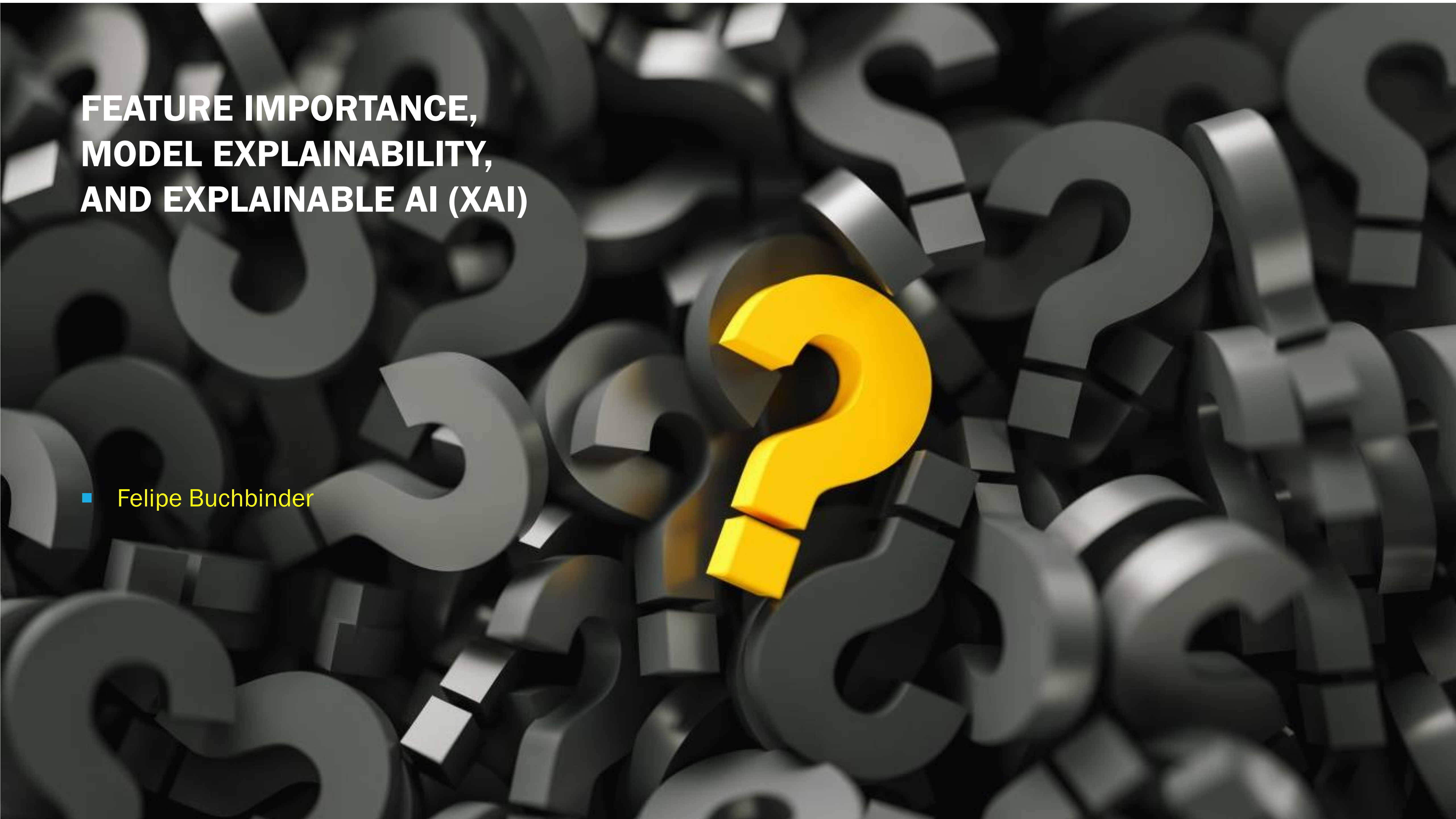
FEATURE IMPORTANCE, MODEL EXPLAINABILITY, AND EXPLAINABLE AI (XAI)

- Felipe Buchbinder



AI IS BEING USED FOR SOME PRETTY IMPORTANT STUFF…

Hiring

Sentencing

Student loans

Health insurance premiums

HOW COULD IT POSSIBLY GO WRONG?

