# Agreement-on-the-Line: Predicting the Performance of Neural Networks under Distribution Shift

***Presented by*** **Christina Baek**
kbaek@andrew.cmu.edu

**Yiding Jiang**[1] **Aditi Raghunathan**[1] **Zico Kolter**[1,2]

[1]**Carnegie Mellon University** [2] **Bosch Center for AI**

# Problem

In practice, machines often need to perform well on distributions that are different from what it has been trained on.

Estimating **out-of-distribution (OOD)** performance is hard because labeled data is expensive.

However, unlabeled data is easier to obtain…

# Problem

Say we are given

- a collection of models $\mathscr{H} = \{h_1, h_2, \ldots, h_n\}$ trained on in-distribution (ID) data $X_{train}, y_{train} \sim \mathscr{D}_{ID}$
- *labeled* ID validation data and *unlabeled* OOD test data

$$X_{val}, y_{val} \sim \mathscr{D}_{ID} \text{ and } X_{test} \sim \mathscr{D}_{OOD}$$

**Can we predict OOD performance of models in $\mathscr{H}$ with only unlabeled data?**

# Characterizing the Shift

**Corollary 1. (Garg, et al. 2022)**

Absent assumptions on the classifier $f$, no method of estimating accuracy will work in all scenarios, i.e., for different nature of distribution shifts.

**Simple proof:**

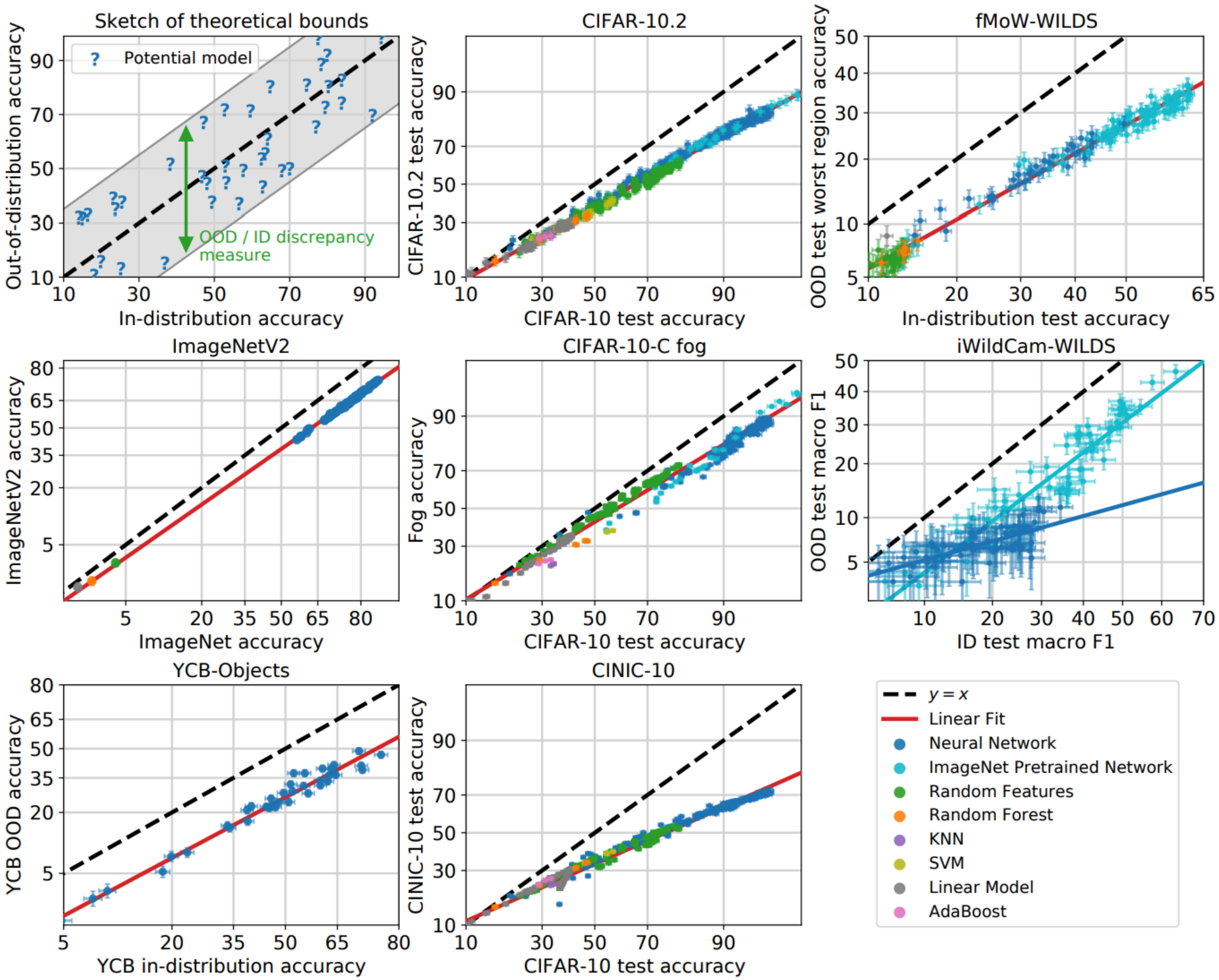
If the classifier has no assumptions, accuracy is only identifiable <u>IFF</u> $p_t(y \mid x)$ is uniquely identified given $p_s(x, y)$ and $p_t(x)$.

