DATAS^3: DATASET SUBSET SELECTION FOR SPECIALIZATION

Anonymous authors
Paper under double-blind review

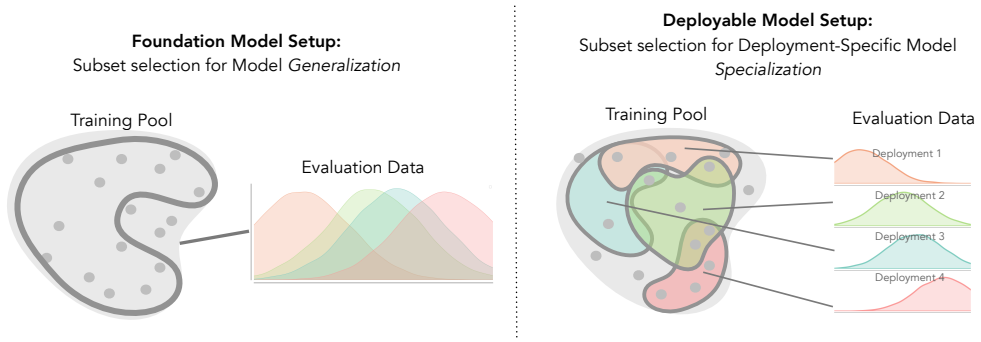

Figure 1: Foundation model training aims for broad generalization, by using all data available, usually from massive internet-scale datasets. In practice, we find these models are often suboptimal for specific deployments, which may exhibit different distributions over categories or data characteristics from the general training data pool. Dataset subset selection for specialization seeks to identify model training subsets closely aligned with the target deployment, achieving superior performance under the given distribution and attribute shifts.

ABSTRACT

In many real-world machine learning applications (e.g. detecting broken bones in x-rays or species in camera traps), models need to perform well on specific deployments (e.g. a specific hospital or national park) rather than the domain broadly. However, deployments often have imbalanced, unique data distributions. Discrepancies between training and deployment distributions lead to suboptimal performance, highlighting the need to curate training data for *specialized models for specific deployment needs*. We formalize **dataset subset selection for specialization (DS3)**: given a training set drawn from a general distribution and a (potentially unlabeled) query set drawn from a deployment-specific distribution, the goal is to select a subset of the training data that optimizes deployment performance.

We introduce DATAS^3; the first dataset and benchmark designed specifically for the DS3 problem. DATAS^3 encompasses five *real-world* application domains, each with a set of distinct deployments to specialize in. We conduct a comprehensive study evaluating different state-of-the-art data curation algorithms and find that methods trained on general distributions consistently fail to perform optimally on deployment tasks. Additionally, we demonstrate the existence of expert-curated (deployment-specific) subsets that outperform training on all available data by up to 51.3%. Our benchmark highlights the critical role of tailored dataset curation in enhancing performance and training efficiency on deployment-specific distributions, which we posit will only become more important as global, public datasets become available across domains and ML models are deployed in the real world.

1 BACKGROUND AND MOTIVATION

Machine learning models are typically trained on large datasets with the assumption that the training distribution closely matches the distribution of the deployment where the model will be applied. However, in real-world applications, deployment data distributions often diverge from general and/or

global training set distributions (Shen et al., 2024; Taori et al., 2020). Selecting relevant data subsets aligned with specific deployments is crucial to maximize field performance. The problem of *data subset selection for specialization* (DS3) is thus critical: given all available training data for a domain and a small (usually unlabeled) query set that represents the desired deployment, the goal is to identify a subset of the training data, such that training the ML model on this subset maximizes performance on the deployment distribution.

Real world example. Consider a wildlife ecologist who aims to build a classifier to detect the presence of invasive species in camera trap images collected at the Channel Islands. Existing labeled training data in this context is limited, thus training a classifier from scratch is likely to be unsuccessful. A common approach is to finetune a general pre-trained model (such as ViT or CLIP) on all *relevant* camera trap data. But *what does "relevant data" mean?* Would using similar species data from other camera trap locations (perhaps on the mainland) improve performance, or introduce noise? What about including data from non-similar species at that location? While adding data to a training set can sometimes improve performance, it can also decrease individual subgroup performance in a biased way (Compton et al., 2023) and introduce spurious correlations that can enable models to learn potentially dangerous "shortcuts," resulting in biased predictions, shown across various domains (Geirhos et al., 2020; Badgeley et al., 2018; Wang et al., 2021; Beery et al., 2022a).

Our contributions. Our key contributions are the following:

(i) We are the first to identify and formalize the challenge of sub-selecting training data to specialize models to new deployments (dataset subset selection for specialization).

(ii) We propose DATAS^3: A novel benchmark that enables the AI community to investigate and make progress on DS3. DATAS^3 reformulates, adapts, and adds to five diverse datasets, each from a different application domain. We worked directly with domain experts throughout the curation and reformulation process to ensure that DATAS^3 accurately reflects **(1)** real-world dataset distribution challenges that require model specialization (i.e., covariate shifts, subpopulation shifts, and long-tailed distributions), and **(2)** evaluation settings (test splits) representative of real-world deployment scenarios in each domain.

(iii) We show that a well-curated subset can consistently outperform models trained on the entire dataset for each deployment.

(iv) We also conduct an extensive experimental study comparing current SOTA subset selection methods on DATAS^3. After training a suite of baselines, our results clearly show that current subset selection methods fail on DS3, highlighting the need future research to solve the DS3 problem on DATAS^3.

(v) We release a codebase, python package, and public leaderboard for submission to the benchmark, available at `datas3-benchmark.github.io`

2 PROBLEM STATEMENT

DS3 problem formulation. Let X be a pool of data points, $T \subset X$ be a given *training set* drawn from a training (pool) distribution P_T over X, and let $Q \subset X$ be a *query set* drawn from the desired **deployment-specific distribution** P_Q over X. Given a model θ, the objective of **dataset subset selection for specialization (DS3)**, is to design an algorithm `SubsetSelection-ALG`, which takes T (the training set) and Q (the deployment representative query set) as input, and outputs a subset $S^* \subset T$ that minimizes the expected loss of θ trained on S^* over the desired deployment-specific distribution P_Q. More formally:

$$S^* = \arg\min_{S \subset T} \mathbb{E}_{q \sim P_Q} \left[\mathcal{L}(\theta(S), q) \right], \tag{1}$$

where $\theta(S)$ denotes the model trained on the subset $S \subset T$, and $\mathcal{L}(\theta(S), q)$ is the loss function evaluated on a single point q sampled from P_Q and the trained model $\theta(S)$. The term $\mathbb{E}_{q \sim P_Q}$ denotes the expected value over the distribution P_Q. Hence, the algorithm `SubsetSelection-ALG` outputs S^*, the subset of T that minimizes the expected loss of the entire desired deployment distribution P_Q. Notably, `SubsetSelection-ALG` can only access the desired deployment-specific distribution via the query set Q. Unlike complementary lines of work such as active domain

adaptation (ADA), which assumes real-time compute and focuses on actively/iteratively selecting and then collecting labels for data within the deployment during the specialization process, DS3 selects data in a single-shot approach prior to specialization on an already available pool of data (potentially for use in resource-constrained applications).

