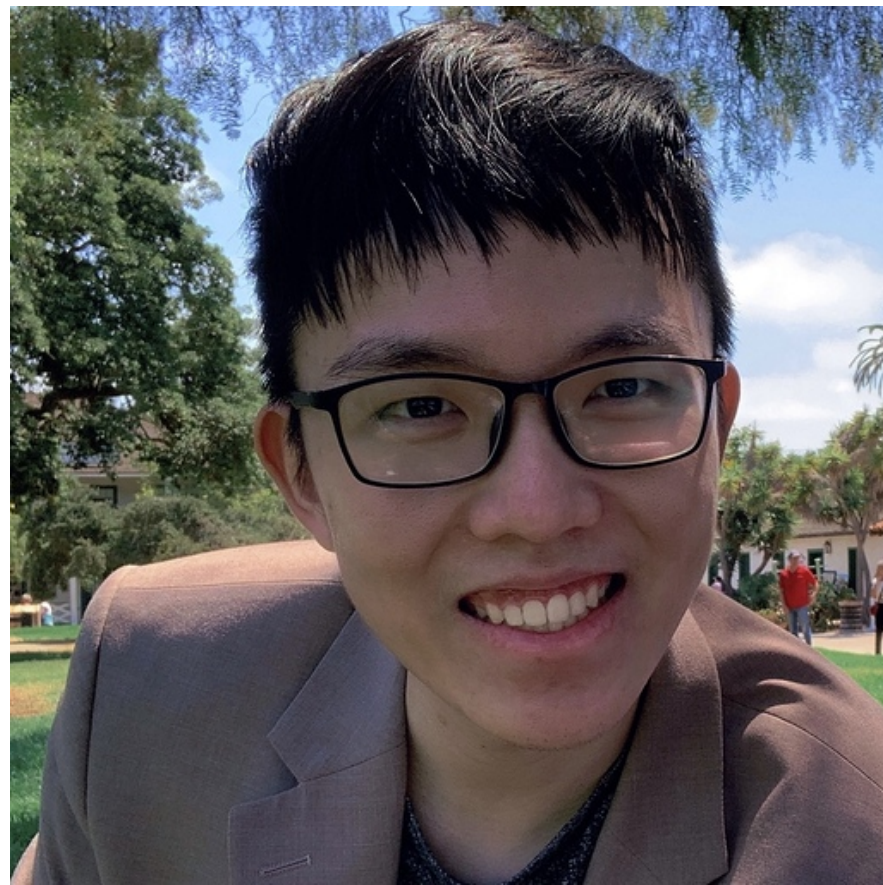
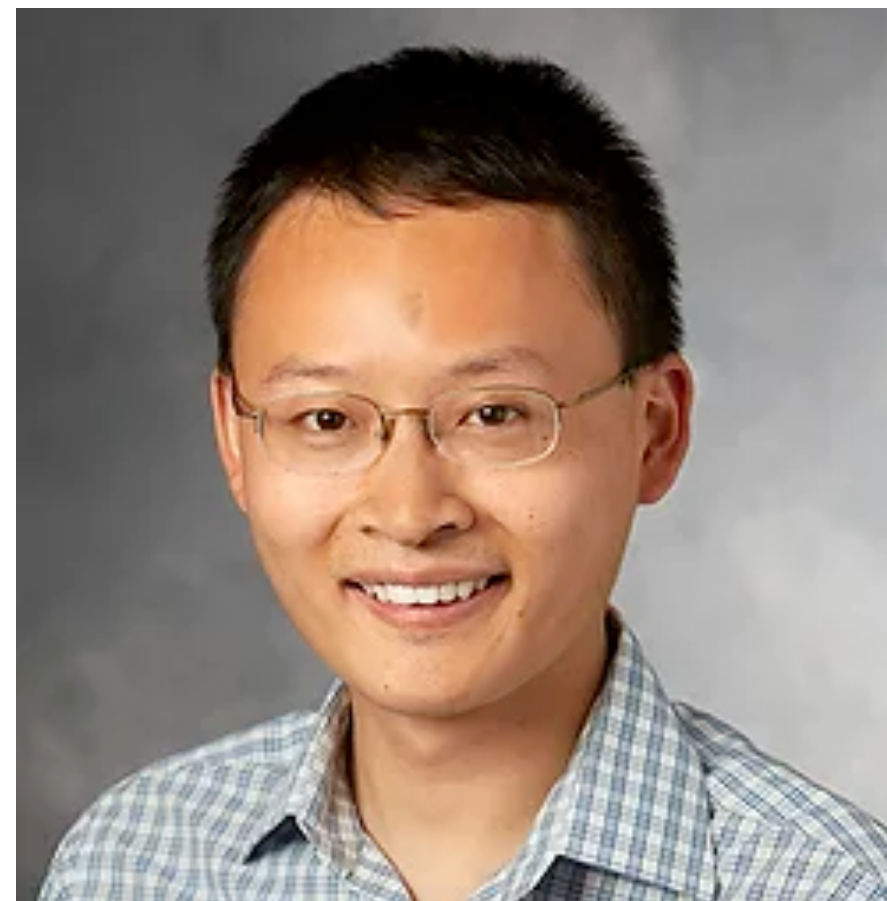