# Characterizing the Shift

1. What are reasonable assumptions we can make about the **distribution shift** and the **behavior of the classifier**?

2. Is there an easy way to **“check”** whether these assumptions hold?

# Accuracy on the Line (Miller, et al. 2021)



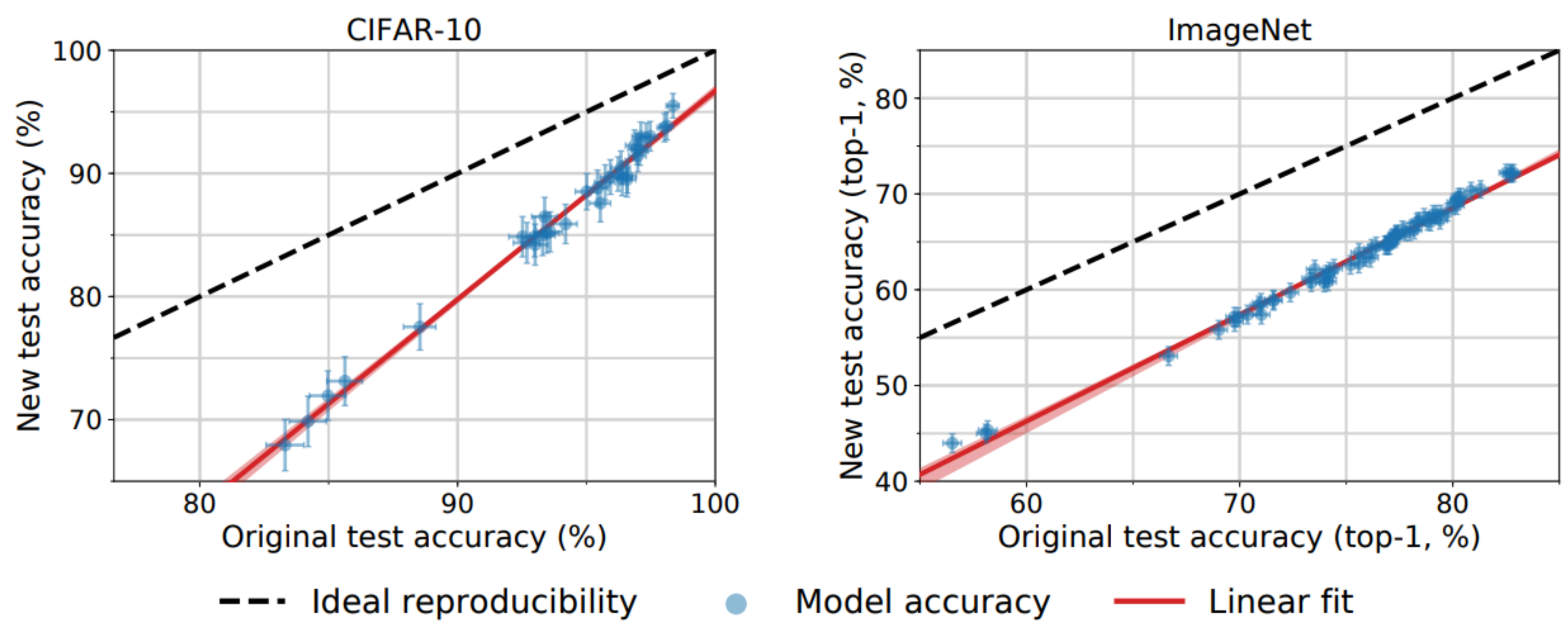

In popular OOD benchmarks, **ID** and **OOD test accuracy** are **strongly linearly correlated**

*They first scale the accuracies by probit transform $\Phi^{-1}(\cdot)$

# Accuracy on the Line (Miller, et al. 2021): OOD Benchmarks

## Dataset reproduction

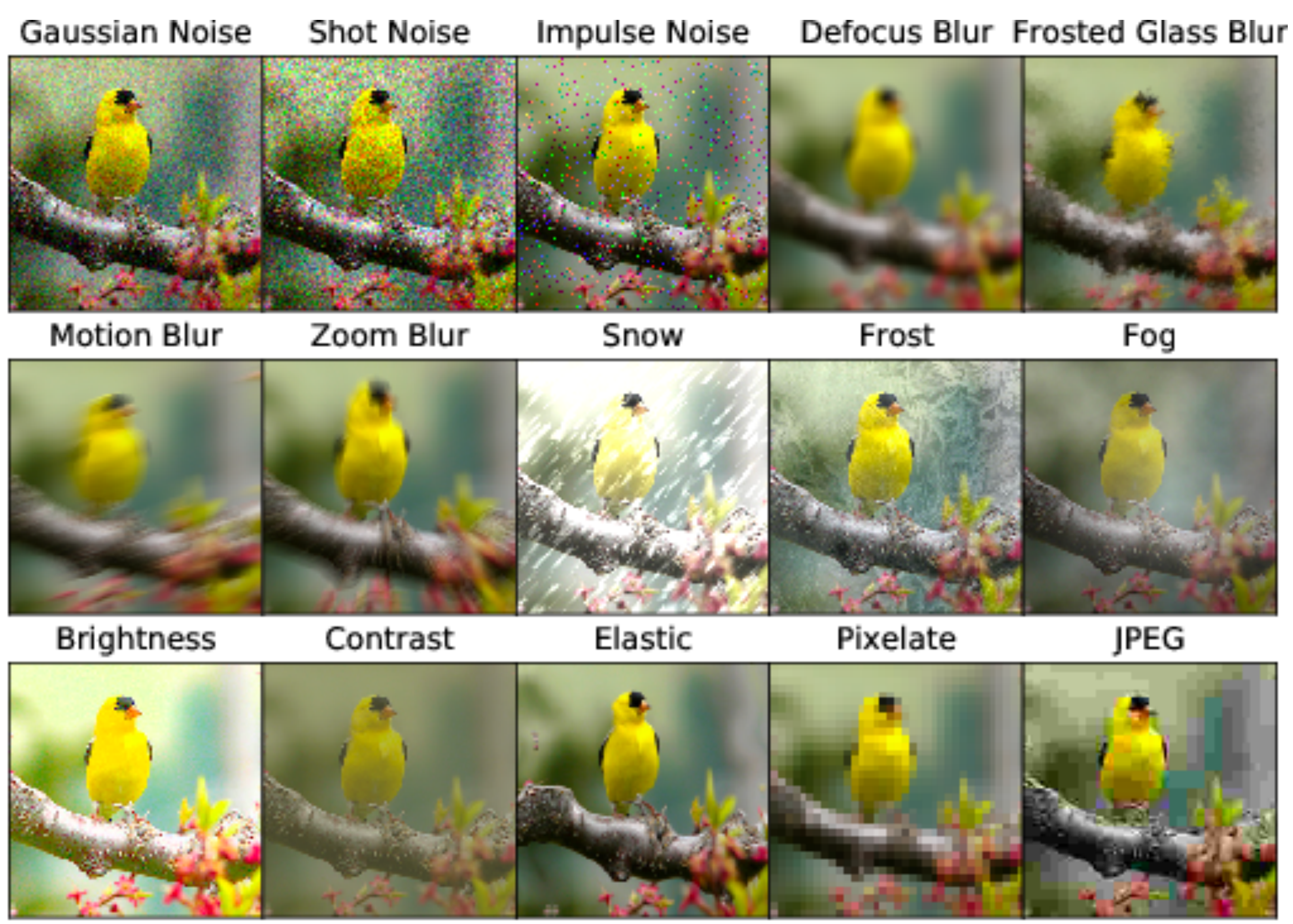
- CIFAR10.1, ImageNetV2 [Recht, et al.  2019]
- CIFAR10.2 [Lu, et al. 2020]