Is the query set labeled? This formalization can be divided into two cases. In the first, the query set Q is annotated with a set of labels: Q is a set of $m > 0$ pairs $Q = \{(q_1, y_1), \cdots, (q_m, y_m)\}$, where for every $i \in [m]$, q_i is the ith feature vector describing the ith input, and y_i is it corresponding label/annotation. In this case the algorithm `SubsetSelection-ALG` has access to the set of labels $\{y_1, \cdots, y_n\}$. In the second scenario, no labels are provided for Q, meaning that the `SubsetSelection-ALG` does not have access to the set $\{y_1, \cdots, y_n\}$ and consequently $Q = \{q_1, \cdots, q_m\}$. In a real-world example, Q can be thought of as the data collected from a deployment thus far, enabling additional selection from a larger database (the training pool). Annotating Q for any specific deployment is quite expensive, requiring time, money, and expertise, so progress on methods without query labels would helpful for real-world applications.

Is `SubsetSelection-ALG` model agnostic? Similarly, this formalization can be approached in two different ways: one where the computation of S^* depends on a given specific model θ, i.e., `SubsetSelection-ALG` is model dependent, and has access to the model θ we wish to train on. Ideally, a well-performing, robust method should work well for multiple models, and will be more generalizable than a model-dependent algorithm. We test several different models on our benchmark for various `SubsetSelection-ALG` baselines to test this.

Should `SubsetSelection-ALG` be sample efficient? The goal of our benchmark is to specialize on a desired deployment distribution. Unlike standard subset selection, where subset size is often a primary concern, our focus is on selecting subsets based on relevance based on a particular deployment that yield highest performance evaluated on that deployment. Smaller subsets offer many advantages, such as training efficiency, lower memory/storage, etc; we analyze these tradeoffs in Appendix C.

3 Related Work

It has become increasingly clear that data work is equally important to architecture design for increased model performance (Compton et al., 2023). Data curation for better quality training pools has been identified as an important line of research within this field. Many methods have been proposed for data curation and subset selection – we provide a comprehensive overview of these methods in Appendix B. Current benchmarks for data curation include Gadre et al. (2024), Mazumder et al. (2023) and Feuer et al. (2024). However, these benchmarks focus on data curation for a single higher quality training pool meant for better performance across many different downstream tasks, in contrast to specialization for a particular deployment. Additionally, data selection methods (Killamsetty et al., 2021b; Tukan et al., 2023) are often evaluated on standard CIFAR10/100 (Krizhevsky et al., 2009) or ImageNet (Deng et al., 2009b) datasets, where test and validation sets have similar distribution to their training sets. No existing benchmarks focus on the DS3 challenge. DataS^3 is the first benchmark specifically designed to evaluate subset selection methods for *deployment-specific specialization*, rather than generalization, where the training and testing data exhibit distributional shifts representative of real-world deployment challenges (Figure 1).

4 The DataS^3 Benchmark

Datasets. We describe each DataS^3 dataset. Our benchmark includes five datasets, each capturing a unique and diverse application of ML: AutoArborist for tree classification (Beery et al., 2022b), iWildCam for camera trap species identification (Beery et al., 2021), GeoDE for diverse object classification (Ramaswamy et al., 2023), NuScenes for autonomous driving footage steering regression (Caesar et al., 2020), and FishDetection for underwater video fish detection (Dawkins et al., 2017). To make the DataS^3 datasets usable, we have made considerable changes to them to better highlight deployment challenges, augment with additional data, or preprocess the data for use with standard ML pipelines. For each dataset, we provide a proof-of-concept "oracle" / knowledge-driven subset that demonstrates the usefulness of subset selection, with improvements over using training on all data. These subsets were created using information that benchmark users are not provided (e.g. metadata, GPS location, region, etc). Additional details about each dataset and can be found in Apdx. E.

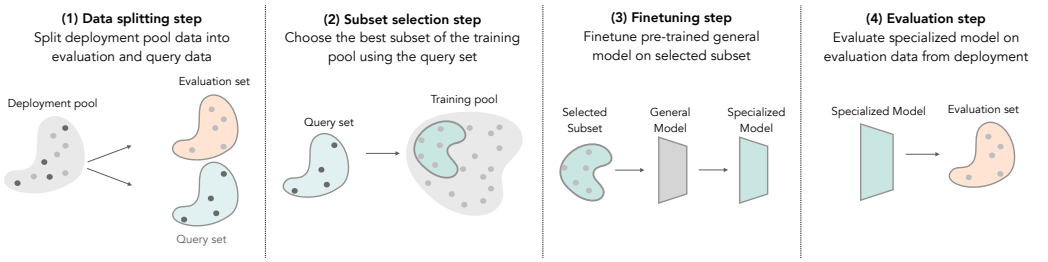

Figure 2: DATAS^{3}benchmark process, involving dataset splitting, subset selection, model specialization/finetuning, and then evaluation.

4.1 iWildCam

Motivation: Animal populations have declined by 68% on average since 1970 (Staub, 2020). To monitor this biodiversity loss, ecologists deploy camera traps—motion-activated cameras placed in the wild (Wearn & Glover-Kapfer, 2017)—and process the data with machine learning models (Norouzzadeh et al., 2019; Beery et al., 2019). However, variations in illumination, camera angle, background, vegetation, color, and animal frequencies across different locations cause these models to generalize poorly to new deployments. To specialize models for specific locations, selecting appropriate data subsets for deployment-specific (in this case location) specialization becomes essential.

