Machine learning

Javier Garcia-Bernardo
Assistant Professor, Department of Methodology & Statistics
j.garciabernardo@uu.nl

Machine learning

”A computer program is said **to learn from experience *E*** with respect to some class of **tasks *T*** and **performance measure *P*** if its performance at task *T*, as measured by *P*, improves with experience *E*.” (Samuel/Mitchell, 1959)

- Experience: Data
- Task: Goal
- Performance measure: Accuracy, R^2, etc

Machine learning

Supervised learning: Output is available. Performance = discrepancy between predicted output and real output

- Regression
- Classification

Unsupervised learning: No labels/output. Performance = reduction of some error

- Clustering (e.g. cluster points so they are as close as possible within clusters, as far as possible between clusters)
- Dimensionality reduction (e.g. combine variables to maximize the amount of variability explained)

Machine learning

- Typically focuses on large, **high-dimensional** datasets with interactions between features

- **Output-driven**:
  - Typically aims to solve a problem (rather than to test a hypothesis)
  - Emphasizes predictive accuracy: Uses theory (to build new features) if it improves accuracy

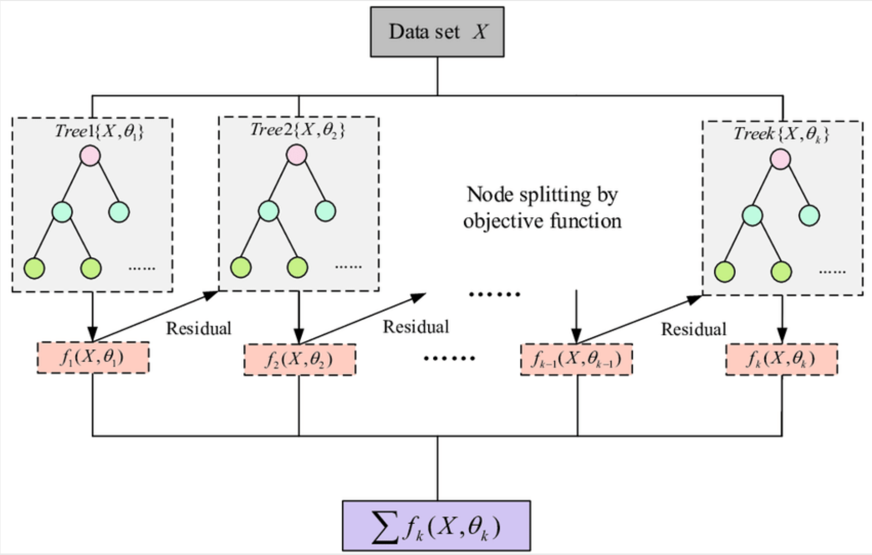
Cannot and should not replace thinking about causation

Algorithm of the day: XGBoost

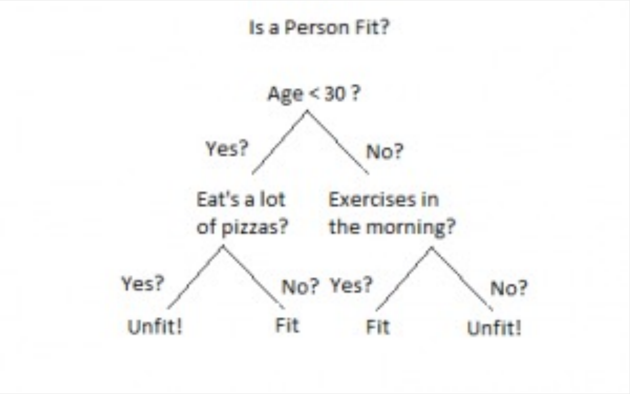
Army of weak learners → Strong predictors (wisdom of the crowd)

Boosting: Incrementally build trees to fix observations previously miscategorized

Boosting

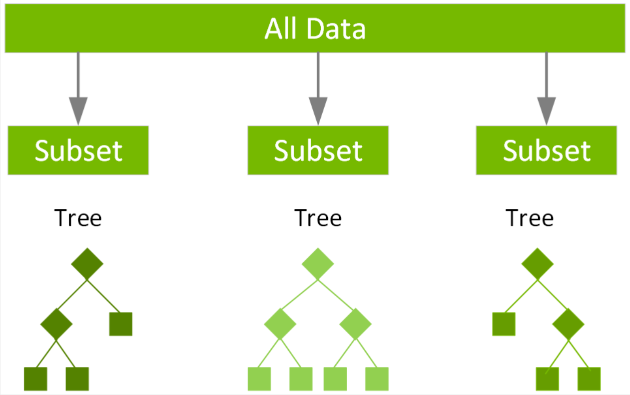


Decision tree



Xoriant.com

Ensemble (e.g. random forest)



NVIDIA

Guo et al, 2020

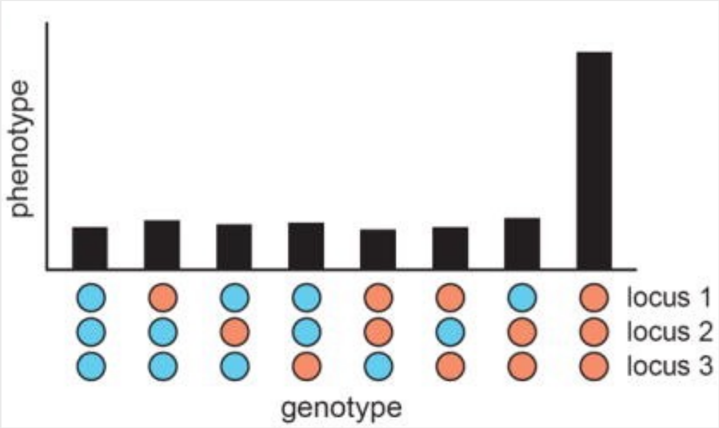
Some uses of ML in epidemiology

Missing data imputation

Prediction of outcomes with interpretability at the individual level (e.g., LIME, SHAP)

Theory building:

- ML can detect higher-order interactions and other complicated responses
- Run the ML model on the full data → Does it increase prediction compared to the traditional model?
    - You may be missing an important variable
    - Your model may be missing interactions
- Evaluate model using interpretability tools



