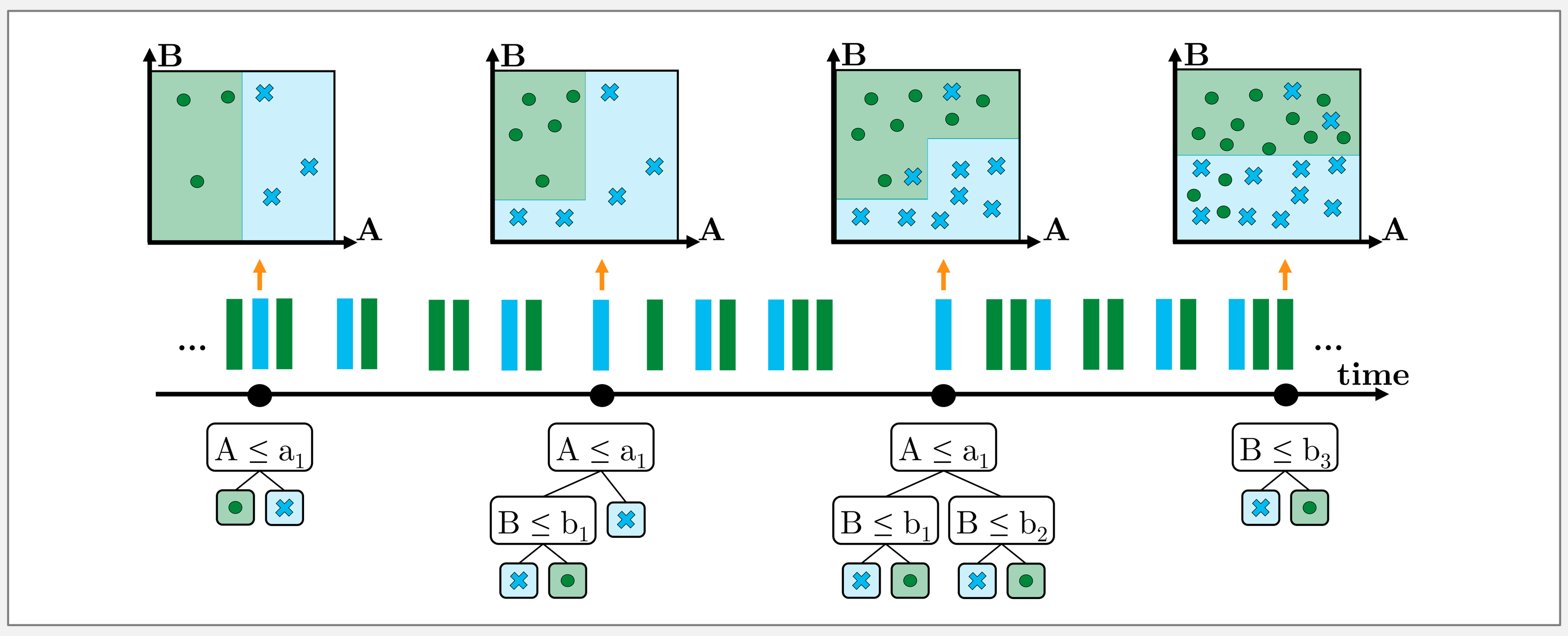
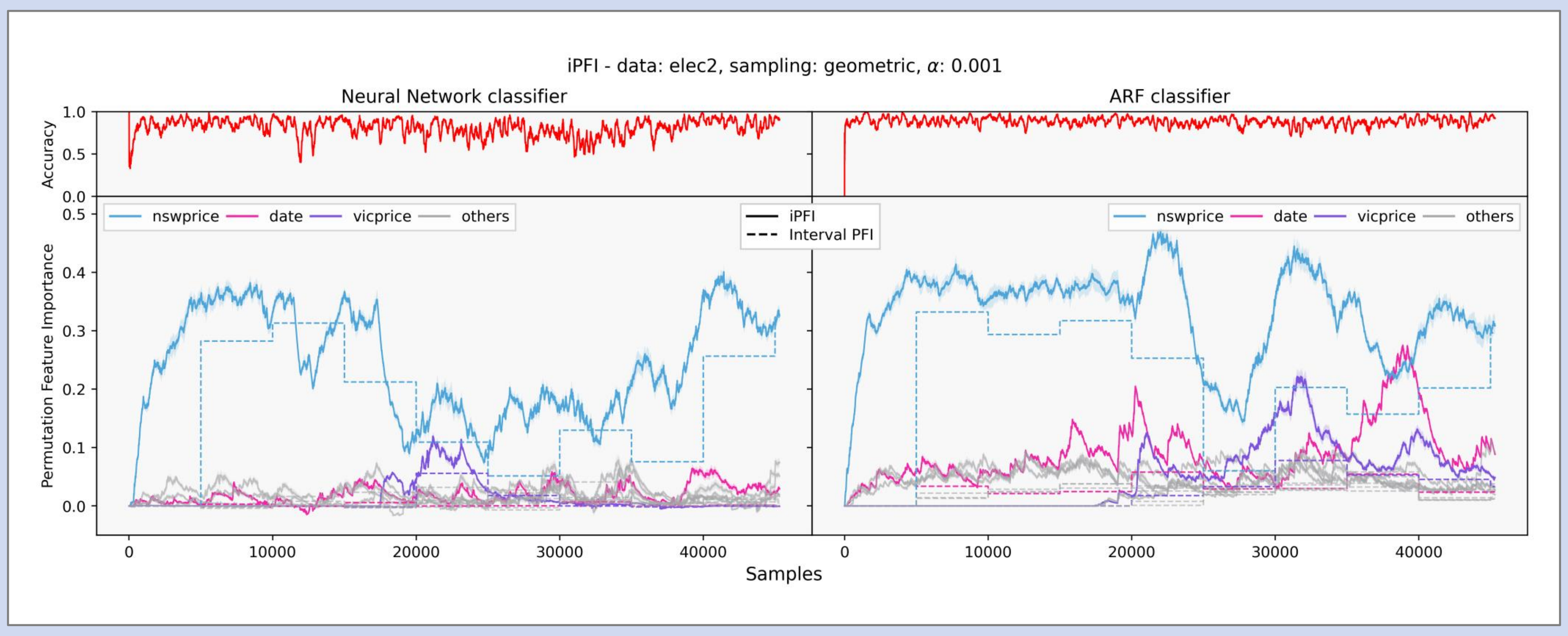
Incremental Permutation Feature Importance (iPFI): Towards Online Explanations on Data Streams

