Friedrich-Alexander-Universität
Technische Fakultät

Advanced Deep Learning

Interpretable & Causal Deep Learning

K. Breininger, V. Christlein

Artificial Intelligence in Medical Imaging + Pattern Recognition Lab,

Friedrich-Alexander-Universität Erlangen-Nürnberg SoSe 2023

Slide Credit: Narang Ishita, Rudresh Veeresh

1. Introduction

2. Interpretable DL

3. Neural Network Interpretation

4. Causal DL

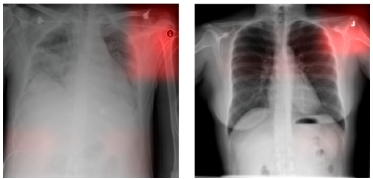
Can you trust just high accuracy?

Can you trust just high accuracy?

Husky classified as wolf. Source [17].

Can you trust just high accuracy?

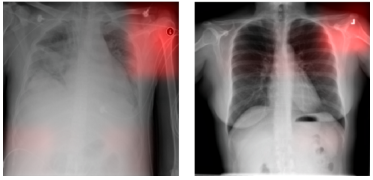
Husky classified as wolf. Source [17].



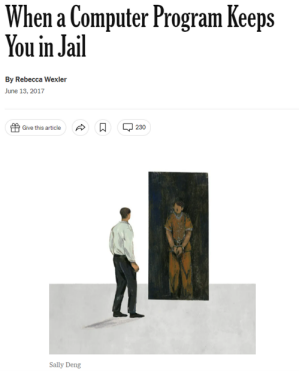
Medical confounders. Souce [16].

Can you trust just high accuracy?

Husky classified as wolf. Source [17].



Medical confounders. Souce [16].



Automated parole decisions. Image source: Sally Deng.

A single metric may not be enough

Opinion

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

- Proprietary system to identify re-offenders for offering parole, based on 137 variables [11]
- Blacks twice as often as false positive re-offenders[1]
- Different "fairness" criteria[2,3]
- Not better than "untrained humans"
- Two parameters (age, # of convictions) can achieve same performance [11]
- Other proprietary systems for DNA analysis, firearm analysis, fingerprint analysis, etc.

Source: Image source: Sally Deng, Article: NYTimes article

Different fairness criteria:

[1] ProPublica Article

[2] Opinion in the American Council on Science and Health [pro-industry]

[3] Article in the Washington Post

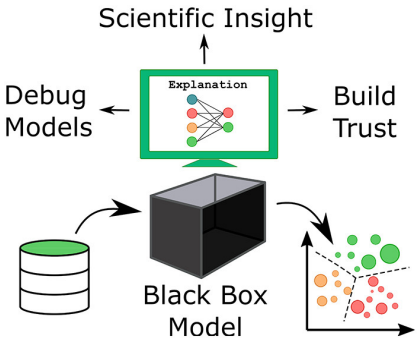
A Right to an Explanation?

In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.
GDPR of the European Union, Recital 71.4

Interpretability and Explainability

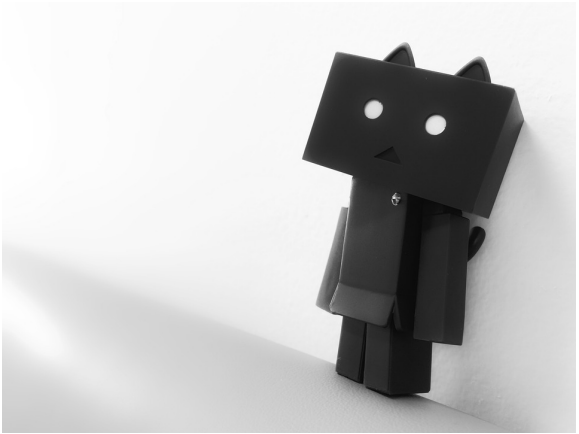
- **Human curiosity and learning**
- **Safety measures**
- **Detecting biases**
- **Interventionability**

→ Social acceptance & Human-AI collaboration



Source: Oviedo et al. 2022 [18]

Looking insight a Blackbox?



Source: https://pixabay.com/photos/danbo-nyangbo-figure-doll-1206478/

1. Introduction

2. Interpretable DL

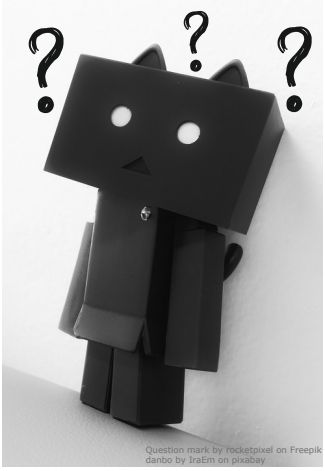
3. Neural Network Interpretation

4. Causal DL

Terminology of Interpretability

Some definitions for interpretability:

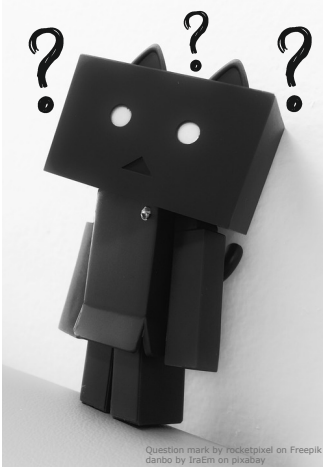
- “Interpretability is the degree to which a human can understand the cause of a decision.” [2]



Terminology of Interpretability

Some definitions for interpretability:

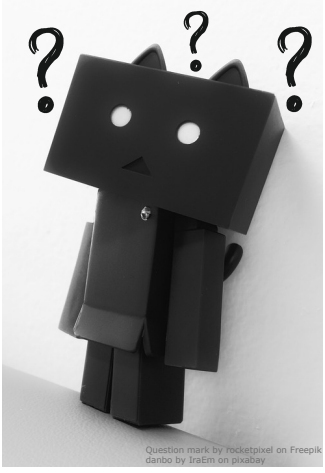
- “Interpretability is the degree to which a human can understand the cause of a decision.” [2]
- “Interpretability is the degree to which a human can consistently predict the model’s result.” [3]



Terminology of Interpretability

Some definitions for interpretability:

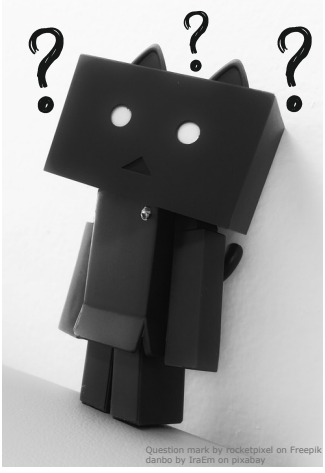
- “Interpretability is the degree to which a human can understand the cause of a decision.” [2]
- “Interpretability is the degree to which a human can consistently predict the model’s result.” [3]
- “[Interpretability means] to explain or to present in understandable terms to a human.” [4]



Terminology of Interpretability

Some definitions for interpretability:

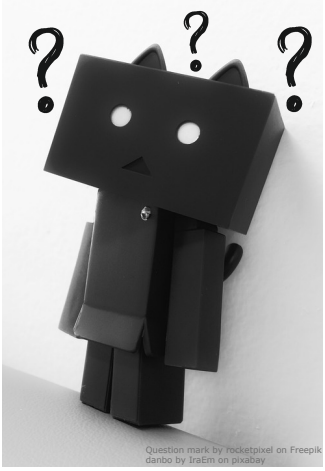
- “Interpretability is the degree to which **a human** can understand the cause of a decision.” [2]
- “Interpretability is the degree to which **a human** can consistently predict the model’s result.” [3]
- “[Interpretability means] to explain or to present in understandable terms to **a human**.” [4]



Terminology of Interpretability

Some definitions for interpretability:

- “Interpretability is the degree to which **a human** can understand the cause of a decision.” [2]
- “Interpretability is the degree to which **a human** can consistently predict the model’s result.” [3]
- “[Interpretability means] to explain or to present in understandable terms to **a human**.” [4]
- “What in the model structure explains its functioning?” [5]

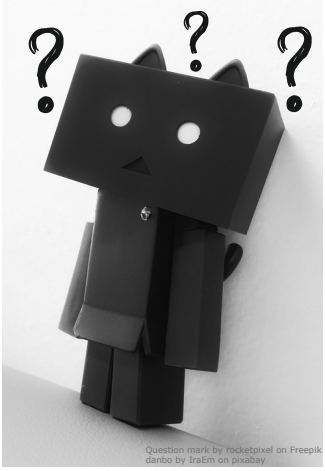


Terminology of Interpretability

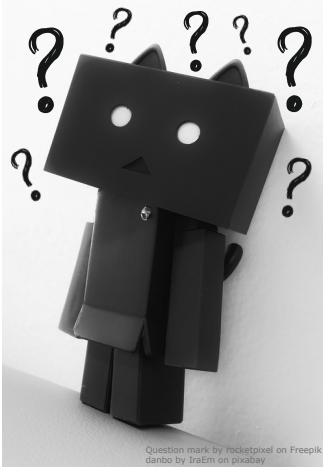
Some definitions for interpretability:

- “Interpretability is the degree to which **a human** can understand the cause of a decision.” [2]
- “Interpretability is the degree to which **a human** can consistently predict the model’s result.” [3]
- “[Interpretability means] to explain or to present in understandable terms to **a human**.” [4]
- “What in the model structure explains its functioning?” [5]

→ Often model-based, i.e., a model is inherently interpretable (or not), focus on the human

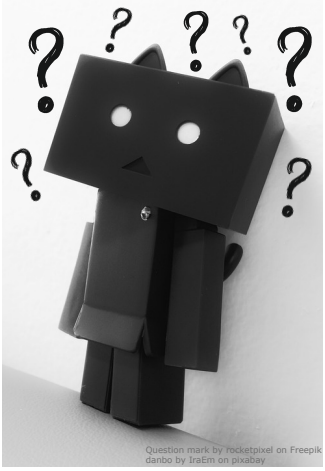


And Explainability?