Main issues in Machine Learning

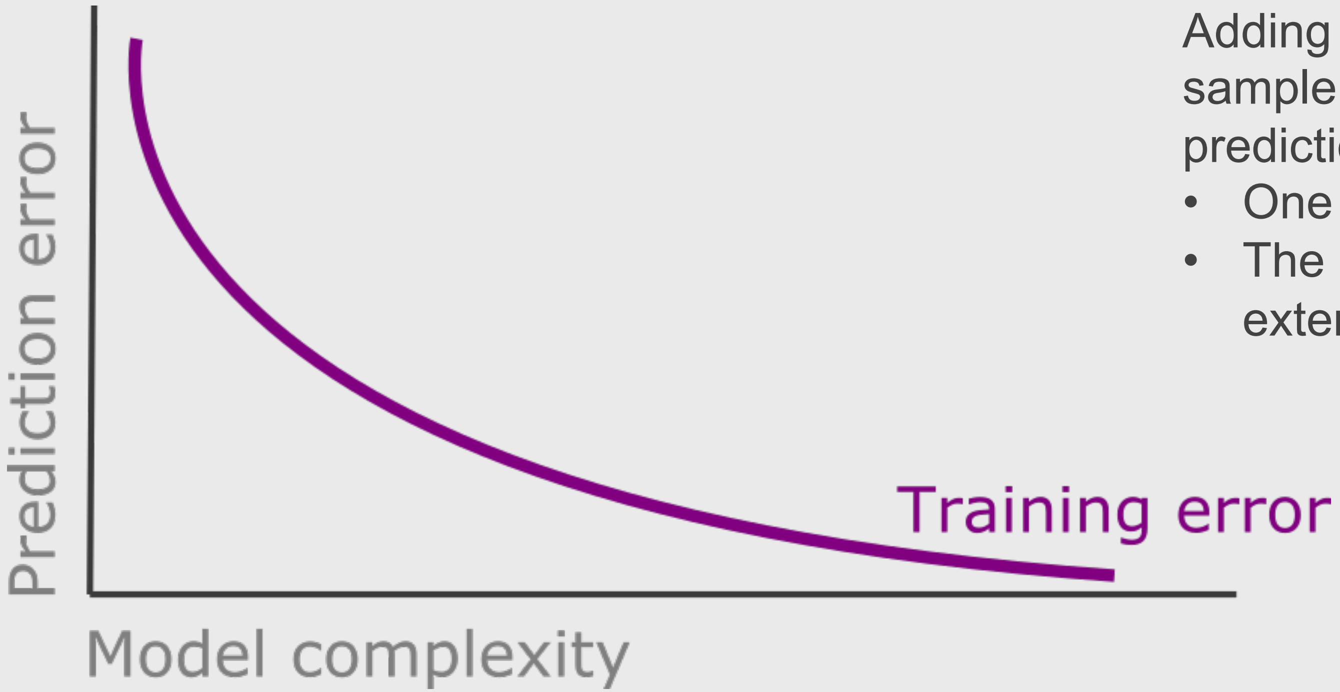
Issue 1: Overfitting

- Lots of data and features → can be a recipe for disaster
- Evaluation of overfitting: cross-validation
- Prevention of overfitting: Regularization, weak learners, dropout, etc

Issue 2: Interpretability of the model

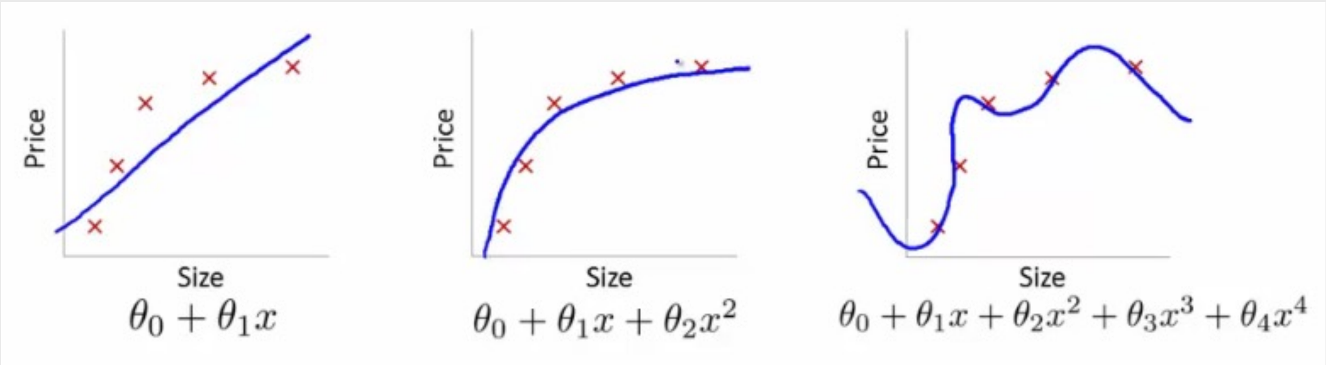
- Complex models are more difficult to interpret
- New measures of interpretability

Issue 1: Overfitting



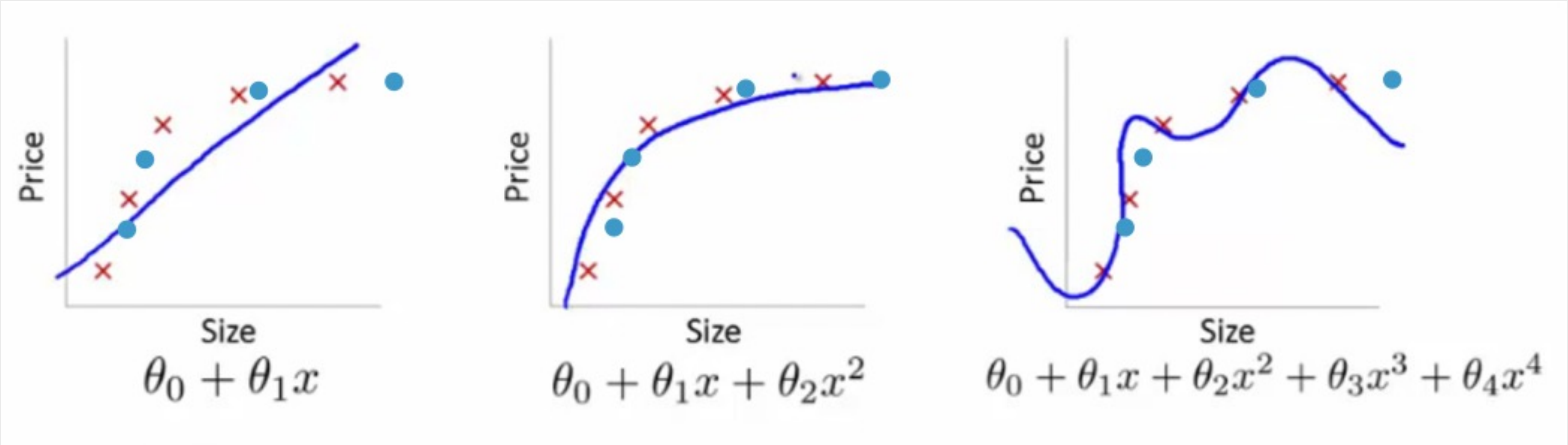
Adding more variables always decreases R^2 (within-sample prediction error). How to estimate real prediction error?

