Enhancing Variable Importance Interpretation in Machine Learning with Conditional Permutation Importance

Ahmad CHAMMA[1]

[1]PhD graduate from Paris-Saclay University/Inria-Saclay

PSI Talk | June 17, 2024

Example I: Population Imaging with UK Biobank

- In multimodal clinical research, prospective epidemiology $\Longrightarrow$ Heterogeneous sources of data, *UK Biobank* dataset

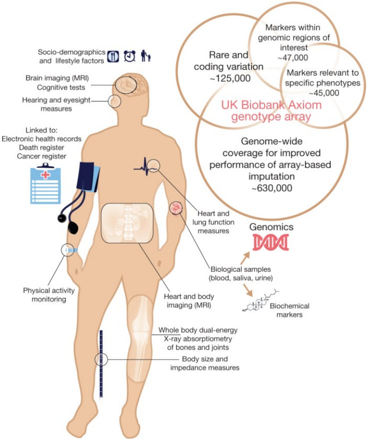


Figure: Taken from (Bycroft et al., Nature, 2018)

Example I: Population Imaging with UK Biobank

- In multimodal clinical research, prospective epidemiology $\Longrightarrow$ Heterogeneous sources of data, *UK Biobank* dataset

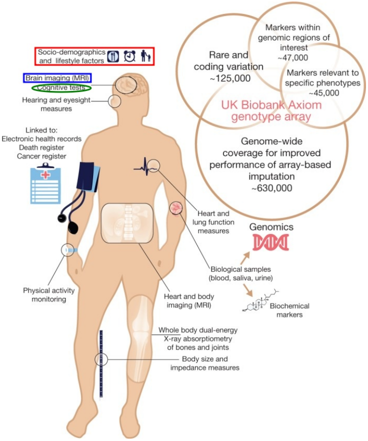
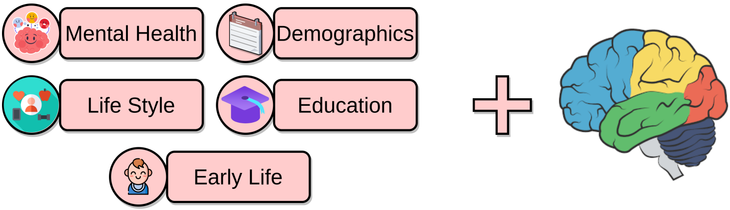


Figure: Taken from (Bycroft et al., Nature, 2018)

Example I: Population Imaging with UK Biobank

- Given socio-demographics data, brain variables provide additional information on top of it for predicting cognitive tests?

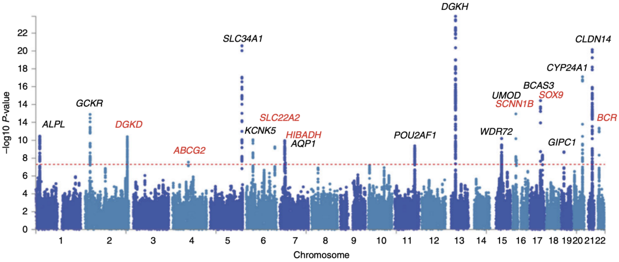


Example II: Genetic Data

- Given a genomic dataset and kidney stone disease diagnosis:

Example II: Genetic Data

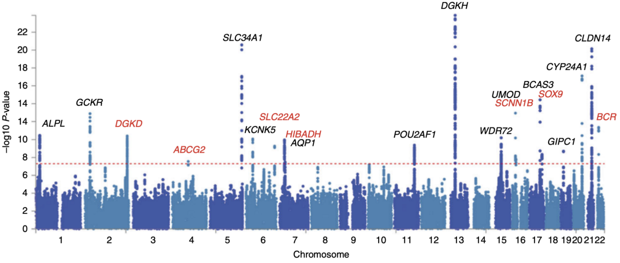
- Given a genomic dataset and kidney stone disease diagnosis:
  - What genes add information w.r.t the outcome *given* **all others**?



Taken from (Howles et al., Nature, 2019)

Example II: Genetic Data

- Given a genomic dataset and kidney stone disease diagnosis:
  - What genes add information w.r.t the outcome *given* **all others**?

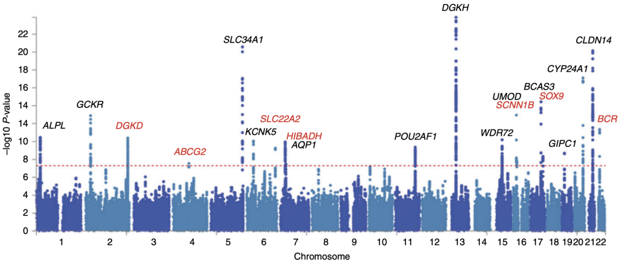


Taken from (Howles et al., Nature, 2019)

Warning!

- Genes are highly correlated locally (High-correlation)
- Assessing the added information of specific genes is crucial!

Example II: Genetic Data



Taken from (Howles et al., Nature, 2019)

Warning!

- Genes are highly correlated locally (High-correlation)
- Assessing the added information of specific genes is crucial!

Definition of Variable Importance

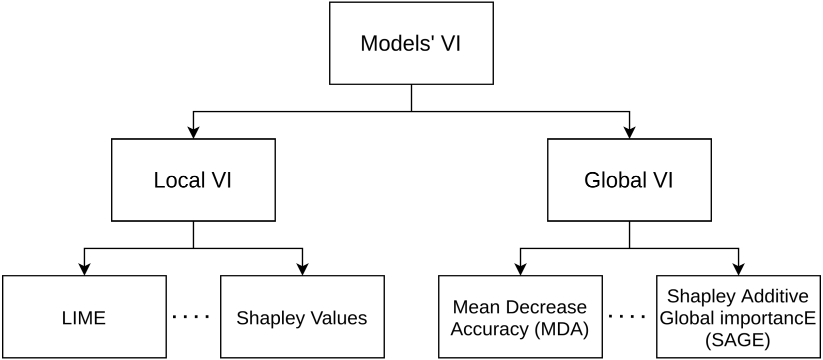
Variable Importance [Hooker et al. 2018; Zien et al. 2009]

