

UNIVERSITY OF OSLO

Uncertainty Quantification in Explainable AI

PhD Trial Lecture

Anders Dahl Hjort

February 28th, 2025



Black box machine learning

Machine learning (ML) are increasingly applied to make [high-stakes decisions](#):

- Does this person likely have a life-threatening illness based on an MR scan?
- Should this person be granted a loan?
- Will this prisoner likely re-offend if released?

Black box machine learning

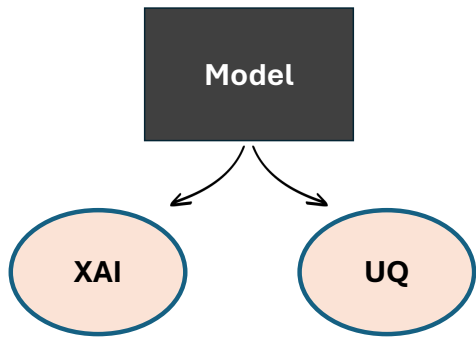
Machine learning (ML) are increasingly applied to make **high-stakes decisions**:

- Does this person likely have a life-threatening illness based on an MR scan?
- Should this person be granted a loan?
- Will this prisoner likely re-offend if released?

Many ML models are too big and complex to comprehend for humans

⇒ **Why** and **when** should we trust them?

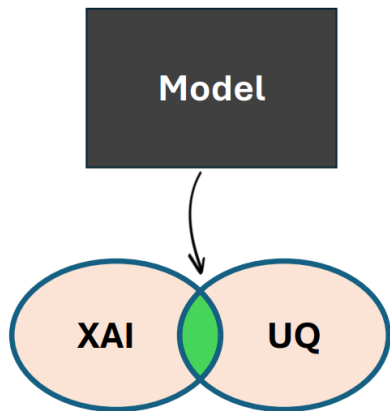
Black box ML has inspired new areas of research



Explainable AI:
*Why did the model give
me this prediction?*

Uncertainty Quantification:
*How certain is the model in
this prediction?*

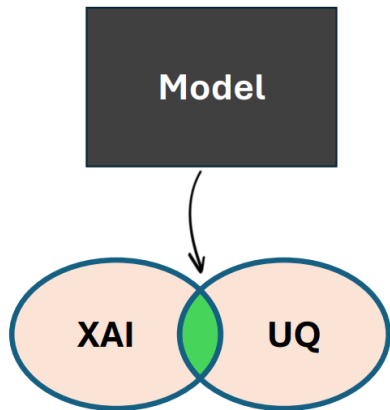
Combining XAI and UQ



Two questions:

- **Explaining the uncertainty:**
Why is the model uncertain?
- **Uncertainty in the explanation:**
How certain are we about the explanation?

Combining XAI and UQ



Two questions:

- **Explaining the uncertainty:**
Why is the model uncertain?
- **Uncertainty in the explanation:**
How certain are we about the explanation?

Agenda

- 1 Explainable AI
- 2 Quantifying uncertainty in Explainable AI
 - XAI model uncertainty
 - Prediction uncertainty
 - Model class uncertainty
- 3 Conclusion

Table of Contents

- 1 Explainable AI
- 2 Quantifying uncertainty in Explainable AI
 - XAI model uncertainty
 - Prediction uncertainty
 - Model class uncertainty
- 3 Conclusion

Context

- Supervised learning setting with $\mathbf{x} \in \mathbb{R}^d$ covariates, $y \in \mathbb{R}$ a response (or $y \in [1, \dots, K]$ if classification), where $\hat{f}$ is a machine learning model, for instance:
 - A neural network for image classification
 - A random forest or gradient boosted tree for prediction with tabular data
 - A mixture of different models

Context

- Supervised learning setting with $\mathbf{x} \in \mathbb{R}^d$ covariates, $y \in \mathbb{R}$ a response (or $y \in [1, \dots, K]$ if classification), where $\hat{f}$ is a machine learning model, for instance:
 - A neural network for image classification
 - A random forest or gradient boosted tree for prediction with tabular data
 - A mixture of different models
- **✗ Challenge:** Most of these models are not intrinsically interpretable, how can we trust them?
- **✓ Solution:** Explainable AI!

