Carnegie Mellon University

RLSBench: Domain Adaptation Under Relaxed Label Shift

Saurabh Garg+

Nick Erickson*

James Sharpnack*

Alex Smola*

Siva Balakrishnan+

Zachary Lipton+

+Carnegie Mellon University

*AWS, Amazon AI

Carnegie Mellon University

Success of ML under IID setting

- Huge **success** in standard i.i.d. supervised machine learning, standard ML
- Inspired applications, e.g., in medical domain

nature

Letter | Published: 25 January 2017

Dermatologist-level classification of ski cancer with deep neural networks

Andre Esteva, Brett Kuprel, Ro

Swetter, Helen M. Blau & Sebastian

ARTICLES | DECEMBER 01 2021

Single-Examination Ris Prematurity

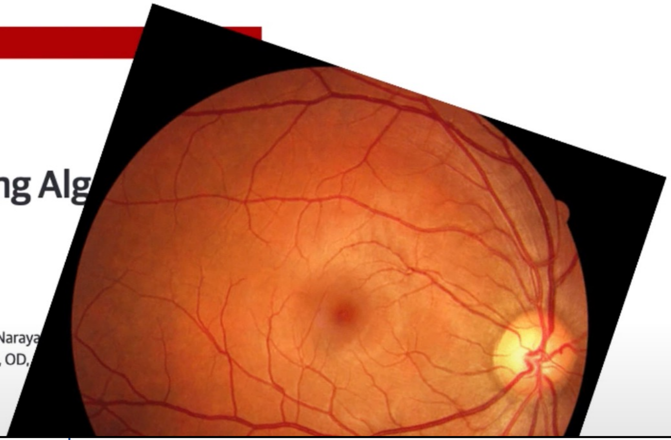
Aaron S. Coyner, PhD; Jimmy S. Chen, BS; Praveer Singh, PhD; Robert L. Schelonka, MD; Brian K. Jordan, MD, PhD; Cindy T. McEvoy, MD; Jamie E. Anderson, BS; R.V. Paul Chan, MD, MSc; Kemal Sonmez, PhD; Deniz Erdogmus, PhD; Michael F. Chiang, MD, MA; Jayashree Kalpathy-Cramer, PhD; J. Peter Campbell, MD, MPH on behalf of the Imaging and Informatics in Retinopathy of Prematurity Consortium

Research

JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Alg for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Coram, PhD; Martin C. Stumpe, PhD; Derek Wu, BS; Arunachalam Naraya Subhashini Venugopalan, MS; Kasumi Widner, MS; Tom Madams, MEng; Jorge Cuadros, OD, PhD; Ramasamy Kim, OD, Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessica L. Mega, MD, MPH; Dale R. Webster, PhD



ML is not Robust under Distribution shift

- Huge **success** in standard i.i.d. supervised machine learning, standard ML
- Inspired applications, e.g., in medical domain
- However, standard ML **breaks** under **distribution shift**

ML is not Robust under Distribution shift

Analysis | Open Access | Published: 15 March 2021

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans

Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, AIX-COVNET, James H. F. Rudd, Evis Sala & Carola-Bibiane Schönlieb

Nature Machine Intelligence **3**, 199–217(2021) | Cite this article

“Our review finds that none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases.”

Failure in medical diagnosis under

Source: Jayashree’s talk at PODS ICML Workshop 2022

ML is not Robust under Distribution shift

- Despite huge **success** in standard i.i.d. supervised machine learning, standard ML **breaks** under **distribution shift**
- Lack of **rigor** in benchmark driver empirical research

Different Learning Scenarios

- Distribution shift problems can appear in different scenarios, e.g.,
  - domain generalization
  - transfer learning with (small amount of) labeled target data
  - domain adaptation
- These settings differ in "what data is available during training"
- In this work, we focus on **domain adaptation problems**

Domain Adaptation

- Problem setup

Data (Observed)

Source Distribution
$P_s(x, y)$

Distribution shift

Target Distribution
$P_t(x, y)$

x_n, y_n

x_1, y_1

x_m

x_1

Labeled data $P_s(x, y)$

Unlabeled data $P_t(x)$

Objective

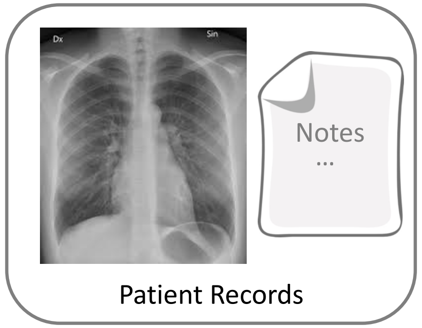
Predict well on target (test) distribution **WITHOUT** seeing any labels from target

Shifts due to Changing Class Prevalence

- In this talk, we will consider distribution shift problems due to changing class prevalence

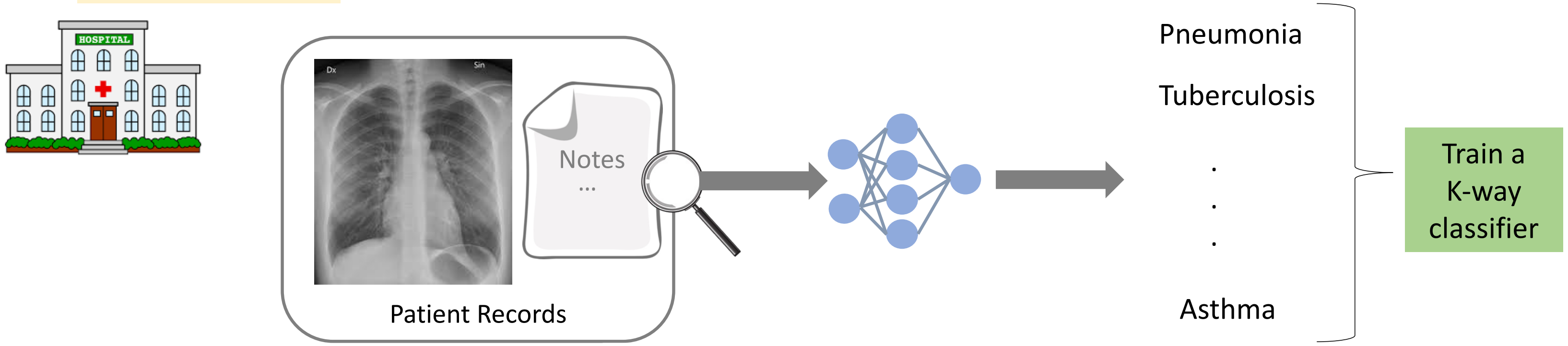
Archetype Example: Disease Diagnosis

Hospital 1

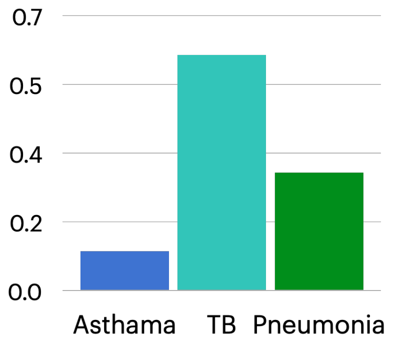


Archetype Example: Disease Diagnosis

Hospital 1



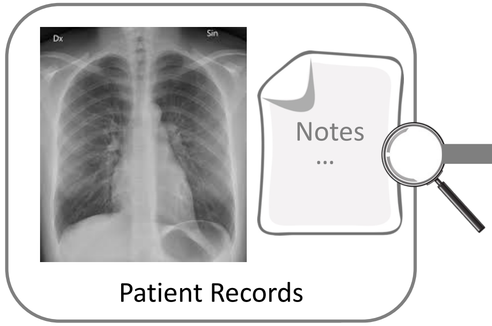
Labeled Source Data ~ $\{P_s(x, y)\}$^n

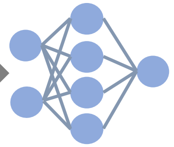


Carnegie Mellon University

Archetype Example: Disease Diagnosis

Hospital 1





Pneumonia

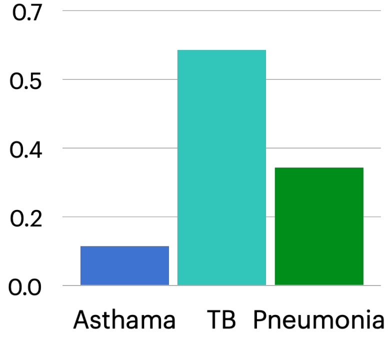
Tuberculosis

.
.
.

