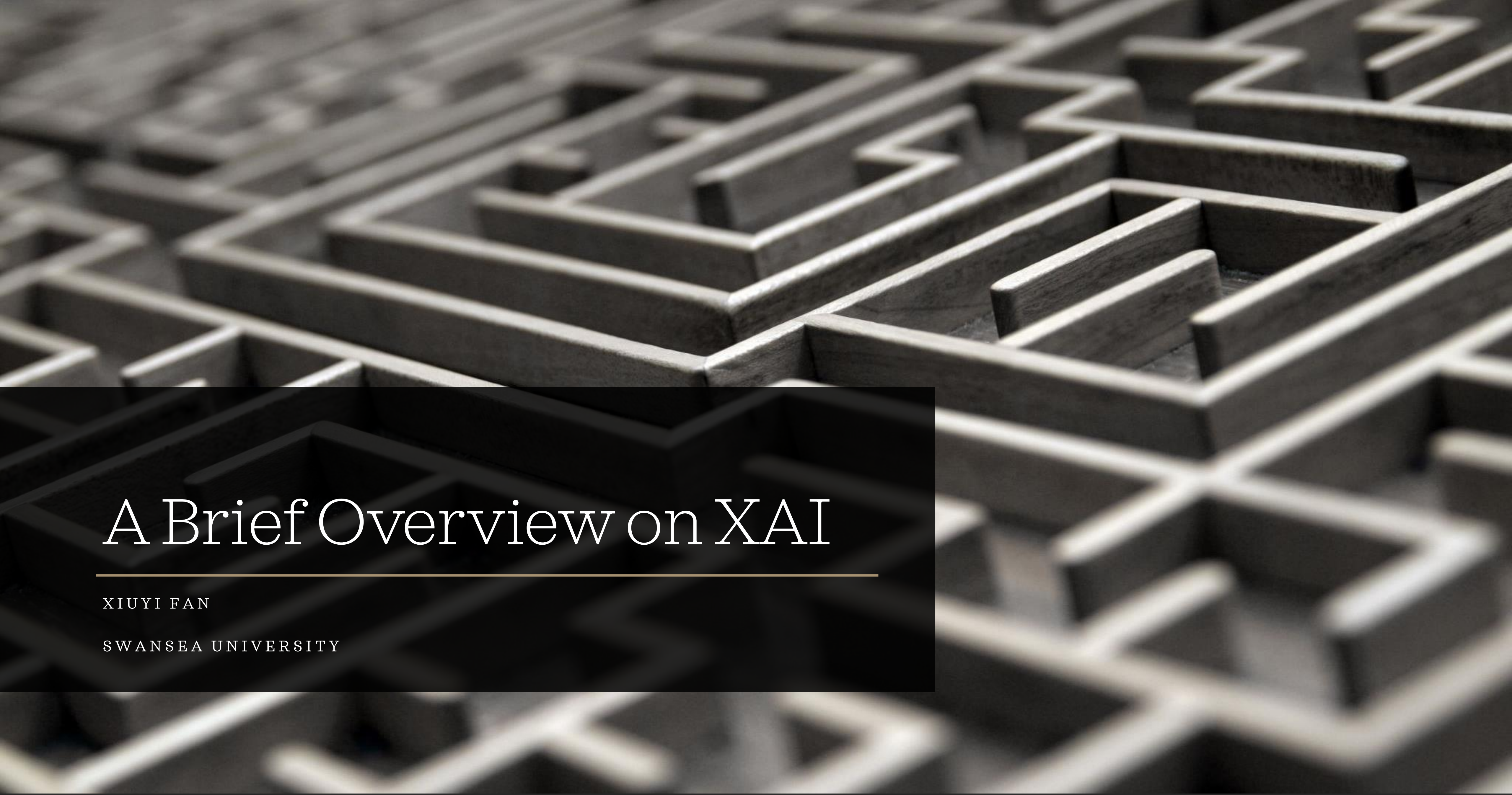


A Brief Overview on XAI

XIUYI FAN

SWANSEA UNIVERSITY

Definitions

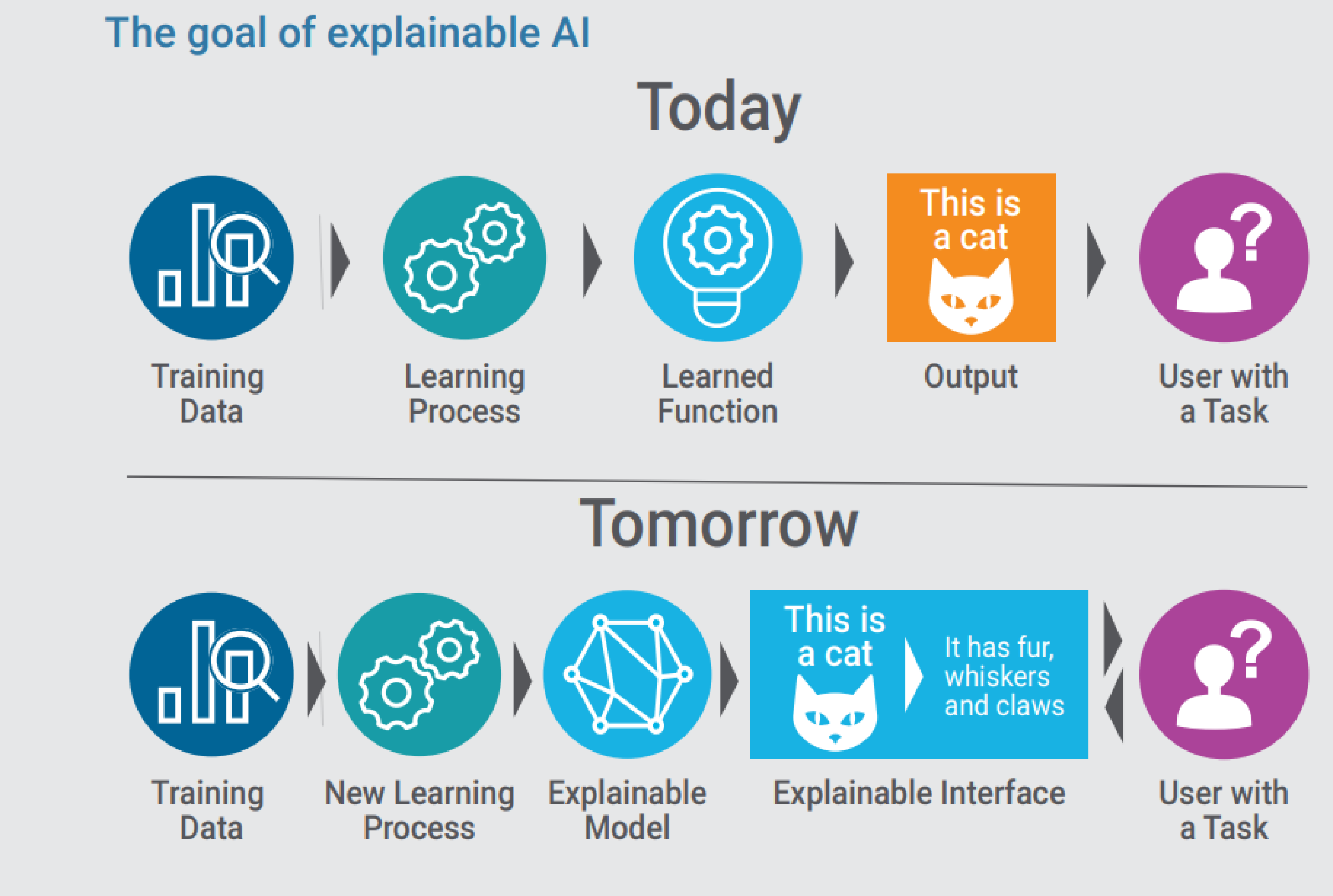
Interpretability is the degree to which a human can understand the cause of a decision.

