CodeGEO Workshop

DPG Coding/ML group

Feature importance

Bram Droppers

Workshop

- Three sections
  - Impurity feature importance
  - Permutation feature importance
  - SHAP feature importance

Workshop

- Three sections
  - Impurity feature importance
  - Permutation feature importance
  - SHAP feature importance
- Per section
  - Small exercise
  - Presentation and questions

Coding/ML
group

Why feature importance?

Coding/ML
group

Why feature importance?

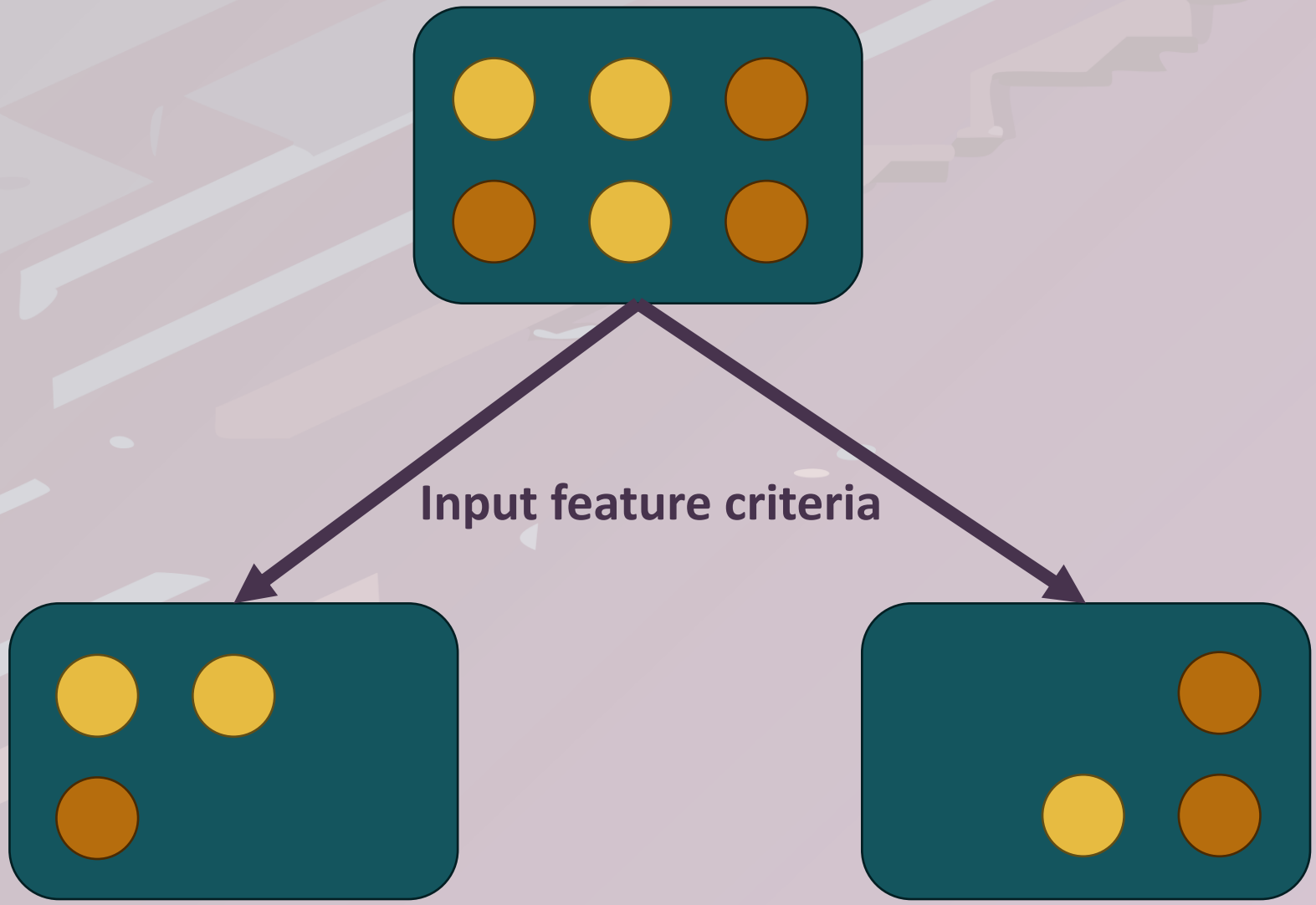
- Common sense check
- Uncertainty analysis
- Reducing model size and complexity

