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

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

Train (mixture of domains)

\[ x = \text{monkey} \]
\[ y = \text{monkey} \]
\[ d = \text{camera 1} \]
\[ \text{drawn from } P_{\text{cam1}} \]

Test (unseen domains)

\[ x = \text{elephant} \]
\[ y = \text{elephant} \]
\[ d = \text{camera 245} \]
\[ \text{drawn from } P_{\text{cam245}} \]

\[ x = \text{curassow} \]
\[ y = \text{curassow} \]
\[ d = \text{camera 246} \]
\[ \text{drawn from } P_{\text{cam246}} \]

\[ x = \text{cow} \]
\[ y = \text{cow} \]
\[ d = \text{camera 324} \]
\[ \text{drawn from } P_{\text{cam324}} \]

\[ \text{macro F1 } = 47.0\% \]

\[ \text{macro F1 } = 31.0\% \]

**domain generalization**: the goal is to generalize to unseen domains

Beery et al., 2020
Shift across regions in land use classification on satellite imagery

\[ y = \text{mall} \quad x = \text{d} = \text{Americas} \]

drawn from \( P_{\text{Americas}} \)

\[ y = \text{residential} \quad x = \text{d} = \text{Africa} \]

drawn from \( P_{\text{Africa}} \)

\[ y = \text{rec facility} \quad x = \text{d} = \text{Americas} \]

drawn from \( P_{\text{Americas}} \)

\[ y = \text{school} \quad x = \text{d} = \text{Africa} \]

drawn from \( P_{\text{Africa}} \)

Test (Americas) accuracy = 55.7%

Test (Africa) accuracy = 32.3%

worst-region accuracy = 32.3%

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**

- photo
  - PACS (Li et al., 2017)
- sketch
  - BREEDS (Santurkar et al., 2020)
- Source
- Target
  + 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

As a Christian, I will not be patronizing any of those businesses.

What do Black and LGBT people have to do with bicycle licensing?
WILDS: A benchmark of in-the-wild distribution shifts

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

Train example

- What do Black and LGBT people have to do with bicycle licensing?
- 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!

Test example

- Overall a solid package that has a good quality of construction for the price.
- "import numpy as np
... 
  norm=np.___

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**

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

**Train example**

![Train example images](image1.png)

**Test example**

![Test example images](image2.png)

**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

**Subpopulation shift**

- What do Black and LGBT people have to do with bicycle licensing?
- 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.

**Domain generalization + subpopulation shift**

- Overall a solid package that has a good quality of construction for the price.
- I import numpy as np
  ...
  norm=np.___
  ...

---

```python
import numpy as np
...
norm=np.____
```
### iWildCam: shifts across cameras in animal classification

<table>
<thead>
<tr>
<th>Train</th>
<th>Test (OOD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(d = \text{Location 1})</td>
<td>Vulturine Guineafowl</td>
</tr>
<tr>
<td>(d = \text{Location 2})</td>
<td>African Bush Elephant</td>
</tr>
<tr>
<td>(d = \text{Location 245})</td>
<td>unknown</td>
</tr>
<tr>
<td>(d = \text{Location 246})</td>
<td>Wild Horse</td>
</tr>
<tr>
<td>Cow</td>
<td>Cow</td>
</tr>
<tr>
<td></td>
<td>Southern Pig-Tailed Macaque</td>
</tr>
<tr>
<td></td>
<td>Great Curassow</td>
</tr>
</tbody>
</table>
## WILDS: A benchmark of in-the-wild distribution shifts

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

### Train example
- [Image of train example with various inputs and outputs]
- **Question:** What do Black and LGBT people have to do with bicycle licensing?

### Test example
- [Image of test example with various inputs and outputs]
- **Question:** As a Christian, I will not be patronizing any of those businesses.