# MetaShift: A Dataset of Datasets for Evaluating Contextual Distribution Shifts and Training Conflicts

<https://metashift.readthedocs.io/>



**Weixin Liang**



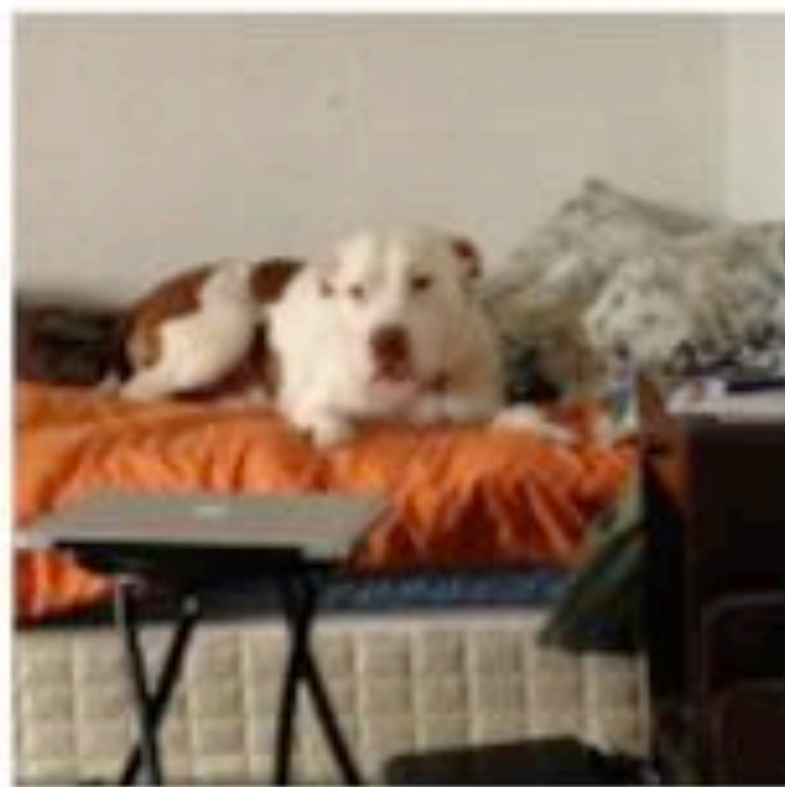
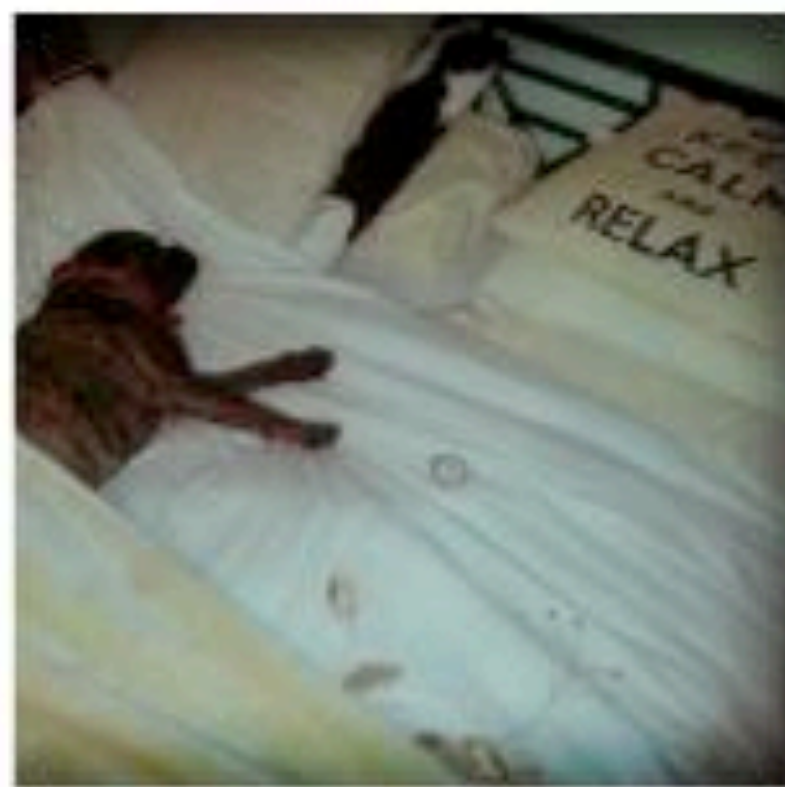
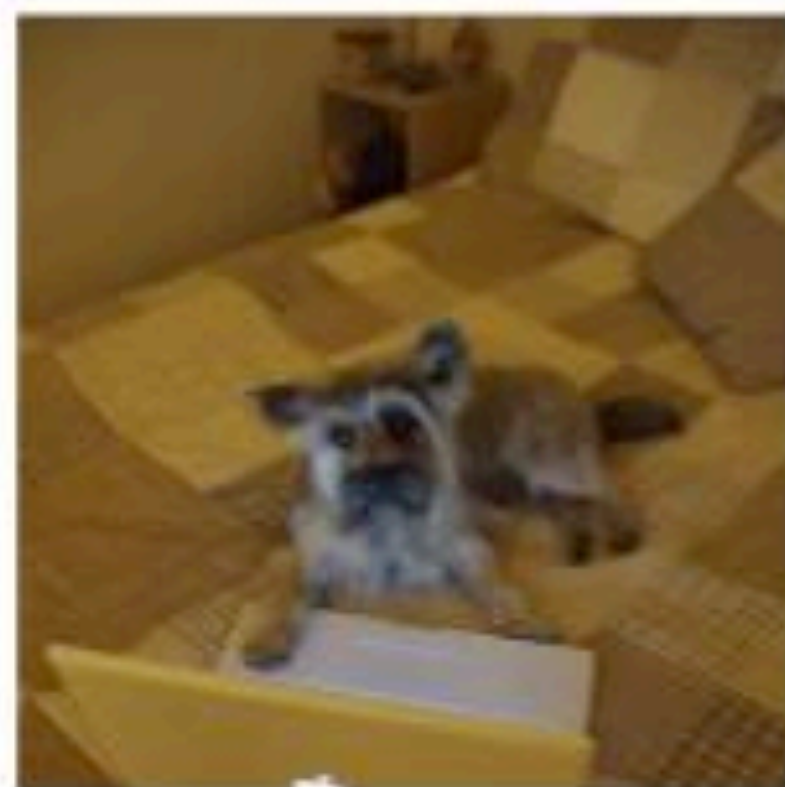
**James Zou**



