eXplainable Artificial Intelligence: A Literature Review

Alessandro Leite & Marc Schoenauer

What can AI do?

Play (and win) games

AlphaGo

DeepMind

Answer queries

IBM

Watson

Debate

Project Debater

IBM

Recognise speech

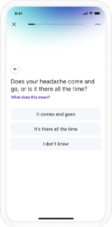
Hey Siri, call Mum

Recognise faces

facebook

Translate across languages

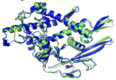
Google Translate



Detect & Diagnose Diseases

babylon

Predict protein structures



Drive vehicles

WAYMO

Vacuum clean

dyson

Why do we need explanation?

Explanations

- reflect an attempt to communicate an understanding[a]

[a] Frank C Keil. "Explanation and understanding". In: *Annu. Rev. Psychol.* 57 (2006), pp. 227–254.
[b] Tania Lombrozo. "Explanation and abductive inference". In: (2012).

Why do we need explanation?

Explanations

- reflect an attempt to communicate an understanding[a]
- create trajectories, expanding individuals' understanding in real-time

[a] Frank C Keil. "Explanation and understanding". In: *Annu. Rev. Psychol.* 57 (2006), pp. 227–254.
[b] Tania Lombrozo. "Explanation and abductive inference". In: (2012).

Why do we need explanation?

Explanations

- reflect an attempt to communicate an understanding[a]
- create trajectories, expanding individuals' understanding in real-time
- may highlight incompleteness

[a] Frank C Keil. "Explanation and understanding". In: *Annu. Rev. Psychol.* 57 (2006), pp. 227–254.
[b] Tania Lombrozo. "Explanation and abductive inference". In: (2012).

Why do we need explanation?

Explanations

- reflect an attempt to communicate an understanding[a]
- create trajectories, expanding individuals' understanding in real-time
- may highlight incompleteness
- may provide common sense mechanisms

[a] Frank C Keil. "Explanation and understanding". In: *Annu. Rev. Psychol.* 57 (2006), pp. 227–254.
[b] Tania Lombrozo. "Explanation and abductive inference". In: (2012).

Why do we need explanation?

Explanations

- reflect an attempt to communicate an understanding[a]
- create trajectories, expanding individuals' understanding in real-time
- may highlight incompleteness
- may provide common sense mechanisms
- relate the event being explained to principles, invoking causal relations[b]

[a] Frank C Keil. "Explanation and understanding". In: *Annu. Rev. Psychol.* 57 (2006), pp. 227–254.
[b] Tania Lombrozo. "Explanation and abductive inference". In: (2012).

Why do we need explanation?

Explanations

- reflect an attempt to communicate an understanding[a]
- create trajectories, expanding individuals' understanding in real-time
- may highlight incompleteness
- may provide common sense mechanisms
- relate the event being explained to principles, invoking causal relations[b]
- answer a "why question" justifying an event

[a] Frank C Keil. "Explanation and understanding". In: *Annu. Rev. Psychol.* 57 (2006), pp. 227–254.
[b] Tania Lombrozo. "Explanation and abductive inference". In: (2012).

Prediction is the most common reason for explanation[1]

[1] Fritz Heider. *The psychology of interpersonal relations*. Wiley, 1958.

Explainable AI

Interpretability

- It describes the internals of a system in a way that is understandable to humans[a]

[a] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning". In: *arXiv:1702.08608* (2017).

Explainable AI

Interpretability

- It describes the internals of a system in a way that is understandable to humans[a]
- It must employ a vocabulary that is meaningful for a human observer

[a] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning". In: *arXiv:1702.08608* (2017).

Explainable AI

Interpretability

- It describes the internals of a system in a way that is understandable to humans[a]
- It must employ a vocabulary that is meaningful for a human observer

[a] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning". In: *arXiv:1702.08608* (2017).

Explainability

- A characteristic of a model, agnostic w.r.t. the type of model

[a] Alejandro Barredo Arrieta et al. "Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115.

Explainable AI

Interpretability

- It describes the internals of a system in a way that is understandable to humans[a]
- It must employ a vocabulary that is meaningful for a human observer

[a] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning". In: *arXiv:1702.08608* (2017).

Explainability

- A characteristic of a model, agnostic w.r.t. the type of model
- Provide the reasons for the behavior of a given machine learning model[a]

[a] Alejandro Barredo Arrieta et al. "Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115.

Explainable AI

Interpretability

- It describes the internals of a system in a way that is understandable to humans[a]
- It must employ a vocabulary that is meaningful for a human observer

[a] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning". In: *arXiv:1702.08608* (2017).

Explainability

- A characteristic of a model, agnostic w.r.t. the type of model
- Provide the reasons for the behavior of a given machine learning model[a]
- Any action taken with the intent of providing an explanation of a model to a human observer

[a] Alejandro Barredo Arrieta et al. "Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115.

XAI and the social sciences

“looking at how humans explain to each other can serve as a useful starting point for explanation in artificial intelligence” – Miller (2019)[2]

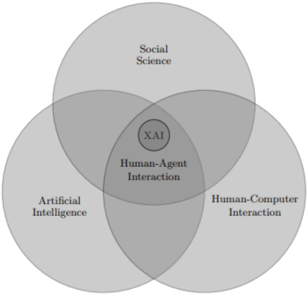


Figure 1: Scope of explainable AI

[2] Tim Miller. “Explanation in artificial intelligence: Insights from the social sciences”. In: *Artificial intelligence* 267 (2019), pp. 1–38,

Principles of explainable AI

Assumptions

- Human observers can query the AI system whenever they want

Principles of explainable AI

Assumptions

- Human observers can query the AI system whenever they want
- The output is the answer of a query

Principles of explainable AI

Assumptions

- Human observers can query the AI system whenever they want
- The output is the answer of a query
- Output varies by type of task

Principles of explainable AI

Assumptions

- Human observers can query the AI system whenever they want
- The output is the answer of a query
- Output varies by type of task
- Human observers have different knowledge and beliefs

Principles of explainable AI

Assumptions

- Human observers can query the AI system whenever they want
- The output is the answer of a query
- Output varies by type of task
- Human observers have different knowledge and beliefs
- The system knows (by some way) the profile of which human observer

Explanation

AI system provides evidences for each of its outputs

- The focus is on the capacity to **provide** an explanation, **not** on its:
  - validity
  - correctness
  - intelligibility
- No metric or evaluation
- Unaware of observers' profiles

