TRANSFER LEARNING FOR GLOBAL FEATURE IMPORTANCE MEASUREMENTS

Xin Li, Siqi Li, Qiming Wu & Kunyu Yu
Center for Quantitative Medicine, Duke-NUS Medical School, Singapore

ABSTRACT

Understanding feature importance is crucial for conducting interpretable clinical decision-making. However, the reliability of such analyses can be heavily impacted by the available sample size, placing sites with lower data quality and smaller sample sizes at inherent disadvantages. To address the challenge, we propose a model-agnostic transfer learning-based approach for feature importance measurement and evaluate its effectiveness using real-world heterogeneous electronic health records.

1 INTRODUCTION

Feature selection is essential for clinical applications that demand model parsimony, aiming to use as few features as possible while maintaining good prediction performance (Sanchez-Pinto et al., 2018). For instance, in scoring systems applied to patients for quick risk stratification at emergency departments (ED) (Vincent et al., 1996; Forni et al., 2013), the number of features should be limited enough for quick manual calculation while ensuring accuracy. Consequently, feature selection becomes an essential step for such applications. However, feature importance measurements (FIM) can be significantly influenced by data quality and the available sample size. Take logistic regression as an example, the required sample size for robust model fitting increases with the number of features included in the model (van Smeden et al., 2019). Another example is data imbalance, in which majority classes may impact FIM (Zhou & Wong, 2021).

Transfer learning (TL) encompasses a set of strategies for transferring information from one model (source) to another (target), proving effective in enhancing model performance, especially in scenarios where target data has a limited sample size (Chiu et al., 2020; Dhruba et al., 2018). To the best of our knowledge, existing TL applications have all focused on the modeling process, overlooking its potential in data pre-processing steps such as FIM. In this work, we employ TL to assist target sites with small sample sizes in acquiring information from sources with larger sample sizes to improve their FIM. We demonstrate the effectiveness of our method by applying it to real-world electronic health records (EHRs), showing that the performance of the model using features selected by TL-based FIM is superior to those selected by the target's local FIM under the same degree of model parsimony. The implementation of our proposed method is available at this GitHub link.

2 METHOD

2.1 SAGE-TL

Our proposed method builds upon Shapley additive global importance (SAGE) (Covert et al., 2020), a pre-existing model-agnostic global feature importance measurement. Our main contribution is integrating TL into the original SAGE framework, enabling it to leverage pre-existing knowledge for targeting specific data. In the original SAGE approximation, (x, y) is sampled uniformly from the sample data, which may be insufficient if the data is unrepresentative of the true population or has a small sample size. To address this issue, we incorporate a privacy-preserving TL strategy for instance re-weighting into SAGE.

Traditional TL methods, such as the kernel mean matching (KMM) method (Huang et al., 2006) and the Kullback-Leibler importance estimation procedure (KLIEP) (Sugiyama et al., 2007), necessitate simultaneous access and manipulation of both source and target datasets when determining the

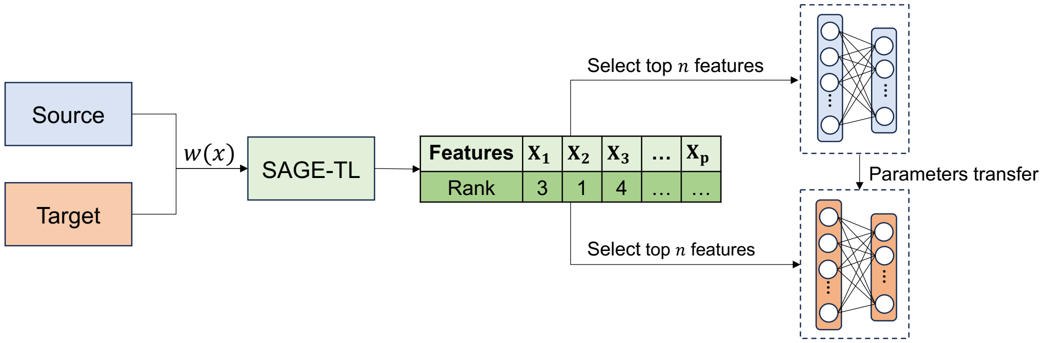

Figure 1: Workflow of SAGE-TL: Calculate instance weights $w(x)$ with uLSIF, then derive SAGE-TL feature importance scores. Use rankings to select the top n features for downstream prediction.