# Motivation: Evaluating Distribution Shift

Understanding the performance across diverse data distributions

Train: indoor dog images

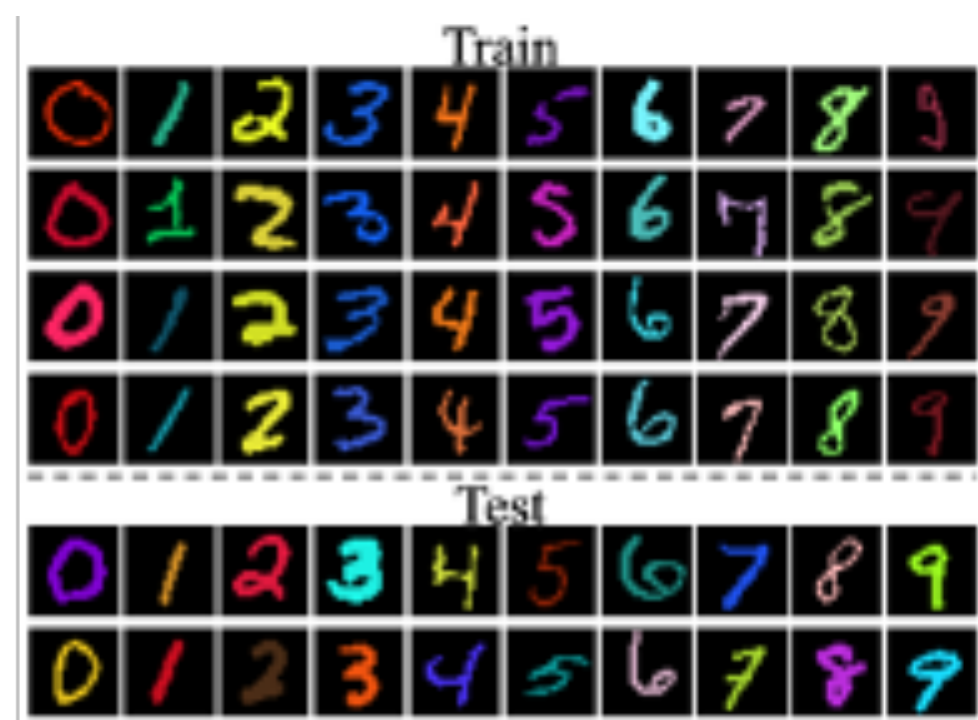


Test: outdoor dog images

