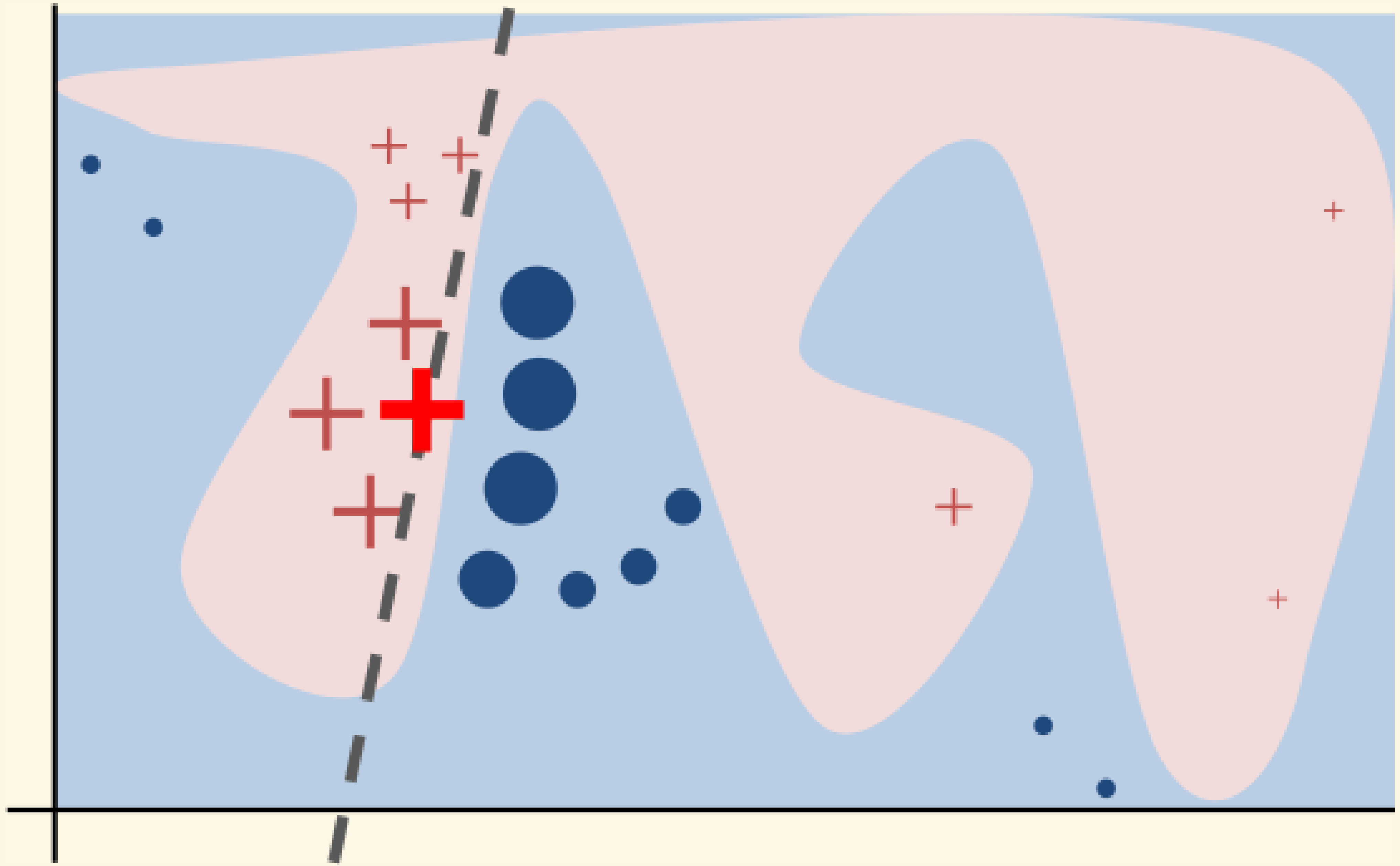
Case Study: Surrogates

The Universal Explainers

Kacper Sokol

LIME: Local Interpretable Model-agnostic Explanations



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

Benefits

- **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

No Free Lunch

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

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 are inherently interpretable. This Perspective clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare and computer vision.

No Free Lunch

Perspective | Published: 13 May 2019

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

Cynthia Rudin

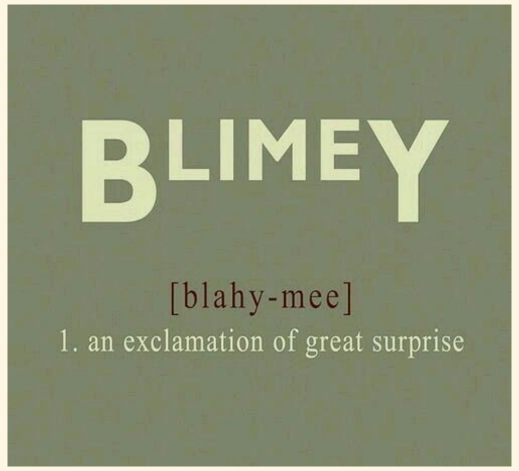
Nature Machine Intelligence **1**, 206–215 (2019) | Cite this article

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 are inherently interpretable. This Perspective clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare and computer vision.

- Post-hoc explainers have poor fidelity
- A **generic** eXplainable Artificial Intelligence process is *beyond our reach* at the moment

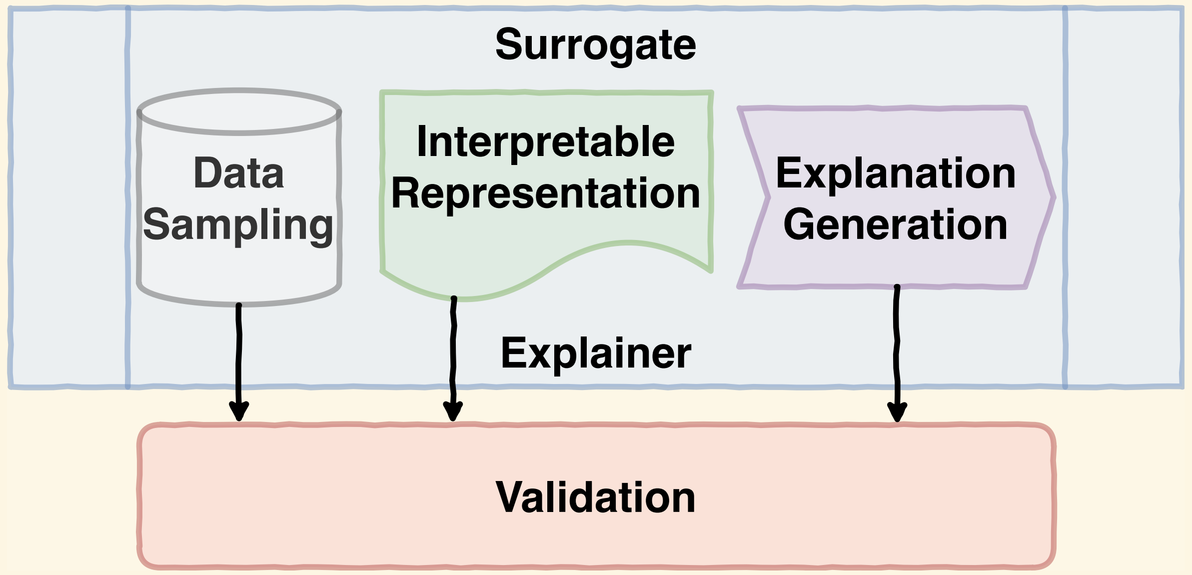
bLIMEy, there has to be a better way...



bLIMEy → build LIME yourself
(*Sokol et al., 2019. bLIMEy: Surrogate prediction explanations beyond LIME*)

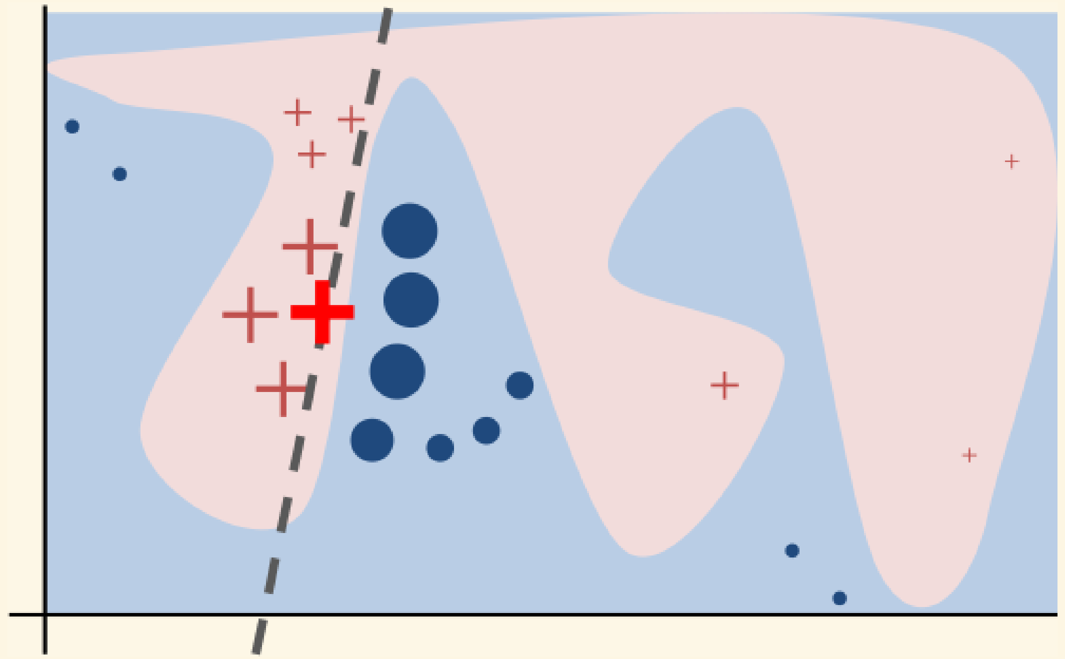
- Framework for building surrogate explainers
- Meta-algorithm for operationalising them
- Accompanied by analysis of surrogate building blocks (akin to a user guide)
- Practical recommendations

