Explainable AI

Tian Zheng

Interpretable Machine Learning

A Guide for Making Black Box Models Explainable
Christoph Molnar

2022-03-29

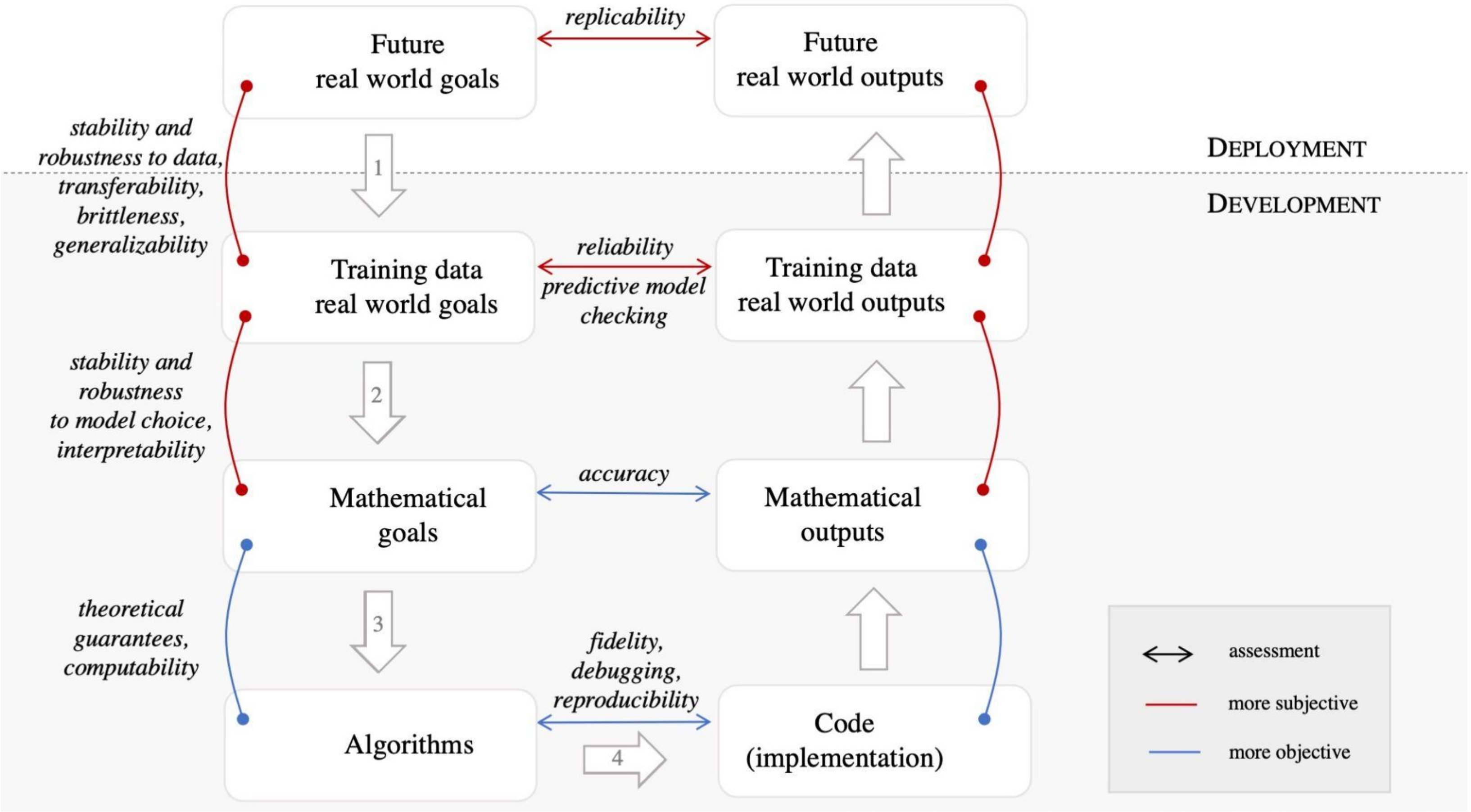
https://christophm.github.io/interpretable-ml-book/

What is Explainable AI

- A “black box” model: how to understand its properties by looking at its parameters
  - As opposed to “white box” models
  - [Recommended] diagnostics of linear models
- Machine learning algorithm is built upon data, features, learning goals, etc.
  - Interpretability, transparency
  - *“The running hypothesis is that by building more transparent, interpretable, or explainable systems, users will be better equipped to understand and therefore* ***trust*** *the intelligent agents”*
    (Miller 2019; https://doi.org/10.1016/j.artint.2018.07.007)