Problem Setting & Data: To study this problem, we use the iWildCam 2020 dataset, comprising of $203,029$ images from 323 different camera traps spread across multiple countries in different parts of the world. The task is multi-class species classification from 182 different animal species. Performance is measured by overall classification accuracy for species identification. The original camera trap data comes from the Wildlife Conservation Society (link).

Deployments: Our deployments were defined to be split across camera trap locations to simulate the common scenario of researchers setting up new cameras within a region, with poor model generalization on the new cameras (Wearn & Glover-Kapfer, 2017). Our train/test split was done randomly across the 200 locations, with the five downstream test tasks created by clustering by the latitude and longitude of camera GPS location in 4 deployments: (1) Central America, (2) Eastern Africa, (3) Southern Africa, and (4) Southeast Asia. Similar to most other camera trap datasets, iWildCam has significant long-tailed label distributions, with variation in species and backgrounds between locations, as can be seen in Figure 3.

Knowledge-driven Subset: These subsets were created by only choosing training data from camera locations that are within 100km of the camera locations in the deployments (the relevant geographical area) and eliminating irrelevant classes that are not present in the deployment.

4.2 GeoDE

Motivation: Object classification datasets are often constructed by scraping images from the web but contain geographical biases (Shankar et al., 2017). Instead of scraping images from the web, GeoDE (Ramaswamy et al., 2023) crowdsources a dataset that is roughly balanced across 40 different objects and six world regions, showing that common objects (stoves, bicycles, etc), vary in appearance across the world. Accordingly, specializing models to different regions becomes useful when the objects have strong covariate shift.

Problem setting & Data: GeoDE is a diverse dataset of 61,490 images comprising 40 different objects collected from 6 world regions (Africa, Americas, East Asia, Europe, Southeast Asia, West Asia). The associated task is multiclass classification, where the goal is to predict the object depicted in each image.

Deployments: We propose 4 different deployments: (1) objects in Indonesia, (2) objects in Nigeria, (3) indoor objects, and (4) outdoor objects, as shown in Figure 3. Nigeria and Indonesia were selected as the two countries with the poorest performance, and the indoor/outdoor deployment tasks were

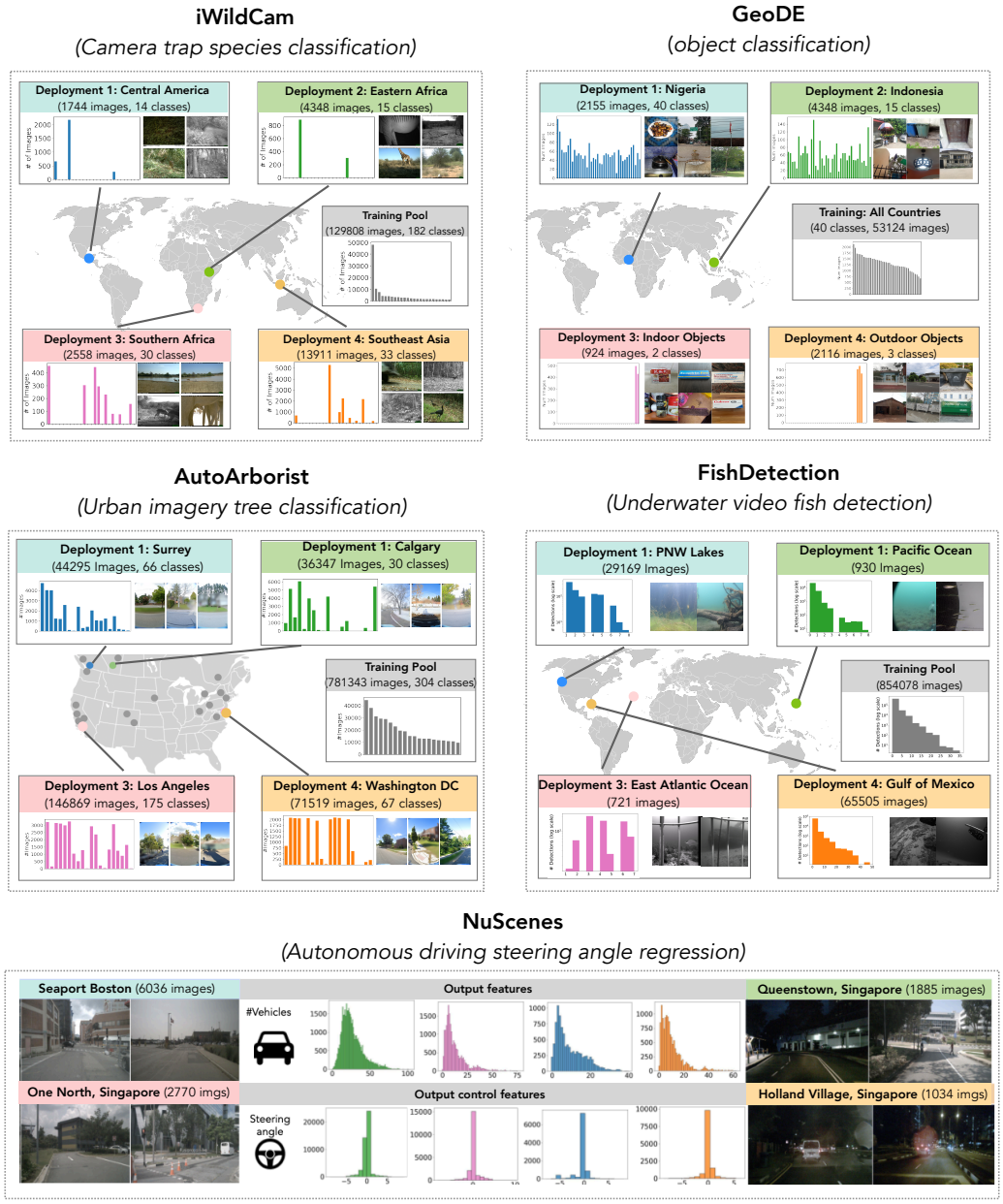

Figure 3: The five datasets in our benchmark: iWildCam, GeoDE, AutoArborist, FishDetection, and NuScenes each have real-world applications in deployment. In iWildCam, GeoDE, and AutoArborist, we show the class distributions of each deployment; in FishDetection, the number of detections per image is shown, and in NuScenes environment features. These diagrams show that each dataset has unique challenges in the deployments that lead to a need for model specialization, including long-tailedness (AutoArborist, iWildCam), covariate shift (all), subpopulation shifts (GeoDE, FishDetection), and more. These axes of variation are described in depth in Section 4 and further in Apdx E.

selected for enabling model specialization. The training dataset includes images from all countries, and the test data contains only images from Nigeria and Indonesia.

