

Prof. Anne Schwerk

EXPLAINABLE AI:

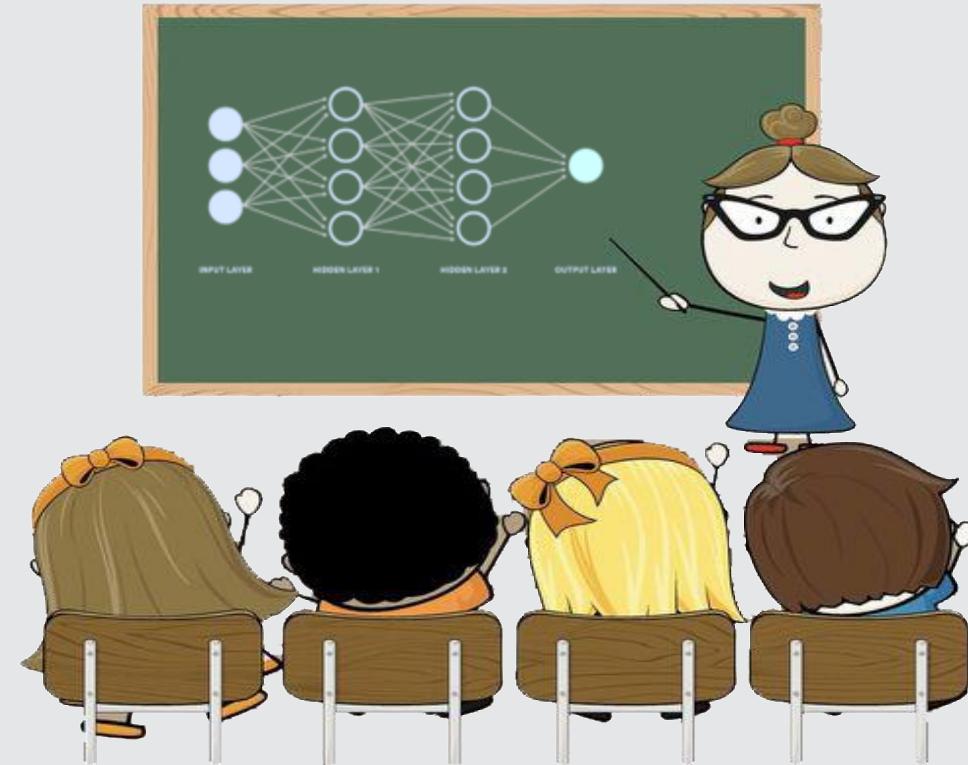
**FAIRNESS,
ROBUSTNESS,
AND SUSTAINABILITY**

 ecoCompute



#1

WHY DO WE NEED EXPLAINABLE AI (XAI)

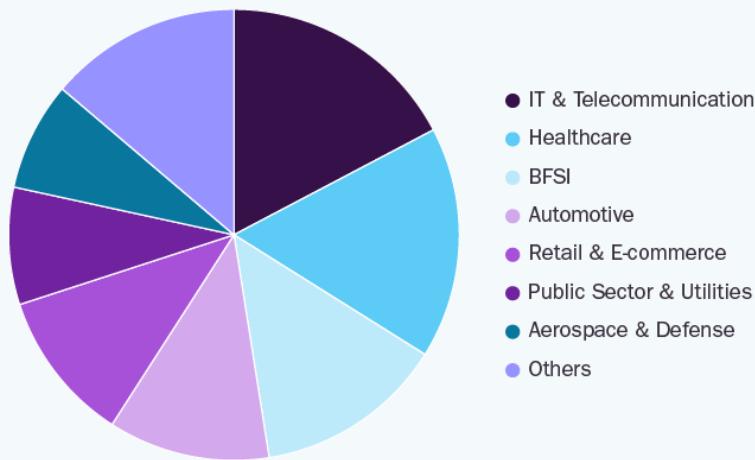


EXPLAINABLE AI MARKET SIZE

Growing Market Size And Healthcare Market Share

Explainable AI Market

Share, by End Use, 2024 (%)

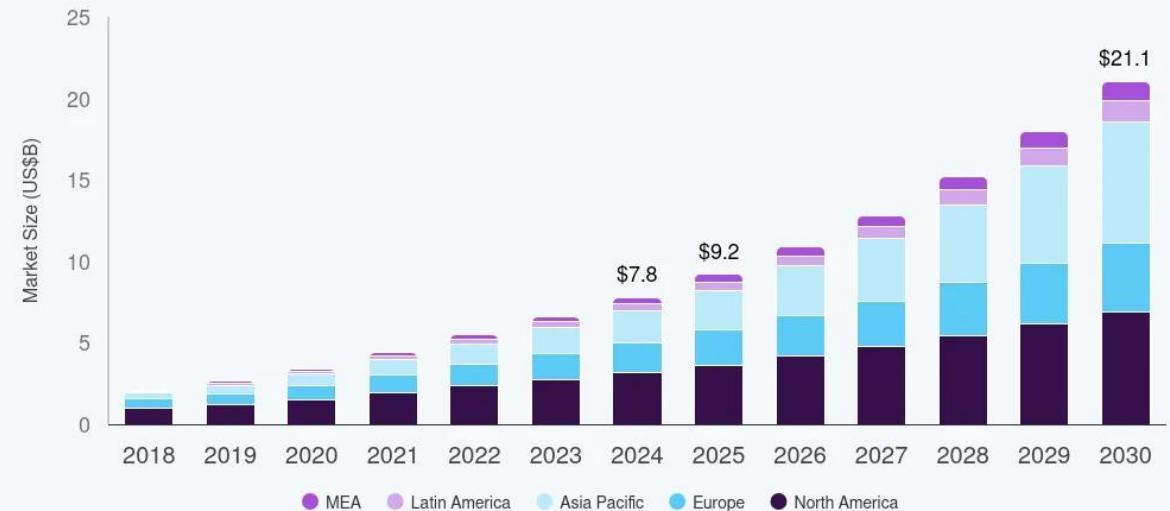


\$7.8B
Global Market Size,
2024

Source:
www.grandviewresearch.com

Explainable AI Market

Size, by Region, 2018 - 2030



GOOGLE GENDER BIAS

Google search results for "IT AND professor". The search bar shows the query. Below it, a row of filters includes All, Images, News, Videos, Books, Products, Finance, More, Tools, and Saved. A series of cards below the filters show various search results, including "Email", "Gilbert strang", "Read literature", "Digital health", "Teaching", "Gpt", "Cyber security", "Student", "College professors", "Assistant professor", and "Hdm stuttg...".

RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Propagation of societal gender inequality by internet search algorithms

Edited by Susan Fiske, Princeton University, Princeton, NJ; received March 14, 2022; accepted May 27, 2022

July 12, 2022 | 119 (29) e2204529119 | <https://doi.org/10.1073/pnas.2204529119>

Madalina Vlasceanu and David M. Amodio [Authors Info & Affiliations](#)

Fachhochschule Dortmund
Fachhochschule Dortmund

Boise State University
Start Talking to Your Professor ...

ZipRecruiter
Professor: What Is It? and How to...

College Data
Relationships with Your Professor...

Nachrichten aus der Wissenschaft » ...
Engineering Secure Devices: Prof...

carrer start bw
The habilitation proc...

smartsciencecareer.com
Is being a professor ...

Kritik
Dr. Scott Johnson ...

Reddit
Prof. Gilbert Stran...

College Data
Relationships with Your Professor...

CHATGPT GENDER BIAS

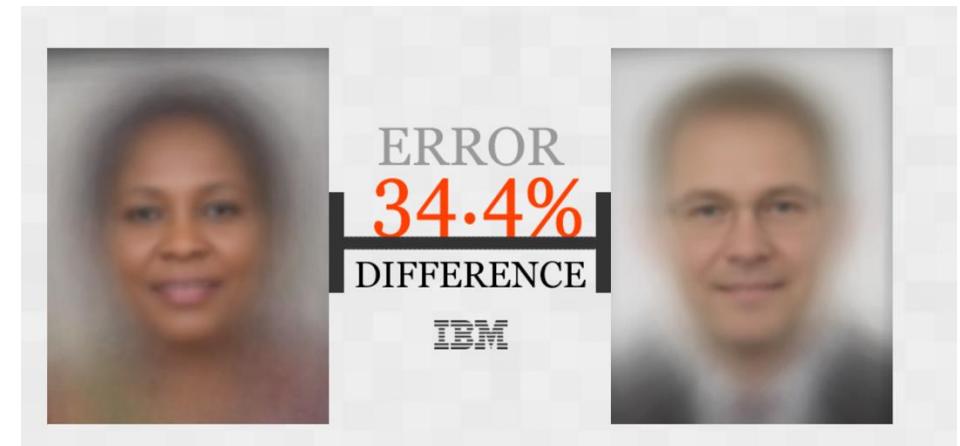
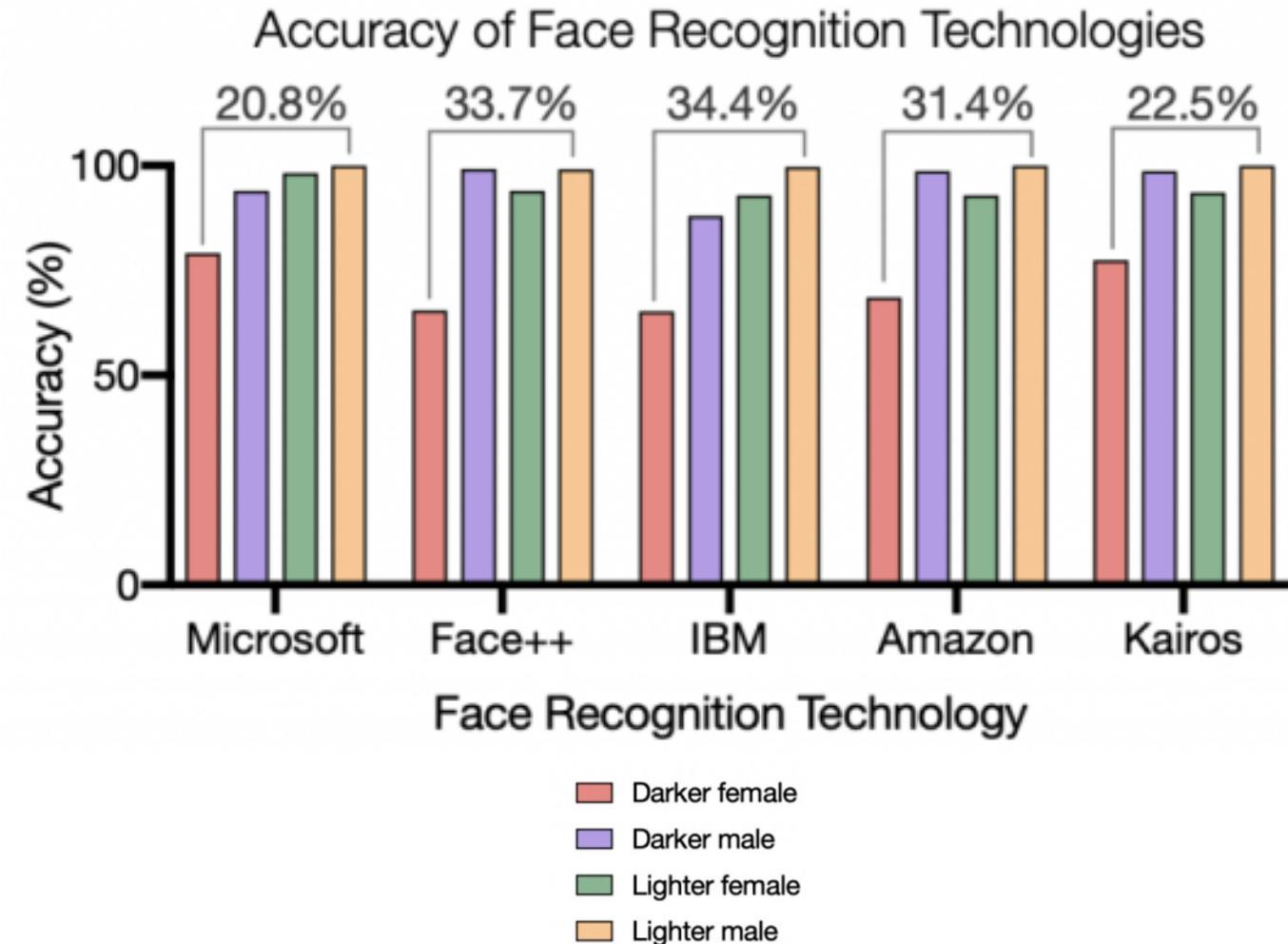
Design 4 images depicting a CEO of an IT company



Design 4 images depicting a professor



BIAS: THE GENDER SHADES PROJECT AUDITS FIVE FACE RECOGNITION TECHNOLOGIES



- **Darker faces:** 93.6% misgendered by Microsoft
- **Female faces:** 95.9% misgendered by Face++

AI-BASED UNDERDIAGNOSIS OF CHEST X-RAY IMAGES

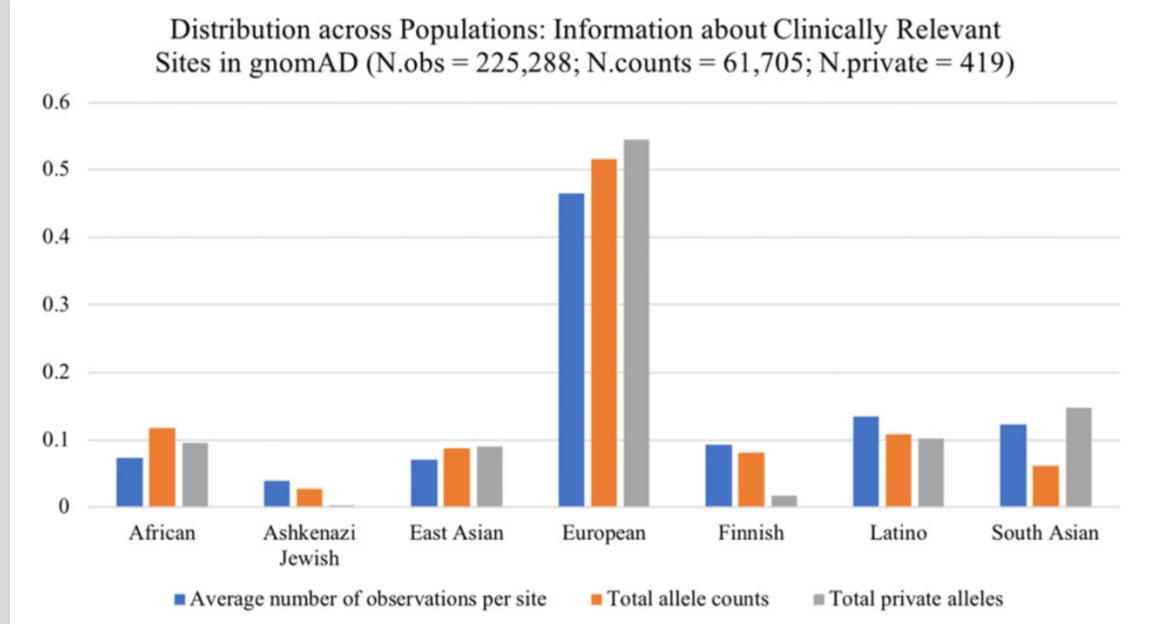
- False prediction of health status in underserved populations:
 - Females
 - Patients < 20 years
 - Afroamericans & Hispanics
 - Medicaid recipients
- State-of-the-art computer vision techniques (121-layer DenseNet)
- Three large publicly-available radiology datasets (MIMIC-CXR, CheXpert, ChestX-ray)
- Lack of real-world testing & evaluation on different demographic groups is common practice (e.g. Epic Sepsis Model)



ETHNIC DISCRIMINATION IN GENOMICS-BASED DIAGNOSIS FOR CARDIOMYOPATHY

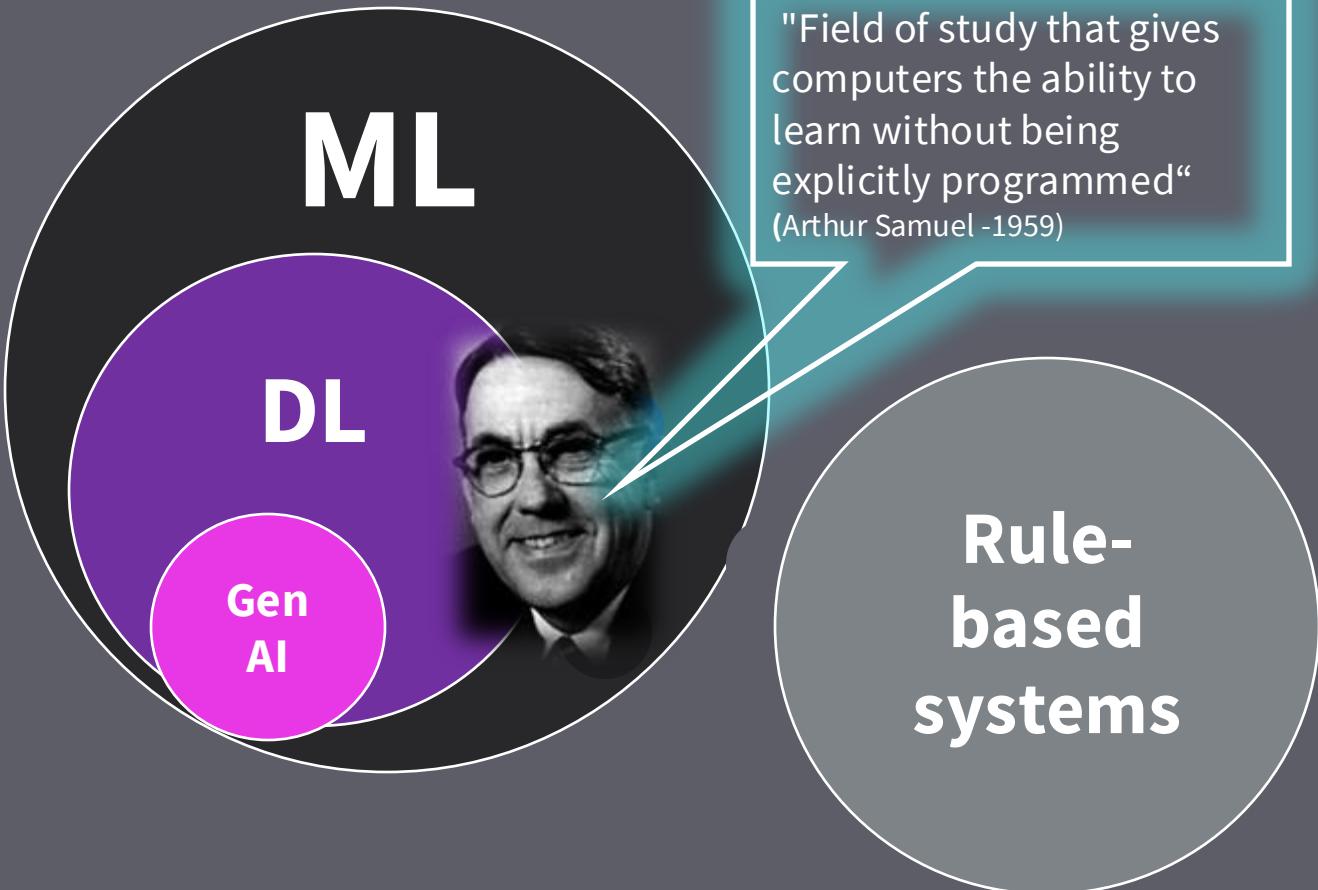
Genetic Misdiagnoses and the Potential for Health Disparities

Arjun K. Manrai, Ph.D., Birgit H. Funke, Ph.D., Heidi L. Rehm, Ph.D., Morten S. Olesen, Ph.D., Bradley A. Maron, M.D., Peter Szolovits, Ph.D., David M. Margulies, M.D., Joseph Loscalzo, M.D., Ph.D., and Isaac S. Kohane, M.D., Ph.D.

