

Lecture 1. Introduction

eXplainable Artificial Intelligence

Let's get to know each other :)

What is your name?

What do you do in life?

What projects are you working on or want to work on?

What are your expectations from this course?

What do you like about data science?

Main events

- 10 classes + 2 tests + 1 project in a team

[illegible]

How to maintain your Zen

[illegible]

Course assessment

1. Tests

70%

- a. Test 1 35%
- b. Test 2 35%

2. Project

30%

- a. Tabular dataset analysis
 - i. Causal inference 5%
 - ii. LIME 3%
 - iii. SHAP 3%
- b. Images
 - i. LIME or SHAP 5%
 - ii. Gradients, signals, attribution methods 9%
 - iii. Counterfactual examples 5%

"All models are wrong but some are useful"

George Box, 1978

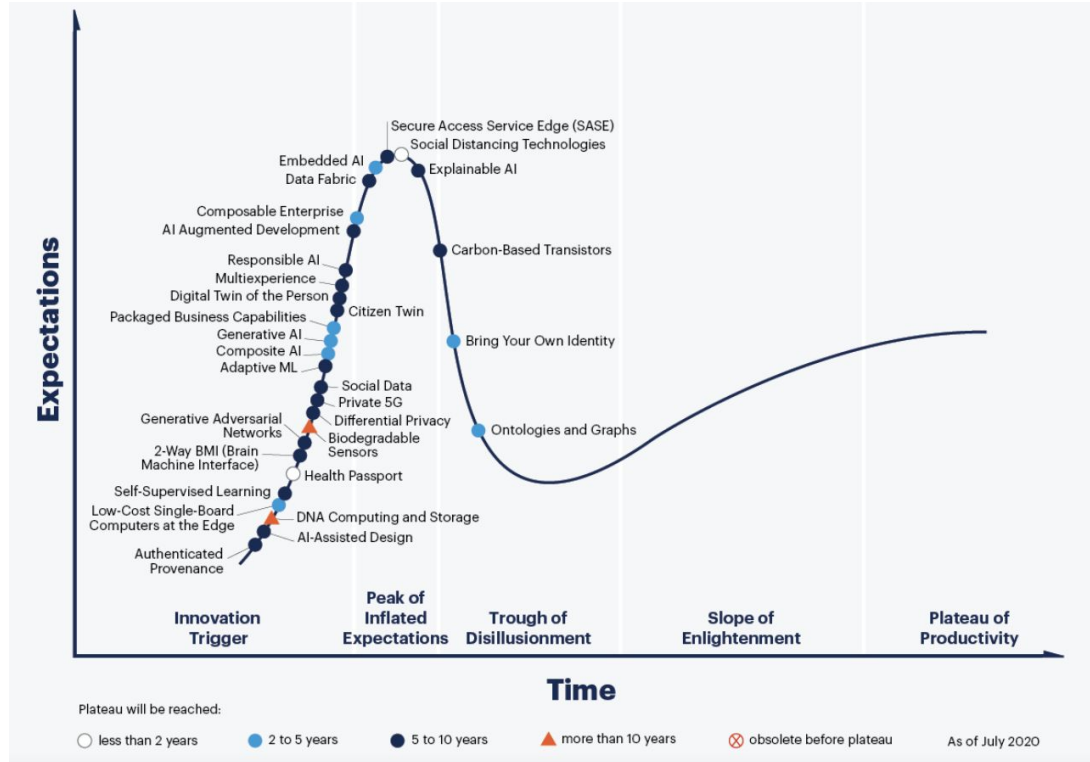
Part 1. General introduction to XAI

Popularity of XAI



Google trends popularity* index for “Explainable Artificial intelligence” (red) and “Explainable AI” (blue) over 01/2009 - 08/2021 in the category “Science”

Hype cycle for emerging technologies, 2020



Emerging technology trends*:

- Digital me
- Composite architectures
- Formative AI

Formative AI is a type of AI capable of dynamically changing to respond to a situation. There are a variety of types, ranging from AI that can dynamically adapt over time to technologies that can generate novel models to solve specific problems

Enterprises looking to explore the boundaries of AI should consider AI-assisted design, AI augmented development, ontologies and graphs, small data, composite AI, adaptive ML, self-supervised learning, generative AI and generative adversarial networks.

- Algorithmic trust

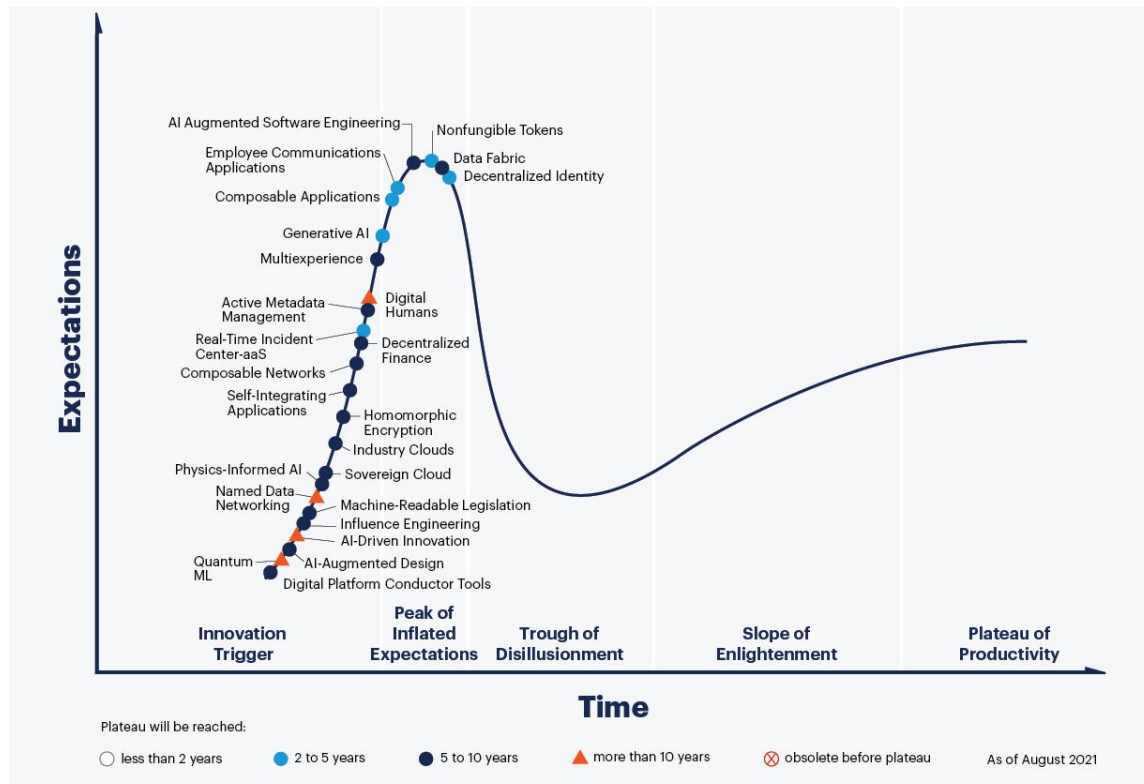
Emerging technologies tied to algorithmic trust include secure access service edge (SASE), differential privacy, authenticated provenance, bring your own identity, responsible AI and explainable AI.

