Lazy Estimation of Variable Importance for Large Neural Networks

Yue Gao, Abby Stevens, Garvesh Raskutti, Rebecca Willett

ICML 2022

Variable Importance (VI)

In distribution-free settings

Data $(\mathbf{X}, y)$
$\mathbf{X} = (X_1, \ldots, X_p)$

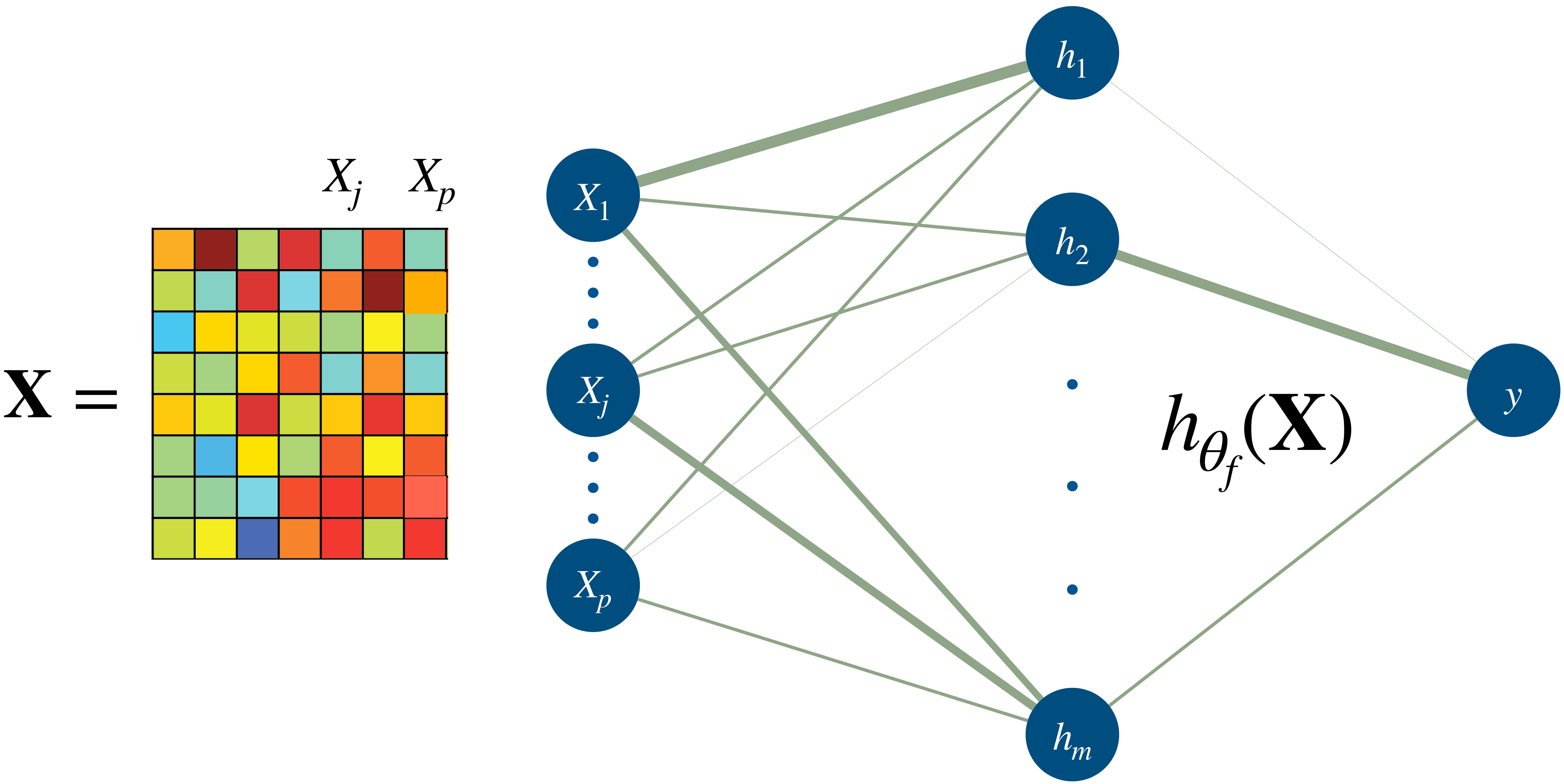
Goal: estimate the **importance** of X_j in **predicting** y with large neural networks h_θ

