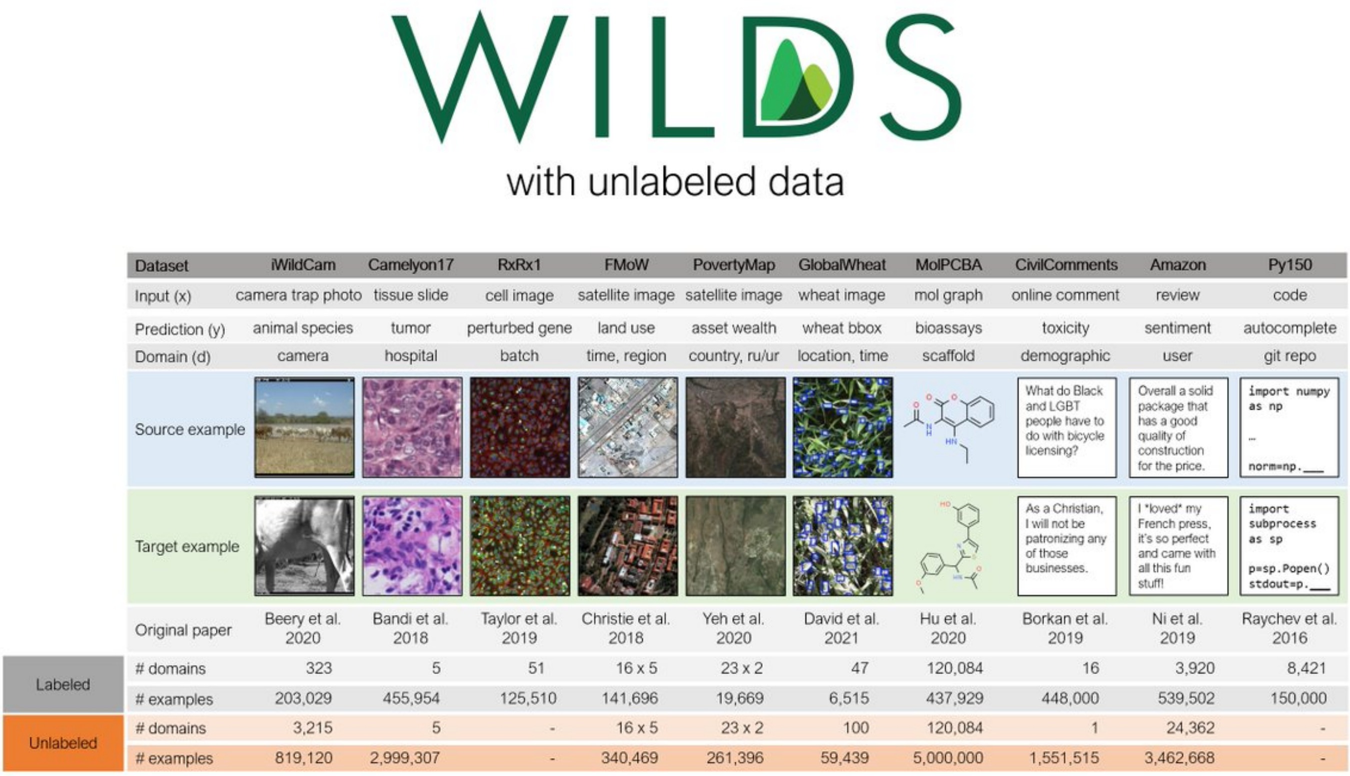
Twitter Thread by Pang Wei Koh

Pang Wei Koh
@PangWeiKoh

We're excited to announce WILDS v2.0, which adds unlabeled data to 8 datasets! This lets us benchmark methods for domain adaptation & representation learning. All labeled data & evaluations are unchanged.