- Beyond silicon

Gartner, July 2020

* <https://www.gartner.com/smarterwithgartner/5-trends-drive-the-gartner-hype-cycle-for-emerging-technologies-2020>

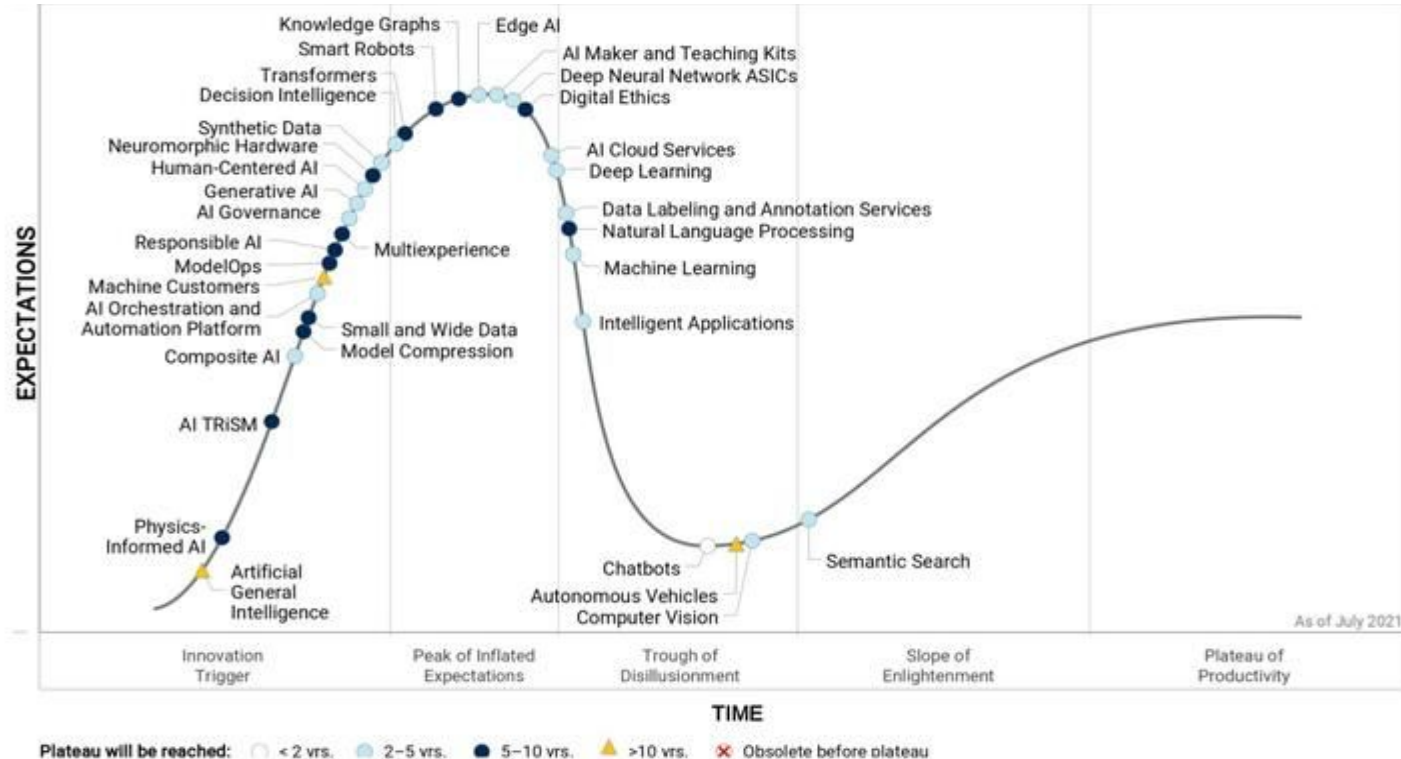
Hype cycle for emerging technologies, 2021



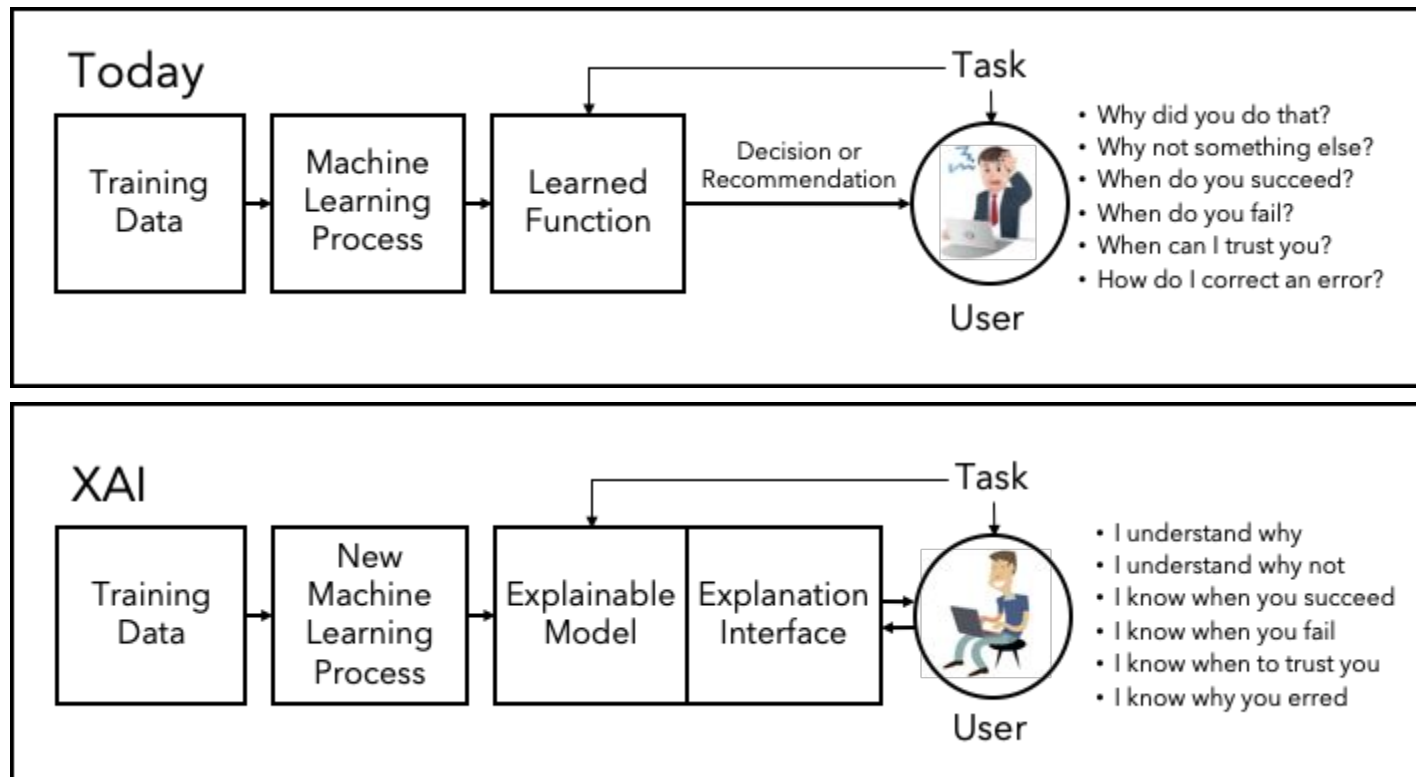
Emerging technology trends:

- Engineering trust
- Accelerating growth
- Sculpting change

Hype cycle for artificial intelligence, 2021



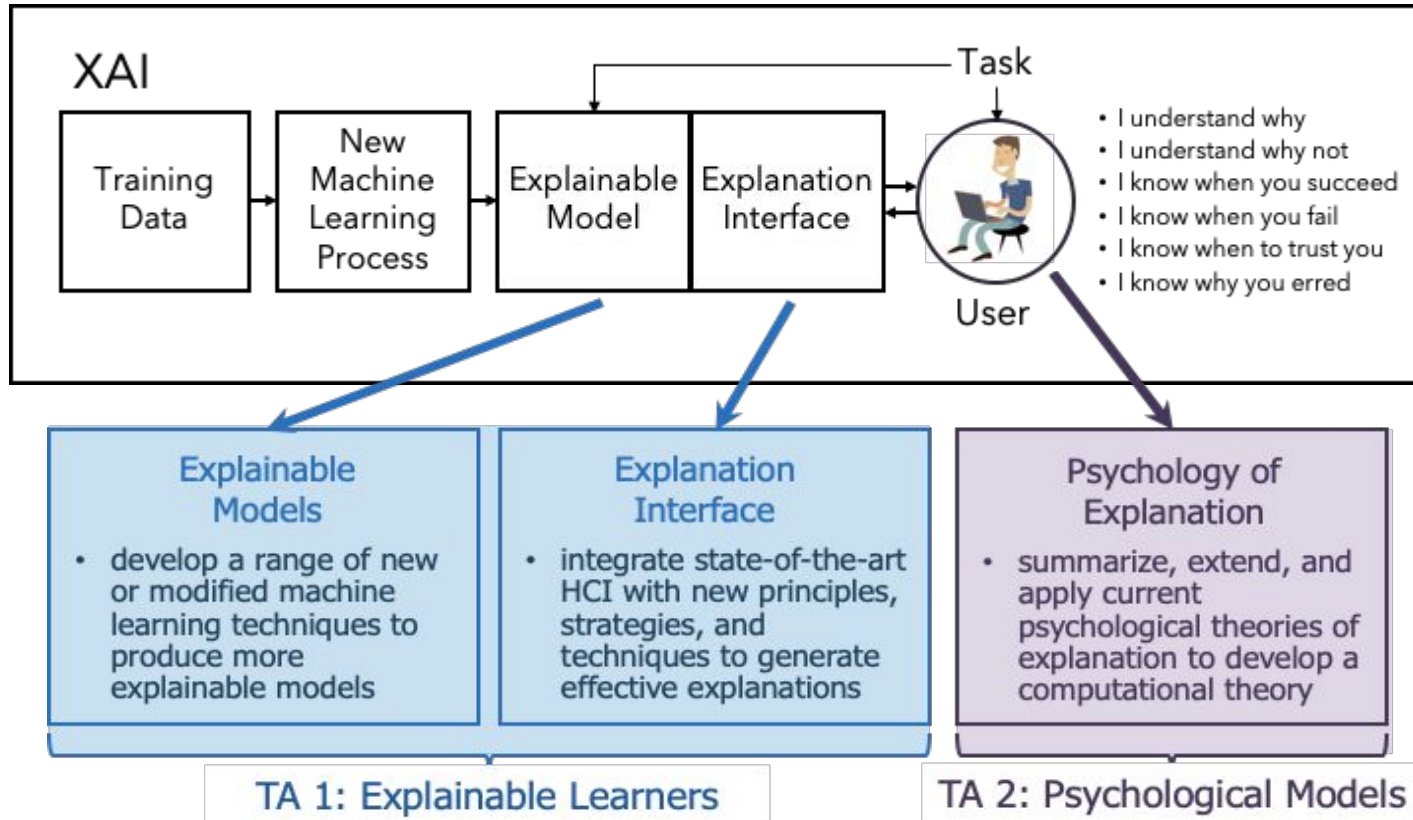
XAI concept from DARPA*



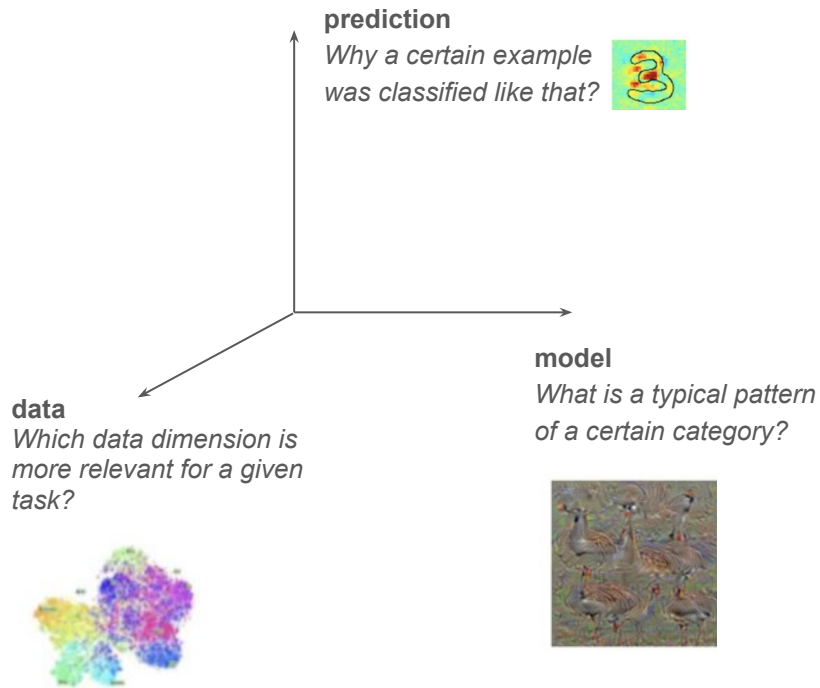
* Defense Advanced Research Project Agency

Credit: https://www.darpa.mil/attachments/XAIIndustryDay_Final.pptx, 2016

XAI challenges



Dimensions of explainability



Why do we need explanation?

XAI refers to the methods that **help** human to **understand** why a black-box model takes a particular **solution**.

What are particular goals?

- debugging
- making interpretation
- making right conclusions
- trusting



Components of trustworthiness

- Stability
- Robustness
- Reproducibility
- Confidence
- Interactivity
- Interpretability
- Explainability



Components of trustworthiness

Stability: the output does not change when a small perturbation is applied on the input or on the model

Example: small perturbation in stop signals for autonomous cars. See example in [1] about generation of visual adversarial perturbations under different physical conditions. In the paper it was proposed a general attack algorithm called Robust Physical Perturbations (RP2)

Robustness: a model that can withstand adversarial attacks

Example: some adversarial attacks for traffic signs recognition models, i.e., adding some imperceptible noise into the image such that human eyes cannot differentiate the real image and the modified one, may fool the model in a consistent manner [2]

Reproducibility: a model repeatedly obtains similar results being run several times on the same dataset

Confidence: the ability of a model to assess how unusual is a prediction vector compared to other points in the validation dataset

Example: assessment how far an analysed example from the input data distribution, assessing confidence of the prediction by probability the example in the input data

1. Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning visual classification." Proceedings of the IEEE conference on computer vision and pattern recognition (2018)
2. Morgulis, Nir, et al. "Fooling a real car with adversarial traffic signs." arXiv preprint arXiv:1907.00374 (2019)

Components of trustworthiness

Explainability: an **active** characteristic of a model, denoting any **action or procedure taken by a model** with the intent of classifying or detailing its internal functions [1,2]

Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand [2]