Meaningful

AI system provides explanations that are understandable by the recipient

- How to evaluate the meaningfulness of an explanation?
  - the receipt can understand it (can be difficult to assess)
  - (s)he can use it to complete a task (utility, . . . , how to know?)
  - feedback loop (e.g., how useful was this explanation?)
  - Psychological differences influence how people interpret and judge how meaningful an explanation is
  - Meaningful changes as people's experiences evolve
- A receipt can represent groups (e.g., data scientists, developers, regulators, judges, . . .)
- System must know who is querying . . .
- Meaningful is influenced by receipt's knowledge, experiences, and mental process

Explanation accuracy

AI system's explanations correctly reflect system's process for generating the output

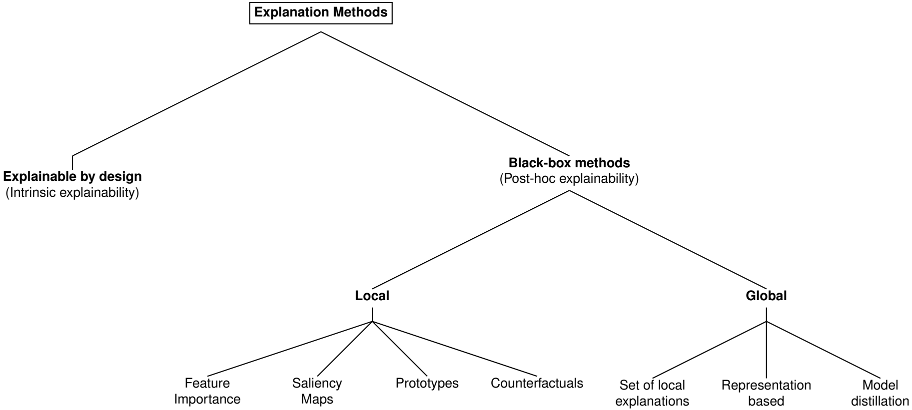
- It is:
  - observer-dependent
  - different from **decision accuracy**
  - measured accord to some pre-defined metrics (e.g., few works on this topic)
  - without overlap with the meaningful principle
- Explanation accuracy increases when the system can generate multiple types of explanations
- Generator/discriminator approach

Knowledge limits

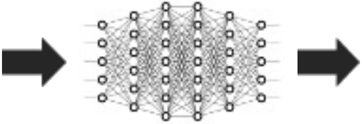
AI systems are aware of the **cases which they were not designed** or allowed to **operate on**, or on which their **answers** are **unreliable**

- The system includes in its explanations its confidence level (i.e., silence is not an answer)
- May prevent misleading, dangerous, outputs
- Need to be queried. Therefore, . . .
- It may change according to the query
  - Is there a bird in this photo?
  - What is the family of the bird in this photo?

Current explainable approaches

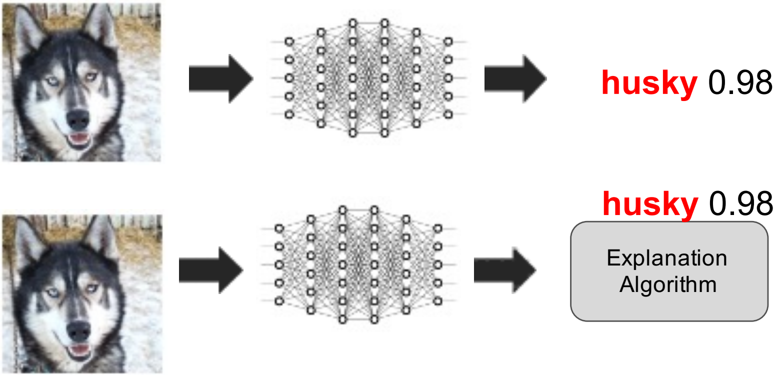


Post-hoc explainability

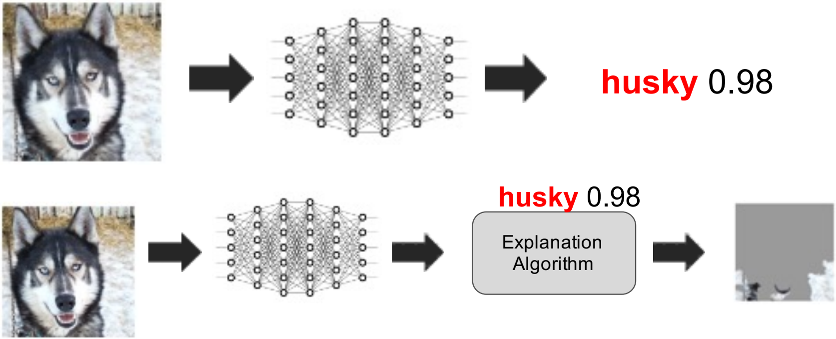


husky 0.98

Post-hoc explainability



Post-hoc explainability



Local vs global explanations

Local explanations

- Explain individual predictions

Global explanations

Local vs global explanations

Local explanations

- Explain individual predictions

Global explanations

- Explain the behavior of a model

Local vs global explanations

Local explanations

- Explain individual predictions
- Help in unearthing biases in the neighborhood of a given sample

Global explanations

- Explain the behavior of a model

Local vs global explanations

Local explanations

- Explain individual predictions
- Help in unearthing biases in the neighborhood of a given sample

Global explanations

- Explain the behavior of a model
- Highlight biases affecting larger subgroups

Local vs global explanations

Local explanations

- Explain individual predictions
- Help in unearthing biases in the neighborhood of a given sample
- Help in checking out if individual predictions are correctly being made

Global explanations

- Explain the behavior of a model
- Highlight biases affecting larger subgroups

Local vs global explanations

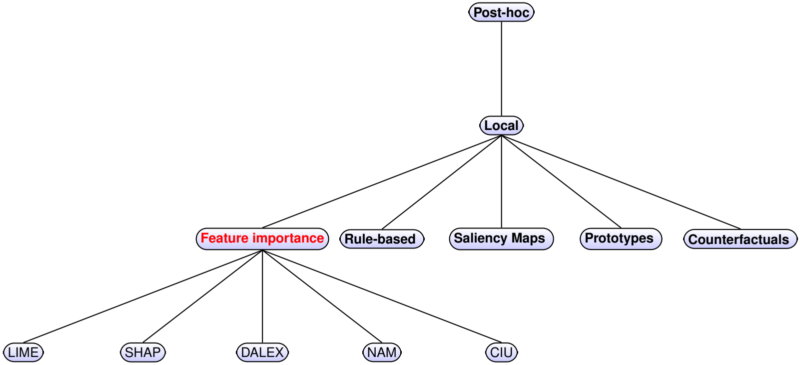
Local explanations

- Explain individual predictions
- Help in unearthing biases in the neighborhood of a given sample
- Help in checking out if individual predictions are correctly being made

