~~DALL·E mini~~

Alex Gagliano

Additive Feature Attribution Methods

ML Journal Club, June 22

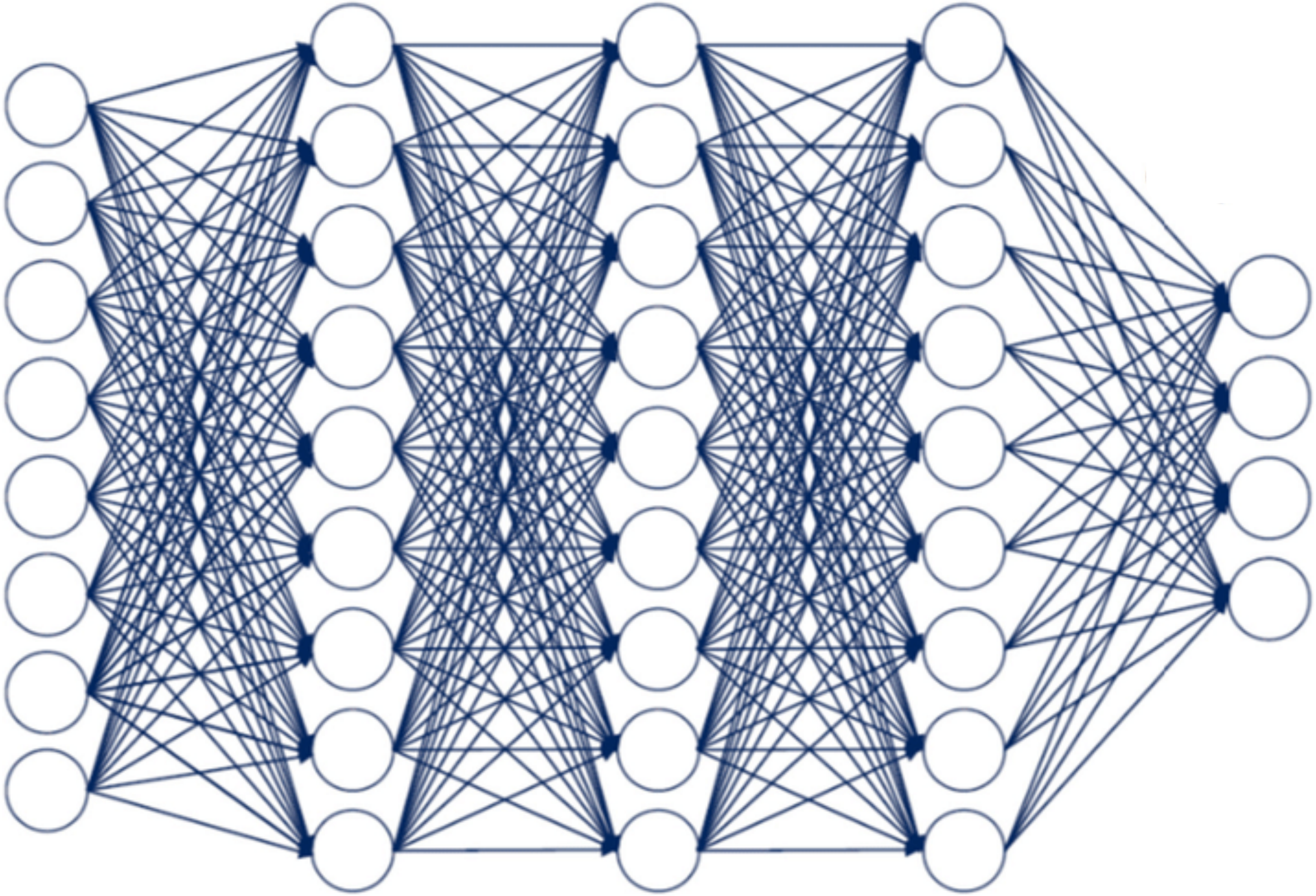
Run

You have a model you want to explain.

Model

$f(x)$

Features



What if we constructed a simpler model connecting inputs to outputs?

