RLSBENCH: Domain Adaptation Under Relaxed Label Shift

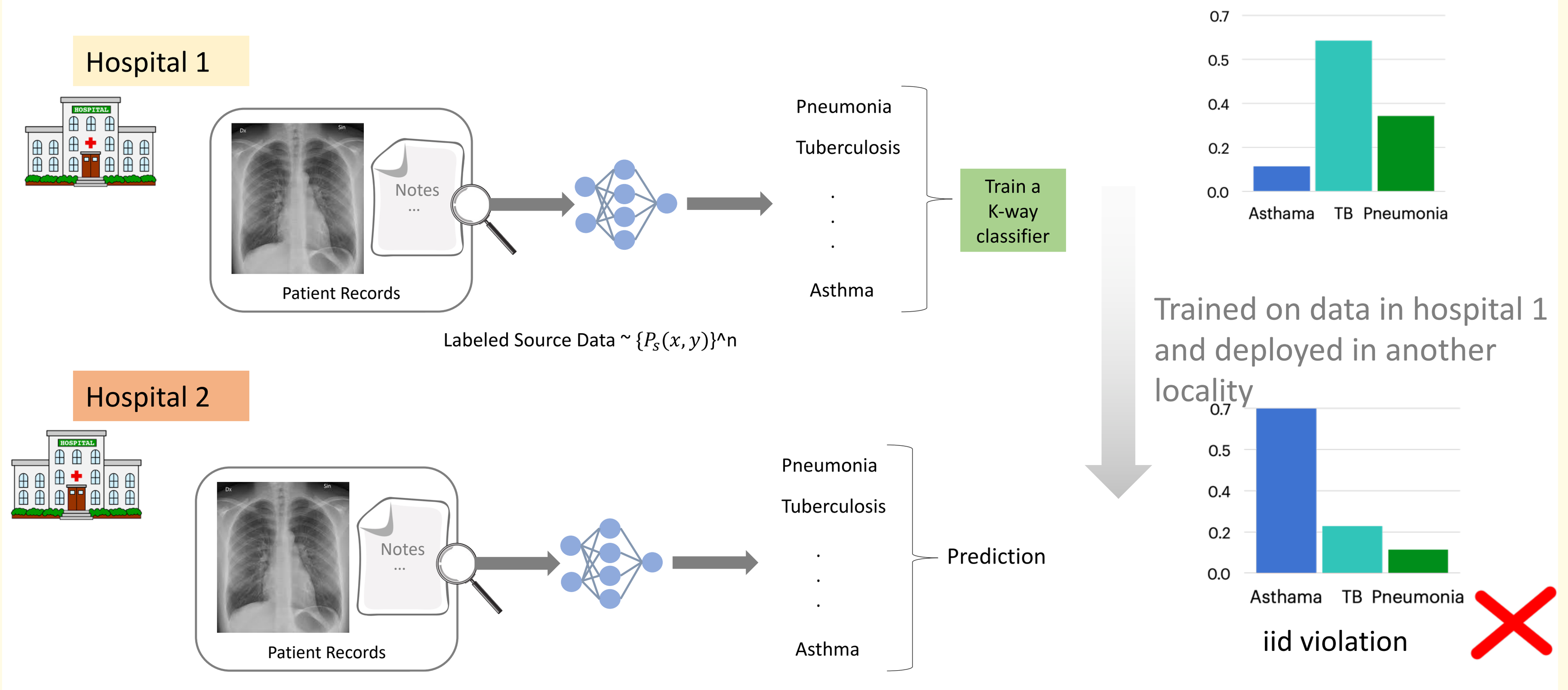
Carnegie Mellon University

Saurabh Garg, Nick Erickson, James Sharpnack, Alex Smola, Sivaraman Balakrishnan, Zachary C. Lipton

amazon

ML is not Robust Under Distribution Shift

- Despite huge **success** in standard i.i.d. supervised machine learning, standard ML practices **break** under **distribution shift**



Relaxed Label Shift

- Two key assumptions in label shift: (i) class overlap in source and target; (ii) p(x|y) remains invariant
- However, label shift assumption can be violated in practice
- **Relaxed Label Shift**: label distribution can shift arbitrarily but that $p(x|y)$ varies between source and target in some *comparatively restrictive way* (e.g., shifts arising naturally in the real-world), i.e.,

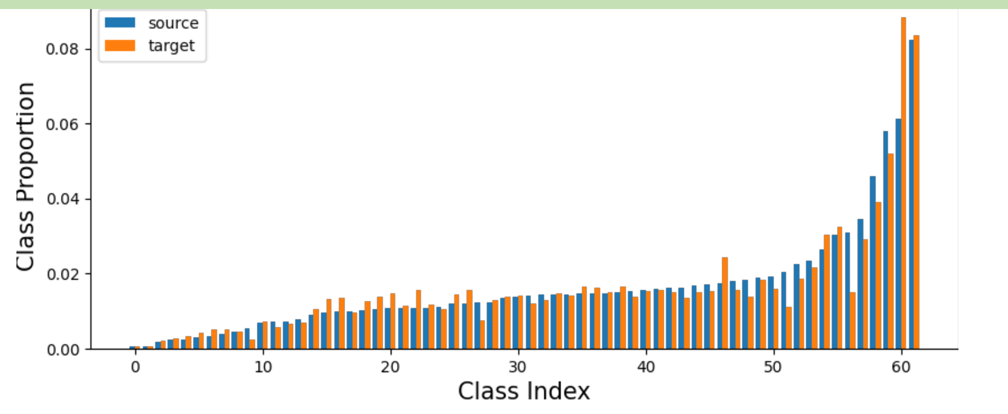
$$\max_y D\big(p_s(x|y), p_t(x|y)\big) < \epsilon$$

- 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

Issues with Prior Work

- 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

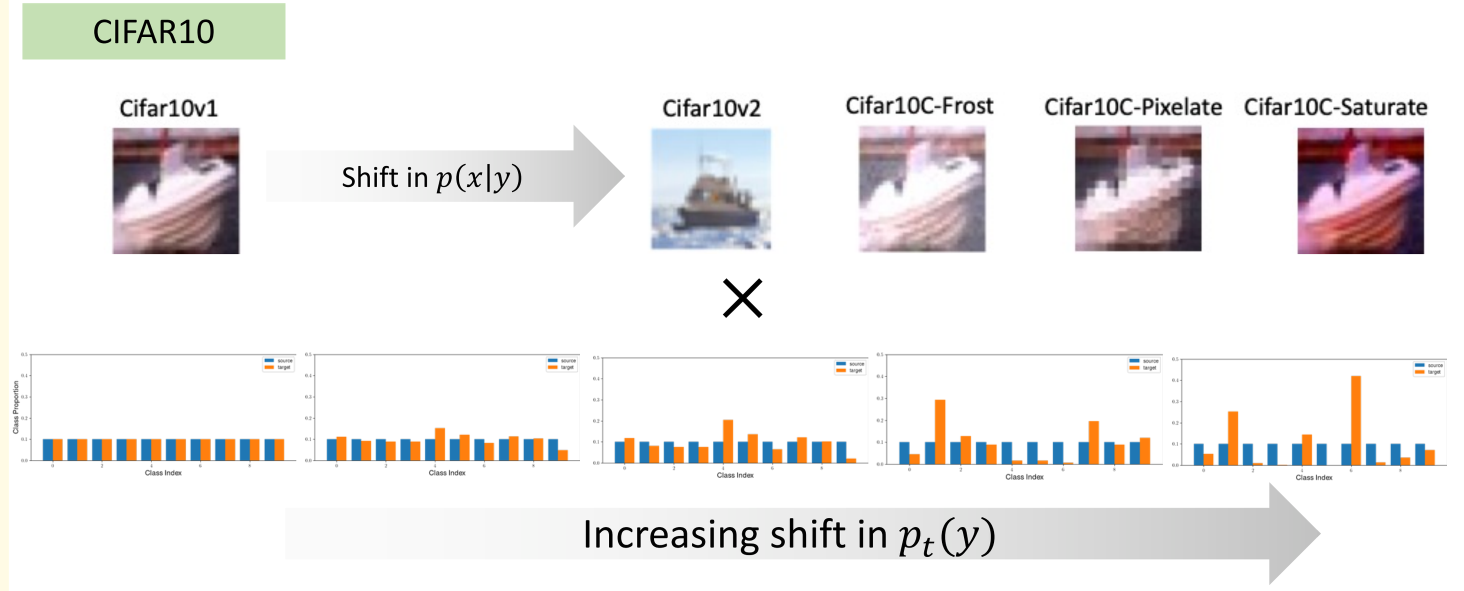


Pitfalls of Current Evaluation Practices

- 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 acc.)
  II. Datasets (e.g., different datasets in different papers)
  III. Baselines (e.g., missing simple label shift correction baselines)
  IV. Model Selection criteria (e.g., peeking at target validation performance)
- Overall, fair and realistic comparison is missing.

RLSbench: Relaxed Label Shift Benchmark

- Consists of **>500 distribution shift** pairs with varying severity of shift in target class proportions across 14 multi-domain datasets



Vision

CIFAR10	Nonliving26
CIFAR100	Officehome
Camelyon	Visda
FMoW	Domainnet
Entity13	
Entity30	
Living17	

NLP

CivilComments

Tabular

Retiring Adults

Mimic readmissions

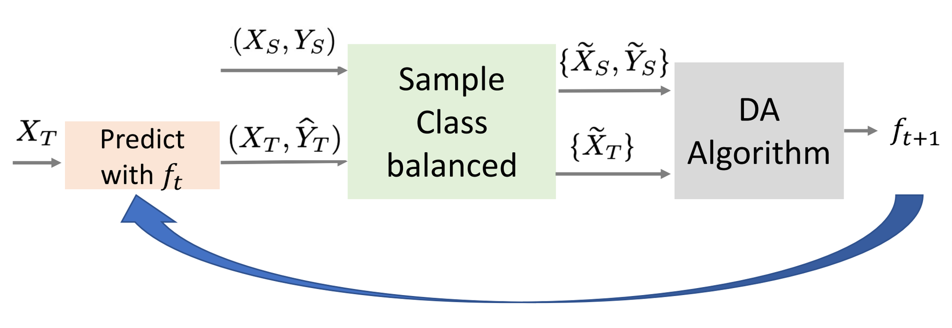
- We evaluate a collection of **12 popular DA methods**
  - Domain invariant learning, e.g., DANN, CDANN, IW-CDANN
  - Self-training, e.g., PseudoLabel, FixMatch, NoisyStudent, SENTRY
  - Test-time adaptation, e.g., TENT, BN-adapt, CORAL
- **Overall,** we train >30k models in our testbed

Proposed Meta-Algorithm to Handle Class Proportion Shift

- We implement two simple general-purpose corrections

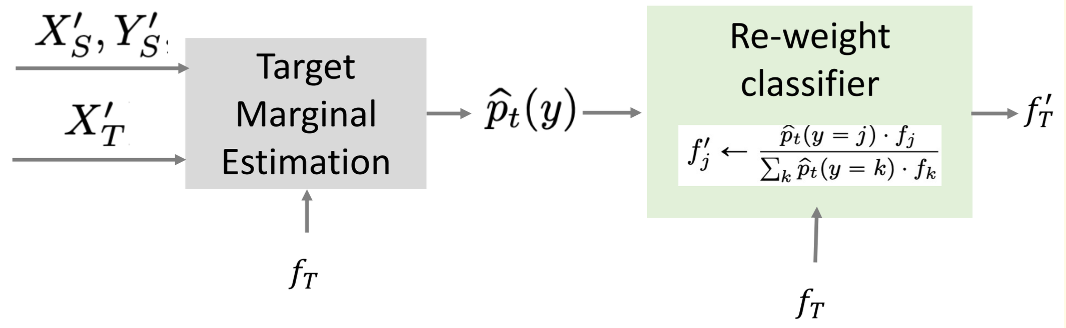
Re-sampling

- Balanced source data
- Use target pseudolabels to perform pseudo class-balanced re-sampling



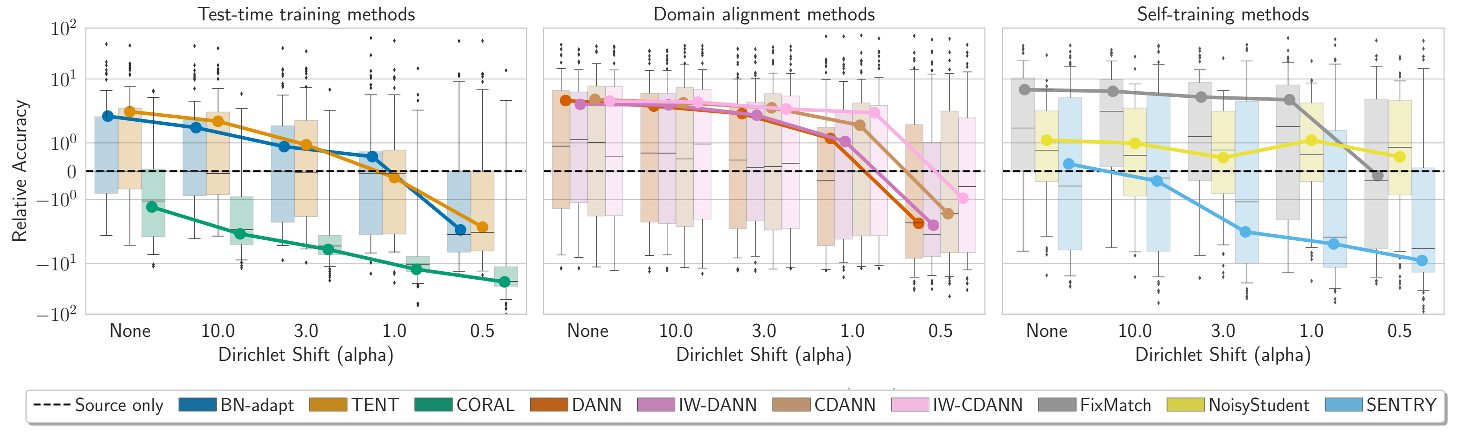
Re-weighting

- Estimate target label marginal with label shift estimation methods (e.g. BBSE, MLLS)
- Post-hoc re-weight the classifier

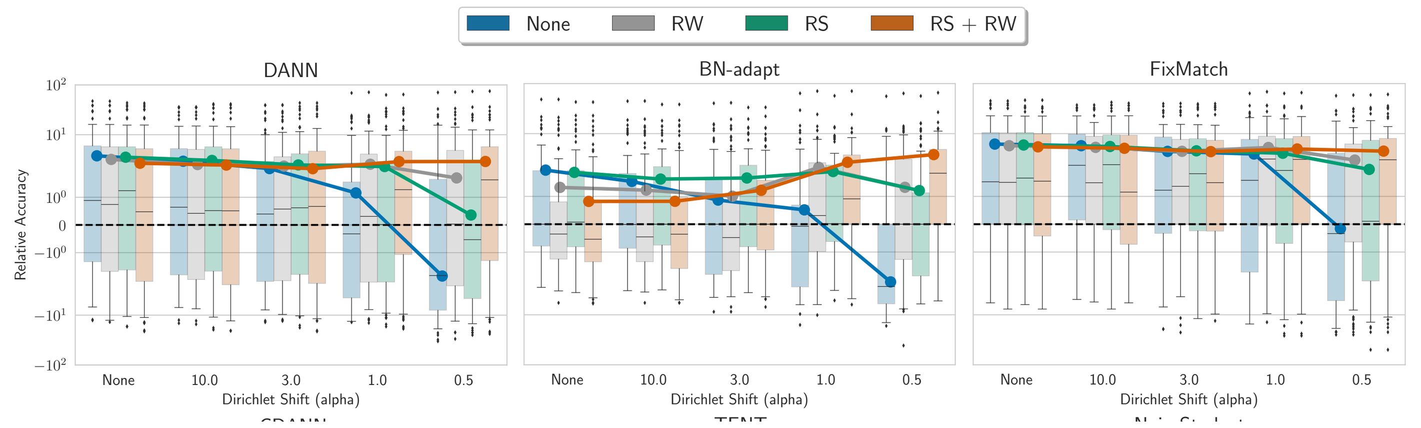


Main Results and Takeaways

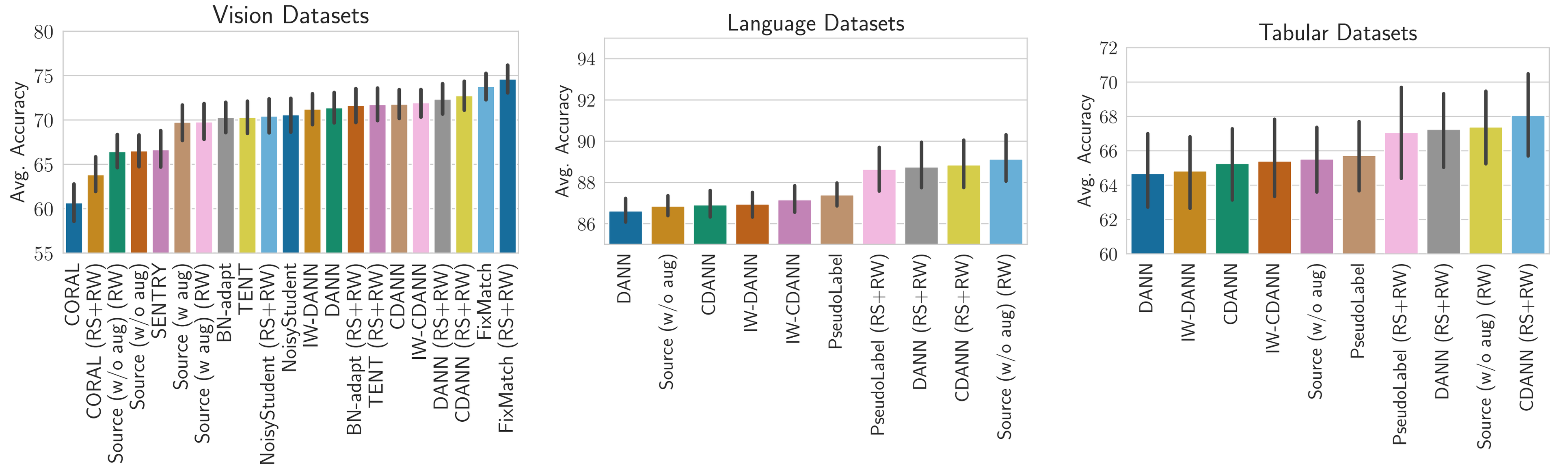
- **Takeaway-1:** Popular deep DA methods without any correction falter



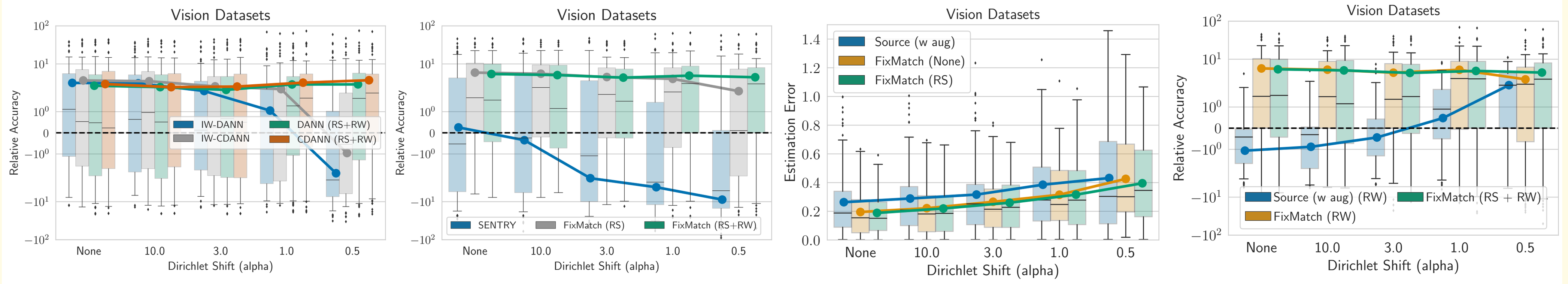
- **Takeaway-2:** Re-sampling to pseudo balance target often helps all DA methods
- **Takeaway-3:** Benefits of post-hoc re-weighting of the classifier depends on shift severity and the underlying DA algorithm.



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



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



- **Takeaway-6:** Deep DA heuristics often improve target label marginal estimation on tabular and vision modalities.
- **Takeaway-7:** With increasing severity of label distribution shift, the accuracy difference with source and target early stopping criterion increases