# Accuracy on the Line (Miller, et al. 2021):
# OOD Benchmarks

**Synthetic corruptions**

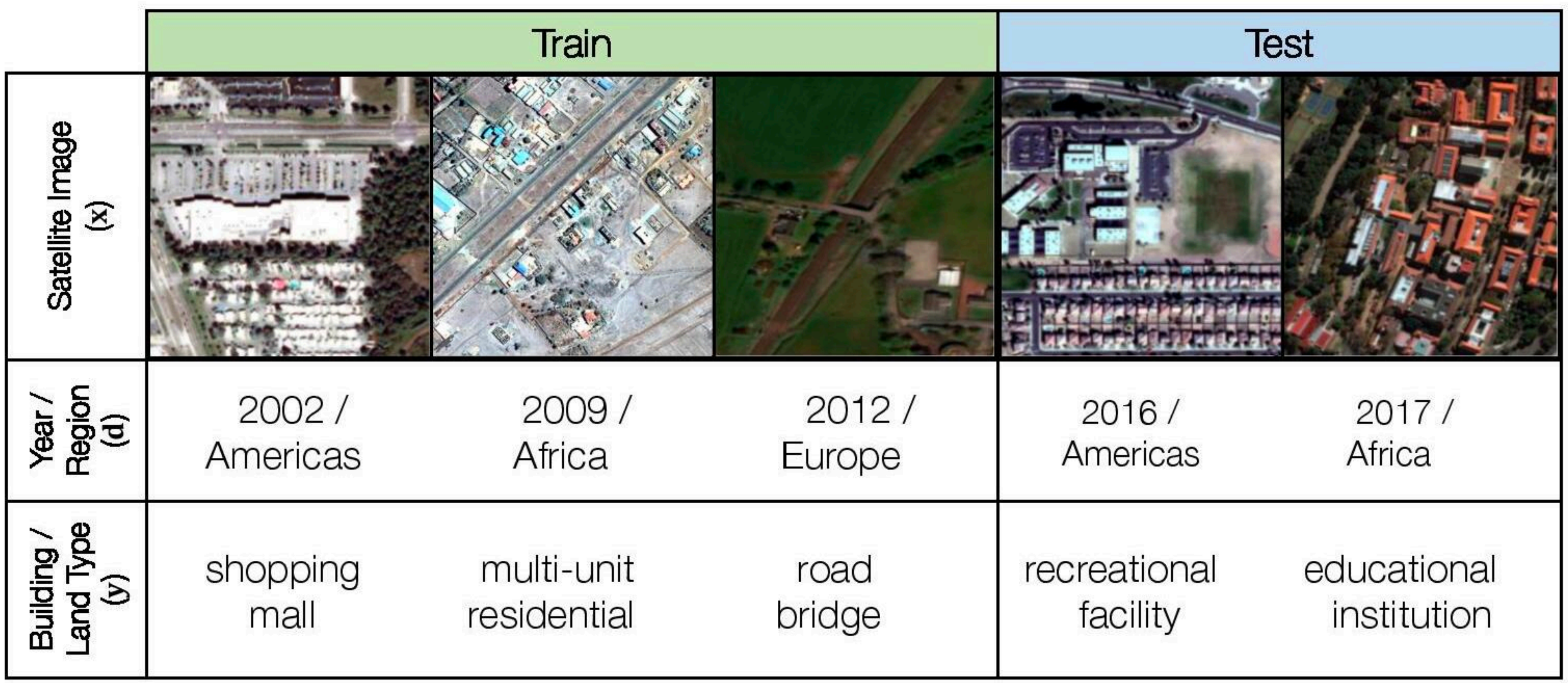
- CIFAR10C [Hendrycks and Dietterich, 2019]



