Explainable Artificial Intelligence

Interrogating the AI systems

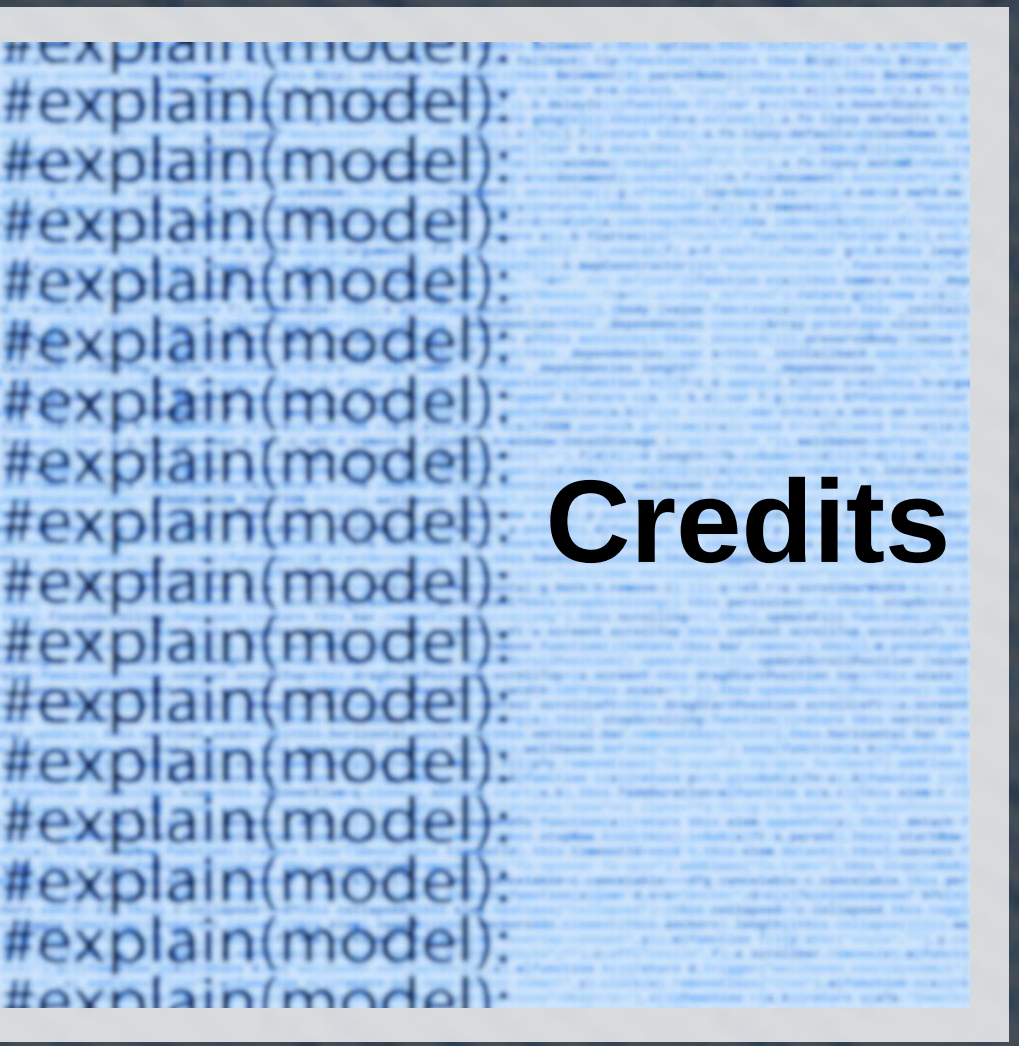
Nick Bassiliades, Professor
School of Informatics
Aristotle University of Thessaloniki, Greece
nbassili@csd.auth.gr

Talk @ Rules: Logic and Applications, 17/12/2019

A few words about the speaker

Nick Bassiliades
(Νικόλαος Βασιλειάδης)
http://intelligence.csd.auth.gr/people/bassiliades

- Professor, School of Informatics, Aristotle University of Thessaloniki, Greece
- Scientific specialization: Knowledge Systems
  - Knowledge Representation & Reasoning
    - Rule-based systems, Logic programming, Defeasible Reasoning, Knowledge-based / expert systems
  - Semantic Web
    - Ontologies, Linked Open Data, Semantic Web Services
  - Multi-agent systems
    - Reputation/trust, knowledge-based interaction (argumentation, negotiation, brokering)
  - Applications on e-Learning, e-Government, e-Commerce, Electric Vehicles



Ioannis Mollas, PhD student

- Explainable AI (mostly works on explainable Machine Learning)
- Credits for the most part of the talk, including the fancy looks
- Funded by project AI4EU (A European AI On Demand Platform and Ecosystem)
  - EU Horizon 2020 research and innovation programme under grant agreement No 825619
  - https://www.ai4eu.eu/

A few words about my institution

- Aristotle University of Thessaloniki, Greece
  - Largest (?) University in Greece and South-East Europe
  - Since 1925, 41 Schools, ~1.8K faculty, ~45K students
- School of Informatics
  - Since 1992, 30 faculty, 3 departments, 6 research labs, ~1500 undergraduate students, ~180 MSc students, ~110 PhD students, ~160 PhD graduates, >5000 pubs
- Intelligent Systems Laboratory (http://intelligence.csd.auth.gr)
  - 4 faculty, 18 PhD students, 4 post-graduate affiliates, 19 PhD graduates
  - Research on Artificial Intelligence, Machine Learning / Data Mining, Knowledge Representation & Reasoning / Semantic Web, Planning, Multi-Agent Systems
  - >470 publications, >40 projects

ARISTOTLE UNIVERSITY OF THESSALONIKI

SCHOOL OF INFORMATICS

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 825619.

