CONSTRUCTING EXPLAINABILITY
TRR 318

UNIVERSITÄT BIELEFELD

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# iSAGE: An Incremental Version of SAGE for Online Explanation on Data Streams


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

maximilian.muschalik@lmu.de
ffumagalli@techfak.uni-bielefeld.de

[1] LMU Munich, [2] Bielefeld University, [*] equal contribution

# Collaboration

Maximilian Muschalik [1,*]

Fabian Fumagalli [2,*]

Barbara Hammer [2]

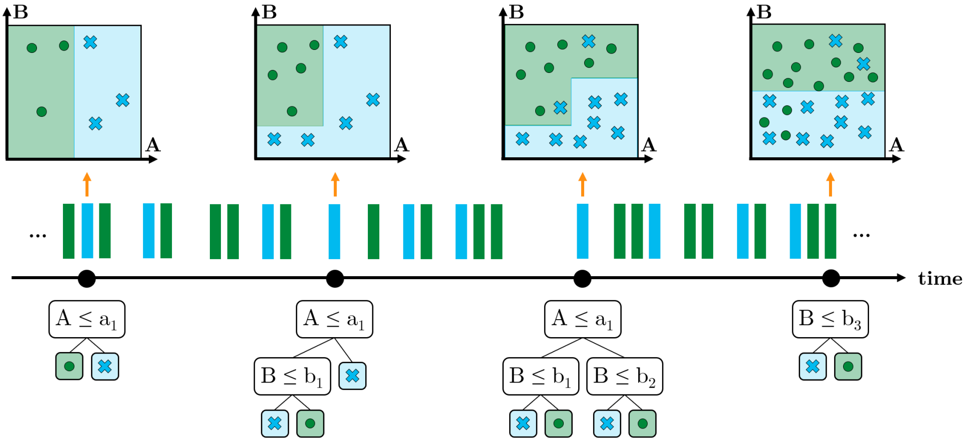
Eyke Hüllermeier [1]

1 LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

2 UNIVERSITÄT BIELEFELD

* denotes equal contribution

# Models in Flux: Incremental Learning from Data Streams




**Various applications**: Bifet and Gavaldà (2007), Gama et al. (2014), Davari et al. (2021), etc.

# Examples of Models in Flux

Fraud Detection

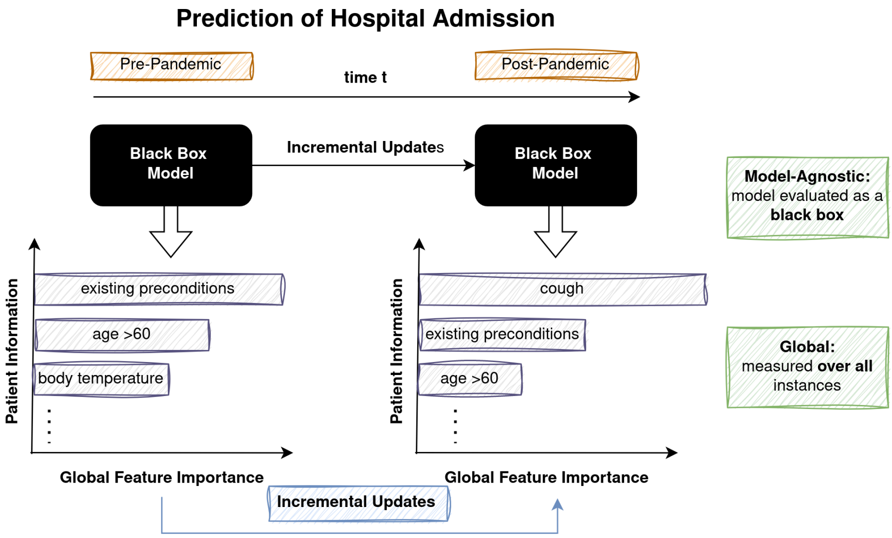
Sensor Networks

Automotive Industry

Predictive Maintenance

Images generated with `Leonardo.ai`.

