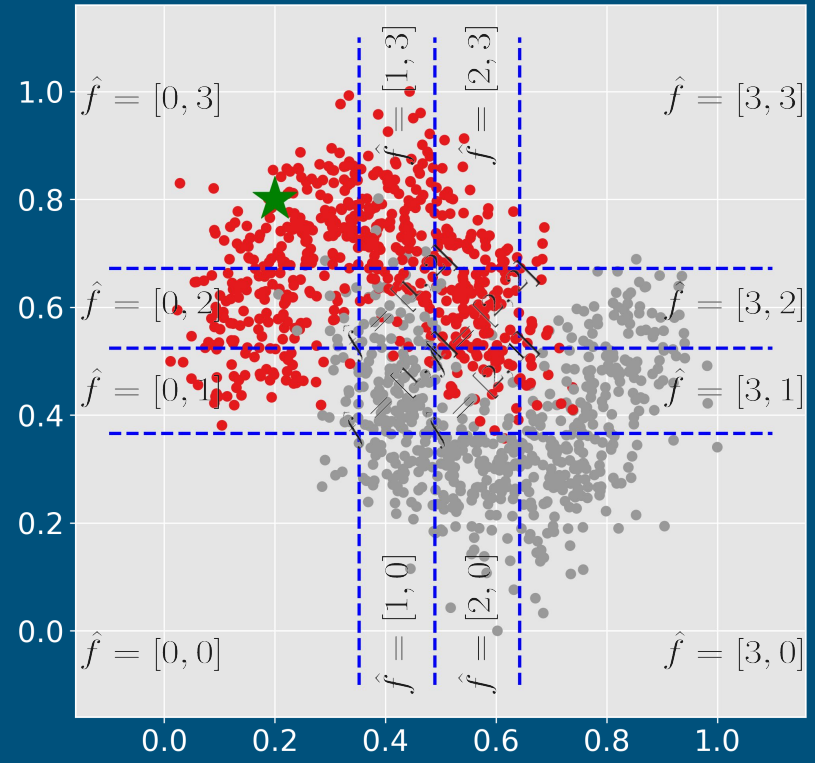
$$\mathring{x} = [0.2, 0.8]$$





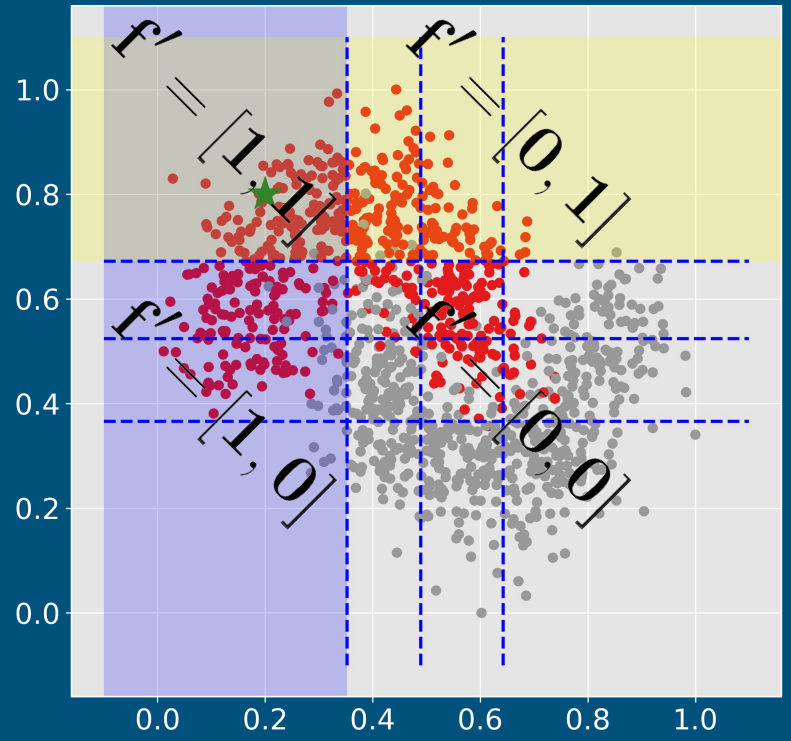
Interpretable Representation: (Q) Discretisation

$\hat{\mathring{x}} = [0, 3]$



Interpretable Representation: (Q) Binarisation

$$\mathring{x}' = [1, 1]$$



Interpretable Representation: (Q) *Complexity*

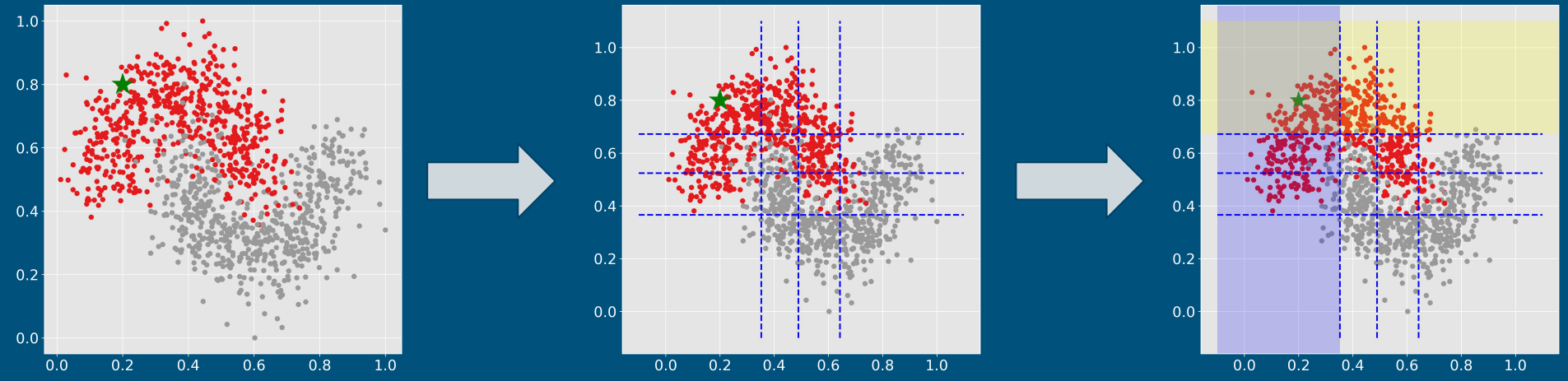
What happens in higher dimensions?