Fabian Fumagalli[1,*], Maximilian Muschalik[2,*], Eyke Hüllermeier[2], Barbara Hammer[1]

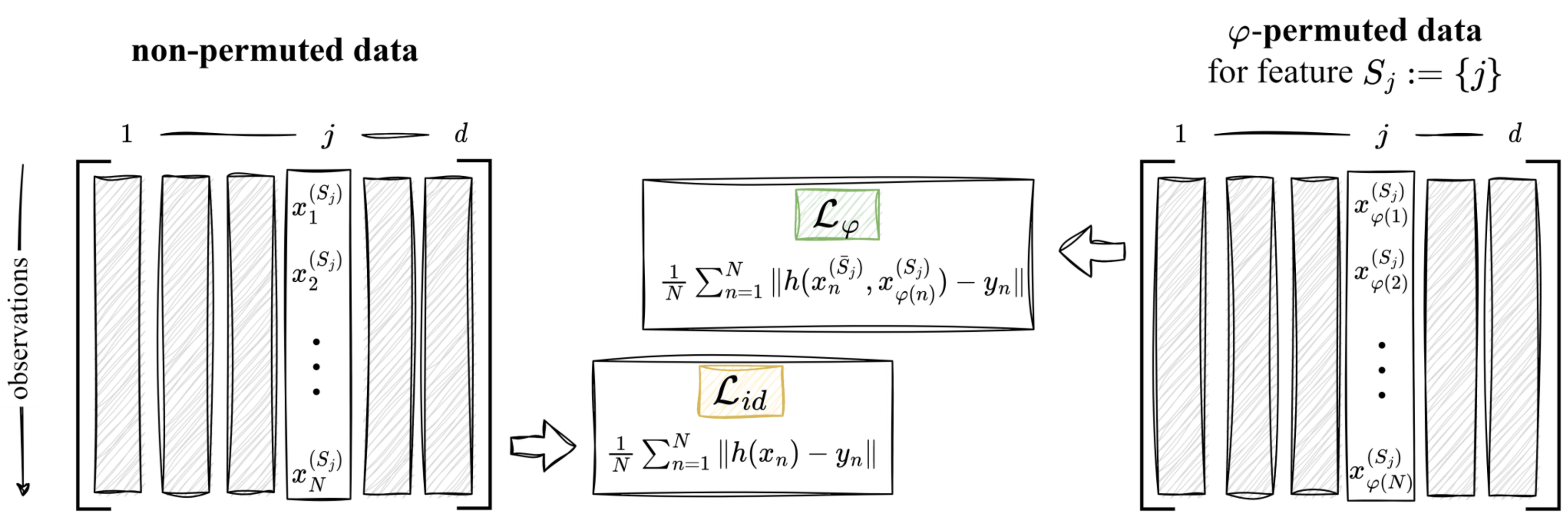
The Problem: Changing Black Box Models



A Solution: Incremental Model-Agnostic Global FI



Permutation Feature Importance (PFI)



Permutation Feature Importance – (Empirical) PFI

Sample permutations $\varphi_1, \ldots, \varphi_M$ uniformly and compute loss increase $\hat{\phi}_\varphi^{(S_j)} := \mathcal{L}_\varphi - \mathcal{L}_{id}$

$$\textbf{(Empirical) PFI: } \hat{\phi}^{(S_j)} := \frac{N}{N-1}\frac{1}{M}\sum_{m=1}^{M} \hat{\phi}^{(S_j)}_{\varphi_m}$$

Global Feature Importance

$\bar{S}_j := D \setminus S_j$

Global Feature Importance (Global FI)

Let $f_{S_j}(x^{(\bar{S}_j)}, y) := \mathbb{E}\left[\|h(x^{(\bar{S}_j)}, X^{(S_j)}) - y\|\right]$ then global FI is defined as

$$\phi^{(S_j)}(h) := \underbrace{\mathbb{E}_{(X,Y)}\left[f_{S_j}(X^{(\bar{S}_j)}, Y)\right]}_{\text{marginalized risk over } S_j} - \underbrace{\mathbb{E}_{(X,Y)}\left[\|h(X) - Y\|\right]}_{\text{risk}}$$

Model Reliance (Fisher, Rudin, and Dominici 2019)

$$\bar{\phi}^{(S_j)} = \frac{1}{N(N-1)}\sum_{n=1}^{N}\sum_{m\neq n}\|h(x_n^{(\bar{S}_j)}, x_m^{(S_j)}) - y_n\| - \frac{1}{N}\sum_{n=1}^{N}\|h(x_n) - y_n\|$$

- is a U-statistic, in particular an unbiased estimator of global FI
- is asymptotically Normal with $\mathbb{V}[\bar{\phi}^{(S_j)}] = \mathcal{O}(1/N)$ and finite sample boundaries

Linking Model Reliance and PFI

Theorem (PFI and Model Reliance are directly linked)

Model reliance is the expectation of PFI over uniformly drawn permutations from $\mathfrak{S}_N$

$$\bar{\phi}^{(S_j)} = \mathbb{E}_{\varphi\sim unif(\mathfrak{S}_N)}[\hat{\phi}^{(S_j)}] = \frac{N}{N-1}\mathbb{E}_{\varphi\sim unif(\mathfrak{S}_N)}\left[\hat{\phi}^{(S_j)}_\varphi\right].$$

PFI $\hat{\phi}^{(S_j)}$

- Properly scaled permutation tests (Breiman 2001)
- Easy to compute in $\mathcal{O}(N)$
- Difficult to analyze theoretically due to dependence on permutations
- Unbiased estimator of $\bar{\phi}^{(S_j)}$
- **Used for computation**

Expected PFI $\bar{\phi}^{(S_j)} = \mathbb{E}_\varphi[\hat{\phi}^{(S_j)}]$

- Expectation of PFI over uniformly sampled permutations
- Difficult to compute in $\mathcal{O}(N^2)$
- U-statistic with strong theoretical guarantees
- Unbiased estimator of global FI
- **Used for theoretical analysis**

References

Fisher A., Rudin C., & Dominici F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. Journal of Machine Learning Research, 20(177), 1–81.
Breiman, L. (2001). Random Forrests. Machine Learning, 45. 5-32.
Vitter, J. S. (1985). Random Sampling with a Reservoir. ACM Transactions on Mathematical Software, 11(1), 37-57.