What is Explainable AI

- Design interpretable machine learning workflow: *how well a human could understand the decisions of the workflow*, i.e., **interpretability** or explainability
  - Consistently predict the model's result
  - Perfect accuracy is not a requirement for trust
  - Most concerning the entire model
- Create explicitly **explanation** of derived AI decisions
  - Most concerning individual model outputs

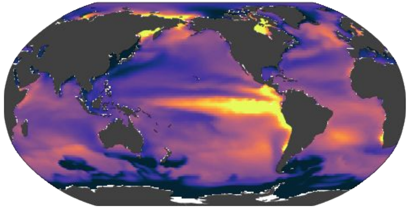


Broderick, T., Gelman, A., Meager, R., Smith, A. L., & Zheng, T. (2021).
Toward a Taxonomy of Trust for Probabilistic Machine Learning. arXiv:2112.03270

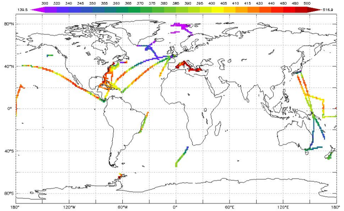
Machine learning workflows require decisions

Estimate how much carbon the ocean absorbs, at each location in space, over time, from sparse data

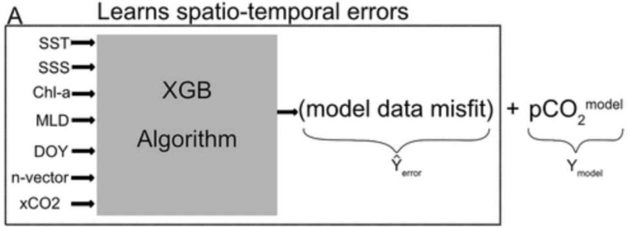
global ocean biogeochemical modelsç