XAI taxonomies

- **Global explanation** or **local explanation**?
 - *"Feature x_1 is most important for the entire model"*
 - *"The most important factor for your prediction was that $x_3 = 10$ "*

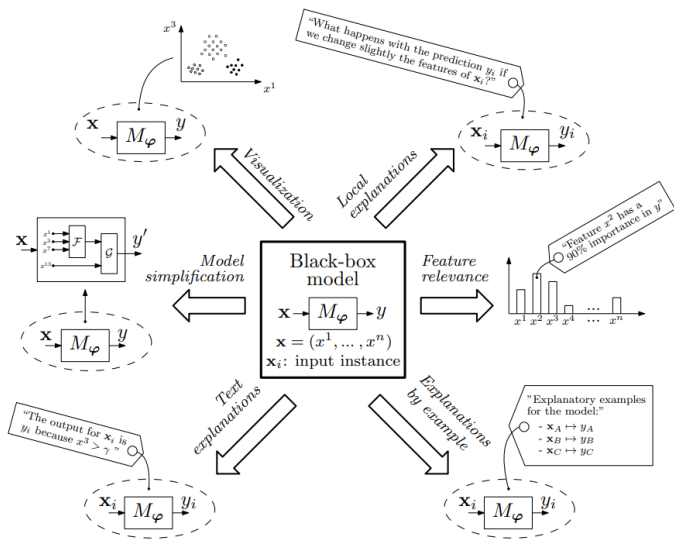
XAI taxonomies

- **Global explanation** or **local explanation**?
 - *"Feature x_1 is most important for the entire model"*
 - *"The most important factor for your prediction was that $x_3 = 10$ "*
- **Model-agnostic** or **model-specific**?

XAI taxonomies

- **Global explanation** or **local explanation**?
 - *"Feature x_1 is most important for the entire model"*
 - *"The most important factor for your prediction was that $x_3 = 10$ "*
- **Model-agnostic** or **model-specific**?
- **Factual** or **counterfactual** explanation?
 - *"The prediction would change by this much if the features changed by that much"*

Overview of methods (Barredo Arrieta et al. 2020)



Example: COMPAS data set

COMPAS: Correctional Offender Management Profiling for Alternative Sanctions

- Proprietary algorithm used to estimate the likelihood of **recidivism** (re-offending) among criminal defendants in Broward County, Florida, USA.

Example: COMPAS data set

COMPAS: Correctional Offender Management Profiling for Alternative Sanctions

- Proprietary algorithm used to estimate the likelihood of **recidivism** (re-offending) among criminal defendants in Broward County, Florida, USA.
- Each row consists demographic information, criminal history, and a two-year recidivism outcome.

Example: COMPAS data set

COMPAS: Correctional Offender Management Profiling for Alternative Sanctions

- Proprietary algorithm used to estimate the likelihood of **recidivism** (re-offending) among criminal defendants in Broward County, Florida, USA.
- Each row consists demographic information, criminal history, and a two-year recidivism outcome.
- Frequently studied in XAI research and controversial due to an analysis by ProPublica showing that it overestimates recidivism risk for black defendants (Larson et al. 2016).

What contributes to a high risk?

Example: COMPAS data set

COMPAS: Correctional Offender Management Profiling for Alternative Sanctions

- Proprietary algorithm used to estimate the likelihood of **recidivism** (re-offending) among criminal defendants in Broward County, Florida, USA.
- Each row consists demographic information, criminal history, and a two-year recidivism outcome.
- Frequently studied in XAI research and controversial due to an analysis by ProPublica showing that it overestimates recidivism risk for black defendants (Larson et al. 2016).

What contributes to a high risk?



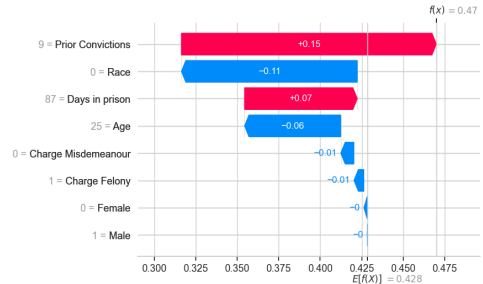
A **global** explanation for a random forest model trained on the COMPAS data set.

Example: COMPAS data set

COMPAS: Correctional Offender Management Profiling for Alternative Sanctions

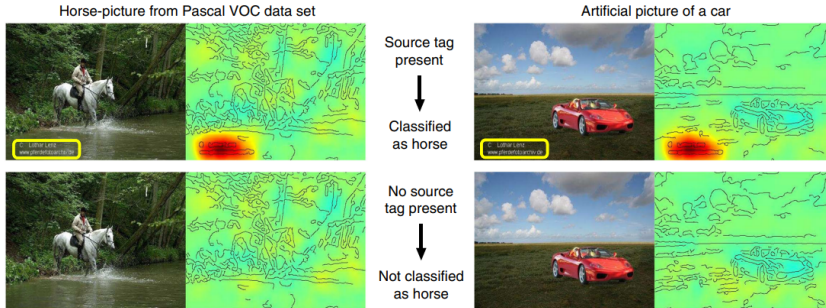
- Proprietary algorithm used to estimate the likelihood of **recidivism** (re-offending) among criminal defendants in Broward County, Florida, USA.
- Each row consists demographic information, criminal history, and a two-year recidivism outcome.
- Frequently studied in XAI research and controversial due to an analysis by ProPublica showing that it overestimates recidivism risk for black defendants (Larson et al. 2016).