Global explanations

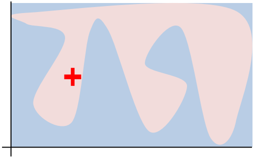
- Explain the behavior of a model
- Highlight biases affecting larger subgroups
- Help in determining if the model is in someway ready for deployment

Post-hoc explainability: feature importance methods



Local Interpretable Model-Agnostic Explanations (LIME)[3]

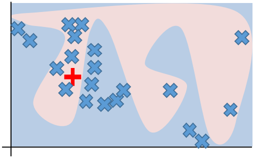
- Model agnostic explanation method based on feature importance



[3] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why should i trust you?" Explaining the predictions of any classifier". In: *22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144,

Local Interpretable Model-Agnostic Explanations (LIME)[3]

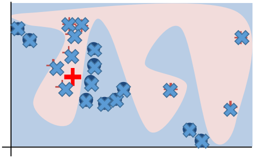
- Model agnostic explanation method based on feature importance
- Draw a perturbed sample of weighted instances $\{z \in \mathbb{R}^d\}$ around a point x_i by exploiting a proximity measure π_x



[3] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “”Why should i trust you?” Explaining the predictions of any classifier”. In: *22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144,

Local Interpretable Model-Agnostic Explanations (LIME)[3]

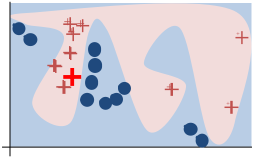
- Model agnostic explanation method based on feature importance
- Draw a perturbed sample of weighted instances $\{z \in \mathbb{R}^d\}$ around a point x_i by exploiting a proximity measure π_x
- Fed them to the black-box model $b(z)$ to predict the output for each sample



[3] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why should i trust you?" Explaining the predictions of any classifier". In: *22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144,

Local Interpretable Model-Agnostic Explanations (LIME)[3]

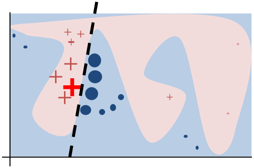
- Model agnostic explanation method based on feature importance
- Draw a perturbed sample of weighted instances $\{z \in \mathbb{R}^d\}$ around a point x_i by exploiting a proximity measure π_x
- Fed them to the black-box model $b(z)$ to predict the output for each sample
- Weights the samples according to the distance to x_i



[3] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why should i trust you?" Explaining the predictions of any classifier". In: *22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144,

Local Interpretable Model-Agnostic Explanations (LIME)[3]

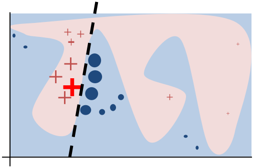
- Model agnostic explanation method based on feature importance
- Draw a perturbed sample of weighted instances $\{z \in \mathbb{R}^d\}$ around a point x_i by exploiting a proximity measure π_x
- Fed them to the black-box model $b(z)$ to predict the output for each sample
- Weights the samples according to the distance to x_i
- Train an explanation model $g(\cdot)$: sparse linear model on the weighted samples



[3] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why should i trust you?" Explaining the predictions of any classifier". In: *22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144,

Local Interpretable Model-Agnostic Explanations (LIME)[3]

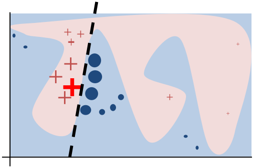
- Model agnostic explanation method based on feature importance
- Draw a perturbed sample of weighted instances $\{z \in \mathbb{R}^d\}$ around a point x_i by exploiting a proximity measure π_x
- Fed them to the black-box model $b(z)$ to predict the output for each sample
- Weights the samples according to the distance to x_i
- Train an explanation model $g(\cdot)$: sparse linear model on the weighted samples
- Use $g(\cdot)$ to explain



[3] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why should i trust you?" Explaining the predictions of any classifier". In: *22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144,

Local Interpretable Model-Agnostic Explanations (LIME)[3]

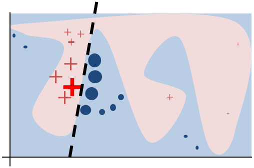
- Model agnostic explanation method based on feature importance
- Draw a perturbed sample of weighted instances $\{z \in \mathbb{R}^d\}$ around a point x_i by exploiting a proximity measure π_x
- Fed them to the black-box model $b(z)$ to predict the output for each sample
- Weights the samples according to the distance to x_i
- Train an explanation model $g(\cdot)$: sparse linear model on the weighted samples
- Use $g(\cdot)$ to explain
- The explanation are the weights of the linear model



[3] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why should i trust you?" Explaining the predictions of any classifier". In: *22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144,

Local Interpretable Model-Agnostic Explanations (LIME)[3]

- Model agnostic explanation method based on feature importance
- Draw a perturbed sample of weighted instances $\{z \in \mathbb{R}^d\}$ around a point x_i by exploiting a proximity measure π_x
- Fed them to the black-box model $b(z)$ to predict the output for each sample
- Weights the samples according to the distance to x_i
- Train an explanation model $g(\cdot)$: sparse linear model on the weighted samples
- Use $g(\cdot)$ to explain
- The explanation are the weights of the linear model
- There are various to overcome LIME's limitations: KL-LIME, DLIME, ILIME, ALIME



[3] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why should i trust you?" Explaining the predictions of any classifier". In: *22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144,

SHapley Additive exPlanations[4]

- Local and global model-agnostic explanation method

[4] Scott M Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, 4765–4774.

SHapley Additive exPlanations[4]

- Local and global model-agnostic explanation method
- Can be employed as a local or global explainer

[4] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, 4765–4774.

SHapley Additive exPlanations[4]

- Local and global model-agnostic explanation method
- Can be employed as a local or global explainer
- Can produce various additive feature attribution methods

$$g(z') = \phi_0 + \sum_{i}^{M} \phi_i z'_i$$

[4] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, 4765–4774.