Asthma

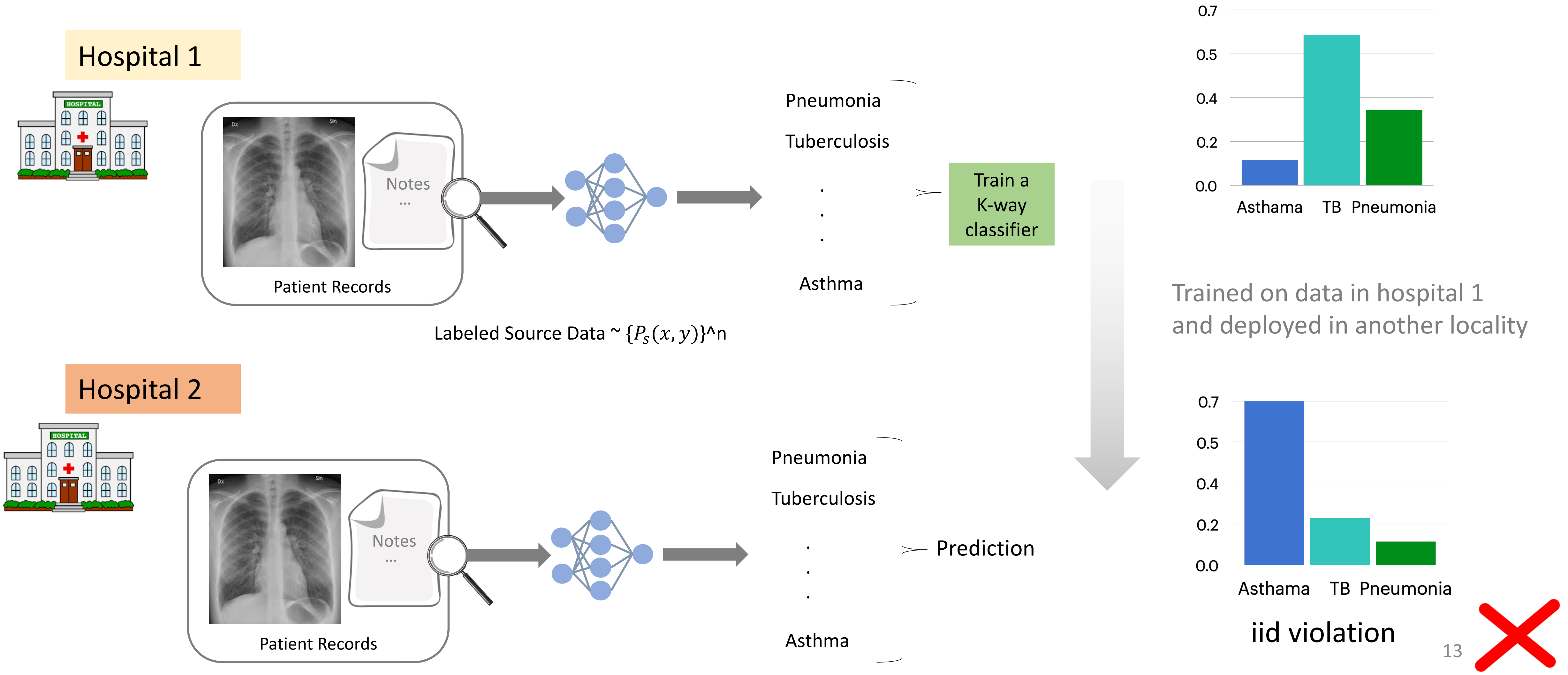
Train a K-way classifier

Labeled Source Data ~ $\{P_S(x,y)\}$^n



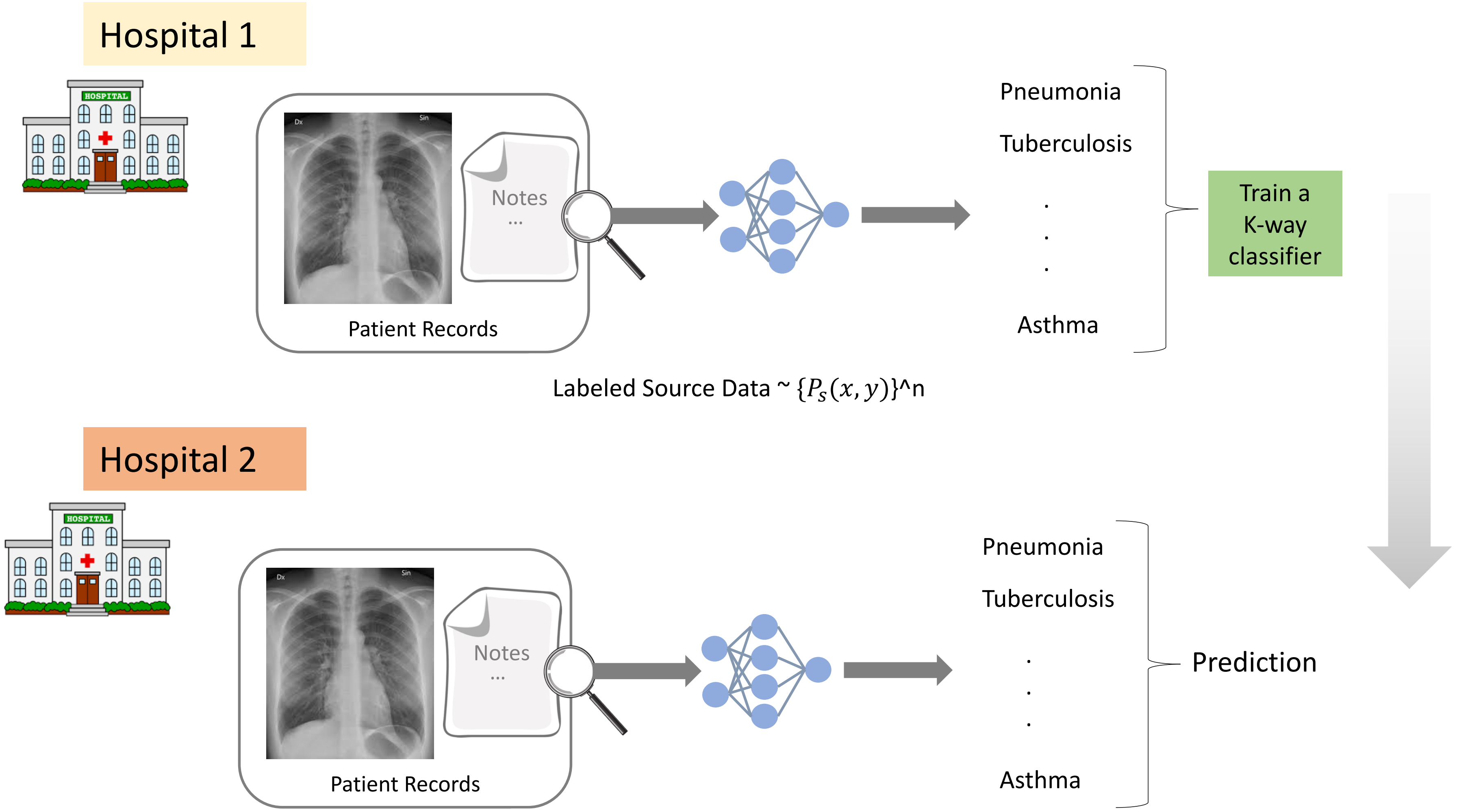
Trained on data in hospital 1 and deployed in another locality

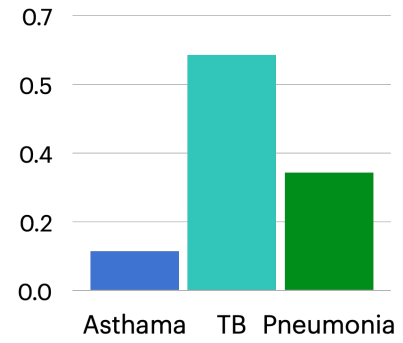
Archetype Example: Disease Diagnosis



Carnegie Mellon University

Archetype Example: Disease Diagnosis





Trained on data in hospital 1 and deployed in another locality

Can we trust models' prediction ??

iid violation

Carnegie Mellon University

Label Shift Setting

- Assume $p(y)$ can change but **the conditional $\boldsymbol{p(x|y)}$ doesn't change**
- Under this assumption, we obtain a well-posed setting
- **Goal:** (i) Estimate the target label marginal $p_t(y)$; and (ii) adapt source classifier f to target data

