Designing Data: Proactive Data Collection and Iteration for Machine Learning Using Reflexive Planning, Monitoring, and Density Estimation

Aspen Hopkins [1 2] **Fred Hohman** [2] **Luca Zappella** [2] **Xavier Suau Cuadros** [2] **Dominik Moritz** [2]

Abstract

Lack of diversity in data collection has caused significant failures in machine learning (ML) applications. While ML developers perform post-collection interventions, these are time intensive and rarely comprehensive. Thus, new methods to track & manage data collection, iteration, and model training are necessary for evaluating whether datasets reflect real world variability. We present *designing data*, an iterative approach to data collection connecting HCI concepts with ML techniques. Our process includes (1) Pre-Collection Planning, to reflexively prompt and document expected data distributions; (2) Collection Monitoring, to systematically encourage sampling diversity; and (3) Data Familiarity, to identify samples that are unfamiliar to a model using density estimation. We apply designing data to a data collection and modeling task. We find models trained on "designed" datasets generalize better across intersectional groups than those trained on similarly sized but less targeted datasets, and that data familiarity is effective for debugging datasets.

1. Introduction

Curating representative training and testing datasets is fundamental to developing robust, generalizable machine learning (ML) models. However, understanding what is representative for a specific task is an iterative process. ML practitioners need to change data, models, and their associated processes as they become more familiar with their modeling task, as the state of the world evolves, and as products are updated or maintained. Iteration directed by this evolving understanding seeks to improve model performance, often editing datasets to ensure desired outcomes.

[1] Massachusetts Institute of Technology [2]Apple. Correspondence to: Aspen Hopkins <dataspen@mit.edu>.

AI & HCI Workshop at the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Failure to effectively recognize data quality and coverage needs can lead to biased ML models (Mitchell et al., 2020). Such failures are responsible for the perpetuation of systemic power and access differentials and the deployment of inaccessible or defective product experiences. Yet building representative datasets is an arduous, historically difficult undertaking (Tayi & Ballou, 1998; Sambasivan et al., 2021) that relies on the efficacy of human-specified data requirements.

To ensure a dataset covers all, or as many, characteristics as possible, specifications must be the result of a comprehensive enumeration of possible dimensions—an open and hard problem that few have practically grappled with in the context of ML. Further contributing to this difficulty is the realization that it is not enough for the training datasets to be aligned with expected distributions: they must also include enough examples from conceptually harder or less common categories if said categories are to be learned (Asudeh et al., 2019). Failure to sufficiently consider both the critical dimensions of data and their relative complexity can have troubling consequences. Instances of such missteps span issues of justice, healthcare, hiring practices, voice and face recognition, and loan qualifications, wherein biases of data and algorithms limit technological use and cause harm (Buolamwini & Gebru, 2018; Asudeh et al., 2019; Palanica et al., 2019; Angwin et al., 2016; Noble, 2018). Yet understanding these data requirements even after training is difficult; knowing them *a priori* is exceptionally so.

Rather than emphasize tools that enable better collection and data iteration practices—that *design better data*—research in fairness and machine learning has largely focused on prescriptive "how-to" frameworks, definitions of fairness, and post-collection analysis techniques (Amershi et al., 2019; Yang et al., 2020). While there are exceptions to this (Hohman et al., 2020b), the hidden technical debt (Sculley et al., 2015) accumulated from poor data design remains an under explored space. To reduce this technical debt and encourage diverse datasets, methods of externalizing data collection, iteration, and training are necessary checks for ensuring datasets reflect diverse experiences and are robust when deployed in real-world models.

Contributions We present *designing data*, an iterative, bias

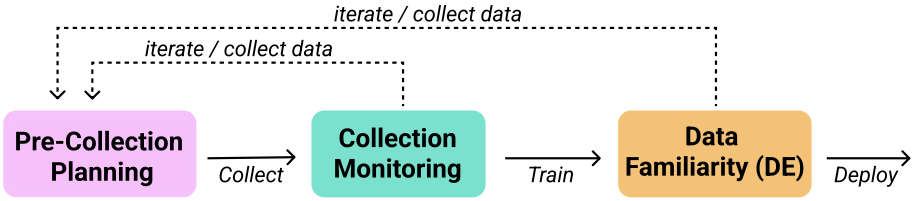

Figure 1. Designing data includes (I) **Pre-Collection Planning**, (II) **Collection Monitoring**, & (III) **Data Familiarity**. Each supplements the other, and should be used iteratively.

mitigating approach to data planning, collection, and ML development; we implement this approach in an interactive dashboard that scaffolds each step for practitioner use. Motivated by 24 formative interviews 3, designing data is a structured, holistic parallel to the current standards for developing production-quality datasets (Hohman et al., 2020b). Each step proactively introduces interventions to improve models prior to deployment:

1. ***Pre-Collection Planning*** prioritizes reflexive consideration for domain and data needs prior to modeling, documents expected distributions of attributes, and highlights potential biases through questions related to class or characteristic representation within data.

2. ***Collection Monitoring*** communicates insight into dataset evolution, allowing users to make targeted adjustments to collection processes based on new insight or disparities between expected distributions and existing data.

