

# New frontiers in machine learning interpretability

## Mihaela van der Schaar

John Humphrey Plummer Professor of Machine Learning, Artificial Intelligence and Medicine, University of Cambridge //  
Director, Cambridge Center for AI in Medicine // Turing Faculty Fellow, The Alan Turing Institute



van\_der\_Schaar  
LAB

vanderschaar-lab.com



UNIVERSITY OF  
CAMBRIDGE



[mv472@cam.ac.uk](mailto:mv472@cam.ac.uk)



[@MihaelaVDS](https://twitter.com/MihaelaVDS)



[linkedin.com/in/  
mihaela-van-der-schaar/](https://www.linkedin.com/in/mihaela-van-der-schaar/)

# Our research team

<https://www.vanderschaar-lab.com/>  
→ Research Team



Fergus Imrie



Alan Jeffares



Alex Chan



Alicia Curth



Alihan Hüyük



Boris van Breugel



Dan Jarrett



Hao Sun



Jeroen Berrevoets



Jonathan Crabbé



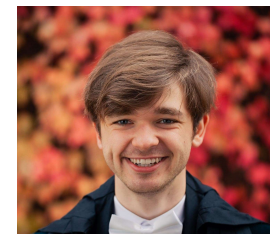
Krzysztof Kacprzyk



Nicolas Huynh



Nabeel Seedat



Paulius Rauba



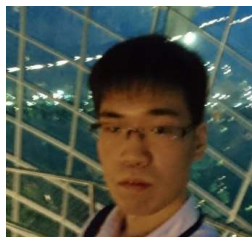
Sam Holt



Tennison Liu



Yangming Li



Yuchao Qin



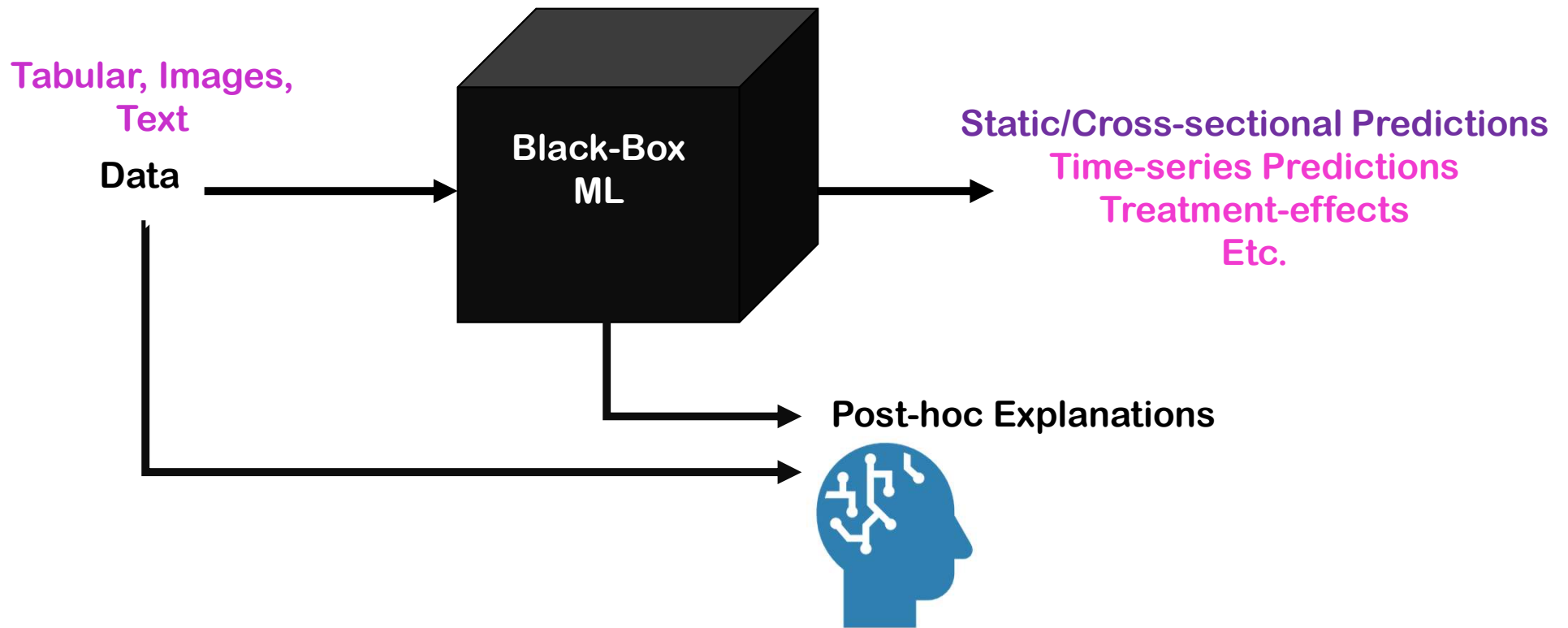
Zhaozhi Qian

## Machine learning interpretability is essential

- ***Understanding.*** Users need to understand, quantify and manage risk
- ***Transparency.*** Users need to comprehend how the model makes certain predictions
- ***Trustworthiness.*** Users can debug the model based on their knowledge
- ***Discovery.*** Users need to distil insights and new knowledge from the learned model
- ***Avoid implicit bias.*** Users need to be able to check whether the model does not learn biases



# We need to go beyond interpretability of static prediction models



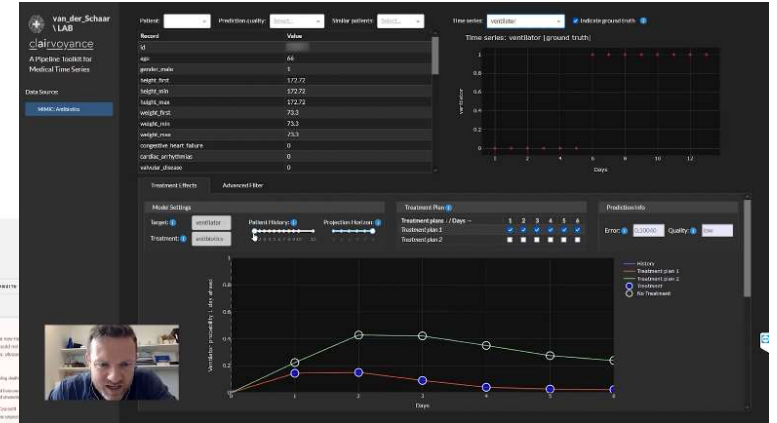
# What do clinicians want from an explanation?



## Adjutorium app demonstration



*If you have any questions/comments, please post them into the general Zoom chat; the earlier, the better!*



[www.vanderschaar-lab.com/making-machine-learning-interpretable-a-dialog-with-clinicians/](http://www.vanderschaar-lab.com/making-machine-learning-interpretable-a-dialog-with-clinicians/)

## 5 classes of explanation methods

---

Explanation class	Definition
Feature-based	Provides the importance of each feature to model predictions
Example-based	Explains model predictions with reference to other examples
Concept-based	Explains model predictions with reference to a human-defined concept
Model-based	Explains model predictions via an auxiliary meta-model
Counterfactual	Explains model predictions by generating synthetic example(s) that are similar but with a different prediction

---



# Today's talk: Four types of interpretability

---

Explanation class	Definition
Feature-based	Provides the importance of each feature to model predictions
Example-based	Explains model predictions with reference to other examples
Concept-based	Explains model predictions with reference to a human-defined concept
Model-based	Explains model predictions via an auxiliary meta-model
Counterfactual	Explains model predictions by generating synthetic example(s) that are similar but with a different prediction

---



# Interpretability Resources

Overview of our lab's work related to interpretability

[vanderschaar-lab.com/](https://vanderschaar-lab.com/)  
→ Research pillars  
→ Interpretable ML

The screenshot shows the website for vander\_schaar LAB. The main heading is "Interpretable machine learning". Below this, there are several text blocks and lists. A "JOIN US" button is visible on the right side, and a "2021 open house" banner is at the bottom right. The text on the page includes:

*Machine learning is capable of enabling truly personalized healthcare; this is what our lab calls "bespoke medicine."*

More info on bespoke medicine can be found here.

*Interpretability is essential to the success of the machine learning and AI models that will make bespoke medicine a reality. Despite its acknowledged importance and value, the actual concept of interpretability has resisted definition and is not well understood.*

*Our lab has conducted field-leading research into a variety of forms of interpretability for years, and has developed a unique and cohesive framework for categorizing and developing interpretable machine learning models. Our framework is presented on this page, alongside much of the accompanying research. In the hope of advancing the discussion on this crucial topic and inspiring readers to engage in new projects and research.*

*The content of this page is designed to be accessible and useful to a wide range of stakeholders.*

**On this page:**

- Interpretability: a concept with clear value but an unclear definition
- Type 1 Interpretability: feature importance
- Type 2 Interpretability: similarity classification
- Type 3 Interpretability: unswept rules and laws
- Type 4 Interpretability: transparent risk equations
- Peering into the ultimate black box
- Find out more and get involved
  - Lecture on Interpretability at The Alan Turing Institute and related blog post
  - Roundtables on Interpretability with clinicians
  - Our engagement sessions

This page is one of several introductions to areas that we see as "research pillars" for our lab. It is a living document, and the content here will evolve as we continue to reach out to the machine learning and healthcare communities, building a shared vision for the future of healthcare.

It should be noted that transparent risk equations can be applied to the other three types of interpretability listed above. Using patient features as inputs and risk as outputs, we can identify variable importance, classify similarities, discover variable interactions, and enable hypothesis induction.

**Peering into the ultimate black box**

The bulk of this page has been dedicated to exploring what it means to make machine learning models "interpretable" and showing how this can be done in a variety of ways. In our view, this is still premised on a relatively blinkered view that ignores some very exciting possibilities for interpretability and machine learning—namely, for humans to use interpretability to understand our own decision-making process.

This possibility is at the heart of quantitative epistemology, a new and transformationally significant research pillar pioneered by our lab. The purpose of this research is to develop a strand of machine learning aimed at understanding, supporting, and improving human decision-making. We aim to do so by building machine learning models of decision-making, including how humans acquire and learn from new information, establish and update their beliefs, and act on the basis of their cumulative knowledge. Because our approach is driven by observational data in studying knowledge as well as using machine learning methods for supporting and improving knowledge acquisition and its impact on decision-making, we call this "quantitative epistemology".

We develop machine learning models that capture how humans acquire new information, how they pay attention to such information, how their beliefs may be represented, how their internal models may be structured, how these different levels of knowledge are leveraged in the form of actions, and how such knowledge is learned and updated over time. Our methods are aimed at studying human decision-making, identifying potential suboptimalities in beliefs and decision processes (such as cognitive biases, selective attention, imperfect retention of past experience, etc.), and understanding risk attitudes and their implications for learning and decision-making. This would allow us to construct decision support systems that provide humans with information pertinent to their intended actions, their possible alternatives and counterfactual outcomes, as well as other evidence to empower better decision-making.

You can learn more about quantitative epistemology and explore some of our first papers in this area in the article below.



van\_der\_Schaar  
LAB

[vanderschaar-lab.com](https://vanderschaar-lab.com)



UNIVERSITY OF  
CAMBRIDGE



# Interpretability Resources

## Explainers

Different model architectures can require different interpretability models, or "Explainers". Below are all the explainers included in this repository, with links to their source code and the papers that introduced them. SimplEx, Dynamask, shap, and Symbolic Pursuit have a common python interface implemented for them for ease of implementation (see [Interface](#) above and [Implementation](#) and [Notebooks](#) below). But any of the other methods can also be implemented by using the code in the GitHub column of the table below.

Explainer	Affiliation	GitHub	Paper	Date of Paper
Concept Activation Regions (CARs)	<a href="#">van der Schaar Lab</a>	<a href="#">CARs source Code</a>	<a href="#">CARs Paper</a>	2022
ITErpretability	<a href="#">van der Schaar Lab</a>	<a href="#">ITErpretability Source Code</a>	<a href="#">ITErpretability Paper</a>	2022
Label-Free XAI	<a href="#">van der Schaar Lab</a>	<a href="#">Label-Free XAI Source Code</a>	<a href="#">Label-Free XAI Paper</a>	2022
SimplEx	<a href="#">van der Schaar Lab</a>	<a href="#">SimplEx Source Code</a>	<a href="#">SimplEx Paper</a>	2021
Dynamask	<a href="#">van der Schaar Lab</a>	<a href="#">Dynamask Source Code</a>	<a href="#">Dynamask Paper</a>	2021
Symbolic Pursuit	<a href="#">van der Schaar Lab</a>	<a href="#">Symbolic Pursuit Source Code</a>	<a href="#">Symbolic Pursuit Paper</a>	2020
INVASE	<a href="#">van der Schaar Lab</a>	<a href="#">INVASE Source Code</a>	<a href="#">INVASE Paper</a>	2019
SHAP	University of Washington	<a href="#">SHAP Source Code</a> (pytorch implementation: <a href="#">Captum</a> <a href="#">GradientShap</a> )	<a href="#">SHAP Paper</a>	2017

## Open Source Code

[github.com/vanderschaarlab/Interpretability](https://github.com/vanderschaarlab/Interpretability)

## Selecting an Interpretability Method

Figure 3 shows a flowchart to help with the process of selecting the method that is most appropriate for your project.

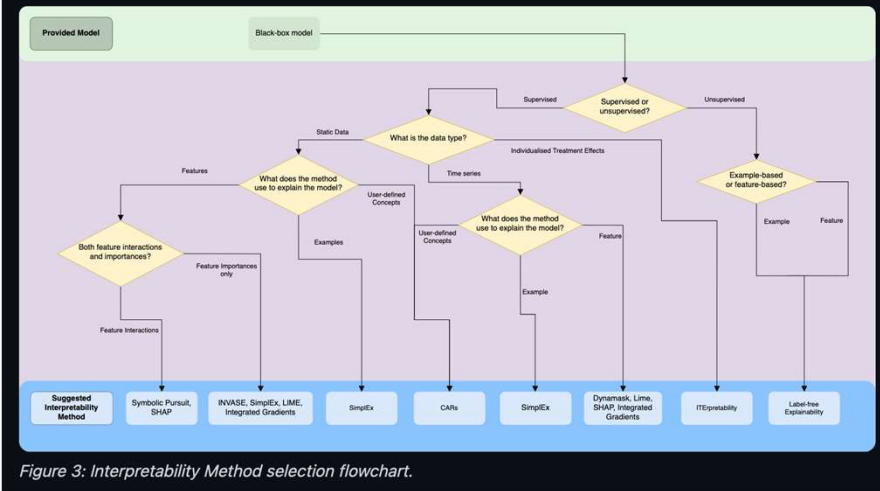


Figure 3: Interpretability Method selection flowchart.