- Why its non-trivial? Recall that we **do not** observe target labels

Estimation and Correction under Label Shift

- Effective methods that are applicable in deep learning regimes exists
- Yield consistent estimates of the target label marginal [Garg et al. 2020, Lipton et al. 2018, Azizzadenesheli et al., 2019]
  - $O\left({}^{1}/_{\sqrt{n}}\right)$ convergence rates with interpretable error bounds
- Principled ways to on-the-fly update the source classifier
  - Importance re-weighted correction

$$[f_t(x)]_y = \frac{w(y)[f_s(x)]_y}{\sum_j w(j)[f_s(x)]_j}$$

SG, YW, SB, ZL. A Unified View of Label Shift Estimation **(NeurIPS 2020)**

Extending the Label Shift Setting

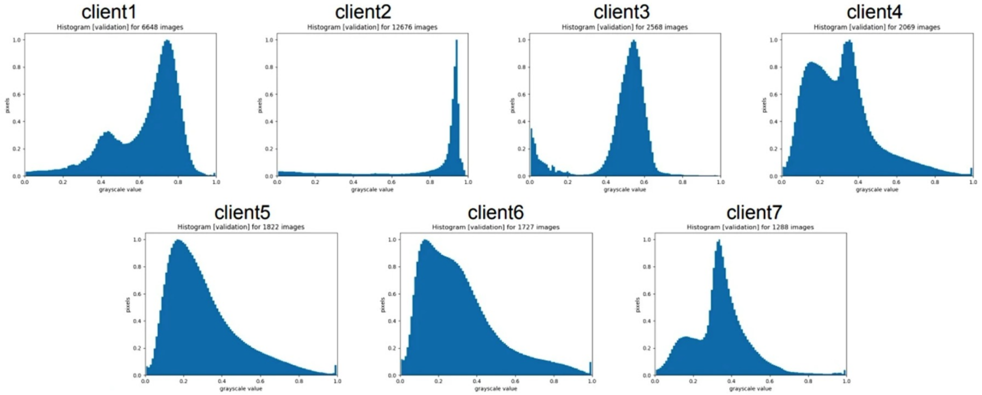
- Two key assumptions in label shift: (i) class overlap in source and target; (ii) p(x|y) remains invariant
- However, these label shift assumptions can be violated in practice
- Our past work on PU learning and Open Set Label Shift (OSLS) relaxes the class overlap assumption [Garg et al. 2021, Garg et al. 2022]
- In this work, we take a step in **relaxing the latter assumption** (i.e., p(x|y) remains invariant)

[1] SG, YW, AS, SB, ZL. Mixture Proportion Estimation and PU Learning: A Modern Approach **(NeurIPS 2021)**

[2] SG, SB, ZL. Domain Adaptation under Open Set Label Shift **(NeurIPS 2022)**

Motivation: Relaxed Label Shift

- In medical domain, along with changing prevalence of diseases, p(x|y) can drift from location A to location B.

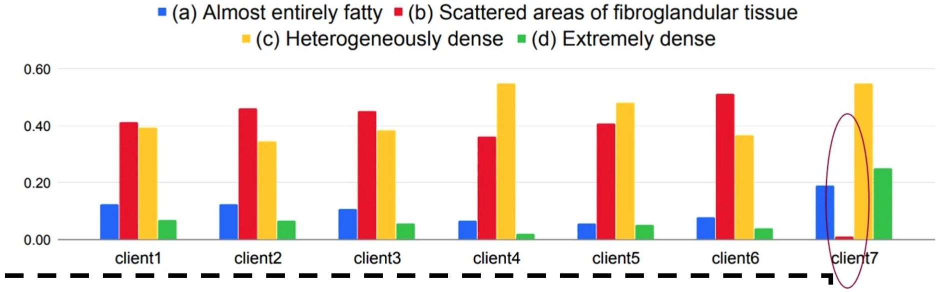


Intensity histograms show substantial differences between sites

Hospital 1

Trained on data in hospital 1
and deployed in another locality

Hospital 2

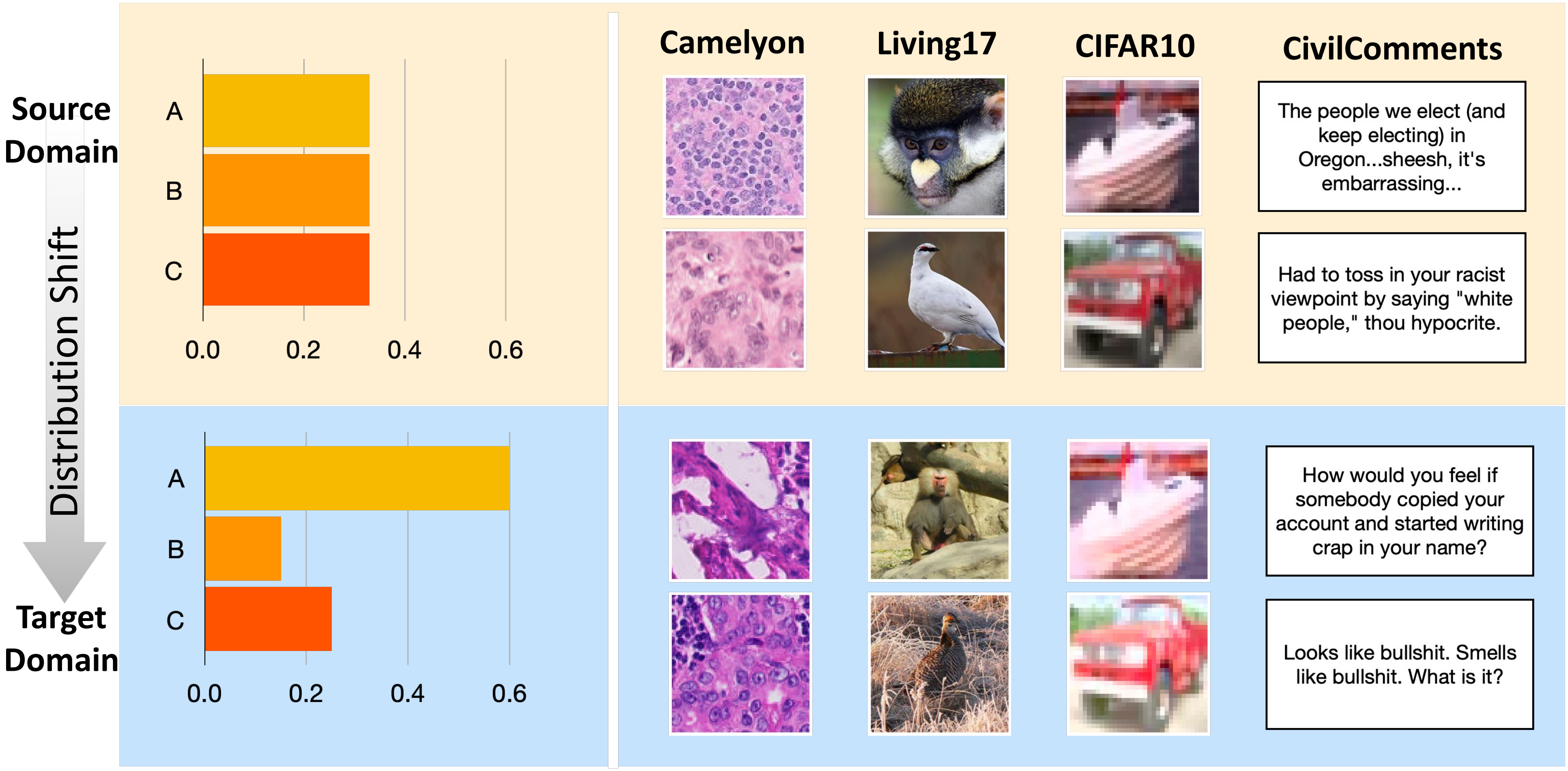


Distribution of classes is different among sites

Source: Jayashree's talk at PODS ICML Workshop 2022

Relaxed Label Shift

$p(y)$ can shift arbitrarily and $p(x|y)$ can shift in seemingly natural ways



Relaxed Label Shift

- Assume that the label distribution can shift from source to target arbitrarily
- **But** that $p(x|y)$ varies between source and target in some *comparatively restrictive way, i.e.,*

$$\max_y D\big(p_s(x|y), p_t(x|y)\big) < \epsilon$$

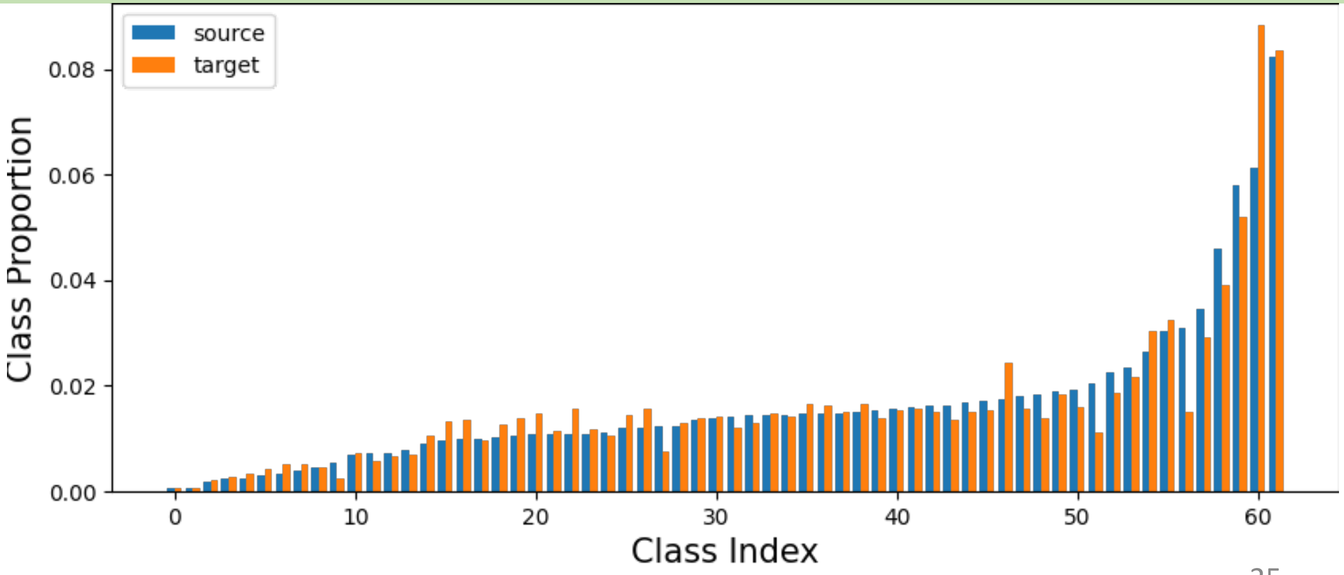
- Lack of rigorous characterization of the sense in which those shifts arise in the wild
- Our work focuses on empirical evaluation with real-world datasets
- **Goal:** (i) Estimate the target label marginal $p_t(y)$; and (ii) adapt source classifier f to target data

Carnegie Mellon University

Issues with Prior Work

- Motivated by the kinds of problems arise in practice, several benchmarks exist (e.g., OfficeHome, DomainNet, WILDS)
- However, most academic benchmarks exhibit little or no shift in the label distribution
- Consequently, benchmark driven research produced heuristics that implicitly assume no shift in class proportions

Eg: Default shift in target label marginal in FMoW-WILDS is small



Issues with Prior Work

- Several works aim to tackle relaxed label shift settings [Tan et al., 2020; Tachet des Combes et al., 2020; Prabhu et al., 2021]
- However, it is difficult to assess the state of the field owing to inconsistencies among relevant papers
    I. Evaluation criteria (e.g., per-class average performance instead of target accuracy)
    II. Datasets (e.g., different datasets in different papers)
    III. Baselines (e.g., missing simple and important baselines)
    IV. Model Selection criteria (e.g., peeking at target validation performance)

- **Overall**, fair and realistic comparison is missing

RLSBENCH: Relaxed Label Shift Benchmark

- We introduce RLSBENCH a large-scale benchmark for relaxed label shift
- Consists of **>500 distribution shift** pairs with varying severity of shift in target class proportions across 14 multi-domain datasets
- We evaluate a collection of **12 popular DA methods** based on domain invariant representation learning, self-training, and test-time adaptation
- **Overall,** we train >30k models in our testbed

Carnegie Mellon University

Datasets

Vision

CIFAR10
CIFAR100
Camelyon
FMoW
Entity13
Entity30
Living17
Nonliving26
Officehome
Visda
Domainnet

NLP

CivilComments

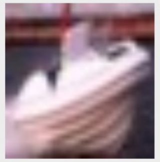
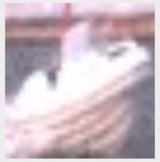
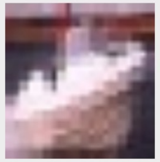
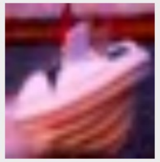
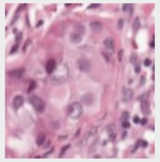
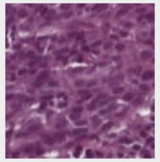
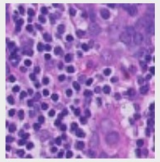
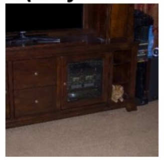
Tabular

Retiring Adults
Mimic readmissions

Across these 12 datasets we obtain **56 source and target pairs** with minor to no shift in class prevalence

Datasets

We show 5 (out of 14) multi-domain dataset in the table next

Dataset	Domains				
CIFAR10	Cifar10v1 	Cifar10v2 	Cifar10C-Frost 	Cifar10C-Pixelate 	Cifar10C-Saturate 
FMoW	Years 2002-’13 	Year 2013-’16 	Year 2016-’18 		
Camelyon	Hospital 1-3 	Hospital 4 	Hospital 5 		
Entity13	v1 	v1 (disjoint sub.) 	v2 	v2 (disjoin sub.) 	
Visda	Rendering 	Real -1 	Real - 2 		

Simulating a Shift in Target Marginal

- We simulate shift by **altering target label marginal**, keeping source fixed
- Sample the target label marginal from a *Dirichlet distribution* with a parameter $\alpha \in \{0.5, 1.0, 3.0, 10.0, \infty\}$ multiplier to the original target marginal
- The Dirichlet parameter α controls the severity of shift
- Intuitively, as α decreases, the severity in shift increases
- After simulating shift, we obtain **560 pairs** of different source and target datasets

Simulating a Shift in Target Marginal

CIFAR10

Cifar10v1

Shift in $p(x|y)$

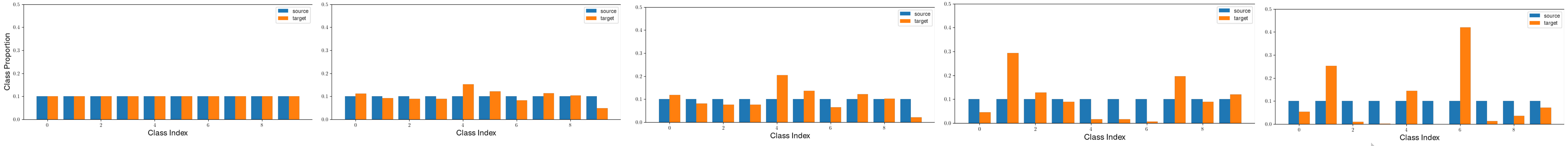
Cifar10v2

Cifar10C-Frost

Cifar10C-Pixelate

Cifar10C-Saturate

$\times$



Increasing shift in $p_t(y)$

Domain Adaptation Methods

- Source only Model
  - with augmentations
  - with adversarial training
- Domain Alignment Methods
  - DANN, CDANN
  - IWDAN, IWCDAN
- Self-training methods
  - FixMatch
  - NoisyStudent
  - SENTRY
- Test-time training methods
  - CORAL/DARE
  - BN-adapt
  - TENT

Carnegie Mellon University

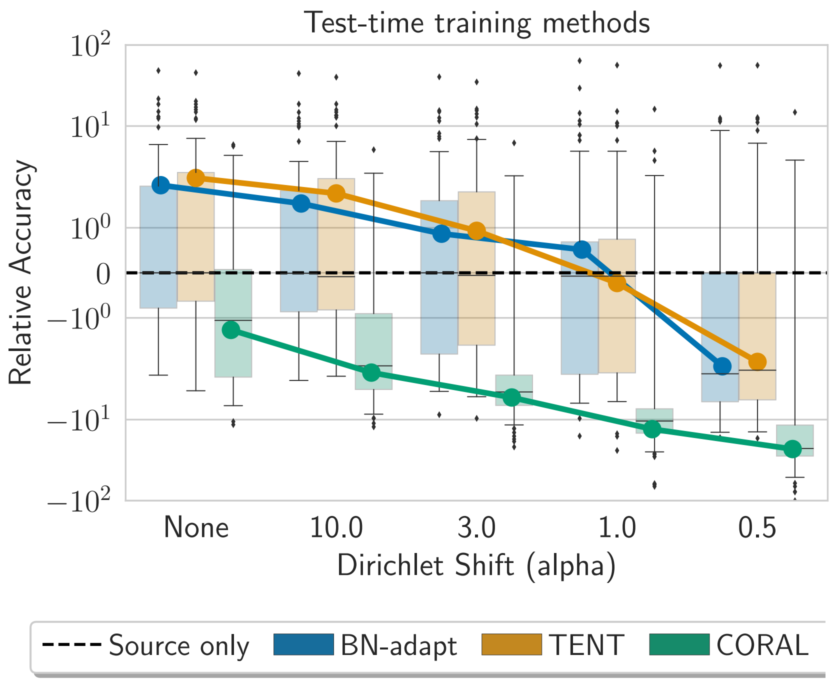
Other Choices for Fair and Realistic Evaluation

- Re-implemented all methods with consistent design choices
- **Model selection criteria and hyperparameter choices**
  - Source hold-out performance
  - DA method specific hyperparameters fixed across datasets incorporating the suggestions made in corresponding papers
- **Architectural and pretraining details**
  - Different architectures (e.g., DenseNet121, Resenet18, Resnet50, DistillBert)
  - Bert pretraining, Imagenet-pretraining and randomly initialized models
- **Data Augmentation**
  - Strong augmentation technique with RandAug on vision datasets

Carnegie Mellon University

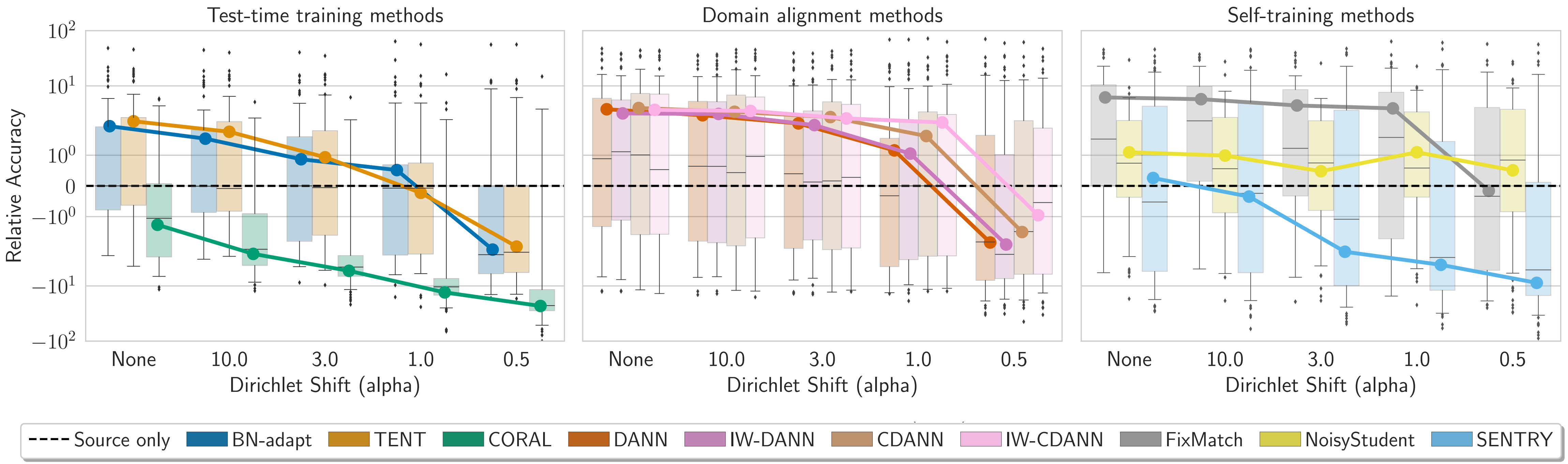
Empirical Results

- **Takeaway-1:** Popular deep DA methods falter under severe shifts in target label proportions



Empirical Results

- **Takeaway-1:** Popular deep DA methods falter under severe shifts in target label proportions



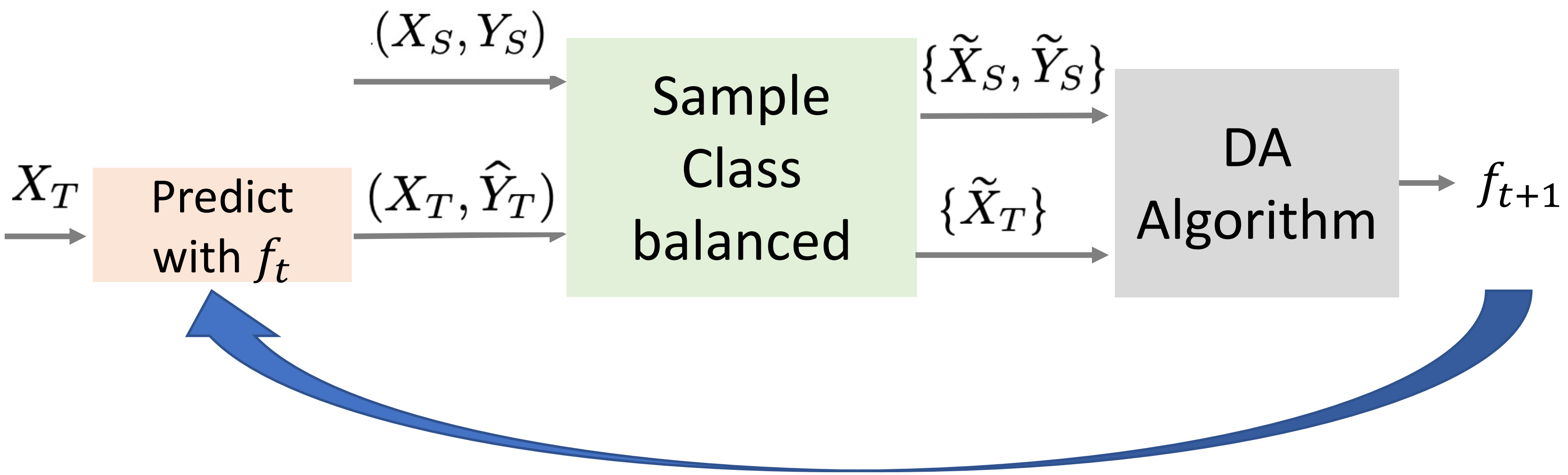
Meta Algorithm to Handle Class Proportion Shift

- Performance of DA methods can falter
- Known failure of domain adversarial training methods [Wu et al., 2019; Zhao et al., 2019]
- We show that this failure is not limited to domain adversarial training methods, **but is common with all the popular DA methods**

Carnegie Mellon University

Meta Algorithm to Handle Class Proportion Shift

- We implement two simple general-purpose corrections
- **Re-sampling**
  - Balanced source data
  - Leverage pseudolabels for target data to perform pseudo class-balanced re-sampling



Carnegie Mellon University

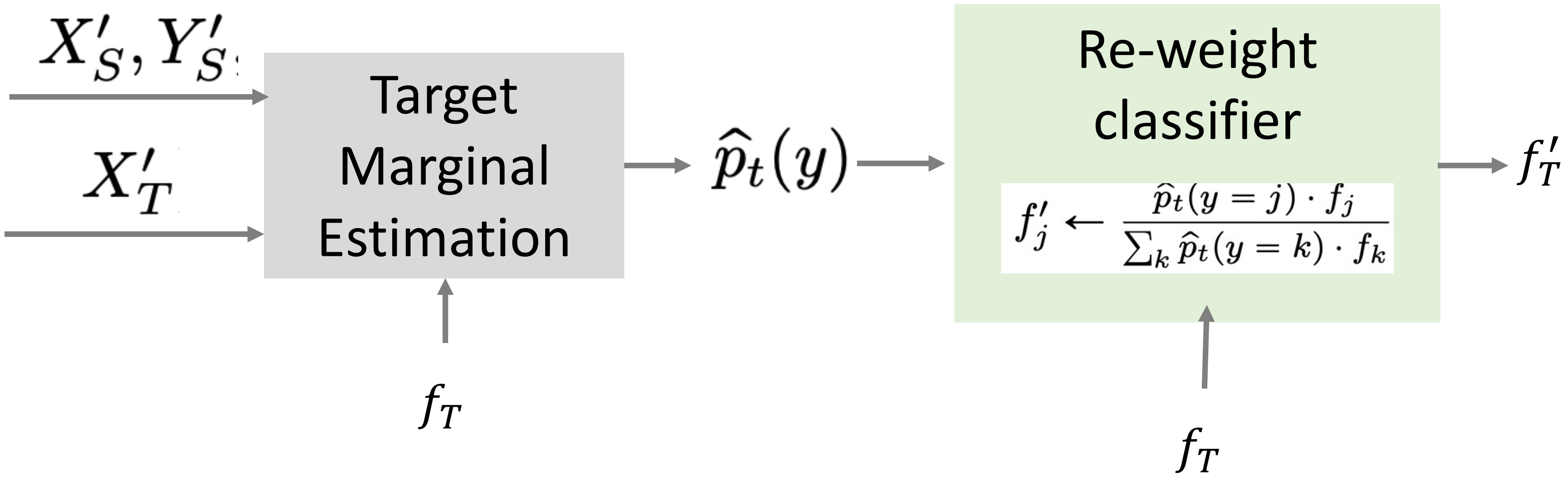
Meta Algorithm to Handle Class Proportion Shift

- We implement two simple general-purpose corrections
- **Re-sampling**
  - Balanced source data
  - Leverage pseudolabels for target data to perform pseudo class-balanced re-sampling
- With re-sampling, we can hope to train f on a mixture of balanced source and balanced target datasets in an ideal case
- Still leaves open the problem of adapting f to the original target label distribution (which is not available)

Carnegie Mellon University

Meta Algorithm to Handle Class Proportion Shift

- **Re-weighting**
  - Estimate target label marginal with label shift estimation methods (e.g. BBSE, MLLS)
  - Use on-the-fly re-weighting of the classifier



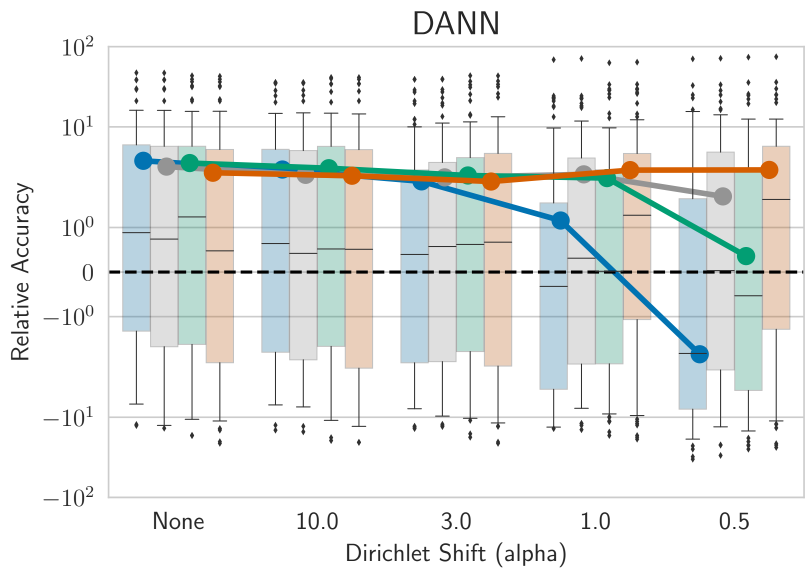
Meta Algorithm to Handle Class Proportion Shift

- **Re-weighting**
  - Estimate target label marginal with label shift estimation methods (e.g. BBSE, MLLS)
  - Use on-the-fly re-weighting of the classifier
- Different DA methods give different plugin f
- Relaxed label shift scenario violates the conditions required for consistency of label shift estimation techniques
- Nonetheless employ these techniques and empirically evaluate efficacy of these methods in our testbed

Carnegie Mellon University

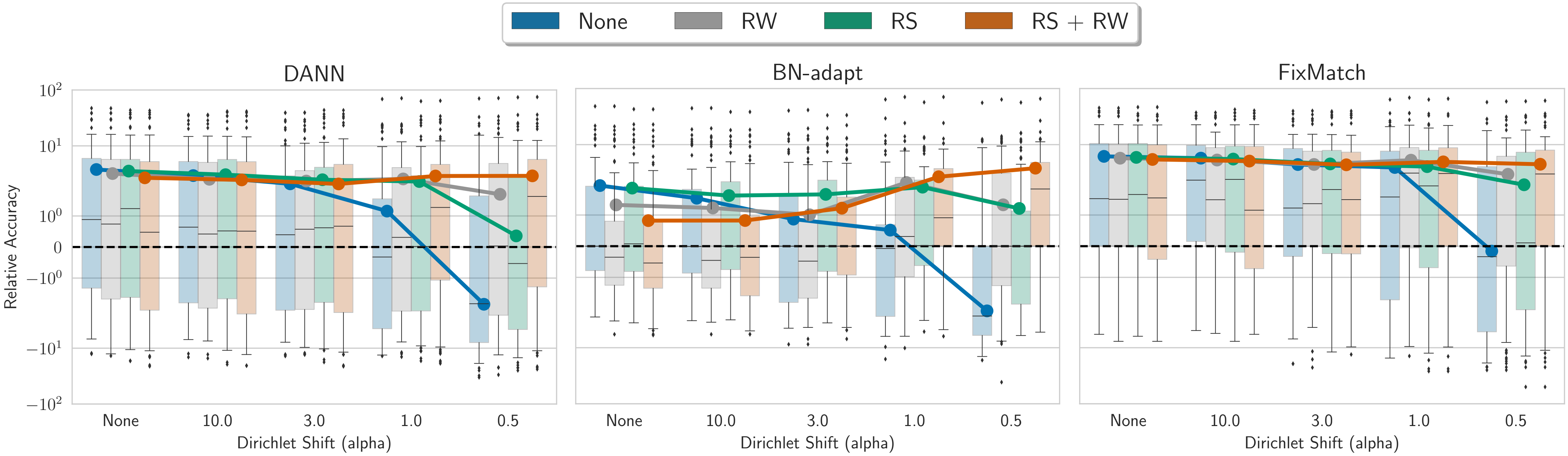
Takeaways

- **Takeaway-2:** Re-sampling to pseudo balance target often helps all DA methods
- **Takeaway-3:** Benefits of post-hoc re-weighting of the classifier depends on shift severity and the underlying DA algorithm.



Takeaways

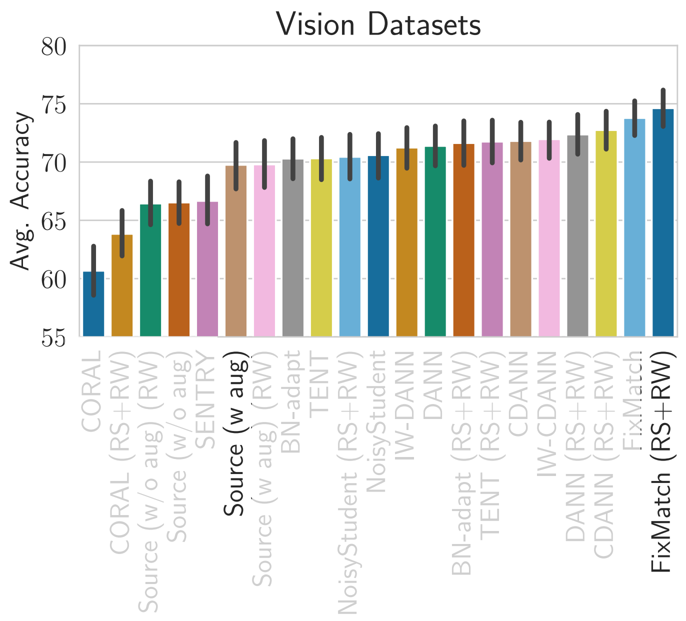
- **Takeaway-2:** Re-sampling to pseudo balance target often helps all DA methods
- **Takeaway-3:** Benefits of post-hoc re-weighting of the classifier depends on shift severity and the underlying DA algorithm.

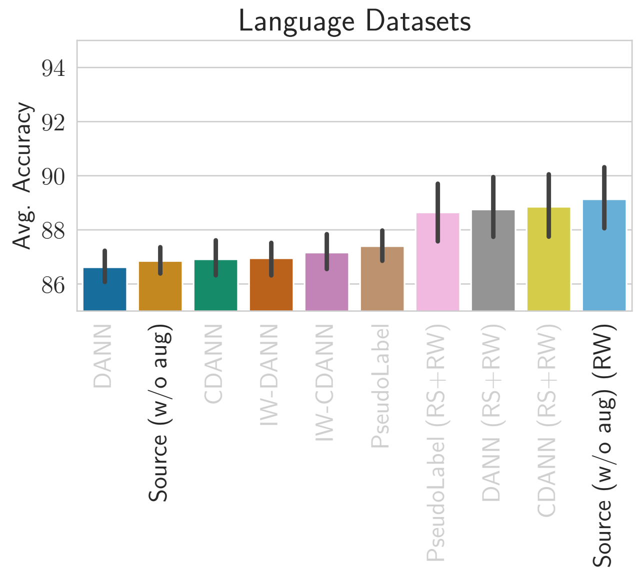


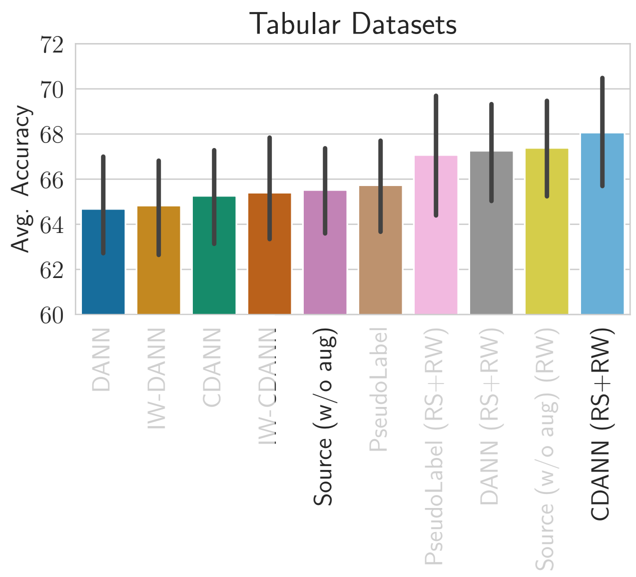
Carnegie Mellon University

Takeaways

- **Takeaway-4:** DA methods paired with our meta-algorithm often improve over source-only classifier but no one method consistently performs the best

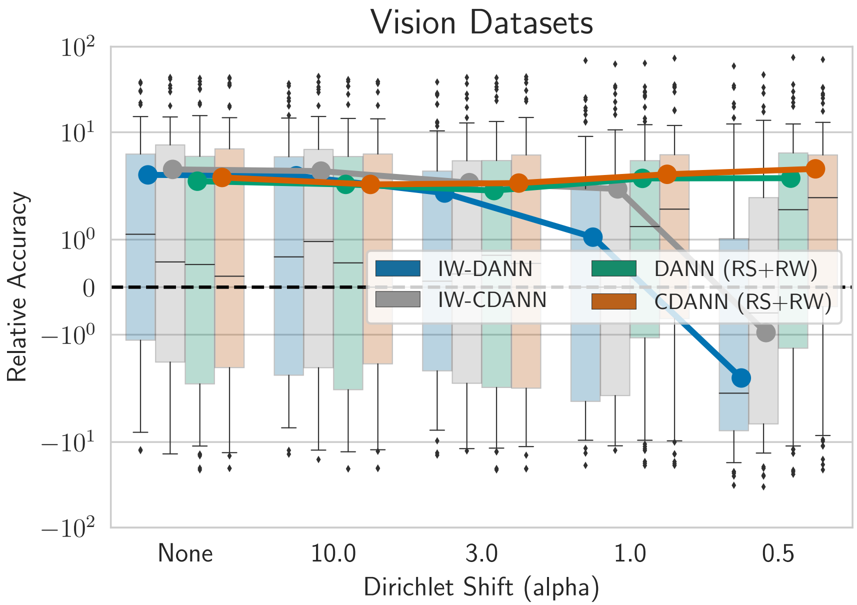


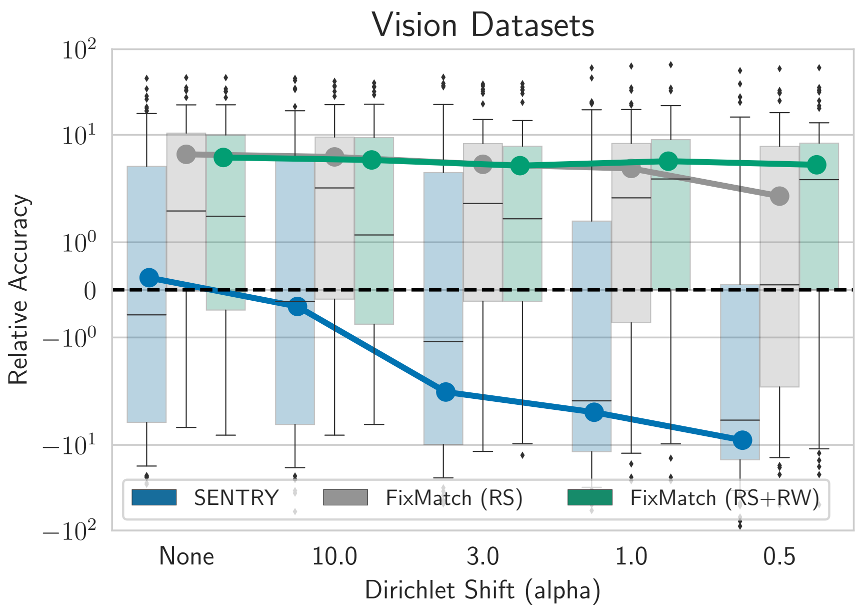




Takeaways

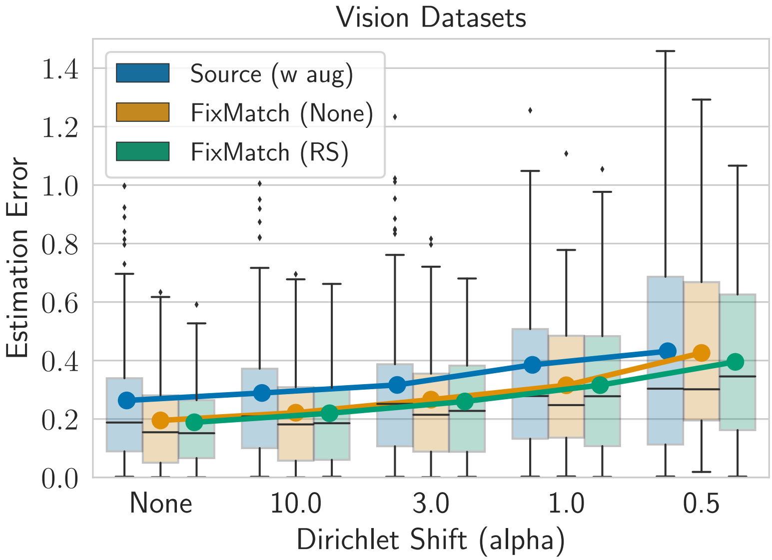
- **Takeaway-5:** Existing DA methods when paired with our meta-algorithm significantly outperform other DA methods specifically proposed for relaxed label shift.

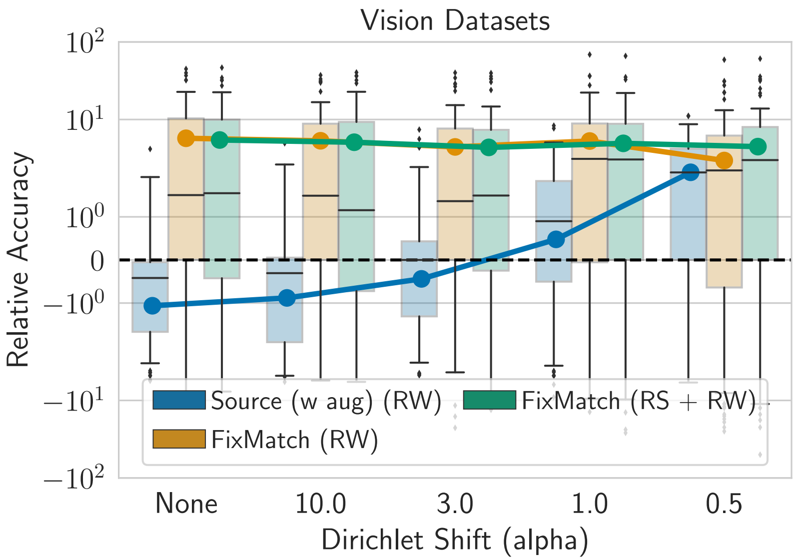




Takeaways

- **Takeaway-6:** Deep DA heuristics often improve target label marginal estimation on tabular and vision modalities.