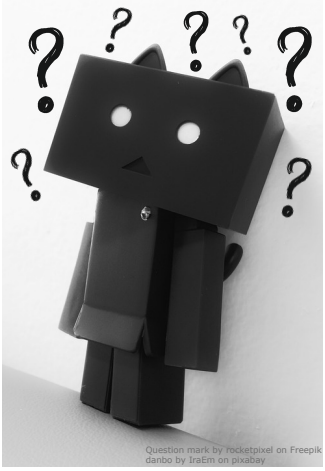
And Explainability?

- “Interpretability” and “Explainability” used interchangeably but **explanation** for individual explanations [2]



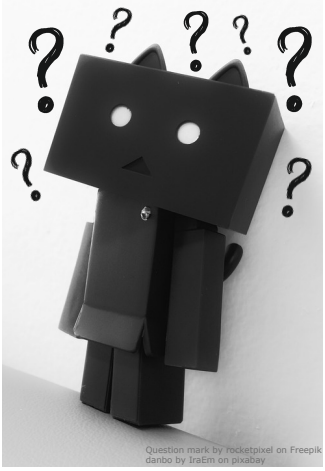
And Explainability?

- “Interpretability” and “Explainability” used interchangeably but **explanation** for individual explanations [2]
- “What is the rationale behind the decision made?” [5]



And Explainability?

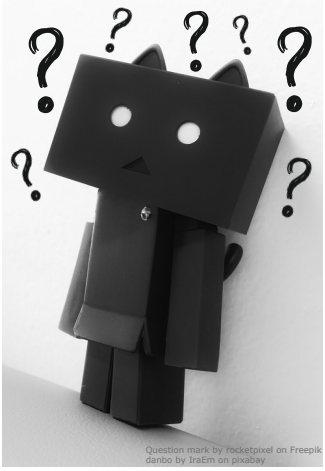
- “Interpretability” and “Explainability” used interchangeably but **explanation** for individual explanations [2]
- “What is the rationale behind the decision made?” [5]
- includes methods for “post-hoc” explanations of models



And Explainability?

- “Interpretability” and “Explainability” used interchangeably but **explanation** for individual explanations [2]
- “What is the rationale behind the decision made?” [5]
- includes methods for “post-hoc” explanations of models

➔ Often used for explanation methods, model-agnostic➔ Domain- and user-specific



Types of Explanations

- ***Intrinsic vs Post-Hoc*** : Does the model yield explanations directly due to it's simple structure or by applying methods that analyze the model after training? (aka interpretable versus explainable)

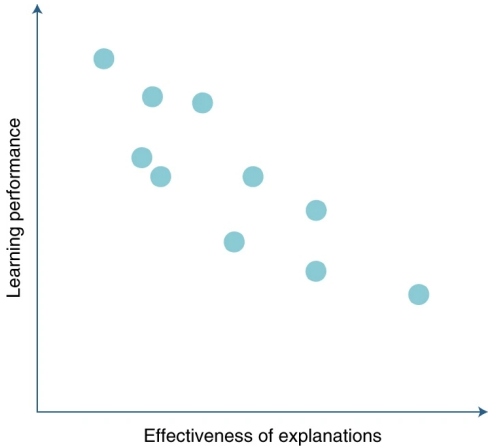
Types of Explanations

- ***Intrinsic vs Post-Hoc*** : Does the model yield explanations directly due to it's simple structure or by applying methods that analyze the model after training? (aka interpretable versus explainable)
- ***Global vs Local Interpretations*** : Do they explain the model behavior on the entire data set or only a small subset near a single data point?

Types of Explanations

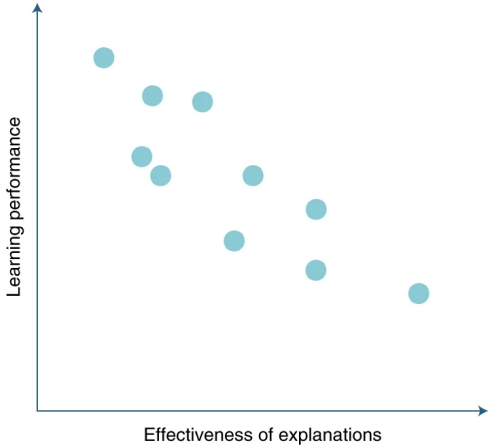
- ***Intrinsic vs Post-Hoc*** : Does the model yield explanations directly due to it's simple structure or by applying methods that analyze the model after training? (aka interpretable versus explainable)
- ***Global vs Local Interpretations*** : Do they explain the model behavior on the entire data set or only a small subset near a single data point?
- ***Model-Specific vs Agnostic Methods*** : Can explanations be obtained only for a specific type of model or for any type?

Model Interpretability vs. Model Performance Trade-Off

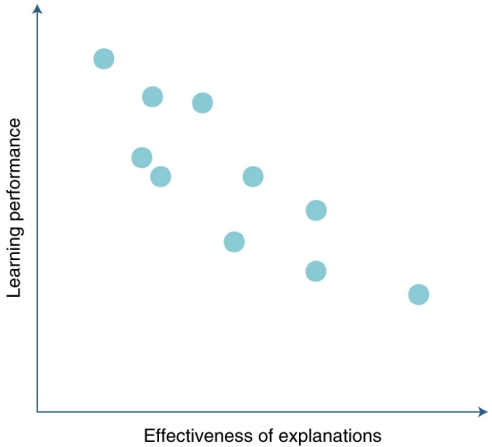


Model Interpretability vs. Model Performance Trade-Off

How to deal with this trade-off?



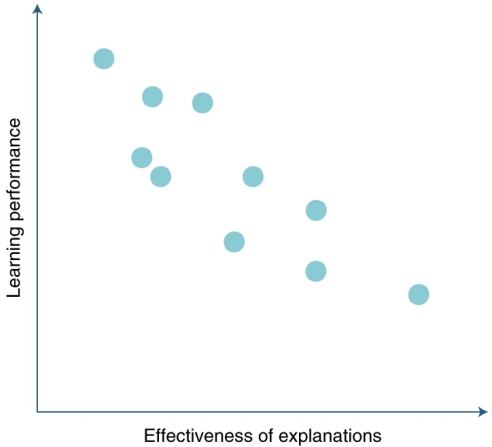
Model Interpretability vs. Model Performance Trade-Off



How to deal with this trade-off?

Making inherently interpretable models more accurate.

Model Interpretability vs. Model Performance Trade-Off



How to deal with this trade-off?

Making inherently interpretable models more accurate.

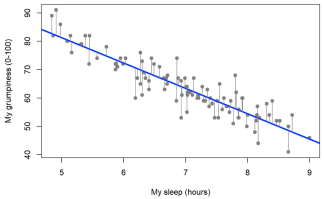
OR

Generating good explanations for accurate black-box models.

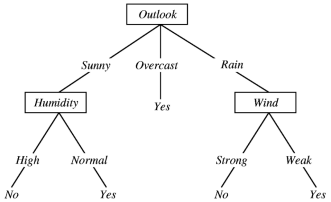
Source: Graphic from [15], in turn from [19]

Intrinsically interpretable models

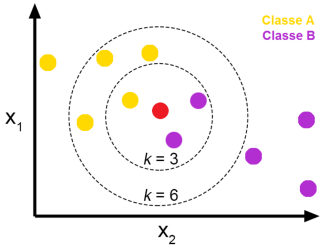
Examples of models that are interpretable by design:



Linear Regression. Source [20].

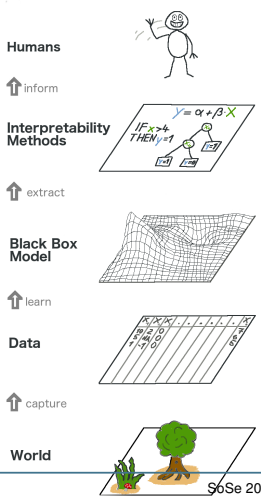


Decision tree. Source [21].