## Implementation and Notebooks

This repository includes a common python interface for the following interpretability methods: SimplEx, Dynamask, shap, and Symbolic Pursuit. The interface provides the same methods for each of the methods such that you can use the same python methods in your scripts to set up an explainer for each interpretability method. The methods that are:

- `init`: Instantiate the class of explainer of your choice.
- `fit`: Performs and training for the explainer (This is not required for Shap explainers).
- `explain`: Provide the explanation of the data provided.
- `summary_plot`: Visualize the explanation.

There are also Notebooks in this GitHub repository to demonstrate how each create the explainer object. These explainers can be saved and uploaded into the Interpretability Suite user interface.

Interpretability Suite
Share ☆ ☰

Interpretability Suite

- SimplEx
- Dynamask
- Shap
- Symbolic Pursuit

# SimplEx

SimplEx is a case-based interpretability method. It can work with either tabular or time series data. You can read more about it in the [paper](#).

For clinically focussed examples go to the bespoke [SimplEx Demonstrator](#). And for further information, [here](#) is a video demonstration of the clinical SimplEx app.

[Examples](#) Upload your own Explainer

Data type:  
Tabular

Dataset:  
Iris

Model:  
MLP

Test record:

0 20

### Test record:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	Test Prediction	Test Label
Test Record	7.7000	3.0000	6.1000	2.3000	2	2

### Corpus:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	Example Importance	Corpus Prediction	Corpus Label
Corpus member 0	7.600000	3.000000	6.600000	2.100000	0.9999	2	2
Corpus member 1	7.700000	2.800000	6.700000	2.800000	0.9999	2	2
Corpus member 2	7.700000	2.600000	6.900000	2.300000	0.9999	2	2
Corpus member 3	7.700000	3.800000	6.700000	2.200000	0.9999	2	2
Corpus member 4	7.200000	3.600000	6.100000	2.500000	0.9999	2	2

github.com/vanderschaarlab/Interpretability

# Our Resources to go Further



## Our Papers

[vanderschaar-lab.com/interpretable-machine-learning/](https://vanderschaar-lab.com/interpretable-machine-learning/)

## Our Code

[github.com/vanderschaarlab/Interpretability](https://github.com/vanderschaarlab/Interpretability)



van\_der\_Schaar  
\ LAB

[vanderschaar-lab.com](https://vanderschaar-lab.com)



# Four types of interpretability

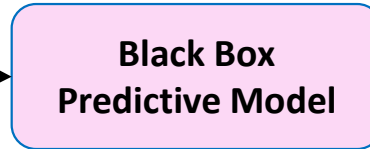
1. **Feature-based interpretability**
  - Static (global/personalized)
  - Time-series
  - Causal effect inference



# From Global to Individual Feature Importance



Age, Gender,  
Diabetes,  
Hypertension,  
SBP, ....



**Mortality due to  
Covid-19: 0.78**



van\_der\_Schaar  
\ LAB

[vanderschaar-lab.com](http://vanderschaar-lab.com)



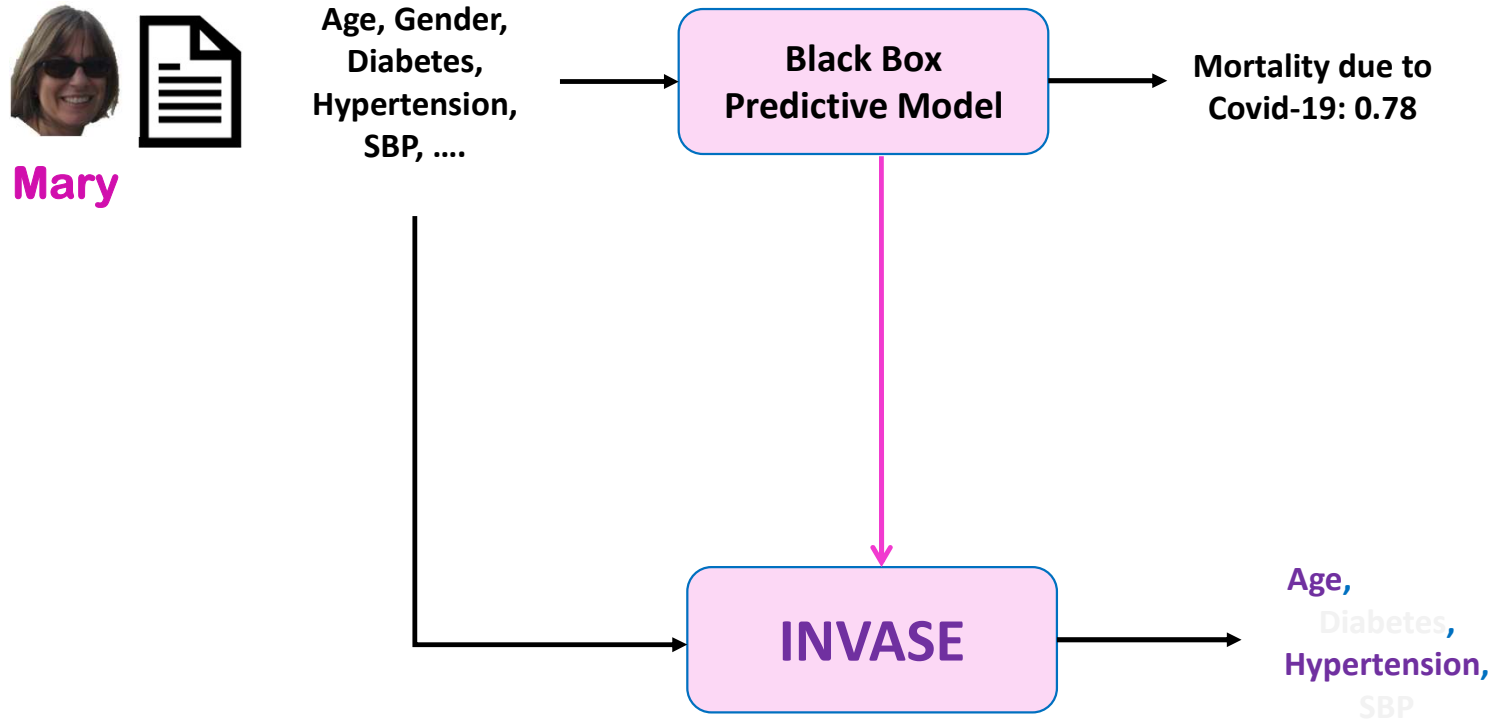
UNIVERSITY OF  
CAMBRIDGE

# Limitations of other methods for model interpretability

Method	Feature importance	Individualized feature importance	Model-independent	Identifying the set of relevant features for each instance
<b>LASSO</b> [Tibshirani, 1996]	✓		✓	
<b>Knock-off</b> [Candes et al, 2016]	✓		✓	
<b>L2X</b> [Chen et al, 2018]	✓	✓	✓	
<b>LIME</b> [Ribeiro et al, 2016]	✓	✓	✓	
<b>SHAP</b> [Lundberg et al, 2017]	✓	✓	✓	
<b>DeepLIFT</b> [Shrikumar et al, 2017]	✓	✓		
<b>Saliency</b> [Simonyan et al, 2013]	✓	✓		
<b>TreeSHAP</b> [Lundberg et al, 2018]	✓	✓		
<b>Pixel-wise</b> [Batch et al, 2015]	✓	✓		
<b>INVASE</b> [Yoon, Jordon and van der Schaar, 2019]	✓	✓	✓	✓

**INVASE discovers the number of relevant features for each instance**

# Which features of an individual are relevant for a prediction?



[Yoon, Jordon, vdS, ICLR 2019]



van\_der\_Schaar  
\ LAB

[vanderschaar-lab.com](http://vanderschaar-lab.com)



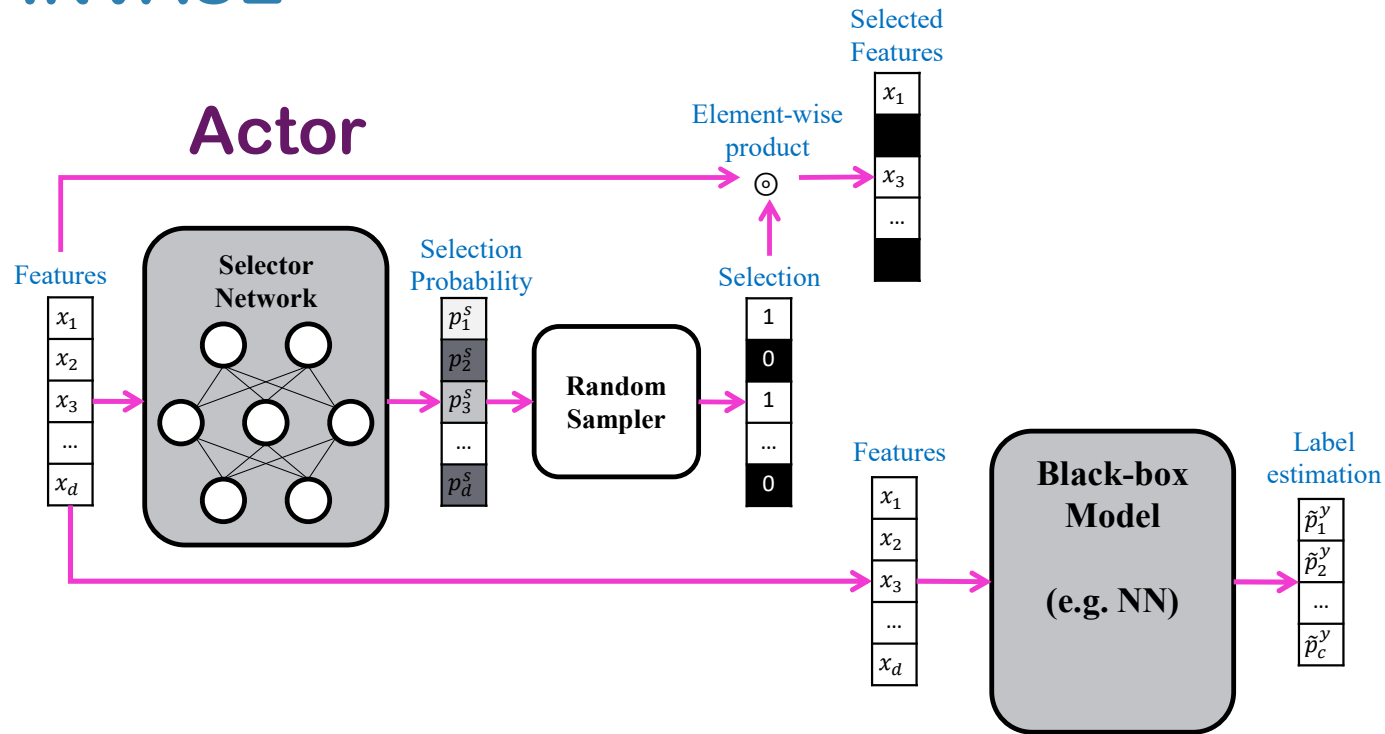
UNIVERSITY OF  
CAMBRIDGE

## INVASE [Yoon, Jordon, vdS, ICLR 2019]

- **How can we learn individualized feature importance?**
- **Key idea: Use Reinforcement Learning (RL)**
  - Make observations
  - Select “actions” on the basis of these observations
  - Determine “rewards” for these actions
  - Ultimately learn a policy which selects the best actions
    - i.e. actions that maximize rewards given observations
- **We use the Actor-Critic approach to RL**

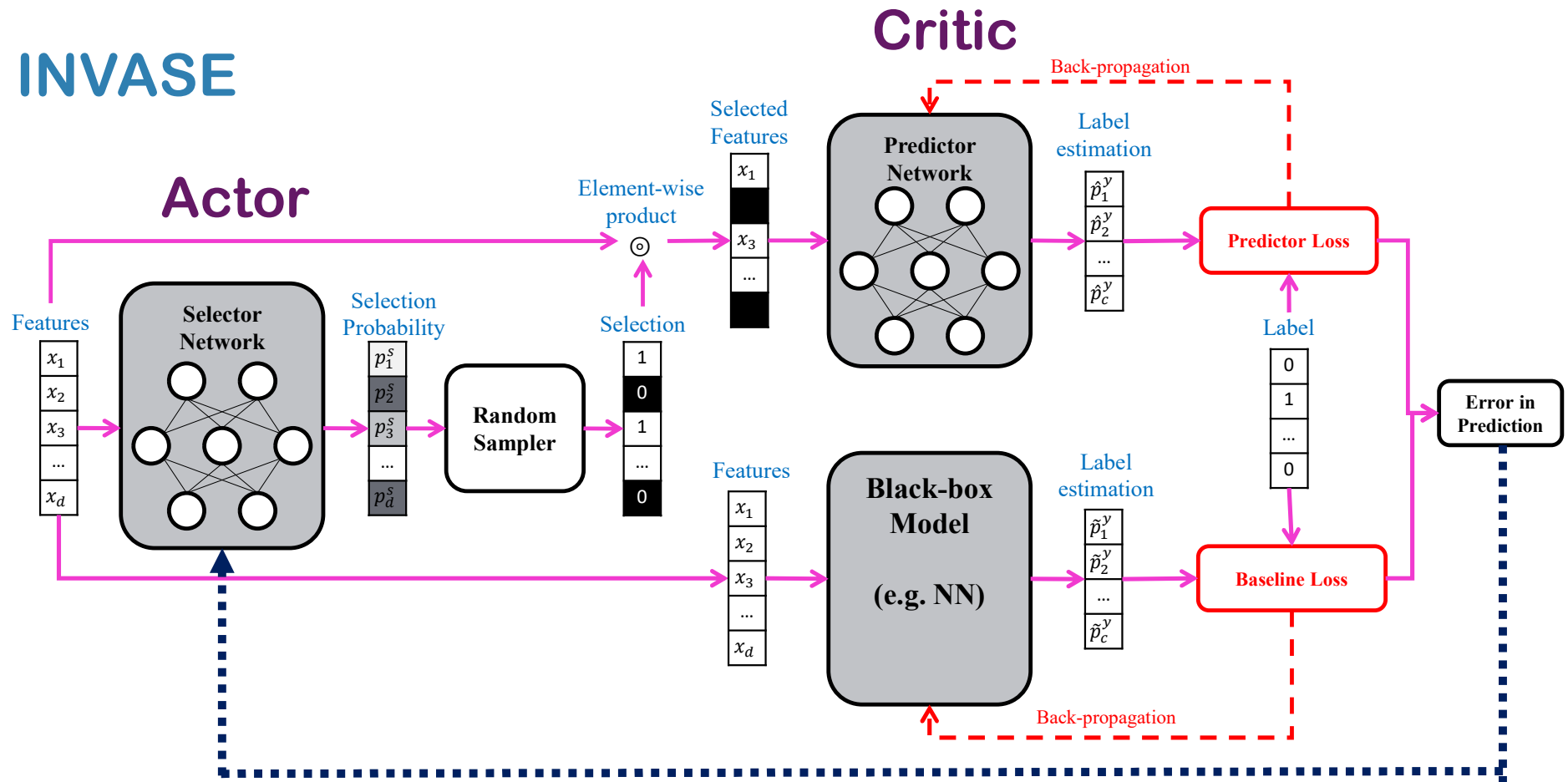


# INVASE



- **Selector network (actor)** takes instances and outputs vector of selection probabilities.