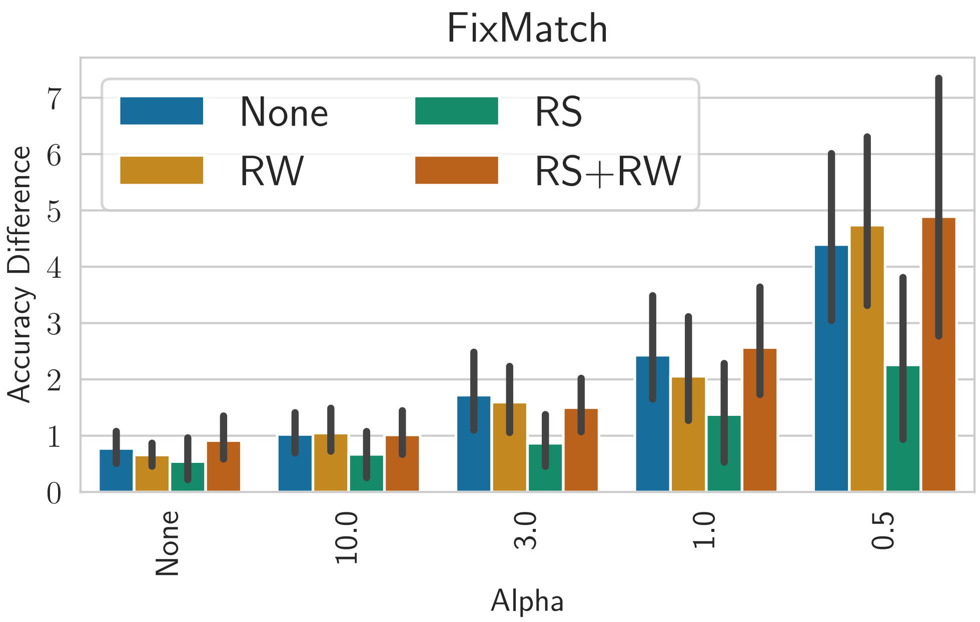


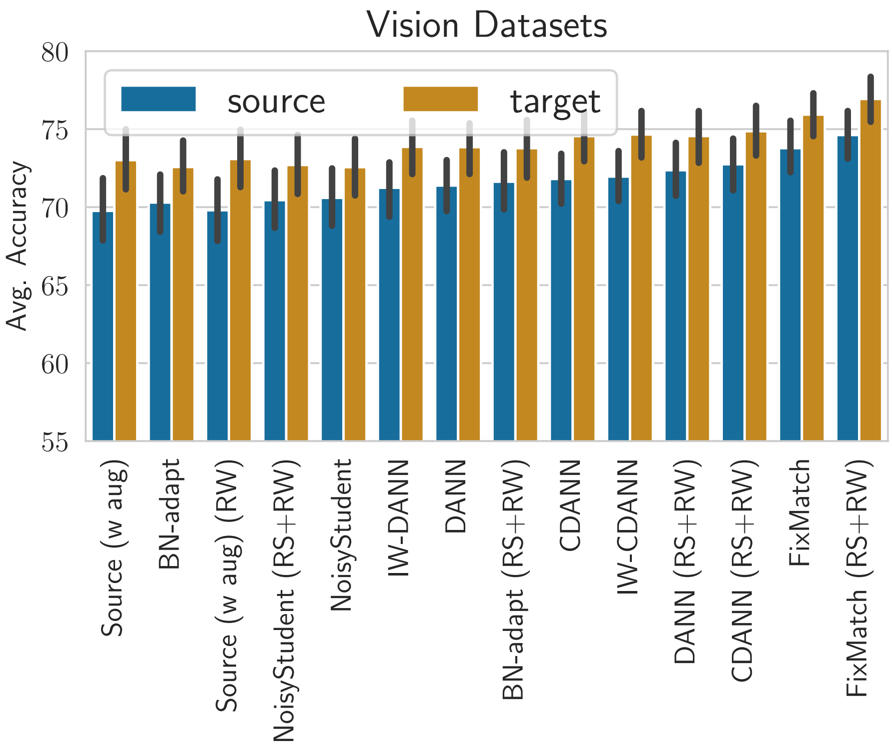


Carnegie Mellon University

Takeaways

- **Takeaway-7:** With increasing severity of label distribution shift, the accuracy difference with source and target early stopping criterion increases





Carnegie Mellon University

Theoretical Result

- We can show that label shift estimates degrade gracefully with shifts in $p(x|y)$

Theoretical Result

Theorem (Estimation under RLS)

In population, the BBSE estimates of importance degrades as

$$||\widehat{w} - w^*||_2 \leq \sqrt{k}\, ||w||_2 \kappa \boxed{TV\left(p_s(f(x)|y), p_t(f(x)|y)\right),}$$

where κ is the condition number of the confusion matrix C_f.

Concluding Remarks

- RLSBENCH provides **sensitivity analysis** to measure the robustnessof label shift methods under various sorts of perturbations
- Relative to the benchmark driven DA literature, RLSBENCH provides **comprehensive and standardized suite**
- One step closer to exhibiting the sort of diversity that we should expect to encounter when deploying models in the wild
- Caution: While promising, given underspecified nature of the problem, benchmark results should be taken with grain of salt

Carnegie Mellon University

Future Work

- Incorporate self-supervised methods that learn representations by training on a union of unlabeled data from source and target [Gidaris et al., 2018; He et al., 2022, Caron et al., 2020; Chen et al., 2020]

- Characterizing the behavior of label shift estimation techniques when the label shift assumption is violated

Thanks!

Questions?

- Paper: https://arxiv.org/abs/2302.03020
- Code: https://github.com/acmi-lab/RLSbench/
- Website: https://sites.google.com/view/rlsbench/

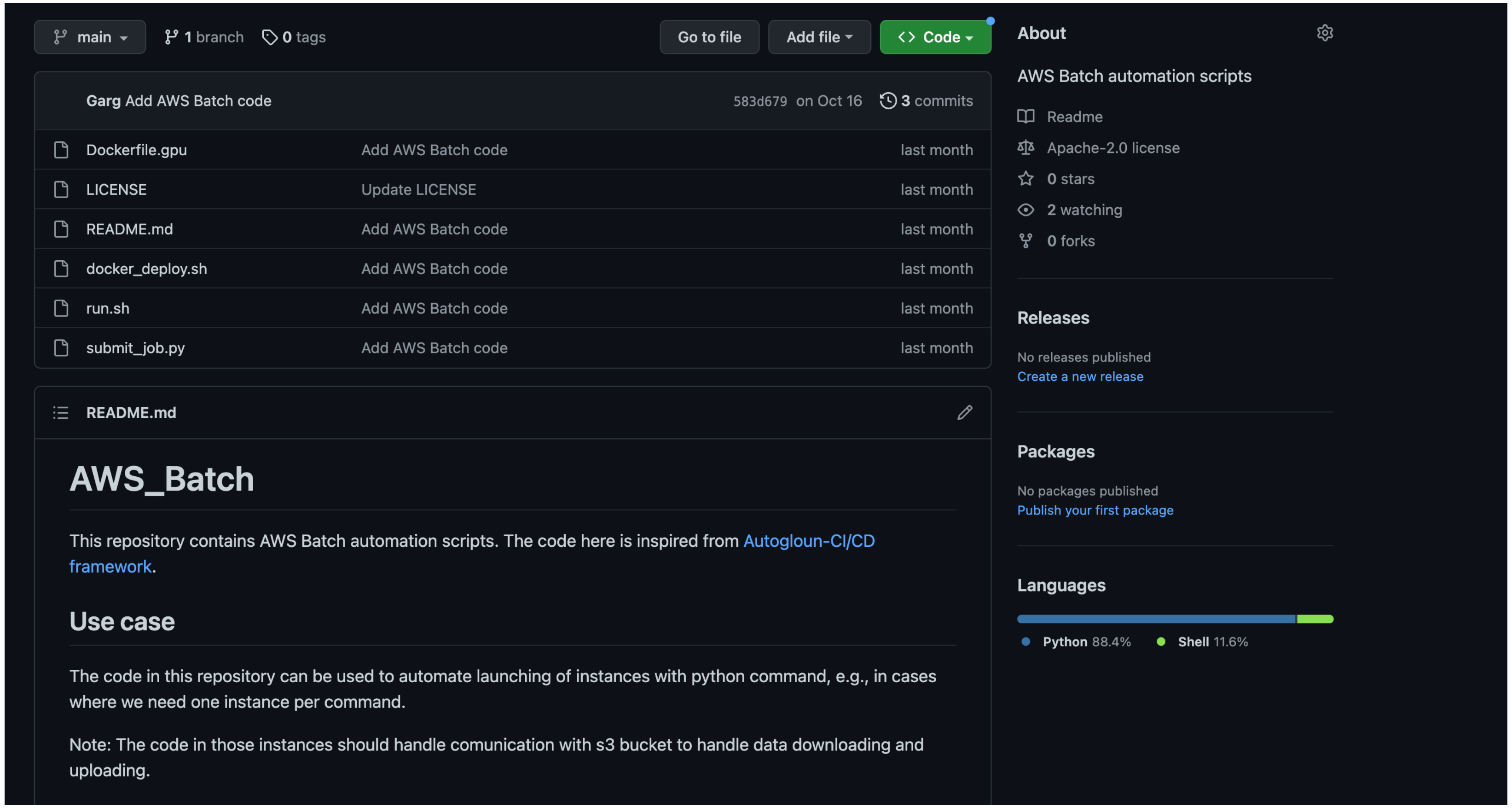
@saurabh_garg67

sgarg2@andrew.cmu.edu

http://saurabhgarg1996.github.io/

AWS Batch

- Launch experiments at scale with simple AWS Batch setup



AWS Batch

- Launch experiments at scale with simple AWS Batch setup
- At a high level, we would need to:
  (i) create a docker image with all the code and setup that we can use to launch our ec2 instances;
  (ii) configure AWS batch setup that can launch EC2 instances with the docker image;
  (iii) local scripts that will trigger, monitor and terminate the launch.