ETHICAL IMPLICATIONS

AI IS BEING USED FOR SOME PRETTY IMPORTANT STUFF...

Hiring

Sentencing

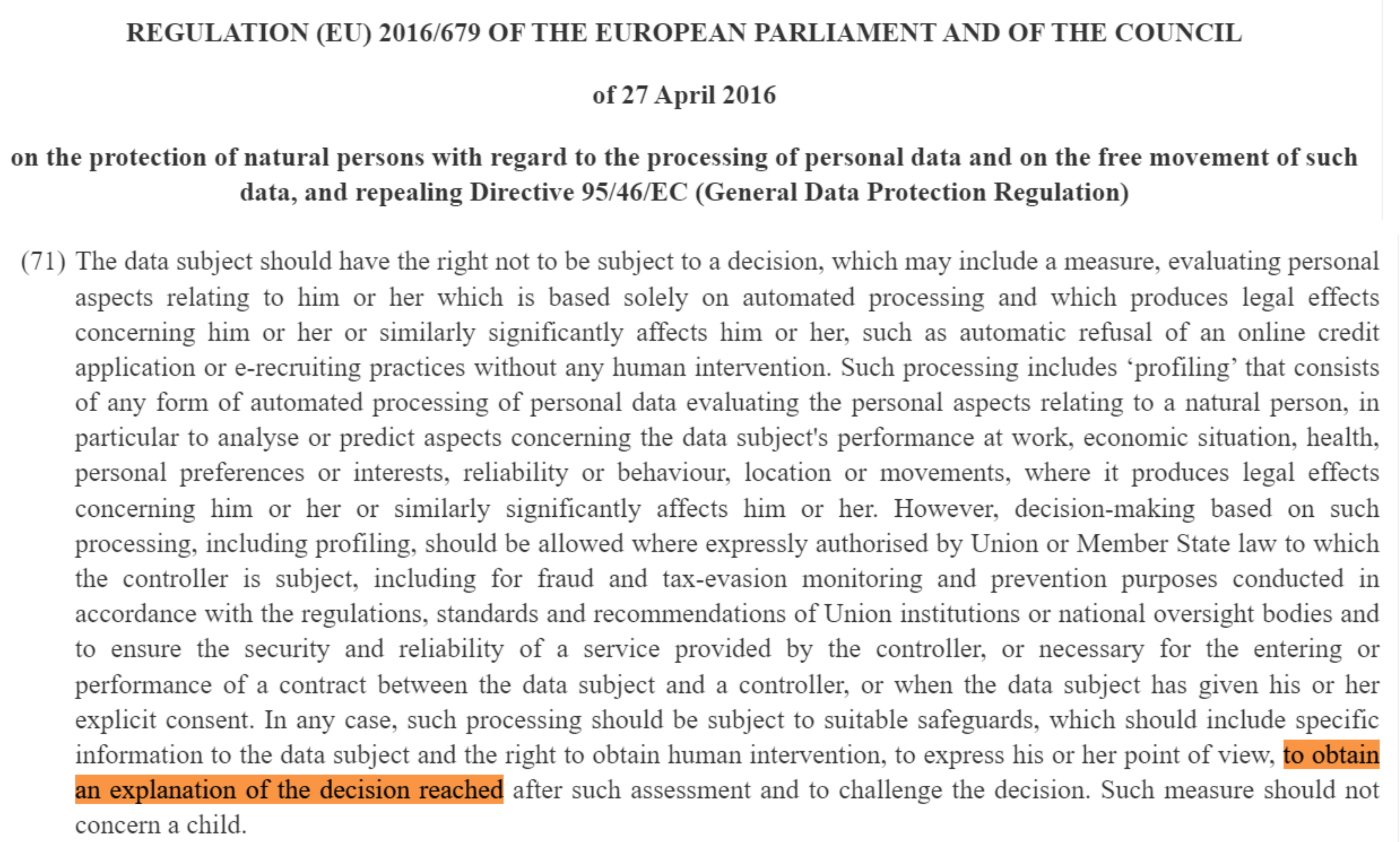
Student loans

Health insurance premiums

STUDENT LOAN

EUROPEAN UNION’S GENERAL DATA PROTECTION AND REGULATION

(ENACTED 2016 AND EFFECTIVE AS OF 2018)



REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

(71) The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention. Such processing includes ‘profiling’ that consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her. However, decision-making based on such processing, including profiling, should be allowed where expressly authorised by Union or Member State law to which the controller is subject, including for fraud and tax-evasion monitoring and prevention purposes conducted in accordance with the regulations, standards and recommendations of Union institutions or national oversight bodies and to ensure the security and reliability of a service provided by the controller, or necessary for the entering or performance of a contract between the data subject and a controller, or when the data subject has given his or her explicit consent. In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision. Such measure should not concern a child.