# INVASE



- **Predictor network (critic)** receives the selected features, makes predictions and provides **feedback** to the actor.

# Feature-based explanation – in medicine, we need to go beyond interpretability of static predictions

Time-series forecasting - **Dynamask [ICML 2021]**

Unsupervised learning methods – **Label-free explainability [ICML 2022]**

Causal effect inference – **ITErpretability [NeurIPS 2022]**



van\_der\_Schaar  
\ LAB

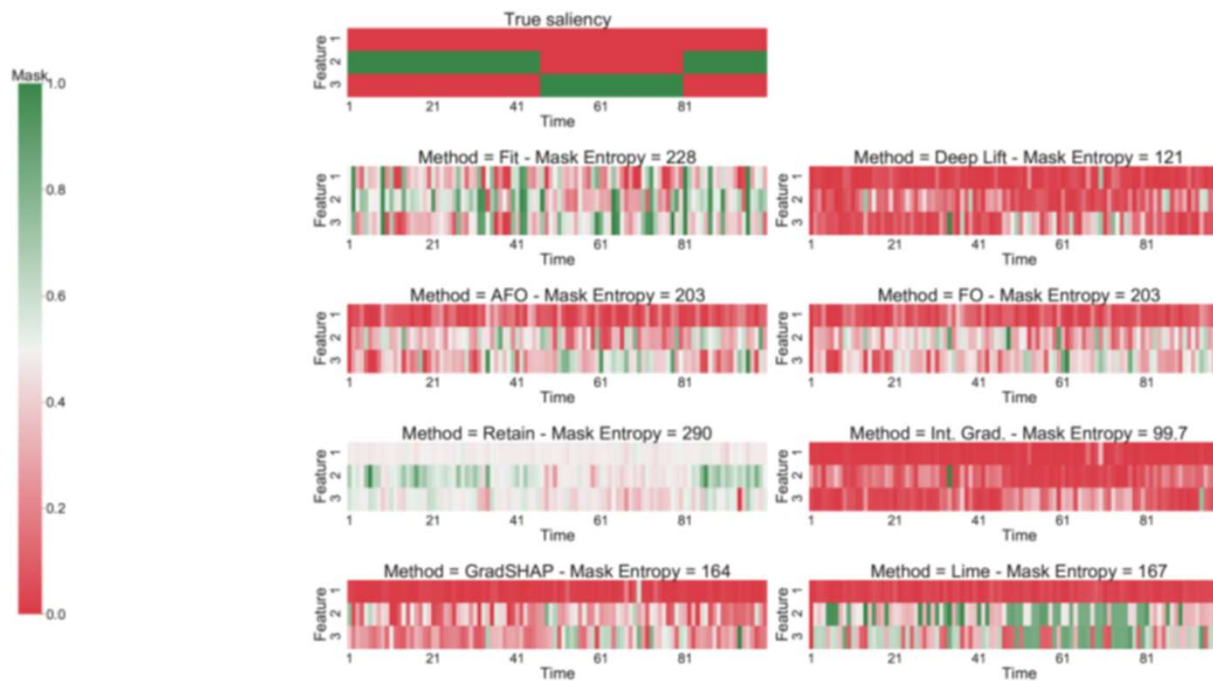
[vanderschaar-lab.com](http://vanderschaar-lab.com)



UNIVERSITY OF  
CAMBRIDGE

# Time-series forecasting – Do standard interpretability methods work?

**NO!** [Ismail et al., NeurIPS 2020]



van\_der\_Schaar  
LAB

[vanderschaar-lab.com](http://vanderschaar-lab.com)



UNIVERSITY OF  
CAMBRIDGE

# How to take the time context into account? [Crabbé, vdS, ICML 2021]

## Challenge: Time context matters!

Standard methods treat each input  $x_{t,i}$  as a feature

⇒ Time dependency is **ignored**

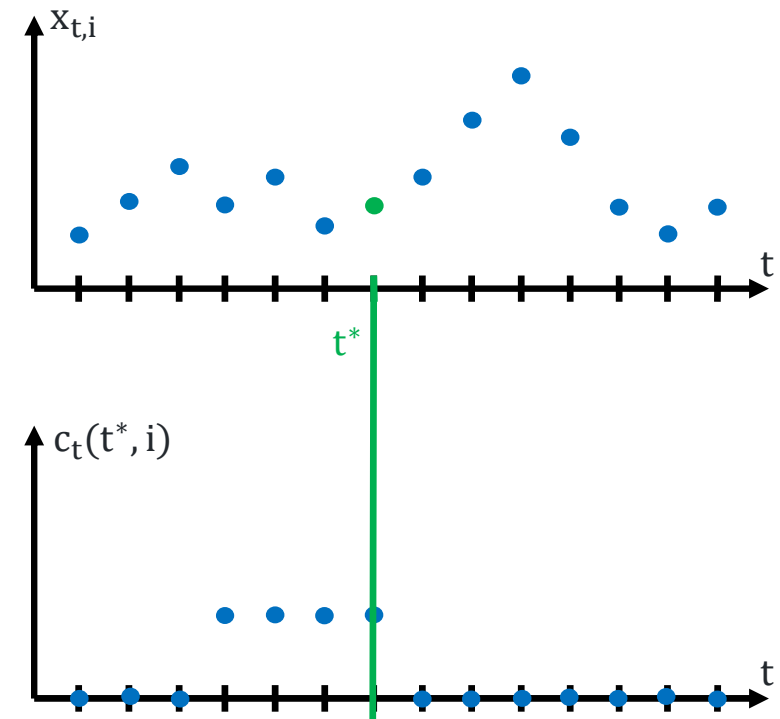
## Dynamic Perturbation Operator

Idea: perturb each  $x_{t^*,i}$  by using **neighbouring times**:

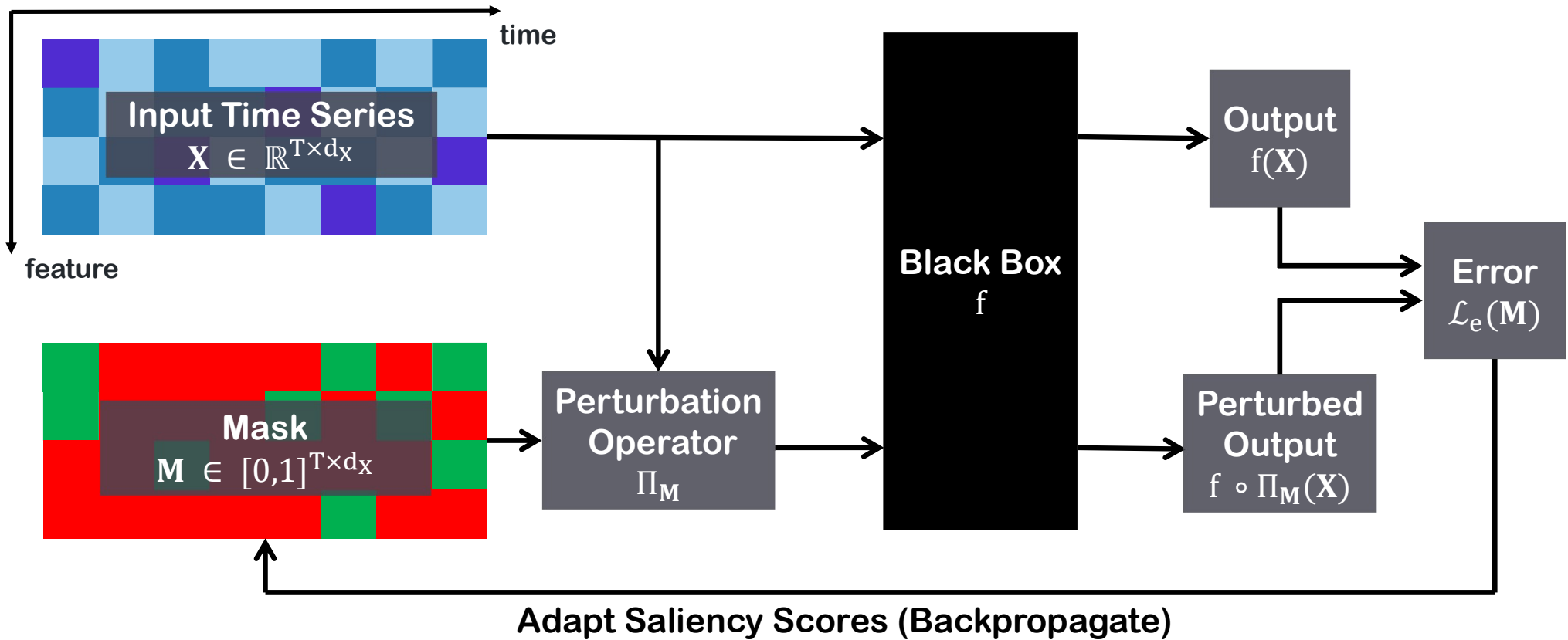
$$\text{Perturbed input } \pi(x_{t^*,i}; t^*, i) = \sum_{t=t^*-W_1}^{t^*+W_2} \text{Linear combination } c_t(t^*, i) \times x_{t,i}$$

⇒ Time dependency is **integrated** in perturbation

Past window perturbation:



# Dynamask [Crabbé, vdS, ICML 2021]



# We need “parsimonious” explanations

What do we mean by **parsimonious**?

Masks should **not** highlight more features than necessary

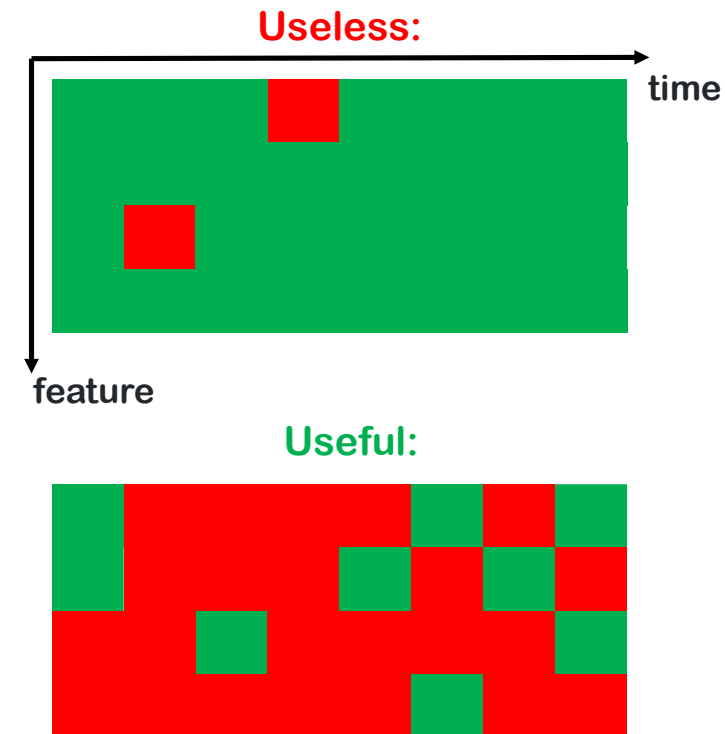
⇒ **Feature selection**

How to enable parsimony?

User selects desired fraction  $\alpha$  of most important features

Dynamask adds a regularization to enforce sparsity:

$$\mathcal{L}_\alpha(\mathbf{M}) = \|\text{vecsort}(\mathbf{M}) - \mathbf{r}_\alpha\|^2$$



van\_der\_Schaar  
\ LAB

vanderschaar-lab.com



UNIVERSITY OF  
CAMBRIDGE

# We need “congruous” explanations

What do we mean by **congruous**?

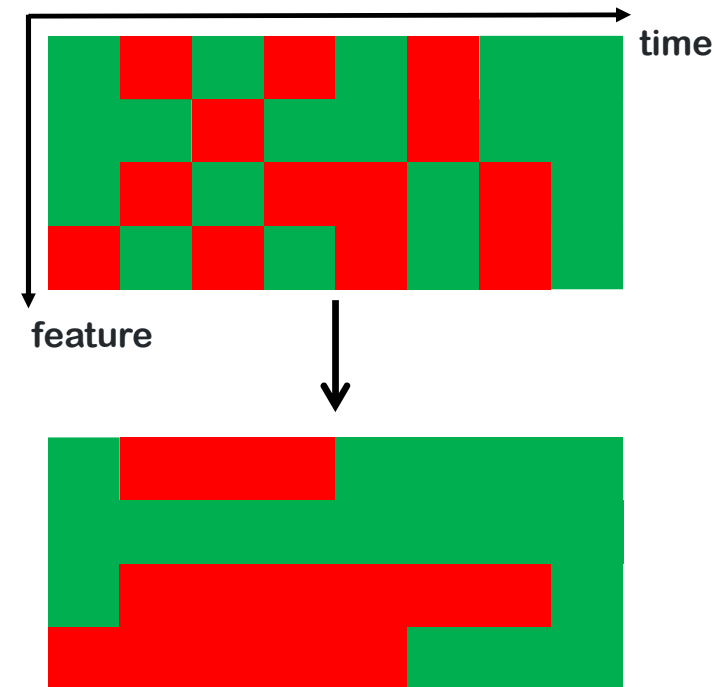
Masks should **avoid quick time variations** of the saliency

(Robustness)

How to enable congruity?