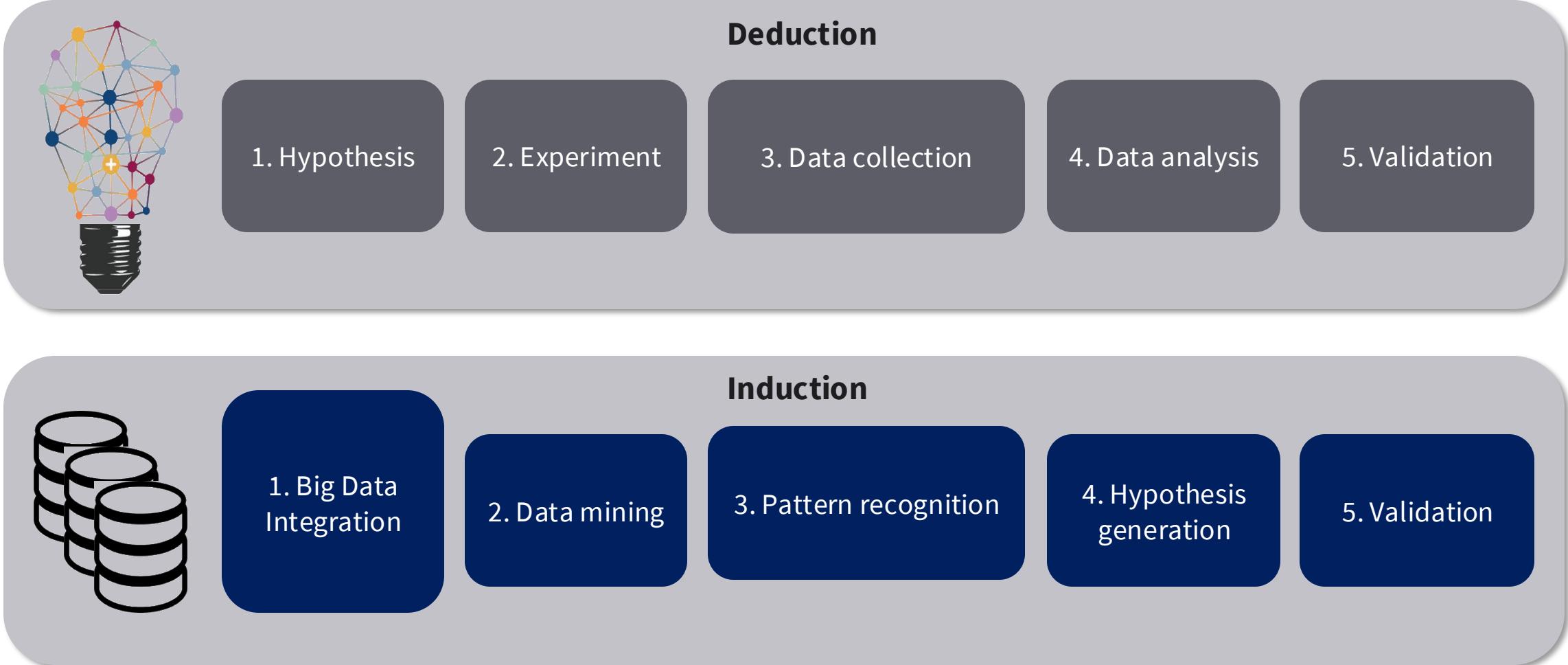


AI AND ML – DATA DRIVEN ENABLERS

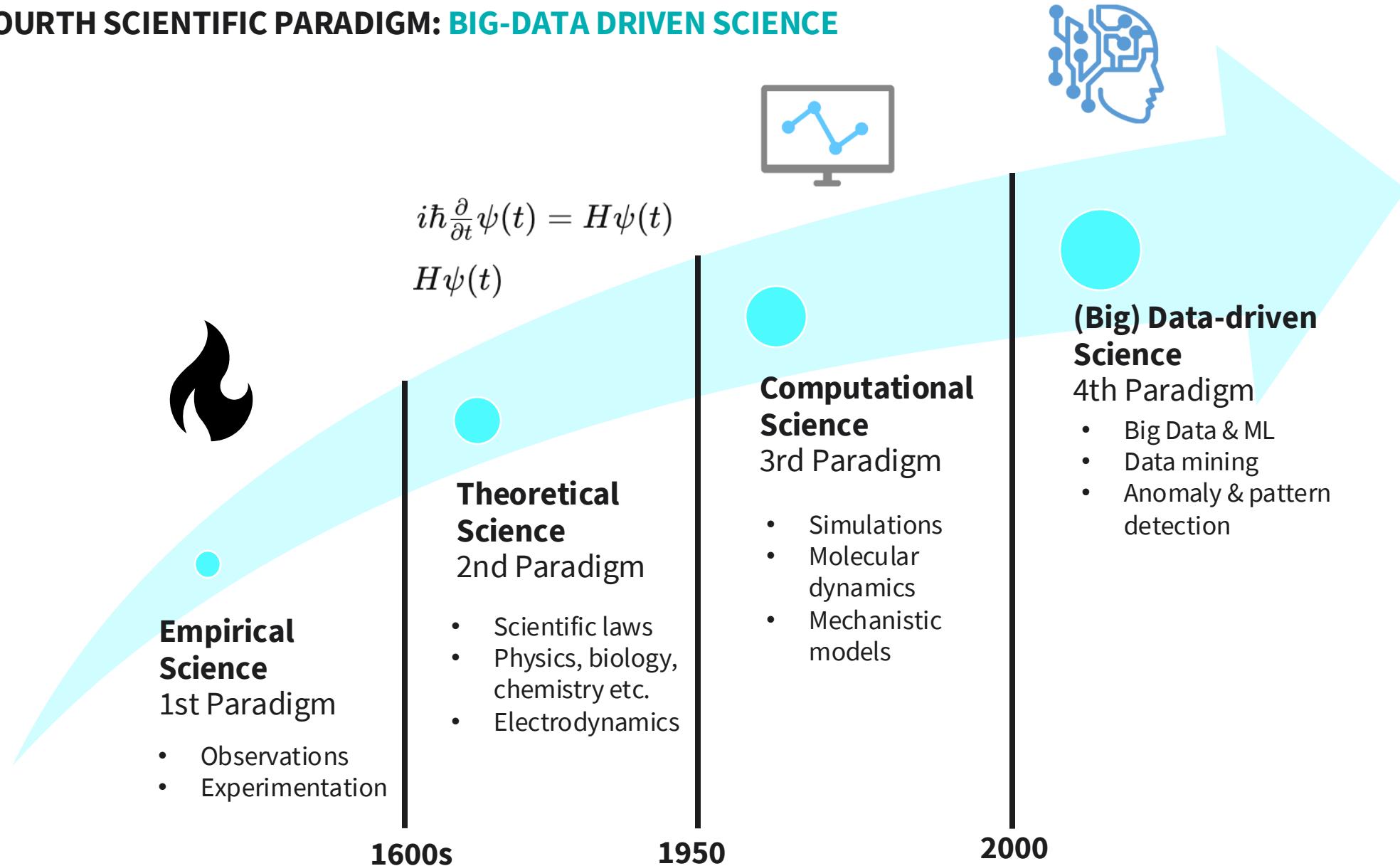
Artificial Intelligence: IT systems with human-like behavior – based on statistical & mathematical models



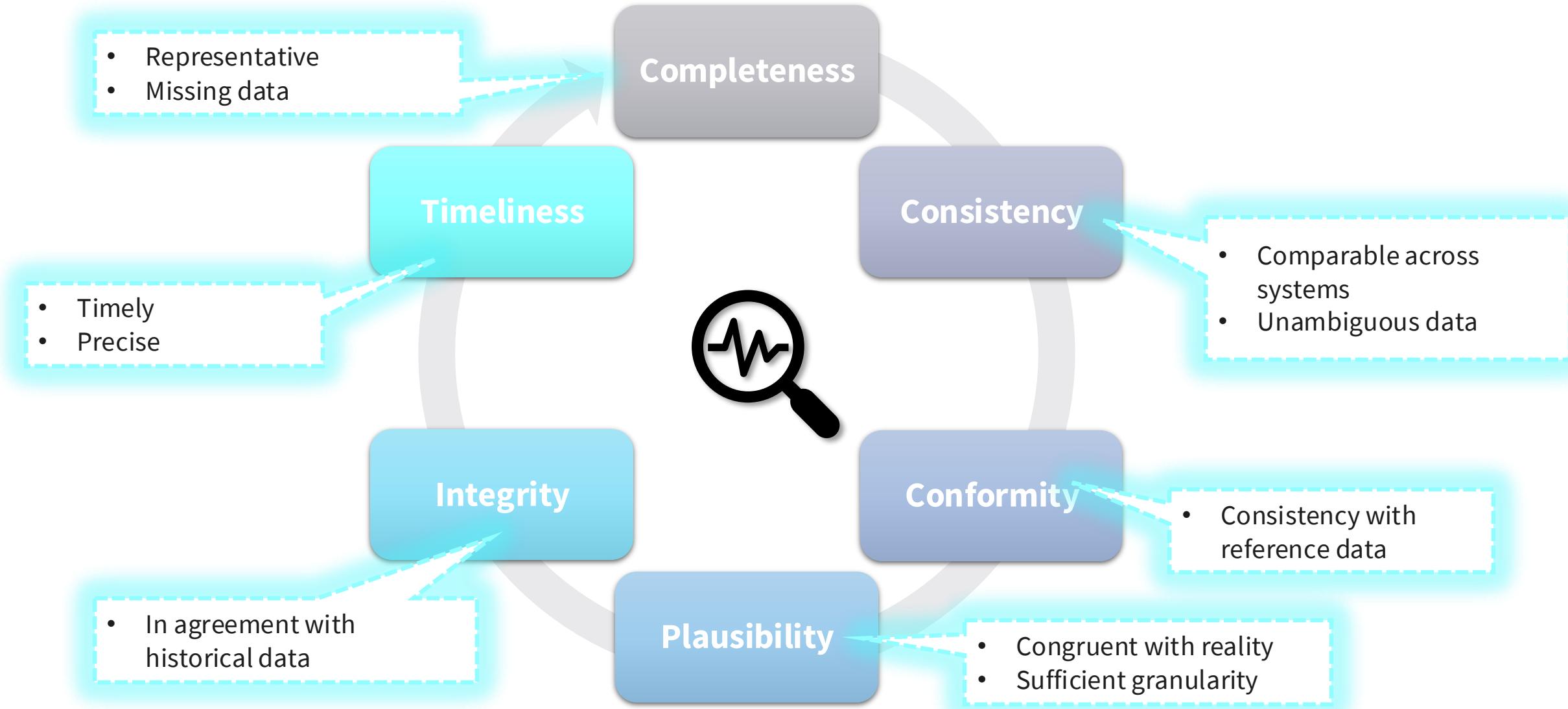
THE END OF THEORY – FROM DEDUCTION TO INDUCTION



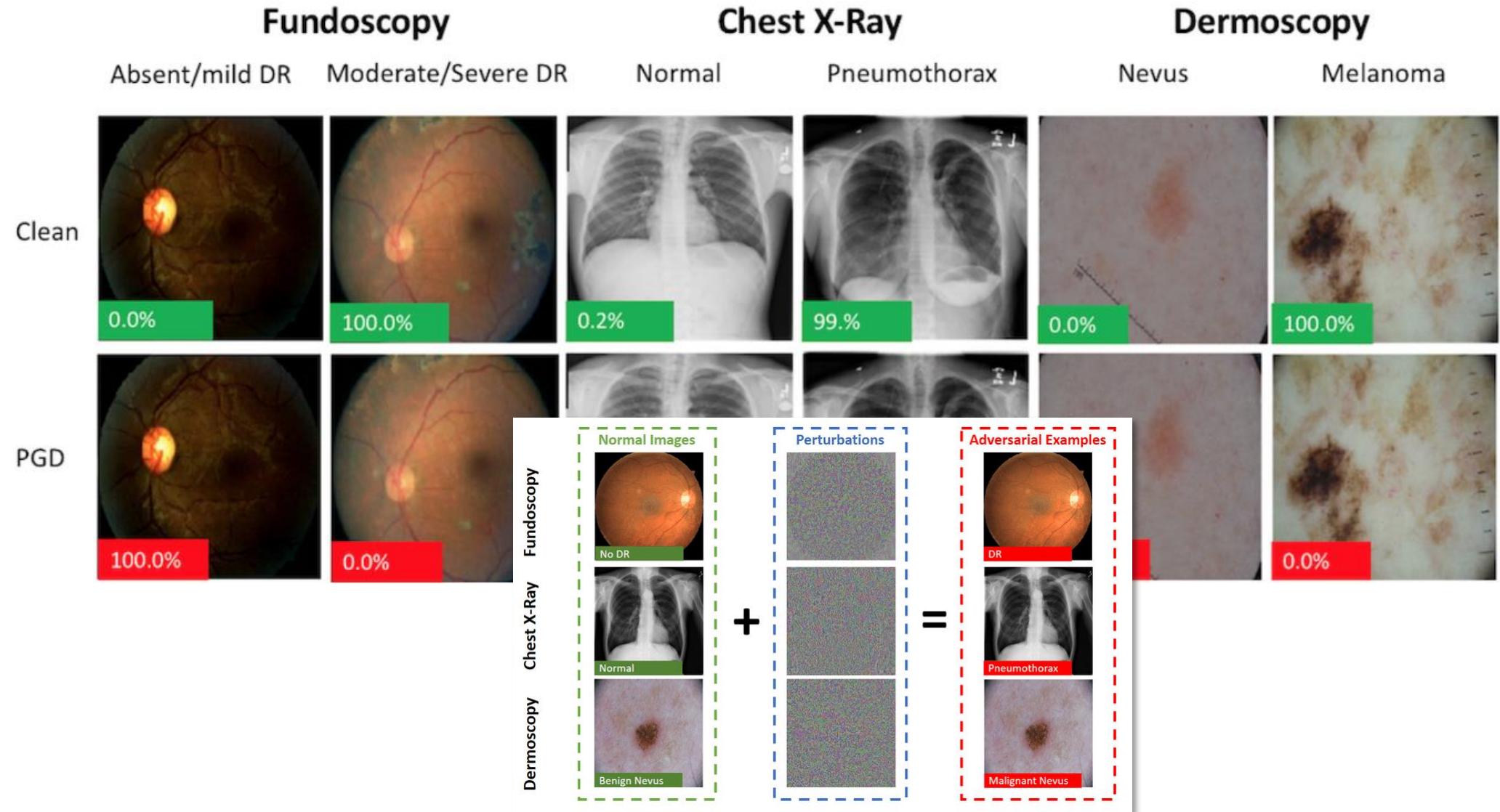
THE FOURTH SCIENTIFIC PARADIGM: BIG-DATA DRIVEN SCIENCE



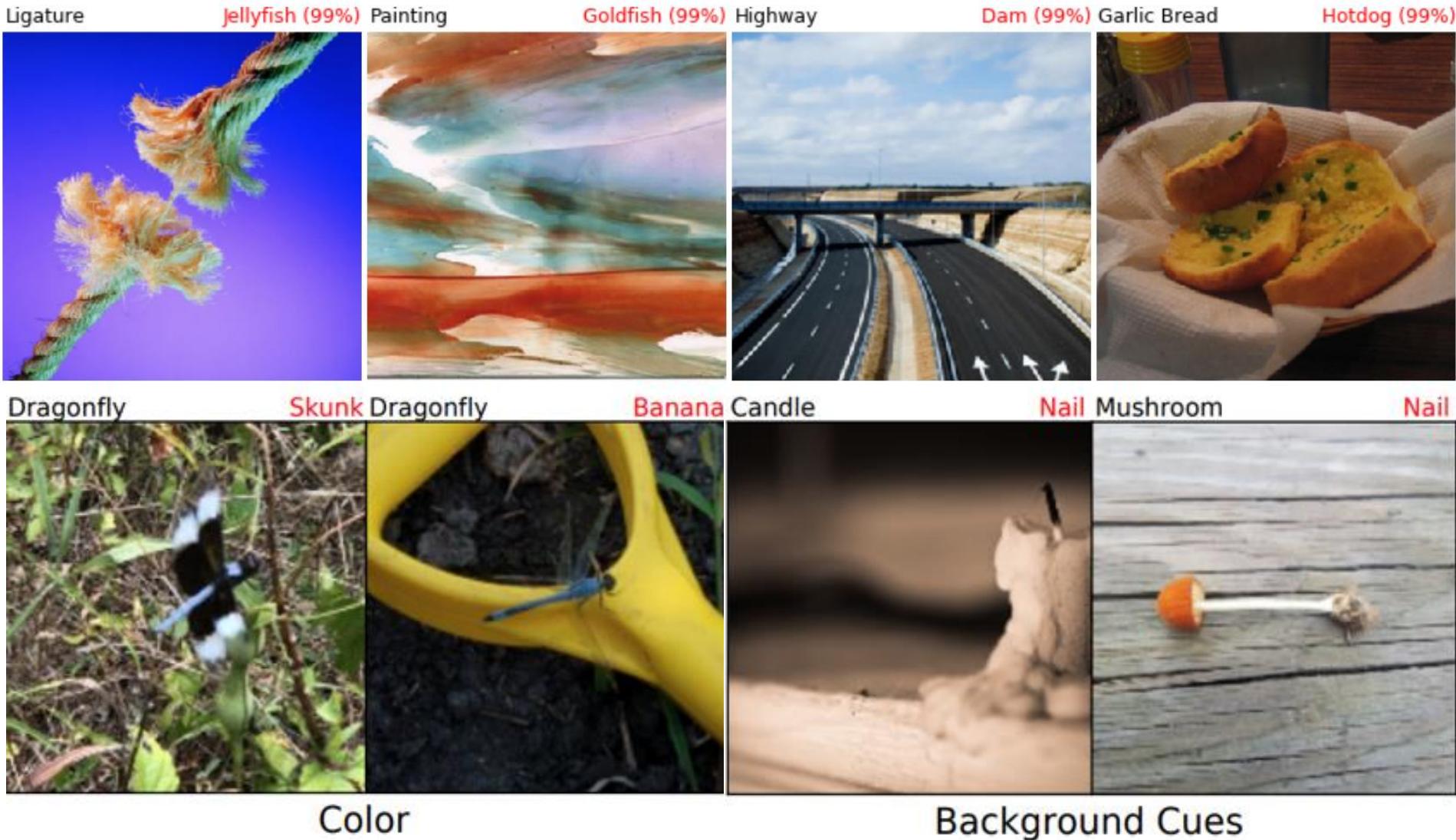
DATA QUALITY



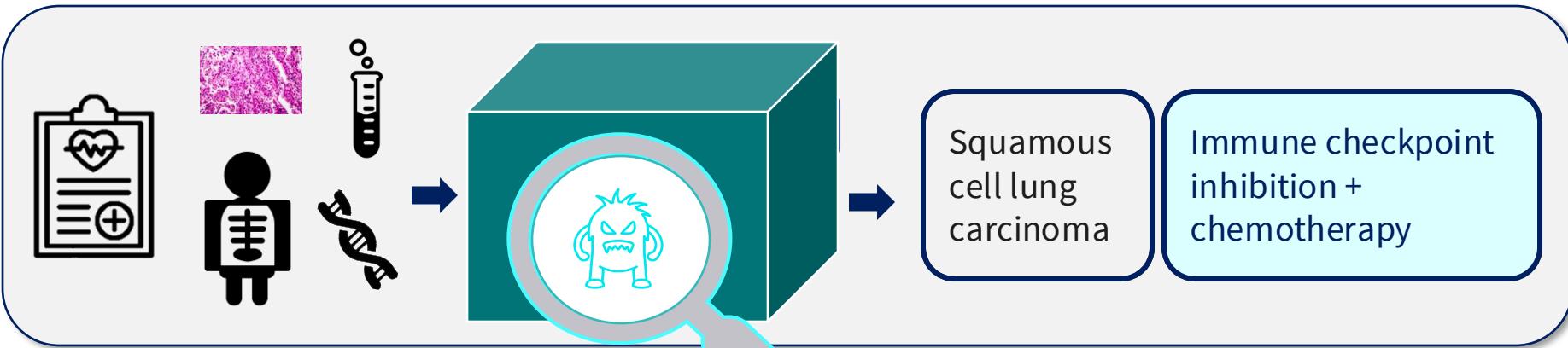
ADVERSERIAL ATTACKS



THE NEED FOR XAI: BIASED MODELS



THE BLACKBOX PROBLEM: MEDICAL DIAGNOSIS



- Responsibility?
- Liability?
- Control?
- Role of GP?