SHapley Additive exPlanations[4]

- Local and global model-agnostic explanation method
- Can be employed as a local or global explainer
- Can produce various additive feature attribution methods

$$g(z') = \phi_0 + \sum_{i}^{M} \phi_i z'_i$$

- Main properties

[4] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, 4765–4774.

SHapley Additive exPlanations[4]

- Local and global model-agnostic explanation method
- Can be employed as a local or global explainer
- Can produce various additive feature attribution methods

$$g(z') = \phi_0 + \sum_{i}^{M} \phi_i z'_i$$

- Main properties
  - local accuracy: $g(x) = b(x)$

[4] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, 4765–4774.

SHapley Additive exPlanations[4]

- Local and global model-agnostic explanation method
- Can be employed as a local or global explainer
- Can produce various additive feature attribution methods

$$g(z') = \phi_0 + \sum_{i}^{M} \phi_i z'_i$$

- Main properties
  - local accuracy: $g(x) = b(x)$
  - missingness: no effect on SHAP values

[4] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, 4765–4774.

SHapley Additive exPlanations[4]

- Local and global model-agnostic explanation method
- Can be employed as a local or global explainer
- Can produce various additive feature attribution methods

$$g(z') = \phi_0 + \sum_{i}^{M} \phi_i z'_i$$

- Main properties
  - local accuracy: $g(x) = b(x)$
  - missingness: no effect on SHAP values
  - consistency: model changing lead to both different marginal feature values and SHAP values

[4] Scott M Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, 4765–4774.

SHapley Additive exPlanations[4]

- Local and global model-agnostic explanation method
- Can be employed as a local or global explainer
- Can produce various additive feature attribution methods

$$g(z') = \phi_0 + \sum_{i}^{M} \phi_i z'_i$$

- Main properties
  - local accuracy: $g(x) = b(x)$
  - missingness: no effect on SHAP values
  - consistency: model changing lead to both different marginal feature values and SHAP values
- Different strategies: Kernel, Linear, Tree, Gradient, and Deep explainer

[4] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, 4765–4774.

DALEX[5]

- Local and global model-agnostic explanation method

[5] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021.

DALEX[5]

- Local and global model-agnostic explanation method
- Local explanations

[5] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021.

DALEX[5]

- Local and global model-agnostic explanation method
- Local explanations
  - Employ variable attribution decomposition to quantify the contribution of each feature

[5] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021.

DALEX[5]

- Local and global model-agnostic explanation method
- Local explanations
  - Employ variable attribution decomposition to quantify the contribution of each feature
  - What-if analysis through its *ceteris-paribus* profile (no causality involved)

[5] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021.

DALEX[5]

- Local and global model-agnostic explanation method
- Local explanations
  - Employ variable attribution decomposition to quantify the contribution of each feature
  - What-if analysis through its *ceteris-paribus* profile (no causality involved)
- Global explanations:

[5] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021.

DALEX[5]

- Local and global model-agnostic explanation method
- Local explanations
  - Employ variable attribution decomposition to quantify the contribution of each feature
  - What-if analysis through its *ceteris-paribus* profile (no causality involved)
- Global explanations:
  - model performance measures

[5] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021.

DALEX[5]

- Local and global model-agnostic explanation method
- Local explanations
  - Employ variable attribution decomposition to quantify the contribution of each feature
  - What-if analysis through its *ceteris-paribus* profile (no causality involved)
- Global explanations:
  - model performance measures
  - variable importance

[5] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021.

DALEX[5]

- Local and global model-agnostic explanation method
- Local explanations
  - Employ variable attribution decomposition to quantify the contribution of each feature
  - What-if analysis through its *ceteris-paribus* profile (no causality involved)
- Global explanations:
  - model performance measures
  - variable importance
  - residual diagnoses

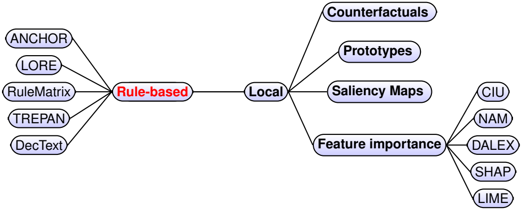
[5] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021.

DALEX[5]

- Local and global model-agnostic explanation method
- Local explanations
  - Employ variable attribution decomposition to quantify the contribution of each feature
  - What-if analysis through its *ceteris-paribus* profile (no causality involved)
- Global explanations:
  - model performance measures
  - variable importance
  - residual diagnoses
  - partial dependence plot

[5] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021.

Post-hoc explainability: rule-based methods



- Use decision rules to explain the reasons that lead to a specific prediction

ANCHOR[6]

- Model agnostic rule-based explanation method

[6] Ribeiro, Singh, and Guestrin, "Anchors: high-precision model-agnostic explanations".

ANCHOR[6]

- Model agnostic rule-based explanation method
- Output rules are named *anchors*

[6] Ribeiro, Singh, and Guestrin, "Anchors: high-precision model-agnostic explanations".

ANCHOR[6]

- Model agnostic rule-based explanation method
- Output rules are named *anchors*
- Given a sample x_i, r is an anchor if $r(x_i) = b(x)$

[6] Ribeiro, Singh, and Guestrin, “Anchors: high-precision model-agnostic explanations”.

ANCHOR[6]

- Model agnostic rule-based explanation method
- Output rules are named *anchors*
- Given a sample x_i, r is an anchor if $r(x_i) = b(x)$
- Build a perturbed sample from x_i

[6] Ribeiro, Singh, and Guestrin, "Anchors: high-precision model-agnostic explanations".

ANCHOR[6]

- Model agnostic rule-based explanation method
- Output rules are named *anchors*
- Given a sample x_i, r is an anchor if $r(x_i) = b(x)$
- Build a perturbed sample from x_i
- Extract all anchors with precision greater than a defined threshold

[6] Ribeiro, Singh, and Guestrin, "Anchors: high-precision model-agnostic explanations".

ANCHOR[6]

- Model agnostic rule-based explanation method
- Output rules are named *anchors*
- Given a sample x_i, r is an anchor if $r(x_i) = b(x)$
- Build a perturbed sample from x_i
- Extract all anchors with precision greater than a defined threshold
- Employs a multi-armed bandit algorithm

[6] Ribeiro, Singh, and Guestrin, "Anchors: high-precision model-agnostic explanations".