Knowledge-driven Subset: These subsets were generated by selecting data from the relevant countries/categories in the training data, ie. only selecting African subcontient data for the Nigeria deployment, Asian subcontinent data for the Indonesia deployment, and indoor/outdoor objects within the training pool for these deployments.

4.3 AUTOARBORIST

Motivation: Ecological imagery for environmental monitoring, such as automated tree classification, provides policymakers with critical, data-driven insights to support climate adaptation, urban planning, and more (Brandt et al., 2016). This task is associated with fundamental challenges such as noisy labels, non-iid data, fine-grained and long-tailed class distribution, and geospatial distribution shift. These challenges lead to a need for specialization of models where general-purpose models fail.

Problem Setting & Data: The AutoArborist dataset is a multi-view, fine-grained visual tree categorization dataset containing street-level images of over 1 million public zone trees from 300 genus-level categories across 23 major cities in the US and Canada.

Deployments: Deployments in AutoArborist correspond to the development models for use by individual cities. The deployment cities of (1) Surrey with 66 distinct tree genus classes, (2) Calgary with 30 classes, (3) Los Angeles with 175 classes, and (4) Washington DC with 67 classes were chosen due to their diverse climates, species distributions, and urban structures, as seen in Figure 3. Surrey and Calgary were treated as our in-distribution (ID) deployments, with some of these cities data in the training pool. Washington DC and LA were the out-of-distribution deployments, with no city data in the training pool.

Knowledge-driven Subset: We used the relevant data from Surrey and Calgary in the training pool for these ID deployments. Accordingly, we used data from San Francisco and San Jose for Los Angeles and Charlottesville, Pittsburgh, and New York for Washington DC. Label distribution shift and covariate shift are visualized in Figure 10 and 9, respectively.

4.4 FISHDETECTION

Motivation: Climate change, pollution, and overfishing continue to threaten marine biodiversity across the globe (United Nations, 2023; Di Lorenzo et al., 2022). Marine imagery is an increasingly common resource to monitor fish stocks and biodiversity. However, ML methods are difficult to apply across various environmental settings due to differences in lighting, turbidity, species, vegetation, camera sensors, etc. (Borremans et al., 2024; Jerlov, 1976; Akkaynak & Treibitz, 2019), creating a need for specialized models.

Problem Setting & Data: We use the public VIAME FishTrack23 dataset (Dawkins et al., 2017) consisting of 854,078 images across various environmental settings, ranging from freshwater rivers to deeper benthos. Specifically, the task is to predict bounding box localizations around every fish present in each image. Performance is measured by mAP across various IoU thresholds. Most of the images across all datasets are taken from video streams, and can be grouped as such, that were deployed primarily on camera traps, both baited and unbaited.

Deployments: Deployments are split according to the subsets of the VIAME dataset, which roughly correspond to geographic regions. Train, test and subset splits are either taken as provided or randomly sampled frames from each dataset, roughly corresponding to: (1) freshwater Pacific Northwest lakes; (2) Pacific Ocean; (3) East Atlantic Ocean; and (4) Gulf of Mexico.

Knowledge-driven Subset: For each deployment, we use the subset in the relevant geographical area (e.g., images from Gulf of Mexico for the Gulf of Mexico deployment).

4.5 NUSCENES

Motivation: End-to-end autonomous driving systems streamline vehicle control by directly mapping sensory inputs, such as images, to control outputs like steering angles (Wang et al., 2024). Adapting these systems to specialize in particular streets or environments is made easier as a single model

encompasses the full system. Thus, training this model to specialize in a specific environment brings advantages, capturing detailed local road layouts, traffic patterns, area-specific obstacles, and more.

Problem Setting & Data: We explore vision-based control for self-driving across diverse environments (e.g., different city areas) and driving scenarios (e.g., pedestrians crossing, construction zones), formulated as a regression task. This dataset includes $88,461$ images from the NuScenes dataset, subsampled from the image sweeps at a rate of 2. The images were captured from a video stream recorded while driving a car. Each image is paired with a steering angle control from the CAN bus, synchronized with the sensor timestamps of both the camera and CAN bus data. The model's goal is to predict a single scalar value representing the car's steering angle. Performance is evaluated in an open-loop manner using metrics like mean squared error.

Deployments: Deployments are organized by the geographic locations where the data was collected, including (1) Boston Seaport, (2) Singapore Holland Village, (3) Singapore One-North, and (4) Singapore Queenstown. While all tasks are based on expert demonstrations of driving and general driving behaviors, each location presents varying environmental features—such as vegetation, road types, roadside infrastructure, and weather—as well as differences in driving style and road regulations. Train/test splits are randomly sampled within each deployment.

Knowledge-driven subset: Since this training pool is a combination of the four deployment locations, we simply use the relevant location's data as the training subset. For example, we use the subset of the training pool with Boston Seaport data for the Boston Seaport deployment.

4.6 Benchmark Pipeline

To compete on our benchmark, models must select relevant data from the training pool and then finetune models on that relevant data. Explicitly, **(i)** given a small query set representing the deployment data (we consider both labeled and unlabeled query sets), curate a subset of data from the training for a specific deployment, **(ii)** finetune/train a fixed model on the chosen subset from the training pool and **(iii)** evaluate on the deployment (test) set (Figure 2).

For each dataset, we fix the training procedure for all subsets, fixing model architecture, optimizers, and loss functions. We match the label distribution of the query set to the deployment/test set as closely as possible using stratified sampling, but from each class of the training pool, we sample uniformly at random. We run a small hyperparameter sweep for each training subset across batch sizes $\{32, 64, 128\}$ and learning rates $\{0.01, 0.001, 0.0001\}$ for each deployment. For all classification/regression datasets, we use ResNet50 for full-finetuning (He et al., 2015) (table 1) and a ViT for LoRA finetuning (Apdx Table 3), as well as a ViT (Dosovitskiy et al., 2020) for linear probes (Apdx Table 2). For the detection dataset, we use a YOLOv8n model, using default parameters, though we subsample images to 640p. Full details are in Apdx D.

