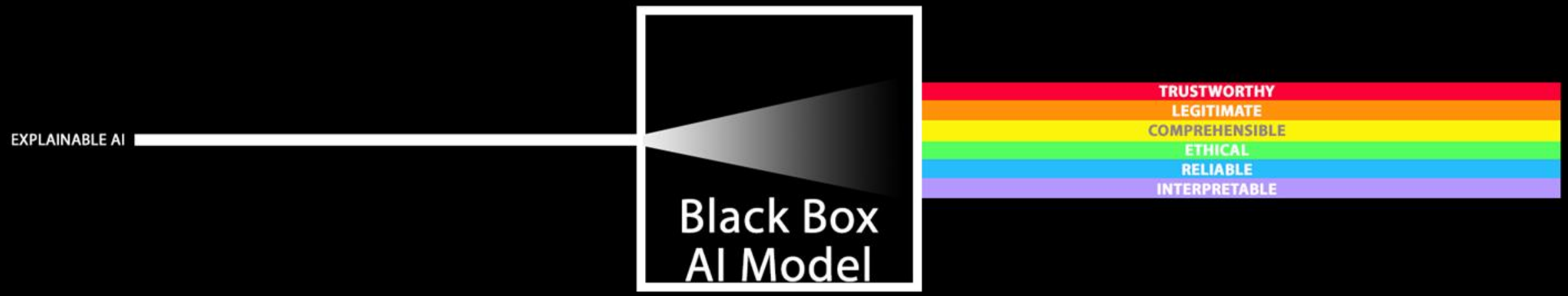
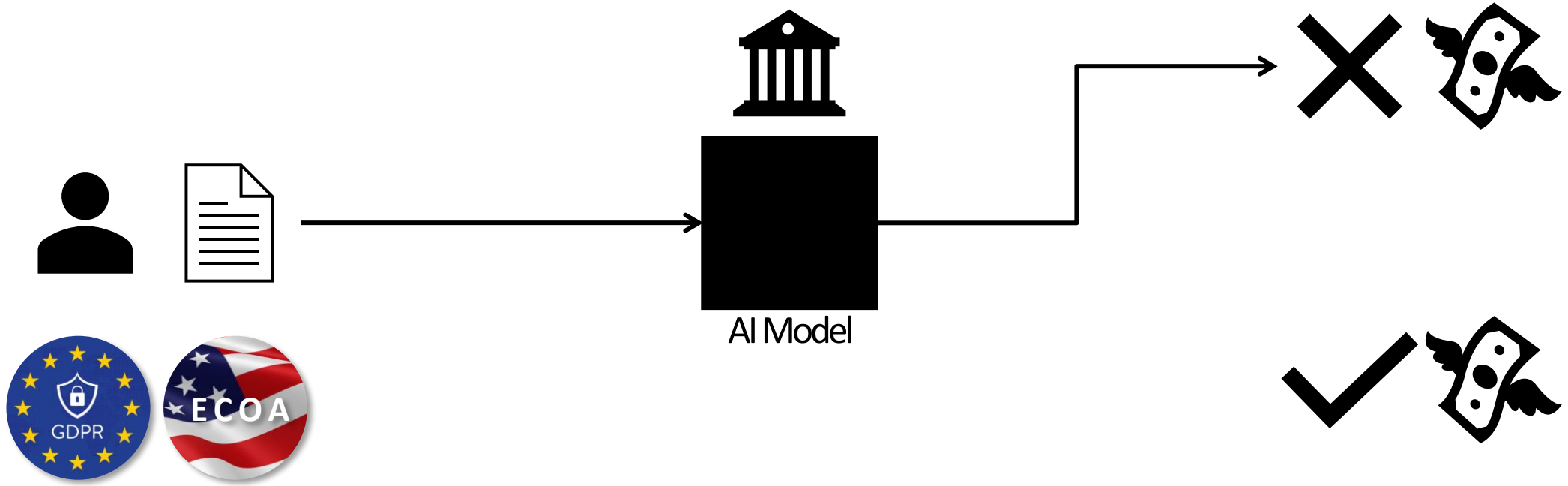