Estimating the influence of a given input feature to the prediction made by a model

Types of Interpretations: Local vs Global

Variable Importance [Hooker et al. 2018; Zien et al. 2009]

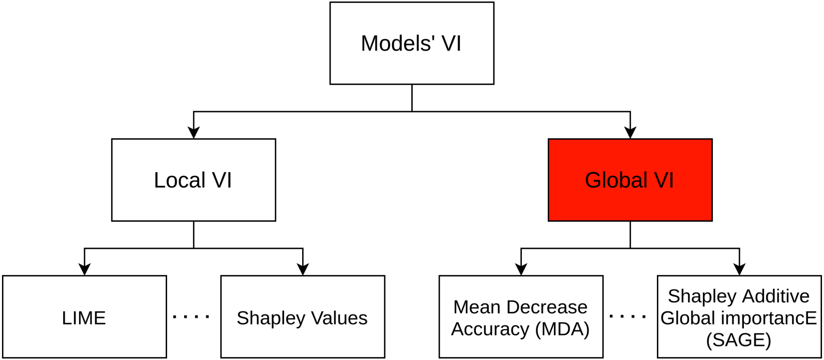
Estimating the influence of a given input feature to the prediction made by a model



Types of Interpretations: Local vs Global

Variable Importance [Hooker et al. 2018; Zien et al. 2009]

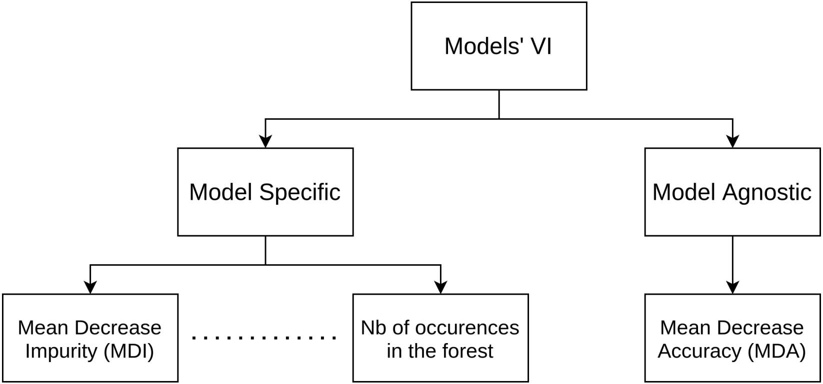
Estimating the influence of a given input feature to the prediction made by a model



Types of Interpretations: Model- Specific vs Agnostic

Variable Importance [Hooker et al. 2018; Zien et al. 2009]

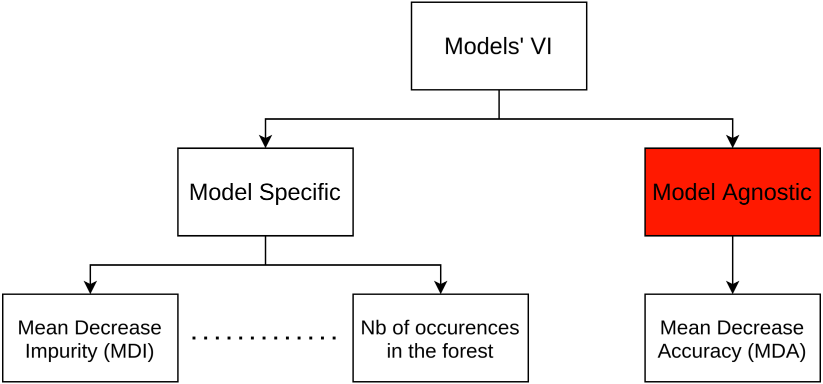
Estimating the influence of a given input feature to the prediction made by a model



Types of Interpretations: Model- Specific vs Agnostic

Variable Importance [Hooker et al. 2018; Zien et al. 2009]

Estimating the influence of a given input feature to the prediction made by a model



Statistical Guarantees for Variable Importance

Variable Importance [Hooker et al. 2018; Zien et al. 2009]

Estimating the influence of a given input feature to the prediction made by a model

Statistical guarantees are important!

- Find as many relevant variables while controlling false positives (Candes et al., Journal of the Royal Statistical Society, 2017)
- I) Essential for scientific discovery
- II) Control the risk in study design
- E.g. in genetic analysis, the cost of examining a falsely selected gene may be intolerable (Zhao et al., AAAI, 2022)

Scientific Question/Challenge

Current *state-of-the-art*

- Current inference tools in deep learning merely perform sensitivity analysis
  - Interpretation is unclear (Adebayo et al. 2018)
  - No statistical control
- Assess the impact of the removal of a given variable $\Longrightarrow$ Noisy & Too expensive

Roadmap!

- Find effectively the relevant predictors for the prediction of a biomedical outcome
  - Statistical error control $\Longrightarrow$ Controlling the rate of true non-relevant variables detected as relevant
  - High-correlation settings
  - High-dimensional settings

Scientific Question/Challenge

Current *state-of-the-art*

- Current inference tools in deep learning merely perform sensitivity analysis
  - Interpretation is unclear (Adebayo et al. 2018)
  - No statistical control
- Assess the impact of the removal of a given variable $\Longrightarrow$ Noisy & Too expensive

Roadmap!

- Find effectively the relevant predictors for the prediction of a biomedical outcome
  - Statistical error control $\Longrightarrow$ Controlling the rate of true non-relevant variables detected as relevant
  - High-correlation settings
  - High-dimensional settings

Scientific Question/Challenge

Current *state-of-the-art*

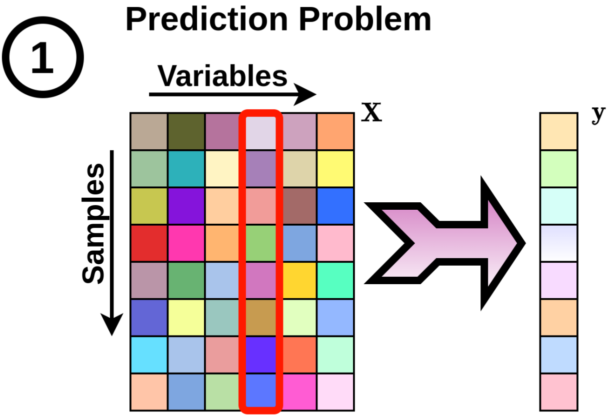
- Current inference tools in deep learning merely perform sensitivity analysis
  - Interpretation is unclear (Adebayo et al. 2018)
  - No statistical control
- Assess the impact of the removal of a given variable $\Longrightarrow$ Noisy & Too expensive

Roadmap!