4.7 Metrics

Participants are evaluated across 12 deployments from five datasets, as outlined in Section 4. For the classification task datasets of GeoDE, AutoArborist, and iWildCam, we report accuracy for each deployment, for the regression task dataset NuScenes, we report mean squared error, and for the detection task FishDetection, we report mAP50. For each deployment, we evaluate participants of the benchmark on overall accuracy of training subset; we also report subset size – while the less data used the better, we mainly focus on optimal performance, in line with the DS3 formulation.

5 Baselines

We compare performance of dataset subset selection algorithms across our benchmark, across different scenarios: (a) access to an unlabeled query set, and (b) access to a labeled query set. We also curate a third category, (c), which leverages domain expertise to generate expert-selected subsets, in order to demonstrate the existence of better-than-all subsets for these deployments.

Non-subset comparisons:
No filtering: Performance of a model trained on the entire training pool, without any filtering.

Query Sets: As a comparison, we also include performance of a model trained directly on the labeled query set for each deployment. Note that this would require access to query labels, which are not always available. When labels are available, performance of models trained on the small query sets are often poor, hence the value of learning from larger-scale general-pool data. As a logistical point, none of the baselines we show in our results train on query set data.

Expert-Driven Subsets: We contribute curated, "expert knowledge" subsets using domain knowledge and/or metadata. We find these knowledge-guided subsets often outperform using all samples in the training pool (no filtering). The creation of these subsets is described per-dataset in section 4.

Unlabeled-query baselines:
Image-alignment (Image-Align): We take the cosine similarity between the training and query embedding space, using examples that exceed a threshold for at least x samples, where x is a hyperparameter chosen from {1,10,100}.

Nearest neighbors features (Near-Nbors): To better align our method with the downstream deployment, we explore using examples whose embedding space overlaps with the query set of data. To do so, we cluster image embeddings extracted by an OpenAI ViT model for each image into 1000 clusters using Faiss (Johnson et al., 2019). Then, we find the nearest neighbor clusters for every query set example and keep the training cluster closest to each query set cluster. This method was inspired by the similar DataComp baseline (Gadre et al., 2024).

Labeled-query baselines:
CLIP score filtering (CLIP-score): We also experiment with CLIP score filtering, using examples that exceed a threshold for cosine similarity between CLIP image and text similarity. Text for each image was created with manual captioning (e.g. for iWildCam, *"This is a camera trap image of a lion taken at time 10-2-2016 at 04:26:13 in Nigeria"*). We select the subset that exceeds a threshold of CLIP-score similarity, with the threshold calculated for subsets that make up $25\%, 50\%$, 75%, and 90% of the dataset.

Matching relative frequency (Match-Dist): We explore having access to the relative frequency of each label in the downstream deployment. For example, a domain expert at a national park might know the relative frequency of species (deployment-specific domain knowledge). We create subsets by sampling 25%, 50%, 75%, and 90% of the training pool to match the label distribution of the deployment.

Matching labels (Match-Label): Similarly, a domain expert may know the classes present in the downstream deployment. For example, a domain expert at a national park might know the species present (deployment-specific knowledge) that we can utilize for subset selection. For these subsets, we simply remove the classes present in the training pool that are not present in the testing pool.

6 RESULTS

Well chosen subsets outperform training on all data. The knowledge-driven subsets in Table 1 show that deployment-specific well-chosen subsets of the data can significantly outperform models trained on all the data, with improvements in deployment accuracy up to 3.6% for GeoDE, 11.9% for iWildCam, 51.3% for AutoArborist, a 0.03 reduction in MSE for NuScenes, and 0.13 increase in mAP50 for FishDetection. Even when the knowledge-driven subsets underperform all training data, as in NuScenes Deployment 2, there exist subsets from other baselines that outperform using all the data. In Appendix E, we provide a additional breakdown of the key factors that contributed to performance gain on these knowledge-driven subsets.

Training on more data has diminishing returns. For all deployments, we see that we can achieve near-optimal performance with subsets of the data. The knowledge-driven subsets are significantly smaller than the total training data size, with the average percentage of the total training pool used being 4% for GeoDE, 11% for iWildCam, 8% for AutoArborist, 10% for NuScenes, and 20% for FishDetection. Appendix C shows that even 25% of the data can perform near-optimally in some cases, with little performance loss with 50% of the data on the algorithmic baselines. Overall, these results demonstrate that greater efficiency for training specialized ML models is possible, potentially reducing computational and data storage burdens in deployable settings.