distance disparity. Such approaches raise potential privacy concerns for scenarios where users do not own both source and target data. In consideration of potential privacy constraints in real-world practice, we choose the unconstrained Least-Squares Importance Fitting (uLSIF) (Kanamori et al., 2009) TL strategy (details available in Algorithm1) for our proposed method. We use uLSIF to calculate sample weight ratios of source data $\{x_i^s\}_{i=1}^{n_s}$ to target data $\{x_i^t\}_{i=1}^{n_t}$.

$$\{x_i^s\}_{i=1}^{n_s} \sim p_s(x) \qquad \{x_i^t\}_{i=1}^{n_t} \sim p_t(x)$$

The importance function $w(x)$ is denoted as $w(x) = \frac{p_s(x)}{p_t(x)}$. The key point in our proposed method is that we use $w(x)$ to update the weight of each instance (details available in Algorithm2).

2.2 Experiments

We consider the MIMIC-IV-ED dataset (a large database of emergency department admissions at the Beth Israel Deaconess Medical Center between 2011 and 2019) (Johnson et al., 2023) and choose inpatient mortality as our major outcome for proof-of-concept. We form a cohort of 8728 observations and 30 features and heterogeneously partition it into target and source data with a sample size of 1409 and 7319. A detailed cohort formation description is available in Appendix A.1.

We utilize a neural network consisting of three layers to predict inpatient mortality. The hyper-parameters are determined through empirical testing and fine-tuning to identify suitable values for achieving convergence (see details in Appendix A.2). Sets of features with different values are selected based on the importance ranking obtained from SAGE-TL. We compare the performance of inpatient mortality prediction using feature selection by SAGE-TL and the baseline method (SAGE) using the Area Under the Receiver Operating Characteristic (AUROC) analysis.

3 Results And Conclusion

Table 1: Model performance comparison (AUROC values) using different numbers of selected features

Numbers of selected top n features	8	10	12	14	16
Target local	0.685	0.643	0.672	0.621	0.564
Target TL	0.691	0.750	0.752	0.718	0.683

As shown in Table 1, given the same degree of model parsimony, the models built upon features selected by our proposed method have better prediction performance compared to those built using the baseline method.

Our work is among the first few studies exploring the usage of TL in non-modeling processes, and our results show its potential with studies in clinical and other domains that suffer from relatively small sample sizes and have a strong need for model parsimony.

URM Statement

The authors acknowledge that all authors of this work meet the URM criteria of ICLR 2024 Tiny Papers Track.

References

Nashat Alrefai, Othman Ibrahim, Hafiz Muhammad Faisal Shehzad, Abdelrahman Altigani, Waheeb Abu-ulbeh, Malek Alzaqebah, and Mutasem K Alsmadi. An integrated framework based deep learning for cancer classification using microarray datasets. *Journal of Ambient Intelligence and Humanized Computing*, 14(3):2249–2260, 2023.

Huan-Jung Chiu, Tzuu-Hseng S Li, and Ping-Huan Kuo. Breast cancer–detection system using pca, multilayer perceptron, transfer learning, and support vector machine. *IEEE Access*, 8:204309–204324, 2020.

Ian Covert, Scott M Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33:17212–17223, 2020.

Saugato Rahman Dhruba, Raziur Rahman, Kevin Matlock, Souparno Ghosh, and Ranadip Pal. Application of transfer learning for cancer drug sensitivity prediction. *BMC bioinformatics*, 19(17):51–63, 2018.

Lui G Forni, Thomas Dawes, Hamish Sinclair, Elizabeth Cheek, Vivien Bewick, Mark Dennis, and Richard Venn. Identifying the patient at risk of acute kidney injury: a predictive scoring system for the development of acute kidney injury in acute medical patients. *Nephron clinical practice*, 123(3-4):143–150, 2013.