### Adapting 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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>iWildCam</th>
<th>Camelyon17</th>
<th>RXRx1</th>
<th>OGB-MoPCBA</th>
<th>GlobalWheat</th>
<th>CivilComments</th>
<th>FMoW</th>
<th>PovertyMap</th>
<th>Amazon</th>
<th>Py150</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input (x)</td>
<td>photo</td>
<td>tissue slide</td>
<td>cell image</td>
<td>molecular graph</td>
<td>wheat image</td>
<td>online comment</td>
<td>satellite image</td>
<td>satellite image</td>
<td>product review</td>
<td>code</td>
</tr>
<tr>
<td>Prediction (y)</td>
<td>animal species</td>
<td>tumor</td>
<td>perturbed gene</td>
<td>bioassays</td>
<td>wheat bbox</td>
<td>toxicity</td>
<td>land use</td>
<td>asset wealth</td>
<td>sentiment</td>
<td>autocomplete</td>
</tr>
<tr>
<td>Domain (d)</td>
<td>camera</td>
<td>hospital</td>
<td>batch</td>
<td>scaffold</td>
<td>location, time</td>
<td>demographic</td>
<td>time, region</td>
<td>location</td>
<td>user</td>
<td>git repository</td>
</tr>
</tbody>
</table>

### Train example
- [Image]
- [Image]
- [Image]
- [Image]
- [Image]
- [Image]
- [Image]
- [Image]
- [Image]

### Test example
- [Image]
- [Image]
- [Image]
- [Image]
- [Image]
- [Image]
- [Image]
- [Image]
- [Image]

### 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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>iWildCam</th>
<th>Camelyon17</th>
<th>RxRx1</th>
<th>OGB-MolPCBA</th>
</tr>
</thead>
</table>

#### Subpopulation shift

<table>
<thead>
<tr>
<th>Domain generalization + subpopulation shift</th>
<th>GlobalWheat</th>
<th>CivilComments</th>
<th>FMoW</th>
<th>PovertyMap</th>
<th>Amazon</th>
<th>Py150</th>
</tr>
</thead>
</table>

#### Input (x)

<table>
<thead>
<tr>
<th>Prediction (y)</th>
<th>animal species</th>
<th>tumor</th>
<th>perturbed gene</th>
<th>bioassays</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain (d)</td>
<td>camera</td>
<td>hospital</td>
<td>batch</td>
<td>scaffold</td>
</tr>
<tr>
<td></td>
<td>photo</td>
<td>tissue slide</td>
<td>cell image</td>
<td>molecular graph</td>
</tr>
<tr>
<td></td>
<td>wheat image</td>
<td>online comment</td>
<td>satellite image</td>
<td>satellite image</td>
</tr>
<tr>
<td></td>
<td>wheat bbox</td>
<td>toxicity</td>
<td>land use</td>
<td>asset wealth</td>
</tr>
<tr>
<td></td>
<td>location</td>
<td>time</td>
<td>demographic</td>
<td>time, region</td>
</tr>
<tr>
<td></td>
<td>location</td>
<td>time</td>
<td>user</td>
<td>git repository</td>
</tr>
</tbody>
</table>

#### Train example

- [Image of tissue slides]
- [Image of a chemical structure]
- [Image of a satellite image]
- [Image of a code snippet]

#### Test example

- [Image of tissue slides]
- [Image of a chemical structure]
- [Image of a satellite image]
- [Image of a code snippet]

#### 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

## Dataset
<table>
<thead>
<tr>
<th>Domain generalization</th>
<th>Subpopulation shift</th>
<th>Domain generalization + subpopulation shift</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>iWildCam</td>
<td>Camelyon17</td>
<td>RxRx1</td>
</tr>
<tr>
<td>OGB-MolPCBA</td>
<td>GlobalWheat</td>
<td>CivilComments</td>
</tr>
<tr>
<td>FMoW</td>
<td>PovertyMap</td>
<td>Amazon</td>
</tr>
<tr>
<td>Py150</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Input (x)
- photo
- tissue slide
- cell image
- molecular graph
- wheat image

## 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
- Tissue slide
- Satellite image
- Online comment
- Molecular graph

## Test example
- Tissue slide
- Satellite image
- Online comment
- Molecular graph

## 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

---

```python
import numpy as np
norm=np.___

Overall a solid package that has a good quality of construction for the price.
```

---

```python
import subprocess as sp
p=sp.Popen() stdout=p.___

As a Christian, I will not be patronizing any of those businesses.
```

---

```python
import numpy as np
...

What do Black and LGBT people have to do with bicycle licensing?
```

---

```python
import numpy as np
norm=np.___

As a Christian, I will not be patronizing any of those businesses.
```

---

```python
import subprocess as sp
p=sp.Popen() stdout=p.___

I *loved* my French press, it’s so perfect and came with all this fun stuff!
```
## WILDS: A benchmark of in-the-wild distribution shifts

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

### Train example
- What do Black and LGBT people have to do with bicycle licensing?
- As a Christian, I will not be patronizing any of those businesses.