ANCHOR[6]

- Model agnostic rule-based explanation method
- Output rules are named *anchors*
- Given a sample x_i, r is an anchor if $r(x_i) = b(x)$
- Build a perturbed sample from x_i
- Extract all anchors with precision greater than a defined threshold
- Employs a multi-armed bandit algorithm
- Uses a bottom-up and beam search to explore the anchors

[6] Ribeiro, Singh, and Guestrin, "Anchors: high-precision model-agnostic explanations".

LOcal Rule-based Explainer (LORE)[7]

- Local model-agnostic method

[7] Guidotti et al., “Local rule-based explanations of black box decision systems”.

LOcal Rule-based Explainer (LORE)[7]

- Local model-agnostic method
- Provides explanations in the form of counterfactuals rules

[7] Guidotti et al., “Local rule-based explanations of black box decision systems”.

LOcal Rule-based Explainer (LORE)[7]

- Local model-agnostic method
- Provides explanations in the form of counterfactuals rules
- Only works with tabular data

[7] Guidotti et al., "Local rule-based explanations of black box decision systems".

LOcal Rule-based Explainer (LORE)[7]

- Local model-agnostic method
- Provides explanations in the form of counterfactuals rules
- Only works with tabular data
- Uses a genetic algorithm to generate a synthetic set Z of neighbors of a sample x_i

[7] Guidotti et al., "Local rule-based explanations of black box decision systems".

LOcal Rule-based Explainer (LORE)[7]

- Local model-agnostic method
- Provides explanations in the form of counterfactuals rules
- Only works with tabular data
- Uses a genetic algorithm to generate a synthetic set Z of neighbors of a sample x_i
  - Use the black-box model on Z to obtain the labels

[7] Guidotti et al., “Local rule-based explanations of black box decision systems”.

LOcal Rule-based Explainer (LORE)[7]

- Local model-agnostic method
- Provides explanations in the form of counterfactuals rules
- Only works with tabular data
- Uses a genetic algorithm to generate a synthetic set Z of neighbors of a sample x_i
  - Use the black-box model on Z to obtain the labels
  - Train a decision tree classifier $g(\cdot)$

[7] Guidotti et al., "Local rule-based explanations of black box decision systems".

LOcal Rule-based Explainer (LORE)[7]

- Local model-agnostic method
- Provides explanations in the form of counterfactuals rules
- Only works with tabular data
- Uses a genetic algorithm to generate a synthetic set Z of neighbors of a sample x_i
  - Use the black-box model on Z to obtain the labels
  - Train a decision tree classifier $g(\cdot)$
  - Optimize the output of $g(\cdot)$ w.r.t. the black-box

[7] Guidotti et al., "Local rule-based explanations of black box decision systems".

LOcal Rule-based Explainer (LORE)[7]

- Local model-agnostic method
- Provides explanations in the form of counterfactuals rules
- Only works with tabular data
- Uses a genetic algorithm to generate a synthetic set Z of neighbors of a sample x_i
  - Use the black-box model on Z to obtain the labels
  - Train a decision tree classifier $g(\cdot)$
  - Optimize the output of $g(\cdot)$ w.r.t. the black-box
- From g retrieves the explanations:

[7] Guidotti et al., "Local rule-based explanations of black box decision systems".

LOcal Rule-based Explainer (LORE)[7]

- Local model-agnostic method
- Provides explanations in the form of counterfactuals rules
- Only works with tabular data
- Uses a genetic algorithm to generate a synthetic set Z of neighbors of a sample x_i
  - Use the black-box model on Z to obtain the labels
  - Train a decision tree classifier $g(\cdot)$
  - Optimize the output of $g(\cdot)$ w.r.t. the black-box
- From g retrieves the explanations:
  - **factual decision rules**: path on the decision tree

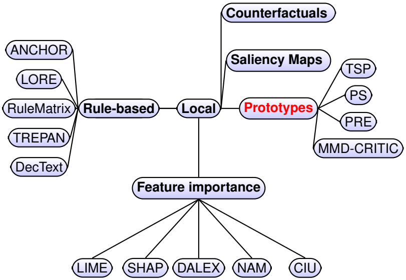
[7] Guidotti et al., "Local rule-based explanations of black box decision systems".

LOcal Rule-based Explainer (LORE)[7]

- Local model-agnostic method
- Provides explanations in the form of counterfactuals rules
- Only works with tabular data
- Uses a genetic algorithm to generate a synthetic set Z of neighbors of a sample x_i
  - Use the black-box model on Z to obtain the labels
  - Train a decision tree classifier $g(\cdot)$
  - Optimize the output of $g(\cdot)$ w.r.t. the black-box
- From g retrieves the explanations:
  - **factual decision rules**: path on the decision tree
  - **counterfactual rules**: which values of x_i lead to different outputs

[7] Guidotti et al., "Local rule-based explanations of black box decision systems".

Post-hoc explainability: prototypes methods



- Explain a model using a synthetic or natural example:
  - from the training set close to the a sample x_i
  - a centroid of a cluster for which x_i belongs to
  - generated by some ad-hoc process
- Humans observers usually understand a model's reasoning by looking at similar cases

Prototypes

- **Influence functions**[8]: identify instances in the training set that are responsible for the prediction of a given test instance

[8] Pang Wei Koh and Percy Liang. "Understanding black-box predictions via influence functions". In: *International Conference on Machine Learning*. 2017, pp. 1885–1894.

[9] Anh Nguyen, Jason Yosinski, and Jeff Clune. "Understanding neural networks via feature visualization: A survey". In: *Explainable AI: interpreting, explaining and visualizing deep learning*. 2019, pp. 55–76.

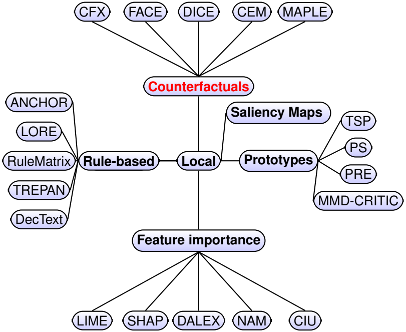
Prototypes

- **Influence functions**[8]: identify instances in the training set that are responsible for the prediction of a given test instance
- **activation maximization**[9]: Identify examples that strongly activate a function of interest

[8] Pang Wei Koh and Percy Liang. "Understanding black-box predictions via influence functions". In: *International Conference on Machine Learning*. 2017, pp. 1885–1894.