Andre L Holder, Supreeth P Shashikumar, Gabriel Wardi, Timothy G Buchman, and Shamim Nemati. A locally optimized data-driven tool to predict sepsis-associated vasopressor use in the icu. *Critical Care Medicine*, 49(12):e1196–e1205, 2021.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19, 2006.

Sangwon Hwang, Chanwoo Gwon, Dong Min Seo, Jooyoung Cho, Jang-Young Kim, Young Uh, et al. A deep neural network for estimating low-density lipoprotein cholesterol from electronic health records: Real-time routine clinical application. *JMIR Medical Informatics*, 9(8):e29331, 2021.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445, 2009.

Jin Li, Yu Tian, Runze Li, Tianshu Zhou, Jun Li, Kefeng Ding, and Jingsong Li. Improving prediction for medical institution with limited patient data: Leveraging hospital-specific data based on multicenter collaborative research network. *Artificial Intelligence in Medicine*, 113:102024, 2021.

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173, 2022.

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning in large-scale gaussian graphical models with false discovery rate control. *Journal of the American Statistical Association*, 118(543):2171–2183, 2023.

L Nelson Sanchez-Pinto, Laura Ruth Venable, John Fahrenbach, and Matthew M Churpek. Comparison of variable selection methods for clinical predictive modeling. *International journal of medical informatics*, 116:10–17, 2018.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in neural information processing systems*, 20, 2007.

Turki Turki, Zhi Wei, and Jason TL Wang. A transfer learning approach via procrustes analysis and mean shift for cancer drug sensitivity prediction. *Journal of bioinformatics and computational biology*, 16(03):1840014, 2018.

Maarten van Smeden, Karel GM Moons, Joris AH de Groot, Gary S Collins, Douglas G Altman, Marinus JC Eijkemans, and Johannes B Reitsma. Sample size for binary logistic prediction models: beyond events per variable criteria. *Statistical methods in medical research*, 28(8):2455–2474, 2019.

J L Vincent, Rui Moreno, Jukka Takala, Sheila Willatts, Arnaldo De Mendonça, Hajo Bruining, CK Reinhart, PeterM Suter, and Lambertius G Thijs. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure: On behalf of the working group on sepsis-related problems of the european society of intensive care medicine (see contributors to the project in the appendix), 1996.

Xuan Wang, Harrison G Zhang, Xin Xiong, Chuan Hong, Griffin M Weber, Gabriel A Brat, Clara-Lea Bonzel, Yuan Luo, Rui Duan, Nathan P Palmer, et al. Survmaximin: robust federated approach to transporting survival risk prediction models. *Journal of biomedical informatics*, 134: 104176, 2022.

Feng Xie, Jun Zhou, Jin Wee Lee, Mingrui Tan, Siqi Li, Logasan S/O Rajnthern, Marcel Lucas Chee, Bibhas Chakraborty, An-Kwok Ian Wong, Alon Dagan, et al. Benchmarking emergency department prediction models with machine learning and public electronic health records. *Scientific Data*, 9(1):658, 2022.

Xiangzhou Zhang, Kang Liu, Borong Yuan, Hongnian Wang, Shaoyong Chen, Yunfei Xue, Weiqi Chen, Mei Liu, and Yong Hu. A hybrid adaptive approach for instance transfer learning with dynamic and imbalanced data. *International Journal of Intelligent Systems*, 37(12):11582–11599, 2022.

Pei-Yuan Zhou and Andrew KC Wong. Explanation and prediction of clinical data with imbalanced class distribution based on pattern discovery and disentanglement. *BMC medical informatics and decision making*, 21(1):1–15, 2021.