OTHER NOTABLE PIECES OF LEGISLATURE

- Brazil:
  - Lei Geral de Proteção de Dados Individuais (LGPD) : Lei 13.709, 14th August 2018
  - No mention of rights to explanation
- USA:
  - Equal Credit Opportunity Act (Regulation B of the Code of Federal Regulations), Title 12, Chapter X, Part 1002, §1002.9
  - Rights to explanation in the case of denied credit.
  - No legislation exists for explainability in general
- France:
  - *Loi* pour *une République numérique,* 2016
  - where there is "a decision taken on the basis of an algorithmic treatment", the rules that define that treatment and its “principal characteristics” must be communicated to the citizen upon request

INHERENTLY INTERPRETABLE MODELS

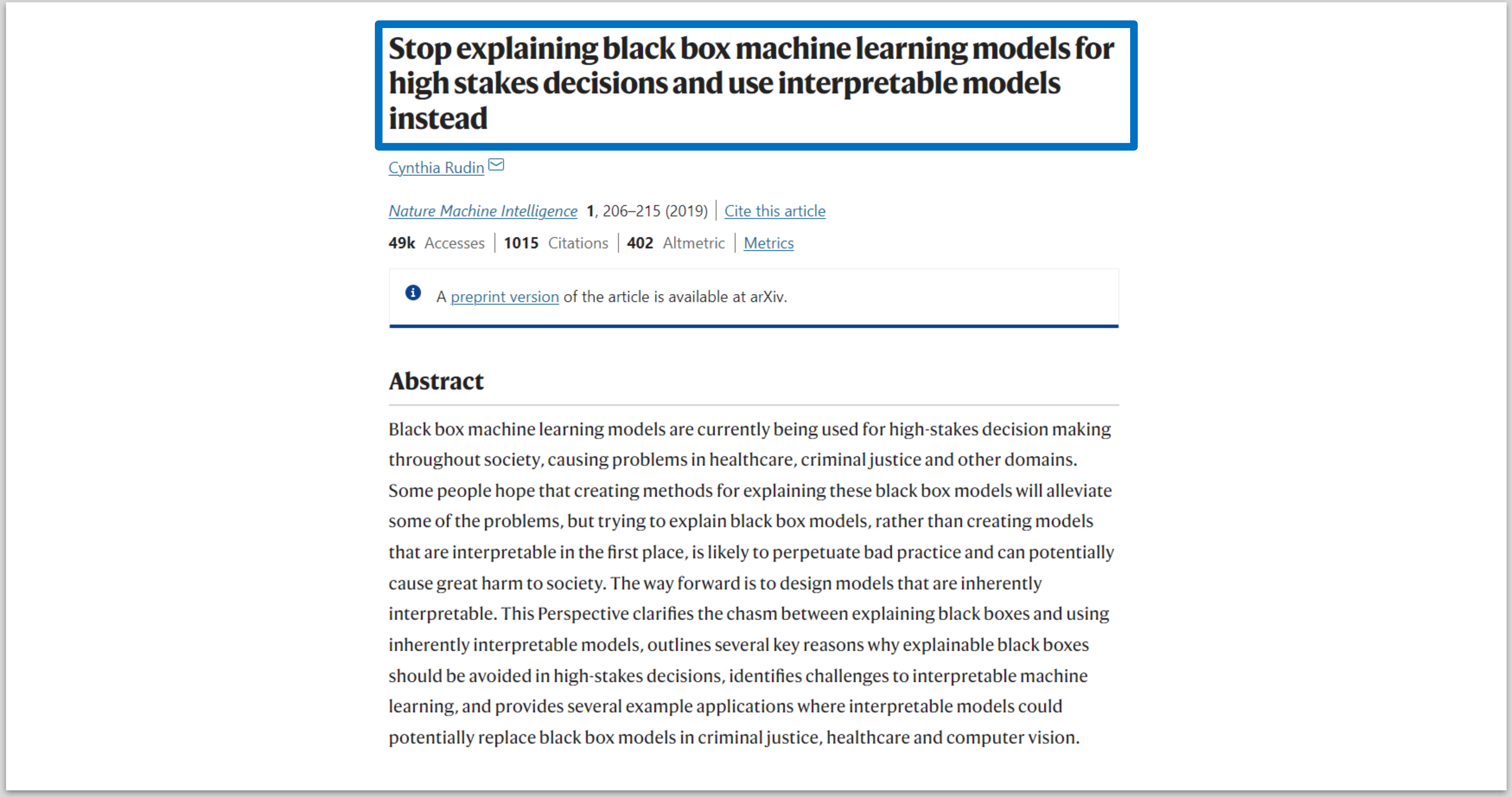
FELIPE BUCHBINDER

CATEGORICAL IMPERATIVE

All other things equal, an inherently interpretable model is always better than a model that needs to be explained.

In other words…

Using a model that's interpretable from the start is better than trying to explain a non-interpretable model using the techniques we're going to see next.



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

49k Accesses | **1015** Citations | **402** Altmetric | Metrics

A preprint version of the article is available at arXiv.

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.

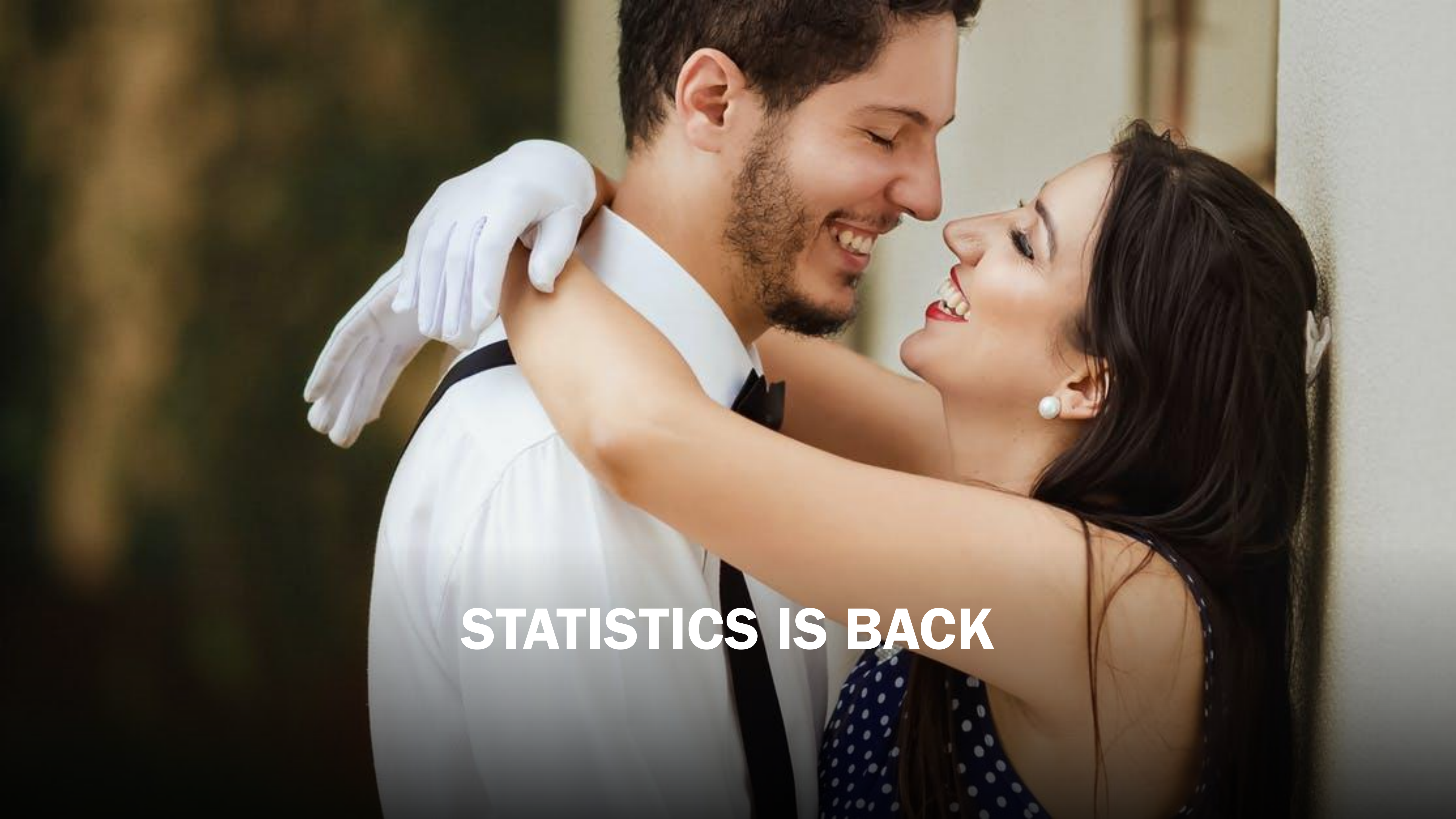
WHICH MODELS ARE INHERENTLY INTERPRETABLE?