Explainable AI – Interpretable ML

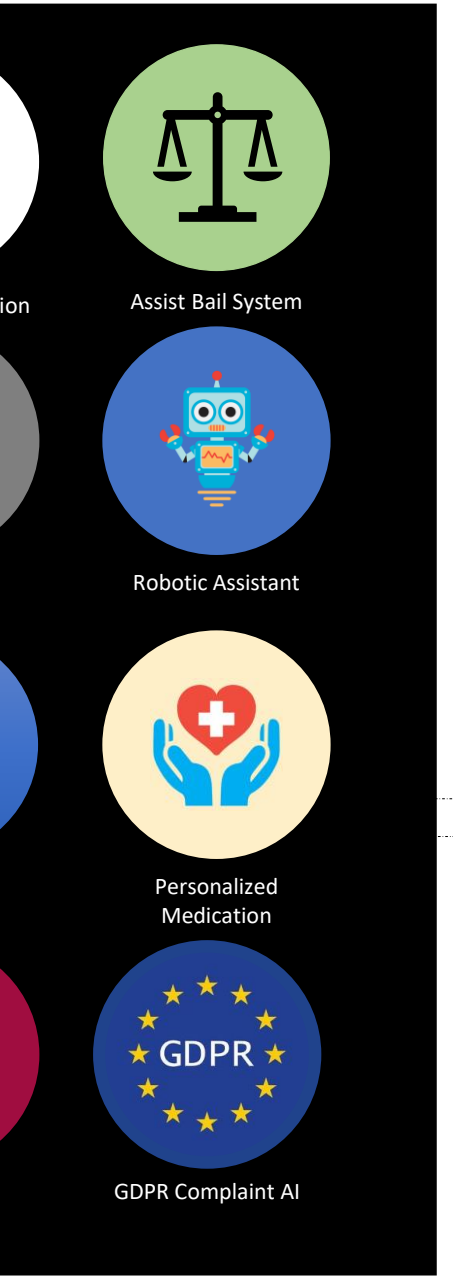
Ioannis Mollas, iamollas@csd.auth.gr, PhD Student
Supervisors: Nick Bassiliades, Grigorios Tsoumakas

Intelligent Systems Laboratory
CSD AUTH


GDPR
ECOA
AI Model

AI APPLICATIONS

Intelligent Systems Laboratory



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

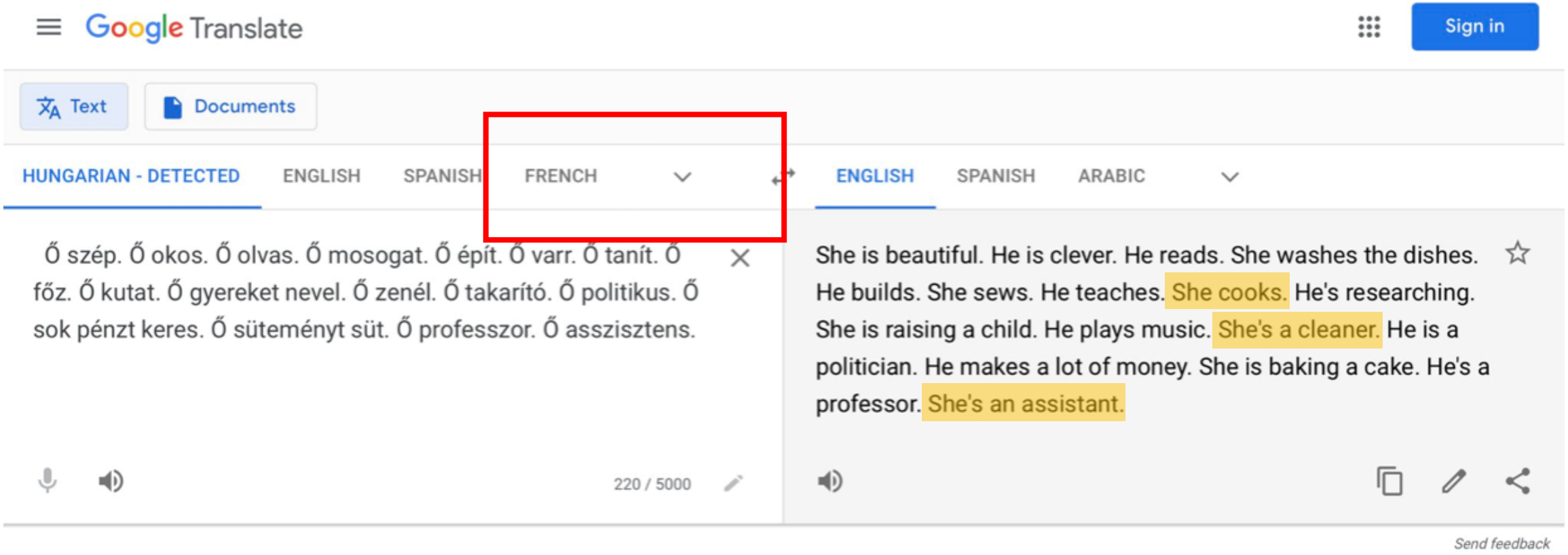
- Acquisition of new knowledge
- Robust systems
- Trustworthy systems

