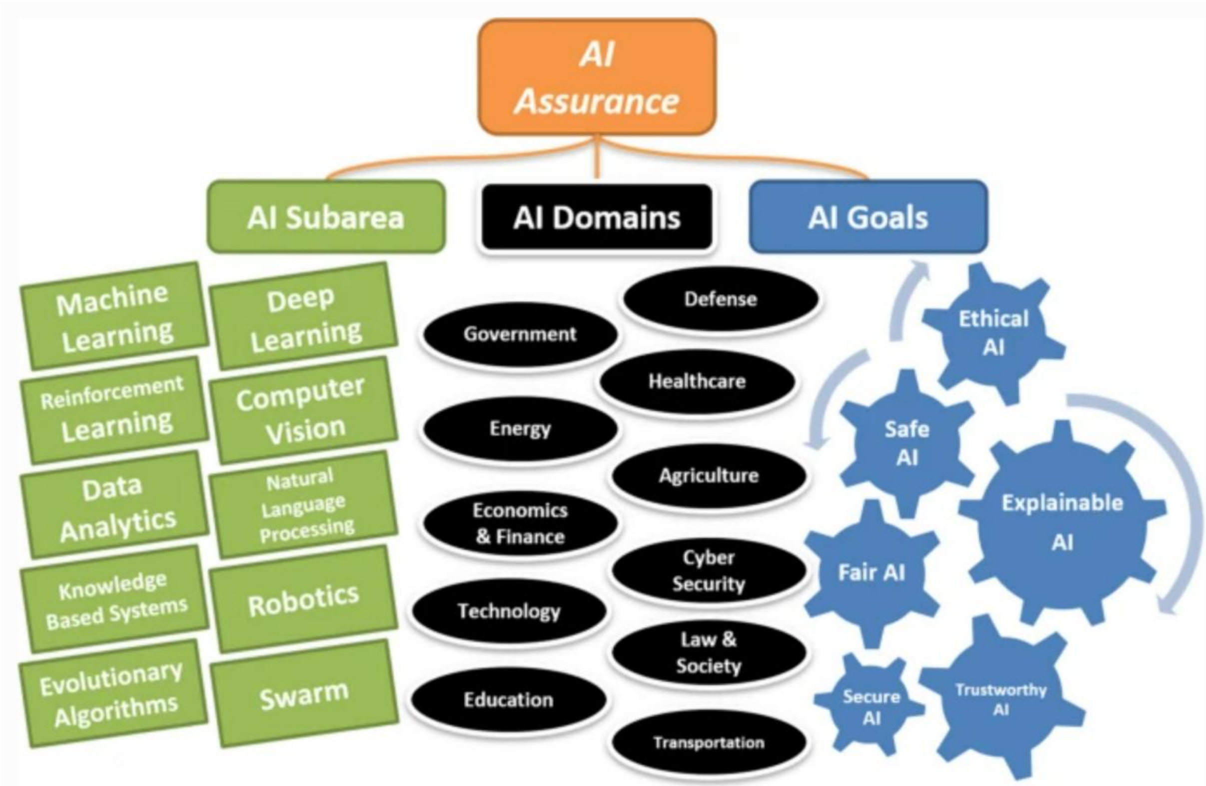
Explainable AI
(XAI)

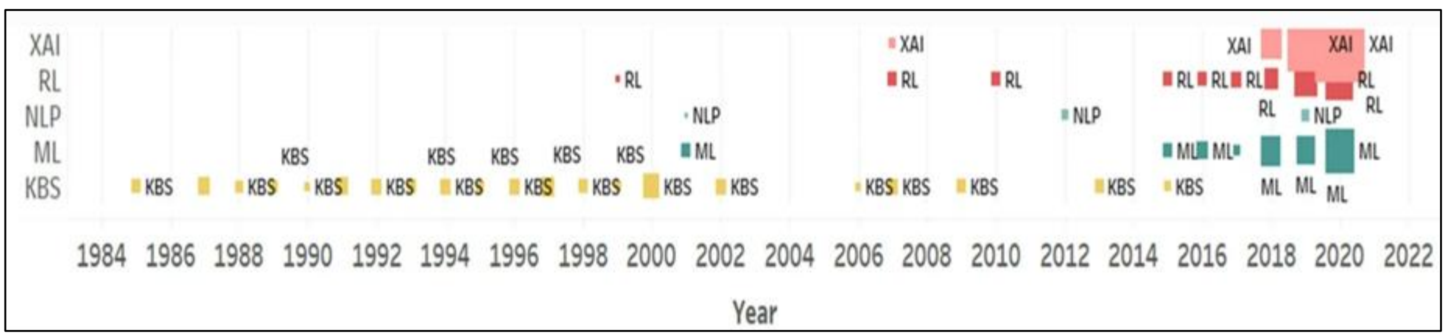
Kim Hyun Joo
Hanyang University
Mobile&Network & AI Lab.
2023/04/07