- Why?!
- How accurate?
- Which confidence interval?
- Under which conditions can I trust?

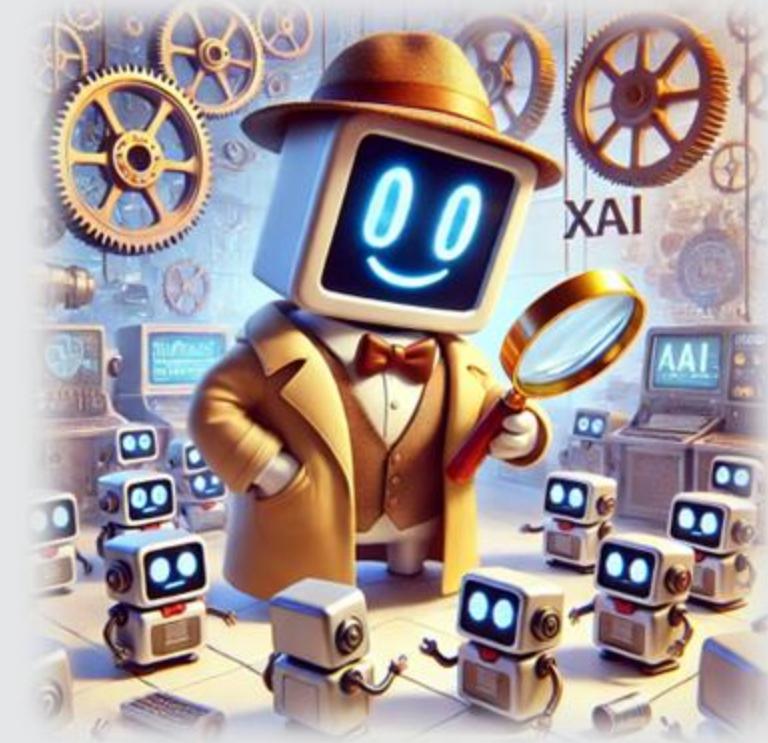


FOUR PRIMARY ASPECTS

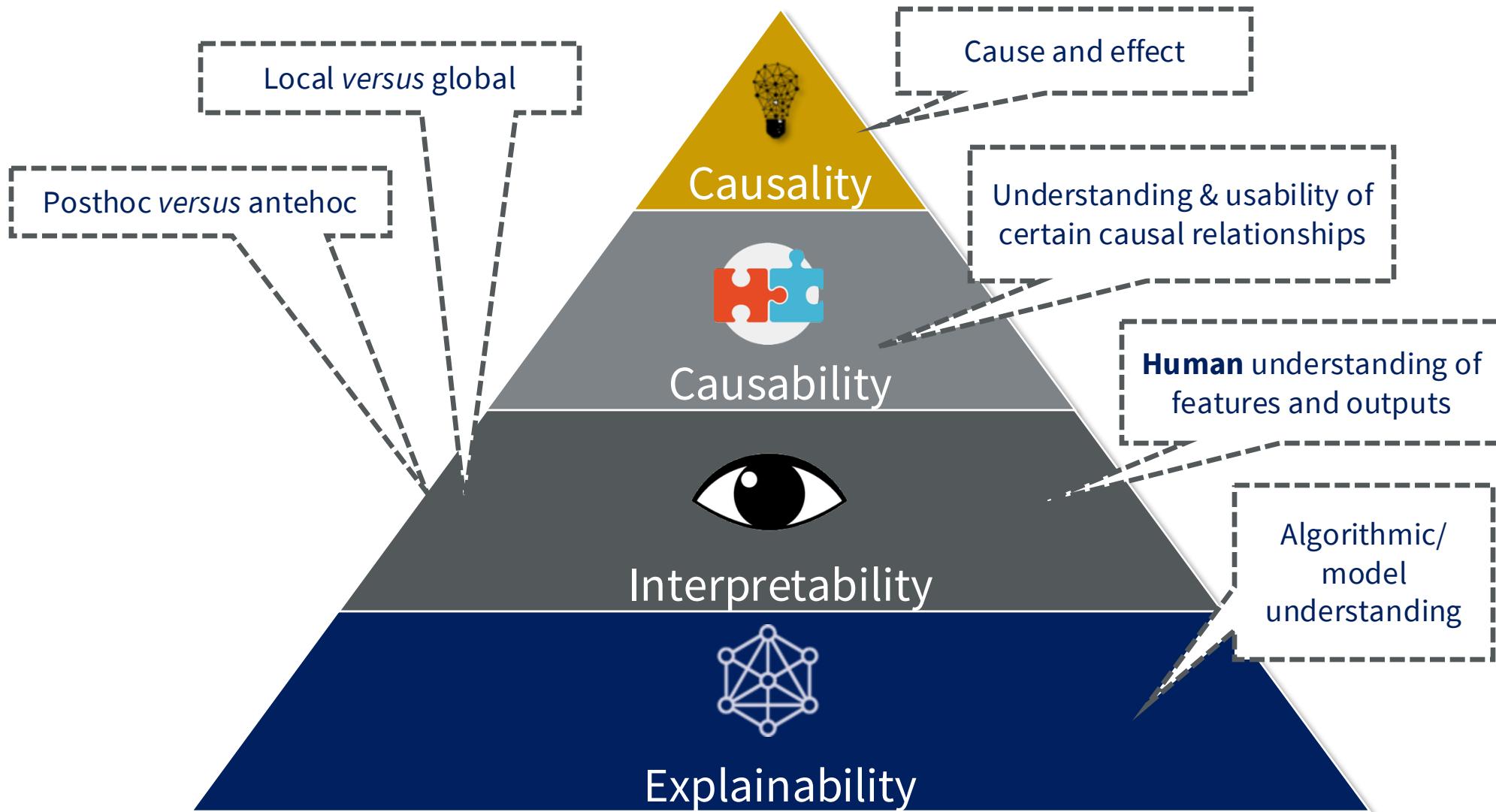


#2

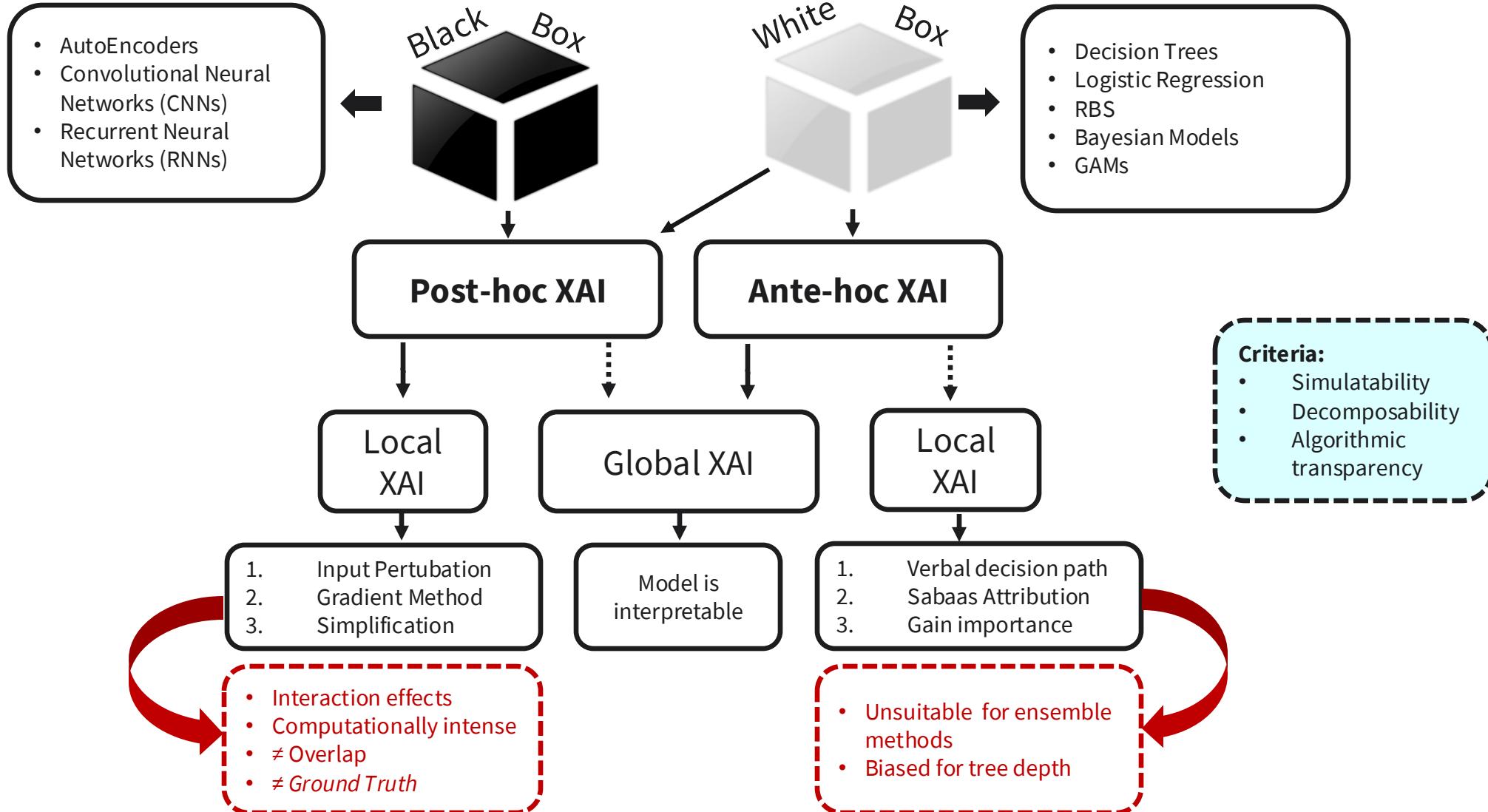
XAI CONCEPTS



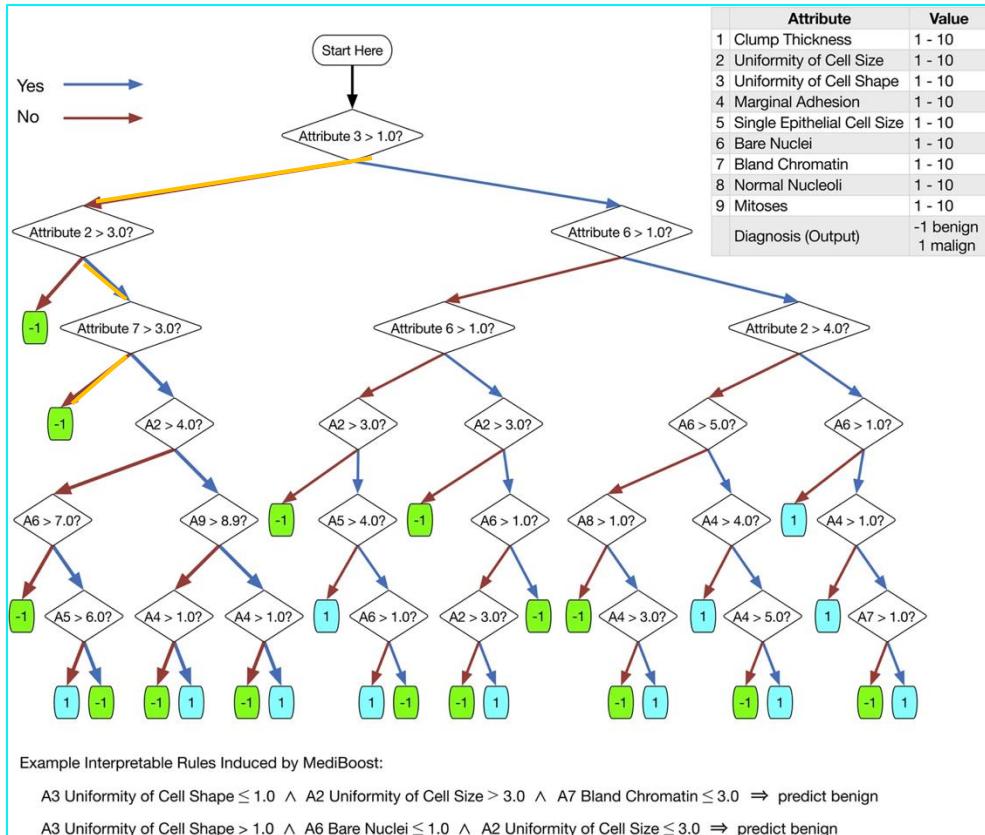
XAI: DEFINITIONS



POSTHOC VERSUS ANTEHOC XAI

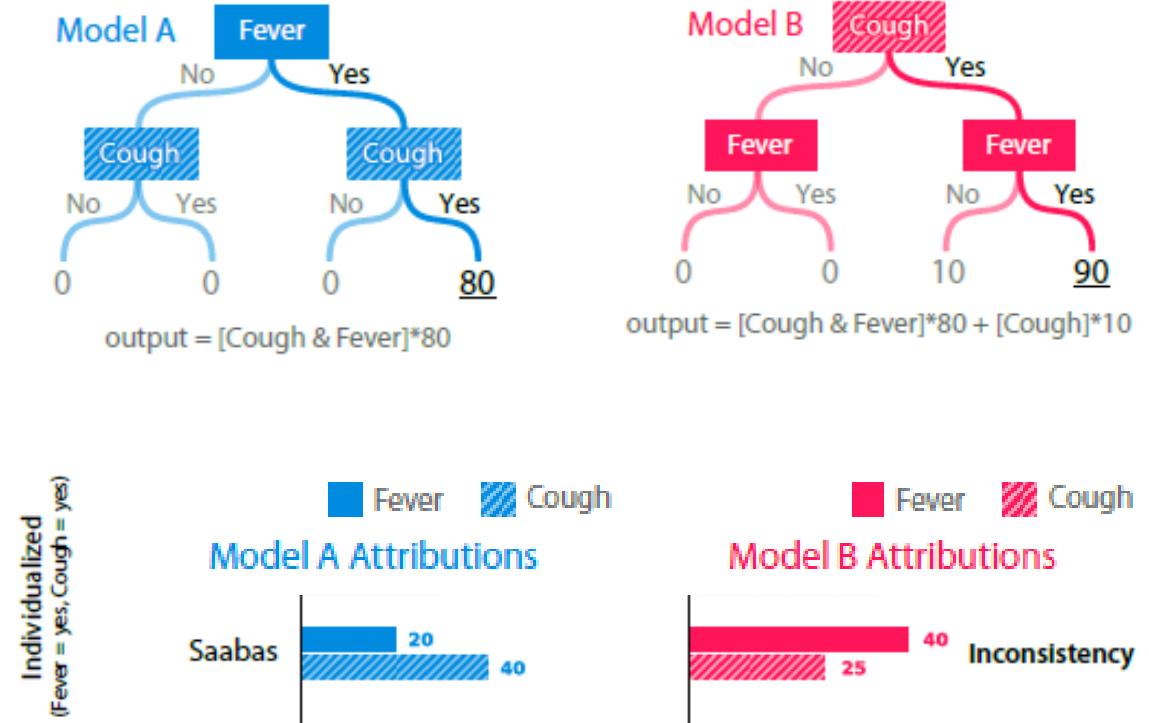


ANTEHOC XAI: DECISION TREES



Decision Path:

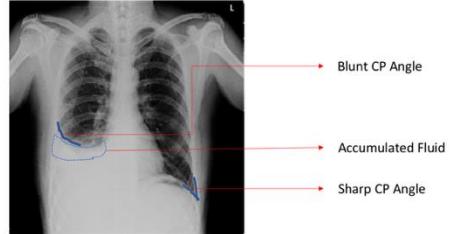
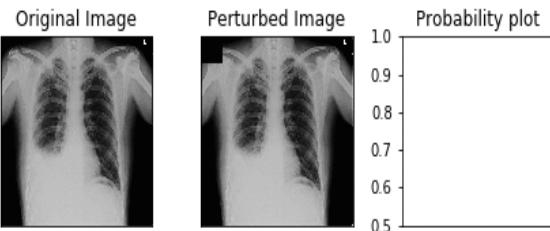
Tumor shows no uniform cell shapes → but >3 cells with uniform cell sizes → no bland chromatin → benign breast cancer



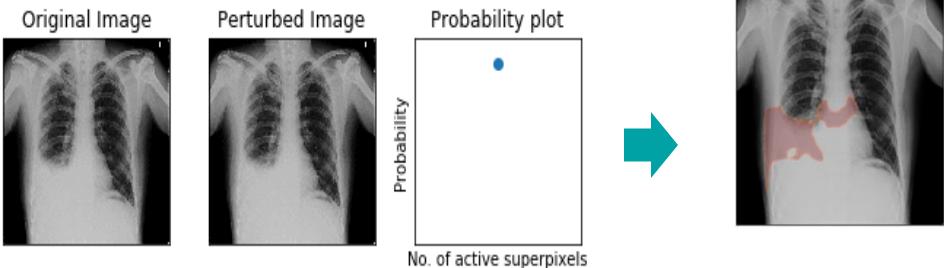
- Only 2 local XAI:
 - Sabaas:** Inconsistency of feature attribution methods
 - Decision path:** Inadequate for multiple trees