Real case examples

”Google Translate automatically chooses the gender for you”

Real case examples

”Google Translate automatically [redacted]]



Source: https://twitter.com/DoraVargha/status/1373211762108076034?s=20
Related Publication: https://link.springer.com/article/10.1007/s00521-019-04144-6

Intelligent Systems Laboratory

Real case examples

”Google Translate automatically chooses the gender for you”

“Inappropriate patient treatments by Watson AI”

Real case examples

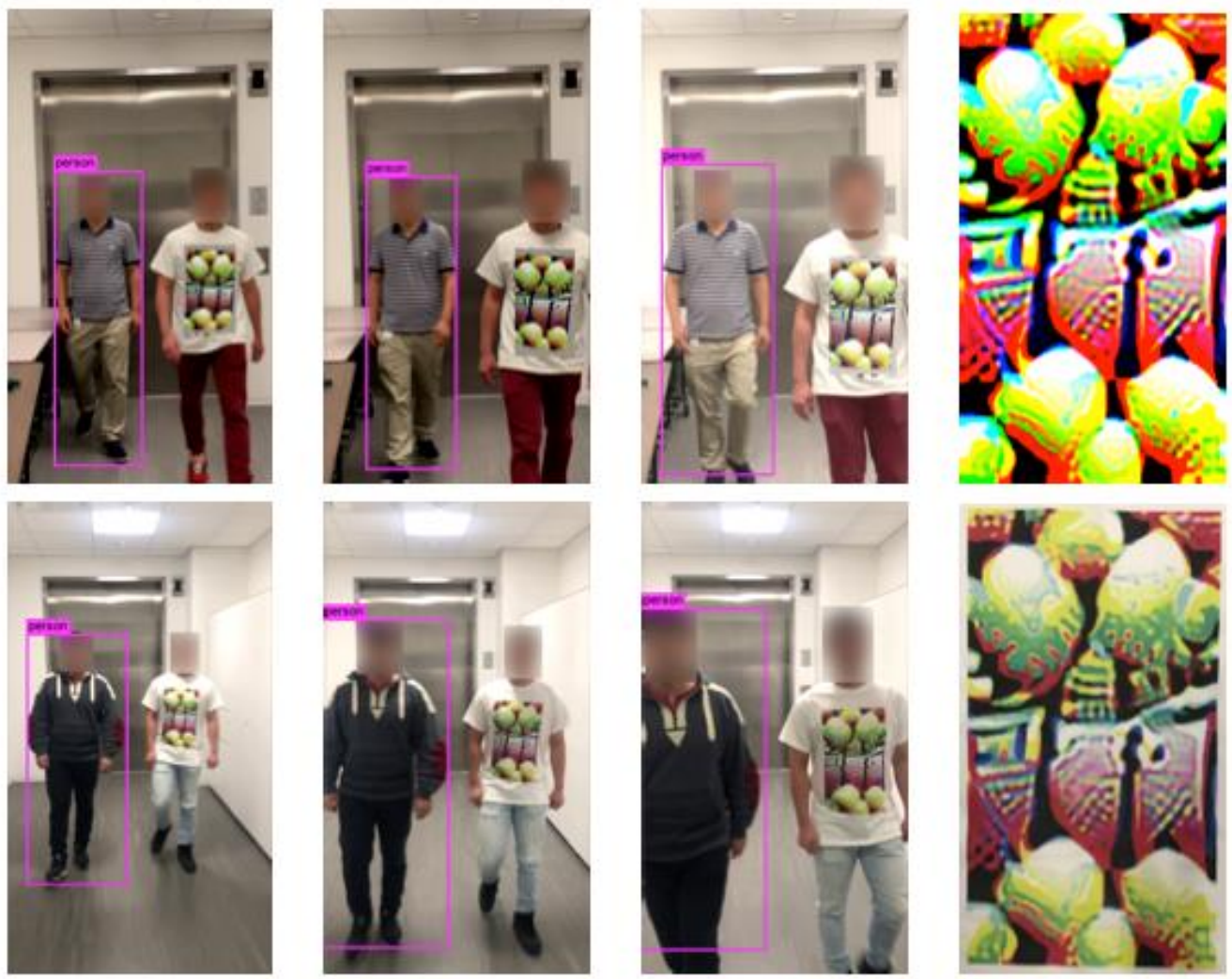
”Google Translate automatically chooses the gender for you”