- One option: adjust for the degrees of freedom
- The “ML” option: evaluate prediction accuracy in an external dataset

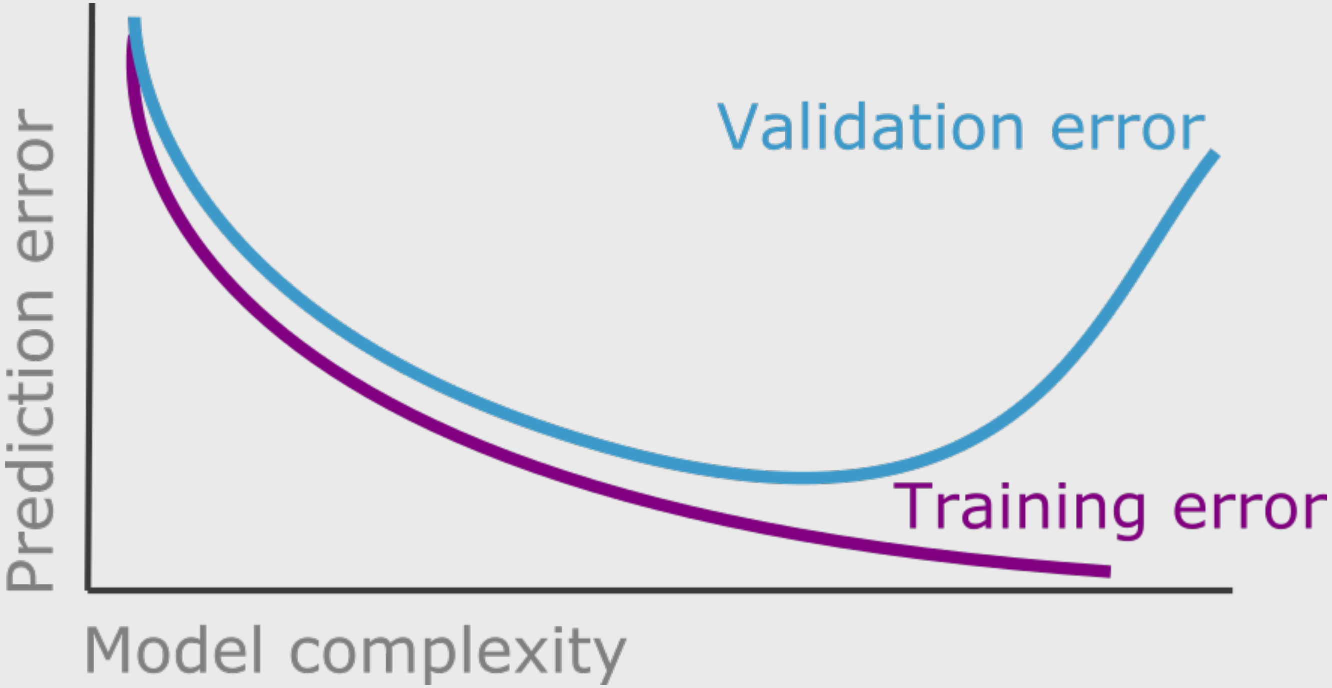


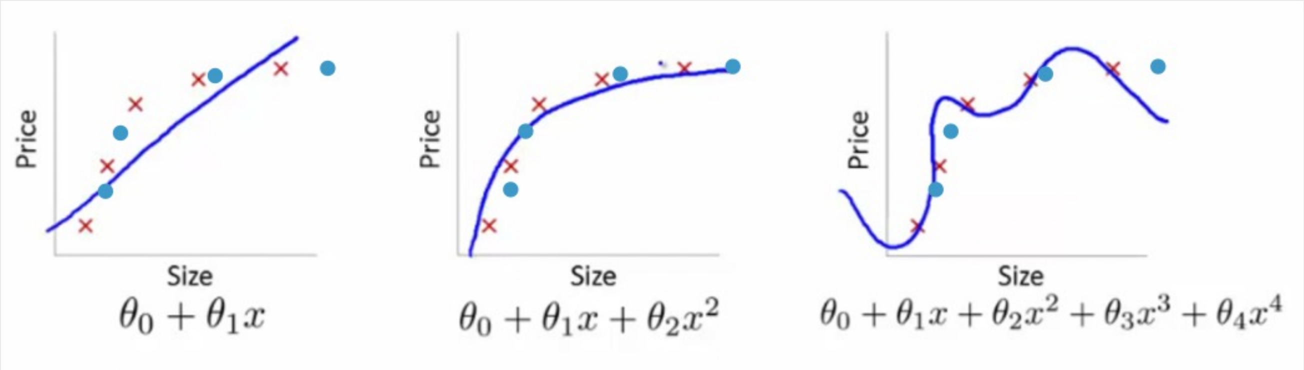
1: Evaluate overfitting using a validation dataset

Training dataset → Use to train different models
Validation dataset → Evaluate out-of-sample prediction error
(Test dataset) → Evaluate out-of-sample prediction error of final model



1: Evaluate overfitting using a validation dataset



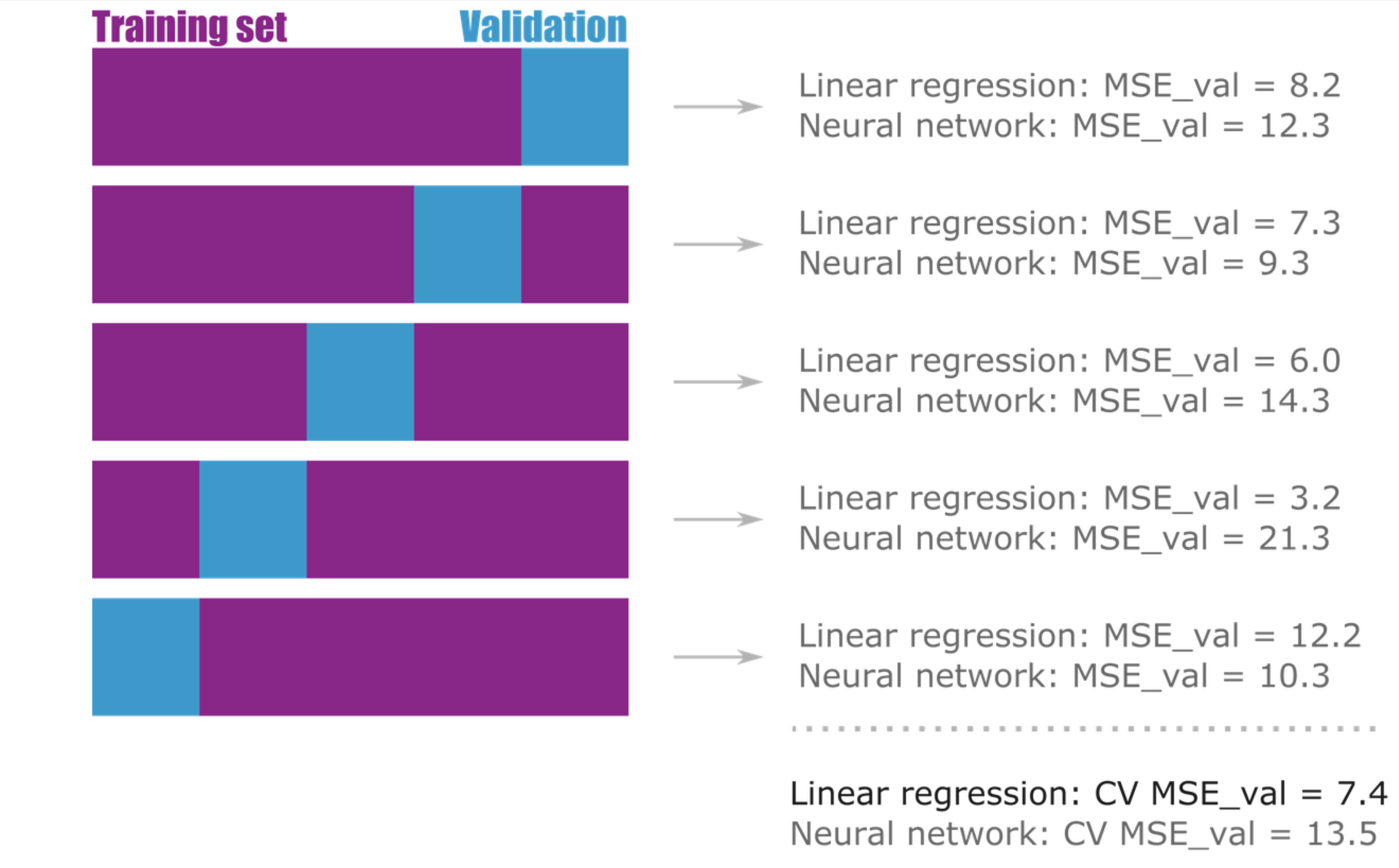


But with this:

- We reduce the training dataset (number of observations)
- We validate on a small dataset (maybe not representative)

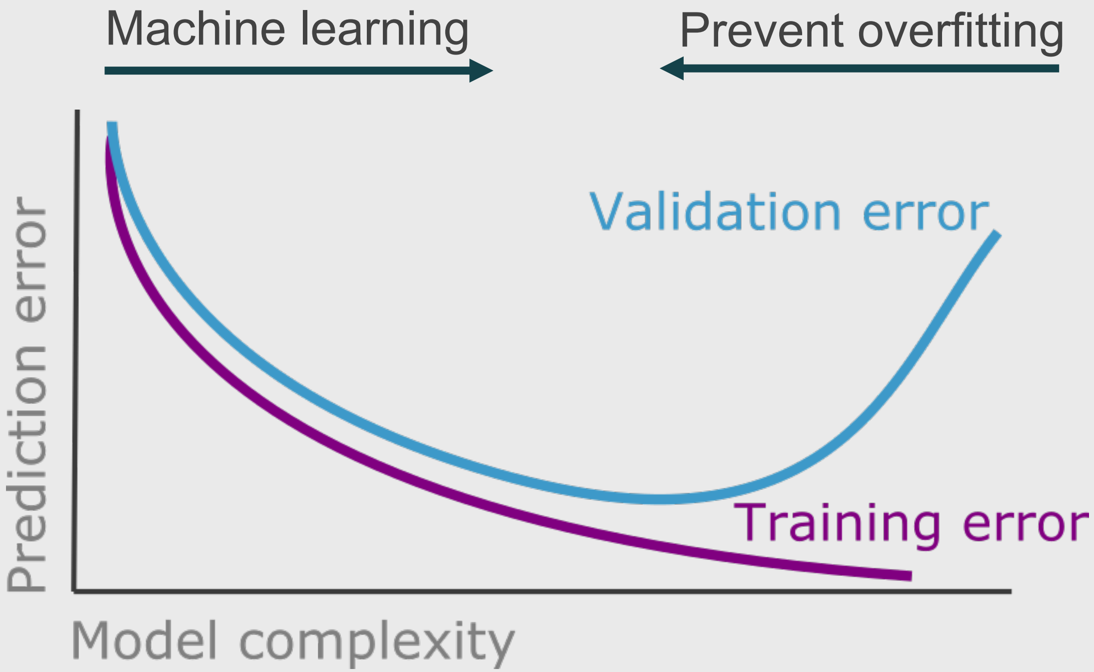
Solution → Cross-validation

1: Evaluate overfitting using cross-validation



Do you want to understand the error due to the splitting? → Run this procedure several times with random splittings

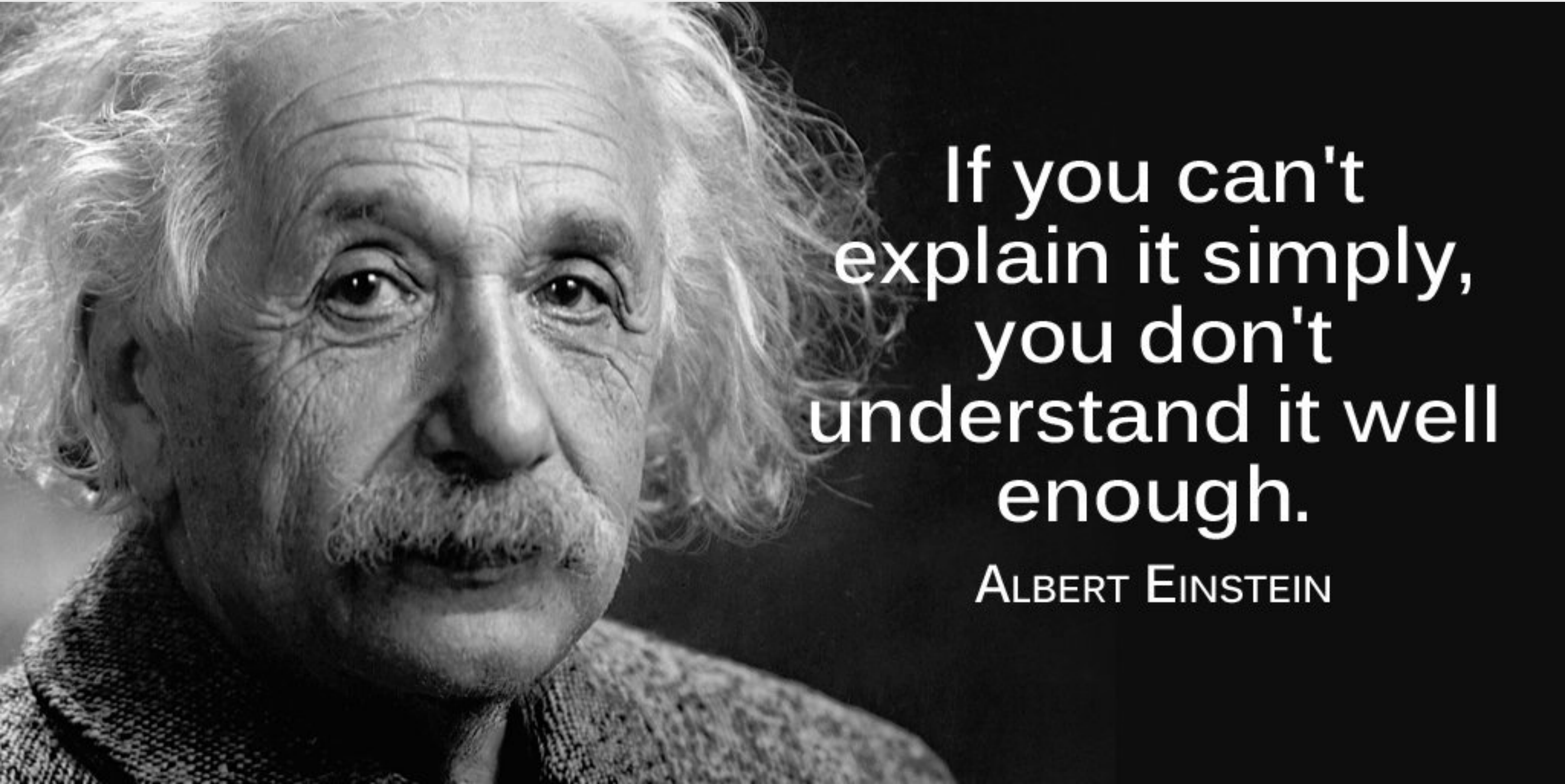
1: Hyperparameter tuning



Hyperparameter tuning using cross-validation →
Balance between flexibility and overfitting:

- **Regularization** (e.g. sum of |coefs| < l)
- Ensembles:
  - Train trees with different data
    - Bootstrap
    - Subset of predictors
  - Use shallow trees
- Neural networks:
  - Train disabling neurons (dropout)
- Early stopping

Issue 2: Interpretability



2: Interpretability

Being Right for the Right Reasons

Choices:

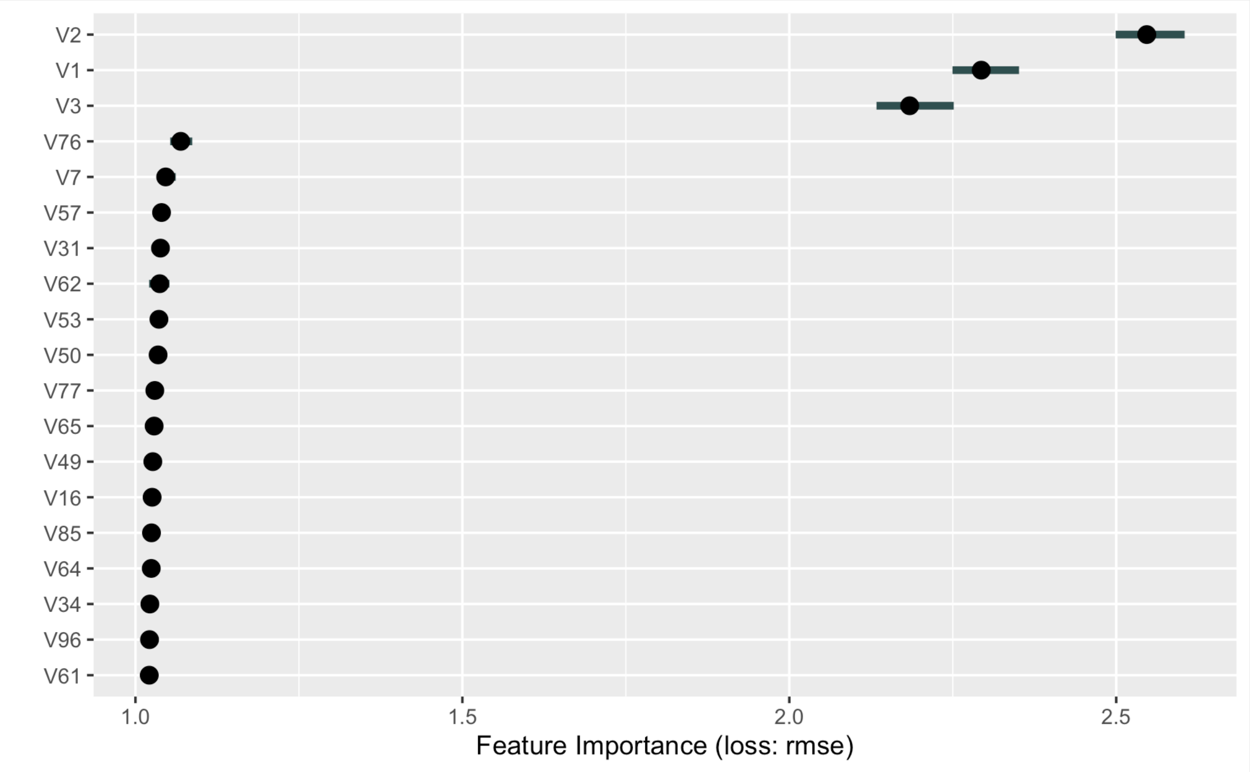
- Global vs Local interpretability
- Model-dependent vs model-agnostic interpretability

2: Global interpretability

Goal: Understand what are the main features/interactions in the model

Example 1: Feature importance: Increase in model error when the information is removed

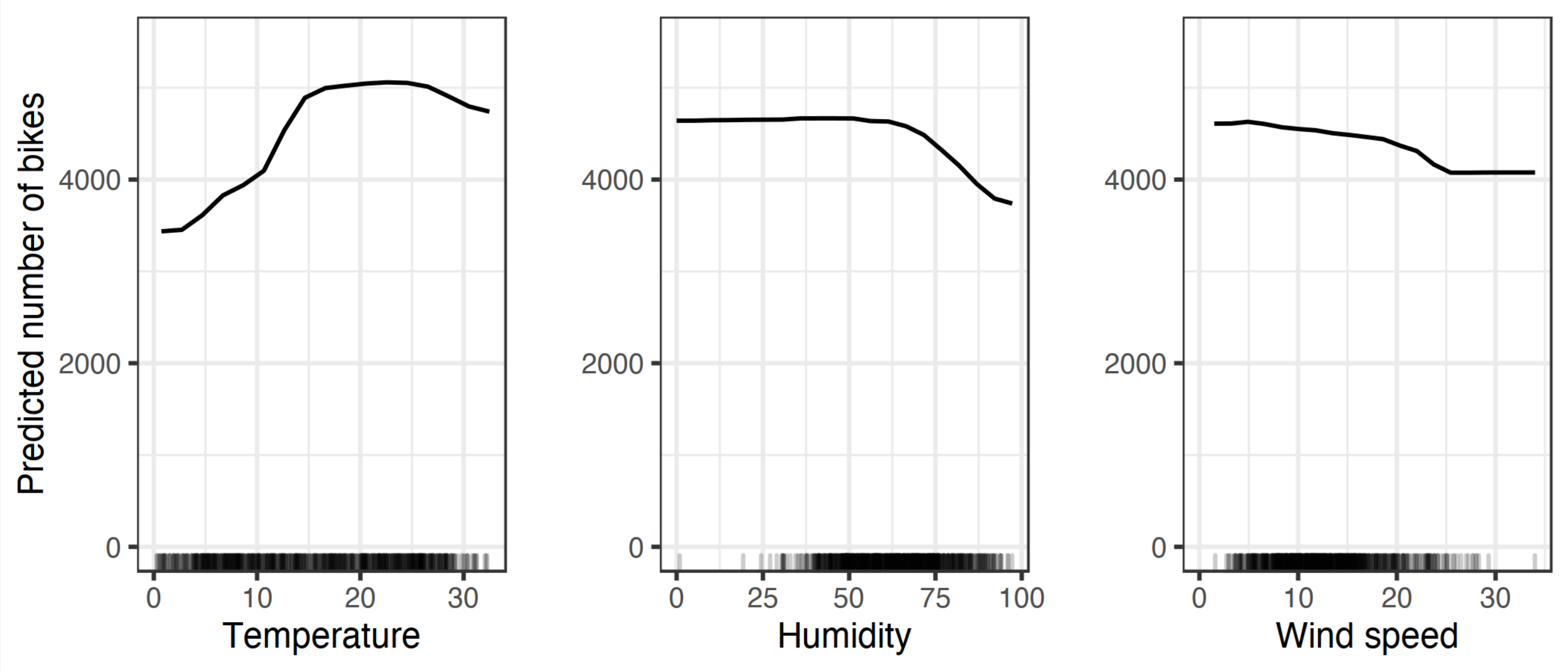
- No need to retrain the model
- Correlated variables are problematic