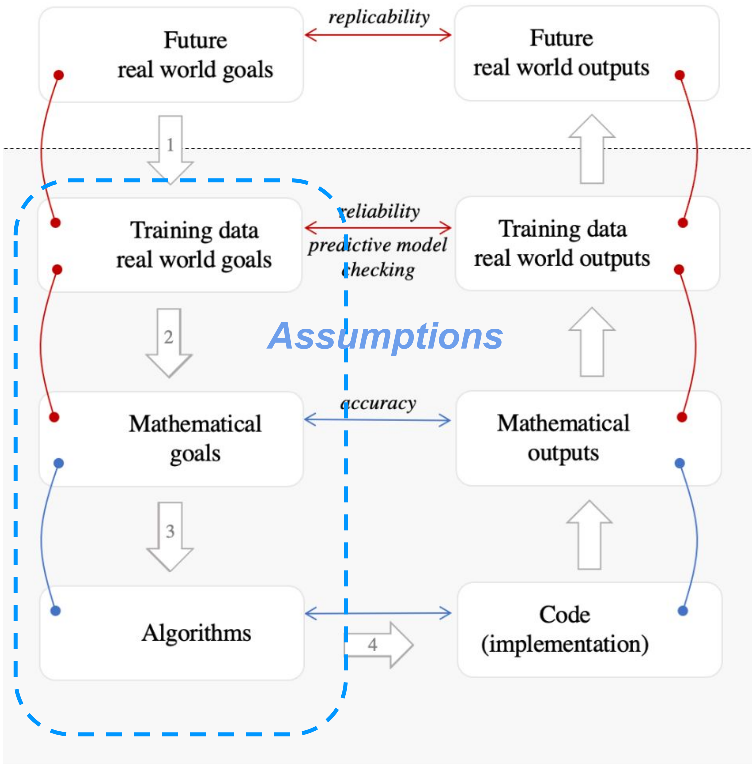


observational-based data products

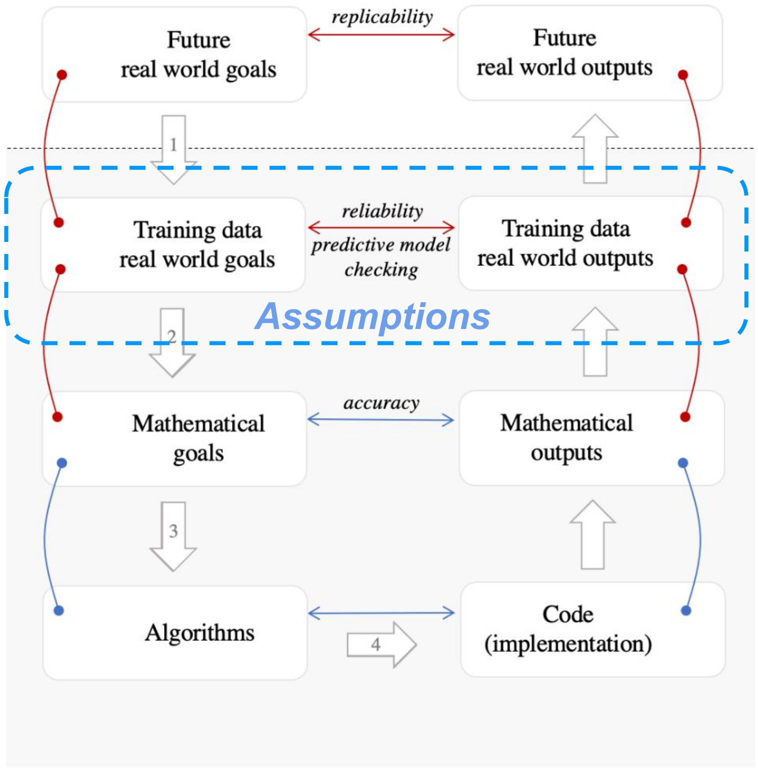


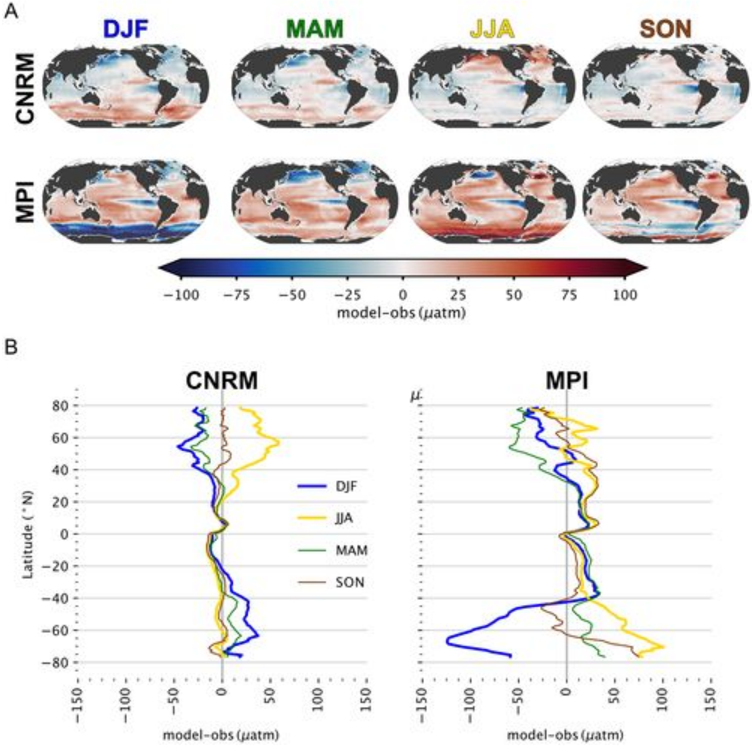
Learn a non-linear relationship between model-data mismatch and observed predictors





Interpretable results drive science forward





Interpretation methods

- Feature summaries and visualizations (e.g., partial dependence)
- Model coefficients
- Data prototypes
- Interpretable models
- Model-agnostic tools
- Local vs. global

Interpretable models - Linear Models

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad (*)$$

$$E(\varepsilon_i) = 0, \quad Var(\varepsilon_i) = \sigma^2$$

and ε_i, ε_j are uncorrelated.

$$\begin{cases} b_1 & = & \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\ b_0 & = & \frac{1}{n}\left(\sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} X_i\right) = \bar{Y} - b_1 \bar{X} \end{cases}$$