Model outputs

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i \; z'_i$$

Binary encoding of included features

What properties should a local linear model satisfy?

1. Local accuracy

Model

$$f(x) = g(z')$$

Features

The function should approximate the model output for the input features.

2. Missingness

$$z_i' = 0 \rightarrow \phi_i = 0$$

An absent feature should not influence the output.

3. Consistency

x : The feature set including feature *i*

y : The feature set excluding feature *i*

$$f_1(x) - f_1(y) > f_2(x) - f_2(y)$$

$$\downarrow$$

$$\phi_i(f_1, x) > \phi_i(f_2, x)$$

If model 1 output increases with feature i more than model 2, its coefficient for feature i in the linear model should be larger.

A *single* function satisfies all properties:

(Lundberg & Lee, 2017)

$$\phi_i(g, x) = \frac{1}{N!} \sum_{R} \left[v(x_R) - v(y_R) \right]$$

All permutations of set x

All permutations of set y

Shapley values: *the average expected marginal contribution of a feature given all combinations*

Shapley 1951, The Value of an n-Person Game

$$\phi_i(g, x) = \frac{1}{N!} \sum_{R} \left[v(x_R) - v(y_R) \right]$$

Shapley values: *the average expected marginal contribution of a single player given all combinations*

SHAP (Shapley Additive exPlanation) Values in Practice

$$\phi_0 + \sum_{i=1}^{M} \phi_i \, z_i'$$

Binary encoding of included features

1. Computationally infeasible to calculate all permutations of all inputs

2. Model must be evaluated relative to some featureless baseline

Article | Published: 10 October 2018

Explainable machine-learning predictions for the prevention of hypoxaemia during surgery

Scott M. Lundberg, Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Ada
David E. Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim & Su-In Lee

Article | Published: 17 January 2020

From local explanations to global understanding with explainable AI for trees

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz,

The Shapley Value in Machine Learning

Benedek Rozemberczki[1], Lauren Watson[2], Péter Bayer[3], Hao-Tsung Yang[2], Oliver Kiss[4], Sebastian Nilsson[1] and Rik Sarkar[2]

Open Access | Published: 02 May 2020

Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions

Raquel Rodríguez-Pérez & Jürgen Bajorath

GPUTreeShap: Massively Parallel Exact Calculation of SHAP Scores for Tree Ensembles

Rory Mitchell[1], Eibe Frank[2], and Geoffrey Holmes[2]

[1]**Nvidia Corporation**

Explaining individual predictions when features are dependent: More accurate approximations to Shapley values ☆

Jullum ✉, Anders Løland ✉

Article

Shapley variable importance c interpretable machine learnin

Yilin Ning [1], Marcus Eng Hock Ong [2, 3, 4], Bibhas Chakraborty [1, 2, 5],
Daniel Shu Wei Ting [1, 7, 8], Roger Vaughan [1, 2], Nan Liu [1, 2, 3, 8, 9, 10]