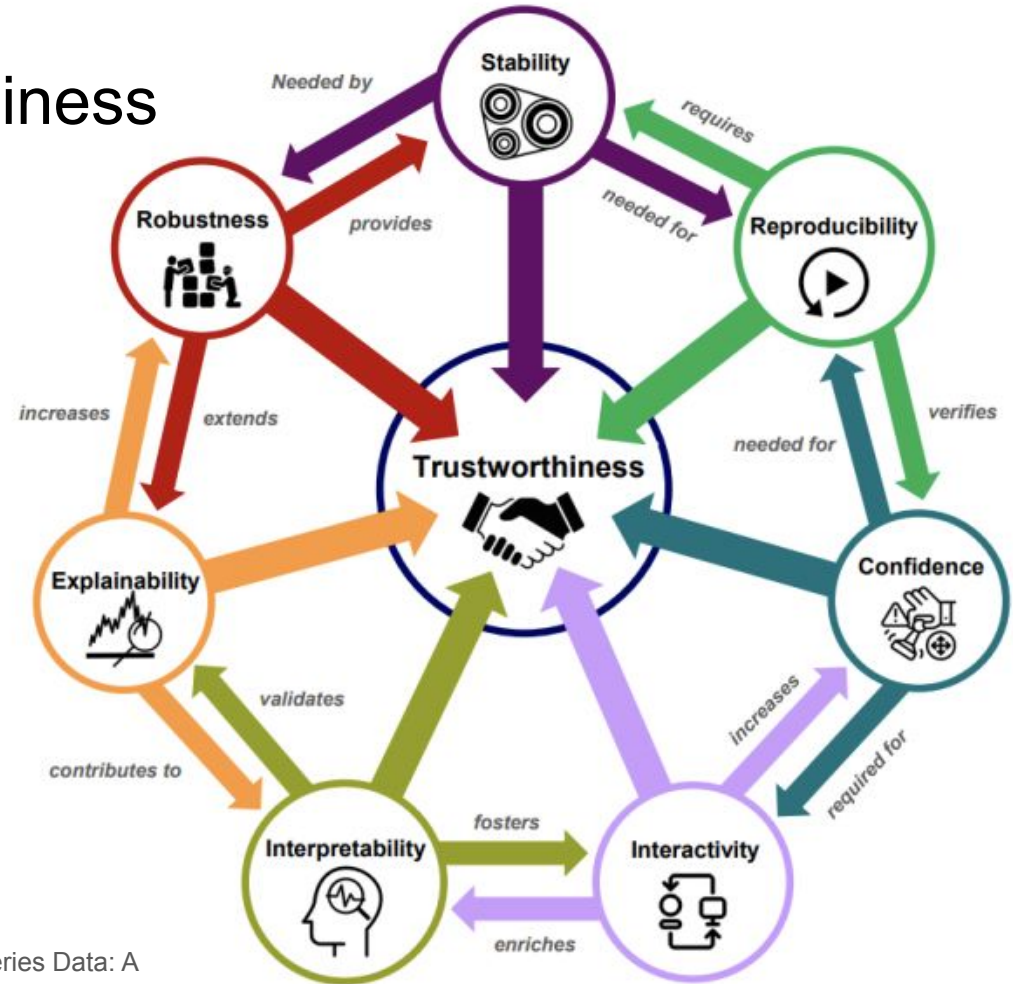
Interpretability: the **passive** characteristic of a model referring to the **level** at which a given **model makes sense** for a human observer [1]

An interpretable system is a system where a user cannot only see, but also study how inputs are mathematically mapped to outputs [3]

1. Tjoa, E., and C. Guan. "A survey on explainable artificial intelligence (XAI): Towards medical XAI. arXiv 2019." arXiv preprint arXiv:1907.07374
2. Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." Information Fusion 58 (2020): 82-115
3. Doran, Derek, Sarah Schulz, and Tarek R. Besold. "What does explainable AI really mean? A new conceptualization of perspectives." arXiv preprint arXiv:1710.00794 (2017)

Components of trustworthiness

- Stability
- Robustness
- Reproducibility
- Confidence
- Interactivity
- Interpretability
- Explainability



Shades of XAI and its target audience

Trustworthiness	Domain experts, users of the model affected by decisions
Causality	Domain experts, managers and executive board members, regulatory entities/agencies
Transferability	Domain experts, data scientists
Informativeness	All
Confidence	Domain experts, developers, managers, regulatory entities/agencies
Fairness	Users affected by model decisions, regulatory entities/agencies
Accessibility	Product owners, managers, users affected by model decisions
Interactivity	Domain experts, users affected by model decisions
Privacy awareness	Users affected by model decisions, regulatory entities/agencies

EU development concept for XAI

The main objective is a **right deployment of AI in the society**

The later is related to the following aspects:

- **Transparency of models:** it relates to the **documentation** of the AI processing chain, including the technical principles of the model, and the description of the data used for the conception of the model. This also encompasses elements that provide a **good understanding of the model**, and related to the **interpretability and explainability** of models
- **Reliability of models:** it concerns the capacity of the models to **avoid failures or malfunction**, either because of edge cases or because of malicious intentions. The main vulnerabilities of AI models have to be identified, and technical solutions have to be implemented to make sure that autonomous systems will not fail or be manipulated by an adversary
- **Protection of data in models:** the **security of data** used in AI models needs to be preserved. In the case of sensitive data, for instance personal data, the risks should be managed by the application of proper organisational and technical controls

EU development concept for XAI

*“On many aspects however, AI systems that are currently under development are **far from achieving the minimal requirements of safety and security** that would be expected from autonomous systems.”*

“As of now, several avenues for reflection could be considered to undertake the implementation of standards in AI technologies, and of security and reliability certifications of AI components embedded in real systems. These avenues include:

*— developing a **methodology to evaluate the impacts of AI systems on society** built on the model of the Data Protection Impact Assessments (DPIA) introduced in the GDPR, that would provide an assessment of the risks involved in the usage of AI models to the users and organisations;*

*— introducing **standardized tests to assess the robustness** of AI models, in particular to determine their field of action with respect to the data that have been used for the training, the type of mathematical model, and the context of use, amongst others factors;*

*— raising **awareness** among AI practitioners through the publication of **good practices regarding to known vulnerabilities** of AI models, and technical solutions to address them;*

*— promoting **transparency** in the conception of machine learning models, emphasizing the need of an **explainability-by-design** approach for AI systems with potential negative impacts on fundamental rights of users.”*

