

Explainable AI – Interpretable ML

Ioannis Mollas, iamollas@csd.auth.gr, PhD Student
Supervisors: Nick Bassiliades, Grigorios Tsoumakas

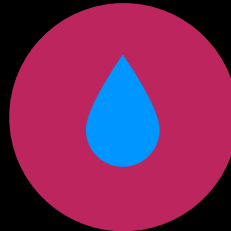




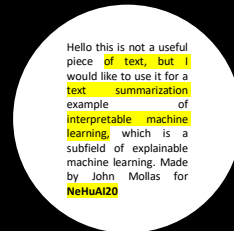
AI APPLICATIONS



Self Driving



Water Waste



Text Summarization



Assist Bail System



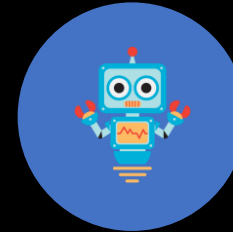
Credit Score



Fall Detection



Drug Design



Robotic Assistant



Security



Data Analysis



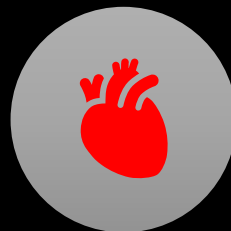
Prescriptive Maintenance



Personalized Medication



Natural Disaster



Heart Disease



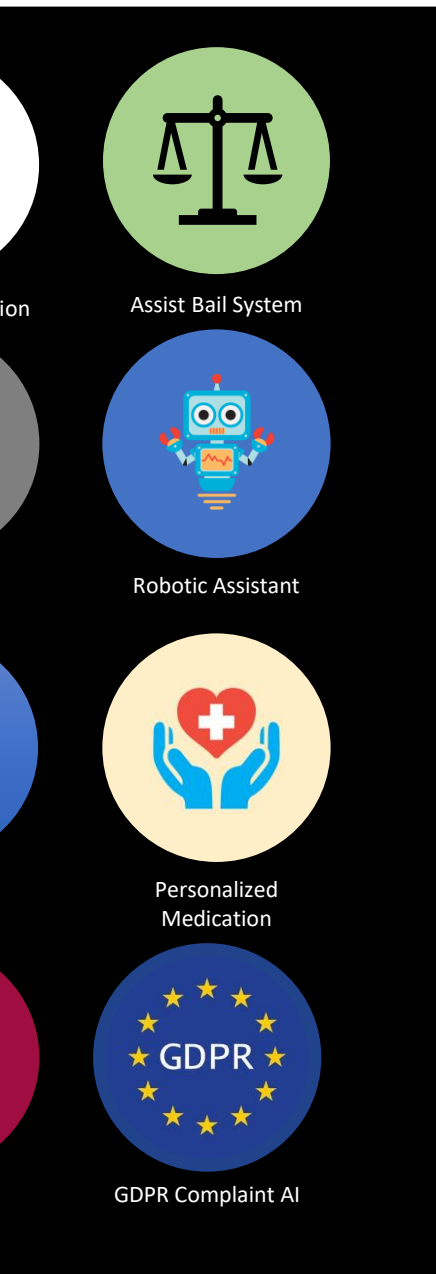
Automated Recruitments



GDPR Complaint AI

XAI

Explainable AI (XAI) refers to methods and techniques in the applications of artificial intelligence such that will shed light to reasons of the decision-making of such systems



What we want to tackle or achieve with XAI?