# Accuracy on the Line (Miller, et al. 2021):
# OOD Benchmarks

**Real world shifts** (Environmental changes, human activity)

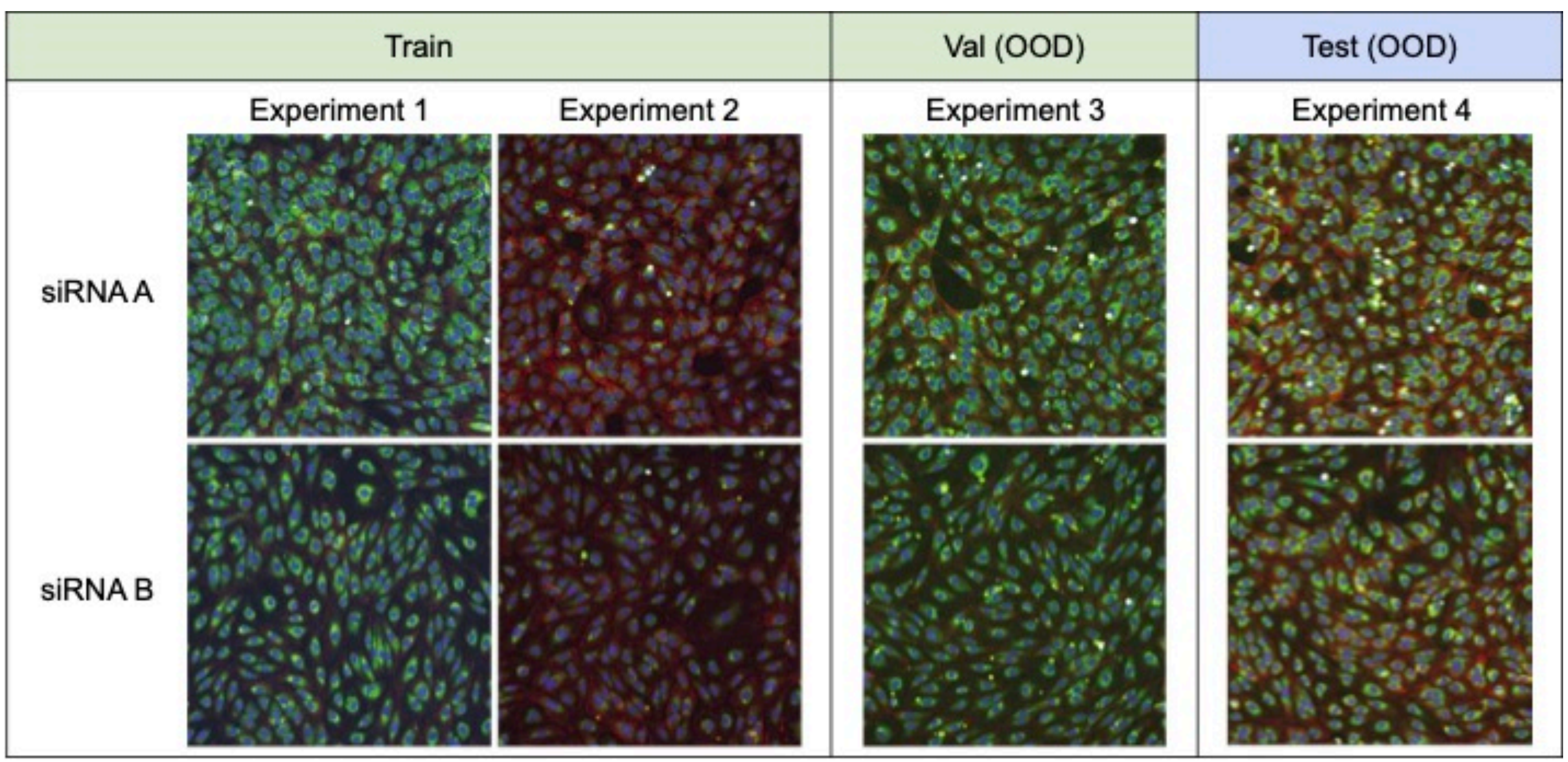
- **fMoW-wilds**
- RxRx1-wilds
- Camelyon17-wilds
- iWildCam-wilds



# Accuracy on the Line (Miller, et al. 2021): OOD Benchmarks

**Real world shifts** (Environmental changes, human activity)

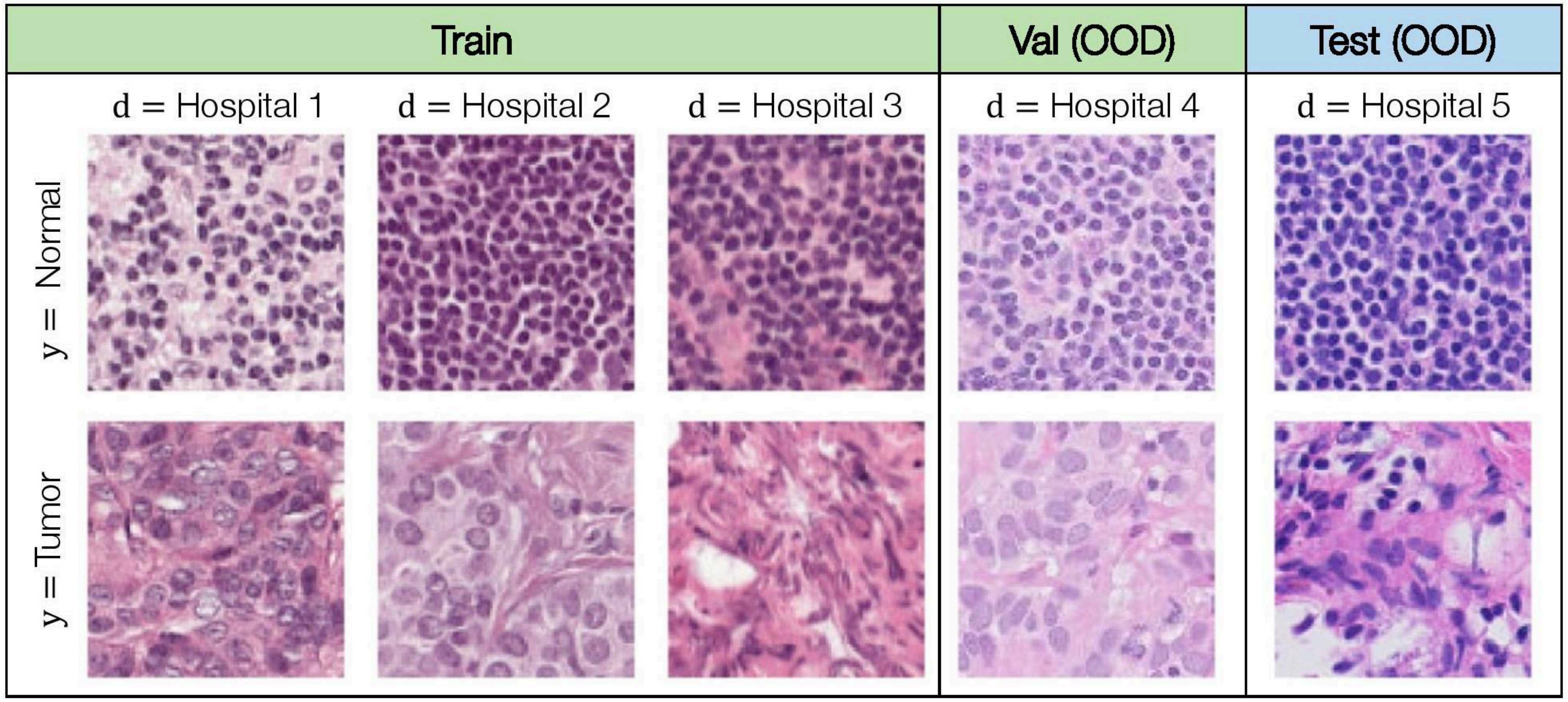
- fMoW-wilds
- **RxRx1-wilds**
- Camelyon17-wilds
- iWildCam-wilds