Our Team

Ioannis Vlahavas
Professor

Nick Bassiliades
Professor

Grigorios Tsoumakas
Assistant Professor

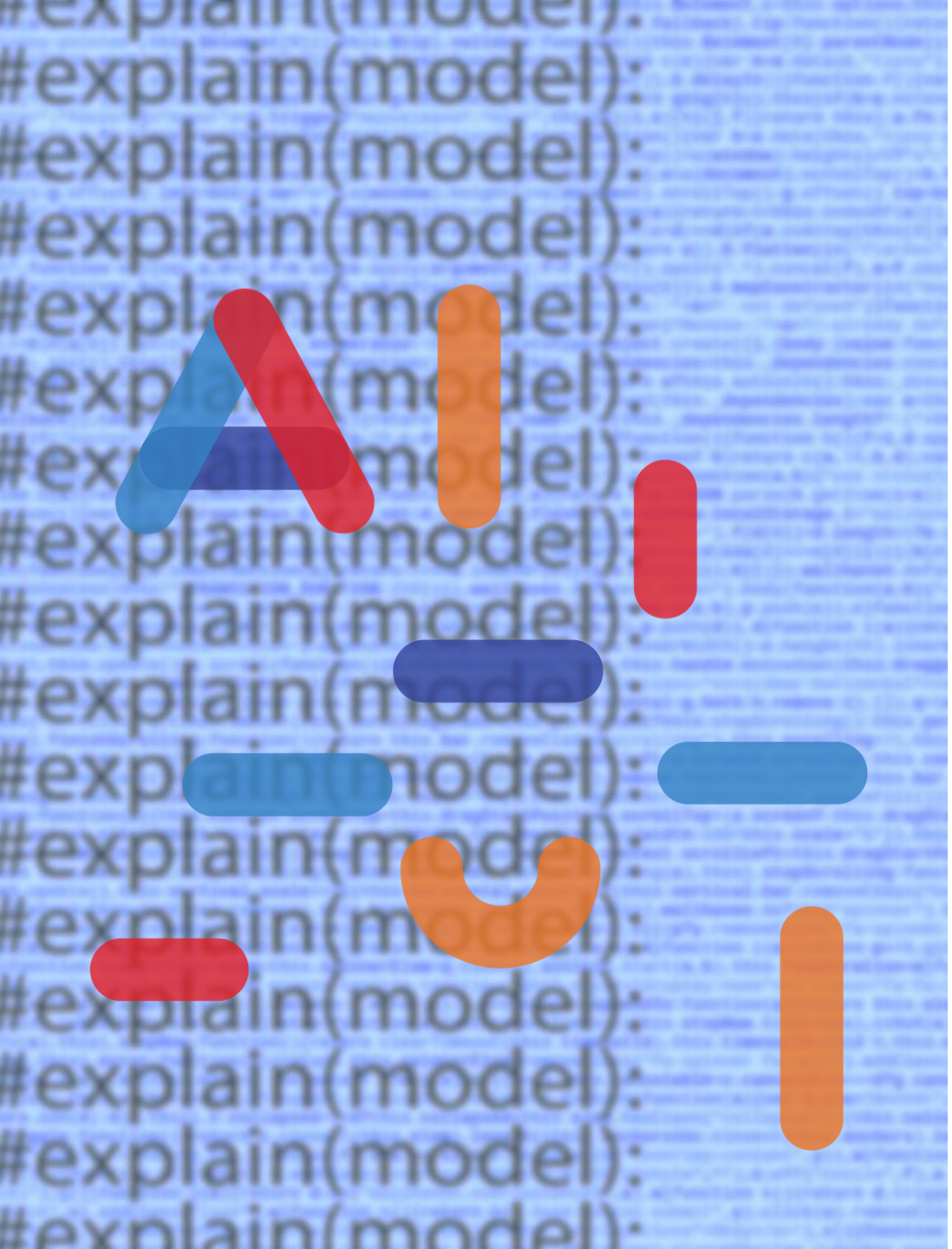
Ioannis Mollas
PhD Student

AI4EU

ARISTOTLE UNIVERSITY OF THESSALONIKI

Intelligent Systems

SCHOOL OF INFORMATICS



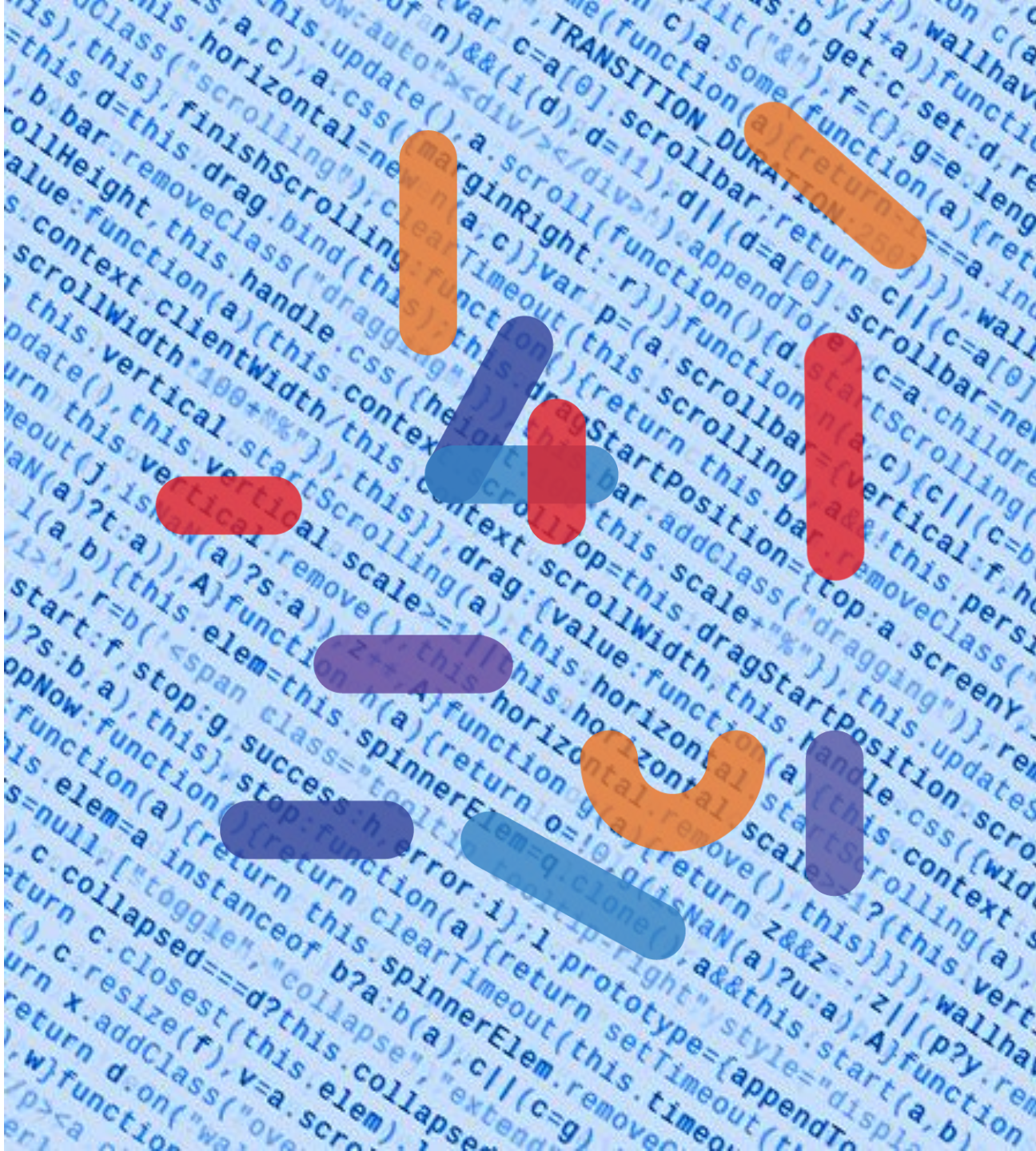
About AI4EU

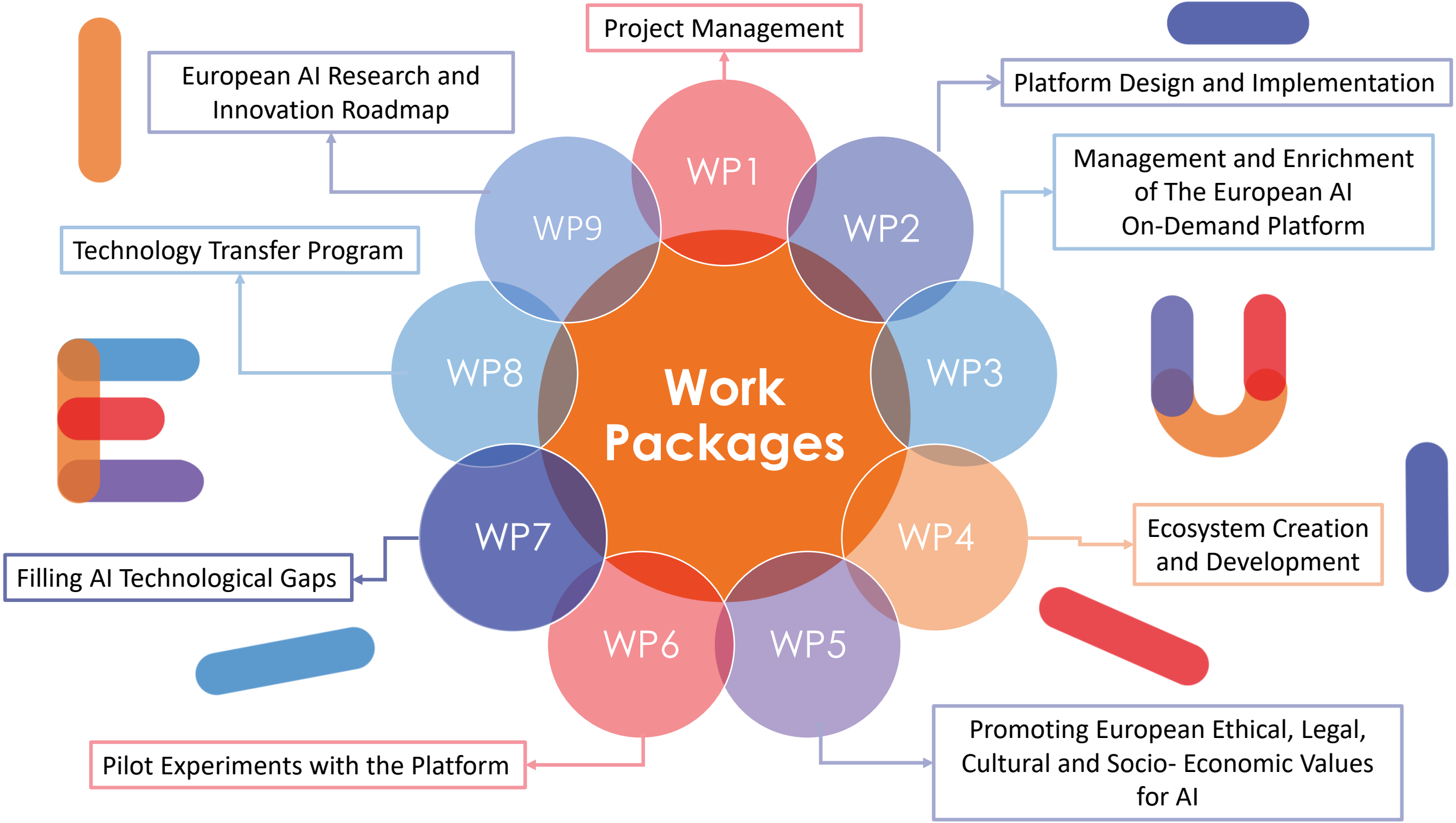
- European Union’s landmark Artificial Intelligence project
- AI ecosystem containing knowledge, algorithms, tools and resources
- 80 partners, 21 countries
- 3 years project
- €20m budget

Goals

Create a leading collaborative AI European platform

- Open and sustainable **AI On-Demand-Platform**
- Unite stakeholders via **high-profile conferences and virtual events**
- Develop a **Strategic Agenda for European AI**
- Establish an **Ethics Observatory** to ensure development of human-centred AI
- Roll out of €3m in **Cascade Funding**




Project Management
European AI Research and Innovation Roadmap
Platform Design and Implementation
Management and Enrichment of The European AI On-Demand Platform
Technology Transfer Program
WP1
WP2
WP3
WP4
WP5
WP6
WP7
WP8
WP9
Work Packages
Ecosystem Creation and Development
Filling AI Technological Gaps
Pilot Experiments with the Platform
Promoting European Ethical, Legal, Cultural and Socio- Economic Values for AI

WP7

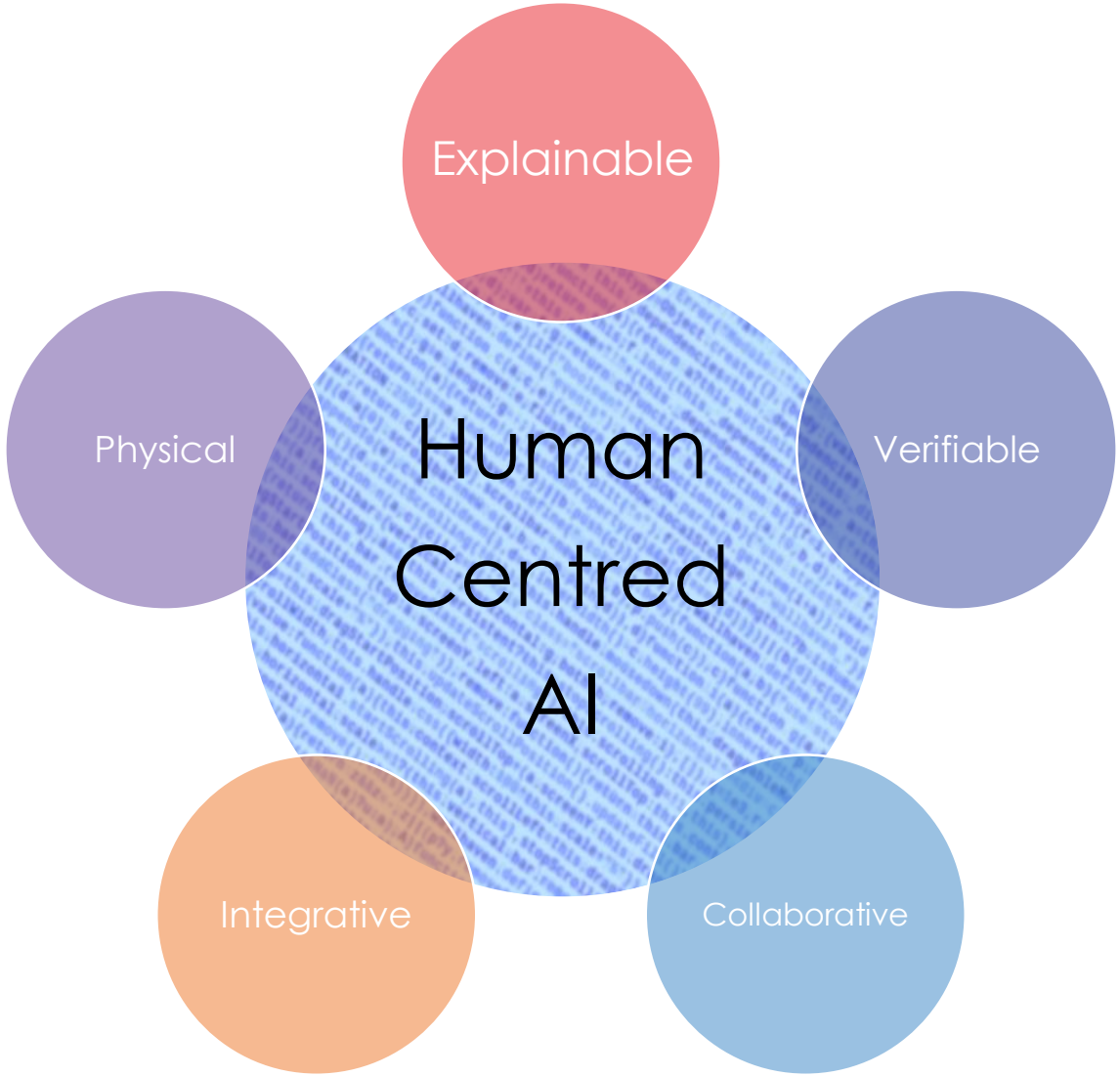
Filling AI Technological Gaps

This WP will reinforce European *excellence* and *leading* position worldwide in major AI research and application domains, through research and innovation efforts to fill important technology gaps.