Table of Contents

Introduction to XAI

- Three-dimensional AI assurance by subarea, domain, and goal



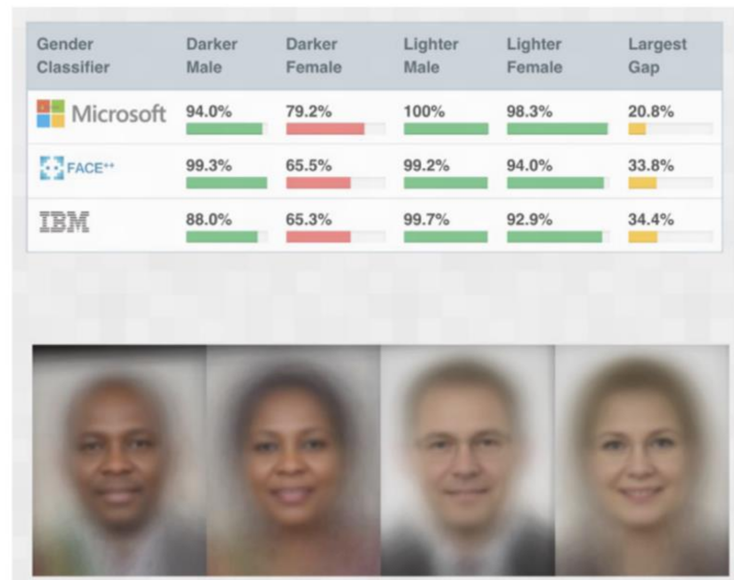


Feras et al,. A survey on artificial intelligence assurance, 2021

한양대학교 HANYANG UNIVERSITY

❑ Need for explainability in machine learning

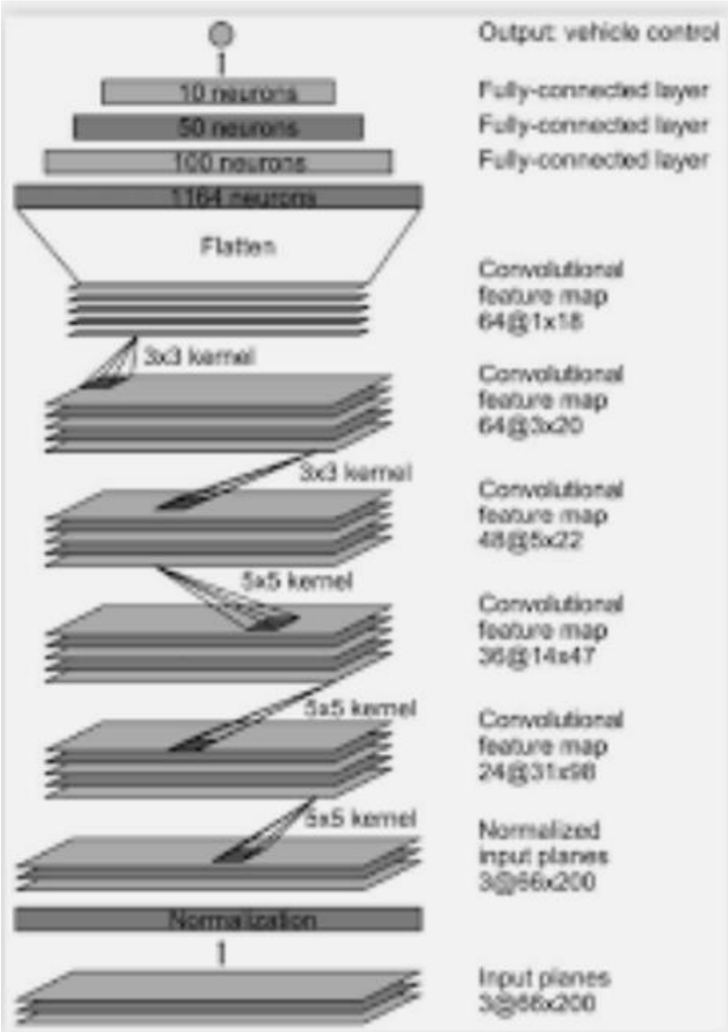
- Essential for critical systems, e.g. autonomous steering, healthcare…
- Legal reason : responsibility, confidentiality, discriminability of ML system
- For help to debug, improve algorithms



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Joy Buolamwini, Timnit Gebru: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. FAT 2018: 77-91

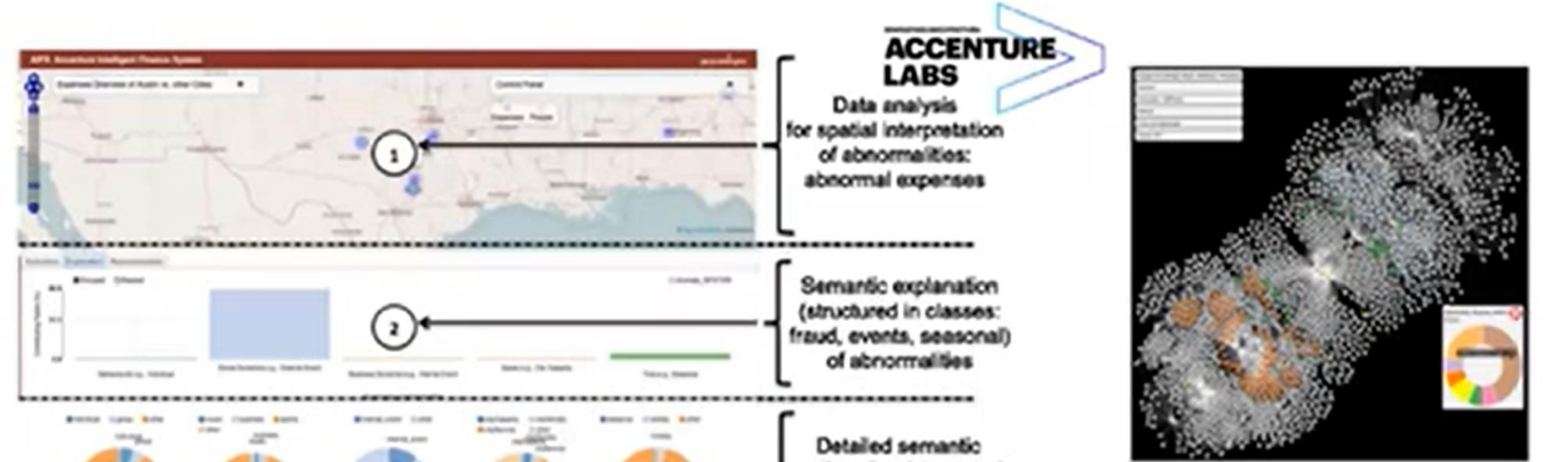
❑ PilotNet architecture



Nvidia/Google research, 2017

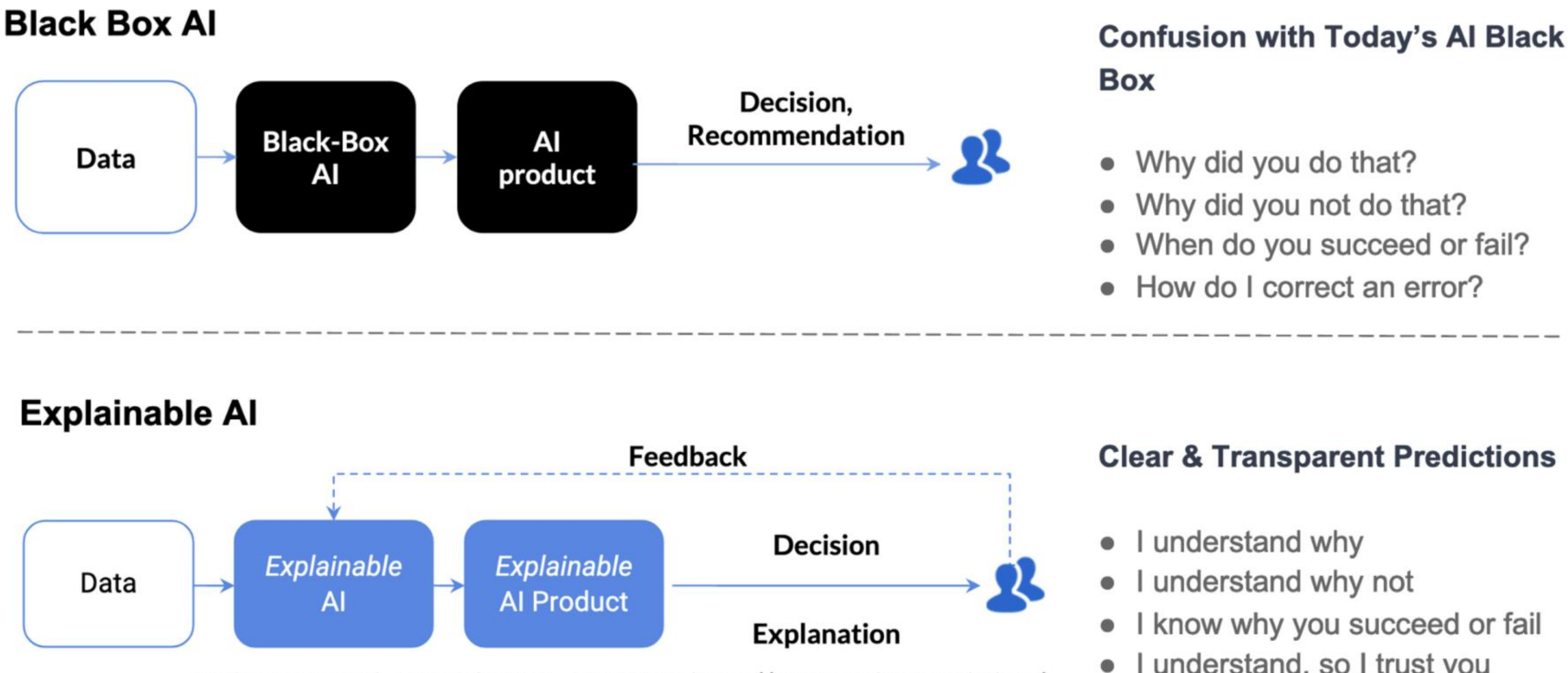
Explaining the decisions of PilotNet

❑ Explainable Anomaly Detection - Finance



Freddy Lécué, Jiewen Wu: Explaining and predicting abnormal expenses at large scale using knowledge graph based reasoning. J. Web Sem. 44: 89-103 (2017)