Dynamask adds a regularization to penalize saliency jumps over time:

$$\mathcal{L}_c(\mathbf{M}) = \sum_{t=1}^{T-1} \sum_{i=1}^{d_X} |m_{t+1,i} - m_{t,i}|$$



van\_der\_Schaar  
\ LAB

vanderschaar-lab.com



UNIVERSITY OF  
CAMBRIDGE



# Dynamask enables the saliency map to be “legible”

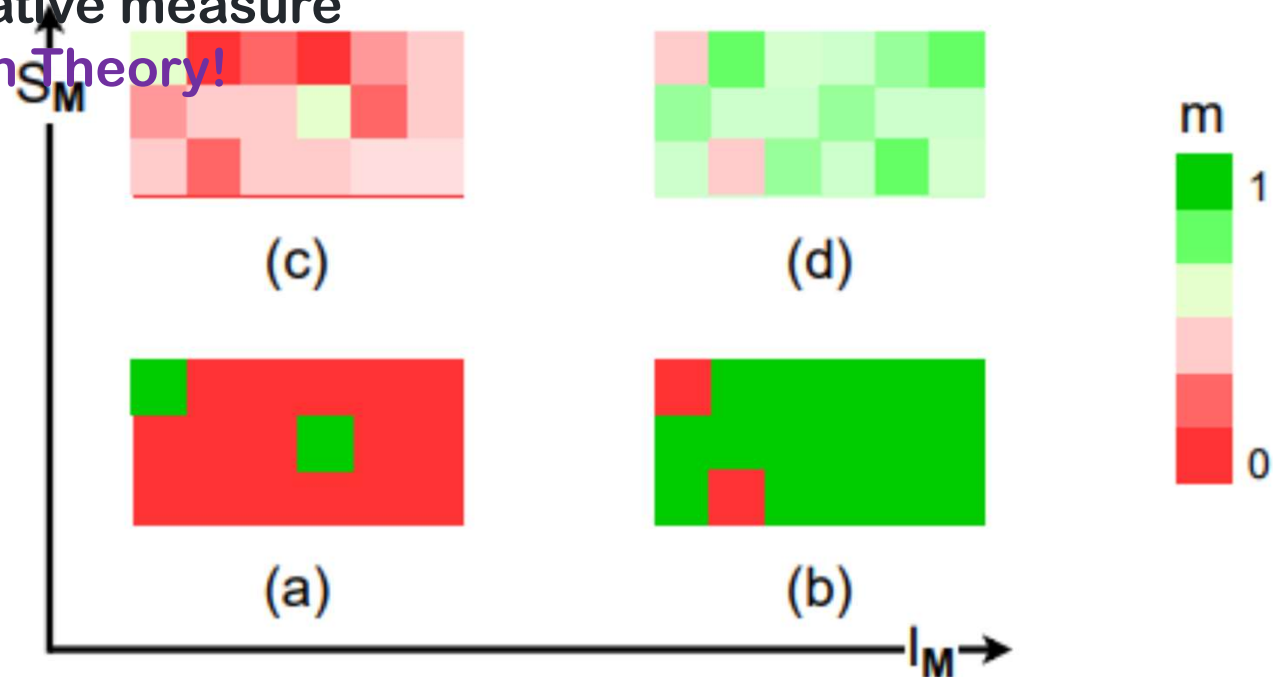
[Crabbé, vdS, ICML 2021]

How do we know if the “legibility” is achieved by an interpretability method?

We need a quantitative measure

We use Information Theory!

Introduced  
Mask information  
&  
Mask entropy



van\_der\_Schaar  
LAB

vanderschaar-lab.com



UNIVERSITY OF  
CAMBRIDGE

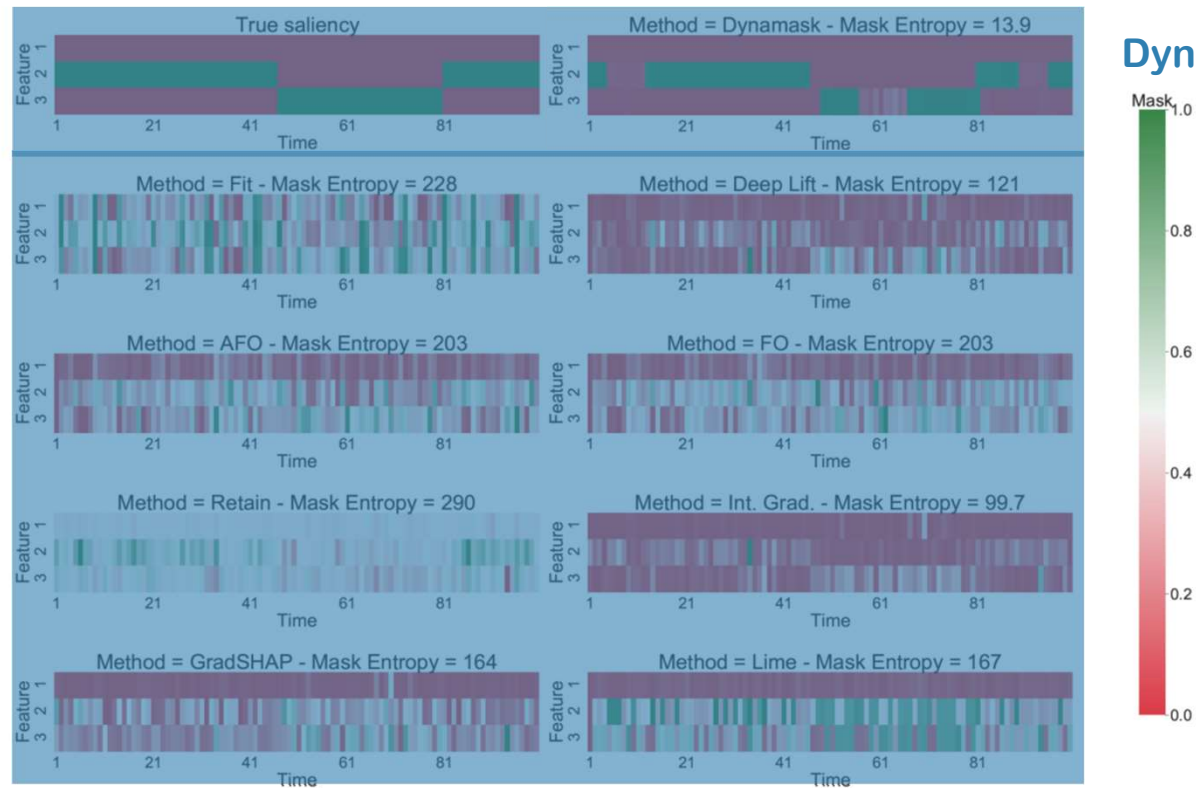
# Dynamask - Example

[Crabbé, vdS, ICML 2021]

Example number 5

True saliency

Dynamask saliency



Baseline saliency



van\_der\_Schaar  
\ LAB

[vanderschaar-lab.com](http://vanderschaar-lab.com)



# Explaining *Unsupervised* Models

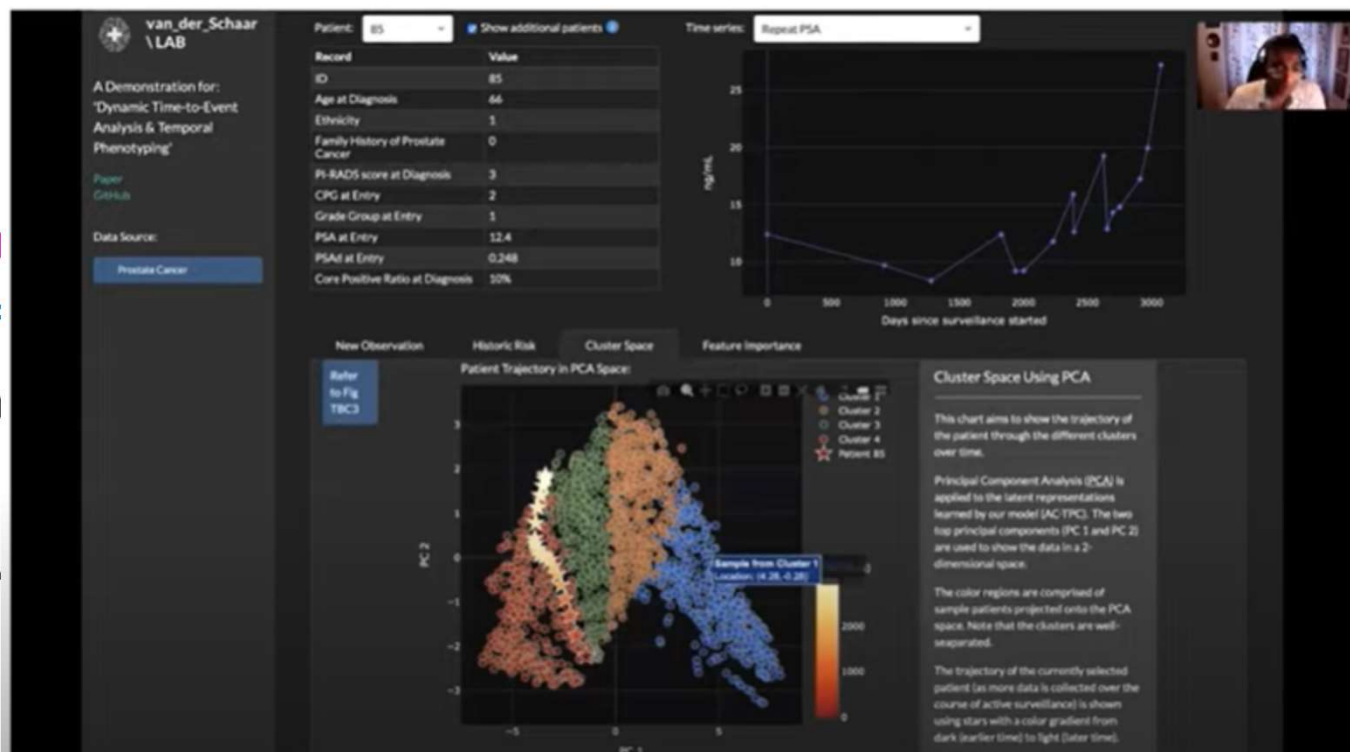
[Crabbé, vdS, ICML 2022]

- Unsupervised learning: e.g. clustering/phenotyping

- Self-supervised

## Desiderata

- ✓ Both f
- ✓ Under
- ✓ Work
- ✓ Work v



[Crabbé, vdS, ICML 2022]

[Crabbé, vdS, ICML 2022]

# Four types of interpretability

1. Feature-based interpretability
2. Example-based interpretability



van\_der\_Schaar  
\ LAB

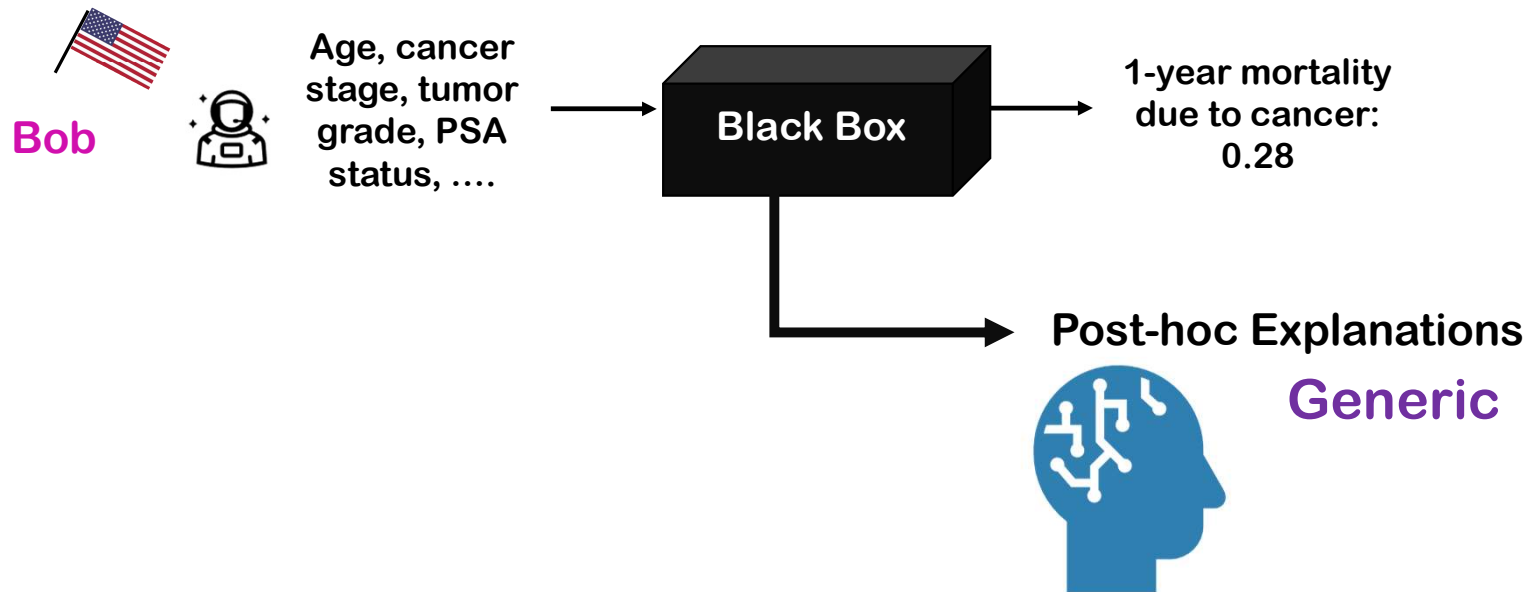
[vanderschaar-lab.com](http://vanderschaar-lab.com)



UNIVERSITY OF  
CAMBRIDGE

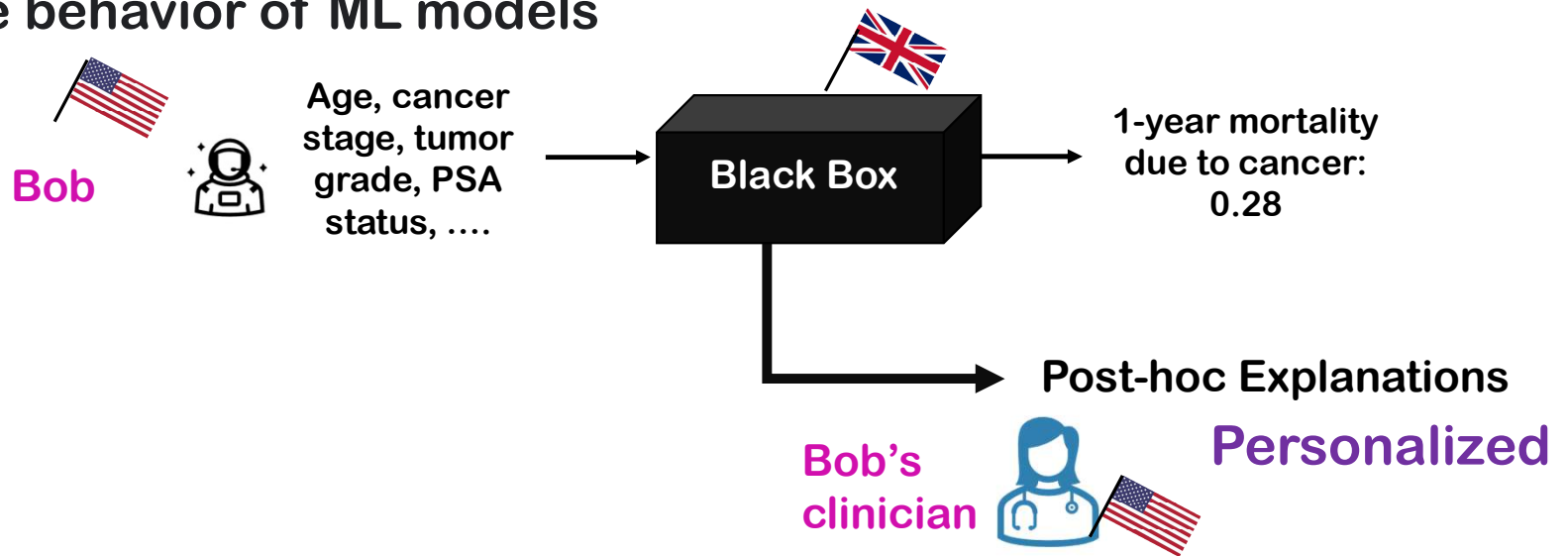
# Example-based explanations

- select particular instances of **the dataset** to explain the behavior of ML models



# Personalized example-based explanations –

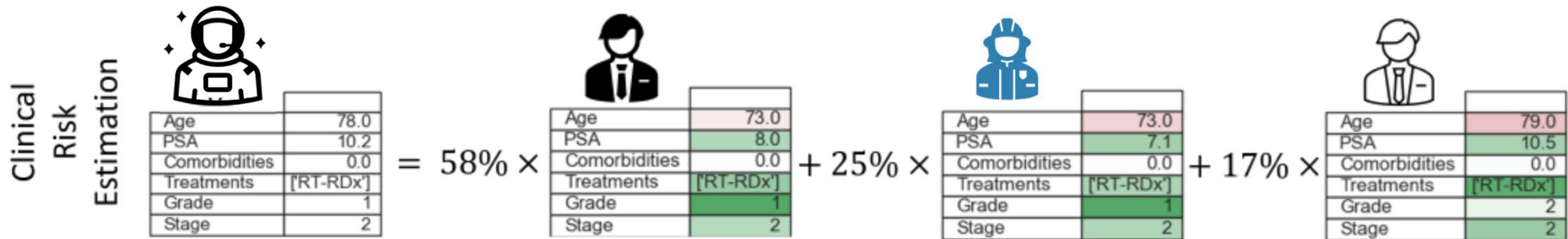
- select particular instances of a **dataset selected by the user (a corpus)** to explain the behavior of ML models



# Desiderata

Personalized explanations with reference to a freely selected set of examples, called the **corpus**

- ✓ Which **corpus examples** explain the prediction issued for a given test example?
- ✓ What **features** of these corpus examples are relevant for the model to relate them to the test example?



