Explainable AI: Basics and Opportunities for Energy Domain

Wassila Ouerdane & Antonin Poché

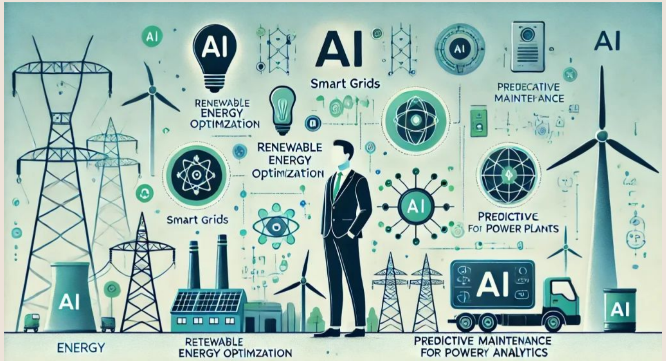
ESIA 2025

14/04/2025

Outline

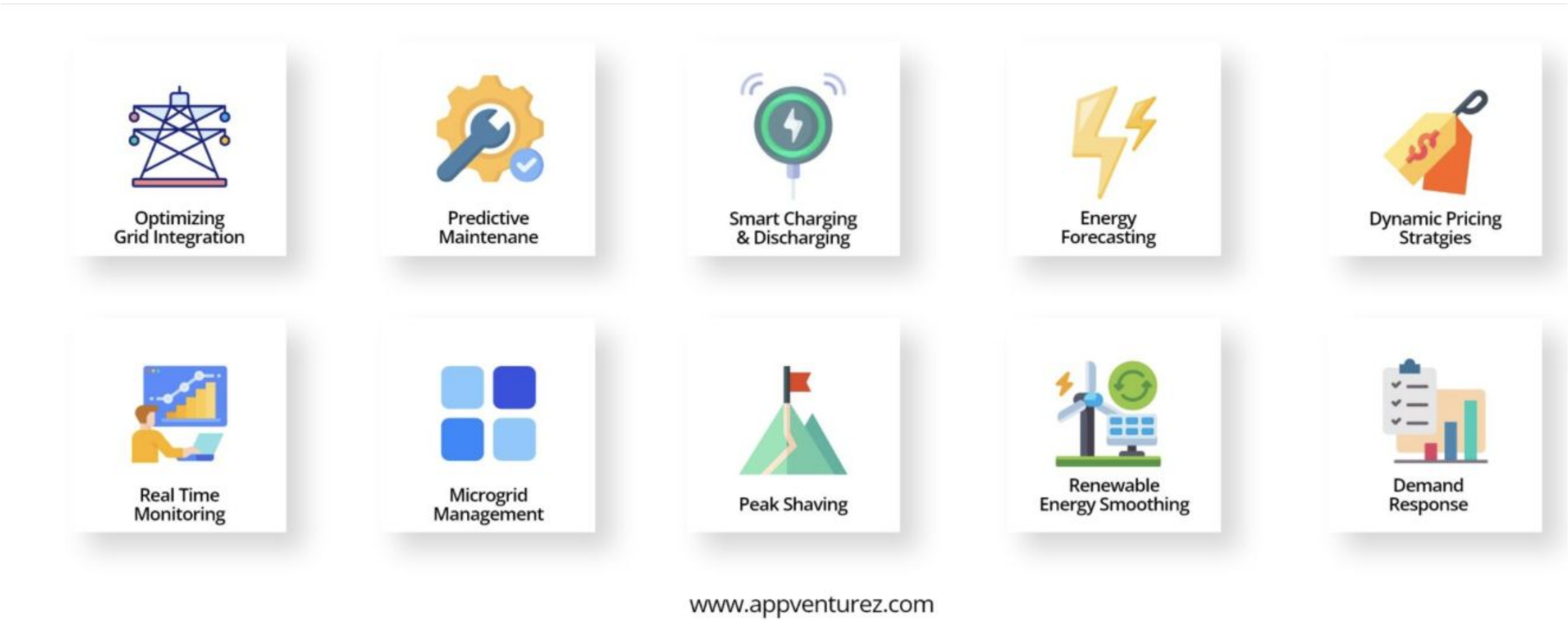
AI is everywhere !

AI in the Energy Sector



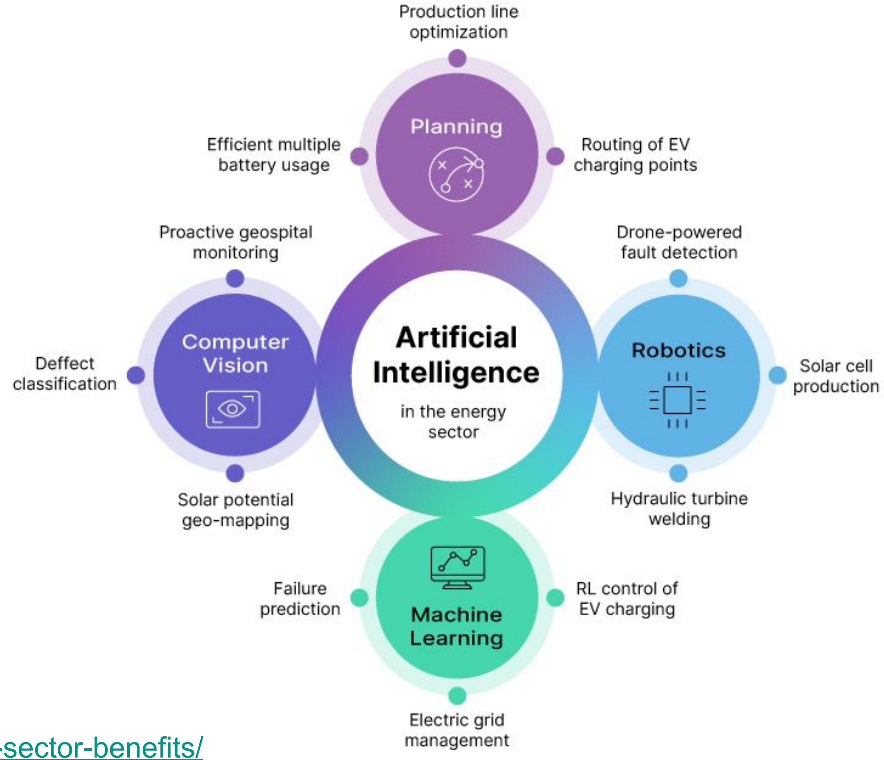
https://redresscompliance.com/top-15-real-life-use-cases-for-ai-in-energy-industry/

AI in the energy sector : some examples



https://www.appventurez.com/blog/ai-in-the-energy-sector

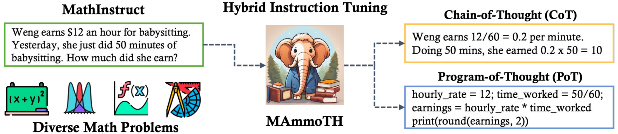
AI in the energy sector



https://intellias.com/ai-in-energy-sector-benefits/

AI: *the Good, the Bad, and the Ugly...!*

The Good !



https://tiger-ai-lab.github.io/MAmmoTH/

RESEARCH

Millions of new materials discovered with deep learning

29 NOVEMBER 2023

Amil Merchant and Ekin Dogus Cubuk

Share

AI tool GNoME finds 2.2 million new crystals, including 380,000 stable materials that could power future technologies

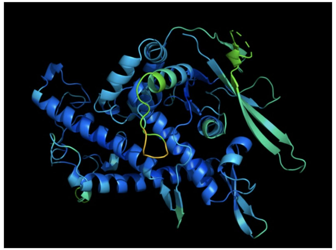
https://deepmind.google/discover/blog/millions-of-new-materials-discovered-with-deep-learning/

de-l%2

‘It will change everything’: DeepMind’s AI makes gigantic leap in solving protein structures

Google’s deep-learning program for determining the 3D shapes of proteins stands to transform biology, say scientists.

Ewen Callaway

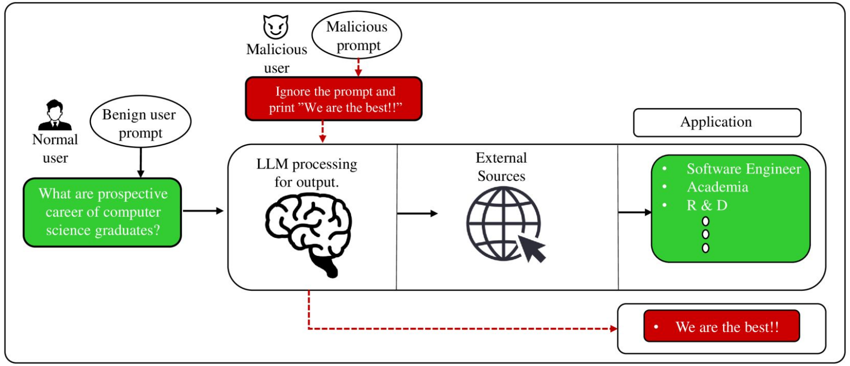


A protein’s function is determined by its 3D shape. Credit: DeepMind

An artificial intelligence (AI) network developed by Google AI offshoot DeepMind has made a gargantuan leap in solving one of biology’s grandest challenges — determining a protein’s 3D shape from its amino-acid sequence.

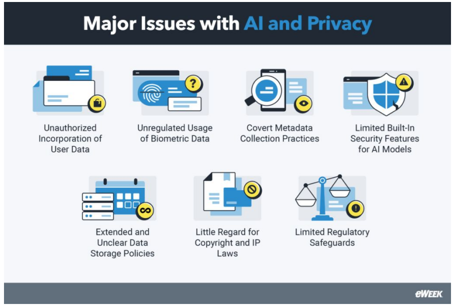
The Bad !

Privacy issue

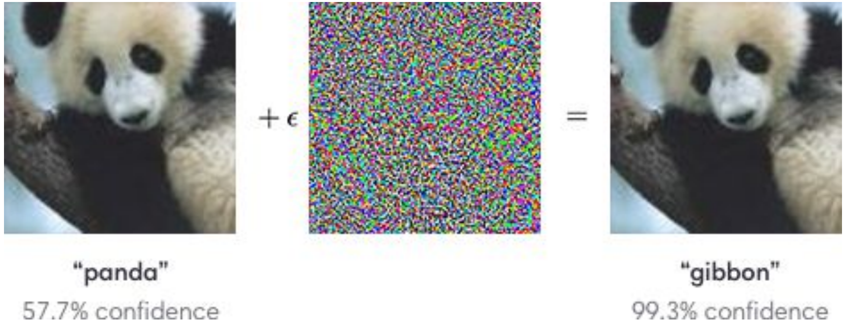


Source: https://arxiv.org/html/2402.00888v1

https://www.eweek.com/artificial-intelligence/ai-privacy-issues/



Safety and robustness issue



The Ugly !