$$\widehat{VI}_j := V(h_{\theta_f}(\mathbf{X}), y) - V(h_{\theta_{-j}}(\mathbf{X}_{-j}), y)$$

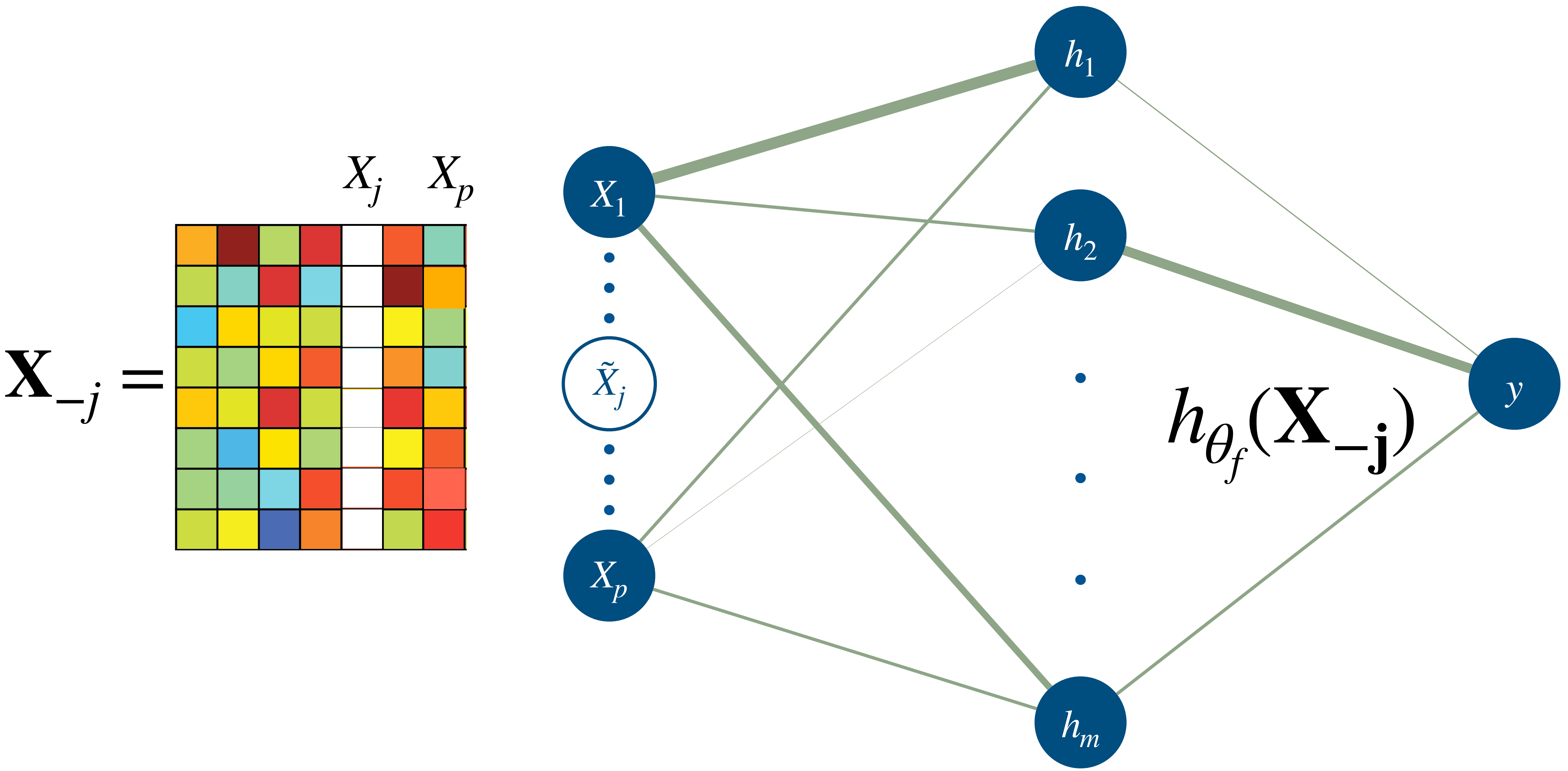
"full model" "reduced model"