- **Discrete** → exponential growth of partitions -- $\mathbf{n^d}$, e.g., 4^2=16 for quartile discretisation in our toy example.
  - Original data with **d** numerical features.
  - Each numerical feature split into **n** sectors.
  - Categorical attributes not affected.
- **Binary** → reduces this number to $\mathbf{2^d}$, e.g., 2^2=4 for binarisation of our toy example.

Sparsity problem -- 11 numerical features yield 4^{11}=**4,194,304** hyper-rectangles and 2^{11}=**2,048** binary encodings.

Interpretable Representation: (★) Purity

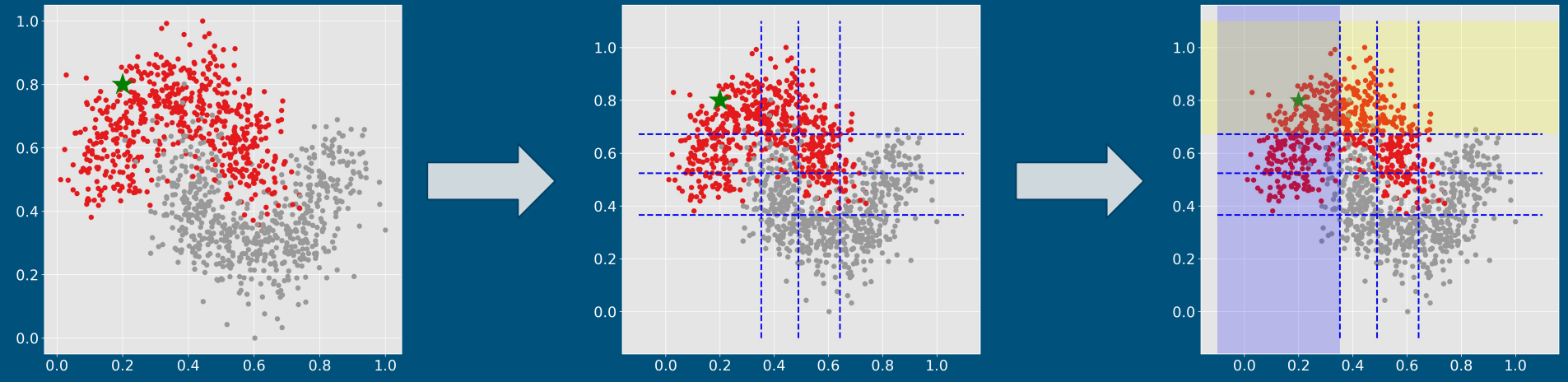
- Imagine this is a **sample around the explained data point**.
- The colour of markers encodes a **class predicted by a black box**.
- **Pure hyper-rectangles** imply good approximation of a black-box boundary.



Interpretable Representation: (★) Purity

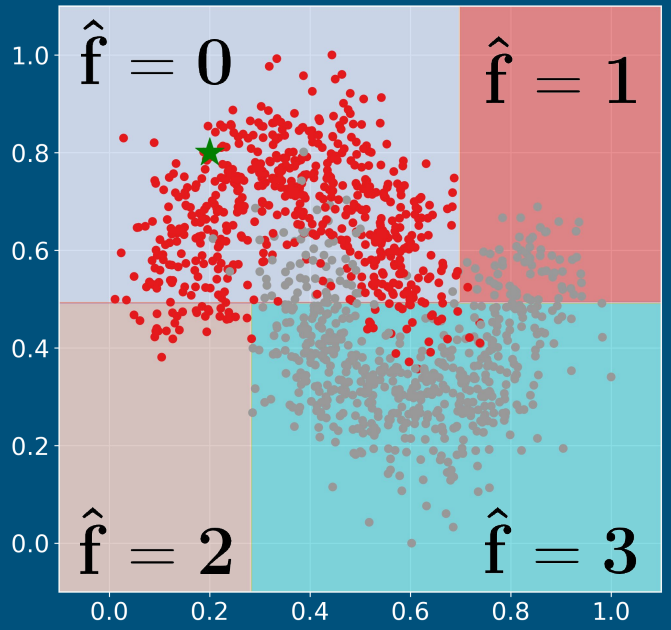
Purity score weighted by the number of instances in each partition:

- crisp classification → **Gini Index/Impurity**;
- probabilistic classification or regression → **Mean Squared Error**.