Good news: A means to build flexible, faithful, interactive, ... surrogates
Not so good news: It requires effort



Operationalising surrogates

- To **use** surrogates, we need to understand
  - their provenance
  - how to (correctly) interpret their explanations
- To **build** surrogates, we should
  - choose suitable building blocks
  - evaluate & validate these



Surrogate Image Explainers

Image surrogates (LIME)

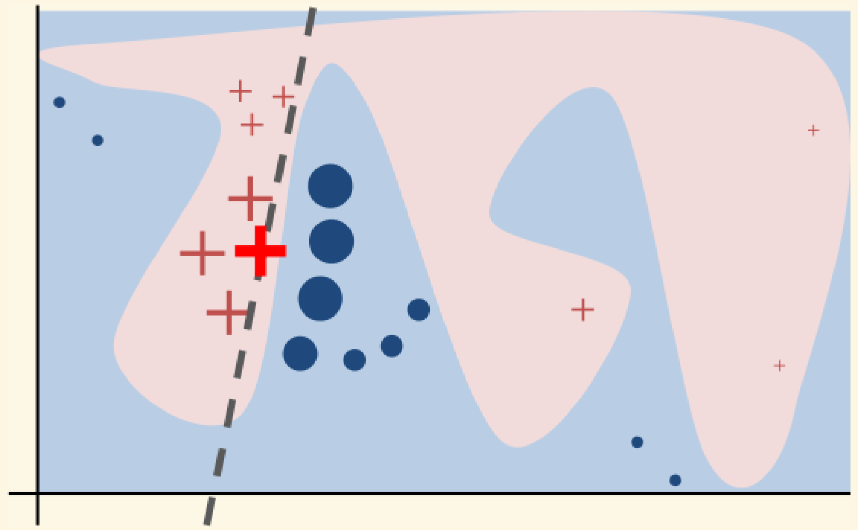
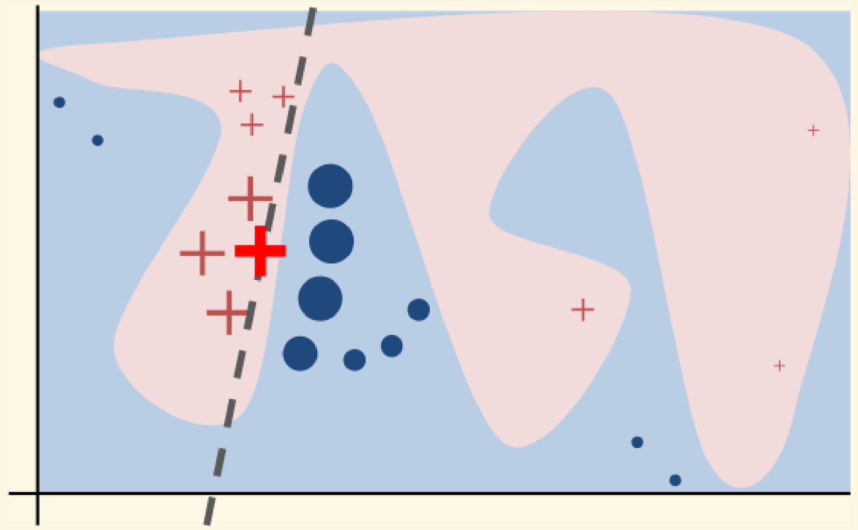
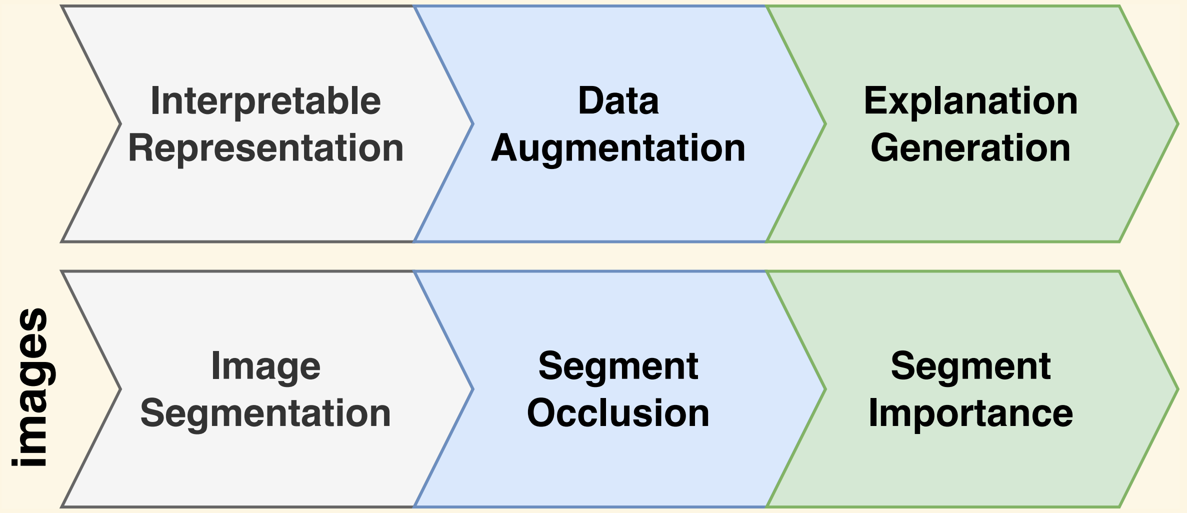
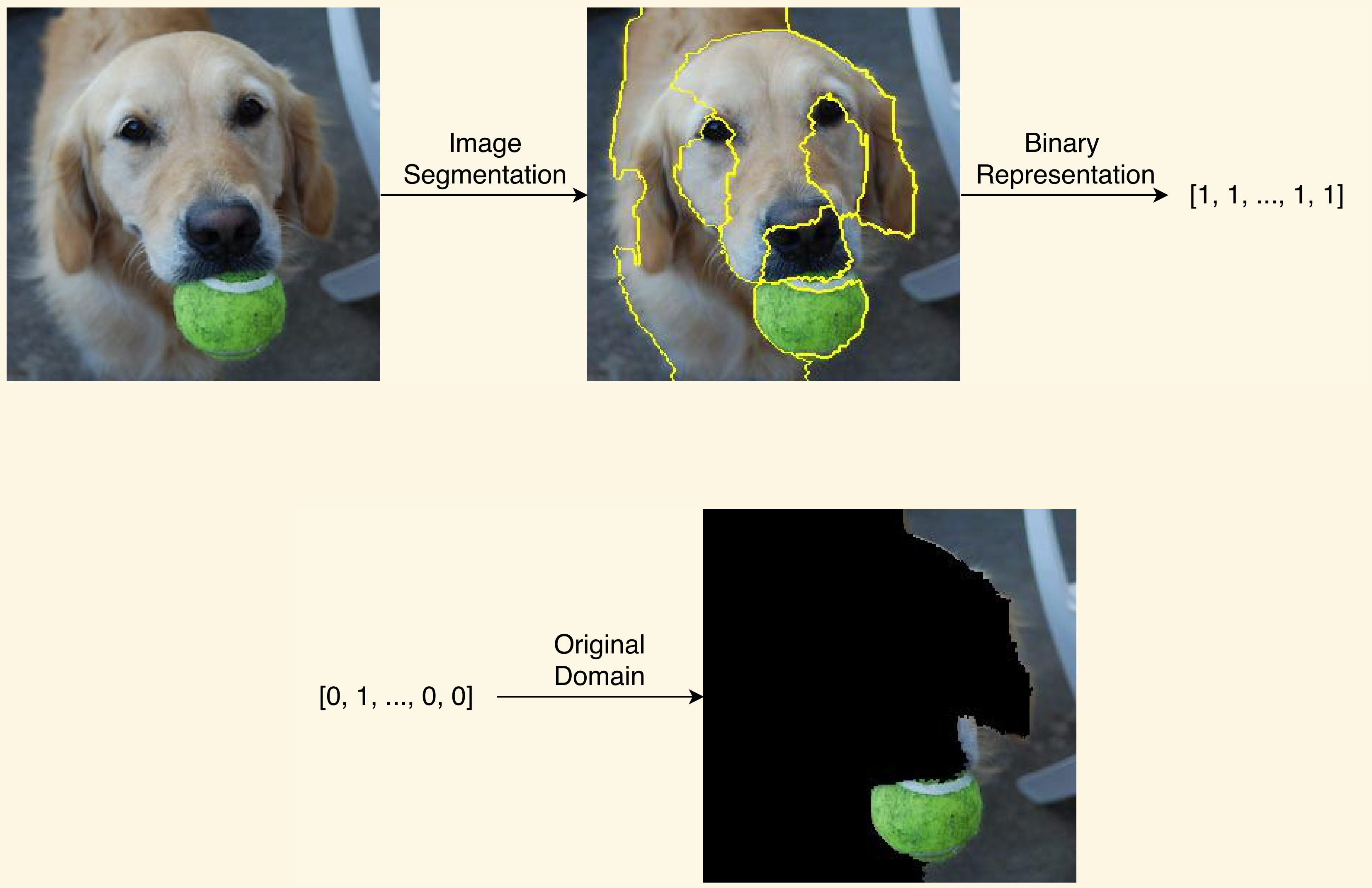