Learning to Estimate Shapley Values with Vision Transformers

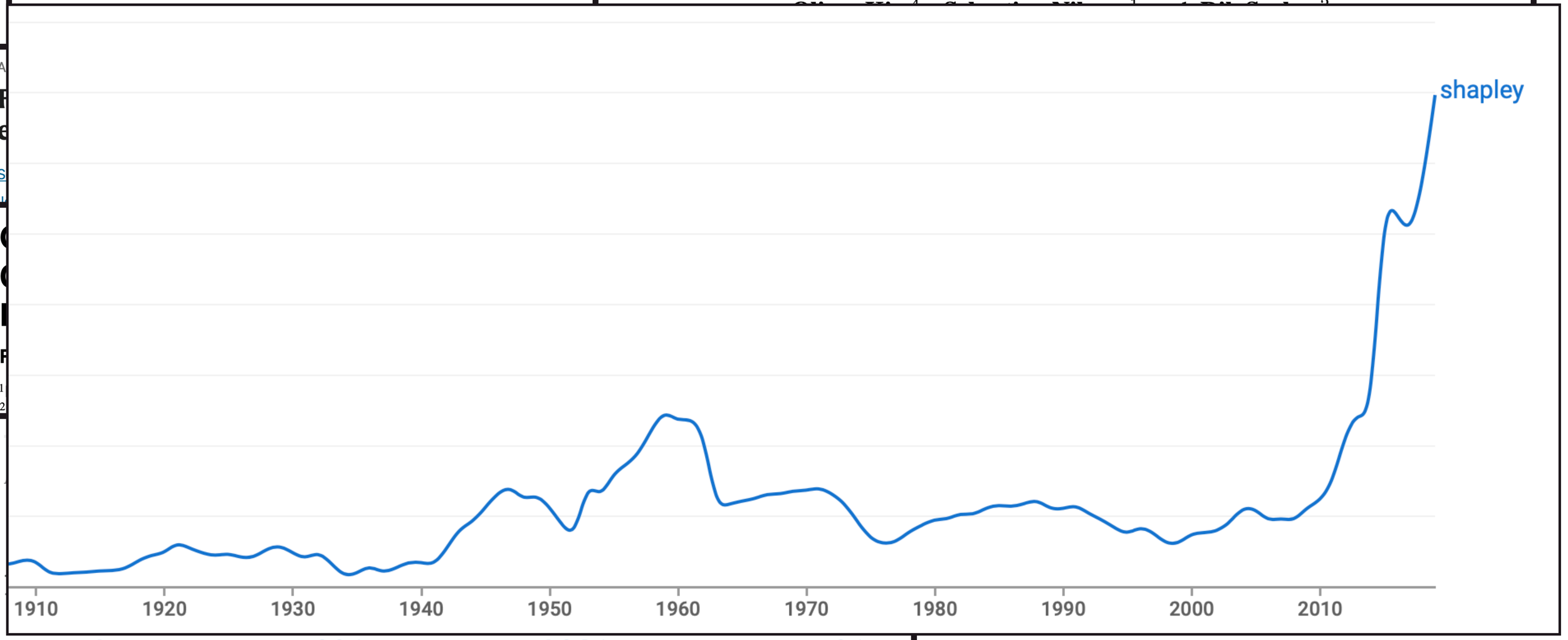
Article | Published: 10 October 2018

Explainable machine-learning predictions for the prevention of hypoxaemia during surgery