- Find effectively the relevant predictors for the prediction of a biomedical outcome
  - Statistical error control $\Longrightarrow$ Controlling the rate of true non-relevant variables detected as relevant
  - High-correlation settings
  - High-dimensional settings

What is the *permutation* approach?

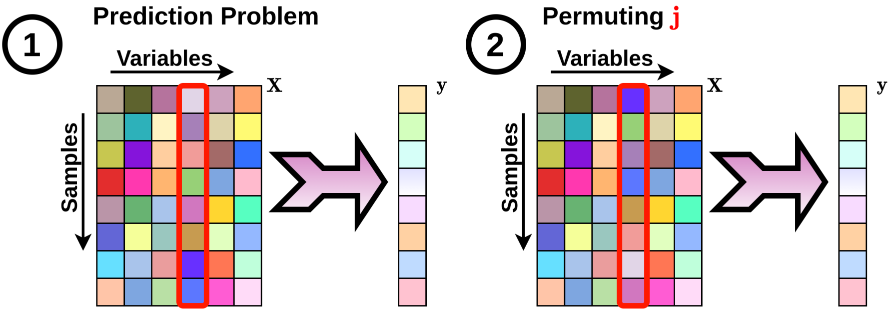
According to (Breiman, Machine Learning, 2001)



Is variable j important for the prediction?

What is the *permutation* approach?

According to (Breiman, Machine Learning, 2001)

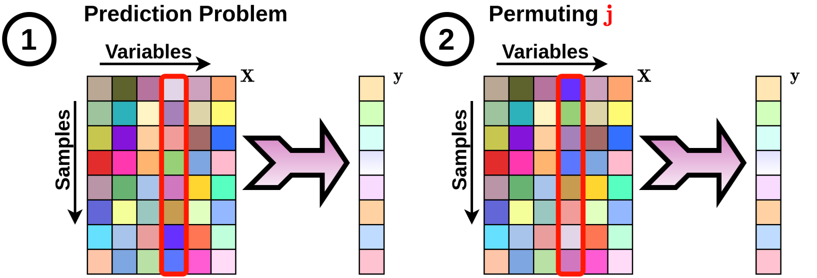


Is variable **j** important for the prediction?

How does perturbing variable **j** impact prediction?

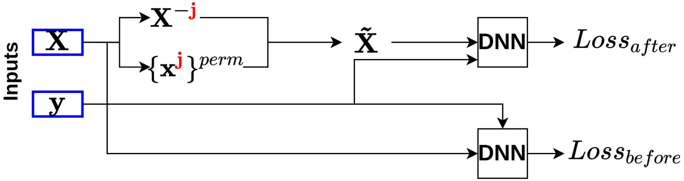
What is the *permutation* approach?

According to (Breiman, Machine Learning, 2001)



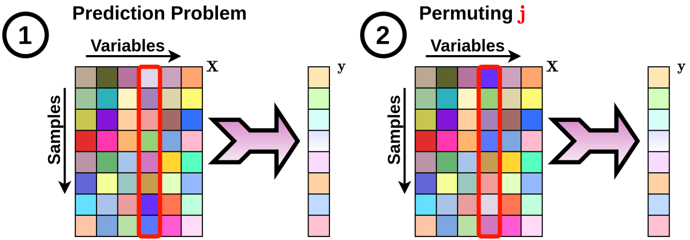
Is variable j important for the prediction?

How does perturbing variable j impact prediction?



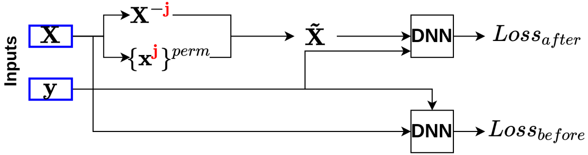
What is the *permutation* approach?

According to (Breiman, Machine Learning, 2001)



Is variable j important for the prediction?

How does perturbing variable j impact prediction?

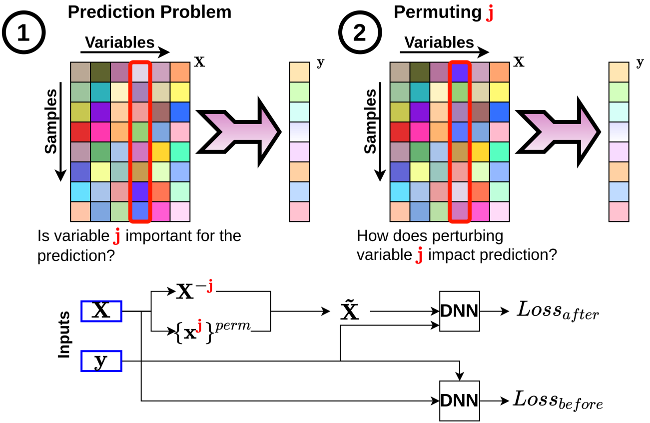


After B permutations, obtain statistical evidence on the significance of variable j

Problem: <u>Does not control errors</u> if variables are correlated

What is the *permutation* approach?

According to (Breiman, Machine Learning, 2001)



After B permutations, obtain statistical evidence on the significance of variable j

Problem: Does not control errors if variables are correlated

Proposed Solution - *Sampling* the variable of interest[1]

- Let $\epsilon^{\mathbf{j}} = \mathbf{x}^{\mathbf{j}}$ - $\hat{\mathbf{x}}^{\mathbf{j}}$ with $\hat{\mathbf{x}}^{\mathbf{j}} = \mathbb{E}(\mathbf{x}^{\mathbf{j}}|\mathbf{X}^{-\mathbf{j}})$

[1] Ahmad Chamma, Denis A. Engemann, and Bertrand Thirion (2023). "Statistically Valid Variable Importance Assessment through Conditional Permutations". In: *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS), New Orleans, USA*. DOI: 10.48550/arXiv.2309.07593.