“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’”

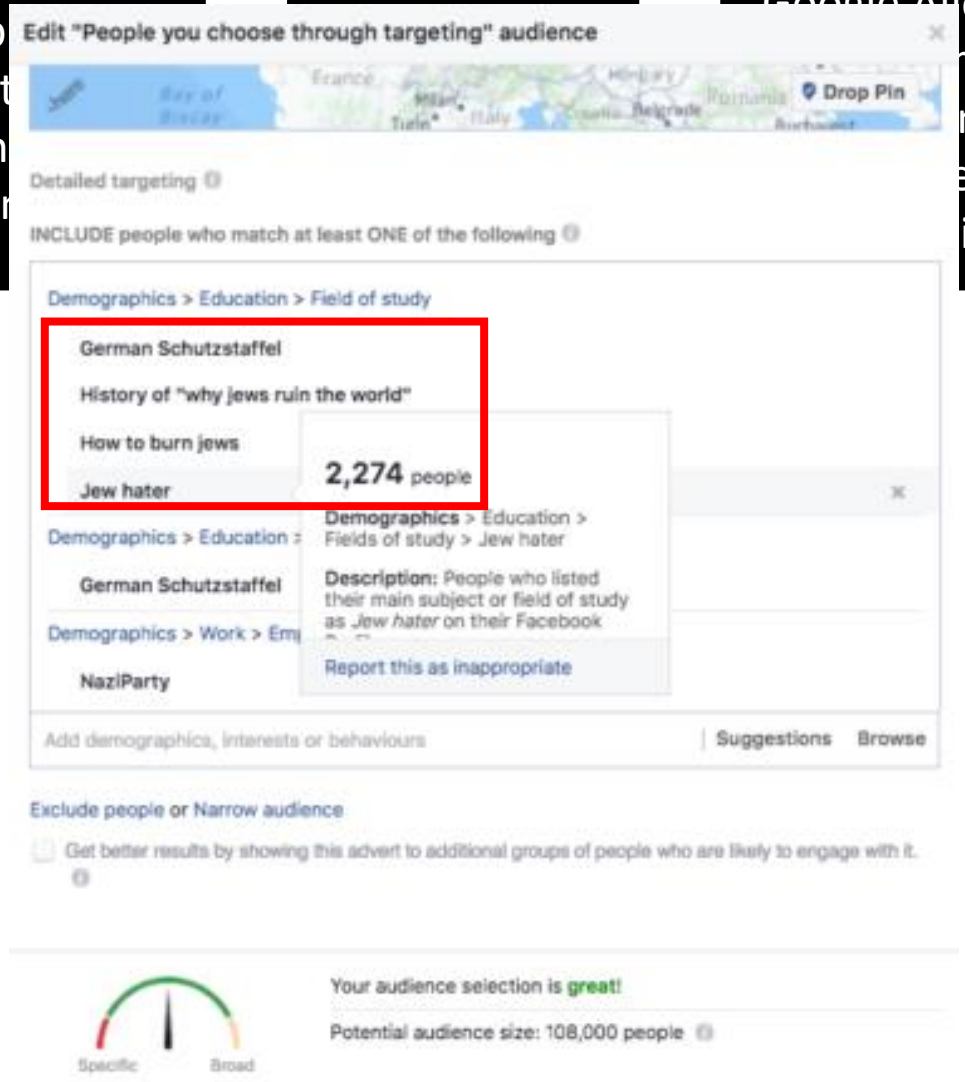
Real case examples

”Google Translate automatically chooses the gender for you”

“Inapp patient treatm Watson

“Google Allo man moji e to i”

“Facebook Enabled Advertisers to Reach ‘Jew Haters’”

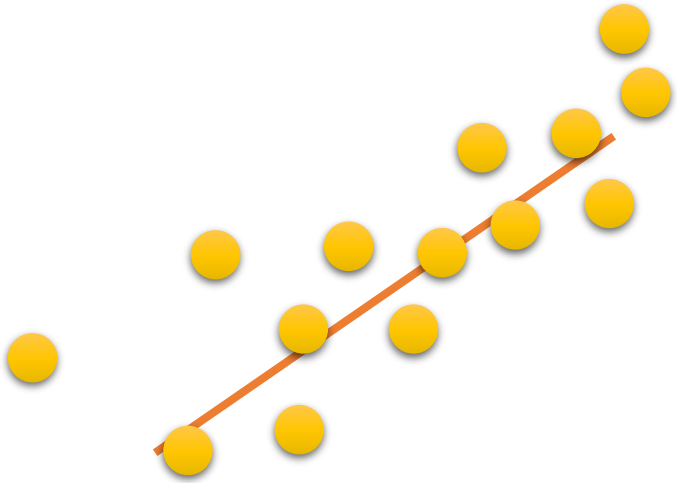


Source: https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters

What about IML?