- Linear / Logistic Regression
- Decision Trees
- Rule-based models
- All classical statistical models

FOR EXAMPLE…

$$\text{Income} = 2.000 + 100 \cdot \text{YearsExperience} + 200 \cdot \text{HasCollegeDegree} + 300 \cdot \text{HasMastersDegree} + 0 \cdot \text{HasPhD}$$

- Does having a college degree matter?
- How much does a person’s income increase, on average, with each year of experience?
- How much more can a person expect to earn by having a master’s degree?
- Does having a PhD matter?



STATISTICS IS BACK

WHICH MODELS ARE *NOT* INHERENTLY INTERPRETABLE?

(A.K.A. BLACK BOX MODELS)

- Support Vector Machines (SVMs)
- Neural Networks
- All ensemble models, including:
  - Random Forests
  - XGBoost
  - Adaboost
  - GBM
  - LightGBM
- All Deep Learning models

WHAT IF INTERPRETABLE MODELS AREN’T ACCURATE ENOUGH AND WE *NEED* TO USE BLACK-BOX MODELS?

A FEW OPTIONS YOU HAVE…

- Surrogate models
- Change in fit metrics
- Permutation feature importance
- SHAP values

SURROGATE MODEL

FELIPE BUCHBINDER

SURROGATE MODEL

A surrogate model is an inherently interpretable model used to explain the prediction of a black-box model

SURROGATE MODEL

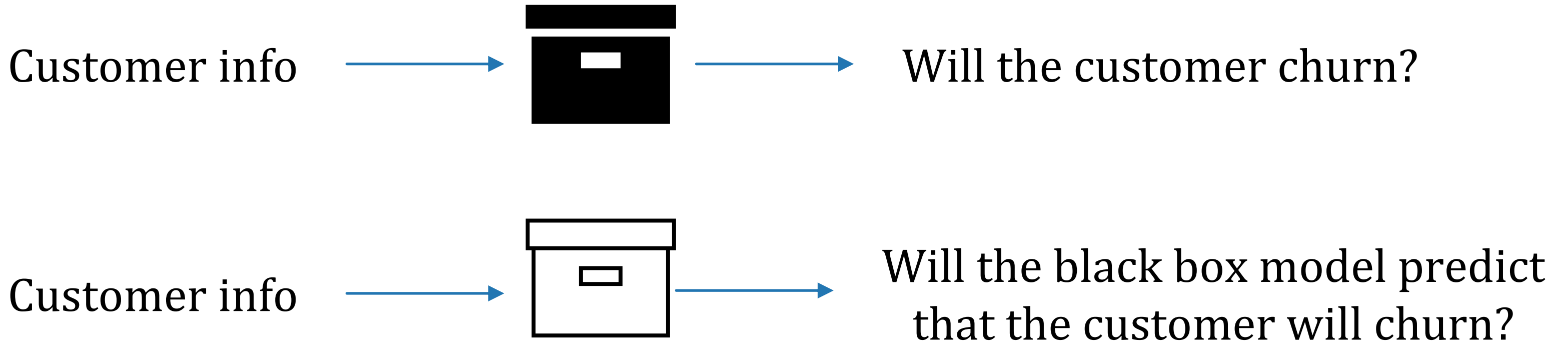
normally a decision tree
(why not a linear regression?)