Dataset	Metric	Deploy #	Non subset		Knowledge-driven	Unlabeled query set		Labeled query set		
			Query-set	All-data		Image-Align	Near-Nbors	CLIP-score	Match-Label	Match-Dist
GeoDE	Acc (#)	Deploy 1	0.87 (500)	0.89 (53k)	**0.92 (2.9k)**	0.88 (26k)	0.88 (48k)	0.89 (40k)	0.88 (53k)	0.89 (48k)
		Deploy 2	0.45 (500)	0.89 (53k)	**0.91 (2.6k)**	0.90 (26k)	0.89 (48k)	0.90 (40k)	0.90 (53k)	0.88 (27k)
		Deploy 3	0.95 (500)	0.82 (53k)	**0.85 (1.4k)**	0.85 (24k)	0.76 (48k)	0.84 (40k)	0.83 (1.4k)	0.88 (48k)
		Deploy 4	0.83 (500)	0.83 (53k)	**0.85 (2.6k)**	0.79 (24k)	0.78 (48k)	0.83 (40k)	0.84 (2.6k)	0.83 (13k)
iWildCam	Acc (#)	Deploy 1	0.70 (301)	0.66 (130k)	0.65 (8.5k)	0.56 (36k)	0.50 (117k)	0.50 (97k)	0.74 (8.1k)	**0.74 (117k)**
		Deploy 2	0.78 (302)	0.34 (130k)	0.35 (9.2k)	0.44 (45k)	0.47 (98k)	0.46 (97k)	0.35 (55k)	**0.49 (65k)**
		Deploy 3	0.44 (301)	0.72 (130k)	0.75 (19k)	0.54 (24k)	0.45 (98k)	0.42 (97k)	0.72 (60k)	**0.75 (117k)**
		Deploy 4	0.46 (309)	0.66 (130k)	0.67 (21k)	0.60 (22k)	0.60 (33k)	0.29 (97k)	0.69 (57k)	**0.74 (33k)**
AutoArborist	Acc (#)	Deploy 1	0.16 (1.5k)	0.35 (781k)	**0.86 (70k)**	0.38 (44k)	0.39 (391k)	0.38 (47k)	0.67 (368k)	0.74 (703k)
		Deploy 2	0.20 (1.5k)	0.48 (781k)	**0.86 (123k)**	0.11 (49k)	0.14 (703k)	0.14 (47k)	0.65 (532k)	0.56 (391k)
		Deploy 3	0.12(1.5k)	0.16 (781k)	**0.38 (35k)**	0.16 (46k)	0.10 (703k)	0.17 (47k)	0.16 (534k)	0.23 (703k)
		Deploy 4	0.12 (1.5k)	0.14 (781k)	**0.39 (26k)**	0.10 (48k)	0.11 (391k)	0.11 (47k)	0.10 (527k)	0.23 (195k)
NuScene	MSE (#)	Deploy 1	0.063 (6.0k)	0.050 (100k)	**0.029 (20k)**	0.040 (35k)	0.040 (90k)	0.073 (31k)	-	-
		Deploy 2	0.070 (1.0k)	0.021 (100k)	0.049 (4.6k)	0.15 (17k)	0.042 (90k)	**0.032(31k)**	-	-
		Deploy 3	0.089 (2.7k)	0.068 (100k)	**0.038 (10k)**	0.049 (28k)	0.13 (90k)	0.071 (31k)	-	-
		Deploy 4	0.12 (1.9k)	0.048 (100k)	**0.039 (7.0k)**	0.086 (26k)	0.39 (90k)	0.050 (31k)	-	-
FishDetection	mAP50 (#)	Deploy 1	0.22 (500)	0.68 (841k)	**0.69 (179k)**	0.50 (630k)	0.60 (103k)	-	-	-
		Deploy 2	0.26 (600)	0.32 (841k)	**0.45 (152k)**	0.31 (630k)	0.40 (120k)	-	-	-
		Deploy 3	0.13 (541)	0.32 (841k)	**0.39 (6.0k)**	0.28 (630k)	0.23 (204k)	-	-	-
		Deploy 4	0.079 (519)	0.59 (841k)	**0.60 (320k)**	0.54 (630k)	0.39 (45k)	-	-	-

Table 1: Best-performing subsets across hyperparameters for baseline methods across all datasets and deployments (abbreviated as Deploy) for YOLOv8 full-finetuning for FishDetection and ResNet50 full-finetuning for the rest. Accuracy is reported for the classification tasks of GeoDE, iWildCam, and Auto Arborist, mAP50 for FishDetection (greater is better), and MSE for NuScenes (smaller is better). We include subset size in parentheses. We include results for **ViT LoRA finetuning and ViT linear probes** in Appendix C in Table 3 and Table 2, which display similar trends. Match-Dist and Match-Label are not applicable for NuScenes, as it is a regression task and does not have clear classes/labels for these methods. FishDetection only uses the unlabeled query set, as the ground truth is bounding boxes, rather than labels themselves. Baselines are distinguished from one another by their access to information, with each baseline having access to expert knowledge, or a labeled/unlabeled query set. We do not report the random baseline in this table, but demonstrate results in Appendix C as it mainly refers to subset size. For each deployment, there exists a subset that outperforms training on all data, indicated in bold.

Methods without access to supervision perform poorly. While the knowledge-driven subsets in Table 1 demonstrate that a well-chosen subset *does exist* for all deployments, finding this subset without extra knowledge is still an open problem. Some of our baselines require access to query labels, this requirement can in many cases be unrealistic in the deployable ML setting (labels can be expensive or difficult to collect). The two unsupervised baselines, the nearest neighbors and image alignment methods, do not perform optimally on the deployments, often underperforming using all the training data. Our benchmark opens up the line of research for potential unsupervised methods for this data subset selection process.

7 Discussion and Conclusions

DataS^3 is the first benchmark to promote the development of dataset subset selection methods capable of specialization to diverse real-world deployments. The benchmark is both open-source and easy to use, lowering the barrier to entry for this new important problem.

DataS^3 highlights open challenges for the research community. In our experimental study, we show that there is no winning baseline that performs well across multiple domains/datasets. Additionally, while some methods perform well when given access to labeled query sets, no methods perform well in the unsupervised setting. Finally, some datasets are more challenging than others– methods may need to specifically target different types of distribution shifts.

DataS^3 has value beyond subset selection. In addition to the DS3 problem, DataS^3 can be used as a testbed for various other complementary lines of work, such as domain adaptation, active learning, coreset selection, and more. We highlight these relevant methods in Appendix B.

Extensions to other domains. Model specialization for deployments isn't limited to the domains we include. We are open to expanding this benchmark to capture more scientific domains, and welcome further dataset contributions from the broader ML and scientific research community.

REFERENCES

Derya Akkaynak and Tali Treibitz. Sea-thru: A method for removing water from underwater images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1682–1691, 2019.

Anonymous. When less is more: Investigating data pruning for pretraining LLMs at scale. In *NeurIPS Workshop on Attributing Model Behavior at Scale*, 2023. URL https://openreview.net/forum?id=XUIYn3jo5T.

Koichiro Asano, Akira Hebisawa, Takashi Ishiguro, Noboru Takayanagi, Yasuhiko Nakamura, Junko Suzuki, Naoki Okada, Jun Tanaka, Yuma Fukutomi, Shigeharu Ueki, Koichi Fukunaga, Satoshi Konno, Hiroto Matsuse, Katsuhiko Kamei, Masami Taniguchi, Terufumi Shimoda, and Tsuyoshi Oguma. New clinical diagnostic criteria for allergic bronchopulmonary aspergillosis/mycosis and its validation. *The Journal of allergy and clinical immunology*, 2020. URL https://api.semanticscholar.org/CorpusID:221673224.

Marcus A. Badgeley, John R. Zech, Luke Oakden-Rayner, Benjamin Scott Glicksberg, Manway Liu, William Gale, Michael V. McConnell, Bethany Percha, Thomas M. Snyder, and Joel T. Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digital Medicine*, 2, 2018. URL https://api.semanticscholar.org/CorpusID:53250171.

Fred Bane, Celia Soler Uguet, Wiktor Stribiżew, and Anna Zaretskaya. A comparison of data filtering methods for neural machine translation. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pp. 313–325, Orlando, USA, September 2022. Association for Machine Translation in the Americas. URL https://aclanthology.org/2022.amta-upg.22.