Image surrogates (LIME)





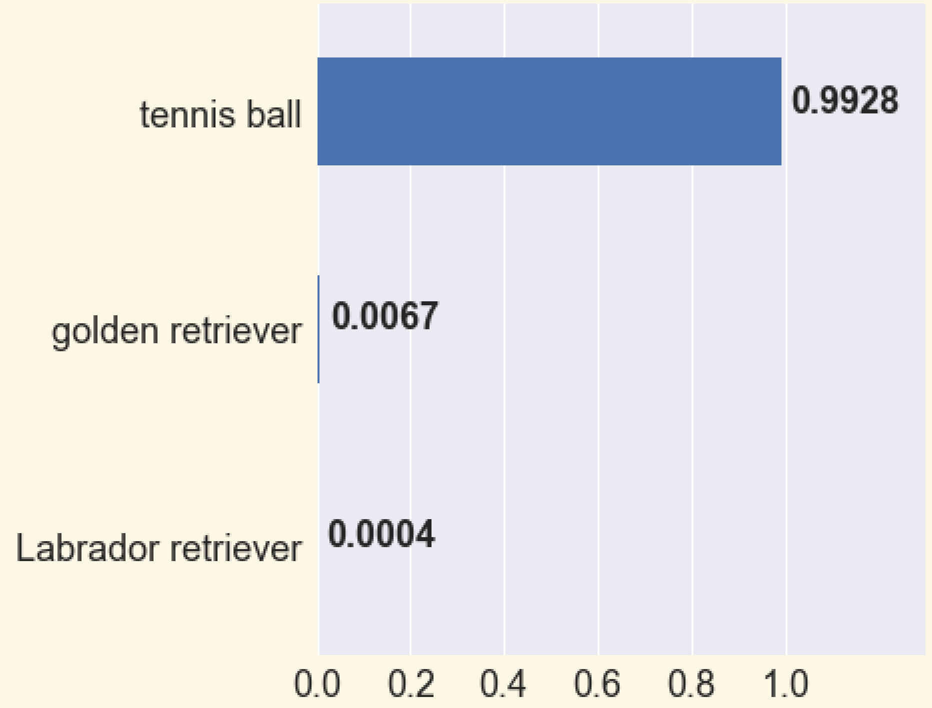
Segmentation-based interpretable representation

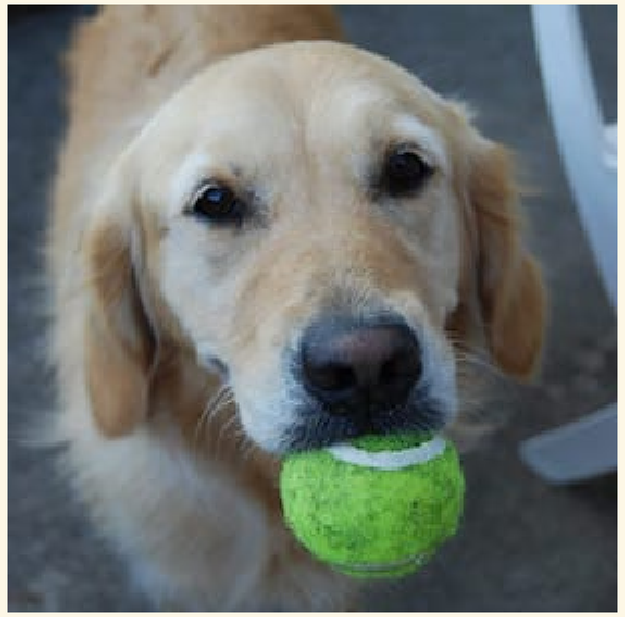


Black-box prediction

In [6]:
```
classification
```

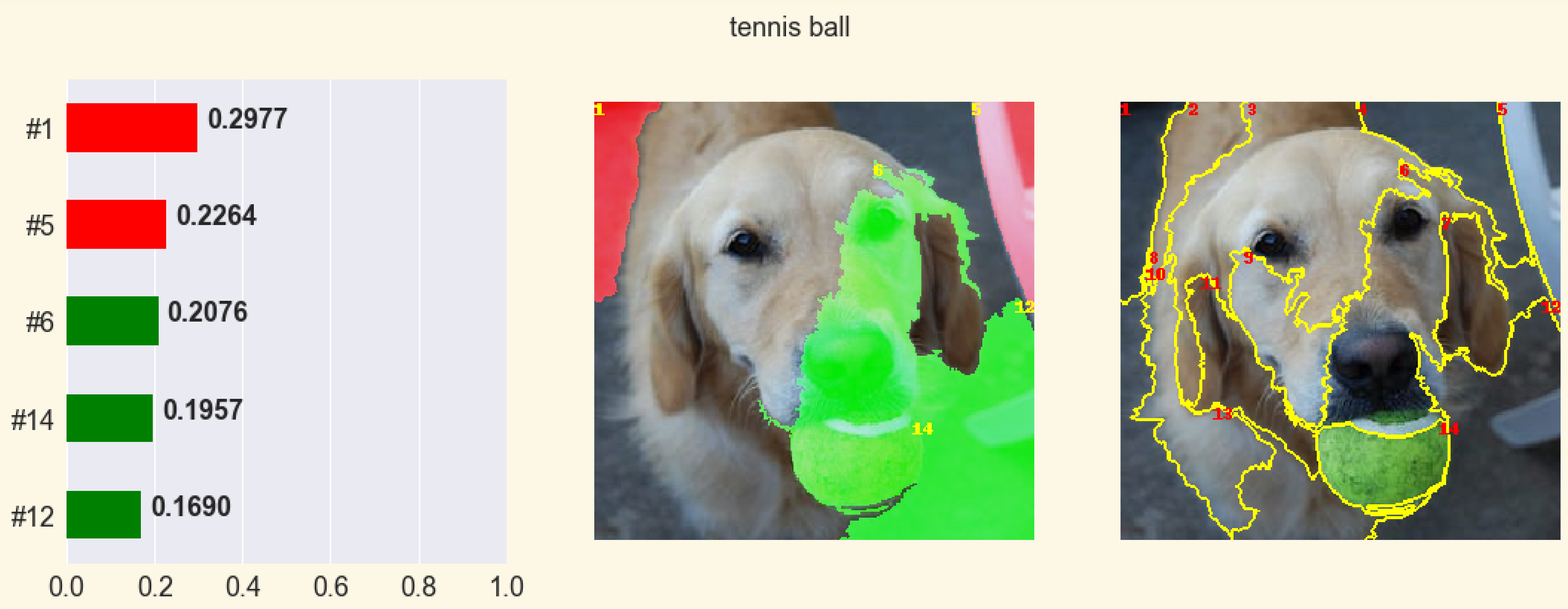
Out[6]:





Prediction explanation

In [8]:
```
exo.plot_image_explanation(blimey_image, explain_classes[0])
```



Prediction explanation

In [9]:
```
exo.plot_image_explanation(blimey_image, explain_classes[1])
```

Explainer demo

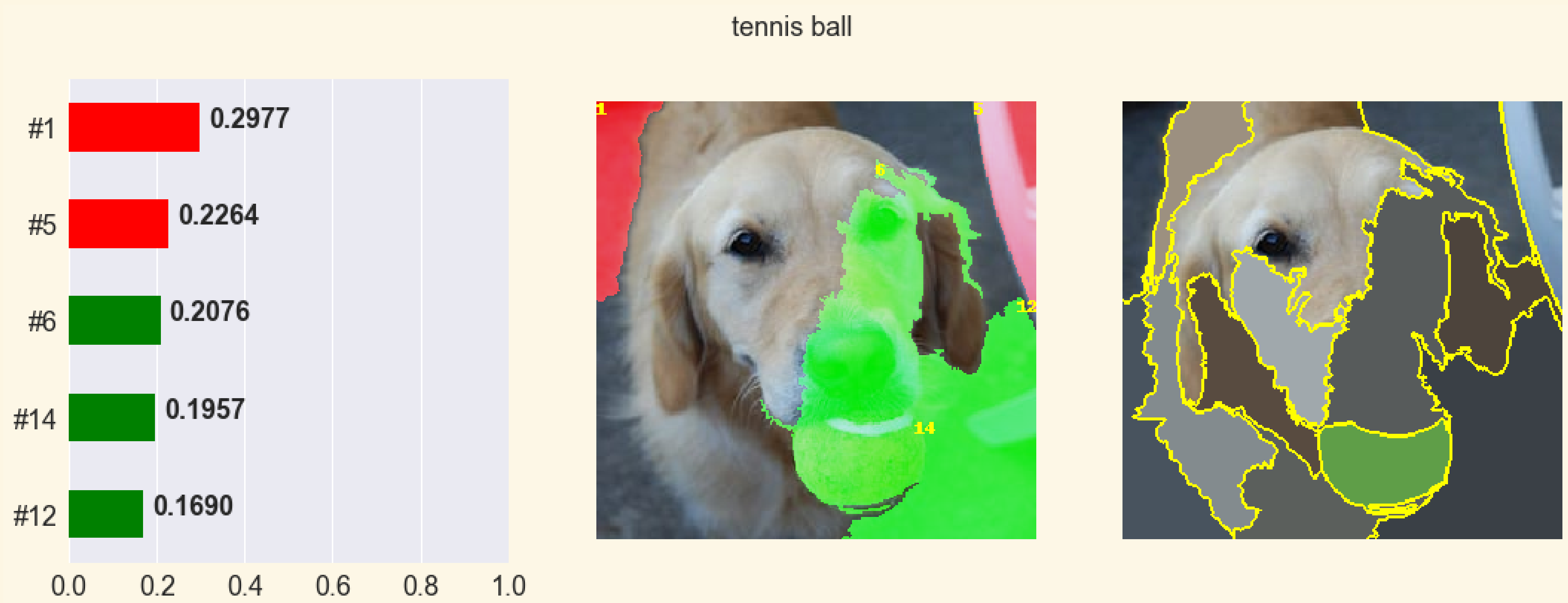
In [14]:
```
surrogate_image_explainer
```