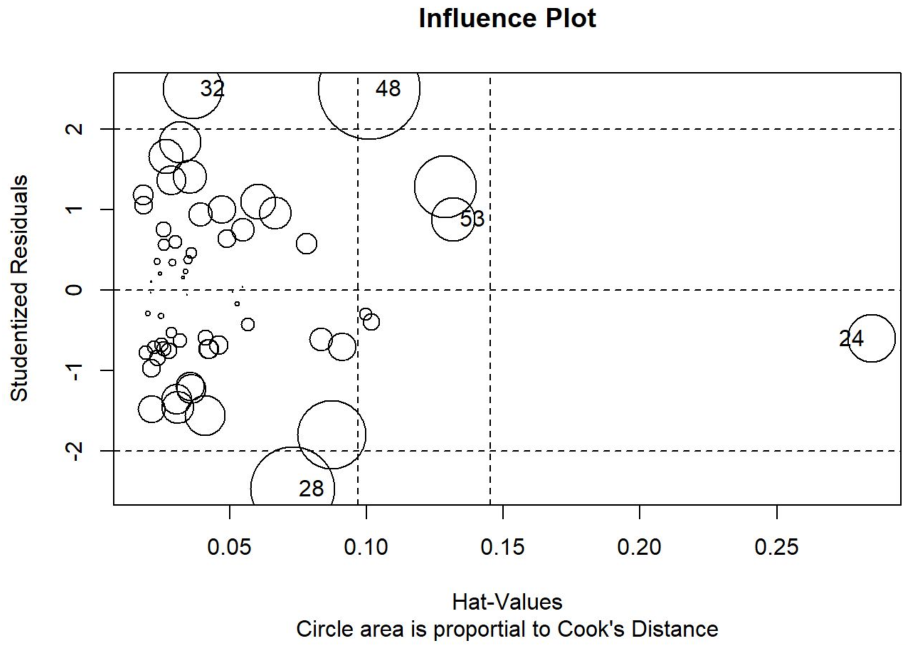
Interpretable models - Linear Models

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})\,Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \sum_{i=1}^{n} K_i Y_i, \quad \text{where } K_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

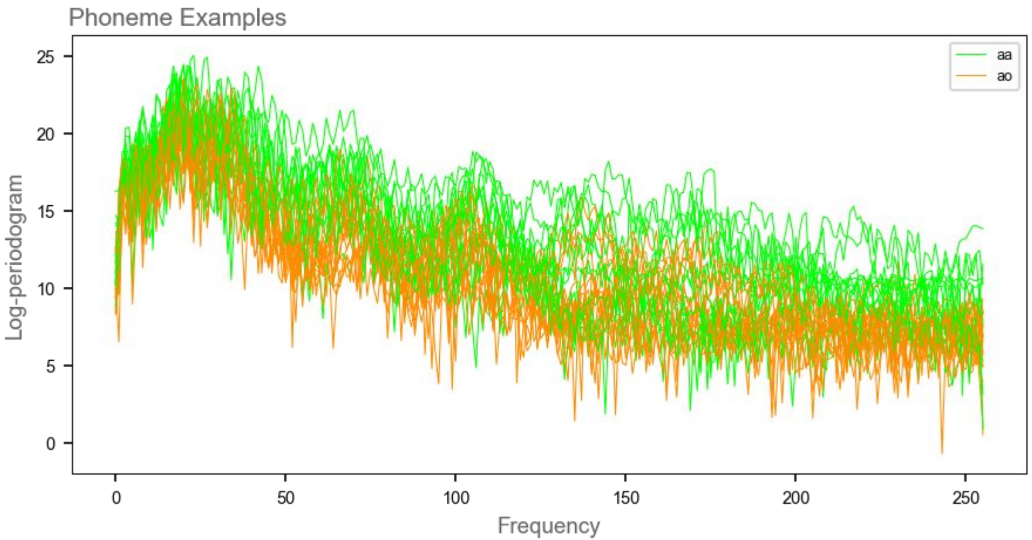
$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y},$$