K-Nearest Neighbor. Source [22]

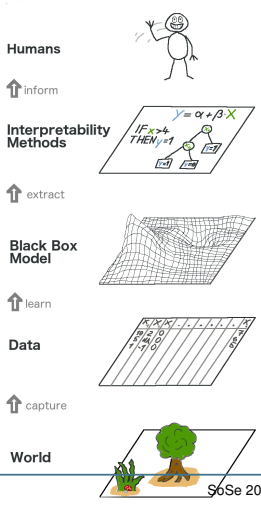
Post-hoc explanation methods



Post-hoc explanation methods

Advantages:

- Allows use of complex ML models
- Enable flexible and comparative use of different ML models
- Highly exploratory data analysis



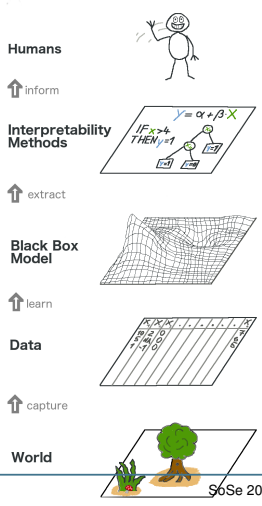
Post-hoc explanation methods

Advantages:

- Allows use of complex ML models
- Enable flexible and comparative use of different ML models
- Highly exploratory data analysis

Disadvantages:

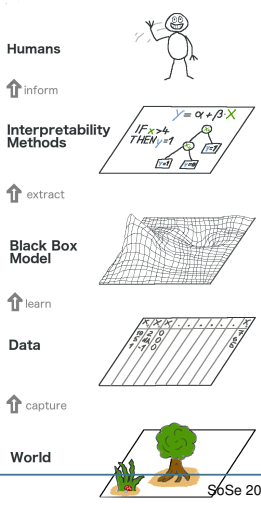
- Explain the model only locally or along single or few input axes
- Often difficult/highly approximative at a global level



Examples for post-hoc explanations

Global post-hoc methods:

- Partial Dependence Plots
- Feature Interaction
- Global Surrogates



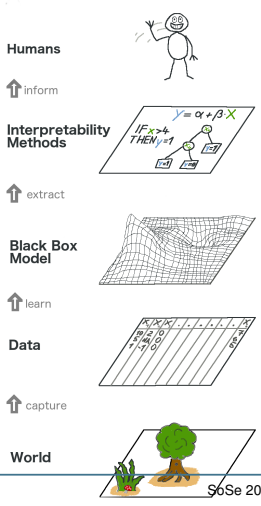
Examples for post-hoc explanations

Global post-hoc methods:

- Partial Dependence Plots
- Feature Interaction
- Global Surrogates

Local post-hoc methods:

- Local surrogates (LIME)
- Counterfactual explanations
- Shapley values
- SHapley Additive exPlanations (SHAP)



Examples for post-hoc explanations

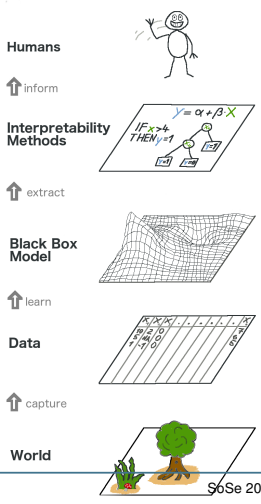
Global post-hoc methods:

- Partial Dependence Plots
- Feature Interaction
- Global Surrogates

Local post-hoc methods:

- Local surrogates (LIME)
- Counterfactual explanations
- Shapley values
- SHapley Additive exPlanations (SHAP)

➔ Refer to [13] for an excellent discussion of different advantages and disadvantages.



Examples for post-hoc explanations

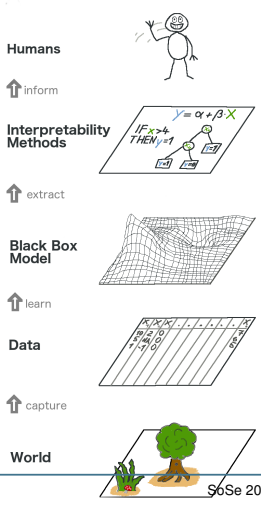
Global post-hoc methods:

- **Partial Dependence Plots**
- Feature Interaction
- Global Surrogates

Local post-hoc methods:

- **Local surrogates (LIME)**
- Counterfactual explanations
- Shapley values
- SHapley Additive exPlanations (SHAP)

➔ Refer to [13] for an excellent discussion of different advantages and disadvantages.



Partial Dependence Plots (PDPs)

- Example use case: Bike rental
  - Features: Temperature, humidity, wind speed, weekday, holiday season, ...
  - Target: Number of rented bikes

Partial Dependence Plots (PDPs)

- Example use case: Bike rental
  - Features: Temperature, humidity, wind speed, weekday, holiday season, ...
  - Target: Number of rented bikes
- A PDP shows the marginal effect of features on the model output

Partial Dependence Plots (PDPs)

- Example use case: Bike rental
  - Features: Temperature, humidity, wind speed, weekday, holiday season, ...
  - Target: Number of rented bikes
- A PDP shows the marginal effect of features on the model output
- PDPs can show whether the relationship between the target and a feature is linear, monotonic or more complex.

Partial Dependence Plots (PDPs)

- Example use case: Bike rental
  - Features: Temperature, humidity, wind speed, weekday, holiday season, ...
  - Target: Number of rented bikes
- A PDP shows the marginal effect of features on the model output
- PDPs can show whether the relationship between the target and a feature is linear, monotonic or more complex.
- Partial dependence function for regression:

$$\hat{f}_S(x_S) = E_{X_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, X_C)\, d\mathbb{P}(X_C)$$

Partial Dependence Plots (PDPs)

- Example use case: Bike rental
  - Features: Temperature, humidity, wind speed, weekday, holiday season, ...
  - Target: Number of rented bikes
- A PDP shows the marginal effect of features on the model output
- PDPs can show whether the relationship between the target and a feature is linear, monotonic or more complex.
- Partial dependence function for regression:

$$\hat{f}_S(x_S) = E_{X_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, X_C)\, d\mathbb{P}(X_C)$$

where:
  - S - Set containing features of interest
  - C - Set that contains all other features not in set S

Partial Dependence Plots (PDPs)

- Example use case: Bike rental
  - Features: Temperature, humidity, wind speed, weekday, holiday season, ...
  - Target: Number of rented bikes
- A PDP shows the marginal effect of features on the model output
- PDPs can show whether the relationship between the target and a feature is linear, monotonic or more complex.
- Partial dependence function for regression:

$$\hat{f}_S(x_S) = E_{X_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, X_C)\, d\mathbb{P}(X_C)$$

where:

- S - Set containing features of interest
- C - Set that contains all other features not in set S
- x_S - Features for which the partial dependence function should be plotted

Partial Dependence Plots (PDPs)

- Example use case: Bike rental
  - Features: Temperature, humidity, wind speed, weekday, holiday season, ...
  - Target: Number of rented bikes
- A PDP shows the marginal effect of features on the model output
- PDPs can show whether the relationship between the target and a feature is linear, monotonic or more complex.
- Partial dependence function for regression:

$$\hat{f}_S(x_S) = E_{X_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, X_C)\, d\mathbb{P}(X_C)$$

where:

- S - Set containing features of interest
- C - Set that contains all other features not in set S
- x_S - Features for which the partial dependence function should be plotted
- X_C - Other features used in the machine learning model $\hat{f}$

Partial Dependence Plots (PDPs)

- The partial function $\hat{f}_S$ is estimated by calculating averages in the training data:

$$\hat{f}_S(x_S) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}\left(x_S, x_C^{(i)}\right)$$

where:

- $x_C^{(i)}$ - Feature values in which we are not interested
- n - Number of instances in the dataset

Partial Dependence Plots (PDPs)

- The partial function $\hat{f}_S$ is estimated by calculating averages in the training data:

$$\hat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}\left(x_S, x_C^{(i)}\right)$$

  where:
  - $x_C^{(i)}$ - Feature values in which we are not interested
  - n - Number of instances in the dataset
- We marginalize the output over all but one (or two) features of interest.

Partial Dependence Plots (PDPs)

- The partial function $\hat{f}_S$ is estimated by calculating averages in the training data:

$$\hat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}\left(x_S, x_C^{(i)}\right)$$

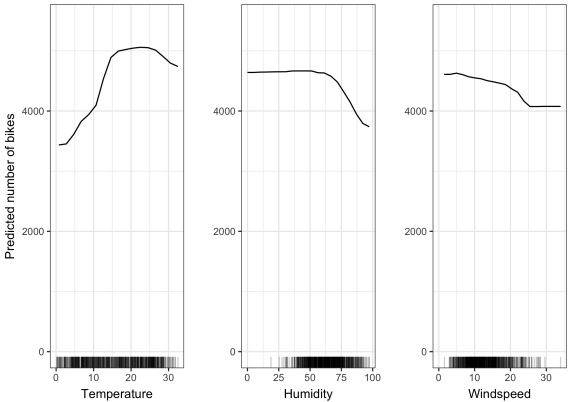
  where:
  - $x_C^{(i)}$ - Feature values in which we are not interested
  - n - Number of instances in the dataset
- We marginalize the output over all but one (or two) features of interest.
- **Core assumption**: Features of interest are independent with the others.

Example: Bike Rental Dataset

Features of interest (individually): **temperature**, **humidity**, and **windspeed**

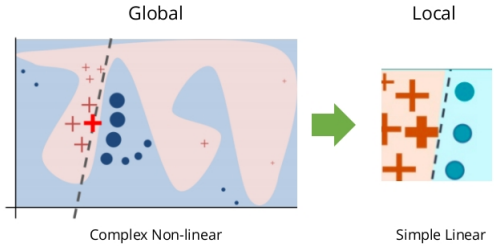
Example: Bike Rental Dataset

Features of interest (individually): **temperature**, **humidity**, and **windspeed**



PDPs for the bicycle count prediction model and temperature, humidity and wind speed.