A Appendix

A.1 Data And Cohort Formation

MIMIC IV Emergency Department (MIMIC-IV-ED) (Johnson et al., 2023) is a public dataset that contains over 400,000 ED visit episodes, and we follow the data extraction pipelines by Xie et al. (2022) to obtain a master dataset, from which the cohort used in this study is formed by excluding observations with missing values. Specifically, we form a cohort of 8728 samples by filtering the master dataset to include only Emergency Department (ED) admissions of Asian patients. We remove observations with missing values in candidate features, including age, gender, pulse (beats/min), respiration (times/min), peripheral capillary oxygen saturation (SpO_2; %), diastolic blood pressure (mm Hg), systolic blood pressure (mm Hg), pain scale, weight loss, depression, temperature (Celsius), fever/chills, drug abuse, ED visit in the past 3 months (times), blood loss anemia, coagulopathy and comorbidities including myocardial infarction, congestive heart failure, stroke, dementia, chronic pulmonary disease, rheumatoid disease, peptic ulcer disease, kidney disease, paralysis, diabetes, peripheral vascular disease and renal disease.

We assign each observation a unique probability (depending on the age feature) of being selected as a source or target domain. As a result, patients with higher age values also have a higher probability of being assigned to the target domain, thus realizing the difference in data distribution between source and target data.

A.2 Hyperparameters

Table 2: Hyperparameter values for fine-tuning with chosen values highlighted

Hyperparameter	**Candidate values**
Learning Rate	**0.01**, 0.005, 0.001
Momentum	0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, **0.9**
Weight Decay	**0**, 0.001, 0.0001

A.3 Pseudo Code

```
Algorithm 1: uLSIF (Kanamori et al., 2009)
Input  : $\{x_i^t\}_{i=1}^{n_t}$ and $\{x_j^s\}_{j=1}^{n_s}$
Output: $\hat{w}(x)$
$b \leftarrow \min(100, n_s)$; $n \leftarrow \min(n_t, n_s)$
Randomly choose $b$ centers $\{c_\ell\}_{\ell=1}^{b}$ from $\{x_j^s\}_{j=1}^{n_s}$ without replacement;
for each candidate of Gaussian width $\sigma$ do
    $\hat{H}_{\ell,\ell'} \leftarrow \frac{1}{n_t}\sum_{i=1}^{n_t} \exp(-\frac{||x_i^t - c_\ell||^2 + ||x_i^t - c_{\ell'}||^2}{2\sigma^2})$ for $\ell, \ell' = 1, 2, ..., b$;
    $\hat{h}_\ell \leftarrow \frac{1}{n_s}\sum_{j=1}^{n_s} \exp(-\frac{||x_j^s - c_\ell||^2}{2\sigma^2})$ for $\ell = 1, 2, ..., b$;
    $X_{\ell,i}^t \leftarrow \exp(-\frac{||x_i^t - c_\ell||^2}{2\sigma^2})$ for $i = 1, 2, ..., n$ and $\ell = 1, 2, ..., b$;
    $X_{\ell,i}^s \leftarrow \exp(-\frac{||x_i^s - c_\ell||^2}{2\sigma^2})$ for $i = 1, 2, ..., n$ and $\ell = 1, 2, ..., b$;
    for each candidate of regularization parameter $\lambda$ do
        $\hat{B} \leftarrow \hat{H} + \frac{\lambda(n_t - 1)}{n_t} I_b$;
        $B_0 \leftarrow \hat{B}^{-1}\hat{h}1_n^\top + \hat{B}^{-1}X^t \mathrm{diag}\left(\frac{\hat{h}^\top \hat{B}^{-1} X^t}{n_t 1_n^\top - 1_b^\top (X^t * \hat{B}^{-1} X^t)}\right)$;
        $B_1 \leftarrow \hat{B}^{-1}X^s + \hat{B}^{-1}X^t \mathrm{diag}\left(\frac{1_b^\top (X^s * \hat{B}^{-1} X^t)}{n_t 1_n^\top - 1_b^\top (X^t * \hat{B}^{-1} X^t)}\right)$;
        $B_2 \leftarrow \max\left(O_{b \times n}, \frac{n_t - 1}{n_t(n_s - 1)}(n_s B_0 - B_1)\right)$;
        $w_t \leftarrow (1_b^\top (X^t * B_2))^\top$; $w_s \leftarrow (1_b^\top (X^s * B_2))^\top$;
        $\mathrm{LOOCV}(\sigma, \lambda) \leftarrow \frac{w_t^\top w_t}{2n} - \frac{1_n^\top w_s}{n}$;
    end
end
$(\hat{\sigma}, \hat{\lambda}) \leftarrow \arg\min_{(\sigma,\lambda)} \mathrm{LOOCV}(\sigma, \lambda)$;
$\tilde{H}_{\ell,\ell'} \leftarrow \frac{1}{n_t}\sum_{i=1}^{n_t} \exp(-\frac{||x_i^t - c_\ell||^2 + ||x_i^t - c_{\ell'}||^2}{2\hat{\sigma}^2})$ for $\ell, \ell' = 1, 2, ..., b$;
$\tilde{h}_\ell \leftarrow \frac{1}{n_s}\sum_{j=1}^{n_s} \exp(-\frac{||x_j^s - c_\ell||^2}{2\hat{\sigma}^2})$ for $\ell = 1, 2, ..., b$;
$\hat{\alpha} \leftarrow \max(0_b, (\tilde{H} + \hat{\lambda} I_b)^{-1}\tilde{h})$;
$\hat{w}(x) \leftarrow \sum_{\ell=1}^{b} \hat{\alpha}_\ell \exp(-\frac{||x - c_\ell||^2}{2\hat{\sigma}^2})$;
```