Incremental Permutation Feature Importance

iPFI Estimator

Online Learning on Data Streams

- Unlimited data stream $(x_0, y_0), \ldots, (x_t, y_t), \ldots$
- Incrementally updated model: $h_{t+1} \leftarrow \text{incrementalUpdate}(h_t, x_t, y_t)$

Incremental PFI (iPFI) $\hat{\phi}_t^{(S_j)}$

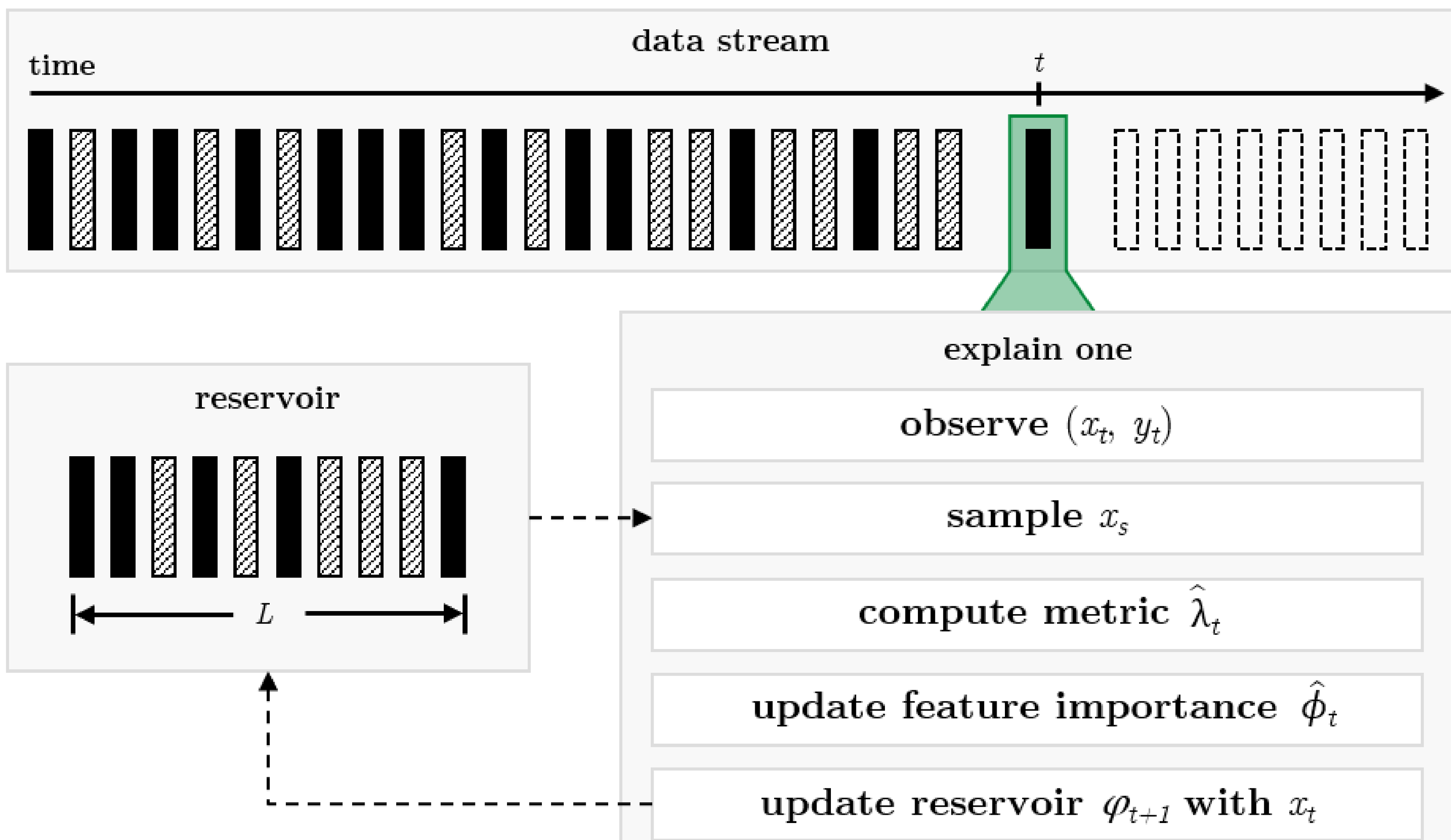
With a **sampling strategy** $\varphi_t : \Omega \to \{0, \ldots, t-1\}$

$$\hat{\lambda}_t^{(S_j)}(x_t, x_{\varphi_t}, y_t) := \|h(x_t^{(\bar{S}_j)}, x_{\varphi_t}^{(S_j)}) - y_t\| - \|h(x_t) - y_t\|$$