Segmentati... low

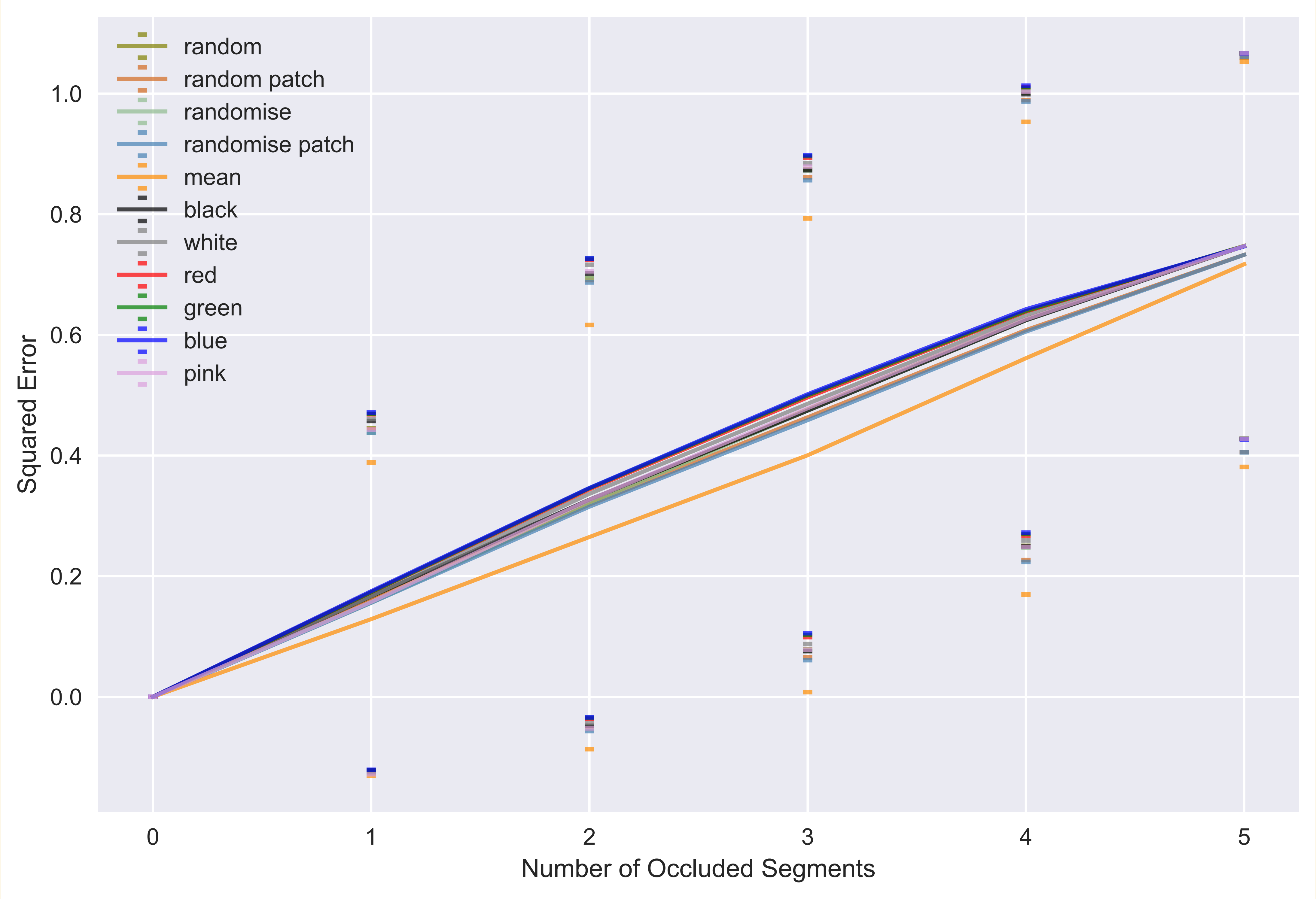
Occlusion c... mean | black | white | randomise-patch | green

Explained cl... tennis ball | golden retriever | Labrador retriever

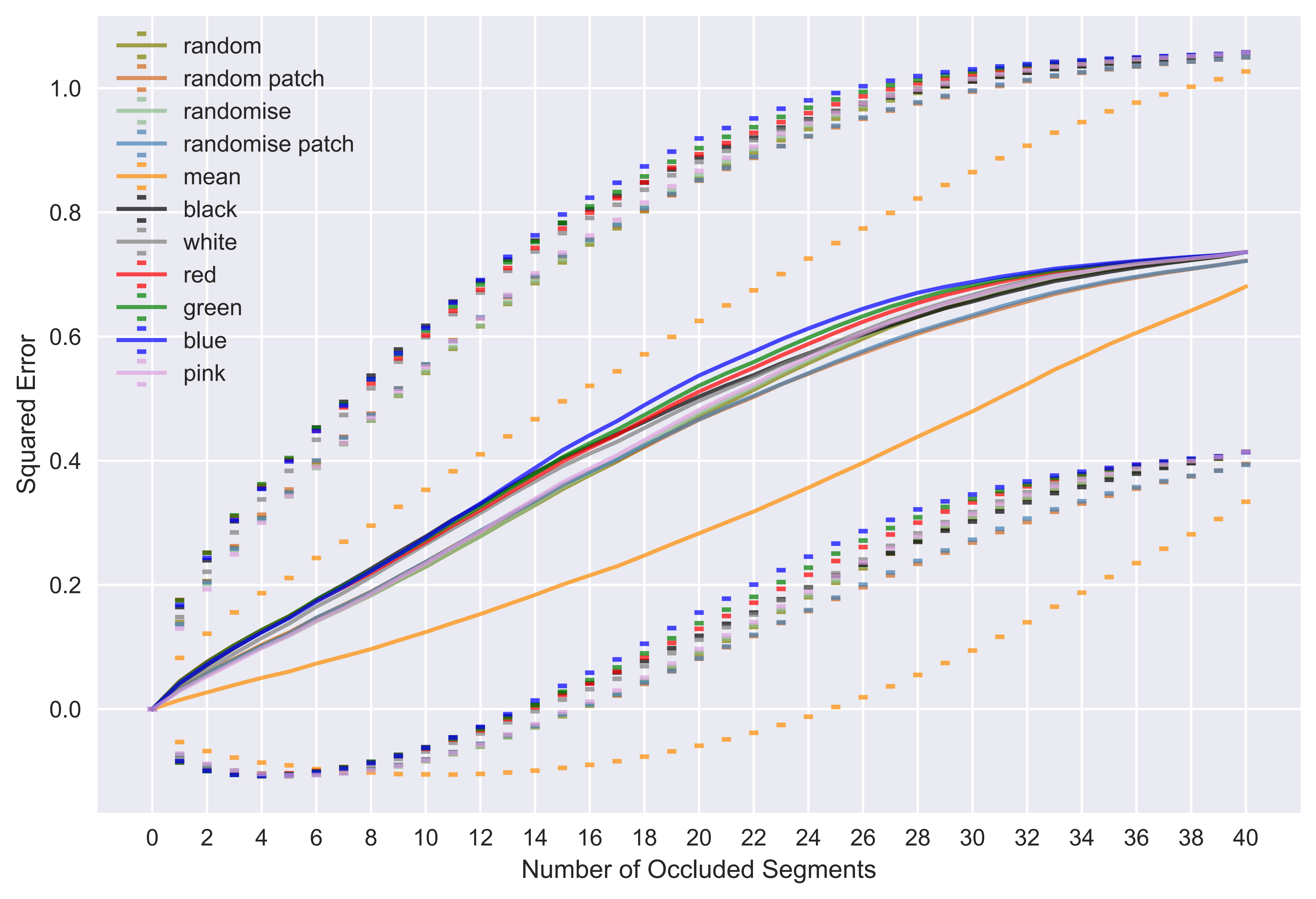
Explain!