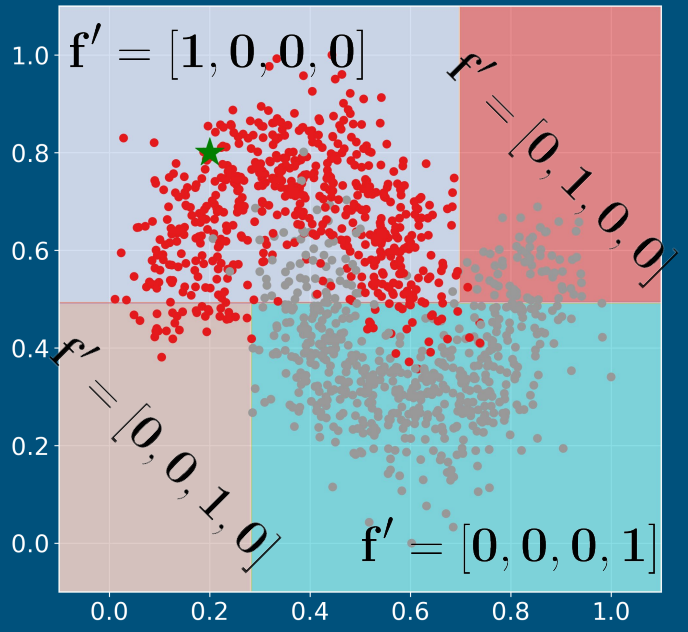
Interpretable Representation: (T) Discretisation

$$\hat{\mathring{x}} = 0$$



Interpretable Representation: (T) Binarisation

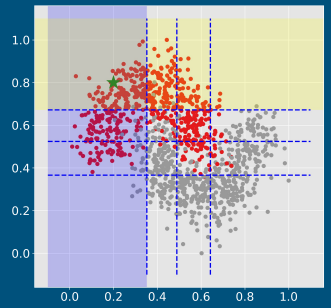
$$\mathring{x}' = [1, 0, 0, 0]$$



Data Sampling

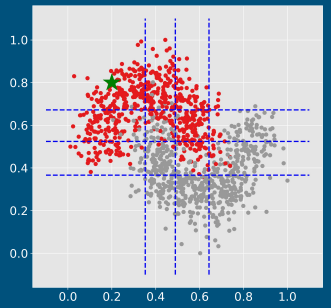
Sample from:

binary interpretable representation



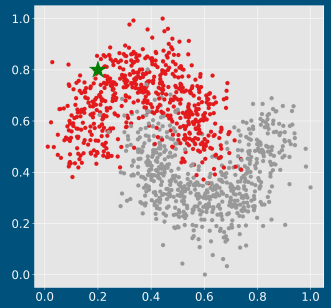

```
randint(low=0, high=1,
        shape=(3, 2))
```

discrete intermediate representation



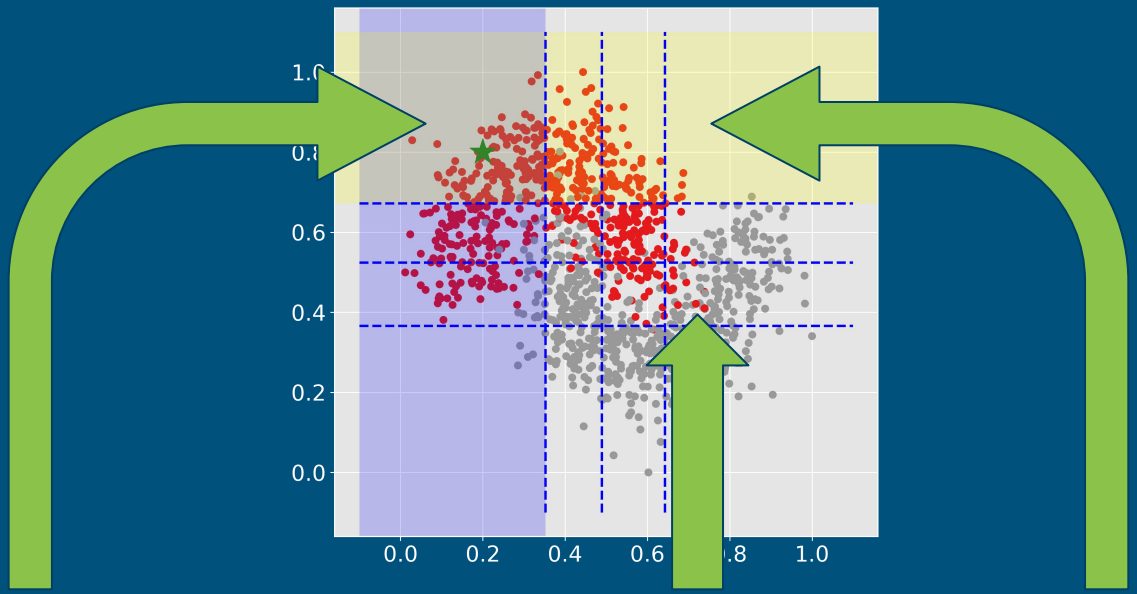

```
randint(low=0, high=3,
        shape=(3, 2))
```

original data domain



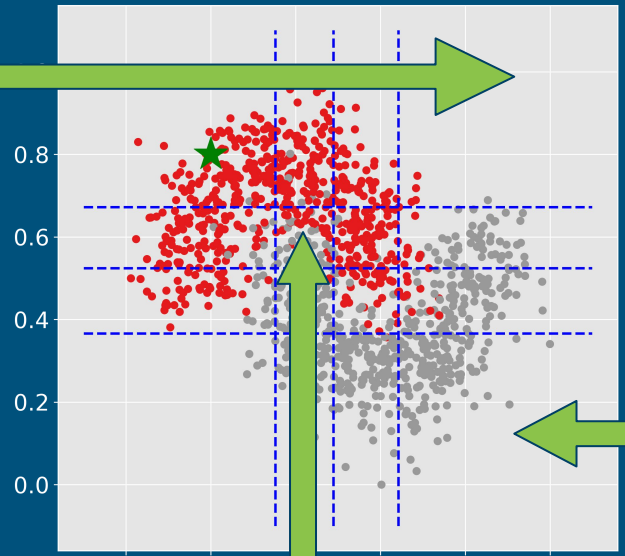
...

