

CSCI 3210  
Computational Game Theory


Explainable AI (XAI)  
and Game Theory

Mohammad T. Irfan  
Web: <https://mtirfan.com>

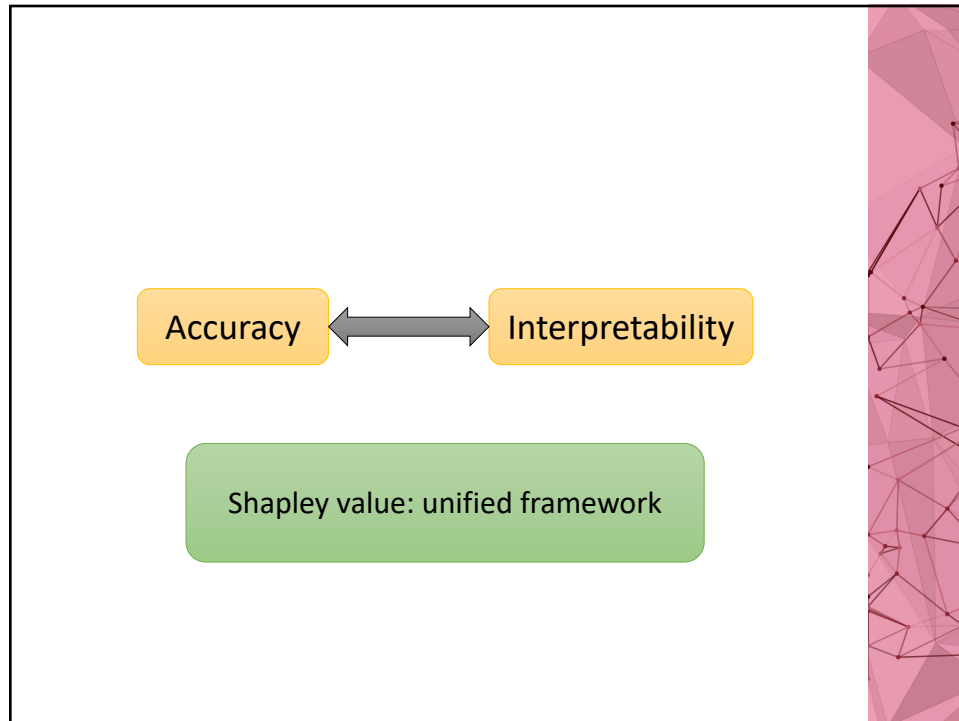
1

References

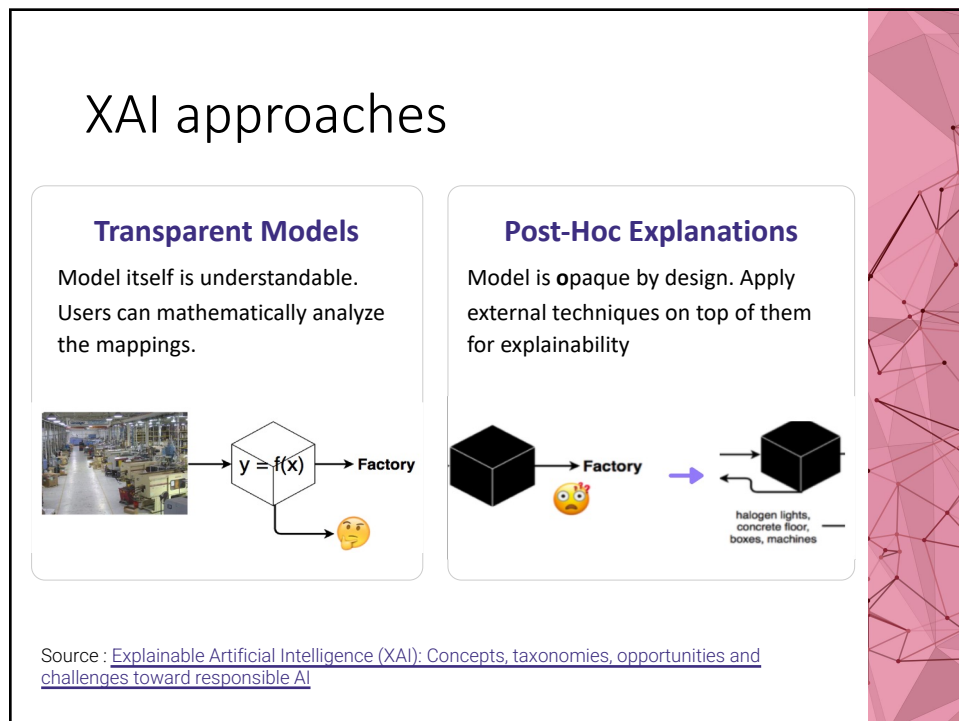
- Canvas →
- Modules →
- Topics for Final Project/Presentation →
- Papers on XAI