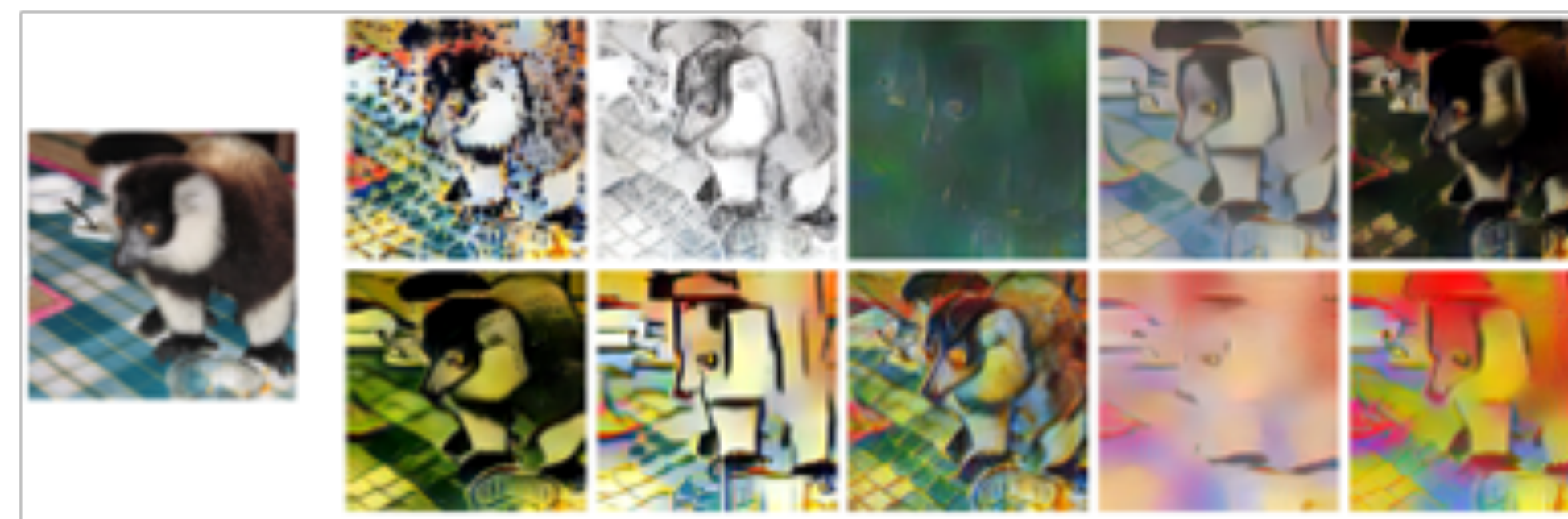


# Existing Datasets for Distribution Shift

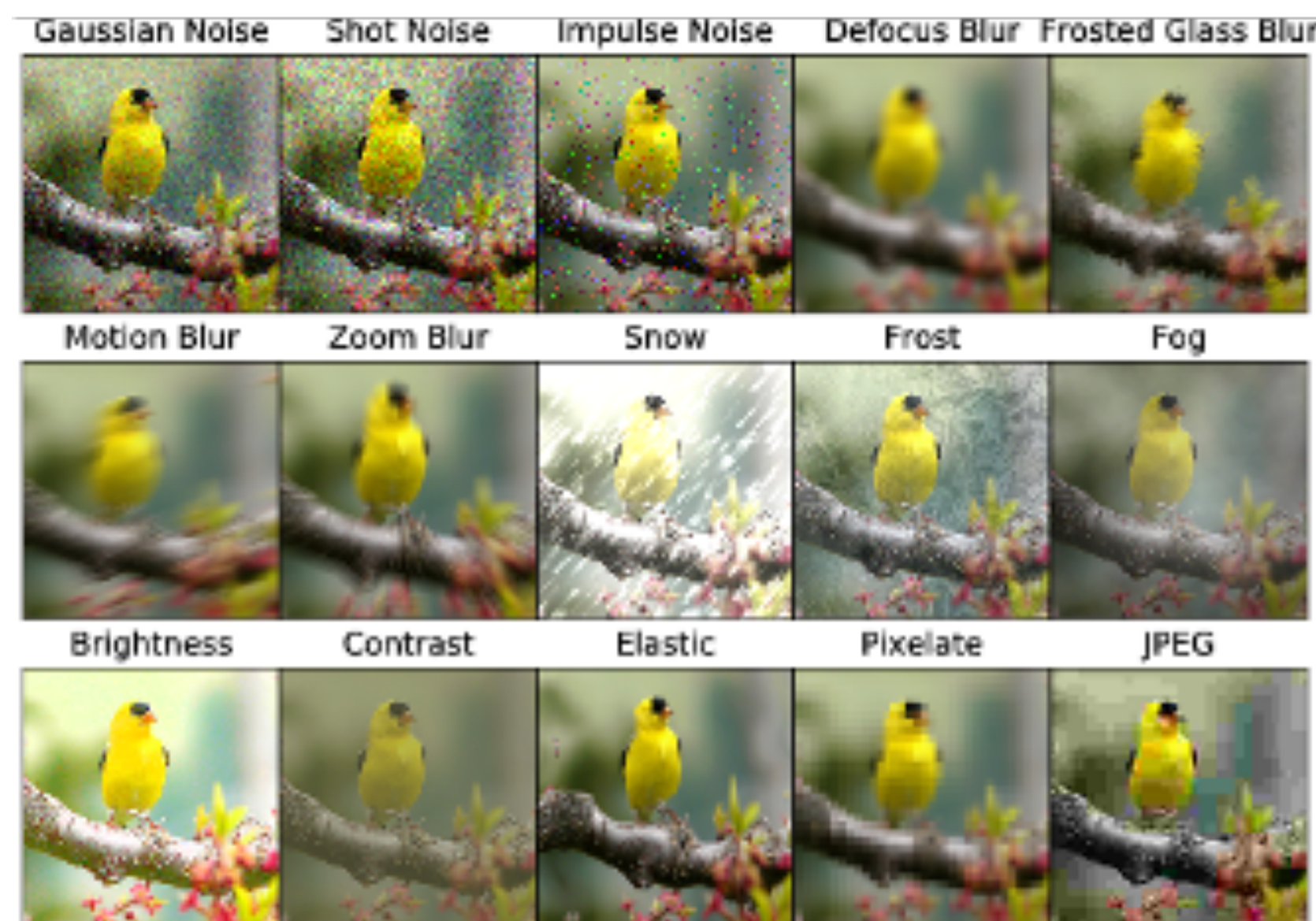
Mostly synthetic, and limited in a small number of shifts