A surrogate model is an inherently interpretable model used to explain the prediction of a black-box model

SURROGATE MODEL

normally a decision tree
(why not a linear regression?)

A surrogate model is an inherently interpretable model used to explain the prediction of a black-box model



Customer info → Will the customer churn?

Customer info → Will the black box model predict that the customer will churn?

RUDIN’S CRITIQUE TO THE USE OF SURROGATE MODEL’S

- Now you need to trust not only your original model, but your surrogate model as well!

CHANGE IN FIT METRICS

CHANGE IN FIT METRICS

Remove a feature and see how much fit measures (e.g. accuracy, R^2,…) change

- Not always evident how big a change should be in order to be meaningful
- Comparing models with different number of features can be misleading, as some metrics increase / decrease by the mere removal of a feature (e.g. R^2). One way to deal with this is to use permutation importance.

PERMUTATION IMPORTANCE

PERMUTATION IMPORTANCE

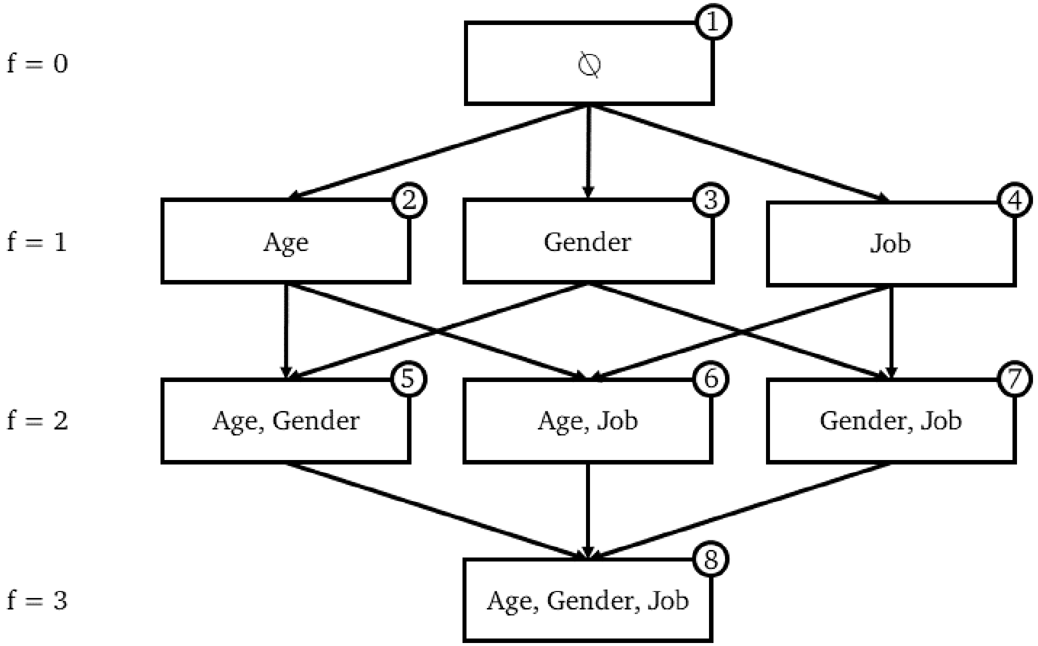
Scramble a feature randomly and see how much fit measures (e.g. accuracy, R^2,...) change

- Better than simply removing the feature (why?)
- Can be misleading if the scrambled feature is highly correlated with another feature (why?). Hence, permutation importance is inaccurate in the presence of multicollinearity.

SHAP VALUES

SHAP VALUES (LUNDBERG AND LEE, 2017)

- SHAP is a method to explain individual predictions
- Based on Collaborative Game Theory: the idea is that features collaborate to make the prediction and we want to assess the contribution of each feature to each individual prediction
- An intuitive way to understand the Shapley value is the following illustration: The feature values enter a room in random order. All feature values in the room participate in the game (= contribute to the prediction). The Shapley value of a feature value is the average change in the prediction that the coalition already in the room receives when the feature value joins them.