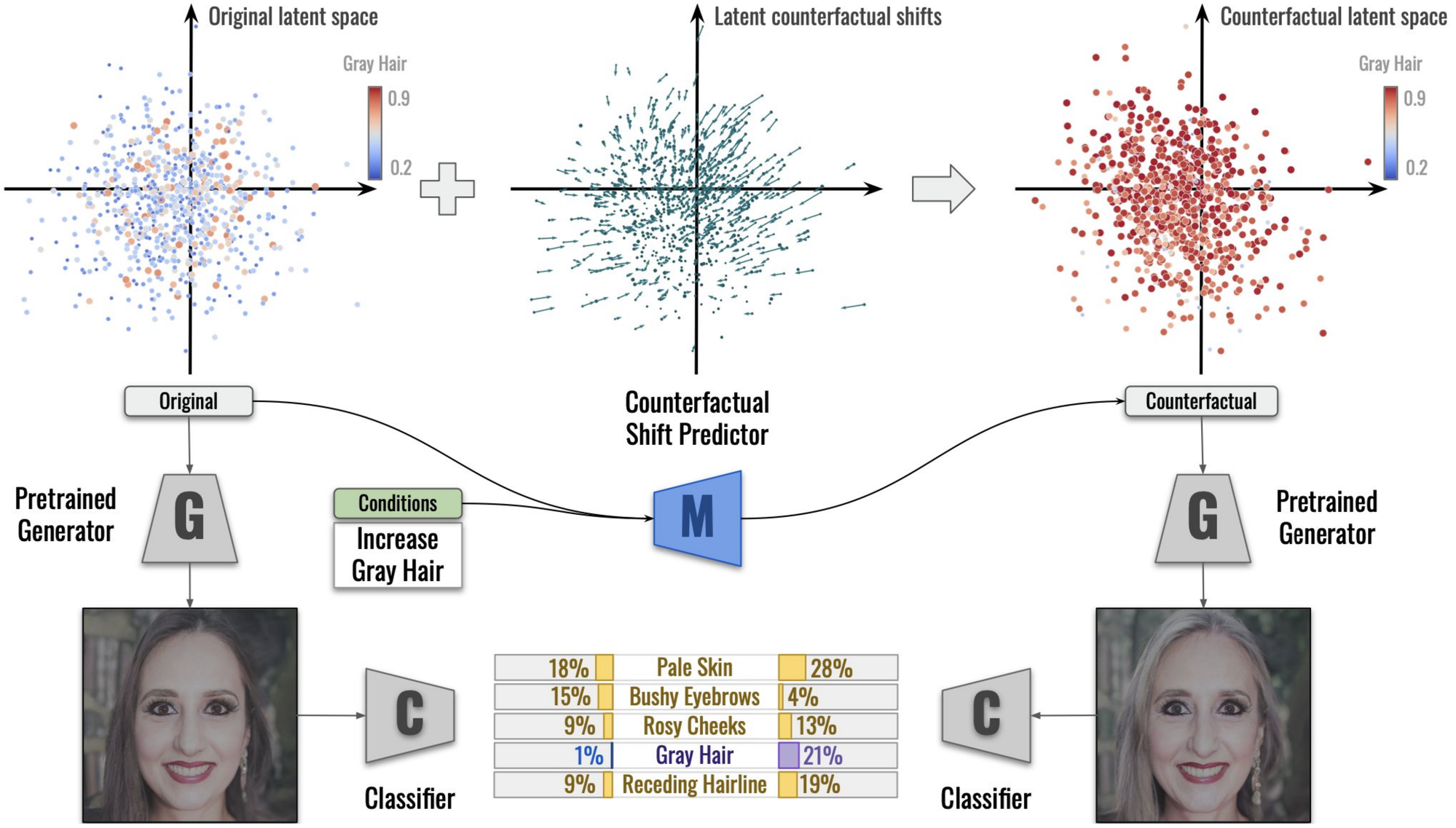
Yilin Ning [1], Marcus Eng Hock Ong [2, 3, 4], Bibhas Chakraborty [1, 2, 5, 6], Benjamin Alan Goldstein [2, 6], Daniel Shu Wei Ting [1, 7, 8], Roger Vaughan [1, 2], Nan Liu [1, 2, 3, 8, 9, 10]

The Shapley Value in Machine Learning

Benedek Rozemberczki[1] , Lauren Watson[2] , Péter Bayer[3] , Hao-Tsung Yang[2] ,