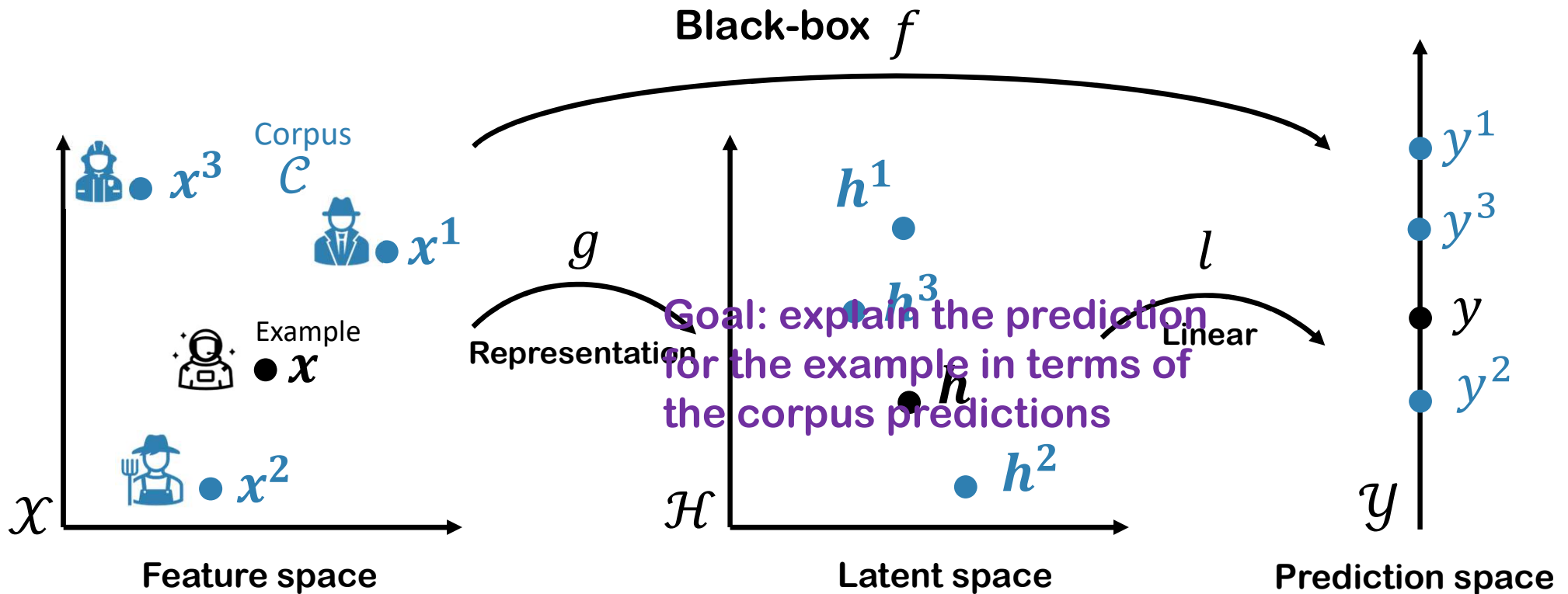
## Our solution: SimplEx [Crabbe, Qian, Imrie, vdS, NeurIPS 2021]

- ✓ **SimplEx** – able to reconstruct the test latent representation as a mixture of corpus latent representations
- ✓ Novel approach (**Integrated Jacobian**) allows SimplEx to make explicit the contribution of each corpus feature in the mixture
  - ✓ Bridge between feature importance & example-based explanations
- ✓ SimplEx gives the user freedom to **choose** the corpus of examples to explain model predictions in a **user-centric** way
- ✓ SimplEx provides user-centric explanations for any ML methods on **diverse data** (tabular, imaging, time-series, multi-modal)

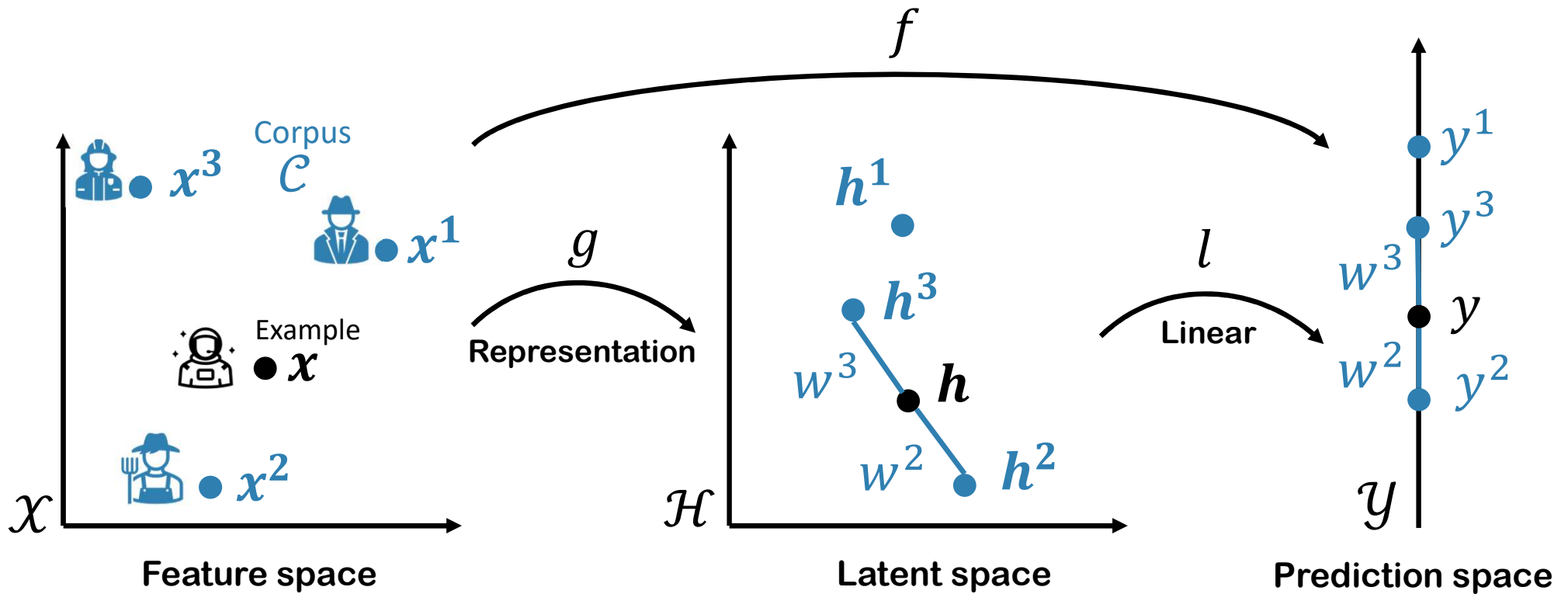




# SimplEx: Problem set-up



# SimplEx: Key idea



# Corpus Decomposition

- Find the best corpus decomposition of the example

$$\hat{\mathbf{h}} = \arg \min \|\mathbf{h} - \tilde{\mathbf{h}}\|_{\mathcal{H}} \quad s.t. \quad \tilde{\mathbf{h}} \in \mathcal{CH}(\mathcal{C})$$



# How to transfer corpus explanations in the input space?

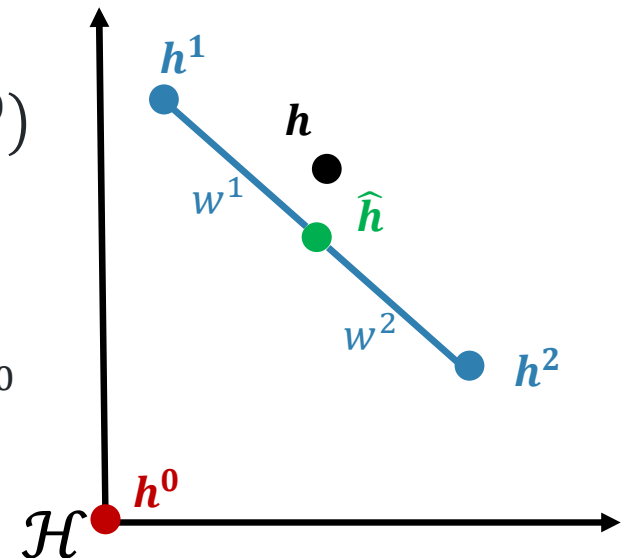
## Idea:

fix a baseline input  $x^0$  with representation  $h^0 = g(x^0)$

$$h - h^0 \approx \sum_{c=1}^C w^c \underline{(h^c - h^0)}$$

Compare each corpus member  $h^c$  to the baseline  $h^0$

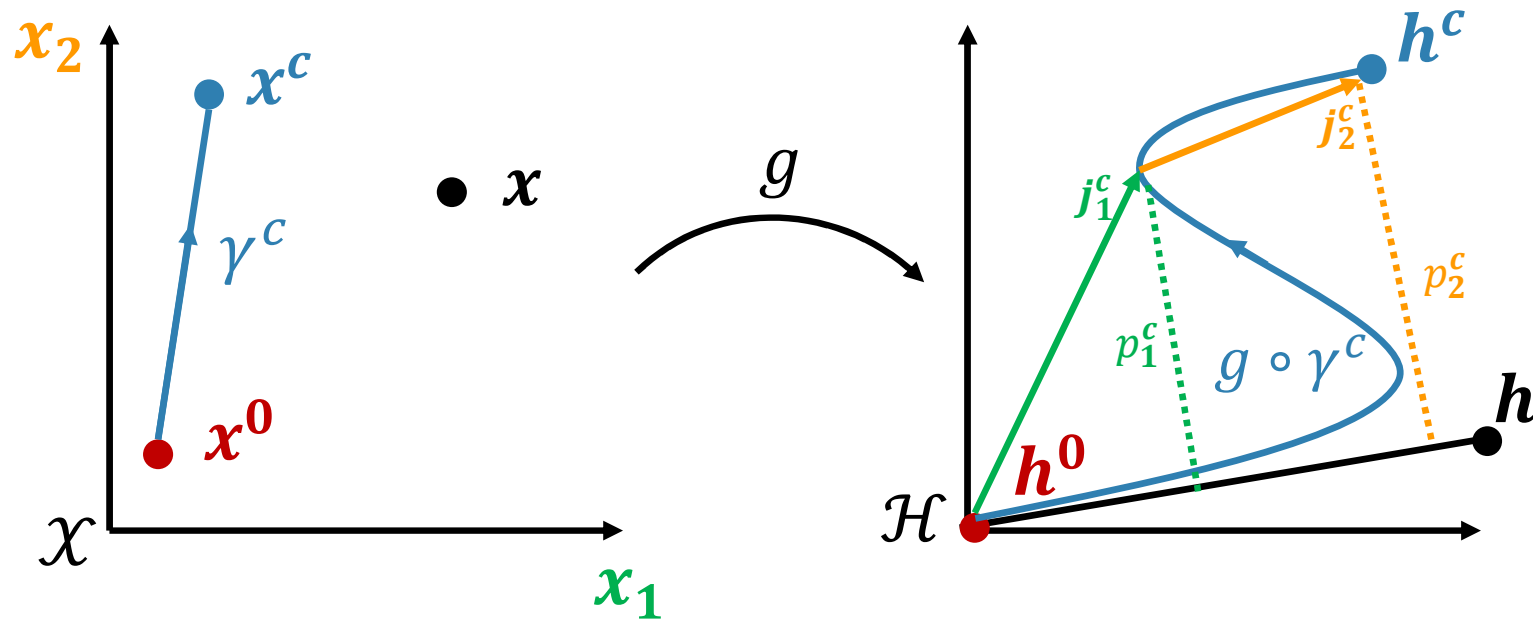
Understand total shift in latent space in terms of **individual contributions** from each corpus member