Local Interpretable Model-Agnostic Explanations (LIME) [23]

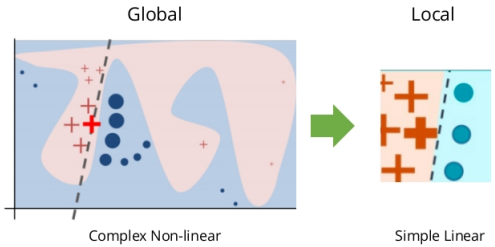


LIME provides a local explanation

Source: https://www.kdnuggets.com/2019/12/interpretability-part-3-lime-shap.html

Local Interpretable Model-Agnostic Explanations (LIME) [23]

- LIME aims to create a local approximation of our complex, arbitrary model for a specific input

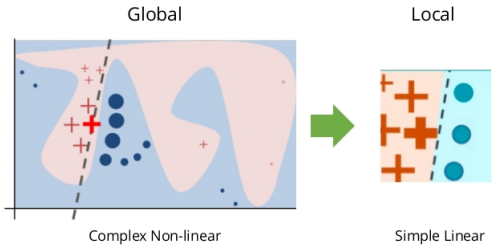


LIME provides a local explanation

Source: https://www.kdnuggets.com/2019/12/interpretability-part-3-lime-shap.html

Local Interpretable Model-Agnostic Explanations (LIME) [23]

- LIME aims to create a local approximation of our complex, arbitrary model for a specific input
- Model internals are hidden/unknown, data-type agnostic

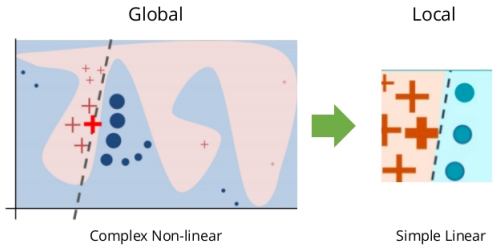


LIME provides a local explanation

Source: https://www.kdnuggets.com/2019/12/interpretability-part-3-lime-shap.html

Local Interpretable Model-Agnostic Explanations (LIME) [23]

- LIME aims to create a local approximation of our complex, arbitrary model for a specific input
- Model internals are hidden/unknown, data-type agnostic
- Using prior knowledge we can validate the explanations and create trust



LIME provides a local explanation

Source: https://www.kdnuggets.com/2019/12/interpretability-part-3-lime-shap.html

The math in LIME

Training local surrogates requires balancing the local fidelity and the model complexity of the surrogate model:

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

The math in LIME

Training local surrogates requires balancing the local fidelity and the model complexity of the surrogate model:

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

where:

- f - Complex black-box model

The math in LIME

Training local surrogates requires balancing the local fidelity and the model complexity of the surrogate model:

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

where:

- f - Complex black-box model
- g - Surrogate model, chosen from a set G
- G - Family of Interpretable models

The math in LIME

Training local surrogates requires balancing the local fidelity and the model complexity of the surrogate model:

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

where:

- f - Complex black-box model
- g - Surrogate model, chosen from a set G
- G - Family of Interpretable models
- L - Loss of g with respect to f

The math in LIME

Training local surrogates requires balancing the local fidelity and the model complexity of the surrogate model:

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

where:

- f - Complex black-box model
- g - Surrogate model, chosen from a set G
- G - Family of Interpretable models
- L - Loss of g with respect to f
- π - proximity of the local data instance x to be explained

The math in LIME

Training local surrogates requires balancing the local fidelity and the model complexity of the surrogate model:

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

where:

- f - Complex black-box model
- g - Surrogate model, chosen from a set G
- G - Family of Interpretable models
- L - Loss of g with respect to f
- π - proximity of the local data instance x to be explained
- Ω - Complexity of the surrogate model

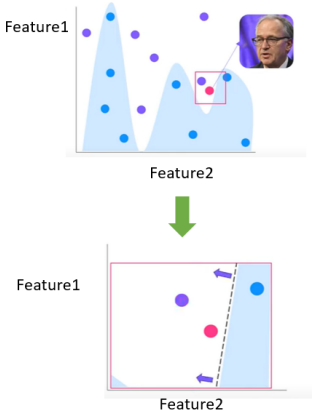
How to train the surrogate model

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

x =

Age	BMI	Avg_glucose_level	work_type	gender	hypertension	Heart_disease

- Select your instance of interest
- Perturb your dataset and get the black box predictions for these new points
- Weight the new samples according to their proximity to the instance of interest
- Train a weighted, interpretable model on the dataset with the variations
- Explain the prediction by interpreting the local model



LIME on stroke dataset

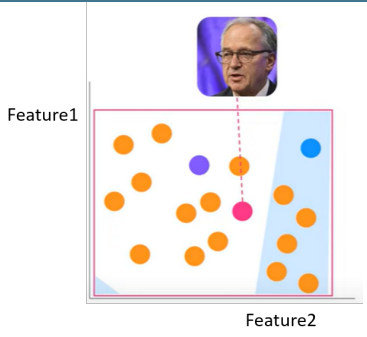
How to train the surrogate

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

x =

Age	BMI	Avg_glucose_level	work_type	gender	hypertension	Heart_disease

- Select your instance of interest.
- Perturb your dataset and get the black box predictions for these new points
- Weight the new samples according to their proximity to the instance of interest
- Train a weighted, interpretable model on the dataset with the variations
- Explain the prediction by interpreting the local model.



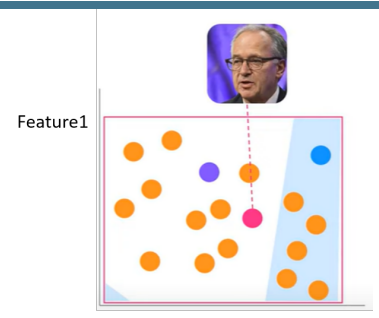
How to train the surrogate

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

x =

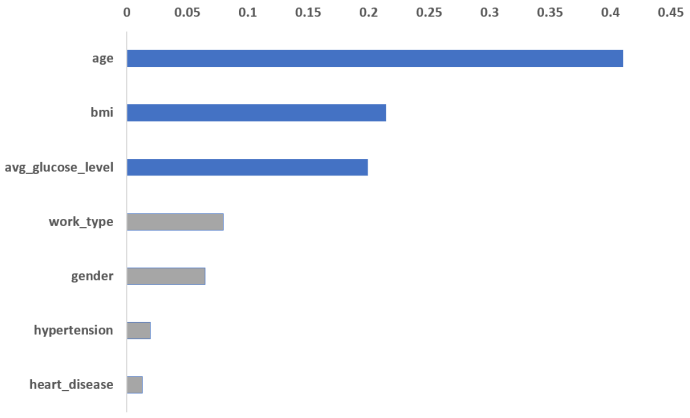
Age	BMI	Avg_glucose_level	work_type	gender	hypertension	Heart_disease

- Select your instance of interest.
- Perturb your dataset and get the black box predictions for these new points
- Weight the new samples according to their proximity to the instance of interest
- Train a weighted, interpretable model on the dataset with the variations
- Explain the prediction by interpreting the local model.



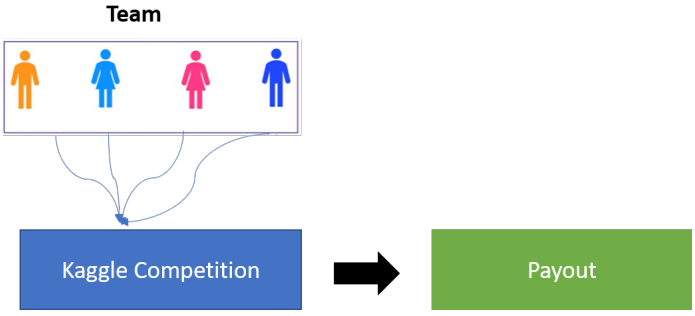
New dataset
Labels: Prediction of complex models
Features: Newly generated datapoints

Understanding Feature Importance using LIME



Feature Importance: Predicting stroke using LIME

Shapley values

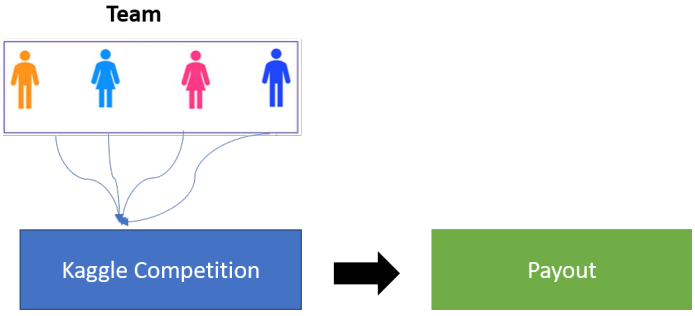


Example: Shapley value

Source: Molnar 2018 [13]

Shapley values

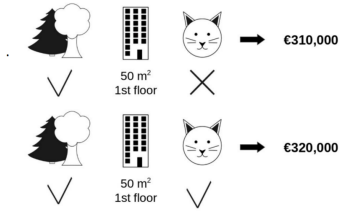
A concept from coalitional game theory – Tell us how to fairly distribute the “payout” among the players (=features).