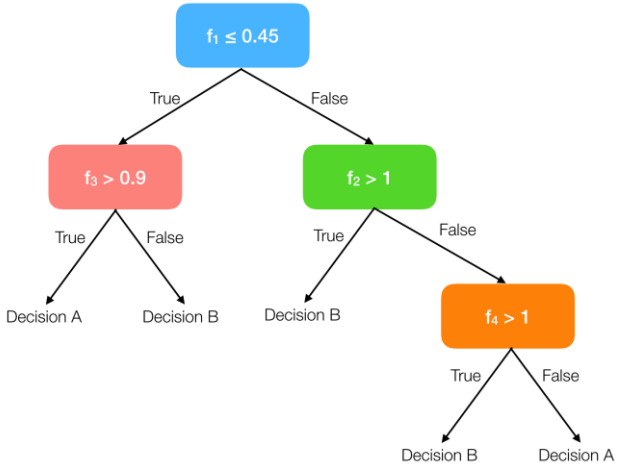
Interpretable ML (**IML**) refers to the ability of machine learning models to justify their decisions in a way people understand.

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



Regression Model: $f(x_1) = w_1 x_1 + bias$

#transparentModel

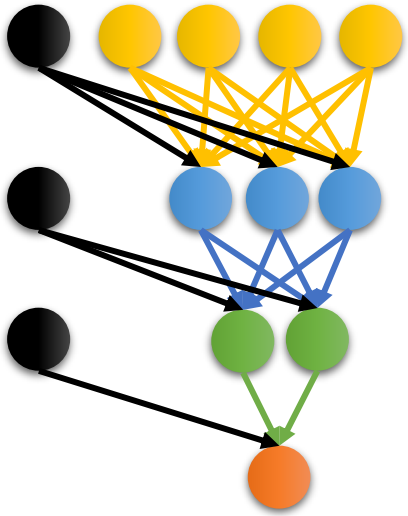


Decision Tree: $if\ f_1 \leq 0.45\ and\ f_3 > 0.9\ then\ Decision\ A$

#transparentModel

Random Forests: ?