Proposed Solution - *Sampling* the variable of interest[1]

- Let $\epsilon^{\mathbf{j}} = \mathbf{x}^{\mathbf{j}}$ - $\hat{\mathbf{x}}^{\mathbf{j}}$ with $\hat{\mathbf{x}}^{\mathbf{j}} = \mathbb{E}(\mathbf{x}^{\mathbf{j}}|\mathbf{X}^{-\mathbf{j}})$

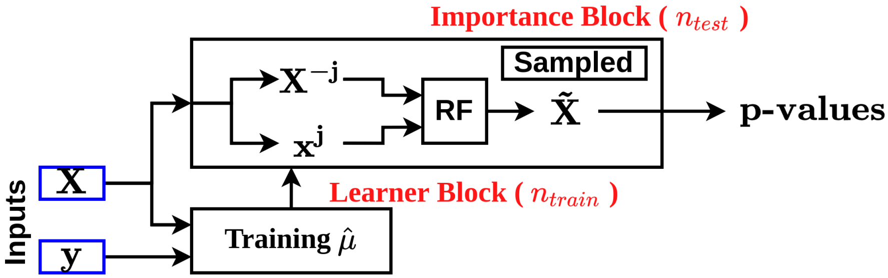
Samling $\mathbf{x}^{\mathbf{j}}$ from the conditional distribution

$$\tilde{\mathbf{x}}^{\mathbf{j}} = \hat{\mathbf{x}}^{\mathbf{j}} + \{\epsilon^{\mathbf{j}}\}^{perm}$$

Why? The dependency between the variable of interest and the remaining variables is preserved.

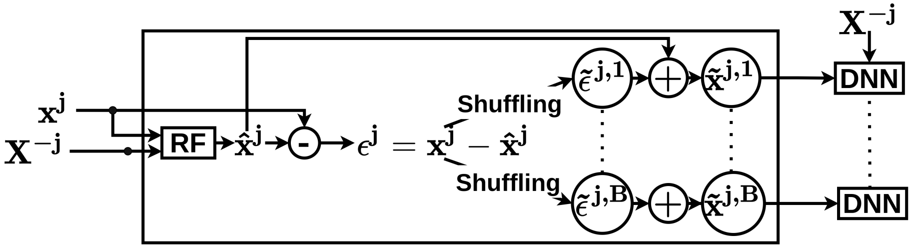
[1] Ahmad Chamma, Denis A. Engemann, and Bertrand Thirion (2023). "Statistically Valid Variable Importance Assessment through Conditional Permutations". In: *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS), New Orleans, USA*. DOI: 10.48550/arXiv.2309.07593.

Proposed Solution - Novel Implementation[1]

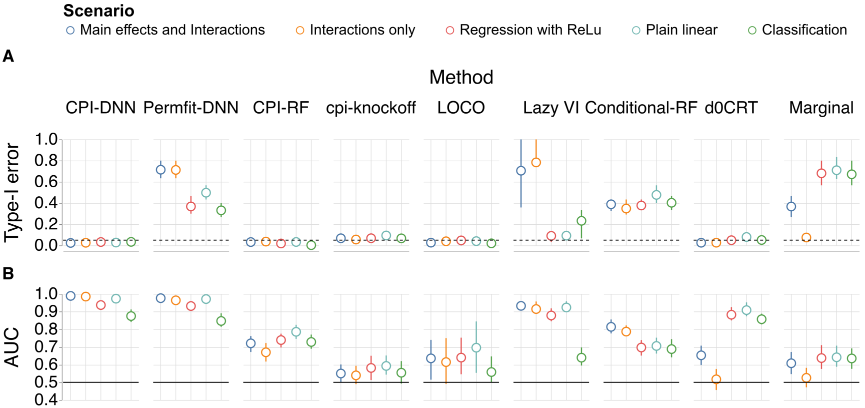


[1] Ahmad Chamma, Denis A. Engemann, and Bertrand Thirion (2023). "Statistically Valid Variable Importance Assessment through Conditional Permutations". In: *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS), New Orleans, USA*. DOI: 10.48550/arXiv.2309.07593.

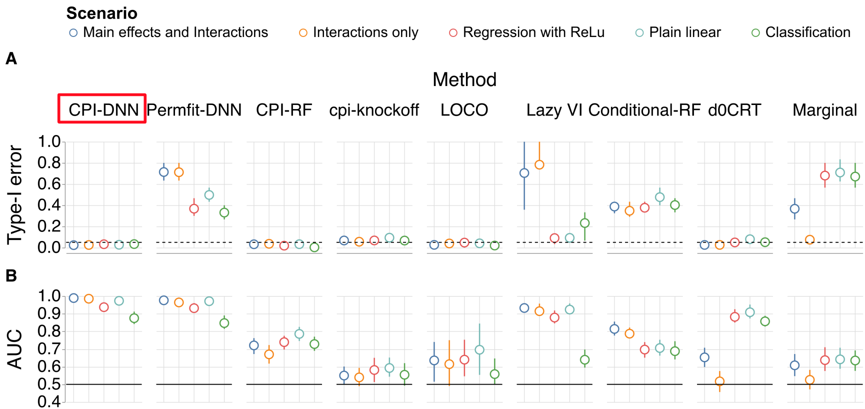
Proposed Solution - *Sampling* Phase



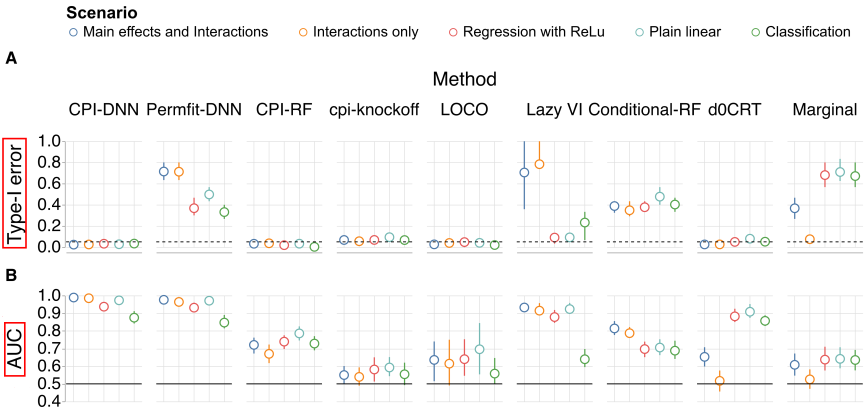
New Benchmark for State-of-the-art Methods