Need to estimate reduced model for each variable/subset of variables

Full model

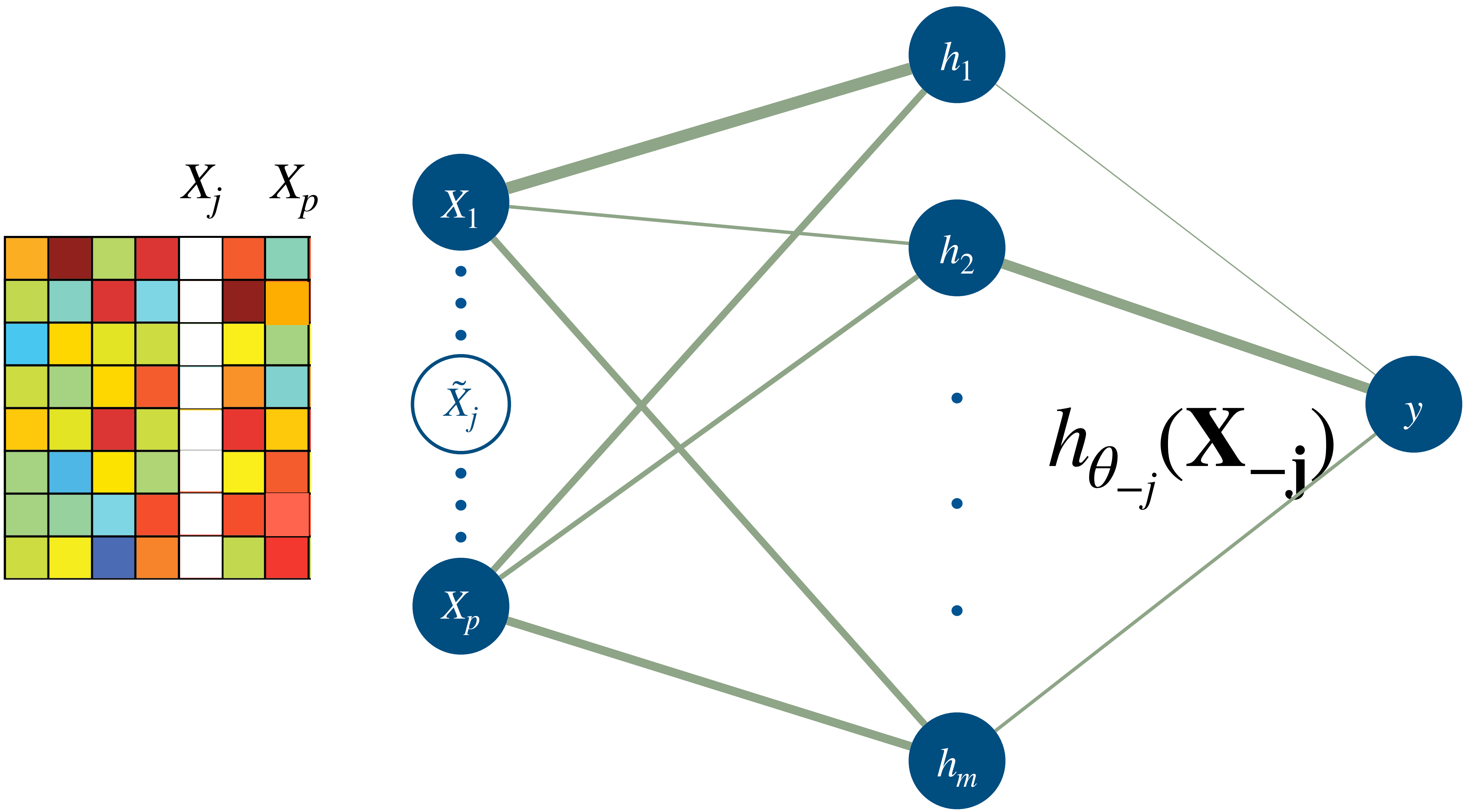


"Dropout"



Reduced model

$$\mathbf{X}_{-j} =$$



Weights have moved, but only slightly

Our contribution: LazyVI

$$h_{\theta_{-j}}(X_{-j}) \approx h_{\theta_f}(X_{-j}) + \nabla_\theta h_\theta(X_{-j})|^T_{\theta=\theta_f}(\theta - \theta_f)$$

Our contribution: LazyVI

$$h_{\theta_{-j}}(X_{-j}) \approx h_{\theta_f}(X_{-j}) + \underbrace{\nabla_\theta h_\theta(X_{-j})\,|^T_{\theta=\theta_f}}_{Z_j}\,\underbrace{(\theta - \theta_f)}_{\Delta\theta_j}$$

1. Linearly estimate
$$\Delta\theta_j = \arg\min_{\omega} \left\{ \frac{1}{n} \left\| \left(y - h_{\theta_f}(X_{-j}) \right) - \omega^T Z_j \right\|_2^2 + \lambda \|\omega\|_2^2 \right\}$$

2. Update parameters: $h_{\theta_{-j}}(\mathbf{X}_{-j}) \approx h_{\theta_f + \Delta\theta_j}(\mathbf{X}_{-j})$

3. $\widehat{VI}_j^{LAZY} = V(h_{\theta_f}(\mathbf{X}), y) - V(h_{\theta_f + \Delta\theta_j}(\mathbf{X}_{-j}), y)$

*We estimate reduced models **linearly** instead of by **fully retraining** a new network*

Theoretical guarantee

Assuming regularity conditions and a sufficiently large regularization parameter $\lambda = O(n^{1/2})$, we show

$$\|h_{\theta_f + \Delta\theta_j}(\mathbf{X}_{-j}) - \mathbb{E}(\mathbf{X}_{-j} \,|\, y)\|_2 = O_P(n^{1/4})$$