Tackle

- Gender biases
- Racial and religious discrimination
- Data errors and misclassifications
- Loss of life

Achieve

- Acquisition of new knowledge
- Robust systems
- Trustworthy systems

Real case examples

"Google Translate automatically chooses the gender for you"

Real case examples

"Google
Translate
automatically

The screenshot shows the Google Translate web interface. At the top, there's a navigation bar with the Google Translate logo and a 'Sign in' button. Below this, there are tabs for 'Text' and 'Documents'. The main area is divided into two sections: the source language (Hungarian - DETECTED) and the target language (English). The source text is: "Ő szép. Ő okos. Ő olvas. Ő mosogat. Ő épít. Ő varr. Ő tanít. Ő főz. Ő kutat. Ő gyereket nevel. Ő zenél. Ő takarító. Ő politikus. Ő sok pénzt keres. Ő süteményt süt. Ő professzor. Ő asszisztens." The target text is: "She is beautiful. He is clever. He reads. She washes the dishes. He builds. She sews. He teaches. She cooks. He's researching. She is raising a child. He plays music. She's a cleaner. He is a politician. He makes a lot of money. She is baking a cake. He's a professor. She's an assistant." The 'FRENCH' language option is highlighted with a red box. At the bottom right, there is a 'Send feedback' link.

Source: <https://twitter.com/DoraVargha/status/1373211762108076034?s=20>

Related Publication: <https://link.springer.com/article/10.1007/s00521-019-04144-6>

Real case examples

"Google Translate automatically chooses the gender for you"

"Inappropriate patient treatments by Watson AI"

Real case examples

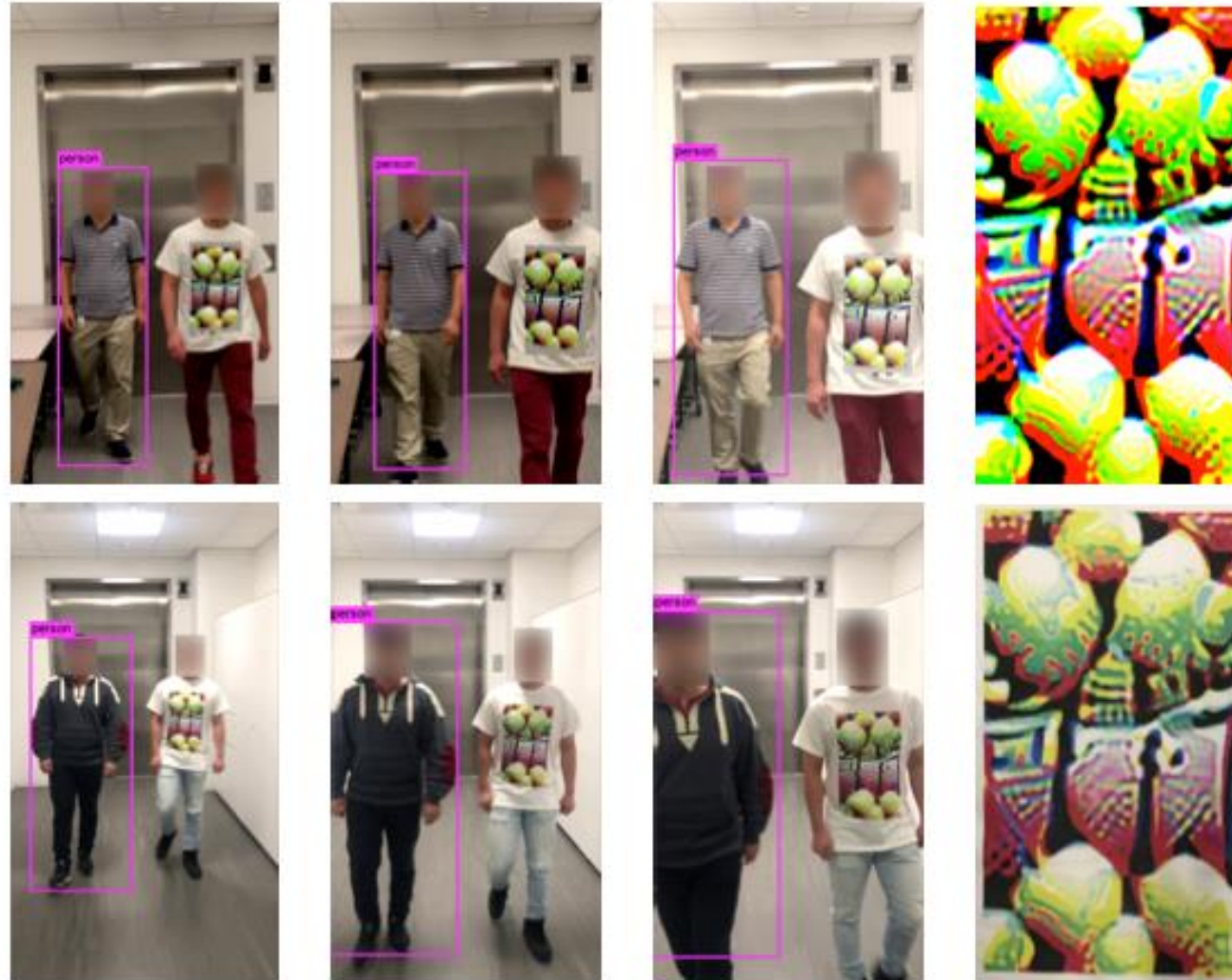
"Google Translate automatically chooses the gender for you"