Data Sampling: Interpretable (Binary)



$$x'_0 = [1, 1] \qquad x'_1 = [0, 0] \qquad x'_2 = [0, 1]$$

Data Sampling: Intermediate (Discrete)

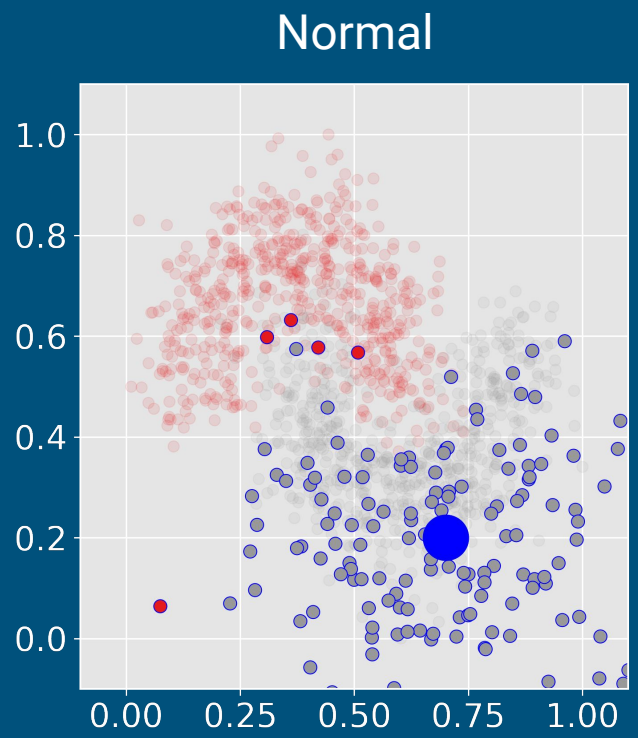


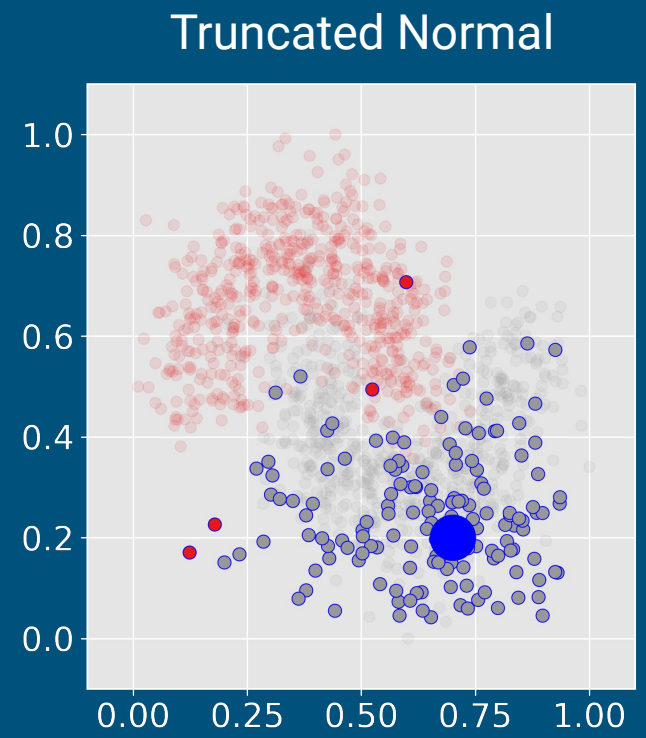
$\hat{x}_0 = [3, 3]$

$\hat{x}_1 = [1, 2]$

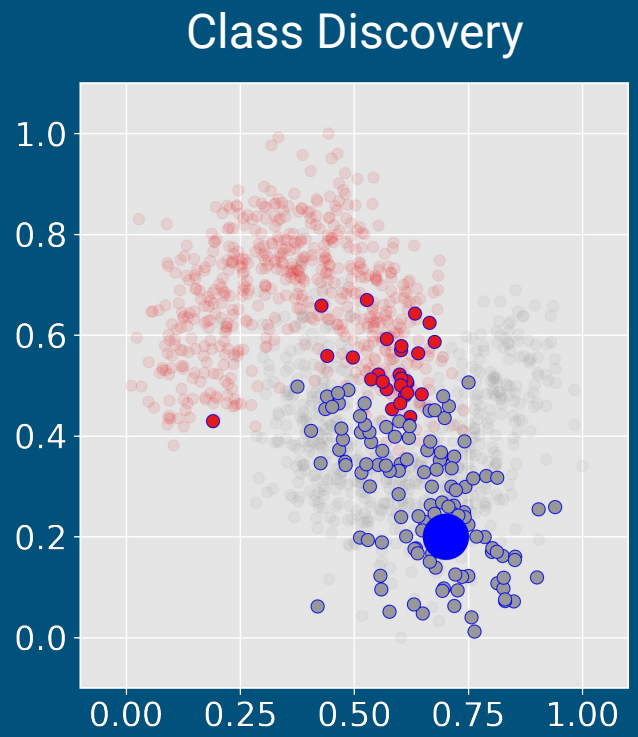
$\hat{x}_2 = [3, 0]$

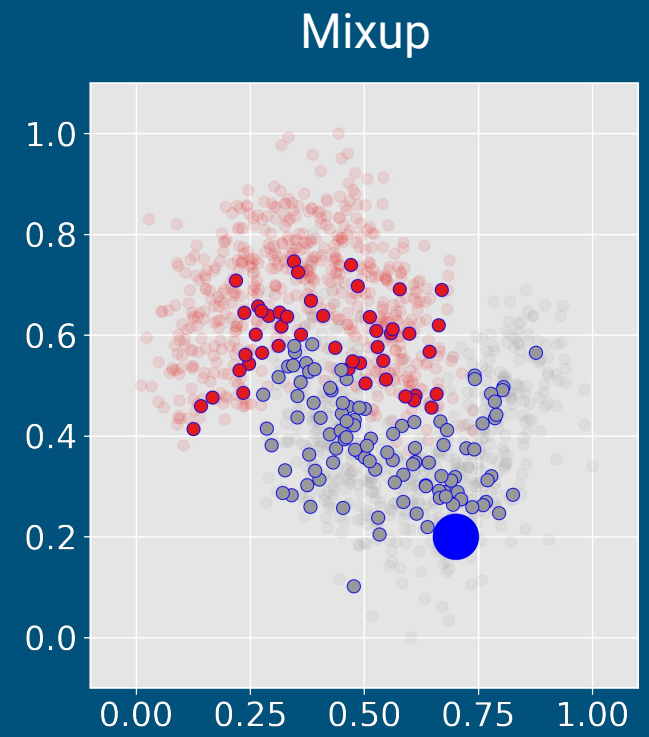
Data Sampling: Original





Data Sampling: Original (ctd.)



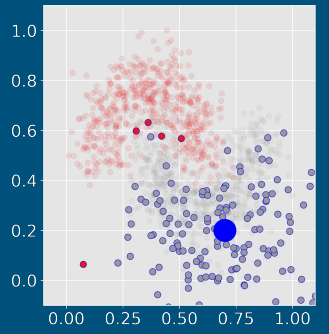


Zhang et al., 2017. mixup: Beyond empirical risk minimization.

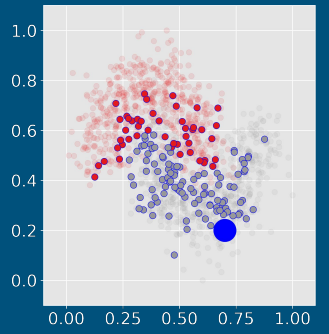
Data Sampling -- *Comparison*

Diverse sample → good chance of approximating black-box decision boundary.