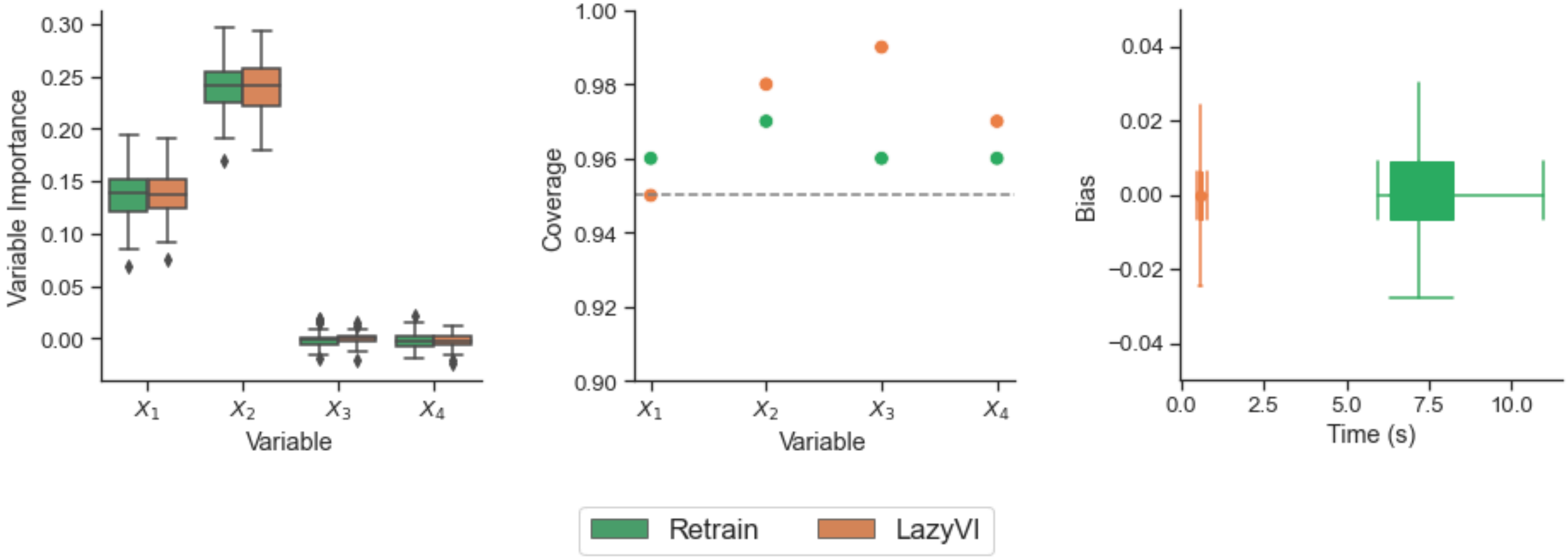
which implies $\widehat{VI}^{LAZY}$ is asymptotically normal and efficient (leveraging framework from (Williamson et al, 2022))

Brian D. Williamson, Peter B. Gilbert, Noah R. Simon & Marco Carone (2022) A General Framework for Inference on Algorithm-Agnostic Variable Importance, Journal