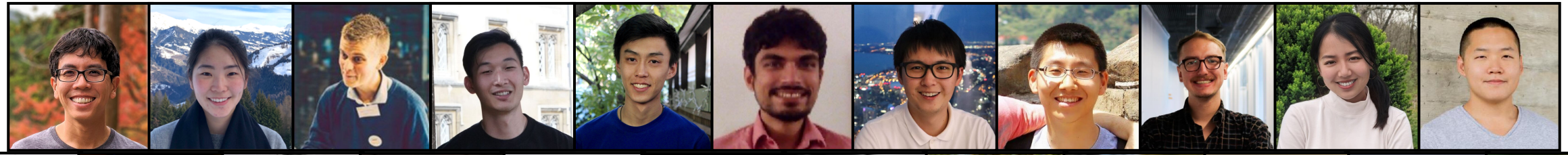
WILDS

A Benchmark of in-the-Wild Distribution Shifts

Pang Wei Koh* **Shiori Sagawa*** Henrik Marklund Michael Xie Marvin Zhang Akshay Balsubramani Weihua Hu Michihiro Yasunaga Richard Phillips Irena Gao Tony Lee



Etienne David Ian Stavness Wei Guo Berton Earnshaw Imran Haque Sara Beery Jure Leskovec Anshul Kundaje Emma Pierson Sergey Levine Chelsea Finn Percy Liang

Stanford, UC Berkeley, Cornell, Caltech, Microsoft Research, University of Tokyo, INRAE, University of Saskatchewan, Recursion