❑ Black-box vs explainable models

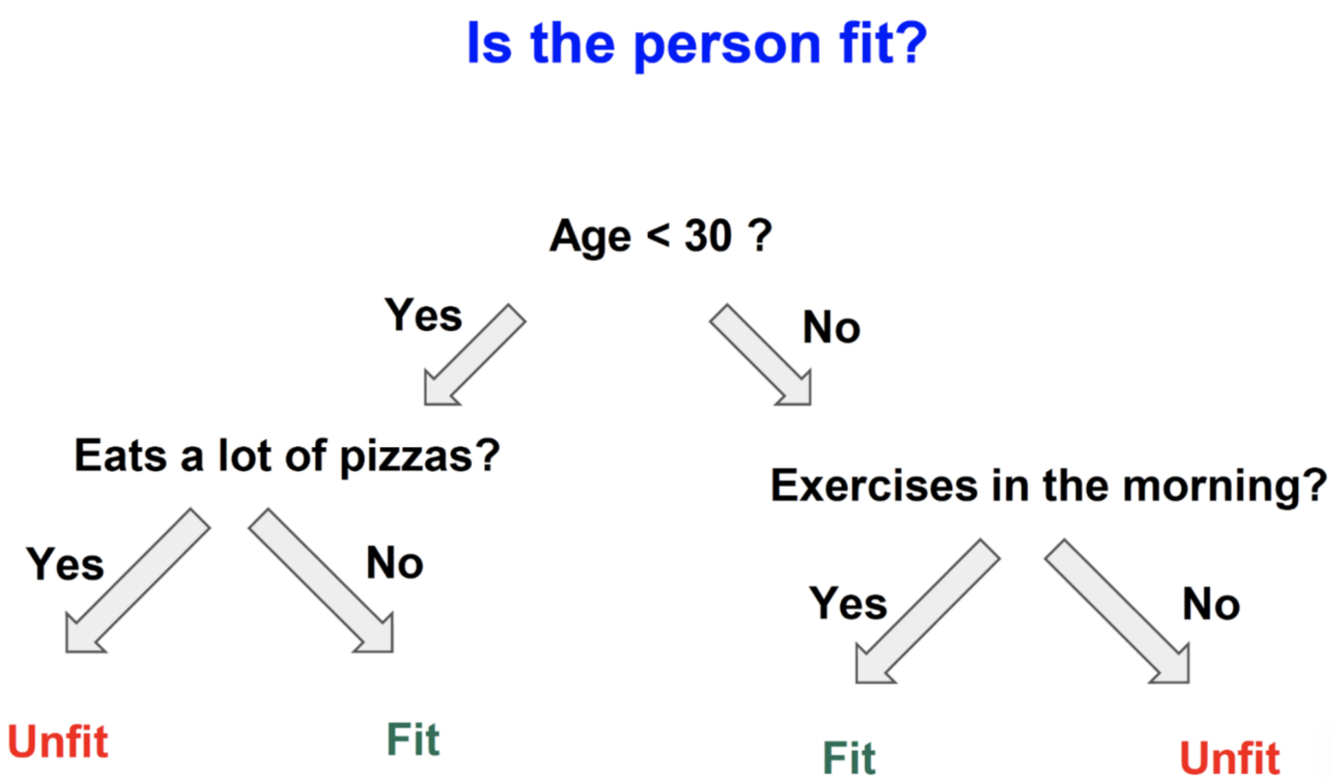


Credit: Lecue et al., Tutorial on XAI. AAAI 2020. https://xaitutorial2020.github.io/

Introduction to XAI

❑ Some ML models naturally explainable

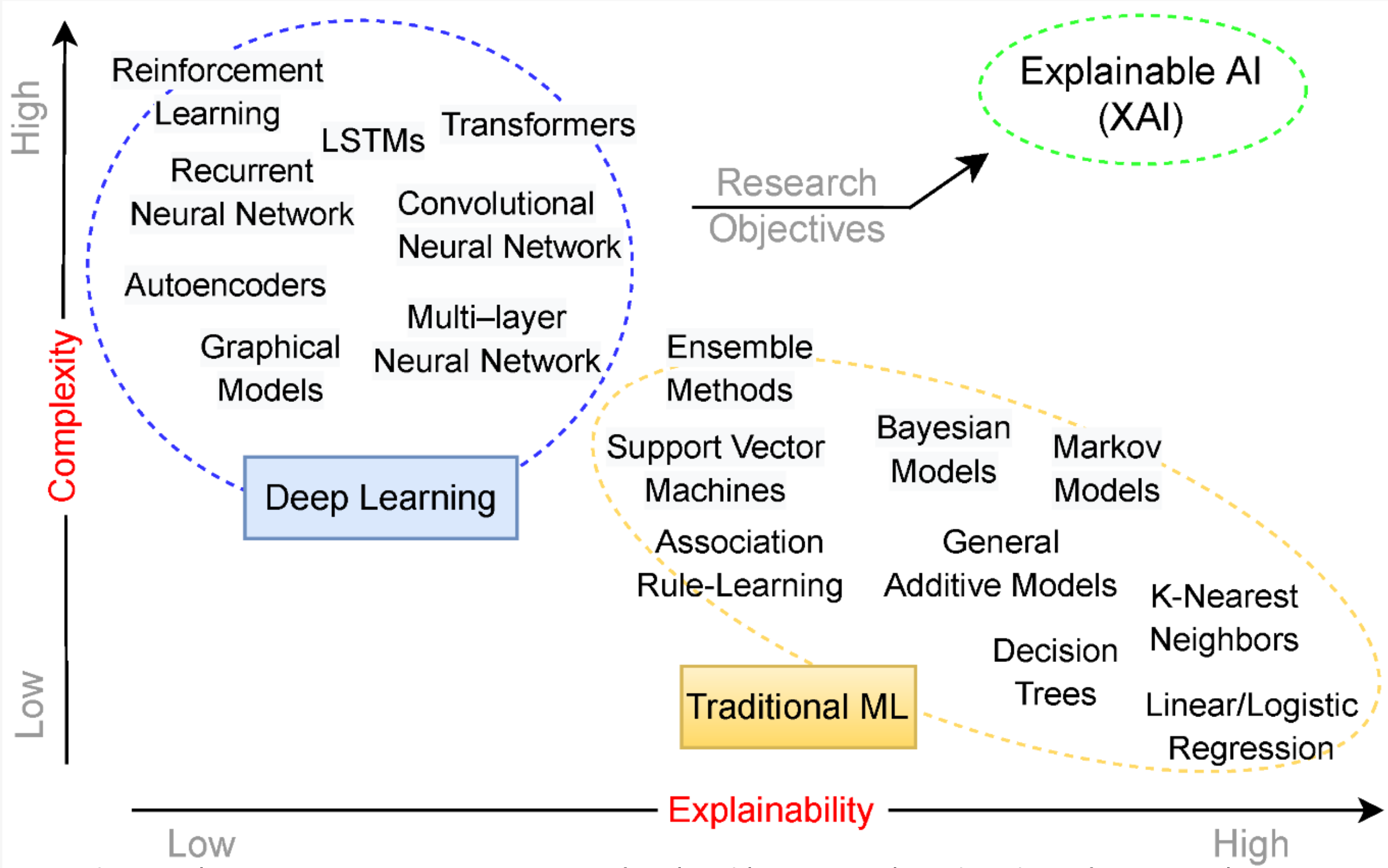
- Decision trees, Lists, and Sets and rules
- (Generalized) Linear models, (generalized) additive models, k-NN



KDD 2019 Tutorial on Explainable AI in Industry - https://sites.google.com/view/kdd19-explainable-ai-tutorial

❑ Explainable models

Figure 1. Classification of AI models according to their level of complexity, explainability, and their potential in modern AI applications.