of the American Statistical Association, DOI: 10.1080/01621459.2021.2003200

Simulation

Setting from (Williamson et al, 2022):

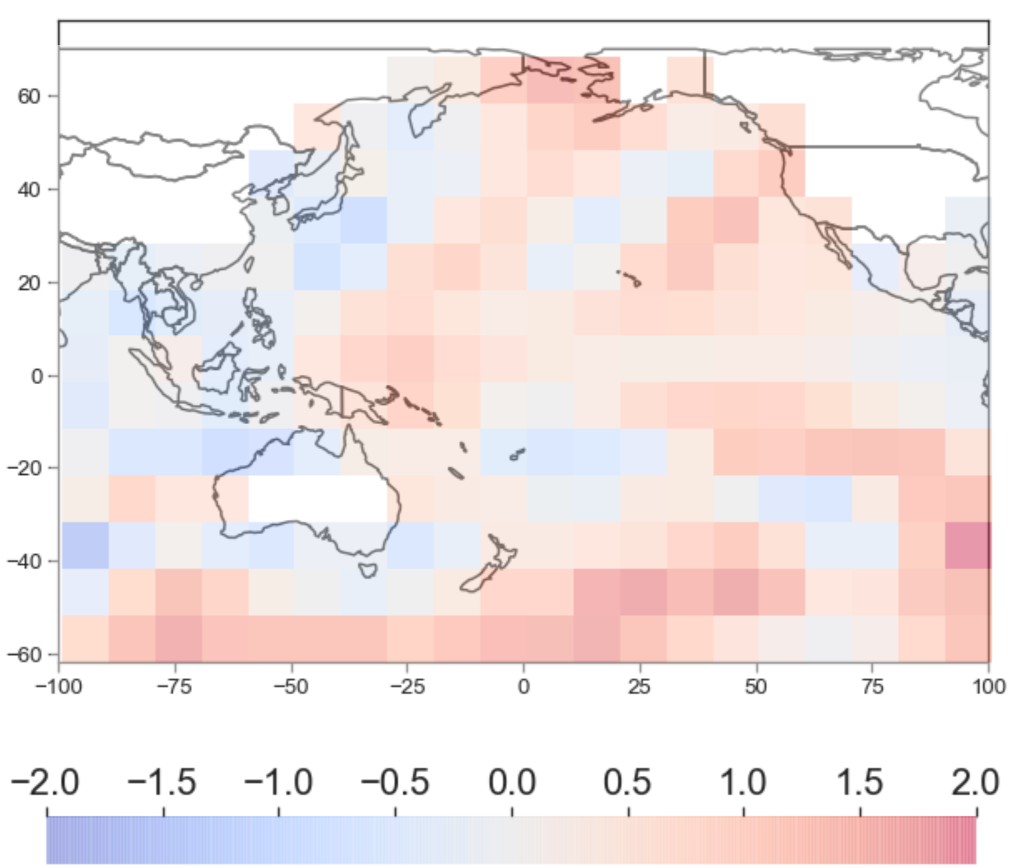
$$\mathbf{X} = (X_1, X_2, X_3, X_4) \quad y = \begin{cases} 1 & \text{if} \quad 2.5X_1 + 3.5X_2 + \epsilon > 0 \\ 0 & \text{otherwise} \end{cases}$$