Cenk Baykal, Lucas Liebenwein, Igor Gilitschenski, Dan Feldman, and Daniela Rus. Sensitivity-informed provable pruning of neural networks. *SIAM J. Math. Data Sci.*, 4(1):26–45, 2022.

Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review, 2019. URL https://arxiv.org/abs/1907.06772.

Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset, 2021.

Sara Beery, Guanhang Wu, Trevor Edwards, Filip Pavetic, Bohdan S. Majewski, Shreyasee Mukherjee, Stanley Chan, John Morgan, Vivek Rathod, and Jonathan Huang. The auto arborist dataset: A large-scale benchmark for multiview urban forest monitoring under domain shift. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21262–21275, 2022a. URL https://api.semanticscholar.org/CorpusID:250476238.

Sara Meghan Beery, Guanhang Wu, Trevor Edwards, Filip Pavetić, Bo Majewski, Shreyasee Mukherjee, Stan Chan, John Morgan, Vivek Mansing Rathod, and Jonathan Chung-kuan Huang. The auto-arborist dataset: A large-scale benchmark for generalizable, multimodal urban forest monitoring. 2022b.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 95–136, virtual+Dublin, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bigscience-1.9. URL https://aclanthology.org/2022.bigscience-1.9.

Catherine Borremans, Jennifer Durden, Timm Schoening, Emma Curtis, Luther Adams, Alexandra Branzan Albu, Aurélien Arnaubec, Sakina-Dorothée Ayata, Reshma Baburaj, Corinne Bassin, et al. Report on the marine imaging workshop 2022. *Research Ideas and Outcomes*, 10:e119782, 2024.

Leslie A. Brandt, Abigail Derby Lewis, Robert T. Fahey, Lydia Scott, Lindsay E. Darling, and Christopher W. Swanston. A framework for adapting urban forests to climate change. *Environmental Science & Policy*, 66:393–402, 2016. URL https://api.semanticscholar.org/CorpusID:156304225.

Vladimir Braverman, Dan Feldman, and Harry Lang. New frameworks for offline and streaming coreset constructions. *arXiv preprint arXiv:1612.00889*, 2016.

Vladimir Braverman, Petros Drineas, Cameron Musco, Christopher Musco, Jalaj Upadhyay, David P. Woodruff, and Samson Zhou. Near optimal linear algebra in the online and sliding window models. In *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS*, pp. 517–528, 2020.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. URL https://api.semanticscholar.org/CorpusID:218971783.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Ke Chen. On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. *SIAM J. Comput.*, 39(3):923–947, 2009.

Rachit Chhaya, Anirban Dasgupta, and Supratim Shit. On coresets for regularized regression. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, 2020.

Kenneth L. Clarkson. Coresets, sparse greedy approximation, and the frank-wolfe algorithm. *ACM Trans. Algorithms*, 6(4):63:1–63:30, 2010.

Michael B. Cohen, Cameron Musco, and Christopher Musco. Input sparsity time low-rank approximation via ridge leverage score sampling. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pp. 1758–1777, 2017.

Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic, Chris Schwiegelshohn, and Omar Ali Sheikh-Omar. Improved coresets for euclidean k-means. *Advances in Neural Information Processing Systems*, 35:2679–2694, 2022.

Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations*, 2019.

Rhys Compton, Lily H. Zhang, Aahlad Manas Puli, and Rajesh Ranganath. When more is less: Incorporating additional datasets can hurt performance by introducing spurious correlations. *ArXiv*, abs/2308.04431, 2023. URL https://api.semanticscholar.org/CorpusID:265819574.

Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W. Mahoney. Sampling algorithms and coresets for l_p regression. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pp. 932–941, 2008.

Matthew Dawkins, Linus Sherrill, Keith Fieldhouse, Anthony Hoogs, Benjamin Richards, David Zhang, Lakshman Prasad, Kresimir Williams, Nathan Lauffenburger, and Gaoang Wang. An open-source platform for underwater image and video analytics. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 898–906, 2017. doi: 10.1109/WACV.2017.105.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009a. doi: 10.1109/CVPR.2009.5206848.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009b. doi: 10.1109/CVPR.2009.5206848.

Emanuele Di Lorenzo, Christian Lønborg, Jesper H Andersen, Elva G Escobar-Briones, Michelle Jillian Devlin, Angel Borja, Marius Nils Müller, Carol Robinson, Alex Ford, Anna Milena Zivian, et al. *Sustainable Development Goal 14-Life Below Water: Towards a Sustainable Ocean*. Frontiers Media SA, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. URL https://api.semanticscholar.org/CorpusID:225039882.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7949–7962, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.638. URL https://aclanthology.org/2020.emnlp-main.638.

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia. A brief review of domain adaptation, 2020. URL https://arxiv.org/abs/2010.03978.

Benjamin Feuer, Jiawei Xu, Niv Cohen, Patrick Yubeaton, Govind Mittal, and Chinmay Hegde. Select: A large-scale benchmark of data curation strategies for image classification. *arXiv preprint arXiv:2410.05057*, 2024.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665 – 673, 2020. URL https://api.semanticscholar.org/CorpusID:215786368.

Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pp. 291–300, 2004.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015. URL https://api.semanticscholar.org/CorpusID:206594692.

Duojun Huang, Jichang Li, Weikai Chen, Jun Steed Huang, Zhenhua Chai, and Guanbin Li. Divide and adapt: Active domain adaptation via customized learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7651–7660, 2023. URL https://api.semanticscholar.org/CorpusID:260067836.

Lingxiao Huang and Nisheeth K. Vishnoi. Coresets for clustering in euclidean spaces: importance sampling is nearly optimal. In *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC*, pp. 1416–1429, 2020.

Nils Gunnar Jerlov. *Marine optics*, volume 14. Elsevier, 1976.

Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, January 2023. URL https://github.com/ultralytics/ultralytics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Ibrahim Jubran, Murad Tukan, Alaa Maalouf, and Dan Feldman. Sets clustering. In *International Conference on Machine Learning*, pp. 4994–5005. PMLR, 2020.

KrishnaTeja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, Abir De, and Rishabh K. Iyer. GRAD-MATCH: gradient matching based data subset selection for efficient deep model training. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, pp. 5464–5474, 2021a.

KrishnaTeja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh K. Iyer. GLISTER: generalization based data subset selection for efficient and robust learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, 2021b.

