eXplainable AI

From a model, to an explanation, to a user

Algorithms, algorithmic kernels, and shared problems

Why practical lessons?

The field is **highly multidisciplinary**, and AI advancements often prompt adaptation.

Lessons from the literature and practical applications on two projects:

- XAI (ERC, 2018 onward)
- FAIR (nation-wide project, 2023 onward)

XAI

Mapping complex *machine learning models* to *human-understandable* models.

Human understandable...?

Rather than the model, let's start from the user: who are they?

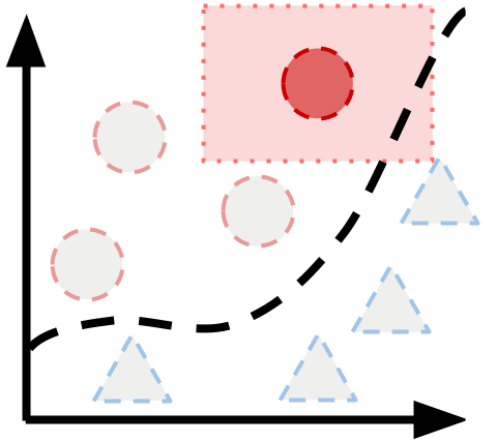
- user: they want to **properly act** on its prediction
- developer: they want to **debug** and **improve** the model
- auditor: they want to **inspect** the model

When, Where, What, How

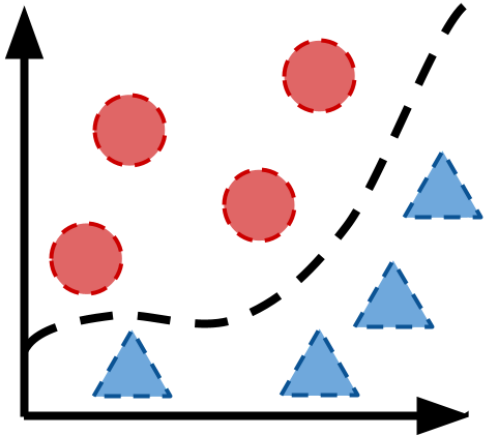
When does something happen?

Describing locally (on one instance) or globally (on multiple instances).

Locally



Globally

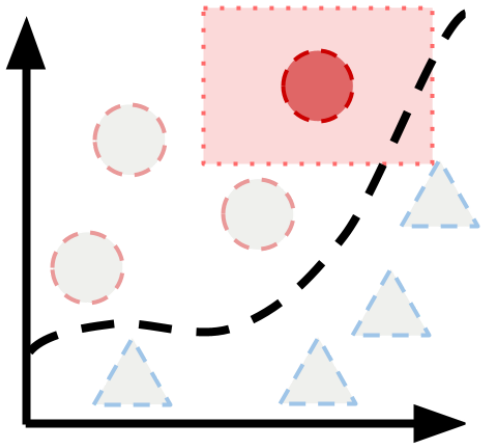


When does something happen?

Decision rules

```
if age >= 30 and  salary > 2.5k then grant loan
if age >= 50 and  salary < 2k then deny loan
```

- Can be extracted and validated programmatically
- Can leverage existing literature on rule extraction and decision tree induction

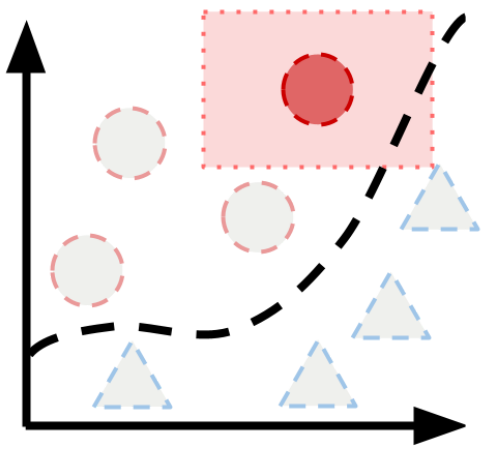


Factual and Counterfactual Explanations for Black-Box Decision Making, Guidotti et al.
Learning Certifiably Optimal Rule Lists, Angelino et al.