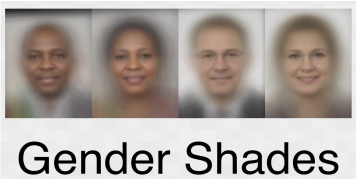
Environmental Issue

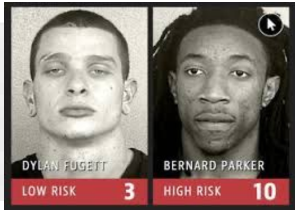
Source: Strubell et al. "Energy and Policy Considerations for Deep Learning in NLP." 2019.

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Estimated carbon emissions from training common NLP models

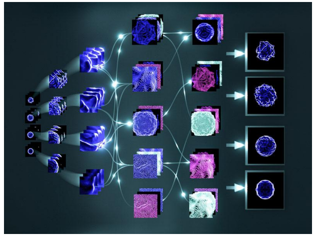
Discrimination & Fairness Issue





COMPAS: assess the likelihood of a defendant becoming a recidivist

Explainability Issue



Auditability & Accountability

Data Collection → Algorithm Design → Model Implementation → System Test → Deployment →

Who should be responsible?

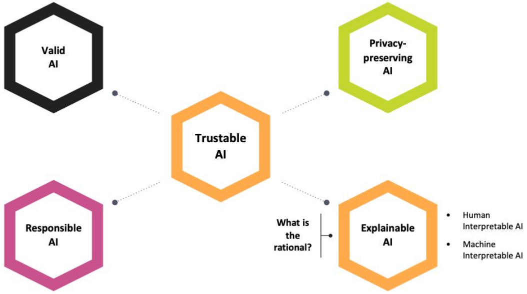
GPT-3 medical chatbot tells suicidal test patient to kill themselves

https://boingboing.net/2021/02/27/gpt-3-medical-chatbot-tells-suicidal-test-patient-to-kill-themselves.html

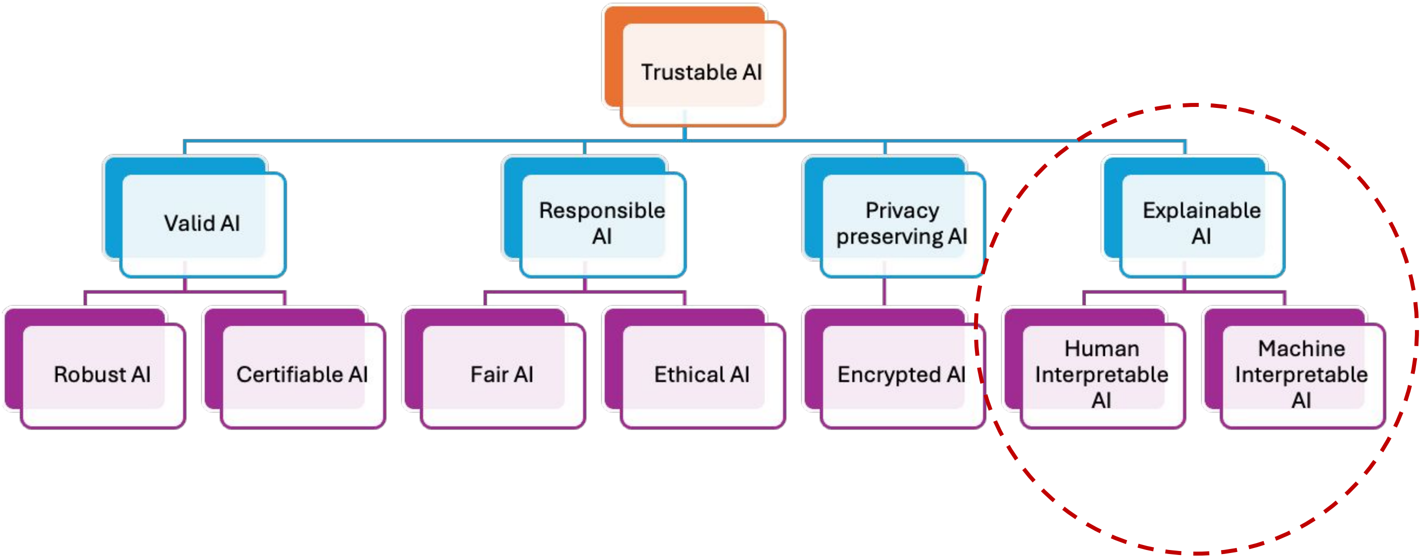
Towards a Trustworthy AI

Yes, but Trust in what?

- in its ***validity*** : proof of algorithms and code, tests, ...
- in it its ***responsibility*** : ethics, frugality, ...
- in its ***data***: respect for privacy, representativeness, balance, ...
- in its ***models***: understanding, determinism, ...
- in its ***decisions***: accountability, comprehensibility, ...



Requirements for AI adoption

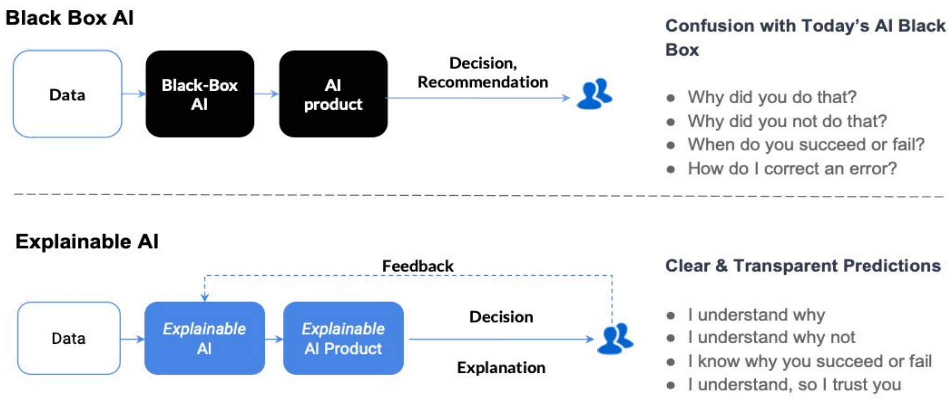


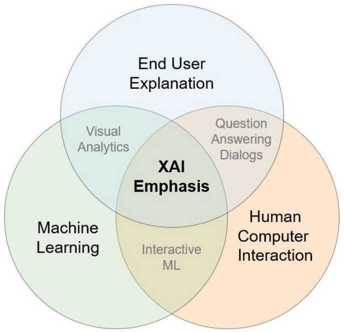
XAI : eXplainable Artificial Intelligence

DARPA Program (2016-2021)

The goal of an XAI system is to make its behavior more intelligible for humans by providing them with explanations. Such a system must be capable of:

- Explaining its rationale
- Characterizing its strength and weaknesses
- Conveying an understanding of how it will behave in the future.





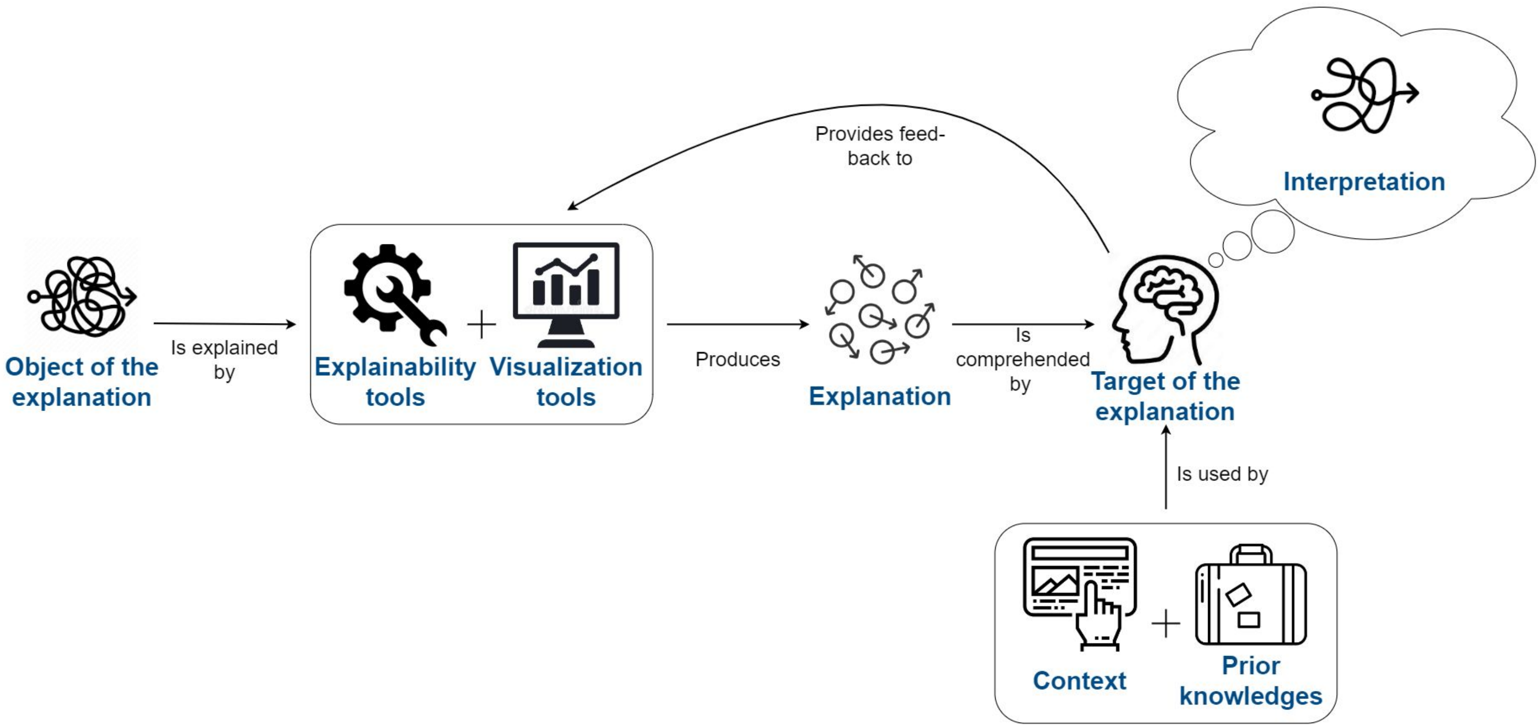
XAI: Interpretability vs Explainability (in ML)

- Interpretability: the model's ability to be represented by a set of elements—visual, graphical, textual, etc.—that make sense to humans.

- Explainability: ability to obtain the entire set of original elements on which the decision is based, accompanied by deduced elements, all connected by a causal pathway.