Results on Simulated Data - *CPI-DNN*

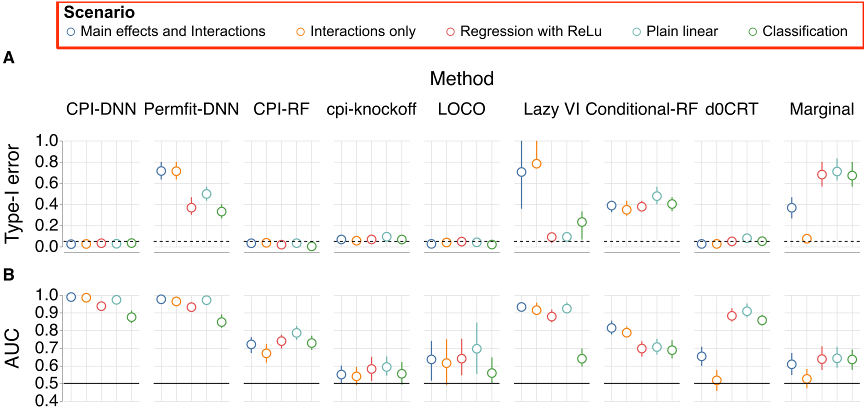


Results on Simulated Data - *Type-I error/AUC score*

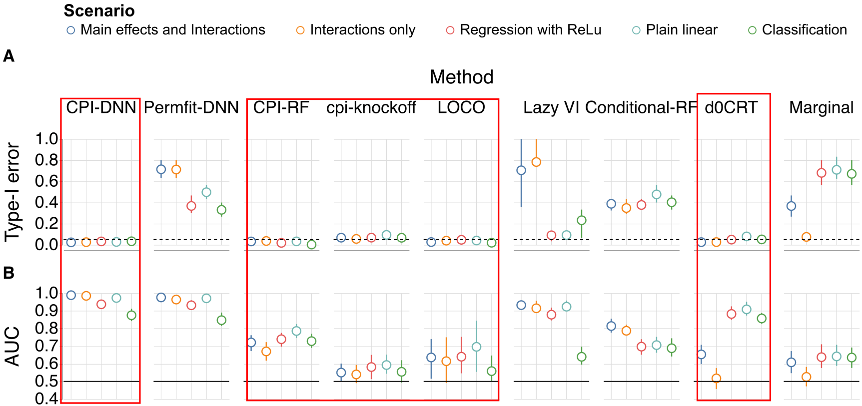


- *AUC score*: Correct significant variables ordering
- *Type-I error*: Rate of true non-significant variables detected as relevant