The augment-then-explain kernel

Idea. Create a dataset `X, y` of neighbors, then train an explainable model on it.

- How to study the local feature distribution?
- How to generate a (synthetic) neighborhood?



```
X = neighborhood(x)
y = query(model)

explanation = Explainable_model(X, y)
```

Factual and Counterfactual Explanations for Black-Box Decision Making, Guidotti et al.

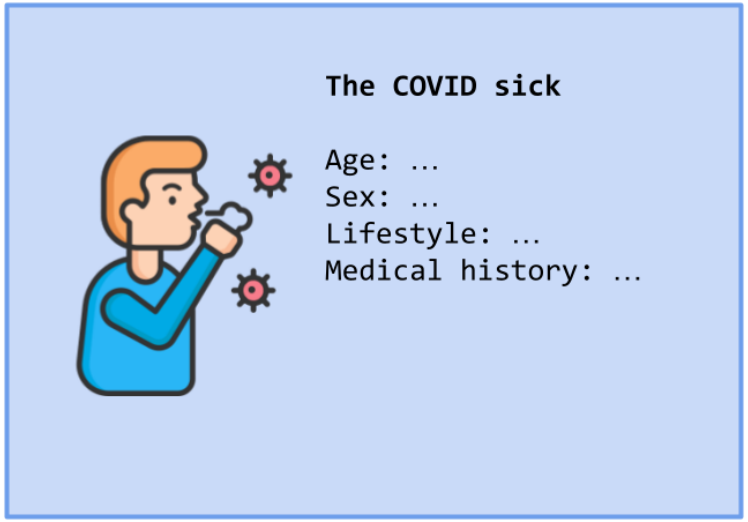
Anchors: High Precision Model-Agnostic Explanations, Ribeiro et al.

When does something happen?

Prototypes

```
Person

Age: 50
Sex: Male
Smoker
Sedentary
```



Provide a (set of) prototypical example(s) for a desired class.

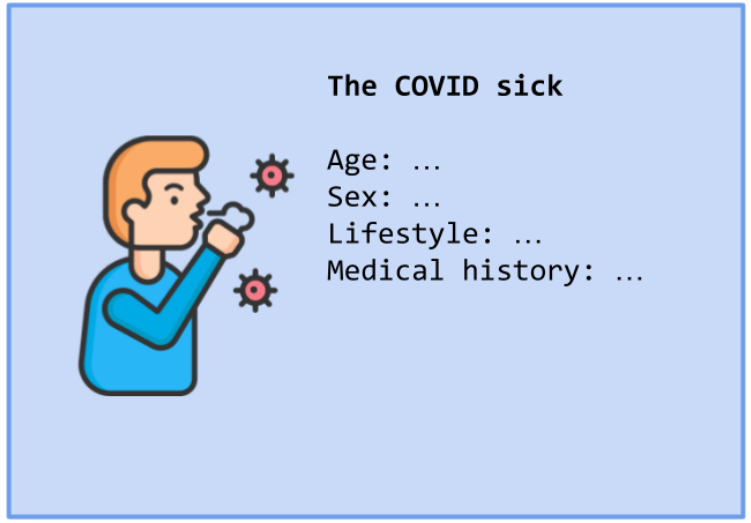
This Looks Like That: Deep Learning for Interpretable Image Recognition, Chen et al.
Explaining image classifiers generating exemplars and counter-exemplars from latent representations, Guidotti et al.

The extract-then-compare kernel

Idea Extract prototypes, then create a case-based reasoning explainable by design model.

```
P = prototypes(X, y)

explainable_model = Model(X, P, y)
```



- How to select the prototypes?
- What kind of reasoning to provide?

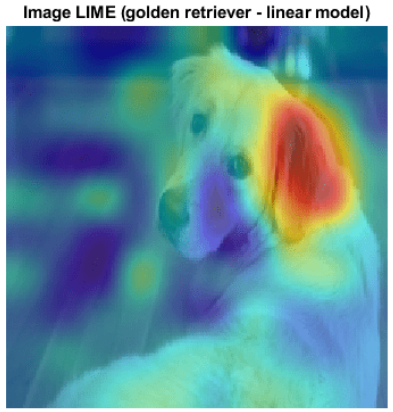
This Looks Like That: Deep Learning for Interpretable Image Recognition, Chen et al.
Explaining image classifiers generating exemplars and counter-exemplars from latent representations, Guidotti et al.