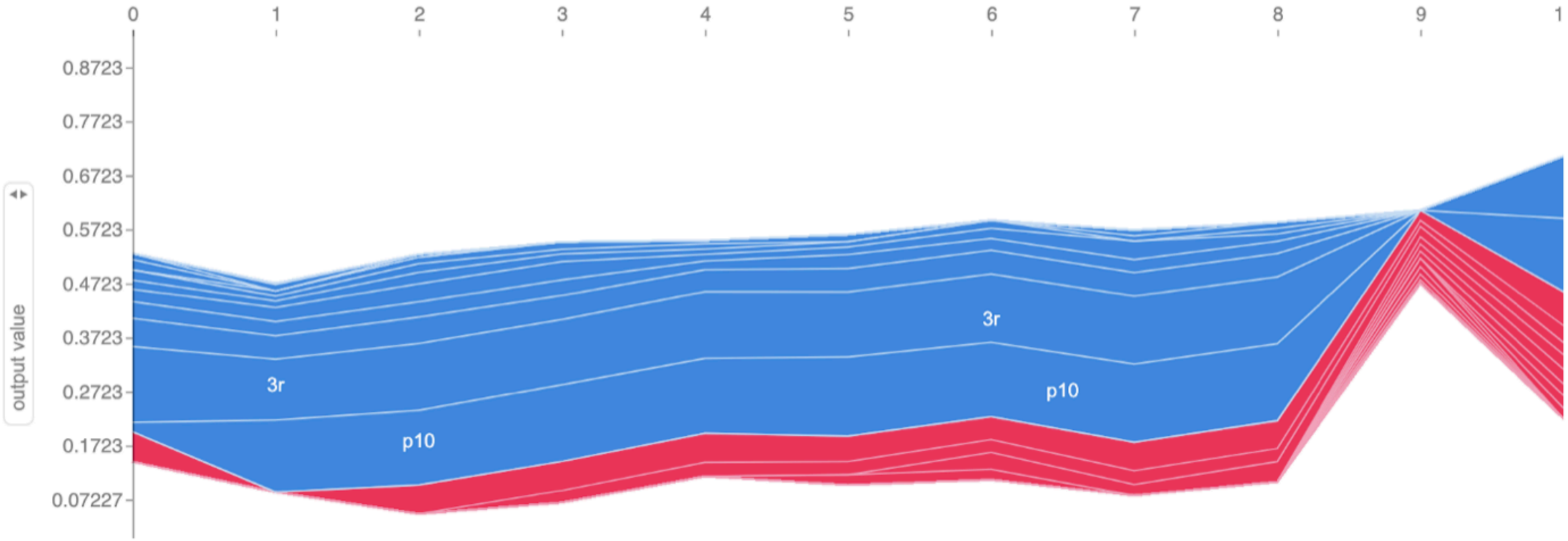


Generative Models Trained on Shapley Values (Yesterday!)



https://arxiv.org/abs/2206.07087

My Goal: Hack Shapley Values for Supernova Classification

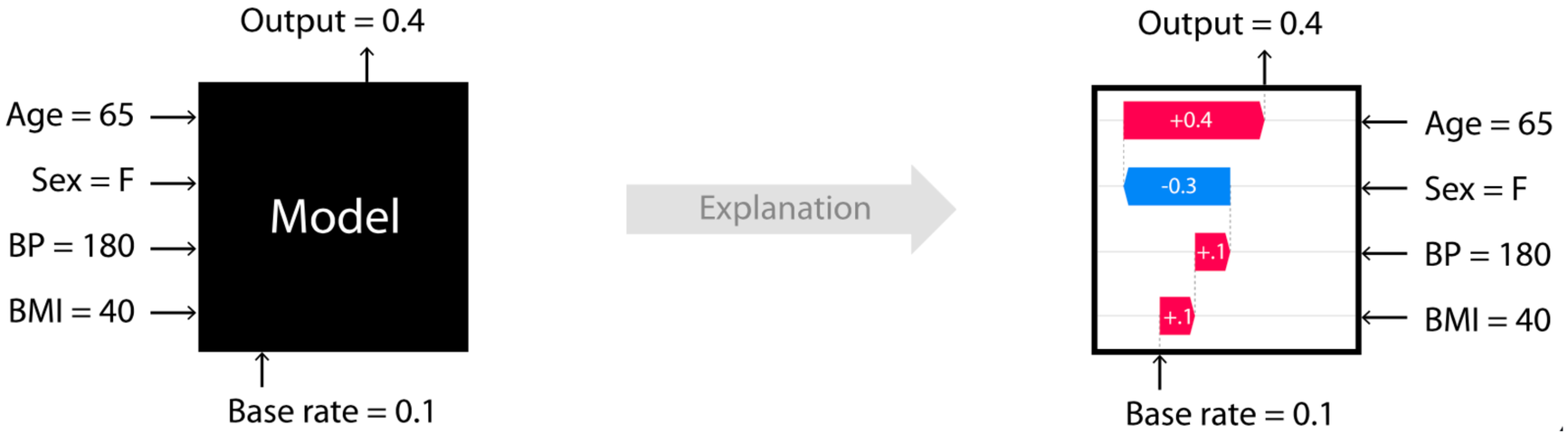


Feature importance along an ALS patient's time series. The border between the red shade (output increasing features) and the blue shade (output decreasing features) represents the model's output for each timestamp.

https://towardsdatascience.com/interpreting-recurrent-neural-networks-on-multivariate-time-series-ebec0edb8f5a

SHAP Walkthroughs

SHAP



https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

https://shap.readthedocs.io/en/latest/index.html