# 8

# *Datasets*

It's become commonplace to point out that machine learning models are only as good as the data they're trained on. The old slogan "garbage in, garbage out" no doubt applies to machine learning practice, as does the related catchphrase "bias in, bias out". Yet, these proverbs still understate—and somewhat misrepresent—the significance of data for machine learning.

It's not only the output of a learning algorithm that may suffer with poor input data. A dataset serves many other vital functions in the machine learning ecosystem. The dataset itself is an integral part of the problem formulation. It implicitly sorts out and operationalizes what the problem is that practitioners end up solving. Datasets have also shaped the course of entire scientific communities in their capacity to measure and benchmark progress, support competitions, and interface between researchers in academia and practitioners in industry.

If so much hinges on data in machine learning, it might come as a surprise that there is no simple answer to the question of what makes data good for what purpose. The collection of data for machine learning applications has not followed any established theoretical framework, certainly not one that was recognized a priori.

In this chapter, we take a closer look at popular datasets in the field of machine learning and the benchmarks that they support. We will use this to tease apart the different roles datasets play in scientific and engineering contexts. Then we will review the harms associated with data and discuss how they can be mitigated based on the dataset's role. We will conclude with several broad directions for improving data practices.

We limit the scope of this chapter in some important ways. Our focus will be largely on publicly available datasets that support training and testing purposes in machine learning research and applications. Our focus excludes large swaths of industrial data collection, surveillance, and data mining practices. It also excludes data purposefully collected to test specific scientific hypotheses, such as, experimental data gathered in a medical trial.

## *A tour of datasets in different domains*

The creation of datasets in machine learning does not follow a clear theoretical framework. Datasets aren't collected to test a specific scientific hypothesis. In fact, we will see that there are many different roles data plays in machine learning. As a result, it makes sense to start by looking at a few influential datasets from different domains to get a better feeling for what they are, what motivated their creation,

how they organized communities, and what impact they had.

*TIMIT*

Automatic speech recognition is a machine learning problem of significant commercial interest. Its roots date back to the early 20th century.[1]

Interestingly, speech recognition also features one of the oldest benchmarks data sets, the TIMIT (Texas Instruments/Massachusetts Institute for Technology) data. The creation of the dataset was funded through a 1986 DARPA program on speech recognition. In the mid-eighties, artificial intelligence was in the middle of a "funding winter" where many governmental and industrial agencies were hesitant to sponsor AI research because it often promised more than it could deliver. DARPA program manager Charles Wayne proposed that a way around this problem was establishing more rigorous evaluation methods. Wayne enlisted the National Institute of Standards and Technology to create and curate shared datasets for speech, and he graded success in his program based on performance on recognition tasks on these datasets.

Many now credit Wayne's program with kick starting a revolution of progress in speech recognition.[2][3][4] According to Kenneth Ward Church,

> It enabled funding to start because the project was glamour-and-deceit-proof, and to continue because funders could measure progress over time. Wayne's idea makes it easy to produce plots which help sell the research program to potential sponsors. A less obvious benefit of Wayne's idea is that it enabled hill climbing. Researchers who had initially objected to being tested twice a year began to evaluate themselves every hour.

A first prototype of the TIMIT dataset was released in December of 1988 on a CD-ROM. An improved release followed in October 1990. TIMIT already featured the training/test split typical for modern machine learning benchmarks. There's a fair bit we know about the creation of the data due to its thorough documentation.[5]

TIMIT features a total of about 5 hours of speech, composed of 6300 utterances, specifically, 10 sentences spoken by each of 630 speakers. The sentences were drawn from a corpus of 2342 sentences such as the following.

```
She had your dark suit in greasy wash water all year. (sa1)
Don't ask me to carry an oily rag like that. (sa2)
This was easy for us. (sx3)
Jane may earn more money by working hard. (sx4)
She is thinner than I am. (sx5)
Bright sunshine shimmers on the ocean. (sx6)
Nothing is as offensive as innocence. (sx7)
```

The TIMIT documentation distinguishes between 8 major dialect regions in the United States, documented as *New England, Northern, North Midland, South Midland,*

*Southern, New York City, Western, Army Brat (moved around).* Of the speakers, 70% are male and 30% are female. All native speakers of American English, the subjects were primarily employees of Texas Instruments at the time. Many of them were new to the Dallas area where they worked.

Racial information was supplied with the distribution of the data and coded as "White", "Black", "American Indian", "Spanish-American", "Oriental", and "Unknown". Of the 630 speakers, 578 were identified as White, 26 as Black, 2 as American Indian, 2 as Spanish-American, 3 as Oriental, and 17 as unknown.

Table 1: Demographic information about the TIMIT speakers

| | Male | Female | Total (%) |
|---|---|---|---|
| White | 402 | 176 | 578 (91.7%) |
| Black | 15 | 11 | 26 (4.1%) |
| American Indian | 2 | 0 | 2 (0.3%) |
| Spanish-American | 2 | 0 | 2 (0.3%) |
| Oriental | 3 | 0 | 3 (0.5%) |
| Unknown | 12 | 5 | 17 (2.6%) |

The documentation notes:

> In addition to these 630 speakers, a small number of speakers with foreign accents or other extreme speech and/or hearing abnormalities were recorded as "auxiliary" subjects, but they are not included on the CD-ROM.

It comes to no surprise that early speech recognition models had significant demographic and racial biases in their performance.

Today, several major companies, including Amazon, Apple, Google, and Microsoft, all use speech recognition models in a variety of products from cell phone apps to voice assistants. There is no longer a major open benchmark that would support training models competitive with the industrial counterparts. Industrial speech recognition pipelines are generally complex and use proprietary data sources that we don't know a lot about. Nevertheless, today's speech recognition systems continue to exhibit performance disparities along racial lines.[6]

### *UCI Machine Learning Repository*

The UCI Machine Learning Repository currently hosts more than 500 datasets, mostly for different classification and regression tasks. Most datasets are relatively small, consisting of a few hundred or a few thousand instances. The majority are structured tabular data sets with a handful or a few tens of attributes.

The UCI Machine Learning Repository contributed to the adoption of the train-test paradigm in machine learning in the late 1980s. Pat Langley recalls:

> The experimental movement was aided by another development. David Aha, then a PhD student at UCI, began to collect data sets for use in empirical studies of machine learning. This grew into the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/), which he made available to the community by FTP in 1987. This was rapidly adopted by many researchers because it was easy to use and because it let them compare their results to previous findings on the same tasks.[7]

The most popular dataset in the repository is the Iris Data Set containing taxonomic measurements of 150 iris flowers, 50 from each of 3 species. The task is to classify the species given the measurements.

As of October 2020, the second most popular dataset in the UCI repository is the *Adult* dataset. Extracted from the 1994 Census database, it features nearly 50,000 instances describing individuals in the United States, each having 14 attributes. The task is to classify whether an individual earns more than 50,000 US dollars or less. The Adult dataset remains popular in the algorithmic fairness community, largely because it is one of the few publicly available datasets that features demographic information including *gender* (coded in binary as male/female), as well as *race* (coded as Amer-Indian-Eskimo, Asian-Pac-Islander, Black, Other, and White).

Unfortunately, the data has some idiosyncrasies that make it less than ideal for understanding biases in machine learning models. Due to the age of the data, and the income cutoff at $50,000, almost all instances labeled *Black* are below the cutoff, as are almost all instances labeled *female*. Indeed, a standard logistic regression model trained on the data achieves about 85% accuracy overall, while the same model achieves 91% accuracy on Black instances, and nearly 93% accuracy on female instances. Likewise, the ROC curves for the latter two groups enclose actually more area than the ROC curve for male instances. This is an atypical situation: more often, machine learning models perform worse on historically disadvantaged groups.

### *MNIST*

The MNIST dataset contains images of handwritten digits. Its most common version has 60,000 training images and 10,000 test images, each having 28x28 black and white pixels.

MNIST was created by researchers Burges, Cortes, and Lecun from an earlier dataset released by the National Institute of Standards and Technology (NIST). The dataset was introduced in a research paper in 1998 to showcase the use of gradient-based deep learning methods for document recognition tasks.[8] Since then cited over 30,000 times, MNIST became a highly influential benchmark in the computer vision community. Two decades later, researchers continue to use the data actively.

The original NIST data had the property that training and test data came from two different populations. The former featured the handwriting of two thousand American Census Bureau employees, whereas the latter came from five hundred

Figure 1: A sample of MNIST digits

American high school students.[9] The creators of MNIST reshuffled these two data sources and split them into training and test set. Moreover, they scaled and centered the digits. The exact procedure to derive MNIST from NIST was lost, but recently reconstructed by matching images from both data sources.[10]

The original MNIST test set was of the same size as the training set, but the smaller test set became standard in research use. The 50,000 digits in the original test set that didn't make it into the smaller test set were later identified and dubbed *the lost digits*.[10]

From the beginning, MNIST was intended to be a benchmark used to compare the strengths of different methods. For several years, LeCun maintained an informal leaderboard on a personal website that listed the best accuracy numbers that different learning algorithms achieved on MNIST.