Tobias et al., XAIR: A Systematic Metareview of Explainable AI (XAI) Aligned to the Software Development Process 2023

❑ Explainability in different data types

Table of baby-name data
(baby-2010.csv)

name	rank	gender	year
Jacob	1	boy	2010
Isabella	1	girl	2010
Ethan	2	boy	2010
Sophia	2	girl	2010
Michael	3	boy	2010

Field names

One row (4 fields)

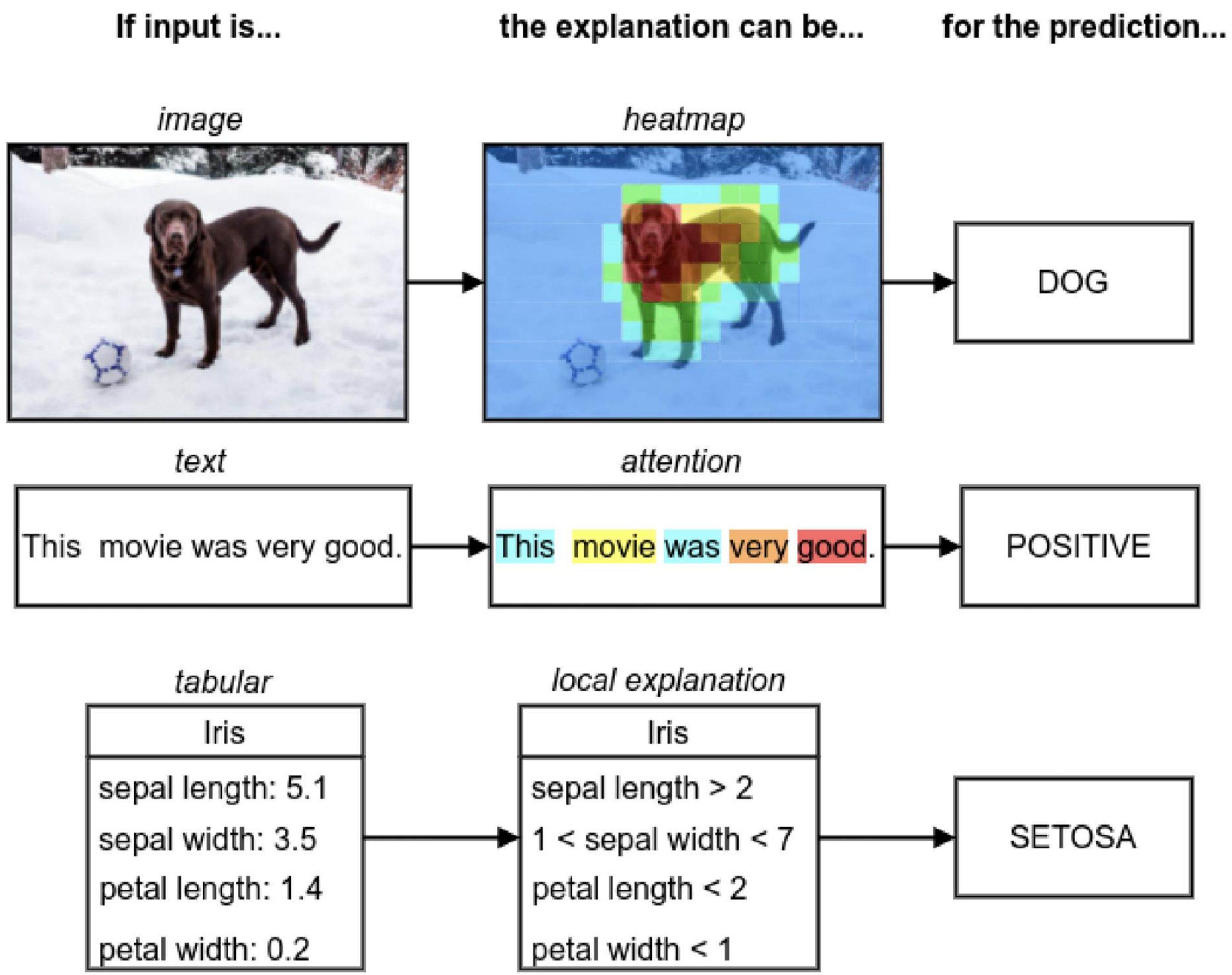
2000 rows all told

Tabular

Images

Text

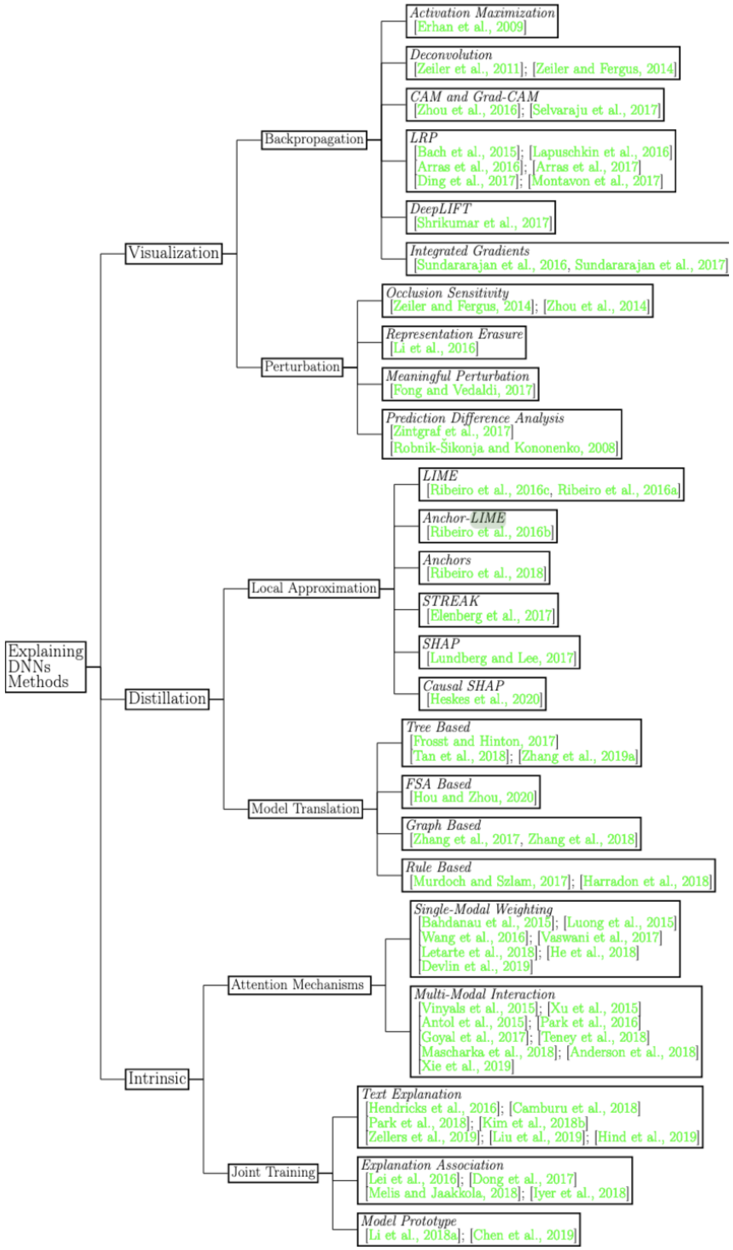
❑ Explainability in different data types