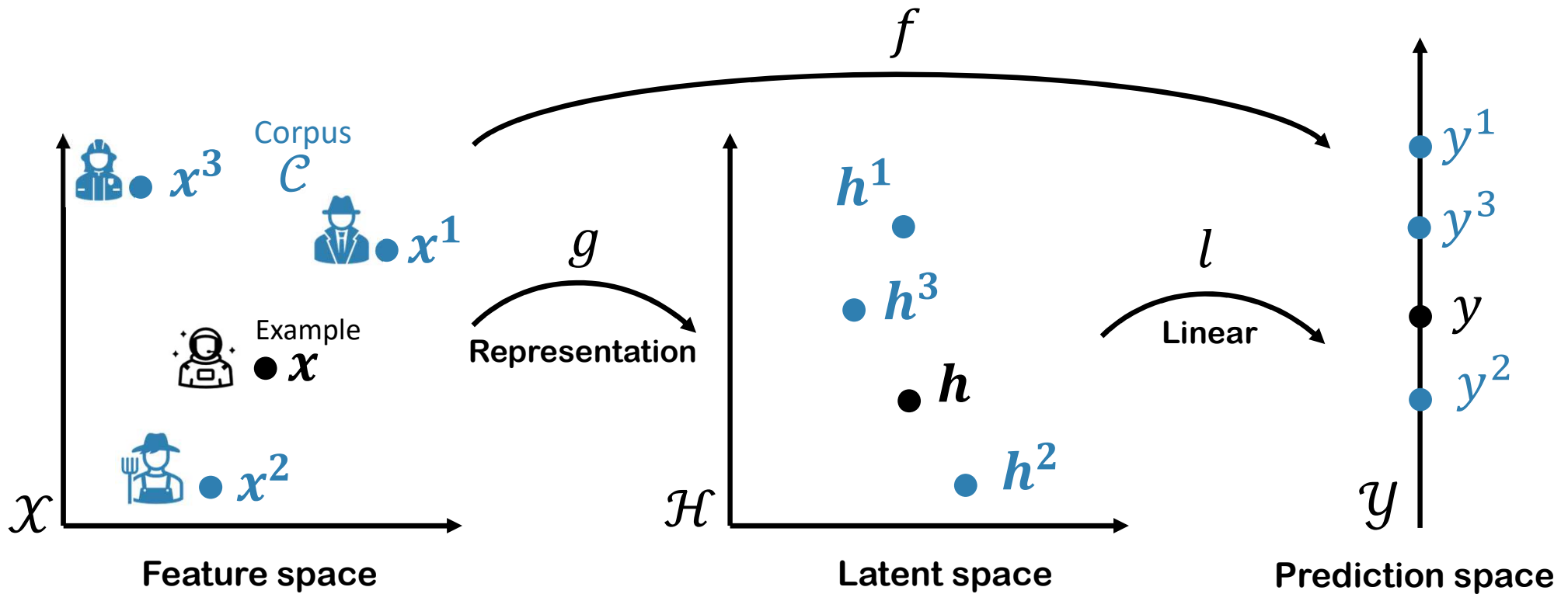
# Integrated Jacobian & Projection

$$j_i^c = \int_0^1 \frac{\partial g \circ \gamma^c}{\partial x_i}(t) dt$$

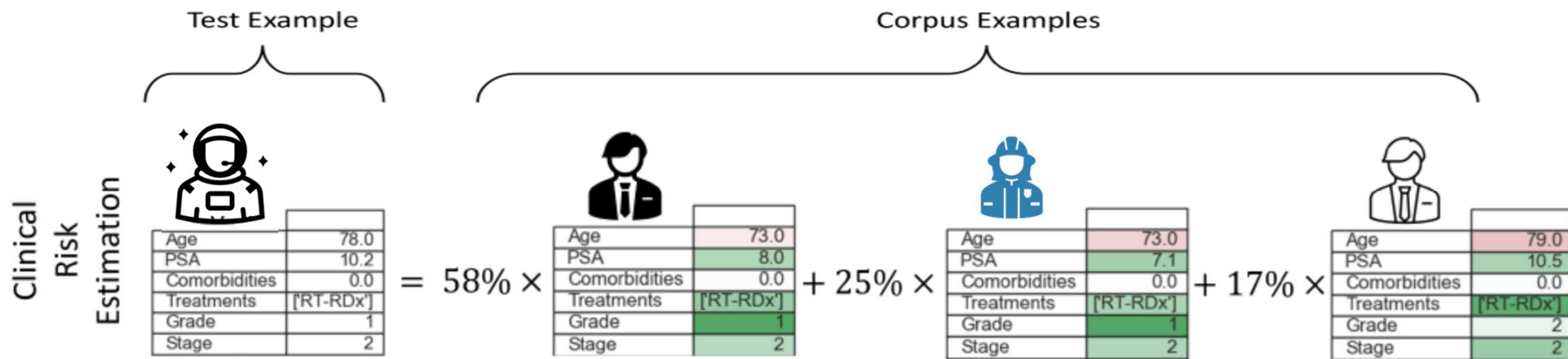
$$p_i^c = \frac{\langle h - h^0, j_i^c \rangle}{\langle h - h^0, h - h^0 \rangle}$$



# SimplEx: Feature sensitivity analysis



# SimplEx Explanations: Going beyond current interpretability



Expanding the picture: SimplEx **unifies** example and feature-based explanations

Enhancing the picture: SimplEx **captures insights** from the model's **latent space**



van\_der\_Schaar  
\ LAB

[vanderschaar-lab.com](http://vanderschaar-lab.com)



UNIVERSITY OF  
CAMBRIDGE

# Four types of interpretability

1. Feature-based interpretability
2. Example-based interpretability
3. Concept-based interpretability



van\_der\_Schaar  
\ LAB

[vanderschaar-lab.com](http://vanderschaar-lab.com)





# What do we mean by *concept*?

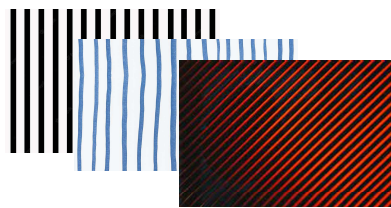
A concept **is**

Defined by the user with concept positive and negative examples

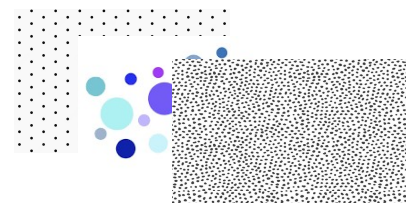
A binary human annotation on the examples fed to ML models

Deducible from the ML model input features

## Stripe Concept



Concept Positives



Concept Negatives

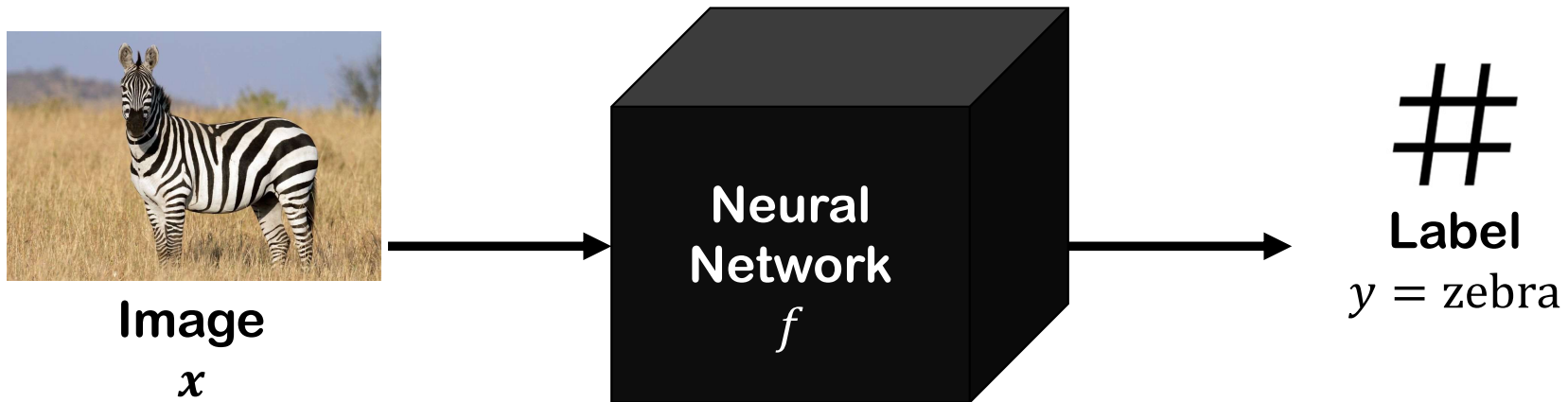


van\_der\_Schaar  
\ LAB

[vanderschaar-lab.com](http://vanderschaar-lab.com)



# Concept-Based Explainability



Is the prediction sensitive to the stripe concept?

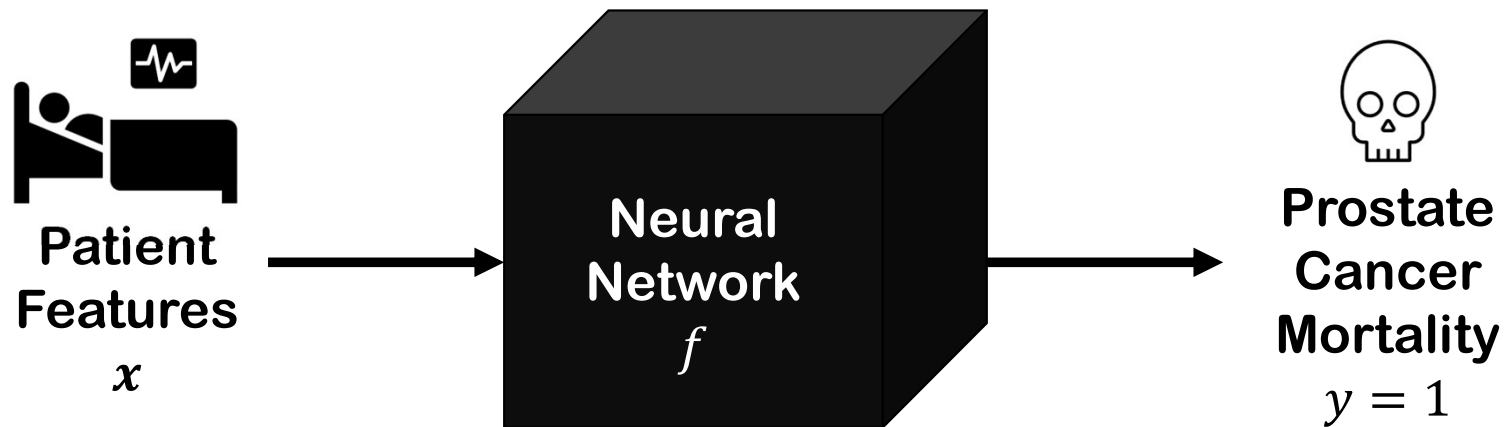


van\_der\_Schaar  
\ LAB

[vanderschaar-lab.com](http://vanderschaar-lab.com)



# Concept-Based Explainability



Is the prediction sensitive to the prostate cancer grading system?

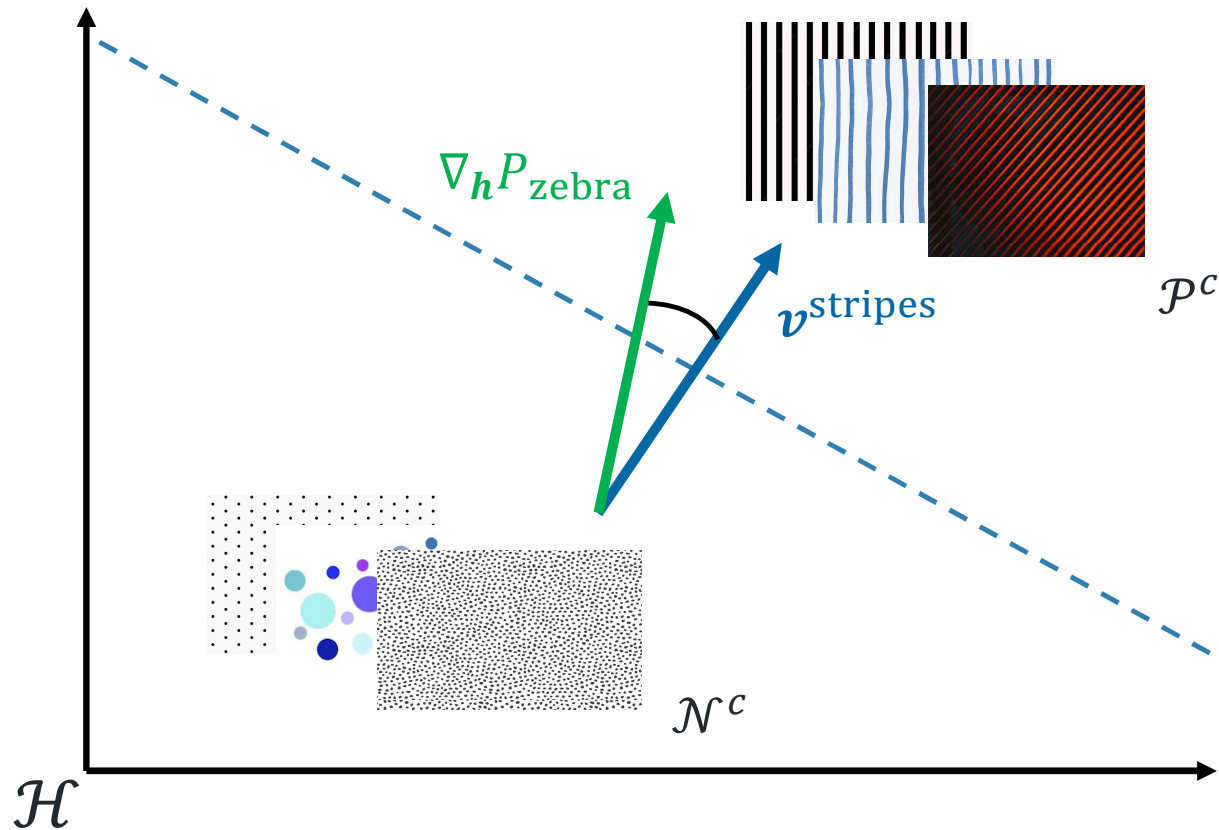


van\_der\_Schaar  
\ LAB

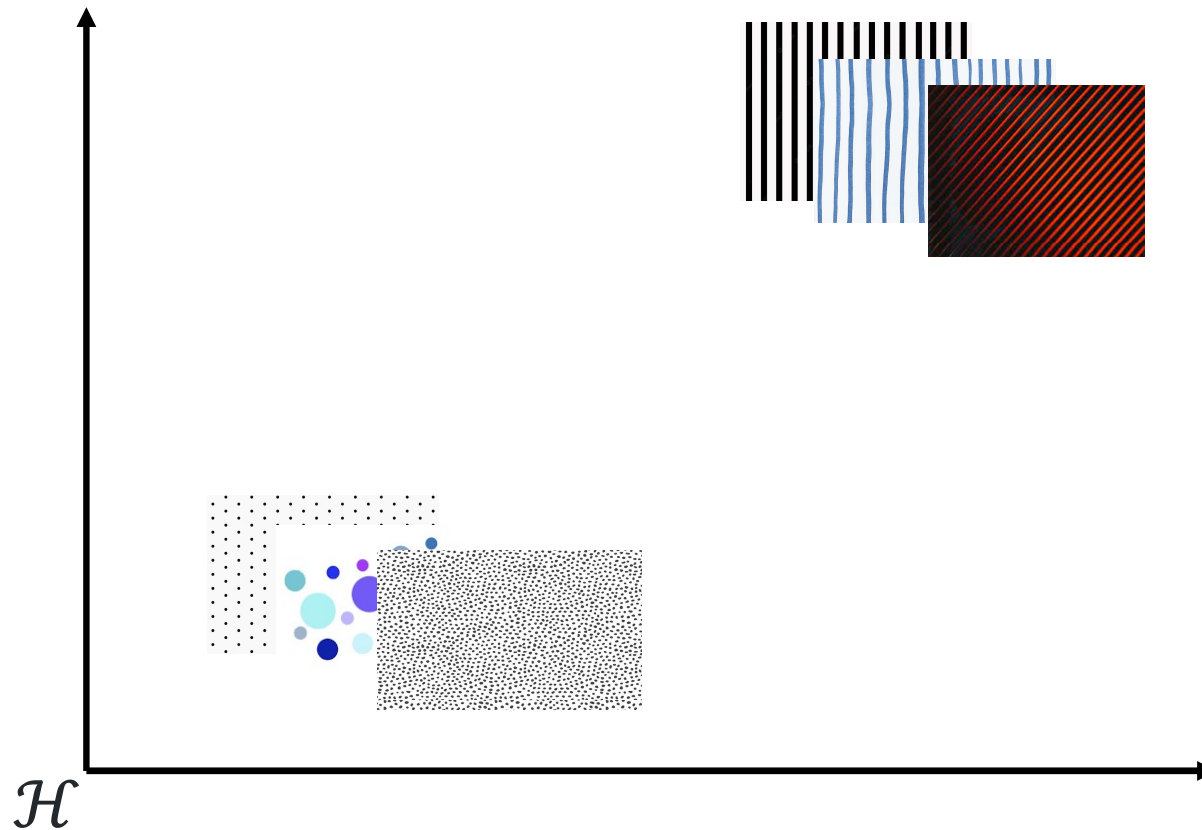
[vanderschaar-lab.com](http://vanderschaar-lab.com)