Climate forecasting

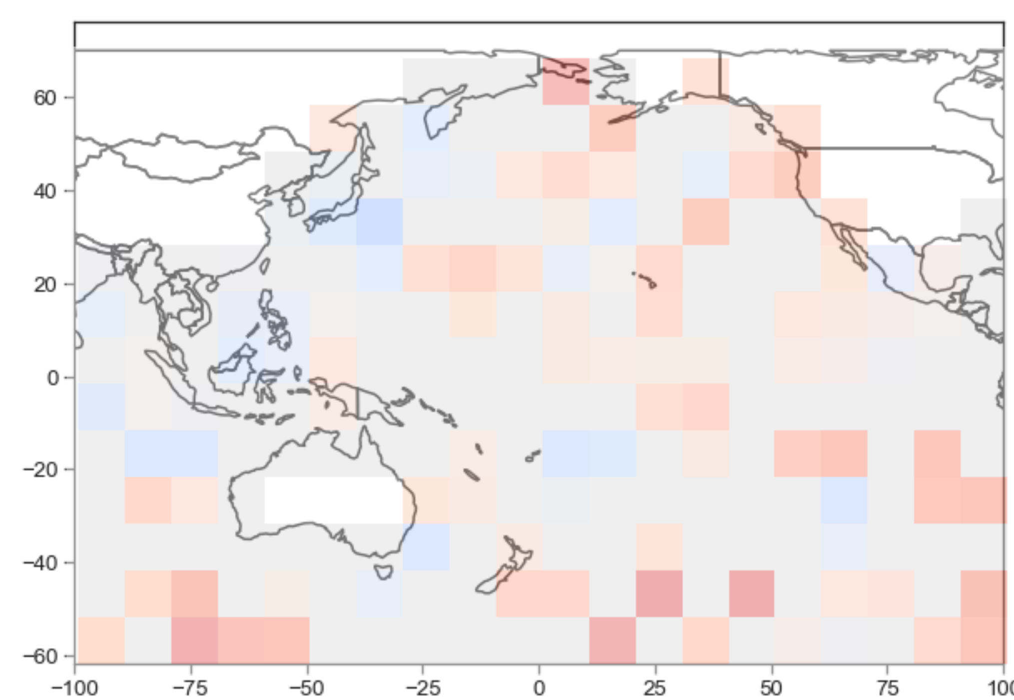
Predict winter precipitation over the Southwestern US using summer Pacific Sea Surface Temperatures

Pacific SSTs



$R^2 = 0.48$

50% removed



Method	**Time (s)**	R^2
Dropout	0	0.35
Retraining	5.6	0.47
LazyVI	0.9	0.46

Conclusion

- We've developed a new method for estimating VI with large neural networks that
  - is **computationally efficient**
  - has **statistical performance guarantees**
  - can be used in a **model-agnostic** and **distribution-free** setting

Thank you!