Impurity feature importance

Coding/ML
group

Impurity feature importance

- Also called:
  - Gini importance
  - Mean decrease impurity
- Decrease in node impurity, weighted by the probability of reaching that node



Coding/ML
group

Impurity feature importance

- Also called:
  - Gini importance
  - Mean decrease impurity
- Decrease in node impurity, weighted by the probability of reaching that node

\+ Already calculated

\- Only for random-forest models

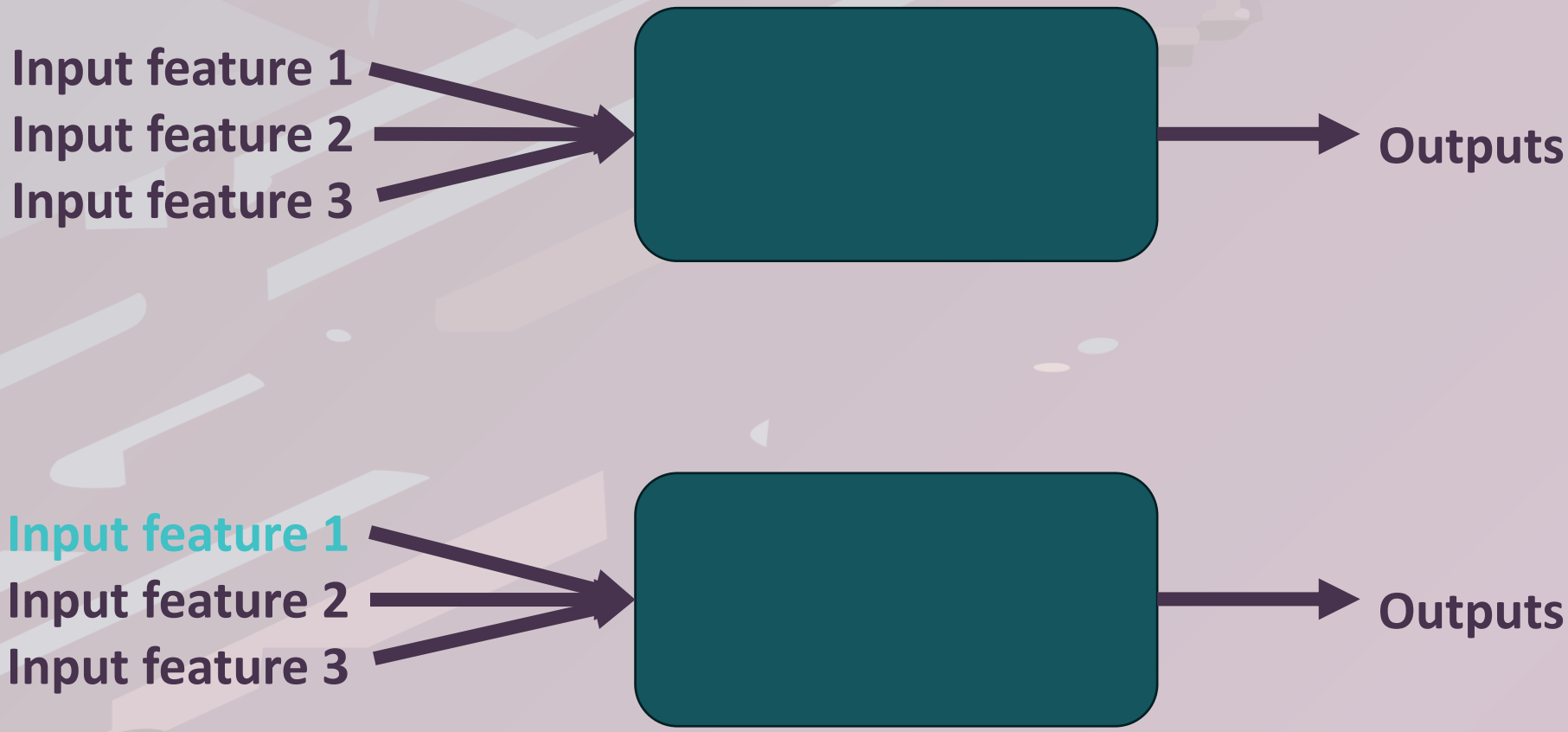