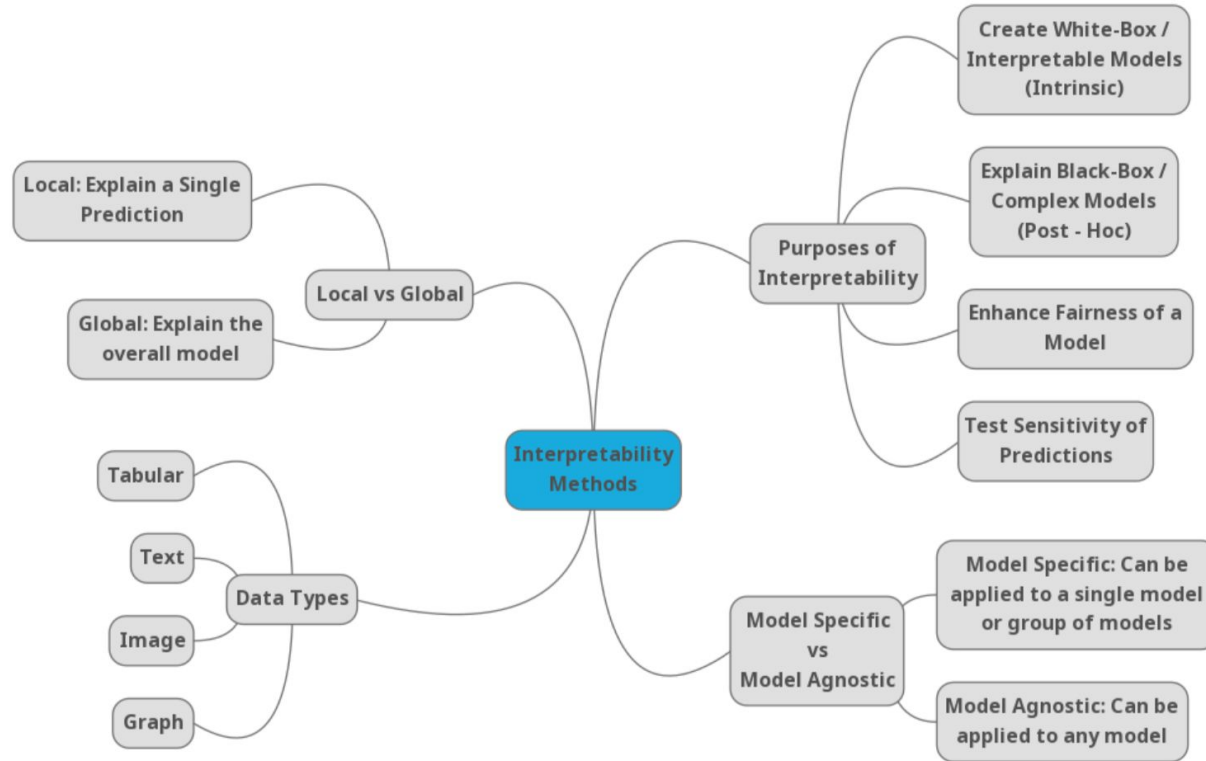
EU rights for explanation

*“The data subject shall have the right **not to be subject to a decision based solely on automated processing**, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her”*

The latter is related to **the rights of the data subject** for

- understanding a certain decision
- contesting it
- altering to get the desired result

Taxonomy of ML techniques for explainability

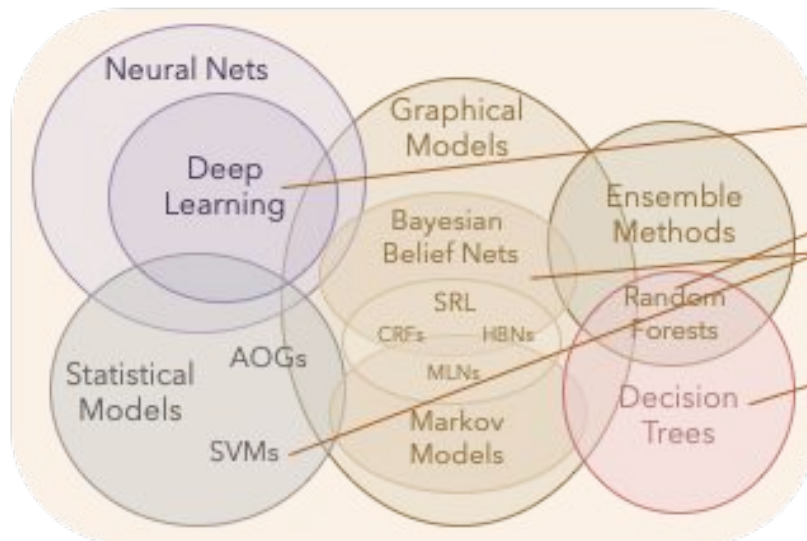


Accuracy vs Explainability

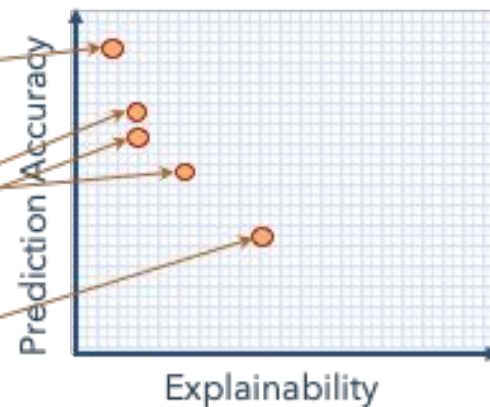
New Approach

Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



Explainability (notional)



Accuracy vs Explainability. Some arguments against XAI

- Searching iteratively for meaningful features
 - Clock prediction¹
- Explanation models have errors w.r.t. the explainable model
 - low-fidelity
 - the features may be used in explanation model differently than in the original model
- Instead of explanation the explanation models show the trends in how predictions are related to the features
 - COMPAS example

1. <https://towardsdatascience.com/training-neural-net-to-read-clock-time-9473175171e3>

The COMPAS algorithm case

The COMPAS recidivism algorithm has been developed to predict whether someone **will be arrested** within a certain time **after being released from jail/prison**

The main features for such kind of models might be **age** and **criminal history**. Moreover, the **dependencies** between the risk and these important features might be non-linear

However, in some datasets important features (e.g., **age** and **criminal history**) might **correlate** with **sensitive features** (e.g., **race**)

Thus, with high risk, an explanation model may conclude that “These persons are predicted to be arrested because they are black”

The COMPAS algorithm case

ProPublica conclusion: a linear explanation model where the recidivism risk **depends on race**, conditioned on age and criminal history [1,2]

C. Rudin et al. conclusion: COMPAS seems to be nonlinear. It is entirely possible that it **does not depend on race** (beyond its correlations with age and criminal history) [3]

1. Angwin J, Larson J, Mattu S, Kirchner L. Machine Bias. ProPublica; 2016. [Online] <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

2. Larson J, Mattu S, Kirchner L, Angwin J. How We Analyzed the COMPAS Recidivism Algorithm. ProPublica; 2016. [Online] <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>

3. Rudin C, Wang C, Coker B. The age of secrecy and unfairness in recidivism prediction. arXiv e-prints 1811.00731 [applied statistics]. 2018 Nov.

COMPAS vs CORELS

CORELS rule list

IF	age between 18-20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21-23 and 2-3 prior offenses	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest.	

Comparison with COMPAS

COMPAS	CORELS
black box 130+ factors might include socio-economic info expensive (software licence) within software used in US Justice System	white box only age, priors, (optional) gender no other information free, transparent

Intellectual property vs benefit to society

Shifting the business model:

- reducing the industrial participation and replacing opaque under the Intellectual property rights
- making focus on transparency and making them appealing to academics and charitable organizations

White- vs black-box models

Arguments supporting white-box models:


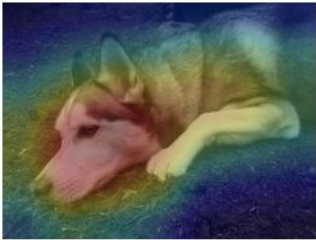

1. Discovering meaningful features through iterative process may allow to build white-box models of a quality comparable with the black-box models
2. Explainable ML methods provide explanations that are not faithful to what the original model computes
3. Explanations often do not make sense (or is not detailed enough) to understand what the black box is doing

Arguments supporting black-box models:

1. Supporting a true concurrency for products ranked by a recommendation system
2. Counterfactual explanation might be a good tool for explaining black-box models

Explanation does not explain

Attention map may provide essentially the same explanation for multiple classes!