# Model-Agnostic Explanations with Global Feature Importance



# SAGE: Global Feature Importance

$(X, Y) \sim \mathbb{P}$ data distribution on $\mathcal{X} \times \mathcal{Y}$ $f : \mathcal{X} \to \mathcal{Y}$ black box model $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ loss function

**Explanation Goal: Difference between Model Loss *with* Features and *without***

$$\nu(D) := \underbrace{\mathbb{E}_Y\left[\ell(\bar{y}, Y)\right]}_{\text{no feature information}} - \underbrace{\mathbb{E}_{(X,Y)}\left[\ell(f(X), Y)\right]}_{\text{with feature information}} \quad \text{with mean prediction } \bar{y} := \mathbb{E}_X[f(X)]$$

# SAGE: Global Feature Importance

$(X, Y) \sim \mathbb{P}$ data distribution on $\mathcal{X} \times \mathcal{Y}$ $f : \mathcal{X} \to \mathcal{Y}$ black box model $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ loss function

## Explanation Goal: Difference between Model Loss *with* Features and *without*

$$\nu(D) := \underbrace{\mathbb{E}_Y\left[\ell(\bar{y}, Y)\right]}_{\text{no feature information}} - \underbrace{\mathbb{E}_{(X,Y)}\left[\ell(f(X), Y)\right]}_{\text{with feature information}} \quad \text{with mean prediction } \bar{y} := \mathbb{E}_X[f(X)]$$

## Requirement: Restricted Improvement in Loss given $S \subset D$

$$\nu(S) := \mathbb{E}_Y[\ell(\bar{y}, Y)] - \mathbb{E}_{(X,Y)}\left[\ell(f(X, S), Y)\right] \quad \text{with restricted model } f(x, S)$$

# SAGE: Global Feature Importance

$(X, Y) \sim \mathbb{P}$ data distribution on $\mathcal{X} \times \mathcal{Y}$ $\quad f : \mathcal{X} \to \mathcal{Y}$ black box model $\quad \ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ loss function

**Explanation Goal: Difference between Model Loss *with* Features and *without***

$$\nu(D) := \underbrace{\mathbb{E}_Y\left[\ell(\bar{y}, Y)\right]}_{\text{no feature information}} - \underbrace{\mathbb{E}_{(X,Y)}\left[\ell(f(X), Y)\right]}_{\text{with feature information}} \quad \text{with mean prediction } \bar{y} := \mathbb{E}_X[f(X)]$$

**Requirement: Restricted Improvement in Loss given $S \subset D$**

$$\nu(S) := \mathbb{E}_Y[\ell(\bar{y}, Y)] - \mathbb{E}_{(X,Y)}\left[\ell(f(X, S), Y)\right] \quad \text{with restricted model } f(x, S)$$

**SAGE values $\phi$ of feature $i \in D$, i.e. Shapley values (Shapley 1953)**

$$\phi(i) := \sum_{S \subset D \setminus \{i\}} \frac{1}{d} \binom{d-1}{|S|}^{-1} \left[\nu(S \cup \{i\}) - \nu(S)\right]$$

# Restricted Model

→ **Requirement**: access model with **partial information** (without $\bar{S} := D \setminus S$)

## Interventional SAGE

$$f^{\text{int}}(x, S) := \mathbb{E}\left[f(x^{(S)}, X^{(\bar{S})})\right]$$

*"true to the model"*

## Observational SAGE

$$f^{\text{obs}}(x, S) := \mathbb{E}\left[f(x^{(S)}, X^{(\bar{S})}) \mid X^{(S)} = x^{(S)}\right]$$

*"true to the data"*

**sampling of replacements**

$$\tilde{x}_m^{(\bar{S})}$$

**computation in practice**

$$\hat{f}(x, S) := \frac{1}{M} \sum_{m=1}^{M} f(x^{(S)}, \tilde{x}_m^{(\bar{S})})$$

**Discussion:** Janzing, Minorics, and Blöbaum (2020), Chen et al. (2020), Aas, Jullum, and Løland (2021)

# SAGE: Computation

## SAGE estimator by Covert, Lundberg, and Lee (2020)

$$\hat{\phi}^{\text{SAGE}}(i) := \frac{1}{N} \sum_{n=1}^{N} \underbrace{\ell(\hat{f}(x_n, u_i^-(\pi_n)), y_n) - \ell(\hat{f}(x_n, u_i^+(\pi_n), y_n))}_{\Delta_n(i)}$$

**Illustration** of Shapley Permutation Sampling by Castro, Gómez, and Tejada (2009)

$\pi_n :$ 3 1 5 2 4

# SAGE: Computation

**SAGE estimator by Covert, Lundberg, and Lee (2020)**

$$\hat{\phi}^{\text{SAGE}}(i) := \frac{1}{N}\sum_{n=1}^{N} \underbrace{\ell(\hat{f}(x_n, u_i^-(\pi_n)), y_n) - \ell(\hat{f}(x_n, u_i^+(\pi_n), y_n))}_{\Delta_n(i)}$$

**Illustration** of Shapley Permutation Sampling by Castro, Gómez, and Tejada (2009)

$$\pi_n : \quad \underbrace{3}_{u_3^+} \quad 1 \quad 5 \quad 2 \quad 4$$

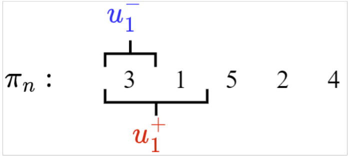
$$\Delta_n(3) = \ell(\hat{f}(x_n, \{\emptyset\}), y_n) - \ell(\hat{f}(x_n, \{3\}, y_n))$$

# SAGE: Computation

## SAGE estimator by Covert, Lundberg, and Lee (2020)

$$\hat{\phi}^{\text{SAGE}}(i) := \frac{1}{N} \sum_{n=1}^{N} \underbrace{\ell(\hat{f}(x_n, u_i^-(\pi_n)), y_n) - \ell(\hat{f}(x_n, u_i^+(\pi_n), y_n))}_{\Delta_n(i)}$$

**Illustration** of Shapley Permutation Sampling by Castro, Gómez, and Tejada (2009)



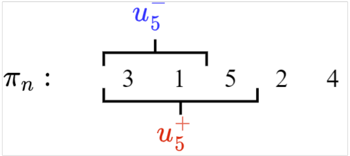

$$\Delta_n(1) = \ell(\hat{f}(x_n, \{3\}), y_n) - \ell(\hat{f}(x_n, \{3, 1\}, y_n))$$

# SAGE: Computation

## SAGE estimator by Covert, Lundberg, and Lee (2020)

$$\hat{\phi}^{\text{SAGE}}(i) := \frac{1}{N} \sum_{n=1}^{N} \underbrace{\ell(\hat{f}(x_n, u_i^-(\pi_n)), y_n) - \ell(\hat{f}(x_n, u_i^+(\pi_n), y_n))}_{\Delta_n(i)}$$

**Illustration** of Shapley Permutation Sampling by Castro, Gómez, and Tejada (2009)




$$\Delta_n(5) = \ell(\hat{f}(x_n, \{3, 1\}), y_n) - \ell(\hat{f}(x_n, \{3, 1, 5\}, y_n))$$

## SAGE estimator by Covert, Lundberg, and Lee (2020)

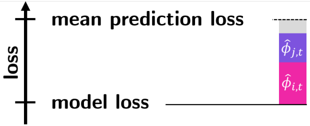
$$\hat{\phi}^{\text{SAGE}}(i) := \frac{1}{N} \sum_{n=1}^{N} \underbrace{\ell(\hat{f}(x_n, u_i^-(\pi_n)), y_n) - \ell(\hat{f}(x_n, u_i^+(\pi_n)), y_n))}_{\Delta_n(i)}$$

**Illustration** of Shapley Permutation Sampling by Castro, Gómez, and Tejada (2009)

$$\pi_n : \quad \underbrace{\overbrace{3 \quad 1}^{u_5^-} \quad 5}_{u_5^+} \quad 2 \quad 4$$

$$\Delta_n(5) = \ell(\hat{f}(x_n, \{3, 1\}), y_n) - \ell(\hat{f}(x_n, \{3, 1, 5\}, y_n))$$

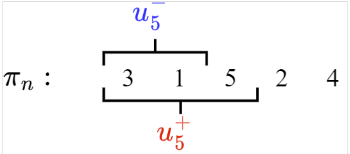
SAGE values follow **efficiency** criterion:



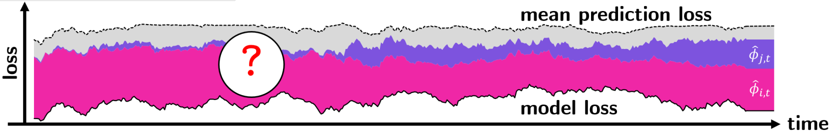
## SAGE estimator by Covert, Lundberg, and Lee (2020)

$$\hat{\phi}^{\text{SAGE}}(i) := \frac{1}{N}\sum_{n=1}^{N} \underbrace{\ell(\hat{f}(x_n, u_i^-(\pi_n)), y_n) - \ell(\hat{f}(x_n, u_i^+(\pi_n), y_n))}_{\Delta_n(i)}$$

**Illustration** of Shapley Permutation Sampling by Castro, Gómez, and Tejada (2009)




$$\Delta_n(5) = \ell(\hat{f}(x_n, \{3, 1\}), y_n) - \ell(\hat{f}(x_n, \{3, 1, 5\}, y_n))$$



# Incremental SAGE (iSAGE) for Explaining Models in Flux

## Online Learning on Data Streams

- unlimited data stream $(x_0, y_0), \ldots, (x_t, y_t), \ldots$
- incrementally updated model $f_{t+1} \leftarrow \text{IncrementalUpdate}(f_t, x_t, y_t)$

**calculation at time t**

$$\Delta_t(i) := \ell(\hat{f}_t(x_t, u_i^-(\pi_t)), y_t) - \ell(\hat{f}_t(x_t, u_i^+(\pi_t)), y_t)$$

**initial computation**

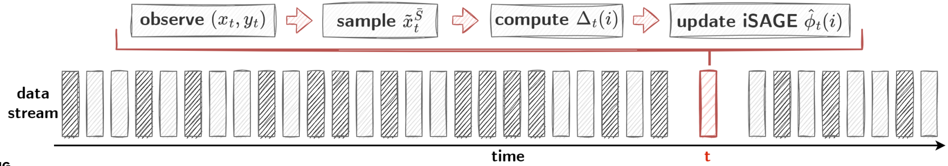
$$\hat{\phi}_{t_0-1}(i) := 0 \text{ for } t \geq t_0 > 0$$

**incremental update to iSAGE**

$$\text{iSAGE: } \hat{\phi}_t(i) = (1-\alpha) \cdot \hat{\phi}_{t-1}(i) + \alpha \cdot \Delta_t(i)$$

**smoothing parameter**

$$\alpha \in (0, 1)$$



# Incremental Sampling Mechanisms

## Interventional iSAGE

**restricted model**

$$f^{\text{obs}}(x, S) := \mathbb{E}\left[f(x^{(S)}, X^{(\bar{S})})\right]$$

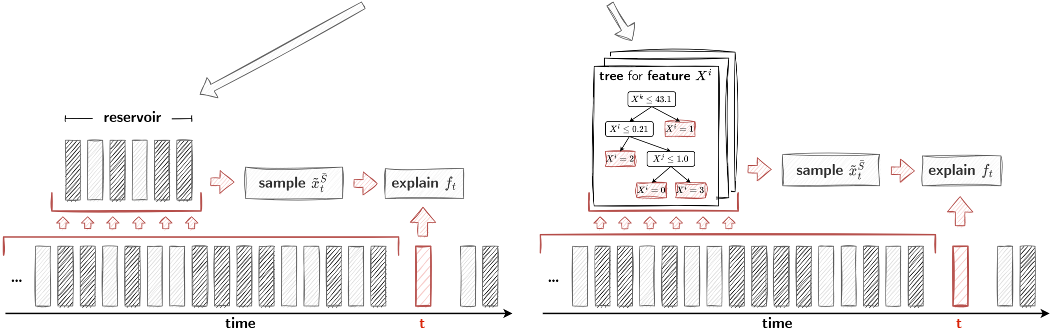
## Observational iSAGE

**restricted model**

$$f^{\text{obs}}(x, S) := \mathbb{E}\left[f(x^{(S)}, X^{(\bar{S})}) \mid X^{(S)} = x^{(S)}\right]$$

**approximation**

$$\hat{f}(x, S) := \frac{1}{M}\sum_{m=1}^{M} f(x^{(S)}, \tilde{x}_m^{(\bar{S})})$$




**Reservoir Sampling** — **Incremental Partition Trees**

# Theoretical Guarantees

**Assumptions:** static model $f_t \equiv f$ and data generating process $(X_t, Y_t) \sim \mathbb{P}_t \equiv \mathbb{P}$

**Theorem (Convergence)**

*For iSAGE $\hat{\phi}_t(i) \to \phi_t(i)$ for $M \to \infty$ and $t \to \infty$.*

**Theorem (Variance)**

*The variance of iSAGE is controlled by $\alpha$, i.e. $\mathbb{V}[\hat{\phi}_t(i)] = \mathcal{O}(\alpha)$.*

**Theorem (Confidence Bounds)**

*Given the SAGE estimator $\hat{\phi}_t^{SAGE}(i)$ computed at time $t$ over all previously observed data points, it holds for iSAGE with $M \to \infty$, $\alpha = \frac{1}{t}$ and every $\epsilon > (1-\alpha)^{t-t_0+1}$ that*

$$\mathbb{P}\left(|\hat{\phi}_t(i) - \hat{\phi}_t^{SAGE}(i)| > \epsilon\right) = \mathcal{O}(\frac{1}{t}).$$

