Distribution Shifts

Medha Agarwal and Scott Geng

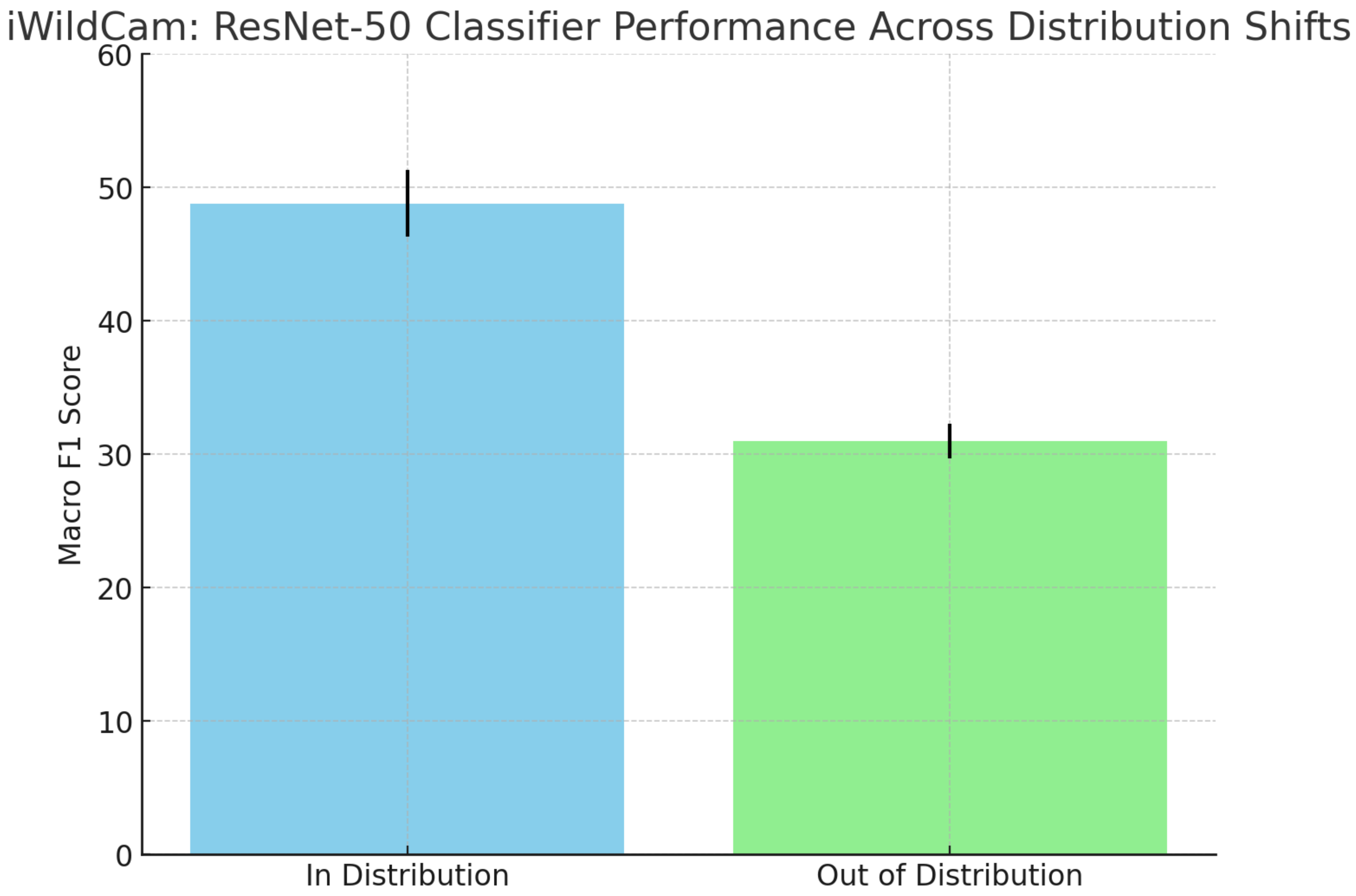
A friendly husky in the WILDS

WILDS: A Benchmark of In-the-Wild Distribution Shifts

Great Curassow

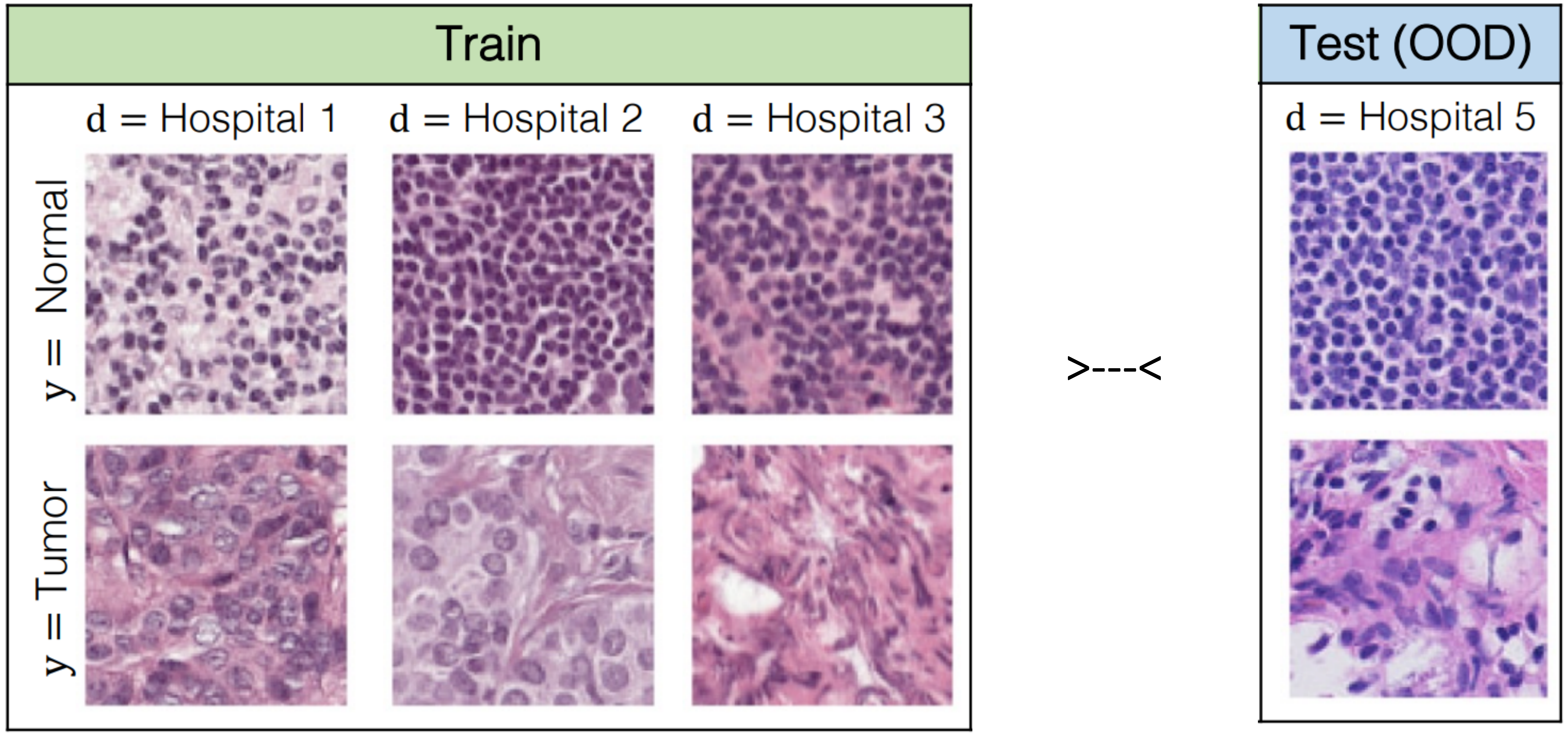
Spurfowl

ML models often fail in the presence of distribution shifts…



[blame ChatGPT for squished text]

…and these failures can have severe real-world ramifications.