[9] Anh Nguyen, Jason Yosinski, and Jeff Clune. "Understanding neural networks via feature visualization: A survey". In: *Explainable AI: interpreting, explaining and visualizing deep learning*. 2019, pp. 55–76.

Post-hoc explainability: counterfactuals methods



- Prototypes' opposite
- Counterfactual explainers:
  - **exogeneous**: synthetically
  - **endogeneous**: from reference sample
  - **instance-based**: exploits a distance function to detected the decision boundary

Counterfactuals methods

- Contrastive explanation method (CEM)[10]
  - Local explanation method for neural network

[10] Amit Dhurandhar et al. "Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives". In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018, pp. 1–12.

[11] Emanuele Albini et al. "Relation-based counterfactual explanations for Bayesian network classifiers". In: *Twenty-Ninth International Joint Conference on Artificial Intelligence*. 2020.

Counterfactuals methods

- Contrastive explanation method (CEM)[10]
  - Local explanation method for neural network
  - Given x to explain, CEM considers $x_1 = x + \delta$

[10] Amit Dhurandhar et al. “Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives”. In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018, pp. 1–12.

[11] Emanuele Albini et al. “Relation-based counterfactual explanations for Bayesian network classifiers”. In: *Twenty-Ninth International Joint Conference on Artificial Intelligence*. 2020.

Counterfactuals methods

- Contrastive explanation method (CEM)[10]
  - Local explanation method for neural network
  - Given x to explain, CEM considers $x_1 = x + \delta$
  - Separate positive (δ^p) and negative (δ^n) perturbations w.r.t. label

[10] Amit Dhurandhar et al. "Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives". In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018, pp. 1–12.

[11] Emanuele Albini et al. "Relation-based counterfactual explanations for Bayesian network classifiers". In: *Twenty-Ninth International Joint Conference on Artificial Intelligence*. 2020.

Counterfactuals methods

- Contrastive explanation method (CEM)[10]
  - Local explanation method for neural network
  - Given x to explain, CEM considers $x_1 = x + \delta$
  - Separate positive (δ^p) and negative (δ^n) perturbations w.r.t. label
  - Use an autoencoder to explore the boundary between both regions

---

[10] Amit Dhurandhar et al. "Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives". In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018, pp. 1–12.

[11] Emanuele Albini et al. "Relation-based counterfactual explanations for Bayesian network classifiers". In: *Twenty-Ninth International Joint Conference on Artificial Intelligence*. 2020.

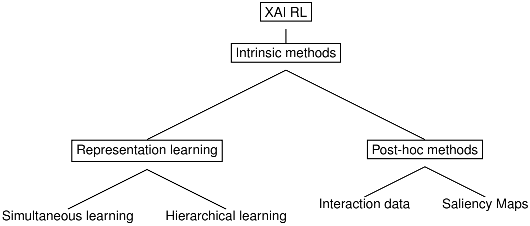
Counterfactuals methods

- Contrastive explanation method (CEM)[10]
  - Local explanation method for neural network
  - Given x to explain, CEM considers $x_1 = x + \delta$
  - Separate positive (δ^p) and negative (δ^n) perturbations w.r.t. label
  - Use an autoencoder to explore the boundary between both regions
- CFX[11]
  - Local explanation method for Bayesian network classifiers
  - Explanations are built from relations of influence between variables, indicating the reasons for the classification

[10] Amit Dhurandhar et al. "Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives". In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018, pp. 1–12.

[11] Emanuele Albini et al. "Relation-based counterfactual explanations for Bayesian network classifiers". In: *Twenty-Ninth International Joint Conference on Artificial Intelligence*. 2020.

Explainable Reinforcement Learning (XAI RL)[12]



[12] Erika Puiutta and Eric MSP Veith. “Explainable reinforcement learning: A survey”. In: *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. 2020, pp. 77–95; Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. “Explainability in deep reinforcement learning”. In: *Knowledge-Based Systems* 214 (2021), pp. 1–13.

Summary of reviewed literature on explainable RL (XRL) and deep RL (DRL).

Reference	Task/Environment	Decision process	Algorithm(s)	Explanation type (Level)	Target
Relational Deep RL [21]	Planning + strategy games (Box-World/ Starcraft II)	POMDP	IMPALA	Images (Local)	Experts
Symbolic RL with Common Sense [22]	Game (object retrieval)	POMDP	SRL+CS, DQL	Images (Global)	Experts
Decoupling feature extraction from policy learning [23]	Robotics (grasping), and navigation	MDP	PPO	Diagram (state plot & image slider (Local)	Experts
Explainable RL via Reward Decomposition [24]	Game (grid and landing)	MDP	HRA, SARSA, Q-learning	Diagrams (Local)	Experts, Users, Executives
Explainable RL Through a Causal Lens [25]	Games (OpenAI benchmark and Starcraft II)	Both	PG, DQN, DDPG, A2C, SARSA	Diagrams, Text (Local)	Experts, Users, Executives
Shapley Q-value: A Local Reward Approach to Solve Global Reward Games [26]	Multiagents (Cooperative Navigation, Prey-and-Predator and Traffic Junction)	POMDP	DDPG	Diagrams (Local)	Experts
Dot-to-Dot: Explainable HRL For Robotic Manipulation [27]	Robotics (grasping)	MDP	DDPG, HER, HRL	Diagrams (Global)	Experts, Developers
Self-Educated Language Agent With HER For Instruction Following [28]	Instruction Following (MiniGrid)	MDP	Textual HER	Text (Local)	Experts, Users, Developers
Commonsense and Semantic-guided Navigation [29]	Room navigation	POMDP	–	Text (Global)	Experts
Boolean Task Algebra [30]	Game (grid)	MDP	DQN	Diagrams	Experts
Visualizing and Understanding Atari [31]	Games (Pong, Breakout, Space Invaders)	MDP	A3C	Images (Global)	Experts, Users, Developers
Interestingness Elements for XRL through Introspection [32, 33]	Arcade game (Frogger)	POMDP	Q-Learning	Images (Local)	Users
Composable DRL for Robotic Manipulation [34]	Robotics (pushing and reaching)	MDP	Soft Q-learning	Diagrams (Local)	Experts
Symbolic-Based Recognition of Contact States for Learning Assembly Skills [35]	Robotic grasping	POMDP	HMM, PAA, K-means	Diagrams (Local)	Experts
Safe Reinforcement Learning with Model Uncertainty Estimates [36]	Collision avoidance	POMDP	Monte Carlo Dropout, bootstrapping	Diagrams (Local)	Experts

Figure 2: Summary of explainable RL and deep RL[13]

[13] Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. "Explainability in deep reinforcement learning". In: *Knowledge-Based Systems* 214 (2021), pp. 1–13.