Example: Shapley value

Source: Molnar 2018 [13]

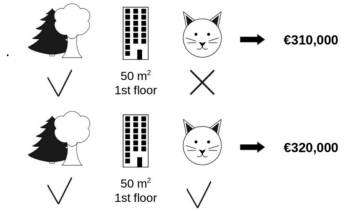
Local Explanation from Shapley values



Prediction with and without “Cat Banned” feature

Local Explanation from Shapley values

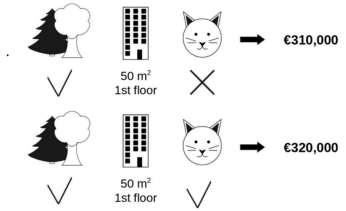
- For one chosen feature, replace its value by the value of a randomly chosen (training) data instance
- Observe how the prediction changes



Prediction with and without “Cat Banned” feature

Local Explanation from Shapley values

- For one chosen feature, replace its value by the value of a randomly chosen (training) data instance
- Observe how the prediction changes
- Repeat this for all possible coalitions of the other features

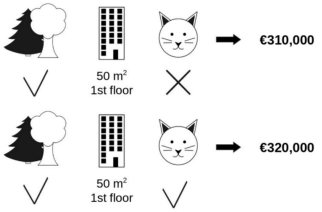


Prediction with and without “Cat Banned” feature

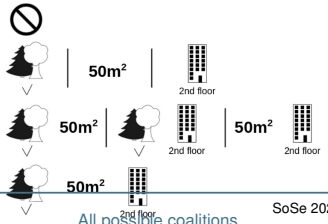
Local Explanation from Shapley values

- For one chosen feature, replace its value by the value of a randomly chosen (training) data instance
- Observe how the prediction changes
- Repeat this for all possible coalitions of the other features

→ Predict the outcome "with" and "without" the feature of interest



Prediction with and without “Cat Banned” feature



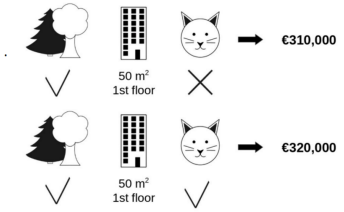
All possible coalitions

Local Explanation from Shapley values

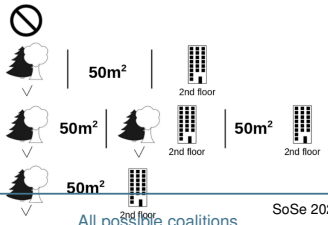
- For one chosen feature, replace its value by the value of a randomly chosen (training) data instance
- Observe how the prediction changes
- Repeat this for all possible coalitions of the other features

→ Predict the outcome "with" and "without" the feature of interest

- Features that are not in a coalition are set to values of randomly chosen data instances



Prediction with and without “Cat Banned” feature



All possible coalitions

Calculating Shapley Values

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!\,(M - |z'| - 1)!}{M!} \left[f_x\left(z'\right) - f_x\left(z' \backslash i\right) \right]$$

X=	Age=56	BMI=30	Avg_glucose_level=100	Work_type=Doctor	Gender=F	Hypertension=No	Heart_disease=Yes

Calculating Shapley Values

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!\,(M - |z'| - 1)!}{M!} \left[f_x\left(z'\right) - f_x\left(z' \backslash i\right) \right]$$

X= | Age=56 | BMI=30 | Avg_glucose_level=100 | Work_type=Doctor | Gender=F | Hypertension=No | Heart_disease=Yes |

where:

- ϕ_i - Shapley value for feature *i*
- *f* - Black-box model

Source: Lundberg and Lee 2017 [24]

Calculating Shapley Values

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!\,(M - |z'| - 1)!}{M!} \left[f_x\left(z'\right) - f_x\left(z' \backslash i\right) \right]$$

X= | Age=56 | BMI=30 | Avg_glucose_level=100 | Work_type=Doctor | Gender=F | Hypertension=No | Heart_disease=Yes |

where:

- ϕ_i - Shapley value for feature *i*
- *f* - Black-box model

Source: Lundberg and Lee 2017 [24]

Calculating Shapley Values

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!\left(M - |z'| - 1\right)!}{M!}\left[f_x\left(z'\right) - f_x\left(z' \backslash i\right)\right]$$

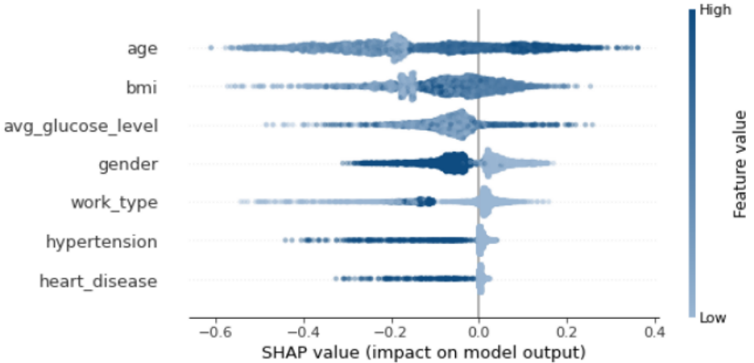
X=	Age=56	BMI=30	Avg_glucose_level=100	Work_type=Doctor	Gender=F	Hypertension=No	Heart_disease=Yes

where:

- ϕ_i - Shapley value for feature i
- f - Black-box model
- $|z'|$ - Number of non-zero entries in z'
- $z' \subseteq x'$ - All z' vectors where the non-zero entries are a subset of the non-zero entries in x'
- M - Total no. of features

Source: Lundberg and Lee 2017 [24]

Example of a Local Explanation Based on Shapley Values



SHAP summary plot

Source: https://www.kaggle.com/code/joshuaswords/predicting-a-stroke-shap-lime-explainer-eli5/notebook

Neural Network Interpretation

- Why do we need interpretability methods specific to Neural Networks?

Neural Network Interpretation

- Why do we need interpretability methods specific to Neural Networks?
  - Neural networks learn features and concepts in their hidden layers
  - NNs are differentiable – gradient information is useful for interpretation

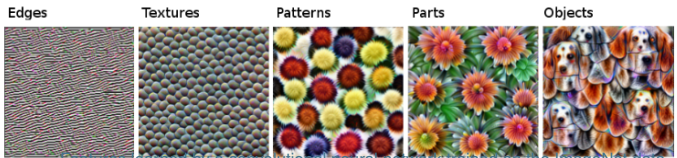
Neural Network Interpretation

- Why do we need interpretability methods specific to Neural Networks?
  - Neural networks learn features and concepts in their hidden layers
  - NNs are differentiable – gradient information is useful for interpretation

- Some common methods are [13]:
  - **Learned Features**: What features has the NN learned?
  - **Saliency Maps**: How did each pixel contribute to a particular prediction?
  - **Concepts**: Which more abstract concepts has the NN learned?
  - **Adversarial Examples**: Can we trick the network?

Learned Features

- Convolutional neural networks learn abstract features and concepts from raw image pixels

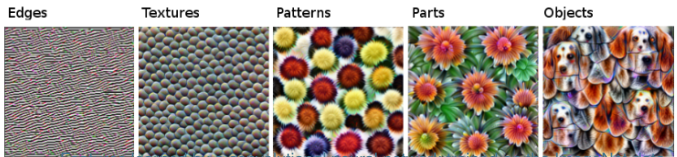


Features learned by a convolutional neural network trained on the ImageNet data

molnar2022

Learned Features

- Convolutional neural networks learn abstract features and concepts from raw image pixels



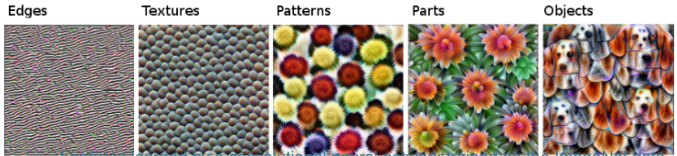
Features learned by a convolutional neural network trained on the ImageNet data

- **Feature Visualization** visualizes the learned features by activation maximization

molnar2022

Learned Features

- Convolutional neural networks learn abstract features and concepts from raw image pixels



Features learned by a convolutional neural network trained on the ImageNet data

- **Feature Visualization** visualizes the learned features by activation maximization
- **Network Dissection** labels neural network units (e.g. channels) with human concepts

molnar2022

From Visualization to Interpretation

The channels of a convolutional neural network learn new features and Feature Visualization helps visualize those features.

From Visualization to Interpretation

The channels of a convolutional neural network learn new features and Feature Visualization helps visualize those features.

BUT...

From Visualization to Interpretation

The channels of a convolutional neural network learn new features and Feature Visualization helps visualize those features.

BUT...

- What concepts have been learned inside?
- How well can a unit detect specific objects, e.g., a cat ?
- How to compare the internal representations?