What contributes to a high risk?



A **local** explanation for a prediction for one prisoner, based on SHAP values (Lundberg and Lee 2017). Red indicates feature contribution towards high risk, blue indicates contribution towards low risk.

XAI helps us identify artifacts in the ML model



The “Clever Hans effect” from Lapuschkin et al. 2019

... but XAI also have some challenges

- **Hard to evaluate:** How to validate the explanations?

... but XAI also have some challenges

- **Hard to evaluate:** How to validate the explanations?
- **Deviating explanations:** Different methods give vastly different explanations, which to trust?

... but XAI also have some challenges

- **Hard to evaluate:** How to validate the explanations?
- **Deviating explanations:** Different methods give vastly different explanations, which to trust?
- **Inconsistency:** Same methods give different explanations when repeated (Visani et al. 2022)

... but XAI also have some challenges

- **Hard to evaluate:** How to validate the explanations?
- **Deviating explanations:** Different methods give vastly different explanations, which to trust?
- **Inconsistency:** Same methods give different explanations when repeated (Visani et al. 2022)
- **Feature dependencies:** Some XAI frameworks are known to struggle with dependent features (Aas et al. 2021)

... but XAI also have some challenges

- **Hard to evaluate:** How to validate the explanations?
- **Deviating explanations:** Different methods give vastly different explanations, which to trust?
- **Inconsistency:** Same methods give different explanations when repeated (Visani et al. 2022)
- **Feature dependencies:** Some XAI frameworks are known to struggle with dependent features (Aas et al. 2021)
- **Easy to fool:** XAI methods are prone to adversarial attacks (Baniecki and Biecek 2024), and it is easy to fool humans with false explanations from XAI (Lakkaraju and Bastani 2020)

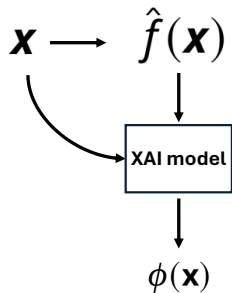
So how certain are we about this?



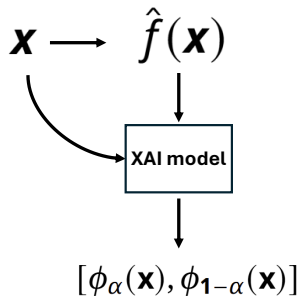
Table of Contents

- 1 Explainable AI
- 2 Quantifying uncertainty in Explainable AI**
 - XAI model uncertainty
 - Prediction uncertainty
 - Model class uncertainty
- 3 Conclusion

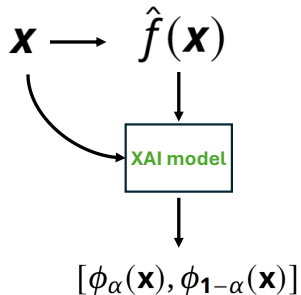
Framework



Desired framework

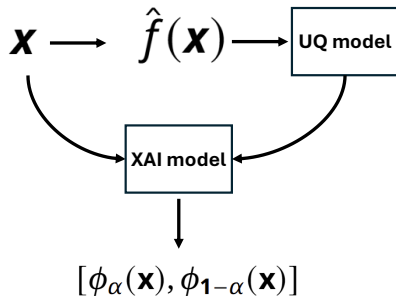


Types of uncertainty



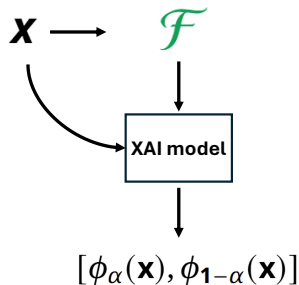
1. Uncertainty about the XAI model

Types of uncertainty



1. Uncertainty about the XAI model
2. Uncertainty about the prediction

Types of uncertainty



1. Uncertainty about the XAI model
2. Uncertainty about the prediction
3. Uncertainty about the model class

Table of Contents

- 1 Explainable AI
- 2 Quantifying uncertainty in Explainable AI
 - XAI model uncertainty
 - Prediction uncertainty
 - Model class uncertainty
- 3 Conclusion

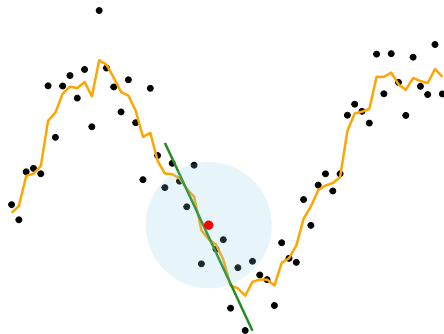
Inconsistency of LIME

LIME: Locally Interpretable Model-Agnostic Explanations (Ribeiro et al. 2016)