Segmentation granularity and occlusion colour – 5 segments

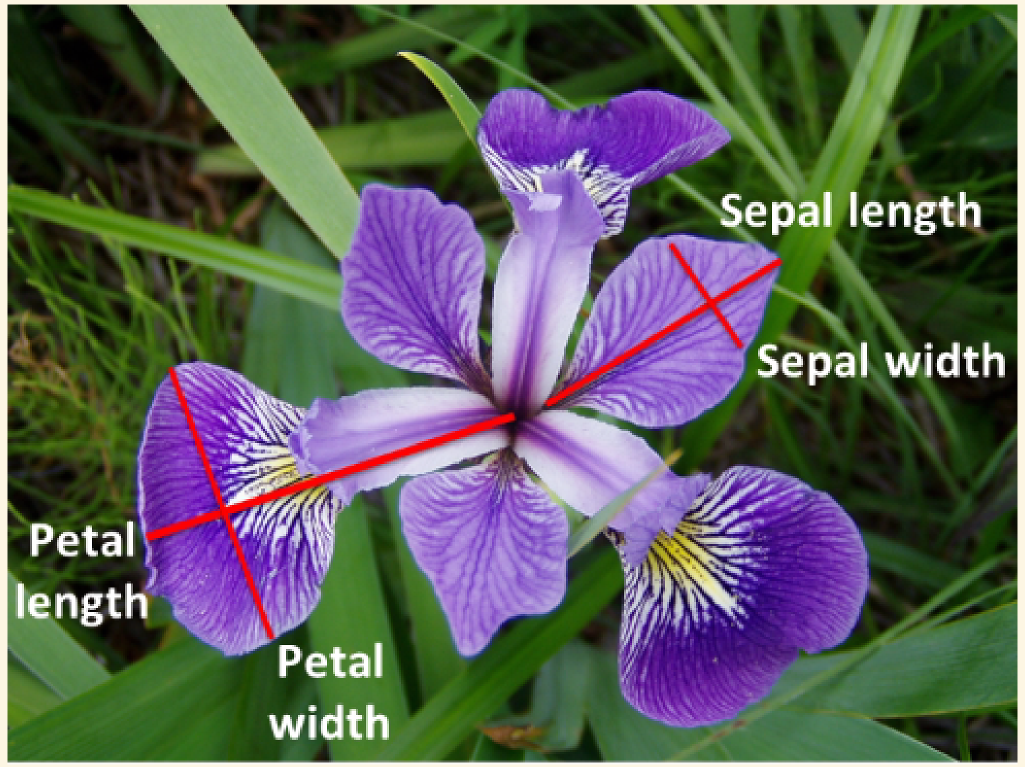


Segmentation granularity and occlusion colour – 40 segments

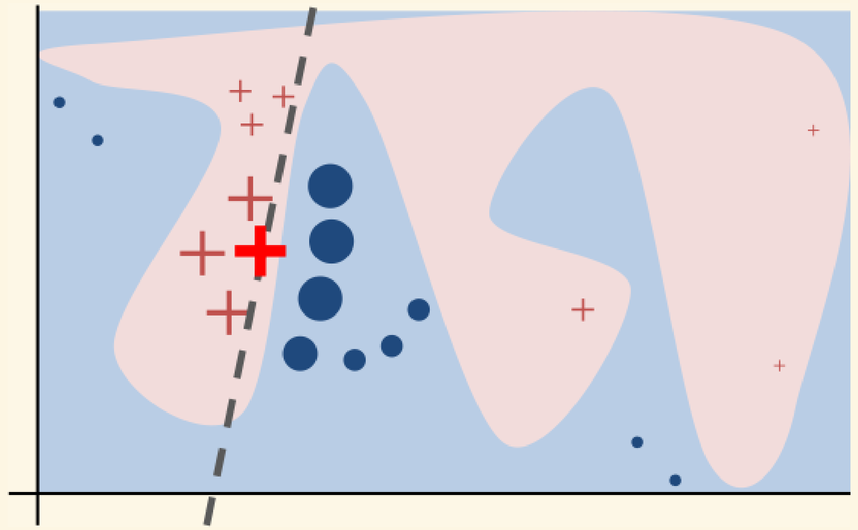


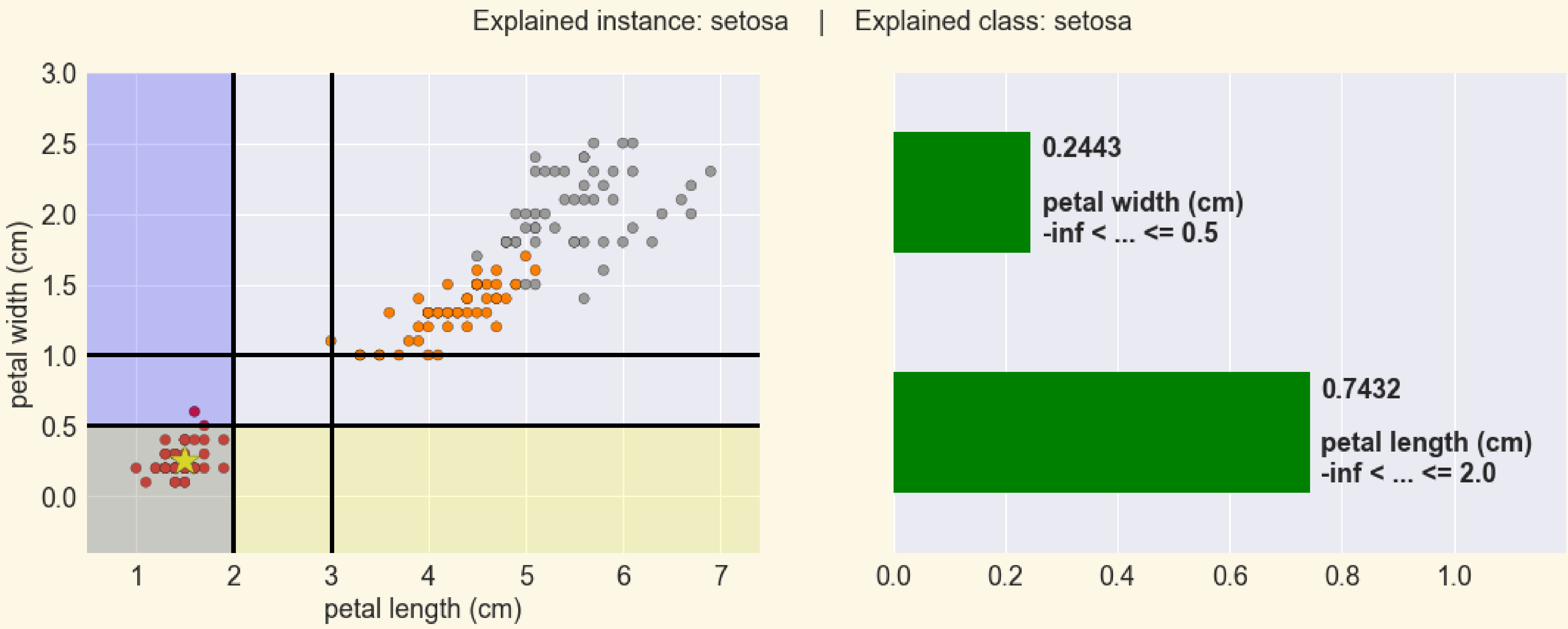
Surrogate Explainers of Tabular Data

Classifying iris flowers

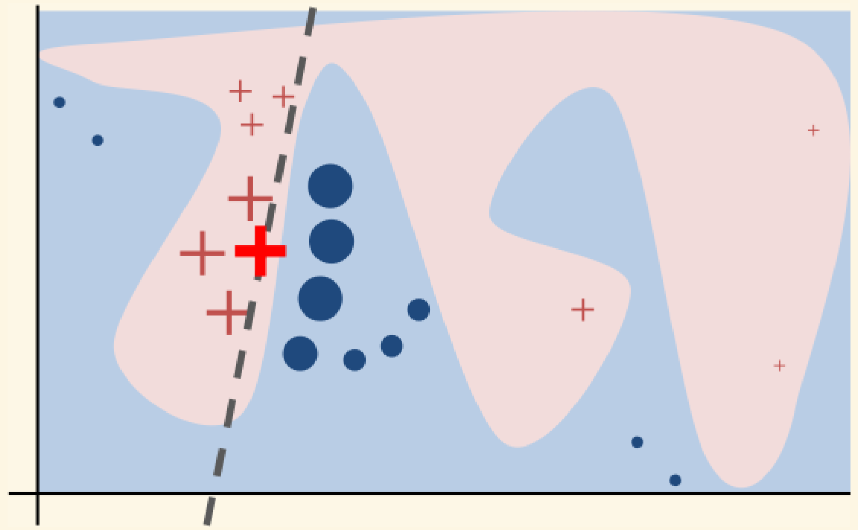


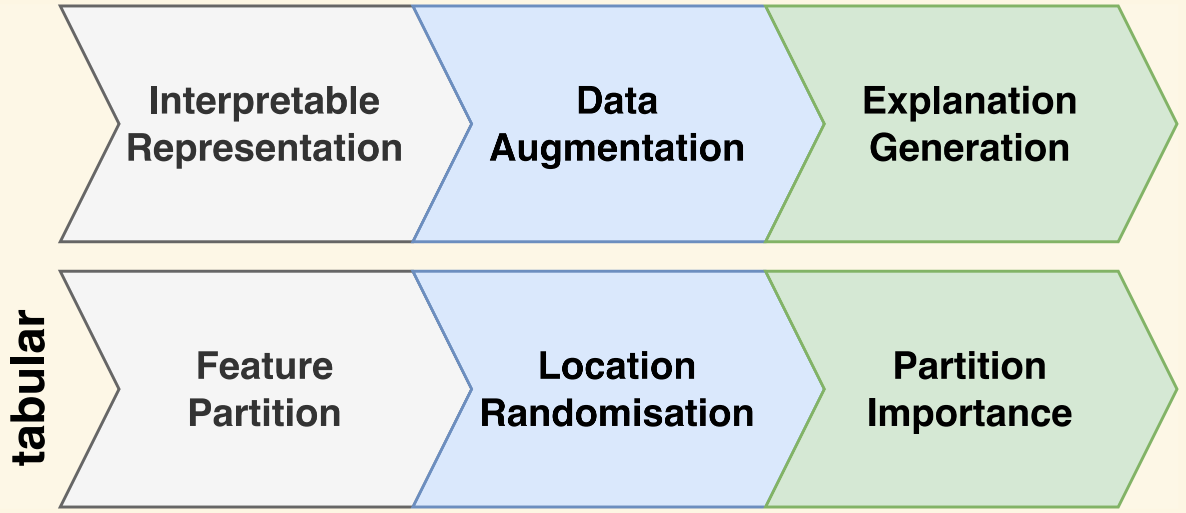
Tabular surrogates (LIME)