-- Tim Miller, Explanation in artificial intelligence: Insights from the social sciences,

Artificial Intelligence, 2019

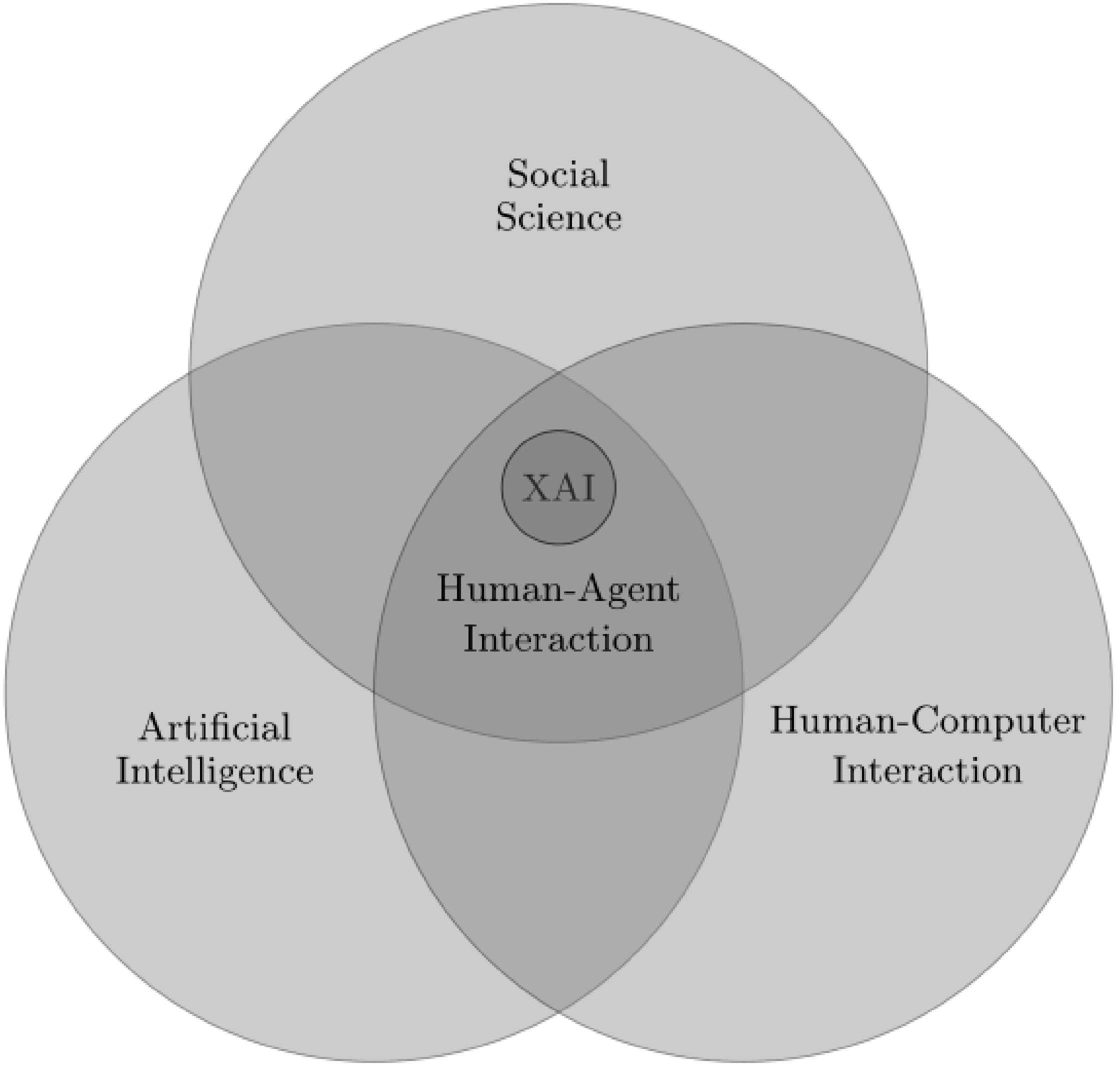
Interpretability is the degree to which a human can consistently predict the model's result.

Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! Criticism for interpretability." Advances in Neural Information Processing Systems (2016).



A Machine Learning Perspective …

(DARPA XAI Project)



XAI as a Field

Tim Miller, Explanation in artificial intelligence: Insights from the social sciences

Artificial Intelligence, 2019

Two main directions

Knowledge Driven (first part of this mini-module)

- Starting point: knowledge representation (logic, ontology, Bayesian nets, etc. - structured)
- Outputs: key elements / core inference processes
- Open questions: reaching explanation / usefulness / knowledge sources

Data Driven (second part of this mini-module)

- Starting point: data (tabular forms, images, etc. - unstructured)
- Outputs: key features / core inputs
- Open questions: shallow explanation / ground truth for correctness / overall effectiveness

Plan for this mini-module

Lecture 1: Introduction

Lecture 2: Assumption-based Argumentation

Lecture 3: Argumentation-based Explanation

Lecture 4: Shapley Value and SHAP

Lecture 5: Practical session on SHAP (??)

Zoom Poll

Why XAI? – Trust & Transparency

The running hypothesis is that by building more transparent, interpretable, or explainable systems, users will be better equipped to understand and therefore trust the intelligent agents.

Explanation in Artificial Intelligence: Insights from the Social Sciences

Tim Miller, 2019

Why XAI? – Insight & Knowledge

Knowing the 'why' can help you learn more about the problem, the data and the reason why a model might fail.

Interpretable Machine Learning, A Guide for Making Black Box Models Explainable
Christoph Molnar, 2019

Why XAI?
(Christoph Molnar, 2019)

Human curiosity and learning

- Why is something the case?

Why XAI?
(Christoph Molnar, 2019)

Human curiosity and learning

- Why is something the case?

Find meaning in the world

- Resolve inconsistencies between elements of our knowledge structures

Why XAI?
(Christoph Molnar, 2019)

Human curiosity and learning

- Why is something the case?

Find meaning in the world

- Resolve inconsistencies between elements of our knowledge structures

Goal of science

- To gain knowledge, which does not always exist in black box methods

Why XAI?

(Christoph Molnar, 2019)

Human curiosity and learning

- Why is something the case?

Find meaning in the world

- Resolve inconsistencies between elements of our knowledge structures

Goal of science

- To gain knowledge, which does not always exist in black box methods

Safety measures

- Understanding the internal mechanisms of algorithms

Why XAI?
(Christoph Molnar, 2019)

Detecting bias

- Identify biases from the training data

Why XAI?
(Christoph Molnar, 2019)

Detecting bias

- Identify biases from the training data

Increase social acceptance

- Make algorithms more human-friendly and acceptable

Why XAI?
(Christoph Molnar, 2019)

Detecting bias

- Identify biases from the training data

Increase social acceptance

- Make algorithms more human-friendly and acceptable

Debug and audit

- Verify and validate algorithms

… …

Scope of Interpretability
(algorithmic perspective)

Algorithm Transparency

How does the algorithm create the model?

- *Focus on specific algorithms*

Scope of Interpretability
(algorithmic perspective)

Algorithm Transparency

How does the algorithm create the model?

- *Focus on specific algorithms*

Global Model Interpretability

How does the trained model make predictions?

- *Which features are important in the model?*

Scope of Interpretability
(algorithmic perspective)

Algorithm Transparency

How does the algorithm create the model?

- *Focus on specific algorithms*

Global Model Interpretability

How does the trained model make predictions?

- *Which features are important in the model?*

Model Interpretability on a Modular Level

How do parts of the model affect predictions?

- *Which "components" in a model are responsible for a decision?*

Scope of Interpretability
(algorithmic perspective)

Algorithm Transparency

How does the algorithm create the model?