Algorithm 2: SAGE-TL (proposed, based on original SAGE algorithm(Covert et al., 2020))

```
Input: data $\{x^i, y^i\}_{i=1}^N$, model $f$, loss function $\ell$, outer samples $n$, inner samples $m$
Initialize $\hat{\phi}_1 = 0, \hat{\phi}_2 = 0, ..., \hat{\phi}_d = 0$
marginalPred = $\frac{1}{N}\sum_{i=1}^N f(x_i)$
for $i = 1$ to $n$ do
    Sample $(x, y)$ from $\{x^i, y^i\}_{i=1}^N$ using $w(x) = \frac{p_s(x)}{p_t(x)}$ obtained from Algorithm1
    Sample $\pi$, a permutation of $D$
    $S = \emptyset$
    lossPrev = $\ell(\text{marginalPred}, y)$
    for $j = 1$ to $d$ do
        $S = S \cup \{\pi[j]\}$
        $y = 0$
        for $k = 1$ to $m$ do
            Sample $x_{\bar{S}}^k \sim q(x_{\bar{S}}|X_S = x_S)$
            $y = y + f(x_S, x_{\bar{S}}^k)$
        end
        $\bar{y} = \frac{y}{m}$
        loss= $\ell(\bar{y}, y)$
        $\Delta$ =lossPrev $-$ loss
        $\hat{\phi}_{\pi[j]} = \hat{\phi}_{\pi[j]} + \Delta$
        lossPrev $=$ loss
    end
end
return $\frac{\hat{\phi}_1}{n}, \frac{\hat{\phi}_2}{n}, ..., \frac{\hat{\phi}_d}{n}$
```