#blackBoxModel



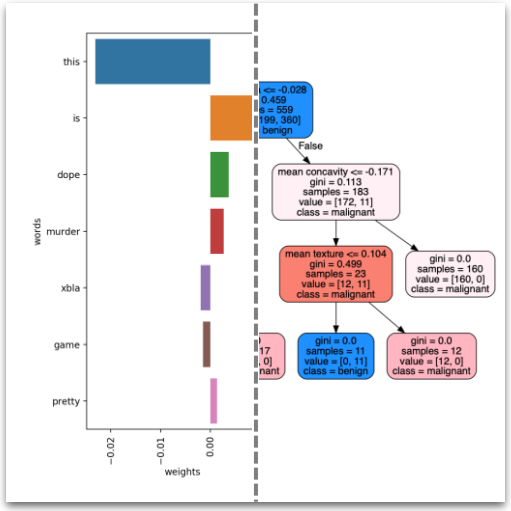
Neural Network: ?

#blackBoxModel

Shape of Interpretations

IF Proline>=990.0 THEN Wine=1
IF Color intensity<=3.85 AND Color intensity<=3.52 THEN Wine=2
IF Flavanoids<=1.41 AND Proline>=470.0 THEN Wine=3

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



Textual Form

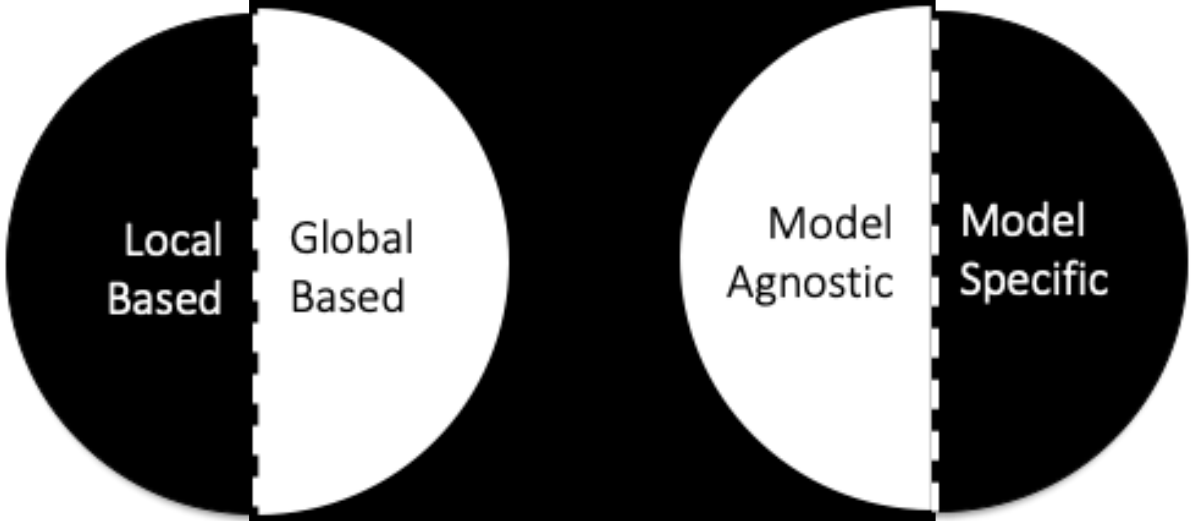
Visual Form

Graphical Form

Dialectical Form

Intelligent Systems Laboratory

Dimensions of Interpretability Techniques



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

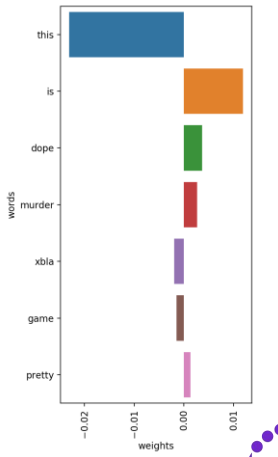
This is a black box

Decision

Input Data:
(in)dependent variables

Decisions
(dependent variables)

Data without
dependent variables



This is a
transparent model

Intelligent Systems Laboratory

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)

Intelligent Systems Laboratory

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”