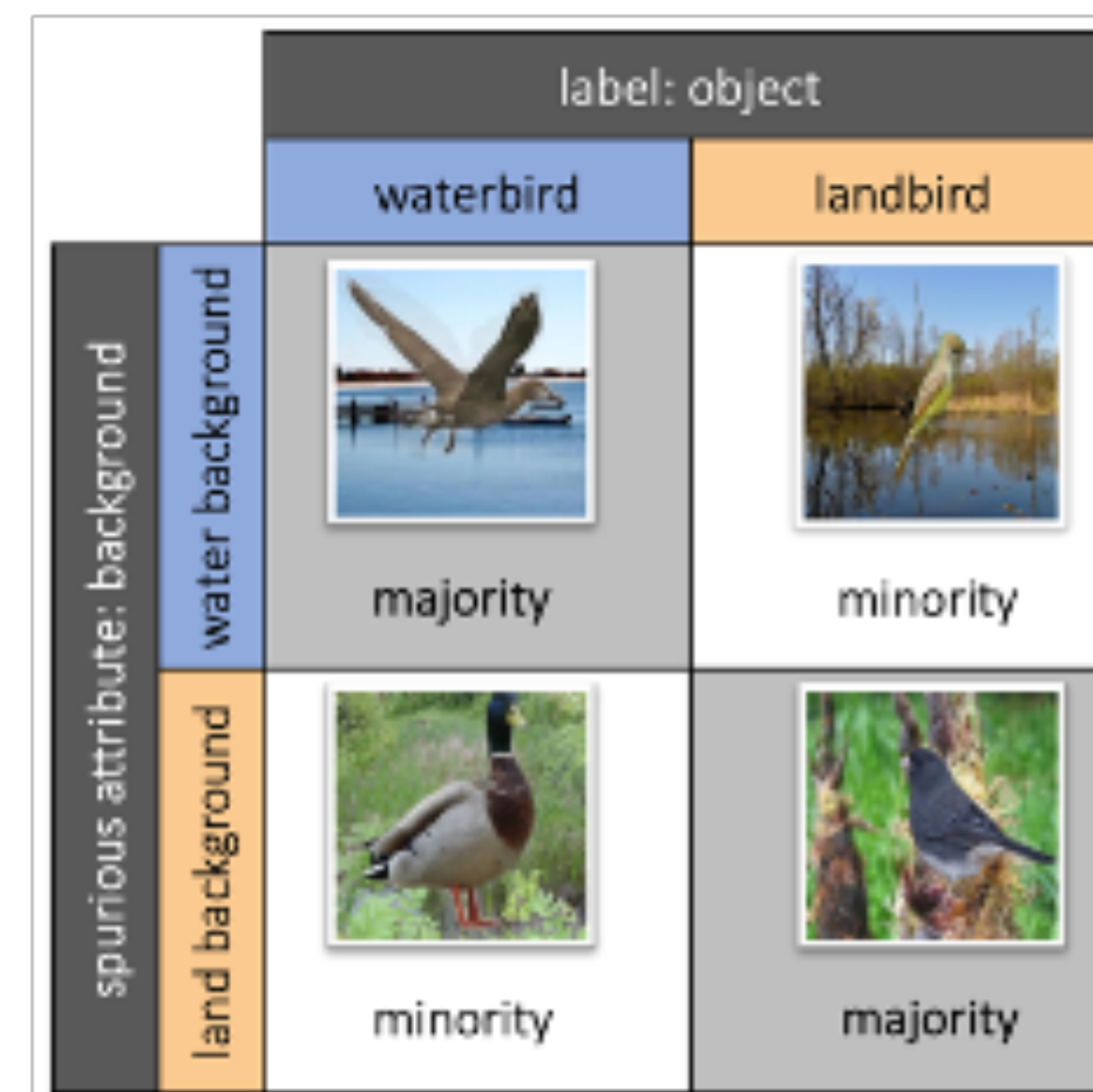
Colored MNIST



Stylized-ImageNet



ImageNet-C



WaterBirds

# MetaShift: collection of 12,868 sets of natural images w/ annotated contexts



# Meta-graph: an annotation graph that explains the similarity/distance between two subsets




cat  
(sink)