Evaluation measures

- **Fidelity**: how good is $f(\cdot)$ at mimicking the $b(\cdot)$?

Evaluation measures

- **Fidelity**: how good is $f(\cdot)$ at mimicking the $b(\cdot)$?
- **Stability**: how consistent are the explanations for similar samples?

Evaluation measures

- **Fidelity**: how good is $f(\cdot)$ at mimicking the $b(\cdot)$?
- **Stability**: how consistent are the explanations for similar samples?
- **Faithfulness**: how are the relevance scores indicating the true importance features?

Evaluation measures

- **Fidelity**: how good is $f(\cdot)$ at mimicking the $b(\cdot)$?
- **Stability**: how consistent are the explanations for similar samples?
- **Faithfulness**: how are the relevance scores indicating the true importance features?
- **Monoticity**: how is the accuracy of be improved when new a new important feature is added?

Fidelity and faithfulness metrics

Table 1: Comparison of fidelity and faithfulness metrics of four explanation methods[14]

Dataset	Black-Box	Fidelity				Faithfulness	
		LIME	SHAP	ANCHOR	LORE	LIME	SHAP
adult	LG	0.979	0.613	**0.989**	0.984	0.099 (0.30)	**0.38** (0.37)
	XGB	0.977	0.877	0.978	**0.982**	0.030 (0.32)	**0.36** (0.49)
	CAT	0.96	0.777	0.988	**0.989**	0.077 (0.32)	**0.44** (0.37)
german	LG	**0.984**	0.910	0.730	0.983	**0.23** (0.60)	0.19 (0.63)
	XGB	**0.999**	0.821	0.802	0.982	0.16 (0.26)	**0.44** (0.21)
	CAT	0.979	0.670	0.620	**0.981**	0.34 (0.33)	**0.43** (0.32)

Higher is better. Logistic Regression (LG), XGBoot (XGB), and CatBoost (CAT)
Adult census income data set
German credit data set

[14] Francesco Bodria et al. "Benchmarking and survey of explanation methods for black box models". In: *arXiv:2102.13076* (2021).

Stability metric

Table 2: Comparison of the stability metric of four explanation methods[15]

Dataset	Black-Box	LIME	SHAP	ANCHOR	LORE
adult	LG	24.37 (2.74)	$\boxed{1.52\ (4.49)}$	22.36 (8.37)	21.76 (11.80)
	XGB	10.16 (6.48)	$\boxed{2.17\ (2.18)}$	26.53 (13.08)	30.01 (20.52)
	CAT	0.35 (0.43)	$\boxed{0.03\ (0.01)}$	6.51 (4.40)	27.80 (70.05)
german	LG	$\boxed{18.87\ (0.73)}$	19.01 (23.44)	101.07 (62.75)	622.12 (256.70)
	XGB	$\boxed{26.08\ (14.50)}$	38.43 (30.66)	121.40 (98.43)	725.81 (337.26)
	CAT	$\boxed{2.49\ (9.91)}$	15.92 (10.71)	123.79 (76.86)	756.70 (348.21)

Lower is better

[15] Francesco Bodria et al. "Benchmarking and survey of explanation methods for black box models". In: *arXiv:2102.13076* (2021).

Fragility: post-hoc explanations can be manipulated

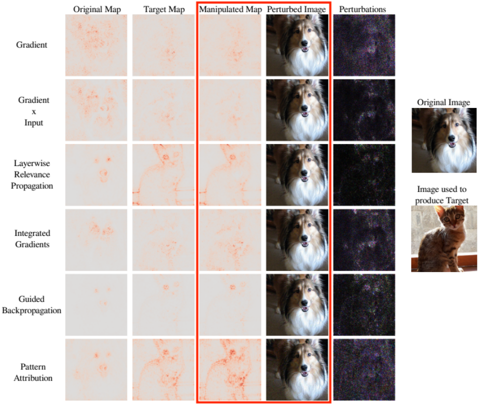
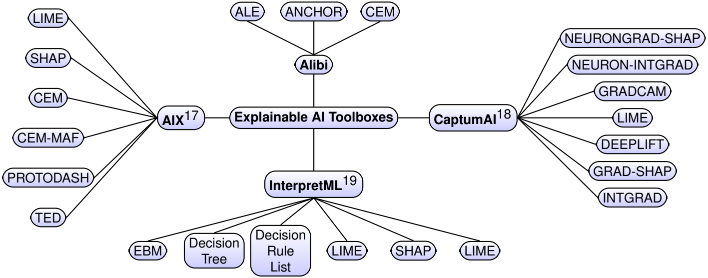


Figure 3: Explanation map of a cat is used as the target of a perturbed dog image[16]

[16] Ann-Kathrin Dombrowski et al. "Explanations can be manipulated and geometry is to blame". In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019.

Explainability toolboxes



[17] Vijay Arya et al. "One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques". In: *arXiv:1909.03012* (2019).

[18] PyTorch limited

[19] Harsha Nori et al. "InterpretML: a unified framework for machine learning interpretability". In: *arXiv:1909.09223* (2019).

In conclusion

"Understanding a phenomena is not simply a matter of reducing the "fundamental incomprehensibilities", but of seeing connections, common patterns, in what initially appeared to be different situations" – Kitcher (1989)[a]

[a] Philip Kitcher. "Explanatory unification and the causal structure of the world". In: *Scientific Explanation*. Ed. by P. Kitcher and W.C. Salmon. University of Minnesota Press, 1989, pp. 410–505.

- Various explainable AI methods have be developed over the last years

In conclusion

"Understanding a phenomena is not simply a matter of reducing the "fundamental incomprehensibilities", but of seeing connections, common patterns, in what initially appeared to be different situations" – Kitcher (1989)[a]

---

[a] Philip Kitcher. "Explanatory unification and the causal structure of the world". In: *Scientific Explanation*. Ed. by P. Kitcher and W.C. Salmon. University of Minnesota Press, 1989, pp. 410–505.