Table 2: A snapshot of the original MNIST leaderboard from February 2, 1999. Source: Internet Archive (Retrieved: December 4, 2020)

| Method | Test error (%) |
|---|---|
| linear classifier (1-layer NN) | 12.0 |
| linear classifier (1-layer NN) [deskewing] | 8.4 |
| pairwise linear classifier | 7.6 |
| K-nearest-neighbors, Euclidean | 5.0 |
| K-nearest-neighbors, Euclidean, deskewed | 2.4 |
| 40 PCA + quadratic classifier | 3.3 |

| Method | Test error (%) |
|---|---|
| 1000 RBF + linear classifier | 3.6 |
| K-NN, Tangent Distance, 16x16 | 1.1 |
| SVM deg 4 polynomial | 1.1 |
| Reduced Set SVM deg 5 polynomial | 1.0 |
| Virtual SVM deg 9 poly [distortions] | 0.8 |
| 2-layer NN, 300 hidden units | 4.7 |
| 2-layer NN, 300 HU, [distortions] | 3.6 |
| 2-layer NN, 300 HU, [deskewing] | 1.6 |
| 2-layer NN, 1000 hidden units | 4.5 |
| 2-layer NN, 1000 HU, [distortions] | 3.8 |
| 3-layer NN, 300+100 hidden units | 3.05 |
| 3-layer NN, 300+100 HU [distortions] | 2.5 |
| 3-layer NN, 500+150 hidden units | 2.95 |
| 3-layer NN, 500+150 HU [distortions] | 2.45 |
| LeNet-1 [with 16x16 input] | 1.7 |
| LeNet-4 | 1.1 |
| LeNet-4 with K-NN instead of last layer | 1.1 |
| LeNet-4 with local learning instead of ll | 1.1 |
| LeNet-5, [no distortions] | 0.95 |
| LeNet-5, [huge distortions] | 0.85 |
| LeNet-5, [distortions] | 0.8 |
| Boosted LeNet-4, [distortions] | 0.7 |

In its capacity as a benchmark, it became a showcase for the emerging kernel methods of the early 2000s that temporarily achieved top performance on MNIST.[11] Today, it is not difficult to achieve less than 0.5% classification error with a wide range of convolutional neural network architectures. The best models classify all but a few pathological test instances correctly. As a result, MNIST is widely considered too easy for today's research tasks.

MNIST wasn't the first dataset of handwritten digits in use for machine learning research. Earlier, the US Postal Service (USPS) had released a dataset of 9298 images (7291 for training, and 2007 for testing). The USPS data was actually a fair bit harder to classify than MNIST. A non-negligible fraction of the USPS digits look unrecognizable to humans,[12] whereas humans recognize essentially all digits in MNIST.

### *ImageNet*

ImageNet is a large repository of labeled images that has been highly influential in computer vision research over the last decade. The image labels correspond to nouns from the WordNet lexical database of the English language.[13] WordNet groups nouns into cognitive synonyms, called *synsets*. The words *car* and *automobile*, for example, would fall into the same synset. On top of these categories WordNet

provides a hierarchical tree structure according to a super-subordinate relationship between synsets. The synset for *chair*, for example, is a child of the synset for *furniture* in the wordnet hierarchy. WordNet existed before ImageNet and in part inspired the creation of Imagenet.

The initial release of ImageNet included about 5000 image categories, each corresponding to a synset in WordNet. These ImageNet categories averaged about 600 images per category.[14] ImageNet grew over time and its Fall 2011 release had reached about 32,000 categories.

The construction of ImageNet required two essential steps: retrieving candidate images for each synset, and labeling the retrieved images. This first step utilized online search engines and photo sharing platforms with a search interface, specifically, Flickr. Candidate images were taken from the image search results associated with the synset nouns for each category.

For the second labeling step, the creators of ImageNet turned to Amazon's Mechanical Turk platform (MTurk). MTurk is an online labor market that allows individuals and corporations to hire on-demand workers to perform simple tasks. In this case, MTurk workers were presented with candidate images and had to decide whether or not the candidate image was indeed an image corresponding to the category that it was putatively associated with.

It is important to distinguish between this ImageNet database and a popular machine learning benchmark and competition, called ImageNet Large Scale Visual Recognition Challenge (ILSVRC), that was derived from it.[15] The competition was organized yearly from 2010 until 2017, reaching significant notoriety in both industry and academia, especially as a benchmark for emerging deep learning models.

When machine learning practitioners say "ImageNet" they typically refer to the data used for the image classification task in the 2012 ILSVRC benchmark. The competition included other tasks, such as object recognition, but image classification has become the most popular task for the dataset. Expressions such as "a model trained on ImageNet" typically refer to training an image classification model on the benchmark data set from 2012.

Another common practice involving the ILSVRC data is *pre-training*. Often a practitioner has a specific classification problem in mind whose label set differs from the 1000 classes present in the data. It's possible nonetheless to use the data to create useful features that can then be used in the target classification problem. Where ILSVRC enters real-world applications it's often to support pre-training.

This colloquial use of the word ImageNet can lead to some confusion, not least because the ILSVRC-2012 dataset differs significantly from the broader database. It only includes a subset of 1000 categories. Moreover, these categories are a rather skewed subset of the broader ImageNet hierarchy. For example, of these 1000 categories only three are in the *person* branch of the WordNet hierarchy, specifically, *groom*, *baseball player*, and *scuba diver*. Yet, more than 100 of the 1000 categories correspond to different dog breeds. The number is 118, to be exact, not counting wolves, foxes, and wild dogs that are also present among the 1000 categories.

What motivated the exact choice of these 1000 categories is not entirely clear.

The apparent canine inclination, however, isn't just a quirk either. At the time, there was an interest in the computer vision community in making progress on prediction with many classes, some of which are very similar. This reflects a broader pattern in the machine learning community. The creation of datasets is often driven by an intuitive sense of what the technical challenges are for the field. In the case of ImageNet, another important consideration was scale, both in terms of the number of images and the number of classes.

The large scale annotation and labeling that went into Imagenet falls into a category of labor that Gray and Suri call *ghost work* in their book of the same name.[16] They point out:

> MTurk workers are the AI revolution's unsung heroes.

Indeed, ImageNet was labeled by about 49,000 MTurk workers from 167 countries over the course of multiple years.

### *The Netflix Prize*

The Netflix Prize was one of the most famous machine learning competitions. Starting on October 2, 2006, the competition ran for nearly three years ending with a grand prize of $1M, announced on September 18, 2009. Over the years, the competition saw 44,014 submissions from 5169 teams.

The Netflix training data contained roughly 100 million movie ratings from nearly 500 thousand Netflix subscribers on a set of 17770 movies. Each data point corresponds to a tuple `<user, movie, date of rating, rating>`. At about 650 megabytes in size, the dataset was just small enough to fit on a CD-ROM, but large enough to be pose a challenge at the time.

The Netflix data can be thought of as a matrix with $n = 480189$ rows and $m = 17770$ columns. Each row corresponds to a Netflix subscriber and each column to a movie. The only entries present in the matrix are those for which a given subscriber rated a given movie with rating in $\{1, 2, 3, 4, 5\}$. All other entries—that is, the vast majority—are missing. The objective of the participants was to predict the missing entries of the matrix, a problem known as matrix completion, or collaborative filtering somewhat more broadly. In fact, the Netflix challenge did so much to popularize this problem that it is sometimes called the Netflix problem. The idea is that if we could predict missing entries, we'd be able to recommend unseen movies to users accordingly.

The hold out data that Netflix kept secret consisted of about three million ratings. Half of them were used to compute a running leaderboard throughout the competition. The other half determined the final winner.

The Netflix competition was hugely influential. Not only did it attract significant participation, it also fueled much academic interest in collaborative filtering for years to come. Moreover, it popularized the competition format as an appealing way for companies to engage with the machine learning community. A startup called Kaggle, founded in April 2010, organized hundreds of machine learning

competitions for various companies and organizations before its acquisition by Google in 2017.

But the Netflix competition became infamous for another reason. Although Netflix had replaced usernames by pseudonymous numbers, researchers Narayanan and Shmatikov were able to re-identify some of the Netflix subscribers whose movie ratings were in the dataset[17] by linking those ratings with publicly available movie ratings on IMDB, an online movie database. Some Netflix subscribers had also publicly rated an overlapping set of movies on IMDB under their real identities. In the privacy literature, this is called a *linkage attack* and it's one of the ways that seemingly anonymized data can be de-anonymized.[18]

What followed were multiple class action lawsuits against Netflix, as well as an inquiry by the Federal Trade Commission over privacy concerns. As a consequence, Netflix canceled plans for a second competition, which it had announced on August 6, 2009.

To this day, privacy concerns are a legitimate obstacle to public data release and dataset creation. Deanonymization techniques are mature and efficient. There provably is no algorithm that could take a dataset and provide a rigorous privacy guarantee to all participants, while being useful for all analyses and machine learning purposes. Dwork and Roth call this the Fundamental Law of Information Recovery: *"overly accurate answers to too many questions will destroy privacy in a spectacular way."*[19]

## *Roles datasets play*

In machine learning research and engineering, datasets play a different and more prominent set of roles than they do in most other fields. We have mentioned several of these above but let us now examine them in more detail. Understanding these is critical to figuring out which technical and cultural aspects of benchmarks are essential, how harms arise, and how to mitigate them.

### *A source of real data*

Edgar Anderson was a botanist and horticulturist who spent much of the 1920s and '30s collecting and analyzing data on Irises to study biological and taxonomic questions. The Iris dataset in the UCI machine learning repository mentioned above is the result of Anderson's labors — or a tiny sliver of them, as most of the observations in the dataset came from a single day of field work. The dataset contains 50 observations each of 3 iris plants; the task is to distinguish the species based on 4 physical attributes (sepal length and width; petal length and width). Most of the tens of thousands of researchers who have used this dataset are not interested in taxonomy, let alone irises. What, then, are they using the dataset for?

Although the data was collected by Anderson, it was actually published in the paper "The use of multiple measurements in taxonomic problems" by Ronald Fisher, who was a founder of modern statistics as well as a eugenicist.[20] The eugenics connection is not accidental: other central figures in the development of modern

statistics such as Francis Galton and Karl Pearson were algo eugenicists.[21,22] Fisher was Anderson's collaborator. Although Fisher had some interest in taxonomy, he was primarily interested in using the data to develop statistical techniques (with an eye toward applications for eugenics). In the 1936 paper, Fisher introduces Linear Discriminant Analysis (LDA) and shows that it performs well on this task.

The reason the Iris dataset proved to be a good application of LDA is that there exists a linear projection of the four features which seems to result in a mixture of Gaussians (one for each of the three species), and the means of the three distributions are relatively far apart; one of the species is in fact perfectly separable from the other two. Every learning algorithm implicitly makes assumptions about the data-generating process: without assumptions, there is no basis for making predictions on unseen points.[23] If we could perfectly mathematically describe the data generating process behind the physical characteristics of irises (or any other population), we wouldn't need a dataset — we could mathematically work out how well an algorithm would perform. In practice, for complex phenomena, such perfect mathematical descriptions rarely exist. Different communities place different value on attempting to discover the true data generating process. Machine learning places relatively little emphasis on this goal.[24] Ultimately, the usefulness of a learning algorithm is established by testing it on real datasets.

The reliance on benchmark datasets as a source of real data was a gradual development in machine learning research. For example, Rosenblatt's perceptron experiments in the 1950s used two artificial stimuli (the characters E and X), with numerous variants of each created by rotation and other transformations.[25] The controlled input was considered useful to understand the behavior of the system. Writing in 1988, Pat Langley advocates for a hybrid approach, pointing out that "successful runs on a number of different natural domains provide evidence of generality" but also highlighting the use of artificial data for better understanding.[26] Especially after the establishment of the UCI repository around this time, it has become common to evaluate new algorithms on widely-used benchmark datasets as a way of establishing that the researcher is not "cheating" by picking contrived inputs.

To summarize, when a researcher seeks to present evidence that an algorithmic innovation is useful, the use of real dataset as opposed to artificial data ensures that the researcher didn't make up data to suit the algorithm. Further, the use of prominent benchmark datasets wards off skepticism that the researcher may have cherry picked a dataset with specific properties that makes the algorithm effective. Finally, the use of multiple benchmark datasets from different domains suggests that the algorithm is highly general.

Perversely, domain ignorance is treated almost as a virtue rather than a drawback. For example, researchers who achieve state-of-the-art performance on (say) Chinese-to-English translation may point out that none of them speak Chinese. The subtext is that they couldn't have knowingly or unknowingly picked a model that works well only when the source language is linguistically similar to Chinese.

*A catalyst and measure of domain-specific progress*

Algorithmic innovations that are highly portable across domains, while important, are rare. Much of the progress in machine learning is instead tailored to specific domains and problems. The most common way to demonstrate such progress is to show that the innovation in question can be used to achieve "state of the art" performance on a benchmark dataset for that task.

The idea that datasets spur algorithmic innovation bears some explanation. For example, the Netflix Prize is commonly credited as responsible for the discovery of the effectiveness of matrix factorization in recommender systems (often attributed to Simon Funk, a pseudonymous contestant[27]). Yet, the technique had been proposed in the context of movie recommendation as early as 1998[28] and for search as early as 1990.[29] However, it was not previously apparent that it outperformed neighborhood-based methods and that it could discover meaningful latent factors. The clarity of the Netflix leaderboard and the credibility of the dataset helped establish the significance of matrix factorization.[30]

Somewhat separately from the role of spurring algorithmic innovation, benchmark datasets also offer a convenient way to measure its results (hence the term benchmark). The progression of state-of-the-art accuracy on a benchmark dataset and task can be a useful indicator. A relatively flat curve of accuracy over time may indicate that progress has stalled, while a discontinuous jump may indicate a breakthrough. Reaching an error rate that is close to zero or at least lower than the "human error" for perception tasks is often considered a sign that the task is "solved" and that it is time for the community to move on to a harder challenge.

While these are appealing heuristics, there are also pitfalls. In particular, a statement such as "the state of the art accuracy for image classification is 95%" is not a scientifically meaningful claim that can be assigned a truth value, because the number is highly sensitive to the data distribution.

A notable illustration of this phenomenon comes from a paper by Recht, Roelofs, Schmidt, and Shankar. They carefully recreated new test sets for the CIFAR-10 and ImageNet classification benchmarks according to the very same procedure as the original test sets.[31] They then took a large collection of representative models proposed over the years and evaluated all of them on the new test sets. All models suffered a significant drop in performance on the new test set, corresponding to about 5 years of progress in image classification. They found that this is because the new test set represents a slightly different distribution. This is despite the researchers' careful efforts to replicate the data collection procedure; we should expect that test sets created by different procedures should result in much greater performance differences.

The same graphs also provide a striking illustration of why benchmark datasets are a practical necessity for performance comparison in machine learning. Consider a hypothetical alternative approach analogous to the norm in many other branches of science: a researcher evaluating a claim (algorithm) describes in detail their procedure for sampling the data; other researchers working on the same problem sample their own datasets based on the published procedure. Some reuse of

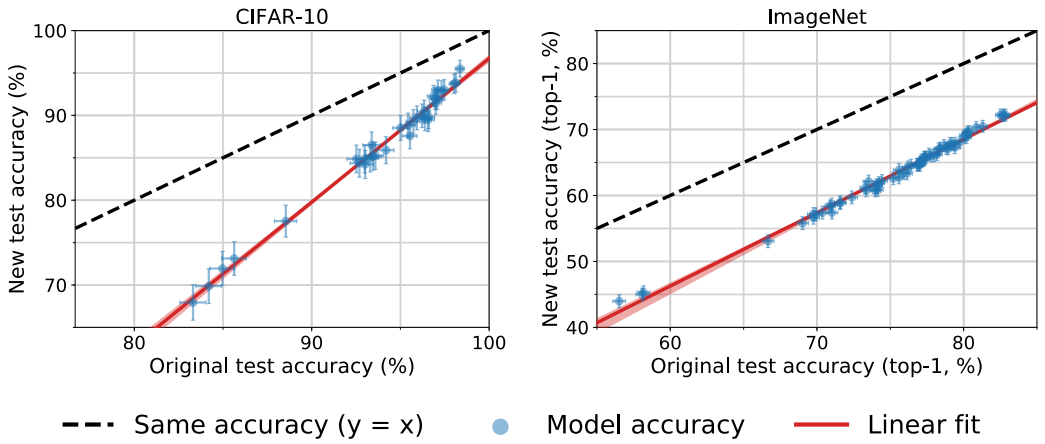


Figure 2: Model accuracy on the original test sets vs. new test sets for CIFAR-10 and ImageNet. Each data point corresponds to one model in a test bed of representative models (shown with 95% Clopper-Pearson confidence intervals). The plots reveal two main phenomena: (i) There is generally a significant drop in accuracy from the original to the new test sets. (ii) The model accuracies closely follow a linear function, meaning that models that perform well on the old test set also tend to perform well on the new test set. The narrow shaded region is a 95% confidence region for the linear fit.

datasets occurs, but there is no standardization. The graphs show that even extremely careful efforts to sample a new dataset from the same distribution would shift the distribution sufficiently to make performance comparison hopeless.

In other words, reported accuracy figures from benchmark datasets do not constitute generalizable scientific knowledge, because they don't have external validity beyond the specific dataset. While the Recht et al. paper is limited to image classification, it seems scientifically prudent to assume a lack of external validity for other machine learning tasks as well, unless there is evidence to the contrary. Yet the two graphs above hint at a different type of knowledge that seems to transfer almost perfectly to the new test set: the *relative* performance of models. Indeed, another paper showed evidence that relative performance is stable on many datasets across a much wider range of distribution shifts, with strong correlations between in-domain and out-of-domain performance.[32]

The relative performance of models for a given task is a very useful type of practitioner-oriented knowledge that can be gained from benchmark leaderboards. A question that practitioners often face is, "which class of models should I use for [given task] and how should I optimize it"? A benchmark dataset (together with the associated task definition) can be seen as a proxy for answering this question in a constrained setting, analogous to laboratory studies in other branches of science. The hope is that algorithms (and model classes or architectures) identified as state of the art based on benchmark evaluation are also the ones that will be effective on the practitioner's test set. In other words, practitioners can outsource the laborious

task of model selection to the benchmark leaderboard.

To be clear, this is an oversimplification. Practitioners have many concerns in addition to accuracy such as the computational cost (of both training and prediction), interpretability, and, increasingly, fairness and environmental cost. Thus, benchmark performance is useful to practitioners but far from the only consideration for model selection.

We can imagine a spectrum of how similar the new test set is to the benchmark set. At the one extreme, if the new test set is truly a new sample from the exact same distribution, then the ranking of model classes should be the same for the two sets. At the other extreme, the distributions may be so different that they constitute essentially different tasks, so that performance on one is not a useful guide to performance on the other. In between these extremes is a big grey area that is not well understood, and it is currently more art than science.

The lack of clarity on how much we can generalize from one or a few benchmarks is associated with well known controversies. For example, support vector machines were competitive with neural networks on earlier-generation benchmarks such as NIST digit recognition,[33] which was one reason why interest in neural networks dwindled in the 1990s. The clear superiority of neural networks on newer benchmarks such as ImageNet was only belatedly recognized.[1]

### *A source of (pre-)training data*

Above, we have envisioned that practitioners use the benchmark leaderboard as a guide to model selection but then train the selected models from scratch on their own (often proprietary) data sources. But practitioners often can and do go further.

In some cases, it may be possible to train on a benchmark dataset and directly use the resulting model in one's application. This depends on the domain and the task, and is more suitable when the distribution shift is minimal and the set of class labels is stable. For example, it is reasonable to deploy a digit recognizer pretrained on MNIST, but not so much an image classifier pretrained on ILSVRC (without some type of adaptation to the target domain). Indeed, ILSVRC consists of a rather arbitrary subset of 1,000 classes of ImageNet, and a pretrained model is correspondingly limited in the set of labels it is able to output. The ImageNet Roulette project was a telling demonstration of what happens when a model trained on the (full) ImageNet dataset is applied to a different test distribution, one consisting primarily of images of people. The results were grotesque. The demonstration has been discontinued, but many archived results may be found in articles about the project.[35] Finally, consider a recommendation system benchmark dataset. There is no way to even attempt to use it directly as training data because the users about whom one wants to make predictions are highly unlikely to be present in the training set.

In most cases, the creators of benchmark datasets do not intend them to be used as a source of training data, although benchmark datasets are often misused for

[1] The difficulty of ascertaining the extent to which a study's findings generalize beyond the studied population bedevils all of the statistical sciences. See, for instance.[34]

this purpose. A rare exception is The Pile: a large (800 GB) English text corpus that is explicitly targeted at training language models. To improve the generalization capabilities of models trained on this corpus, the authors included diverse text from 22 different sources.[36]

Even when benchmark datasets are not useful as training data for the above-mentioned reasons, they can be useful as pre-training data for transfer learning. Transfer learning refers to using an existing model as a starting point for building a new model. A new model may be needed because the data distribution has shifted compared to what the existing model was optimized for, or because it aims to solve a different task altogether. For example, a model pre-trained on ImageNet (or ILSVRC) may be adapted via further training for recognizing different species (distribution shift) or as part of an image captioning model (a different task).

There are different intuitions to explain why transfer learning is often effective. One is that the final layers of a neural network correspond to semantically high-level representations of the input. Pre-training is a way of learning these representations that tend to be useful for many tasks. Another intuition is that pre-training is a way of initializing weights that offers an improvement over random initialization in that it requires fewer samples from the target domain for convergence.

Pretraining offers the practical benefit of being able to share the knowledge contained in a dataset without releasing the raw data. Many datasets, especially those created by companies using customer data, cannot be published due to privacy or confidentiality concerns. The release of pretrained models is thus an important avenue of knowledge sharing from industry to academia. Sharing pretrained models is also helpful to users for whom training from scratch is cost prohibitive. However, privacy and data protection concerns surface in the context of sharing pretrained models due to the possibility that personal data used for training can be recontructed from the pretrained model.[37]

Let's wrap up our analysis of the roles of benchmark datasets. We identified six distinct roles: (1) providing data sampled from real-world occurring distributions that enables largely domain-agnostic investigations of learning algorithms; (2) enabling domain-specific progress by providing datasets that are representative of real-world tasks in that domain yet abstract away unnecessary detail; (3) providing a convenient albeit crude numerical way to track scientific progress on a problem; (4) enabling model comparison and allowing practitioners to outsource model selection to public leaderboards; (5) providing a source of pre-training data for representation learning, weight initialization, etc; (6) providing a source of training data. The progression of these six roles is generally toward increasing domain- and task-specificity, and from science-oriented to practice-oriented.

### *The scientific basis of machine learning benchmarks*

Now we examine a seeming mystery: whether and why the benchmark approach works despite the practice of repeated testing on the same data.

Methodologically, much of modern machine learning practice rests on a variant of trial and error, which we call the train-test paradigm. Practitioners repeatedly

build models using any number of heuristics and test their performance to see what works. Anything goes as far as training is concerned, subject only to computational constraints, so long as the performance looks good in testing. Trial and error is sound so long as the testing protocol is robust enough to absorb the pressure placed on it. We will examine to what extent this is the case in machine learning.

From a theoretical perspective, the best way to test the performance of a classifier is to collect a sufficiently large fresh dataset and to compute the average error on that test set. Data collection, however, is a difficult and costly task. In most applications, practitioners cannot sample fresh data for each model they would like to try out. A different practice has therefore become the de-facto standard. Practitioners split their dataset into typically two parts, a *training set* used for training a model, and a *test set* used for evaluating its performance.[2] Often the split is determined when the dataset is created. Datasets used for benchmarks in particular have one fixed split persistent throughout time. A number of variations on this theme go under the name *holdout method*.

Machine learning competitions have adopted the same format. The company Kaggle, for example, has organized hundreds of competitions since it was founded. In a competition, a holdout set is kept secret and is used to rank participants on a public leaderboard as the competition unfolds. In the end, the final winner is whoever scores highest on a separate secret test set not used to that point.

In all applications of the holdout method the hope is that the test set will serve as a fresh sample that provides good performance estimates for all the models. The central problem is that practitioners don't just use the test data once only to retire it immediately thereafter. The test data are used incrementally for building one model at a time while incorporating feedback received previously from the test data. This leads to the fear that eventually models begin to *overfit* to the test data. This type of overfitting is sometimes called *adaptive overfitting* or *human-in-the-loop overfitting*.

Duda, Hart, and Stork summarize the problem aptly in their 1973 textbook:[38]

> In the early work on pattern recognition, when experiments were often done with very small numbers of samples, the same data were often used for designing and testing the classifier. This mistake is frequently referred to as "testing on the training data." A related but less obvious problem arises when a classifier undergoes a long series of refinements guided by the results of repeated testing on the same data. This form of "training on the testing data" often escapes attention until new test samples are obtained.

Nearly half a century later, Hastie, Tibshirani, and Friedman still caution in the 2017 edition of their influential textbook:[39]

> Ideally, the test set should be kept in a "vault," and be brought out only at the end of the data analysis. Suppose instead that we use the test-set

[2]Sometimes practitioners divide their data into multiple splits, e.g., training, validation, and test sets. However, for our discussion here that won't be necessary.

> repeatedly, choosing the model with smallest test-set error. Then the test set error of the final chosen model will underestimate the true test error, sometimes substantially.

While the suggestion to keep the test data in a "vault" is safe, it couldn't be further from the reality of modern practice. Popular test datasets often see tens of thousands of evaluations.

Yet adaptive overfitting doesn't seem to be happening. Recall the scatter plots by Recht et al. above: the plots admit a clean linear fit with positive slope. In other words, the better a model is on the old test set, the better it is on the new test set. But notice that newer models, i.e., those with higher performance on the original test set, had *more* time to adapt to the test set and to incorporate more information about it. Nonetheless, the better a model performed on the old test set the better it performs on the new set. Moreover, on CIFAR-10 we even see clearly that the absolute performance drops diminishes with increasing accuracy on the old test set. In particular, if our goal was to do well on the new test set, seemingly our best strategy is to continue to inch forward on the old test set.

The theoretical understanding of why machine learning practice has not resulted in overfitting is still catching up. Here, we highlight one of many potential explanations, called the leaderboard principle. It is a subtle effect in which publication biases force researchers to chase state-of-the-art results, and they only publish models if they see significant improvements over prior models. This cultural practice can be formalized by Blum & Hardt's *Ladder algorithm*. For each given classifier, it compares the classifier's holdout error to the previously smallest holdout error achieved by any classifier encountered so far. If the error is below the previous best by some margin, it announces the holdout error of the current classifier and notes it as the best seen so far. Importantly, if the error is not smaller by a margin, the algorithm releases the previous best (rather than the new error). It can be proven that the Ladder algorithm avoids overfitting in the sense that it accurately measures the error of the best performing classifier among those encountered.[40]

### *Benchmark praxis and culture*

The above discussion hints at the importance of cultural practices for a full understanding of benchmark datasets. Let us now discuss these in more detail, highlighting both dataset creators and users. These practices have helped make the benchmark-oriented approach successful but also impact the harms associated with data. Let's start with creators.

Benchmark creators define the task. This involves, among other things, selecting the high-level problem, defining the target variable, the procedure for sampling the data, and the scoring function. If manual annotation of the data is necessary, the dataset creator must develop a codebook or rubric for doing so and orchestrate crowd-work if needed. Data cleaning to ensure high-quality labels is usually required.

In defining the task, benchmark developers navigate a tricky balance: a task that is seen as too easy using existing techniques will not spur innovation while a

task that is seen as too hard may be demotivating. Finding the sweet spot requires expertise, judgment, and some luck. If the right balance is achieved, the benchmark drives progress on the problem. In this way, benchmark creators play an outsized role in defining the vision and agenda for machine learning communities. The selection of tasks in benchmarks is known to affect the ranking of models, which influences and biases the direction of progress in the community.[41] This effect may be getting more pronounced over time due to increasing concentration on fewer datasets.[42]

As an example of the kinds of decisions benchmark developers must make, and how they influence the direction of research, consider MNIST. As discussed above, it was derived from a previous dataset released by NIST in which the training and test set were drawn from different sources, but MNIST eliminated this distribution shift. The MNIST creators argued that this was necessary because

> Drawing sensible conclusions from learning experiments requires that the result be independent of the choice of training set and test among the complete set of samples.

In other words, if an algorithm performs well on NIST it is unclear how much of this due to its ability to learn the training distribution and how much of it is due to its ability to ignore the differences between the train and test distributions. MNIST allows researchers to focus selectively on the former question. This was a fruitful approach in 1995. Decades later, when problems like MNIST classification are effectively solved, the attention of benchmark dataset creators has turned towards methods for handling distribution shift that LeCun et al. justifiably chose to ignore.[43]

Another tricky balance is between abstracting away domain details so that the task is approachable for a broad swath of machine learning experts, and preserving enough details so that the methods that work in the benchmark setting will translate to production settings. One reason the Netflix Prize was so popular is because the data is just a matrix, and it is possible to achieve good performance (in the sense of beating Netflix's baseline) without really thinking about what the data means. No understanding of film or user psychology was necessary — or helpful, as it turned out. It is possible that domain expertise would have proved essential if the problem had been formulated differently — say, to require explainability or emphasize good performance even for users with very few previous ratings.

Another challenge for dataset creators is to avoid leakage. In an apocryphal story from the early days of computer vision, a classifier was trained to discriminate between images of Russian and American tanks with seemingly high accuracy, but it turned out that this was only because the Russian tanks had been photographed on a cloudy day and the American ones on a sunny day.[44] Data leakage refers to a spurious relationship between the feature vector and the target variable that is an artifact of the data collection or sampling strategy. Since the spurious relationship won't be present when the model is deployed, leakage usually leads to inflated estimates of model performance. Kaufman et al. present an overview of leakage in machine learning.[45]

Yet another critical responsibility of benchmark dataset creators is to implement a train-test framework. Most contests have various restrictions in place in an attempt to prevent both accidental overfitting to the leaderboard test set and intentional reverse engineering. Although, as we described above, benchmark praxis differs from the textbook version of the holdout method, practitioners have arrived at a set of techniques that have worked in practice, even if our theoretical understanding of why they work is still catching up.

Taking a step back, in any scientific endeavor there are the difficult tasks of framing the problem, ensuring that the methods have internal and external validity, and interpreting the results. Benchmark dataset creators handle as many of these hard tasks as possible, simplifying the goal of dataset users to the point where if a researcher beats the state-of-the-art performance, there is a good chance that there is a scientific insight somewhere in the methods, although extracting what this insight is may still require work. Further simplifying things for dataset users is the fact that there are no restrictions other than computational constraints on how the researcher uses the training data, as long as performance on the test set looks good.

To be clear, this approach has many pitfalls. Researchers rarely perform the statistical hypothesis tests needed to have confidence in the claim that one model performs better than another.[46] Our understanding of how to account for the numerous sources of variance in these performance measurements is still evolving; a 2021 paper that aims to do so argues that many of the claims of State-of-the-Art performance in natural language performance and computer vision don't hold up when subjected to such tests.[47]

There have long been articles noting the limitations of what researchers and practitioners can learn from benchmark performance evaluation.[48,49] David Aha, co-creator of the UCI repository, recalls that these limitations were well understood as early as 1995, just a few years after the repository was established.[50]

While it is important to acknowledge the limitations, it is also worth highlighting that this approach works at all. One reason for this success is that the scientific questions are primarily about algorithms and not the populations that the datasets are sampled from.

Indeed, there is a case to be made that other scientific communities should adopt the machine learning community's approach, sometimes called the Common Task Method.[2] Diverse scientific fields including economics, political science, psychology, genetics, and many others have seen an infusion of machine learning methods alongside a new focus on maximizing predictive accuracy as a research objective. These shifts have been accompanied by a rash of reproducibility failures, with large fractions of published papers falling prey to pitfalls such as data leakage.[51] Use of the benchmark dataset approach could have avoided most of these pitfalls.

Now let us transition to dataset users. Benchmark users have embraced the freedom afforded by the approach. As a result, the community of users is large — for example, the data science platform Kaggle has over 5 million registered users of whom over 130,000 have participated in a competition. There is less gatekeeping in machine learning research than in other disciplines. Many prominent findings bypass peer review. If a technique performs well on the leaderboard, that is

considered to speak for itself. Many people who contribute these findings are not formally affiliated with research institutions.

Overall, the culture of progress in machine learning combines the culture of academic scholarship, engineering, and even gaming, with a community of hobbyists and practitioners sharing tips and tricks on forums and engaging in friendly competition. This freewheeling culture may seem jarring to some observers, especially given the sensitivity of some of the datasets involved. The lack of gatekeeping means fewer opportunities for ethical training.

There is another aspect of benchmark culture that amplifies the harms associated with data: collecting data without informed consent and distributing it widely without adequate context. Many modern datasets, especially in computer vision and natural language processing, are scraped from the web. In such cases, it is infeasible to obtain informed consent from the individual authors of the content. What about a dataset such as the Netflix Prize where a company releases data from its own platform? Even if companies disclose in their terms of service that data might be used for research, it is doubtful that informed consent has been obtained since few users read and understand Terms of Service documents and because of the complexity of the issues involved.

When an individual's data becomes part of a benchmark dataset, it gets distributed widely. Popular benchmark datasets are downloaded by thousands of researchers, students, developers, and hobbyists. Scientific norms also call for the data to be preserved indefinitely in the interest of transparency and reproducibility. Thus, not only might individual pieces of data in these datasets be distributed and viewed widely, they are viewed in a form that strips them of their original context. A joke in bad taste written on social media and later deleted may be captured alongside documents from the library of congress.

## *Harms associated with data*

Now we will discuss a few important types of harms associated with benchmark datasets and how to mitigate them. We don't mean to imply that all of these harms are the "fault" of dataset creators, but understanding how data plays into these harms will bring clarity on how to intervene.

### *Downstream and representational harms*

A dataset's downstream harms are those that arise from the models trained on it. This is a type of harm that readily comes to mind: bad data may lead to bad models which can cause harm to the people they purportedly serve. For instance, biased criminal risk prediction systems disproportionately harm Black, minority, and overpoliced populations among others.

Properties of datasets that sometimes (but not always, and not in easily predictable ways) propagate downstream include imbalance, biases, stereotypes, and categorization. By imbalance we mean unequal representation of different groups.

For example, Buolamwini and Gebru pointed out that two facial analysis benchmarks, IJB-A and Adience, overwhelmingly featured lighter-skinned subjects.[52] By dataset biases we mean incorrect associations, especially those corresponding to social and historical prejudices. For example, a dataset that measures arrests as a proxy for crime may reflect the biases of policing and discriminatory laws. By stereotypes we mean associations that accurately reflect a property of the world (or a specific culture at a specific point in time) that is thought to be the result of social and historical prejudice. For example, gender-occupation associations can be called stereotypes. By categorization we mean assigning discrete (often binary) labels to complex aspects of identity such as gender and race.

Representational harms occur when systems reinforce the subordination of some groups along the lines of identity. Representational harms could be downstream harms — such as when models apply offensive labels to people from some groups — but they could be inherent in the dataset. For example, ImageNet contains numerous slurs and offensive labels inherited from WordNet and pornographic images of people who did not consent to their inclusion in the dataset.[53,54]

While downstream and representational harms are two categories that have drawn a lot of attention and criticism, there are many other harms that often arise including the environmental cost of training models on unnecessarily large datasets[55] and the erasure of the labor of subjects who contributed the data[50] or the annotators who labeled it.[16] For an overview of ethical concerns associated with datasets, see the survey by Paullada et al.[56]

### *Mitigating harms: an overview*

Approaches for mitigating the harms associated with data are quickly developing. Here we review a few selected ideas.

One approach targets the fact that many machine learning datasets are poorly documented, and details about their creation are often missing. This leads to a range of issues from lack of reproducibility and concerns of scientific validity to misuse and ethical concerns. In response, *datasheets for datasets* is a template and initiative by Gebru et al. to promote more detailed and systematic annotation for datasets.[57] A datasheet requires the creator of a dataset to answer questions relating to several areas of interest: Motivation, composition, collection process, preprocessing/cleaning/labeling, uses, distribution, maintenance. One goal is that process of creating a datasheet will help anticipate ethical issues with the dataset. But datasheets also aim to make data practices more reproducible, and help practitioners select more adequate data sources.

Going a step beyond datasheets, Jo and Gebru[58] draw lessons from archival and library sciences for the construction and documentation of machine learning datasets. These lessons draw attention to issues of consent, inclusivity, power, transparency, ethics and privacy.

Other approaches stay within the paradigm of minimally curated data collection but aim to modify or sanitize content deemed problematic in datasets. The

ImageNet creators have made efforts to remove slurs and harmful terms as well as categories considered non-imageable, or unable to be characterized using images. "Vegetarian" and "philanthropist" are two such categories that were removed.[54] The REVISE tool aims to partially automate the process of identifying various kinds of biases in visual datasets.[59]

### *Mitigating harms by separating the roles of datasets*

Our analysis of the different roles datasets play allows greater clarity in mitigating harms while preserving benefits. This analysis is not intended as an alternative to the many approaches that have already been proposed for mitigating harms. Rather, it can sharpen our thinking and strengthen other harm-mitigation strategies.

Our main observation is that the reuse of scientific benchmark datasets in engineering pipelines complicates efforts to address biases and harms. Attempts to address harms in such dual-use datasets leaves creators with a conundrum. On the one hand, benchmark datasets need to be long-lived: many benchmark datasets created decades ago continue to be useful and widely used today. Thus, modifying a dataset down the line when new harms become known will compromise its scientific utility, as performance on the modified dataset may not be meaningfully comparable to performance on the older dataset.

On the other hand, attempting to anticipate all possible harms during dataset creation is infeasible if the dataset is going to be used as training or pre-training data. Experience shows that datasets turn out to be useful for an ever-expanding set of downstream tasks, some of which were not even conceived of at the time of dataset creation.

Better tradeoffs are possible if there is a clear separation between scientific benchmarks and production-oriented datasets. In cases where the same dataset can be potentially useful for both purposes, creators should consider making two versions or forks of the data, because many of the harm mitigation strategies that apply to one don't apply to the other, and vice versa.

To enforce this separation, benchmark dataset creators should consider avoiding the use of the dataset in production pipelines by explicitly prohibiting it in the terms of use. Currently the licenses of many benchmark datasets prohibit commercial uses. This restriction has a similar effect, but it is not the best way to make this distinction. After all, production models may be noncommercial: they may be built by researchers or governments, with the latter category having an especially high potential for harm. At the same time, prohibiting commercial uses is arguably too strict, as it prohibits the use of the dataset as a guide to model selection, a use that does not raise the same risks of downstream harm.

One reason why there are fairness interventions applicable to scientific benchmark datasets but not production datasets is that, as we've argued, most of the scientific utility of benchmarks is captured by the *relative* performance of models. The fact that interventions that hurt absolute performance may be acceptable gives greater leeway for harm mitigation efforts. Consider image classification benchmarks. We hypothesize that the relative ranking of models will be only minimally

affected if the dataset is modified to remove all images containing people (keeping high-level properties including the number of classes and images the same). Such an intervention would avoid a wide swath of the harms associated with datasets while preserving much of its scientific utility.

Conversely, one reason why there are fairness interventions applicable to production datasets but not scientific benchmarks is that interventions for production datasets can be strongly guided by an understanding of their downstream impacts in specific applications. Language and images, in particular, capture such a variety of cultural stereotypes that sanitizing all of them has proved infeasible.[60] It is much easier to design interventions once we fix an application and the cultural context(s) in which it will be deployed. Different interventions may be applicable to the same dataset used in different applications. Unlike scientific benchmarks, dataset standardization is not necessary in engineering settings.

In fact, the best locus of intervention even for dataset biases may be downstream of the data. For example, it has been observed for many years that online translation systems perpetuate gender stereotypes when translating gender-neutral pronouns. The text "O bir doctor. O bir hemşire." may be translated from Turkish to English as "He is a doctor. She is a nurse." Google Translate mitigated this by showing multiple translations in such cases.[61,62] Compared to data interventions, this has the benefit of making the potential bias (or, in some cases, erroneous translation) more visible to the user.

Our analysis points to many areas where further research could help clarify ethical implications. In particular, the pre-training role of benchmark datasets occupies a grey area where it is not clear when and to what extent data biases propagate to the target task/domain. Research on this area is nascent;[63] this research is vital because the (mis)use of scientific benchmarks for pre-training in production pipelines is common today and unlikely to cease in the near future.

Datasets should not be seen as static, neutral technical artifacts. The harms that could arise from a dataset depend not just on its contents but also the rules, norms, and culture surrounding its usage. Thus, modifying these cultural practices is one potential way to mitigate harms. As we discussed above, lack of domain knowledge by dataset users has come to be seen almost as a virtue in machine learning. This attitude should be reconsidered as it has a tendency to accentuate ethical blind spots.

Datasets require stewardship, whether by the dataset creator or by another designated entity or set of entities. Consider the problem of derivatives: popular benchmark datasets are often extended by other researchers with additional features, and these derived datasets can introduce the possibility of harms not present in the original (to the same extent). For example, the Labeled Faces in the Wild (LFW) dataset of faces was annotated by other researchers with characteristics as race, gender, and attractiveness.[64,65] Regardless of the ethics of LFW itself, the derived dataset enables new applications that classify people by appearance in harmful ways.[3] Of course, not all derivatives are ethically problematic. Adjudicat-

[3]The intended purpose of the derived dataset is to enable searching corpora of face images by describable attributes.

ing and enforcing such ethical distinctions is only possible if there is a governance mechanism in place.

## *Beyond datasets*

In this final section, we discuss important scientific and ethical questions that are relevant to datasets but also go beyond datasets, pervading machine learning: validity, problem framing, and limits to prediction.

### *Lessons from measurement*

Measurement theory is an established science with ancient roots. In short, measurement is about assigning numbers to objects in the real world in a way that reflects relationships between these objects. Measurement draws an important distinction between a *construct* that we wish to measure and the measurement procedure that we used to create a numerical representation of the construct.

For example, we can think of a well-designed math exam as measuring the mathematical abilities of a student. A student with greater mathematical ability than another is expected to score higher on the exam. Viewed this way, an exam is a *measurement procedure* that assigns numbers to students. The *mathematical ability* of a student is the construct we hope to measure. We desire that the ordering of these numbers reflects the sorting of students by their mathematical abilities. A measurement procedure operationalizes a construct.

Every prediction problem has a target variable, the thing we're trying to predict.[4] By viewing the target variable as a construct, we can apply measurement theory to understand what makes a good target variable.

The choice of a poor target variable cannot be ironed out with additional data. In fact, the more data we feed into our model, the better it gets at capturing the flawed target variable. Improved data quality or diversity are no cure either.

All formal fairness criteria that involve the target variable, separation and sufficiency being two prominent examples[5], are either meaningless or downright misleading when the target variable itself is the locus of discrimination.

But what makes a target variable good or bad? Let's get a better grasp on this question by considering a few examples.

1. Predicting the value of the Standard and Poor 500 Index (S&P 500) at the close of the New York Stock Exchange tomorrow.
2. Predicting whether an individual is going to default on a loan.
3. Predicting whether an individual is going to commit a crime.

[4]Recall that in a prediction problem we have covariates $X$ from which we're trying to predict a variable $Y$. This variable $Y$ is what we call the *target variable* in our prediction problem.

[5]Recall from Chapter 3 that separation requires the protected attribute to be independent of the prediction conditional on the target variable. Sufficiency requires the target variable to be independent of the protected attribute given the prediction.

The first example is rather innocuous. It references a fairly robust target variable, even though it relies on a number of social facts.

The second example is a common application of statistical modeling that underlies much of modern credit scoring in the United States. At first sight a default event seems like a clean cut target variable. But the reality is different. In a public dataset by the Federal Reserve[66] default events are coded by a so-called *performance* variable that measures a *serious delinquency in at least one credit line of a certain time period*. More specifically, the report states that the

> measure is based on the performance of new or existing accounts and measures whether individuals have been late 90 days or more on one or more of their accounts or had a public record item or a new collection agency account during the performance period.[6]

Our third example runs into the most concerning measurement problem. How do we determine if an individual committed a crime? What we can determine with certainty is whether or not an individual was arrested and found guilty of a crime. But this depends crucially on who is likely to be policed in the first place and who is able to maneuver the criminal justice system successfully following an arrest.

Sorting out what a good target variable is, in full generality, can involve the whole apparatus of measurement theory. The scope of measurement theory, however, goes beyond defining reliable and valid target variables for prediction. Measurement comes in whenever we create features for a machine learning problem and should therefore be an essential part of the data creation process.[67]

Judging the quality of a measurement procedure is a difficult task. Measurement theory has two important conceptual frameworks for arguing about what makes measurement *good*. One is *reliability*. The other is *validity*.

Reliability describes the differences observed in multiple measurements of the same object under identical conditions. Thinking of the measurement variable as a random variable, reliability is about the variance between independent identically distributed measurements. As such, reliability can be analogized with the statistical notion of variance.

*Validity* is concerned with how well the measurement procedure in principle captures the concept that we try to measure. If reliability is analogous to variance, it is tempting to see validity as analogous to bias. But the situation is a bit more complicated. There is no simple formal criterion that we could use to establish validity. In practice, validity is based to a large extent on human expertise and subjective judgments.

One approach to formalize validity is to ask how well a score predicts some external criterion. This is called *external validity*. For example, we could judge a measure of creditworthiness by how well it predicts default in a lending scenario. While external validity leads to concrete technical criteria, it essentially identifies good measurement with predictive accuracy. However, that's certainly not all there is to validity.

[6] Quote from the Federal Reserve report.

Construct validity is a framework for discussing validity that includes numerous different types of evidence. Messick highlights six aspects of construct validity:

- Content: How well does the content of the measurement instrument, such as the items on a questionnaire, measure the construct of interest?
- Substantive: Is the construct supported by a sound theoretical foundation?
- Structural: Does the score express relationships in the construct domain?
- Generalizability: Does the score generalize across different populations, settings, and tasks?
- External: Does the score successfully predict external criteria?
- Consequential: What are the potential risks of using the score with regards to bias, fairness, and distributive justice?

Of these different criteria, external validity is the one most familiar to the machine learning practitioner. But machine learning practice would do well to embrace the other, more qualitative, criteria as well. The consequential criterion has been controversial, but Messick forcefully defends its inclusion as an aspect of validity.[68] Ultimately, measurement forces us to grapple with the often surprisingly uncomfortable question: What are we even trying to do when we predict something?

### *Problem framing: comparisons with humans*

A long-standing ambition of artificial intelligence research is to match or exceed human cognitive abilities by an algorithm. This desire often leads to comparisons between humans and machines on various tasks. Judgments about human accuracy often also enter the debate around when to use statistical models in high stakes decision making settings.

The comparison between human decision makers and statistical models is by no means new. For decades, researchers have compared the accuracy of human judgments with that of statistical models.[69]

Even within machine learning, the debate dates way back. A 1991 paper by Bromley and Sackinger explicitly compared the performance of artificial neural networks to a measure of human accuracy on the USPS digits dataset that predates the famous MNIST data.[12] A first experiment put the human accuracy at 2.5%, a second experiment found the number 1.51%, while a third reported the number 2.37%.[70]

Comparison with so-called human baselines has since become widely accepted in the machine learning community. The Electronic Frontier Foundation (EFF), for example, hosts a major repository of AI progress measures that compares the performance of machine learning models to reported human accuracies on numerous benchmarks.

For the ILSVRC 2012 data, the reported human accuracy is 5.1%.[7] This often quoted number corresponds to the performance of a single human annotator who

[7] To be precise, this number is referring to the fraction of times that the correct image label was not contained in the top 5 predicted labels of the model or human.

was "trained on 500 images and annotated 1500 test images".[15] A second annotator who was "trained on 100 images and then annotated 258 test images" achieved an accuracy of 12%. Based on this number of 5.1%, researchers announced in 2015 that their model was "the first to surpass human-level performance".[71] Not surprisingly, this claim received significant attention throughout the media.

However, a later more careful investigation into "human accuracy" on ImageNet revealed a very different picture.[72] The researchers found that only models from 2020 are actually on par with the strongest human labeler. Moreover, when restricting the data to 590 object classes out of 1000 classes in total, the best human labeler performed much better at less than 1% error than even the best predictive models. Recall, that the ILSVRC 2012 data featured 118 different dog breeds alone, some of which are extremely hard to distinguish for anyone who is not a trained dog expert. In fact, the researchers had to consult with experts from the American Kennel Club (AKC) to disambiguate challenging cases of different dog breeds. Simply removing dog classes alone increases the performance of the best human labeler to less than 1.3% error.

There is another troubling fact. Small variations in the data collection protocol turn out to have a significant effect on the performance of machine classifiers: "the accuracy scores of even the best image classifiers are still highly sensitive to minutiae of the data cleaning process."[31]

These results cast doubt not only on how me measure human accuracy, but also on the validity of the presumed theoretical construct of "human accuracy" itself. However, the machine learning community has adopted a rather casual approach to measuring human accuracy. Many researchers assume that the construct of *human accuracy* exists unambiguously and it is whatever number comes out of some ad-hoc testing protocol for some set of human beings. These ad-hoc protocols often result in anecdotal comparisons of questionable scientific value.

Invalid judgments about human performance relative to machines are not just a scientific error, they also have the potential to create narratives that support poor policy choices in high stakes policy questions around the use of predictive models in consequential decisions. For example, criminal justice policy is being driven by claims that statistical methods are superior to judges at predicting risk of recidivism or failure to appear in court. However, these comparisons are dubious because judges are not solving pure prediction problems but rather incorporate other factors such as leniency towards younger defendants.[73]

### *Problem framing: focusing on a single optimization objective*

Real-life problems rarely involve optimizing a single objective and more commonly involve some kind of tradeoff between multiple objectives. How best to formulate this as a statistical optimization problem is both an art and a science. However, benchmark tasks, especially those with leaderboards, tend to pick a single objective. For high-profile benchmarks, the resulting "overfitting to the problem formulation" may result in scientific blind spots and limit the applicability of published findings to practical settings.

For example, it was well known at the time Netflix launched its Prize that recommendation is not just a matter of maximizing predictive accuracy and, even to the extent that it is, there isn't one single measure that's always appropriate.[74] Yet the contest focused purely on prediction accuracy evaluated by a single metric. A few years after the contest ended, Netflix revealed that most of the work that went into the leaderboard had not translated to production models. Part of the reason was that the contest did not capture the range of Netflix's objectives and constraints: the tight dependence of recommendations on the user interface; the fact that "users" are typically households made of members with differing tastes; explainability; freshness, and many more.[75]

If many of the insights from the leaderboard did not even generalize to Netflix's own production setting, the gap between Netflix and other recommendation-oriented platforms is far greater. Notably, as a movie platform, Netflix is unusual in that it has a relatively static inventory compared to those with user-generated content such as YouTube or Facebook. When the content pool is dynamic, a different class of algorithms is needed. The pull that the Netflix Prize exerted on recommender systems research may have diverted attention away from the latter type of algorithm for many years, although it is hard to know for sure because the counterfactual is unobservable.

Formal machine learning competitions, even if they cause blind spots due to the need to pick a single optimization objective, are at least carefully structured to promote scientific progress in some narrow sense. Arguably more damaging are the informal competitions that seems to inevitably emerge in the presence of a prominent benchmark dataset, resulting in unfortunate outcomes such as insightful papers being rejected because they failed to beat the state of the art, or unoriginal papers being published because they did beat the state of the art by (scientifically insignificant) application of greater computing power.

Another downside to a field oriented around one-dimensional, competitive pursuit is that it becomes structurally difficult to address biases in models and classifiers. If a contestant takes steps to prevent dataset bias from propagating to their models, there will be an accuracy drop (because accuracy is judged on a biased dataset) and fewer people will pay attention to the work.

As fairness issues in machine learning have gained prominence, fairness-focused benchmarks datasets have proliferated, such as the Pilot Parliamentarians Benchmark for facial analysis[52] and the Equity Evaluation Corpus for sentiment analysis.[76] An advantage of this approach is that the scientific and cultural machinery of benchmark-oriented innovation can be repurposed for fairness research. A potential danger is Goodhart's law, which states, in its broad form, "When a measure becomes a target, it ceases to be a good measure." As we've emphasized in this book, fairness is multifaceted, and benchmarks can capture only narrow notions of fairness. While these can be useful diagnostics, if they are misconstrued as targets in their own right, then research that is focused on optimizing for these benchmarks may not result in fairness in a more substantive sense. In addition, the construction of these datasets has often been haphazard, without adequate attention to issues of validity.[77]

In addition to creating fairness-focused benchmarks, the algorithmic fairness community has also repurposed earlier benchmarks toward the study of fairness questions. Consider the Census dataset from the UCI repository discussed earlier. It originally gained popularity as a source of real-world data. Its use is acceptable for studying algorithmic questions such as, say, the relative strengths of decision trees and logistic regression. We expect the answers to be insensitive to issues like the cultural context of the data. But now it is being used for studying fairness questions such as how classification accuracy tends to vary by race or gender. For such questions, the answers are sensitive to the details of the subpopulations. Further, the classification task associated with the benchmark (prediction of income treated as a binary variable) is artificial and does not correspond to any real-life application. Thus, accuracy disparities (and other fairness-related measurements) may look different for a different task, or if the data had been sampled differently, or if it came from a different time or place. Using benchmark datasets to make generalizable claims about fairness requires careful attention to issues of context, sampling, and validity. Bao et al. question whether benchmark datasets for socio-technical systems like criminal justice are useful. They point out that benchmark culture — where the focus is on methods, with the dataset being secondary and the context ignored — is at odds with the actual needs of fairness and justice, where attention to context is paramount.[78]

### *Limits of data and prediction*

Machine learning fails in many scenarios and it's important to understand the failure cases as much as the success stories.

The Fragile Families Challenge was a machine learning competition based on the Fragile Families and Child Wellbeing study (FFCWS).[79] Starting from a random sample of hospital births between 1998 and 2000, the FFCWS followed thousand of American families over the course of 15 years, collecting detailed information, about the families' children, their parents, educational outcomes, and the larger social environment. Once a family agreed to participate in the study, data were collected when the child was born, and then at ages 1, 3, 5, 9, and 15.

The Fragile Families Challenge concluded in 2017. The underlying dataset for the competition contains 4242 rows, one for each family, and 12943 columns, one for each variable plus an ID number of each family. Of the 12942 variables, 2358 are constant (i.e., had the same value for all rows), mostly due to redactions for privacy and ethics concerns. Of the approximately 55 million (4242 x 12942) entries in the dataset, about 73% do not have a value. Missing values have many possible reasons, including non-response of surveyed families, drop out of study participants, as well as logical relationships between features that imply certain fields are missing depending on how others are set. There are six outcome variables, measured at age 15: *1) child grade point average (GPA), 2) child grit, 3) household eviction, 4) household material hardship, 5) caregiver layoff, and 6) caregiver participation in job training.*

The goal of the competition was to predict the value of the outcome variables at age 15 given the data from age 1 through 9. As is common for competitions, the

challenge featured a three-way data split: training, leaderboard, and test sets. The training set is publicly available to all participants, the leaderboard data support a leaderboard throughout the competition, and the test set is used to determine a final winner.

The outcome of the prediction challenge was disappointing. Even the winning models performed hardly better than a simple baseline their predictions didn't differ much compared to predicting the mean of each outcome.

What caused the poor performance of machine learning on the fragile families data? One obvious possibility is that none of the contestants hit upon the right machine learning techniques for this task. But the fact that 160 teams of motivated experts submitted thousands of models over the course of five months makes this highly unlikely. Besides, models from disparate model classes all made very similar (and equally erroneous) predictions, suggesting that learning algorithms weren't the limitation.[8]

There are a few other technical possibilities that could explain the disappointing performance, including the sample size, the study design, and the missing values.

But there is also a more fundamental reason that remains plausible. Perhaps the dynamics of life trajectories are inherently unpredictable over the six year time delay between measurement of the covariates and measurement of the outcome. This six year gap, for example, included the Great Recession, a period of economic shocks and decline between 2007 and 2009, that might have changed trajectories in unforeseeable ways.

In fact, there's an important reason why even the performance of models in the challenge, dismal as they were, may overestimate what we can expect in a real-world setting. That's because the models were allowed to peek into the future, so to speak. The training and test sets were drawn from the same distribution and, in particular, the same time period, as is the standard practice in machine learning research. Thus, the data already incorporates information about the effect of the Great Recession and other global shocks during this period. In a real application, models must be trained on data from the past whereas predictions are about the future. Thus, there is always some drift — a change in the relationship between the covariates and the outcome. This puts a further limit on model performance.

Machine learning works best in a static and stable world where the past looks like the future. Prediction alone can be a poor choice when we're anticipating dynamic changes, or when we are trying to reason about the effect that hypothetical actions would have in the real world.

## *Summary*

Benchmark datasets are central to machine learning. They play many roles including enabling algorithmic innovation, measuring progress, and providing training

[8] This highlights an advantage of the benchmark dataset approach over one with less standardization: even when there is a failure to make substantial progress on prediction, we can learn something valuable from that failure.

data. Since its systematization in the late 1980s, performance evaluation on benchmarks has gradually become a ubiquitous practice because it makes it harder for researchers to cheat intentionally or unintentionally.

But an excessive focus on benchmarks brings many drawbacks. Researchers spend prodigious amounts of effort optimizing models to achieve state of the art performance. The results are often both scientifically uninteresting and of little relevance to practitioners because benchmarks omit many real-world details. The approach also amplifies the harms associated with data including downstream harms, representational harms, and privacy violations.

As we write this book, the benchmark approach is coming under scrutiny because of these ethical concerns. While the benefits and drawbacks of benchmarks are both well known, our overarching goal in this chapter has been to provide a single framework that can help analyze both. Our position is that the core of the benchmark approach is worth preserving, but we envision a future where benchmarks play a more modest role as one of many ways to advance knowledge. To mitigate the harms associated with data, we believe that substantial changes to the practices of dataset creation, use, and governance are necessary. We have outlined a few ways to do this, adding to the emerging literature on this topic.

## *Chapter notes*

This chapter was developed and first published by Hardt and Recht in the textbook *Patterns, Predictions, and Actions: Foundations of Machine Learning*.[81] With permission from the authors, we include a large part of the original text here with only slight modifications. We removed a significant amount of material on adaptive data analysis and the problem of overfitting in machine learning benchmarks. We added new material on the roles that datasets play, as well as discussion about fairness and ethical concerns relating to datasets.

Adaptivity in holdout reuse was studied by Dwork et al.[82] and there has been subsequent work in the area of adaptive data analysis. Similar concerns go under the name of *inference after selection* in the statistics community.

The collection and use of large ad-hoc data sets (once referred to as "big data") has been scrutinized in several important works, see, for example, boyd and Crawford,[83] as well as Tufekci.[84,85] More recently, Couldry and Mejias[86] use the term *data colonialism* to emphasize the processes by which data are appropriated and marginalized communities are exploited through data collection. Olteanu et al.[87] discuss biases, methodological pitfalls, and ethical questions in the context of social data analysis. In particular, the article provides taxonomies of biases and issues that can arise in the sourcing, collection, processing, and analysis of social data. Bowker and Star's classic text explains why categorization is a morally laden activity.[88] For a discussion of the harms of category systems embedded in machine learning datasets, see *Atlas of AI*.[89]

The benefits of the benchmark dataset approach are discussed in a talk by Mark Liberman, who calls it the common task method.[90] Paullada, Raji, Ben-

der, Denton, and Hanna survey dataset development and use cases in machine learning research.[91] A survey by Fabris, Messina, Silvello, and Susto lists and discusses numerous datasets uses throughout the fairness literature.[92] Denton, Hanna, Amironesei, Smart and Nicole provide a genealogy of ImageNet through a critical lens.[93] Raji, Bender, Paullada, Denton and Hanna give an overview of concerns arising from basing our understanding of progress on a small collection of influential benchmarks.[94] The EFF AI metrics project is available at: https://www.eff.org/ai/metrics.

For an introduction to measurement theory, not specific to the social sciences, see the books by Hand.[95,96] The textbook by Bandalos[97] focuses on applications to the social science, including a chapter on fairness. Liao, Taori, Raji and Schmidt provide a taxonomy of evaluation failures across many subfields of machine learning, encompassing both internal and external validity issues.[98]

# *Bibliography*


[1] Xiaochang Li and Mara Mills. Vocal features: from voice identification to speech recognition by machine. *Technology and culture*, 60(2):S129–S160, 2019.

[2] Mark Liberman. Fred Jelinek. *Computational Linguistics*, 36(4):595–599, 2010.

[3] Kenneth Ward Church. Emerging trends: A tribute to Charles Wayne. *Natural Language Engineering*, 24(1):155–160, 2018.

[4] Mark Liberman and Charles Wayne. Human language technology. *AI Magazine*, 41(2), 2020.

[5] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93:27403, 1993.

[6] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

[7] Pat Langley. The changing science of machine learning, 2011.

[8] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[9] Patrick J Grother. Nist special database 19. *Handprinted forms and characters database, National Institute of Standards and Technology*, page 10, 1995.

[10] Chhavi Yadav and Léon Bottou. Cold case: The lost mnist digits. *arXiv preprint arXiv:1905.10498*, 2019.

[11] Dennis DeCoste and Bernhard Schölkopf. Training invariant support vector machines. *Machine learning*, 46(1):161–190, 2002.

[12] J Bromley and E Sackinger. Neural-network and k-nearest-neighbor classifiers. *Rapport technique*, pages 11359–910819, 1991.

[13] George A Miller. *WordNet: An electronic lexical database*. MIT Press, 1998.

[14] 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*, pages 248–255. Ieee, 2009.

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[16] Mary L Gray and Siddharth Suri. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books, 2019.

[17] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

[18] Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4:61–84, 2017.

[19] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[20] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[21] Richard Evans. Ra fisher and the science of hatred, 2020.

[22] Francisco Louçã. Emancipation through interaction–how eugenics and statistics converged and diverged. *Journal of the History of Biology*, 42(4):649–684, 2009.

[23] Tom M Mitchell. *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research . . . , 1980.

[24] Leo Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

[25] Frank Rosenblatt. Perceptron simulation experiments. *Proceedings of the IRE*, 48(3):301–309, 1960.

[26] Pat Langley. Machine learning as an experimental science, 1988.

[27] Simon Funk. Try this at home. *http://sifter.org/˜simon/journal/2006*, 2006.

[28] Daniel Billsus, Michael J Pazzani, et al. Learning collaborative information filters. In *Icml*, volume 98, pages 46–54, 1998.

[29] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[30] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[31] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.

[32] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR, 2021.

[33] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[34] Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010.

[35] Kate Crawford and Trevor Paglen. Excavating ai: the politics of training sets for machine learning. *Excavating AI (www.excavating.ai)*, 2019.

[36] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[37] Michael Veale, Reuben Binns, and Lilian Edwards. Algorithms that remember: model inversion attacks and data protection law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180083, 2018.

[38] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification and scene analysis*, volume 3. Wiley New York, 1973.

[39] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2017.

[40] Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning*, pages 1006–1014. PMLR, 2015.

[41] Mostafa Dehghani, Yi Tay, Alexey A Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.

[42] Bernard Koch, Emily Denton, Alex Hanna, and Jacob G Foster. Reduced, reused and recycled: The life of a dataset in machine learning research. *arXiv preprint arXiv:2112.01716*, 2021.

[43] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. *arXiv preprint arXiv:2012.07421*, 2020.

[44] Gwern Branwen. The neural net tank urban legend. 2011.

[45] Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):1–21, 2012.

[46] Benjamin Marie, Atsushi Fujita, and Raphael Rubino. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. *arXiv preprint arXiv:2106.15195*, 2021.

[47] Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3, 2021.

[48] Lorenza Saitta and Filippo Neri. Learning in the "real world". *Machine learning*, 30(2):133–163, 1998.

[49] Steven L Salzberg. On comparing classifiers: A critique of current research and methods. *Data mining and knowledge discovery*, 1(1):1–12, 1999.

[50] Joanna Radin. "digital natives": How medical and indigenous histories matter for big data. *Osiris*, 32(1):43–64, 2017.

[51] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in ml-based science. *arXiv preprint arXiv:2207.07048*, 2022.

[52] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability, and Transparency*, pages 77–91, 2018.

[53] Vinay Uday Prabhu and Abeba Birhane. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*, 2020.

[54] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In *Conference on Fairness, Accountability, and Transparency*, pages 547–558, 2020.

[55] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

[56] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *arXiv preprint arXiv:2012.05345*, 2020.

[57] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

[58] Eun Seo Jo and Timnit Gebru. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Conference on Fairness, Accountability, and Transparency*, pages 306–316, 2020.

[59] Angelina Wang, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. In *European Conference on Computer Vision*, pages 733–751. Springer, 2020.

[60] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*, 2019.

[61] James Kuczmarski. Reducing gender bias in google translate. *Google Blog*, 6, 2018.

[62] Melvin Johnson. A scalable approach to reducing gender bias in google translate. *Google Blog*, 2020.

[63] Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. In *Conference on Fairness, Accountability, and Transparency*, pages 701–713, 2021.

[64] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[65] Neeraj Kumar, Alexander Berg, Peter N Belhumeur, and Shree Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011.

[66] The Federal Reserve Board. Report to the congress on credit scoring and its effects on the availability and affordability of credit. https://www.federalreserve.gov/boarddocs/rptcongress/creditscore/, 2007. Accessed: 2018-05-29.

[67] Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. In *Conference on Fairness, Accountability, and Transparency*, pages 375–385, 2021.

[68] Samuel Messick. Test validity: A matter of consequence. *Social Indicators Research*, 45(1):35–44, 1998.

[69] Robyn M Dawes, David Faust, and Paul E Meehl. Clinical versus actuarial judgment. *Science*, 243(4899):1668–1674, 1989.

[70] Ibrahim Chaaban and Michael R Scheessele. Human performance on the usps database. *Report, Indiana University South Bend*, 2007.

[71] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *International Conference on Computer Vision*, pages 1026–1034, 2015.

[72] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning*, pages 8634–8644. PMLR, 2020.

[73] Megan T Stevenson and Jennifer L Doleac. Algorithmic risk assessment in the hands of humans. *Available at SSRN*, 2022.

[74] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[75] Xavier Amatriain and Justin Basilico. Netflix recommendations: Beyond the 5 stars (part 1). *Netflix Tech Blog*, 6, 2012.

[76] Svetlana Kiritchenko and Saif Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. In *Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[77] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, 2021.

[78] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. *arXiv preprint arXiv:2106.05498*, 2021.

[79] Nancy E Reichman, Julien O Teitler, Irwin Garfinkel, and Sara S McLanahan. Fragile families: Sample and design. *Children and Youth Services Review*, 23(4-5):303–326, 2001.

[80] Matthew J Salganik, Ian Lundberg, Alexander T Kindel, Caitlin E Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M Altschul, Jennie E Brand, Nicole Bohme Carnegie, Ryan James Compton, et al. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15):8398–8403, 2020.

[81] Moritz Hardt and Benjamin Recht. *Patterns, predictions, and actions: Foundations of machine learning*. Princeton University Press, 2022.

[82] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

[83] Danah Boyd and Kate Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679, 2012.

[84] Zeynep Tufekci. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Conference on Web and Social Media*, volume 8, 2014.

[85] Zeynep Tufekci. Engineering the public: Big data, surveillance and computational politics. *First Monday*, 2014.

[86] Nick Couldry and Ulises A Mejias. Data colonialism: Rethinking big data's relation to the contemporary subject. *Television & New Media*, 20(4):336–349, 2019.

[87] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, 2019.

[88] Geoffrey C Bowker and Susan Leigh Star. *Sorting things out: Classification and its consequences*. MIT Press, 2000.

[89] Kate Crawford. *The Atlas of AI*. Yale University Press, 2021.

[90] Marc Liberman. Reproducible research and the common task method. *Simmons Foundation Lecture https://www. simonsfoundation. org/lecture/reproducible-research-and-thecommon-task-method*, 2, 2015.

[91] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.

[92] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. Algorithmic fairness datasets: the story so far. *arXiv preprint arXiv:2202.01711*, 2022.

[93] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. On the genealogy of machine learning datasets: A critical history of imagenet. *Big Data & Society*, 8(2):20539517211035955, 2021.

[94] Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. Ai and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*, 2021.

[95] David J. Hand. *Measurement Theory and Practice: The World Through Quantification*. Wiley, 2010.

[96] David J Hand. *Measurement: A very short introduction*. Oxford University Press, 2016.

[97] Deborah L Bandalos. *Measurement theory and applications for the social sciences*. Guilford Publications, 2018.

[98] Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.