Objectives:

1. Develop new AI tools and techniques (to be included in the AI4EU Platform)

2. Consolidate and strengthen excellence in AI in EU


Explainable
Physical
Human Centred AI
Verifiable
Integrative
Collaborative

Task 7.1 Explainable AI

Explainable
Interpretable
Comprehensible
Intelligible, Justifiable
Understandable

Definition: An AI system should allow humans to understand the reasons behind its recommendations or decisions. It should be possible to know the data, rationale and arguments that lead to a result, to question them and to correct them

Tasks:

- ✓ Community creation and aggregation activities
  - ❖ Summer School covering explainable AI
  - ❖ Workshops and tutorials
  - ❖ …
- ✓ Research activities
  - ❖ Survey for explainable AI
  - ❖ Methodology and software components
  - ❖ …

Partners:

AUT — Intelligent Systems, Department of Informatics, Aristotle University of Thessaloniki

ORU — Örebro University

TRT — THALES

UCC — University College Cork, Ireland, Coláiste na hOllscoile Corcaigh

USO — Sorbonne Université

Back to XAI

XAI: Explainable Artificial Intelligence

Explainable AI (**XAI**) refers to methods and techniques in the application of artificial intelligence technology (AI) such that the results of the solution can be understood by human experts.

Outline

- ❑ Why XAI?
- ❑ Interpretable Machine Learning
  - ✓ Transparent and Black Box Models
  - ✓ Explainable ML
  - ✓ Types of Data
  - ✓ Opening the Black Box
  - ✓ Implementations

Where we need Explainable AI

Health Care &
Personalized Medicine

Self Driving
Cars & Smart Cities

Credit Scores &
Financial Information

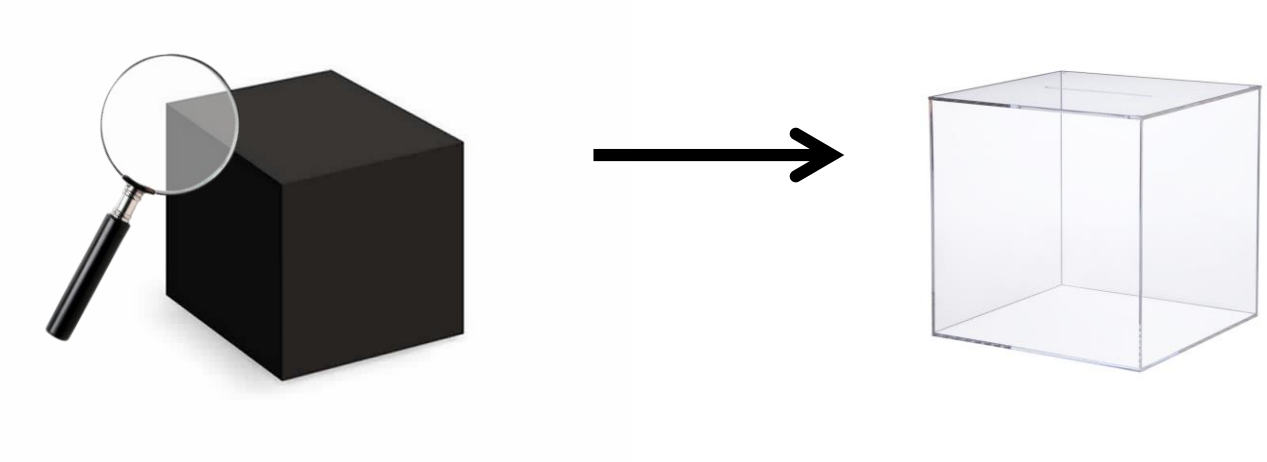
Robotic Assistants

GDPR Regulation:
'Right to Explanation"

User's Trust,
Miss Classification,
Result Understanding

«At the individual level, designers have both ethical and legal responsibilities to provide such justification for decisions that could result in death, financial loss, or denial of parole.»,
Don Monroe 2018[3]

Interpretable Machine Learning



"A way to present the result of a black box model in human readable terms"[1]

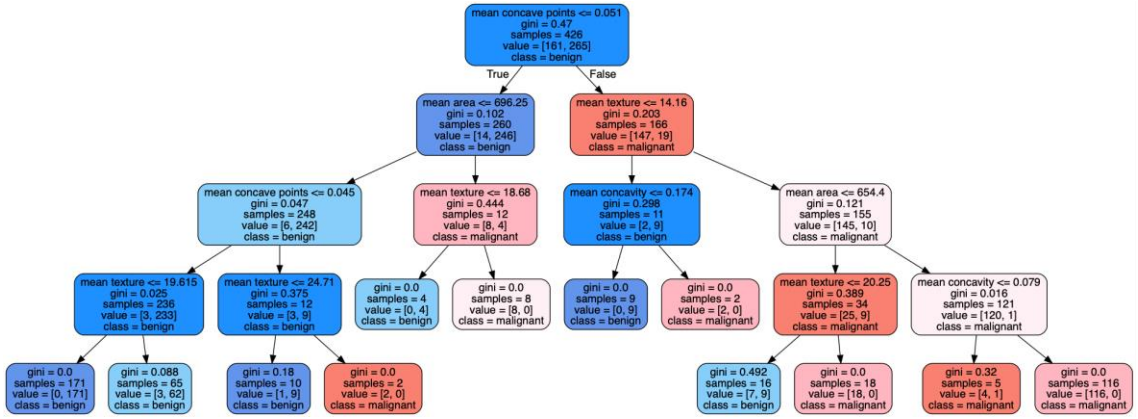
Each community addresses this meaning from a different perspective

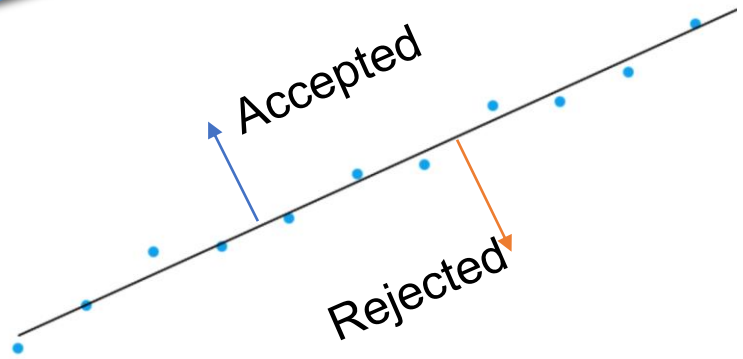
Recognized Interpretable Models

Or Transparent Boxes:

- Decision Trees
- Rule Based Systems
- Decision Tables
- Linear Models
- Decision Sets
- K-NN

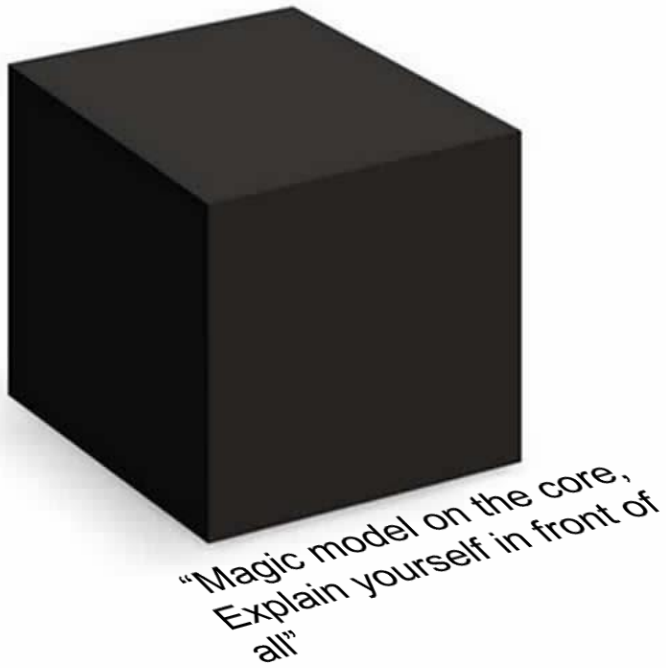
IF Proline>=990.0 THEN Wine=1
IF Color intensity<=3.85 AND Color intensity<=3.52 THEN Wine=2
IF Flavanoids<=1.41 AND Proline>=470.0 THEN Wine=3
IF Proline>=680.0 AND Alcohol>=12.93 THEN Wine=1
IF Hue>=0.69 THEN Wine=2
IF TRUE THEN Wine=2





	Rule 1	Rule 2	Rule 3	Rule 4
Conditions				
Valid User Name	F	T	T	T
Valid Password	-	F	T	T
Adequate balance in the account	-	-	F	T
Actions				
Login accepted	F	F	T	T
Amount transferred	-	-	F	T

Black Box Predictor... Means?