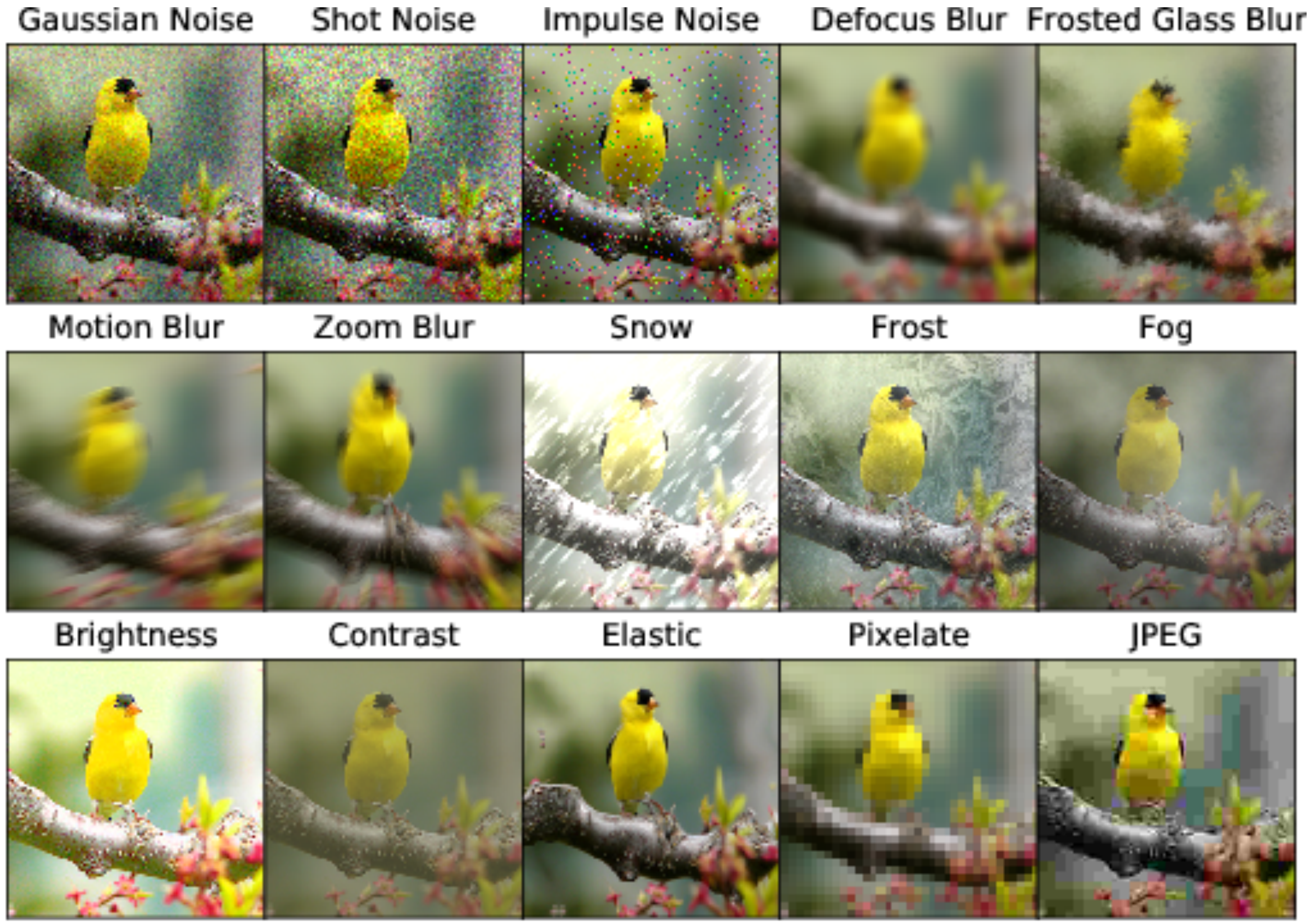
We hope to build models that can generalize well across distribution shifts.

Q: What sorts of ✨datasets✨ have ML researchers used to study this problem?

Colored MNIST

[Arjovsky et al. 2019, Kim et al. 2019]

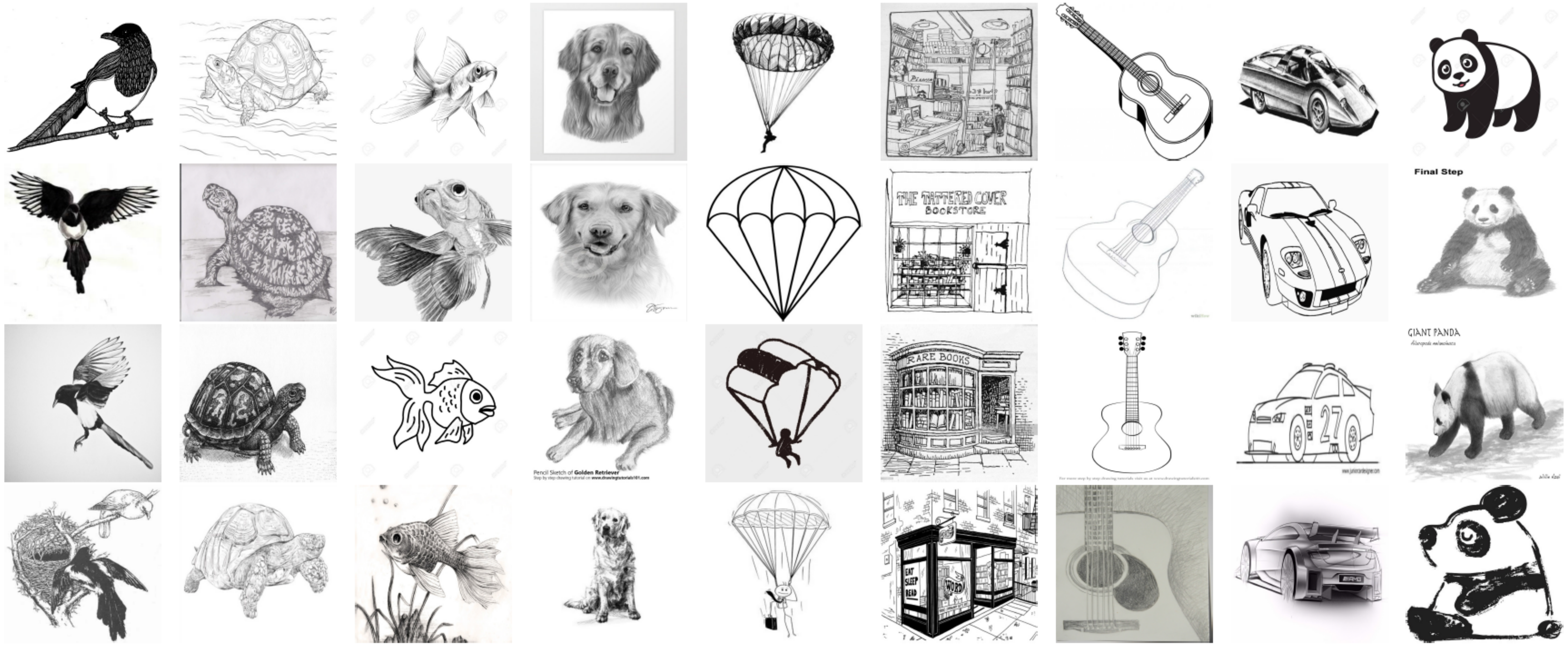
Q: What sorts of ✨datasets✨ have ML researchers used to study this problem?



ImageNet-C

[Hendrycks and Dietterich, 2019]

Q: What sorts of ✨datasets✨ have ML researchers used to study this problem?



ImageNet-Sketch

[Wang et al. 2019]

Q: What sorts of ✨datasets✨ have ML researchers used to study this problem?