- A linear **surrogate model** that approximates the **black box** locally around a reference point x :

$$\min_{\phi} \left\{ \sum_z \pi_x(z) \left(\hat{f}(z) - \phi^\top z \right)^2 \right\}$$

- Sample points in neighborhood (z) , weight by distance to reference $(\pi_x(z))$



Inconsistency of LIME

LIME: Locally Interpretable Model-Agnostic Explanations (Ribeiro et al. 2016)

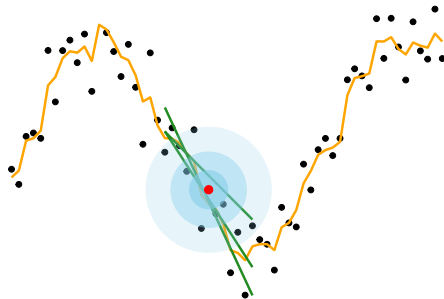
- A linear **surrogate model** that approximates the **black box** locally around a reference point x :

$$\min_{\phi} \left\{ \sum_z \pi_x(z) \left(\hat{f}(z) - \phi^\top z \right)^2 \right\}$$

- Sample points in neighborhood (z) , weight by distance to reference $(\pi_x(z))$

Challenge:

- How many points to sample?
- What is meant by “locally”?



Bayesian LIME

Bayesian LIME (Slack et al. 2021; Zhao et al. 2020):

Assume that

$$\hat{f}(z) = \phi^\top z + \varepsilon$$

with

$$\phi \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$$

$$\sigma^2 \sim \text{Inv-}\chi^2(n_0, \sigma_0^2)$$

$$\varepsilon | \sigma^2 \sim \mathcal{N}\left(0, \frac{\sigma^2}{\pi_x(z)}\right)$$

$\implies$ obtain the **posterior LIME distribution**
 $p(\phi | \text{data})$

Bayesian LIME

Bayesian LIME (Slack et al. 2021; Zhao et al. 2020):

Assume that

$$\hat{f}(z) = \phi^\top z + \varepsilon$$

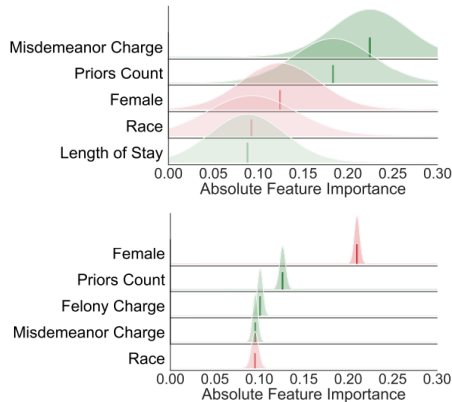
with

$$\phi \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$$

$$\sigma^2 \sim \text{Inv-}\chi^2(n_0, \sigma_0^2)$$

$$\epsilon | \sigma^2 \sim \mathcal{N}\left(0, \frac{\sigma^2}{\pi_x(z)}\right)$$

$\implies$ obtain the **posterior LIME distribution**
 $p(\phi | \text{data})$



The posterior LIME distribution $p(\phi | \text{data})$ for the COMPAS data set with $N = 100$ and $N = 2\,000$ samples.

Bayesian LIME: Why is it useful?

- Sample from the posterior distribution of ϕ to get a **credible interval**
 - More samples $\implies$ more certain in our explanations

Bayesian LIME: Why is it useful?

- Sample from the posterior distribution of ϕ to get a **credible interval**
 - More samples $\implies$ more certain in our explanations
- Can be used to determine the **number of samples**
 - "In order to reduce the posterior variance sufficiently, we need $N = 123$ samples."

Bayesian LIME: Why is it useful?

- Sample from the posterior distribution of ϕ to get a **credible interval**
 - More samples $\implies$ more certain in our explanations
- Can be used to determine the **number of samples**
 - "In order to reduce the posterior variance sufficiently, we need $N = 123$ samples."
- Can be used to do **focused sampling**:
 - Sample points strategically where the surrogate model struggle to mimic the black box

Table of Contents

- 1 Explainable AI
- 2 Quantifying uncertainty in Explainable AI
 - XAI model uncertainty
 - Prediction uncertainty
 - Model class uncertainty
- 3 Conclusion

Motivation

A binary classification model predicts “High risk” with probability $p = 0.8$.