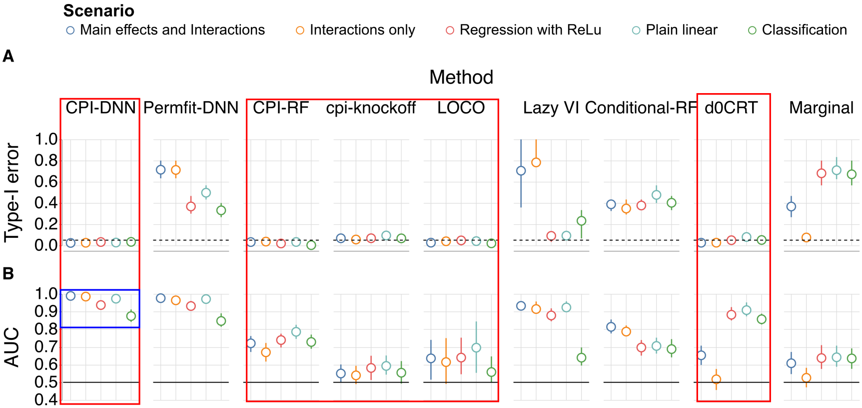
Results on Simulated Data - *Scenarios*



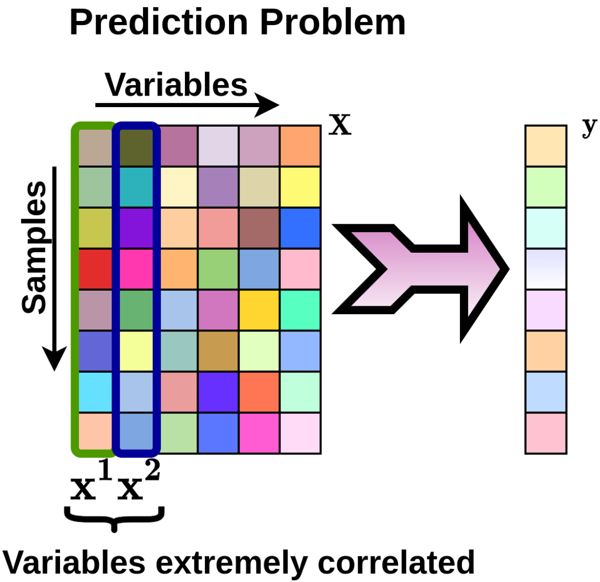
Results on Simulated Data



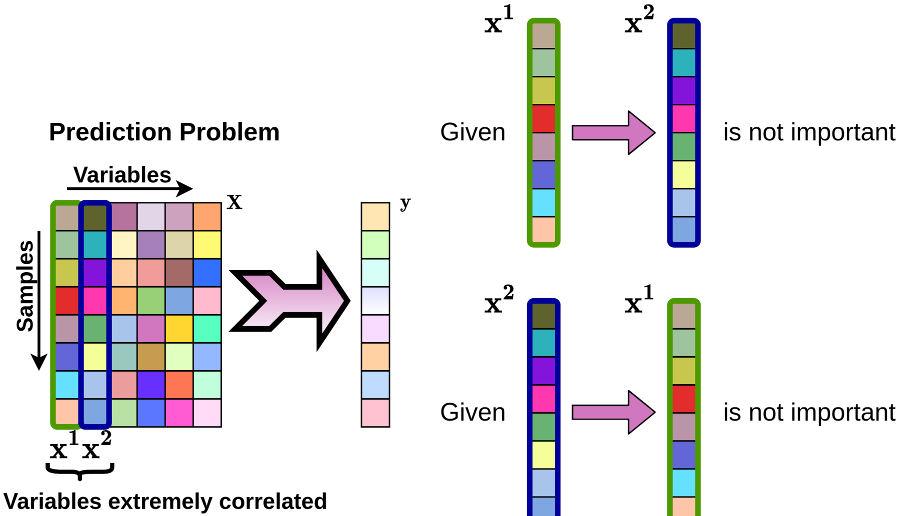
Results on Simulated Data



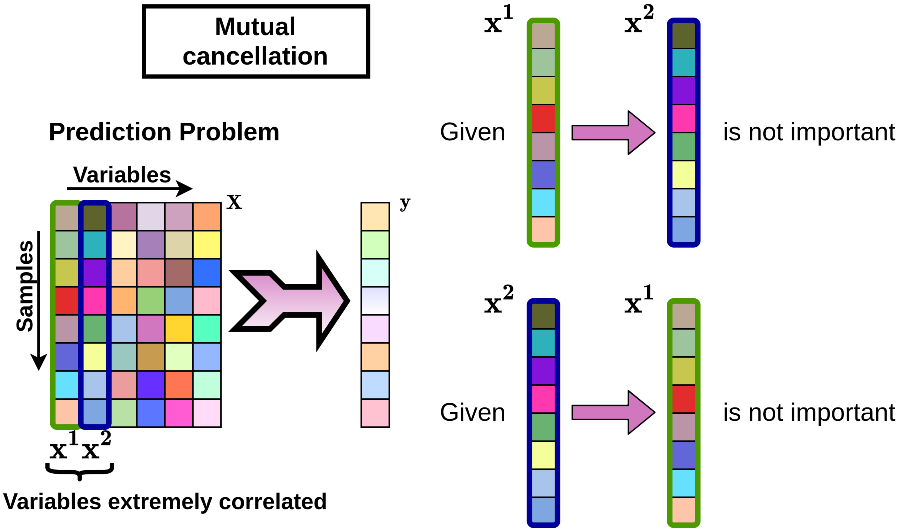
Limitations of the Conditional Inference



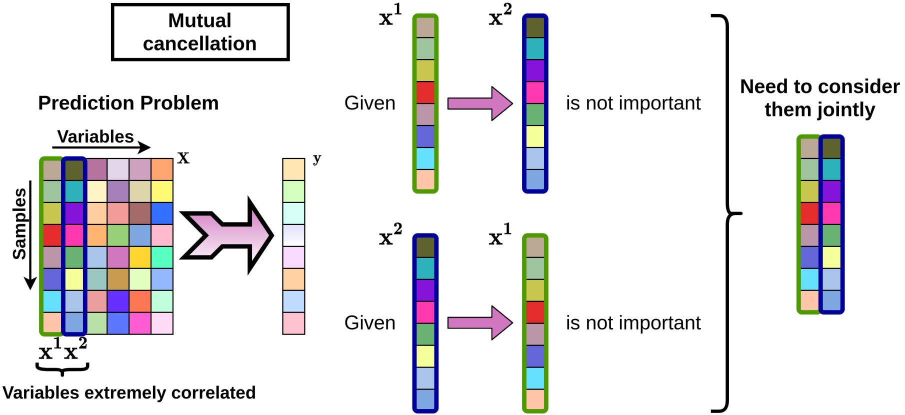
Limitations of the Conditional Inference



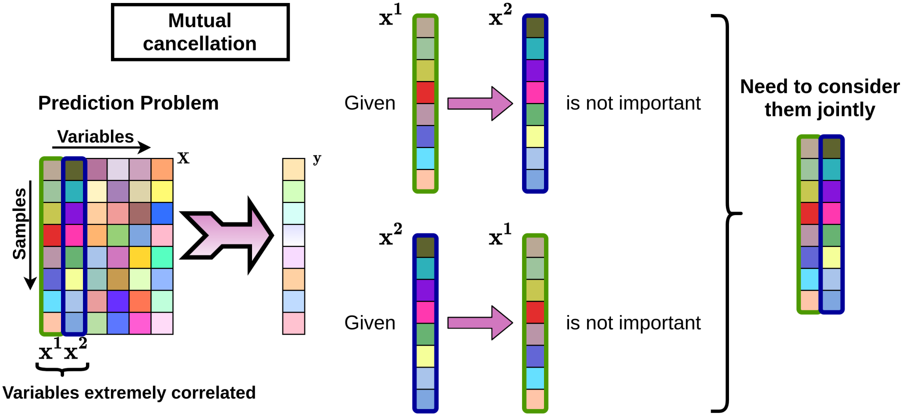
Limitations of the Conditional Inference



Limitations of the Conditional Inference

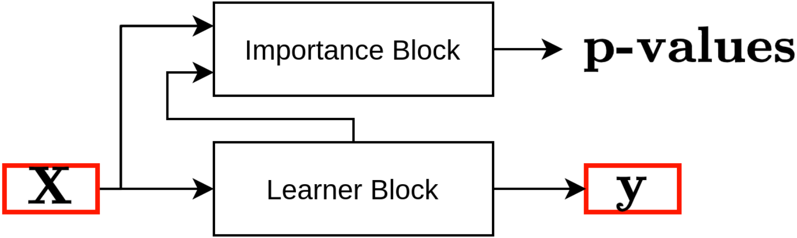


Limitations of the Conditional Inference



- Grouped interpretations might be interesting for high-dimensional settings with hundreds or thousands of features (example in Neuroimaging)