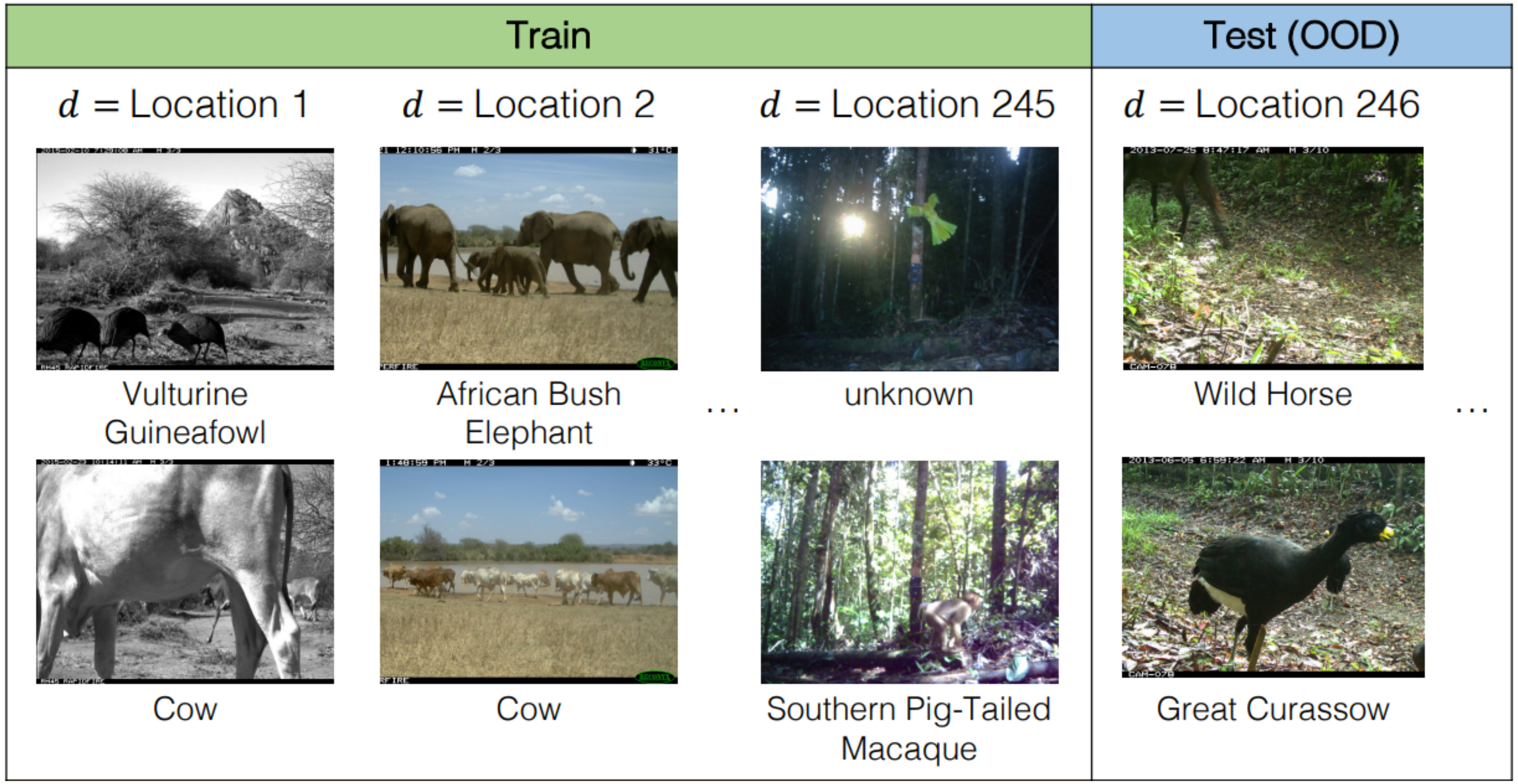
A: ML researchers have predominantly studied datasets of artificial distribution shifts.

Synthetic transformations.

1. Colored MNIST
2. ImageNet-C
3. Waterbirds
4. ……

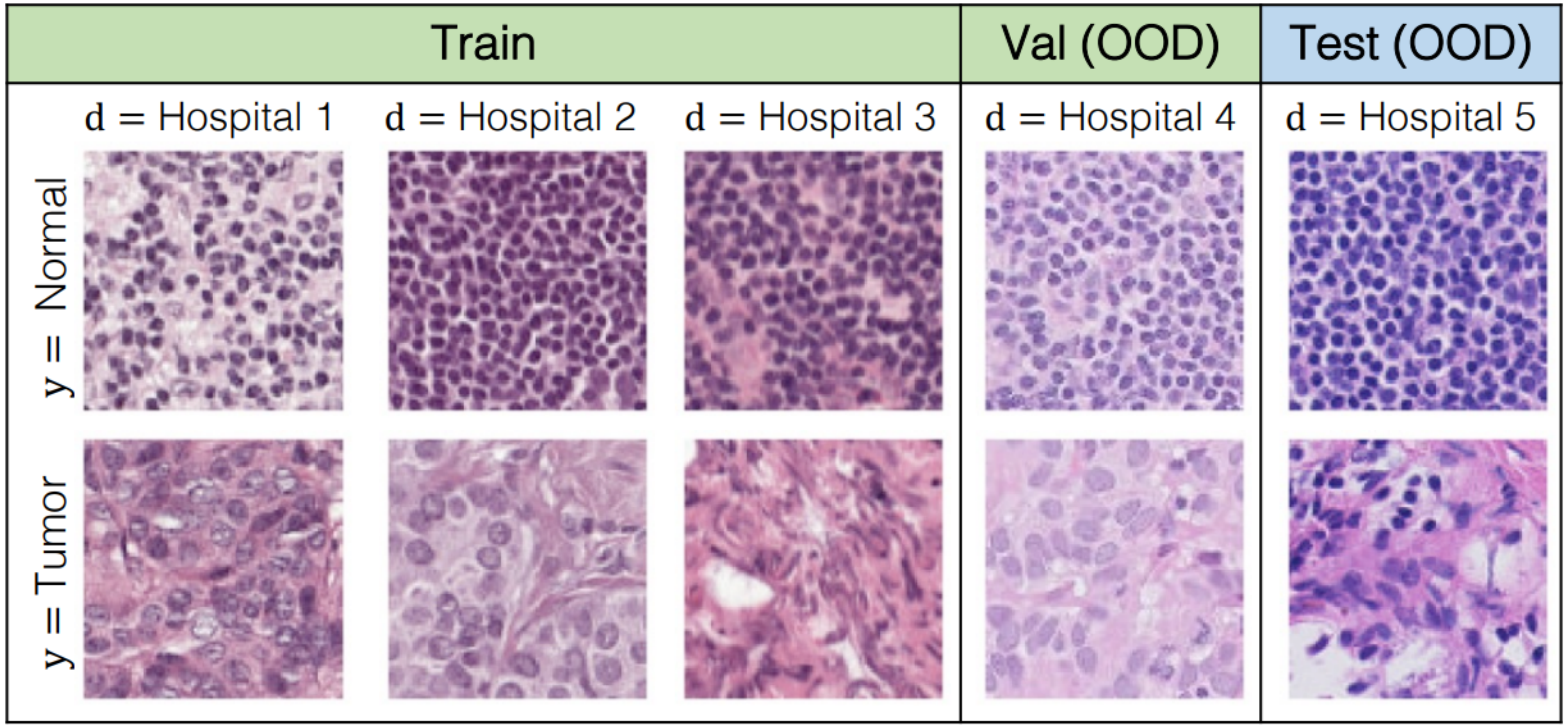
Artificially disparate data splits.

1. ImageNet-Sketch
2. ImageNet-Rendition
3. PACS
4. ……



WILDS

[Koh et al. 2021]



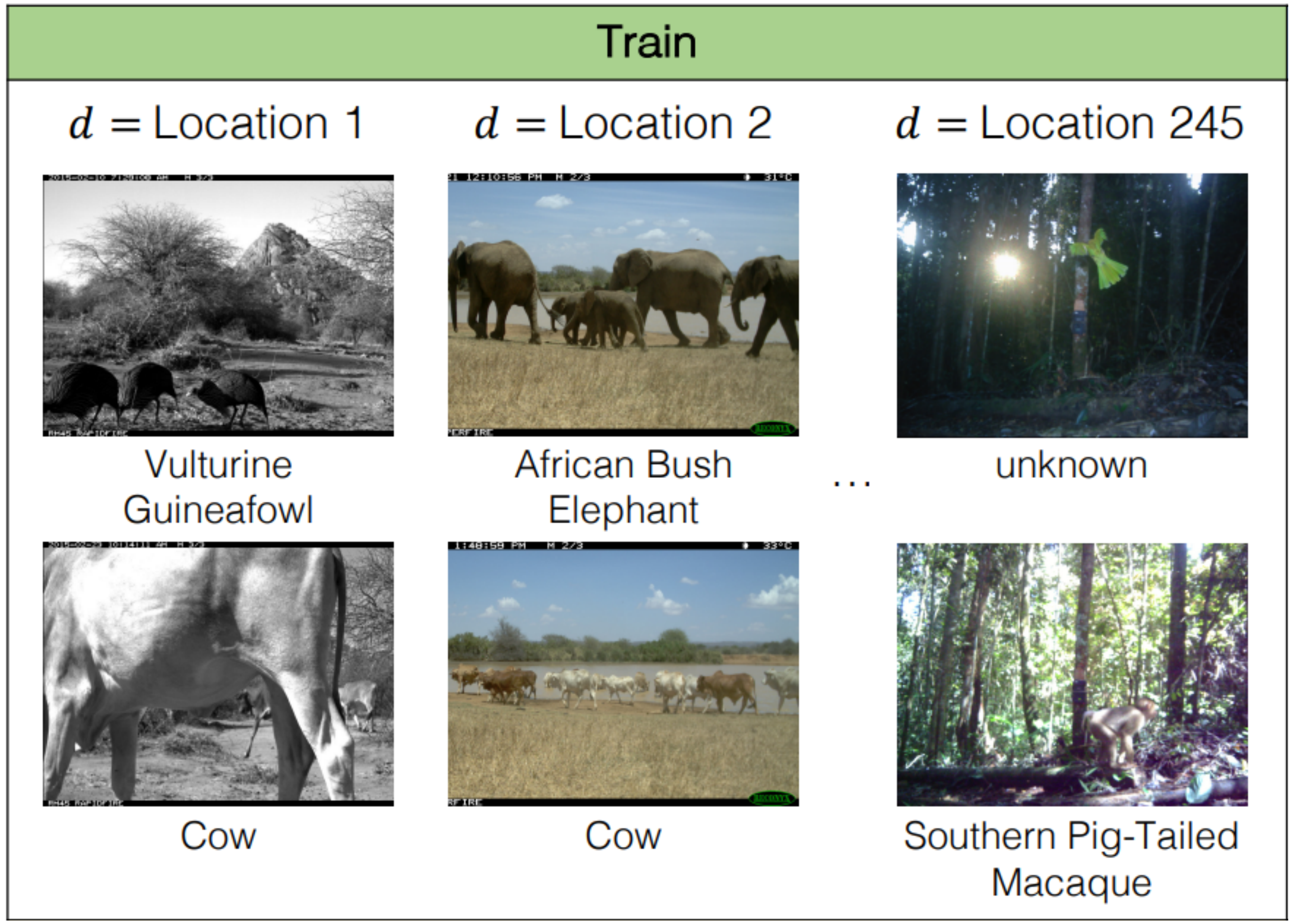
[Koh et al. 2021]