- **Is the model calibrated?** Is the model correct 80% of the times it predicts $p = 0.8$?
 - If not: Why explain? (Scafarto et al. 2022; H. Löfström et al. 2023)

Motivation

A binary classification model predicts “High risk” with probability $p = 0.8$.

- **Is the model calibrated?** Is the model correct 80% of the times it predicts $p = 0.8$?
 - If not: Why explain? (Scafarto et al. 2022; H. Löfström et al. 2023)
- Methods exist for **post-hoc calibration**
- Some calibration methods (e.g., Venn predictors, Vovk et al. 2004) return a calibrated **probability interval** (PI) for each class, e.g. $[0.78, 0.82]$

Motivation

A binary classification model predicts “High risk” with probability $p = 0.8$.

- **Is the model calibrated?** Is the model correct 80% of the times it predicts $p = 0.8$?
 - If not: Why explain? (Scafarto et al. 2022; H. Löfström et al. 2023)
- Methods exist for **post-hoc calibration**
- Some calibration methods (e.g., Venn predictors, Vovk et al. 2004) return a calibrated **probability interval** (PI) for each class, e.g. $[0.78, 0.82]$

Can we utilize this to reason about the confidence in the explanation?

Calibrated local explanations

Idea: Measure the effect of feature permutation on PIs (T. Löffström et al. 2023; Yapicioglu et al. 2024)

Calibrated local explanations

Idea: Measure the effect of feature permutation on PIs (T. Löfström et al. 2023; Yapicioglu et al. 2024)

1. Use the black box model to get a class probability p
2. Get a calibrated prediction p_{cal} and a PI $[p_{0,\text{cal}}, p_{1,\text{cal}}]$ for each class

Calibrated local explanations

Idea: Measure the effect of feature permutation on PIs (T. Löfström et al. 2023; Yapicioglu et al. 2024)

1. Use the black box model to get a class probability p
2. Get a calibrated prediction p_{cal} and a PI $[p_{0,\text{cal}}, p_{1,\text{cal}}]$ for each class
3. Perturb x_j slightly

Calibrated local explanations

Idea: Measure the effect of feature permutation on PIs (T. Löffström et al. 2023; Yapicioglu et al. 2024)

1. Use the black box model to get a class probability p
2. Get a calibrated prediction p_{cal} and a PI $[p_{0,\text{cal}}, p_{1,\text{cal}}]$ for each class
3. Perturb x_j slightly
4. Predict again, calibrate and get PI

Calibrated local explanations

Idea: Measure the effect of feature permutation on PIs (T. Löfström et al. 2023; Yapicioglu et al. 2024)

1. Use the black box model to get a class probability p
2. Get a calibrated prediction p_{cal} and a PI $[p_{0,\text{cal}}, p_{1,\text{cal}}]$ for each class
3. Perturb x_j slightly
4. Predict again, calibrate and get PI
5. $\phi(x_j)$: The change in the (calibrated) prediction

Calibrated local explanations

Idea: Measure the effect of feature permutation on PIs (T. Löfström et al. 2023; Yapicioglu et al. 2024)

1. Use the black box model to get a class probability p
2. Get a calibrated prediction p_{cal} and a PI $[p_{0,\text{cal}}, p_{1,\text{cal}}]$ for each class
3. Perturb x_j slightly
4. Predict again, calibrate and get PI
5. $\phi(x_j)$: The change in the (calibrated) prediction
6. Confidence set for $\phi(x_j)$: The change in the calibrated PI

Calibrated local explanations

Idea: Measure the effect of feature permutation on PIs (T. Löfström et al. 2023; Yapicioglu et al. 2024)

1. Use the black box model to get a class probability p
2. Get a calibrated prediction p_{cal} and a PI $[p_{0,\text{cal}}, p_{1,\text{cal}}]$ for each class
3. Perturb x_j slightly
4. Predict again, calibrate and get PI
5. $\phi(x_j)$: The change in the (calibrated) prediction
6. Confidence set for $\phi(x_j)$: The change in the calibrated PI

Summarized: Feature j is important if perturbing it leads to large changes in the PI

Example: Calibrated explanation

Regular $\mathbf{x}$:

$$\hat{f}(\mathbf{x}) \implies p = 0.8 \implies p_{\text{cal}} = 0.83 \implies [p_{0,\text{cal}}, p_{1,\text{cal}}] = [0.80, 0.86]$$

Example: Calibrated explanation

Regular $\mathbf{x}$:

$$\hat{f}(\mathbf{x}) \implies p = 0.8 \implies p_{\text{cal}} = 0.83 \implies [p_{0,\text{cal}}, p_{1,\text{cal}}] = [0.80, 0.86]$$