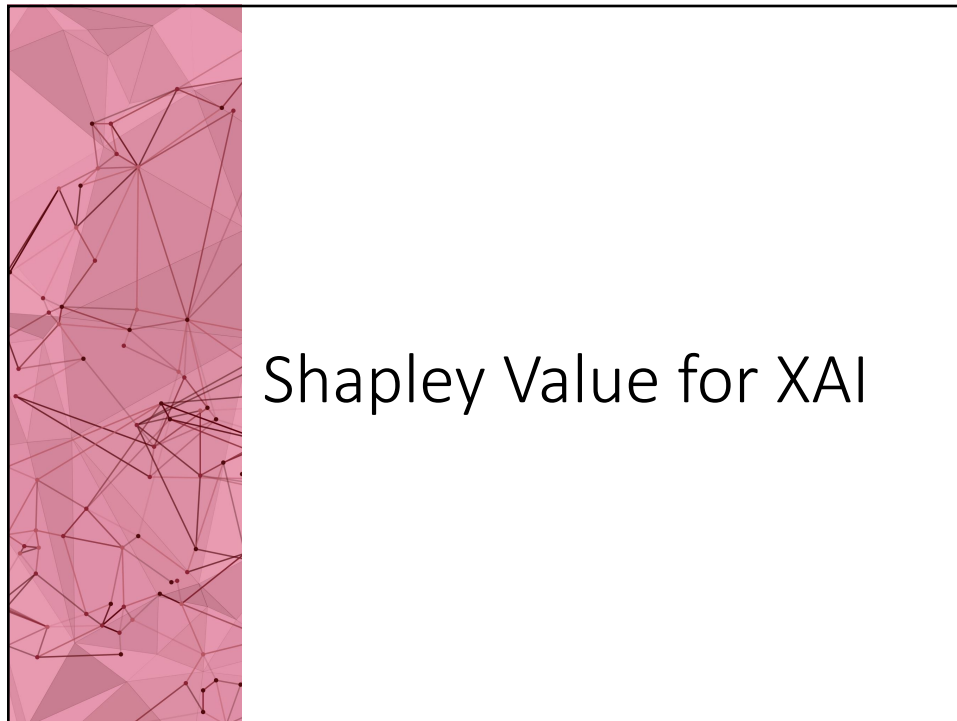
2



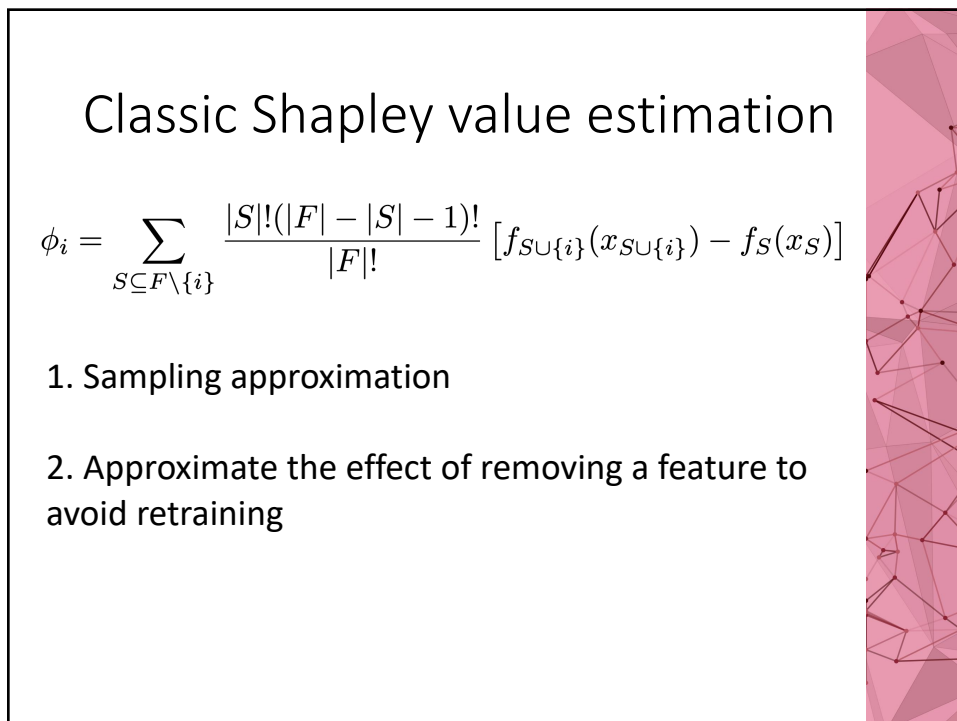
3



4



5


A slide with a decorative background on the right side consisting of a network of red lines and dots. The main content is the title "Classic Shapley value estimation" in a large, black, sans-serif font, centered at the top. Below the title is a mathematical formula for the Shapley value  $\phi_i$ . Below the formula are two numbered steps: "1. Sampling approximation" and "2. Approximate the effect of removing a feature to avoid retraining".