Toxic	Comment Text	Male	Female	LGBTQ	White	Black	…	Christian
0	I applaud your father. He was a good man! We need more like him.	1	0	0	0	0	…	0
0	As a Christian, I will not be patronizing any of those businesses.	0	0	0	0	0	…	1
0	What do Black and LGBT people have to do with bicycle licensing?	0	0	1	0	1	…	0
0	Government agencies track down foreign baddies and protect law-abiding white citizens. How many shows does that describe?	0	0	0	1	0	…	0
1	Maybe you should learn to write a coherent sentence so we can understand WTF your point is.	0	0	0	0	0	…	0

WILDS

[Koh et al. 2021]

Q: With our new dataset, what can we learn?

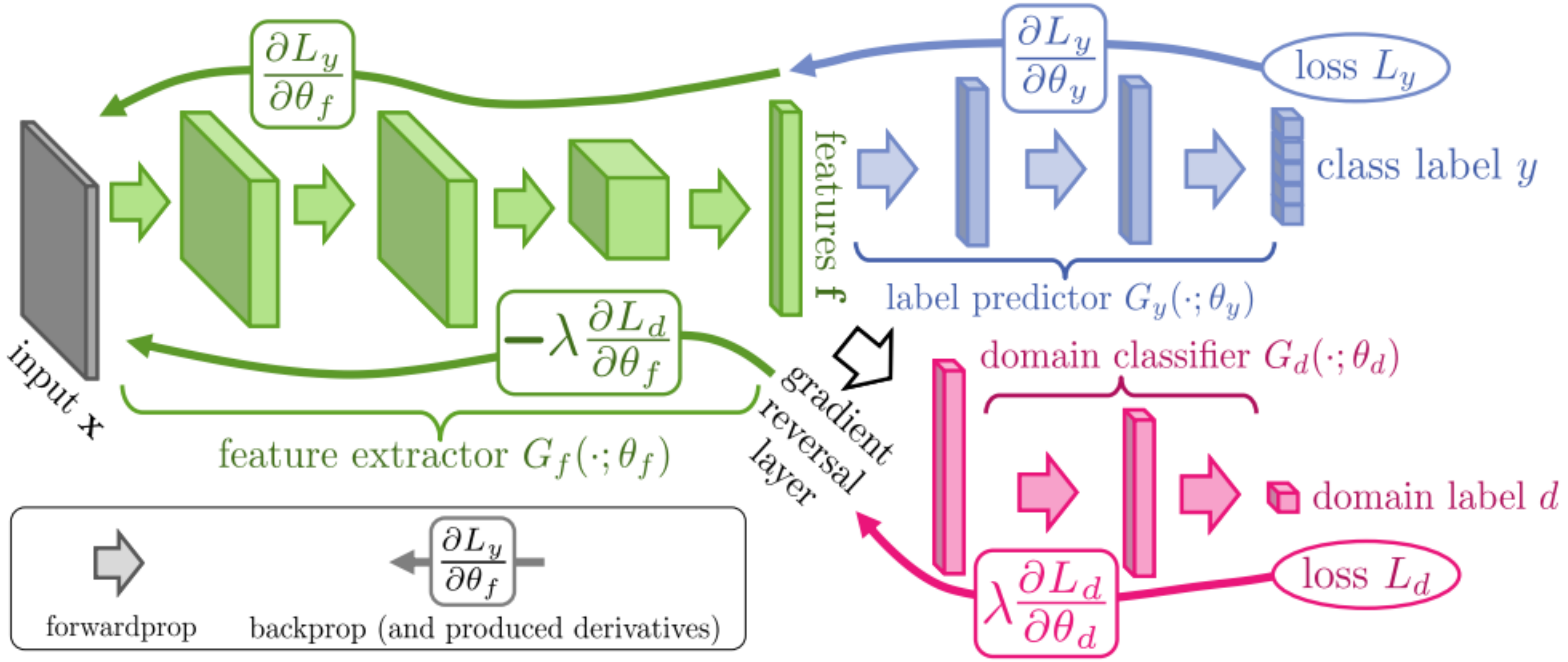


WILDS v2

Additional unlabeled examples
(possibly from test distribution)

[Koh et al. 2021]

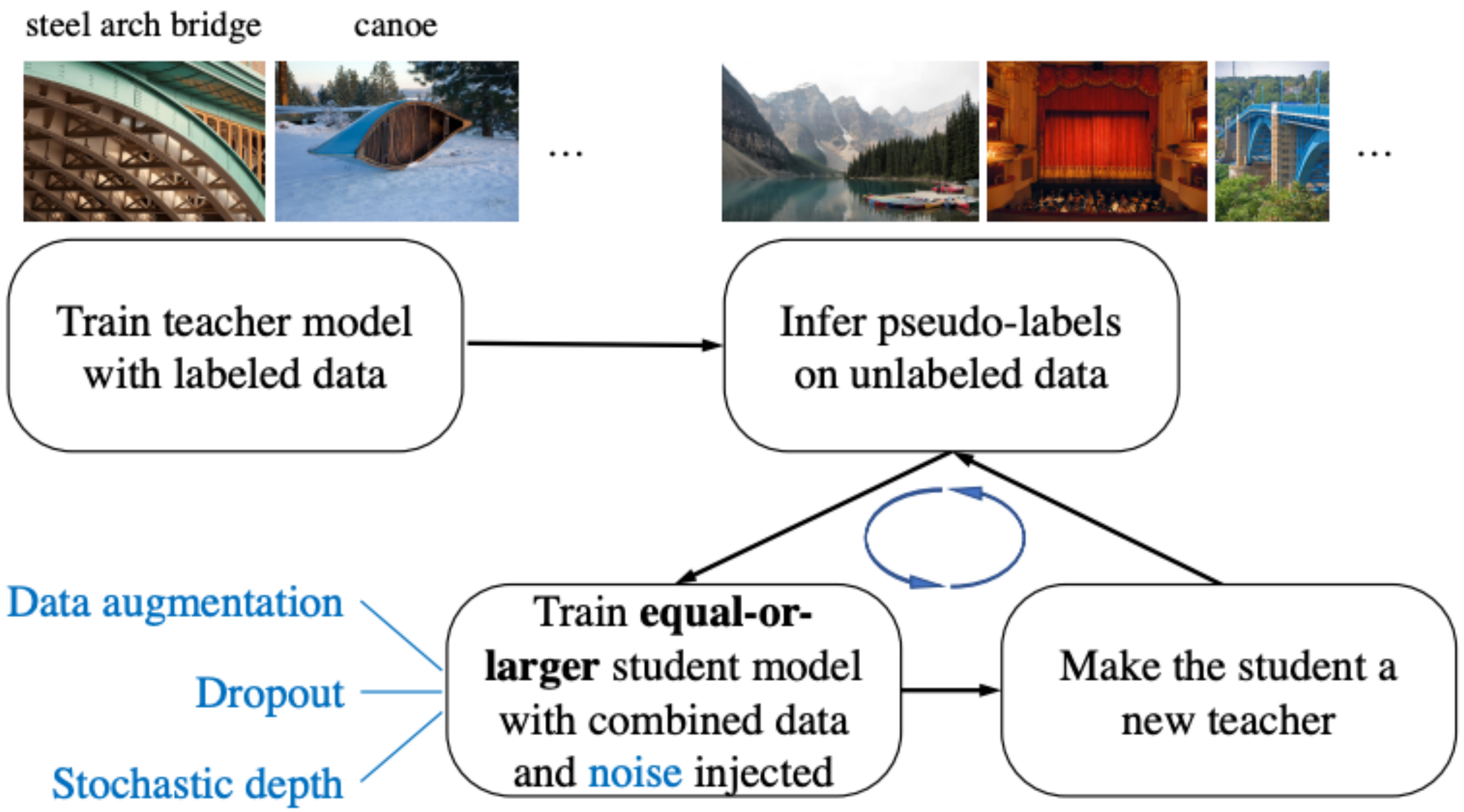
Domain adaptation approach: try to learn features that are invariant across domains.