POSTHOC XAI : OCCLUSION AND LIME FOR CHEST X-RAY OF PLEURA EFFLUX

Occlusion



Lime



- Computationally intense
- Bias of occlusion/superpixel size
- Interaction effects

- Image is divided into superpixels
- Randomly activate superpixel (n times)
- Predict outcome of perturbed data
- Regression model of perturbed data
- Heatmap regression weights (weights= proximity of predictions to output)

POSTHOC XAI: SHAPLEY VALUES

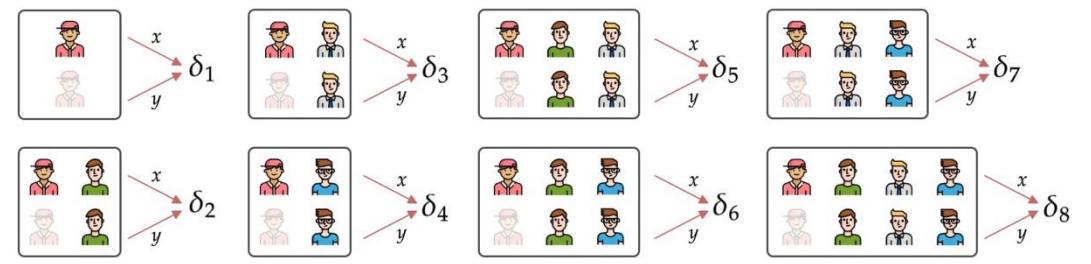
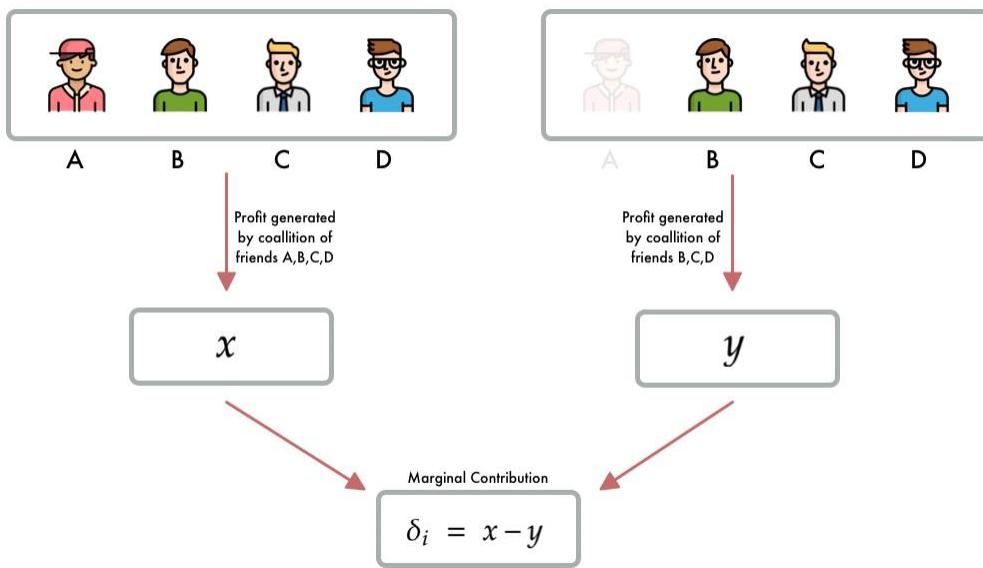
Benefits

- Only XAI that allows a fair effect distribution (interaction effects)
- Allows prediction comparisons to subset (not only average)
- Rooted in Game Theory: average expected marginal contribution to model decision after accounting for all possible combinations
- Model agnostic



Drawbacks

- Independency assumption
- No prediction model
- Data completeness: not suitable for sparse distributions
- NP hard problem (can only be approximated)



The Shapley value for member

is given by:

$$\phi_i = \frac{\delta_1 + \delta_3 + \delta_4 + \delta_5 + \delta_6 + \delta_7 + \delta_8}{8}$$

#3

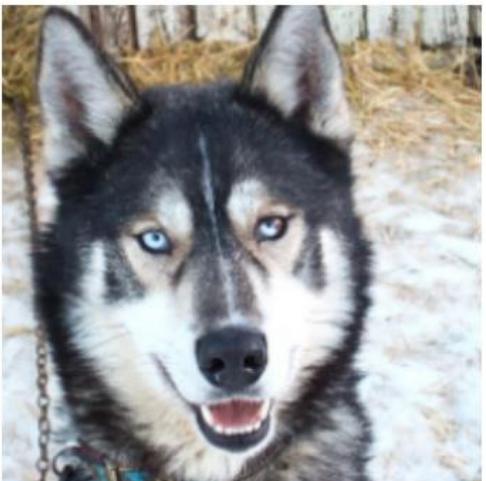
XAI EXAMPLES



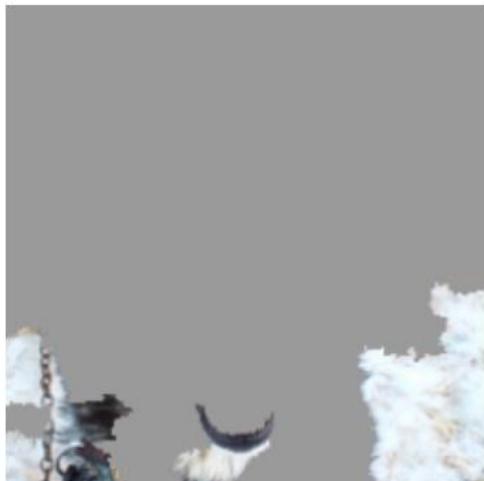
XAI EXAMPLES: MODEL UNDERSTANDING & IMPROVEMENT

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: “Husky vs Wolf” experiment results.



(a) Husky classified as wolf



(b) Explanation

Figure 11: Raw data and explanation of a bad model’s prediction in the “Husky vs Wolf” task.

- Correct prediction but wrong features
- Shows selection of non-representative data for model training
- Non-expert proof (graduate students)
- Allows for specific feature engineering and model improvement
- Leads to generalizability & robustness

XAI EXAMPLES: : MODEL UNDERSTANDING & IMPROVEMENT

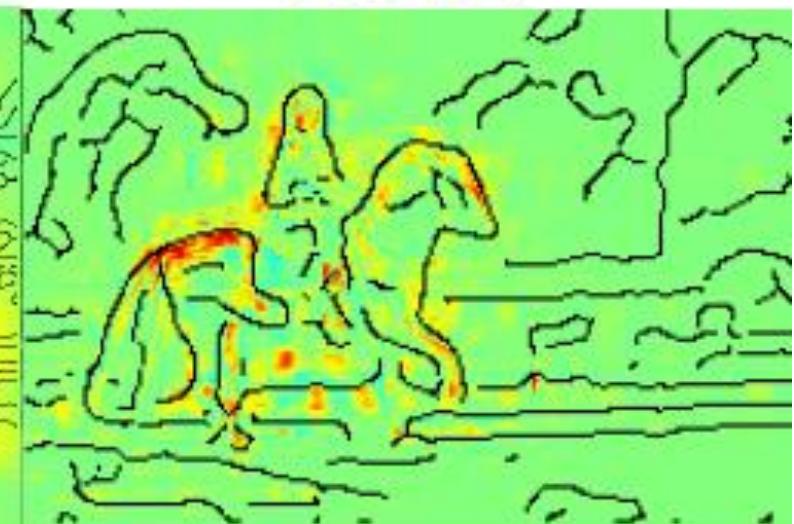
Image



FV



DNN

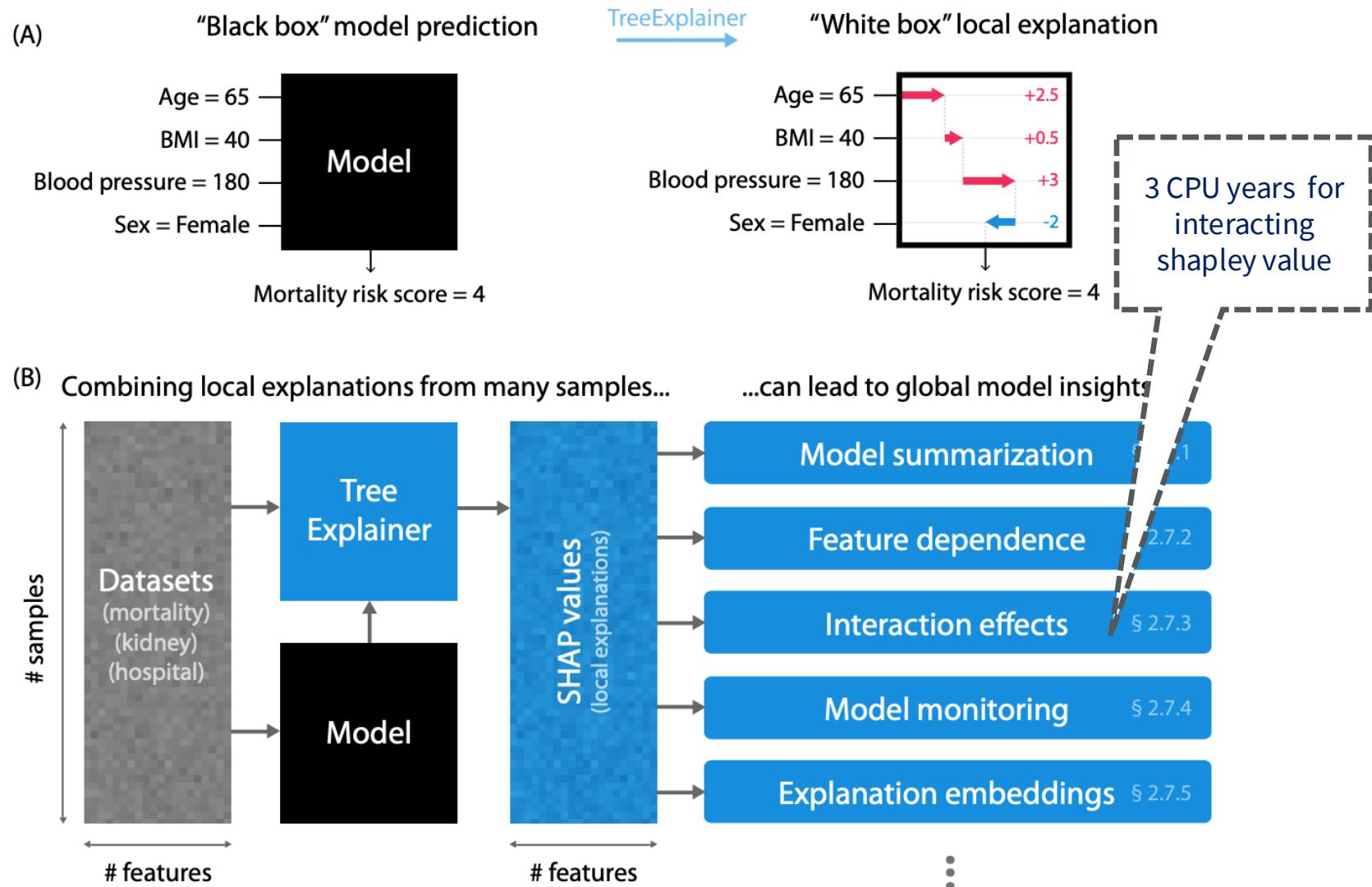


XAI EXAMPLES: KNOWLEDGE GENERATION

Improve interpretability of tree-based models:

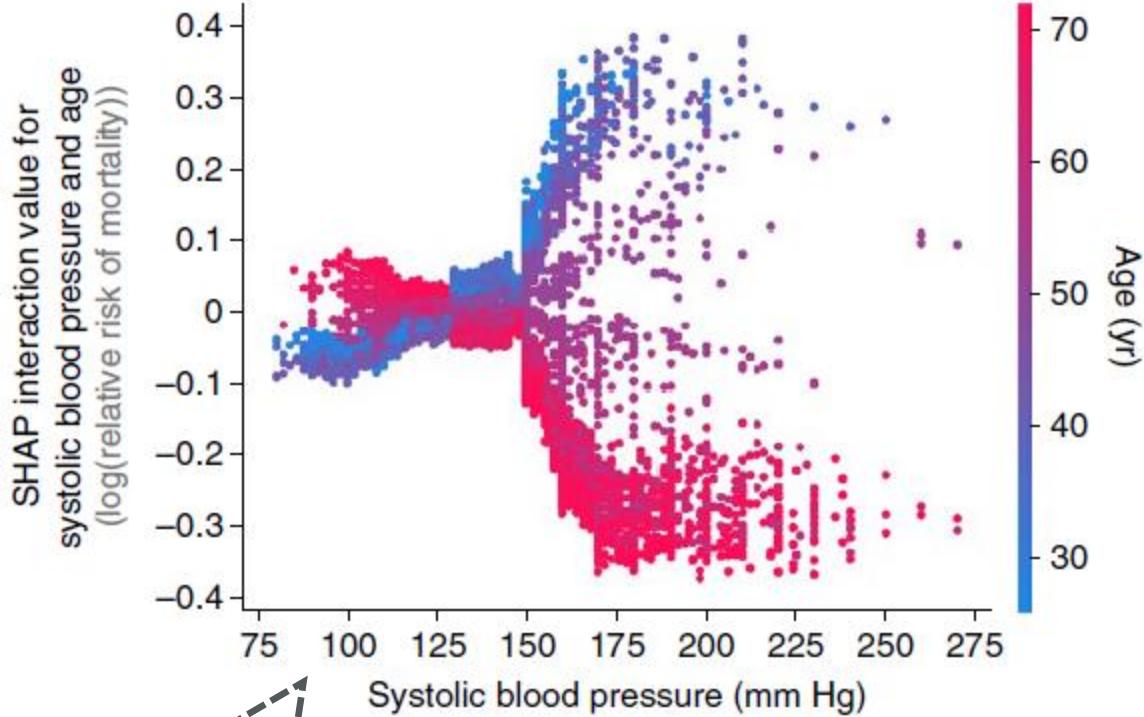
1. Polynomial time algorithm to compute SHAP values
2. XAI for local feature interactions
3. Global XAI through combining many local explanations

- 3 ML models
- 3 medical data sets
- Human consensus
- 15 metrics to evaluate performance



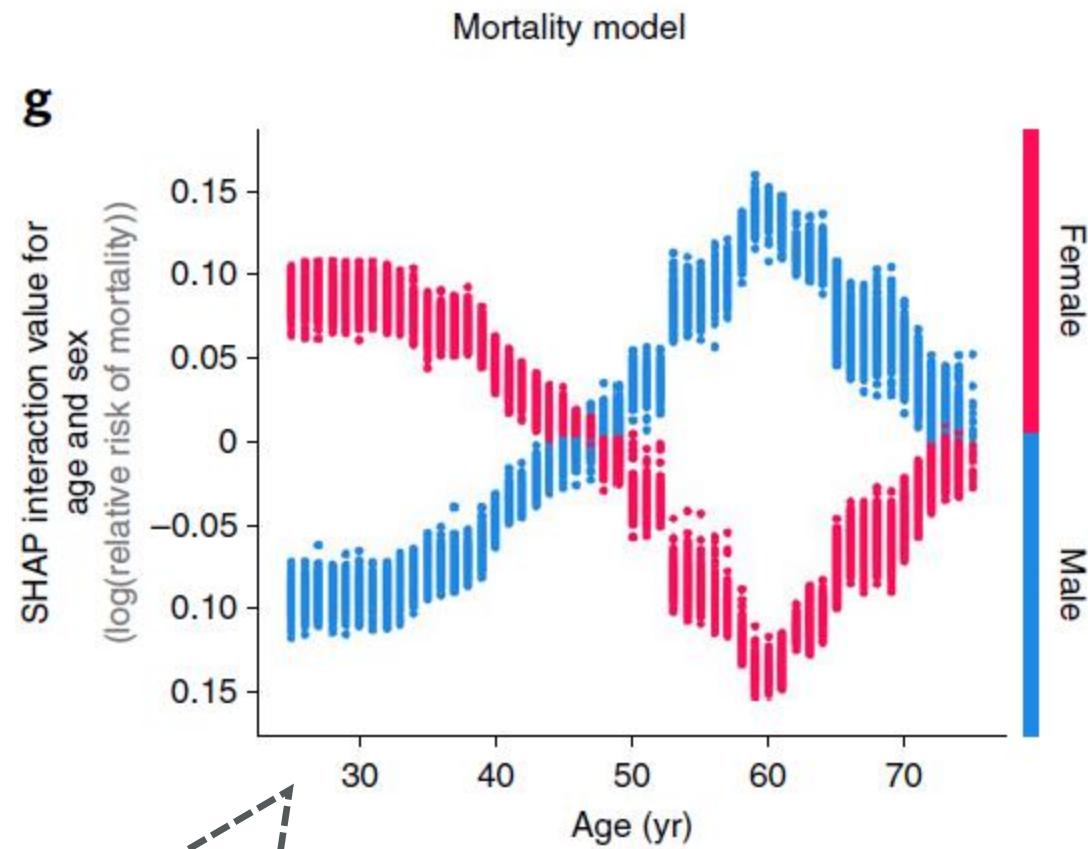
XAI EXAMPLES: SHAP INTERACTION VALUES FOR NOVEL INSIGHTS

d



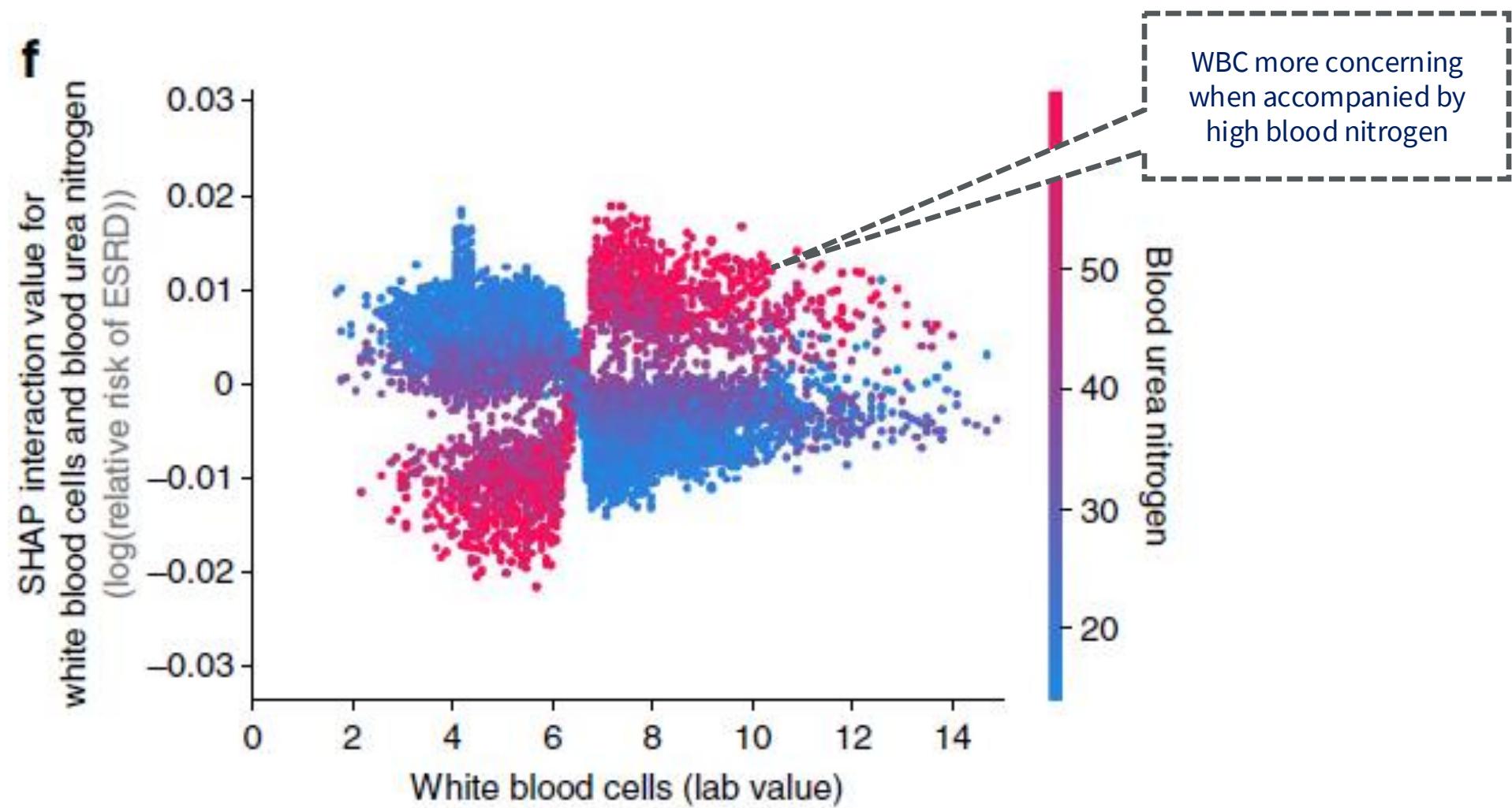
Systolic blood pressure effect on mortality risk varies with age