Permutation feature importance

Coding/ML
group

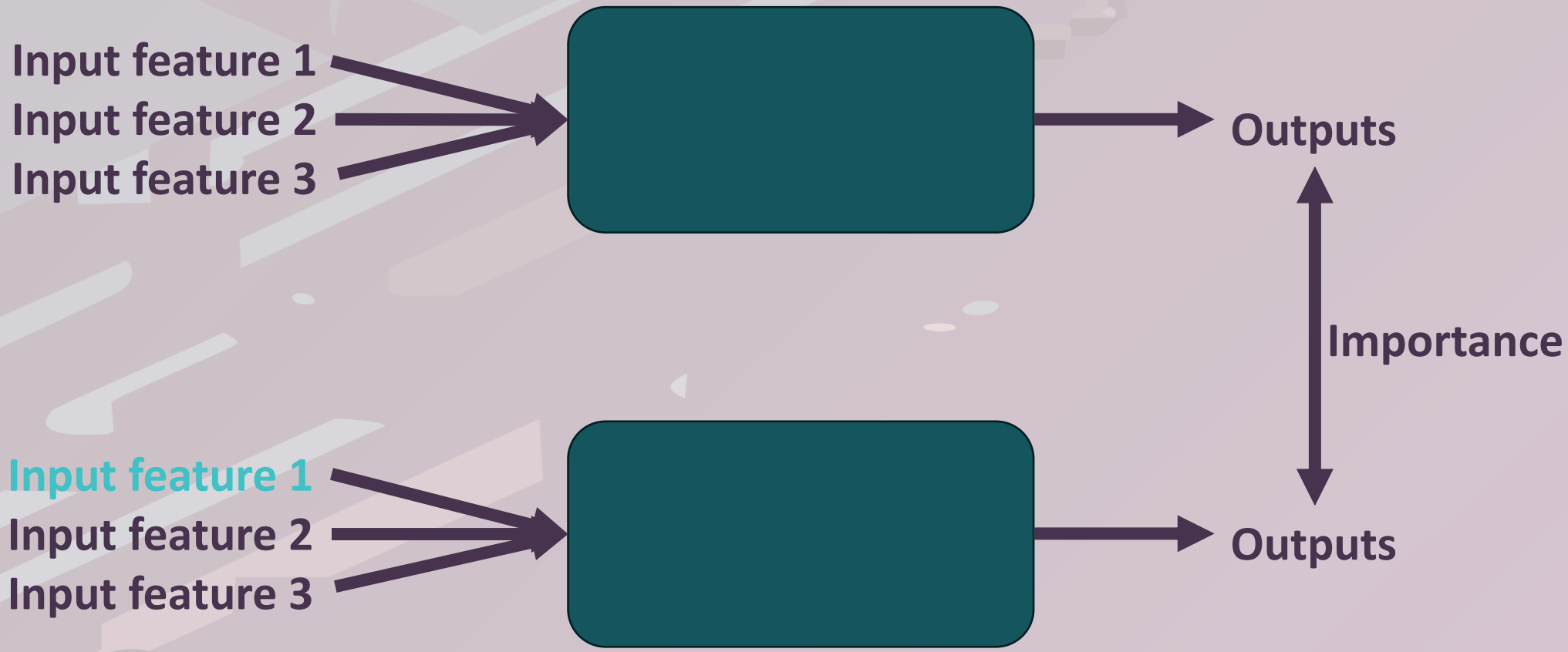
Permutation feature importance

- Difference in model output after permutation of input features

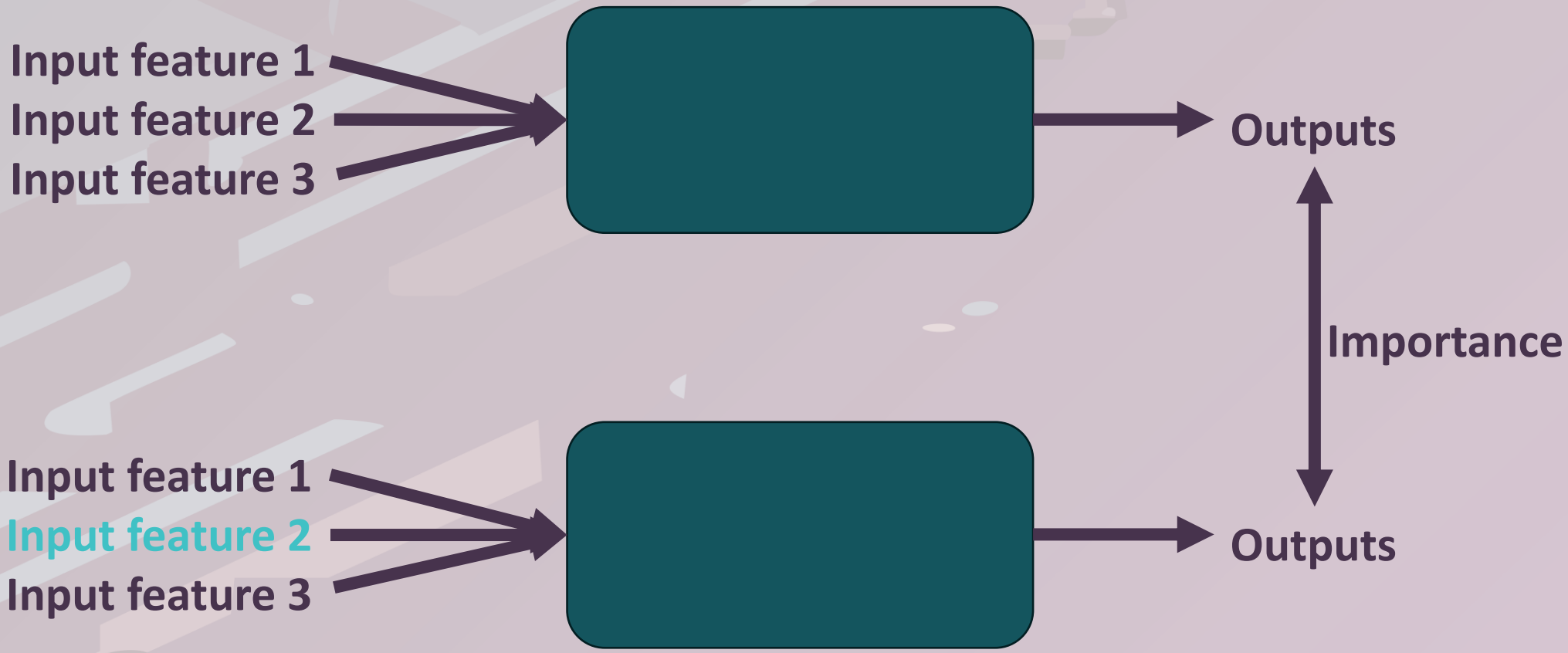
Coding/ML group