# Accuracy on the Line (Miller, et al. 2021):
# OOD Benchmarks

**Real world shifts** (Environmental changes, human activity)

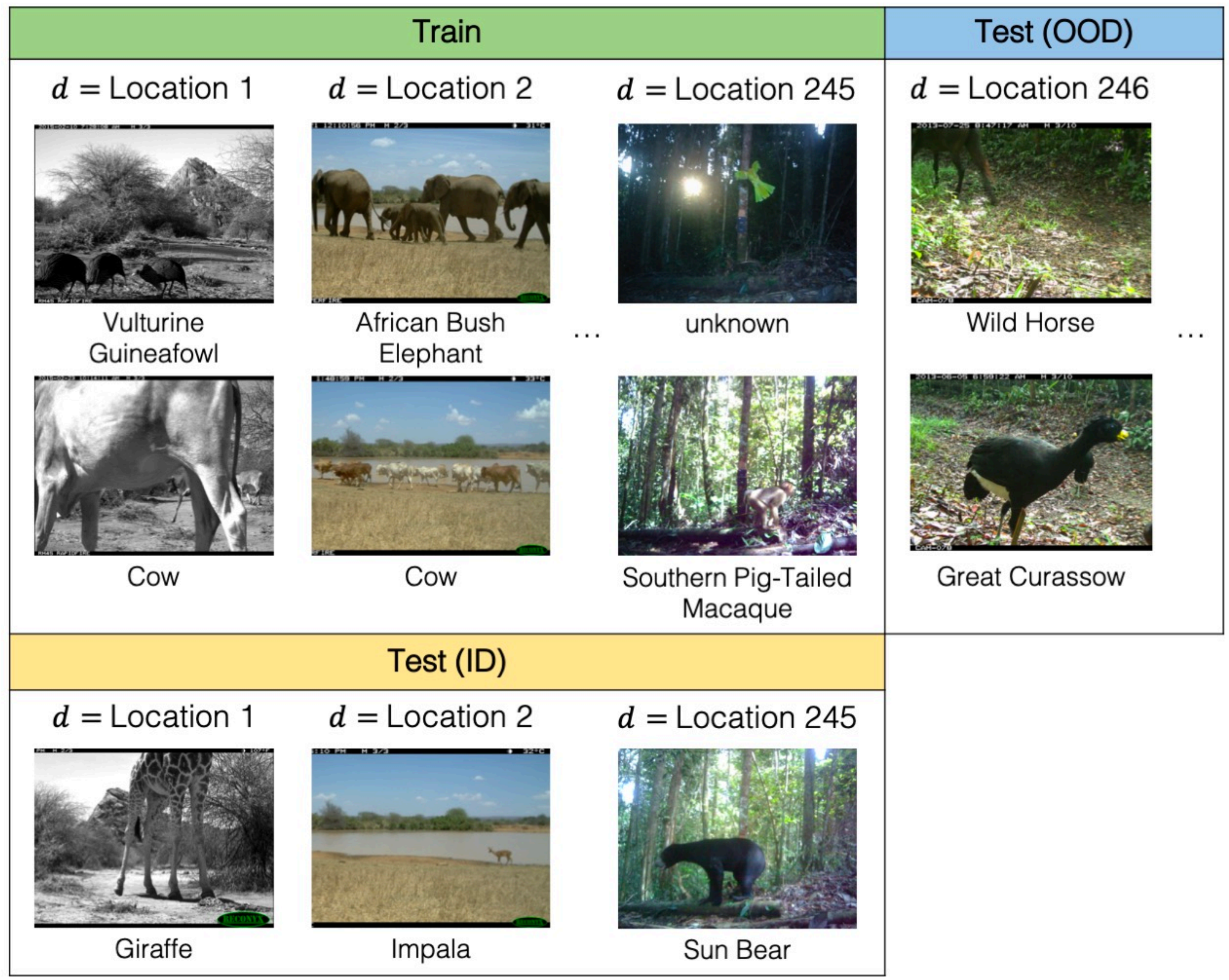
- fMoW-wilds
- RxRx1-wilds
- **Camelyon17-wilds**
- iWildCam-wilds