3. ***Data Familiarity*** borrows from Out-of-Distribution (OOD) methodologies to identify data that a model perceives as unfamiliar, creating new directives for data and model iteration.

We demonstrate designing data's effectiveness through a case study using inertial measurement units (IMU) data–time series data representing X, Y, & Z positioning–to classify hand position while texting. First, we collected data iteratively, using Pre-Collection Planning and Monitoring steps to build a diverse dataset. Then, we use leave-one-out cross-validation to evaluate how these initial steps influence performance. Finally, we evaluate familiarity to first debug a dataset, then to direct collection efforts. Each step is centralized within a dashboard. We find models trained on highly diverse data outperform those trained on less diverse data across intersectional groups.

2. Related Work

ML Documentation A close alternative to *designing data* is model and dataset documentation, such as Model Cards (Mitchell et al., 2019) or Datasheets for Datasets (Gebru et al., 2021). Such work details what to include within said documentation for transparency in downstream model and dataset use (Koesten et al., 2019; Bender & Friedman, 2018; Arnold et al., 2019). This type of documentation has been widely adopted (Hopkins & Booth, 2021). However, documentation on its own is limited, as it typically accompanies a model or dataset upon release rather than shaping its development. For instance, Model Cards do not explicitly guide how to reconcile model weakness or actively direct data collection to improve fairness–rather, they are intended for transparency. Our contribution is embedding the transparency revealed by such documentation into our designing data process and dashboard. Through simple prompts, users actively surface and engage with their priors–which later informs their evaluation.

Reflexivity and Self-Reflection In social science, the practice of *reflexivity* is a way to externalize implicit subjectivity present in data collection and interpretation (Fish & Stark, 2021; Dodgson, 2019). Reflexivity entails deliberately examining practitioners' own assumptions, practices, and belief systems, then *contrasting* them with alternative perspectives. This acknowledges positionality—how differences in social position and power shape identities. While reflexivity is typically practiced retrospectively, (Soedirgo & Glas, 2020) outlined how it could be an active process through *"ongoing reflection about our own social location and [...] our assumptions regarding others' perceptions."* This approach includes recording assumptions of positionality; routinizing reflexivity; including other actors in the process; and communicating reflexive outcomes with data.

Separate from yet related to reflexivity is recent data visualization work prompting viewers to reflect on their individual beliefs of data (Kim et al., 2018; Hohman et al., 2020a). Examples of this include The New York Times "You Draw It" visualizations, where readers draw a projected trendline of what they think the data looks like and then compare that projection with the real data (Aisch et al., 2015; Katz, 2017; Buchanan et al., 2017), and an MIT Tech Review article illustrating the complexity of building fair recidivism models wherein readers change different hyperparameters then contrast their outcomes to existing models (Hao & Stray, 2019; Stray & Hao, 2020). Such visualizations act as powerful tools ensuring self-reflection—an important factor in

building representative datasets.

Learnability and Familiarity Having diverse characteristics with sample size parity in data is one step towards mitigating bias but does not ensure equitable learning across classes. Instead, these approaches ignore data learnability: having an equal number of samples per class neglects the fact that some classes are inherently easier to learn than others (Ben-David et al., 2019; Schapire, 1990; Klawonn et al., 2019). Unfortunately, this complexity might not become apparent until after deployment. In order to build better systems, discovering what has not been learned by a model is critical. In the past, testing data acted as a proxy for this evaluation. But while test datasets offer excellent thresholds for performance expectations, these are not the same measurements: testing accuracy measures whether a model was correct in its classification without consideration for how it reached its conclusion. For this reason, models are often poorly calibrated (Van Calster et al., 2019). To address this, we use density estimation (DE)–a common technique for measuring OOD–to evaluate how familiar a model is to a given sample. Our implementation of DE is most similar to (Lee et al., 2018), who use it to detect anomalous samples *post*-training (e.g., adversarial attacks). Unlike prior work, we extend DE techniques into training to direct data efforts.

Our use of DE to guide data collection is a response to two drawbacks we note from other work seeking to improve data representation and model performance: (1) alternative approaches require *prior awareness* of what facets of data are variable—for data that people are unaccustomed to, this might be impossible—and (2) measuring diversity within a given dataset does not account for what the model actually *learns* (Hooker, 2021; Schmidt et al., 2018). Because it is *model* outputs that we are concerned for, these are important weaknesses to counter.

3. Formative Interviews

To understand ML practitioners' data collection needs, we conducted 24 semi-structured exploratory interviews with individuals possessing extensive machine learning and data collection experience within a large technology company. The interviewees ranged from ML research scientists and designers focused on ML experiences to engineering product managers. Interviews lasted $\sim$ one hour. As interviews progressed, common themes surfaced that directed our attention to issues of data coverage and representation. We describe those themes (T1-T3) below.