Block-based Conditional Permutation Importance (BCPI)[1]

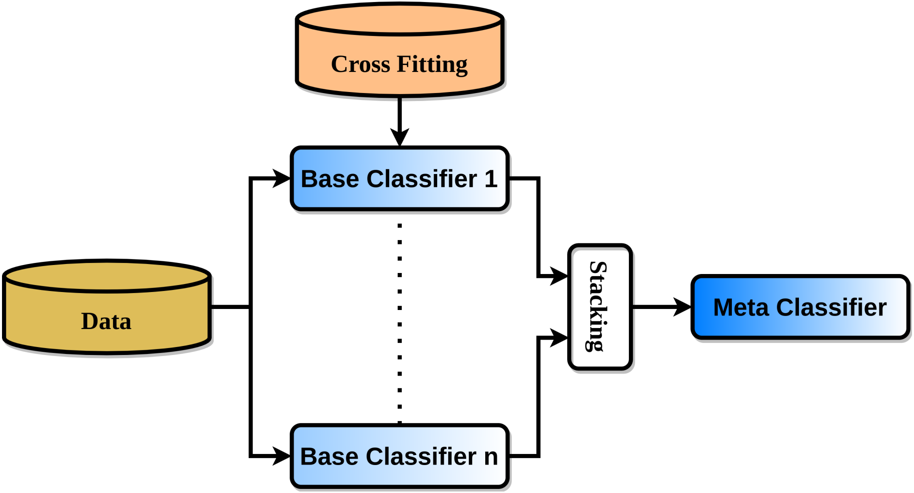


- Learner Block: Performing inference on the data
- Importance Block: Sampling the variable/group of interest
- Any Scikit-learn compatible learners can be used

[1] Ahmad Chamma, Bertrand Thirion, and Denis A. Engemann (2024). "Variable Importance in High-Dimensional Settings Requires Grouping". In: *Proceedings of the 38th Conference of the Association for the Advancement of Artificial Intelligence (AAAI), Vancouver, Canada*. DOI: 10.48550/arXiv.2312.10858.

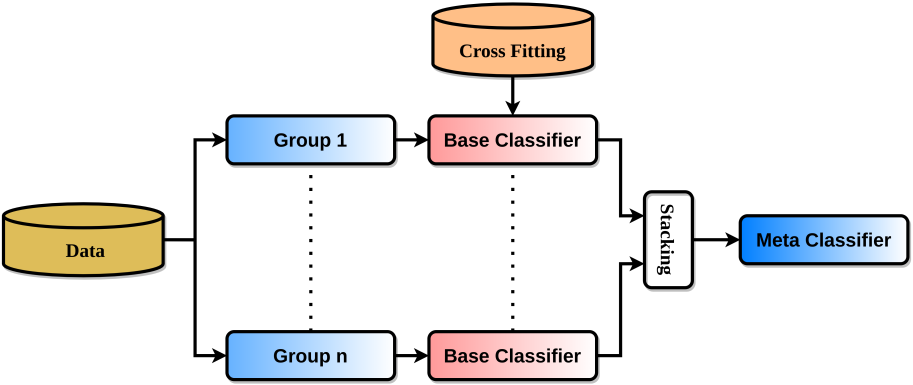
Stacking approach

- Enhancing predictions by stacking multiple prediction models (Wolpert, Neural Networks, 1992)



Stacking approach

- Enhancing predictions by stacking multiple prediction models (Wolpert, Neural Networks, 1992)
- Combine different input domains and groups of variables (Rahim et al., Springer, 2015)

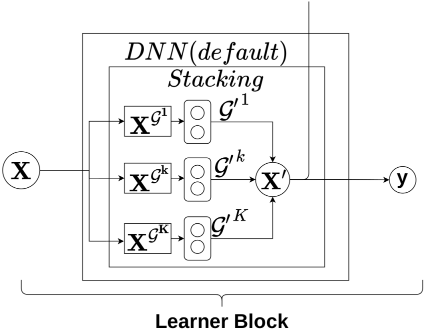


Stacking approach

- Enhancing predictions by stacking multiple prediction models (Wolpert, Neural Networks, 1992)
- Combine different input domains and groups of variables (Rahim et al., Springer, 2015)
- Existing *external* solution $\Longrightarrow$ Making it an *internal* solution

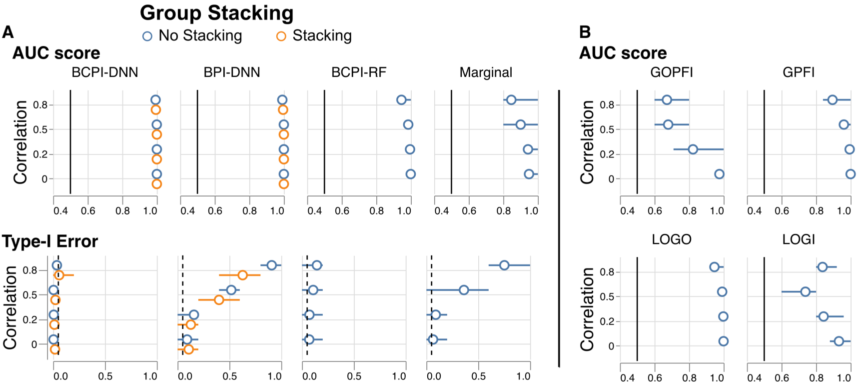
Internal Stacking - Learner Block

G: Original group, **G′**: Linear projected group

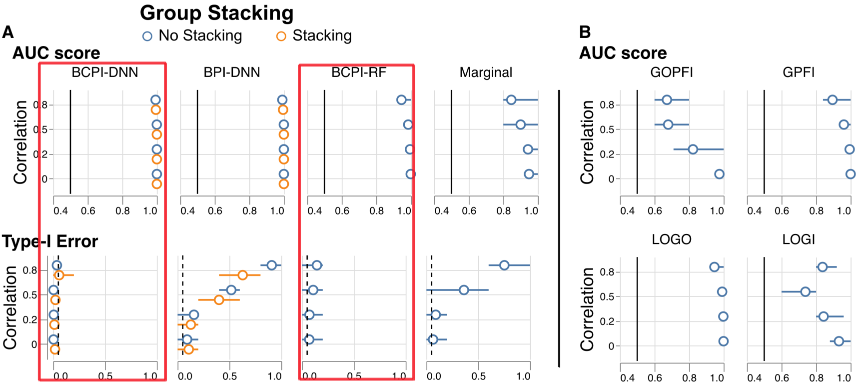


- For a DNN learner $\implies$ Expanding the architecture
- Using sub-linear layers per group followed by a stacking
- Flexible number of output neurons