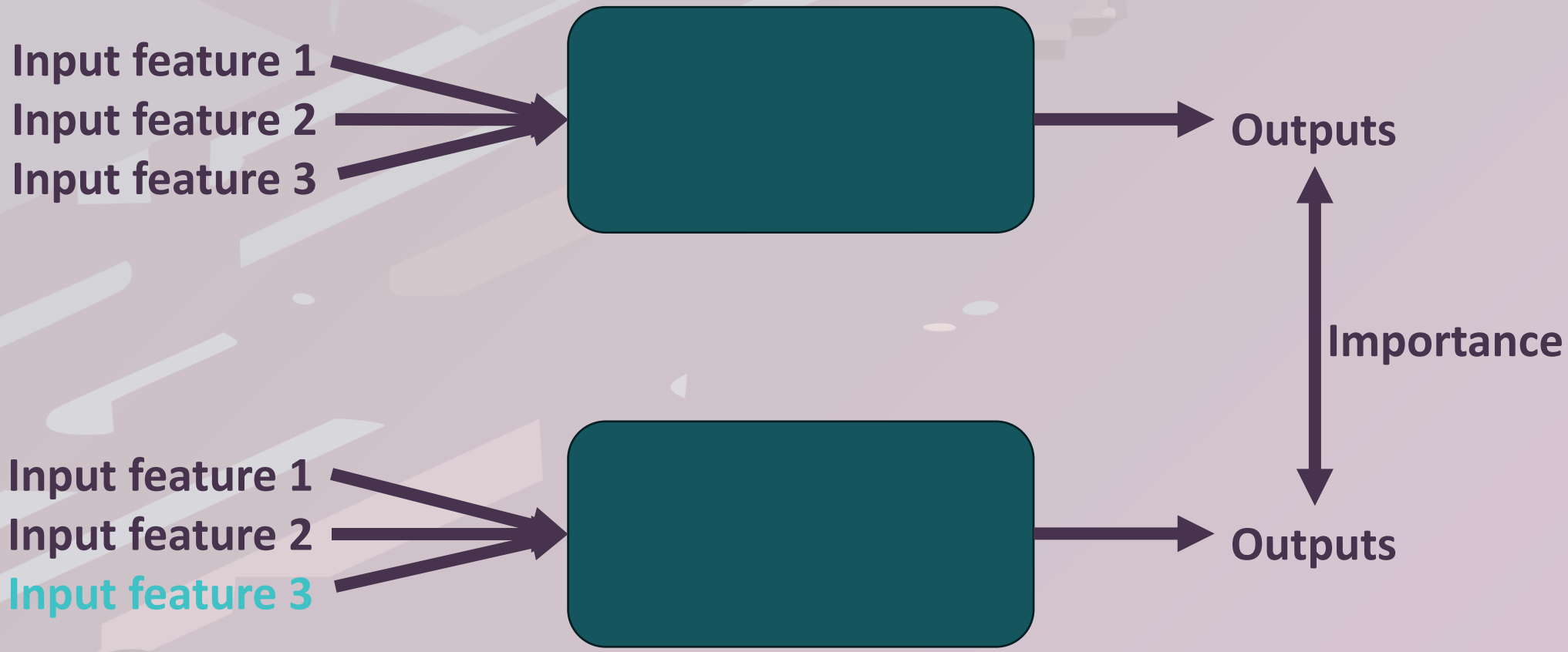
Coding/ML group



Coding/ML group



Coding/ML group



Coding/ML group

Permutation feature importance

- Difference in model output after permutation of input features

+ Applicable to all models

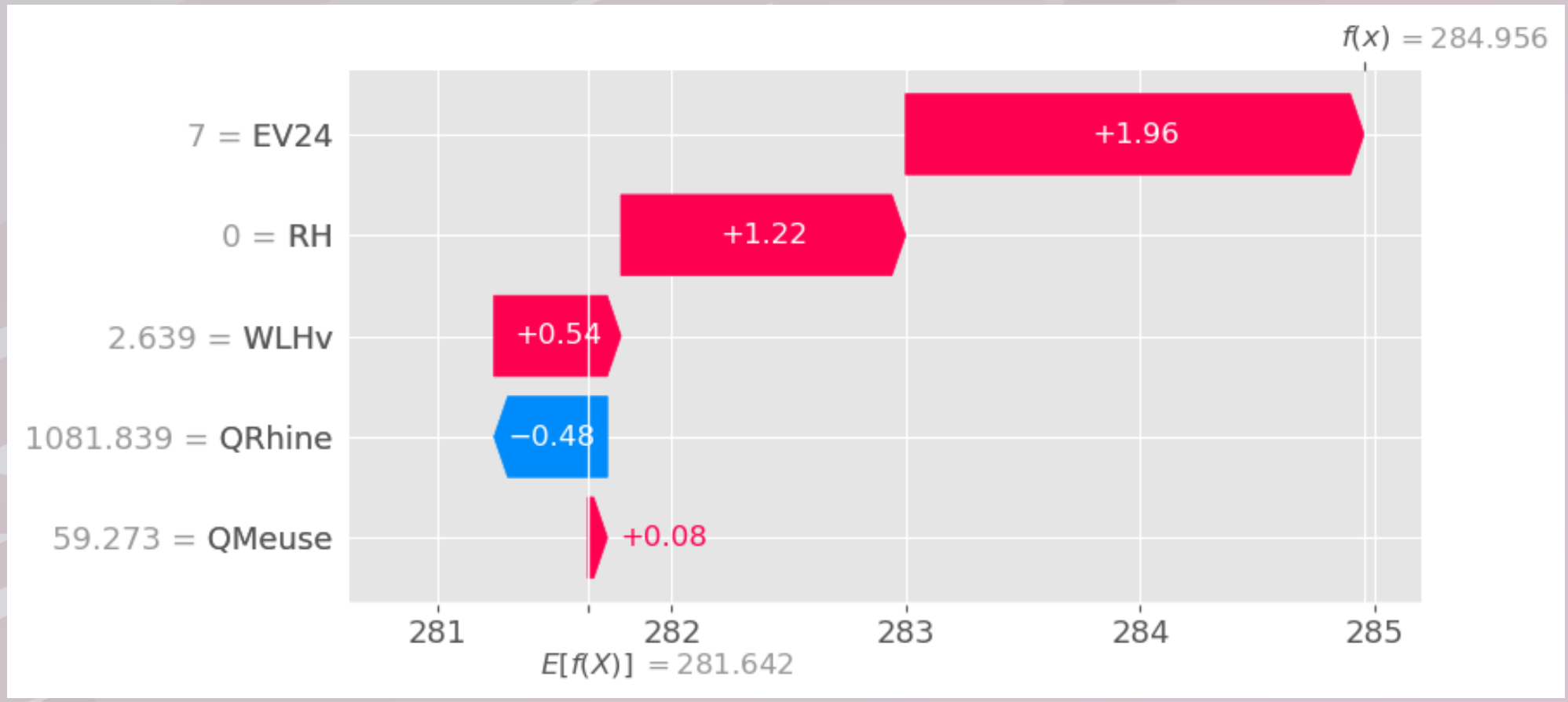
- Slow
- Limited accounting for complex non-linear interactions

SHAP feature importance

Coding/ML
group

SHAP feature importance

- Represents the marginal contribution of a feature's value to the prediction averaged over all possible combinations