(New) paper: https://t.co/9MaYUFluu7
Website: https://t.co/vA5KxsZf6c

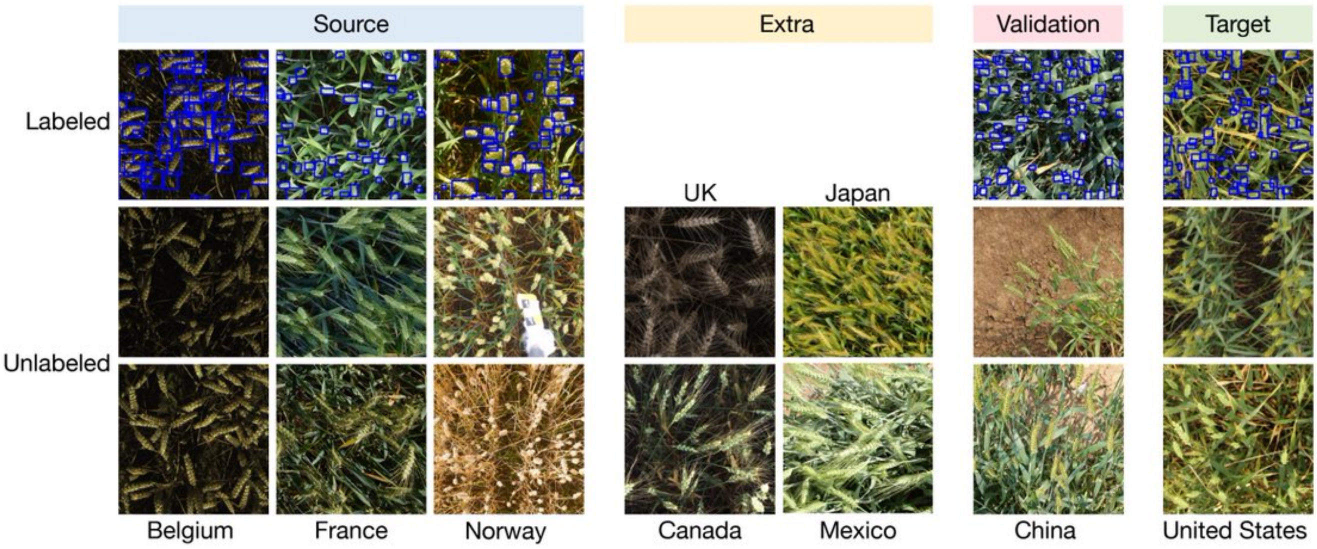
■



	Dataset	iWildCam	Camelyon17	RxRx1	FMoW	PovertyMap	GlobalWheat	MolPCBA	CivilComments	Amazon	Py150
	Input (x)	camera trap photo	tissue slide	cell image	satellite image	satellite image	wheat image	mol graph	online comment	review	code
	Prediction (y)	animal species	tumor	perturbed gene	land use	asset wealth	wheat bbox	bioassays	toxicity	sentiment	autocomplete
	Domain (d)	camera	hospital	batch	time, region	country, ru/ur	location, time	scaffold	demographic	user	git repo
	Source example								What do Black and LGBT people have to do with bicycle licensing?	Overall a solid package that has a good quality of construction for the price.	import numpy as np … norm=np.___
	Target example								As a Christian, I will not be patronizing any of those businesses.	I *loved* my French press, it's so perfect and came with all this fun stuff!	import subprocess as sp p=sp.Popen() stdout=p.___
	Original paper	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Christie et al. 2018	Yeh et al. 2020	David et al. 2021	Hu et al. 2020	Borkan et al. 2019	Ni et al. 2019	Raychev et al. 2016
Labeled	# domains	323	5	51	16 x 5	23 x 2	47	120,084	16	3,920	8,421
Labeled	# examples	203,029	455,954	125,510	141,696	19,669	6,515	437,929	448,000	539,502	150,000
Unlabeled	# domains	3,215	5	-	16 x 5	23 x 2	100	120,084	1	24,362	-
Unlabeled	# examples	819,120	2,999,307	-	340,469	261,396	59,439	5,000,000	1,551,515	3,462,668	-

Unlabeled data can be a powerful source of leverage. It comes from a mixture of:
- source domains (same as the labeled training data)
- target domains (same as the labeled test data)
- extra domains with no labeled data.

We illustrate this for the GlobalWheat dataset:



We evaluated domain adaptation, self-training, & self-supervised methods on these datasets. Unfortunately, many methods did not do better than standard supervised training, despite using additional unlabeled data.

This table shows OOD test performance; higher numbers are better.

	iWildCam2020-wilds (Unlabeled extra, macro F1)		FMoW-wilds (Unlabeled target, worst-region acc)	
	In-distribution	Out-of-distribution	In-distribution	Out-of-distribution
ERM (-data aug)	46.7 (0.6)	30.6 (1.1)	59.3 (0.7)	33.7 (1.5)
ERM	47.0 (1.4)	**32.2** (1.2)	60.6 (0.6)	34.8 (1.5)
CORAL	40.5 (1.4)	27.9 (0.4)	58.9 (0.3)	34.1 (0.6)
DANN	48.5 (2.8)	**31.9** (1.4)	57.9 (0.8)	34.6 (1.7)
Pseudo-Label	47.3 (0.4)	30.3 (0.4)	60.9 (0.5)	33.7 (0.2)
FixMatch	46.3 (0.5)	**31.0** (1.3)	58.6 (2.4)	32.1 (2.0)
Noisy Student	47.5 (0.9)	**32.1** (0.7)	61.3 (0.4)	**37.8** (0.6)
SwAV	47.3 (1.4)	29.0 (2.0)	61.8 (1.0)	36.3 (1.0)
ERM (fully-labeled)	54.6 (1.5)	44.0 (2.3)	65.4 (0.4)	58.7 (1.4)

	Camelyon17-wilds (Unlabeled target, avg acc)		PovertyMap-wilds (Unlabeled target, worst U/R corr)	
	In-distribution	Out-of-distribution	In-distribution	Out-of-distribution
ERM (-data aug)	85.8 (1.9)	70.8 (7.2)	0.65 (0.03)	**0.50** (0.07)
ERM	90.6 (1.2)	82.0 (7.4)	0.66 (0.04)	**0.49** (0.06)
CORAL	90.4 (0.9)	77.9 (6.6)	0.54 (0.10)	0.36 (0.08)
DANN	86.9 (2.2)	68.4 (9.2)	0.50 (0.07)	0.33 (0.10)
Pseudo-Label	91.3 (1.3)	67.7 (8.2)	–	–
FixMatch	91.3 (1.1)	71.0 (4.9)	0.54 (0.11)	0.30 (0.11)
Noisy Student	93.2 (0.5)	86.7 (1.7)	0.61 (0.07)	0.42 (0.11)
SwAV	92.3 (0.4)	**91.4** (2.0)	0.60 (0.13)	**0.45** (0.05)