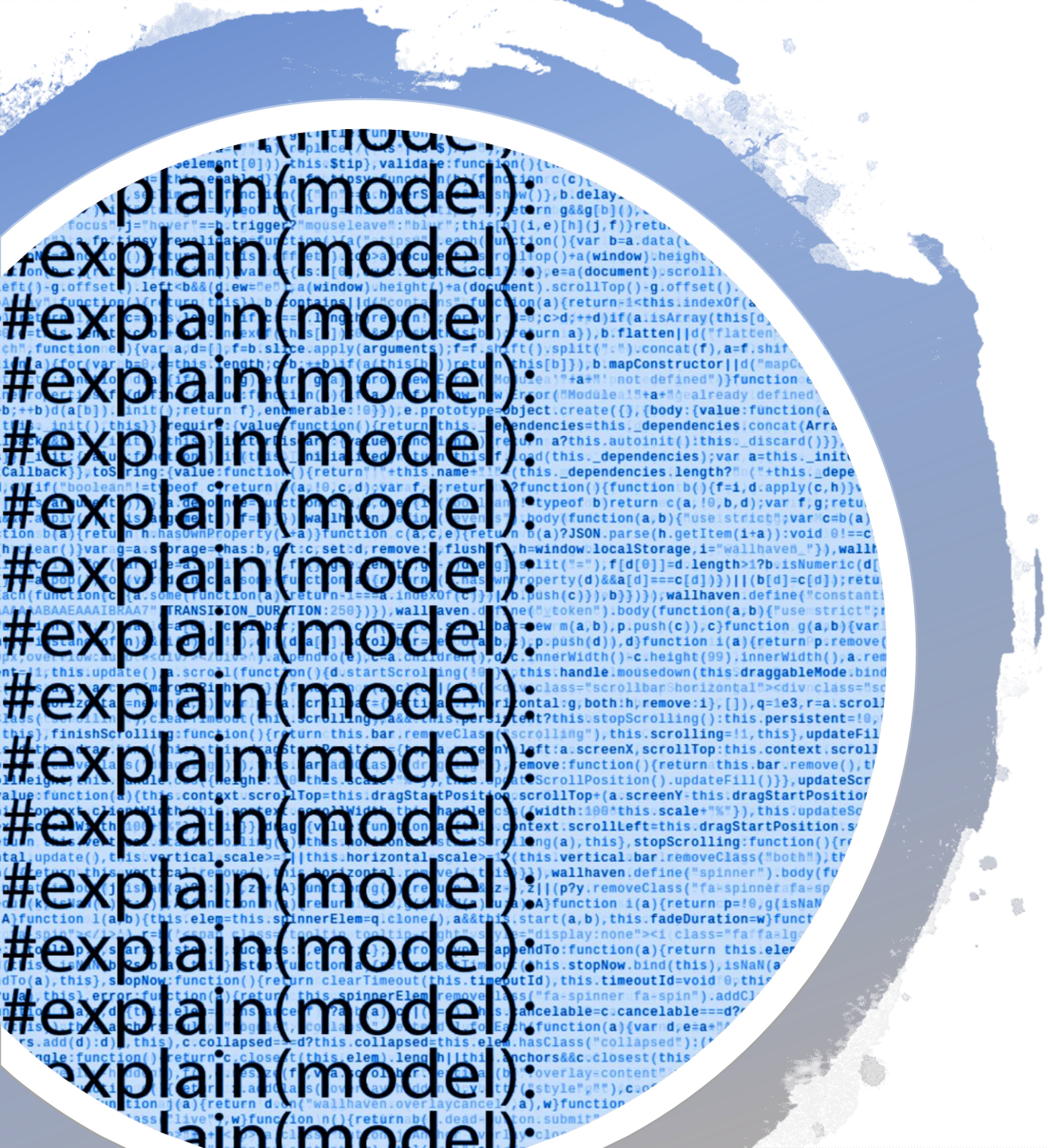


A black box predictor is the result of a machine learning algorithm, whose internals are:

- known to his creator or
- unknown to his creator

and uninterpretable to other people.

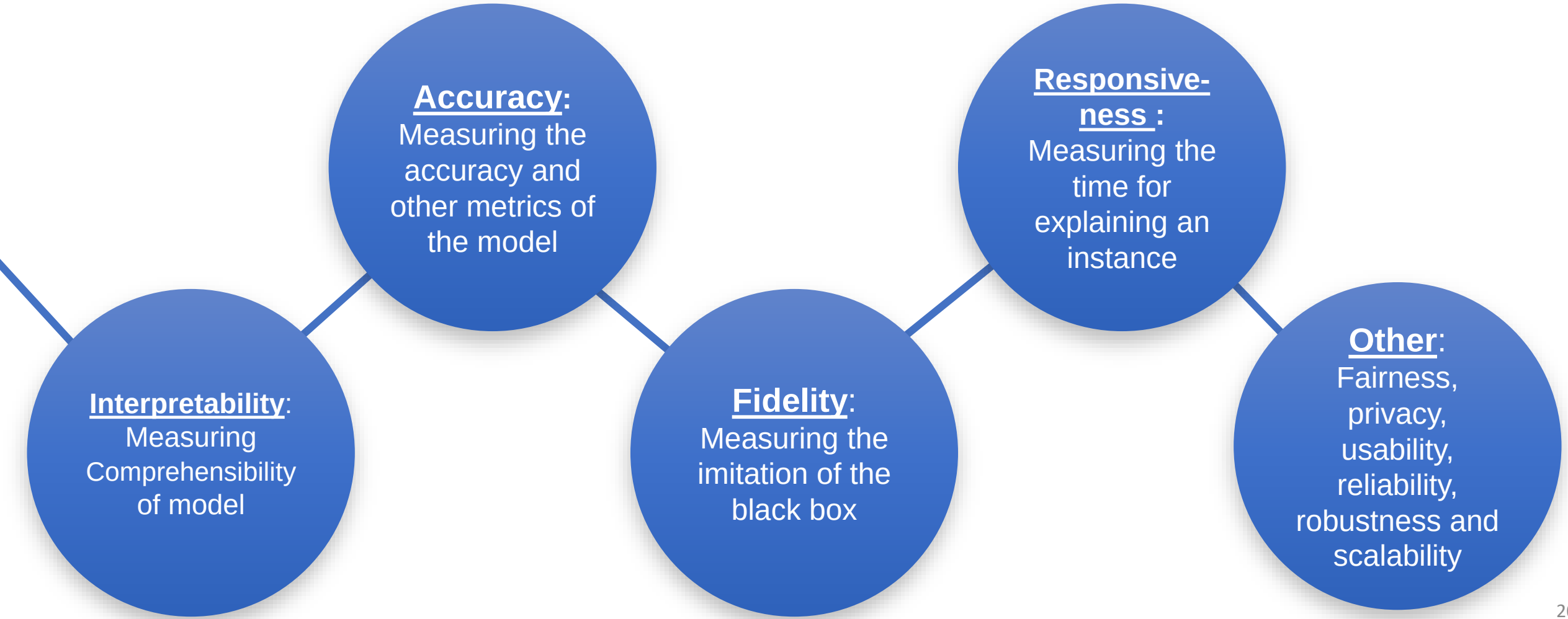
Some known black box models are: Support Vector Machines, Neural Networks, Tree Ensemble, Deep Neural Networks and Non-Linear Models



Dimensions of Interpretability

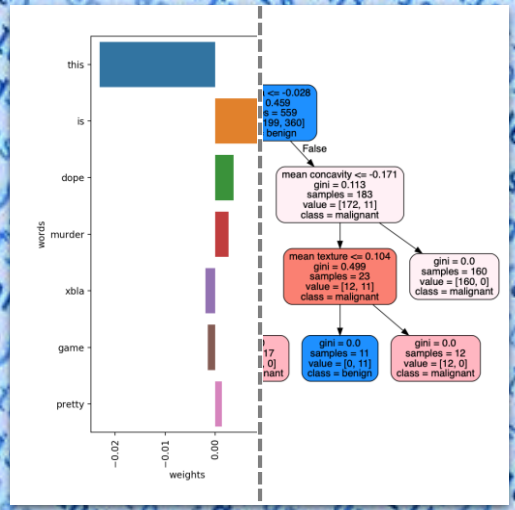
1. Is it Global or Local?
2. Has Time Limitation?
3. Which is the Nature of User Expertise?
4. Which is the Shape of Explanation?

Desired Features of Explainable Model



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

Shapes of Explanations

Comprehensibility Measure[2,6]

- ✓ Human Centered Evaluation
- ✓ Number of regions (parts of the model)
- ✓ Number of conditions (in a rule)
- ✓ Number of non-zero weights (in linear models)
- ✓ Depth of the tree (in a decision tree)

Simplification Methods like pruning are used in order to enhance the comprehensibility and to avoid overfitting.

But do we want to avoid overfitting when explaining?

Comprehensibility Measure[2,6]

If f1 > 20 and f2 <= 123 and f4 > 0.5 and f4 <= 1 and f5 <= 0 and f6 > 9 and f7 <= 10 and f8 <= 1 and f8 > 0.7 and f9 <= 19 and f10 > 10 and f10 <= 100 and f11 <= 1.23 and f12 > 12.1452 and f12 <= 15 and …….. then "class A"

vs

You classified as "class A" because 0.5 < f5 <= 1 and f9 < 19.