# Accuracy on the Line (Miller, et al. 2021):
# OOD Benchmarks

**Real world shifts** (Environmental changes, human activity)

- fMoW-wilds
- RxRx1-wilds
- Camelyon17-wilds
- **iWildCam-wilds**



# Models

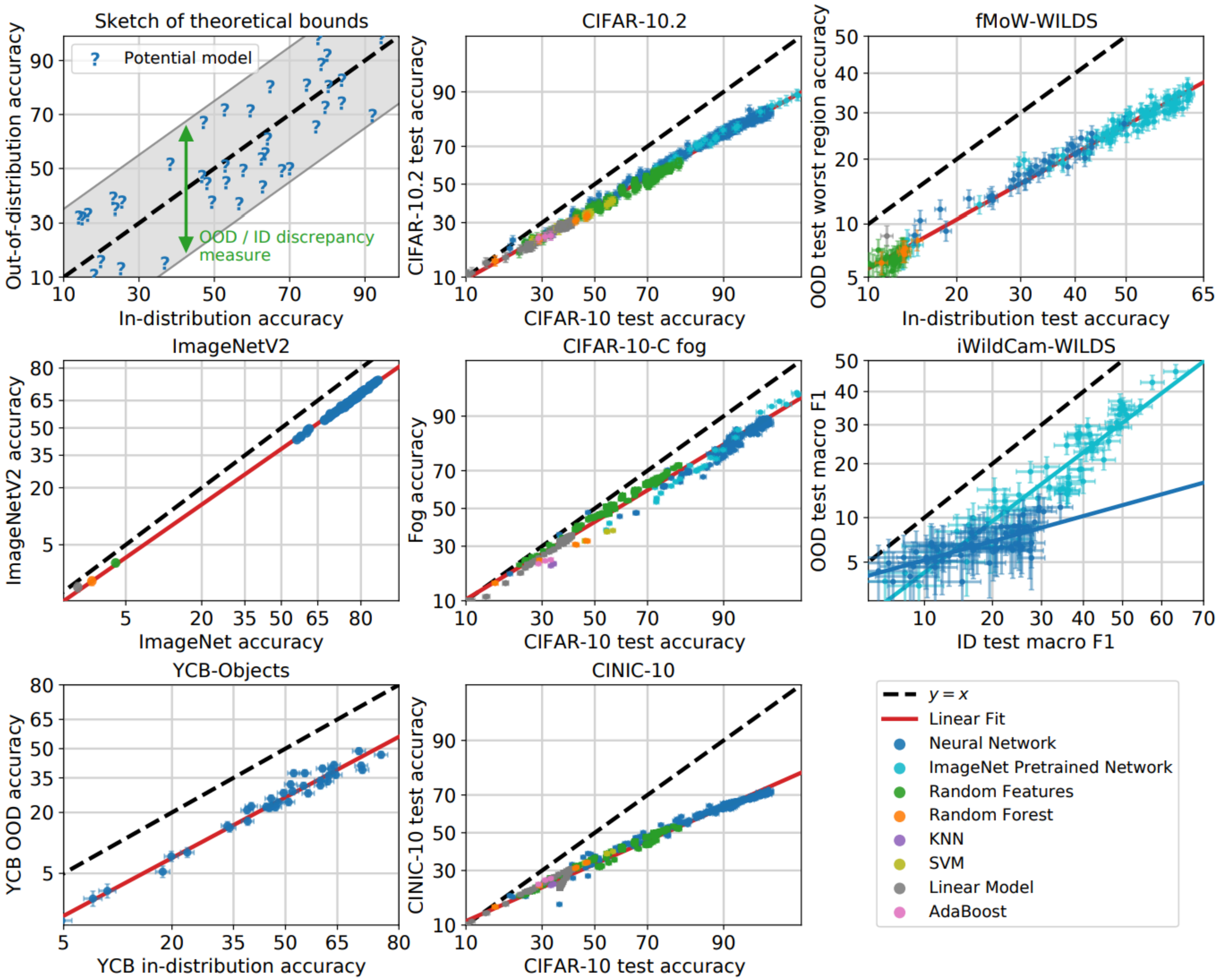
Our model collection $\mathscr{H}$ consists of **CNNs** and **Vision Transformers** trained using different

1. hyperparameter
2. training set size
3. training duration

| Architecture | Number of models |
|---|---|
| Adversarial Inception v3 [46] | 1 |
| AlexNet [45] | 1 |
| BEiT [2] | 1 |
| BoTNet [68] | 1 |
| CaiT [74] | 1 |
| CoaT [78] | 2 |
| ConViT [20] | 3 |
| ConvNeXT [50] | 1 |
| CrossViT [8] | 9 |
| DenseNet [37] | 3 |
| DLA [79] | 10 |
| EfficientNet [72] | 1 |
| HaloNet [75] | 1 |
| NFNet [7] | 1 |
| ResNet [33] | 10 |
| ResNeXT [77] | 1 |
| Inception v3 [71] | 1 |
| VGG [67] | 1 |

| Architecture | Number of models |
|---|---|
| DenseNet121 [37] | 21 |
| DenseNet169 [37] | 8 |
| EfficientNetB0 [72] | 13 |
| ResNet18 [33] | 13 |
| ResNet50 [33] | 18 |
| ResNet101 [33] | 7 |
| PreActResNet18 [32] | 63 |
| PreActResNet34 [32] | 9 |
| PreActResNet50 [32] | 11 |
| PreActResNet101 [32] | 4 |
| ResNeXT $2 \times 64$d [77] | 12 |
| ResNeXT $32 \times 4$d [77] | 8 |
| ResNeXT $4 \times 64$d [77] | 1 |
| RegNet X200 [62] | 11 |
| RegNet X400 [62] | 13 |
| RegNet Y400 [62] | 5 |
| VGG11 [67] | 16 |
| VGG13 [67] | 13 |
| VGG16 [67] | 13 |
| VGG19 [67] | 12 |
| ShuffleNetV2 [53] | 56 |
| ShuffleNetG2 [81] | 13 |
| ShuffleNetG3 [81] | 8 |
| AlexNet [45] | 2 |
| MobileNet [65] | 12 |
| MobileNetV2 [65] | 13 |
| PNASNet-A [48] | 13 |
| PNASNet-B [48] | 13 |
| PNASNet-5-Large [48] | 3 |
| SqueezeNet [39] | 3 |
| SENet18 [42] | 13 |
| GoogLeNet [70] | 20 |
| DPN26 [11] | 8 |
| DPN92 [11] | 2 |
| Myrtlenet [Repo] | 1 |
| Xception [12] | 3 |

# Accuracy on the Line (Miller, et al. 2021)




There is a **structure** to the way distributions commonly shift

…but this fact does not solve the problem of **needing OOD labels for accuracy.**