"Inappropriate patient treatments by Watson AI"

"Trippy T-Shirt Makes You Invisible to AI"

Real case examples

"Google Translate automatically chooses the gender for you"



Source: <https://www.vice.com/en/article/evj9bm/adversarial-design-shirt-makes-you-invisible-to-ai>

Related Publication: <https://arxiv.org/pdf/1910.11099.pdf>

Real case examples

"Google Translate automatically chooses the gender for you"

"Inappropriate patient treatments by Watson AI"

"Trippy T-Shirt Makes You Invisible to AI"

"Google Allo suggested man in turban emoji as response to a gun emoji"

Real case examples

"Google Translate automatically chooses the gender for you"

"Inappropriate patient treatments by Watson AI"

"Trippy T-Shirt Makes You Invisible to AI"

"Google Allo suggested man in turban emoji as response to a gun emoji"

"Facebook Enabled Advertisers to Reach 'Jew Haters'"

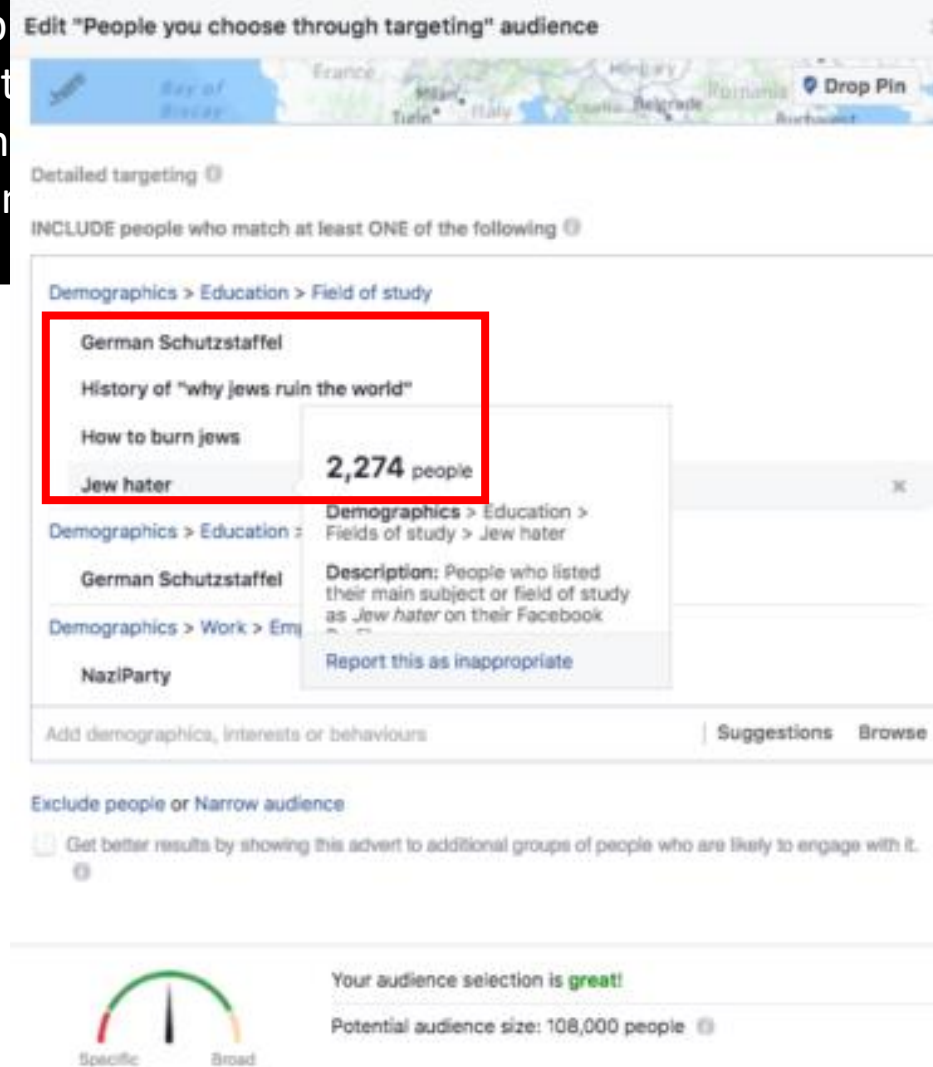
Real case examples

"Google Translate automatically chooses the gender for you"

"Inapp patient treatment Watson"

"Google Allo man emoji to i"

"Facebook Enabled Advertisers to Reach 'Jew Haters'"

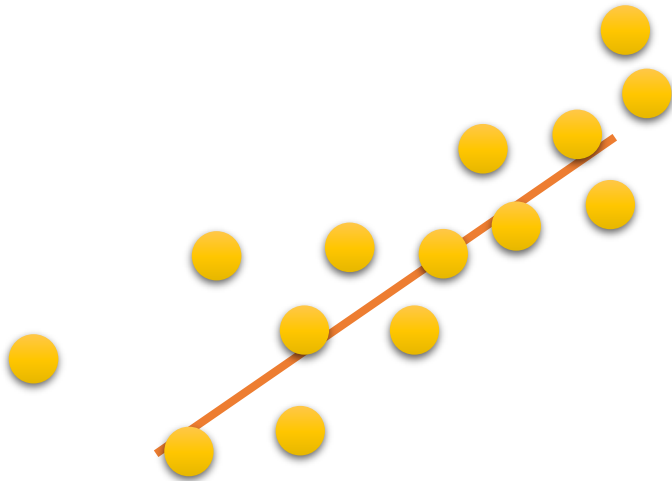


Source: <https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters>

What about IML?

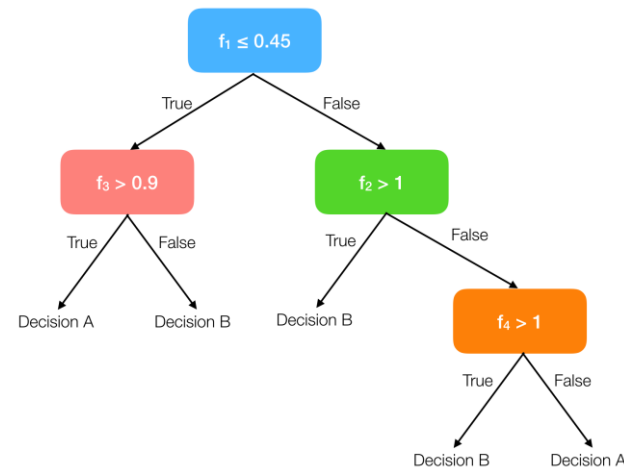
Interpretable ML (IML) refers to the ability of machine learning models to justify their decisions in a way people understand.