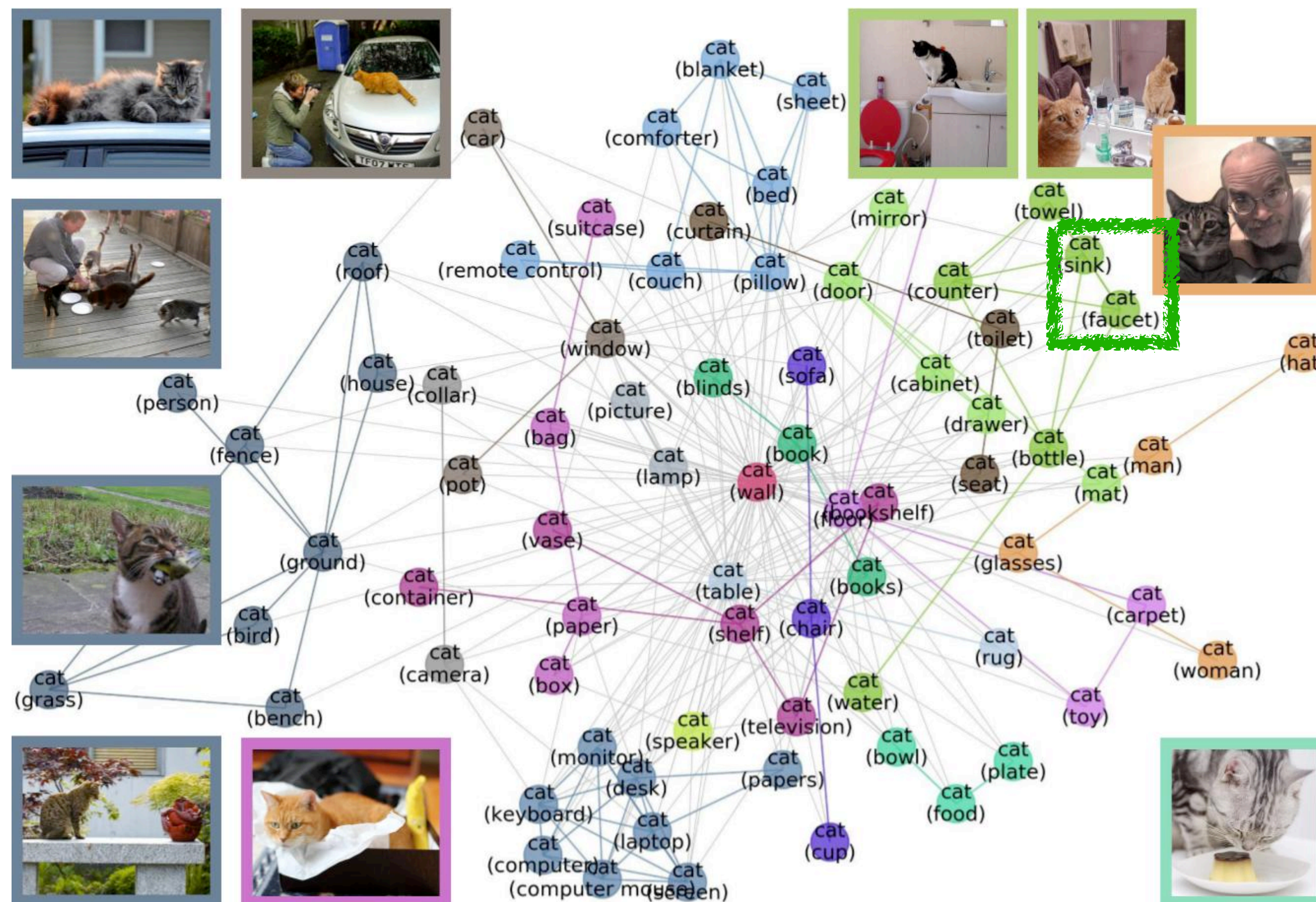
## Node: what is unique about each subset



cat  
(sink)