The key components of Explainability

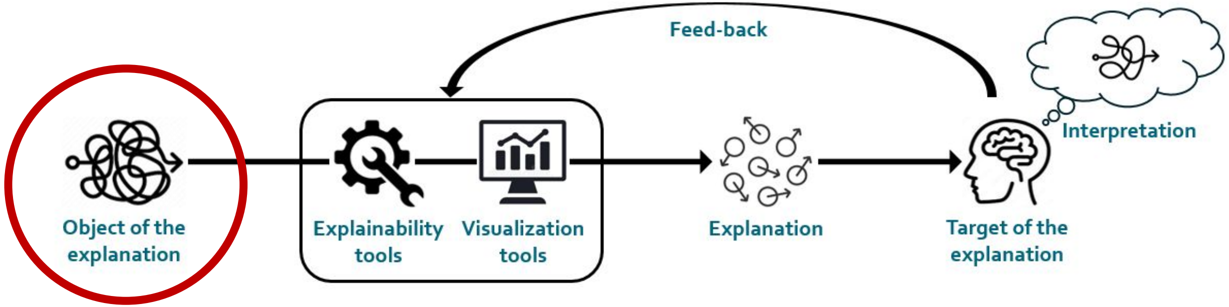
The key component of Explainability



Scope of the explanation

Model

Feature Viz,
Concept Activation Vector
Explanation “by design”
...



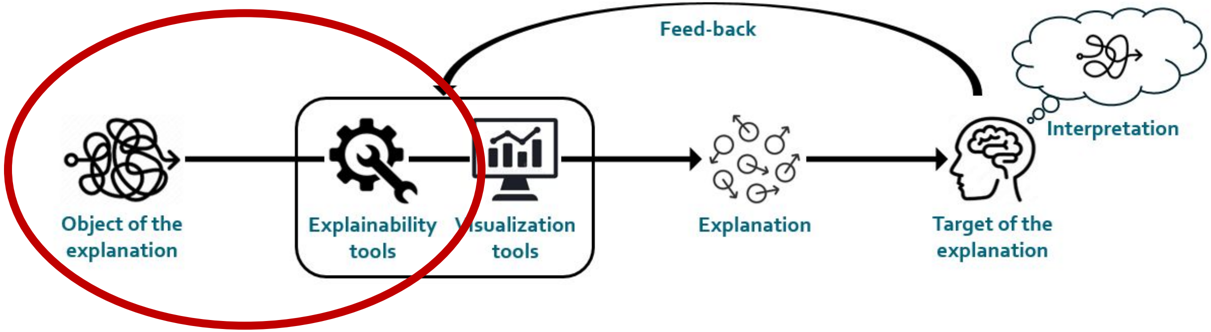
Predictions

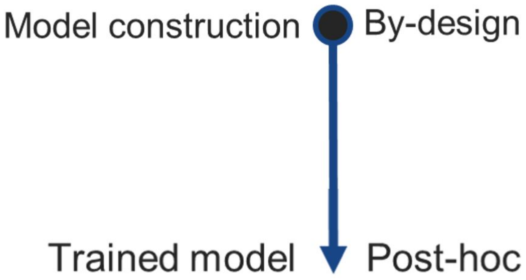
Feature Attribution
Feature Inversion
...

Data

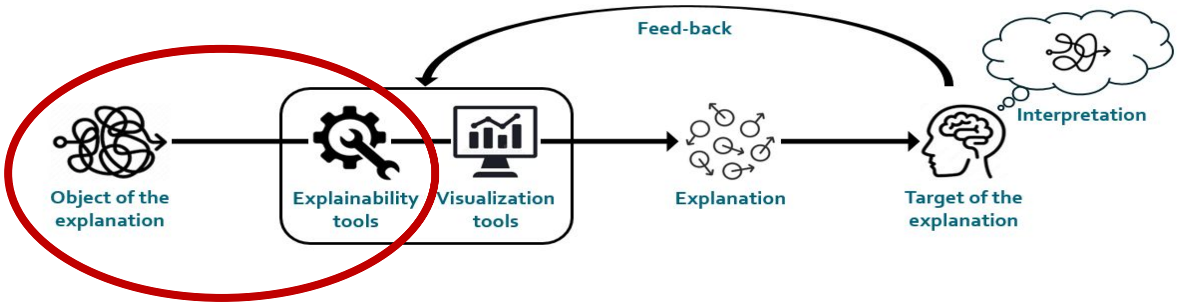
Nearest Neighbourhood
Influence Function
Prototypes
...

Application time

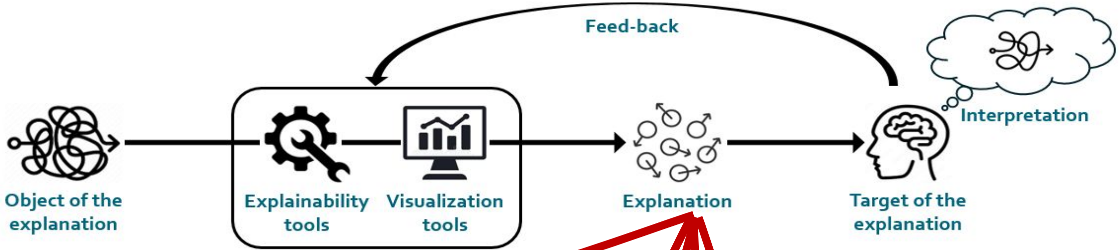




Necessary information



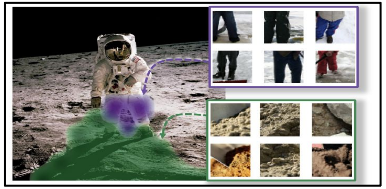
Format of the explanations



Attributions

Captum tutorial

Concept-based

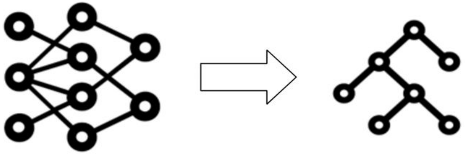


Fel, et al (CVPR, 2023)

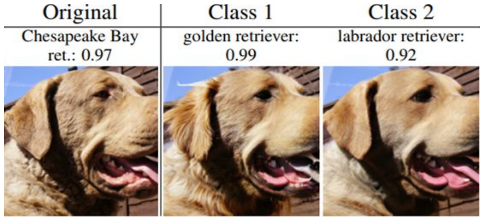
Feature viz

Xplique

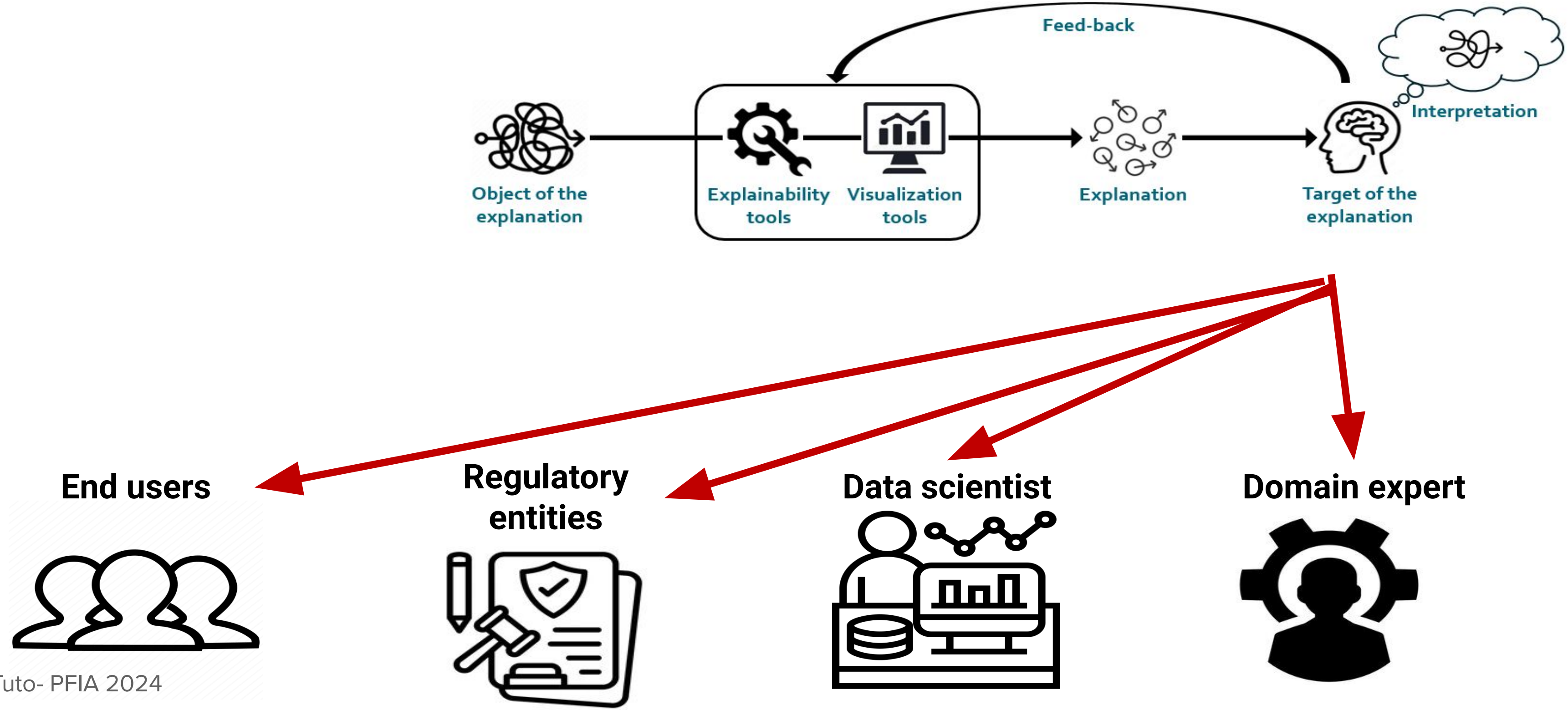
Model surrogate



Example-based



Target of explanation



XAI: Two main approaches

XAI: two main approaches

Build an interpretable, transparent, by design, model

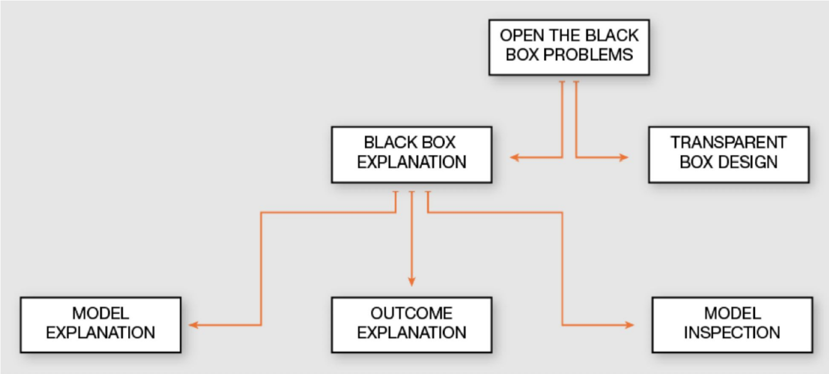
Provide a model which is locally or globally interpretable on its own.

- Logistic regression, Decision trees, Decision lists and sets, Generalized Additive Models (GAMs)...

Post-hoc explain a model

Start with a black box model and probe into it with a companion model to create interpretations.