Krishnateja Killamsetty, Xujiang Zhao, Feng Chen, and Rishabh Iyer. Retrieve: Coreset selection for efficient and robust semi-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021c.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etiene David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Meghan Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2020. URL https://api.semanticscholar.org/CorpusID:229156320.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009.

Alina Kuznetsova, Hassan Rom, Neil Gordon Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4. *International Journal of Computer Vision*, 128:1956 – 1981, 2018. URL https://api.semanticscholar.org/CorpusID:53296866.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. The bigscience roots corpus: A 1.6tb composite multilingual dataset, 2023.

Lucas Liebenwein, Cenk Baykal, Harry Lang, Dan Feldman, and Daniela Rus. Provable filter pruning for efficient neural networks. In *International Conference on Learning Representations*, 2019.

Alaa Maalouf, Ibrahim Jubran, and Dan Feldman. Fast and accurate least-mean-squares solvers. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 8307–8318, 2019.

Alaa Maalouf, Ibrahim Jubran, Murad Tukan, and Dan Feldman. Coresets for the average case error for finite query sets. *Sensors*, 21(19):6689, 2021.

Alaa Maalouf, Gilad Eini, Ben Mussay, Dan Feldman, and Margarita Osadchy. A unified approach to coreset learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Juan Ciro, Lora Aroyo, Bilge Acun, Lingjiao Chen, Mehul Smriti Raje, Max Bartolo, Sabri Eyuboglu, Amirata Ghorbani, Emmett Goodman, Oana Inel, Tariq Kane, Christine R. Kirkpatrick, Tzu-Sheng Kuo, Jonas Mueller, Tristan

Thrush, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen Paritosh, Lilith Bat-Leah, Ce Zhang, James Zou, Carole-Jean Wu, Cody Coleman, Andrew Ng, Peter Mattson, and Vijay Janapa Reddi. Dataperf: Benchmarks for data-centric ai development, 2023. URL https://arxiv.org/abs/2207.10062.

Raphael A. Meyer, Cameron Musco, Christopher Musco, David P. Woodruff, and Samson Zhou. Fast regression for structured inputs. In *The Tenth International Conference on Learning Representations, ICLR*, 2022.

Sören Mindermann, Jan Brauner, Muhammed Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N. Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt, 2022.

Baharan Mirzasoleiman, Jeff A. Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, pp. 6950–6960, 2020a.

Baharan Mirzasoleiman, Kaidi Cao, and Jure Leskovec. Coresets for robust training of deep neural networks against noisy labels. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2020b.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models, 2023.

Mohammad Sadegh Norouzzadeh, Dan Morris, Sara Beery, Neel Joshi, Nebojsa Jojic, and Jeff Clune. A deep active learning system for species identification and counting in camera trap images, 2019. URL https://arxiv.org/abs/1910.09716.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2021.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only, 2023.

Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8485–8494, 2020. URL https://api.semanticscholar.org/CorpusID:224714171.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis and insights from training gopher, 2022.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.

Vikram V. Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron B. Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Geode: a geographically diverse evaluation dataset for object recognition, 2023. URL https://arxiv.org/abs/2301.02560.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115: 211 – 252, 2014. URL https://api.semanticscholar.org/CorpusID:2930547.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*, 2018.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv: Machine Learning*, 2017. URL https://api.semanticscholar.org/CorpusID:26262581.

Judy Hanwen Shen, Inioluwa Deborah Raji, and Irene Y. Chen. The data addition dilemma, 2024. URL https://arxiv.org/abs/2408.04154.

Shoaib Ahmed Siddiqui, Nitarshan Rajkumar, Tegan Maharaj, David Krueger, and Sara Hooker. Metadata archaeology: Unearthing data subsets by leveraging training dynamics, 2022.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *arXiv*, 2022.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning, 2023.

Francis Staub. Living planet report 2020: Bending the curve of biodiversity loss, Sep 2020. URL https://icriforum.org/living-planet-report-2020-bending-the-curve-of-biodiversity-loss/.

Alex Tamkin, Dat Nguyen, Salil Deshpande, Jesse Mu, and Noah Goodman. Active learning helps pretrained models learn the intended task, 2022.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *ArXiv*, abs/2007.00644, 2020. URL https://api.semanticscholar.org/CorpusID:220280805.

Elad Tolochinsky, Ibrahim Jubran, and Dan Feldman. Generic coreset for scalable learning of monotonic kernels: Logistic regression, sigmoid and more. In *International Conference on Machine Learning, ICML*, 2022.

Murad Tukan, Cenk Baykal, Dan Feldman, and Daniela Rus. On coresets for support vector machines. *Theor. Comput. Sci.*, 890:171–191, 2021.

Murad Tukan, Loay Mualem, and Alaa Maalouf. Pruning neural networks via coresets and convex geometry: Towards no assumptions. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022.

Murad Tukan, Samson Zhou, Alaa Maalouf, Daniela Rus, Vladimir Braverman, and Dan Feldman. Provable data subset selection for efficient neural networks training. In *International Conference on Machine Learning*, pp. 34533–34555. PMLR, 2023.

United Nations. Agreement under the United Nations Convention on the Law of the Sea on the Conservation and Sustainable Use of Marine Biological Diversity of Areas Beyond National Jurisdiction. United Nations Treaty Collection, 2023. Opened for signature on 20 September 2023.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35:8052–8072, 2021. URL https://api.semanticscholar.org/CorpusID:232110832.

Tsun-Hsuan Wang, Alaa Maalouf, Wei Xiao, Yutong Ban, Alexander Amini, Guy Rosman, Sertac Karaman, and Daniela Rus. Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6687–6694. IEEE, 2024.

Xinyi Wang, Hieu Pham, Paul Michel, Antonios Anastasopoulos, Jaime Carbonell, and Graham Neubig. Optimizing data usage via differentiable rewards. In *International Conference on Machine Learning (ICML)*, 2020.

Xudong Wang, Long Lian, and Stella X Yu. Unsupervised selective labeling for more effective semi-supervised learning. In *European Conference on Computer Vision*, pp. 427–445. Springer, 2022.

Oliver Wearn and Paul Glover-Kapfer. Camera-trapping for conservation: a guide to best-practices, 10 2017.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4003–4012, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.494.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7935–7948, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.637. URL https://aclanthology.org/2020.emnlp-main.637.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024.