What happens...?

Feature importance

What feature is relevant for the prediction?

```
feature 0: 0.432
...
feature n - 1: 0.016
```



Why Should I Trust You? Explaining the Predictions of Any Classifier, Ribeiro et al.
A Unified Approach to Interpreting Model Predictions, Lundberg et al.

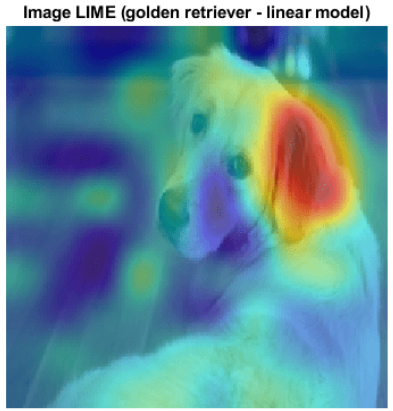
The perturb-then-analyze kernel

Feature importance

Idea Perturb the data, then analyze the perturbation.

```
X' = perturb(x)
y = query(X', y)

explanation = analysis(X', y)
```



- What to perturb?
- How to perturb?

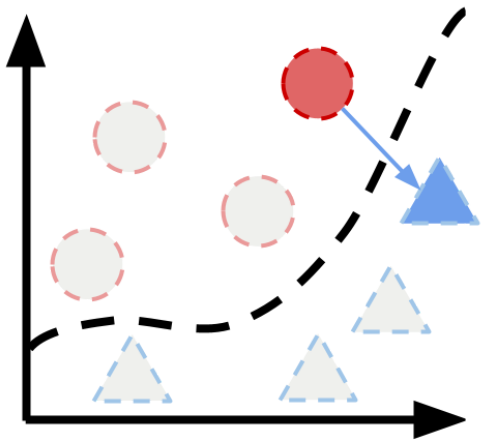
Why Should I Trust You? Explaining the Predictions of Any Classifier, Ribeiro et al.
A Unified Approach to Interpreting Model Predictions, Lundberg et al.

How to change it?

Counterfactual examples and rules

Define a rule for the prediction to change, or directly provide a close instance of different class.

```
if improve  salary  to > 2.5k then grant loan
```



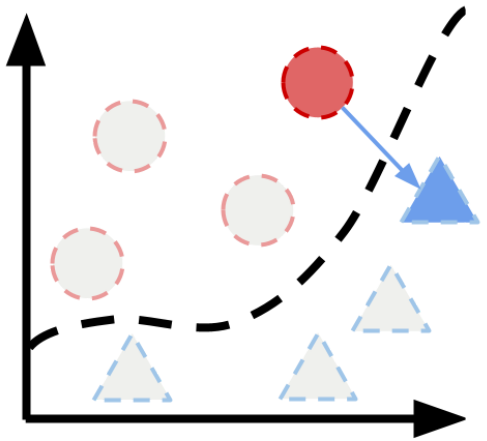
Algorithmic Recourse: from Counterfactual Explanations to Interventions, Karimi et al.
Matching the expert's knowledge via a counterfactual-based feature importance measure, Alfeo et al.

The search-then-optimize kernel

```
direction = find_counterfactual_direction(model)

counterfactual = find_examples(direction)
```

- Is the counterfactual within distribution?
- Does it respect the causal model of the data?



FACE: Feasible and Actionable Counterfactual Explanations, Poyiadzi et al.
Decisions, Counterfactual Explanations and Strategic Behavior, Tsirtsis and Gomez-Rodriguez
Explaining NLP Models via Minimal Contrastive Editing (MICE), Ross et al.

Looking ahead, systemic kernel problems

- **augment-then-explain** How to infer the local feature distribution?
- **perturb-then-analyze** Is the perturbation within distribution?
- **search-then-optimize** Is the search direction within distribution?

Looking ahead, systemic kernel problems

Most algorithms based on a few basic kernels relying on estimated distributions which are

- loose estimates
- often slow
- ignore the data-generating model
- can't easily integrate user knowledge