Introduction to XAI

❑ Explainability in ML/DL

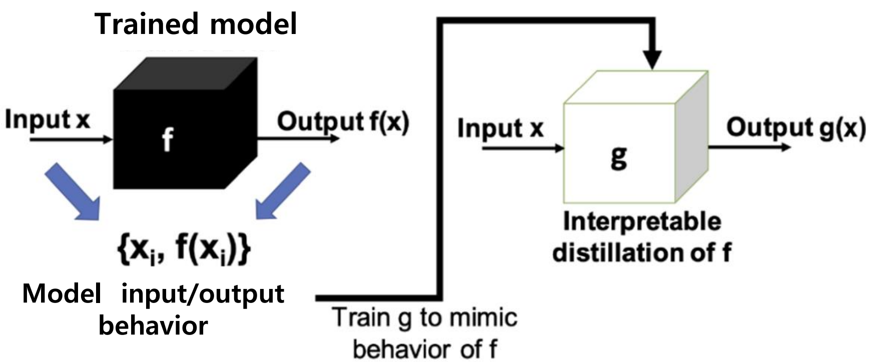
- Visualization: for data with local info (text, audio, images)
- Distillation: approximate non X-AI models with explainable one
- Intrinsic: make the model explicitly explainable

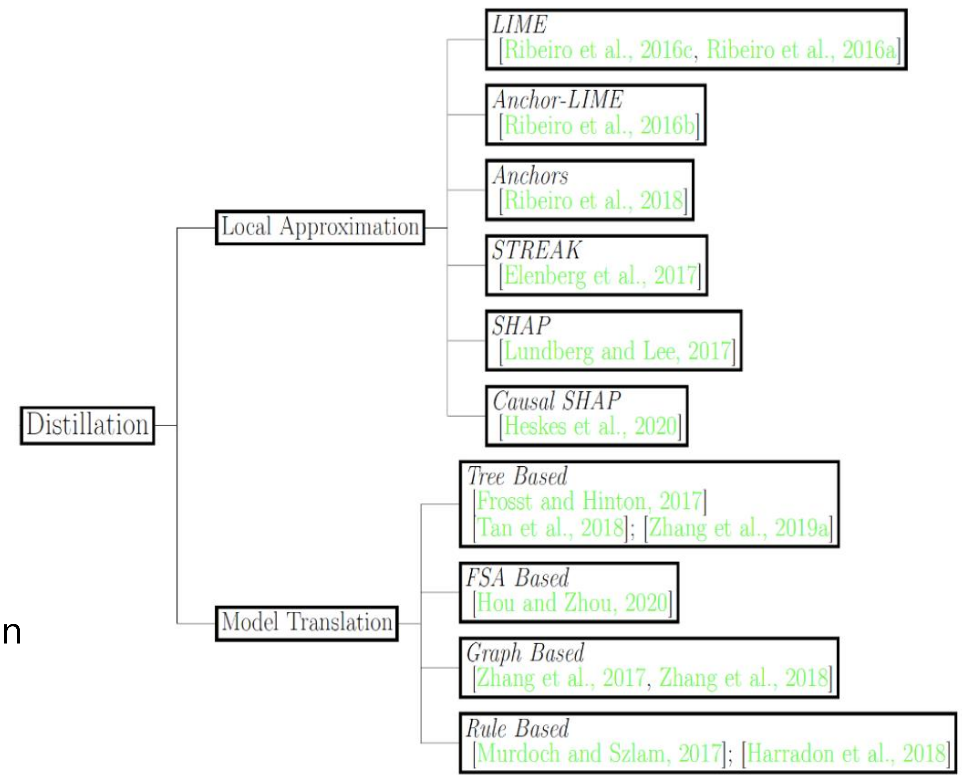


[Xie et al., 2020] Xie, N., Ras, G., van Gerven, M., and Doran, D. (2020). Explainable deep learning: A field guide for the uninitiated.

❑ Distillation methods

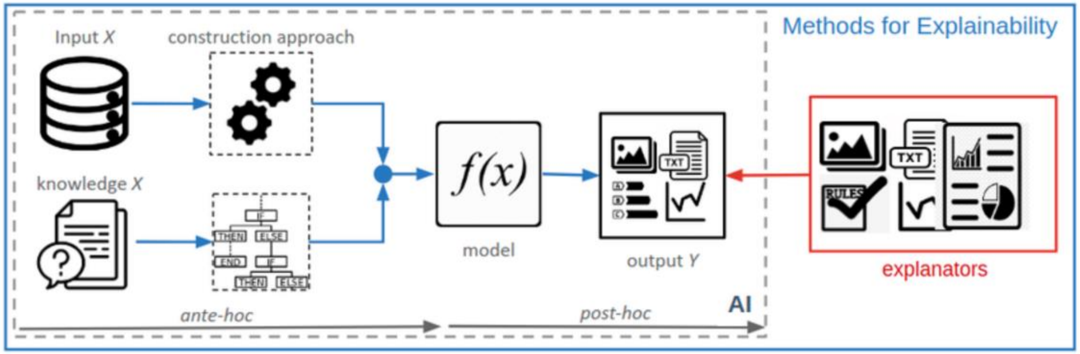
- Complex blackbox model f, simpler explainable one: $g(x) \approx f(x)$
- Perfs of f not necessarily below g



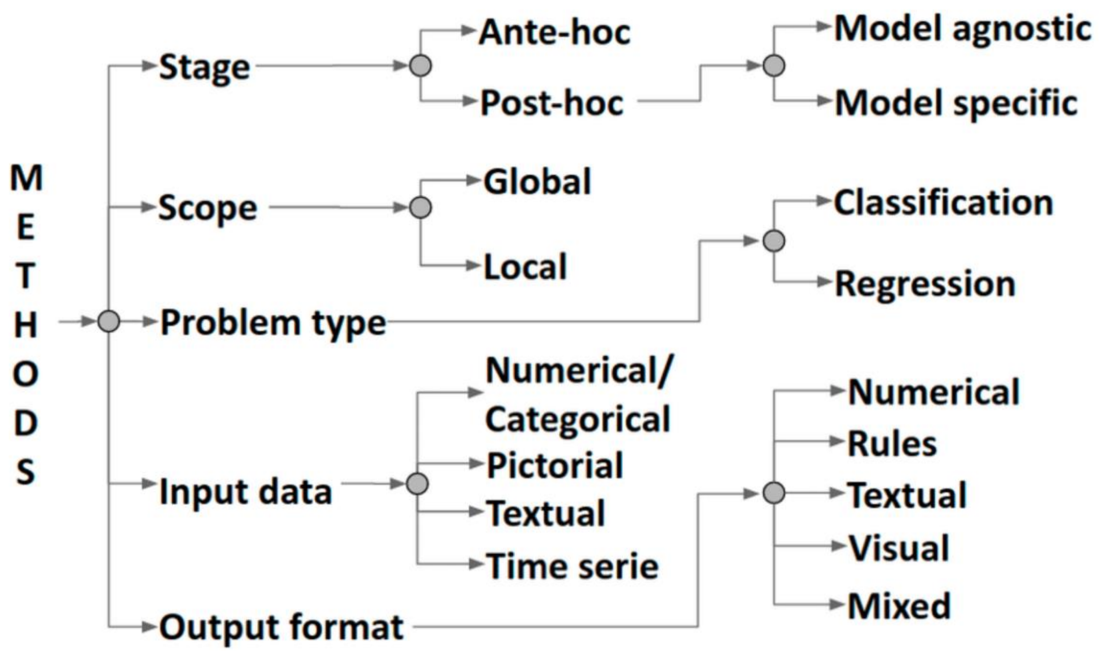


- Global approximation:
  For the whole distribution, a.k.a translation
- Local approximation
  For a subset of input data

❑ Classification of XAI method



Diagrammatic view of how an explainable artificial intelligence (XAI) solution is typically constructed.



https://www.researchgate.net/publication/353687759_Classification_of_Explainable_Artificial_Intelligence_Methods_through_Their_Output_Formats