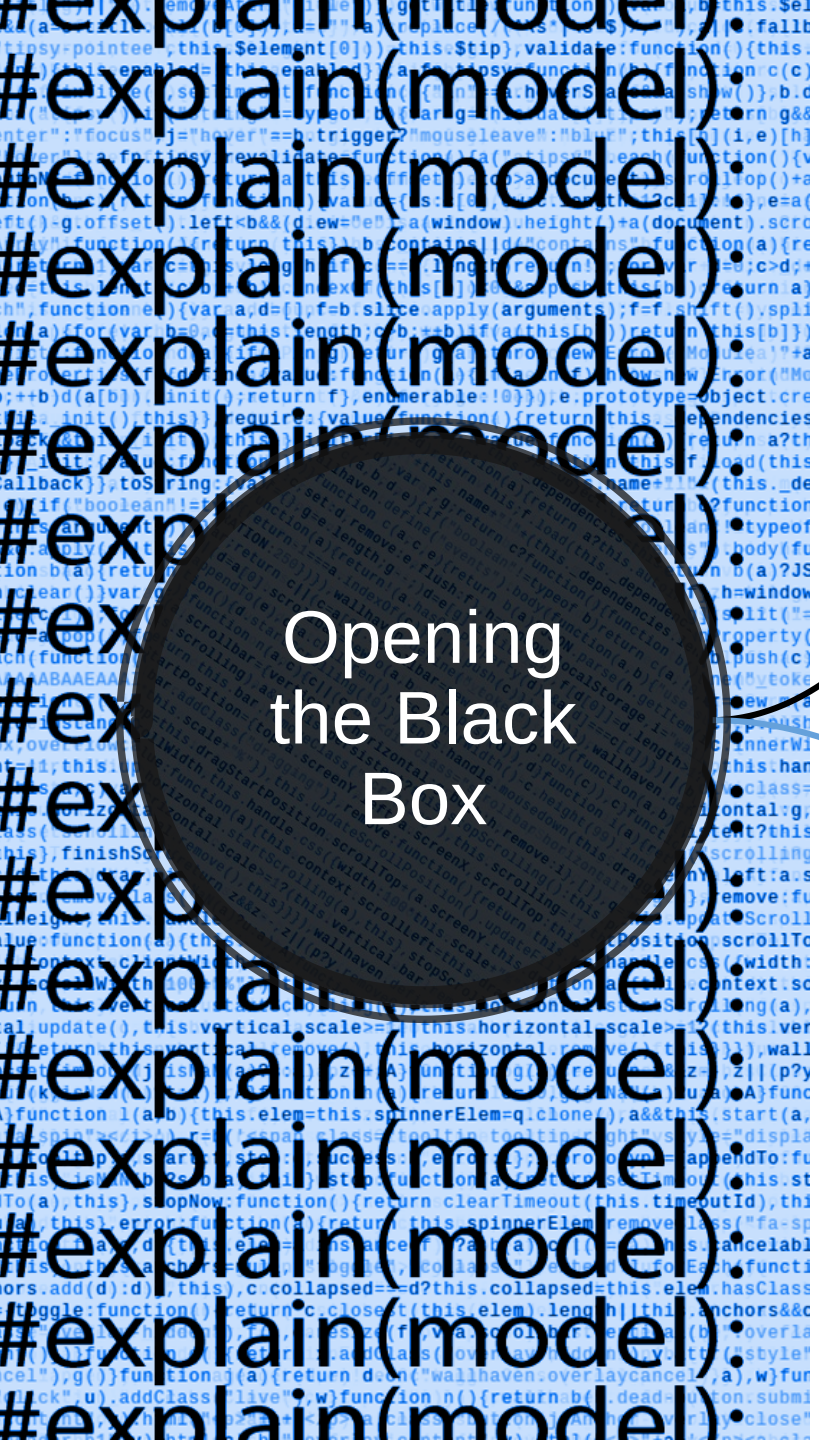
Which one would you prefer?

Types of Data

Tabular Data

Image Data

Text Data



Reverse Engineering on Black Box

- Model Explanation
- Outcome Explanation
- Model Inspection

Explanation for Fixed Dataset **or** Capable to Explain new Examples

Agnostic Explanators?

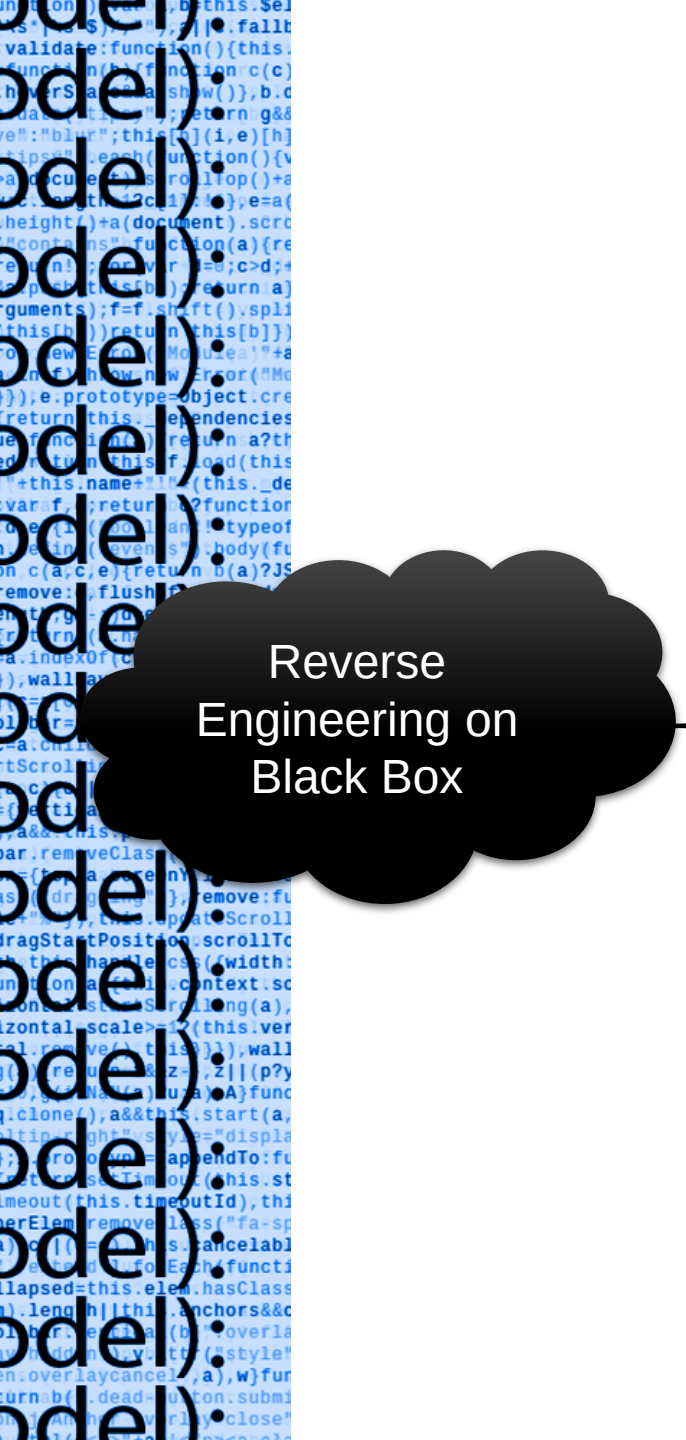
Replace with Transparent Box

- Can explain indifferently any kind of Black Box in most of the papers

- Most of the times, It has a specific type of data (Tabular, Images, Texts, Other), but rarely It could be agnostic on type of data too

Some Advantages of Agnostic Explanators:

- *They can explain models, even without dataset knowledge*
- *They can explain models remotely*



Reverse Engineering on Black Box

Model Explanation:

Is aimed at understanding the overall logic behind the black box. Provides a global explanation, through a transparent box.

Examples: Surrogate Method Explanators, Knowledge Distillation

Outcome Explanation:

Is aimed at understanding the outcome of an instance. Provides a local explanation, through a transparent box.

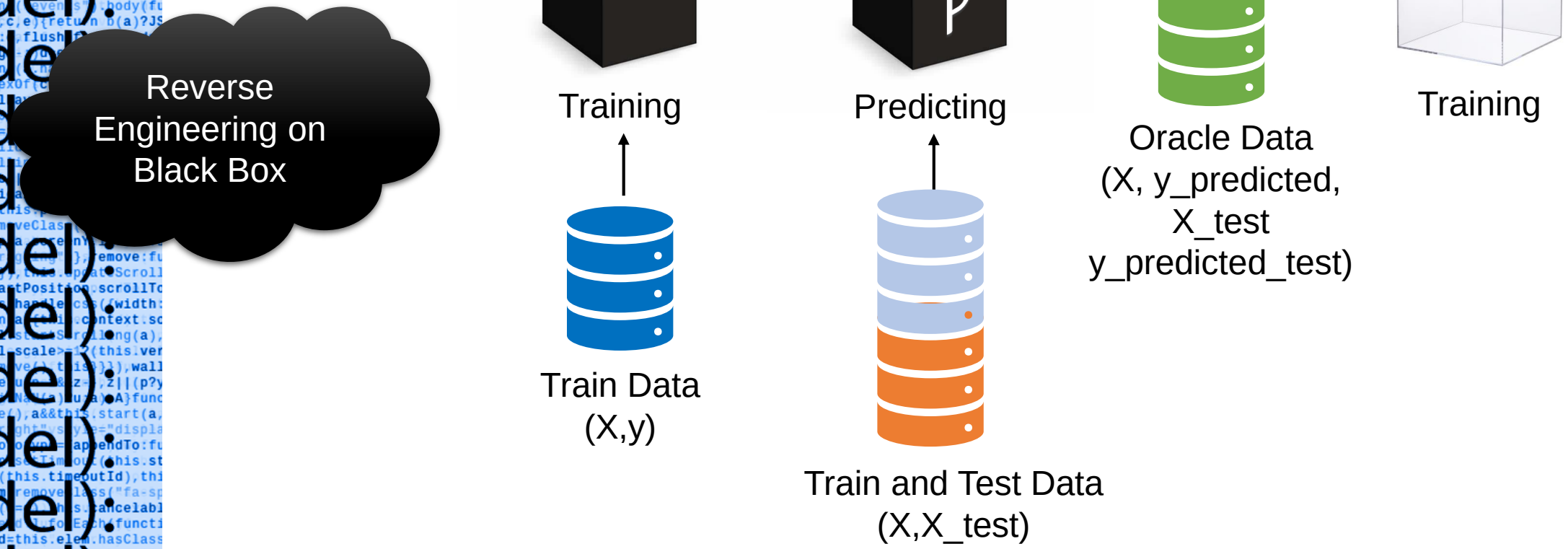
Examples: Lime, Anchors

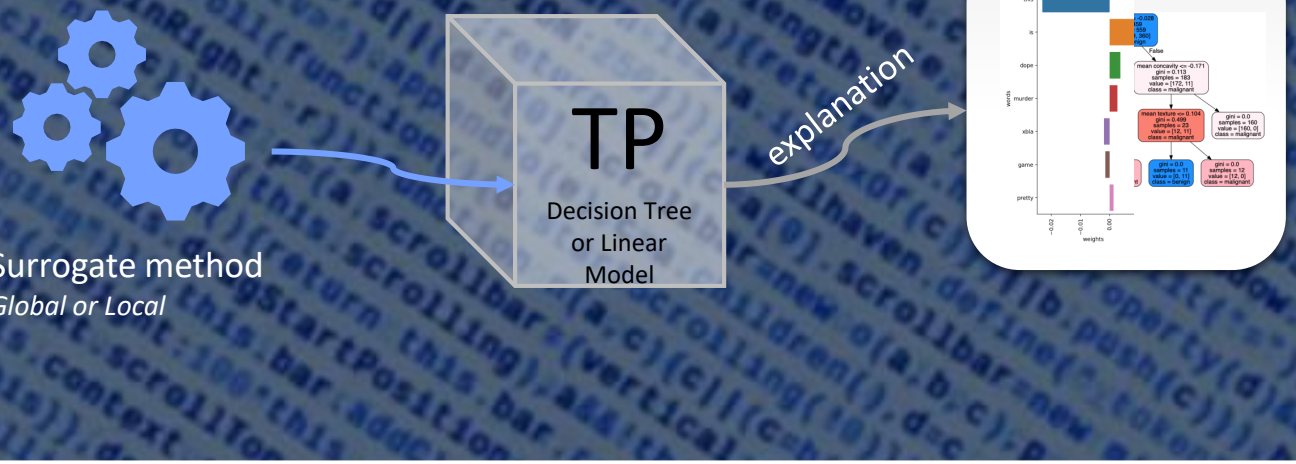
Model Inspection:

Provides a visual or textual representation of some specific properties.