Perturb feature j slightly:

$$\hat{f}(\mathbf{x}^*) \implies p = 0.75 \implies p_{\text{cal}} = 0.77 \implies [p_{0,\text{cal}}, p_{1,\text{cal}}] = [0.75, 0.79]$$

Example: Calibrated explanation

Regular $\mathbf{x}$:

$$\hat{f}(\mathbf{x}) \implies p = 0.8 \implies p_{\text{cal}} = 0.83 \implies [p_{0,\text{cal}}, p_{1,\text{cal}}] = [0.80, 0.86]$$

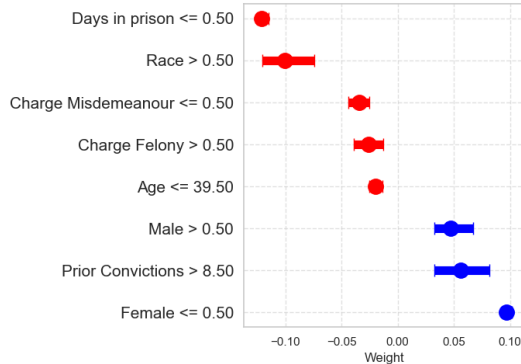
Perturb feature j slightly:

$$\hat{f}(\mathbf{x}^*) \implies p = 0.75 \implies p_{\text{cal}} = 0.77 \implies [p_{0,\text{cal}}, p_{1,\text{cal}}] = [0.75, 0.79]$$

Feature importance:

$$\begin{aligned} \phi(x_j) &= 0.83 - 0.77 = 0.06 \\ [\phi_\alpha(x_j), \phi_{1-\alpha}(x_j)] &= \{0.83 - 0.79, 0.83 - 0.75\} = \{0.04, 0.08\} \end{aligned}$$

Calibrated explanation



Running a local explanation with the calibrated-explanation (T. Löffström et al. 2023) package on the COMPAS data set.

Open questions

- ✓ Promising approach because it relies on **model-agnostic** calibration techniques
- ✗ Currently limited to specific measures of feature importance

Open question: How to extend this to popular off-the-shelf XAI tools like SHAP and LIME?

Table of Contents

- 1 Explainable AI
- 2 Quantifying uncertainty in Explainable AI
 - XAI model uncertainty
 - Prediction uncertainty
 - Model class uncertainty
- 3 Conclusion

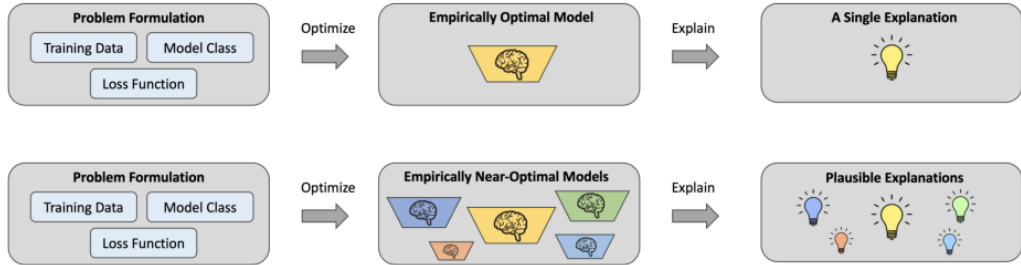
Rashomon effect



“Statistical Modeling: The Two Cultures” (Breiman 2001):

*Rashomon is a wonderful Japanese movie in which four people, from different vantage points, witness an incident in which one person dies and another is supposedly raped. When they come to testify in court, they all report the same facts, but their stories of what happened are very different. What I call the **Rashomon Effect** is that there is often a multitude of different descriptions in a class of functions giving about the same minimum error rate.*

Rashomon effect in XAI



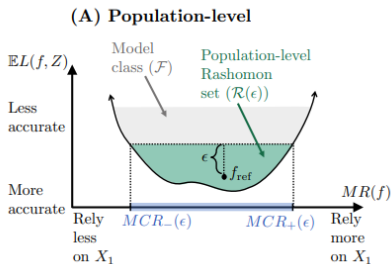
The approach presented in Marx et al. 2023.

How to define the set of near-optimal models?

- Fisher et al. 2019: “Rashomon set” is defined by

$$\mathcal{R}_\epsilon = \left\{ f \in \mathcal{F} : \widehat{\text{Loss}}(f) \leq \widehat{\text{Loss}}(f_{\text{ref}}) + \epsilon \right\}$$

= “All models that are reasonably close to the best model”



From Fisher et al. 2019

How to define the set of near-optimal models?