Standard assumption in machine learning

Train data distribution = Test data distribution

Models perform well

Distribution shifts can cause models to fail

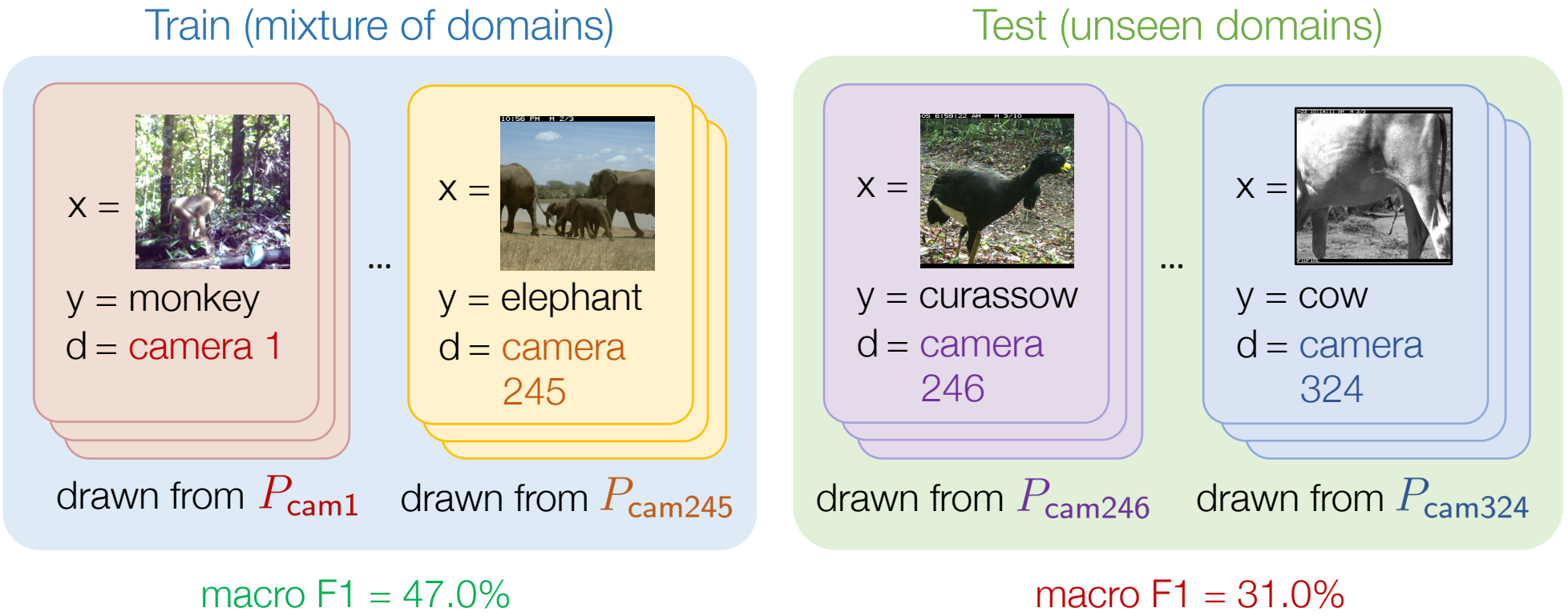
Train data distribution

$\neq$

Test data distribution

Model performance degrades

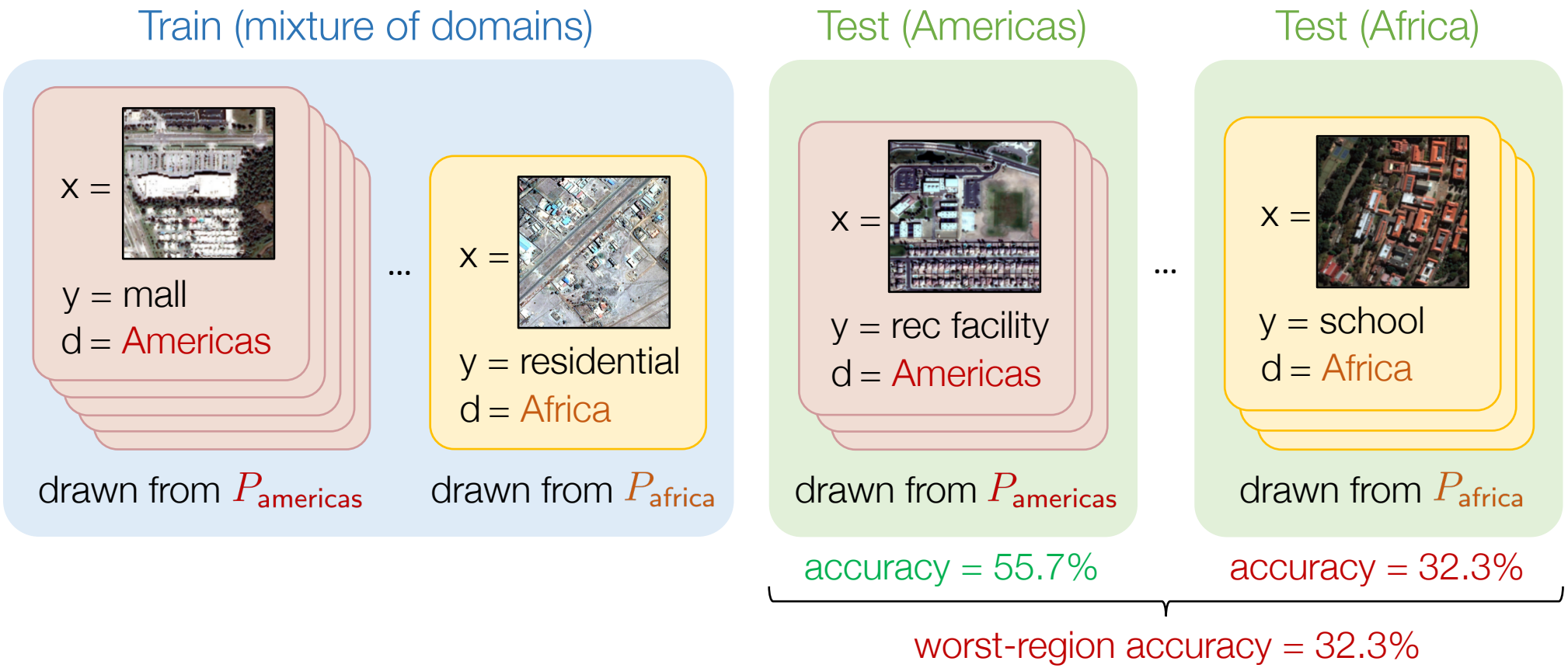
Shift to unseen cameras in animal classification for wildlife conservation



domain generalization: the goal is to generalize to unseen domains

Beery et al., 2020

Shift across regions in land use classification on satellite imagery



subpopulation shift: the goal is to perform well on many subpopulations of the training distribution

Christie et al., 2018

Existing datasets don’t focus on real-world shifts

synthetic perturbations

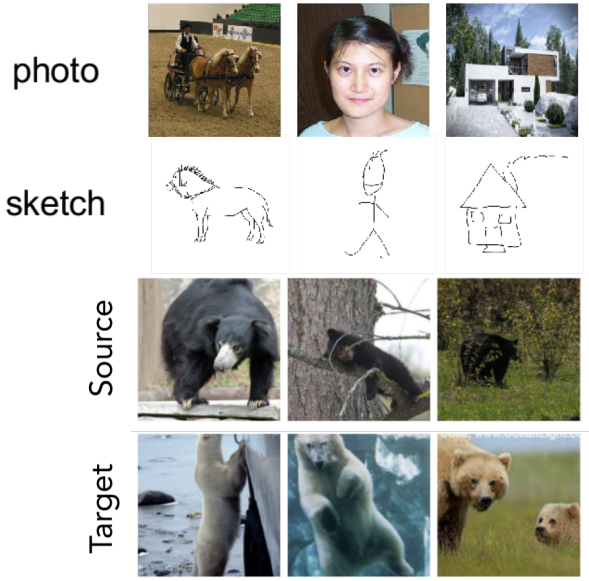
Colored MNIST
(Kim et al., 2018)

ImageNet-C
(Hendrycks et al., 2019)

Waterbirds
(Sagawa et al., 2020)

+ rotated MNIST and CIFAR-10
  Stylized ImageNet (Geirhos et al., 2018)
  the Backgrounds Challenge (Xiao et al., 2020)
  …

disparate data splits



PACS
(Li et al., 2017)

BREEDS
(Santurkar et al., 2020)