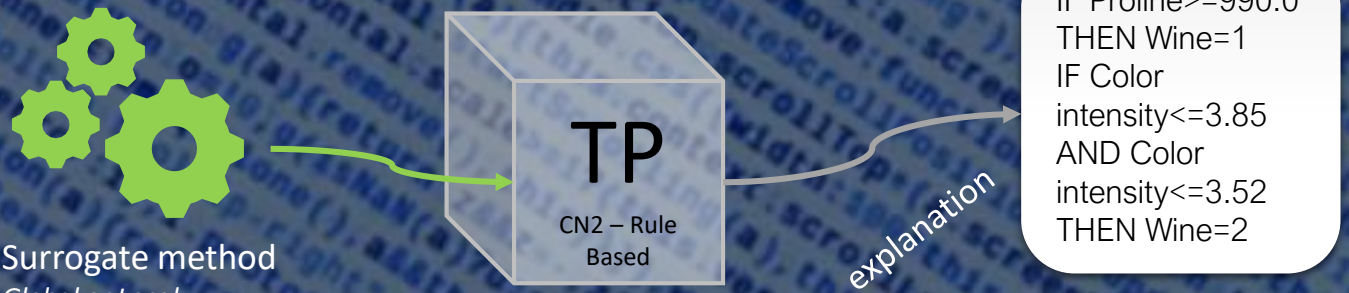
Examples: Permutation Importance, PDP, ICE, SHAP

Surrogate or Meta-Learner Method[7]

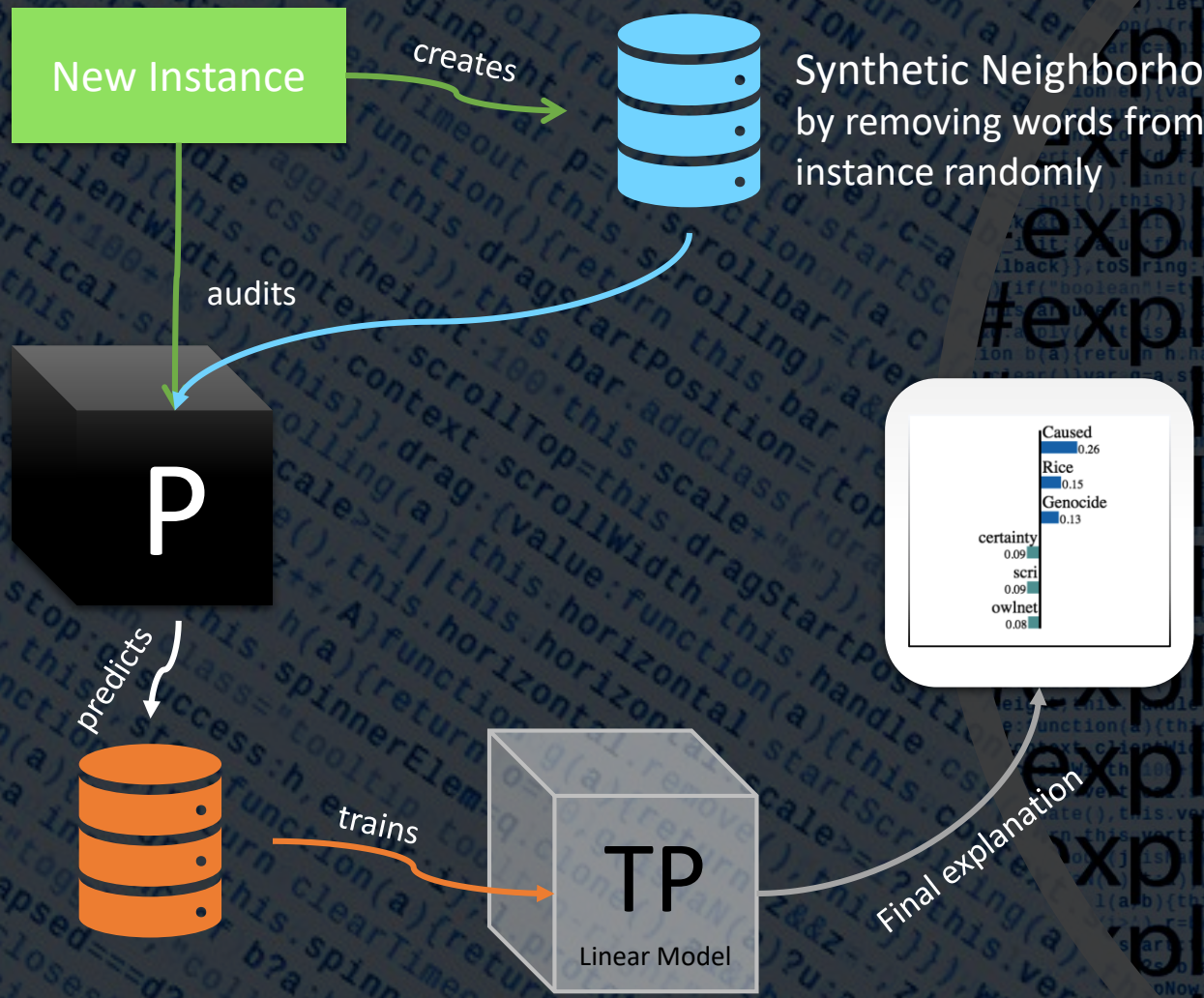




Use of Scikit Learn for Explainable Models
https://scikit-learn.org



Use of Orange for Explainable Models
https://github.com/biolab/orange3



State of the Art

LIME[5]

local *interpretable*
model-agnostic
explanations

https://github.com/marcotcr/lime

Mathematic Formulation

$L(f, g, \pi x) + \Omega(g)$

Disadvantage on Sparse Data!

Why?

LIME can only generate 2^n different neighbours, where n the number of non-zero values.

We proposed LioNets[9]…

State of the Art

LIME[5]

local *interpretable*
model-agnostic
explanations
https://github.com/marcotcr/lime

Mathematic Formulation

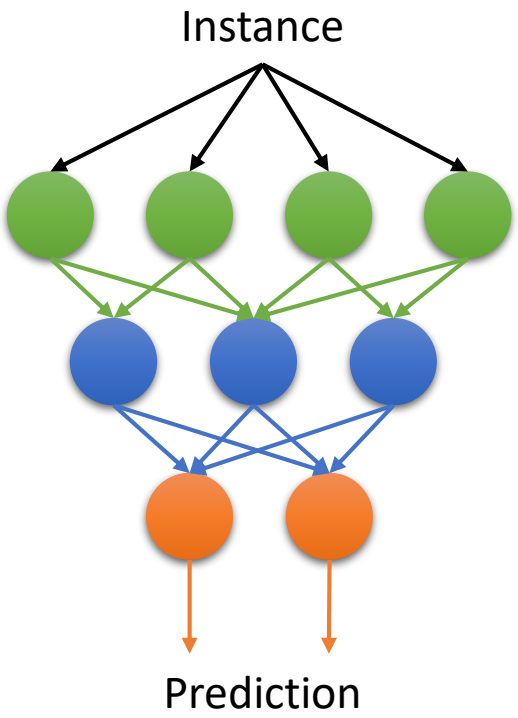
$L(f, g, \pi x) + \Omega(g)$

LioNets Hypothesis

Create neighbours to the abstract space using a decoder trained with the neural network as encoder will lead to better, bigger and more representative neighbourhoods, thus bet*ter explanations*.

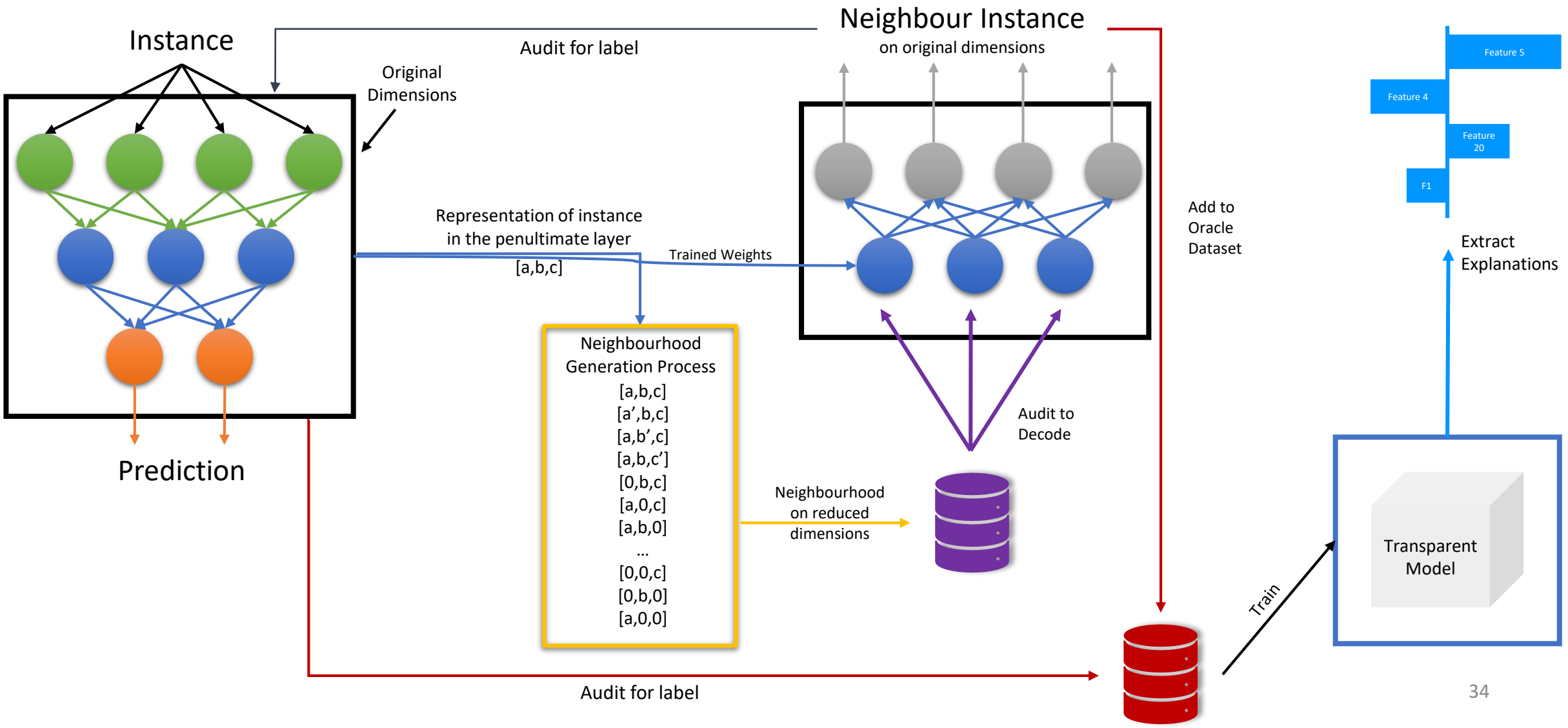
LioNets Architecture

LioNets will try to interpret a neural network's prediction. Thus, it is a **model-specific** outcome explanator