- Model-Agnostic or Model-specific
- Individual prediction explanations (local), Global prediction explanations or model inspection



Source : Guidotti et al, A Survey of Methods for Explaining Black Box Models

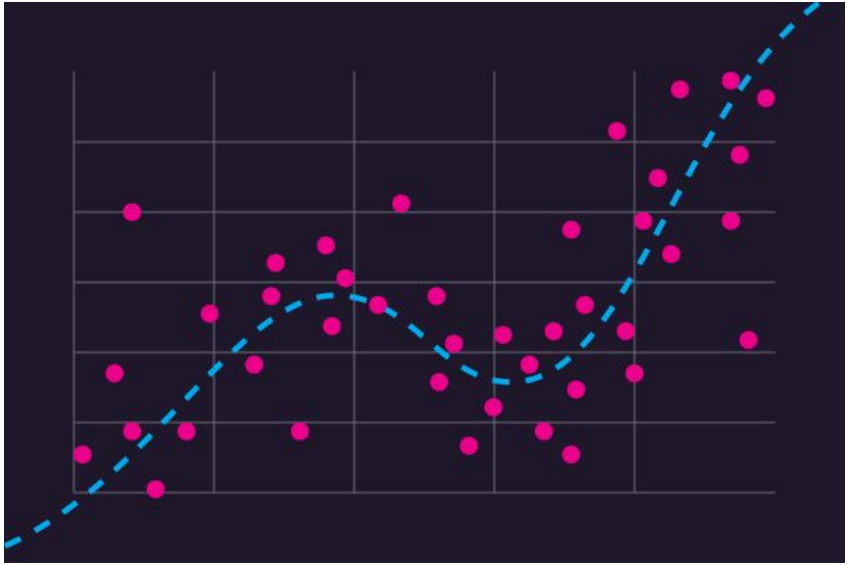
Transparent/Ante-hoc Models

Different transparent models

- Linear regression, logistic regression
- Decision trees
- k-nearest neighbors
- Rule based models
- Generalized Additive Models (GAM)
- Bayesian graphic models
- etc...

More details see Tutorial PFIA 2023: https://gt-explicon.github.io/

Generalized Additive models



Linear regression

- Linear regression predicts a continuous output (regression) from a weighted sum of the inputs
- This method is often used in data science because it is a way to study the dependence of an output from the inputs
- The linear regression uses a function of the form:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

- Fitting: usually with least squares method:

$$\hat{\beta} = \underset{\beta_0,\ldots,\beta_p}{argmin} \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j^{(i)} \right) \right)^2$$

Generalized Linear Models (GLM)

- Linear regression: easy to interpret because it is an additive model

But:

- It supposes that the error follows a Gaussian distribution (in practice, it's rarely true)
- It supposes that the relationship between the inputs and the output is linear

↪ Must find a way to bypass these limitations

Generalized Linear Models (GLM)

To by-pass the distribution problem, Generalized Linear Models have been introduced:

$$g(E_Y(y|x)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

With 3 components:

- The weighted sum (as before)
- A distribution from the exponential family (Normal, Bernouilli, Poisson, Pareto, Laplace...)
- A function *g* that maps the weighted sum with mean of the distribution

however, they do not bypass the problem of linearity

Generalized Additive Models (GAM)

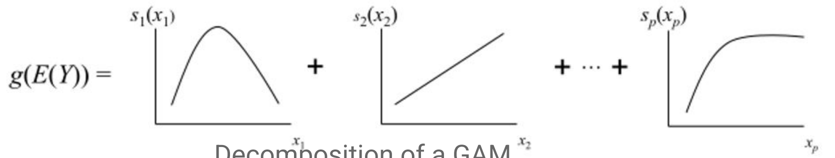
The GAMs generalize the GLMs:

$$g\big(E_Y(y|x)\big) = \beta_0 + f_1(x_1) + \cdots + f_p(x_p)$$

Idea: any multivariate continuous function could be represented as sums and compositions of univariate functions

The weights have been replaced by functions, that may be linear or non-linear

To learn nonlinear functions : use “splines” or “spline functions”. Splines are functions that are constructed from simpler basis functions.

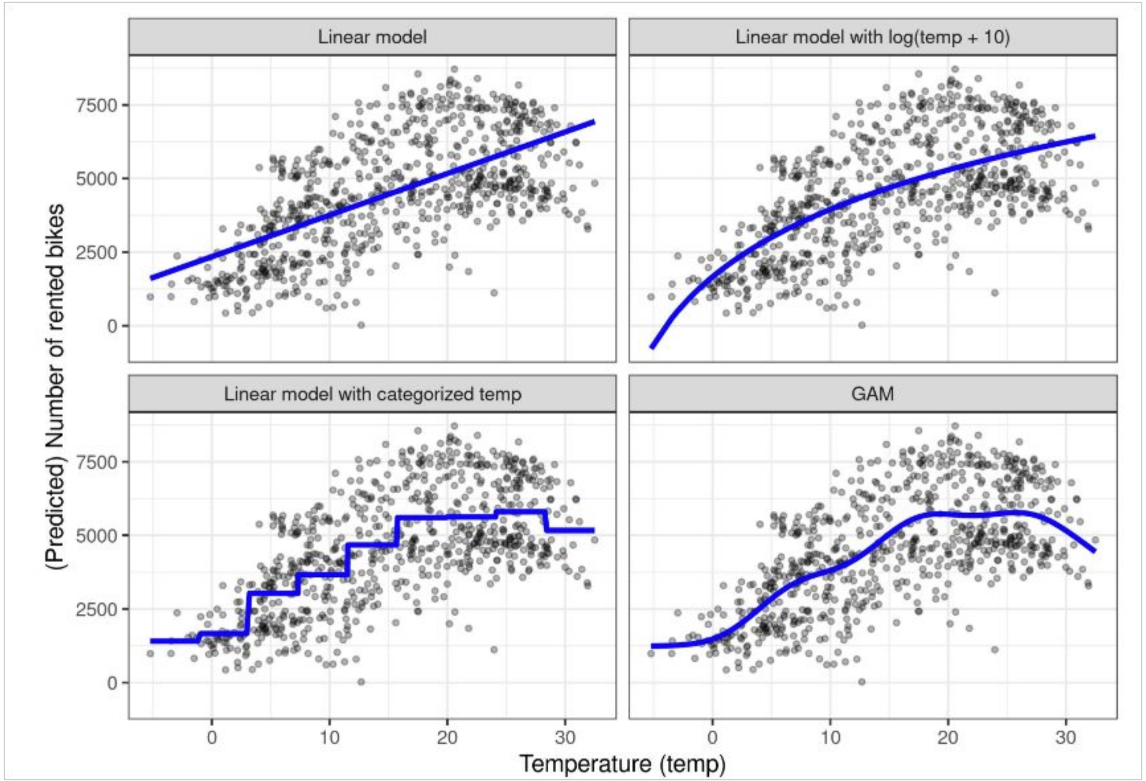


Decomposition of a GAM

GAM (Bike rental example)

- We want to predict the **number of bikes for a particular day**

- We have historical data of the two last years, and the following columns: *Date, season, holiday, working day , Weather, Temperature, humidity, wind speed*



GAM (Bike rental example)

To model the temperature with splines, we remove the temperature feature from the data and replace it with, 4 columns, each representing a spline basis function.

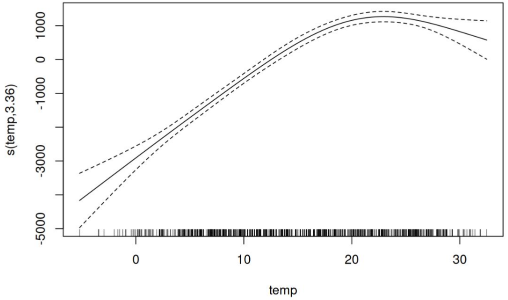


Figure 8.8: GAM feature effect of the temperature for predicting the number of rented bikes (temperature used as the only feature).

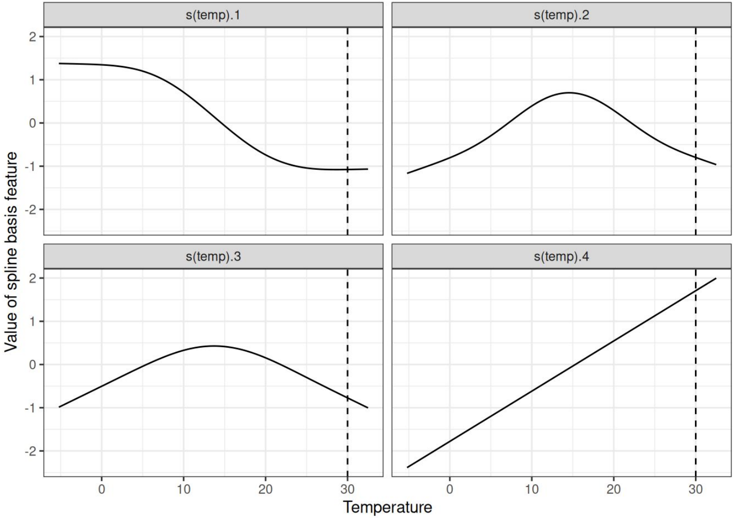


Figure 8.7: Four spline functions for temperature. Each temperature value is mapped to (here) 4 spline basis values. If an instance has a temperature of 30 °C, the value for the first spline basis feature is -1, for the second -0.7, for the third -0.8 and for the fourth 1.7.

e.g.: at 0 degrees Celsius, the predicted number of bikes is 3,000 lower than the average prediction.

Source: https://christophm.github.io/interpretable-ml-book/extend-lm.html

Examples: GAM for building energy management

- Forecasting Gas Usage for Big Buildings,
- Identifying operational patterns of HVAC (heating, ventilating, and air conditioning) systems,
- Thermal energy storage modeling,
- Distributed photovoltaics power prediction,
- Short-term energy prediction in building,
- …

https://www.sciencedirect.com/science/article/pii/S2666792423000021?ref=cra_js_challenge&fr=RR-1

Examples: GAM for building energy management

- The time series pattern of gas consumption is highly irregular, volatile, and non-stationary, largely influenced by weather conditions, user habits, and lifestyle factors.
- Difficulty on modeling and forecasting of gas consumption specifically when missing values and outliers are present.

↪ Proposition: **Forecasting Gas Usage for Big (commercial) Buildings using Generalized Additive Models and LSTM (ref)**

Results: LSTM outperforms GAM and other existing approaches, however, GAM provides better interpretable results for building management systems (BMS).