Example1: LIME for Tabular Data

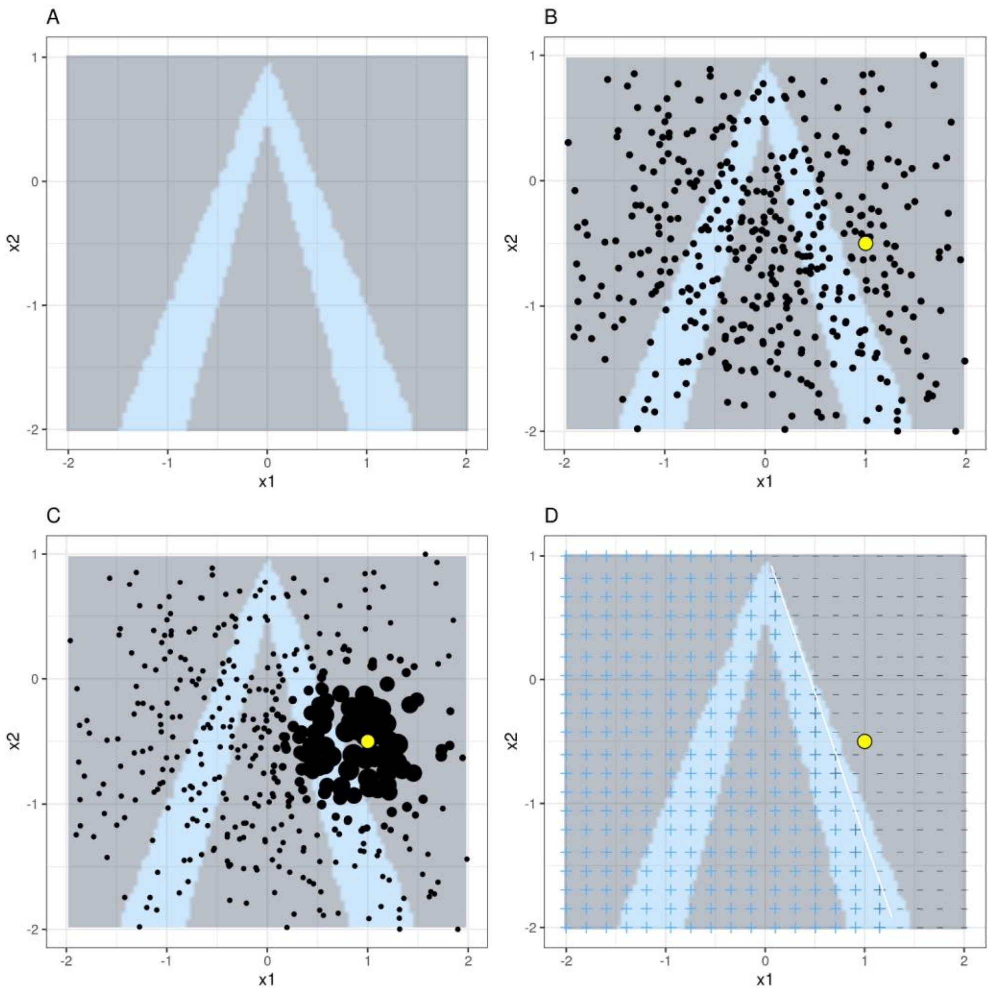


FIGURE 9.5: LIME algorithm for tabular data. A) Random forest predictions given features x1 and x2. Predicted classes: 1 (dark) or 0 (light). B) Instance of interest (big dot) and data sampled from a normal distribution (small dots). C) Assign higher weight to points near the instance of interest. D) Signs of the grid show the classifications of the locally learned model from the weighted samples. The white line marks the decision boundary (P(class=1) = 0.5).

- ❑ 1. Select your instance Of interest for which you want to have an explanation of its black box prediction
- ❑ 2. Weight the new samples according to their proximity to the instance of interest
- ❑ 3. Perturb your dataset and get the black box predictions for these new points.
- ❑ 4. Train a weighted, interpretable model on the dataset with the variations.

LIME Local Interpretable Model-agnostic Explanations

❑ Mathematical formulation

original function

loss function

$$\text{explanation}(x) = \arg\min_{g \in G} [L(f, g, \pi_x) + \Omega(g)]$$

explainable model
(Lasso or Decision Tree)

model complexity

proximity measure
(how large the neighborhood
around instance x?)

Example 2: LIME for Text Data

❑ Classify YouTube comments as spam or normal

	CONTENT	CLASS
267	PSY is a good guy	0
173	For Christmas Song visit my channel! ;)	1

❑ How to perturb

– Randomly remove words and observe the result
– Weight is calculated as 1-(1/# of removed words)

For	Christmas	Song	visit	my	channel!	;)	prob	weight
1	0	1	1	0	0	1	0.17	0.57
0	1	1	1	1	0	1	0.17	0.71
1	0	0	1	1	1	1	0.99	0.71
1	0	1	1	1	1	1	0.99	0.86
0	1	1	1	0	0	1	0.17	0.57

https://christophm.github.io/interpretable-ml-book/lime.html

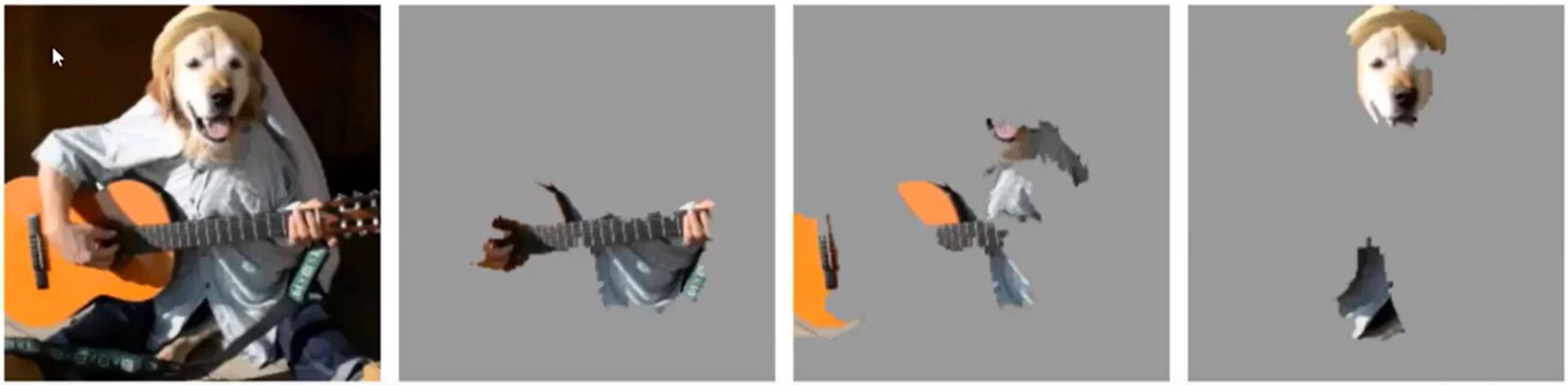
Example 2: LIME for Text Data

❑ Classifiy YouTube comments as spam or normal

case	label_prob	feature	feature_weight
1	0.1701170	PSY	0.000000
1	0.1701170	guy	0.000000
1	0.1701170	good	0.000000
2	0.9939024	channel!	6.180747
2	0.9939024	;)	0.000000
2	0.9939024	visit	0.000000

Example 3: LIME for Images

- Explaining an image classification prediction made by neural Google's Inception neural network



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

- Image regions are selected by the superpixel methods

Pros and Cons for LIME

❑ Pros:

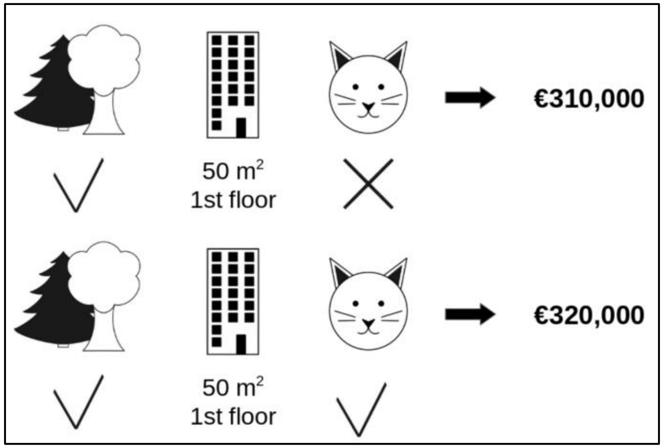
- 1. LIME is model-agnostic
- 2. Explanations are human-friendly
- 3. It works for tabular data, text and images
- 4. The fidelity measure proves the reliability of the interpretable model
- 5. Very easy to use
- 6. Other interpretable features are able to be used instead of original model features.