MATHEMATICALLY

$$\text{Shap}(x_j) = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_j\}} \frac{|S|!\,(p-|S|-1)!}{p!}\left[val\left(S \cup \{x_j\}\right) - val(S)\right]$$

Where S is a subset of features used in the model, x is the vector of feature values of the instance to be explained and p is the number of features.

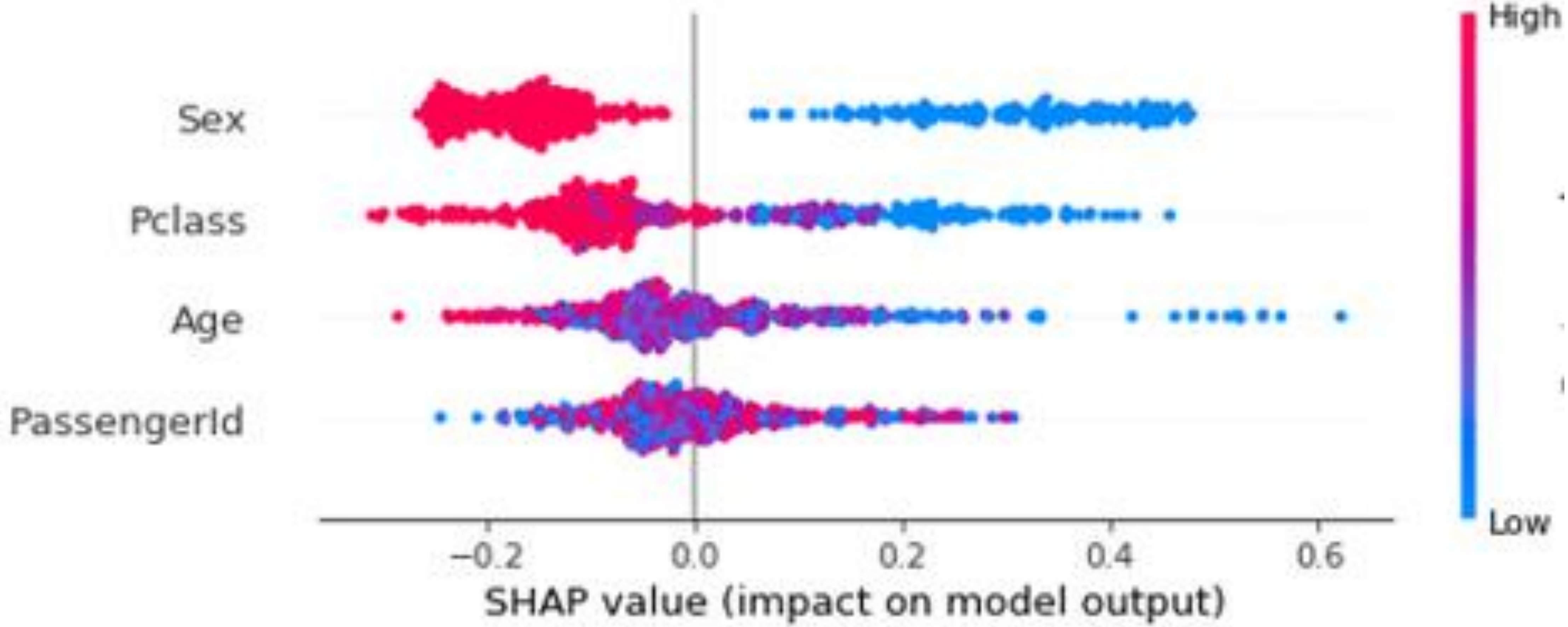
$val(S)$ is the prediction for feature values in set S that are marginalized over features that are not included in set S.