2: Global interpretability

Goal: Understand what are the main features/interactions in the model

Example 2: Partial dependencies plots (average marginal effects)

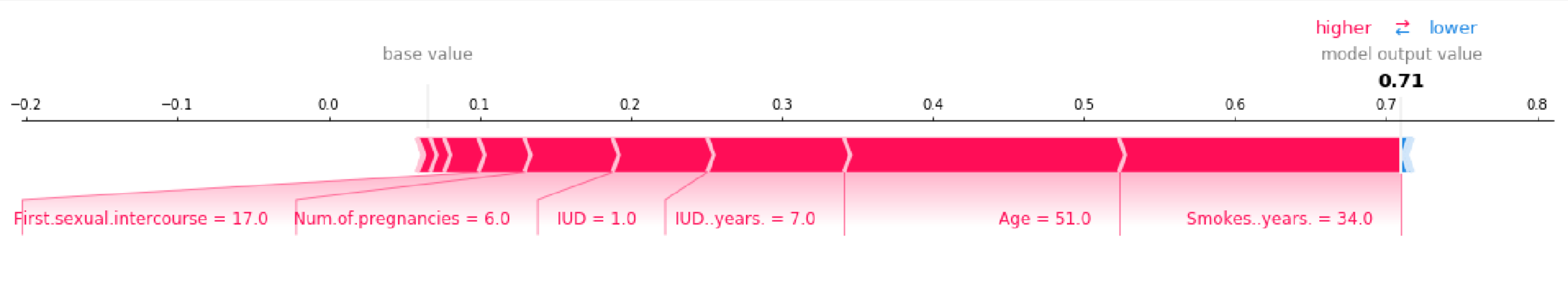


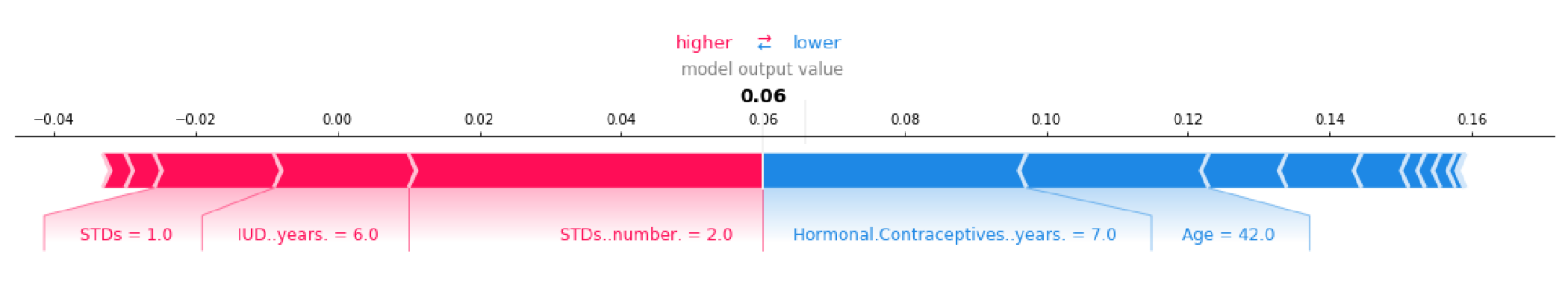
2: Local interpretability

Goal: Inform the human about the factors determining the prediction

Many advances in the last decade (SHAP, LIME, Anchors, Counterfactuals)

Basic idea: Change the observations slightly to observe how the prediction will change





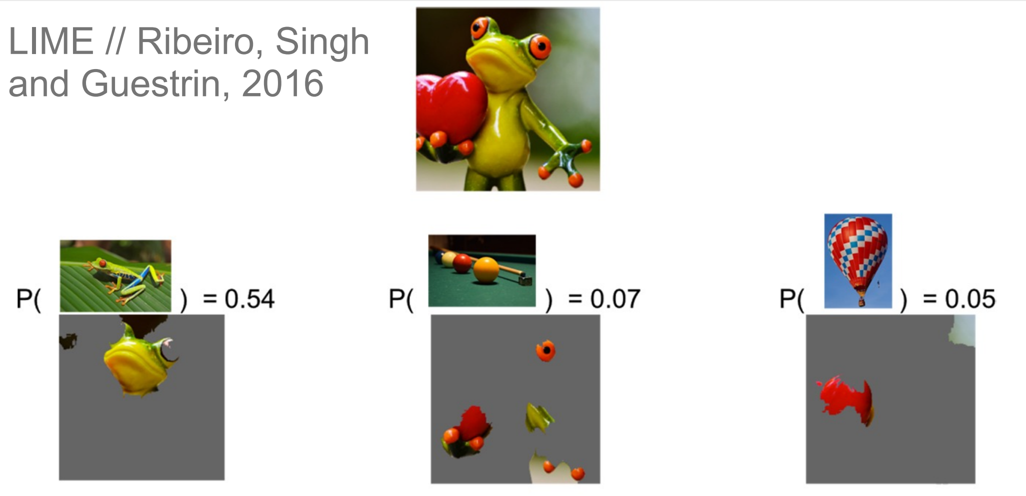
Two examples of women diagnosed with cervical cancer (Interpretable Machine Learning, C. Molnar 2022)

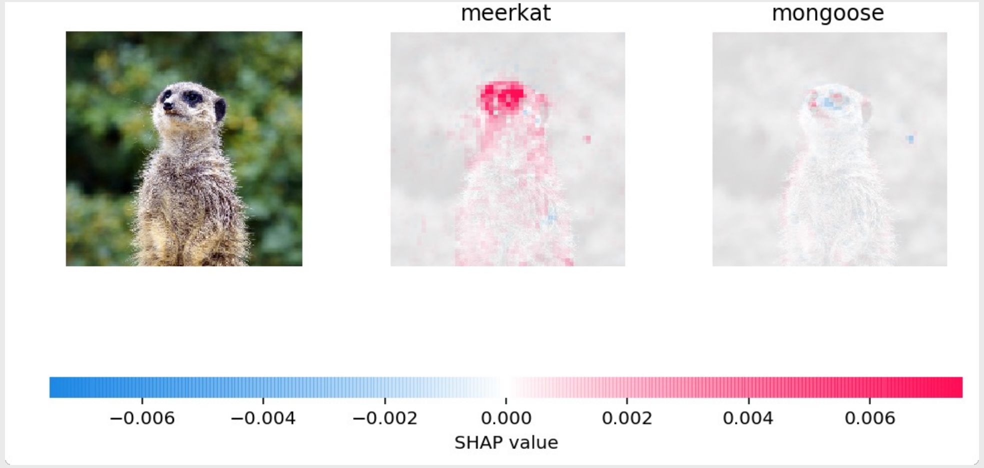
2: Local interpretability

Goal: Inform the human the reason of the prediction

Basic idea: Change the observations slightly to observe how the prediction will change