A.4 SAGE-TL Rank

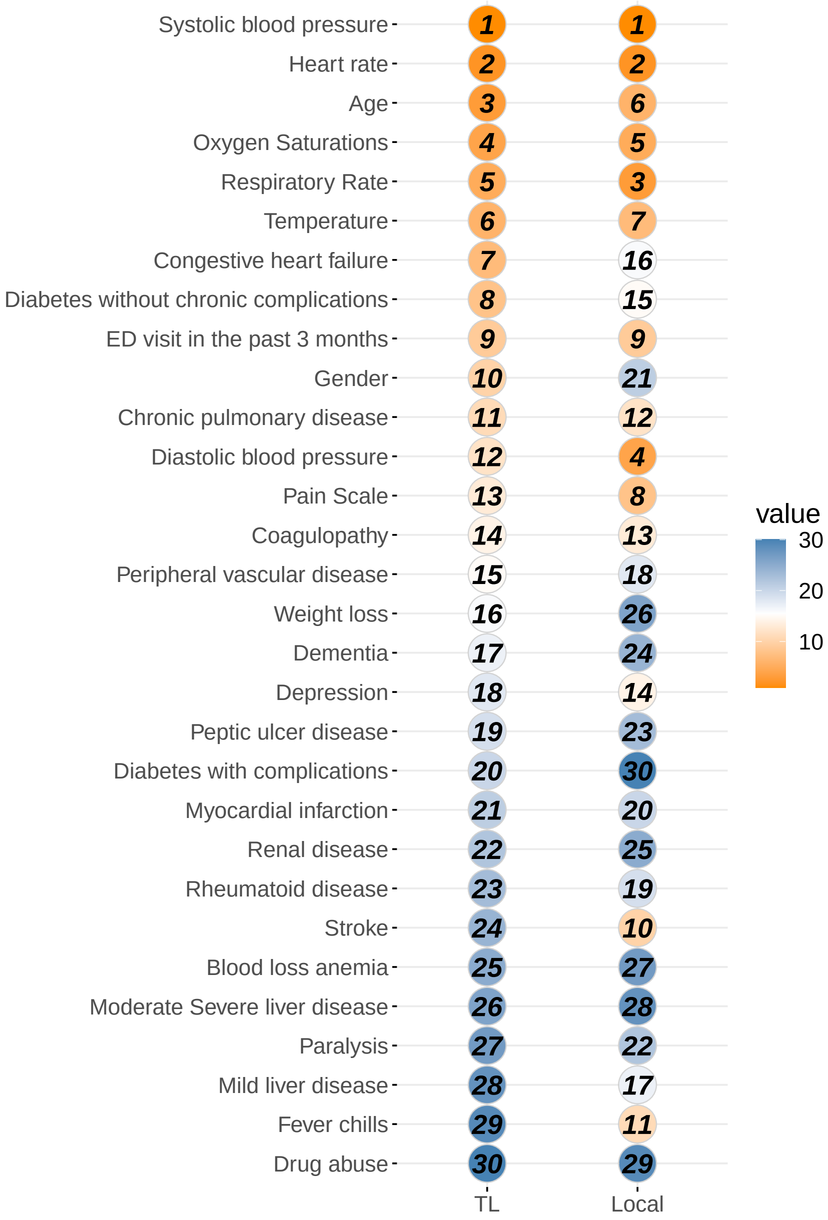

Figure 2: SAGE-TL Rank

A.5 Discussion

A.5.1 Potential Scalability

Although we have only used an EHR dataset for illustration, our SAGE-TL method can be applied to a wide range of clinical datasets, including but not limited to clinical trails and cohort studies of rare diseases, especially those that suffer from insufficient sample size due to study limits and seek to borrow information from pre-established external models. SAGE's model-agnostic nature ensures broad compatibility with different prediction models, from traditional statistical regressions to neural networks, enhancing the method's general applicability and adaptability to new and evolving modeling techniques. We plan to explore the application of SAGE-TL in future work with diverse clinical datasets and prediction tasks.

A.5.2 Clinical Scenarios of Transfer Learning Application

Transfer learning (TL) in clinical and biomedical research can occur across various domains in a range of formats, encompassing but not limited to cross-tissues (Li et al., 2022; 2023), cross-diseases (Alrefai et al., 2023; Turki et al., 2018), and cross-institutions (Holder et al., 2021; Wang et al., 2022).

Our proposed method specifically focuses on addressing distribution shifts occurring between different data sources. We achieve this by artificially partitioning data (based on age) from the same source into two groups with distinct distributions. This mirrors common real-world scenarios, such as hospitals with different patient populations—one with more senior patients and another with fewer. This approach also aligns with existing works (Hwang et al., 2021; Li et al., 2021), where transfer learning involves precisely the same type of clinical data from two sources but with heterogeneity in data distribution.

Furthermore, TL applications in healthcare can also occur within a single data source when researchers posit that the underlying data distribution may dynamically change over time (Zhang et al., 2022). In such cases, TL is applied within one site when researchers believe that the population has significantly shifted over time, using older models to update newer ones.