- Various explainable AI methods have be developed over the last years
- Feature importance is the most widely adopted strategy

In conclusion

"Understanding a phenomena is not simply a matter of reducing the "fundamental incomprehensibilities", but of seeing connections, common patterns, in what initially appeared to be different situations" – Kitcher (1989)[a]

[a] Philip Kitcher. "Explanatory unification and the causal structure of the world". In: *Scientific Explanation*. Ed. by P. Kitcher and W.C. Salmon. University of Minnesota Press, 1989, pp. 410–505.

- Various explainable AI methods have be developed over the last years
- Feature importance is the most widely adopted strategy
- Rule-based explanations are gaining attention due to the logical formalization strategy

In conclusion

"Understanding a phenomena is not simply a matter of reducing the "fundamental incomprehensibilities", but of seeing connections, common patterns, in what initially appeared to be different situations" – Kitcher (1989)[a]

[a] Philip Kitcher. "Explanatory unification and the causal structure of the world". In: *Scientific Explanation*. Ed. by P. Kitcher and W.C. Salmon. University of Minnesota Press, 1989, pp. 410–505.

- Various explainable AI methods have be developed over the last years
- Feature importance is the most widely adopted strategy
- Rule-based explanations are gaining attention due to the logical formalization strategy
- There are still space to make explainable AI stable, as well as understandable by different human observers

In conclusion

"Understanding a phenomena is not simply a matter of reducing the "fundamental incomprehensibilities", but of seeing connections, common patterns, in what initially appeared to be different situations" – Kitcher (1989)[a]

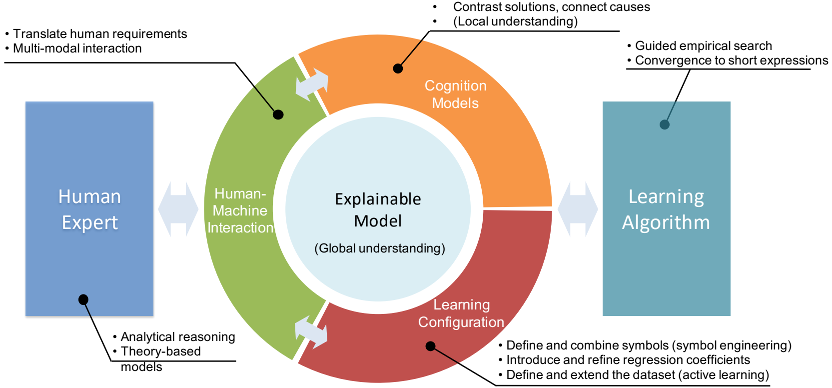
[a] Philip Kitcher. "Explanatory unification and the causal structure of the world". In: *Scientific Explanation*. Ed. by P. Kitcher and W.C. Salmon. University of Minnesota Press, 1989, pp. 410–505.

- Various explainable AI methods have be developed over the last years
- Feature importance is the most widely adopted strategy
- Rule-based explanations are gaining attention due to the logical formalization strategy
- There are still space to make explainable AI stable, as well as understandable by different human observers

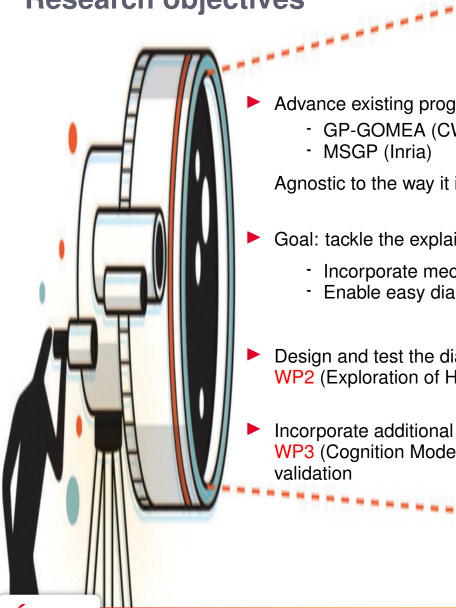
"Explaining black boxes, rather than replace them with interpretable models, can make the problem worse by providing misleading or false characterizations to the black box. – Rudin (2019)[a]*"*

[a] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215.

Building a TRUSTED AI system



Research objectives



- Advance existing program synthesis algorithms
  - GP-GOMEA (CWI)
  - MSGP (Inria)

  Agnostic to the way it is used toward explainability

- Goal: tackle the explainability vs accuracy trade-off
  - Incorporate mechanisms for tunable explainability
  - Enable easy dialogue in view of multi-objective optimization

- Design and test the dialog platform for the different algorithms together with WP2 (Exploration of Human-Machine Interaction)

- Incorporate additional variables (e.g., latent confounders) as proposed by WP3 (Cognition Models for Human-Centric XAI) following experimental validation

Use cases

	Use Case 1	Use Case 2	Use Case 3
Problem / Application	Cancer Treatment (Healthcare)	Time Slot Selection (Retail)	Demand Forecast (Energy)
AI Task	Regression (Predictive)	Selection (Prescriptive)	Regression (Predictive)
Key Features	Risk, Learning from Small Data	Fairness (to multiple stakeholders), Multi-criteria	Distributed Sources of Data, Incremental and Active Learning
Partner	LUMC Leiden University Medical Center	SONAE MC	APINTECH

Project consortium

Research Organizations	Small or Medium-Size Enterprises	Industrial Partners
INESCTEC: Engineering systems institute, with experts in operations management	TAZI: Explainable continuous machine learning platform for insurance, banking and retail sectors	Leiden University Medical Center: Department of radiation oncology, with previous work on mathematics and AI applied to radiotherapy, brachytherapy and image-guidance systems
Inria: Computer science institute, with experts in machine learning and evolutionary optimization	LTP: Analytics-based consultancy for retail, manufacturing and telecommunications	SONAE MC: Large food retailer, with extensive use of analytics (optimization, simulation and AI) in their supply chain operations, both physical and online
University of Tartu: Computational neuroscience lab, with experts in cognitive artificial intelligence	APINTECH: Sensors and IoT solutions, where big data and machine learning methods are built on	
CWI: Mathematics and computer science institute, with experts in medical informatics		

Internships

- Counterfactuals-based explanations (Alex Westbrook)
- XRL explanations (Mathurin Videau)
- MILP (Li Wenhao)


Thank
you!