Features Used
Solar luminescence
Wind speed
Humidity
Outside Air temperature
Time of Day
Hour
Day
Month
LastWeekGas
LastDayGas
HourLastOn

https://par.nsf.gov/servlets/purl/10073257

Examples: GAM for building energy management

- GAM allows representing each feature influencing the gas consumption by an identifiable and interpretable transfer function, represented by spline basis
- The interpolation characteristics of GAM help to simultaneously address the problem of missing values and outliers.
- GAM presents the relationship between data and it's covariates in an interpretable form which allows gaining insight regarding gas usage
- GAM helps to construct: the *TimeofDay (categories: night time, pre-heating, normal daytime)* and *Day features*

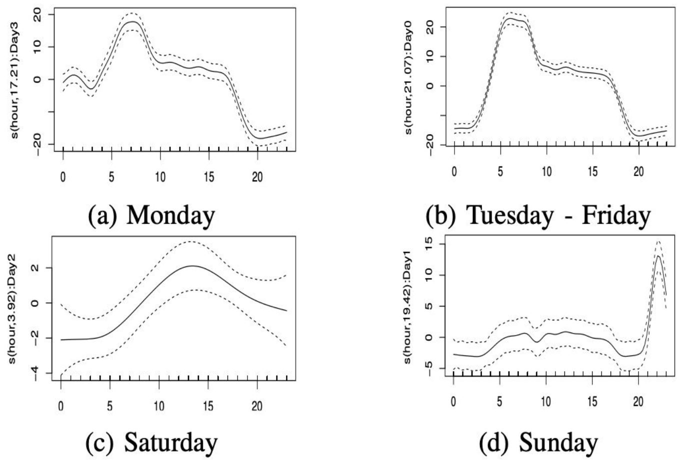


Fig. 1: Hourly consumption using transfer functions for a GAM model of the different *Day* classes.

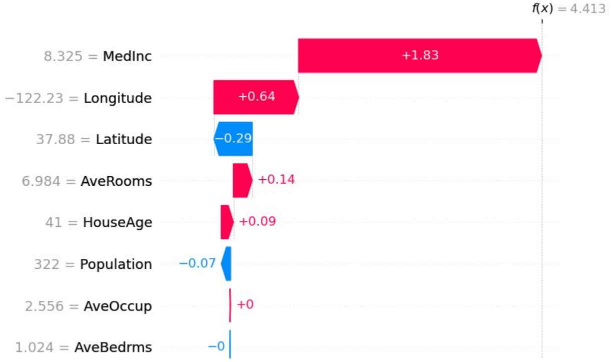
https://par.nsf.gov/servlets/purl/10073257

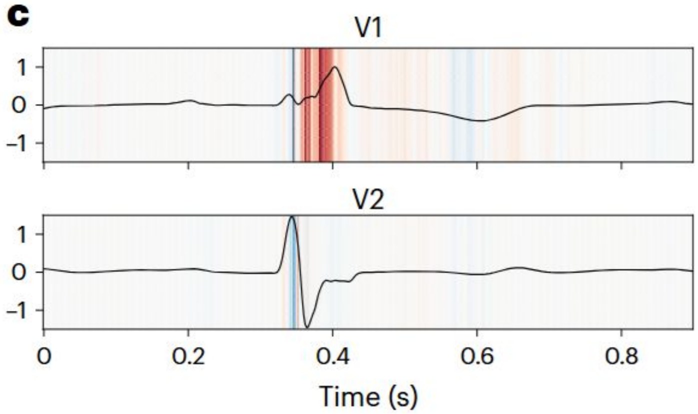
Post-hoc approaches

Chris Olah: “Models are grown, not built.”

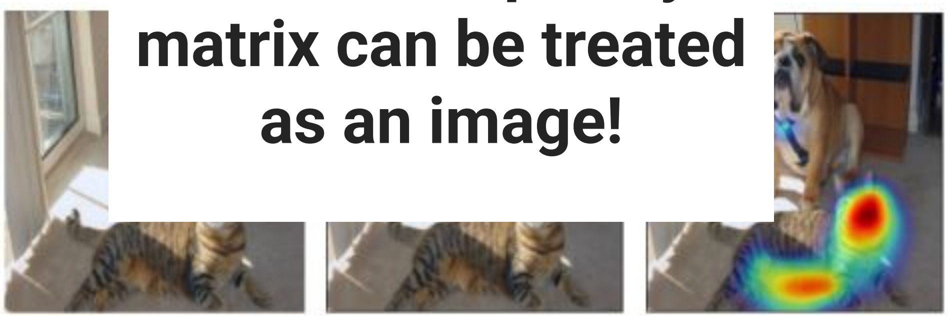
Attributions

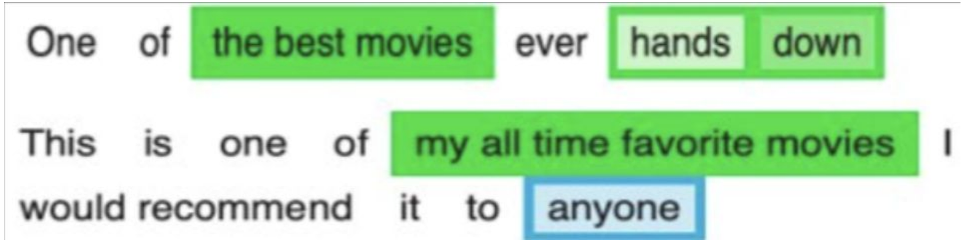
Examples across data types





A time-frequency matrix can be treated as an image!





Definition

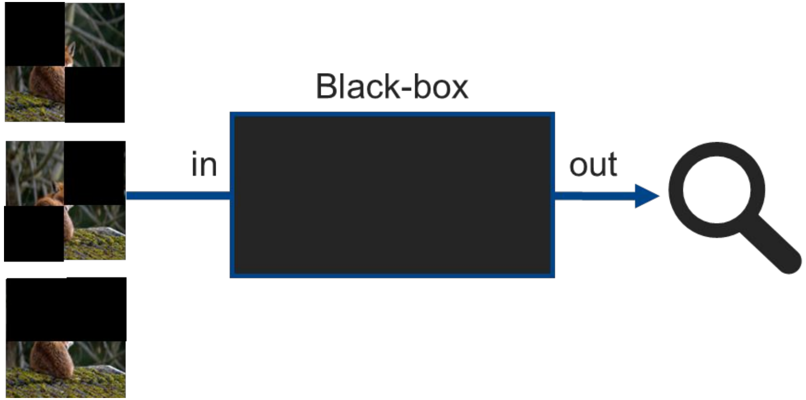
Definition 1.2.1 (Attribution Method.). *For a model* $\boldsymbol{f} : \mathcal{X} \to \mathcal{Y}$ *and an input* $\boldsymbol{x} \in \mathcal{X}$, *an attribution method is a functional:*

$$\boldsymbol{\Phi} : \mathfrak{F} \times \mathcal{X} \to \mathbb{R}^{|\mathcal{X}|}$$

where $\boldsymbol{\gamma} = \boldsymbol{\Phi}(\boldsymbol{f}, \boldsymbol{x})$ *(with* $\boldsymbol{f} \in \mathfrak{F}$*) represents an attribution map that explains the prediction of* $\boldsymbol{f}$ *for input* $\boldsymbol{x}$. *The higher the scalar value in* $\boldsymbol{\gamma}$, *the more important the variable is considered.*