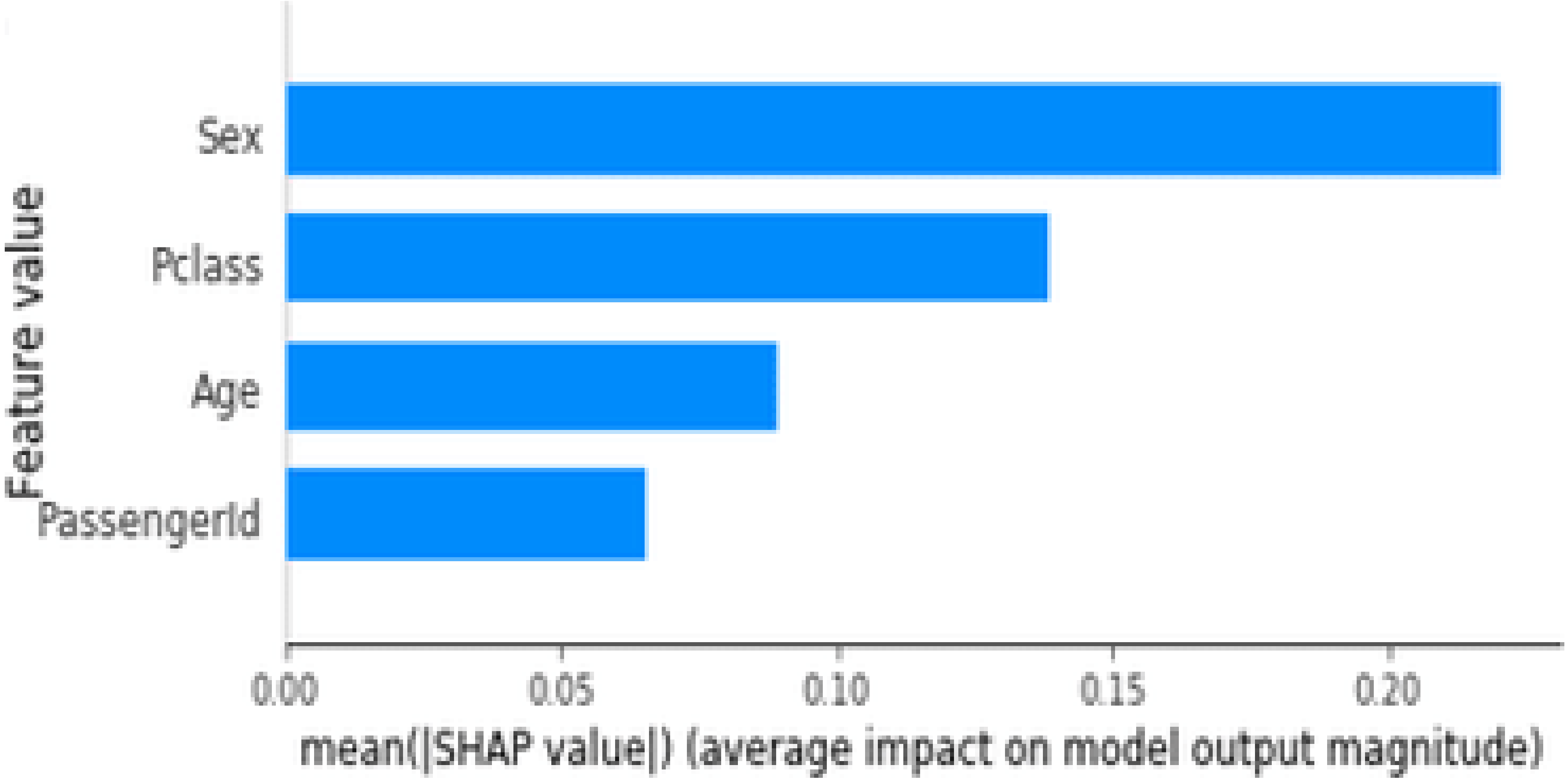
PROPERTIES OF SHAP VALUES

- **Efficiency**: The feature contributions must add up to the difference of prediction for x and the average.
- **Symmetry**: The contributions of two feature values j and k should be the same if they contribute equally to all possible coalitions.
- **Dummy**: A feature j that does not change the predicted value – regardless of which coalition of feature values it is added to – should have a Shapley value of 0.
- **Additivity**: The sum of the feature attributions is equal to the output of the model we are trying to explain.

The SHAP values are the only feature importance metrics with hard theory guaranteeing that it satisfies these (or any) properties

EXAMPLE: PREDICTING SURVIVAL IN THE TITANIC


High
Sex
Pclass
Age
PassengerId
Low
−0.2
0.0
0.2
0.4
0.6
SHAP value (impact on model output)


Feature value
Sex
Pclass
Age
PassengerId
0.00
0.05
0.10
0.15
0.20
mean(|SHAP value|) (average impact on model output magnitude)