- *Focus on specific algorithms*

Global Model Interpretability

How does the trained model make predictions?

- *Which features are important in the model?*

Model Interpretability on a Modular Level

How do parts of the model affect predictions?

- *Which "components" in a model are responsible for a decision?*

Instance Interpretability for a Single Prediction

Why did the model make a certain prediction for an instance?

- *Instance level explanation*

Two main types (for X-ML):

Interpretable Models vs. Model-Agnostic Methods

Interpretable Models

Standard simple / weak models …

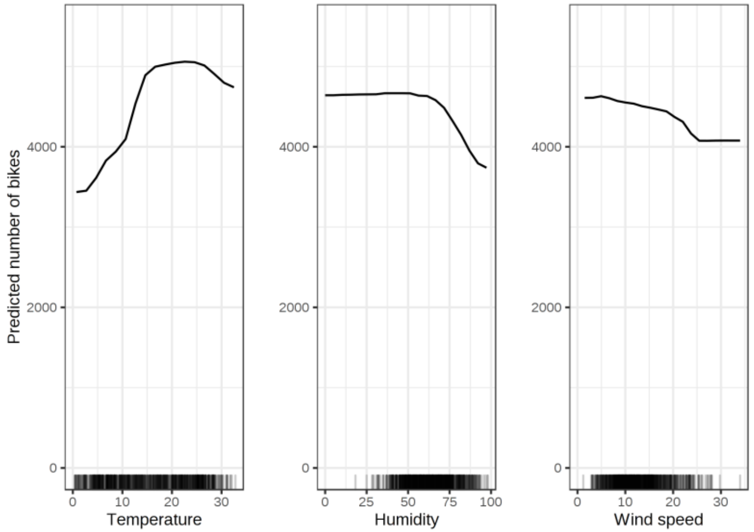
- Linear Regression
- Logistic Regression
- Decision Trees
- Naïve Bayes
- K-nearest Neighbors

Model-Agnostic Methods:
Partial Dependence Plot (PDP)

Global Method

The partial dependence plot shows the marginal effect of one or two features have on the predicted outcome of a machine learning model.

For classification where the machine learning model outputs probabilities, the partial dependence plot displays the probability for a certain class given different values for feature(s) in S.

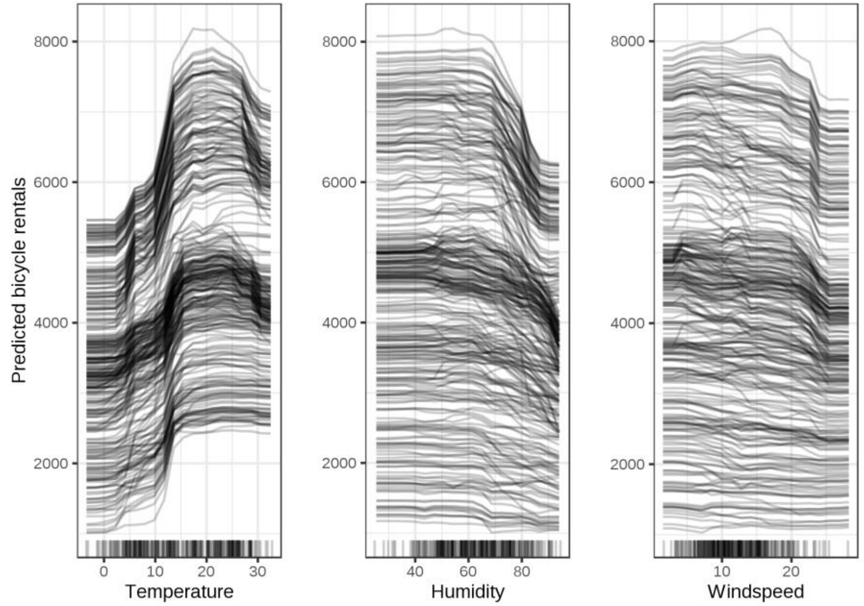


- Friedman, Jerome H. “Greedy function approximation: A gradient boosting machine.” Annals of statistics (2001): 1189-1232.

Model-Agnostic Methods:
Individual Conditional Expectation

Instance version of PDP

Individual Conditional Expectation (ICE) plots display one line per instance that shows how the instance's prediction changes when a feature changes.



- Goldstein, Alex, et al. "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation." Journal of Computational and Graphical Statistics 24.1 (2015): 44-65