# Agreement

Measure the rate at which predictions of two hypotheses agree

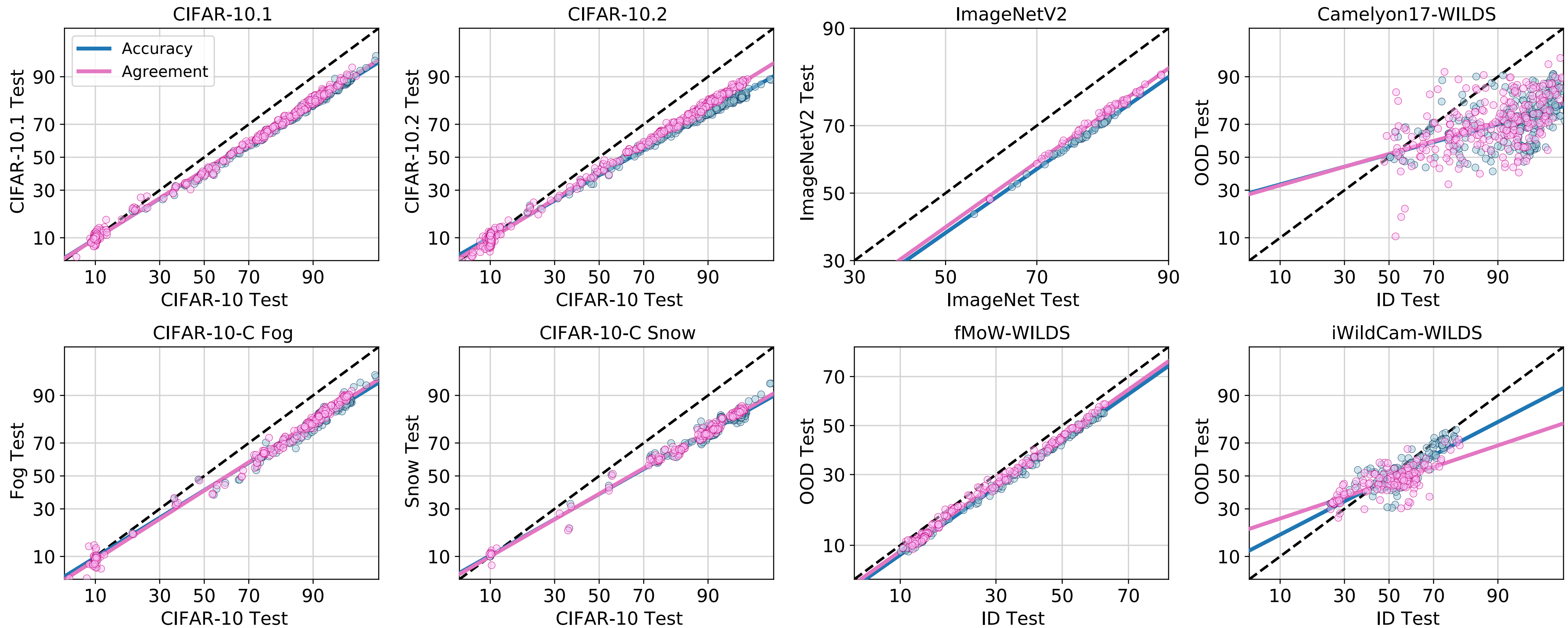
Test Input $(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10})$

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Model 1 Predictions** | 1 | 2 | 2 | 1 | 2 | 1 | 2 | 1 | 1 | 1 |
| **Model 2 Predictions** | 1 | 1 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 1 |

**Agreement: 60%**

**Does not need labels!**

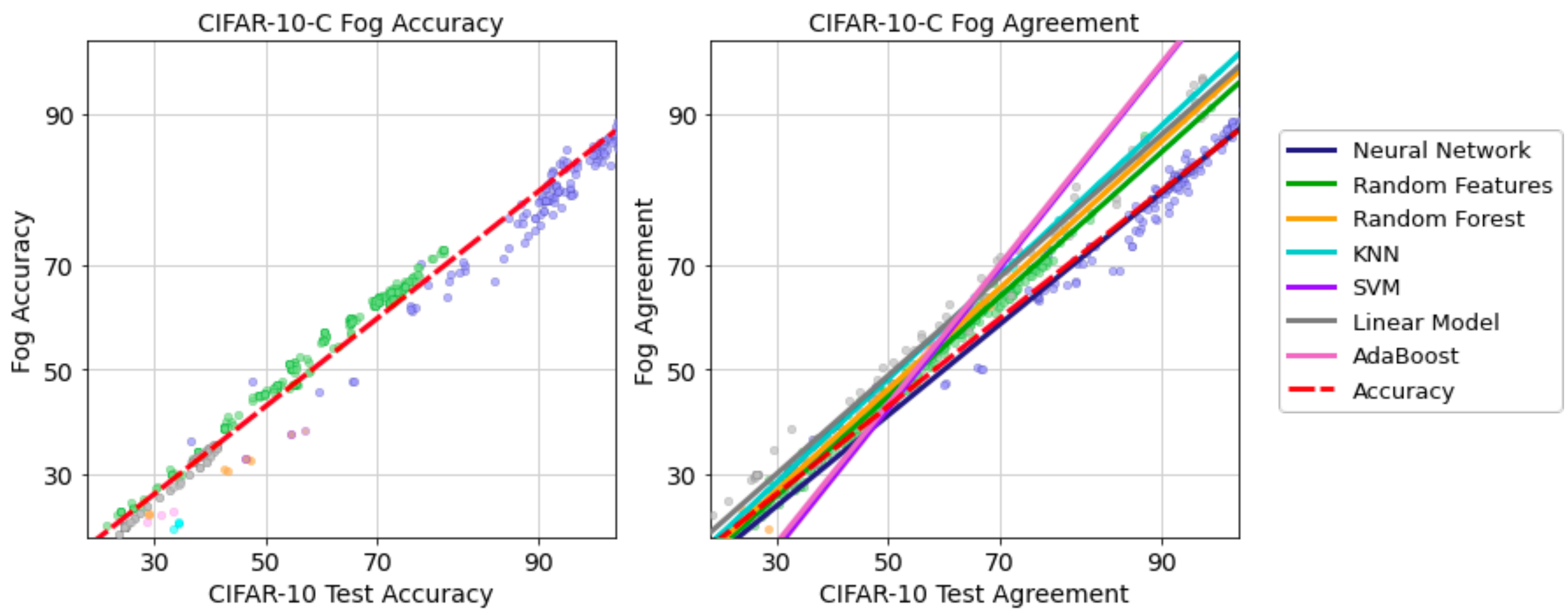
# Agreement-on-the-Line




- **Strong Correlation** When ID vs OOD accuracy is *strongly* linearly correlated (≥ 0.95 $R^2$ values), ID vs OOD agreement is *also* *strongly* linearly correlated. Additionally, these linear correlations have almost the **same slope and bias**.

- **Weak Correlation** When ID vs OOD accuracy is *weakly* linearly correlated (≤ 0.75 $R^2$ values), ID and OOD agreement is *also* *weakly* linearly correlated.