(Whereas interpretable representation should be as pure as possible.)

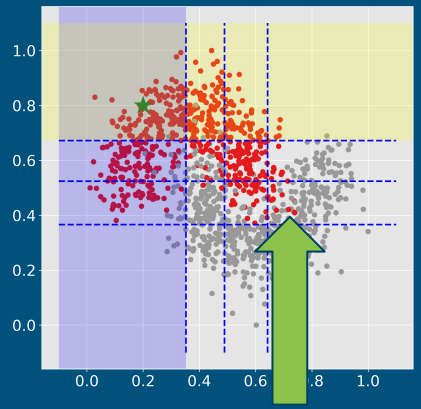


- *Binary/discrete sampling*, in principle, discover classes as they are somewhat global. (Caveat: large number of hyper-rectangles to sample from.)
- *Original data domain sampling* must explicitly discover the neighbourhood.

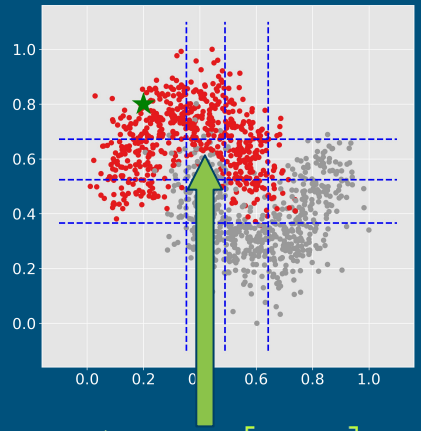


Explanation Generation -- Restoring Sample

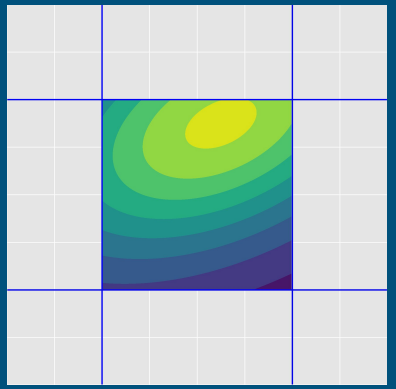
Only required for binary and discrete sampling.



$x'_1 = [0, 0]$

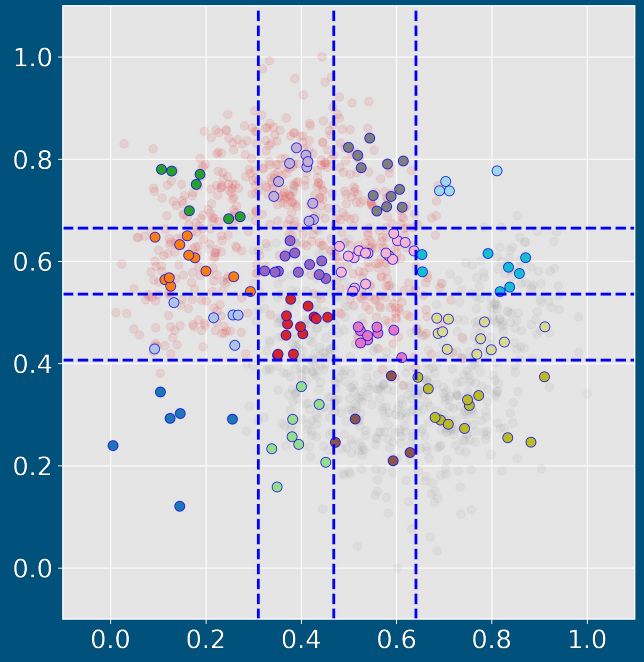


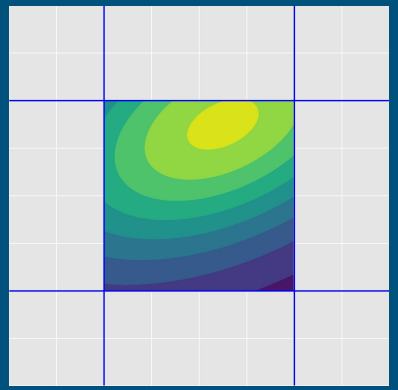
$\hat{x}_1 = [1, 2]$



Explanation Generation -- Restoring Sample

Only required for binary and discrete sampling.



