CIS 8398
Advanced AI Topics in Business

#Explainable AI

Yu-Kai Lin

Agenda

In this session, we will look into an emerging topic in machine learning: **Explainability**

- Importance of Explainability
- Taxonomy of Explainability Methods

[**Acknowledgements**] The materials in the following slides are based on the source(s) below:

- Interpretable Machine Learning by Christoph Molnar
- AAAI 2020 Tutorial on Explainable AI

Explainability in AI and ML

- The vast majority of machine learning (ML) models are designed to make good predictions, and they typically perform really well in that regard
- As artificial intelligence (AI) becomes more advanced, humans are challenged to comprehend and retrace how the algorithm came to a result
- **Explainability** is the degree to which a human can understand the cause of a decision
  - The higher the explainability of a ML model, the easier it is for someone to comprehend why certain decisions or predictions have been made
  - A model is more explainable than another model if its decisions are easier for a human to comprehend than decisions from the other model

Explainability or Interpretability?

- In the context of AI/ML, we often use the terms "explainability" and "interpretability" interchangeably
    - You could argue that interpret and explain have different definitions in the dictionary. But in my view, the differences are really not meaningful here.
- Since AI is more than just ML, explainable AI (XAI) involves more than interpretable ML (IML). However, since we are mainly interested in ML (and not other areas of AI, such as planning, search, robotics, etc.), XAI and IML are equivalent in the scope of this lecture.

The dark secret at the heart of AI

> No one really knows how the most advanced algorithms do what they do.
> That could be a problem.
>
> — Will Knight
> The dark secret at the heart of AI

Glass-box vs. black-box ML models

- Glass-box models are the ones that intrinsically permit humans, at least the experts in the domain, to understand how a prediction was made
- Black-box models, on the other hand, are extremely hard to explain and can hardly be understood even by domain experts
    - How difficult? Check out the article "Visualizing CNN architectures side by side with mxnet" by Joseph Paul Cohen to see some examples
- Aside from certain simple models (regressions, decision trees, etc.), most AI/ML models are black-box models and are not designed to offer explanations

Accuracy vs. Explainability

- There is often a clear trade-off between accuracy vs. explainability
- To increase accuracy, ML models often rely on complex non-linear functions (e.g., deep learning) or combine multiple models (e.g., random forests)
- As a result, these make it very difficult to interpret what is happening inside an algorithm