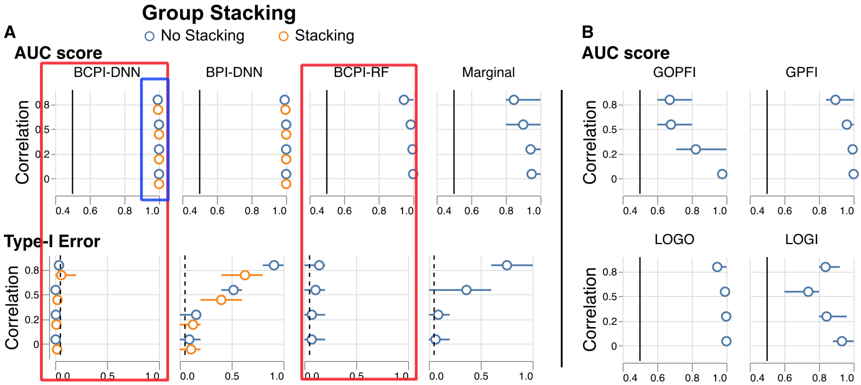
Results on Simulated Data - Group Variable Importance



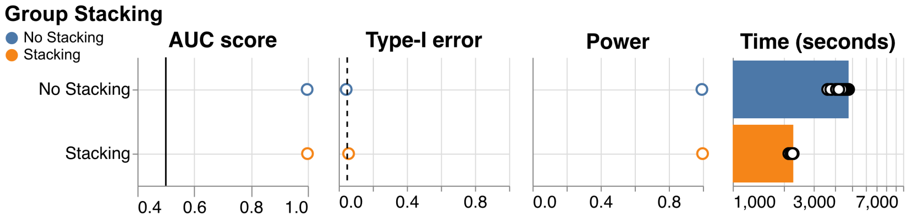
Results on Simulated Data - Group Variable Importance



Results on Simulated Data - Group Variable Importance



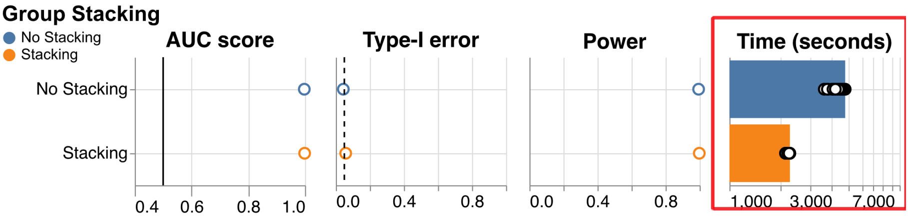
Impact of **Internal Stacking**?



Impact of **Internal Stacking**?

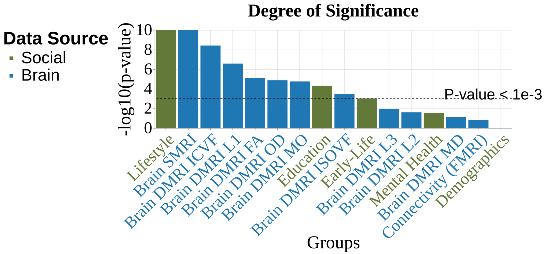
- Applying *Stacking* or *Non Stacking* approaches achieve the same performance (controlling Type-I error)

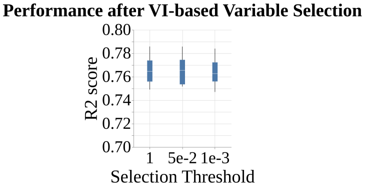
Impact of **Internal Stacking**?



- Applying *Stacking* or *Non Stacking* approaches achieve the same performance (controlling Type-I error)
- The main benefit is an important decrease in time cost

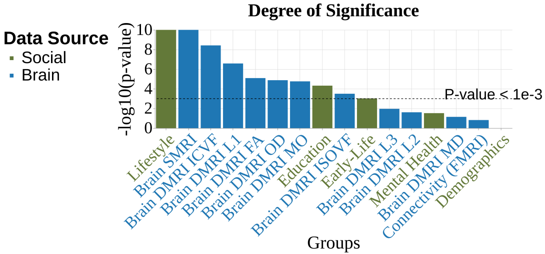
Results on Real Dataset - *UK Biobank*

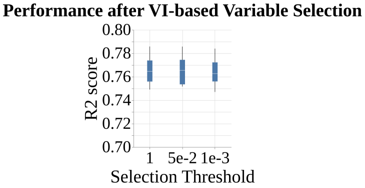




- Left Panel: Degree of significance of pre-defined Brain vs socio-demographic groups
- Right Panel: Performance check after retrieving the non-important groups (having p-value > 0.001)

Results on Real Dataset - *UK Biobank*





- Left Panel: <u>Degree of significance</u> of pre-defined Brain vs socio-demographic groups
- Right Panel: <u>Performance check</u> after retrieving the non-important groups (having p-value > 0.001)

Take-home messages

- *BCPI* provides an indicator of significance under statistical guarantees with a reduced computation time.
- *BCPI* controls type-I error in high-correlation and high-dimensional settings.
- *Internal* stacking maintains the same performance while providing important time savings.
- Deep-learning models are the most accurate for significant variables' assessment.

Thank You for your attention!