Coding/ML group

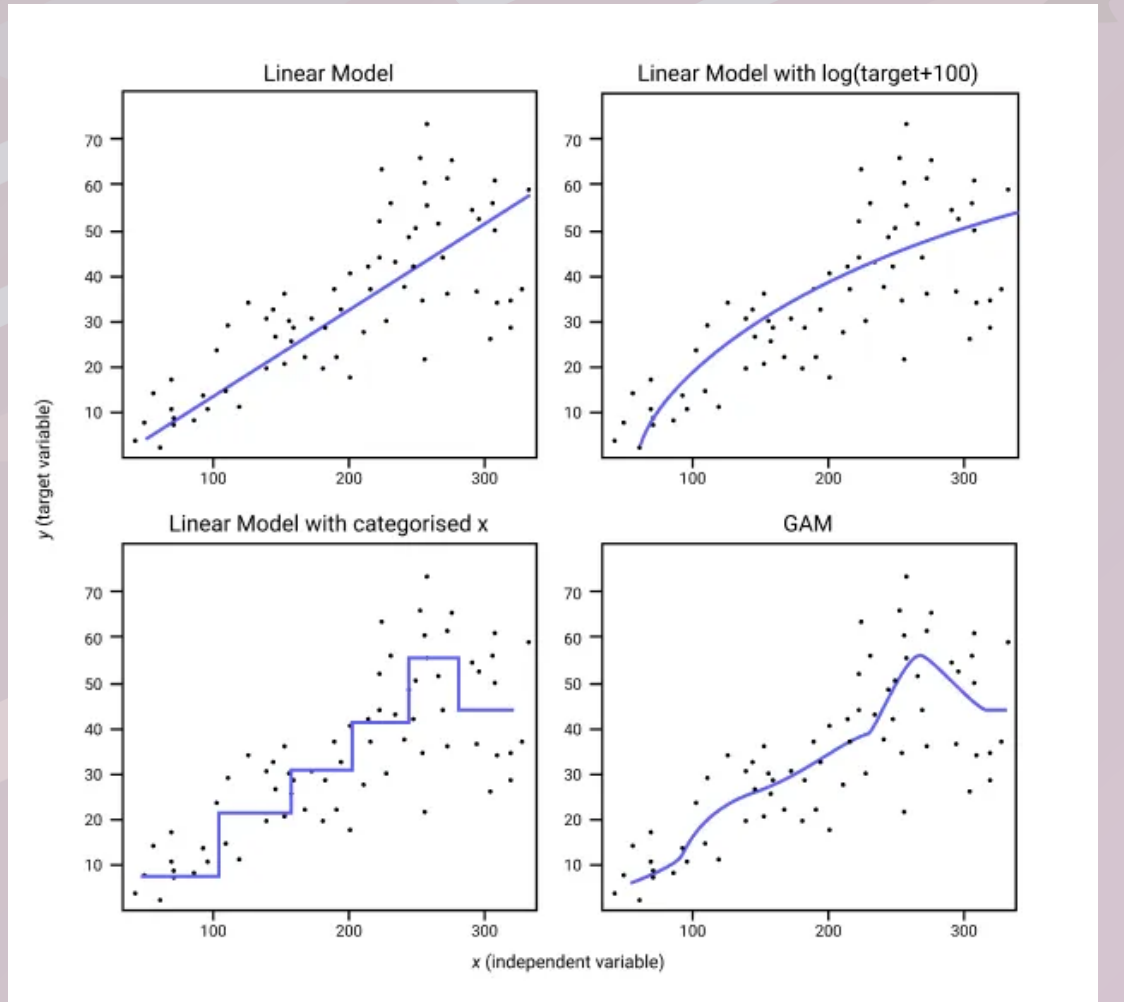
SHAP feature importance

- Represents the marginal contribution of a feature's value to the prediction averaged over all possible combinations
  - Theoretically requires permutation of all features for every timestep

SHAP feature importance

- Represents the marginal contribution of a feature's value to the prediction averaged over all possible combinations
  - Theoretically requires permutation of all features for every timestep
  - In practice only a few permutations and a generalized additive model to estimate relations

Coding/ML group



Coding/ML group

SHAP feature importance

- Represents the marginal contribution of a feature's value to the prediction averaged over all possible combinations
  - Theoretically requires permutation of all features for every timestep
  - In practice only a few permutations and a generalized additive model to estimate relations