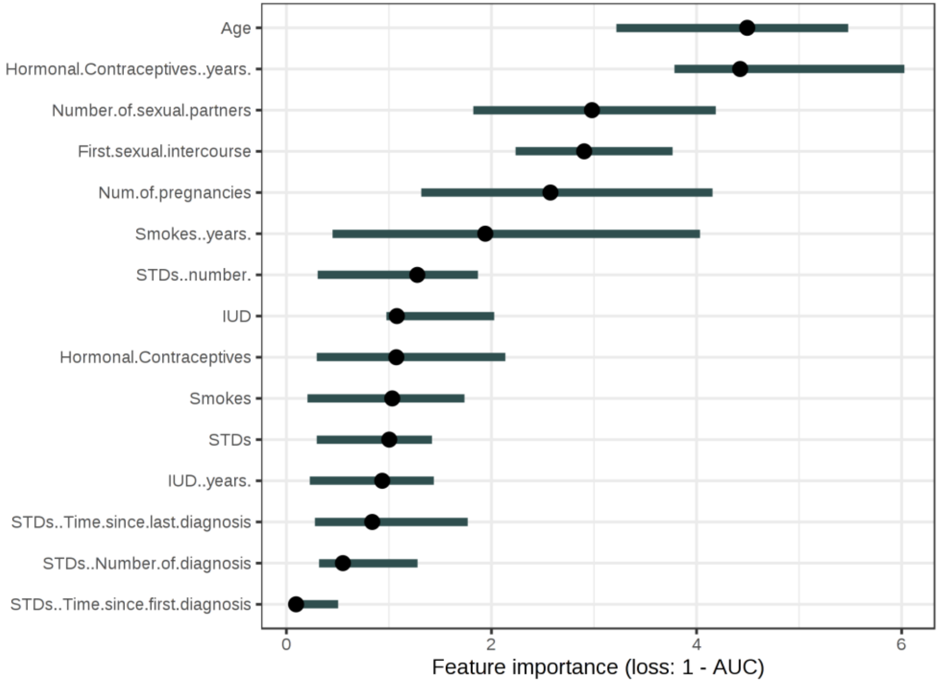
Model-Agnostic Methods
Permutation Feature Importance

Input: Trained model f, feature matrix X, target vector y, error measure L(y,f).

1. Estimate the original model error $e^{orig} = L(y, f(X))$ (e.g. mean squared error)
2. For each feature j = 1,…,p do:
   - Generate feature matrix X^{perm} by permuting feature j in the data X. This breaks the association between feature j and true outcome y.
   - Estimate error $e^{perm} = L(Y,f(X^{perm}))$ based on the predictions of the permuted data.
   - Calculate permutation feature importance $FI^j = e^{perm}/e^{orig}$. Alternatively, the difference can be used: $FI^j = e^{perm} - e^{orig}$
3. Sort features by descending FI.

Global Method

Permutation feature importance measures the increase in the prediction error of the model after we permuted the feature's values, which breaks the relationship between the feature and the true outcome.



Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. "Model Class Reliance: Variable importance measures for any machine learning model class, from the 'Rashomon' perspective." , arxiv 2018

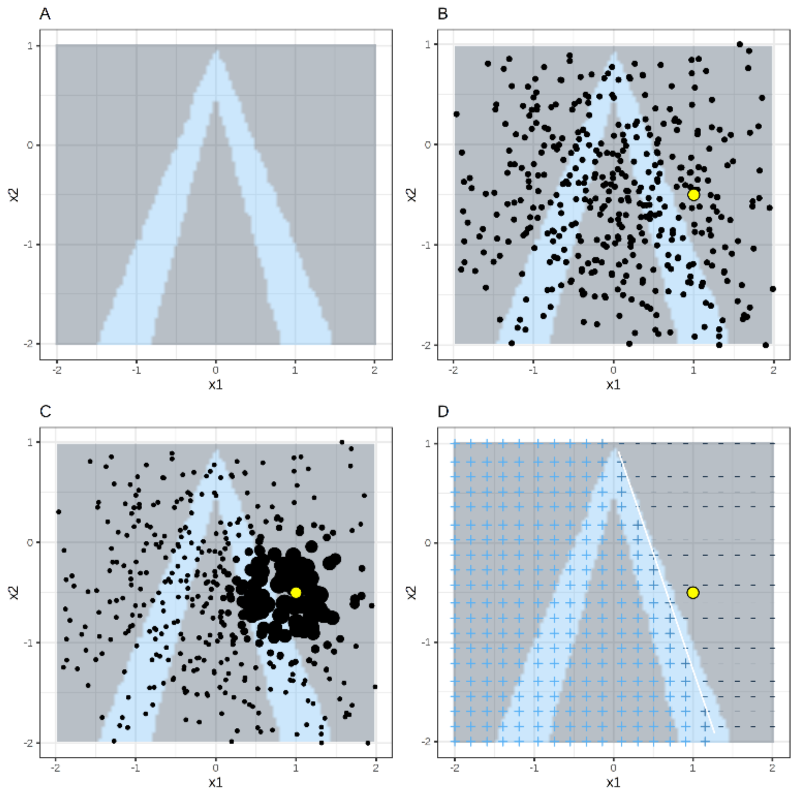
Model-Agnostic Methods

Global Surrogate

Global Method (*)

A global surrogate model is an interpretable model that is trained to approximate the predictions of a black box model. We can draw conclusions about the black box model by interpreting the surrogate model.

Decision Tree -> SVM, etc.



Model-Agnostic Methods

Local Surrogate (LIME)

Instance Method

Local surrogate models are interpretable models that are used to explain individual predictions of black box machine learning models.