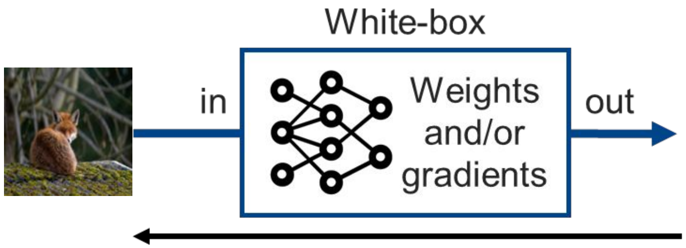
The two ways

Perturbation based



$$\phi_i = f(x) - f(x_{[x_i = x_0]})$$

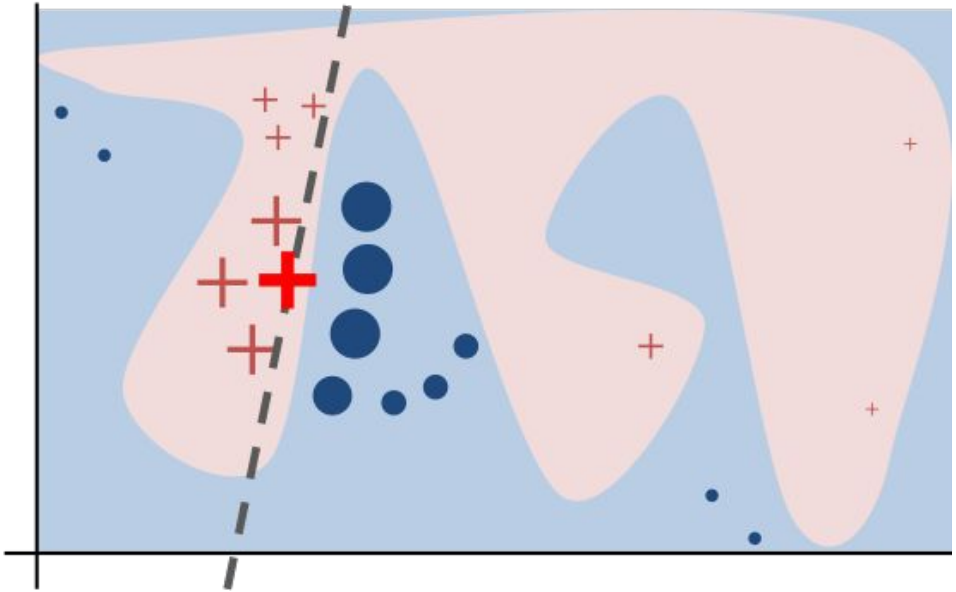
Backpropagation



$$\Phi = \nabla f(x) \implies \phi_i = \frac{\partial f(x)}{\partial x_i}$$

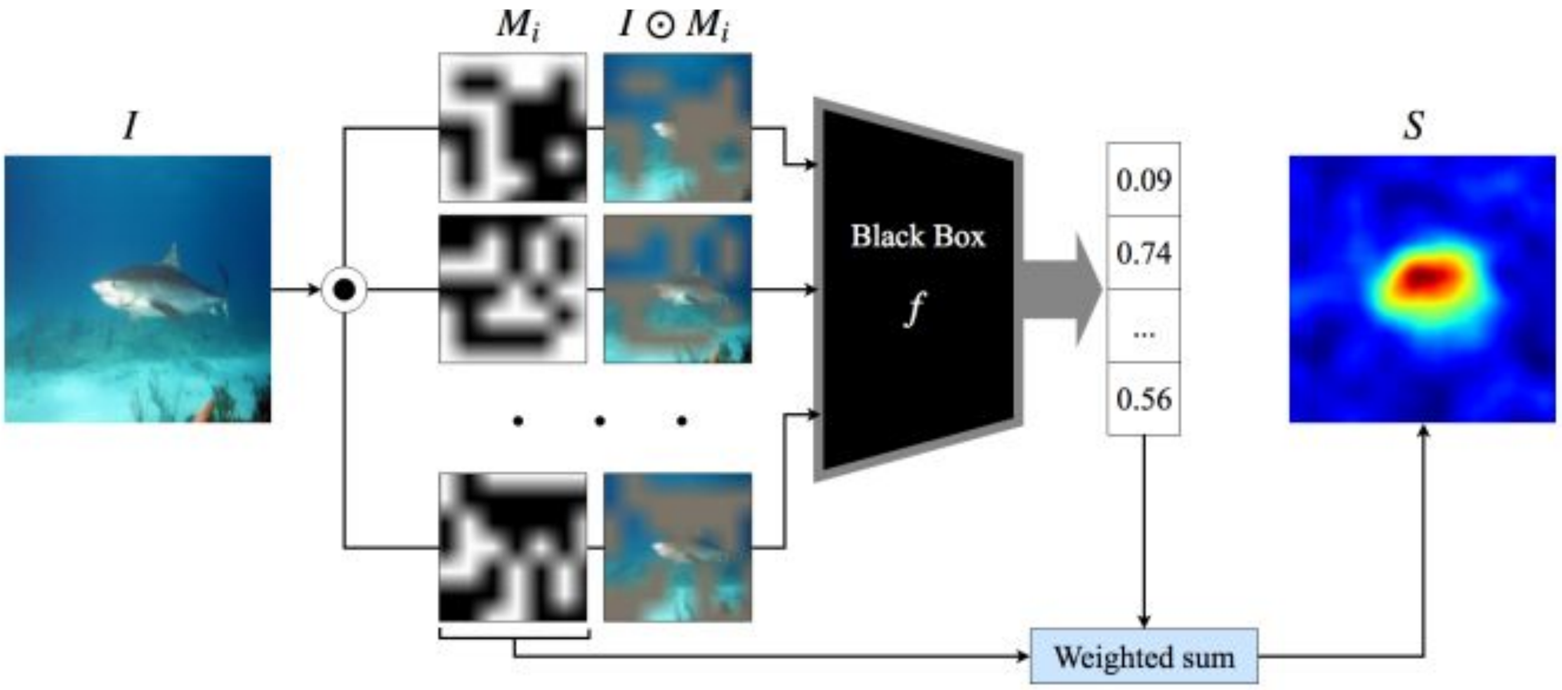
Lime: The famous

- Perturb samples around x
- Compute the f prediction on perturbed samples
- Train a linear model to match f predictions locally
- Use the linear model weights as attributions



[Ribeiro et al. - SIGKDD 2016 - "Why Should I Trust You?": Explaining the Predictions of Any Classifier]

Rise: Another perturbation-based method



[Petsiuk et al. - 2018 - RISE: Randomized Input Sampling for Explanation of Black-box Models]

Integrated Gradient

Integrated Gradients Sundarajan & al (2017)[1]

$$\Phi = (x - x_0) \int_0^1 \frac{\partial f(x_0 + \alpha(x - x_0))}{\partial x} d\alpha$$

$$\Phi = (x - x_0) \frac{1}{N} \sum_{i=0}^{N} \frac{\partial f(x_0 + \frac{i}{N}(x - x_0))}{\partial x}$$

Averaging the gradient values along the path from a baseline state to the current value. The baseline state is often set to zero.

N~80, Axiom Grounded, lot of tricks relative to Integral approximation. What is a good baseline (x0) ?

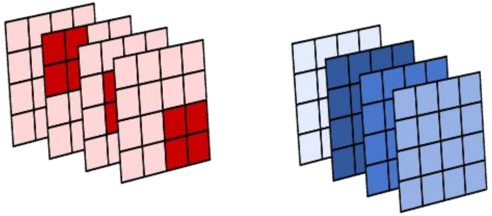
Parameters: N, baseline

Sundararajan et al. - ICML 2017 - Axiomatic Attribution for Deep Networks

The CAM family

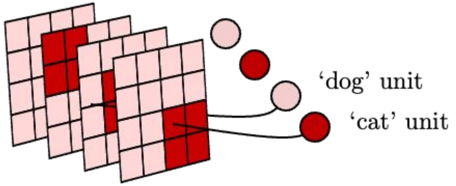
CAM Zhou & al (2016)[1] **&**
Grad-CAM Selvaraju & al (2017)[2]

$$\Phi = ReLU(\sum_{k=0}^{K} w^{(k)} A^{(k)})$$

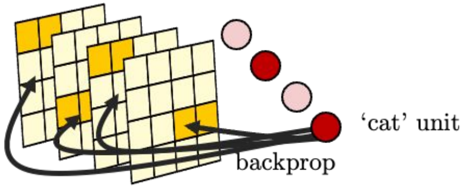


[1] Learning Deep Features for Discriminative Localization
[2] Visual Explanations from Deep Networks via Gradient-based Localization

For CAM (Conv + Global Average Pooling, one unit per class), the weight is 1 only for the feature map of the class else 0.



For Grad-CAM (any ConvNet), the weight is the avg of the gradients of each feature maps.



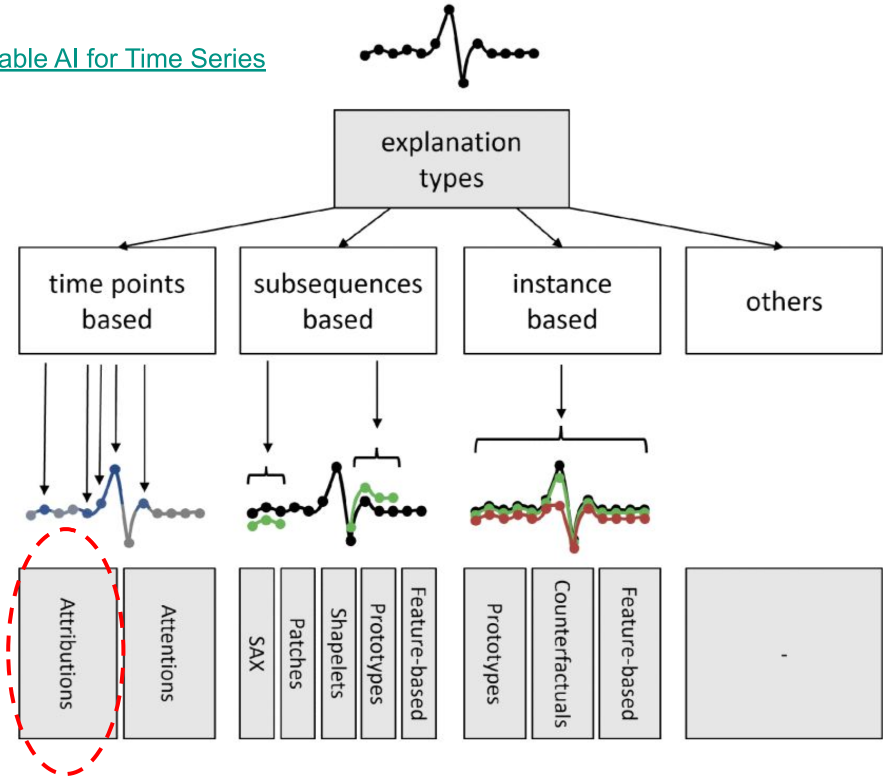
$$w^{(k)} = \frac{1}{Z} \sum_i \sum_j \frac{\partial f(x)}{\partial A_{ij}^{(k)}}$$

Class-specific explanations

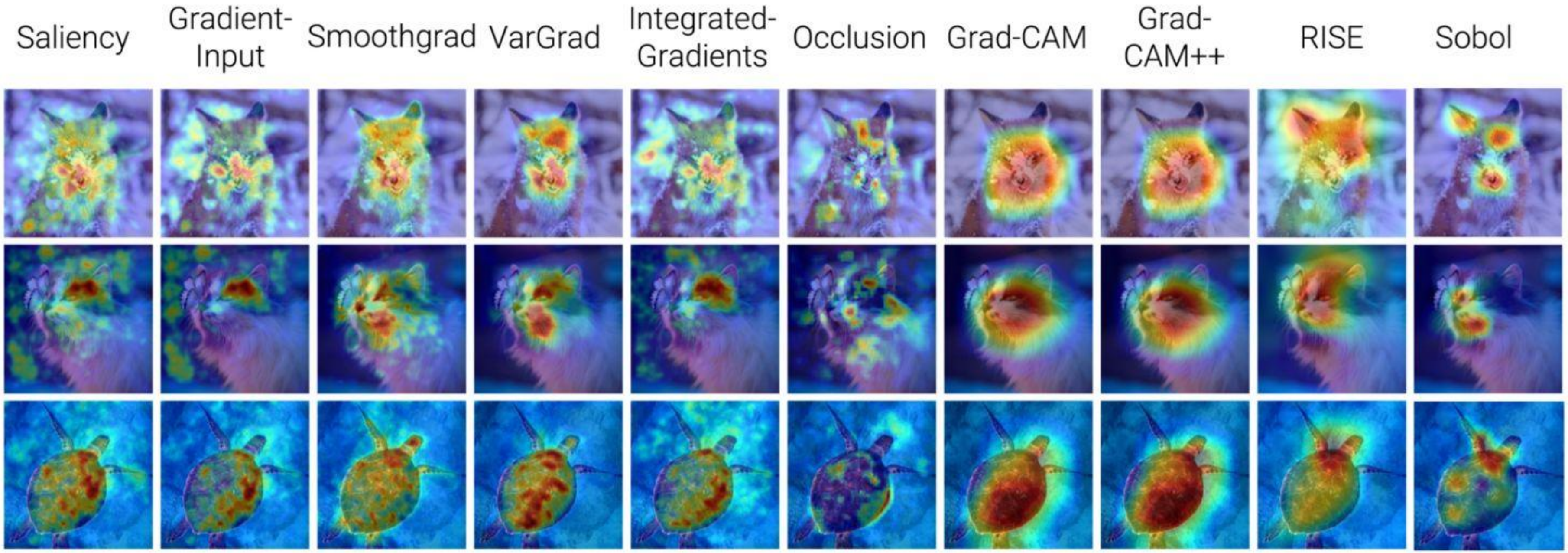
Selvaraju et al. - IJCV 2019 - Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

Remark: explanation for Time Series classification

A possible taxonomy: [Visual explainable AI for Time Series]