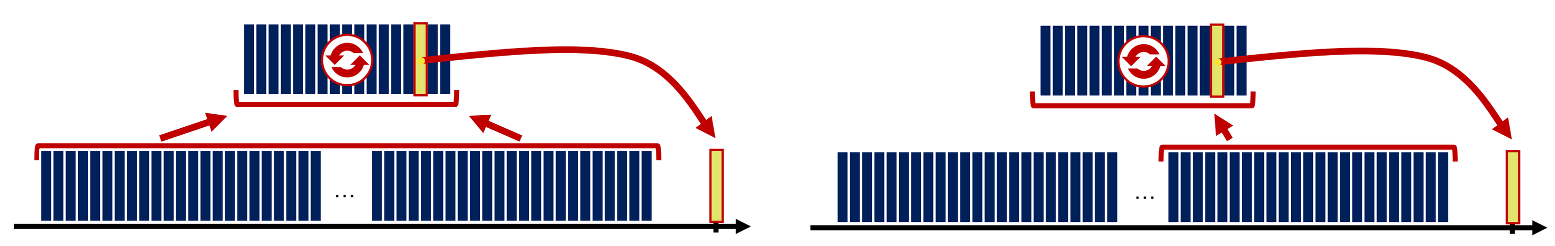
With a **smoothing parameter** $\alpha \in (0,1)$ for $t \geq t_0$ and initial value $\hat{\phi}_{t_0-1}^{(S_j)} := 0$

$$\textbf{iPFI: } \hat{\phi}_t^{(S_j)} := (1-\alpha)\cdot\hat{\phi}_{t-1}^{(S_j)} + \alpha\cdot\hat{\lambda}_t^{(S_j)}(x_t, x_{\varphi_t}, y_t)$$

iPFI Online Explanation Procedure



Incremental Sampling Mechanisms



Theoretical Guarantees

Theorem (Theoretical Guarantees for expected iPFI $\bar{\phi}_t^{(S_j)} := \mathbb{E}_\varphi[\hat{\phi}_t^{(S_j)}]$)

We define a **measure of change** between two timesteps $t_0 \leq s \leq t$ as

$$f_S^\Delta(x^{(\bar{S}_j)}, h_s, h_t) := \mathbb{E}_{\tilde{X}\sim\mathbb{P}_S}[\|h_t(x^{(\bar{S}_j)}, \tilde{X}) - h_s(x^{(\bar{S}_j)}, \tilde{X})\|]$$
$$\Delta_S(h_s, h_t) := \mathbb{E}_X[f_S^\Delta(X, h_s, h_t)] \text{ and } \Delta(h_s, h_t) := \Delta_\emptyset(h_s, h_t).$$

If $\Delta(h_s, h_t) \leq \delta$ and $\Delta_S(h_s, h_t) \leq \delta_S$ for $t_0 \leq s \leq t$ and finite covariances, then

$$|\mathbb{E}[\bar{\phi}_t^{(S_j)}] - \phi^{(S_j)}(h_t)| \leq \delta_S + \delta + \mathcal{O}((1-\alpha)^t) \quad \text{(bias)}$$
$$\mathbb{V}\left[\lim_{t\to\infty}\bar{\phi}_t^{(S_j)}\right] = \mathcal{O}(-\alpha\log(\alpha)) \quad \text{(uniform sampling)}$$
$$\mathbb{V}\left[\lim_{t\to\infty}\bar{\phi}_t^{(S_j)}\right] = \mathcal{O}(\alpha) + \mathcal{O}(1/L) \quad \text{(geometric sampling)}$$

Open Source Implementation: iXAI

- works natively with **riverml.xyz**
- incorporates: iPFI, iSAGE, iPDP, and MDI
- looking for **collaborators**!

ECML PKDD 2023

* denotes equal contribution
1) Bielefeld University, Bielefeld, D-33594, Germany
2) LMU Munich, Munich, D-80539, Germany

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
UNIVERSITÄT BIELEFELD
TRR 318 CONSTRUCTING EXPLAINABILITY
Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation