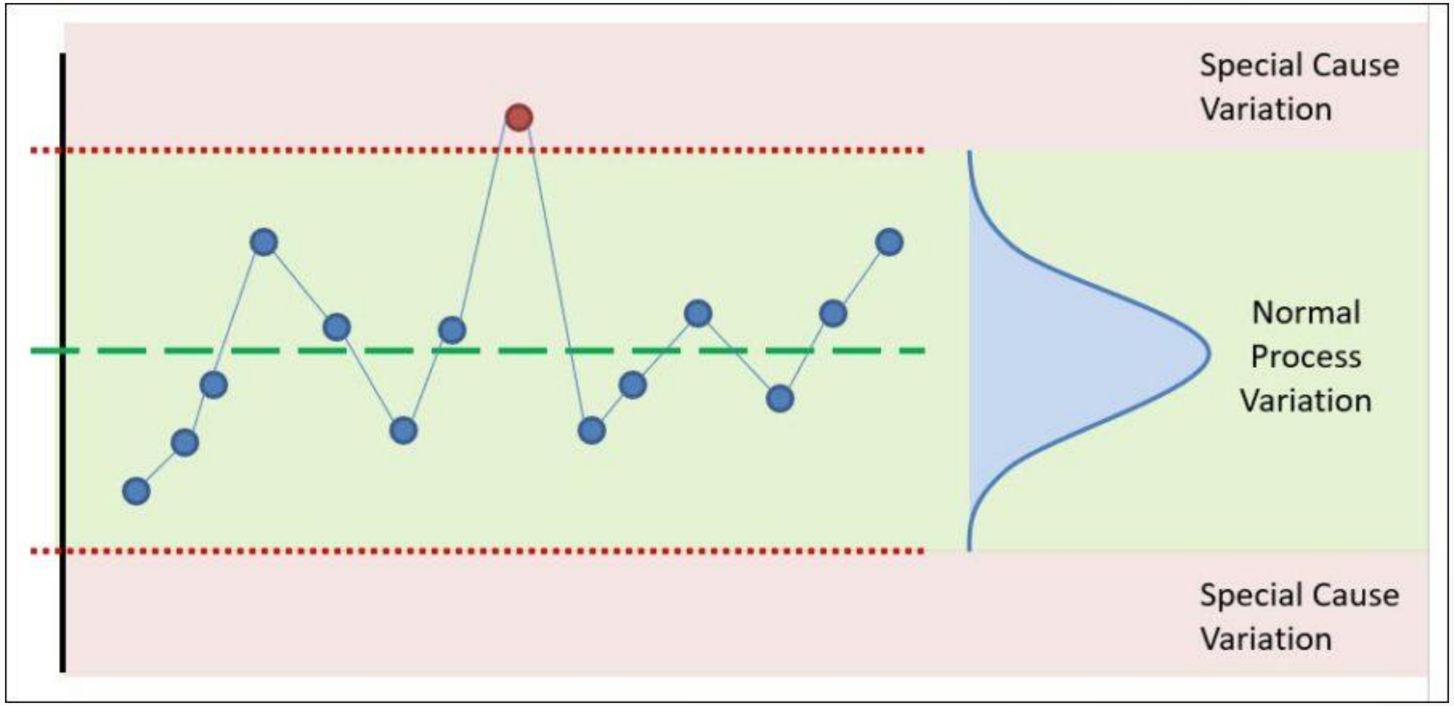
ESTIMATING MODEL PERFORMANCE UNDER COVARIATE SHIFT WITHOUT LABELS

Joint work by Jakub Białek, Juhani Kivimäki, Wojtek Kuberski, and Nikolaos Perrakis

ML model performance monitoring

- If Ground Truth (GT) is available with acceptable lag during inference, one can monitor performance directly.
- Unfortunately, in many cases the GT labels arrive only after substantial lag (when the damage is already done), or in the worst case not at all.
- Can we use model confidence to *estimate* the performance in such cases?



Source: cqeacademy.com

Estimating performance with calibrated confidence scores

- Lets assume that the confidence scores s_i of a binary classifier f, attached to predictions $\hat{y}_i = f(\mathbf{x}_i) \in \{0,1\}$, are *calibrated*. That is, the scores align with the empirical probabilities of the instances belonging to the positive class, formally $P(\hat{y}_i = 1 \mid s_i = s) = s, \ \ \forall s \in [0, 1]$.
- Now, we can treat the prediction as a random variable $\hat{Y}_i = f(X_i)$ that follows a Bernoulli distribution $\hat{Y}_i \sim \mathrm{Bernoulli}(s_i)$.
- A sum of n such (independent) Bernoulli-distributed random variables is a random variable

$$Z = \sum_{i=1}^{n} \hat{Y}_i$$

which follows a Poisson binomial distribution with a probability mass function (PMF)

$$P(Z = k) = \sum_{A \in F_k} \prod_{i \in A} s_i \prod_{j \in A^c} (1 - s_j)$$

Estimating performance with calibrated confidence scores

- Similarly, the expression $1 - s_i$ gives the probability of an instance belonging to the negative class. This enables us to treat the elements of the *confusion matrix* of the classifier for a given dataset or a chunk of data as random variables.
- We can derive the full probability distribution for each of the elements of the confusion matrix
- Then, we can apply a suitable algorithm to estimate any classification metric which can be defined using the confusion matrix.
- This is the CBPE method described in our previous paper.

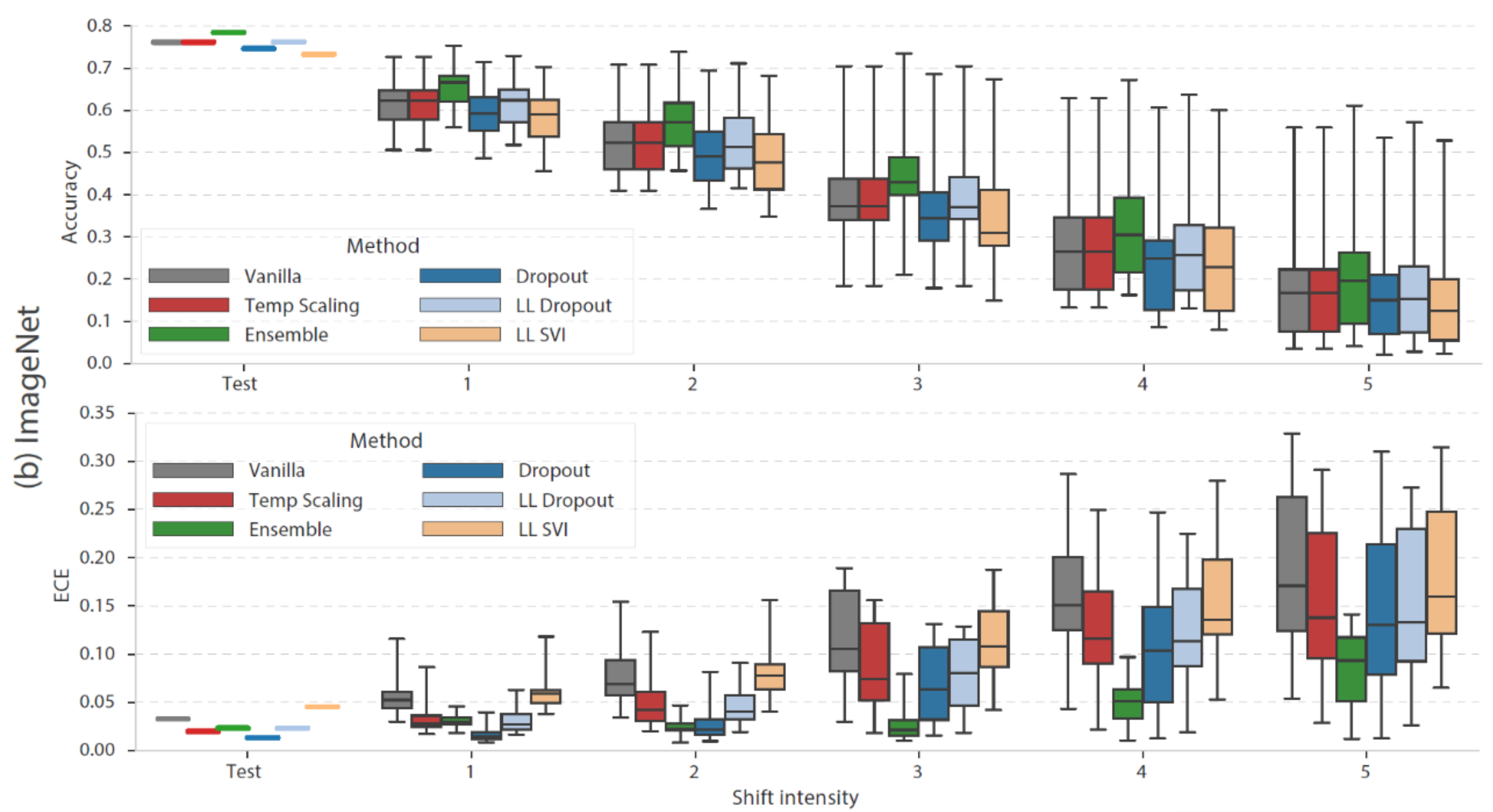
Predicted

Actual

Confusion matrix	Positive	Negative
Positive	X_{TP}	X_{TN}
Negative	X_{FP}	X_{FN}

Problem: Covariate shift hampers calibration

- Typically, model calibration deteriorates under covariate shift. Previous confidence-based methods suffer from this.
- Our method, Probabilistic Adaptive Performance Estimation (PAPE), adapts calibration to the shifted distribution.