DANN: I want my learned features to achieve low classification loss on my labeled data and have high domain identification loss across all data.

[Ganin et al. 2015]

Domain adaptation approach: try to self-train by producing pseudo-labels for our unlabeled data.

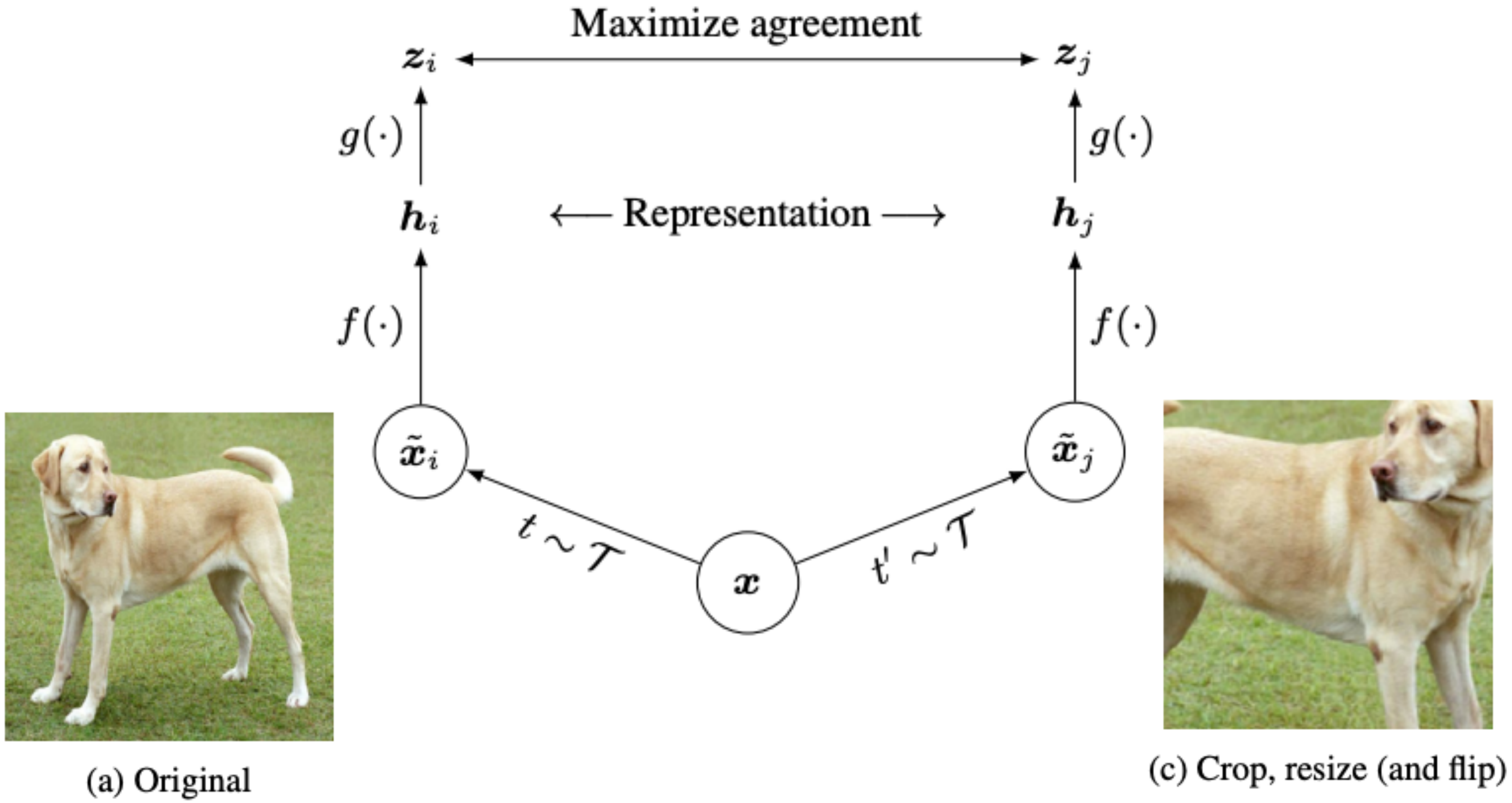


NoisyStudent intuition: **very strong regularization** allows us to avoid overfitting to wrong pseudo-labels.

https://ai.stanford.edu/blog/understanding-self-training/

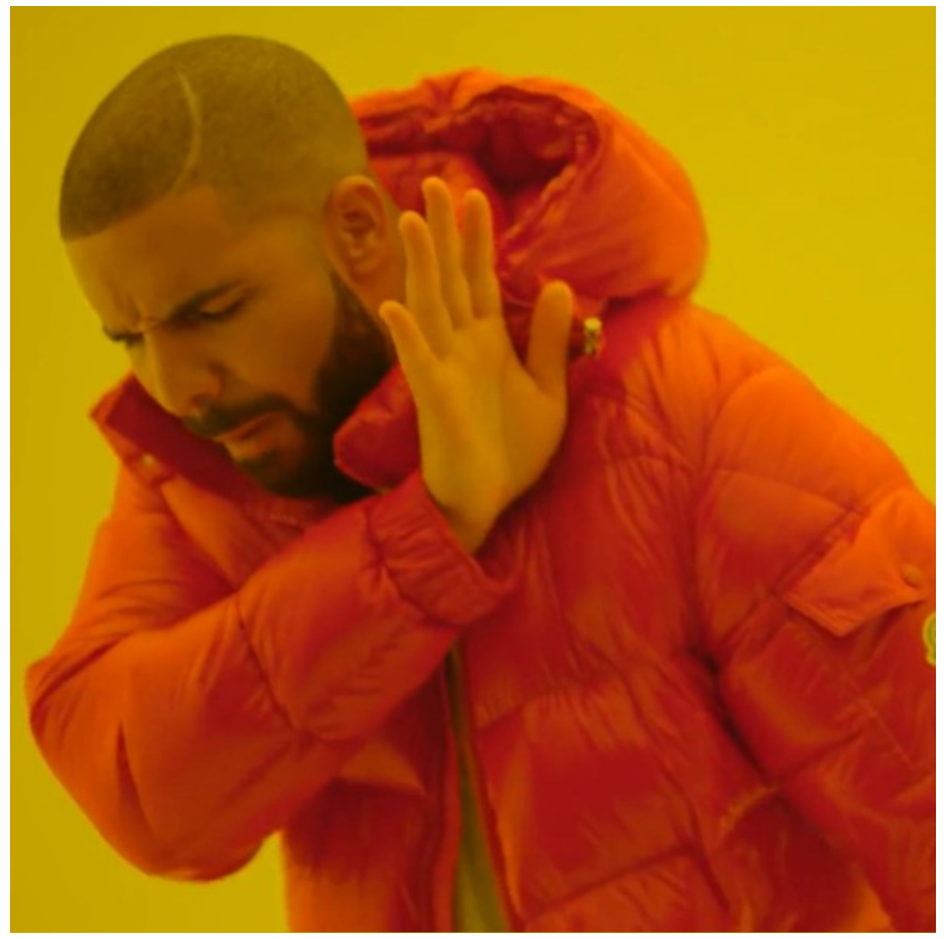
[Xie et al. 2019]