# Edge weight: the similarity/distance between two subsets

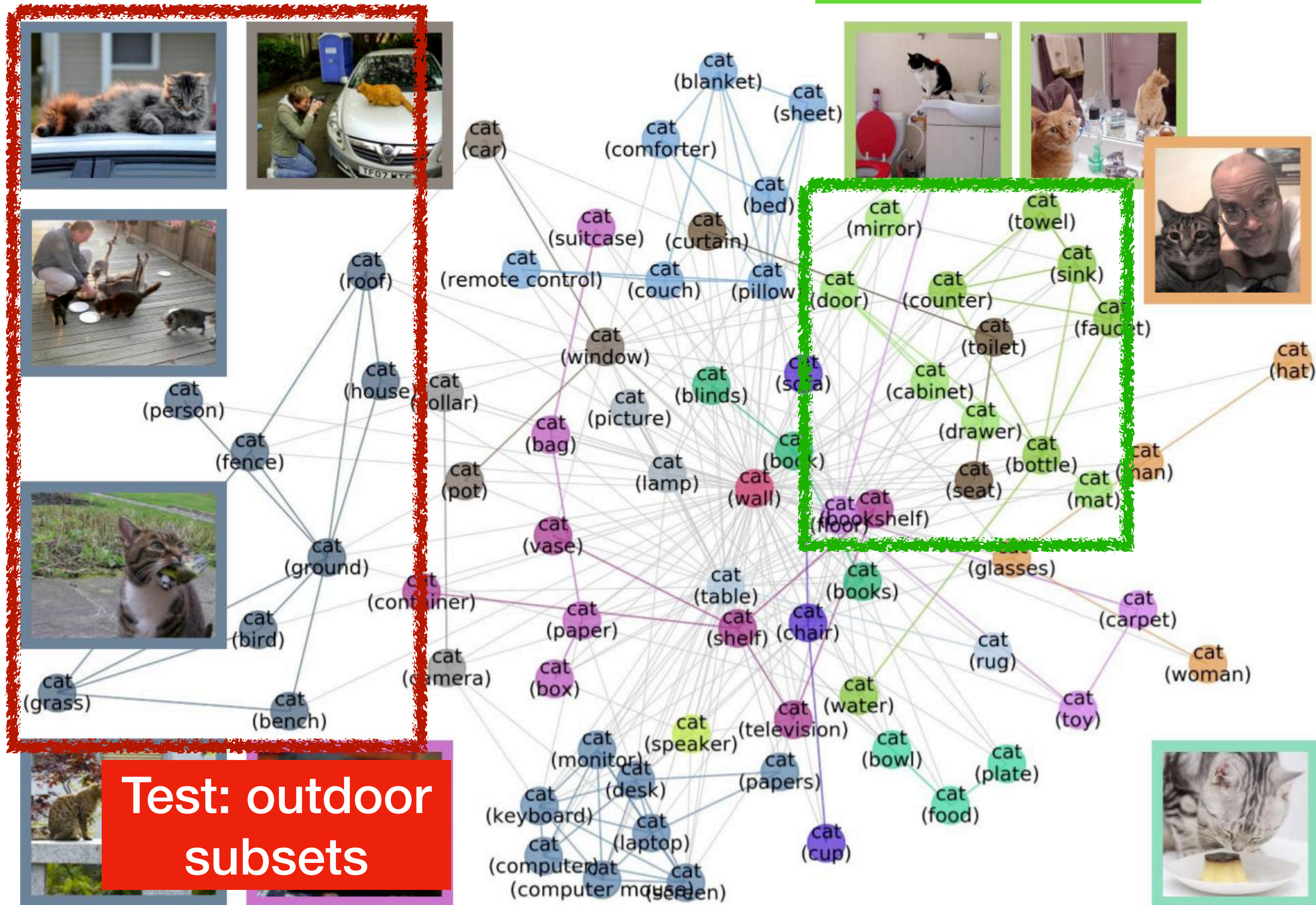


# How Can We Use MetaShift?

# Application 1: Evaluating Distribution Shift

A flexible framework to generate **a large number of real-world distribution shifts** that are **well-annotated** and **controlled**.

# Train: bathroom subsets



# Finding: The simple empirical risk minimization performs the best when shifts are moderate

d: Distance of Distribution Shift

Algorithm	Cat vs. Dog			
	$d=0.44$	$d=0.71$	$d=1.12$	$d=1.43$
ERM	<b>0.844</b>	0.605	0.357	0.240
IRM	0.814	<b>0.628</b>	0.380	<b>0.341</b>
GroupDRO	0.837	0.597	0.434	0.264
CORA	0.798	0.589	<b>0.481</b>	0.302
CDANN	0.729	0.620	0.380	0.326

Algorithm	Elephant vs. Horse			
	$d=0.44$	$d=0.63$	$d=0.89$	$d=1.44$
ERM	<b>0.964</b>	0.821	0.793	0.729
IRM	0.943	0.886	0.764	0.750
GroupDRO	0.936	0.864	0.829	0.743
CORA	0.929	<b>0.900</b>	0.814	0.771
CDANN	0.921	0.857	<b>0.836</b>	<b>0.779</b>

Algorithm	Bus vs. Truck			
	$d=0.81$	$d=1.20$	$d=1.42$	$d=1.52$
ERM	<b>0.950</b>	0.863	0.702	0.609
IRM	0.863	<b>0.901</b>	0.752	0.634
GroupDRO	0.901	0.857	<b>0.770</b>	<b>0.665</b>
CORA	0.925	0.801	0.783	0.640
CDANN	0.944	0.888	0.789	0.584

Algorithm	Bowl vs. Cup			
	$d=0.16$	$d=0.47$	$d=1.03$	$d=1.31$
ERM	<b>0.888</b>	0.768	0.401	0.276
IRM	0.883	<b>0.793</b>	0.426	<b>0.404</b>
GroupDRO	0.829	0.765	0.444	0.303
CORA	0.850	0.734	<b>0.482</b>	0.274
CDANN	0.879	0.752	0.432	0.280

# Finding: No method had a systematic advantage for large shifts

d: Distance of Distribution Shift

Algorithm	Cat vs. Dog			
	$d=0.44$	$d=0.71$	$d=1.12$	$d=1.43$
ERM	<b>0.844</b>	0.605	0.357	0.240
IRM	0.814	<b>0.628</b>	0.380	<b>0.341</b>
GroupDRO	0.837	0.597	0.434	0.264
CORA	0.798	0.589	<b>0.481</b>	0.302
CDANN	0.729	0.620	0.380	0.326