LIME generates a new dataset consisting of permuted samples and the corresponding predictions of the black box model. On this new dataset LIME then trains an interpretable model, which is weighted by the proximity of the sampled instances to the instance of interest.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM (2016)

Model-Agnostic Methods
Shapley Values

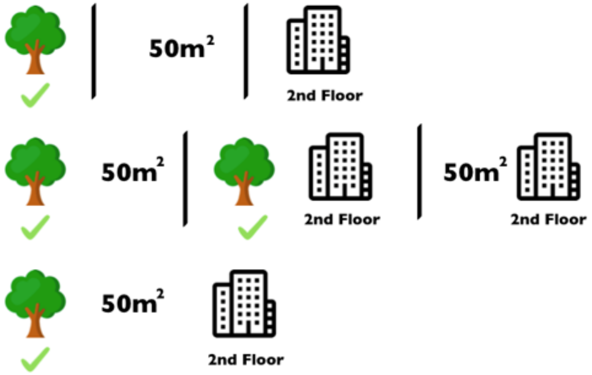
Instance Method

Features are viewed as "players".

Shapley values – a method from coalitional game theory tells us how to fairly distribute the "payout" among the features.

The Shapley value is the average marginal contribution of a feature value across all possible coalitions.

Model-Agnostic Methods
Shapley Values



Instance Method

Features are viewed as "players".

Shapley values – a method from coalitional game theory tells us how to fairly distribute the "pay-out" among the features.

The Shapley value is the average marginal contribution of a feature value across all possible coalitions.

Example-Based Explanations

Counterfactual Explanations

- If X had not occurred, Y would not have occurred

Adversarial Examples

- Adversarial examples are counterfactual examples with the aim to deceive the model, not interpret it.
- Identify cases where the model fails.

… …

Visualize Neural Network

Very important & interesting area

See e.g. papers in I*nterpretable Machine Learning, Christoph Molnar, 2019*

An Example

Using Explainable AI to Understand Impacts of Non-pharmaceutical Control Measures on COVID-19 Transmission for Evidence-based Policy

Funded by Welsh Government
2020-2021

COVID Project

“This project will gain deep insights into the effects of measures used to control the spread of COVID-19. … We will further explore XAI techniques to reveal dynamics between control measures and disease transmission.”

COVID Project

Question: *Which control measures are more effective?*

Two Objectives:

(**O1**) *develop and maintain a streamlined dataset capturing COVID-19 transmission information and implemented control measures, and*

(**O2**) *adapt and apply multiple XAI techniques on the dataset constructed to fulfil O1 to obtain easily interpretable relations between control measures and disease transmission while actively interacting with the policy community.*

International Journal of Information Technology Vol. 26 No. 1 2020

An Investigation of COVID-19 Spreading Factors with Explainable AI Techniques

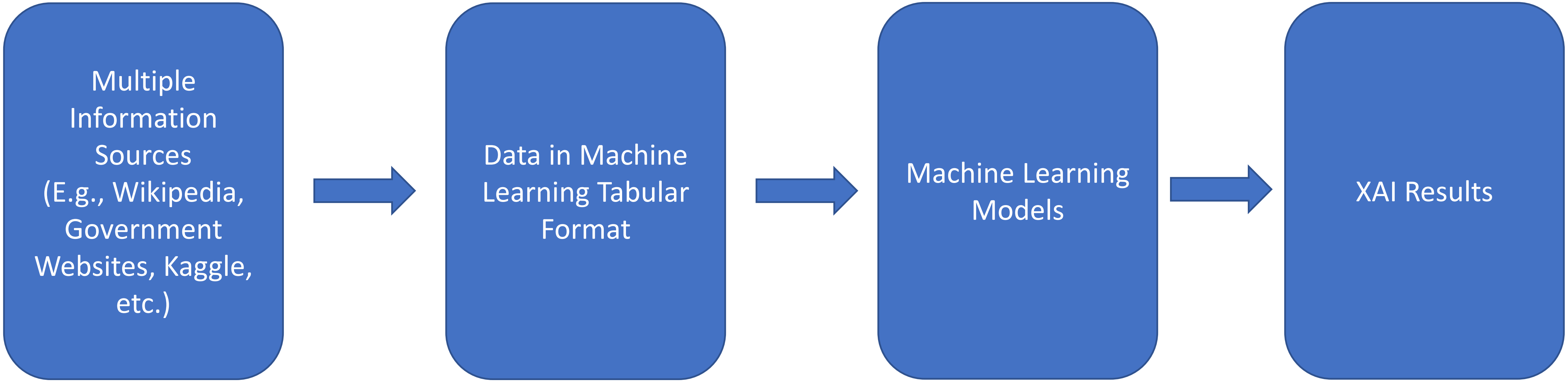
Xiuyi Fan[1], Siyuan Liu[1], Jiarong Chen[2,3,4], Matthew Williams[4], and Thomas C. Henderson[5]

[1]Computer Science Department, Swansea University, United Kingdom
[2]Clinical Experimental Center, Jiangmen Key Laboratory of Clinical Biobanks and Translational Research, Jiangmen Central Hospital, Affiliated Jiangmen Hospital of Sun Yat-sen University, Jiangmen 529030, China
[3]Department of Oncology, Jiangmen Central Hospital, Affiliated Jiangmen Hospital of Sun Yat-sen University, Jiangmen 529039, China
[4]Computational Oncology Group, Imperial College London, United Kingdom
5School of Computing, University of Utah, USA

xiuyi.fan@seansea.ac.uk

Publication

Current Work



Current Work

- France
  - Government Advocation (GA): 12/03/2020
    - On March 12, French President Emmanuel Macron announced on public television that all schools and all universities would close from Monday (March 16) until further notice.
    - https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_France
  - School Closure (SC): 16/03/2020
    - On March 12, French President Emmanuel Macron announced on public television that all schools and all universities would close from Monday (March 16) until further notice.
    - https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_France

- United Kingdom
  - Government Advocation (GA): 01/03/2020
    - By March 1, cases had been detected in England, Wales, Northern Ireland and Scotland. Subsequently, Prime Minister Boris Johnson unveiled the Coronavirus Action Plan, and the government declared the outbreak as "level 4 incident".