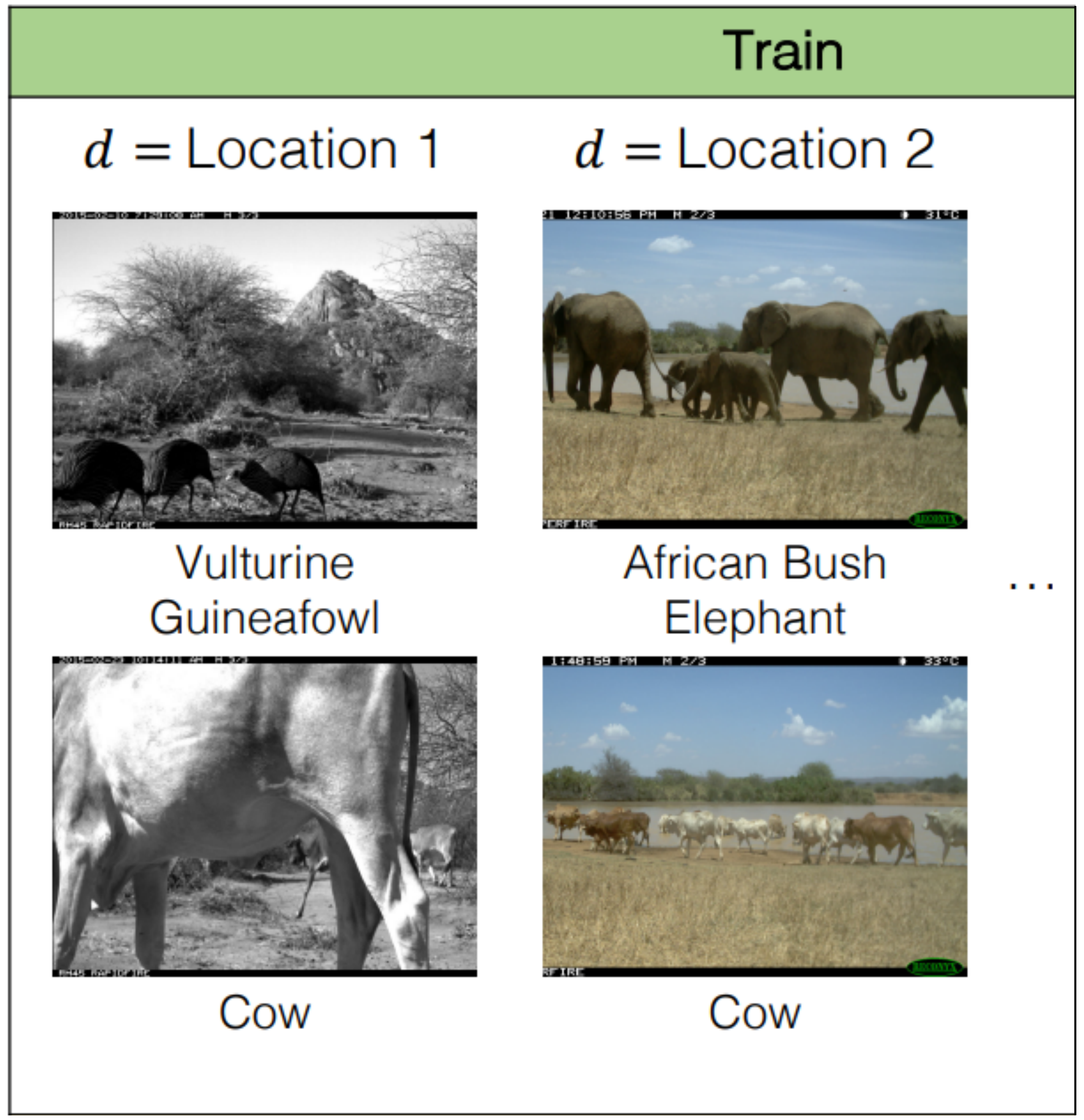
Domain adaptation approach: try to learn from unlabeled data via a self-supervised objective.



[Chen et al. 2020]

Baseline approach: ERM (+/- data augmentation)





Just pretend like our unlabeled data doesn’t exist.

Q: With our new dataset, what can we learn?

SOTA on ImageNet-C 🤯

	iWildCam2020-wilds (Unlabeled extra, macro F1)	
	In-distribution	Out-of-distribution
ERM (-data aug)	46.7 (0.6)	30.6 (1.1)
ERM	47.0 (1.4)	**32.2** (1.2)
CORAL	40.5 (1.4)	27.9 (0.4)
DANN	48.5 (2.8)	**31.9** (1.4)
Pseudo-Label	47.3 (0.4)	30.3 (0.4)
FixMatch	46.3 (0.5)	**31.0** (1.3)
Noisy Student	47.5 (0.9)	**32.1** (0.7)
SwAV	47.3 (1.4)	29.0 (2.0)
ERM (fully-labeled)	54.6 (1.5)	44.0 (2.3)

Q: With our new dataset, what can we learn?

A: Existing domain adaptation methods basically do not work*.

*they largely fail to significantly improve over an ERM baseline on the distribution shifts captured by WILDS.

Takeaway 1: As ML researchers, we should ground our work in (or at least by cognizant of) real-world use.

Takeaway 2: The field of domain adaptation is wide open!

References

- Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *International conference on machine learning*. PMLR, 2020.
- Sagawa, Shiori, et al. "Extending the WILDS benchmark for unsupervised adaptation." *arXiv preprint arXiv:2112.05090* (2021).
- Koh, Pang Wei, et al. "Wilds: A benchmark of in-the-wild distribution shifts." *International Conference on Machine Learning*. PMLR, 2021.
- Ajakan, Hana, et al. "Domain-adversarial neural networks." *arXiv preprint arXiv:1412.4446* (2014).
- Arjovsky, Martin, et al. "Invariant risk minimization." *arXiv preprint arXiv:1907.02893* (2019).
- Xie, Qizhe, et al. "Self-training with noisy student improves imagenet classification." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
- Hendrycks, Dan, and Thomas Dietterich. "Benchmarking neural network robustness to common corruptions and perturbations." arXiv preprint arXiv:1903.12261 (2019).
- Wang, Haohan, et al. "Learning robust global representations by penalizing local predictive power." Advances in Neural Information Processing Systems 32 (2019).
- Kim, Byungju, et al. "Learning not to learn: Training deep neural networks with biased data." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.
- https://ai.stanford.edu/blog/understanding-self-training/

A Theory of Learning from Different Domains

Shai Ben-David · John Blitzer · Koby Crammer ·
Alex Kulesza · Fernando Pereira · Jennifer Wortman Vaughan

CSE 599 Presentation

Medha Agarwal | February 02, 2024

Distribution Shifts

Source Domain $\neq$ Target Domain

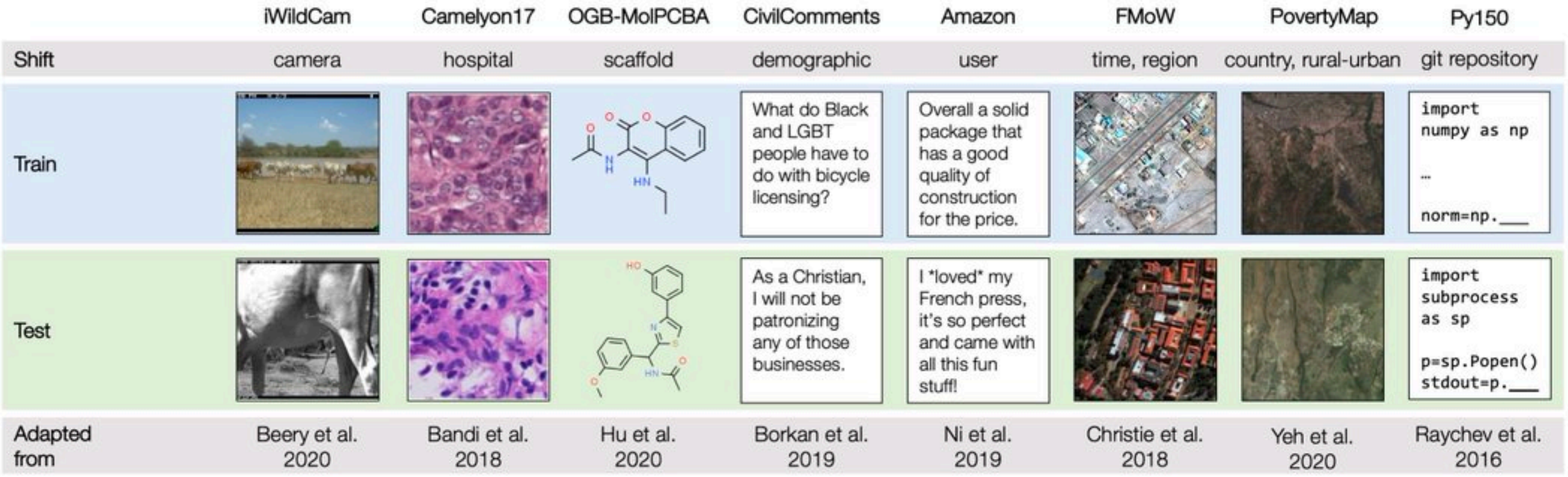


Image credits: Koh, Pang Wei, et al. (2021)

What is Domain?

Binary Classification Setting

Inputs	$\mathscr{X}$
Distribution on inputs	$\mathscr{D}$
Labeling function	$f : \mathscr{X} \to \{0,1\}$
Domain	$(\mathscr{D}, f)$

When source domain $\neq$ target domain, let $(\mathscr{D}_S, f_s)$ = source domain and $(\mathscr{D}_T, f_T)$ = target domain.