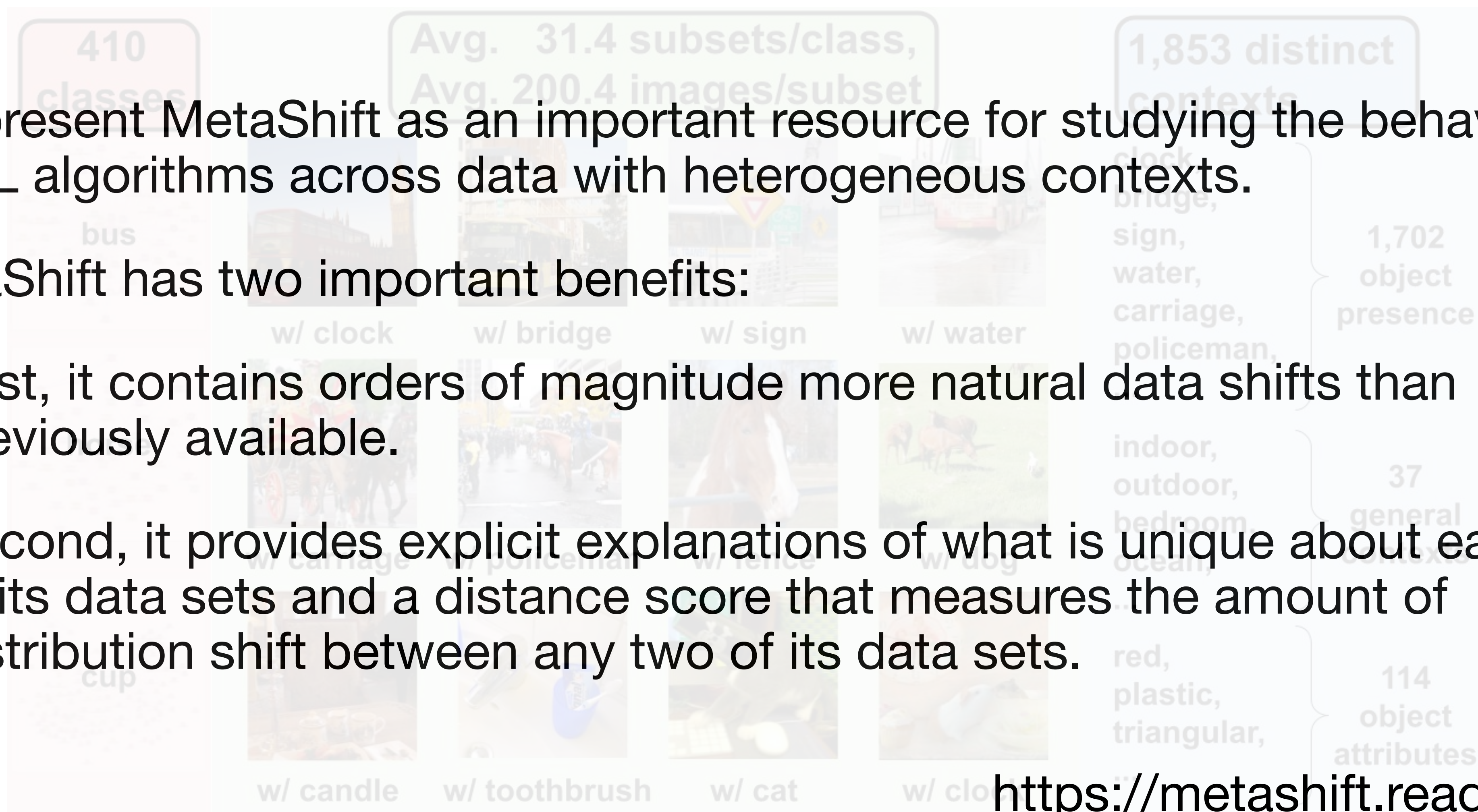
Algorithm	Elephant vs. Horse			
	$d=0.44$	$d=0.63$	$d=0.89$	$d=1.44$
ERM	<b>0.964</b>	0.821	0.793	0.729
IRM	0.943	0.886	0.764	0.750
GroupDRO	0.936	0.864	0.829	0.743
CORA	0.929	<b>0.900</b>	0.814	0.771
CDANN	0.921	0.857	<b>0.836</b>	<b>0.779</b>

Algorithm	Bus vs. Truck			
	$d=0.81$	$d=1.20$	$d=1.42$	$d=1.52$
ERM	<b>0.950</b>	0.863	0.702	0.609
IRM	0.863	<b>0.901</b>	0.752	0.634
GroupDRO	0.901	0.857	<b>0.770</b>	<b>0.665</b>
CORA	0.925	0.801	0.783	0.640
CDANN	0.944	0.888	0.789	0.584

Algorithm	Bowl vs. Cup			
	$d=0.16$	$d=0.47$	$d=1.03$	$d=1.31$
ERM	<b>0.888</b>	0.768	0.401	0.276
IRM	0.883	<b>0.793</b>	0.426	<b>0.404</b>
GroupDRO	0.829	0.765	0.444	0.303
CORA	0.850	0.734	<b>0.482</b>	0.274
CDANN	0.879	0.752	0.432	0.280

# Summary

- We present MetaShift as an important resource for studying the behavior of ML algorithms across data with heterogeneous contexts.
- MetaShift has two important benefits:
  - First, it contains orders of magnitude more natural data shifts than previously available.
  - Second, it provides explicit explanations of what is unique about each of its data sets and a distance score that measures the amount of distribution shift between any two of its data sets.



<https://metashift.readthedocs.io/>