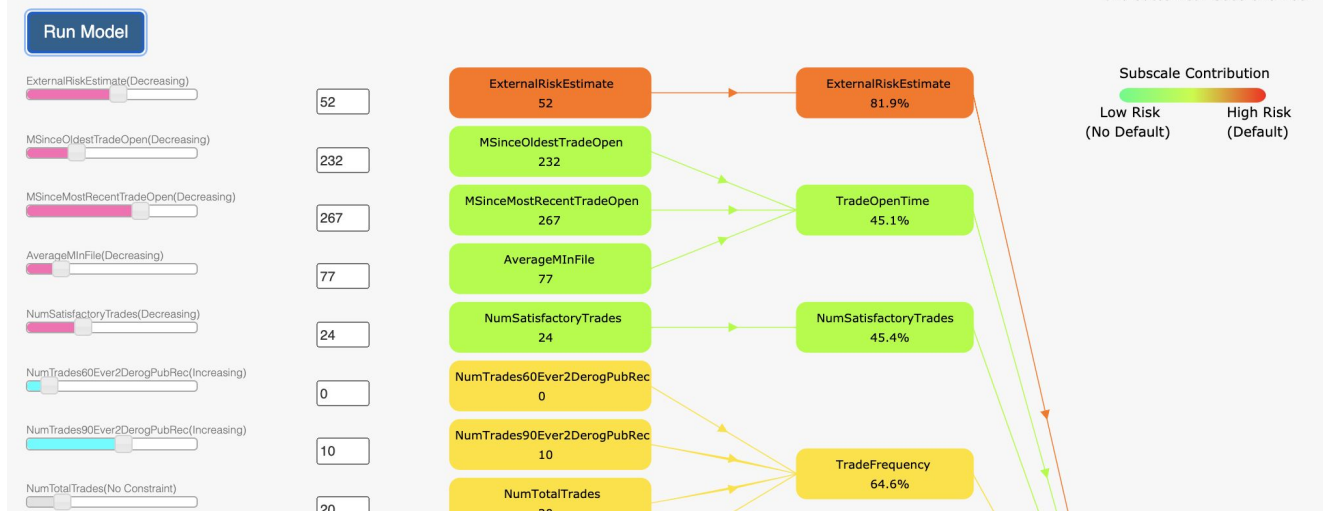
	Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps			

Be careful and check not only the explanation of the true classes but also the other classes

2-layer additive risk model. An interpretable analog of NN

Global Model

Below is our Input Panel. Click on variable names or check [Appendix](#) for more details. Model will take a few seconds to run.



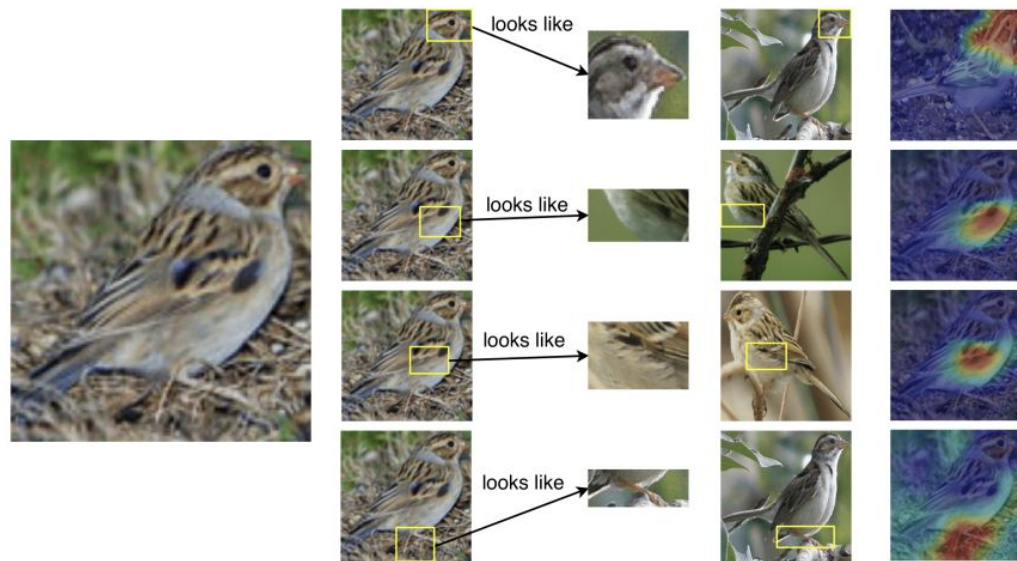
FICO explainable
ML challenge for
predicting credit risk

Detailed description: Chen C, Lin K, Rudin C, Shaposhnik Y, Wang S, Wang T. An Interpretable Model with Globally Consistent Explanations for Credit Risk. In: Proceedings of NeurIPS 2018 Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy; 2018

Source: <http://dukedatasciencefico.cs.duke.edu/models/>

ProtoPNet. An explainable CNN

Idea: imitate the reasoning process which is qualitatively similar to that of humans, namely paying attention to the part of images that are similar to the prototypes



Leftmost: a test image of a clay-colored sparrow

Second column: same test image, each with a bounding box generated by our model -- the content within the bounding box is considered by our model to look similar to the prototypical part (same row, third column) learned by our algorithm

Third column: prototypical parts learned by our algorithm

Fourth column: source images of the prototypical parts in the third column

Rightmost column: activation maps indicating how similar each prototypical part resembles part of the test bird

InfoGAN and Beta-VAE. Disentanglement of the latent space

β -VAE



VAE



InfoGAN

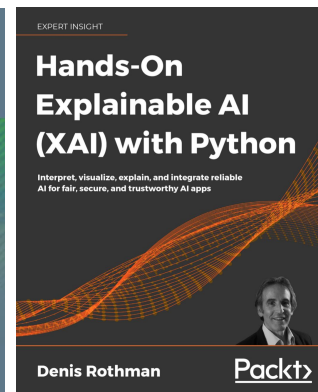
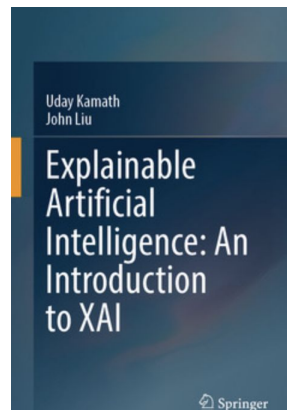
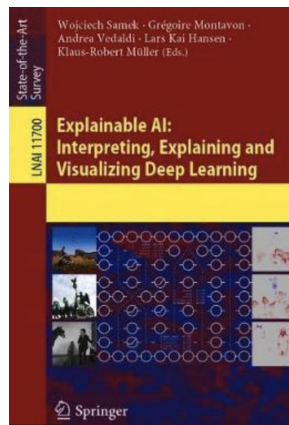
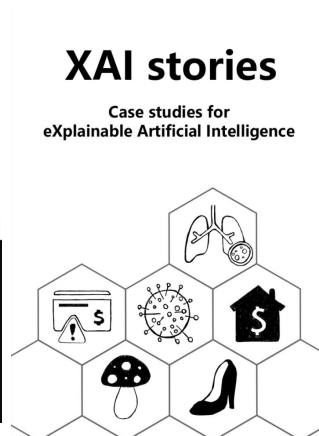
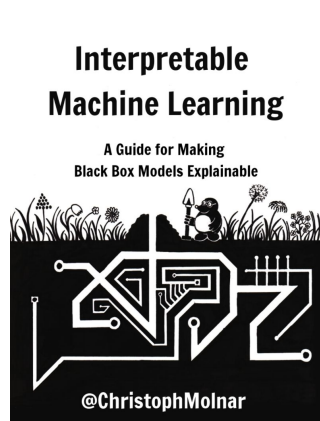


Wrapping up...