LIME // Ribeiro, Singh and Guestrin, 2016

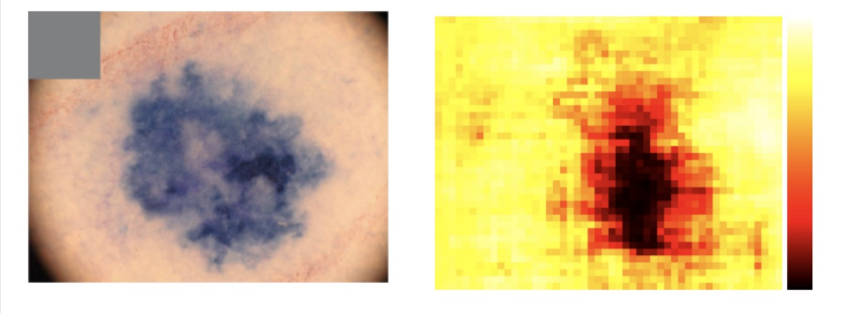




Scott M. Lundberg, Su-In Lee, 2017

(a) Original image (b) Anchor for "beagle"

Anchors // Ribeiro, Singh and Guestrin, 2018

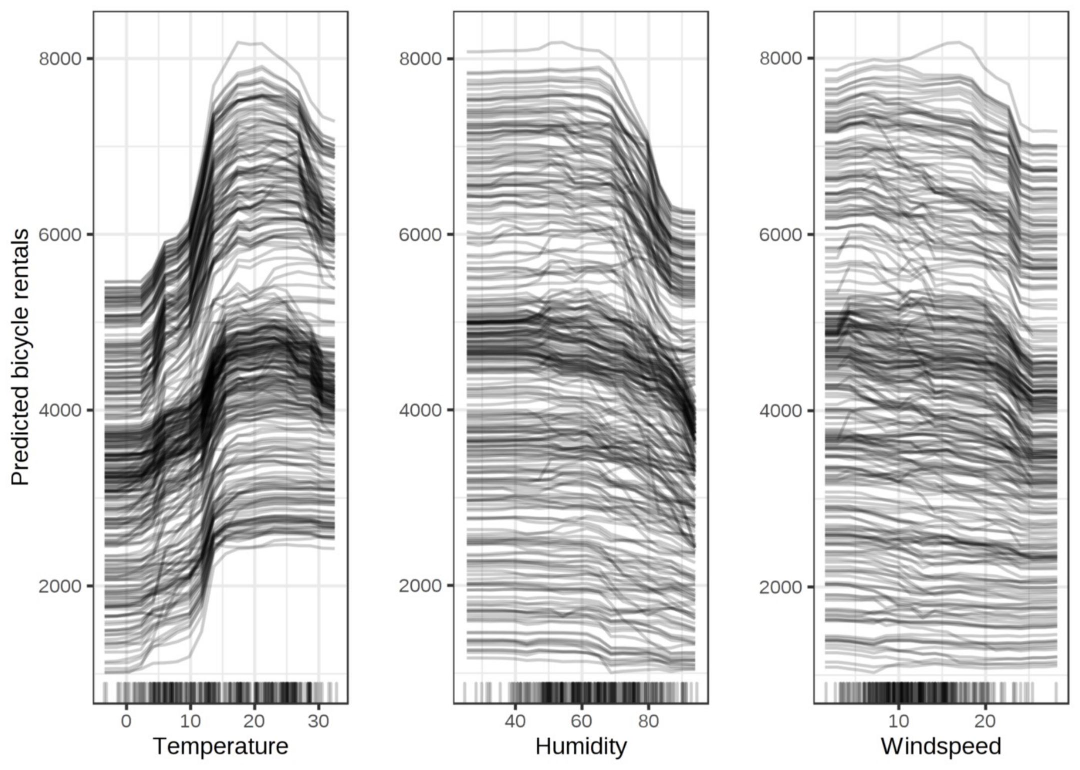


https://silverpond.com.au/2018/04/17/an-ai-tells-us-what-it-knows-when-we-poke-it-in-the-eye/