- \+ Applicable to all models
- \- Even slower

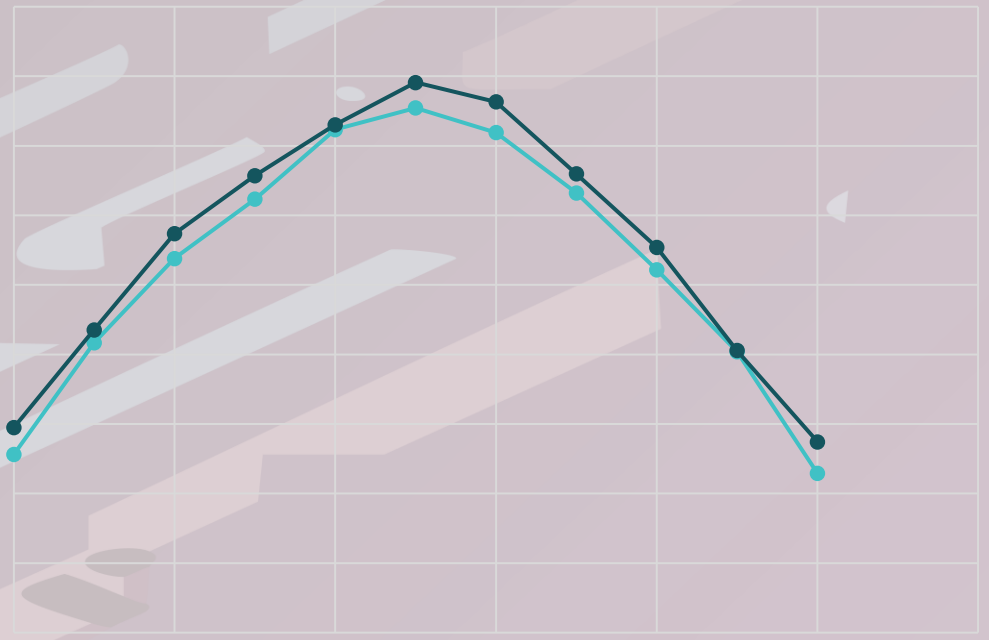
Coding/ML group

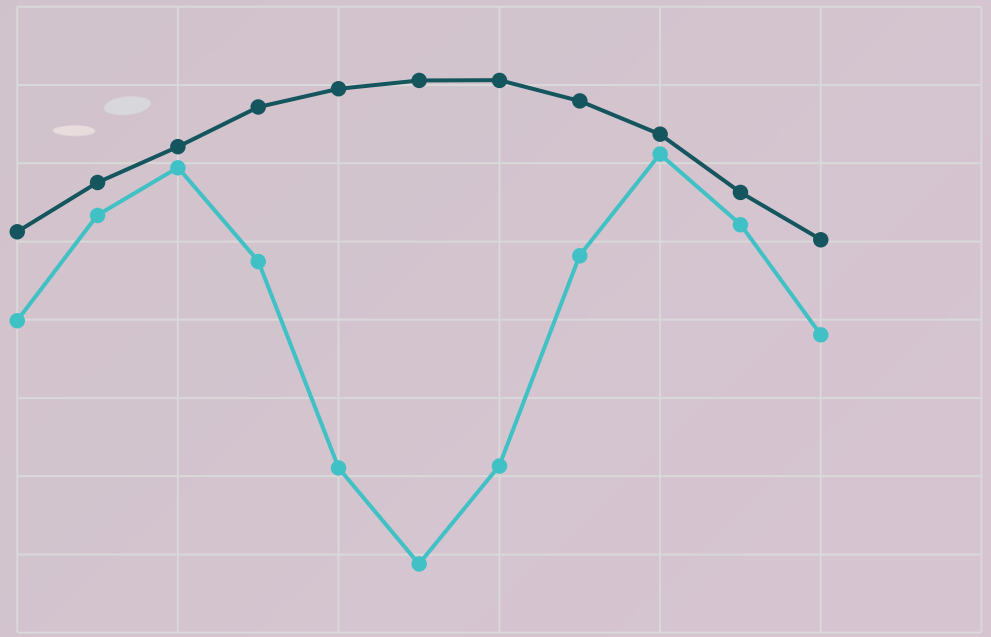
Input feature correlation

Coding/ML
group

Input feature correlation

- We do not know how our models handle correlated input features
  - Ignore one
  - Use both





Coding/ML group

Input feature correlation

- We do not know how our models handle correlated input features
  - Ignore one
  - Use both
- This is reflected in the feature importance analysis

Input feature correlation

- Combine correlated features
- Omit correlated features
- **Design a better train-test set**

CodeGEO Workshop

DPG Coding/ML group

Feature importance

Bram Droppers