Table 1: Implementation dates of control measures at 18 countries and regions.

Countries and Regions	Government Advocation (GA)	Mask Use (MU)	School Closure (SC)	City Lockdown (CL)	Mass Testing (MT)	International Travel Ban (ITB)	Contact Tracing (CT)
Australia	13/03/2020					01/02/2020	
France	12/03/2020		16/03/2020	17/03/2020		16/03/2020	
Germany	28/01/2020		26/02/2020	16/03/2020		28/01/2020	
Italy	31/01/2020		04/03/2020	08/03/2020		31/01/2020	
Japan	24/01/2020	22/01/2020	02/03/2020			01/02/2020	25/02/2020
Singapore	22/01/2020	01/02/2020			24/01/2020	29/01/2020	23/01/2020
South Korea	22/01/2020	22/01/2020	22/01/2020		31/01/2020	02/02/2020	22/01/2020
Spain	14/03/2020		12/03/2020	14/03/2020		10/03/2020	
United Kingdom	01/03/2020		20/03/2020	21/03/2020			
Beijing	24/01/2020	07/02/2020	22/01/2020	24/01/2020	24/01/2020	28/03/2020	24/01/2020
California	04/03/2020		13/03/2020	19/03/2020		02/02/2020	
Guangdong	23/01/2020	26/01/2020	22/01/2020	24/01/2020	23/01/2020	28/03/2020	23/01/2020
Hong Kong	04/01/2020	08/01/2020	22/01/2020		04/01/2020	27/01/2020	04/01/2020
Hubei	20/01/2020	22/01/2020	22/01/2020	23/01/2020	05/02/2020	23/01/2020	03/02/2020
Macau	31/12/2019	03/02/2020	22/01/2020		20/02/2020	28/01/2020	
New York	07/03/2020		15/03/2020	20/03/2020	13/03/2020	02/02/2020	
Taiwan	20/01/2020	31/01/2020	22/01/2020		01/02/2020	23/01/2020	27/01/2020
Washington	29/02/2020		13/03/2020	23/03/2020	17/03/2020	02/02/2020	

Current Work

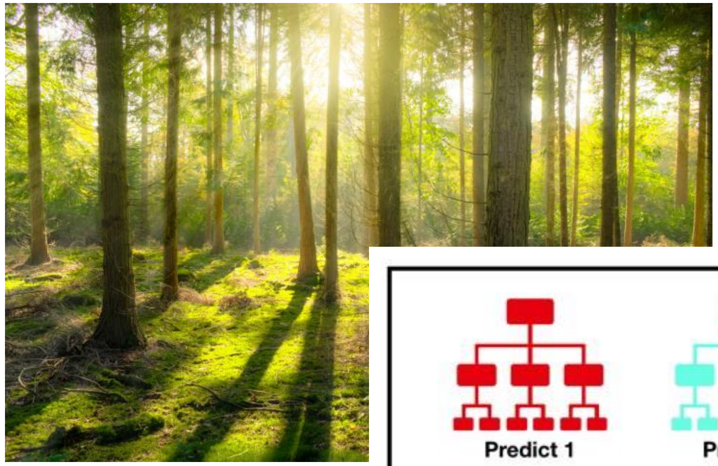
Table 2: An illustration of the data set with four data entries (Singapore, 12/02/2020, Japan, 26/03/2020, Germany, 26/03/2020, South Korea, 16/03/2020, and Guangdong, 08/02/2020). NC = New Case, GA = Government Advocation, MU = Mask Use, SC = School Closure, CL = City Lockdown, MT = Mass Testing, ITB = International Travel Ban, CT = Contact Tracing, T = Temperature, and H = Humidity.