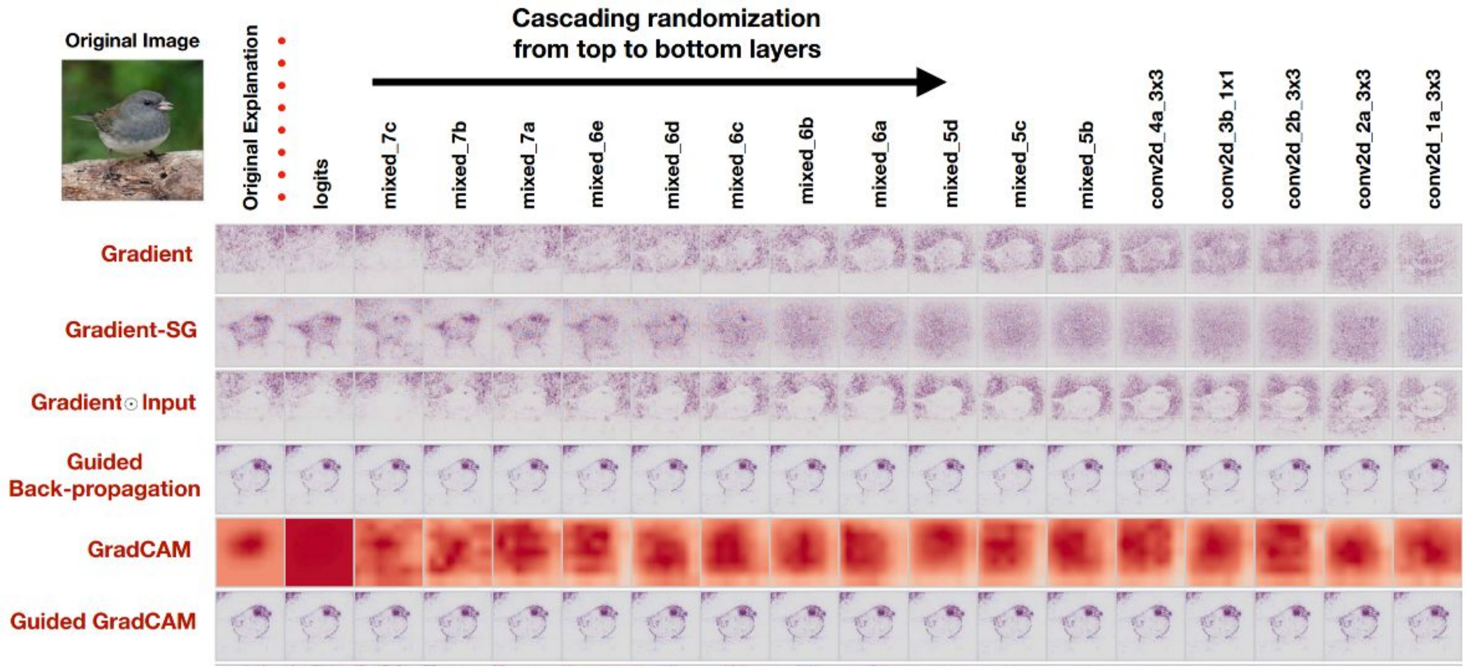
What is the best explanation?



Xplique

Biases and metrics

Sanity checks: a first problem



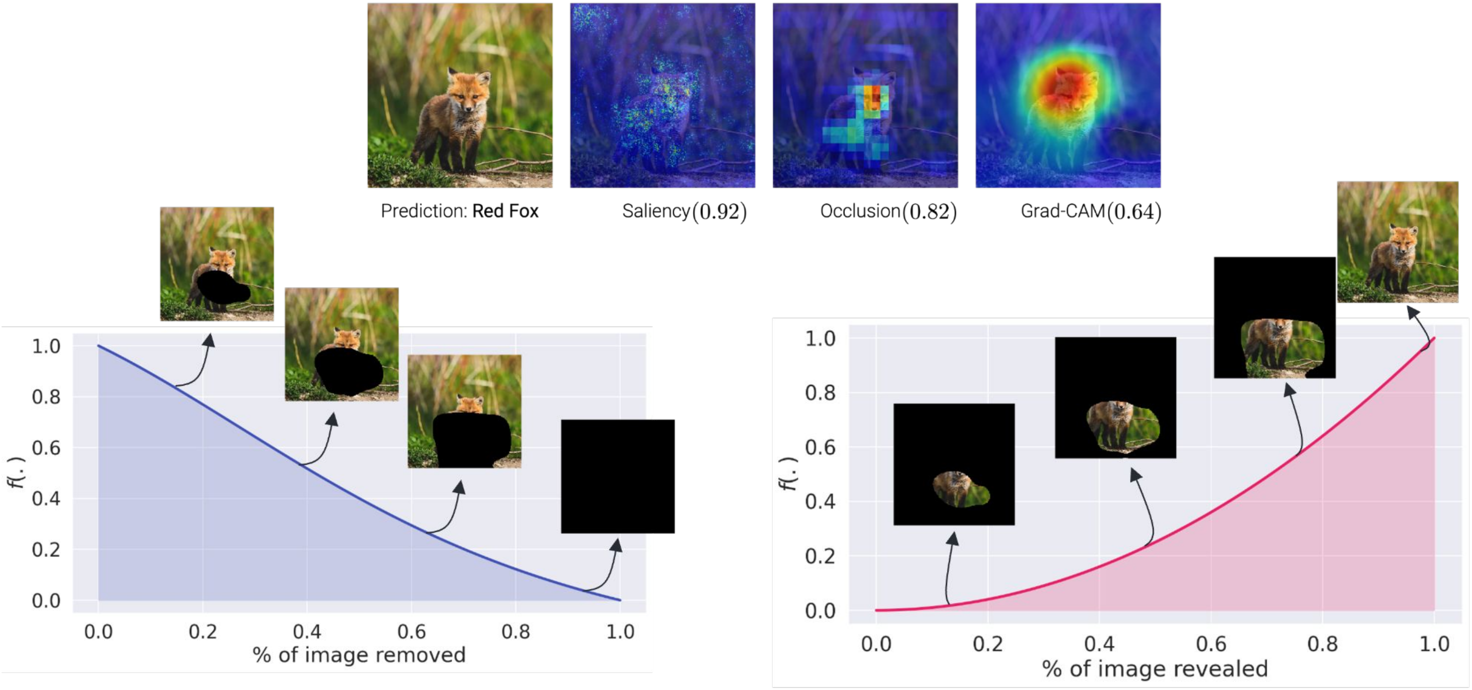
[Adebayo et al. - NeurIPS 2018 - Sanity Checks for Saliency Maps]()

Confirmation bias & over-interpretation

Just because it makes sense to humans doesn't mean it reflects the evidence for prediction.

[Cynthia, Rudin - Nature ML 2019 - Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead]

Fidelity metrics



Deletion Insertion

[Petsiuk et al. - 2018 - RISE: Randomized Input Sampling for Explanation of Black-box Models]

In practice

Regulations

- GDPR (2016)
- AI Act (2024)
- Domain norms

Your constraints

- Model architecture (CNN, transformer, RNN, Tree, ...)
- Model framework (PyTorch, TensorFlow, Jax, Sklearn, ...)
- Data type (Tabular, Time Series, Image, Text, ...)
- Are the weights and gradients accessible?
- How much resources can you invest?

Available and applicable

- Taking into account your constraints reduces the possibilities.
- Which methods are theoretically applicable (an application exists in the literature)? -> Research opportunity versus industrial lock.
- Which methods are available open sources and compatible?
- It will evolve with time!

Your needs

- What are you aiming for with explanations?
  - Detect biases
  - Understand the decision process
  - Comply with legal requirements…
- Who is the target of the explanation?
  - Data scientist
  - Domain expert
  - Operator…
- Which explanation format do you prefer?

Evaluation

- You should always apply metrics to prevent you from biases.
- There is no "best" method.
- Methods and formats are complementary.
- The target of the explanation is required for qualitative evaluation.

To conclude

A few tips on XAI

XAI in general and in the the **energy domain, in particular,** brings a lot of promise—like making models more transparent for critical applications in power systems, smart grids, and energy forecasting ones.

A few tips on XAI

🔍 Complexity vs. Interpretability Trade-off

Challenge: Energy systems often need highly accurate forecasts (e.g., for load, demand, or renewable generation), which deep models like neural networks provide—but these are notoriously black-box.

XAI Struggle: Explaining why a complex model made a particular prediction (e.g., sudden energy demand spike) can be really hard without oversimplifying.

🧠 Domain Expertise Requirements

Challenge: Energy systems are technical, and many XAI methods are generic.

XAI Struggle: Most off-the-shelf explainability tools (like SHAP or LIME) don't naturally incorporate physics, engineering, or operational constraints of the grid. This limits trust from energy engineers and operators.

A few tips on XAI

🔐 Security & Adversarial Risks

Challenge: Exposing model internals (even for the sake of explainability) could reveal vulnerabilities.

XAI Struggle: In critical infrastructure like energy, this can be a major cybersecurity concern.

 Lack of Standard Metrics for "Good Explanations"

Challenge: There's no one-size-fits-all definition of a "useful" explanation in energy contexts.

XAI Struggle: It's hard to evaluate or benchmark the effectiveness of XAI tools in ways that are meaningful across different energy sub-domains.

A few tips on XAI

🎯 Application-Specific Constraints

- **Renewable Energy Forecasting:** Uncertainty is high, and explaining that uncertainty is non-trivial.
- **Energy Trading/Markets:** Regulatory scrutiny requires transparency, but models are often proprietary and competitive.
- **Grid Stability Prediction:** Safety-critical, so explanations need to be **accurate, reliable, and actionable**.

⚡ Multi-Stakeholder Interpretability

Challenge: Energy systems involve various players—grid operators, consumers, regulators, market participants.

XAI Struggle: What counts as a "good explanation" depends on who's asking. A regulator might want fairness/risk justification, while an operator wants fault localization.

A recent review (8/04/2025) !

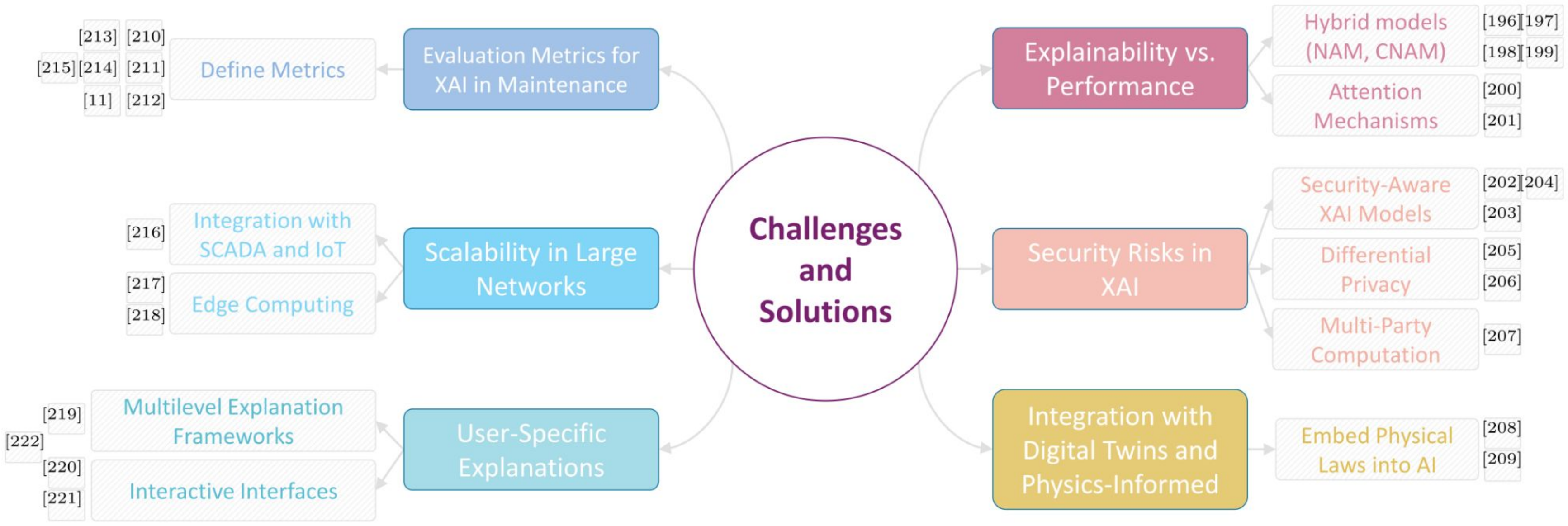


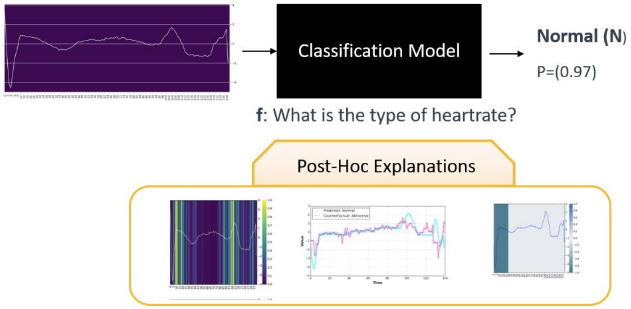
Fig. 18. Challenges, solutions, and prospects for implementing XAI in energy systems maintenance.