Classic Shapley value estimation

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)]$$

1. Sampling approximation
2. Approximate the effect of removing a feature to avoid retraining

6



# SHAP Algorithm

[Paper: A Unified Approach to Interpreting Model Predictions](#)

Thanks to Rose Xi '22 for the examples!

7

$$\phi_i(f, x) = \sum_{z' \subseteq x} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)]$$

input data point  $x$

<b>Age</b> 42	<b>Yrs of Educ</b> 13	<b>Marital Status</b> Separated	<b>Race</b> White	<b>Occupation</b> Sales	...
------------------	--------------------------	------------------------------------	----------------------	----------------------------	-----

$f(x)$ : How likely is the person to earn at least \$50K?  
 $f_x(z')$ : later

8

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)]$$

Shapley value for feature i      input data point

example input data point  $x$

Age 42	Yrs of Educ 13	Marital Status Separated	Race White	Occupation Sales	...
-----------	-------------------	-----------------------------	---------------	---------------------	-----

↑  
feature i that we are interested in calculating Shapley value for

9

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)]$$

blackbox model      Shapley value for feature i      input data point

example input data point  $x$

Age 42	Yrs of Educ 13	Marital Status Separated	Race White	Occupation Sales	...
-----------	-------------------	-----------------------------	---------------	---------------------	-----

↑  
feature i that we are interested in calculating Shapley value for

$x \rightarrow$  **black box model**  $\rightarrow f(x) = 70\%$  likely to have income >50k

$f(x) =$  How likely is the person to earn at least \$50K?

10

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)]$$

input data point      simplified data point

example input data point $x$	Age 42	Yrs of Educ 13	Marital Status Separated	Race White	Occupation Sales	...
simplified data point $x'$	Age	Yrs of Educ	Marital Status	Race	Occupation	...

$x'$  = **simplified data point (yes or no)**  
instead of details like "13"

11

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)]$$

input data point      subset of simplified data point      simplified data point      subset z

example input data point $x$	Age 42	Yrs of Educ 13	Marital Status Separated	Race White	Occupation Sales	...
simplified data point $x'$	Age	Yrs of Educ	Marital Status	Race	Occupation	...
subset $z'$	Age	Yrs of Educ	Marital Status	Race	Occupation	...

$z'$  is a subset of the simplified data

12

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)]$$

input data point  $x$       subset of simplified data point  $z' \subseteq x'$       simplified data point  $z'$       subset  $z$       subset  $z$  excluding the feature  $i$  we're interested in  $z' \setminus i$

example input data point $x$	Age 42	Yrs of Educ 13	Marital Status Separated	Race White	Occupation Sales	...
simplified data point $x'$	Age	Yrs of Educ	Marital Status	Race	Occupation	...
subset $z'$	Age	Yrs of Educ	Marital Status	Race	Occupation	...
$z' \setminus i$	Age	Yrs of Educ	Marital Status	Race	Occupation	...

$z' \setminus i$  is the subset  $z'$  excluding the feature  $i$  for which we are calculating the Shapley value

13

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)]$$

example input data point  $x$

Age 42	Yrs of Educ 13	Marital Status Separated	Race White	Occupation Sales	...
-----------	-------------------	-----------------------------	---------------	---------------------	-----

simplified data point  $x'$

Age	Yrs of Educ	Marital Status	Race	Occupation	...
-----	-------------	----------------	------	------------	-----

subset  $z'$

Age	Yrs of Educ	Marital Status	Race	Occupation	...
-----	-------------	----------------	------	------------	-----

$z' \setminus i$

Age	Yrs of Educ	Marital Status	Race	Occupation	...
-----	-------------	----------------	------	------------	-----

$f_x(z')$  maps the features  $z' = (\text{Age, Yrs of Educ, Race})$  to their values in  $x = (42, 13, \text{White})$  and applies that to  $f(\cdot)$ . Same goes with  $f_x(z' \setminus i)$ .

70%      40%


14

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)]$$

70%
40%
  
70% - 40% = 30% = contribution of feature i

example input data point $x$	Age 42	Yrs of Educ 13	Marital Status Separated	Race White	Occupation Sales	...
simplified data point $x'$	Age	Yrs of Educ	Marital Status	Race	Occupation	...
subset $z'$	Age	Yrs of Educ	Marital Status	Race	Occupation	...
$z' \setminus i$	Age	Yrs of Educ	Marital Status	Race	Occupation	...

15



# Demos

16



## Census data

[Link](#)

Homepage:

<https://shap.readthedocs.io/en/latest/overviews.html>

17

## Titanic survival prediction

<https://titanicexplainer.herokuapp.com/classifier/>

18