Network Dissection

- The Network Dissection quantifies the interpretability of a unit of a convolutional neural network

David Bau et al. “Network Dissection: Quantifying Interpretability of Deep Visual Representations”. In: *CoRR* abs/1704.05796 (2017). arXiv: 1704.05796

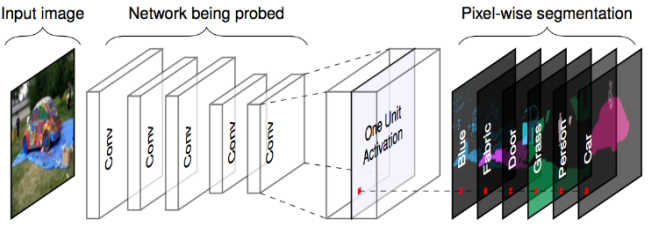
Network Dissection

- The Network Dissection quantifies the interpretability of a unit of a convolutional neural network
- Network Dissection has three steps:

David Bau et al. “Network Dissection: Quantifying Interpretability of Deep Visual Representations”. In: *CoRR* abs/1704.05796 (2017). arXiv: 1704.05796

Network Dissection

- The Network Dissection quantifies the interpretability of a unit of a convolutional neural network
- Network Dissection has three steps:
  - Identify a broad set of human-labeled visual concepts

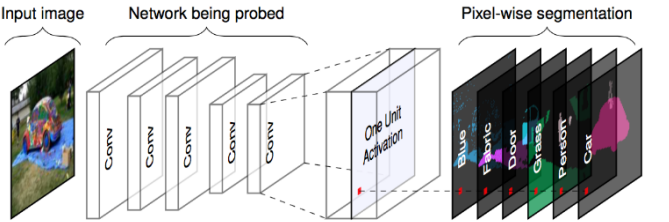


Network Dissection Algorithm

David Bau et al. "Network Dissection: Quantifying Interpretability of Deep Visual Representations". In: *CoRR* abs/1704.05796 (2017). arXiv: 1704.05796

Network Dissection

- The Network Dissection quantifies the interpretability of a unit of a convolutional neural network
- Network Dissection has three steps:
  - Identify a broad set of human-labeled visual concepts
  - Gather hidden variables' response to known concepts

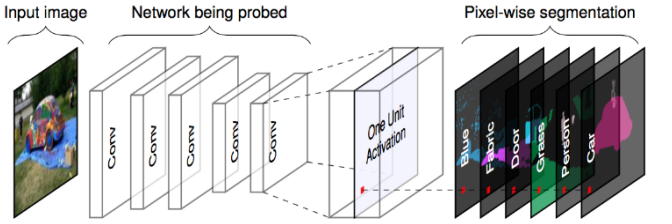


Network Dissection Algorithm

David Bau et al. "Network Dissection: Quantifying Interpretability of Deep Visual Representations". In: *CoRR* abs/1704.05796 (2017). arXiv: 1704.05796

Network Dissection

- The Network Dissection quantifies the interpretability of a unit of a convolutional neural network
- Network Dissection has three steps:
  - Identify a broad set of human-labeled visual concepts
  - Gather hidden variables' response to known concepts
  - Quantify alignment of hidden variable-concept pairs



Network Dissection Algorithm

David Bau et al. "Network Dissection: Quantifying Interpretability of Deep Visual Representations". In: *CoRR* abs/1704.05796 (2017). arXiv: 1704.05796

Scoring Unit Interpretability

Intersection over Union $IoU_{k,c}$ can be interpreted as the accuracy with which unit k detects concept c.

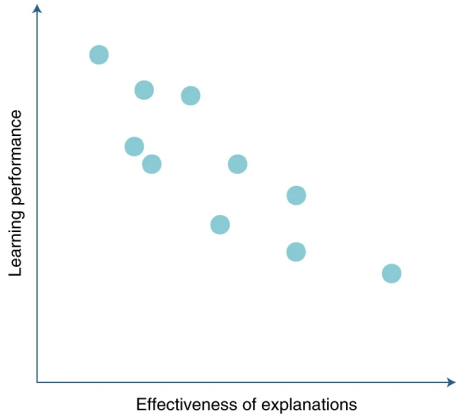
$$\mathbf{IoU_{k,c}} = \frac{\sum |\mathbf{M_k}(\mathbf{x}) \cap \mathbf{L_c}(\mathbf{x})|}{\sum |\mathbf{M_k}(\mathbf{x}) \cup \mathbf{L_c}(\mathbf{x})|}$$

$|.|$ denotes cardinality of a set

IOU is computed by comparing the human ground truth annotation and the top activated pixels

Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* 2nd. USA: Basic Books, Inc., 2018

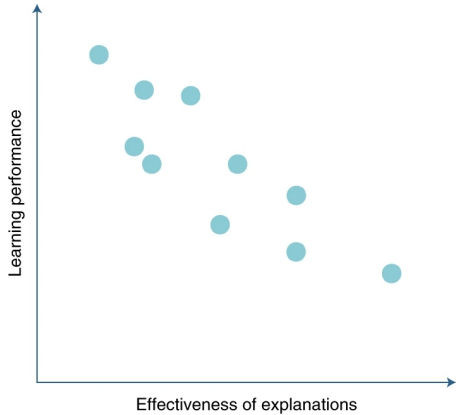
Back to the initial assumption: Accuracy vs. explainability trade-off



How to deal with this trade-off?

Making inherently interpretable models more accurate.

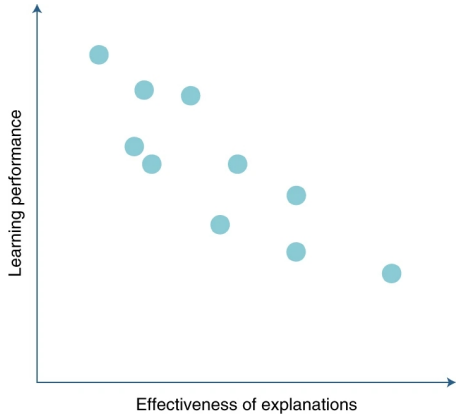
Back to the initial assumption: Accuracy vs. explainability trade-off



How to deal with this trade-off?

Making inherently interpretable models more accurate.

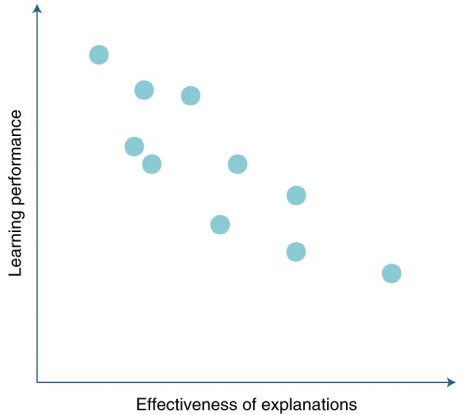
Back to the initial assumption: Accuracy vs. explainability trade-off



How to deal with this trade-off?

Making inherently interpretable models more accurate.

Back to the initial assumption: Accuracy vs. explainability trade-off



How to deal with this trade-off?

Making inherently interpretable models more accurate.

OR

Generating good explanations for accurate black-box models. → But is this truly the trade-off?

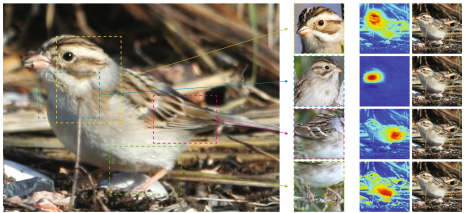
Source: From [15], in turn from [19]

Focussing on inherently interpretable models

- Many “explanation methods” don’t show us actual explanation (just “focus”)
- Confirmation bias for correct predictions
- Black box models in prevent collaboration and flexibility in many cases

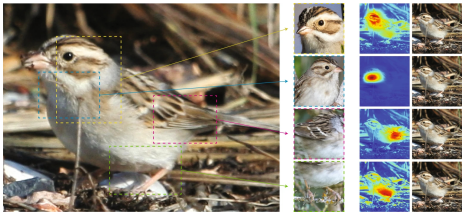
Two recommendations

- Cynthia Rudin [15]: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead



Two recommendations

- Cynthia Rudin [15]: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead



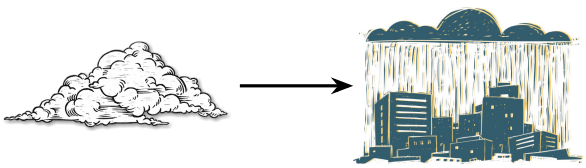
- Chen et al. [14]: Explainable medical imaging AI needs human-centered design
  - Explanations / interpretability always has a use-case
  - There is not one fits-it-all explanation method
  - Human-in-the-loop approaches require proactive design

Summary (for this part)

- Interpretability and explainability catalyze the adoption of machine learning models
- Strong white-box models can be difficult to create especially in Computer Vision and Natural Language Processing
- Many local methods are only model-agnostic and they can be applied to any kind of data
- Potentially more powerful exploratory research with black box models... potentially
- Human-centered design essential for suitable explanations (+ interaction) models

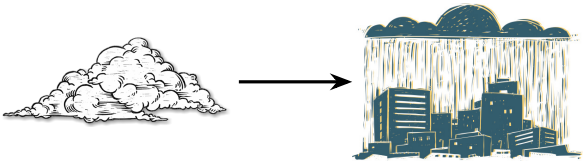
Causality

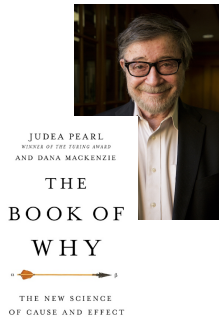
Causality explains the cause of an effect



Causality

Causality explains the cause of an effect





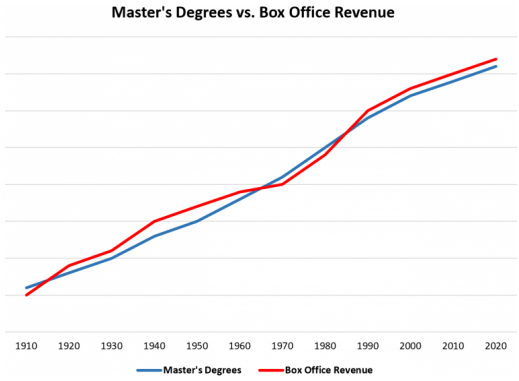
"All the impressive achievements of deep learning amount to just curve fitting."[1]