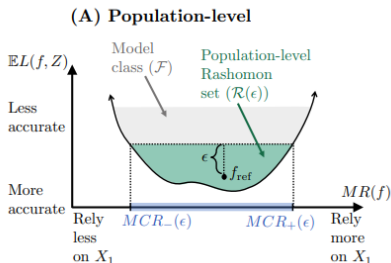
- Fisher et al. 2019: “Rashomon set” is defined by

$$\mathcal{R}_\epsilon = \left\{ f \in \mathcal{F} : \widehat{\text{Loss}}(f) \leq \widehat{\text{Loss}}(f_{\text{ref}}) + \epsilon \right\}$$

= “All models that are reasonably close to the best model”

The **Model Reliance** (MR) on x_1 is

$$\text{MR}(x_1; f) = \frac{\text{Loss of } f \text{ when } x_1 \text{ is perturbed}}{\text{Loss of } f \text{ when } x_1 \text{ is **not** perturbed}}$$



From Fisher et al. 2019

How to define the set of near-optimal models?

- Fisher et al. 2019: “Rashomon set” is defined by

$$\mathcal{R}_\epsilon = \left\{ f \in \mathcal{F} : \widehat{\text{Loss}}(f) \leq \widehat{\text{Loss}}(f_{\text{ref}}) + \epsilon \right\}$$

= “All models that are reasonably close to the best model”

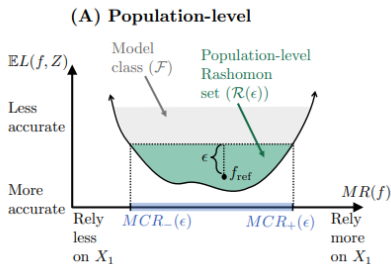
The **Model Reliance** (MR) on x_1 is

$$\text{MR}(x_1; f) = \frac{\text{Loss of } f \text{ when } x_1 \text{ is perturbed}}{\text{Loss of } f \text{ when } x_1 \text{ is **not** perturbed}}$$

and the **Model Class Reliance** is thus

$$\text{MRC}(x_1; \mathcal{R}_\epsilon) = \left[\min_{f \in \mathcal{R}_\epsilon} \text{MR}(x_1; f), \max_{f \in \mathcal{R}_\epsilon} \text{MR}(x_1; f) \right]$$

- Breiman 2001; Marx et al. 2020: **the 1% set**



From Fisher et al. 2019

Model Class Reliance for random forests



Model Class Reliance for random forests at two different ϵ levels on the COMPAS data set. The plot is based on results in Smith et al. 2020.

Table of Contents

- 1 Explainable AI
- 2 Quantifying uncertainty in Explainable AI
 - XAI model uncertainty
 - Prediction uncertainty
 - Model class uncertainty
- 3 Conclusion

Conclusion

- XAI opens up the black box of machine learning, but comes with several challenges
- Adding uncertainty quantification to the XAI methods makes the explanations more trustworthy and useful
- Three approaches studied:
 1. **Bayesian LIME values**
 2. **Calibrated intervals $\implies$ XAI intervals**
 3. **Model class reliance**

Are we adding too many layers?

Ribeiro et al. 2016: XAI methods must “take into account the user’s limitations”.

Are we adding too many layers?

Ribeiro et al. 2016: XAI methods must “take into account the user’s limitations”.

- *“The contribution of x_1 to $\hat{y}$ was in the interval $[1.2, 1.6]$ with 70% confidence.”*
- *“If we changed x_1 by ϵ , we would be 75% confident instead.”*

PERSPECTIVE

<https://doi.org/10.1038/s42256-019-0048-x>nature
machine intelligence

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin 

Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that are inherently interpretable. This Perspective clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare and computer vision.

PERSPECTIVE

<https://doi.org/10.1038/s42256-019-0048-x>nature
machine intelligence

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that are inherently interpretable. This Perspective clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare and computer vision.

Stop explaining black box machine learning models for high stakes decisions **without quantifying the uncertainty!**

References I

-  Aas, Kjersti, Martin Jullum and Anders Løland (2021). 'Explaining individual predictions when features are dependent: More accurate approximations to Shapley values'. In: *Artificial Intelligence* 298, p. 103502.
-  Baniecki, Hubert and Przemyslaw Biecek (2024). 'Adversarial attacks and defenses in explainable artificial intelligence: A survey'. In: *Information Fusion* 107, p. 102303.
-  Barredo Arrieta, Alejandro et al. (2020). 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI'. In: *Information Fusion* 58, pp. 82–115.
-  Breiman, Leo (2001). 'Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)'. In: *Statistical Science* 16.3, pp. 199–231.

References II

-  Fisher, Aaron, Cynthia Rudin and Francesca Dominici (2019). 'All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously'. In: *Journal of Machine Learning Research* 20.
-  Lakkaraju, Himabindu and Osbert Bastani (2020). "'How do I fool you?': Manipulating User Trust via Misleading Black Box Explanations'. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. AIES '20, pp. 79–85.
-  Lapuschkin, Sebastian, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek and Klaus-Robert Müller (2019). 'Unmasking Clever Hans Predictors and Assessing What Machines Really Learn'. In: *Nature Communications* 10.