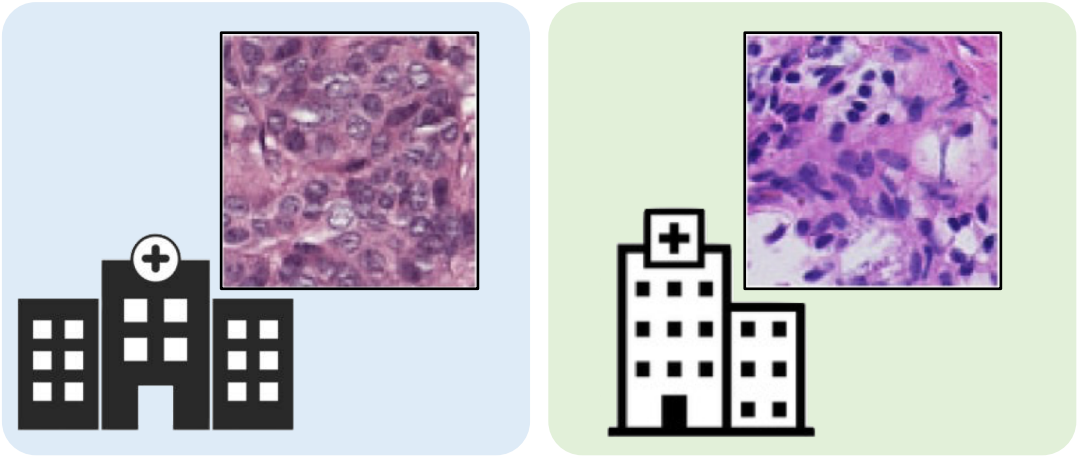
+ ObjectNet (Barbu et al., 2019)
  NICO (He et al., 2020)
  DeepFashion-Remixed (Hendrycks et al., 2020)
  …

WILDS: A benchmark for robustness to distribution shifts

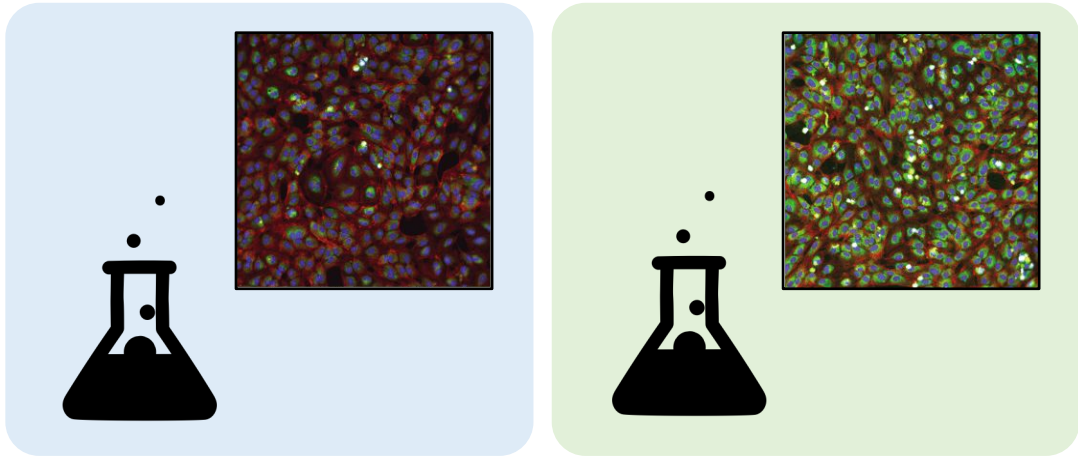
A suite of 10 datasets with…

- Real-world distribution shifts
- Diverse applications

Talking to domain experts → lots of real-world distribution shifts!

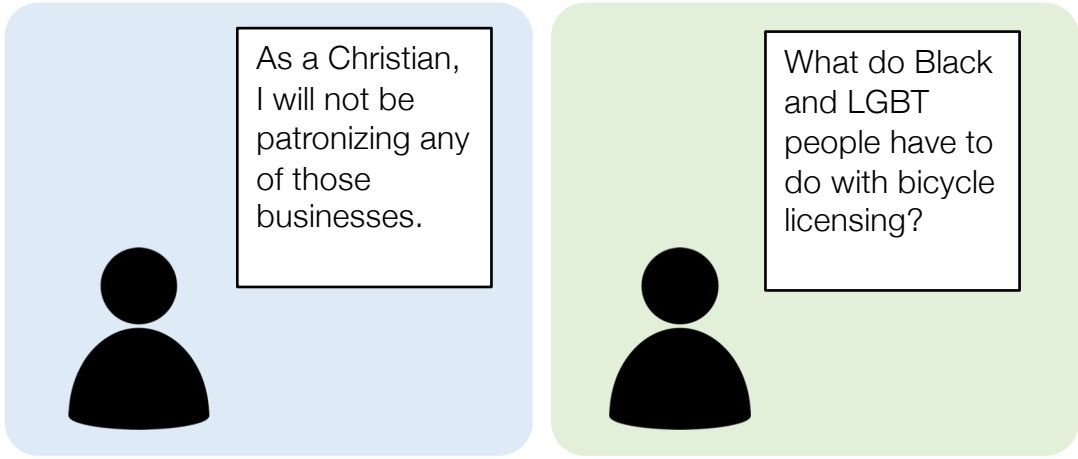


shifts across hospitals in histopathology



shifts across batches in cell imaging experiments

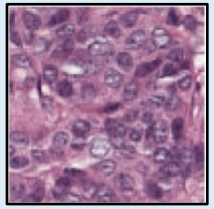
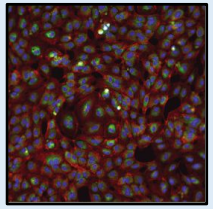
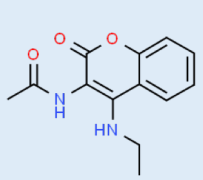
shifts across regions in wheat head detection