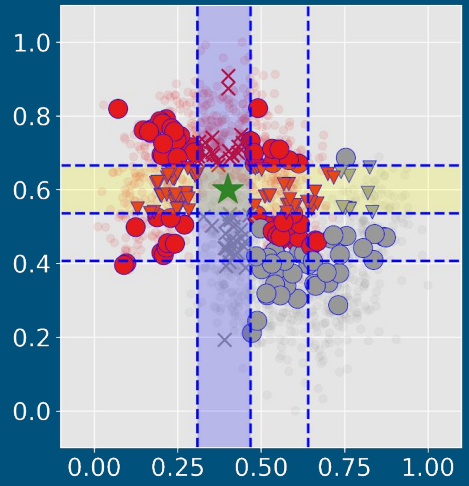


Explanation Generation -- Weighting Sample

But in which domain?

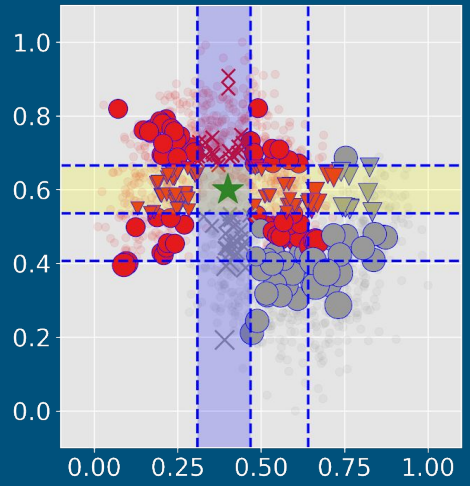
Binary

$$\mathring{x}' = [1, 1]$$



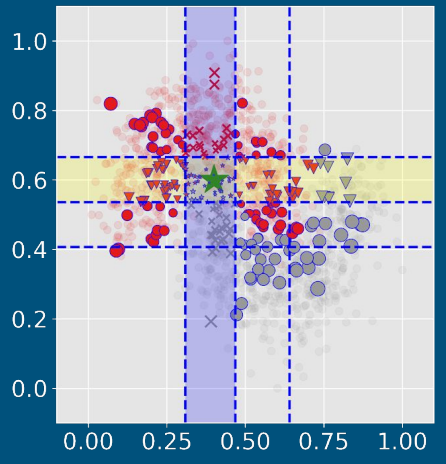
Discrete

$$\hat{\mathring{x}} = [1, 2]$$



Original

$$\mathring{x} = [0.2, 0.8]$$

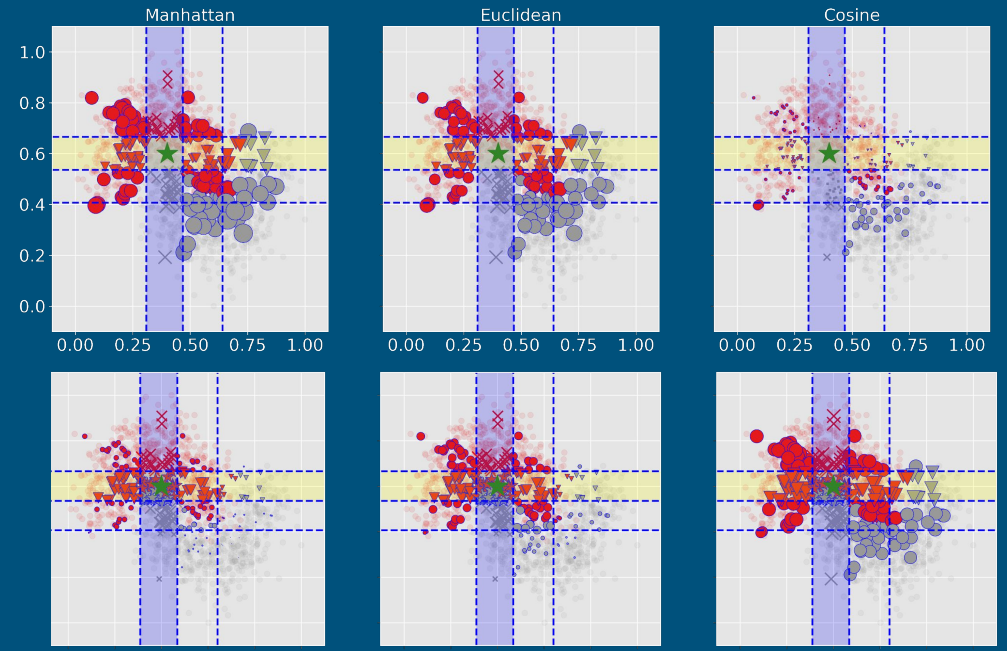


Explanation Generation -- Weighting Sample

Kernelisation... But in which domain?