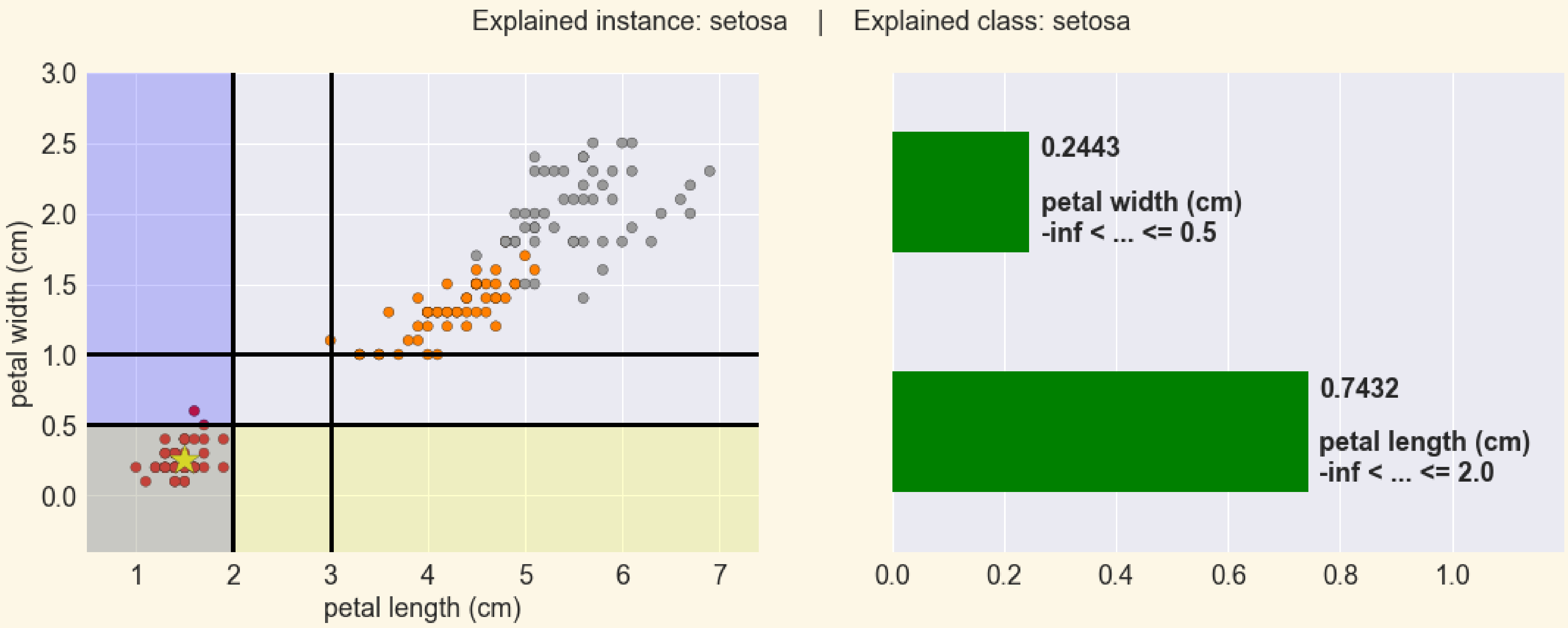




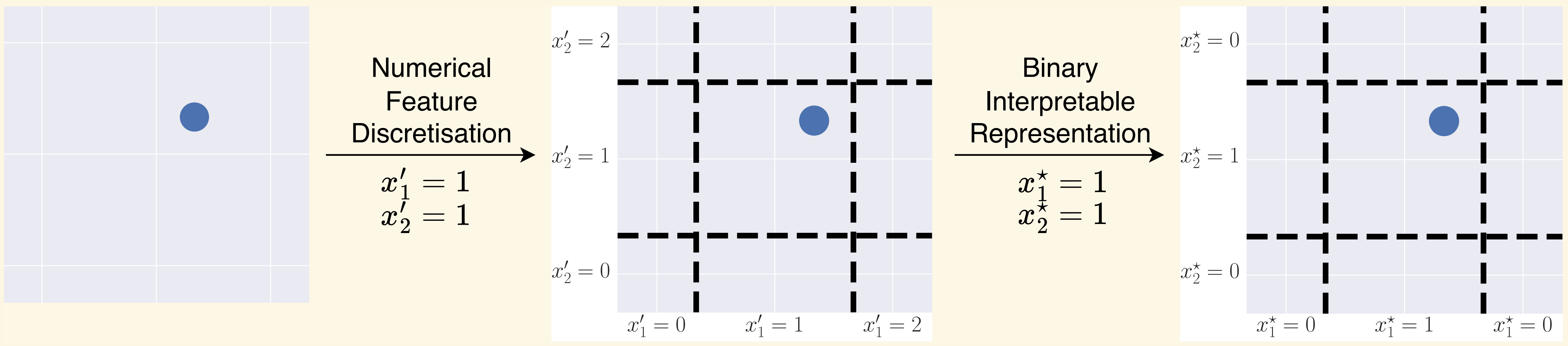
Tabular surrogates (LIME)