g



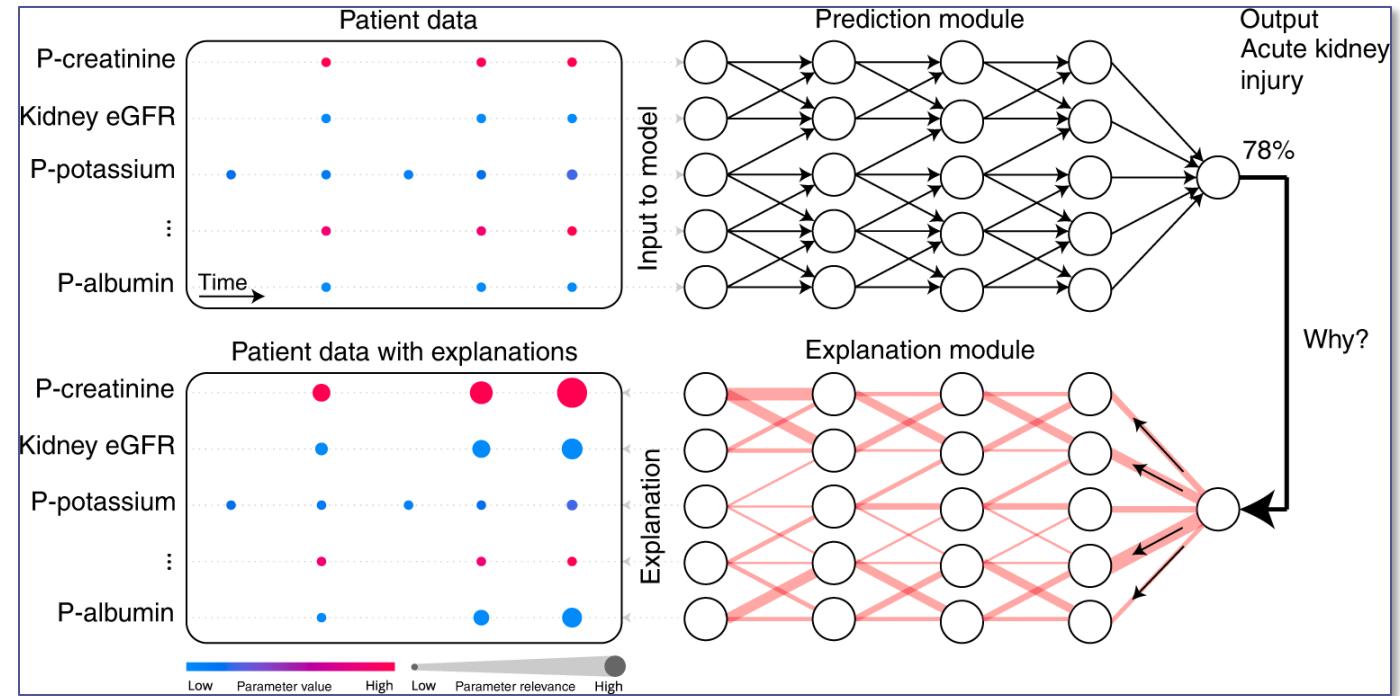
Differential risk of men and women changes over their lifetimes

XAI EXAMPLES: SHAP INTERACTION VALUES FOR NOVEL INSIGHTS



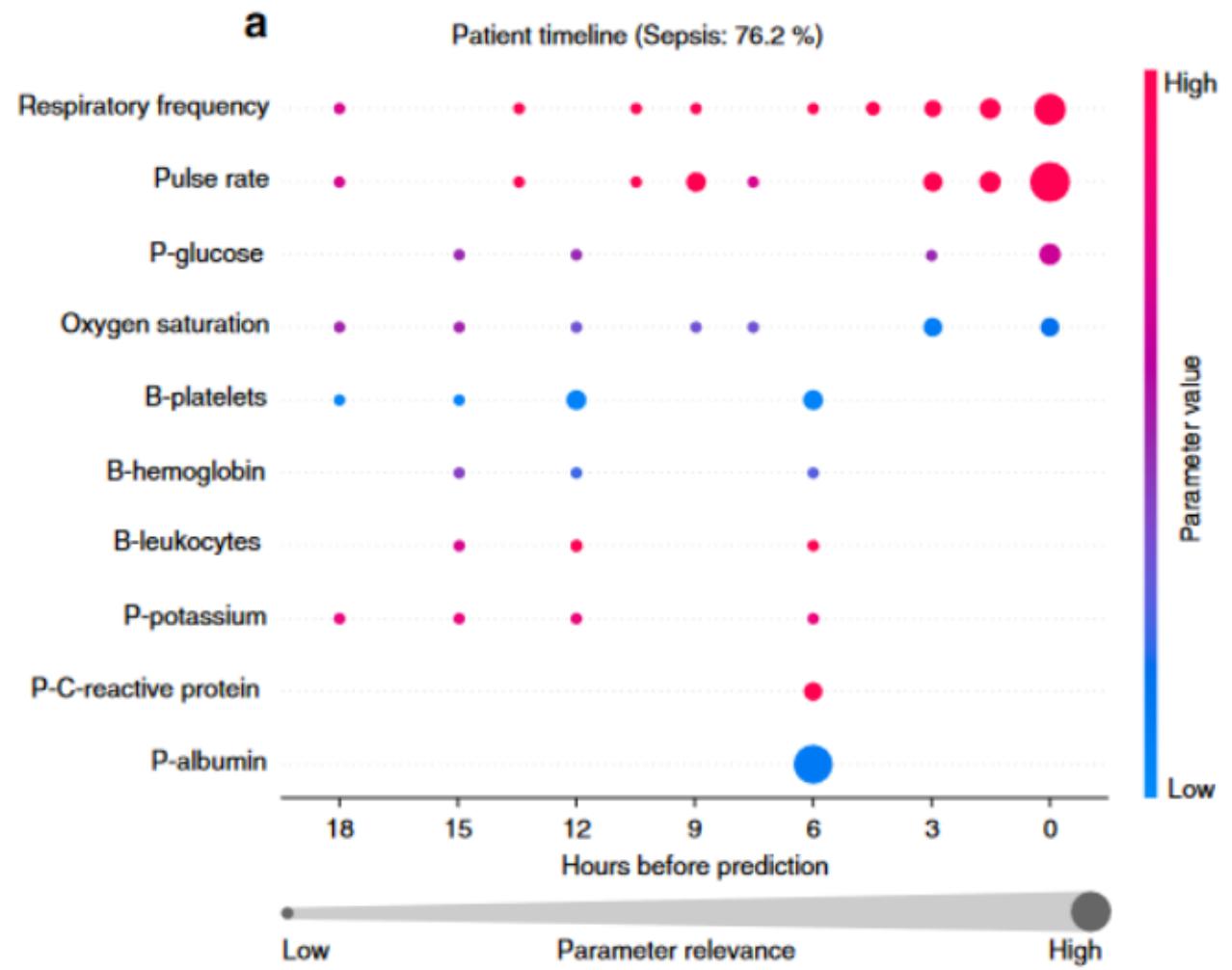
XAI EXAMPLE: TIME DEPENDENT DATA

- EHR Data + vital signs (n= 163.050)
- Real-time assessment
- Accounts for time effects: 24h
- **Prediction:** Temporal convolutional network (TCN)
- **XAI:** Deep Taylor Decomposition
- Global and local XAI (time effects)



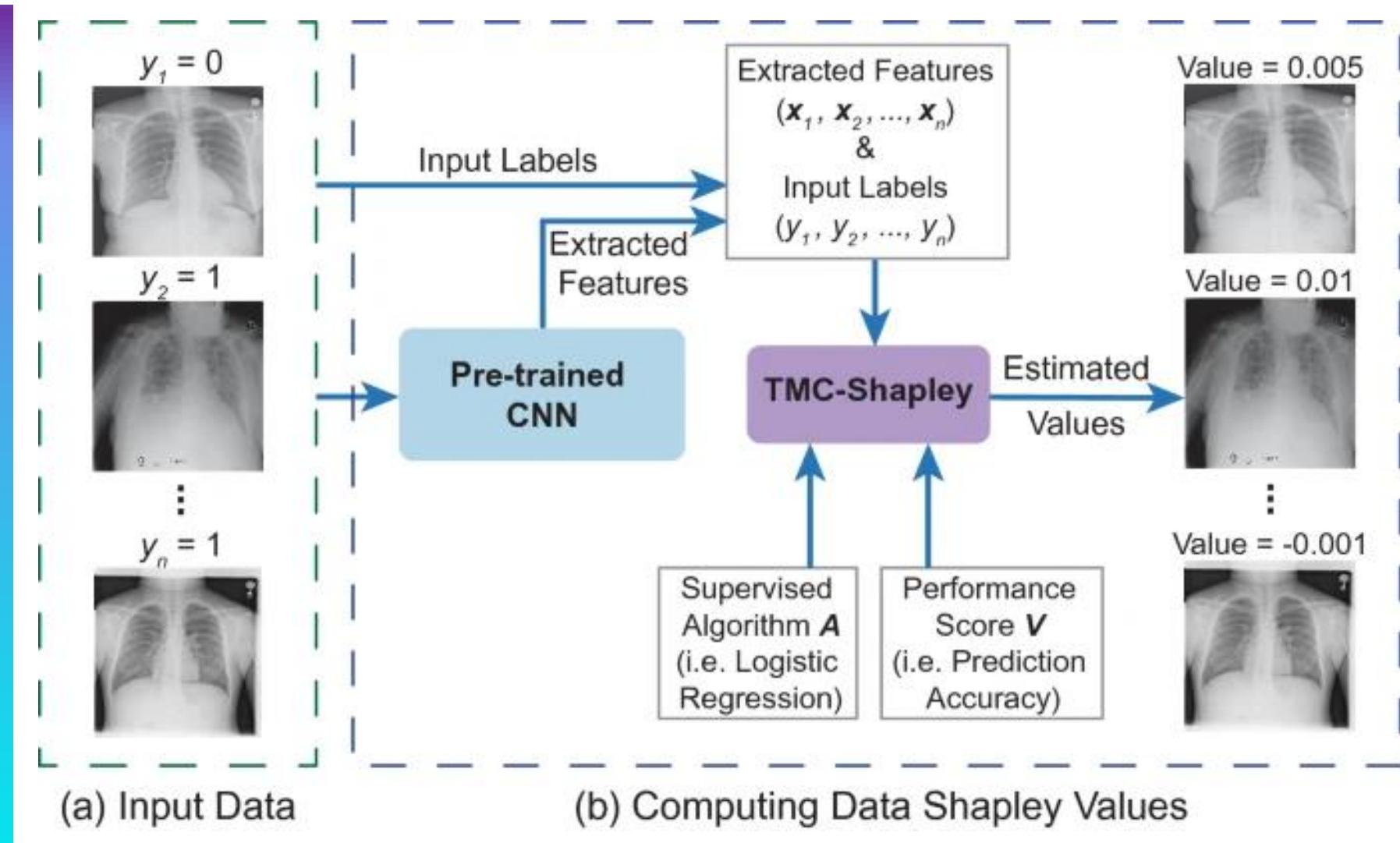
XAI EXAMPLE: TIME DEPENDENT DATA

- EHR Data + vital signs (n= 163.050)
- Real-time assessment
- Accounts for time effects: 24h
- **Prediction:** Temporal convolutional network (TCN)
- **XAI:** Deep Taylor Decomposition
- Global and local XAI (time effects)



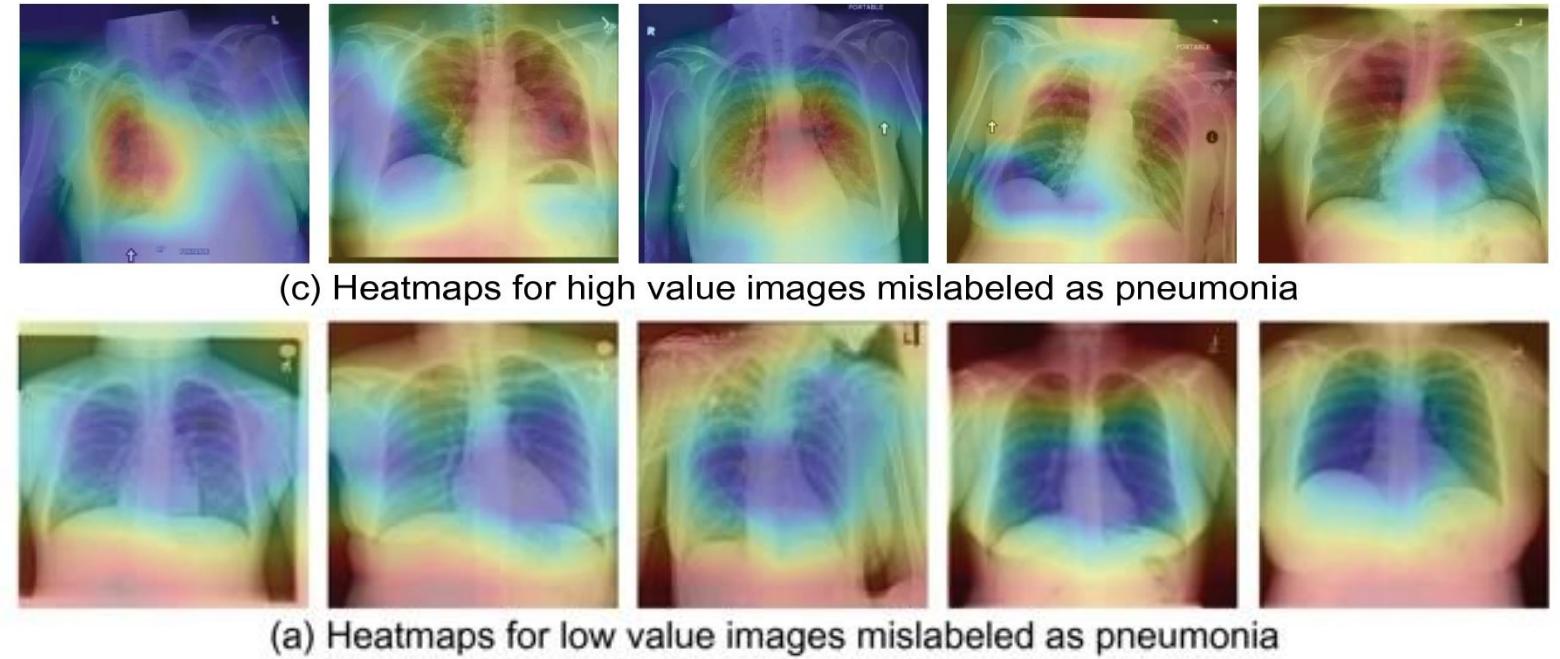
XAI EXAMPLES: SHAPLEY VALUES FOR LEVERAGING DATA QUALITY IN PNEUMONIA PREDICTION

- Feature vectors from pretrained CheXNet CNN
- Approximation of SVs based on Monte-Carlo sampling
- Logistic regression for pneumonia detection
- 3 radiologists for verification
- 2500 chest X ray images from ChestX-ray14



XAI EXAMPLES: SHAPLEY VALUES FOR LEVERAGING DATA QUALITY IN PNEUMONIA PREDICTION

- Removing data with high SVs decreased performance
- Removing data with low SVs improved performance
- Low SVs indicate mislabels and poor image quality



→ Insights into relevant features for model performance
→ Scalable data cleaning

XAI EXAMPLE: AFFECTIVE COMPUTING OF MULTIMODAL DATA



Research Project Phantomatrix:

- Multi-model ML to classify emotions without gender / cultural bias
- VR environments to evoke specific emotions
- FDA cleared wearable for data streams
- **Research Questions:**
 - Can emotions be classified by ML to construct new VR scenes?
 - How do individual differences, such as cultural backgrounds or gender affect emotion classification?
 - Can XAI explain emotion classification from multi-modal and time-dependent data?
 - Can XAI be used to detect gender / culture specific effects?