- Judea Pearl

Source: http://bayes.cs.ucla.edu/jp_home.html

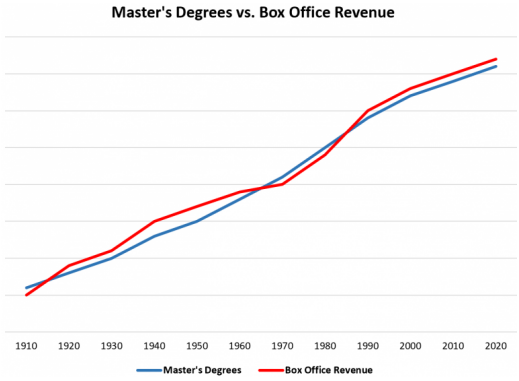
[1] Judea Pearl interview with Kevin Hartnett for Quantamagazine. *To Build Truly Intelligent Machines, Teach Them Cause and Effect*. 2018

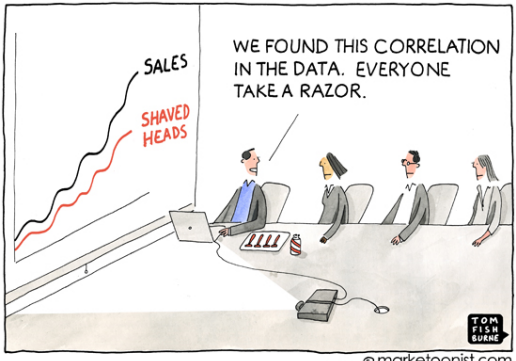
Correlation is not Causation!



Sources:
https://www.statology.org/correlation-does-not-imply-causation-examples/
https://www.statology.org/correlation-does-not-imply-causation-examples/
https://www.marketoonitst.com

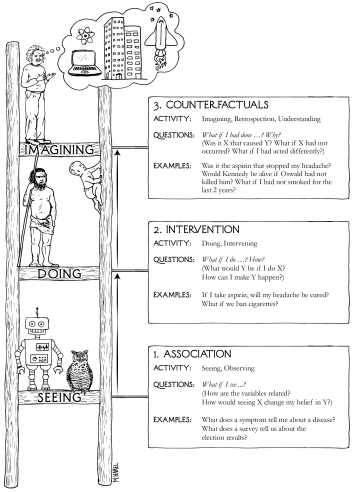
Correlation is not Causation!





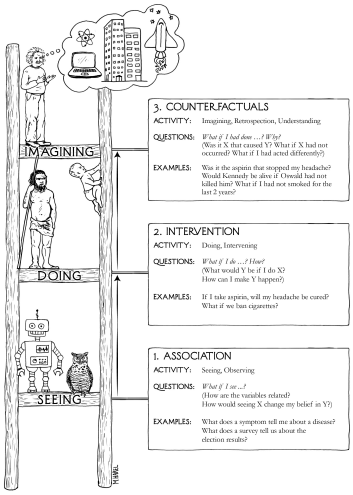
Sources:
https://www.statology.org/correlation-does-not-imply-causation-examples/
https://www.statology.org/correlation-does-not-imply-causation-examples/
https://www.marketoonitst.com

Ladder of Causation



Source: Pearl and Mackenzie 2018 [8]

Ladder of Causation



1. ASSOCIATION

ACTIVITY: Seeing, Observing

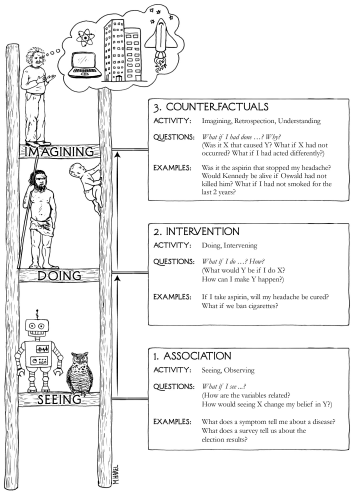
QUESTIONS: *What if I see ...?*
(How are the variables related?
How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease?
What does a survey tell us about the election results?

What if it's dawn?
I should wake up, get ready, and goto work.

Source: Pearl and Mackenzie 2018 [8]

Ladder of Causation



2. INTERVENTION

ACTIVITY: Doing, Intervening

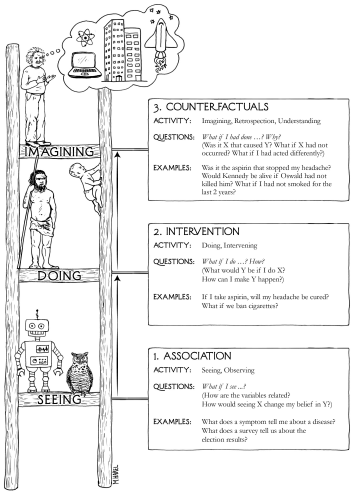
QUESTIONS: *What if I do …? How?*
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?

What if I wake up late at 10AM?

Source: Pearl and Mackenzie 2018 [8]

Ladder of Causation



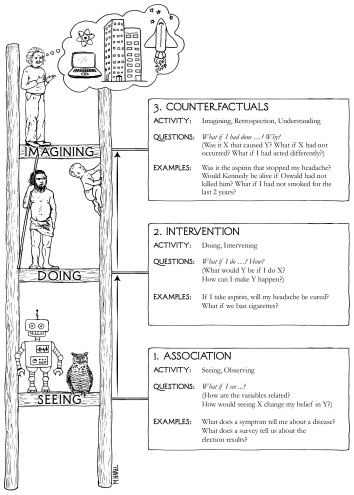
3. COUNTERFACTUALS

ACTIVITY:	Imagining, Retrospection, Understanding
QUESTIONS:	*What if I had done …? Why?* (Was it X that caused Y? What if X had not occurred? What if I had acted differently?)
EXAMPLES:	Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

Today I woke up late and the day was bad.
Would my day have been better if I had woken up early at 5AM?

Source: Pearl and Mackenzie 2018 [8]

Ladder of Causation



Statistical models can only do ASSOCIATION

Only **Statistical Inference**

$P(Fire \mid Smoke)$ $P(Smoke \mid Fire)$

No **Causal Inference**

Fire causes Smoke ~~Smoke causes Fire~~

Source: Pearl and Mackenzie 2018 [8]

Current ML Models

Hugely successful because of

- massive amount of data
- high number of parameters
- high performance computing
- independent and identically distributed (i.i.d.) assumption

Current ML Models

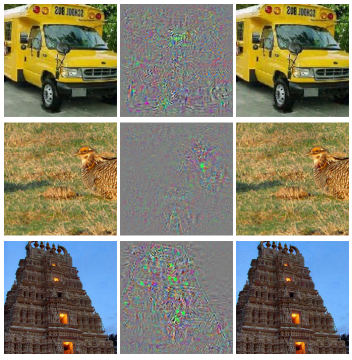
Hugely successful because of

- massive amount of data
- high number of parameters
- high performance computing
- independent and identically distributed (i.i.d.) assumption

☞ In real world, distribution shifts are very common and it is very easy to violate i.i.d. assumption

Adverserial Vulnerability

Violation of i.i.d. assumption ➔ Adverserial vulnerability ➔ Failure in ML models



correct prediction | difference | wrong prediction

Source: Szegedy et al. 2013 [10]

Towards Causality

Causal Learning will help ML models to

Towards Causality

Causal Learning will help ML models to

- Plan and act according to interventions → Robust to distribution shifts

Towards Causality

Causal Learning will help ML models to

- Plan and act according to interventions → Robust to distribution shifts
- Think and reason using counterfactual questions → Learn unseen data

Towards Causality

Causal Learning will help ML models to

- Plan and act according to interventions → Robust to distribution shifts
- Think and reason using counterfactual questions → Learn unseen data

Help towards important goals of ML:

→ Generalization

→ Transfer Learning

→ Explainability

Common Cause Principle

- Explains the connection between causality and statistical dependence
- Postulated by Hans Reichenbach in 1956

Given two observations X and Y are statistically dependent,

$$\text{i.e., } p(X, Y) \neq p(X)\, p(Y)$$

then $\exists Z$ which causally influences both

$$\text{s.t., } X \perp Y \mid Z$$

$$\text{i.e., } p(X, Y|Z) = p(X|Z)\, p(Y|Z)$$

Common Cause Principle

- Explains the connection between causality and statistical dependence
- Postulated by Hans Reichenbach in 1956

Given two observations X and Y are statistically dependent,

$$\text{i.e., } p(X, Y) \neq p(X)\, p(Y)$$

$$\text{then } \exists Z \text{ which causally influences both}$$

$$\text{s.t., } X \perp Y \mid Z$$

$$\text{i.e., } p(X, Y|Z) = p(X|Z)\, p(Y|Z)$$

If Z is not observed, then it is a **Confounder**

Causal Graphical Models (CGMs)

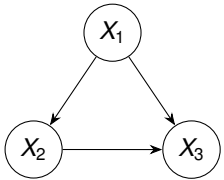
- Directed Acyclic Graphs (DAGs) or Bayesian Networks
- Satisfy **Causal Markov condition**