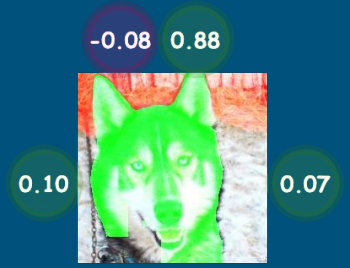
Discrete Distances

Kernelisation

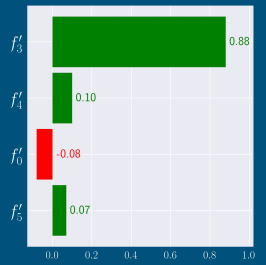


Explanation Generation -- Feature Selection

- Not required for image and text data -- the explanation presentation is as intelligible as the explained instance.

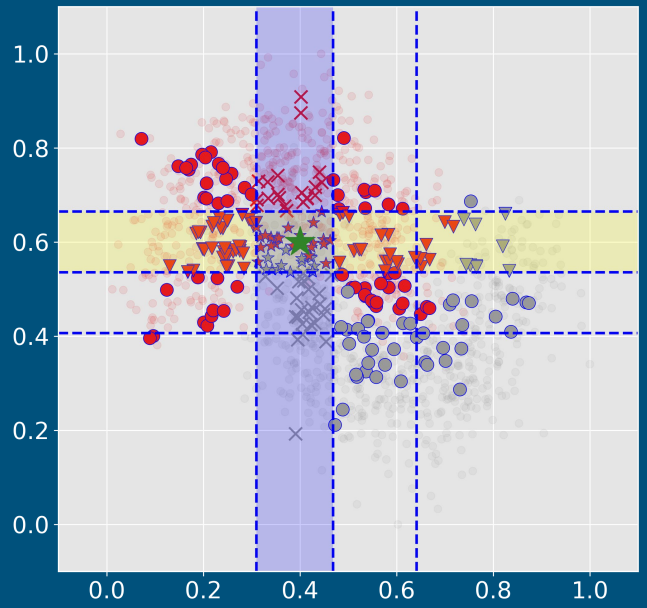


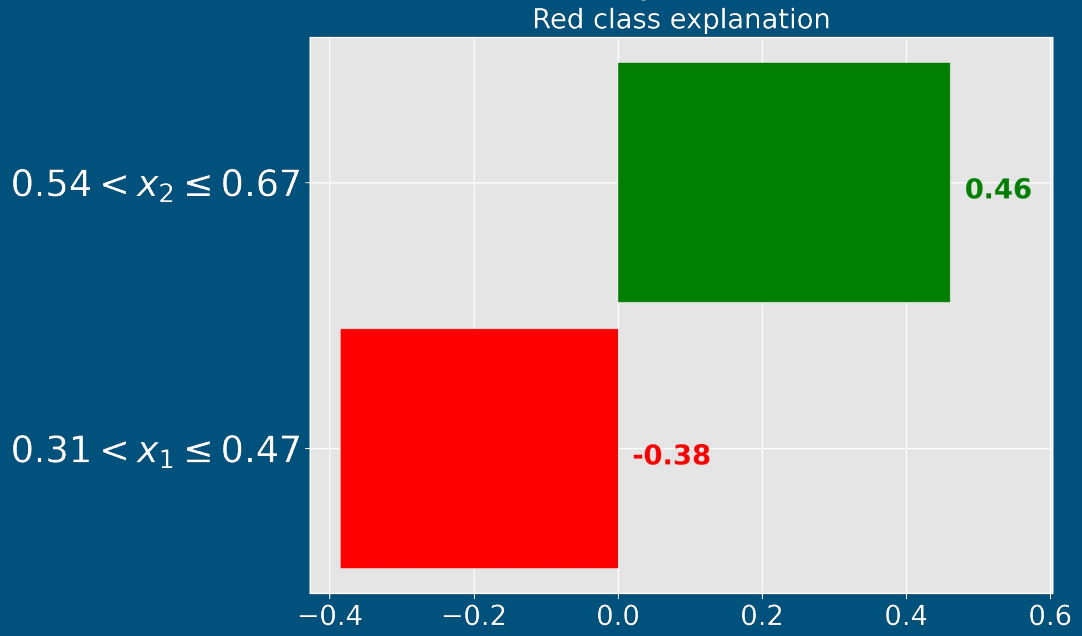
- Can be *computed* for any representation, but *applies* to the binary interpretable -- introducing sparsity to the bar plot.
  - lasso path;
  - forward selection; and
  - highest weights.
- Interpretable representations based on **tree splits** have sparsity mechanisms built in (e.g., impurity reduction and pruning).



Linear Surrogate + Discretisation/Binarisation

Caveats: feature independence and linearity.