	GlobalWheat-wilds (Unlabeled target, avg domain acc)		OGB-MolPCBA (Unlabeled target, avg AP)	
	In-distribution	Out-of-distribution	In-distribution	Out-of-distribution
ERM	77.8 (0.2)	**51.0** (0.7)	–	**28.3** (0.1)
CORAL	–	–	–	26.6 (0.2)
DANN	–	–	–	20.4 (0.8)
Pseudo-Label	73.3 (0.9)	42.9 (2.3)	–	19.7 (0.1)
Noisy Student	78.1 (0.3)	46.8 (1.2)	–	27.5 (0.1)

	CivilComments-wilds (Unlabeled extra, worst-group acc)		Amazon-wilds (Unlabeled target, 10th percentile acc)	
	In-distribution	Out-of-distribution	In-distribution	Out-of-distribution
ERM	89.8 (0.8)	**66.6** (1.6)	72.0 (0.1)	**54.2** (0.8)
CORAL	–	–	71.7 (0.1)	53.3 (0.0)
DANN	–	–	71.7 (0.1)	53.3 (0.0)
Pseudo-Label	90.3 (0.5)	**66.9** (2.6)	71.6 (0.1)	52.3 (1.1)
Masked LM	89.4 (1.2)	**65.7** (2.3)	71.9 (0.4)	**53.9** (0.7)
ERM (fully-labeled)	89.9 (0.1)	69.4 (0.6)	73.6 (0.1)	56.4 (0.8)

In contrast, prior work has shown these methods to be successful on standard domain adaptation tasks such as DomainNet, which we replicate below. This underscores the importance of developing and evaluating methods on a broad variety of distribution shifts.

	In-distribution (real)	Out-of-distribution (sketch)
ERM (-data aug)	82.6 (0.0)	34.9 (0.2)
ERM	82.5 (0.3)	35.9 (0.3)
CORAL	79.1 (0.4)	33.6 (0.6)
DANN	77.8 (0.2)	39.4 (0.8)
Pseudo-Label	79.9 (0.2)	36.1 (0.4)
Pseudo-Label (weak aug)	79.9 (0.6)	32.0 (0.8)
FixMatch	80.8 (0.2)	**50.2** (0.4)
FixMatch (weak aug)	80.1 (0.1)	49.3 (0.2)
Noisy Student	82.0 (0.3)	39.7 (0.2)
SwAV	79.0 (0.3)	38.2 (0.4)

We've added the unlabeled data loaders + method implementations to our Python package: https://t.co/S73kjDxMis. They're easy to use: check out the code snippet below!

We've also updated our leaderboards to accept submissions with and without unlabeled data.

```
from wilds import get_dataset
from wilds.common.data_loaders import get_train_loader
import torchvision.transforms as transforms
# Load the labeled data
dataset = get_dataset(dataset="fmow", download=True)
labeled_subset = dataset.get_subset("train", transform=transforms.ToTensor())
data_loader = get_train_loader("standard", labeled_subset, batch_size=16)
# Load the unlabeled data
dataset = get_dataset(dataset="fmow", unlabeled=True, download=True)
unlabeled_subset = dataset.get_subset("test_unlabeled", transform=transforms.ToTensor())
unlabeled_data_loader = get_train_loader("standard", unlabeled_subset, batch_size=64)
# Train loop
for labeled_batch, unlabeled_batch in zip(data_loader, unlabeled_data_loader):
    x, y, metadata = labeled_batch
            unlabeled_x, unlabeled_metadata = unlabeled_batch
    ...
```

We've uploaded the exact commands and hyperparameters used in our paper, as well as trained model checkpoints, to https://t.co/qI7yvTWGsT. This is thanks to @tonyh_lee, who oversaw all of the experimental infrastructure and made it fully reproducible on @CodaLabWS.

We're grateful to everyone who helped us with WILDS and the v2.0 update: https://t.co/1CAsr8JV99.

We'd also like to thank Jiang et al. for https://t.co/CSIYF8gcFT and Zhang et al. for https://t.co/Kla5i4C9Y9, which were very helpful references for our method implementations.

This was joint work with @shiorisagawa* @tonyh_lee* IrenaGao*, and @sangmichaelxie @kendrick_shen @ananyaku @weihua916 @michiyasunaga HenrikMarklund @sarameghanbeery @EtienneDavid @IanStavness @guowei_net @jure @kate_saenko_ @tatsu_hashimoto @svlevine @chelseabfinn @percyliang.

We'll be presenting this at the DistShift workshop at NeurIPS. Find us at our poster on Dec 13, 1-3pm Pacific Time: https://t.co/gid3wBSqb6

Read our paper for more details and analysis: https://t.co/m95JSY9LbJ