→ Disentangled Factorization

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid PA_i)$$

  - $X_1, \ldots, X_n$ are observations
  - PA_i are parents of X_i

→ **Independent Causal Mechanism (ICM)**



$$p(X_1, X_2, X_3) = p(X_1)\, p(X_2 \mid X_1)\, p(X_3 \mid X_1, X_2)$$

Structural Causal Models (SCMs)

SCMs: conditional probabilities replaced by functional parent-child relationships

$$X_i = f_i(PA_i, U_i), \quad for \quad i = 1, \ldots, n$$

- f_i: deterministic function
- U_i:
  - stochastic noise terms
  - jointly independent → **Causal Sufficiency**

Structural Causal Models (SCMs)

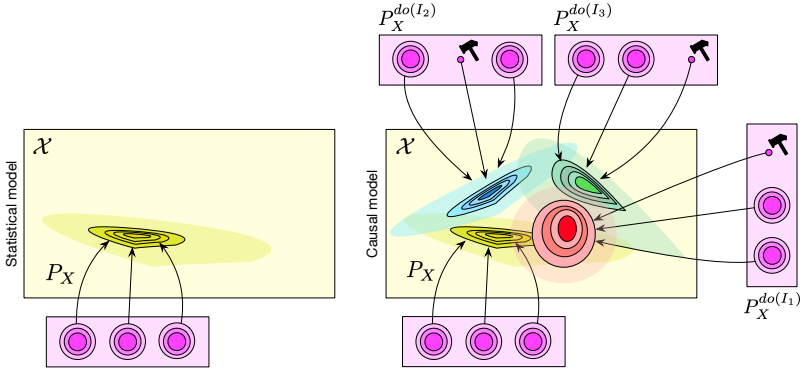
SCMs: conditional probabilities replaced by functional parent-child relationships

$$X_i = f_i(PA_i, U_i), \quad for \quad i = 1, \ldots, n$$

- f_i: deterministic function
- U_i:
  - stochastic noise terms
  - jointly independent → **Causal Sufficiency**

→ Interventions & counterfactuals possible

Statistical vs. Causal Models

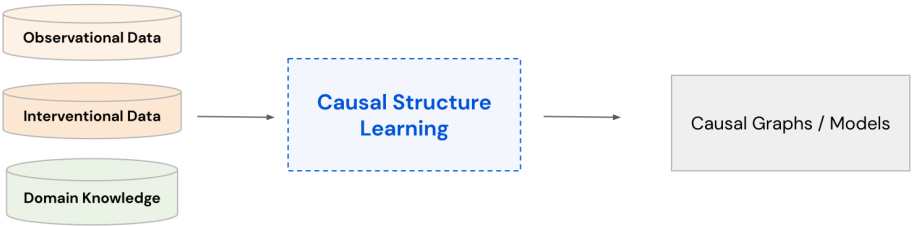


- Statistical model: can only capture one probability distribution
- Causal model: every intervention defines new joint distribution intervention

Source: Schölkopf et al. 2021 [1]

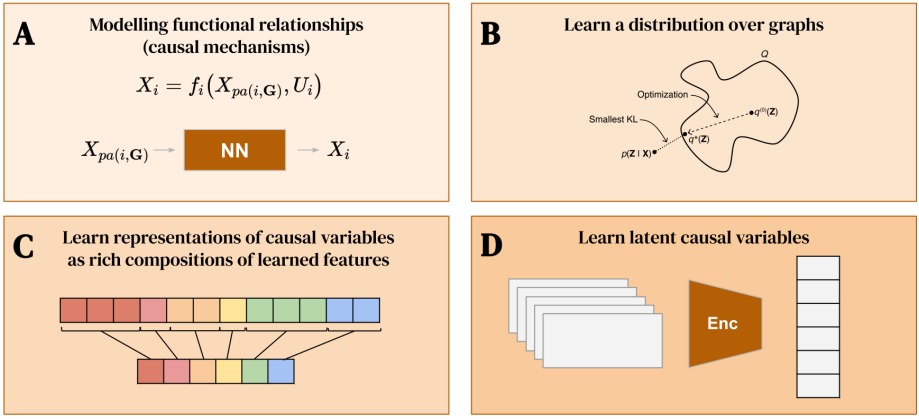
Causal Structure Learning

Inferring underlying causal graphs from data and domain knowledge



Source: https://drive.google.com/file/d/1wNyDm2j03YzVW7g8w5NkFdrMakw5lId5/view

Deep Learning for Causality



Source: https://drive.google.com/file/d/1wNyDm2j03YzVW7g8w5NkFdrMakw5lId5/view

Summary

- Interpretable
- Explainable
- Causal

... help to build **trustworthy and robust** models, might enable to explain importance of different factors

NEXT TIME

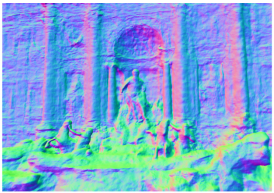
ADVANCED

ON DEEP LEARNING

Neural rendering

Training image

Reconstruction



Normals

Diffuse Albedo

Illumination

Shadows

Novel Lighting

Novel Lighting and Viewpoint

Source: Rudnev et al. 2022 [9]

Comprehensive Questions

- What are main differentiation criteria for explainable methods?
- Is there a difference between interpretability and explainability?
- Discuss explanation methods and potential use cases / examples.
- What are weaknesses of common gradient-based explanation methods?
- Discuss in which situations explanations / interpretable models are (not) needed
- Explain the basics of Causal DL

Further Reading

- Christoph Molnar: Interpretable Machine Learning: A Guide for Making Black Box Models Explainable [13]
  Available at: https://christophm.github.io/interpretable-ml-book/
- Workshop playlist for full-day workshop at Stanford on ML Explainable by Hima Lakkaraju from Harvard
  https://www.youtube.com/playlist?list=PLoROMvodv4rPh6wa6PGcHH6vMG9sEIPxL
- ICML'22 Tutorial on Causal DL by Ke and Bauer:
  https://sites.google.com/view/causalityanddeeplearning/start

References

References I

[1] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. “Toward Causal Representation Learning”. In: *Proceedings of the IEEE* 109.5 (May 2021), pp. 612–634.

[2] Tim Miller. “Explanation in Artificial Intelligence: Insights from the Social Sciences”. In: *CoRR* abs/1706.07269 (2017). arXiv: 1706.07269.

[3] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. “Examples are not enough, learn to criticize! Criticism for Interpretability”. In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc., 2016.

[4] Finale Doshi-Velez and Been Kim. “Considerations for Evaluation and Generalization in Interpretable Machine Learning”. In: *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Cham: Springer International Publishing, 2018, pp. 3–17.

[5] .*Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer International Publishing, 2018.

References II

[6] Gabriëlle Ras, Marcel van Gerven, and Pim Haselager. “Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges”. In: *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Cham: Springer International Publishing, 2018, pp. 19–36.

[7] Judea Pearl interview with Kevin Hartnett for Quantamagazine. *To Build Truly Intelligent Machines, Teach Them Cause and Effect*. 2018.

[8] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. 1st. USA: Basic Books, Inc., 2018.

[9] Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. “NeRF for Outdoor Scene Relighting”. In: *Computer Vision – ECCV 2022*. Cham: Springer Nature Switzerland, 2022, pp. 615–631.

[10] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. “Intriguing properties of neural networks”. In: (2013).

References III

[11] Julia Dressel and Hany Farid. “The accuracy, fairness, and limits of predicting recidivism”. In: *Science Advances* 4.1 (2018), eaao5580. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.aao5580.

[12] Finale Doshi-Velez and Been Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. arXiv: 1702.08608 [stat.ML].

[13] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd. USA: Basic Books, Inc., 2018.

[14] Haomin Chen, Catalina Gomez, Chien-Ming Huang, and Mathias Unberath. “Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review”. In: *npj Digital Medicine* 5.1 (Oct. 2022), p. 156.

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

References IV

[16] John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. “Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study”. In: *PLOS Medicine* 15.11 (Nov. 2018), pp. 1–17.

[17] Kasia Kulma. *Interpretable Machine Learning Using LIME Framework*. Accessed: 21.06.2023. 2017.

[18] Felipe Oviedo, Juan Lavista Ferres, Tonio Buonassisi, and Keith T. Butler. “Interpretable and Explainable Machine Learning for Materials Science and Chemistry”. In: *Accounts of Materials Research* 3.6 (2022), pp. 597–607. eprint: https://doi.org/10.1021/accountsmr.1c00244.

[19] Defense Advanced Research Projects Agency. *Broad Agency Announcement, Explainable Artificial Intelligence (XAI), DARPA-BAA-16-53*. 2016.

[20] Danielle Navarro. *12.2: Estimating a linear regression model*. Oct. 2022.

[21] Prince Yadav. *Decision tree in machine learning*. Sept. 2019.

References V

[22] Italo Jose. *KNN (K-Nearest Neighbors)*. June 2021.

[23] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *CoRR* abs/1602.04938 (2016). arXiv: 1602.04938.

[24] Scott M. Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *CoRR* abs/1705.07874 (2017). arXiv: 1705.07874.

[25] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. "Network Dissection: Quantifying Interpretability of Deep Visual Representations". In: *CoRR* abs/1704.05796 (2017). arXiv: 1704.05796.