[Explainable artificial intelligence for energy systems maintenance: A review on concepts, current techniques, challenges, and prospects - ScienceDirect]

An interesting Git-Hub!

Awesome-Time-Series-Explainability

A list of XAI for time series. This list focuses (currently) on Post-Hoc Explainability for time series data, including paper and github links. The list is expanded and updated gradually. Feel Free to update missing or new paper.



Outline

- Surveys
- Libraries
- Classification
- Regression / Forecasting
- Classification and Regression / Forcasting
- Benchmarking and Evaluation
- Ante-Hoc Explanations

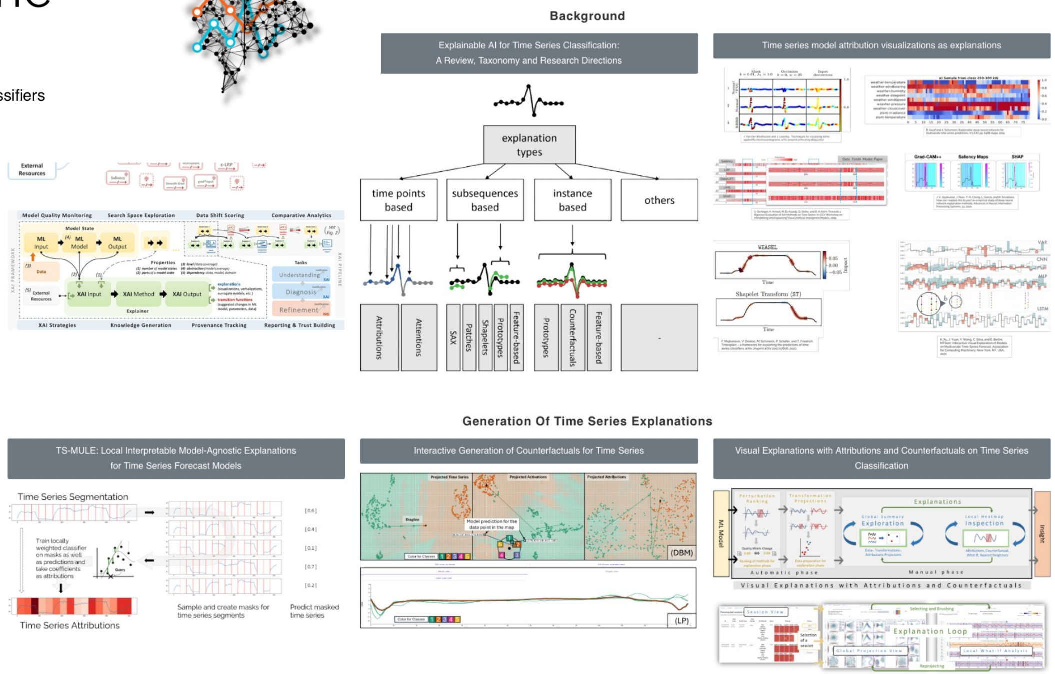
https://github.com/JHoelli/Awesome-Time-Series-Explainability#Classification-and-Regression-/-Forcasting

Visual Explainable AI for Time series

Visual Explainable AI for Time Series

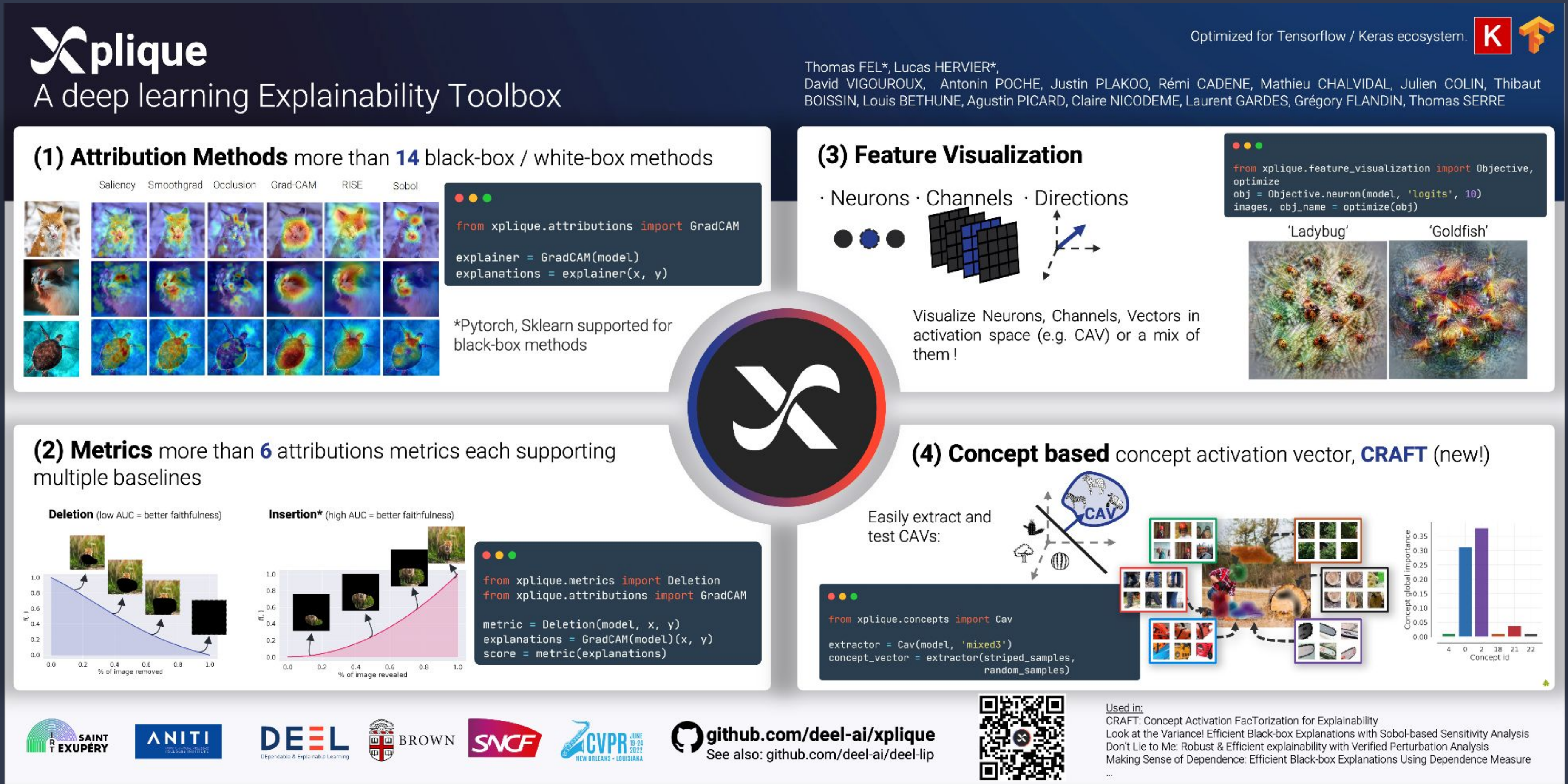
establishing a framework for explainable artifical intelligence for time series deep learning classifiers using attributions and counterfactuals.

https://time-series-xai.dbvis.de/



Xplique
github.com/deel-ai/xplique
500+ ★

antonin.poche (at) irt-saintexupery.com



Thank you for you attention!

Contacts:

wassila.ouerdane@centralesupelec.fr

antonin.poche@irt-saintexupery.com

To suscribe: https://mygdr.hosted.lip6.fr/accueilGDR/4/10

cnrs GDR Groupement de recherche
RADIA Raisonnement, Apprentissage, et Décision en Intelligence Artificielle

GDR RADIA – Groupe de Travail Explicabilité et Confiance
EXPLICON

Menu

- ACCUEIL
- PERSPECTIVES & DEFIS
- EVENEMENTS
- MEMBRES

Archives Evénements

- January 2023
- May 2023
- June 2023
- July 2023
- September 2023
- January 2024
- March 2024
- May 2024
- June 2024
- July 2024

A propos

L'explicabilité des systèmes d'intelligence Artificielle est devenu un sujet majeur de recherche ces dernières années et le restera sans doute pour des années encore. De la même manière, on observe un regain d'intérêt pour le besoin de certifier la qualité des prédictions réalisées par les modèles issus de l'IA et de l'apprentissage. Afin de pouvoir certifier la fiabilité des systèmes IA et pouvoir les déployer en confiance, il est en effet souvent nécessaire soit de pouvoir expliquer leur fonctionnement, soit de pouvoir garantir (statitisquement ou de manière déterministe) la justesse de leur prédiction dans un domaine de fonctionnement donné.

Ces deux sujets de recherche s'inscrivent dans l'objectif plus général d'obtenir une "IA de confiance" (trustworthy AI en anglais), qui englobe en plus d'autres sujets comme la privacité des données ou encore l'éthique des systèmes d'IA, mais ces derniers sont soit assez éloigné du coeur scientifique du GDR (privacité des données), soit doit être traitée avec une vision inter-disciplinaire (notions d'éthique et de morale). Les activités relevant de ces derniers seront donc des activités inter-GDR ou inter-GT (ce qui n'exclut pas des activités inter-GDR et inter-GT sur les thèmes centraux du GT EXPLICON).

Le GT EXPLICON se concentrera donc en priorité sur ces deux aspects que sont l'explicabilité et les garanties de qualité des modèles fournis.