*Few machine learning model, which are either **transparent** or **black boxes***



Regression Model: $f(x_1) = w_1x_1 + \text{bias}$

#transparentModel



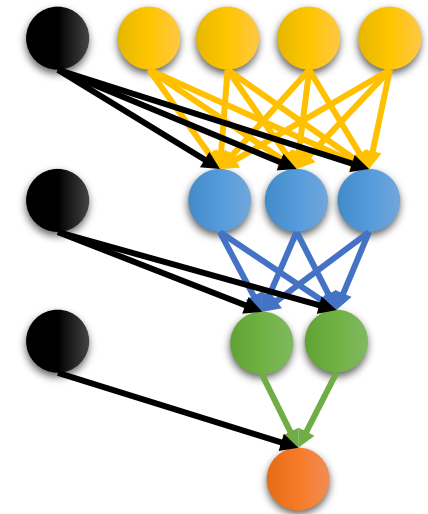
Decision Tree: if $f_1 \leq 0.45$ and $f_3 > 0.9$ then Decision A

#transparentModel



Random Forests: ?

#blackBoxModel



Neural Network: ?

#blackBoxModel

Shape of Interpretations

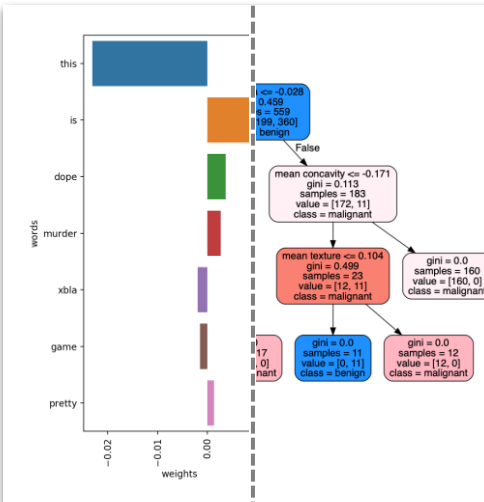
IF Proline \geq 990.0 THEN Wine=1
IF Color intensity \leq 3.85 AND
Color intensity \leq 3.52 THEN
Wine=2
IF Flavanoids \leq 1.41 AND
~~Proline \geq 170.0 THEN Wine=3~~