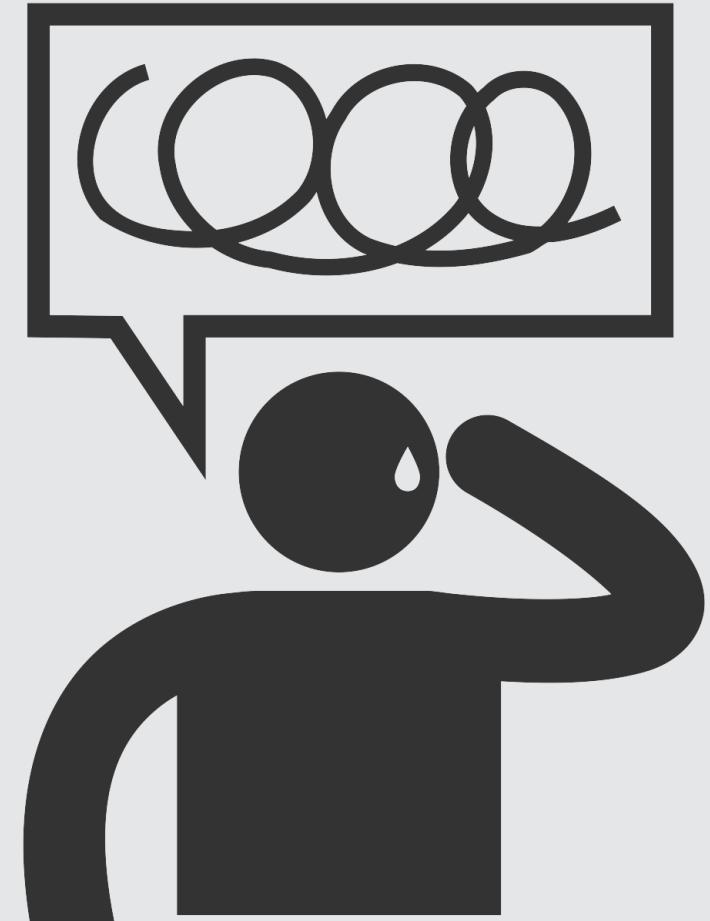
XAI EXAMPLE: AFFECTIVE COMPUTING OF MULTIMODAL DATA



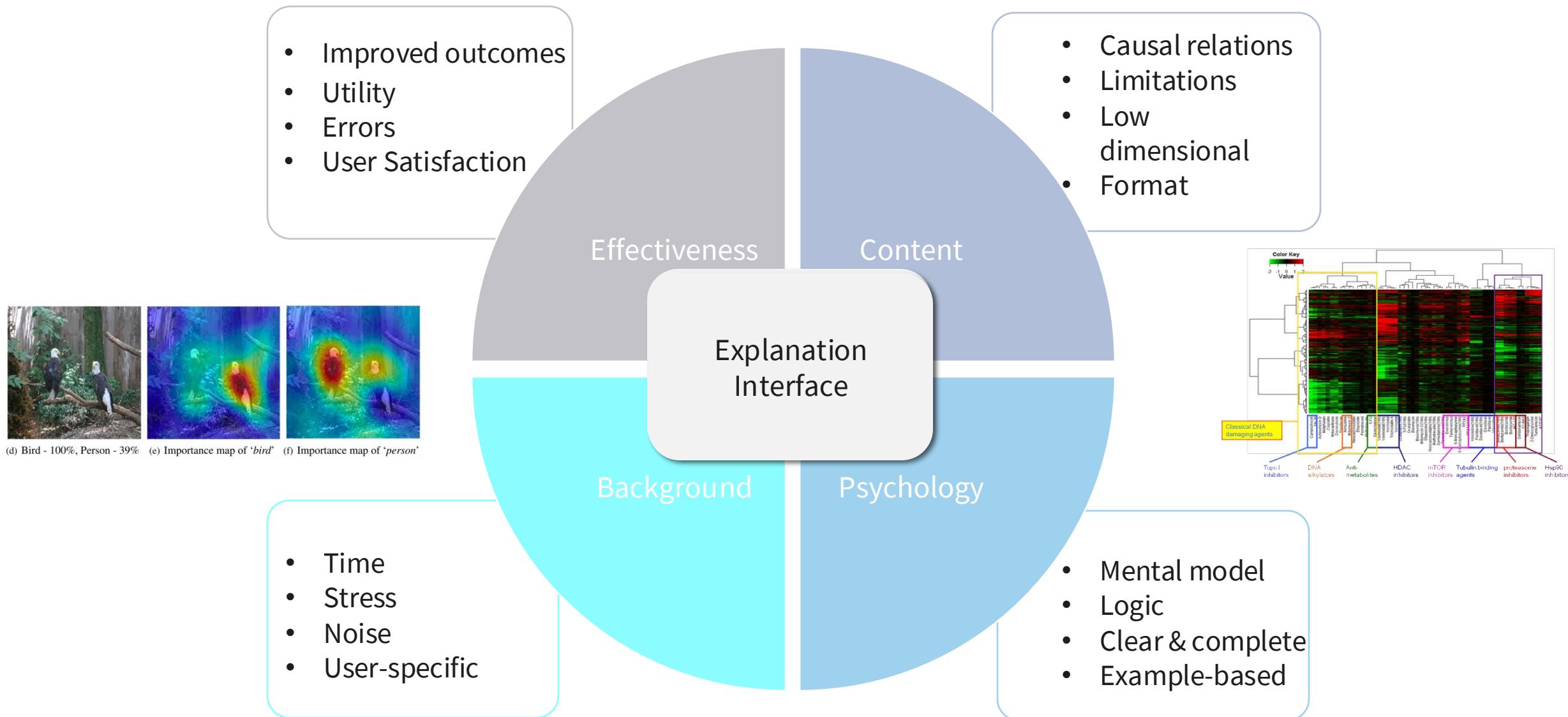
- Emotions are highly dependent on cultural backgrounds and gender
 - XAI can help to:
 - Guide feature engineering
 - Indicate the need to remove bias
 - Finding anomalous cases (clustering)
 - Remove bias: Effect of features like age, race and gender, are summed up and subtracted from the prediction
- Create a bias-free general affect model
- Find an intuitive way to visualize explanations

#4

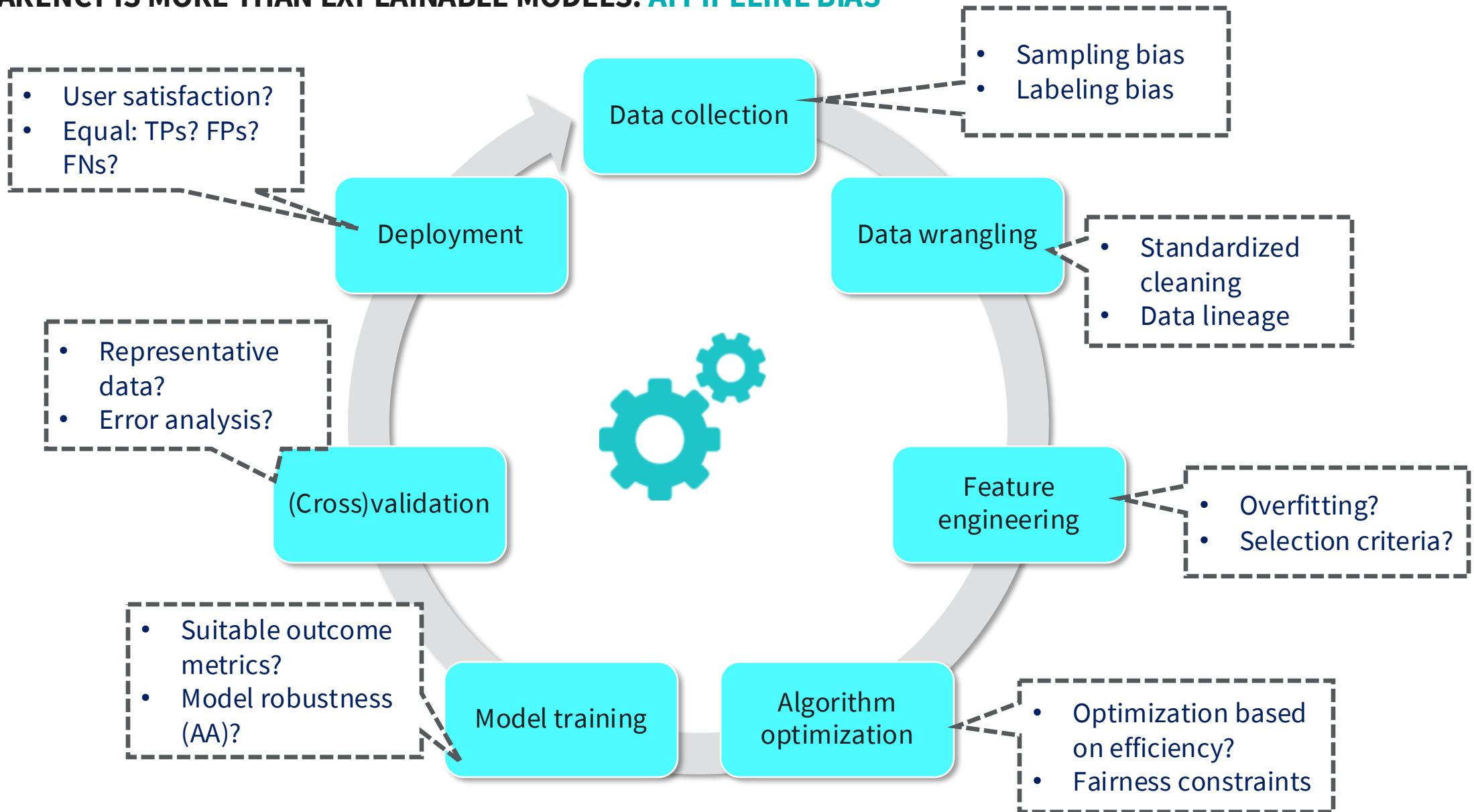
XAI CHALLENGES



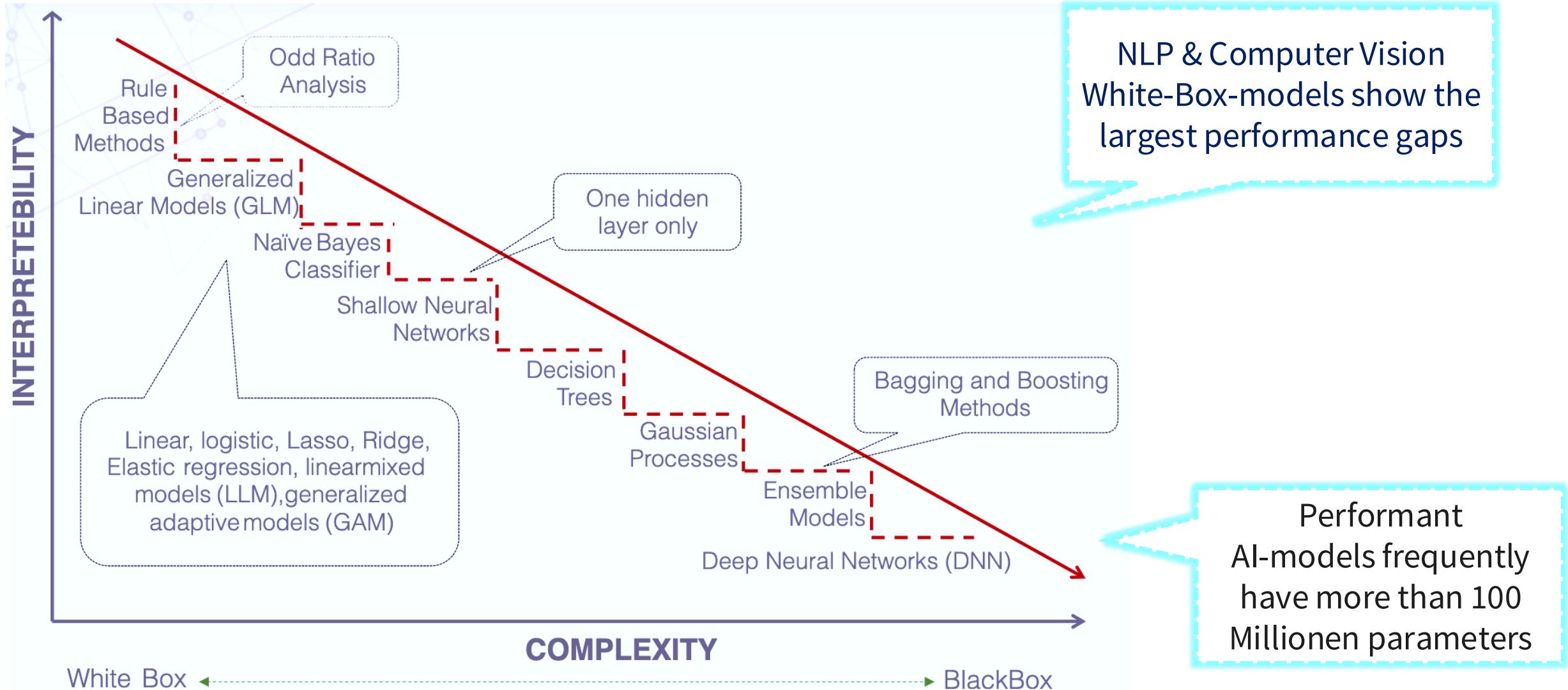
WHAT DEFINES A HUMAN UNDERSTANDABLE EXPLANATION?



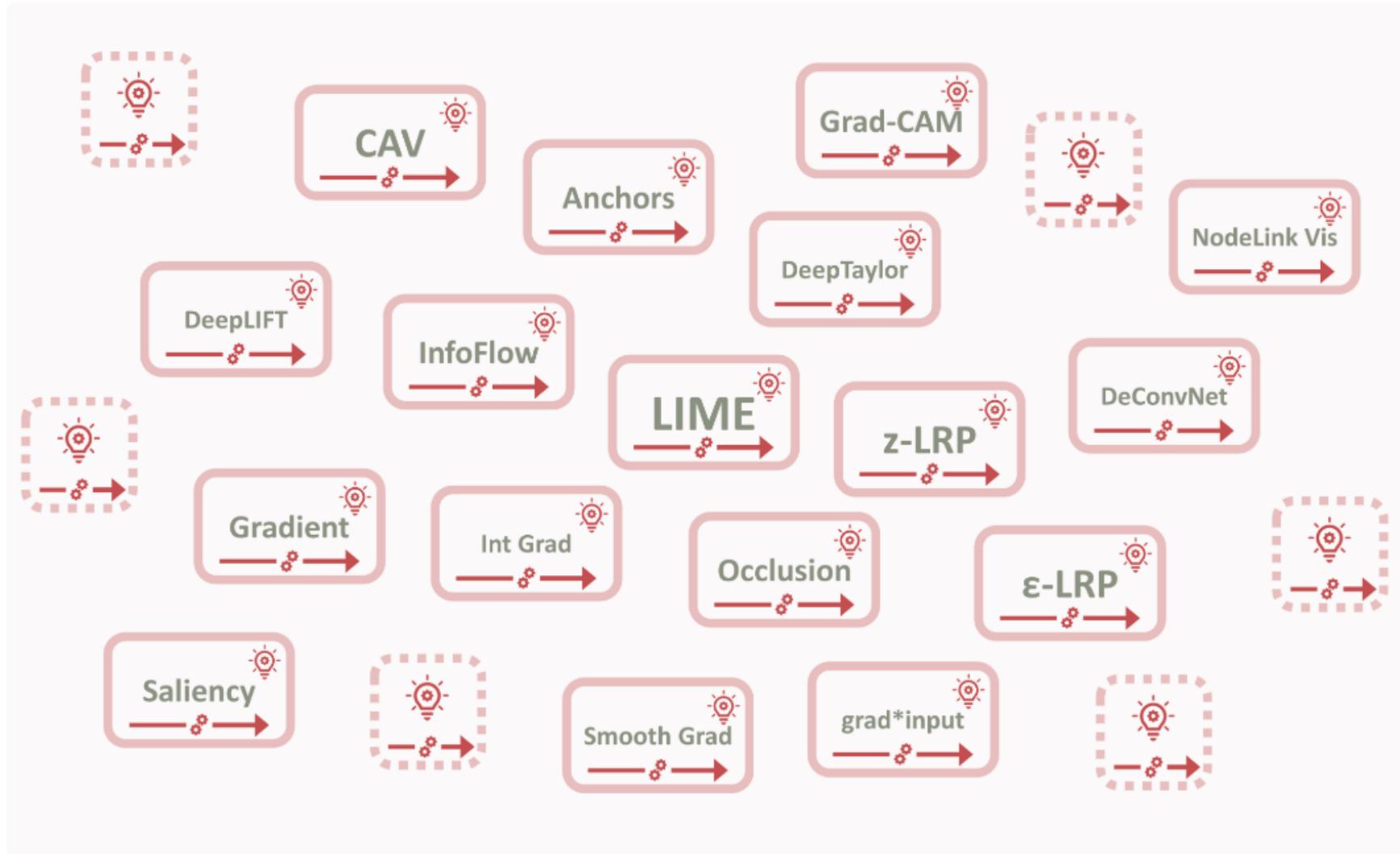
TRANSPARENCY IS MORE THAN EXPLAINABLE MODELS: AI PIPELINE BIAS



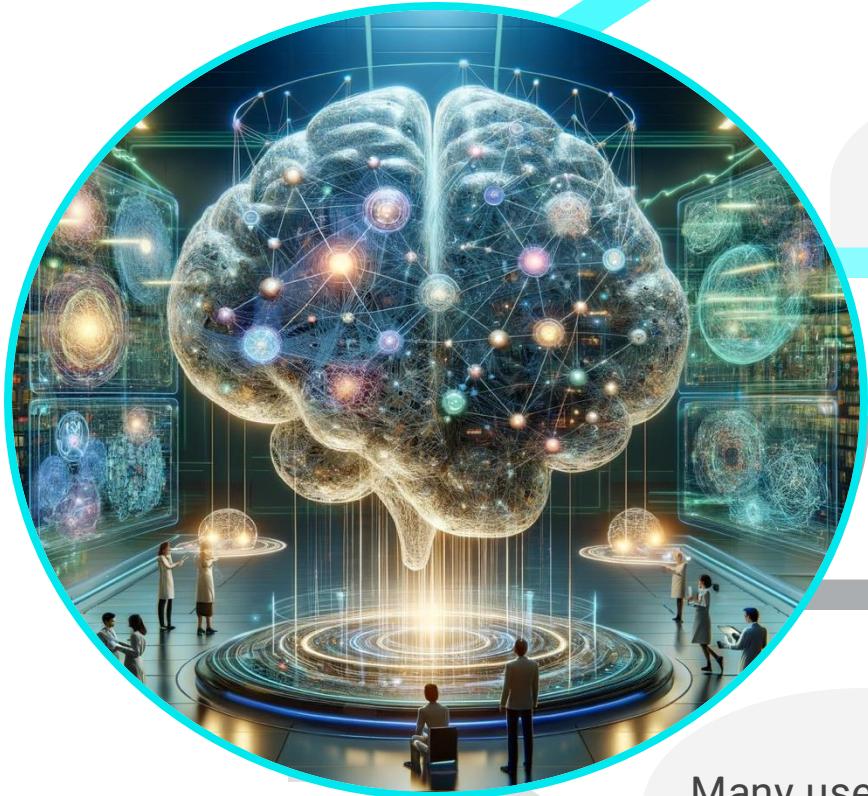
TRADEOFF: INTERPRETABILITY, COMPLEXITY AND ACCURACY



A MULTITUDE OF POSTHOC XAI AND LITTLE OVERLAP



XAI CHALLENGES: LLMS & XAI



Unrevealed training corpora: many sources and unclear weighting

Bias & consistency: biased data, hallucinations, toxic output, unrepeatable content

Many levels of training & alignment: RLHF, unclear which values are universally accepted,

Many endusers, tasks & sustainability:
Many users and tasks: hard to tailor XAI, energy consumption, water and CO₂

XAI CHALLENGE: AGENTIC AI

- **Non-determinism:** The same prompt can lead to different results.
- **Long reasoning chains:** Many intermediate steps and dynamic decisions.
- **Tool/environment black boxes:** Third-party tools, external APIs.
- **Multi-agent dynamics:** Interaction effects, emergent behavior, diffuse responsibilities.
- **Self-modification:** reinforcement learning, tool changes
- **Non-transparent governance layers:** system prompts, guardrails interact unpredictably
- **Real-time decision-making:**
Machine-Speed XAI