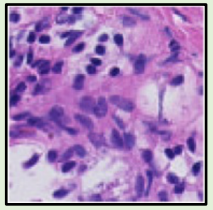
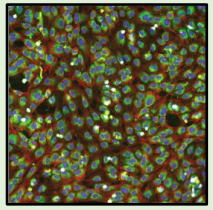
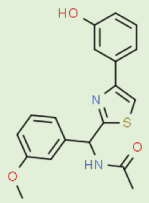


shifts across demographics in toxic comment detection

WILDS: A benchmark of in-the-wild distribution shifts

	Domain generalization					Subpopulation shift	Domain generalization + subpopulation shift			
Dataset	iWildCam	Camelyon17	RxRx1	OGB-MolPCBA	GlobalWheat	CivilComments	FMoW	PovertyMap	Amazon	Py150
Input (x)	photo	tissue slide	cell image	molecular graph	wheat image	online comment	satellite image	satellite image	product review	code
Prediction (y)	animal species	tumor	perturbed gene	bioassays	wheat head bbox	toxicity	land use	asset wealth	sentiment	autocomplete
Domain (d)	camera	hospital	batch	scaffold	location, time	demographic	time, region	location	user	git repository

Train 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.___`

Test 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.___`

Adapted from	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Hu et al. 2020	David et al. 2021	Borkan et al. 2019	Christie et al. 2018	Yeh et al. 2020	Ni et al. 2019	Raychev et al. 2016

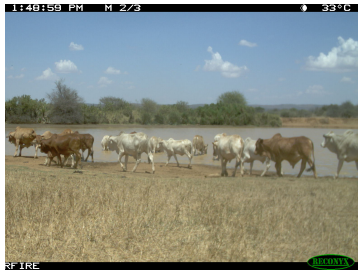
WILDS: A benchmark of in-the-wild distribution shifts

	Domain generalization					Subpopulation shift	Domain generalization + subpopulation shift			
Dataset	**iWildCam**	Camelyon17	RxRx1	OGB-MolPCBA	GlobalWheat	CivilComments	FMoW	PovertyMap	Amazon	Py150
Input (x)	photo	tissue slide	cell image	molecular graph	wheat image	online comment	satellite image	satellite image	product review	code
Prediction (y)	animal species	tumor	perturbed gene	bioassays	wheat bbox	toxicity	land use	asset wealth	sentiment	autocomplete
Domain (d)	camera	hospital	batch	scaffold	location, time	demographic	time, region	location	user	git repository

Train 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.___`

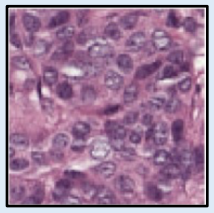
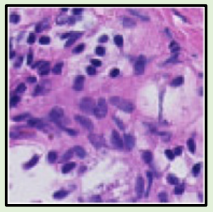
Test example						As a Christian, I will not be patronizing any of those businesses.			I *loved* my French press, it's so perfect and came with all this fun stuff!	`import subprocess as sp` `p=sp.Popen() stdout=p.___`
Adapted from	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Hu et al. 2020	David et al. 2021	Borkan et al. 2019	Christie et al. 2018	Yeh et al. 2020	Ni et al. 2019	Raychev et al. 2016

iWildCam: shifts across cameras in animal classification

Train			Test (OOD)	
d = Location 1	d = Location 2	d = Location 245	d = Location 246	
				
Vulturine Guineafowl	African Bush Elephant	unknown	Wild Horse	…
				
Cow	Cow	Southern Pig-Tailed Macaque	Great Curassow	

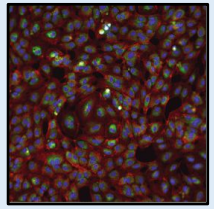
…

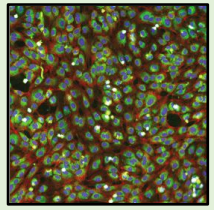
WILDS: A benchmark of in-the-wild distribution shifts

	Domain generalization					Subpopulation shift	Domain generalization + subpopulation shift			
Dataset	iWildCam	Camelyon17	RxRx1	OGB-MolPCBA	GlobalWheat	CivilComments	FMoW	PovertyMap	Amazon	Py150
Input (x)	photo	tissue slide	cell image	molecular graph	wheat image	online comment	satellite image	satellite image	product review	code
Prediction (y)	animal species	tumor	perturbed gene	bioassays	wheat bbox	toxicity	land use	asset wealth	sentiment	autocomplete
Domain (d)	camera	hospital	batch	scaffold	location, time	demographic	time, region	location	user	git repository
Train 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.___`
Test 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.___`
Adapted from	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Hu et al. 2020	David et al. 2021	Borkan et al. 2019	Christie et al. 2018	Yeh et al. 2020	Ni et al. 2019	Raychev et al. 2016