❑ Cons:

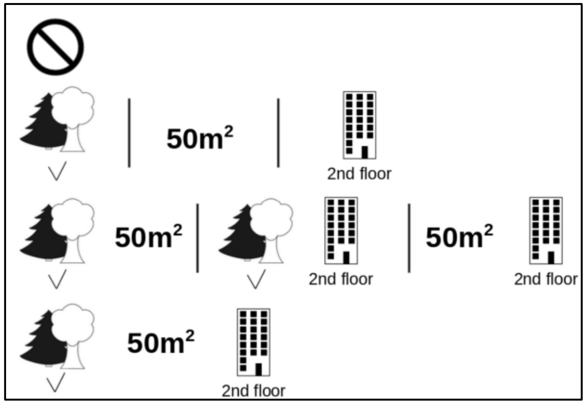
- 1. Finding a good neighborhood is unsolved problem
- 2. Sampling can be wroing(e.g. Gaussian)
- 3. The complexity should be pre-defined
- 4. Explanations can be instable

Shapley Values

- **The Shapley value is the average marginal contribution of a feature value across all possible coalitions.**



No feature values
park-nearby
area-50
floor-2nd
park-nearby+area-50
park-nearby+floor-2nd
area-50+floor-2nd
park-nearby+area-50+floor-2nd.

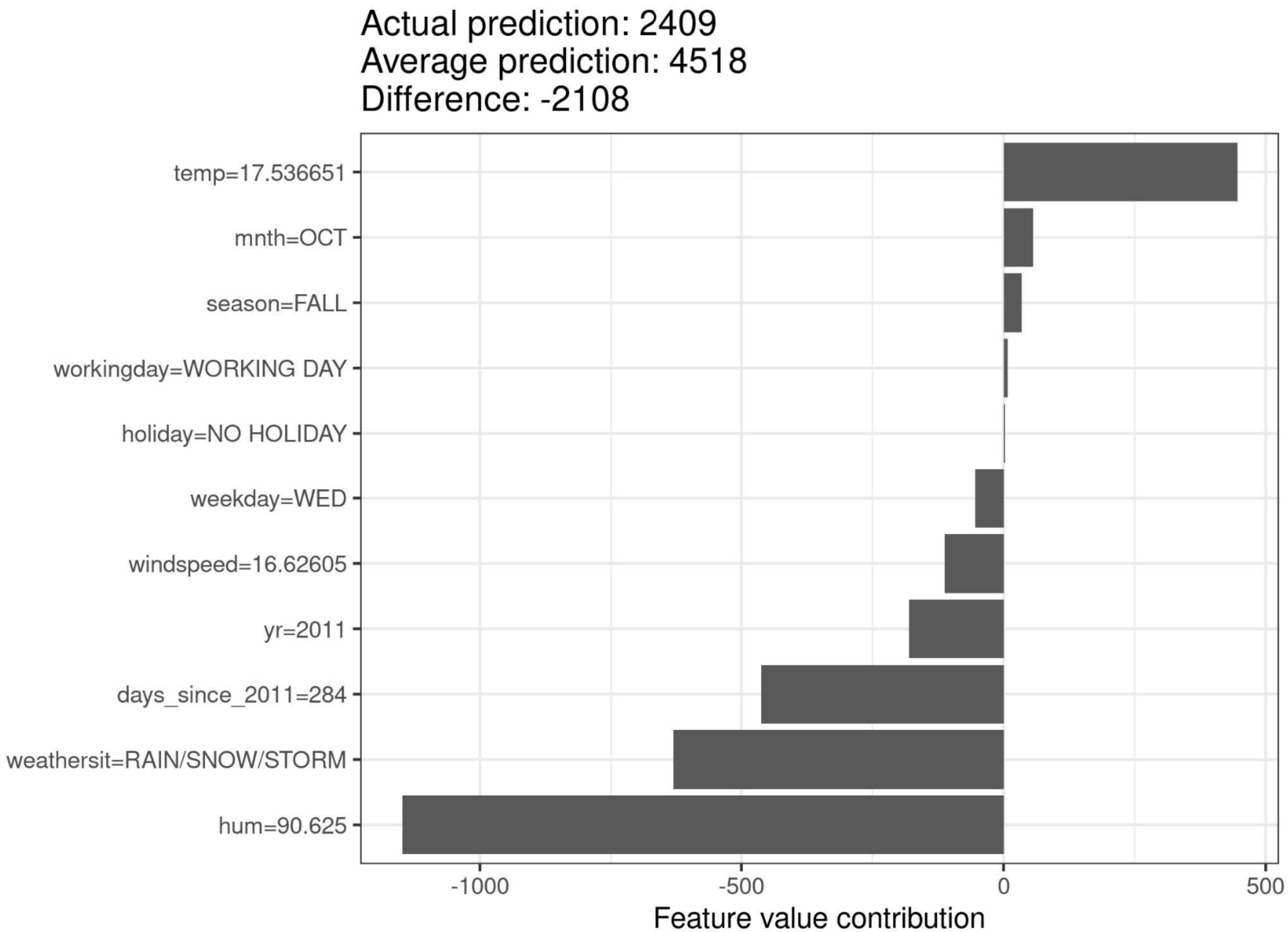


https://christophm.github.io/interpretable-ml-book/shapley.html

한양대학교 HANYANG UNIVERSITY

Shapley Values

❑ Bike Rental Example



The Shapley Value Definition

- The Shapley Value of a feature value is its contribution to the payout, weighted and summed over all possible feature value combinations

$val_x(S)$ is the prediction for feature values in set S that are marginalized over features that are not included in set S:

$$val_x(S) = \int \hat{f}(x_1, \ldots, x_p) d\mathbb{P}_{x \notin S} - E_X(\hat{f}(X))$$

$$\phi_j(val) = \sum_{S \subseteq \{x_1, \ldots, x_p\} \setminus \{x_j\}} \underbrace{\frac{|S|!\,(p - |S| - 1)!}{p!}}_{\text{weight}} \underbrace{(val(S \cup \{x_j\}) - val(S))}_{\text{marginal contribution}}$$

- SHAP : conditional expectation of shapley value

Pros and Cons for Shapley Value

❑ Pros

- 1. The prediction is fairly distributed among the features (no guarantee in LIME)
- 2. Contrastive Explanations are allowed
- 3. The Shapley value is the only explanation method with a solid theory
- 4. It is mind-blowing to explain a prediction as a game

❑ Cons:

- 1. It requires a lot of computing time
- 2. Easy to be misinterpreted (It is NOT a feature value difference after removing the feature)
- 3. Always use all the features, thus not a selective explanation
- 4. Need access to the data
- 5. It suffers from inclusion of unrealistic data instances

SHAP(Shapley Additive exPlanations)

❑ **In SHAP, the Shapley value explanation is represented as an additive feature attribute method, a linear model.**