# iSAGE recovers Ground Truth SAGE Values for Models in Flux

**Changing Models**

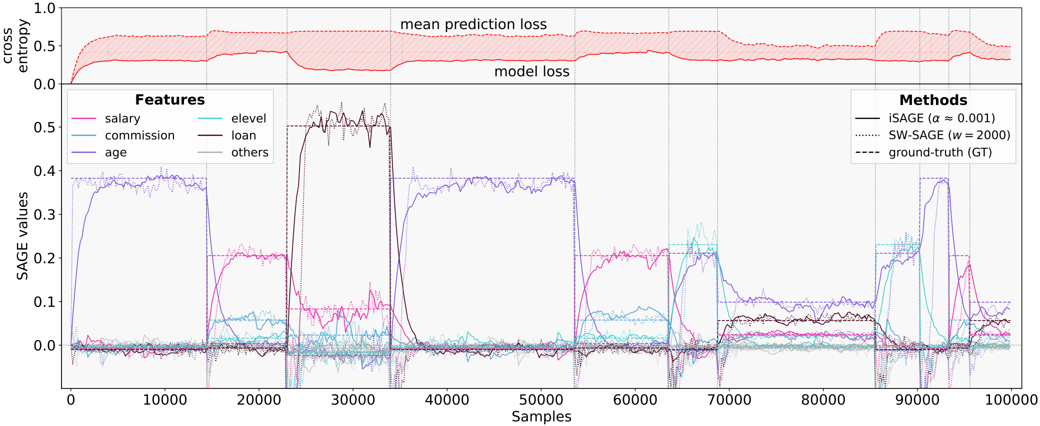
$f_{t_1}$ switch in models
$f_{t_2}$

**Ground Truth (GT)**

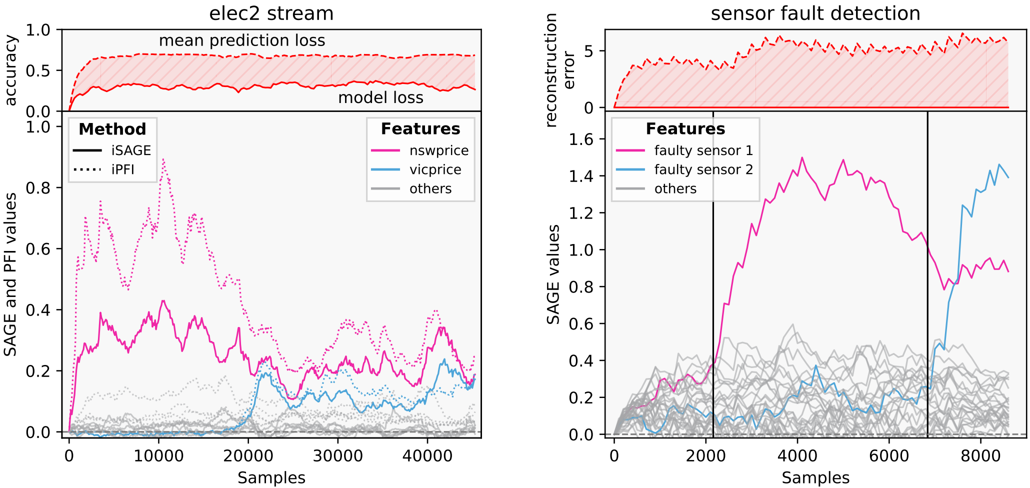
pre-computed ground-truth SAGE values as

**Sliding Window (SW) SAGE**

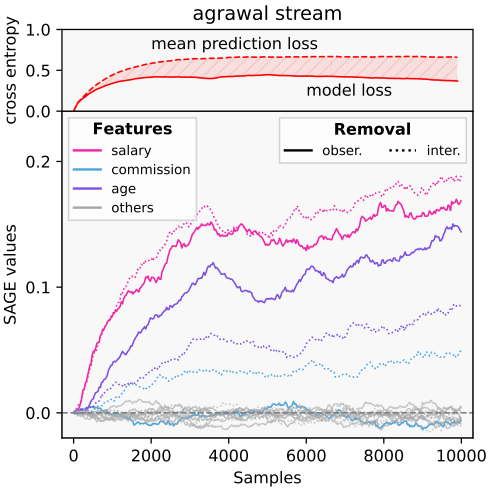
SAGE values computed via a sliding window over the last 2k samples



# Example Applications: Concept Drift Detection



# Observational vs. Interventional iSAGE




**Setting**:

- $X^{\text{com.}}$ depends on $X^{\text{salary}}$
- knowledge about $X^{\text{salary}}$ allows perfect reconstruction of $X^{\text{com.}}$
- target depends indirectly on $X^{\text{com.}}$

**observational** and **interventional** iSAGE retrieve **different** FI scores

- observational iSAGE shows that $X^{\text{com.}}$ is not important
- interventional iSAGE shows that the model has learned to use $X^{\text{com.}}$ (i.e. decision splits exist for $X^{\text{com.}}$)

# The Road Ahead and Open Source Implementation

## Towards Explaining Change.

- iSAGE is a **model-agnostic** XAI method to compute **global SAGE** values for ML models **in flux**.
- Other online XAI methods include **iPFI** (ECMLPKDD'23) and **iPDP** (xAI'23).