WILDS: A benchmark of in-the-wild distribution shifts

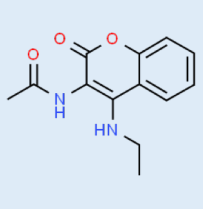
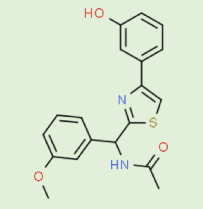
	Domain generalization					Subpopulation shift	Domain generalization + subpopulation shift			
Dataset	iWildCam	Camelyon17	RxRx1	OGB-MolPCBA	GlobalWheat	CivilComments	FMoW	PovertyMap	Amazon	Py150
Input (x)	photo	tissue slide	cell image	molecular graph	wheat image	online comment	satellite image	satellite image	product review	code
Prediction (y)	animal species	tumor	perturbed gene	bioassays	wheat bbox	toxicity	land use	asset wealth	sentiment	autocomplete
Domain (d)	camera	hospital	batch	scaffold	location, time	demographic	time, region	location	user	git repository

Train 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.___`

Test 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.___`

Adapted from	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Hu et al. 2020	David et al. 2021	Borkan et al. 2019	Christie et al. 2018	Yeh et al. 2020	Ni et al. 2019	Raychev et al. 2016

WILDS: A benchmark of in-the-wild distribution shifts

	Domain generalization				Subpopulation shift	Domain generalization + subpopulation shift				
Dataset	iWildCam	Camelyon17	RxRx1	OGB-MolPCBA	GlobalWheat	CivilComments	FMoW	PovertyMap	Amazon	Py150
Input (x)	photo	tissue slide	cell image	molecular graph	wheat image	online comment	satellite image	satellite image	product review	code
Prediction (y)	animal species	tumor	perturbed gene	bioassays	wheat bbox	toxicity	land use	asset wealth	sentiment	autocomplete
Domain (d)	camera	hospital	batch	scaffold	location, time	demographic	time, region	location	user	git repository
Train 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.___`
Test 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.___`
Adapted from	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Hu et al. 2020	David et al. 2021	Borkan et al. 2019	Christie et al. 2018	Yeh et al. 2020	Ni et al. 2019	Raychev et al. 2016

WILDS: A benchmark of in-the-wild distribution shifts

	Domain generalization				Subpopulation shift	Domain generalization + subpopulation shift				
Dataset	iWildCam	Camelyon17	RxRx1	OGB-MolPCBA	GlobalWheat	CivilComments	FMoW	PovertyMap	Amazon	Py150
Input (x)	photo	tissue slide	cell image	molecular graph	wheat image	online comment	satellite image	satellite image	product review	code
Prediction (y)	animal species	tumor	perturbed gene	bioassays	wheat bbox	toxicity	land use	asset wealth	sentiment	autocomplete
Domain (d)	camera	hospital	batch	scaffold	location, time	demographic	time, region	location	user	git repository

Train 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.___`

Test 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.___`
Adapted from	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Hu et al. 2020	David et al. 2021	Borkan et al. 2019	Christie et al. 2018	Yeh et al. 2020	Ni et al. 2019	Raychev et al. 2016

WILDS: A benchmark of in-the-wild distribution shifts

	Domain generalization					Subpopulation shift	Domain generalization + subpopulation shift			
Dataset	iWildCam	Camelyon17	RxRx1	OGB-MolPCBA	GlobalWheat	CivilComments	FMoW	PovertyMap	Amazon	Py150
Input (x)	photo	tissue slide	cell image	molecular graph	wheat image	online comment	satellite image	satellite image	product review	code
Prediction (y)	animal species	tumor	perturbed gene	bioassays	wheat bbox	toxicity	land use	asset wealth	sentiment	autocomplete
Domain (d)	camera	hospital	batch	scaffold	location, time	demographic	time, region	location	user	git repository

Train 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.___`

Test 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.___`
Adapted from	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Hu et al. 2020	David et al. 2021	Borkan et al. 2019	Christie et al. 2018	Yeh et al. 2020	Ni et al. 2019	Raychev et al. 2016

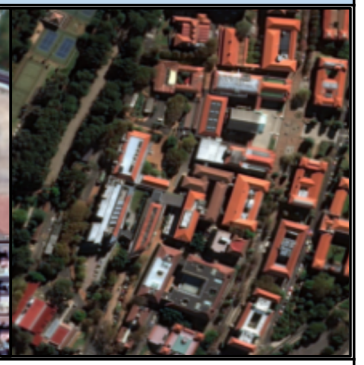
WILDS: A benchmark of in-the-wild distribution shifts

	Domain generalization					Subpopulation shift	Domain generalization + subpopulation shift			
Dataset	iWildCam	Camelyon17	RxRx1	OGB-MolPCBA	GlobalWheat	CivilComments	FMoW	PovertyMap	Amazon	Py150
Input (x)	photo	tissue slide	cell image	molecular graph	wheat image	online comment	satellite image	satellite image	product review	code
Prediction (y)	animal species	tumor	perturbed gene	bioassays	wheat bbox	toxicity	land use	asset wealth	sentiment	autocomplete
Domain (d)	camera	hospital	batch	scaffold	location, time	demographic	time, region	location	user	git repository

Train 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.___`

Test 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.___`
Adapted from	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Hu et al. 2020	David et al. 2021	Borkan et al. 2019	Christie et al. 2018	Yeh et al. 2020	Ni et al. 2019	Raychev et al. 2016

FMoW: hybrid shift across time and region

	Train			Test	
Satellite Image (x)					
Year / Region (d)	2002 / Americas	2009 / Africa	2012 / Europe	2016 / Americas	2017 / Africa
Building / Land Type (y)	shopping mall	multi-unit residential	road bridge	recreational facility	educational institution