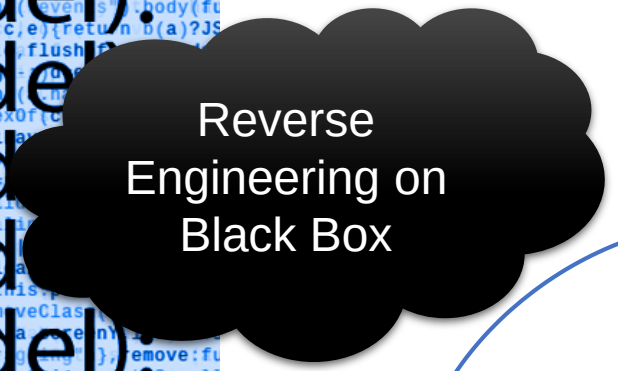


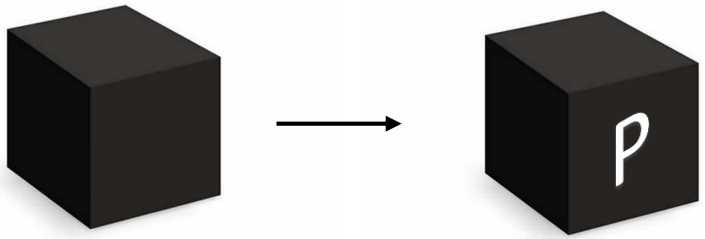
Deep Neural Network Classifier

LioNets Architecture



Permutation Importance[4]



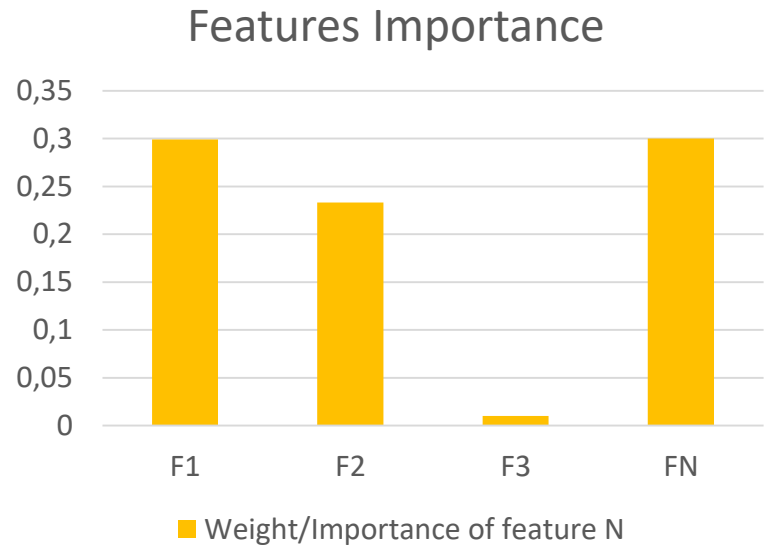


Training

Audit
New Data

Randomize a
feature

Train Data
(X,y)

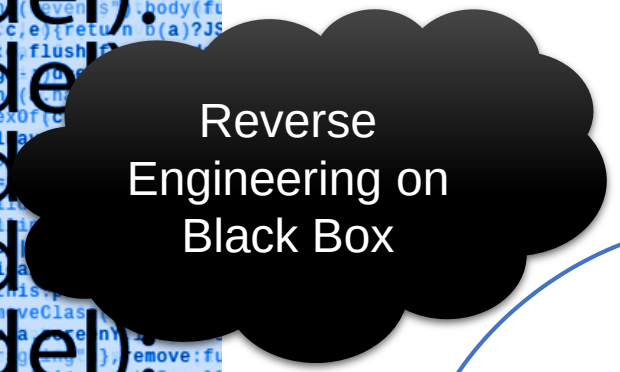


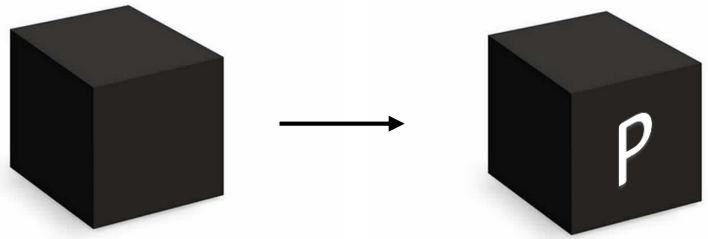
Inst.	F1	F2	F3	FN	Target
x1	0	1	blue	cold	1
x2	1	1	red	warm	0
x3	0	0	blue	cold	1

Inst.	F1	F2	F3	FN	Predicted
x1	0	1	blue	warm	0
x2	1	1	red	cold	1
x3	0	0	blue	cold	1

Inst.	F1	F2	F3	FN	Predicted
x1	0	1	red	cold	1
x2	1	1	blue	warm	0
x3	0	0	blue	cold	1

PDP (Partial Dependence Plot)[4]





Training

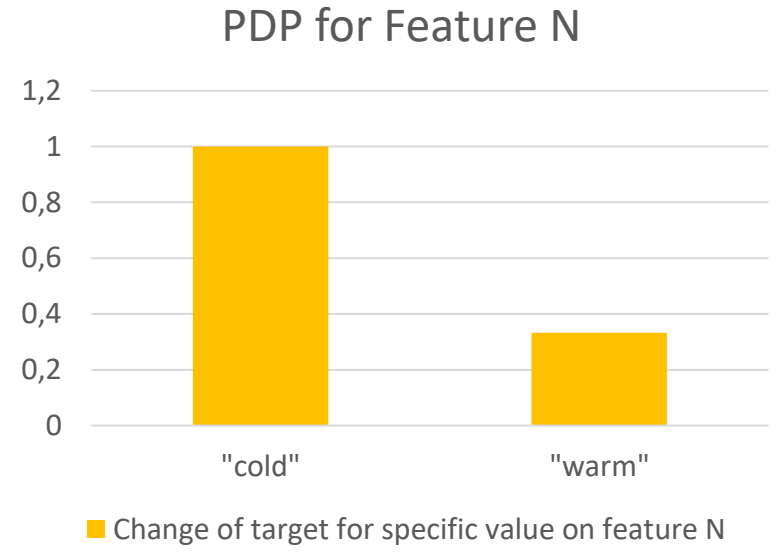
Audit
New Data

Train Data
(X,y)

Inst.	F1	F2	F3	FN	Target
x1	0	1	blue	cold	1
x2	1	1	red	warm	0
x3	0	0	blue	cold	1

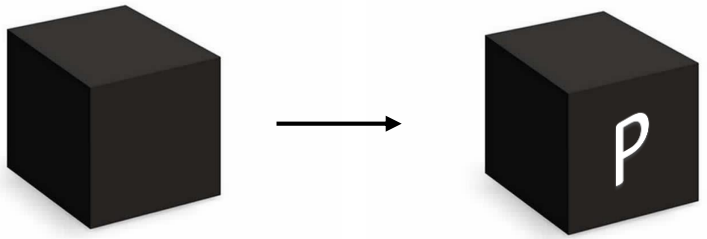
Inst.	F1	F2	F3	FN	Predicted
x1	0	1	blue	cold	1
x2	1	1	red	**cold**	**1**
x3	0	0	blue	cold	1

Inst.	F1	F2	F3	FN	Predicted
x1	0	1	blue	**warm**	**0**
x2	1	1	red	warm	0
x3	0	0	blue	**warm**	1



Mean Prediction

SHAP/Shapley Value [4]



Training

Audit
New Data

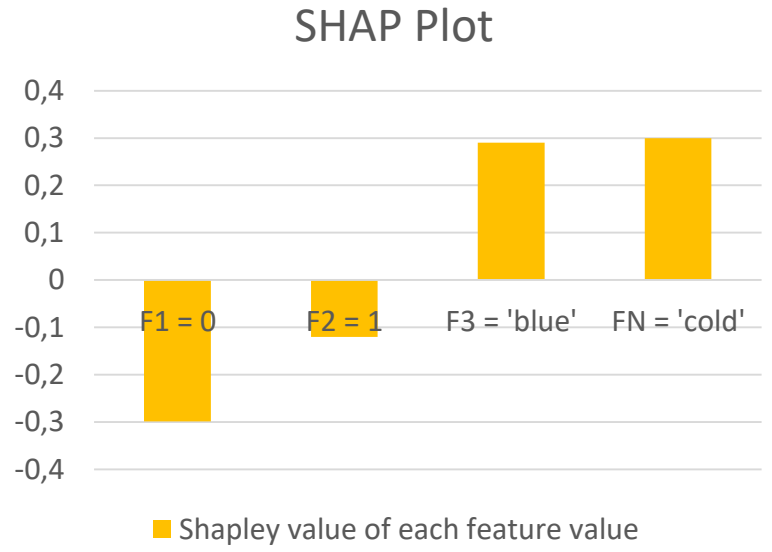
Reverse Engineering on Black Box

Train Data
(X,y)

Inst.	F1	F2	F3	FN	Target
x1	0	1	blue	cold	1
x2	1	1	red	warm	0
x3	0	0	blue	cold	1

Inst.	F1	F2	F3	FN	Predicted
x1	0	1	blue	cold	1
x2	?	?	?	cold	?
x3	?	?	?	cold	?