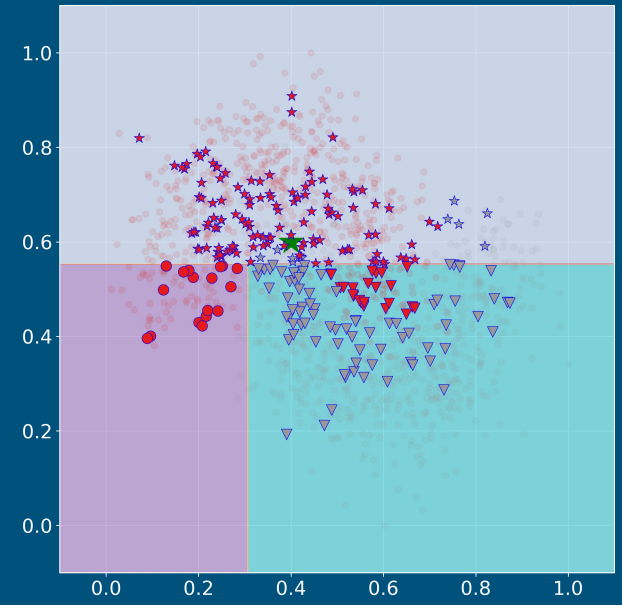


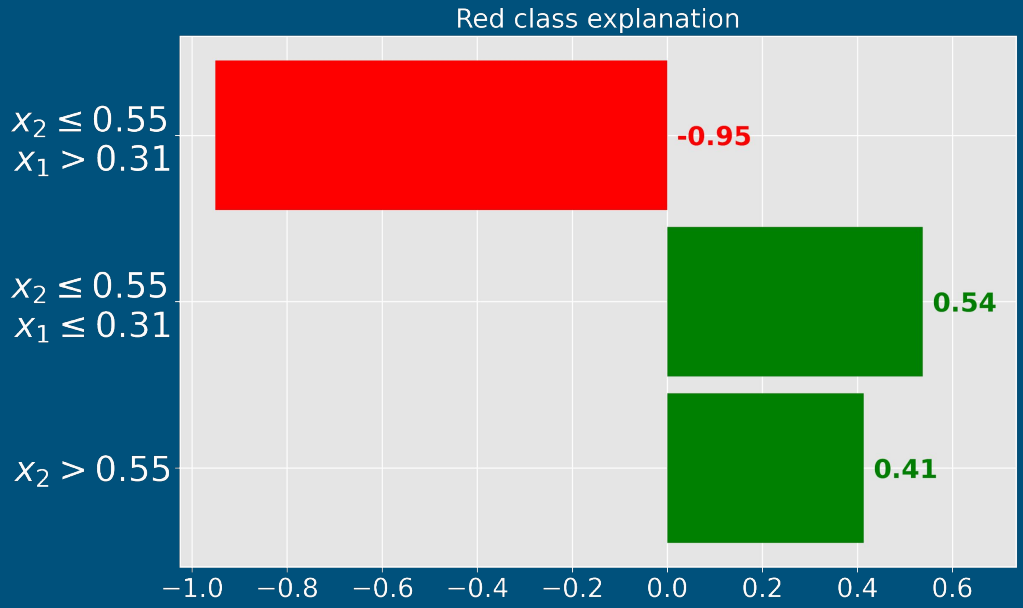


Sokol and Flach, 2020.Towards Faithful and Meaningful Interpretable Representations.

Linear Surrogate with One-hot Tree Features

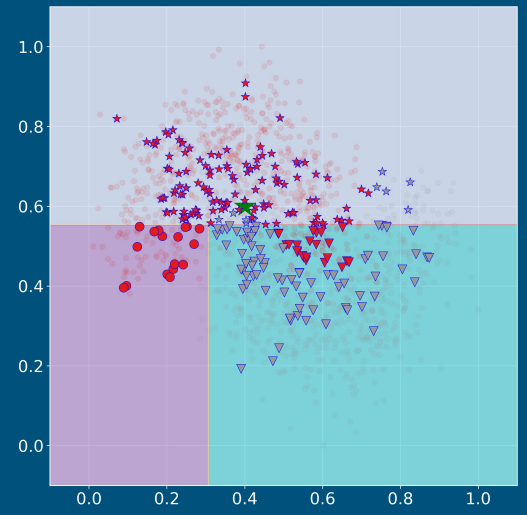
Caveats: feature independence, linearity, possible rule overlap.

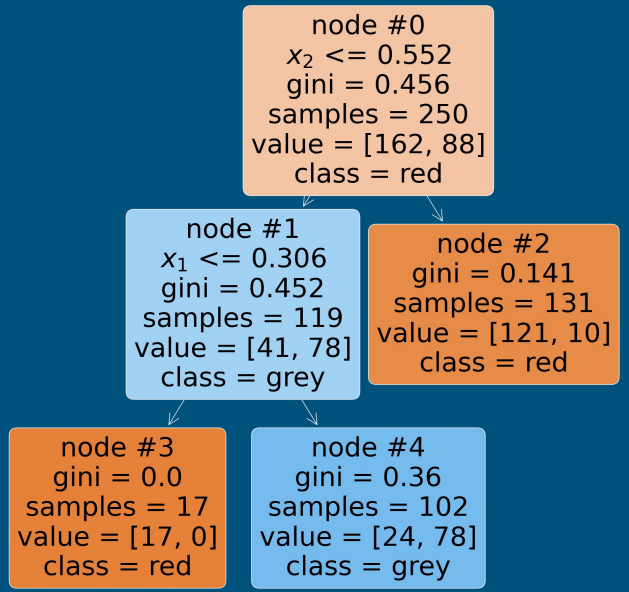


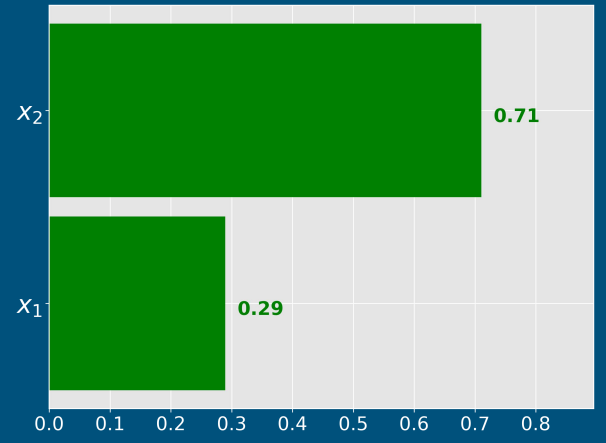


End-to-end Tree-based Surrogate

Caveats: axis-parallel splits.







You Need to Choose Wisely

There is no magic formula -- it takes work.

Next Up

Open Source Interpretability Tools

(Alex Hepburn)