WILDS: A benchmark of in-the-wild distribution shifts

	Domain generalization					Subpopulation shift	Domain generalization + subpopulation shift			
Dataset	iWildCam	Camelyon17	RxRx1	OGB-MolPCBA	GlobalWheat	CivilComments	FMoW	PovertyMap	Amazon	Py150
Input (x)	photo	tissue slide	cell image	molecular graph	wheat image	online comment	satellite image	satellite image	product review	code
Prediction (y)	animal species	tumor	perturbed gene	bioassays	wheat bbox	toxicity	land use	asset wealth	sentiment	autocomplete
Domain (d)	camera	hospital	batch	scaffold	location, time	demographic	time, region	location	user	git repository
Train 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.___`
Test 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.___`
Adapted from	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Hu et al. 2020	David et al. 2021	Borkan et al. 2019	Christie et al. 2018	Yeh et al. 2020	Ni et al. 2019	Raychev et al. 2016

WILDS: A benchmark of in-the-wild distribution shifts

	Domain generalization					Subpopulation shift	Domain generalization + subpopulation shift			
Dataset	iWildCam	Camelyon17	RxRx1	OGB-MolPCBA	GlobalWheat	CivilComments	FMoW	PovertyMap	Amazon	Py150
Input (x)	photo	tissue slide	cell image	molecular graph	wheat image	online comment	satellite image	satellite image	product review	code
Prediction (y)	animal species	tumor	perturbed gene	bioassays	wheat bbox	toxicity	land use	asset wealth	sentiment	autocomplete
Domain (d)	camera	hospital	batch	scaffold	location, time	demographic	time, region	location	user	git repository
Train 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.___`
Test 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.___`
Adapted from	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Hu et al. 2020	David et al. 2021	Borkan et al. 2019	Christie et al. 2018	Yeh et al. 2020	Ni et al. 2019	Raychev et al. 2016

WILDS: A benchmark of in-the-wild distribution shifts

	Domain generalization					Subpopulation shift	Domain generalization + subpopulation shift			
Dataset	iWildCam	Camelyon17	RxRx1	OGB-MolPCBA	GlobalWheat	CivilComments	FMoW	PovertyMap	Amazon	Py150
Input (x)	photo	tissue slide	cell image	molecular graph	wheat image	online comment	satellite image	satellite image	product review	code
Prediction (y)	animal species	tumor	perturbed gene	bioassays	wheat bbox	toxicity	land use	asset wealth	sentiment	autocomplete
Domain (d)	camera	hospital	batch	scaffold	location, time	demographic	time, region	location	user	git repository

Train 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.___`

Test 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.___`
Adapted from	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Hu et al. 2020	David et al. 2021	Borkan et al. 2019	Christie et al. 2018	Yeh et al. 2020	Ni et al. 2019	Raychev et al. 2016

Criteria for selecting datasets

Real-world distribution shifts

Criteria for selecting datasets

Real-world distribution shifts

Potential leverage

- Training data consists of multiple domains
- All points annotated with domain and other metadata

Criteria for selecting datasets

- Real-world distribution shifts
- Potential leverage
- Large performance drops

Large gaps between ID and OOD performance

- Evaluated standard models (e.g., ResNet) trained using ERM on metrics chosen for each application
- Out-of-distribution: WILDS default splits
- In-distribution: performance without distribution shift (on held-out set)

Large gaps between ID and OOD performance

- Evaluated standard models (e.g., ResNet) trained using ERM on metrics chosen for each application
- Out-of-distribution: WILDS default splits
- In-distribution: performance without distribution shift (on held-out set)

Dataset	Metric	In-distribution	Out-of-distribution	Gap
iWildCam2020-wilds	Macro F1	47.0	31.0	16.0
Camelyon17-wilds	Average accuracy	93.2	70.3	22.9
RxRx1-wilds	Average accuracy	39.8	29.9	9.9
OGB-MolPCBA	Average AP	34.4	27.2	7.2
GlobalWheat-wilds	Average domain accuracy	64.8	48.4	16.4
CivilComments-wilds	Worst-group accuracy	92.2	56.0	36.2
FMoW-wilds	Worst-region accuracy	48.6	32.3	16.3
PovertyMap-wilds	Worst-U/R Pearson R	0.60	0.45	0.15
Amazon-wilds	10th percentile accuracy	71.9	53.8	18.1
Py150-wilds	Method/class accuracy	75.4	67.9	7.5

large gaps!

Existing algorithms do not close ID-OOD gaps

- Benchmarked representative algorithms for domain generalization and subpopulation shifts
  - Domain generalization: CORAL (Sun and Saenko, 2016), IRM (Arjovsky et al., 2019)
  - Subpopulation shift: Group DRO (Sagawa et al., 2020)

Dataset	Setting	ERM	CORAL	IRM	Group DRO
iWildCam2020-wilds	Domain gen.	31.0 (1.3)	**32.8 (0.1)**	15.1 (4.9)	23.9 (2.1)
Camelyon17-wilds	Domain gen.	**70.3 (6.4)**	59.5 (7.7)	64.2 (8.1)	68.4 (7.3)
RxRx1-wilds	Domain gen.	**29.9 (0.4)**	28.4 (0.3)	8.2 (1.1)	23.0 (0.3)
OGB-MolPCBA	Domain gen.	**27.2 (0.3)**	17.9 (0.5)	15.6 (0.3)	22.4 (0.6)
GlobalWheat-wilds	Domain gen.	**49.2 (1.5)**	—	—	46.1 (1.6)