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

g	Explanation model
$z' \in \{0,1\}^M$	Coalition vector (e.g. images in super-pixel level)
M	Maximum coalition size
ϕ_j	Feature attribution for a feature j, the Shapley values

SHAP feature importance plot

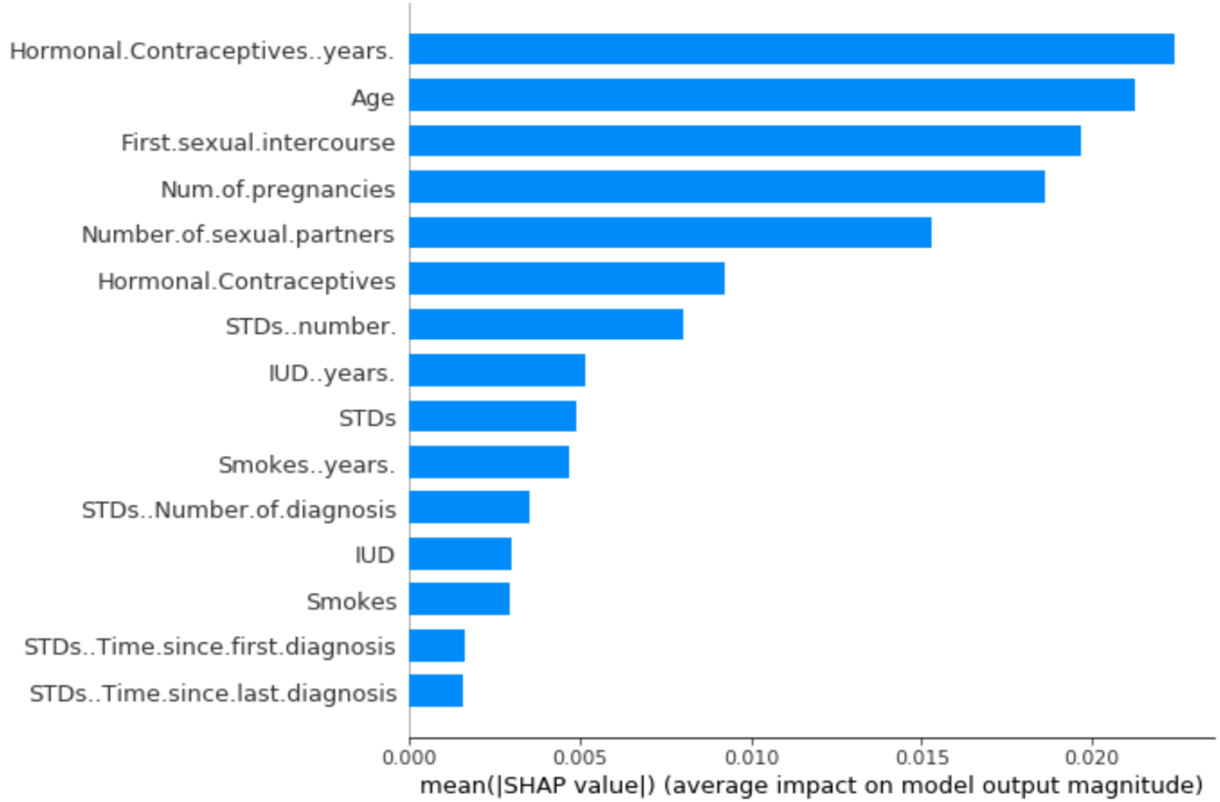


FIGURE 9.25: SHAP feature importance measured as the mean absolute Shapley values. The number of years with hormonal contraceptives was the most important feature, changing the predicted absolute cancer probability on average by 2.4 percentage points (0.024 on x-axis).

SHAP Summary Plot

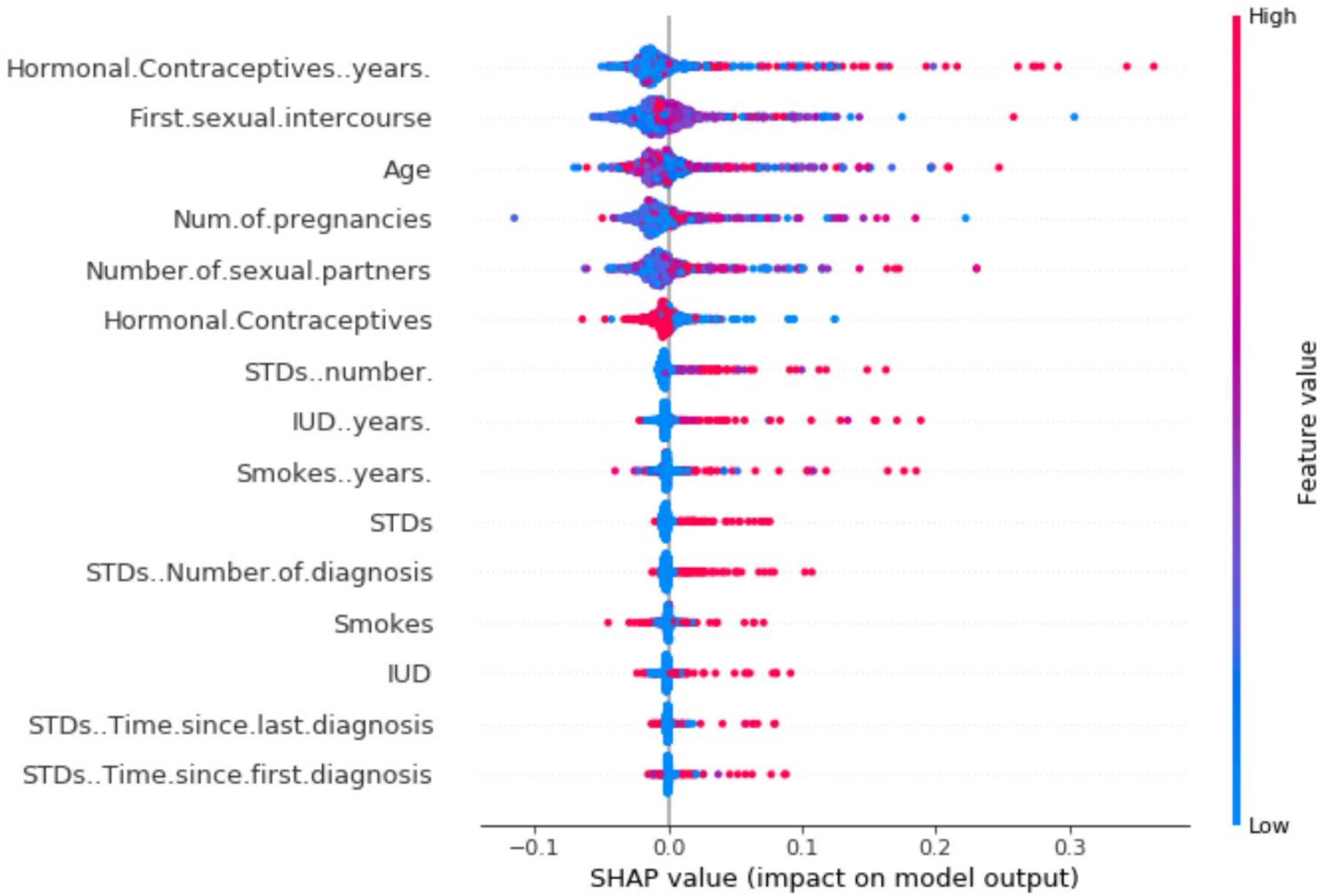


FIGURE 9.26: SHAP summary plot. Low number of years on hormonal contraceptives reduce the predicted cancer risk, a large number of years increases the risk. Your regular reminder: All effects describe the behavior of the model and are not necessarily causal in the real world.