### Test example
- Overall a solid package that has a good quality of construction for the price.
- I *loved* my French press, it's so perfect and came with all this fun stuff!

### 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

### Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>iWildCam</th>
<th>Camelyon17</th>
<th>RxRx1</th>
<th>OGB-MolPCBA</th>
<th>GlobalWheat</th>
<th>CivilComments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input (x)</td>
<td>photo</td>
<td>tissue slide</td>
<td>cell image</td>
<td>molecular graph</td>
<td>wheat image</td>
<td>online comment</td>
</tr>
<tr>
<td>Prediction (y)</td>
<td>animal species</td>
<td>tumor</td>
<td>perturbed gene</td>
<td>bioassays</td>
<td>wheat bbox</td>
<td>toxicity</td>
</tr>
<tr>
<td>Domain (d)</td>
<td>camera</td>
<td>hospital</td>
<td>batch</td>
<td>scaffold</td>
<td>location, time</td>
<td>demographic</td>
</tr>
</tbody>
</table>

### Experiment

<table>
<thead>
<tr>
<th>Train example</th>
<th>Test example</th>
<th>Adapated from</th>
</tr>
</thead>
</table>
| ![Train example images] | ![Test example images] | Beery et al. 2020  
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 |

### Code Snippets

```python
import numpy as np

norm=np.___
```

```
import subprocess as sp

p=sp.Popen(stdout=p.___
```

## Overall a solid package that has a good quality of construction for the price.

## I *loved* my French press, it's so perfect and came with all this fun stuff!

## As a Christian, I will not be patronizing any of those businesses.
FMoW: hybrid shift across time and region

<table>
<thead>
<tr>
<th>Satellite Image (x)</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year / Region (a)</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002 / Americas</td>
<td><img src="image4.png" alt="Image" /></td>
<td>2016 / Americas</td>
</tr>
<tr>
<td>2009 / Africa</td>
<td><img src="image5.png" alt="Image" /></td>
<td>2017 / Africa</td>
</tr>
<tr>
<td>2012 / Europe</td>
<td><img src="image6.png" alt="Image" /></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Building / Land Type (y)</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>shopping mall</td>
<td><img src="image7.png" alt="Image" /></td>
<td>recreational facility</td>
</tr>
<tr>
<td>multi-unit residential</td>
<td><img src="image8.png" alt="Image" /></td>
<td>educational institution</td>
</tr>
<tr>
<td>road bridge</td>
<td><img src="image9.png" alt="Image" /></td>
<td></td>
</tr>
</tbody>
</table>

Adapted from Christie et al. 2018
## WILDS: A benchmark of in-the-wild distribution shifts

### Domain generalization

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

### Experiment example

#### Train example

- What do Black and LGBT people have to do with bicycle licensing?
- 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!

#### Test example

- Overall a solid package that has a good quality of construction for the price.
- I *loved* my French press, it’s so perfect and came with all this fun stuff!

### 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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>iWildCam</th>
<th>Camelyon17</th>
<th>RxRx1</th>
<th>OGB-MolPCBA</th>
<th>GlobalWheat</th>
<th>CivilComments</th>
<th>FMoW</th>
<th>PovertyMap</th>
<th>Amazon</th>
<th>Py150</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input (x)</td>
<td>photo</td>
<td>tissue slide</td>
<td>cell image</td>
<td>molecular graph</td>
<td>wheat image</td>
<td>online comment</td>
<td>satellite image</td>
<td>satellite image</td>
<td>product review</td>
<td>code</td>
</tr>
<tr>
<td>Prediction (y)</td>
<td>animal species</td>
<td>tumor</td>
<td>perturbed gene</td>
<td>bioassays</td>
<td>wheat bbox</td>
<td>toxicity</td>
<td>land use</td>
<td>asset wealth</td>
<td>sentiment</td>
<td>autocomplete</td>
</tr>
<tr>
<td>Domain (d)</td>
<td>camera</td>
<td>hospital</td>
<td>batch</td>
<td>scaffold</td>
<td>location, time</td>
<td>demographic</td>
<td>time, region</td>
<td>location</td>
<td>user</td>
<td>git repository</td>
</tr>
<tr>
<td>Train example</td>
<td><img src="image" alt="Train example" /></td>
<td><img src="image" alt="Train example" /></td>
<td><img src="image" alt="Train example" /></td>
<td><img src="image" alt="Train example" /></td>
<td><img src="image" alt="Train example" /></td>
<td><img src="image" alt="Train example" /></td>
<td><img src="image" alt="Train example" /></td>
<td><img src="image" alt="Train example" /></td>
<td><img src="image" alt="Train example" /></td>
<td><img src="image" alt="Train example" /></td>
</tr>
<tr>
<td>Test example</td>
<td><img src="image" alt="Test example" /></td>
<td><img src="image" alt="Test example" /></td>
<td><img src="image" alt="Test example" /></td>
<td><img src="image" alt="Test example" /></td>
<td><img src="image" alt="Test example" /></td>
<td><img src="image" alt="Test example" /></td>
<td><img src="image" alt="Test example" /></td>
<td><img src="image" alt="Test example" /></td>
<td><img src="image" alt="Test example" /></td>
<td><img src="image" alt="Test example" /></td>
</tr>
</tbody>
</table>