Existing algorithms do not close ID-OOD gaps

- Benchmarked representative algorithms for domain generalization and subpopulation shifts
  - Domain generalization: CORAL (Sun and Saenko, 2016), IRM (Arjovsky et al., 2019)
  - Subpopulation shift: Group DRO (Sagawa et al., 2020)

Dataset	Setting	ERM	CORAL	IRM	Group DRO
iWildCam2020-wilds	Domain gen.	31.0 (1.3)	**32.8 (0.1)**	15.1 (4.9)	23.9 (2.1)
Camelyon17-wilds	Domain gen.	**70.3 (6.4)**	59.5 (7.7)	64.2 (8.1)	68.4 (7.3)
RxRx1-wilds	Domain gen.	**29.9 (0.4)**	28.4 (0.3)	8.2 (1.1)	23.0 (0.3)
OGB-MolPCBA	Domain gen.	**27.2 (0.3)**	17.9 (0.5)	15.6 (0.3)	22.4 (0.6)
GlobalWheat-wilds	Domain gen.	**49.2 (1.5)**	—	—	46.1 (1.6)
CivilComments-wilds	Subpop. shift	56.0 (3.6)	65.6 (1.3)	66.3 (2.1)	**70.0 (2.0)**

Existing algorithms do not close ID-OOD gaps

- Benchmarked representative algorithms for domain generalization and subpopulation shifts
  - Domain generalization: CORAL (Sun and Saenko, 2016), IRM (Arjovsky et al., 2019)
  - Subpopulation shift: Group DRO (Sagawa et al., 2020)

Dataset	Setting	ERM	CORAL	IRM	Group DRO
iWildCam2020-wilds	Domain gen.	31.0 (1.3)	**32.8 (0.1)**	15.1 (4.9)	23.9 (2.1)
Camelyon17-wilds	Domain gen.	**70.3 (6.4)**	59.5 (7.7)	64.2 (8.1)	68.4 (7.3)
RxRx1-wilds	Domain gen.	**29.9 (0.4)**	28.4 (0.3)	8.2 (1.1)	23.0 (0.3)
OGB-MolPCBA	Domain gen.	**27.2 (0.3)**	17.9 (0.5)	15.6 (0.3)	22.4 (0.6)
GlobalWheat-wilds	Domain gen.	**49.2 (1.5)**	—	—	46.1 (1.6)
CivilComments-wilds	Subpop. shift	56.0 (3.6)	65.6 (1.3)	66.3 (2.1)	**70.0 (2.0)**
FMoW-wilds	Hybrid	**32.3 (1.3)**	31.7 (1.2)	30.0 (1.4)	30.8 (0.8)
PovertyMap-wilds	Hybrid	**0.45 (0.06)**	0.44 (0.06)	0.43 (0.07)	0.39 (0.06)
Amazon-wilds	Hybrid	**53.8 (0.8)**	52.9 (0.8)	52.4 (0.8)	53.3 (0.0)
Py150-wilds	Hybrid	**67.9 (0.1)**	65.9 (0.1)	64.3 (0.2)	65.9 (0.1)

No improvements over ERM!

These real-world shifts are still an open problem

WILDS leaderboard

wilds.stanford.edu

WILDS

GET STARTED DATASETS LEADERBOARD TEAM PAPER GITHUB

Algorithm	Contact	FMoW	PovertyMap	iWildCam	Camelyon17	OGB-MolPCBA	Amazon	CivilComments	Py150
		Worst-Region Acc	Rural Pearson r	Macro F1	Avg Acc	Avg Precision	10% Acc	Worst-Group Acc	Method/Class Acc
ERM	WILDS	32.8 (0.45)	0.46 (0.07)	31.0 (1.3)	70.3 (6.4)	27.2 (0.3)	53.8 (0.8)	56.0 (3.6)	67.9 (0.1)
CORAL	WILDS	31.0 (0.35)	0.44 (0.07)	32.8 (0.1)	59.5 (7.7)	17.9 (0.5)	52.9 (0.8)	65.6 (1.3)	65.9 (0.1)
IRM	WILDS	33.5 (1.35)	0.48 (0.04)	15.1 (4.9)	64.2 (8.1)	15.6 (0.3)	52.4 (0.8)	66.3 (2.1)	64.3 (0.2)
Group DRO	WILDS	31.4 (2.1)	0.4 (0.08)	23.9 (2.1)	68.4 (7.3)	22.4 (0.6)	53.3 (0.0)	70.0 (2.0)	65.9 (0.1)

Not just for “distribution shift researchers”:
Distribution shifts are unavoidable in many ML applications

WILDS package (`pip install wilds`)