Source: Ovadia et al. (2019) “Can you trust your model’s uncertainty?”

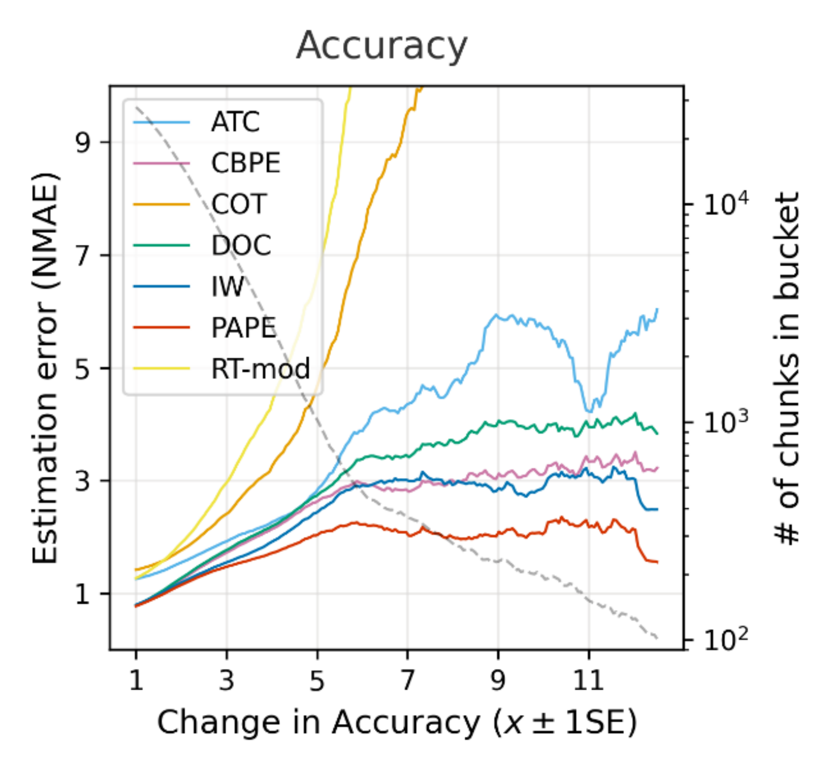
How PAPE fixes calibration of model f under covariate shift?

1. Collect n_s labeled samples (X_i, Y_i) from the source distribution $p_s(\boldsymbol{x}, y)$ and n_t unlabeled samples X_j from the marginal target distribution $p_t(\boldsymbol{x})$.
2. Use X_i and X_j to train a binary classifier h to differentiate between samples from $p_s(\boldsymbol{x})$ and $p_t(\boldsymbol{x})$ (DRE trick).
3. Estimate density ratios $w_{s\to t}(\boldsymbol{x})$ with
$$\hat{w}_{s\to t}(\boldsymbol{x}) = \frac{n_s}{n_t} \cdot \frac{h(\boldsymbol{x})}{1 - h(\boldsymbol{x})}$$
4. Fit a weighted post hoc calibration mapping c to $\{f(X_i), Y_i\}$ using $\hat{w}_{s\to t}(\boldsymbol{x})$ as weights.
5. Use CBPE with $c(f(X_j)$ to estimate performance in $p_t(\boldsymbol{x}, y)$.

PAPE establishes new SotA in performance estimation

- We pitted PAPE against several benchmark methods using over 900 dataset-model combinations from the U.S. census data.
- We compared the estimated performance against actual performance for three metrics; accuracy, F1 score, and AUROC using normalized MAE.
- PAPE provides the best quality estimates out of all tested methods for all three performance metrics.

- PAPE is not guaranteed to work
  - under concept shift ($p_t(y \mid \boldsymbol{x}) \neq p_s(y \mid \boldsymbol{x})$)
  - if $\mathrm{Supp}(p_t(\boldsymbol{x})) \nsubseteq \mathrm{Supp}(p_s(\boldsymbol{x}))$



THANK YOU!

Paper

Code