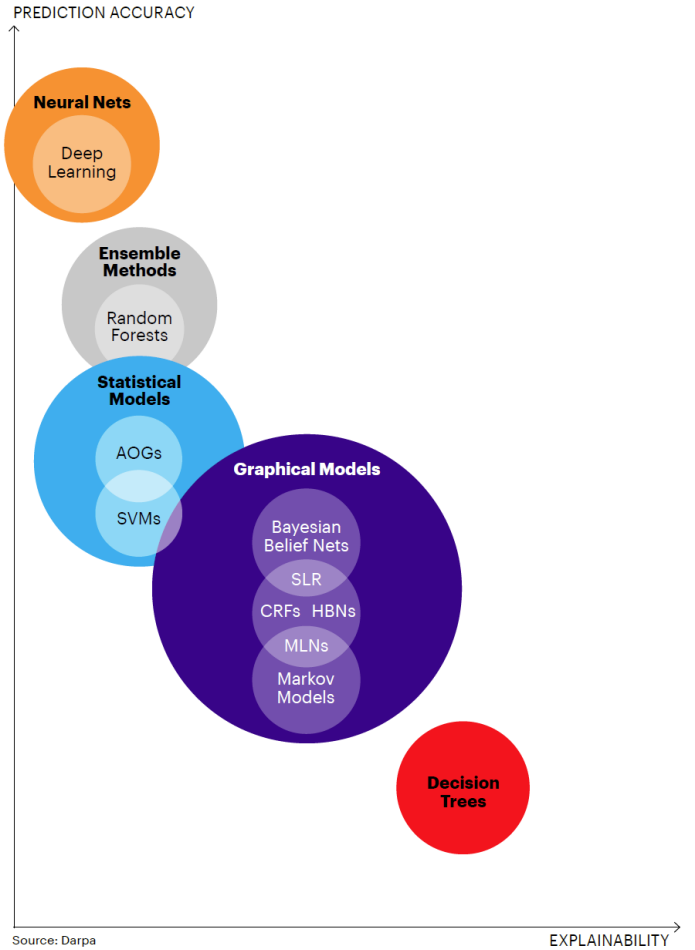


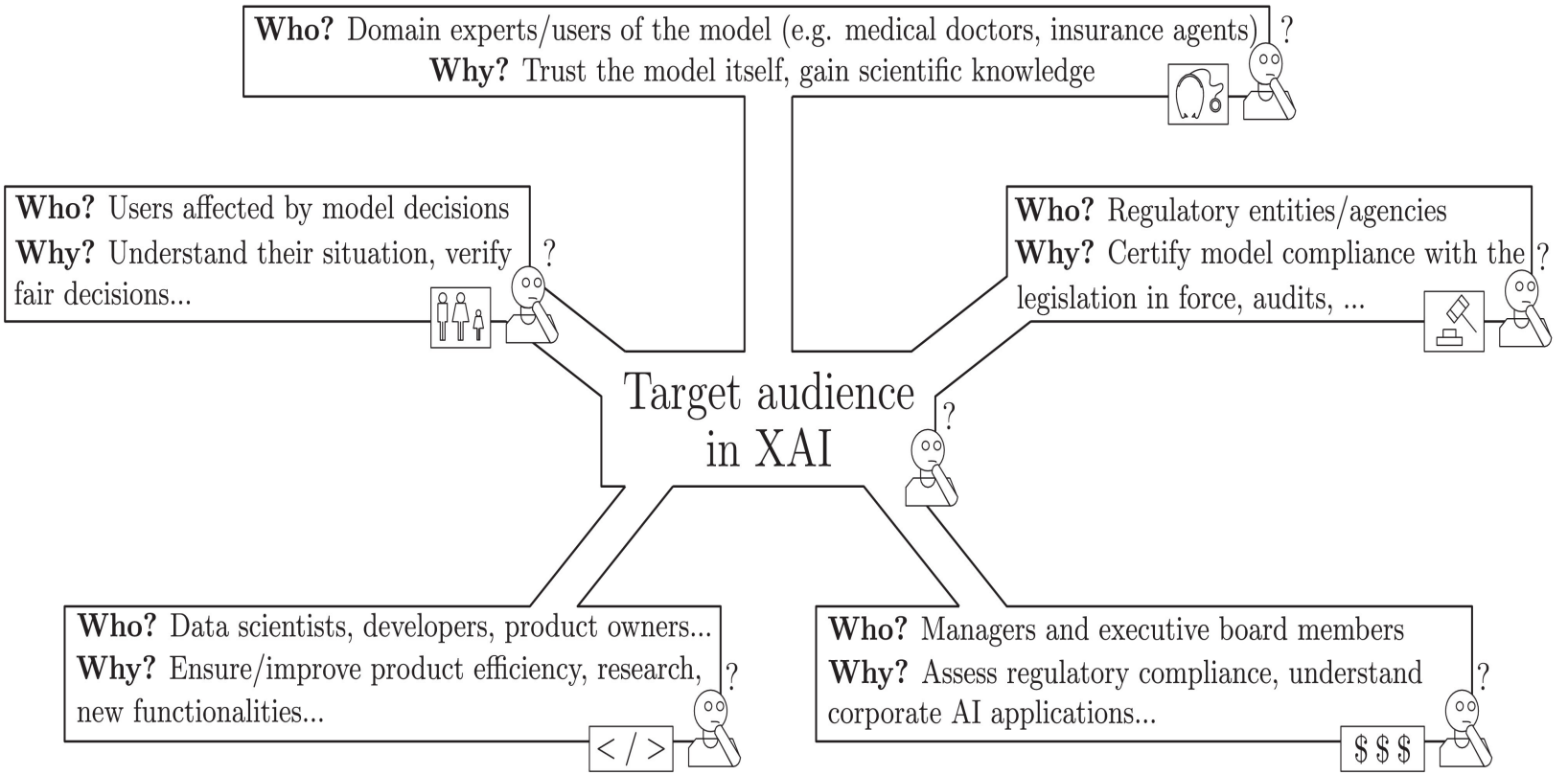
Image source: Explainable AI: The next stage of human-machine collaboration by Accenture

Explainability is essential for ...

- Model debugging - Why did my model make this mistake?
- Feature engineering - How can I improve my model?
- Detecting fairness issues - Does my model discriminate?
- Human-AI cooperation - How can I understand and trust the model's decisions?
- Regulatory compliance - Does my model satisfy legal requirements?
- High-risk applications - Healthcare, finance, judicial, ...

Importance of explainability

XAI enables **model governance** for fairness, accountability, and transparency



Source: Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.

Growing global AI regulation

- **GDPR**: Article 22 empowers individuals with the right to demand an explanation of how an automated system made a decision that affects them
- **Algorithmic Accountability Act of 2019**: Requires companies to provide an assessment of the risks posed by the automated decision system to the privacy or security and the risks that contribute to inaccurate, unfair, biased, or discriminatory decisions impacting consumers (reintroduced in 2022)
- **California Consumer Privacy Act**: Requires companies to rethink their approach to capturing, storing, and sharing personal data (effective on January 1st, 2020)
- **Washington Bill 5116**: Establishing guidelines for government procurement and use of automated decision systems in order to protect consumers, improve transparency, and create more market predictability.
- **Massachusetts House Bill 4512**: Establishing a commission on automated decision-making by government in the commonwealth
- **Illinois House Bill 1811**: States predictive data analytics determining creditworthiness or hiring decisions may not include information that correlates with the applicant race or zip code

Source: AAAI 2020 Tutorial on Explainable AI

Tech giants are embracing XAI

Microsoft

- Model interpretability in Azure Machine Learning
- Fairlearn: A toolkit for assessing and improving fairness in AI