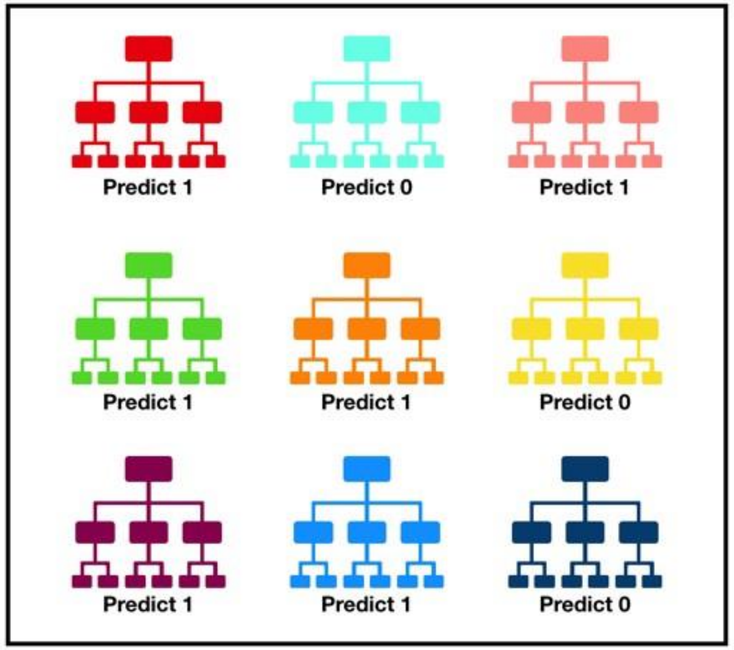
R_t	NC	GA	MU	SC	CL	MT	ITB	CT	T	H
0.31	78	55	55	55	0	46	44	55	3.73	48.47
0.72	53	17	14	18	16	17	0	18	15.89	62.66
1.34	4	22	12	0	0	20	15	21	27.86	83.86
1.91	92	63	65	25	0	0	55	31	17.375	32.75
2.14	5962	59	0	30	11	0	12	0	6.19	39.35

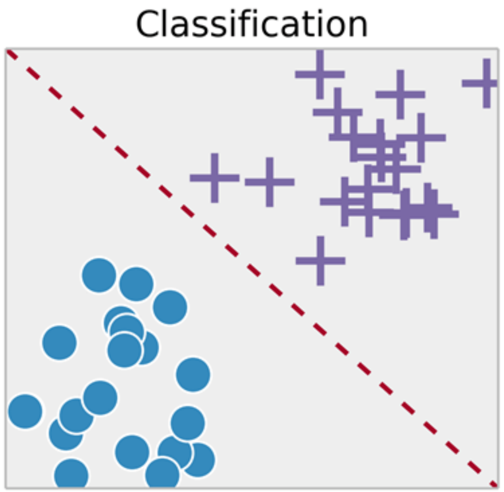
Table 3: Five data entries in Table 2 after discretization. For example, for the first row of Table 2, with $R_t = 0.31$, NC=78, GA=55,MU=55, SC=55,CL=0, MT=46, ITB=44, CT =55, T=3.73, H=48.47, it is discetrized as shown in the first row of this table, with $R_t = 0.31$, NC=1, GA=4, MU=4, SC=4, CL=0, MT=4, ITB=4, CT=4, T=1, H=1.

R_t	NC	GA	MU	SC	CL	MT	ITB	CT	T	H
0.31	1	4	4	4	0	4	4	4	1	1
0.72	1	4	3	4	4	4	0	4	2	1
1.34	0	4	3	0	0	4	3	4	3	2
1.91	1	4	4	4	0	0	4	4	2	0
2.14	2	4	0	4	3	0	3	0	1	0

Current Work

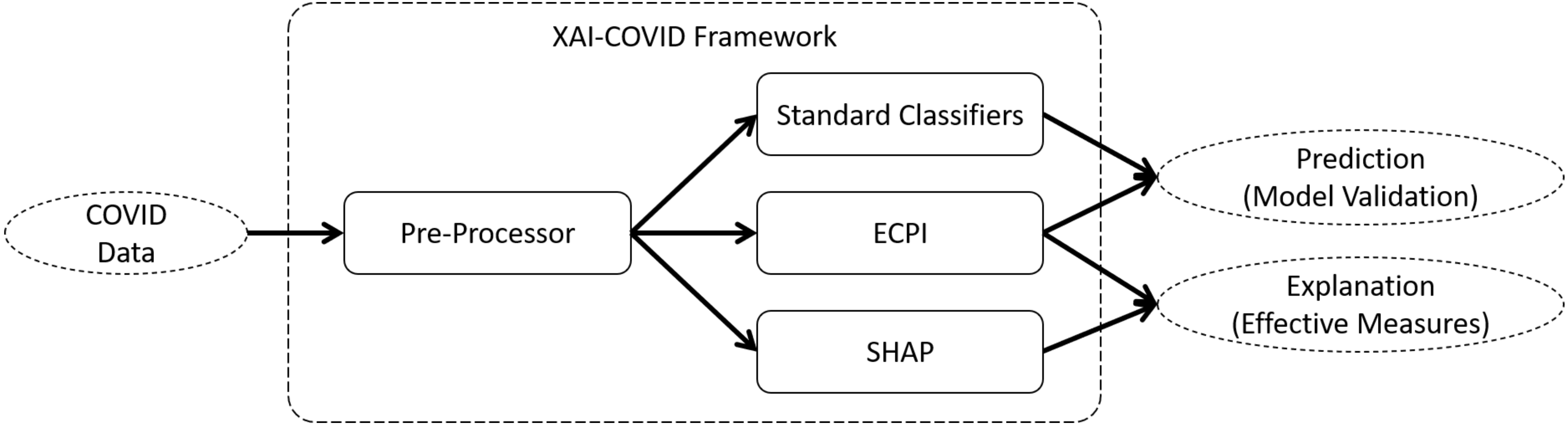






Important Features

Current Work



Current Work

Anti-XAI

Cynthia Rudin, ‘Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead’, Nature Machine Intelligence, 1, 206–215, (May 2019).

In short ...

Motivated by real reasons (many of them)

Many approaches, no dominant method

New area, still large scope to explore