**Domain generalization**

- Batch
- Scaffold
- Location, time
- Demographic

**Subpopulation shift**

- Location, time
- Batch
- Domain (d)

**Domain generalization + subpopulation shift**

- Overall a solid package that has a good quality of construction for the price.
- I "loved" my French press, it’s so perfect and came with all this fun stuff!

**Train example**

- What do Black and LGBT people have to do with bicycle licensing?

**Test example**

- As a Christian, I will not be patronizing any of those businesses.
### WILDS: A benchmark of in-the-wild distribution shifts

#### Domain generalization

<table>
<thead>
<tr>
<th>Dataset</th>
<th>iWildCam</th>
<th>Camelyon17</th>
<th>RxRx1</th>
<th>OGB-MoIPCA</th>
<th>GlobalWheat</th>
<th>CivilComments</th>
<th>FMoW</th>
<th>PovertyMap</th>
<th>Amazon</th>
<th>Py150</th>
</tr>
</thead>
</table>

#### Input (x)

<table>
<thead>
<tr>
<th></th>
<th>photo</th>
<th>tissue slide</th>
<th>cell image</th>
<th>molecular graph</th>
<th>wheat image</th>
<th>online comment</th>
<th>satellite image</th>
<th>satellite image</th>
<th>product review</th>
<th>code</th>
</tr>
</thead>
</table>

#### Prediction (y)

<table>
<thead>
<tr>
<th>prediction</th>
<th>animal species</th>
<th>tumor</th>
<th>perturbed gene</th>
<th>bioassays</th>
<th>wheat bbox</th>
<th>toxicity</th>
<th>land use</th>
<th>asset wealth</th>
<th>sentiment</th>
<th>autocomplete</th>
</tr>
</thead>
</table>

#### Domain (d)

<table>
<thead>
<tr>
<th>domain</th>
<th>camera</th>
<th>hospital</th>
<th>batch</th>
<th>scaffold</th>
<th>location, time</th>
<th>demographic</th>
<th>time, region</th>
<th>location</th>
<th>user</th>
<th>git repository</th>
</tr>
</thead>
</table>

#### Train example

- What do Black and LGBT people have to do with bicycle licensing?
- 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.

#### Test example

- Overall a solid package that has a good quality of construction for the price.

#### 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)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Metric</th>
<th>In-distribution</th>
<th>Out-of-distribution</th>
<th>Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>IWILDCAM2020-WILDS</td>
<td>Macro F1</td>
<td>47.0</td>
<td>31.0</td>
<td>16.0</td>
</tr>
<tr>
<td>CAMELYON17-WILDS</td>
<td>Average accuracy</td>
<td>93.2</td>
<td>70.3</td>
<td>22.9</td>
</tr>
<tr>
<td>RXRX1-WILDS</td>
<td>Average accuracy</td>
<td>39.8</td>
<td>29.9</td>
<td>9.9</td>
</tr>
<tr>
<td>OGB-MOLPCBA</td>
<td>Average AP</td>
<td>34.4</td>
<td>27.2</td>
<td>7.2</td>
</tr>
<tr>
<td>GLOBALWHEAT-WILDS</td>
<td>Average domain accuracy</td>
<td>64.8</td>
<td>48.4</td>
<td>16.4</td>
</tr>
<tr>
<td>CIVILCOMMENTS-WILDS</td>
<td>Worst-group accuracy</td>
<td>92.2</td>
<td>56.0</td>
<td>36.2</td>
</tr>
<tr>
<td>FMoW-WILDS</td>
<td>Worst-region accuracy</td>
<td>48.6</td>
<td>32.3</td>
<td>16.3</td>
</tr>
<tr>
<td>POVERTYMAP-WILDS</td>
<td>Worst-U/R Pearson R</td>
<td>0.60</td>
<td>0.45</td>
<td>0.15</td>
</tr>
<tr>
<td>AMAZON-WILDS</td>
<td>10th percentile accuracy</td>
<td>71.9</td>
<td>53.8</td>
<td>18.1</td>
</tr>
<tr>
<td>PY150-WILDS</td>
<td>Method/class accuracy</td>
<td>75.4</td>
<td>67.9</td>
<td>7.5</td>
</tr>
</tbody>
</table>