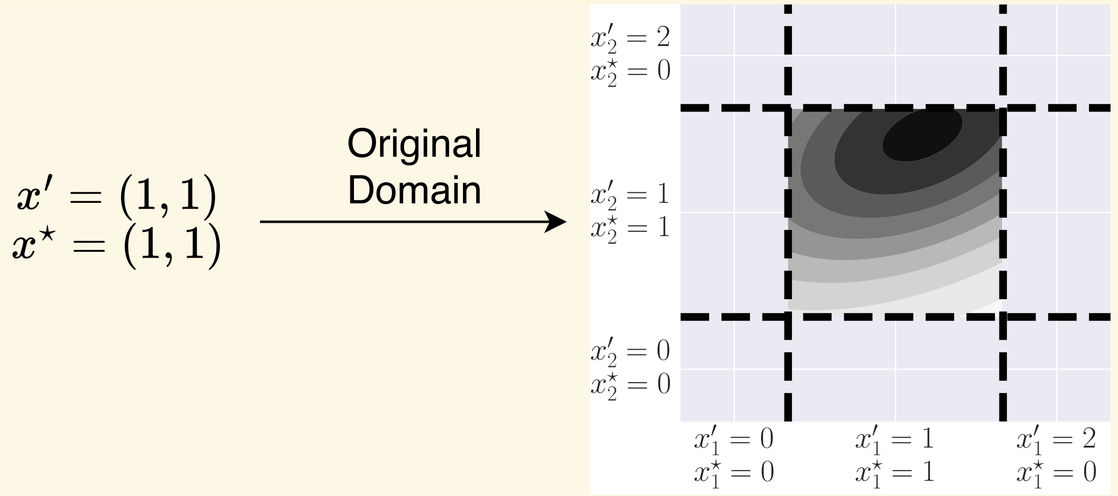






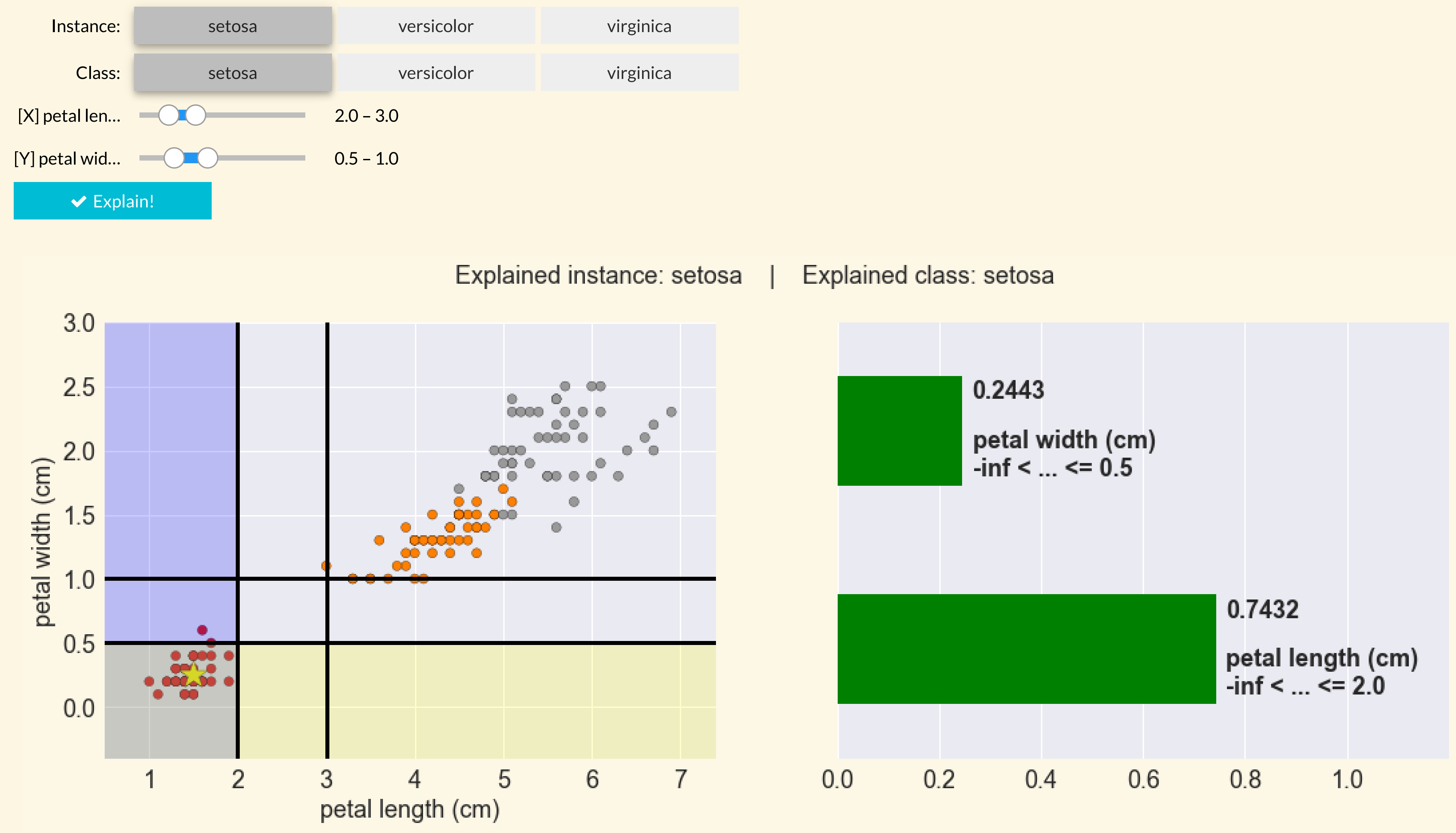
Interpretable representation



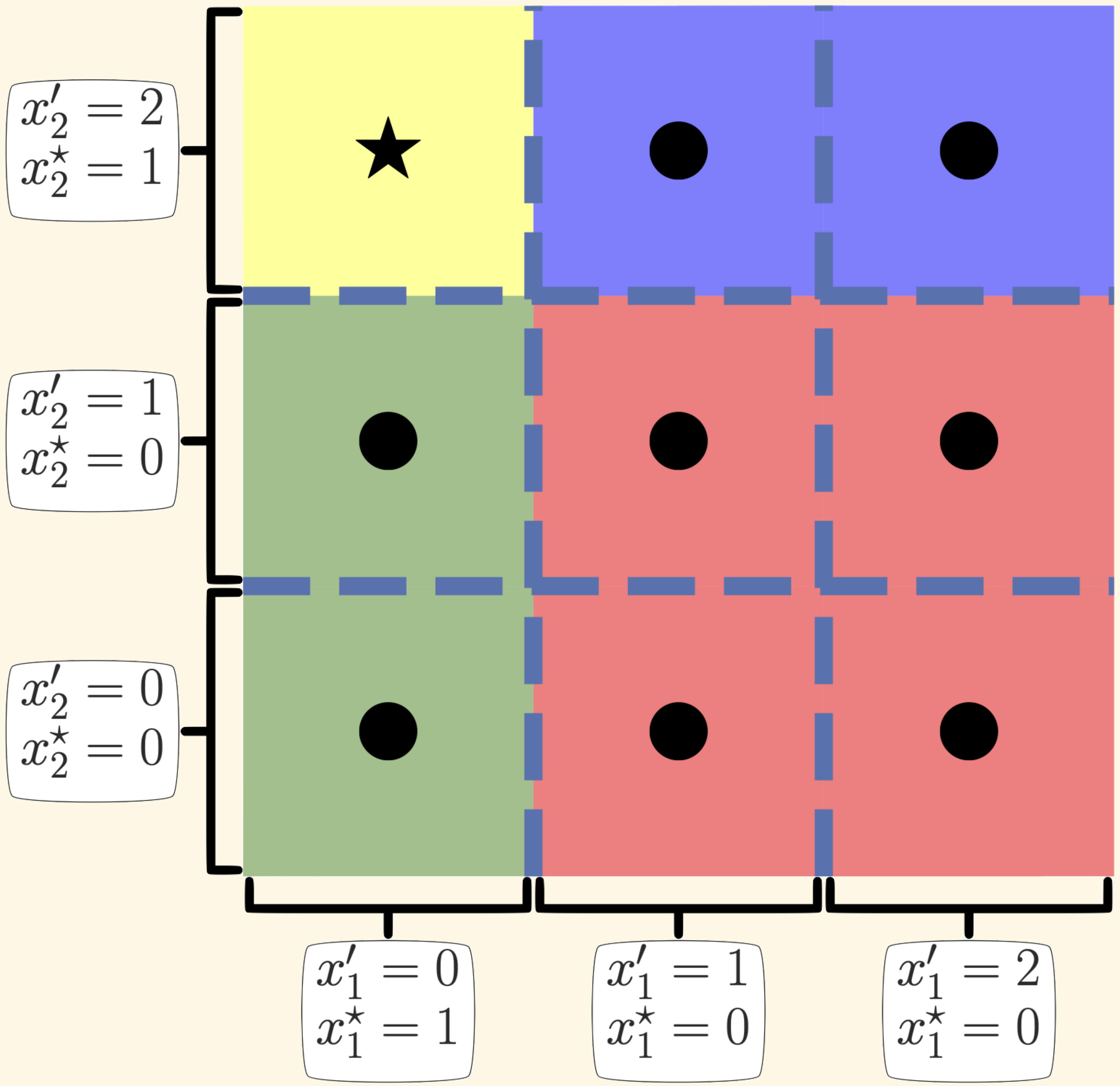


Explainer demo

```
In [18]: surrogate_tabular_explainer
```



But why? Meaning of the explanations



But why? Meaning of the explanations (ctd.)

$$\Theta = \begin{bmatrix} 1 & \frac{w_{11}+w_{10}}{\sum w_{ij}} & \frac{w_{11}+w_{01}}{\sum w_{ij}} \\ 1 & 1 & \frac{w_{11}}{w_{11}+w_{10}} \\ 1 & \frac{w_{11}}{w_{11}+w_{01}} & 1 \end{bmatrix}^{-1} \times \begin{bmatrix} \bar{y}_{\mathcal{W}} \\ \bar{y}_{\mathcal{W}_{11} \cup \mathcal{W}_{10}} \\ \bar{y}_{\mathcal{W}_{11} \cup \mathcal{W}_{01}} \end{bmatrix}$$

Take-home Messages

Explainability algorithms ***are not*** *monolithic entities*

*Explainers need to be **configured** or **tailor-made** for the application at hand*

*These are **diagnostic tools** that only become **explainers** when their provenance, caveats, properties and outputs are well-understood*

Do we really need to use complex methods to solve the problem at hand?

- *AI*
- *ML*
- *DL*
- *[insert the name of a new technology]*

Where to Go from Here

`FAT Forensics` <https://fat-forensics.org/>

- A modular Python toolkit for algorithmic Fairness, Accountability and **Transparency**
- Aimed at both *end-users* and *domain experts*
- Built for *research* and *deployment*

- *Sokol et al., 2020. FAT Forensics: A Python toolbox for implementing and deploying fairness, accountability and transparency algorithms in predictive systems*
- *Sokol et al., 2022. FAT Forensics: A Python toolbox for algorithmic fairness, accountability and transparency*

ECML-PKDD 2020 hands-on explainability tutorial

VIRTUAL CONFERENCE 14-18 SEP. ECML PKDD 2020

Tutorial resources: https://events.fat-forensics.org/2020_ecml-pkdd

---

- *Sokol et al., 2020. What and How of Machine Learning Transparency: Building Bespoke Explainability Tools with Interoperable Algorithmic Components*

Extra resources

- 2021 TAILOR – Summer School session
- University of Bristol Centre for Doctoral Training in Interactive Artificial Intelligence – BIAS Summer School session
- 2021 The Alan Turing Institute's AI UK
- …

- https://events.fat-forensics.org/
- https://github.com/fat-forensics/resources

Self-paced online learning materials

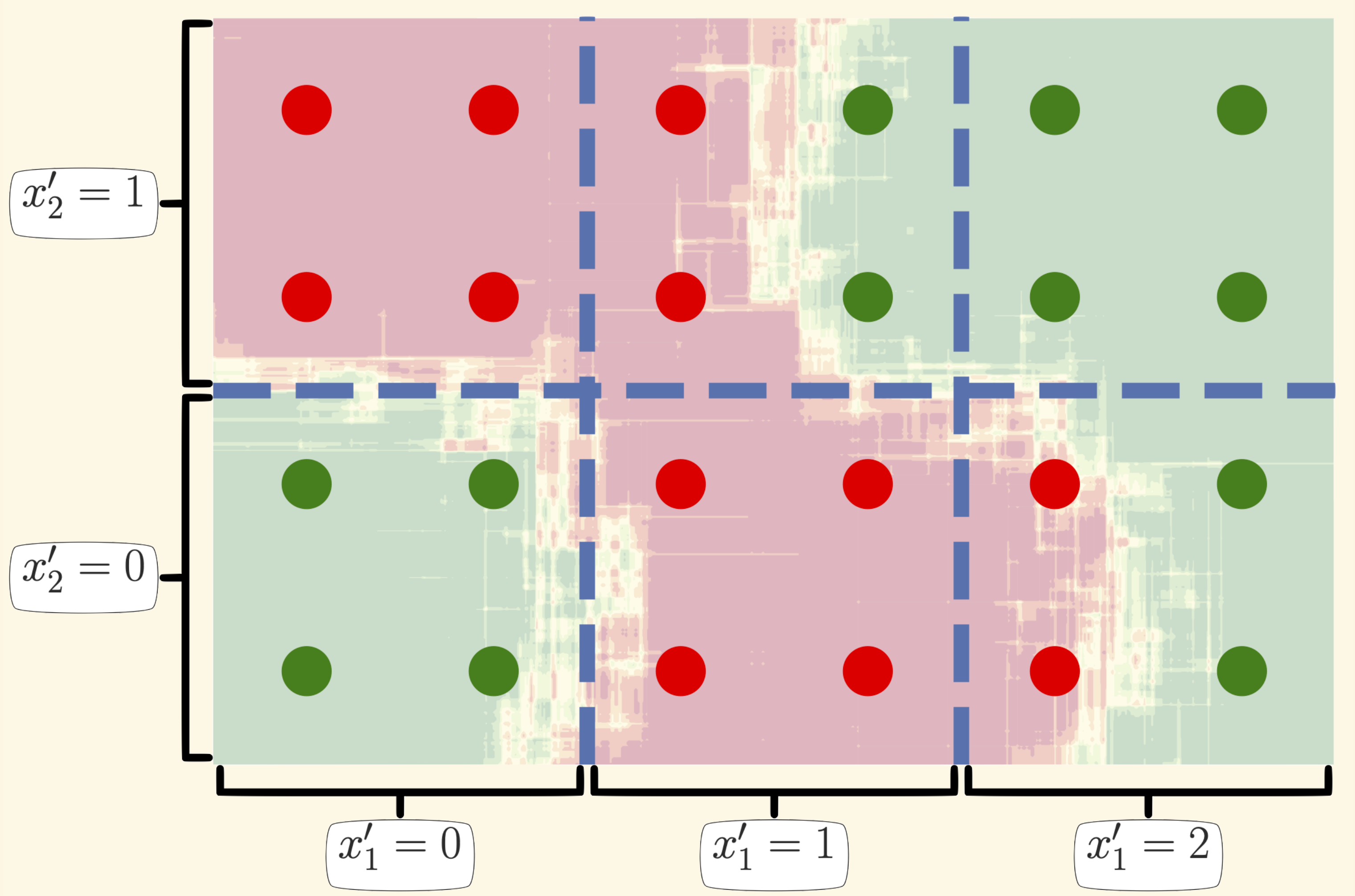
- Interactive online training resources on *interpretability*, *explainability* and *transparency*
- To be published in *late 2022 / early 2023*