## Workshop Friday Afternoon Slot

- Time: **14:00-18:00**
- Room: **PoliTo Room 10i**
- Title: *Explainable Artificial Intelligence: From Static to Dynamic*

docs passing | pypi v0.1.3 | status alpha | License MIT

### Installation

```
pip install ixai
```

### Quickstart

```
>>> for (n, (x, y)) in enumerate(stream, start=1)
...     accuracy.update(y, model.predict_one(x))    # inference
...     incremental_pfi.explain_one(x, y)           # explaining
...     model.learn_one(x, y)                       # learning
```

# References


- Aas, Kjersti, Martin Jullum, and Anders Løland (2021). "Explaining individual predictions when features are dependent: More accurate approximations to Shapley values". In: *Artificial Intelligence* 298, p. 103502. DOI: 10.1016/j.artint.2021.103502.
- Bifet, Albert and Ricard Gavaldà (2007). "Learning from Time-Changing Data with Adaptive Windowing". In: *Proceedings of the Seventh SIAM International Conference on Data Mining (SIAM 2007)*, pp. 443–448. DOI: 10.1137/1.9781611972771.42.
- Castro, Javier, Daniel Gómez, and Juan Tejada (2009). "Polynomial calculation of the Shapley value based on sampling". In: *Computers & Operations Research* 36.5, pp. 1726–1730. DOI: 10.1016/j.cor.2008.04.004.
- Chen, Hugh et al. (2020). "True to the Model or True to the Data?" In: *CoRR* abs/2006.16234. arXiv: 2006.16234.
- Covert, Ian, Scott M. Lundberg, and Su-In Lee (2020). "Understanding Global Feature Contributions With Additive Importance Measures". In: *Advances in Neural Information Processing Systems 33: (NeurIPS 2020)*, pp. 17212–17223.

# References


- Davari, Narjes et al. (2021). "Predictive Maintenance Based on Anomaly Detection Using Deep Learning for Air Production Unit in the Railway Industry". In: *8th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2021)*. IEEE, pp. 1–10. DOI: 10.1109/DSAA53316.2021.9564181.
- Gama, João et al. (2014). "A Survey on Concept Drift Adaptation". In: *ACM Comput. Surv.* 46.4, 44:1–44:37. DOI: 10.1145/2523813.
- Janzing, Dominik, Lenon Minorics, and Patrick Blöbaum (2020). "Feature Relevance Quantification in Explainable AI: A Causal Problem". In: *International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*. Vol. 108. Proceedings of Machine Learning Research. PMLR, pp. 2907–2916. URL: http://proceedings.mlr.press/v108/janzing20a.
- Shapley, L. S. (1953). "A Value for n-Person Games". In: *Contributions to the Theory of Games (AM-28), Volume II*. Princeton University Press, pp. 307–318. ISBN: 9781400881970. DOI: 10.1515/9781400881970-018.

# Complete iSAGE Algorithm

**Algorithm 1** Incremental SAGE (iSAGE)

**Require:** stream $\{x_t, y_t\}_{t=1}^{\infty}$, feature indices $D = \{1, \ldots, d\}$, model $f_t$, loss function $\ell$, and inner samples $m$

Initialize $\hat{\phi}^1 \leftarrow 0, \hat{\phi}^2 \leftarrow 0, \ldots, \hat{\phi}^d \leftarrow 0$, and smoothed mean prediction $y_\emptyset \leftarrow 0$
**for all** $(x_t, y_t) \in$ stream **do**
  Sample $\pi$, a permutation of $D$
  $S \leftarrow \emptyset$
  $y_\emptyset \leftarrow (1-\alpha) \cdot y_\emptyset + \alpha \cdot f(x_t)$ {Udpate mean prediction}
  lossPrev $\leftarrow \ell(y_\emptyset, y_t)$ {Compute mean prediction loss}
  **for** $j = 1$ to $d$ **do** {Iterate over $\pi$}
    $S \leftarrow S \cup \{\pi[j]\}$
    $y \leftarrow 0$
    **for** $k = 1$ to $m$ **do** {Marginalize prediction with $S$}
      Sample $x_k^{(\bar{S})} \sim \mathbb{Q}_t^{(x,S)}$ {interventional (Appendix, Algorithm 2) or observational (Appendix, Algorithm 3)}
      $y \leftarrow y + f_t(x_t^{(S)}, x_k^{(\bar{S})})$
    **end for**
    $\bar{y} \leftarrow \frac{y}{m}$
    loss $\leftarrow \ell(\bar{y}, y_t)$
    $\Delta \leftarrow$ lossPrev $-$ loss
    $\hat{\phi}^{\pi[j]} \leftarrow (1-\alpha) \cdot \phi^{\pi[j]} + \alpha \cdot \Delta$
    lossPrev $\leftarrow$ loss
  **end for**
**end for**
**return** $\phi^1, \phi^2, \ldots, \phi^d$

