nature machine intelligence

PERSPECTIVE

https://doi.org/10.1038/s42256-022-00516-1

Check for updates

Advances, challenges and opportunities in creating data for trustworthy AI

Weixin Liang[1], Girmaw Abebe Tadesse[2], Daniel Ho[3], L. Fei-Fei[1], Matei Zaharia[1], Ce Zhang[4] and James Zou[1,5] ✉

As artificial intelligence (AI) transitions from research to deployment, creating the appropriate datasets and data pipelines to develop and evaluate AI models is increasingly the biggest challenge. Automated AI model builders that are publicly available can now achieve top performance in many applications. In contrast, the design and sculpting of the data used to develop AI often rely on bespoke manual work, and they critically affect the trustworthiness of the model. This Perspective discusses key considerations for each stage of the data-for-AI pipeline—starting from data design to data sculpting (for example, cleaning, valuation and annotation) and data evaluation—to make AI more reliable. We highlight technical advances that help to make the data-for-AI pipeline more scalable and rigorous. Furthermore, we discuss how recent data regulations and policies can impact AI.

Artificial intelligence (AI) has experienced tremendous progress in recent years and is increasingly deployed across domains including healthcare, e-commerce and media[1–3]. As AI matures, AI model building is becoming increasingly turn-key with technologies such as automated machine learning (AutoML), which automates model design and hyperparameter tuning, large public repositories of trained models, and industry-standard platforms such as PyTorch, Tensorflow and so on[4–6]. Companies including Amazon, Google and Microsoft all offer AutoML products, allowing users to build state-of-the-art AI models on their own data without writing any code[7]. For example, a study on three public medical image datasets found that models produced by commercial AutoML demonstrated comparable or even higher performance compared with published bespoke algorithms[8]. All of these resources make it much easier to develop models when the data are provided.

In contrast to the increasing ease of model building, creating datasets for AI remains a major pain point due to the cost of curation and annotation. Surveys report that 96% of enterprises encounter data challenges including data quality and labelling in AI projects[9], and 40% of them lack confidence in ensuring data quality[10]. Data scientists spend nearly twice as much time on data loading, cleansing and visualization than on model training, selection and deployment[11]. Data pipelines can also be very expensive; for example, Flatiron Health, a US data aggregator that employs a network of clinicians to curate the medical records of patients with cancer, was acquired by Roche-Genentech for more than US$2 billion[12].

There is also growing recognition that state-of-the-art AI models often pick up spurious correlations and biases in the development data[13]. Choices made in each step of the data pipeline can greatly affect the generalizability and reliability of the AI model trained on these data, sometimes more than the choice of model. For example, a systematic assessment of three computer-vision AI models for diagnosing malignant skin lesions demonstrated that the models all performed substantially worse on lesions appearing on dark skin compared with light skin—the area under the receiver operating curves (AUROC) dropped by 10–15% across skin tones[14]. The models' poor performance can be attributed to data design—the training data had few dark-skin images—and data annotation errors (most of the training data were annotated for disease by a dermatologist's visual inspection and dermatologists made more mistakes on dark skin). Changing the method of training the model on the original biased data did not reduce the model's disparity across skin tones. However, improving annotation quality and skin tone representations on a set of just a few hundred images effectively closed the performance gap and improved the overall reliability of these models[14] (Fig. 1).

More attention needs to be placed on developing methods and standards to improve the data-for-AI pipeline. Much of the recent research in AI has focused on improving model performance on several standard benchmark datasets such as ImageNet, CIFAR100 (Canadian Institute for Advanced Research, 100 classes), Adult Income, MIMIC (Medical Information Mart for Intensive Care) and so on[15–17]. In such model-centric development, the dataset is typically fixed and given, and the focus is on iterating the model architecture or training procedure to improve the benchmark performance. This has led to substantial research progress in modelling, and now the incremental gains from improving models are diminishing in many tasks[18]. At the same time, as illustrated in the dermatology AI example, relatively small improvements in data can make AI models much more reliable. The data-centric perspective thus highlights the need for systematic methods to evaluate, synthesize, clean and annotate the data used to train and test the AI model.

A data-centric focus is often lacking in current AI research[19]. Furthermore, the data used to train or evaluate AI are often sparsely discussed. For example, a recent survey of 70 dermatology AI papers found that 63 (90%) papers did not present information on the skin tones in their dataset[20]. A review of 130 Food and Drug Administration-approved medical AI devices found that a comparison of the AI's performance on data from multiple locations was not reported for 93 devices[21].

[1]Department of Computer Science, Stanford University, Stanford, CA, USA. [2]IBM Research – Africa, Nairobi, Kenya. [3]Stanford Law School, Stanford University, Stanford, CA, USA. [4]Department of Computer Science, ETH Zurich, Zurich, Switzerland. [5]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. ✉e-mail: jamesz@stanford.edu

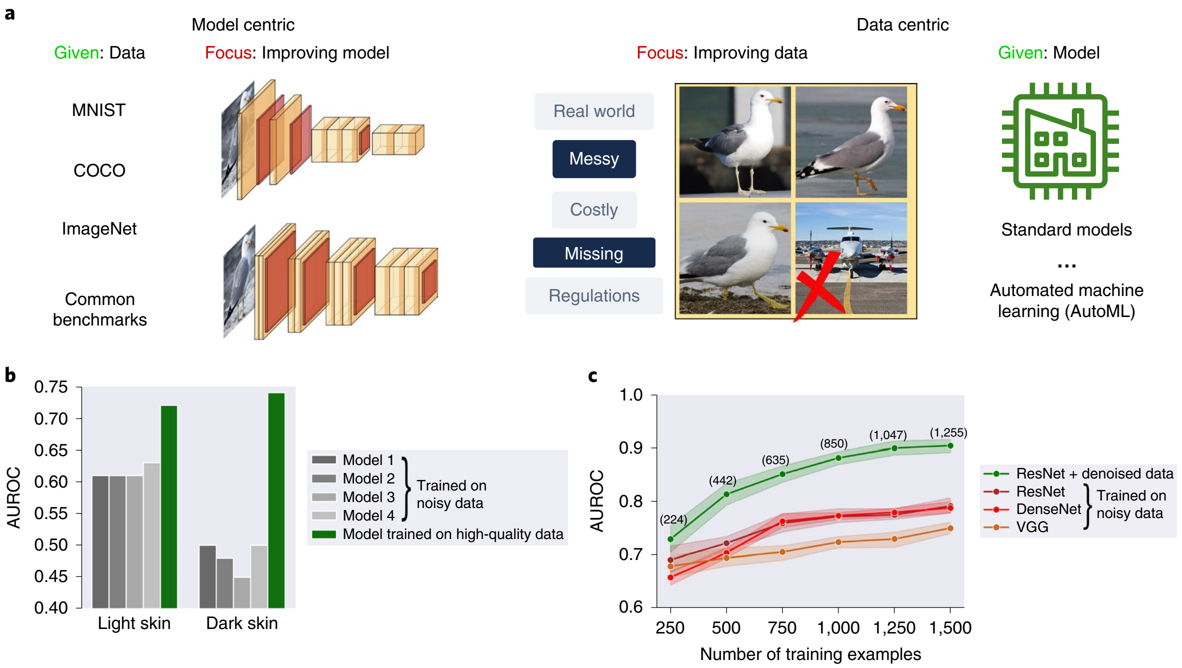


Fig. 1 | Comparison of model-centric versus data-centric approaches in AI. a, Model-centric research typically considers data as given and focuses on improving the model architecture or optimization on this data. Data-centric research focuses on scalable methods to systematically improve the data pipeline with data cleaning, selection, annotation and so on, and may use turn-key model builders. The airplane image indicates a noisy data point in the bird dataset that should be removed. MNIST, COCO and ImageNet are commonly used datasets in AI research. **b**, Skin disease diagnosis test performance on images of light and dark skin tones. Four state-of-the-art models trained on large, previously used dermatology data exhibit poor performance, especially on dark-skin images, due to training data errors. Model 1 trained on smaller higher-quality data is more reliable across skin tones. **c**, Object recognition test performance of different models compared with training a model (ResNet) on a cleaner subset of data after filtering by data Shapley value. The number in parentheses represents the number of training data points left after the filtering out noisy data. Results are aggregated over five random seeds. The shaded area represents the 95% confidence interval. ResNet, DenseNet and VGG are commonly used image classification models. Panel **a** reproduced from ref. [17], Springer Nature Ltd.

Data quality is often more emphasized in other disciplines, such as the social sciences and biomedicine, and there are many insights to borrow[22–24]. At the same time, AI's use of large volumes of heterogeneous unstructured data (for example, videos, audio and free text), often requiring expensive annotations, and the surprising ways in which AI models pick up correlations in the data present new challenges and opportunities[25]. There is thus a tremendous need for new automated or human-in-the-loop approaches to improve AI data pipelines in a systematic and scalable way. In the next sections, we explore some of the critical questions that AI developers should consider and technologies that facilitate creating data pipelines for AI. We organize the discussions to mirror the main steps of the AI data pipeline: data design (the sourcing and documentation of data), data sculpting (data selection, cleaning and annotation), and data strategies for model testing and monitoring (Fig. 2). We also discuss how recent data regulations impact data for AI.

Data design for AI

Once an AI application has been identified, designing the data—namely identifying and documenting the sources of data—to develop the AI model is often one of the first considerations. Data design is critical for mitigating bias and ensuring the generalizability and reliability of the AI model trained on this data[25]. Design should be an iterative process—it is often useful to have pilot data to develop an initial AI model and then collect additional data to patch the model's limitations. A critical design criterion is to ensure that the data are appropriate for the task and have good coverage to represent diverse users and scenarios that the model can encounter in practice. Datasets currently used to develop AI often have limited or biased coverage. For example, commonly used datasets for training facial recognition models are overwhelmingly composed of lighter-skinned subjects[26]. In medical AI, patient data used for developing algorithms were disproportionately collected from California, Massachusetts and New York, with little to no representation from other states or countries[27]. Such coverage gaps introduce bias and limit the generalizability of the AI models to diverse populations[28,29].

One promising approach to improve data coverage is to engage broader communities to participate in citizen-science data creation. As an example, the Common Voice project (https://commonvoice.mozilla.org/) is the largest public dataset of 11,192 hours of speech transcription in 76 languages from more than 166,000 participants[30]. Participants upload short recordings of themselves reading pre-specified text, which are assessed by other participants for scalable and transparent data quality control. Through culturally aware interface design and gamification, Common Voice has substantial coverage from digitally disadvantaged languages and accents—its second-largest language, after English, is Kinyarwanda[31].

When representative data are hard to access, synthetic data can potentially fill some of the coverage gaps. The collection of real human faces often involves privacy issues[32] and sampling biases[26]. Synthetic faces created by deep generative models have been used to reduce data imbalance and reduce bias[33,34]. In healthcare, synthetic medical records can be shared to facilitate knowledge discovery without disclosing actual patient-level information[35]. In robotics, although real-world challenges provide the ultimate testbed, high-fidelity simulated environments are widely used to enable faster, safer learning on complex and long-horizon tasks[36,37].

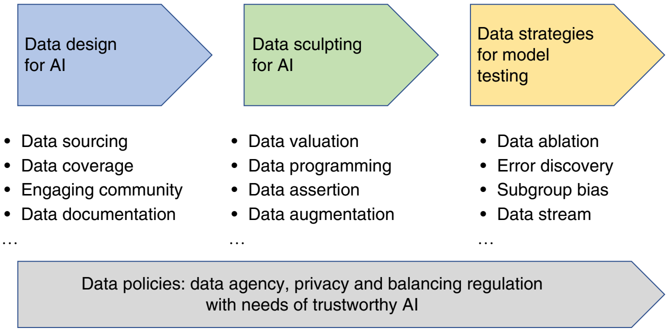


Fig. 2 | Roadmap for data-centric method development from data design to evaluation. Each box represents a major step of the data pipeline: data design for AI, data sculpting for AI, and data to evaluate and monitor AI. Several key methodologies and considerations for improving the data-for-AI pipeline are listed under each step. Data policies can affect each of the steps of developing trustworthy AI.

However, synthetic data have important caveats. There is always a simulation-to-real gap between the synthetic data and reality, so there is often a performance drop when transferring the AI model trained on synthetic data to the real world[38]. Synthetic data can also exacerbate data disparity when digitally disadvantaged groups are not considered in simulator design[39,40].

As an AI model's performance is highly dependent on the context of its training and evaluation data, it is important to document the data design context in standardized and transparent reports. Researchers have created various 'data nutrition labels' to capture metadata about the data design and annotation processes[41–43]. Useful metadata includes statistics on the sex, gender, ethnicity and geographical location of the participants in the dataset, which helps to surface potential issues of underrepresented subpopulations[44]. Additional metadata include data provenance, which tracks where and when a piece of data comes from, and the processes and methodology by which it was produced[41]. It is important for data documentation to capture the lifecycle and sociotechnical context of the data where applicable[45,46]. Data design documentations can enhance trust by discussing how data privacy and ownership are respected during data collection. It is a good practice to upload these documentations to stable and centralized data repositories such as Zenodo (https://www.zenodo.org/).

Data sculpting through selection, cleaning and annotation

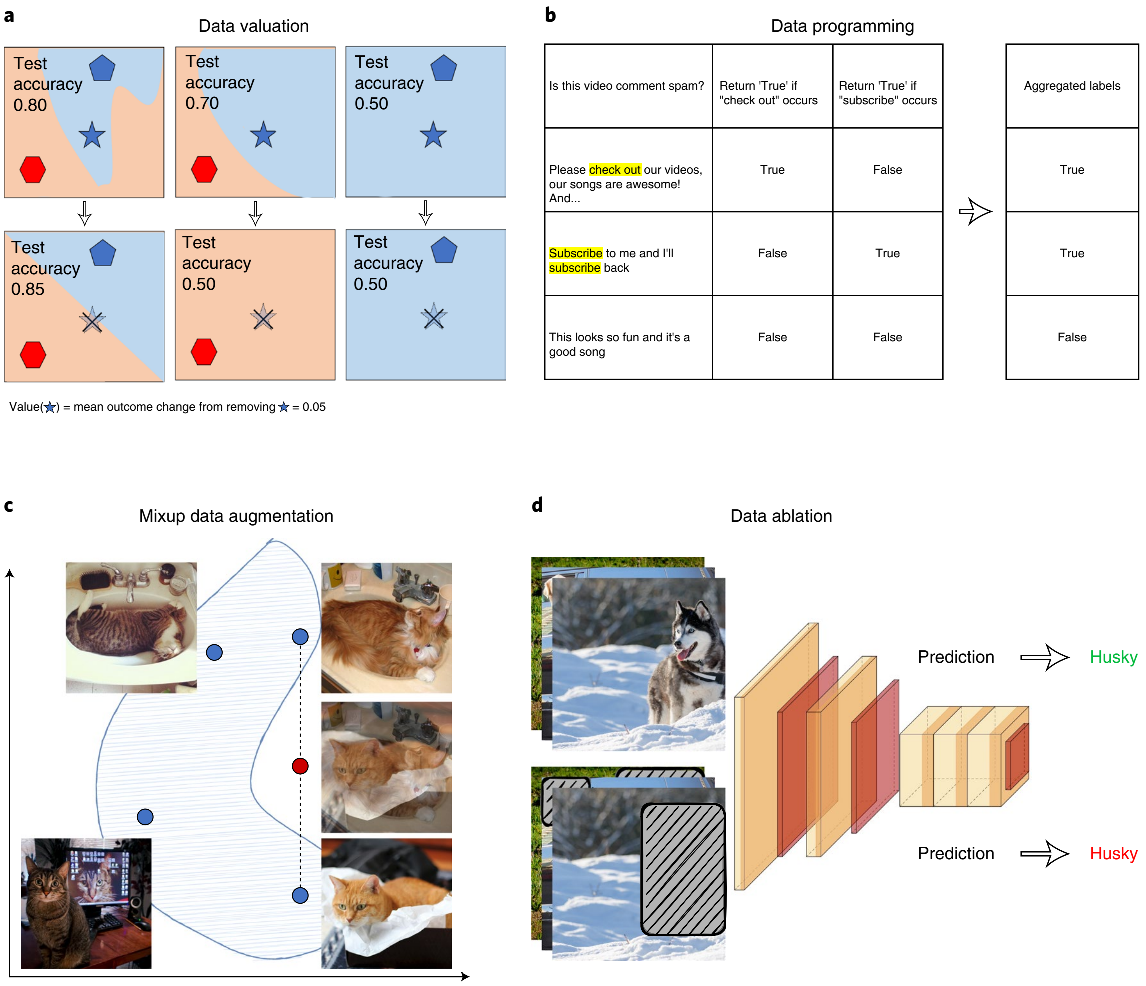
Once an initial dataset is collected, a substantial amount of work is needed to sculpt or refine the data to make it effective for AI development. In particular, careful selection of a subset of data for training can substantially improve the reliability and generalizability of the model than if it is trained on the entire noisy dataset. As an example, three popular deep learning architectures trained to classify animals on a noisy image dataset[47] all have similar suboptimal performances (Fig. 1). Filtering out poor-quality data using the data Shapley score (described below) and training the same model on the remaining data led to substantial improvements.

Examples like this motivate data valuation, which aims to quantify the importance of different data and filter out data that may hurt the model performance because of poor quality or biases[48,49]. One promising class of methods for data valuation is to measure changes in the AI model's behaviour when different data are removed from the training process (Fig. 3a). This can be done effectively using recent methods such as data Shapley scores or influence approximations[48–51]. Substantial progress has been made to efficiently compute these valuations for large AI models[49,52]. A complementary data valuation approach is to leverage prediction uncertainty to detect poor-quality data points. By looking at data points whose human annotation systematically deviates from predictions made by AI models, the Confidence Learning algorithm identified over 3% of the test data in common benchmarks such as ImageNet that are misannotated[53,54]. Filtering out these mistakes can substantially alter the model's performance. However, filtering can result in biases. For instance, a study on a commonly used web crawl corpus found that its keyword-based filtering mechanism disproportionately excludes minority identities (for example, lesbian, gay) and certain dialects (for example, Hispanic-aligned English), while a non-trivial fraction of those filtered out is non-offensive[55]. An alternative to filtering is to systematically clean dirty data to improve data quality. Methods such as ActiveClean improve cleaning efficiency by identifying the subset of the dirty data that are critical for the AI model and hence are important to clean[56].

Data annotation is often a major bottleneck and a source of errors. Although AI models can tolerate some level of random label noise[57], biased errors create biased models. Current annotations often rely on human manual labelling and can be expensive—for example, annotating a single LIDAR scan can cost more than US$30 because the LIDAR scan is three-dimensional and annotators need to draw three-dimensional bounding boxes[58,59]. Labellers on crowdsource platforms such as MTurk need to be carefully calibrated to provide consistent annotations, and the phrasing of the labelling tasks should be optimized to improve crowd response[60]. In medical settings, annotation may require domain expertise or involve sensitive data that could not be crowd-sourced, making labelling even more challenging and expensive.

One approach to reduce annotation costs is data programming[61–63]. In data programming, instead of hand-labelling data points, AI developers write programmatic labelling functions to automatically label the training set, usually based on rules, heuristics or ontologies (Fig. 3b). As the labels are annotated automatically and can be noisy, additional algorithms are used to aggregate multiple labelling functions to reduce noise. Another human-in-the-loop approach to reduce annotation costs is to prioritize the most valuable data for humans to annotate with active learning[64–66]. In active learning, which draws ideas from optimal experimental design[23], an algorithm is given a pool of unlabelled data points and selects the most informative points—for example, points that have high information gain or where the model is uncertain—for humans to annotate. As the learning algorithm can choose the data from which it learns, the number of data needed can be much smaller than in standard supervised learning[64].

Customized annotation interfaces such as eye-tracking can capture additional information during data annotation[67,68]. For instance, when examining chest X-rays, a radiologist's eye movements can contain rich information about parts of the image the expert is paying attention to. Such tracking can act as indirect annotations, also called weak supervision, to train the AI[69]. To ensure that data annotations are consistent and of high quality,



Is this video comment spam?	Return 'True' if "check out" occurs	Return 'True' if "subscribe" occurs	Aggregated labels
Please check out our videos, our songs are awesome! And...	True	False	True
Subscribe to me and I'll subscribe back	False	True	True
This looks so fun and it's a good song	False	False	False

Fig. 3 | Illustrations of methods for data valuation, data programming, data augmentation and data ablation. a, Data Shapley quantifies the value of each data point (star symbol) by measuring how models trained on different subsets of data would change if a particular point were removed from training ('crossed-out' faded star). Colors indicate class labels. **b**, Data programming uses user-defined labelling functions to automatically generate multiple potentially noisy labels for each input and then aggregates across the labels to reduce noise. **c**, Mixup augments the dataset by creating synthetic data interpolating pairs of existing data. Blue dots indicate existing data points in the training set (shaded blue region). The red dot indicates a synthetic data point created by interpolation of two existing data points. **d**, Data ablation assesses whether the model uses spurious artefacts (for example, snowy background) to make predictions. Panels **c** and **d** reproduced from ref. [17], Springer Nature Ltd.

researchers can use data assertions to define warnings that are triggered when annotation errors may have occurred[70]. For example, a warning function could be triggered if the LIDAR sensor detects an object but there is no bounding box annotated in the video frame, or when the bounding box annotations are inconsistent across consecutive video frames[70]. It's often useful to have a small dataset with gold-standard labels to evaluate and control the quality of human or algorithmically generated labels.

When the existing data are limited, data augmentation can be an effective approach to enhance the dataset and improve model reliability. Computer vision data can be augmented by image rotations, flips and other digital transformations[71], and text data can be augmented by automated writing style changes[72]. Recent works have proposed more sophisticated augmentation techniques suchas Mixup, which creates new training data by taking interpolations of pairs of training samples (Fig. 3c)[73,74]. Beyond hand-crafted data augmentation, researchers have also explored automating data augmentation pipelines for AI[75]. Furthermore, when unlabelled data are available, label augmentation can be performed by using an initial model to make predictions (these predictions are called pseudo-labels) and then training a potentially larger model on the combined data with both real and high-confidence pseudo-labels[76,77]. Substantial empirical works have shown the effectiveness of data augmentation. Recent theoretical analyses demonstrate that data augmentation such as Mixup and pseudo-labelling smooths and regularizes AI models, thus improving model robustness and generalizability[77,78].

Data to evaluate and monitor AI models

After a model has been trained, the goal of AI evaluation is to assess its generalizability and trustworthiness. To achieve this goal, the evaluation data should be carefully designed to capture the real-world settings where the model may be used while being sufficiently different from the model's training data. In medical research, for example, AI models are often trained on data from a small number of hospitals. However, when deploying at new hospitals, variations in data collection and processing can degrade model accuracy[79]. To evaluate model generalizability, the evaluation data should be collected from different hospitals with different data-processing pipelines. In other applications, the evaluation data should be collected from different sources and ideally labelled by different annotators as the training data. High-quality, human labels remain paramount for evaluation.

Table 1 | Selected resources for data-centric AI

Resource type	Name	URL/reference	Comment
Data documentation	Data Nutrition Labels	https://datanutrition.org/	Template for reporting datasets
	Frictionless	https://github.com/frictionlessdata/frictionless-py	Open source framework for managing metadata
	Data Version Control	https://dvc.org/	Versioning tool for datasets and machine learning models
Data repositories	Harvard Dataverse	https://dataverse.harvard.edu	Online platform to share, preserve and cite research data
	Zenodo	https://zenodo.org/	Visualization and analysis of image datasets
Data annotation	Label Studio	https://github.com/heartexlabs/label-studio	Data labelling tool for images, text, time series and so on
	Snorkel	https://github.com/snorkel-team/snorkel	Method to translate domain knowledge into labelling functions
	Modular Active Learning	https://github.com/modAL-python/modAL	Suite of active learning tools
Data-quality assurance	Tensorflow Data Validation	https://www.tensorflow.org/tfx/data_validation/get_started	Framework to monitor statistics of data and detect anomalies
	Cleanlab	https://github.com/cleanlab/cleanlab	Tool for finding mislabelled data
	Know Your Data	http://knowyourdata.withgoogle.com/	Visualization and analysis of image datasets
Data selection and improvement	Data Shapley Value	https://github.com/amiratag/DataShapley	Method to quantify the impact of different data points
	EaseML	https://ease.ml/	Toolkit for improving data quality
Data augmentation	Augmentor	https://augmentor.readthedocs.io/en/master/	Suite of data augmentation tools for images
Education resources	nlpaug	https://pypi.org/project/nlpaug	Data augmentation tools for NLP
Data-centric benchmarks	DCBench	https://www.datacentricai.cc/benchmark/	Suite of puzzles for different steps of the data-for-AI pipeline
	CATS4ML	https://cats4ml.humancomputation.com/	Crowdsourcing to curate challenging test data for AI models
	Dynabench	https://dynabench.org/	Dynamic data collection to evaluate NLP models
Education resources	Data-centric AI workshop	https://hai.stanford.edu/agenda-data-centric-ai-workshop	Talks on addressing data-centric challenges in AI
	NeurIPS workshop	https://datacentricai.org	Links to recent research

An important aspect of evaluation is to verify that the AI models do not use 'shortcut' strategies based on spurious correlations in the training data that may not generalize well. For example, in medical imaging, how the data are processed (for example cropping or image compression) can produce spurious correlations (that is, shortcuts) that the model picks up[13]. Though superficially helpful, these shortcuts can fail catastrophically when deployed in slightly different environments. Systematic data ablation is a good approach to check for potential model shortcuts. In data ablations, AI models are trained and tested on ablated inputs to surface signals of spurious correlations (Fig. 3d). For instance, a study on common natural language inference datasets found that AI models trained on only the first half of the text input achieved high accuracy in inferring the logical relationship between the first and the second half of the text, while humans cannot do better than a random guess on the same input[80]. This suggests that the AI models exploited spurious correlations as a shortcut solution for this task. Indeed, the authors found that specific linguistic phenomena such as negation in the text were highly correlated with the labels, which were exploited by the AI models. Data ablation is broadly applicable across domains. In medicine, for example, biologically relevant parts of images can be masked to assess whether the AI learns from spurious background or image quality artefacts[81].

AI evaluations are often limited to comparing the aggregate performance metrics (for example, AUC) over the entire test dataset. Going forwards, we recommend putting more emphasis on understanding the model's error modes on fine-grained subgroups of data. Even if the AI model works well at the aggregate data level, it may still exhibit systematic errors on specific subgroups of data, and characterizing such mistake clusters can provide deeper insights into the model's limitations. When metadata are available, a fine-grained evaluation approach should slice the evaluation data by sex, gender, ethnicity and geographical location of the participants in the dataset to the extent possible—for example, 'elderly Asian man' or 'Native American woman'—and quantify the model's performance on each data subgroup[82]. Multi-accuracy auditing is an algorithmic approach that automates the search for data subgroups where the AI model performs poorly[83]. Here an auditor algorithm is trained to predict and cluster the original model's errors using metadata, providing interpretable insights on where and why the AI model makes mistakes. When metadata are not available, methods such as Domino automatically identify clusters of evaluation data where the model is mistake prone and use text generation to create natural language explanations of these model errors[84].

The above sections have distinguished between data design and data sculpting, as most current AI research projects develop a dataset only once. But real-world AI users generally need to update datasets and models continuously[85]. For example, autonomous-vehicle companies such as Tesla collect millions of hours of data per day, some of which are then annotated by humans and used for training and evaluation of iterations of Tesla's models[86].

Continuous data development poses several challenges. First, both the data and the AI task can change over time: for example, perhaps a new model of vehicle appears on the road (that is, domain shift), or perhaps the AI developer wants to recognize a new class of objects (for example, school buses as distinct from regular buses), which changes the taxonomy of labels. It is wasteful to throw away the millions of hours of old labelled data. In addition, training and evaluation metrics should be carefully designed to weigh the new data and to use the appropriate data for each subtask[87]. Second, to acquire and use data continuously, users will need to automate much of the data-centric AI process. Such automation includes using algorithms to select which data to send to annotators and how to use it to retrain models, and only alerting the model developer if the process appears to be going wrong (for example, when accuracy metrics have dropped). Companies are starting to use tools to automate the machine learning lifecycle as part of the 'MLOps' trend. Example tools include open-source packages such as TFX and MLflow[88,89]. These packages include features to specify expectations about data (for example, rough distribution of classes) and automatically raise alerts if new data fail these expectations or if the evaluation metrics for new models are poor.

Data regulation and policies

Government regulation and data policy will play an important role in promoting, constraining and shaping data-centric and trustworthy AI. New data regulatory requirements support a shift towards the kinds of approaches we have discussed here. Europe's draft AI regulation, for instance, requires that "[t]raining, validation and testing data sets shall be relevant, representative, free of errors and complete" for high-risk AI systems[90]. Developers must conduct 'conformity assessments' of AI products. Although such compliance mandates can add substantial overhead, they could improve data design, sculpting and evaluation.

Some data regulatory requirements may in fact impede developing trustworthy AI. Privacy concerns can impede the curation of representative data with access to sensitive attributes[91,92]. As a consequence, many AI algorithms, especially in the biomedical sciences, are developed using data from only one or a small number of sites, which limits generalizability[21], and evaluation of subgroup performance can be challenging[92]. Privacy's typical reliance on consent and opt-outs can itself contribute to data bias[93]. And companies can hoard expensive private datasets under restrictive licenses[94], leaving academic researchers to scrape the Internet for lower-quality and potentially biased data[95].

AI and data regulation will need to balance privacy and access interests. A promising direction is to focus on data agency to reduce the transaction costs of controlling one's own data and increase the transparency of how the data are used. Public willingness to share de-identified medical data can be high with appropriate design and agency[96]. The UK Biobank, wherein 500,000 volunteers shared genomic and health record data for science, is an exemplar for granting such agency to volunteers, who retain control rights to their medical information[97]. Similarly, researchers need improved mechanisms for securely sharing data, collaboratively curating data and navigating the thicket of data-use agreements via more standardized templates[91]. A promising path lies in regulatory requirements that foster data portability[98,99], which may enable an ecosystem of data intermediaries to improve, streamline, and tailor data sharing and control. Nevertheless, much technical and implementation work remains to be done to realize this potential[100].

In addition, much more work is needed to harmonize aspirational regulatory goals with concrete implementation. As an initial step, VDE (a standards development organization in Europe) has recently released guidance[101,102] aiming to aid compliance with the European Union AI Act through the usage of datasheets and model cards[103]. Algorithmic approaches to protect privacy, for instance, can have important trade-offs and limitations: differential privacy (which adds noise to datasets to prevent leakage of private information) can degrade the ability to assess bias by subgroup[104]; federated learning may still leak private information when model updates are shared across devices[105]. Similarly, data erasure can be challenging to implement, as large AI models memorize user information[106], and selective erasure can magnify coverage gaps in the data[107]. Technologies for cleaning data pipelines can help to adjust for such trade-offs. Machine learning can, for instance, improve record linkage and inferences when sensitive attributes are lacking[29,108,109].

Discussion

As AI model-building matures, creating appropriate data pipelines is often the biggest bottleneck for developing trustworthy algorithms, that is, algorithms that work reliably well across diverse users and settings. Especially because AI applications often involve large-scale and unstructured data, automated or technology-assisted human-in-the-loop methods are needed to systematically address the data challenges. The regulatory landscape will also be critical to enabling trustworthy, data-centric AI and the trade-offs must be artfully managed. This Perspective highlights several promising approaches and relevant considerations at different steps of the data pipeline. It is not scoped to be a comprehensive review because data-centric AI methodology is in its early stages and much more research is needed.

In particular, evaluation of AI datasets is still ad hoc and not often done. For example, the ImageNet dataset has been one of the most popular and impactful benchmark datasets for AI since 2009. It was created by querying image search engines followed by manual curation and annotation[110]. An evaluation of ImageNet, only conducted in 2019, revealed significant limitations in the dataset—annotator biases are prevalent. In particular, 56% of the people categories in ImageNet contained potentially offensive labels and needed to be removed[111]. Important progress is being made to clean ImageNet, but it is less clear how decisions on filtering or rebalancing categories affect representation learning or the downstream performance of models trained on the updated ImageNet. While efforts such as Know Your Data (https://knowyourdata.withgoogle.com) help users to visually inspect image data, how to make data evaluation scalable and systematic is an open challenge. In addition, AI has been mostly built around static datasets such as ImageNet, which do not capture the constantly changing data that models face in deployment. Data streams that are continuously updated, along with methods for evaluating such streams, would be important new resources for AI development.

One approach to make dataset creation, cleaning and evaluation more rigorous is to create benchmarks for each of these tasks. The recently released Data-centric AI Benchmarks, which is a suite of hundreds of self-contained data puzzles, is a step in this direction. Each puzzle contains a dataset, a particular data-pipeline task (for example, data cleaning or data selection) and the ground-truth solution[112]. Researchers are encouraged to compete and submit their methods to tackle these data tasks, which are then automatically tested on a collection of hidden data puzzles. For example, one set of puzzles involves datasets that are intentionally corrupted with noise or with spurious correlations, where researchers can evaluate their methods for removing such artefacts. They can then be used to systematically evaluate methods for cleaning and identifying biases in data. The goal of Data-centric AI Benchmarks is to foster the development of new scalable tools for improving data quality by making it easier to compare the conformance of these tools. While such benchmarks can provide useful quantitative feedback, they currently cover only a fraction of the data-centric challenges, mostly focused on data cleaning, selection and annotation for natural images. More benchmarks need to be developed for other data pipeline tasks, data types and different application domains (for example, clinical data). Expanding data-centric benchmarks is an important direction of new research.

The data-centric challenges discussed here are especially salient in developing regions due to resource limitations and greater data heterogeneity. While this Perspective discusses algorithms to improve the quality and diversity of data, it is important to recognize that there are socio-technical challenges in dataset creation that require broader participation to address[19,45,113]. For example, AI researchers in Africa are starting to extend common AI datasets for their local context. Inspired by the computer vision COCO (Common Object in Context) dataset, COCO-Africa is an object detection, segmentation and captioning dataset containing scenes and objects that are more likely to be observed in Africa[114]. In natural language processing (NLP), MasakhaNER provides a large curated dataset of ten African languages with named-entity annotations[115]. Similarly, Knowledge-for-all and AI for Development led projects developing datasets for under-represented African languages[116]. There is still a large gap in public datasets that require more expert curation in the African context—for example, healthcare images such as computed-tomography scans are largely unavailable[117]. These data-centric efforts will be critical to ensure that advances in AI models generalize to and benefit broader populations.

Topics on how to create high-quality and responsible data pipelines should be incorporated into the AI curriculum. References like the ones listed in Table 1 can be useful teaching resources. It is important for students to appreciate potential pitfalls in the data used to develop their AI models and to understand how to use systematic techniques to improve the data. Improvements in data pipelines and AI models form a positive feedback loop. More reliable and scalable frameworks to sculpt and evaluate datasets and data streams enhance the reliability of the models developed on this data. At the same time, better calibrated AI models can facilitate the detection of anomalies, errors and biases in its development data (for example, by associating greater uncertainty to poor-quality points). A data-centric focus will be integral to the next stage of AI development, especially as we translate models from research sandbox to real-world deployment.

Received: 3 April 2022; Accepted: 30 June 2022;
Published online: 17 August 2022

References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
2. Ouyang, D. et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **580**, 252–256 (2020).
3. Hutson, M. Robo-writers: the rise and risks of language-generating AI. *Nature* **591**, 22–25 (2021).
4. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, 8026–8037 (2019).
5. Abadi, M. et al. TensorFlow: a system for large-scale machine learning. In *Proc. 12th USENIX Symposium on Operating Systems Design and Implementation* 265–283 (USENIX Association, 2016).
6. Zhang, X. et al. Dnnbuilder: an automated tool for building high-performance dnn hardware accelerators for fpgas. In *2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)* 1–8 (IEEE, 2018).
7. Code-free machine learning: AutoML with AutoGluon, Amazon SageMaker, and AWS Lambda. *AWS Machine Learning Blog* https://aws.amazon.com/blogs/machine-learning/code-free-machine-learning-automl-with-autogluon-amazon-sagemaker-and-aws-lambda/ (2020).
8. Korot, E. et al. Code-free deep learning for multi-modality medical image classification. *Nat. Mach. Intell.* **3**, 288–298 (2021).
9. Dimensional Research. What Data Scientists Tell Us About AI Model Training Today. *Alegion* https://content.alegion.com/dimensional-researchs-survey (2019).
10. Forrester Consulting. Overcome Obstacles To Get To AI At Scale. *IBM* https://www.ibm.com/downloads/cas/VBMPEQLN (2020).
11. State of data science 2020. *Anaconda* https://www.anaconda.com/state-of-data-science-2020 (2020).
12. Petrone, J. Roche pays $1.9 billion for Flatiron's army of electronic health record curators. *Nat. Biotechnol.* **36**, 289–290 (2018).
13. Geirhos, R. et al. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).
14. Daneshjou, R. et al. Disparities in dermatology AI: assessments using diverse clinical images. Preprint at http://arxiv.org/abs/2111.08006 (2021).
15. Koch, B., Denton, E., Hanna, A. & Foster, J. G. Reduced, reused and recycled: the life of a dataset in machine learning research. In *NeurIPS 2021 Datasets and Benchmarks Track* 50 (OpenReview, 2021).
16. Coleman, C. et al. DAWNBench: An end-to-end deep learning benchmark and competition. In *NeurIPS MLSys Workshop* 10 (MLSys, 2017).
17. Krishna, R. et al. Visual genome: connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision* **123**, 32–73 (2017).
18. Kiela, D. et al. Dynabench: rethinking benchmarking in NLP. In *Proc. 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 4110–4124 (ACL, 2021).
19. Sambasivan, N. et al. 'Everyone wants to do the model work, not the data work': data cascades in high-stakes AI. In *Proc. 2021 CHI Conference on Human Factors in Computing Systems* (ACM, 2021); https://doi.org/10.1145/3411764.3445518
20. Daneshjou, R., Smith, M. P., Sun, M. D., Rotemberg, V. & Zou, J. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. *JAMA Dermatol.* **157**, 1362–1369 (2021).
21. Wu, E. et al. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat. Med.* **27**, 582–584 (2021).
22. Paullada, A., Raji, I. D., Bender, E. M., Denton, E. & Hanna, A. Data and its (dis)contents: a survey of dataset development and use in machine learning research. *Patterns* **2**, 100336 (2021).
23. Smucker, B., Krzywinski, M. & Altman, N. Optimal experimental design. *Nat. Methods* **15**, 559–560 (2018).
24. Fan, W. & Geerts, F. Foundations of data quality management. *Synth. Lect. Data Manag.* **4**, 1–217 (2012).
25. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **54**, 1–35 (2021).
26. Buolamwini, J. & Gebru, T. Gender shades: intersectional accuracy disparities in commercial gender classification. In *Proc. 1st Conference on Fairness, Accountability and Transparency* 77–91 (PMLR, 2018).
27. Kaushal, A., Altman, R. & Langlotz, C. Geographic distribution of US cohorts used to train deep learning algorithms. *J. Am. Med. Assoc.* **324**, 1212–1213 (2020).
28. Zou, J. & Schiebinger, L. AI can be sexist and racist—it's time to make it fair. *Nature* **559**, 324–326 (2018).
29. Coston, A. et al. Leveraging administrative data for bias audits: assessing disparate coverage with mobility data for COVID-19 policy. In *Proc. 2021 ACM Conference on Fairness, Accountability, and Transparency* 173–184 (ACM, 2021); https://doi.org/10.1145/3442188.3445881
30. Mozilla. Mozilla Common Voice receives $3.4 million investment to democratize and diversify voice tech in East Africa. *Mozilla Foundation* https://foundation.mozilla.org/en/blog/mozilla-common-voice-receives-34-million-investment-to-democratize-and-diversify-voice-tech-in-east-africa/ (2021).
31. Reid, K. Community partnerships and technical excellence unlock open voice technology success in Rwanda. *Mozilla Foundation* https://foundation.mozilla.org/en/blog/open-voice-success-in-rwanda/ (2021).
32. Van Noorden, R. The ethical questions that haunt facial-recognition research. *Nature* **587**, 354–358 (2020).
33. Build more ethical AI. *Synthesis AI* https://synthesis.ai/use-cases/bias-reduction/ (2022).
34. Kortylewski, A. et al. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops* 2261–2268 (IEEE, 2019).
35. Nikolenko, S. I. *Synthetic Data for Deep Learning* Vol. 174 (Springer, 2021).

36. Srivastava, S. et al. BEHAVIOR: Benchmark for Everyday Household Activities in Virtual, Interactive, and Ecological Environments. In *Proc. 5th Annual Conference on Robot Learning* Vol. 164 477–490 (PMLR, 2022).
37. Li, C. et al. iGibson 2.0: object-centric simulation for robot learning of everyday household tasks. In *Proc. 5th Annual Conference on Robot Learning* Vol. 164 455–465 (PMLR, 2022).
38. Höfer, S. et al. Perspectives on Sim2Real transfer for robotics: a summary of the R:SS 2020 workshop. Preprint at http://arxiv.org/abs/2012.03806 (2020)
39. Egger, B. et al. 3D morphable face models—past, present, and future. *ACM Trans. Graph.* **39**, 1–38 (2020).
40. Choi, K., Grover, A., Singh, T., Shu, R. & Ermon, S. Fair generative modeling via weak supervision. *Proc. Mach. Learn. Res.* **119**, 1887–1898 (2020).
41. Holland, S., Hosny, A., Newman, S., Joseph, J. & Chmielinski, K. The dataset nutrition label: a framework to drive higher data quality standards. Preprint at https://arxiv.org/abs/1805.03677 (2018).
42. Gebru, T. et al. Datasheets for datasets. *Commun. ACM* **64**, 86–92 (2021).
43. Bender, E. M. & Friedman, B. Data statements for natural language processing: toward mitigating system bias and enabling better science. *Trans. Assoc. Comput. Linguist.* **6**, 587–604 (2018).
44. Wang, A., Narayanan, A. & Russakovsky, O. REVISE: a tool for measuring and mitigating bias in visual datasets. In *European Conference on Computer Vision* 733–751 (Springer, 2020).
45. Miceli, M. et al. Documenting computer vision datasets: an invitation to reflexive data practices. In *Proc. 2021 ACM on Conference on Fairness, Accountability, and Transparency* 161–172 (2021).
46. Scheuerman, M. K., Hanna, A. & Denton, E. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proc. ACM Hum. Comput. Interact.* **5**, 317:1–317:37 (2021).
47. Liang, W. & Zou, J. MetaShift: a dataset of datasets for evaluating contextual distribution shifts and training conflicts. In *International Conference on Learning Representations* 400 (OpenReview, 2022).
48. Ghorbani, A. & Zou, J. Data Shapley: equitable valuation of data for machine learning. *Proc. Mach. Learn. Res.* **97**, 2242–2251 (2019).
49. Kwon, Y., Rivas, M. A. & Zou, J. Efficient computation and analysis of distributional Shapley values. *Proc. Mach. Learn. Res.* **130**, 793–801 (2021).
50. Jia, R. et al. Towards efficient data valuation based on the Shapley value. *Proc. Mach. Learn. Res.* **89**, 1167–1176 (2019).
51. Koh, P. W. & Liang, P. Understanding black-box predictions via influence functions. *Proc. Mach. Learn. Res.* **70**, 1885–1894 (2017).
52. Kwon, Y. & Zou, J. Beta Shapley: a unified and noise-reduced data valuation framework for machine learning. In *Proc. 25th International Conference on Artificial Intelligence and Statistics* Vol. 151 8780–8802 (PMLR, 2022).
53. Northcutt, C., Jiang, L. & Chuang, I. Confident learning: estimating uncertainty in dataset labels. *J. Artif. Intell. Res.* **70**, 1373–1411 (2021).
54. Northcutt, C. G., Athalye, A. & Mueller, J. Pervasive label errors in test sets destabilize machine learning benchmarks. In *NeurIPS 2021 Datasets and Benchmarks Track* 172 (OpenReview, 2021).
55. Dodge, J. et al. Documenting large webtext corpora: a case study on the Colossal Clean Crawled Corpus. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing* 12861305 (ACL, 2021).
56. Krishnan, S., Wang, J., Wu, E., Franklin, M. J. & Goldberg, K. ActiveClean: interactive data cleaning for statistical modeling. *Proc. VLDB Endow.* **9**, 948–959 (2016).
57. Rolnick, D., Veit, A., Belongie, S. & Shavit, N. Deep learning is robust to massive label noise. Preprint at http://arxiv.org/abs/1705.10694 (2018).
58. Geiger, A., Lenz, P. & Urtasun, R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* 3354–3361 (IEEE, 2012); https://doi.org/10.1109/CVPR.2012.6248074
59. Sun, P. et al. Scalability in perception for autonomous driving: Waymo Open Dataset. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2446–2454 (IEEE, 2020).
60. Park, J., Krishna, R., Khadpe, P., Fei-Fei, L. & Bernstein, M. AI-based request augmentation to increase crowdsourcing participation. *Proc. AAAI Conf. Hum. Comput. Crowdsourcing* **7**, 115–124 (2019).
61. Ratner, A. et al. Snorkel: rapid training data creation with weak supervision. *VLDB J.* **29**, 709–730 (2020).
62. Ratner, A. J., De, Sa,C. M., Wu, S., Selsam, D. & Ré, C. Data programming: creating large training sets, quickly. *Adv. Neural Inf. Process. Syst.* **29**, 3567–3575 (2016).
63. Liang, W., Liang, K.-H. & Yu, Z. HERALD: an annotation efficient method to detect user disengagement in social conversations. In *Proc. 59th Annual Meeting of the Association for Computational Linguistics* 3652–3665 (ACL, 2021).
64. Settles, B. Active Learning Literature Survey. *MINDS@UW* http://digital.library.wisc.edu/1793/60660 (University of Wisconsin-Madison, 2009).
65. Coleman, C. et al. Similarity search for efficient active learning and search of rare concepts. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 36 6402–6410 (2022).
66. Liang, W., Zou, J. & Yu, Z. ALICE: Active Learning with Contrastive Natural Language Explanations. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing* 4380–4391 (ACL, 2020).
67. Hollenstein, N. & Zhang, C. Entity recognition at first sight: improving NER with eye movement information. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 1–10 (ACL, 2019).
68. Valliappan, N. et al. Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nat. Commun.* **11**, 4553 (2020).
69. Saab, K. et al. Observational supervision for medical image classification using gaze data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 603–614 (Springer, 2021).
70. Kang, D., Raghavan, D., Bailis, P. & Zaharia, M. Model assertions for debugging machine learning. In *NeurIPS MLSys Workshop* 23 (MLSys, 2020).
71. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012).
72. Sennrich, R., Haddow, B. & Birch, A. Improving neural machine translation models with monolingual data. In *Proc. 54th Annual Meeting of the Association for Computational Linguistics* 86–96 (ACL, 2016).
73. Zhang, H., Cissé, M., Dauphin, Y. N. & Lopez-Paz, D. mixup: beyond empirical risk minimization. In *Proc. International Conference on Learning Representations* 296 (OpenReview, 2018).
74. Liang, W. & Zou, J. Neural group testing to accelerate deep learning. In *2021 IEEE International Symposium on Information Theory (ISIT)* 958–963 (IEEE, 2021); https://doi.org/10.1109/ISIT45174.2021.9518038
75. Cubuk, E. D., Zoph, B., Shlens, J. & Le, Q. V. Randaugment: practical automated data augmentation with a reduced search space. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* 702–703 (IEEE, 2020).
76. Caron, M., Bojanowski, P., Joulin, A. & Douze, M. Deep clustering for unsupervised learning of visual features. In *Proc. European Conference on Computer Vision (ECCV)* 132–149 (2018).
77. Deng, Z., Zhang, L., Ghorbani, A. & Zou, J. Improving adversarial robustness via unlabeled out-of-domain. *Data. Proc. Mach. Learn. Res.* **130**, 2845–2853 (2021).
78. Zhang, L., Deng, Z., Kawaguchi, K., Ghorbani, A. & Zou, J. How does mixup help with robustness and generalization? In *Proc. International Conference on Learning Representations* 79 (OpenReview, 2021).
79. Zech, J. R. et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med.* **15**, e1002683 (2018).
80. Gururangan, S. et al. Annotation artifacts in natural language inference data. In *Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 107–112 (ACL, 2018).
81. Hughes, J. W. et al. Deep learning evaluation of biomarkers from echocardiogram videos. *EBioMedicine* **73**, 103613 (2021).
82. Tannenbaum, C., Ellis, R. P., Eyssel, F., Zou, J. & Schiebinger, L. Sex and gender analysis improves science and engineering. *Nature* **575**, 137–146 (2019).
83. Kim, M. P., Ghorbani, A. & Zou, J. Y. Multiaccuracy: black-box post-processing for fairness in classification. In *Proc. 2019 AAAI/ACM Conference on AI, Ethics, and Society* 247–254 (ACM, 2019); https://doi.org/10.1145/3306618.3314287
84. Eyuboglu, S. et al. Domino: discovering systematic errors with cross-modal embeddings. In *Proc. International Conference on Learning Representations* 1 (OpenReview, 2022).
85. Karlaš, B. et al. Building continuous integration services for machine learning. In *Proc. 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 2407–2415 (ACM, 2020); https://doi.org/10.1145/3394486.3403290
86. Lambert, F. Tesla is collecting insane amount of data from its full self-driving test fleet. *Electrek* https://electrek.co/2020/10/24/tesla-collecting-insane-amount-data-full-self-driving-test-fleet/ (2020).
87. Azizzadenesheli, K., Liu, A., Yang, F. & Anandkumar, A. Regularized learning for domain adaptation under label shifts. In *Proc. International Conference on Learning Representations* 432 (OpenReview, 2019).
88. Baylor, D. et al. TFX: a TensorFlow-based production-scale machine learning platform. In *Proc. 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1387–1395 (ACM, 2017); https://doi.org/10.1145/3097983.3098021
89. Zaharia, M. et al. Accelerating the machine learning lifecycle with MLflow. *IEEE Data Eng Bull* **41**, 39–45 (2018).
90. *Proposal for a Regulation of the European Parliament and the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts* COM(2021) 206 final (European Commission, 2021); https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN

91. Mello, M. M., Triantis, G., Stanton, R., Blumenkranz, E. & Studdert, D. M. Waiting for data: barriers to executing data use agreements. *Science* **367**, 150–152 (2020).
92. Andrus, M., Spitzer, E., Brown, J. & Xiang, A. What we can't measure, we can't understand: challenges to demographic data procurement in the pursuit of fairness. In *Proc. 2021 ACM Conference on Fairness, Accountability, and Transparency* 249–260 (ACM, 2021).
93. Woolf, S. H., Rothemich, S. F., Johnson, R. E. & Marsland, D. W. Selection bias from requiring patients to give consent to examine data for health services research. *Arch. Fam. Med.* **9**, 1111–1118 (2000).
94. Marshall, E. Is data-hoarding slowing the assault of pathogens? *Science* **275**, 777–780 (1997).
95. Baeza-Yates, R. Data and algorithmic bias in the web. In *Proc. 8th ACM Conference on Web Science* 1 (ACM, 2016).
96. Garrison, N. A. et al. A systematic literature review of individuals' perspectives on broad consent and data sharing in the United States. *Genet. Med.* **18**, 663–671 (2016).
97. Cox, N. UK Biobank shares the promise of big data. *Nature* **562**, 194–195 (2018).
98. *Art. 20 GDPR: Right to Data Portability* https://gdpr-info.eu/art-20-gdpr/ (General Data Protection Regulation, 2021).
99. *TITLE 1.81.5. California Consumer Privacy Act of 2018* https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5 (California Legislative Information, 2018).
100. Krämer, J., Senellart, P. & de Streel, A. *Making Data Portability More Effective for the Digital Economy: Economic Implications and Regulatory Challenges* (CERRE, 2020).
101. Loh, W., Hauschke, A., Puntschuh, M. & Hallensleben, S. *VDE SPEC 90012 V1.0: VCIO Based Description of Systems for AI Trustworthiness Characterisation* (VDE Press, 2022).
102. Can artificial intelligence conform to values? VDE SPEC as the basis for future developments. *VDE Presse* https://www.vde.com/ai-trust (2022).
103. Mitchell, M. et al. Model cards for model reporting. In *Proc. Conference on Fairness, Accountability, and Transparency* 220–229 (ACM, 2019).
104. Bagdasaryan, E., Poursaeed, O. & Shmatikov, V. Differential privacy has disparate impact on model accuracy. *Adv. Neural Inf. Process. Syst.* **32**, 15453–15462 (2019).
105. Lyu, L., Yu, H. & Yang, Q. Threats to federated learning: a survey. Preprint at http://arxiv.org/abs/2003.02133 (2020).
106. Izzo, Z., Smart, M. A., Chaudhuri, K. & Zou, J. Approximate data deletion from machine learning models. *Proc. Mach. Learn. Res.* **130**, 2008–2016 (2021).
107. Johnson, G. A., Shriver, S. K. & Du, S. Consumer privacy choice in online advertising: who opts out and at what cost to industry? *Mark. Sci.* **39**, 33–51 (2020).
108. Wilson, D. R. Beyond probabilistic record linkage: Using neural networks and complex features to improve genealogical record linkage. In *2011 International Joint Conference on Neural Networks* 9–14 (IEEE, 2011); https://doi.org/10.1109/IJCNN.2011.6033192
109. Kallus, N., Mao, X. & Zhou, A. Assessing algorithmic fairness with unobserved protected class using data combination. *Manag. Sci*. https://doi.org/10.1287/mnsc.2020.3850 (2021).
110. Deng, J. et al. Imagenet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
111. Yang, K., Qinami, K., Fei-Fei, L., Deng, J. & Russakovsky, O. Towards fairer datasets: filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In *Proc. 2020 Conference on Fairness, Accountability, and Transparency* 547–558 (ACM, 2020); https://doi.org/10.1145/3351095.3375709
112. DCBench: a benchmark of data-centric tasks from across the machine learning lifecycle. *DCAI* https://www.datacentricai.cc/benchmark/ (2021).
113. Zaugg, I. A., Hossain, A. & Molloy, B. Digitally-disadvantaged languages. *Internet Policy Rev*. https://doi.org/10.14763/2022.2.1654 (2022).
114. Victor, D. COCO-Africa: a curation tool and dataset of common objects in the context of Africa. In *2018 Conference on Neural Information Processing, 2nd Black in AI Workshop* 1 (NeurIPS, 2019).
115. Adelani, D. I. et al. MasakhaNER: Named Entity Recognition for African languages. *Trans. Assoc. Comput. Linguist.* **9**, 1116–1131 (2021).
116. Siminyu, K. et al. AI4D—African language program. Preprint at http://arxiv.org/abs/2104.02516 (2021).
117. Frija, G. et al. How to improve access to medical imaging in low- and middle-income countries? *EClinicalMedicine* **38**, 101034 (2021).

Acknowledgements

We thank T. Hastie, R. Daneshjou, K. Vodrahalli and A. Ghorbani for discussions. J.Z. is supported by a NSF CAREER grant.

Competing interests

M.Z. is a co-founder of Databricks. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-022-00516-1.

Correspondence should be addressed to James Zou.

Peer review information *Nature Machine Intelligence* thanks Emmanuel Kahembwe and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.