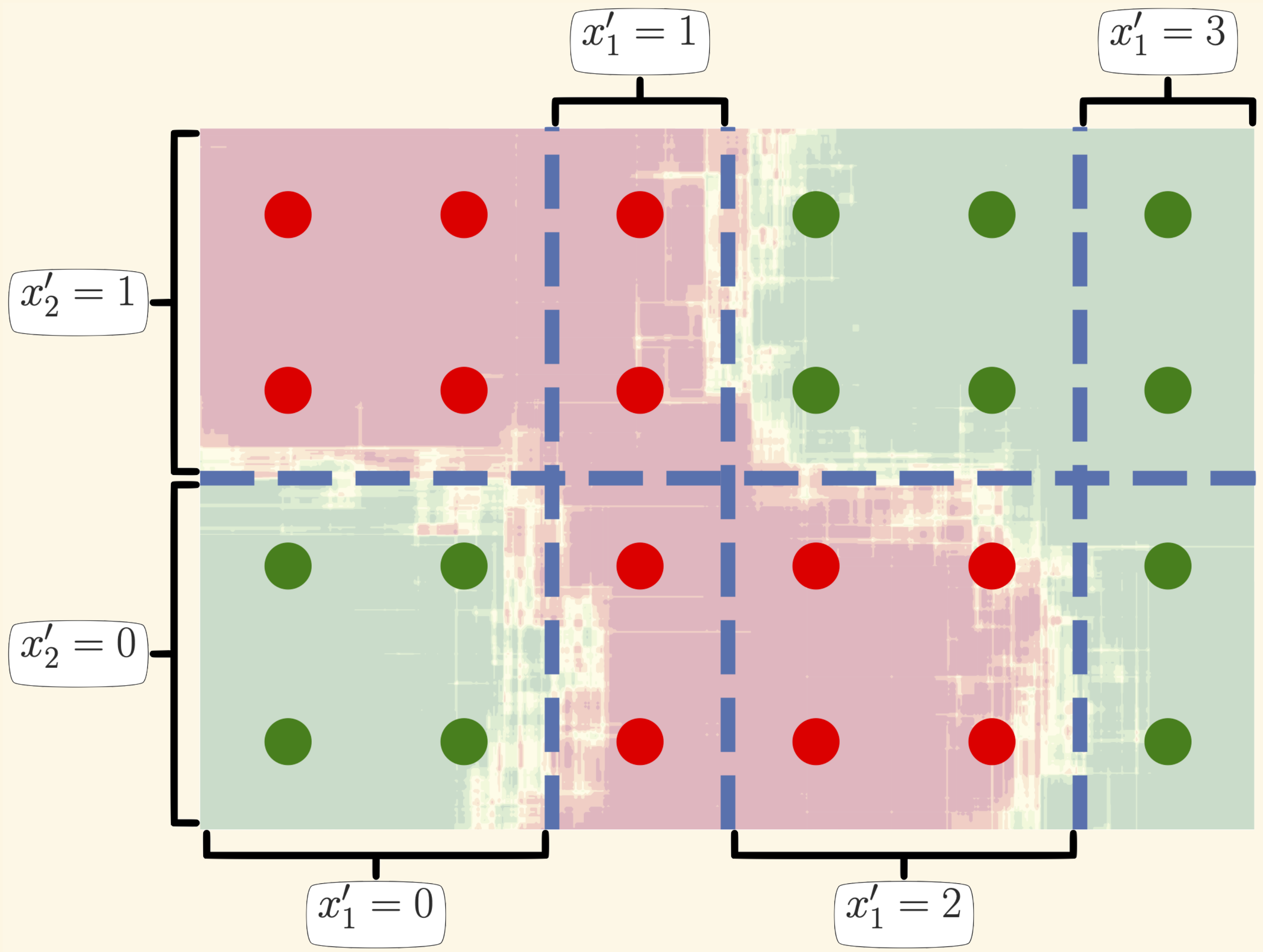
PhD / Master's Course materials

- Comprehensive overview of *interpretability*, *explainability* and *transparency*
- To be published *sometime in 2023*
- (Possibly transformed into a MOOC later in the year)

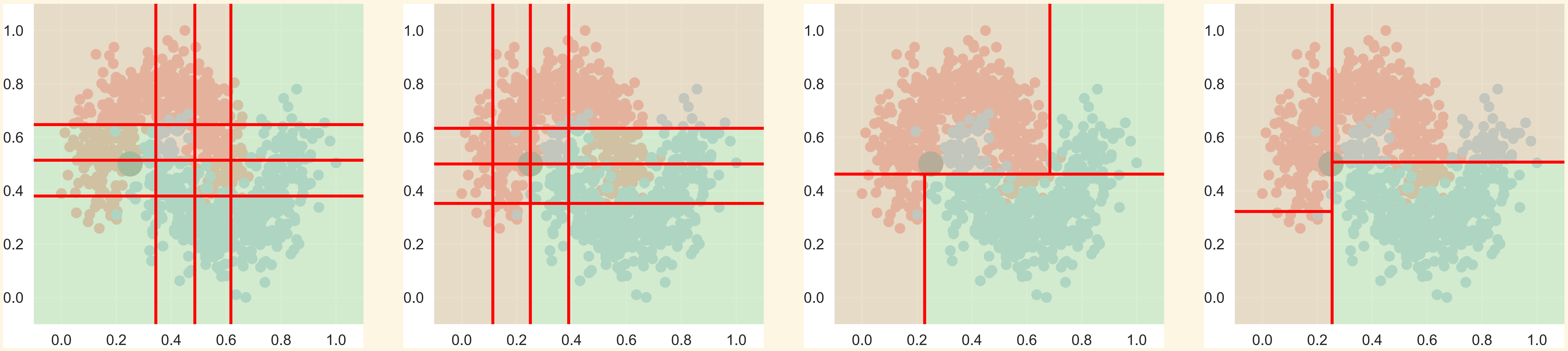
Helpers

Feature partition

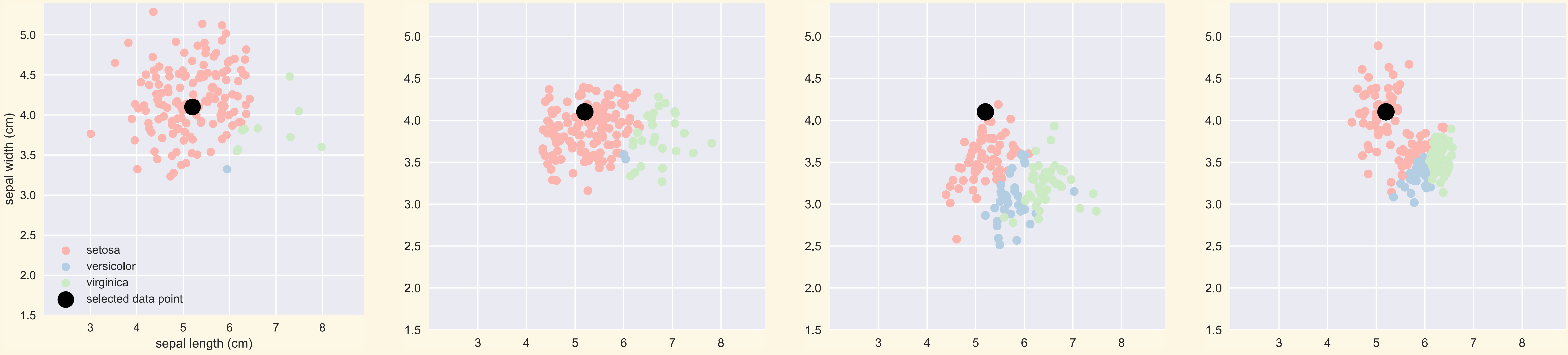




Feature partition (ctd.)



Sampling



sepal width (cm)
sepal length (cm)
setosa
versicolor
virginica
selected data point