# The phenomena only occurs for neural networks



# OOD Accuracy Estimation

1. Estimate slope and bias by linear regression of ID vs OOD agreement
   $$\Phi^{-1}(\mathrm{Agr}_{\mathrm{OOD}}(h,h')) = a \cdot \Phi^{-1}(\mathrm{Agr}_{\mathrm{ID}}(h,h'))+b$$

2. *If the linear correlation is strong*, *we know* approximately
   $$\Phi^{-1}(\mathrm{Acc}_{\mathrm{OOD}}(h)) = a \cdot \Phi^{-1}(\mathrm{Acc}_{\mathrm{ID}}(h))+b$$

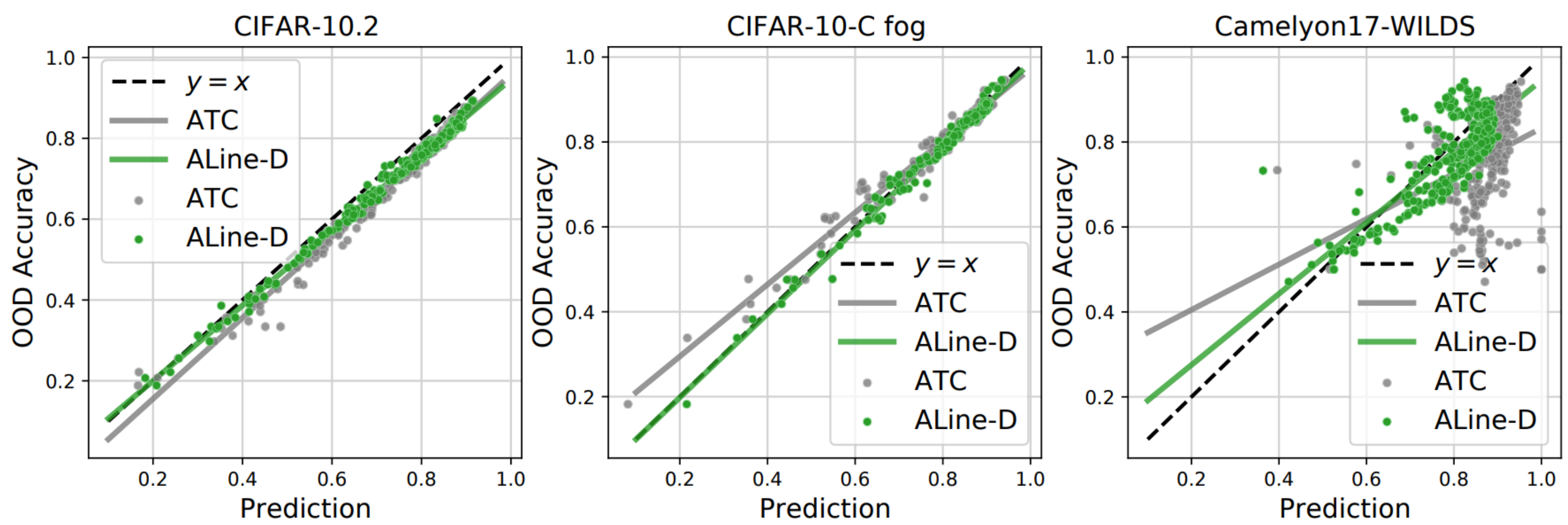
3. From 1 and 2, note that for any two models $h, h' \in \mathscr{H}$

$$\frac{1}{2}\underbrace{\Phi^{-1}(\mathrm{Acc}_{\mathrm{OOD}}(h))}_{\text{unknown}} + \frac{1}{2}\underbrace{\Phi^{-1}(\mathrm{Acc}_{\mathrm{OOD}}(h'))}_{\text{unknown}} = \underbrace{\Phi^{-1}(\mathrm{Agr}_{\mathrm{OOD}}(h,h'))+a \cdot \left(\frac{\Phi^{-1}(\mathrm{Acc}_{\mathrm{ID}}(h)) + \Phi^{-1}(\mathrm{Acc}_{\mathrm{ID}}(h'))}{2} - \Phi^{-1}(\mathrm{Agr}_{\mathrm{ID}}(h,h'))\right)}_{\text{known}}$$

4. Solve system of linear equations for $\Phi^{-1}(\mathrm{Acc}_{\mathrm{OOD}}(h)) \quad \forall h \in \mathscr{H}$

**ALine-S:** Steps 1-2
**ALine-D:** Steps 1-4

# OOD Accuracy Estimation




**Mean Absolute Estimation Error with % as units.**

| Dataset | ALine-D | ALine-S | ATC [Garg '22] | AC [Hendrycks '17] | DOC [Guillory '17] | Agreement |
|---|---|---|---|---|---|---|
| CIFAR-10.1 | **1.11** | 1.17 | 1.21 | 4.51 | 3.87 | 5.98 |
| CIFAR-10.2 | **3.93** | **3.93** | 4.35 | 8.23 | 7.64 | 5.42 |
| ImageNetV2 | 2.06 | 2.08 | **1.12** | 66.2 | 11.50 | 6.70 |
| CIFAR-10C-Fog | **1.45** | 1.75 | 1.78 | 4.47 | 3.93 | 3.47 |
| CIFAR-10C-Snow | **1.32** | 1.97 | **1.31** | 5.94 | 5.49 | 2.57 |
| CIFAR10C-Saturate | **0.41** | 0.77 | 0.69 | 2.03 | 1.51 | 4.14 |
| fMoW-wilds | **1.30** | 1.44 | 1.53 | 2.89 | 2.60 | 8.99 |
| RxRx1-wilds | **0.27** | 0.52 | 2.97 | 2.46 | 0.65 | 8.67 |
| Camelyon17-wilds | **5.47** | 8.31 | 11.93 | 13.30 | 13.57 | 6.79 |
| iWildCam-wilds | 4.95 | 6.01 | 12.12 | **4.46** | 5.02 | 7.53 |

# Along One Trajectory

1. Train a single ResNet18 model on CIFAR-10.

2. Every 5 epochs, save the predictions of the model over CIFAR-10 and CIFAR-10.1 Test.

3. Perform ALine-D