- Explainable AI is a **hot topic** in Computer Science due to the increased need on the part of **different categories** of people to **trust black-box models**
- The ideal scenario is to use **interpretable** models
- Each explanation method has its own flaws, thus should be applied **consciously** (with due regard to shortcomings)
- **Black-box models** as well as methods for their explanation **should be regulated**

Sources and avenues to check

- Avenues: <http://www.wikicfp.com/cfp/call?conference=explainable%20ai>
- Books



Part 2. To practice

Pre-step: exploratory data analysis

- Data summary (quantitative evaluation / statistics)
- Data plotting:
 - 2D projections: principal components analysis (PCA), multidimensional scaling (MDS), t-distributed stochastic neighbor embedding (t-SNE), and autoencoder networks
 - Correlation network graphs [1]
- Data leakage / dataset shift [2]

Do not underestimate its importance!

1. https://github.com/jphall663/corr_graph

2. Dignum, Virginia. "The Myth of Complete AI-Fairness." International Conference on Artificial Intelligence in Medicine. Springer, Cham, 2021

Post-step: getting insights into prediction behaviour of your model

- Partial dependence plot (PDP)
- Individual conditional expectation (ICE)
- Accumulated local effects (ALE)

Partial dependencies plot (PDP)

Goal: to show the marginal effect one or two features have on the predicted outcome of a machine learning model (regression or classification, where the model outputs class probabilities)

$$\hat{f}_{x_S}(x_S) = E_{x_C} \left[\hat{f}(x_S, x_C) \right] = \int \hat{f}(x_S, x_C) d\mathbb{P}(x_C)$$

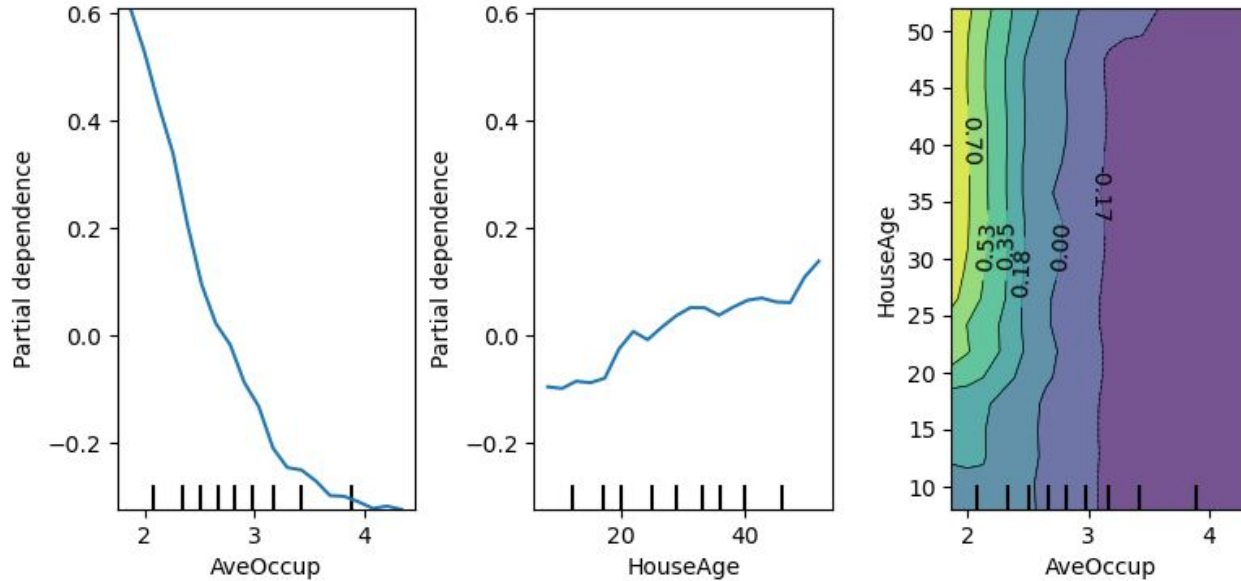
x_S are the features for which the PDP is plotted, and x_C are the remaining features

The Monte Carlo approximation:

$$\hat{f}_{x_S}(x_S) = \frac{1}{n} \sum_{i=1}^n \hat{f}(x_S, x_C^{(i)})$$

Main assumption: features in C are not correlated with the features in S (OOD problem)

Example. The California housing dataset

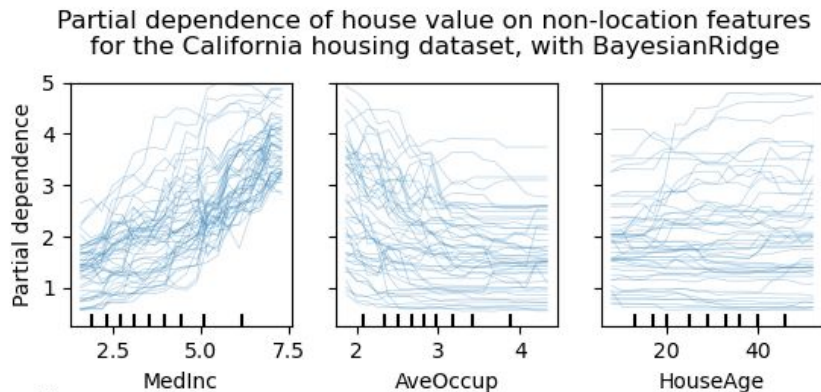


- Credit: https://scikit-learn.org/stable/modules/partial_dependence.html#individual-conditional

Individual conditional expectation (ICE)

$$\hat{f}_{x_S}^{(i)}(x_S) = \hat{f}(x_S, x_C^{(i)})$$

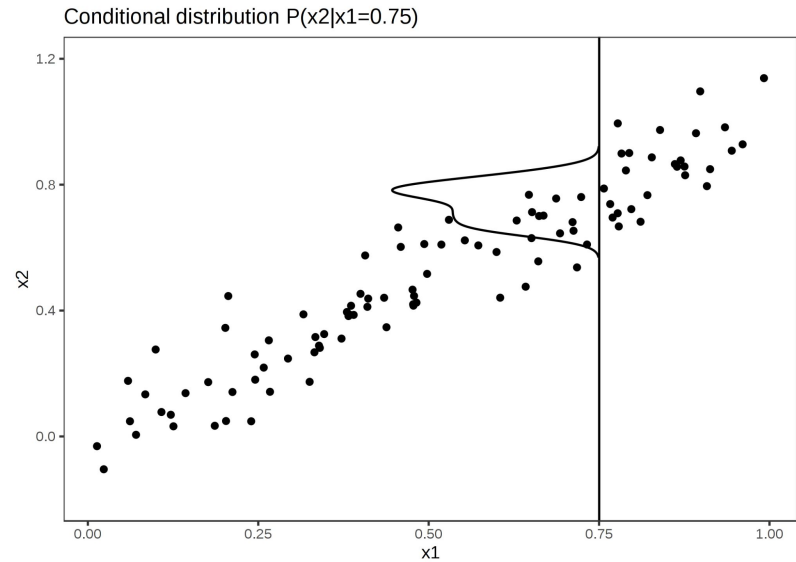
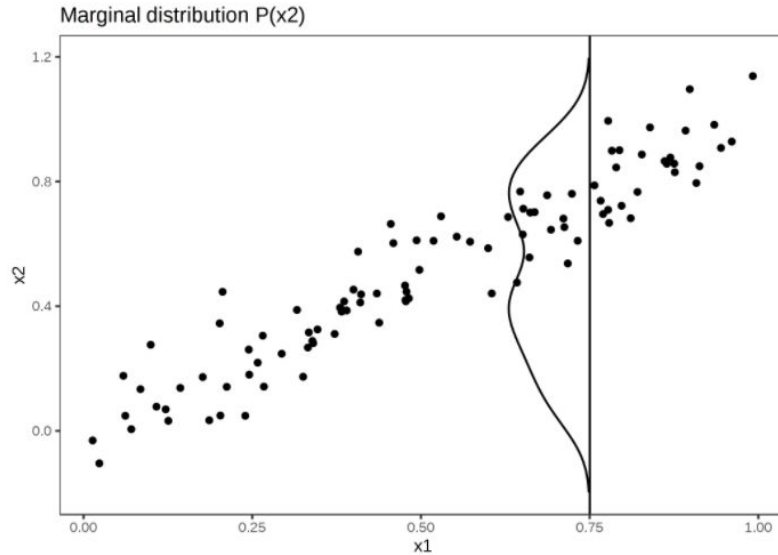
Example:



Credit:

https://scikit-learn.org/stable/modules/partial_dependence.html#individual-conditional

Correlated features. Marginal vs conditional distribution



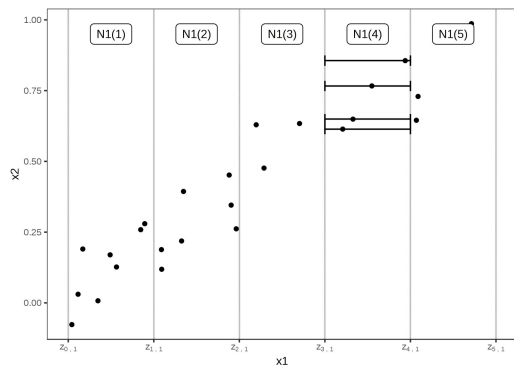
Marginal plots (M-plot)

$$\hat{f}_{x_S, M}(x_S) = E_{X_C | X_S} \left[\hat{f}(X_S, X_C) | X_S = x_S \right] = \int_{x_C} \hat{f}(x_S, x_C) \mathbb{P}(x_C | x_S) dx_C$$

Accumulated local effects (ALE)

Accumulated local effects describe how features influence the prediction of a machine learning model on average.

$$\hat{f}_{x_S, ALE}(x_S) = \int_{z_{0,1}}^{x_S} E_{X_C|X_S} \left[\hat{f}^S(X_S, X_C) | X_S = z_S \right] dz_S - \text{const} = \int_{z_{0,1}}^{x_S} \int_{x_C} \hat{f}^S(z_S, x_C) \mathbb{P}(x_C | z_S) dx_C dz_S - \text{const}$$



How to compute:

1. Define a grid (usually equal-height)
2. For all data points in the interval compute the difference in the prediction in the rightmost and leftmost points of the interval or gradient
3. Accumulate and center all the values

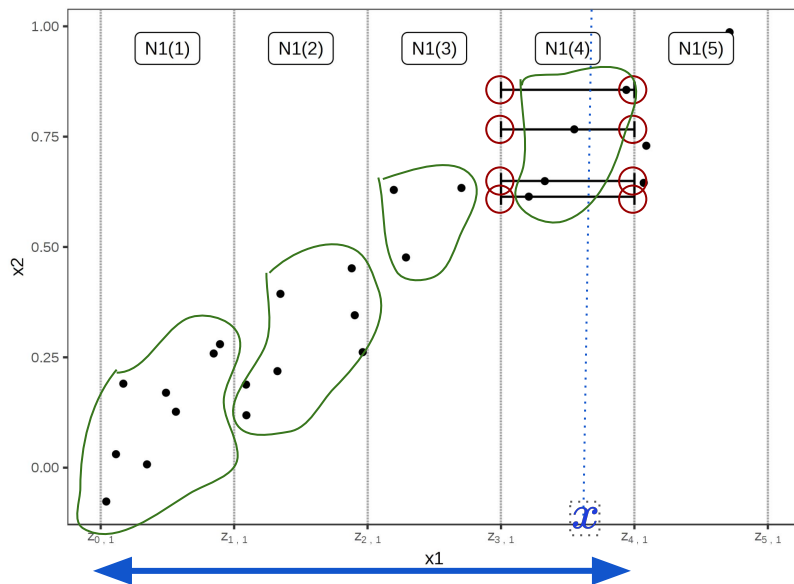
1. Compute the gradient

$$\hat{f}^S(x_s, x_c) = \frac{\delta \hat{f}(x_s, x_c)}{\delta x_s}$$

2. Accumulate and center all the values

Accumulated local effects (ALE). Non-centred estimates

$$\hat{\tilde{f}}_{j,ALE}(x) = \underbrace{\sum_{k=1}^{k_j(x)}}_{\text{accumulated}} \frac{1}{n_j(k)} \underbrace{\sum_{i: x_j^{(i)} \in N_j(k)}}_{\text{local}} \left[\underbrace{f(z_{k,j}, x_{\setminus j}^{(i)}) - f(z_{k-1,j}, x_{\setminus j}^{(i)})}_{\text{effects}} \right]$$



Apley, Daniel W., and Jingyu Zhu.
 “Visualizing the effects of predictor
 variables in black box supervised
 learning models.” Journal of the Royal
 Statistical Society: Series B (Statistical
 Methodology) 82.4 (2020): 1059-1086

Accumulated local effects (ALE). Estimates

Uncentered estimates

$$\hat{\tilde{f}}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i: x_j^{(i)} \in N_j(k)} \left[f(z_{k,j}, x_{\setminus j}^{(i)}) - f(z_{k-1,j}, x_{\setminus j}^{(i)}) \right]$$

Centered estimates

$$\hat{f}_{j,ALE}(x) = \hat{\tilde{f}}_{j,ALE}(x) - \frac{1}{n} \sum_{i=1}^n \hat{\tilde{f}}_{j,ALE}(x_j^{(i)})$$

ALE: pros and cons

Pros:

- ALE works well even with correlated data
- faster than PDP
- ALE shows the relative effect of the feature (difference with the mean prediction)
- the 2D ALE plot only shows the interaction: if two features do not interact, the plot shows nothing

Cons:

- depends on the number of intervals (with small intervals it's more shaky), with large intervals it can be too smooth and blur out the important interactions. Especially 2D plots
- shows averaged effects (ICE are not available)
- computation not really intuitive

Wrapping up

- **Partial Dependence Plots:** “Let me show you what the model predicts on average when each data instance has the value v for that feature. I ignore whether the value v makes sense for all data instances.”
- **M-Plots:** “Let me show you what the model predicts on average for data instances that have values close to v for that feature. The effect could be due to that feature, but also due to correlated features.”
- **ALE plots:** “Let me show you how the model predictions change in a small “window” of the feature around v for data instances in that window.”

Further reading

Feature interaction:

- Friedman's H-statistic [1]
- Variable Interaction Networks [2]

Residual Analysis:

https://github.com/jphall663/interpretable_machine_learning_with_python/blob/master/resid_sens_analysis.ipynb

Etc...

1. Friedman, Jerome H, and Bogdan E Popescu. "Predictive learning via rule ensembles." The Annals of Applied Statistics. JSTOR, 916–54. (2008)
2. Hooker, Giles. "Discovering additive structure in black box functions." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. (2004)
3. Christoph, Molnar "A Guide for Making Black Box Models Explainable", Feature interaction, <https://christophm.github.io/interpretable-ml-book/interaction.html#interaction> (2021)