Standardized datasets and data loaders

Defaults and baselines

Easy evaluation

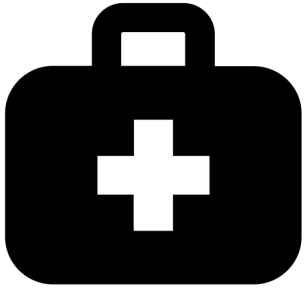
```
>>> from wilds.datasets.iwildcam_dataset import IWildCamDataset
>>> from wilds.common.data_loaders import get_train_loader

>>> dataset = get_dataset(dataset="iwildcam", download=True)
>>> train_data = dataset.get_subset("train")
>>> train_loader = get_train_loader("standard", train_data,
...                                 batch_size=16)

>>> for x, y_true, metadata in train_loader:
...     [Train a model using your algorithm; we provide defaults]

>>> dataset.eval(y_pred, y_true, metadata)
{'macro_recall': 0.66, ...}
```

Other distribution shifts beyond WILDS

- Many other real-world shifts, but challenging to find suitable datasets

Demographic shifts in automatic speech recognition (ASR)

Difficulty finding training data on natural speech with enough demographic diversity



Time and hospital shifts in medicine

Datasets with those shifts also involve *concept drifts* due to changes in label definition, clinical procedures, etc.

Koenecke et al., 2020, Nestor et al., 2019

Other distribution shifts beyond WILDS

- Other datasets had no substantial ID-OOD gap

Day vs. night shift in autonomous driving (BDD100K)

No substantial performance drop if training set is sufficiently diverse

Demographic fairness in weapon possession prediction (SQF)

Substantial disparities across races, but due to biased data instead of the distribution shift

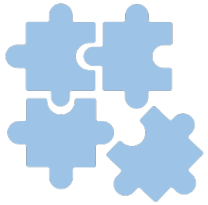
Surveys: fairness, healthcare, genomics, speech, NLP, robotics, education
Additional examples: cell type shift in genomics, shifts in review datasets

Yu et al., 2020, Goel et al., 2016

Many open questions

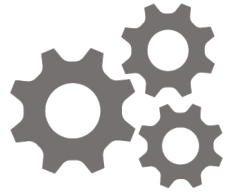
Datasets and shifts

- How can we construct benchmarks for the many applications and shifts for which we don't have suitable datasets?



Theoretical frameworks

- How can we characterize and categorize all of these different shifts?



Algorithms

- How can we train models that are robust due to real-world shifts? (e.g., by incorporating domain annotations and metadata, or using prior knowledge?)

A Benchmark of in-the-Wild Distribution Shifts

Code, paper, leaderboard, and contact info at
https://wilds.stanford.edu

Acknowledgments

Based on datasets from…

P. Bandi, O. Geessink, Q. Manson, M. V. Dijk, M. Balkenhol, M. Hermsen, B. E. Bejnordi, B. Lee, K. Paeng, A. Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 2018.

S. Beery, E. Cole, and A. Gjoka. The iWildCam 2020 competition dataset. *arXiv*, 2020.

D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. *WWW*, 2019.

G. Christie, N. Fendley, J. Wilson, and R. Mukherjee. Functional map of the world. *CVPR*, 2018.

E. David, S. Madec, P. Sadeghi-Tehran, H. Aasen, B. Zheng, S. Liu, N. Kirchgessner, G. Ishikawa, K. Nagasawa, M.A. Badhon, and C. Pozniak. Global Wheat Head Detection (GWHD) dataset: a large and diverse dataset of high-resolution RGB-labelled images to develop and benchmark wheat head detection methods. *Plant Phenomics*, *2020*.

W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec. Open Graph Benchmark: Datasets for Machine Learning on Graphs. *NeurIPS,* 2020.

J. Ni, J. Li, and J. McAuley. Justifying recommendations using distantly-labeled reviews and fine- grained aspects. *EMNLP*, 2019.

C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, 2020.

V. Raychev, P. Bielik and M. Vechev. Probabilistic Model for Code with Decision Trees. *OOPSLA*, 2016.

J. Taylor, B. Earnshaw, B. Mabey, M. Victors, and J. Yosinski. RxRx1: An Image Set for Cellular Morphological Variation Across Many Experimental Batches. *ICLR,* 2019.

Thanks to…

Aditya Khosla, Andreas Schlueter, Annie Chen, Aleksander Madry, Alexander D'Amour, Allison Koenecke, Alyssa Lees, Ananya Kumar, Andrew Beck, Behzad Haghgoo, Charles Sutton, Christopher Yeh, Cody Coleman, Dan Hendrycks, Dan Jurafsky, Daniel Levy, Daphne Koller, David Tellez, Erik Jones, Evan Liu, Fisher Yu, Georgi Marinov, Hongseok Namkoong, Irene Chen, Jacky Kang, Jacob Schreiber, Jacob Steinhardt, Jared Dunnmon, Jean Feng, Jeffrey Sorensen, Jianmo Ni, John Hewitt, John Miller, Kate Saenko, Kelly Cochran, Kensen Shi, Kyle Loh, Li Ji ang, Lucy Vasserman, Ludwig Schmidt, Luke Oakden-Rayner, Marco Tulio Ribeiro, Matthew Lungren, Megha Srivastava, Nelson Liu, Nimit Sohoni, Pranav Rajpurkar, Robin Jia, Rohan Taori, Sarah Bird, Sharad Goel, Sherrie Wang, Shyamal Buch, Stefano Ermon, Steve Yadlowsky, Tatsunori Hashimoto, Tengyu Ma, Vincent Hellendoorn, Yair Carmon, Zachary Lipton, and Zhenghao Chen.

The design of WILDS was inspired by the Open Graph Benchmark (Hu et al., 2020).