Google

- Explainable AI on Google Cloud
- The What-If Tool: Code-free probing of ML models

Facebook

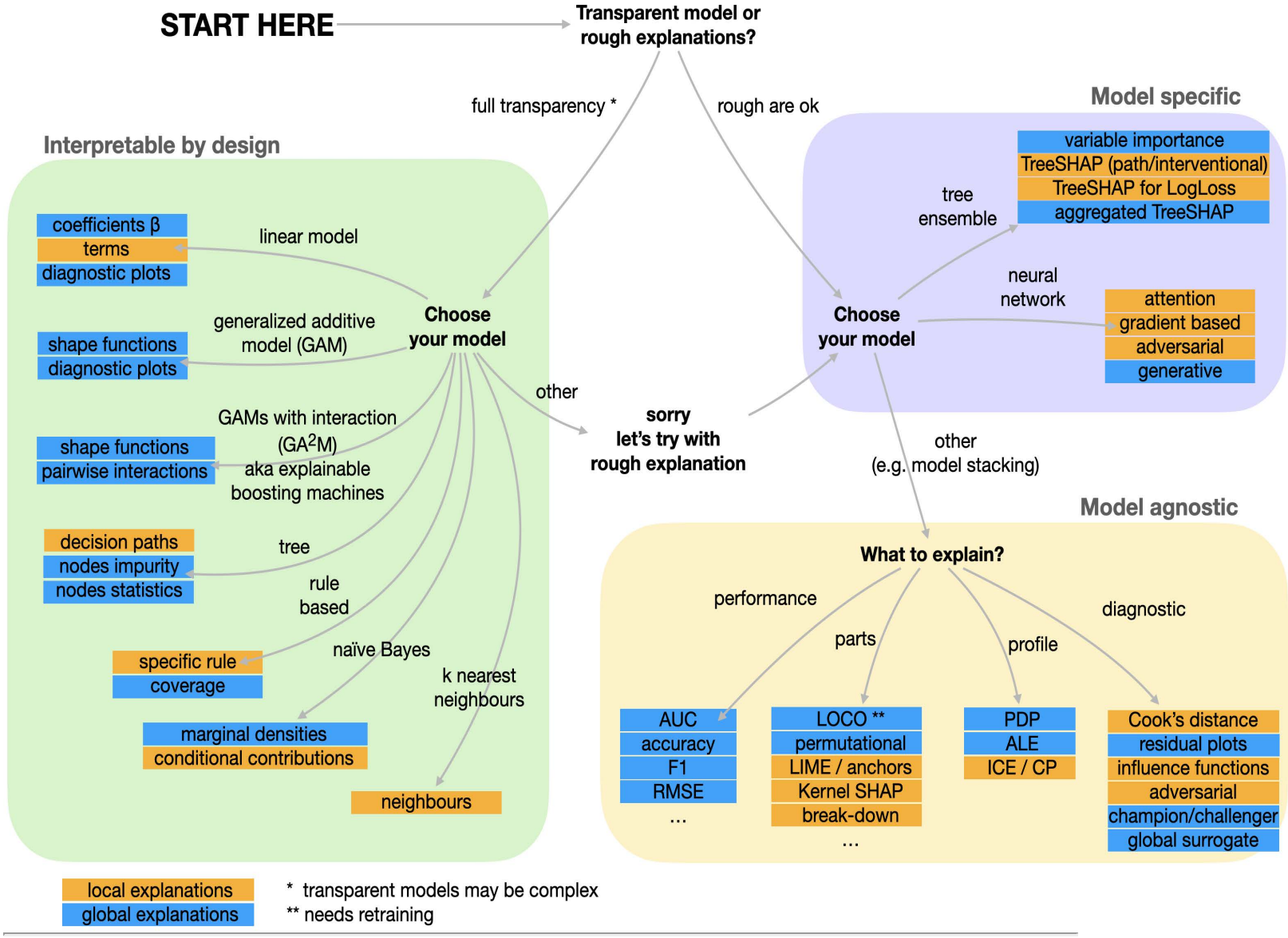
- Captum - A model interpretability library for PyTorch
- Introducing explainability to ad and news feed algorithms

IBM

- AI Explainability 360

Taxonomies of explainability methods

- **Intrinsic or post hoc?**
  - This criteria distinguishes whether interpretability is achieved by restricting the complexity of the ML model (intrinsic) or by applying methods that analyze the model after training (post hoc).
- **Model-specific or model-agnostic?**
  - Model-specific interpretation tools are limited to specific model classes, e.g., xgboostExplainer, randomForestExplainer, and DeepLIFT
  - Model-agnostic tools can be used on any ML model and are applied after the model has been trained (post hoc). These agnostic methods usually work by analyzing feature input and output pairs. By definition, these methods cannot have access to model internals such as weights or structural information.
- **Local or global?**
  - Does the interpretation method explain an individual prediction or the entire model behavior? Or is the scope somewhere in between?



Source: Maksymiuk, S., Gosiewska, A., & Biecek, P. (2020). http://128.84.4.34/abs/2009.13248.