Gartner:

- 2024: <1 % of all enterprise implementations use Agentic AI
- 2028: 33 % will use Agentic AI

#5

XAI & SUSTAINABILITY



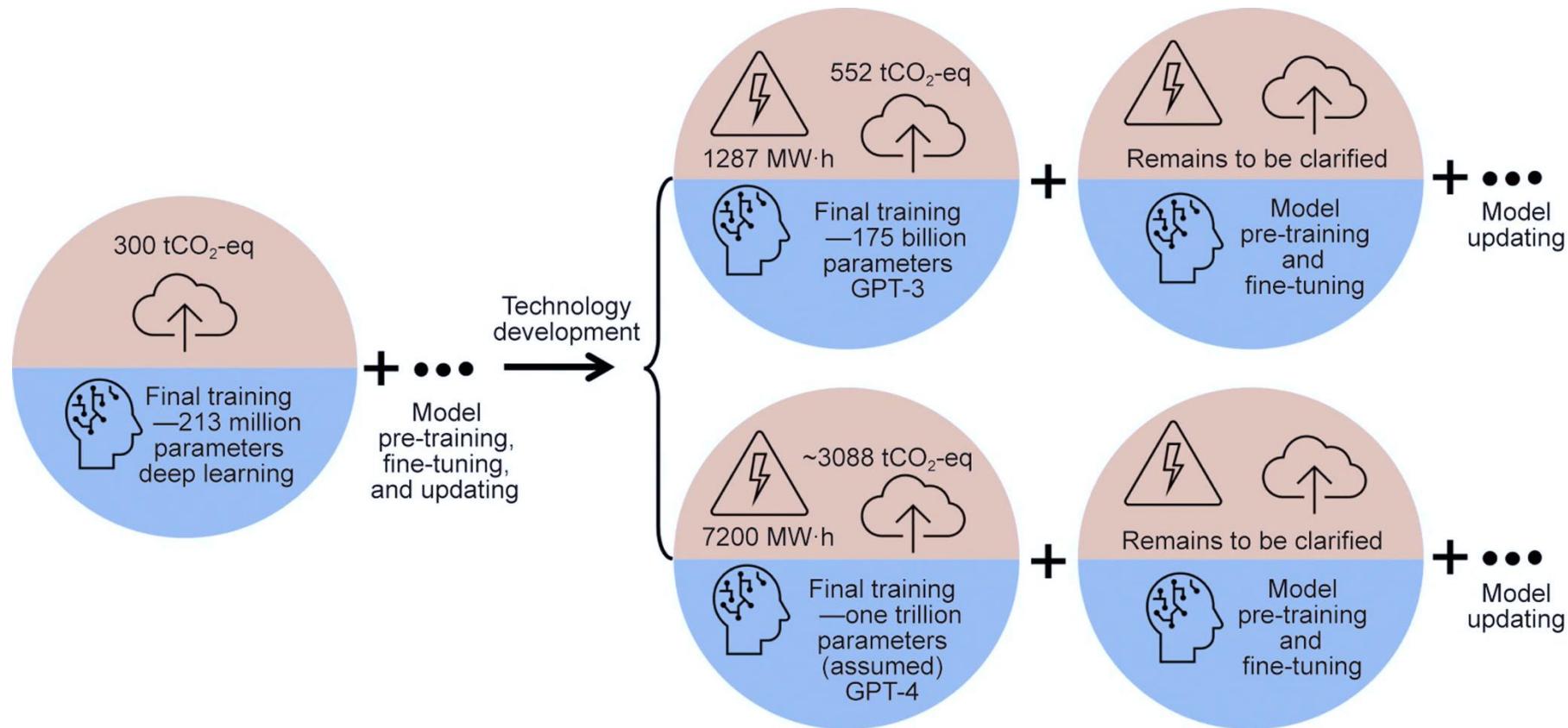
THE SUSTAINABILITY ISSUE OF XAI

XAI algorithms consider the explanation process an additional step alongside the prediction:

- **Computational costs**
(Lundberg et al. 2020: 3 CPU years for interacting SVs)
- **Energy / water consumption**
- **CO₂ emission**
- **Time overhead**



LIFE-CYCLE ENERGY AND CARBON FOOTPRINTS OF LLM



- The final training run of GPT-3 with 175 billion parameters consumes 1287 MW·h of electricity with a carbon footprint of 552 tCO₂-eq
- The final training run of GPT-4 with 1 Trillion parameters consumes 7200 MW·h of electricity with a carbon footprint of 3088 tCO₂-eq

THREE WAYS TO COUNTERACT XAI EFFECTS ON SUSTAINABILITY

Algorithm Optimization

- Computing **exact SV** is NP-hard, as it requires summing over all possible feature subsets (2^M combinations).
→ 30 features => 1 billion subsets
- **Lundberg et al. (2020)**
 - Restructured the computation for tree ensembles, reducing complexity from exponential to low-order polynomial time.
 - Instead of iterating over all subsets, TreeExplainer recursively tracks the proportion of all feature subsets that flow to each leaf node — effectively **simulating** all subsets simultaneously.

1

Hardware optimization

- **GPUTreeShap:** an adaptation of the TreeShap algorithm optimized for massively parallel processing of XAI algorithms on GPUs (Mitchell et al. 2022)
- Hardware architecture tailored to enhance XAI performance in **graph-convolutional networks** using field-programmable gatearrays (**FPGAs**) (Zhou et al. 2022).
- **XALedge:** energy-aware fine-tuned **approximate** computing into the XAI algorithms with parallel hardware acceleration (**TPUs**) (Siddique et al. 2025).

2

Feature engineering

- SHAP-based attributions can identify **redundant** or low-impact features, allowing models to be simplified without significant performance loss — reducing computational costs.
- Detecting feature interactions helps design **leaner models** by removing correlated or redundant inputs, minimizing both training and inference resource use.
- Revealing features that cause **instability**, **bias**, or **overfitting** enables targeted removal or reweighting, improving model robustness and reducing unnecessary retraining cycles.

3

SUSTAINABILITY PREDICTIONS WITH XAI: CARBON FOOTPRINT (CO₂ EMISSIONS) PREDICTION OF VEHICLES USING SHAP

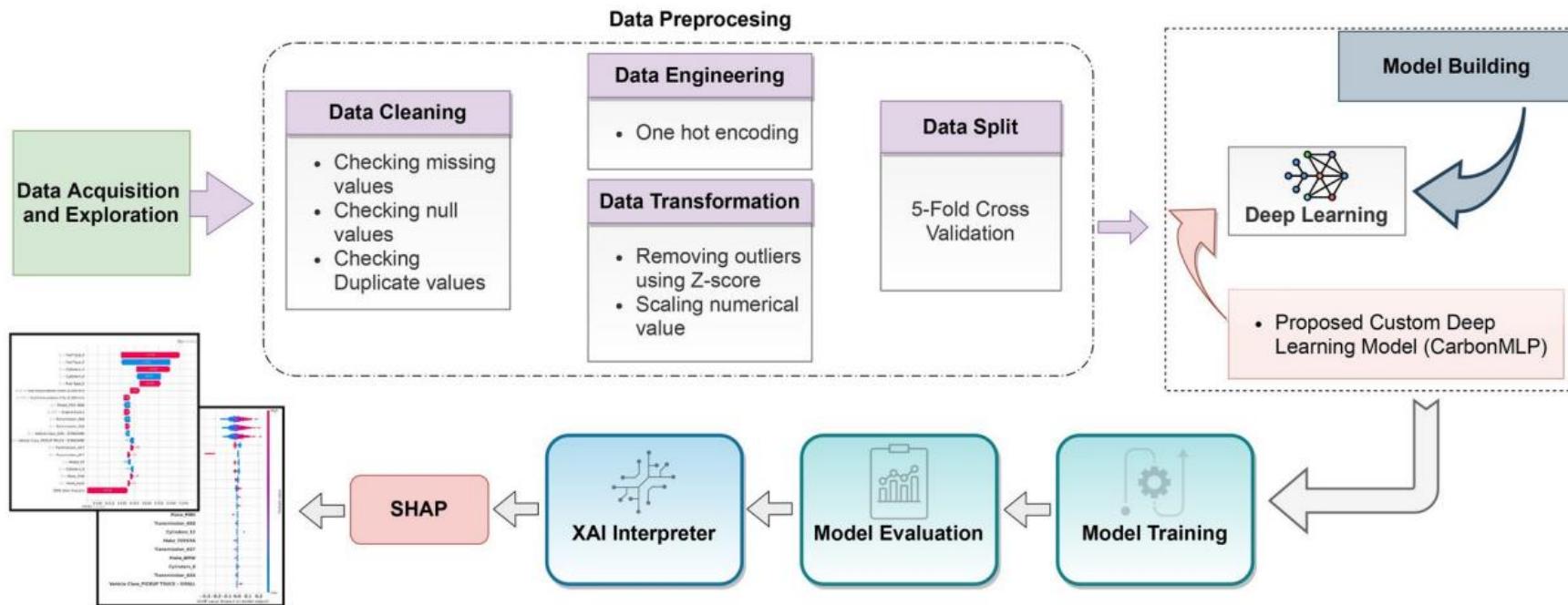


Fig. 2. Methodology Diagram Illustrating the Entire Process Described in the Paper, from Data Collection to Model Evaluation and Interpretation.

Approach



Custom Deep Learning Model (CarbonMLP) to predict the carbon footprint (CO₂ emissions) of vehicles



SHAP values to interpret how vehicle attributes influence emissions

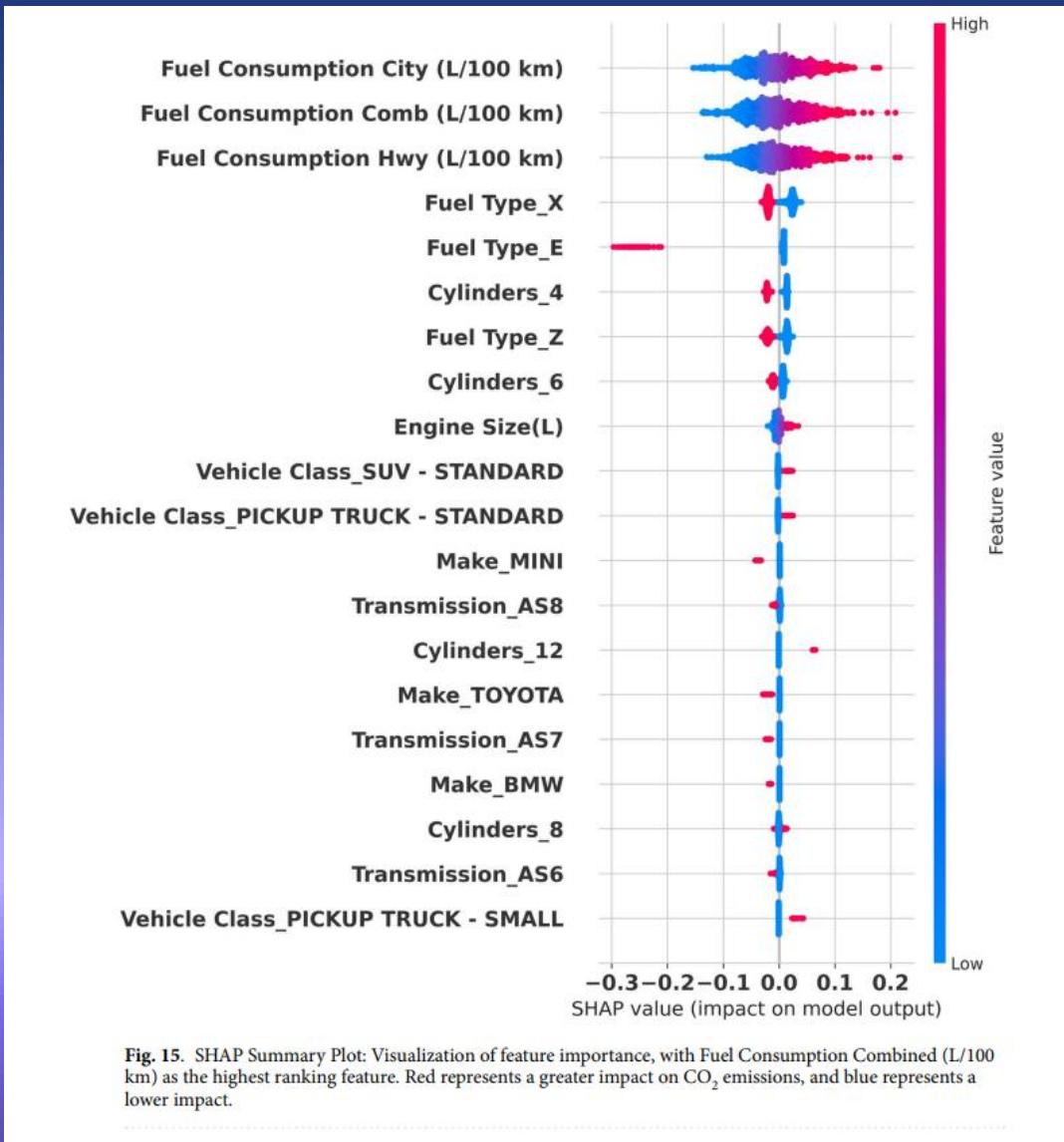


Trained on a dataset of 7,385 vehicles from Canada's open government database



High predictive accuracy: R² of 0.9938, outperforming other models (e.g., LSTM, BiLSTM, XGBoost)

CARBON FOOTPRINT (CO₂ EMISSIONS) PREDICTION OF VEHICLES USING SHAP



SHAP Summary Plot
Global feature importance
across all predictions

→ Fuel consumption had
the strongest positive
impact on CO₂ emissions;
followed by fuel type.

CARBON FOOTPRINT (CO₂ EMISSIONS) PREDICTION OF VEHICLES USING SHAP



SHAP Waterfall Plot

Explains individual predictions for specific vehicles.

Demonstrates how specific features raise or lower a single vehicle's CO₂ prediction relative to a baseline.

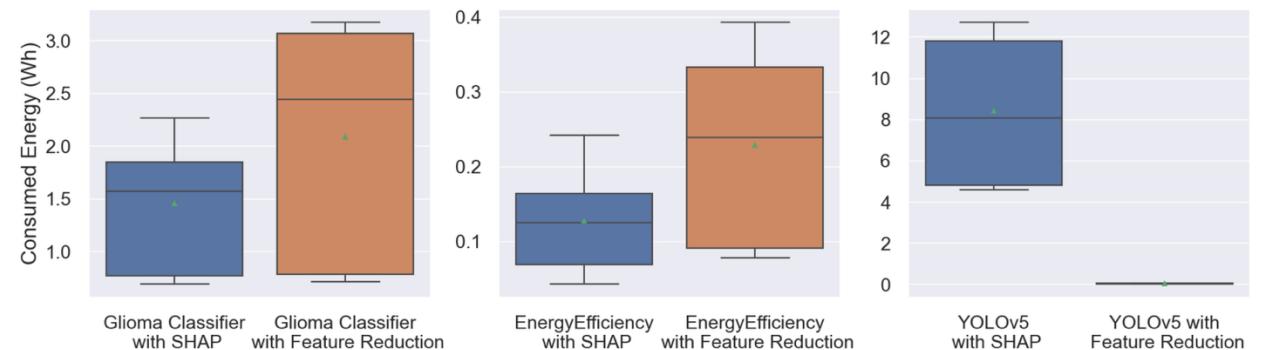
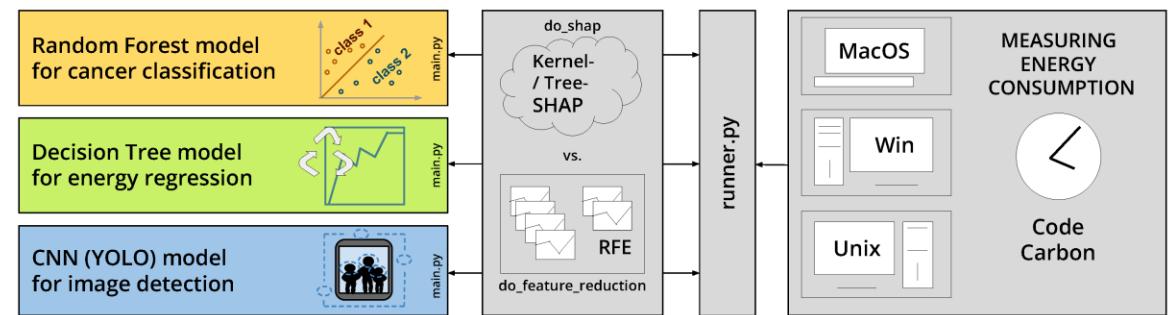
FEATURE ENGINEERING: THE COST OF UNDERSTANDING—XAI ALGORITHMS TOWARDS SUSTAINABLE ML IN THE VIEW OF COMPUTATIONAL COST

Method:

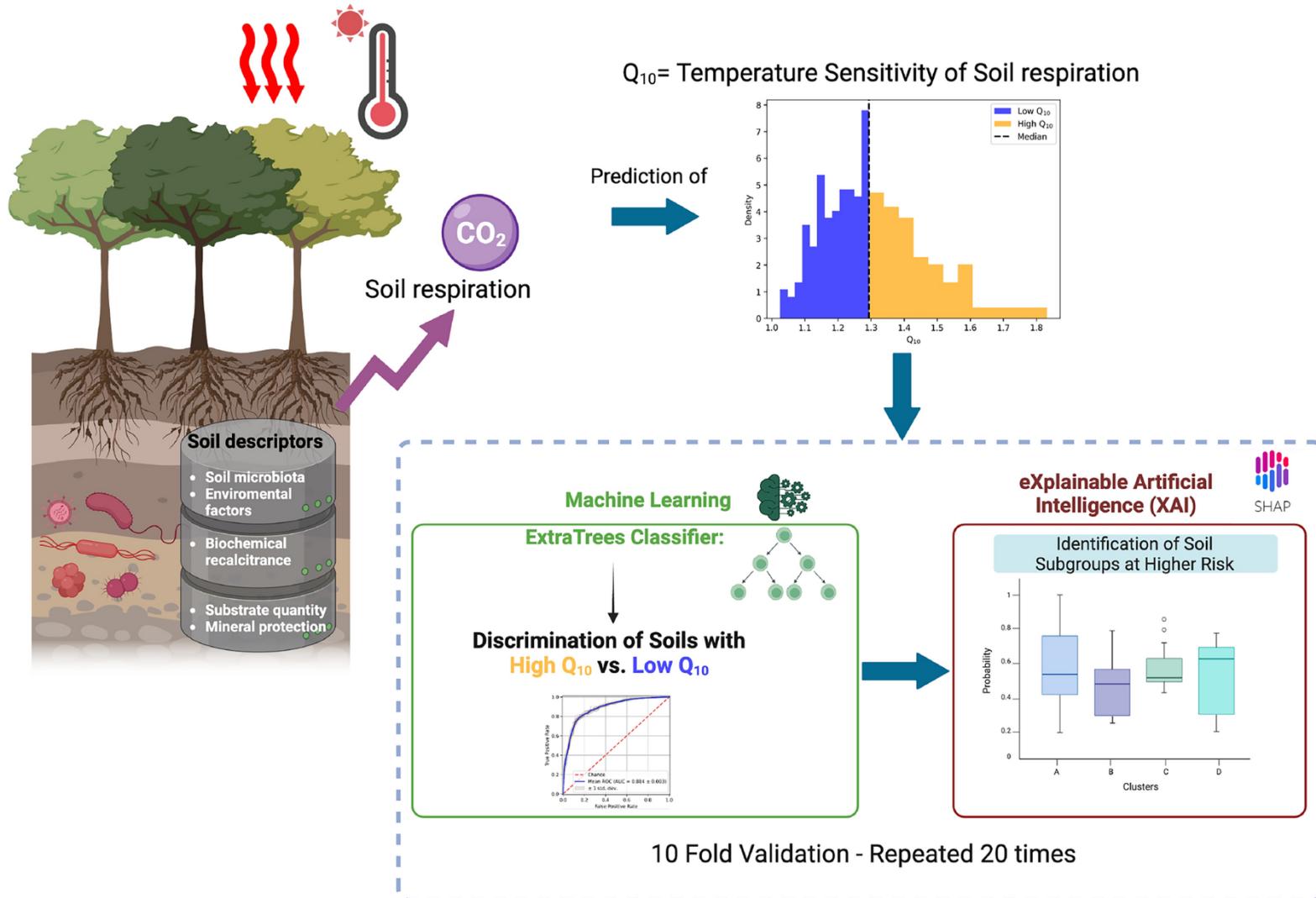
- Compared SHAP-based feature importance with Recursive Feature Elimination (RFE)
- SHAP values: to rank features by model contribution
- The top-ranked features were then retained for training

Outcome:

- **Random Forest & Decision Tree:** SHAP-based selection produced similar or better performance than RFE but with lower energy consumption and shorter training time.
- **Deep learning (YOLOv5):** SHAP XAI were computationally costly because of the need to calculate **localized image attributions**, leading to higher overall energy use.