*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)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Setting</th>
<th>ERM</th>
<th>CORAL</th>
<th>IRM</th>
<th>Group DRO</th>
</tr>
</thead>
<tbody>
<tr>
<td>iWILDCAM2020-WILDS</td>
<td>Domain gen.</td>
<td>31.0 (1.3)</td>
<td><strong>32.8 (0.1)</strong></td>
<td>15.1 (4.9)</td>
<td>23.9 (2.1)</td>
</tr>
<tr>
<td>CAMELYON17-WILDS</td>
<td>Domain gen.</td>
<td><strong>70.3 (6.4)</strong></td>
<td>59.5 (7.7)</td>
<td>64.2 (8.1)</td>
<td>68.4 (7.3)</td>
</tr>
<tr>
<td>RXRX1-WILDS</td>
<td>Domain gen.</td>
<td><strong>29.9 (0.4)</strong></td>
<td>28.4 (0.3)</td>
<td>8.2 (1.1)</td>
<td>23.0 (0.3)</td>
</tr>
<tr>
<td>OGB-MOLPCBA</td>
<td>Domain gen.</td>
<td><strong>27.2 (0.3)</strong></td>
<td>17.9 (0.5)</td>
<td>15.6 (0.3)</td>
<td>22.4 (0.6)</td>
</tr>
<tr>
<td>GLOBALWHEAT-WILDS</td>
<td>Domain gen.</td>
<td><strong>49.2 (1.5)</strong></td>
<td>---</td>
<td>---</td>
<td>46.1 (1.6)</td>
</tr>
</tbody>
</table>
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)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Setting</th>
<th>ERM</th>
<th>CORAL</th>
<th>IRM</th>
<th>Group DRO</th>
</tr>
</thead>
<tbody>
<tr>
<td>iWILDCAM2020-WILDS</td>
<td>Domain gen.</td>
<td>31.0</td>
<td>32.8</td>
<td>15.1</td>
<td>23.9</td>
</tr>
<tr>
<td>CAMELYON17-WILDS</td>
<td>Domain gen.</td>
<td><strong>70.3</strong></td>
<td>59.5</td>
<td>64.2</td>
<td>68.4</td>
</tr>
<tr>
<td>RXRX1-WILDS</td>
<td>Domain gen.</td>
<td><strong>29.9</strong></td>
<td>28.4</td>
<td>8.2</td>
<td>23.0</td>
</tr>
<tr>
<td>OGB-MOLPCBA</td>
<td>Domain gen.</td>
<td>27.2</td>
<td>17.9</td>
<td>15.6</td>
<td>22.4</td>
</tr>
<tr>
<td>GLOBALWHEAT-WILDS</td>
<td>Domain gen.</td>
<td><strong>49.2</strong></td>
<td>—</td>
<td>—</td>
<td>46.1</td>
</tr>
<tr>
<td>CIVILCOMMENTS-WILDS</td>
<td>Subpop. shift</td>
<td>56.0</td>
<td>65.6</td>
<td>66.3</td>
<td><strong>70.0</strong></td>
</tr>
</tbody>
</table>
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)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Setting</th>
<th>ERM</th>
<th>CORAL</th>
<th>IRM</th>
<th>Group DRO</th>
</tr>
</thead>
<tbody>
<tr>
<td>iWildCam2020-WILDS</td>
<td>Domain gen.</td>
<td>31.0 (1.3)</td>
<td><strong>32.8 (0.1)</strong></td>
<td>15.1 (4.9)</td>
<td>23.9 (2.1)</td>
</tr>
<tr>
<td>CAMELYON17-WILDS</td>
<td>Domain gen.</td>
<td><strong>70.3 (6.4)</strong></td>
<td>59.5 (7.7)</td>
<td>64.2 (8.1)</td>
<td>68.4 (7.3)</td>
</tr>
<tr>
<td>RXRX1-WILDS</td>
<td>Domain gen.</td>
<td>29.9 (0.4)</td>
<td>28.4 (0.3)</td>
<td>8.2 (1.1)</td>
<td>23.0 (0.3)</td>
</tr>
<tr>
<td>OGB-MOLPCBA</td>
<td>Domain gen.</td>
<td>27.2 (0.3)</td>
<td>17.9 (0.5)</td>
<td>15.6 (0.3)</td>
<td>22.4 (0.6)</td>
</tr>
<tr>
<td>GLOBALWHEAT-WILDS</td>
<td>Domain gen.</td>
<td><strong>49.2 (1.5)</strong></td>
<td>—</td>
<td>—</td>
<td>46.1 (1.6)</td>
</tr>
<tr>
<td>CIVILCOMMENTS-WILDS</td>
<td>Subpop. shift</td>
<td>56.0 (3.6)</td>
<td>65.6 (1.3)</td>
<td>66.3 (2.1)</td>
<td><strong>70.0 (2.0)</strong></td>
</tr>
<tr>
<td>FMoW-WILDS</td>
<td>Hybrid</td>
<td>32.3 (1.3)</td>
<td>31.7 (1.2)</td>
<td>30.0 (1.4)</td>
<td>30.8 (0.8)</td>
</tr>
<tr>
<td>POVERTYMAP-WILDS</td>
<td>Hybrid</td>
<td>0.45 (0.06)</td>
<td>0.44 (0.06)</td>
<td>0.43 (0.07)</td>
<td>0.39 (0.06)</td>
</tr>
<tr>
<td>AMAZON-WILDS</td>
<td>Hybrid</td>
<td><strong>53.8 (0.8)</strong></td>
<td>52.9 (0.8)</td>
<td>52.4 (0.8)</td>
<td>53.3 (0.0)</td>
</tr>
<tr>
<td>PY150-WILDS</td>
<td>Hybrid</td>
<td>67.9 (0.1)</td>
<td>65.9 (0.1)</td>
<td>64.3 (0.2)</td>
<td>65.9 (0.1)</td>
</tr>
</tbody>
</table>

No improvements over ERM!

These real-world shifts are still an open problem
WILDS leaderboard

wilds.stanford.edu

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Contact</th>
<th>FMoW</th>
<th>PovertyMap</th>
<th>iWildCam</th>
<th>Camelyon17</th>
<th>OGB-MolPCBA</th>
<th>Amazon</th>
<th>CivilComments</th>
<th>Py150</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERM</td>
<td>WILDS</td>
<td>32.8 (0.45)</td>
<td>0.46 (0.07)</td>
<td>31.0 (1.3)</td>
<td>70.3 (6.4)</td>
<td>27.2 (0.3)</td>
<td>53.8 (0.8)</td>
<td>56.0 (3.6)</td>
<td>67.9 (0.1)</td>
</tr>
<tr>
<td>CORAL</td>
<td>WILDS</td>
<td>31.0 (0.35)</td>
<td>0.44 (0.07)</td>
<td>32.8 (0.1)</td>
<td>59.5 (7.7)</td>
<td>17.9 (0.5)</td>
<td>52.9 (0.8)</td>
<td>65.6 (1.3)</td>
<td>65.9 (0.1)</td>
</tr>
<tr>
<td>IRM</td>
<td>WILDS</td>
<td>33.5 (1.35)</td>
<td>0.48 (0.04)</td>
<td>15.1 (4.9)</td>
<td>64.2 (8.1)</td>
<td>15.6 (0.3)</td>
<td>52.4 (0.8)</td>
<td>66.3 (2.1)</td>
<td>64.3 (0.2)</td>
</tr>
<tr>
<td>Group DRO</td>
<td>WILDS</td>
<td>31.4 (2.1)</td>
<td>0.4 (0.08)</td>
<td>23.9 (2.1)</td>
<td>68.4 (7.3)</td>
<td>22.4 (0.6)</td>
<td>53.3 (0.0)</td>
<td>70.0 (2.0)</td>
<td>65.9 (0.1)</td>
</tr>
</tbody>
</table>

Not just for “distribution shift researchers”: Distribution shifts are unavoidable in many ML applications.
WILDS package (pip install wilds)

```python
>>> 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...


Thanks to...


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