# Explanation Procedure

## General Explanation Algorithm

**Algorithm 6** Incremental explanation procedure

**Require:** stream $\{x_t, y_t\}_{t=1}^{\infty}$, model $f(.)$, loss function $\mathcal{L}(.)$

1: **for all** $(x_t, y_t) \in$ stream **do**

2: $\quad \hat{y}_t \leftarrow f_t(x_t)$

3: $\quad \hat{\phi}_t \leftarrow \texttt{explain\_one}(x_t, y_t)$

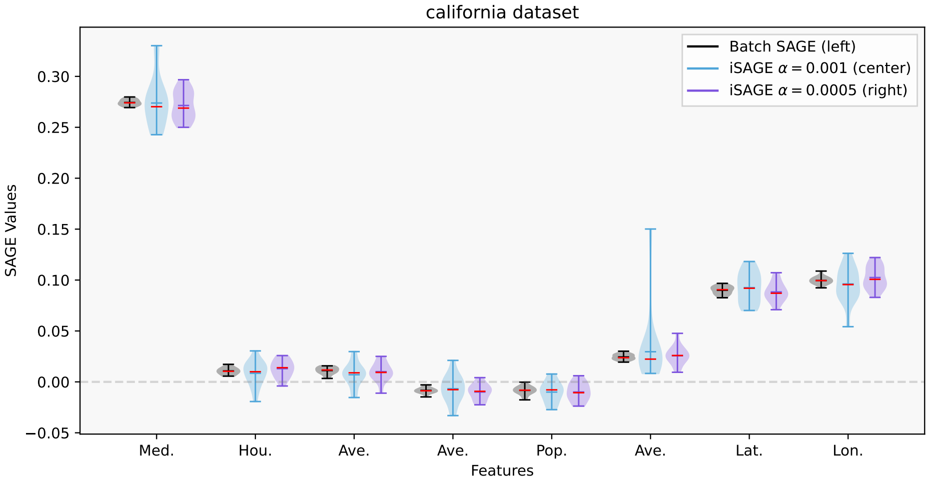
4: $\quad f_{t+1} \leftarrow \texttt{learn\_one}(\mathcal{L}(\hat{y}_t, y_t))$

5: **end for**

- similarly to the **prequential** training: models are explained prequentially.
- data points are used first for explanations (model has not seen the observation, line 3) and then the model is allowed to use it for training (line 4)

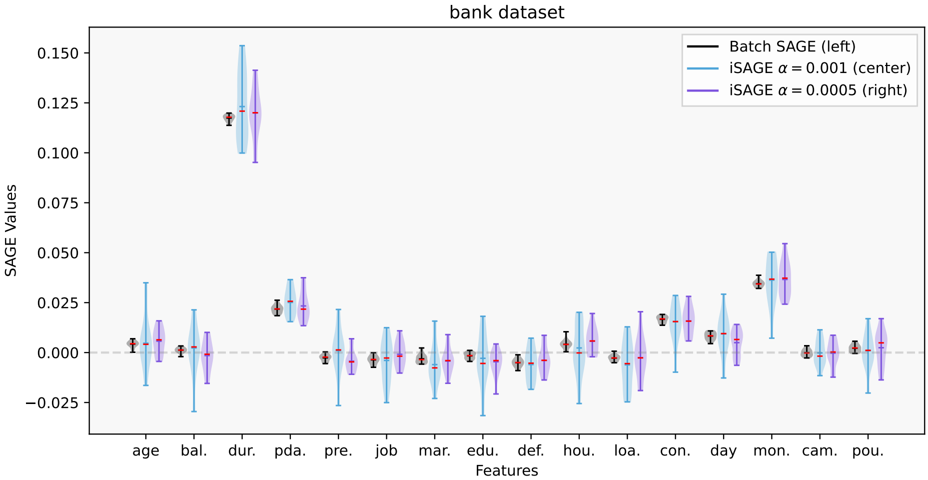
# iSAGE retrieves SAGE values in Static Learning Environments