TRAIN DATA
TEST DATA
$(\mathscr{D}_1, f_1)$
$(\mathscr{D}_2, f_2)$
$(\mathscr{D}_N, f_N)$
$(\mathscr{D}_T, f_T)$
$\{(X_i, y_i)\}_{i=1}^{m_1}$
$\{(X_i, y_i)\}_{i=1}^{m_2}$
...
$\{(X_i, y_i)\}_{i=1}^{m_N}$
$\{(X_i, y_i)\}_{i=1}^{m_T}$
TRAIN DATA
TEST DATA
$(\mathscr{D}_1, f_1)$
$(\mathscr{D}_2, f_2)$
$(\mathscr{D}_N, f_N)$
$(\mathscr{D}_T, f_T)$
$\{(X_i, y_i)\}_{i=1}^{m_1}$
$\{(X_i, y_i)\}_{i=1}^{m_2}$
...
$\{(X_i, y_i)\}_{i=1}^{m_N}$
$\{X_i\}_{i=1}^{m_T}$

Two Questions in Domain Adaptation

Question 1

Under what conditions can a classifier which performs well on a source data be expected to perform well on the target data?

Question 2

Given a small amount of labeled target data, how should we combine it during training with large amounts of labeled source data to achieve lowest target error at test time?

Quick Answers from the Paper

Answer 1

The authors bound a classifier's target domain error in terms of its **source domain error** and a measure of **divergence between the source & target domain**.

Answer 2

Minimize a **convex combination of the empirical source and target error**. The coefficients depend on the divergence between the domains and the size of source & target data.

Related Work - Theoretical

- Crammer et. al (2008) assume $X_1, \ldots, X_N$ follow same distribution but the deterministic labeling functions $f_1, \ldots, f_N$ are different. They minimize (uniformly weighted) source error.
- Blitzer et. al (2008) give error bounds for the hypothesis learned by minimizing weighted combination of source errors for the case of empirical risk minimization.
- Mansour et. al (2008) give theoretical analysis when the target is a mixture of source domains.
- Mansour et. al (2009) provide bounds on test error using a new discrepancy distance and provide generalized bounds for regularization based algorithms.

Related Work - Applications

- Deep Transfer Networks - Long et. al (2014), (2015), (2016)

Related Work - Applications

- Deep Transfer Networks - Long et. al (2014), (2015), (2016)
- Multi-task Learning

Related Work - Modeling Technology

- Deep Transfer Networks - Long et. al (2014), (2015), (2016)
- Multi-task Learning
- Multiple source adaptation model

Related Work - Modeling Technology

- Deep Transfer Networks - Long et. al (2014), (2015), (2016)
- Multi-task Learning
- Multiple source adaptation model
- Adversarial Learning - Cao et. al (2018)

Model for Domain Adaptation

- A hypothesis is a function $h : \mathscr{X} \rightarrow \{0,1\}$.

Model for Domain Adaptation

- A hypothesis is a function $h : \mathscr{X} \to \{0,1\}$.
- The probability according to distribution $\mathscr{D}_S$ that a hypothesis h disagrees with labeling function f is

$$\epsilon_S(h,f) = \mathbb{E}_{X\sim\mathscr{D}_S}[\,|h(X) - f(X)|\,] = \mathbb{P}_{X\sim\mathscr{D}_S}(h(X) \neq f(X))$$

Model for Domain Adaptation

- A hypothesis is a function $h : \mathscr{X} \to \{0,1\}$.
- The probability according to distribution $\mathscr{D}_S$ that a hypothesis h disagrees with labeling function f is

$$\epsilon_S(h,f) = \mathbb{E}_{X\sim\mathscr{D}_S}[\,|h(X) - f(X)|\,] = \mathbb{P}_{X\sim\mathscr{D}_S}(h(X) \neq f(X))$$

- Risk of a hypothesis/source error: $\epsilon_S(h) = \epsilon_S(h,f_s)$

Model for Domain Adaptation

- A hypothesis is a function $h : \mathscr{X} \to \{0,1\}$.
- The probability according to distribution $\mathscr{D}_S$ that a hypothesis h disagrees with labeling function f is

$$\epsilon_S(h,f) = \mathbb{E}_{X\sim\mathscr{D}_S}[\,|h(X) - f(X)|\,] = \mathbb{P}_{X\sim\mathscr{D}_S}(h(X) \neq f(X))$$

- Risk of a hypothesis/source error: $\epsilon_S(h) = \epsilon_S(h,f_s)$
- Empirical source error $\hat{\epsilon}_S(h)$.

Model for Domain Adaptation

- A hypothesis is a function $h : \mathcal{X} \to \{0,1\}$.
- The probability according to distribution $\mathcal{D}_S$ that a hypothesis h disagrees with labeling function f is

$$\epsilon_S(h,f) = \mathbb{E}_{X \sim \mathcal{D}_S}[\,|h(X) - f(X)|\,] = \mathbb{P}_{X \sim \mathcal{D}_S}(h(X) \neq f(X))$$

- Risk of a hypothesis/source error: $\epsilon_S(h) = \epsilon_S(h, f_s)$
- Empirical source error $\hat{\epsilon}_S(h)$.
- Parallel notation for $\epsilon_S(h,f)$, $\epsilon_T(h)$, and $\hat{\epsilon}_T(h)$.

Answer 1

Establishing bounds on target domain performance of a classifier trained on source domain

Some Definitions

The $\mathscr{H}$-divergence

Given a domain $\mathscr{X}$ with two probability distributions $\mathscr{D}$ and $\mathscr{D}'$.
Let $\mathscr{H}$ be a hypothesis class on $\mathscr{X}$, and
$I(h) = \{x \in \mathscr{X} : h(x) = 1\}$.

$$d_{\mathscr{H}}(\mathscr{D}, \mathscr{D}') = 2 \sup_{h \in \mathscr{H}} |\Pr_{\mathscr{D}}[I(h)] - \Pr_{\mathscr{D}'}[I(h)]|$$

Some Definitions

Ideal Joint Hypothesis

$$h^* = \arg\min_{h \in \mathcal{H}} [\epsilon_S(h) + \epsilon_T(h)]$$

and

$$\lambda = \epsilon_S(h^*) + \epsilon_T(h^*)$$

Some Definitions

Ideal Joint Hypothesis

$$h* = \arg\min_{h \in \mathscr{H}} [\epsilon_S(h) + \epsilon_T(h)]$$

and

$$\lambda = \epsilon_S(h*) + \epsilon_T(h*)$$

When this ideal joint hypothesis performs poorly, we cannot expect to learn a good target classifier by minimizing source error.

Some Definitions

Symmetric Difference Hypothesis

For a hypothesis space $\mathcal{H}$, the symmetric difference hypothesis $\mathcal{H}\Delta\mathcal{H} = \{g : g(x) = h(x) \oplus h'(x)\}$ for some $h, h' \in \mathcal{H}$,

Where $\oplus$ is the XOR function

Every hypothesis $g \in \mathcal{H}\Delta\mathcal{H}$ is the set of disagreements between two hypotheses in $\mathcal{H}$.

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}, \mathcal{D}') = 2 \sup_{h,h' \in \mathcal{H}} \left| \Pr_{X\sim\mathcal{D}_S}[h(X) \neq h'(X)] - \Pr_{X\sim\mathcal{D}_T}[h(X) \neq h'(X)] \right|$$

Main Result

Theorem 2 *Let $\mathcal{H}$ be a hypothesis space of VC dimension d. If $\mathcal{U}_S, \mathcal{U}_T$ are unlabeled samples of size m' each, drawn from $\mathcal{D}_S$ and $\mathcal{D}_T$ respectively, then for any $\delta \in (0,1)$, with probability at least $1-\delta$ (over the choice of the samples), for every $h \in \mathcal{H}$:*

$$\epsilon_T(h) \leq \epsilon_S(h) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_S, \mathcal{U}_T) + 4\sqrt{\frac{2d\log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda.$$

Main Result

Theorem 2 *Let $\mathcal{H}$ be a hypothesis space of VC dimension d. If $\mathcal{U}_S, \mathcal{U}_T$ are unlabeled samples of size m' each, drawn from $\mathcal{D}_S$ and $\mathcal{D}_T$ respectively, then for any $\delta \in (0,1)$, with probability at least $1-\delta$ (over the choice of the samples), for every $h \in \mathcal{H}$:*

$$\epsilon_T(h) \leq \epsilon_S(h) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_S, \mathcal{U}_T) + 4\sqrt{\frac{2d\log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda.$$

When λ is small, domain adaptation is relevant => source error and unlabeled $\mathscr{H}\Delta\mathscr{H}$ -divergence are important for bounding target error.