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}.$$

Interpretable models - Linear Models

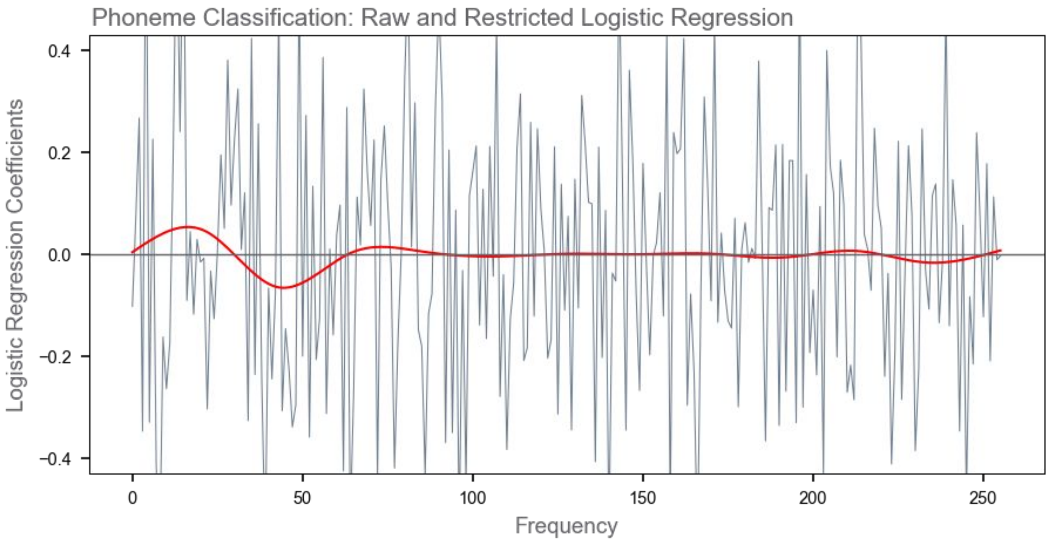


Interpretable models - Linear Models



https://github.com/empathy87/The-Elements-of-Statistical-Learning-Python-Notebooks/blob/master/examples/Phoneme%20Recognition.ipynb

Interpretable models - Linear Models



https://github.com/empathy87/The-Elements-of-Statistical-Learning-Python-Notebooks/blob/master/examples/Phoneme%20Recognition.ipynb

Interpretation Tools - Shapley Values

- ❑ Model agnostic
- ❑ “How much has each feature value contributed to the prediction?”
- ❑ The Shapley value, for assigning payouts to players depending on their contribution to the total payout.
  - ❑ “Game” - the prediction task for one instance
  - ❑ “Gain” - the actual prediction for this instance minus the average prediction for all instances.
  - ❑ “Players” - the feature values of the instance

Interpretation Tools - Shapley Values

- The Shapley value is the average of all the marginal contributions to all possible "coalitions".
- The values of features that are not in a coalition are replaced by values randomly drawn from observed data.

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(val\left(S \cup \{j\}\right) - val(S) \right)$$

$$val_x(S) = \int \hat{f}\,(x_1, \ldots, x_p)\, d\mathbb{P}_{x \notin S} - E_X(\hat{f}\,(X))$$

Interpretation Tools - Shapley Values

Approximate Shapley estimation for single feature value:

- Output: Shapley value for the value of the j-th feature
- Required: Number of iterations M, instance of interest x, feature index j, data matrix X, and machine learning model f
    - For all m = 1,...,M:
        - Draw random instance z from the data matrix X
        - Choose a random permutation o of the feature values
        - Order instance x: $x_o = (x_{(1)}, \ldots, x_{(j)}, \ldots, x_{(p)})$
        - Order instance z: $z_o = (z_{(1)}, \ldots, z_{(j)}, \ldots, z_{(p)})$
        - Construct two new instances
            - With j: $x_{+j} = (x_{(1)}, \ldots, x_{(j-1)}, x_{(j)}, z_{(j+1)}, \ldots, z_{(p)})$
            - Without j: $x_{-j} = (x_{(1)}, \ldots, x_{(j-1)}, z_{(j)}, z_{(j+1)}, \ldots, z_{(p)})$
        - Compute marginal contribution: $\phi_j^m = \hat{f}(x_{+j}) - \hat{f}(x_{-j})$
- Compute Shapley value as the average: $\phi_j(x) = \frac{1}{M}\sum_{m=1}^{M} \phi_j^m$

$$\hat{\phi}_j = \frac{1}{M}\sum_{m=1}^{M}\left(\hat{f}\left(x_{+j}^m\right) - \hat{f}\left(x_{-j}^m\right)\right)$$

Interpretation Tools - Shapley Values

- ❑ Desirable properties and theory
- ❑ Computational intensive
- ❑ Can still be misinterpreted
- ❑ Need access to the data
- ❑ Can still ignore innate correlations between features