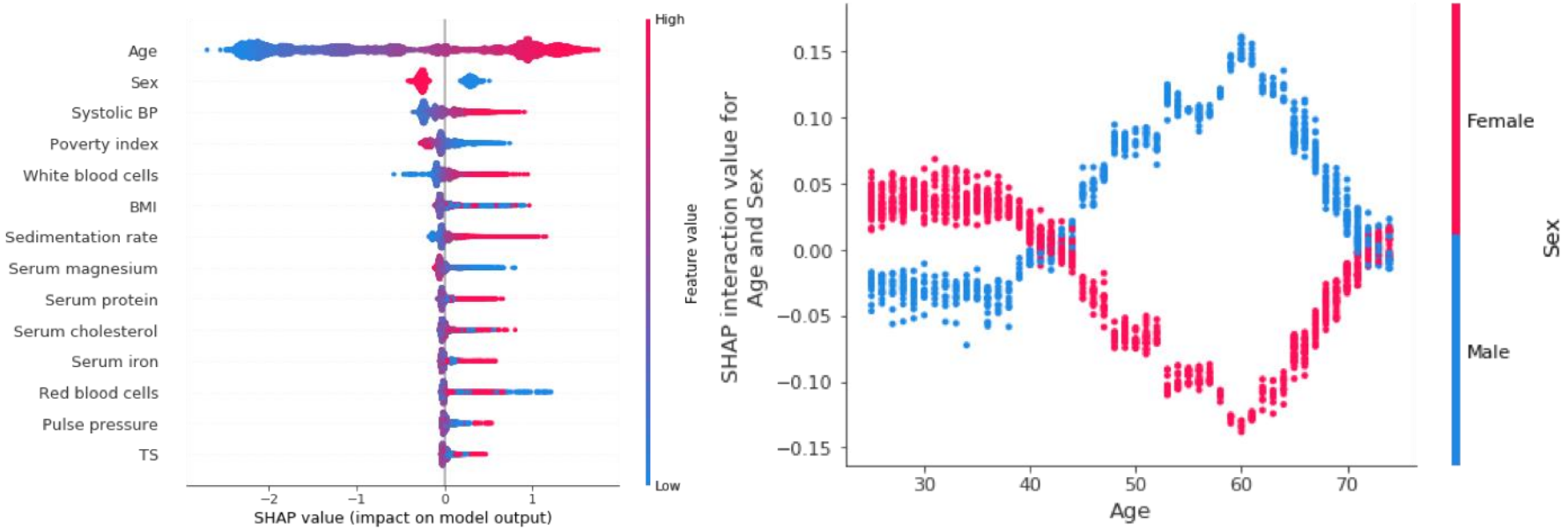
Inst.	F1	F2	F3	FN	Predicted
x1	0	1	blue	cold	1
x2	?	?	blue	?	?
x3	?	?	blue	?	?



SHAP/Shapley Value Examples[4]

Can do anything!

Reverse Engineering on Black Box



Model Inspection via Feature Importance (left) and Partial Dependence Plots (right) (computed using SHAP values)

SHAP/Shapley Value Examples[4]

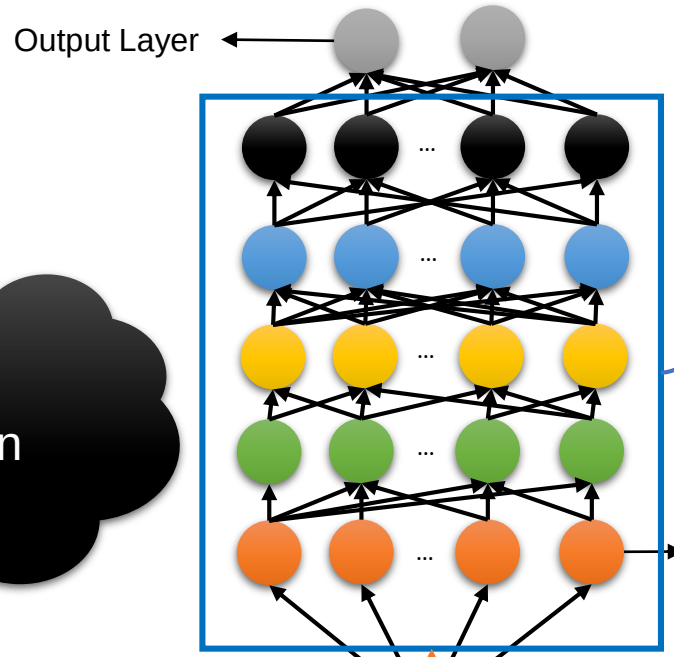
Can do anything!

Reverse Engineering on Black Box

Outcome explanation

Knowledge Distillation[8]

Using Transparent boxes to mimic black boxes' accuracies



Reverse Engineering on Black Box

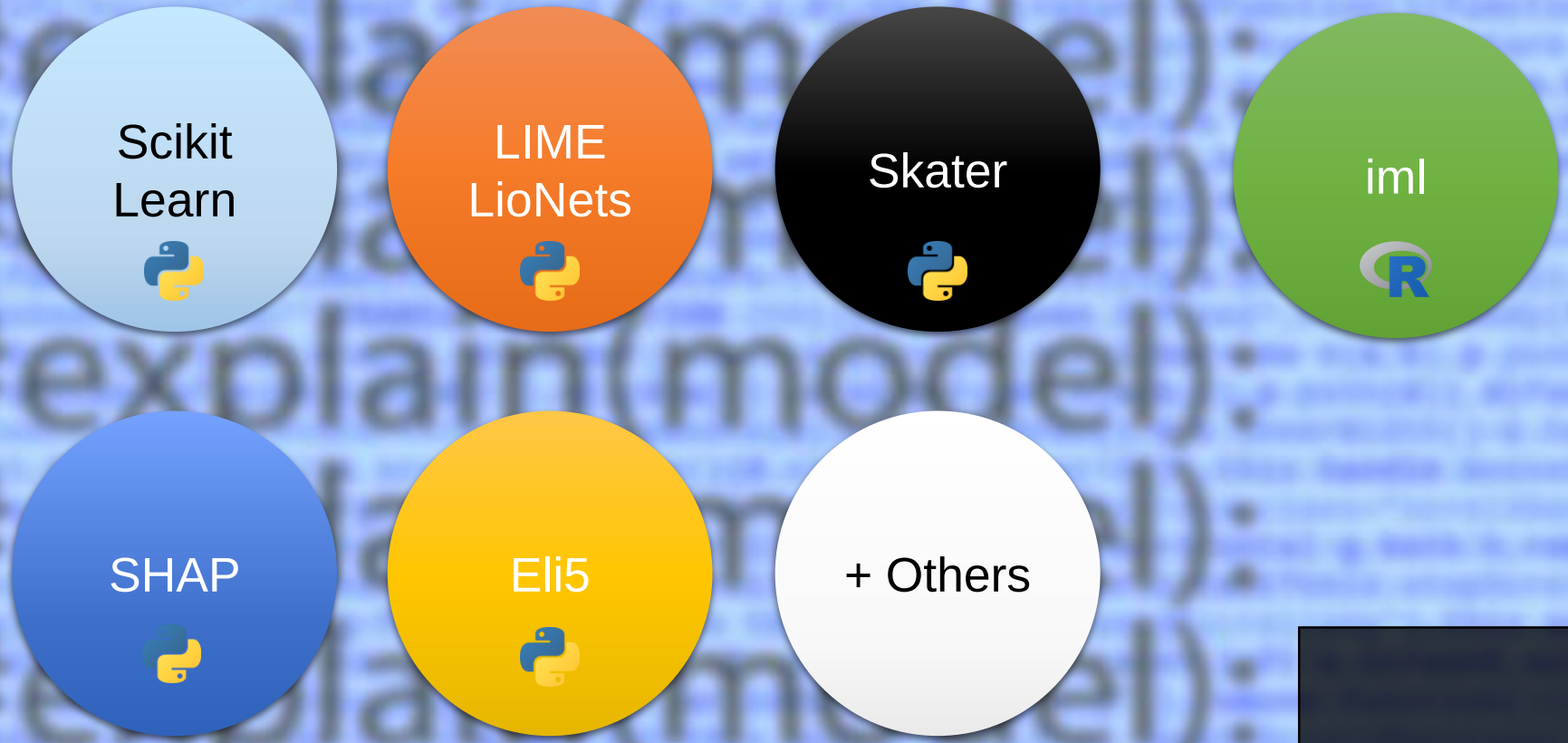
New Train Data from Labeled and Unlabeled Data annotated by the deep neural network are used to train the ***"Student"***

Explainable Model

Training a high-accuracy deep neural network as ***"Teacher"***

Train Data (X,y)

Unlabeled Data



Existing XML Libraries

Bibliography

1. Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, Fosca Giannotti. 2018. A Survey Of Methods For Explaining Black Box Models.
2. Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI).
3. Don Monroe. 2018. AI, Explain Yourself.
4. Christoph Molnar. 2018. Interpretable Machine Learning.
5. Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.
6. Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. 2011. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models.
7. Pedro Domingos. 1998. Knowledge Discovery Via Multiple Models
8. Geoffrey Hinton, Oriol Vinyals, Jeff Dean. 2015. **Distilling the Knowledge in a Neural Network**
9. I. Mollas, N. Bassiliades, G. Tsoumakas. 2019. LioNets: Local Interpretation of Neural Networks through Penultimate Layer Decoding.
10. Čyras, Kristijonas and Satoh, Ken and Toni, Francesca. 2016. Abstract argumentation for case-based reasoning.
11. Čyras, Kristijonas and Satoh, Ken and Toni, Francesca. 2016. Explanation for case-based reasoning via abstract argumentation.
12. Martin Možina, Jure Žarbkar, Ivan Bratko. 2007. Argument based machine learning.

Some say: Future Is Model Agnostic Models

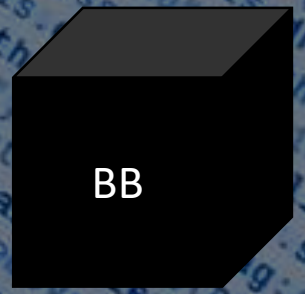
Others say: Future is New Models, Interpretable by Design

Before we go…

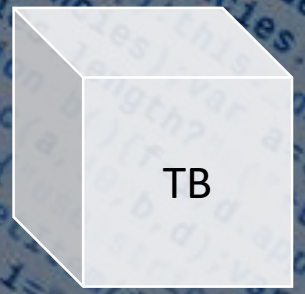
We Say Model Specific Implementations

Explainable Artificial Intelligence

Interrogating the AI systems



"Magic model on the core, Explain yourself in front of all"



Nick Bassiliades

nbassili@csd.auth.gr

The End

ARISTOTLE UNIVERSITY OF THESSALONIKI

Intelligent Systems

SCHOOL OF INFORMATICS