Answer 2

A learning bound combining source and target data

Setup

- Sample $S = (S_S, S_T)$ of m instances.
- S_T consists of βm i.i.d. samples from $\mathscr{D}_T$.
- S_S consists of $(1-\beta)m$ i.i.d. samples from $\mathscr{D}_S$.
- Goal: find a hypothesis h that minimizes $\epsilon_T(h)$.
- When β is small, minimizing empirical target error is not feasible.
- Consider minimizing: $\hat{\epsilon}_\alpha(h) := \alpha\hat{\epsilon}_T(h) + (1-\alpha)\hat{\epsilon}_S(h)$

Main Result

Theorem 3 *Let $\mathcal{H}$ be a hypothesis space of VC dimension d. Let $\mathcal{U}_S$ and $\mathcal{U}_T$ be unlabeled samples of size m' each, drawn from $\mathcal{D}_S$ and $\mathcal{D}_T$ respectively. Let S be a labeled sample of size m generated by drawing βm points from $\mathcal{D}_T$ and $(1-\beta)m$ points from $\mathcal{D}_S$ and labeling them according to f_S and f_T, respectively. If $\hat{h} \in \mathcal{H}$ is the empirical minimizer of $\hat{\epsilon}_\alpha(h)$ on S and $h_T^* = \min_{h \in \mathcal{H}} \epsilon_T(h)$ is the target error minimizer, then for any $\delta \in (0,1)$, with probability at least $1-\delta$ (over the choice of the samples),*

$$\epsilon_T(\hat{h}) \leq \epsilon_T(h_T^*) + 4\sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{2d\log(2(m+1)) + 2\log(\frac{8}{\delta})}{m}}$$

$$+ 2(1-\alpha)\left(\frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_S, \mathcal{U}_T) + 4\sqrt{\frac{2d\log(2m') + \log(\frac{8}{\delta})}{m'}} + \lambda\right).$$

Observations

- When $\alpha = 0$ (ignore target data) and $\alpha = 1$(ignore source data) the bound coincides with known bounds on target error.

- Choosing $\alpha \in (0,1)$ optimally allows us to tradeoff “small” amounts of “good” vs “large” amounts of “less relevant” source data.

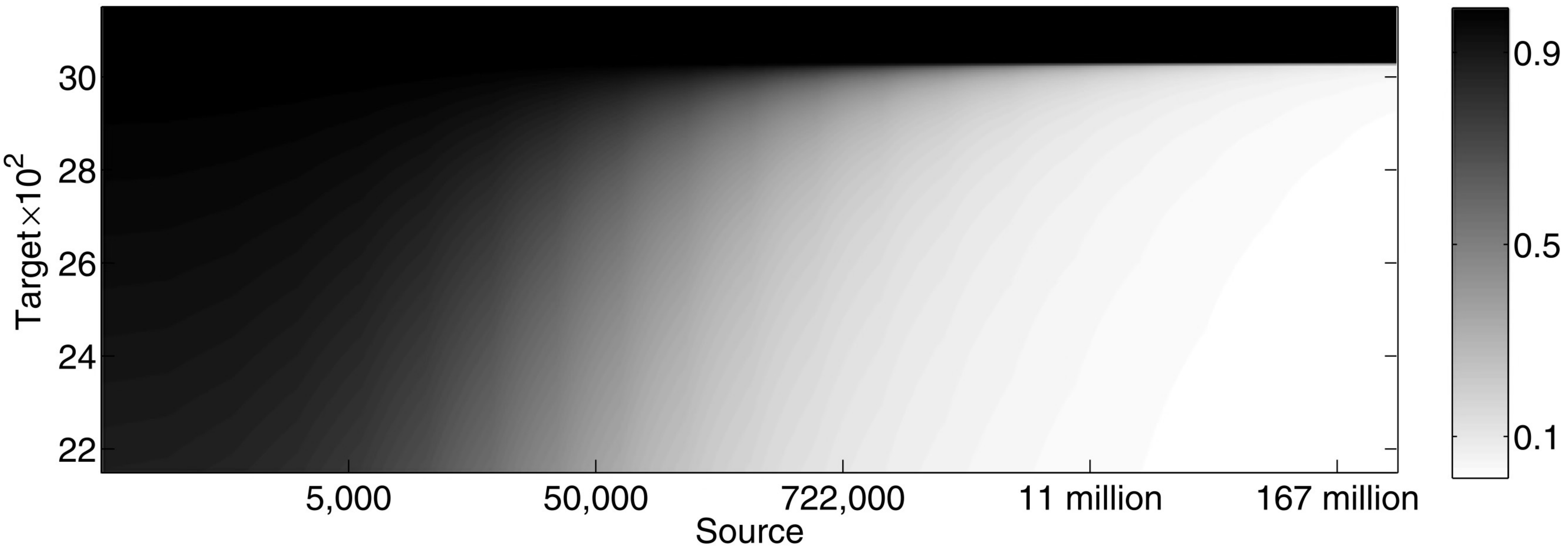
Optimal Mixing

$$\boxed{\begin{aligned} \alpha^*(m_T, m_S; D) &= \begin{cases} 1 & m_T \geq D^2 \\ \min\{1, \nu\} & m_T \leq D^2, \end{cases} \\ \nu &= \frac{m_T}{m_T + m_S}\left(1 + \frac{m_S}{\sqrt{D^2(m_S + m_T) - m_S m_T}}\right). \end{aligned}}$$

$D = \sqrt{d}/A$ where

$$A = \left(\frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_S, \mathcal{U}_T) + 4\sqrt{\frac{2d\log(2m') + \log(\frac{4}{\delta})}{m'}} + \lambda\right)$$

Optimal Mixing Illustration



Thank you!
Question?

Bibliography

Crammer, K., Kearns, M., & Wortman, J. (2008). Learning from multiple sources. Journal of Machine Learning Research, 9, 1757–1774

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Wortman, J. (2008). Learning bounds for domain adaptation. Proceedings of NIPS 2007.

Mansour, Y., Mohri, M., & Rostamizadeh, A. Domain adaptation with multiple sources. Advances in Neural Information Processing Systems (2008).

Mansour, Yishay, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430* (2009).

Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer joint matching for unsupervised domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1410–1417, 2014. [34]

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In International Conference on Machine Learning, pages 97–105, 2015. [35]

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In Advances in Neural Information Processing Systems, pages 136–144, 2016.

Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, ECCV (8), volume 11212 of Lecture Notes in Computer Science, pages 139–155. Springer, 2018. ISBN 978-3-030-01237-3.

Bonus

Combining Data from Multiple Sources

Combining Data from Multiple Sources

- Source data comes from N distinct sources.
- Each source S_j has distribution $\mathcal{D}_j$ over inputs and labeling function f_j.
- Out of total m source samples, $\beta_j m$ are from source S_j.
- Minimizing convex combination of training error from different source using domain weights $\alpha = (\alpha_1, \ldots, \alpha_N)$,

$$\boxed{\hat{\epsilon}_\alpha(h) = \sum_{j=1}^{N} \alpha_j \hat{\epsilon}_j(h) = \sum_{j=1}^{N} \frac{\alpha_j}{\beta_j m} \sum_{x \in S_j} |h(x) - f_j(x)|}$$

A bound using pairwise divergence

for any $\delta \in (0,1)$, *with probability at least* $1-\delta$,

$$\epsilon_T(\hat{h}) \leq \epsilon_T(h_T^*) + 2\sqrt{\left(\sum_{j=1}^{N} \frac{\alpha_j^2}{\beta_j}\right)\left(\frac{d\log(2m) - \log(\delta)}{2m}\right)}$$

$$+ \sum_{j=1}^{N} \alpha_j \left(2\lambda_j + d_{\mathcal{H}\Delta\mathcal{H}}(D_j, D_T)\right),$$

where $\lambda_j = \min_{h\in\mathcal{H}}\{\epsilon_T(h) + \epsilon_j(h)\}$.

A bound using combined divergence

for any $\delta \in (0, 1)$, *with probability at least* $1 - \delta$,

$$\epsilon_T(\hat{h}) \le \epsilon_T(h_T^*) + 4\sqrt{\left(\sum_{j=1}^{N} \frac{\alpha_j^2}{\beta_j}\right)\left(\frac{d\log(2m) - \log(\delta)}{2m}\right)}$$

$$+ 2\gamma_{\boldsymbol{\alpha}} + d_{\mathcal{H}\Delta\mathcal{H}}(D_{\boldsymbol{\alpha}}, D_T),$$

where $\gamma_{\boldsymbol{\alpha}} = \min_h\{\epsilon_T(h) + \epsilon_{\boldsymbol{\alpha}}(h)\} = \min_h\{\epsilon_T(h) + \sum_{j=1}^{N} \alpha_j \epsilon_j(h)\}.$