## California Housing Dataset (Regression) for a Neural Network

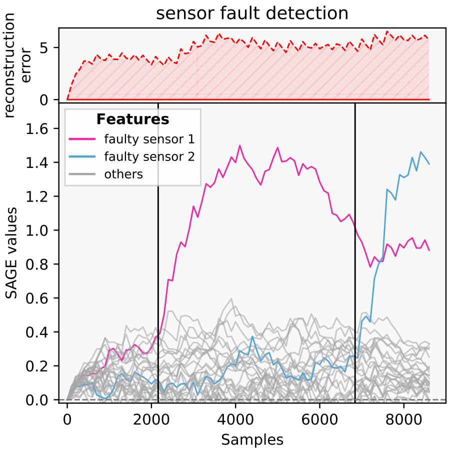


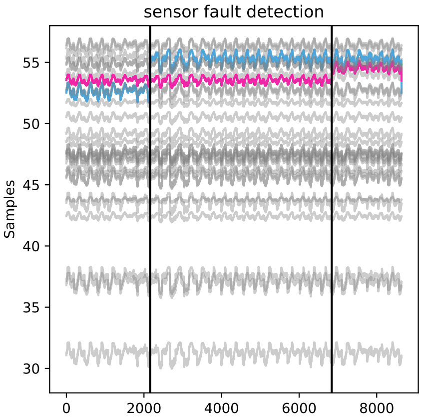

CONSTRUCTING EXPLAINABILITY
TRR 318

## Bank Dataset (Classification) for a Neural Network



# Example Applications: Concept Drift Detection

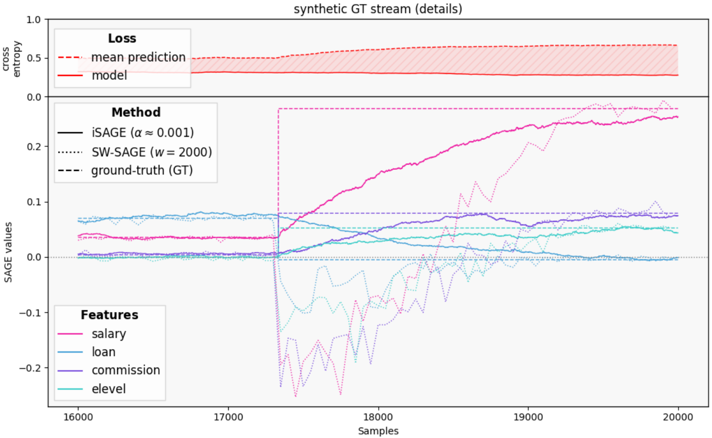



Concept Drift Detection

Raw Sensor Data

# The Problem with Sliding Window (SW) Explanations

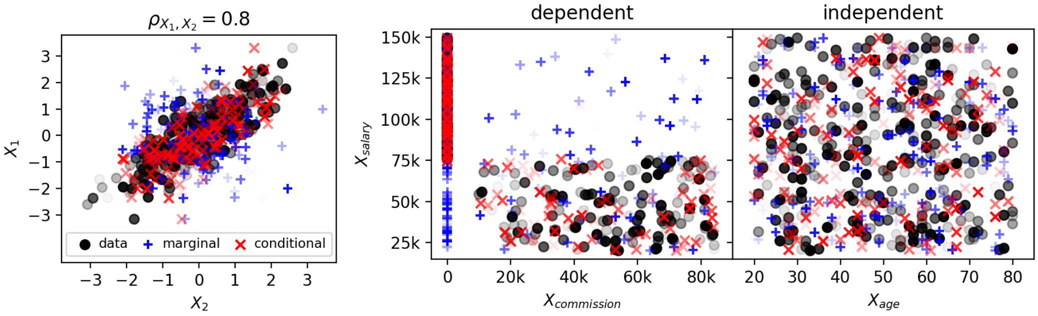



detail view of a ground-truth (GT) data stream

- change point: 17 335
- **before**: iSAGE and SW-SAGE approximate the GT well
- **after**: SW-SAGE recovers more slowly with a high approximation error

! after the change in the model **no previous information** is **useful** anymore

# Observational and Interventional Feature Removal




## Feature Distribution:

- $X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{N}(0,1)$, $X_{\text{age}} \sim \text{unif}([20, 80])$
- $X_{\text{salary}} \sim \text{unif}([20k, 150k])$, and $X_{\text{commission}} = 1(X_{\text{salary}} \leq 75k) \cdot Q$ with $Q \sim \text{unif}([10k, 75k])$