*"You were classified as cat
because you have pointy ears :)"*

Textual Form



Visual Form



Graphical Form

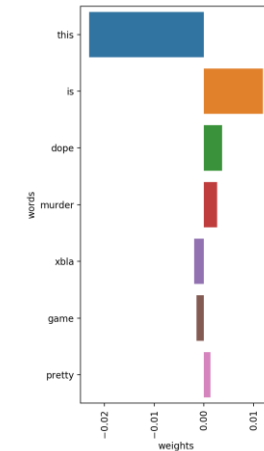
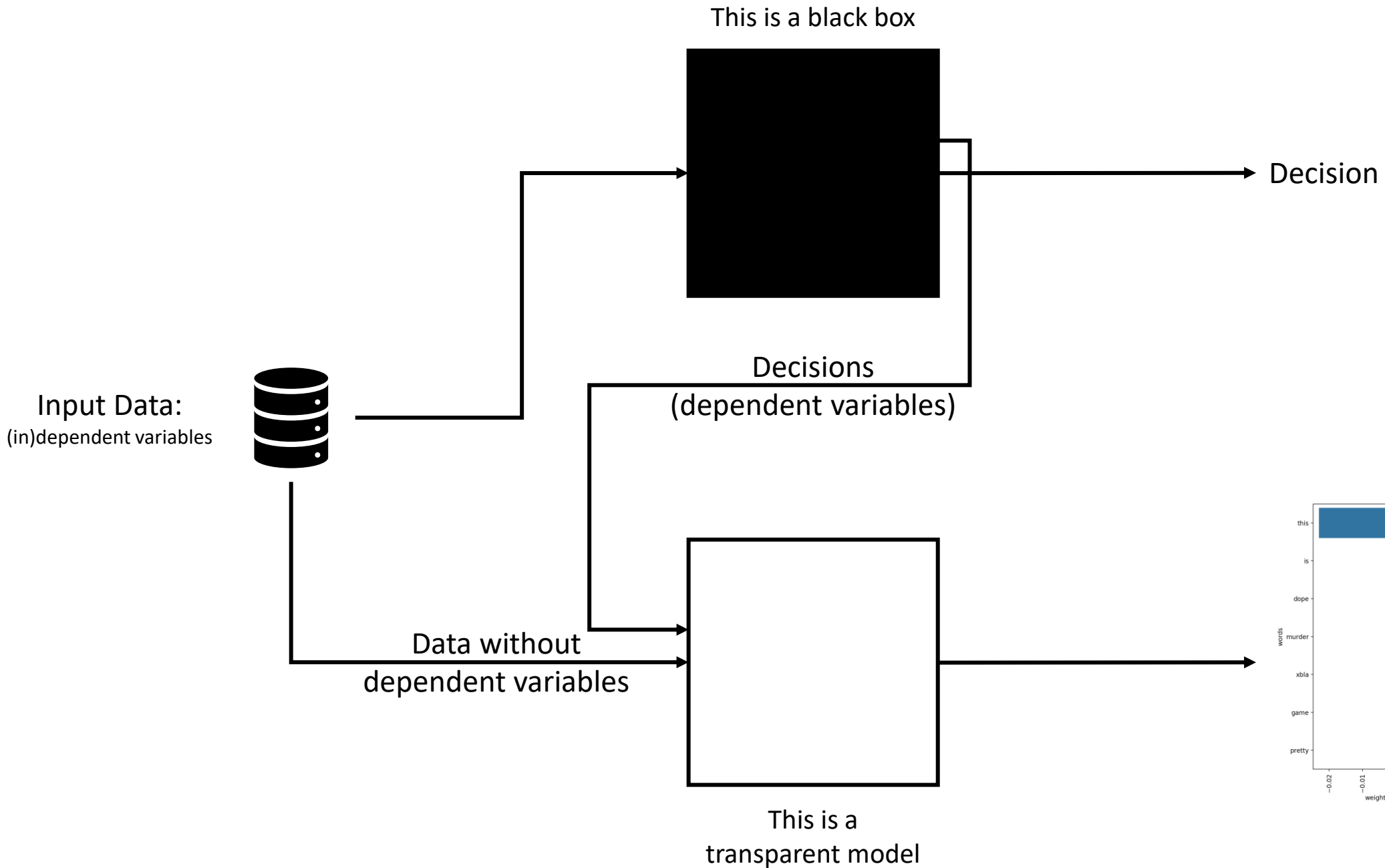


Dialectical Form

Dimensions of Interpretability Techniques



But what if we want to interpret a black box model?



Our work

Find most of our work in our
github repositories:

- <https://git.io/JOqkQ>
- <https://git.io/JOqk7>



Interpreting Neural Networks (LioNets, AutoLioNets, LioNets V2)

Interpreting Random Forests (Preliminary LionForests, LionForests, LionForests Bot)

Interpreting Predictive Maintenance Models (VisioRed)

Visio Red

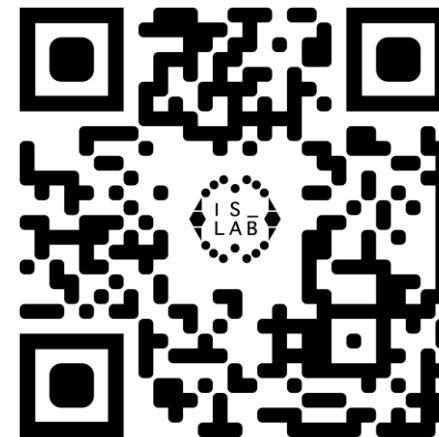
ALTRUIST

Argumentative Explanations & Truthfulness Evaluation (Altruist)

Demo Time

The End

Ioannis Mollas



Git Repository

These resources here:
<https://git.io/JOr3l>

*“Magic model on the core,
Explain yourself in front of all”*