# Concept Activation Vectors (Kim et al, 2017)

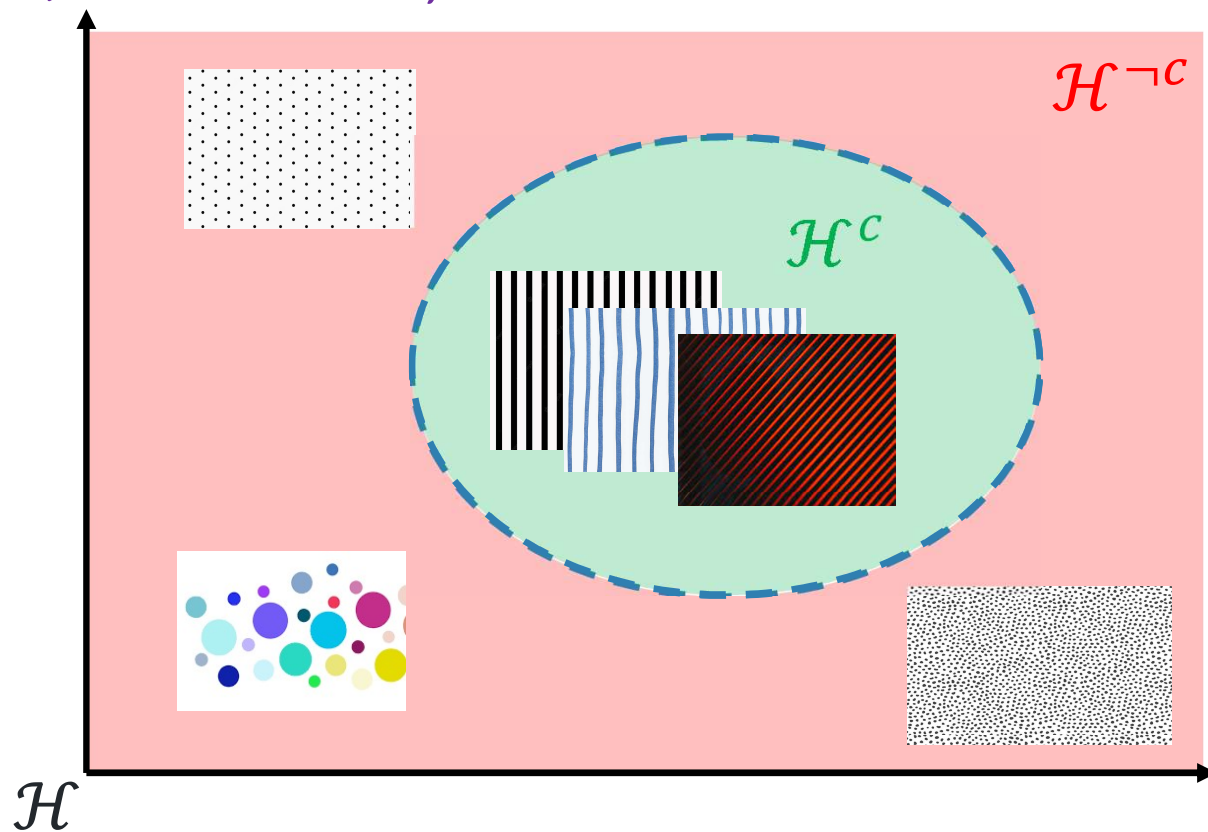


# Concept Activation Vectors (Kim et al, 2017)



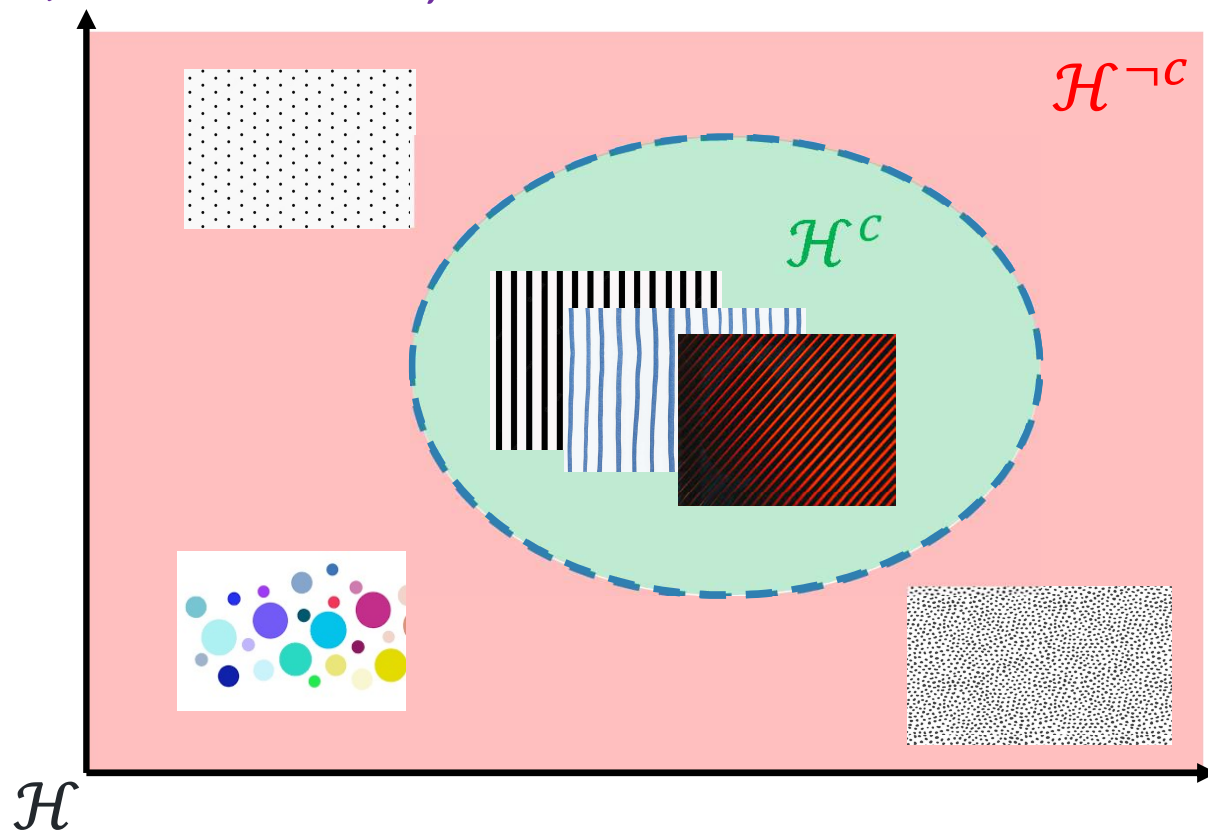
# Concept Activation Regions

(Crabbe, vdS, NeurIPS 2022)



# Concept Activation Regions

(Crabbe, vdS, NeurIPS 2022)



# CAR Formalism

 **Idea.** Borrow the smoothness assumption from semi-supervised learning



# CAR Formalism

 **Idea.** Borrow the smoothness assumption from semi-supervised learning

 A concept  $c$  is well encoded in  $\mathcal{H}$  if we can split  $\mathcal{H} = \mathcal{H}^c \sqcup \mathcal{H}^{-c}$ , where

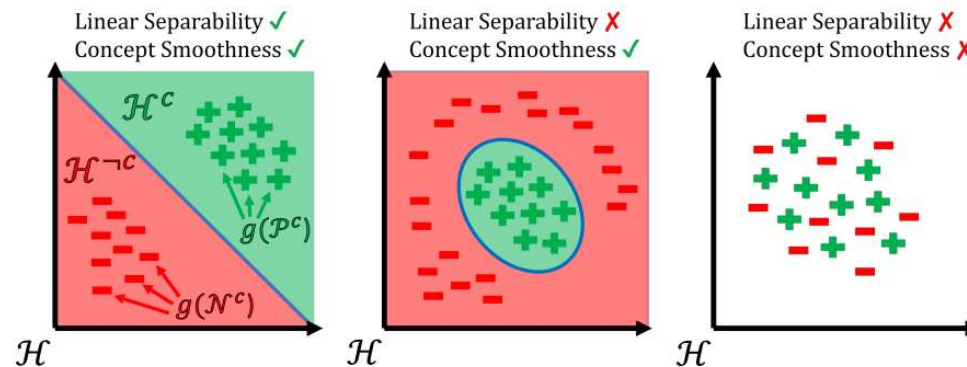
1. The CAR  $\mathcal{H}^c$  mostly overlaps with positives  $\mathcal{P}^c$
2. The region  $\mathcal{H}^{-c}$  mostly overlaps with negatives  $\mathcal{N}^c$
3. If two  $h_1, h_2 \in \mathcal{H}$  are close and in a high-density region, then  $h_1, h_2 \in \mathcal{H}^c$  xor  $h_1, h_2 \in \mathcal{H}^{-c}$

# CAR Formalism

💡 Idea. Borrow the smoothness assumption from semi-supervised learning

📋 A concept  $c$  is well encoded in  $\mathcal{H}$  if we can split  $\mathcal{H} = \mathcal{H}^c \sqcup \mathcal{H}^{-c}$ , where

1. The CAR  $\mathcal{H}^c$  mostly overlaps with positives  $\mathcal{P}^c$
2. The region  $\mathcal{H}^{-c}$  mostly overlaps with negatives  $\mathcal{N}^c$
3. If two  $h_1, h_2 \in \mathcal{H}$  are close and in a high-density region, then  $h_1, h_2 \in \mathcal{H}^c$  xor  $h_1, h_2 \in \mathcal{H}^{-c}$



# CAR Formalism

**Concept Density.** Define a signed density to measure the presence of a concept

$$\rho^c(\mathbf{h}) = \sum_{\mathbf{h}' \in \mathcal{g}(\mathcal{P}^c)} \kappa[\mathbf{h}, \mathbf{h}'] - \sum_{\mathbf{h}' \in \mathcal{g}(\mathcal{N}^c)} \kappa[\mathbf{h}, \mathbf{h}']$$

**Concept Activation Region.** Use concept density with SVMs to infer the CAR  $\mathcal{H}^c$

$$\mathcal{H}^c = (s_{\kappa}^c)^{-1}(1)$$

# CAR Formalism

**Concept Density.** Define a signed density to measure the presence of a concept

$$\rho^c(\mathbf{h}) = \sum_{\mathbf{h}' \in \mathcal{g}(\mathcal{P}^c)} \kappa[\mathbf{h}, \mathbf{h}'] - \sum_{\mathbf{h}' \in \mathcal{g}(\mathcal{N}^c)} \kappa[\mathbf{h}, \mathbf{h}']$$

**Concept Activation Region.** Use concept density with SVMs to infer the CAR  $\mathcal{H}^c$

$$\mathcal{H}^c = (s_{\kappa}^c)^{-1}(1)$$

**Global Explanation.** Measure the relationship between class  $k$  and concept  $c$  with score

$$\text{TCAR}_k^c \equiv \frac{|\mathcal{g}(\mathcal{D}_k) \cap \mathcal{H}^c|}{|\mathcal{D}_k|}$$

# CAR Formalism

**Concept Density.** Define a signed density to measure the presence of a concept

$$\rho^c(\mathbf{h}) = \sum_{\mathbf{h}' \in \mathcal{g}(\mathcal{P}^c)} \kappa[\mathbf{h}, \mathbf{h}'] - \sum_{\mathbf{h}' \in \mathcal{g}(\mathcal{N}^c)} \kappa[\mathbf{h}, \mathbf{h}']$$

**Concept Activation Region.** Use concept density with SVMs to infer the CAR  $\mathcal{H}^c$

$$\mathcal{H}^c = (s_{\kappa}^c)^{-1}(1)$$

**Global Explanation.** Measure the relationship between class  $k$  and concept  $c$  with score

$$\text{TCAR}_k^c \equiv \frac{|\mathcal{g}(\mathcal{D}_k) \cap \mathcal{H}^c|}{|\mathcal{D}_k|}$$

**Feature Importance.** Use any attribution method  $a$  to assign concept importance to features

$$\text{Importance}(x_i) \text{ for } c \equiv a_i(\rho^c \circ \mathbf{g}, \mathbf{x})$$

# CAR Advantages

What do we get by allowing  $\mathcal{H}^c$  and  $\mathcal{H}^{\neg c}$  to be nonlinearly separable?