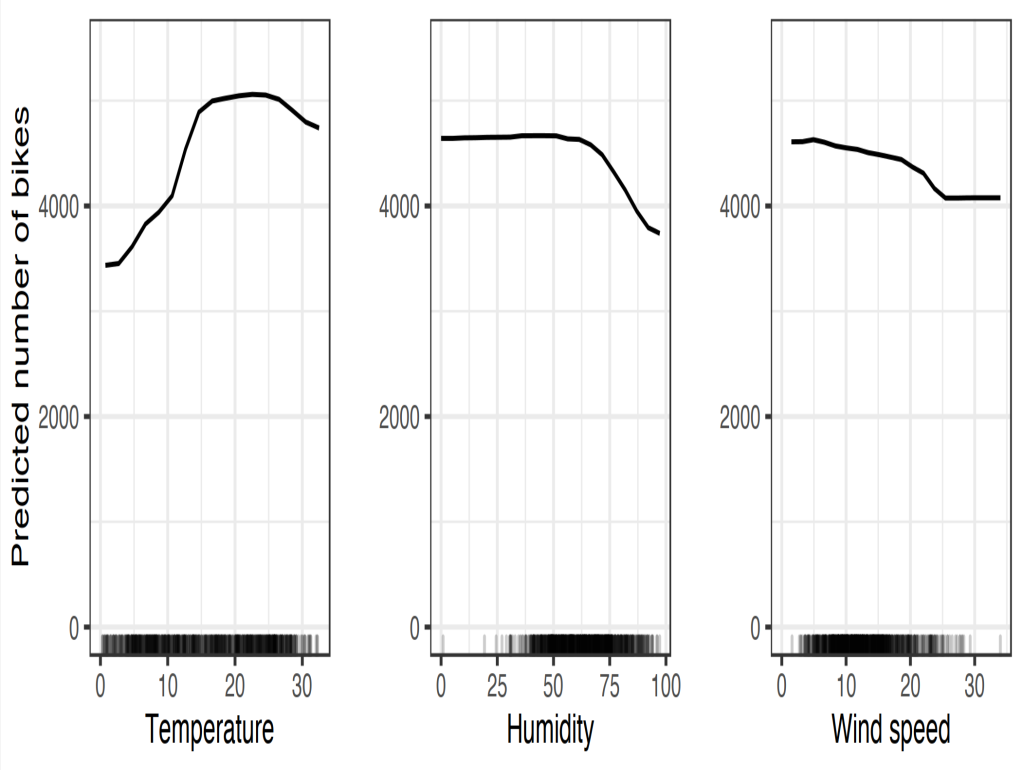
2: Local → Global interpretability

Goal: Understand what are the main features/interactions in the model

Basic idea: Average individual predictions

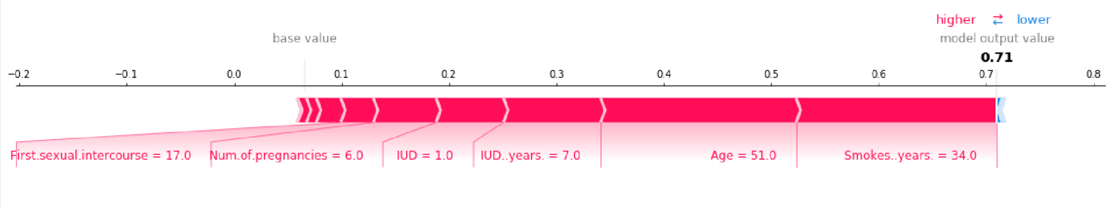


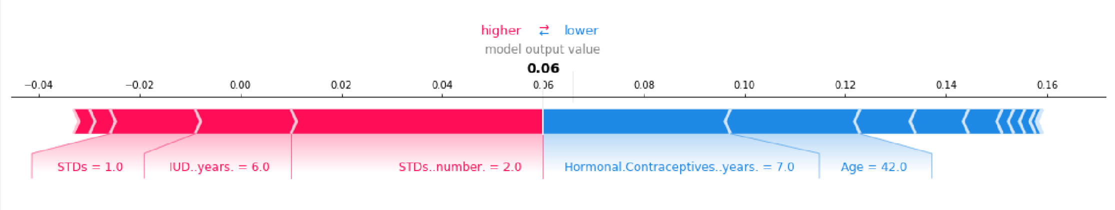
Individual Conditional Expectation (ICE)



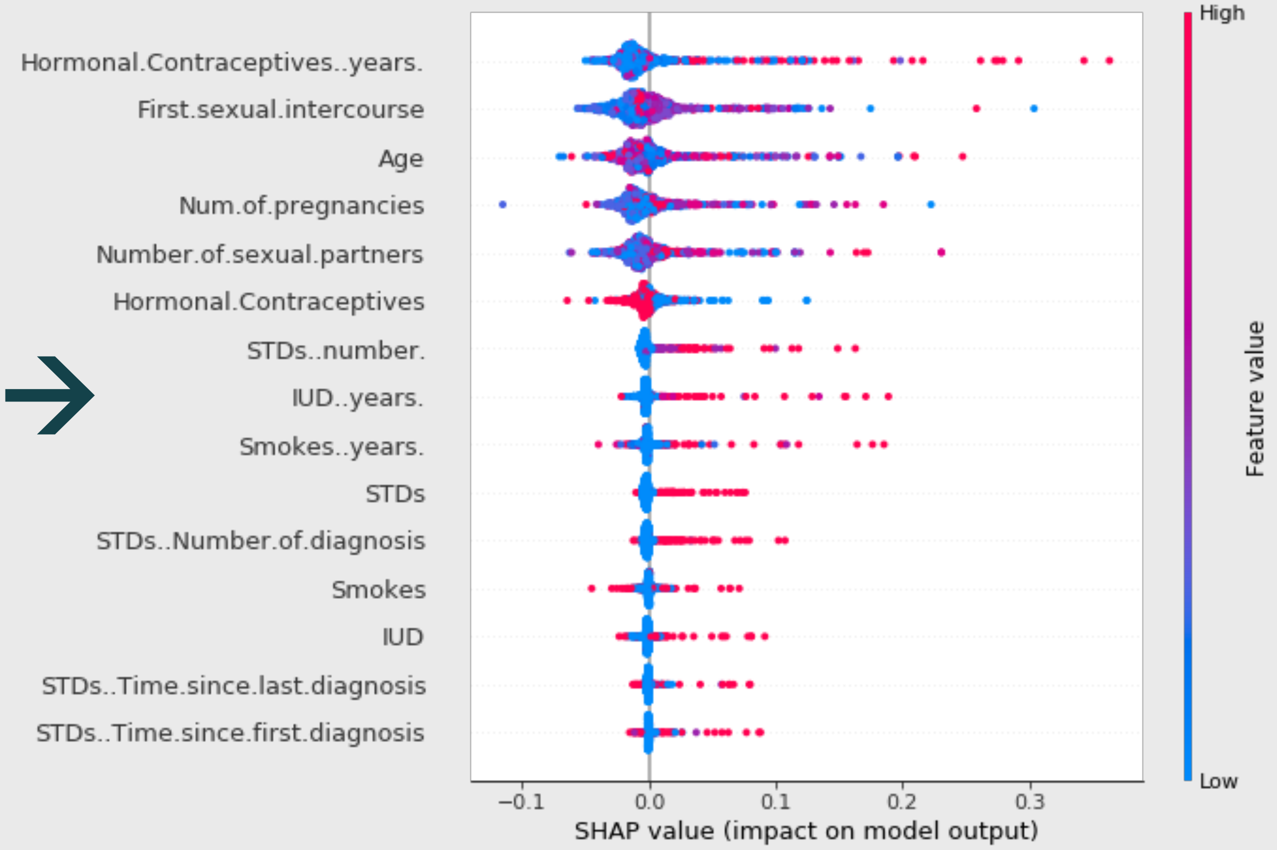
Partial Dependency Plot (PDP)

2: Local → Global interpretability





→



Exercise (javier.science/ml_julius)

Synthetic data:

- X = 105 features (100 quantitative, 5 categorical)
- Y = Quantitative response (depending on interaction between 3 quantitative and 1 categorical)

Goal:

- Try two methods: LASSO regression and XGBoost
- Hyperparameter tuning using cross-validation
- Use feature importance to understand the results of XGBoost

Instructions: (in groups)

- Read the code line by line
- Run de code
- Adapt the code with other possible outputs in the data (e.g. increase number of features)

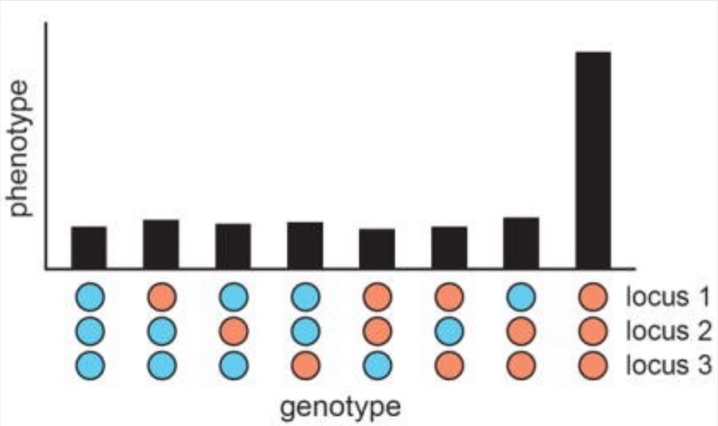
Some uses of ML in epidemiology

Missing data imputation

Prediction of outcomes with interpretability at the individual level (e.g., LIME, SHAP)

Theory building:

- ML can detect higher-order interactions and other complicated responses
- Run the ML model on the full data → Does it increase prediction compared to the traditional model?
  - You may be missing an important variable
  - Your model may be missing interactions
- Evaluate model using interpretability tools



Final remarks

Use the following questions when using ML in epidemiology:

- **How much better is the fancy model?** → Use cross-validation to evaluate. Don't use complicated methods if they are not better than simple methods.
- **What can we learn about the world from the model?** → Interpretability helps here
- **Is our method able to generalize?** → If we use our algorithm to predict new observations, make sure it keeps predicting well (i.e., do not blindly trust it)
- **Is our algorithm fair?** → http://aequitas.dssg.io/audit/_g5htt_b/compas_for_aequitas/

Thanks!