References III

-  Larson, Jeff, Surya Mattu, Lauren Kirchner and Julia Angwin (2016). *How We Analyzed the COMPAS Recidivism Algorithm*. URL: <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm> (visited on 23/02/2025).
-  Löffström, Helena, Tuwe Löffström, Ulf Johansson and Cecilia Sönströd (2023). 'Investigating the impact of calibration on the quality of explanations'. In: *Annals of Mathematics and Artificial Intelligence*.
-  Löffström, Tuwe, Helena Löffström, Ulf Johansson and Cecilia Sönströd (2023). *Calibrated Explanations for Regression*.
-  Lundberg, Scott M and Su-In Lee (2017). 'A Unified Approach to Interpreting Model Predictions'. In: *Advances in Neural Information Processing Systems*. Vol. 30, pp. 4765–4774.

References IV

-  Marx, Charles, Flavio Calmon and Berk Ustun (2020). 'Predictive Multiplicity in Classification'. In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research, pp. 6765–6774.
-  Marx, Charles, Youngsuk Park, Hilaf Hasson, Yuyang Wang, Stefano Ermon and Luke Huan (2023). 'But Are You Sure? An Uncertainty-Aware Perspective on Explainable AI'. In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. Vol. 206. Proceedings of Machine Learning Research, pp. 7375–7391.
-  Ribeiro, Marco Tulio, Sameer Singh and Carlos Guestrin (2016). "'Why Should I Trust You?': Explaining the Predictions of Any Classifier". In: *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16, pp. 1135–1144.

References V

-  Scafarto, Gregory, Nicolas Posocco and Antoine Bonnefoy (2022). *Calibrate to Interpret*. arXiv: 2207.03324 [cs.LG]. URL: <https://arxiv.org/abs/2207.03324>.
-  Slack, Dylan, Sophie Hilgard, Sameer Singh and Himabindu Lakkaraju (2021). 'Reliable Post hoc Explanations Modeling Uncertainty in Explainability'. In: *Neural Information Processing Systems (NeurIPS)*.
-  Smith, Gavin, Roberto Mansilla and James Goulding (Dec. 2020). 'Model Class Reliance for Random Forests'. In: *Advances in Neural Information Processing Systems*. Vol. 33.
-  Visani, Giorgio, Enrico Bagli, Federico Chesani, Alessandro Poluzzi and Davide Capuzzo (2022). 'Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models'. In: *Journal of the Operational Research Society* 73.1, pp. 91–101.

References VI

-  Vovk, Vladimir, Glenn Shafer and Ilia Nouretdinov (Mar. 2004). 'Self-calibrating Probability Forecasting'. In: *Advances In Neural Information Processing Systems* 16.
-  Yapicioglu, Fatima, Alessandra Stramiglio and Fabio Vitali (2024). 'ConformaSight: Conformal Prediction-Based Global and Model-Agnostic Explainability Framework'. In: pp. 270–293.
-  Zhao, Xingyu, Xiaowei Huang, V. Robu and David Flynn (2020). 'BayLIME: Bayesian Local Interpretable Model-Agnostic Explanations'. In: *Conference on Uncertainty in Artificial Intelligence*.

Theoretical guarantees

Marx et al. 2023:

- Create a confidence set of explanations from the Rashomon set of “near-optimal models”
- Can be used on any off-the-shelf explanation framework (e.g., SHAP or LIME)
- Validity guarantees:

$$P(\phi(f) \in \text{Confidence Set}) \geq 1 - \alpha$$

- Requirement: Must be able to sample models from the Rashomon set – not always trivial?

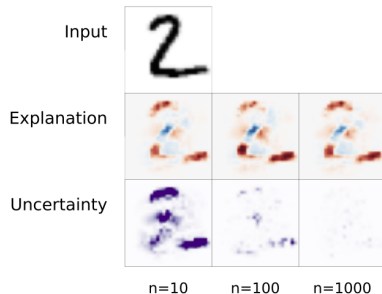
Theoretical guarantees

Marx et al. 2023:

- Create a confidence set of explanations from the Rashomon set of “near-optimal models”
- Can be used on any off-the-shelf explanation framework (e.g., SHAP or LIME)
- Validity guarantees:

$$P(\phi(f) \in \text{Confidence Set}) \geq 1 - \alpha$$

- Requirement: Must be able to sample models from the Rashomon set – not always trivial?



Marx et al. 2023: Uncertainty of a Deep SHAP explanation of a neural network classifier on the MNIST data set.