**More precision.** CAR classifiers better capture how concepts are spread in  $\mathcal{H}$

**Better agreement with humans.** TCAR scores better correlate with human annotations

**Consistent feature importance.** CAR feature importance captures concept associations

# CAR Applications

- Doctors use 5 grades (5 concepts) to determine the likelihood of prostate cancer spreading

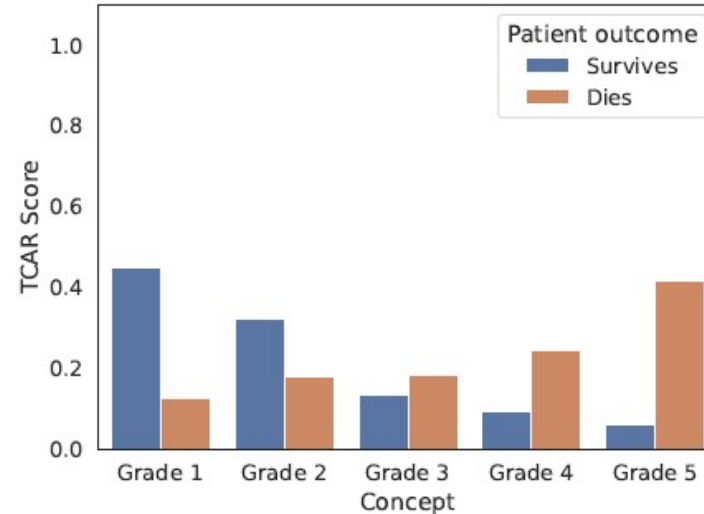
# CAR Applications

- Doctors use 5 grades (5 concepts) to determine the likelihood of prostate cancer spreading
- DNNs implicitly encode prostate grading system (CAR classifiers with  $> 90\%$  ACC)
- In DNNs representations, higher grade is associated with higher mortality



# CAR Applications

- Doctors use 5 grades (5 concepts) to determine the likelihood of prostate cancer spreading
- DNNs implicitly encode prostate grading system (CAR classifiers with  $> 90\%$  ACC)
- In DNNs representations, higher grade is associated with higher mortality



# CAR – Other advantages not covered in this talk

- CAR explanations are invariant to latent isometries
- CAR explanations are robust to adversarial perturbations and background shifts
- CAR explanations can be used to understand abstract concepts discovered
- CAR explanations can be used with a wide variety of modalities (images, time series, tabular)

# Four types of interpretability

1. Feature-based interpretability
2. Example-based interpretability
3. Concept-based interpretability
4. Discovering governing laws - Explicit-functions



van\_der\_Schaar  
\ LAB

[vanderschaar-lab.com](http://vanderschaar-lab.com)



UNIVERSITY OF  
CAMBRIDGE

# Discover the governing models of medicine

- Discover powerful models!

- Why?

Models are needed to

- ✓ *understand* variables, relationships, components
- ✓ *experiment*
- ✓ *act*

We need to go beyond feature & example interpretability

# Discovery of governing models using ML

**Our focus: governing equations –  
compact and closed-form equations**

**Benefits:**

**Concise**

**Generalizable**

**Amenable to further analysis (e.g., identifying stable equilibria)**

**Transparent**

**Interpretable to human experts**



# Clinical Risk Prediction

[Alaa, Gurdasani, Harris, Rashbass & vdS, Nature MI, 2021]

## Example: Predicting breast cancer risk survival (5 years)



Nearly **1 million** patients involved in the analysis.



**NCRAS**  
> 390,000

# Clinical Risk Prediction

[Alaa, Gurdasani, Harris, Rashbass & vdS, Nature MI, 2021]

Example: Predicting breast cancer risk survival (5 years)

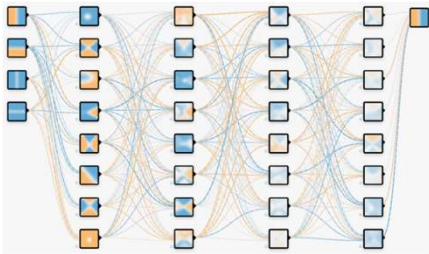


## Metamodeling

Method	AUC-ROC
PREDICT	$0.75 \pm 0.0033$
AutoPrognosis	$0.84 \pm 0.0032$

# Turning black boxes into white boxes using symbolic metamodels [Alaa & vdS, NeurIPS 2019] [Crabbe, Zhang, vdS, NeurIPS 2020]

Black-box ML model



$f(\mathbf{x})$



Symbolic  
Metamodeling

$$g(\mathbf{x}) = G(\mathbf{x}; \theta^*)$$



Explicit function

$$\alpha_1 X_1 + \alpha_2 X_2^2 + \alpha_3 X_1 X_2$$
$$\alpha_4 X_3^3 + \alpha_5 \log(X_4)$$

$g(\mathbf{x})$

$$\theta^* = \arg \min_{\theta \in \Theta} \ell(f(\mathbf{x}), G(\mathbf{x}; \theta))$$

## Metamodels

Operates on a **trained machine learning** model and outputs a symbolic formula describing the model's prediction surface

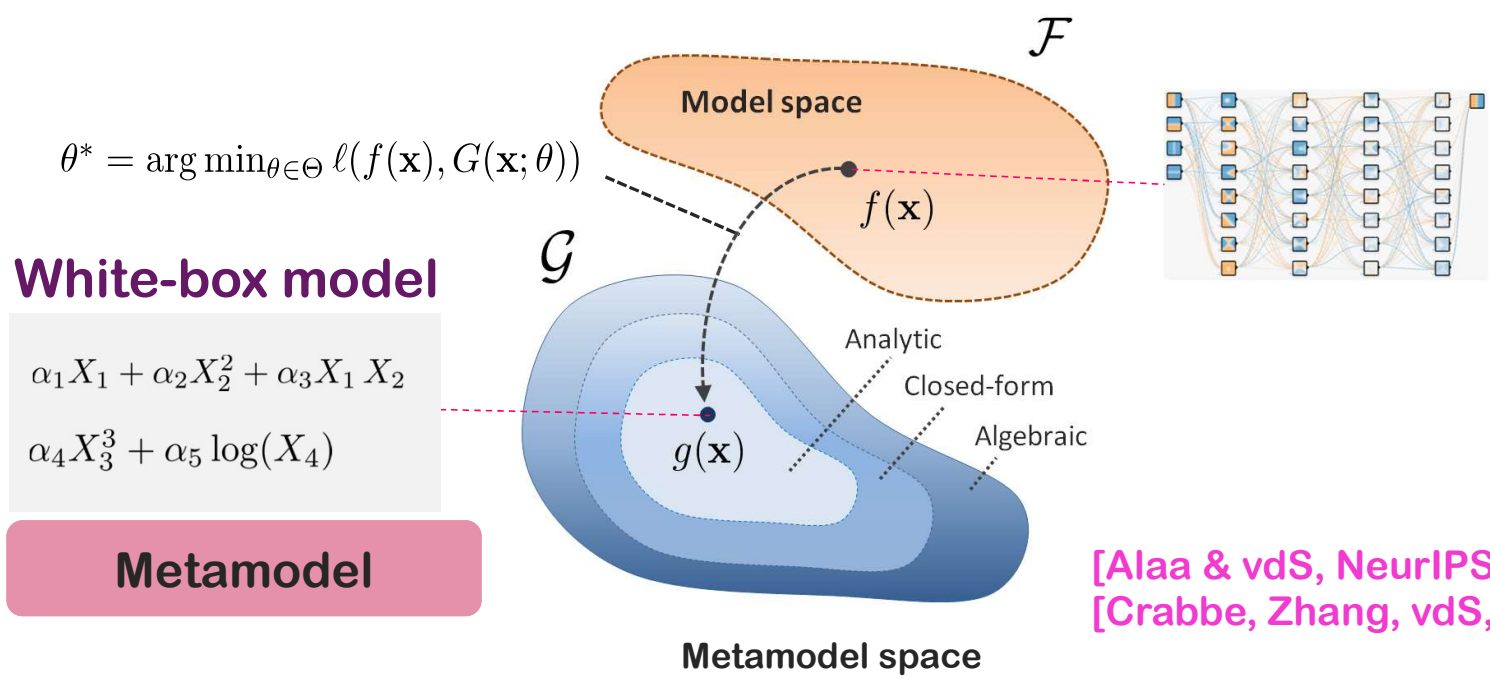


# Building transparent risk equations of black-box ML

Metamodel representation  $g(\mathbf{x}) = G(\mathbf{x}; \theta^*)$

Black-box ML model

Model space (uninterpretable)

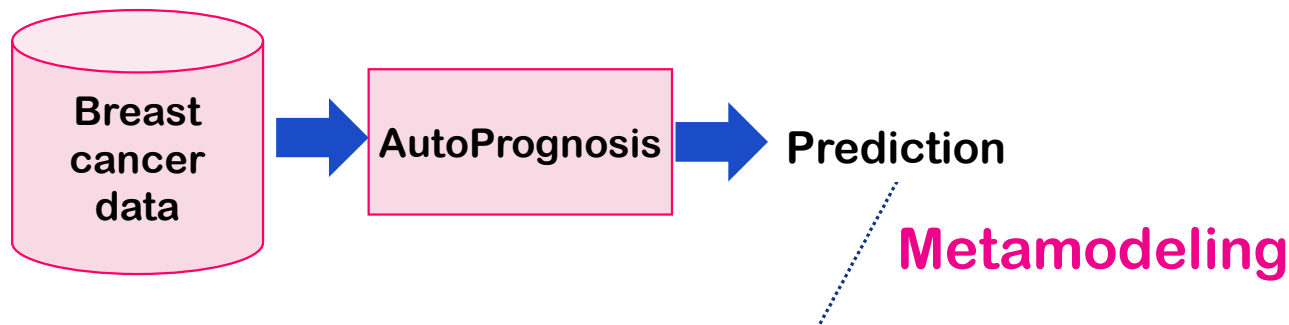


[Alaa & vdS, NeurIPS 2019]  
 [Crabbe, Zhang, vdS, NeurIPS 2020]

# Interpretability using symbolic metamodeling in practice

[Alaa, Gurdasani, Harris, Rashbass & vdS, Nature MI, 2021]

## Example: Predicting breast cancer risk survival (5 years)



### Risk equations

$f(\text{Age}, \text{ER}, \text{HER2}, \text{Tumor size}, \text{Grade}, \text{Nodes}, \text{Screening})$

$$\exp\left(\frac{\text{Age}}{5} - \log\left(\frac{\text{Tumor size}}{100}\right) + \frac{1}{10} \log(\text{Nodes})\right) \times \exp\left(\frac{\text{ER} \cdot \text{Nodes}}{20} + \frac{\text{ER} \cdot \text{Tumor size}}{23}\right)$$



# Interpretability using symbolic metamodeling in practice

[Alaa, Gurdasani, Harris, Rashbass & vdS, Nature MI, 2021]

## Example: Predicting breast cancer risk survival (5 years)



Risk equations

Metamodeling

Method	AUC-ROC
PREDICT	0.75 ± 0.0033
<del>AutoPrognosis</del>	<del>0.84 ± 0.0002</del>
Metamodel	0.83 ± 0.0020

# Illustration

Nancy



Age  
gender  
diabetes  
BMI

ML Model

Risk  
prediction

$g(\mathbf{x})$

$$\alpha_0 \text{ Age} + \alpha_1 \text{ BMI}^2 + \alpha_2 \text{ Age} \cdot \text{BMI} + \alpha_3 \text{ Age} \cdot \text{Gender} \\ \alpha_4 \text{ Gender} \cdot (1 + \alpha_5 \text{ Diabetes}) + \alpha_6 \log(\text{Age} \cdot \text{Diabetes} + 1)$$

Explicit risk  
formulae

Individual-level feature importance

$$\frac{\partial g(\mathbf{x})}{\partial \text{Age}} = \alpha_0 + \alpha_2 \text{ BMI} + \alpha_3 \text{ Gender} + \frac{\alpha_6 \text{ Diabetes}}{\text{Age} + 1}$$



van\_der\_Schaar  
LAB

vanderschaar-lab.com



UNIVERSITY OF  
CAMBRIDGE

# Discovery of governing equations using ML

	Explicit function	Implicit function	Ordinary differential equation	Partial differential equation
Typical form	$y = f(x)$	$f(x, y) = c$	$\frac{dx}{dt} = f(x, t)$	$\frac{\partial u}{\partial t} = f(u, x)$

**Symbolic  
Metamodels**  
[NeurIPS '19, '20]



**D-Code**  
[ICLR '22]



**D-CIPHER**  
[archive]



# Our Resources to go Further



## Our Papers

[vanderschaar-lab.com/interpretable-machine-learning/](https://vanderschaar-lab.com/interpretable-machine-learning/)

## Our Code

[github.com/vanderschaarlab/Interpretability](https://github.com/vanderschaarlab/Interpretability)



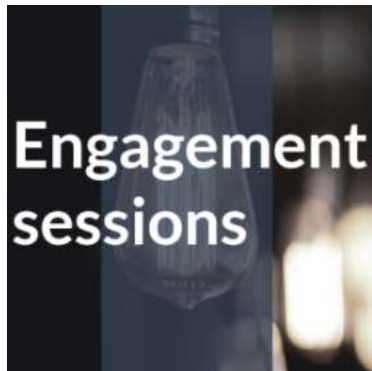
van\_der\_Schaar  
\ LAB

[vanderschaar-lab.com](https://vanderschaar-lab.com)



# Engagement sessions: Inspiration Exchange

Online engagement sessions for  
ML researchers in healthcare;  
themed presentations & Q&A



Inspiration Exchange is a series of engagement sessions aiming to share ideas and discuss topics that will define the future of machine learning in healthcare. These events will target machine learning students, and will emphasize sharing of new ideas and development of new methods, approaches, and techniques.

As a lab, our purpose is to create new and powerful machine learning techniques and methods that can revolutionize healthcare. This doesn't happen in a vacuum. At inception, we are inspired by ideas and discussions; in implementation, we need connections, trust, and partnership to make a real difference.

While you can learn about our work at major conferences in machine learning or in our papers, we think it's a better idea to create a community and keep these conversations going. We're also aware that many people—both in healthcare and machine learning—have questions about what we do, and how they can contribute.

For more information about Inspiration Exchange—and to sign up to join in—please have a look at the sections below, and keep checking for new updates.



**Inspiration Exchange**

Themed discussion sessions specifically for machine learning students (particularly masters, PH.D., and post-docs).

We would like to:

- discuss machine learning models and techniques
- share ideas about how machine learning can revolutionize healthcare
- spark new projects and collaborations
- raise awareness about this unique and exciting area of machine learning.

Standard session format:












- presentations by van der Schaar Lab researchers
- Q&A



<https://www.vanderschaar-lab.com/>  
→ Engagement sessions  
→ Inspiration Exchange



[vanderschaar-lab.com](https://www.vanderschaar-lab.com)

 1-10-22	<b>Inspiration Exchange - time series in healthcare</b> van der Schaar Lab
 1-20-26	<b>Inspiration Exchange - quantitative epistemology</b> van der Schaar Lab
 1-9-51	<b>Inspiration exchange - individualized treatment effect inference (2/2)</b> van der Schaar Lab
 1-04-17	<b>Inspiration exchange - individualized treatment effect inference (1/2)</b> van der Schaar Lab
 5-6-18	<b>Inspiration Exchange - application-oriented projects in machine learning for healthcare</b> van der Schaar Lab
 5-7-55	<b>Inspiration Exchange - synthetic data evaluation</b> van der Schaar Lab
 1-01-49	<b>Inspiration Exchange - synthetic data concepts and approaches</b> van der Schaar Lab
 1-01-40	<b>Inspiration Exchange - recent projects in machine learning for healthcare</b> van der Schaar Lab
 4-8-29	<b>Inspiration Exchange - software packages for automated machine learning</b> van der Schaar Lab
 1-12-49	<b>Inspiration Exchange - automated machine learning pipelines</b> van der Schaar Lab
 1-01-29	<b>Inspiration Exchange - introduction to automated machine learning</b> van der Schaar Lab