FEATURE TRANSFORMATION: LEVERAGING EXPLAINABLE AI TO PREDICT SOIL RESPIRATION SENSITIVITY (Q_{10}) AND ITS DRIVERS FOR CLIMATE CHANGE MITIGATION



Novielli et al., 2025

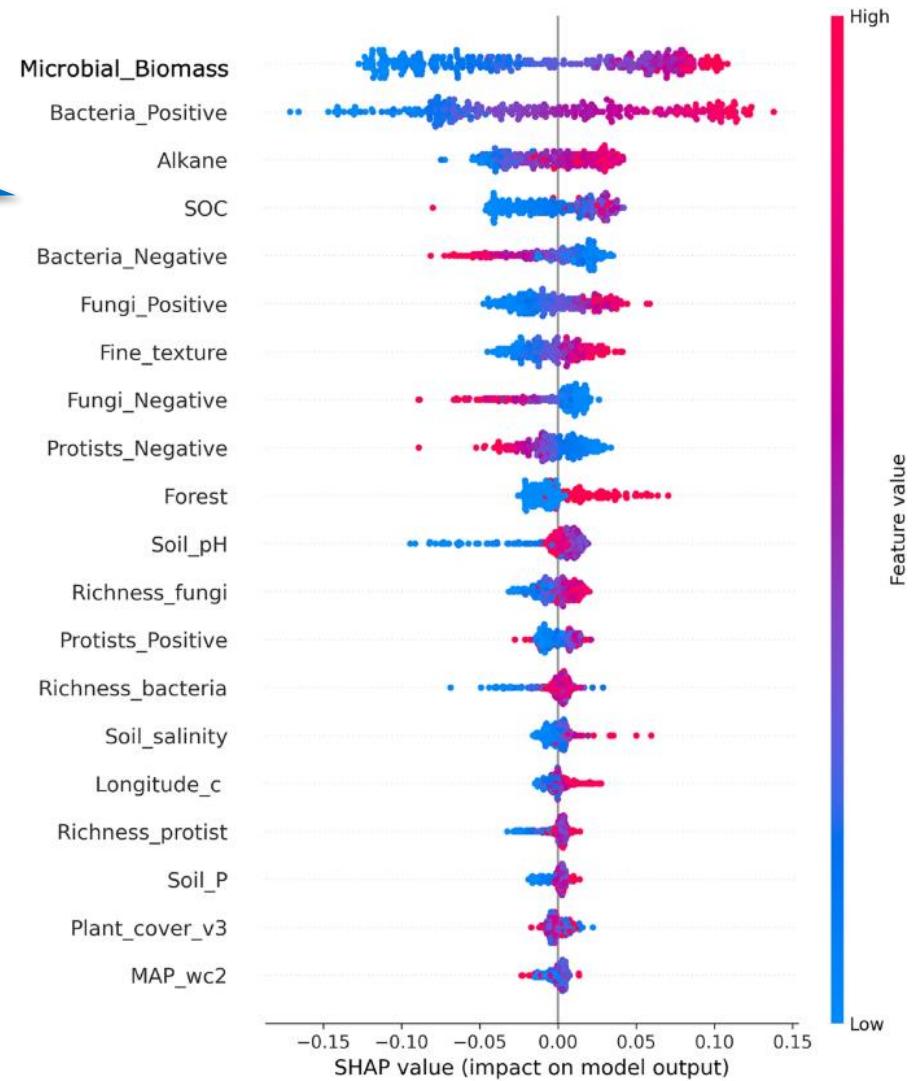
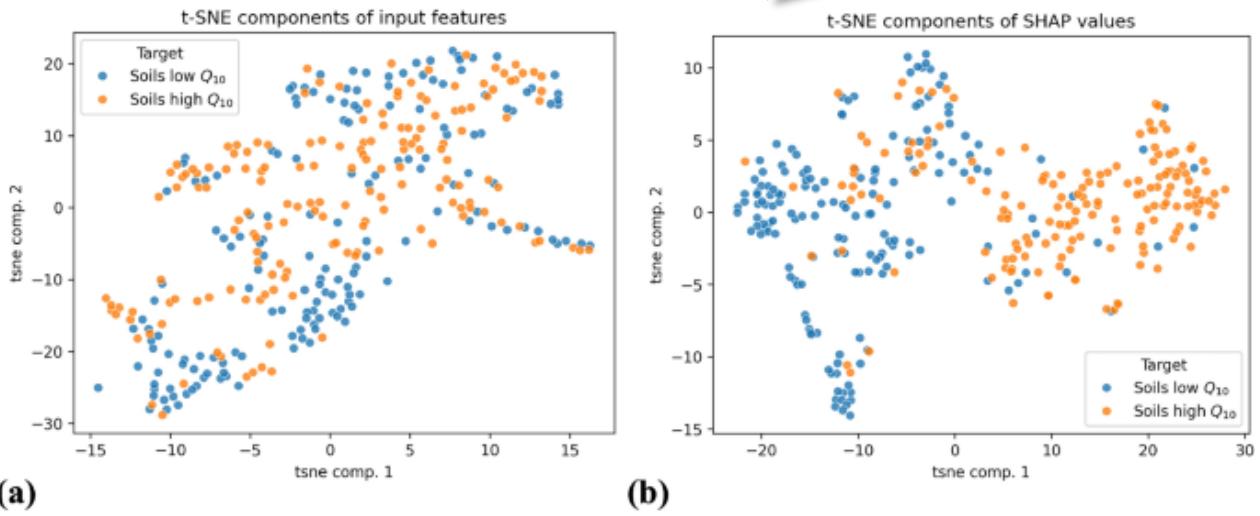
Approach

- **Global dataset** of 332 soil samples from 29 countries containing 27 environmental, biochemical, and microbiome features
- **Extra Trees Classifier** distinguishes between high and low Q_{10} soils
- **SHAP for:**
 1. Revealing key drivers of soil CO_2 sensitivity (e.g., microbial biomass, carbon content) (**post-hoc**)
 2. SHAP-based embeddings, t-SNE visualizations showed clearer separations between high- and low- Q_{10} soils than using raw features.

LEVERAGING EXPLAINABLE AI TO PREDICT SOIL RESPIRATION SENSITIVITY (Q_{10}) AND ITS DRIVERS FOR CLIMATE CHANGE MITIGATION

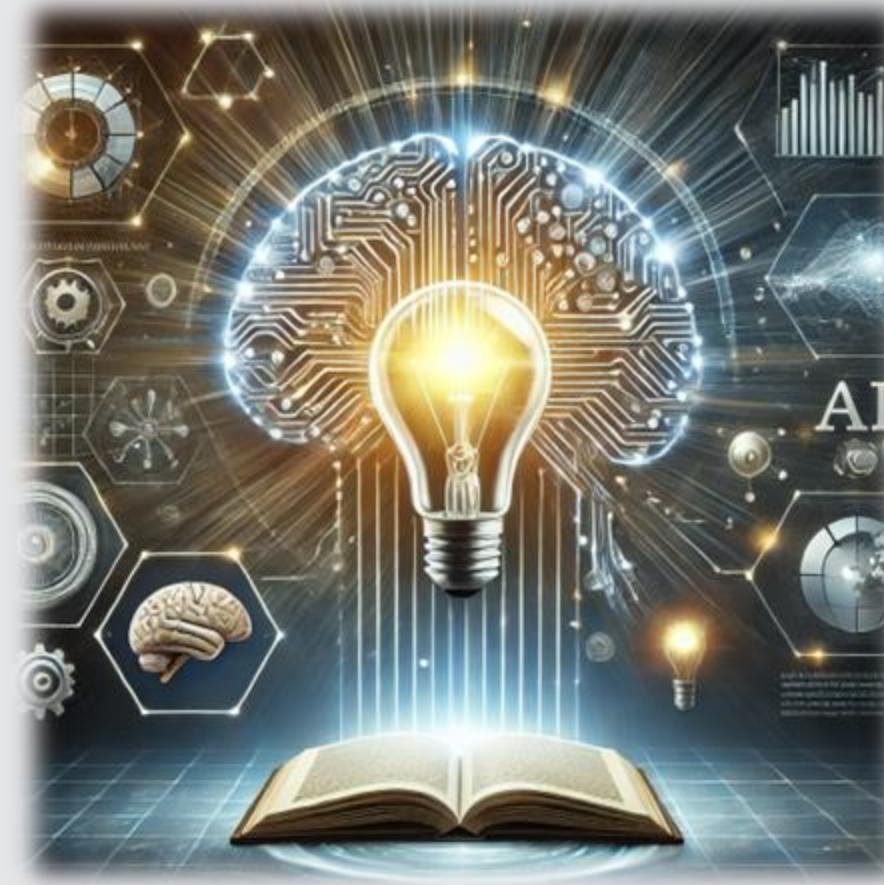
SHAP summary plot of feature importance in predicting Q_{10} sensitivity: most influential predictors : “Bacteria_Positive” and “Microbial_Biomass”

SHAP value-based T-SNE projections showing improved separation of high Q_{10} and low Q_{10} soils



#6

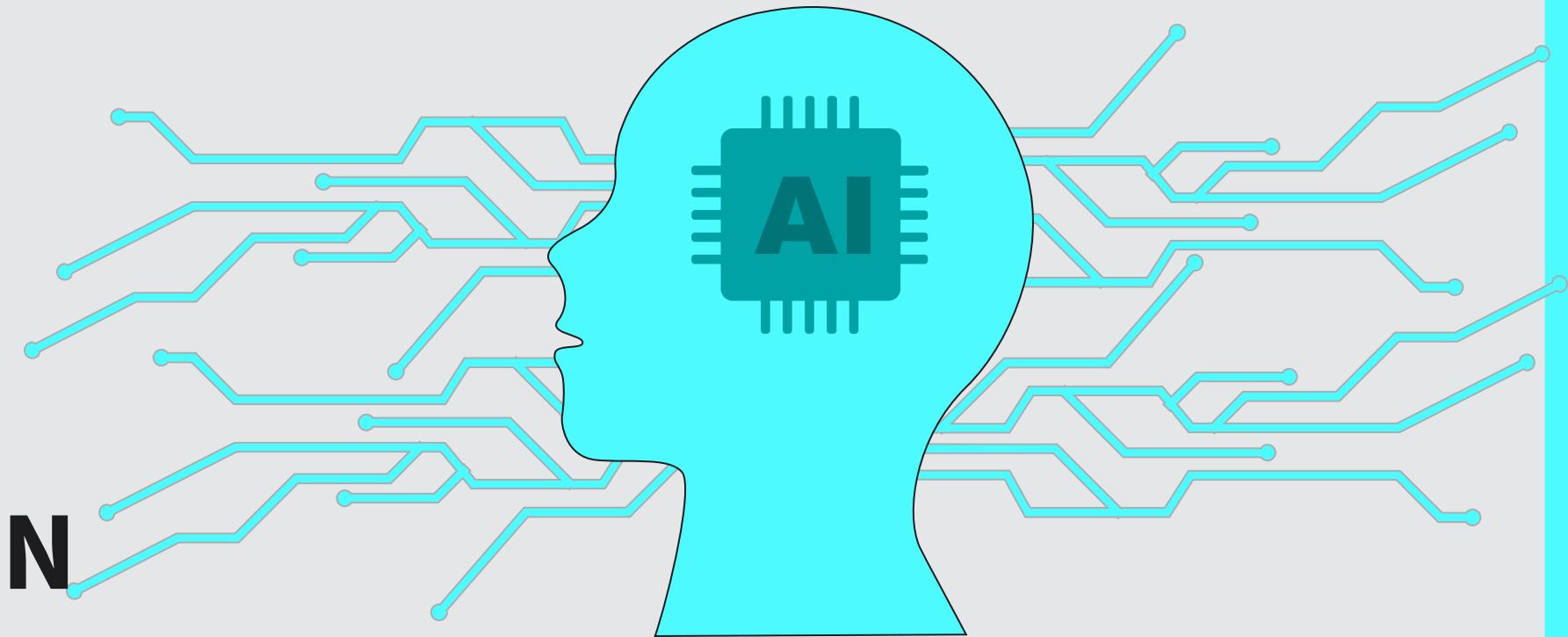
TAKE-HOME- MESSAGE



XAI HAS MANY FACETS



**THANK
YOU FOR
YOUR
ATTENTION**



LIST OF SOURCES

- Barman, K.G., Wood, N. & Pawłowski, P. Beyond transparency and explainability: on the need for adequate and contextualized user guidelines for LLM use. *Ethics Inf Technol* 26, 47 (2024). <https://doi.org/10.1007/s10676-024-09778-2>
- Belle V, Papantonis I. Principles and Practice of Explainable Machine Learning. *Front Big Data*. 2021 Jul 1;4:688969. doi: 10.3389/fdata.2021.688969. PMID: 34278297; PMCID: PMC8281957.
- Bologna, Guido & Hayashi, Yoichi. (2017). Characterization of Symbolic Rules Embedded in Deep DIMLP Networks: A Challenge to Transparency of Deep Learning. *Journal of Artificial Intelligence and Soft Computing Research*. 7. 265. 10.1515/jaiscr-2017-0019.
- Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: A systematic literature review. *Artificial Intelligence Review*, 57(8), 216.
- Gilpin et al. (2019). Explaining Explanations: An Overview of Interpretability of Machine Learning. *arXiv:1806.00069*
- Gunning, D., Vorm, E., Wang, J.Y. and Turek, M. (2021), DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters*, 2: e61. <https://doi.org/10.1002/ail2.61>
- Kassuhn W, Klein O, Darb-Esfahani S, Lammert H, Handzik S, Taube ET, Schmitt WD, Keunecke C, Horst D, Dreher F, George J, Bowtell DD, Dorigo O, Hummel M, Sehouli J, Blüthgen N, Kulbe H, Braicu EI. Classification of Molecular Subtypes of High-Grade Serous Ovarian Cancer by MALDI-Imaging. *Cancers*. 2021; 13(7):1512. <https://doi.org/10.3390/cancers13071512>
- Kim, H., Jung, D. C., & Choi, B. W. (2019). Exploiting the vulnerability of deep learning-based artificial intelligence models in medical imaging: adversarial attacks. *Journal of the Korean Society of Radiology*, 80(2), 259-273.
- Hendrycks, D., Zhao, K., Basart, S., Steinhhardt, J., & Song, D. (2021). Natural adversarial examples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15262-15271).
- S. Jain, B.C. Wallace. Attention is not Explanation. (2019). 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics.
- Kearns, M.; Neel, S.; Roth, A.; Wu, Z.S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 2564–2572
- Kim, H., Jung, D. C., & Choi, B. W. (2019). Exploiting the vulnerability of deep learning-based artificial intelligence models in medical imaging: adversarial attacks. *Journal of the Korean Society of Radiology*, 80(2), 259-273.
- Lapuschkin, S., Wäldchen, S., Binder, A. et al. Unmasking Clever Hans predictors and assessing what machines really learn. *Nat Commun* 10, 1096 (2019). <https://doi.org/10.1038/s41467-019-10897-4>
- Lundberg S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, 2(1), 56-67.
- Hind, M.; Wei, D.; Campbell, M.; Codella, N.C.; Dhurandhar, A.; Mojsilović, A.; Natesan Ramamurthy, K.; Varshney, K.R. TED: Teaching AI to explain its decisions. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, 27–28 January 2019; pp. 123–129.
- Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H. Causability and explainability of artificial intelligence in medicine. *WIREs*
- Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI. Retrieved from <http://arxiv.org/abs/1902.01876>
- Ribiero, et al. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. <https://arxiv.org/abs/1602.04938>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
- C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. (2019). *Nature Machine Intelligence*.
- Tang, S., Ghorbani, A., Yamashita, R., Rehman, S., Dunnmon, J. A., Zou, J., & Rubin, D. L. (2021). Data valuation for medical imaging using Shapley value and application to a large-scale chest X-ray dataset. *Scientific reports*, 11(1), 8366.
- Valdes, G., Luna, J., Eaton, E. et al. MediBoost: a Patient Stratification Tool for Interpretable Decision Making in the Era of Precision Medicine. *Sci Rep* 6, 37854 (2016). <https://doi.org/10.1038/srep37854>
- Vlasceanu, M., & Amodio, D. M. (2022). Propagation of societal gender inequality by internet search algorithms. *Proceedings of the National Academy of Sciences*, 119(29), e2204529119.
- Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for Large Language Models: A Survey. *ACM Trans. Intell. Syst. Technol.* 15, 2, Article 20 (April 2024), 38 pages. <https://doi.org/10.1145/3639372>
- Weitz, Katharina, Hassan, Teena, Schmid, Ute and Garbas, Jens-Uwe. "Deep-learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable AI methods" tm - Technisches Messen, vol. 86, no. 7-8, 2019, pp. 404-412. <https://doi.org/10.1515/teme-2019-0024>
- Zicari CSIG, 2019; World Economic Forum: white paper

© 2022 IU Internationale Hochschule GmbH

This content is protected by copyright. All rights reserved.

This content may not be reproduced and/or electronically edited, duplicated, or distributed in any kind of form without written permission by the IU Internationale Hochschule GmbH.