Critical dimensions of data are hard to know a priori (T1). A proactive approach to data collection requires knowing what axes are important for observation. As one interviewee put it, *"how do we know who and what is missing?"* This was a shared difficulty described in nearly all interviews. While it is typically impossible to have a complete understanding of critical dimensions of the data before starting data collection, there are some common characteristics for human-centric data collection based on existing knowledge of population statistics and power imbalances. How useful that information is depends on the context of how the data is used—for example, a person's accent and speech pathology are important factors in speech recognition, but not for creating a personalized wine recommendation. Unlike summary statistics, surfacing noisy data or missing subsets–essentially "debugging" data– requires significant effort. As another interviewee said, *"fairness analysis is useful to a point"*. Generic tools for surfacing these nuanced limitations of data do not exist, but there is both need and desire for them (Holstein et al., 2019).

Difficulty of collection leads to compromises (T2). Data collection is a difficult process to launch, requiring significant tooling. This difficulty contributes to issues of representation in data, as the emphasis in early data collection is on *how* to collect and structure data rather than building a complete picture of *what* to collect for. Further, early collection efforts tend to prototype, making convenience sampling canon. In human-centric data (e.g. speech or movement), such requirements can encumber diversity across many axes. This has the potential to inhibit robust ML—while there is a natural iteration in datasets stemming from distribution shift (Quionero-Candela et al., 2009) and collectors' evolving understanding, it is a cycle full of forking paths (Kale et al., 2019; Hohman et al., 2020b) and dead-ends.

Model failures are invisible without participation and iteration (T3). Real world failures are only visible when communicated. But that communication comes from those invested enough in the tool, system, or research agenda to make the effort to bridge the communication gap. To quote an interviewee: *"The people who had issues were invisible to the system because they didn't like using it"*. Unlike other domains which employ tools to surface and track issues via user engagement, there are currently no tools that address the gap between a deployed ML product (let alone its early prototype) and a user. Such gaps are widened by language and knowledge barriers, and the products themselves are inaccessible to many who do not align with the priors on which the system is built. As these limitations have substantial downstream impacts, we sought to introduce early, comprehensive data & model checks to facilitate easy pivoting during data collection.

4. Designing Data

Existing bias mitigation and model evaluation approaches attempt to address the themes uncovered in our formative interviews but are not comprehensive to data collection *and* machine learning pipelines. In response to this disparity, we propose an iterative approach to data collection and ma-

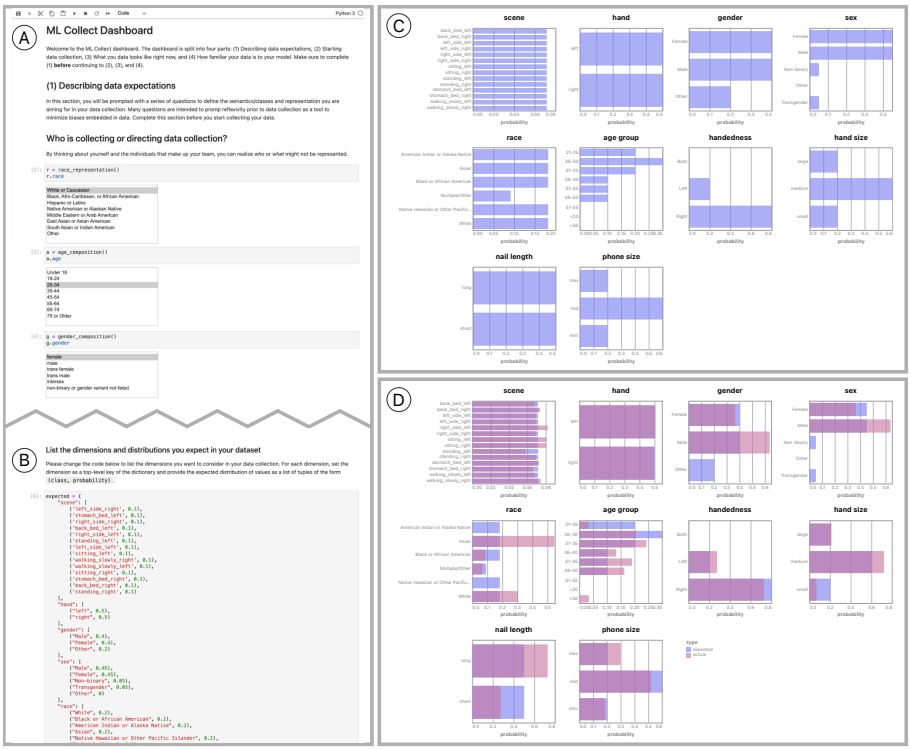

Figure 2. (A) Dashboard prompts practitioners through designing their datasets. (B) Before collection, the dashboard prompts documentation of expected distribution. (C) These are visualized as histograms. (D) As data is collected, true distributions (pink) overlay expected distributions (blue), highlighting divergent patterns.

chine learning that we call *designing data*. Designing data responds to our themes by introducing interventions before, during, and after collection & training. Shown in Figure 1, each step is intended to complement the others, compensating for their limitations to holistically improve the development and deployment process. We describe each below.

I: Pre-Collection Planning To facilitate broader consideration of critical facets of data, data design requires explicit documentation of *what* will be collected—including expected dimensions and distributions of data—*before* collection begins. Our documentation process ensures developers pay close attention to data diversity and coverage early in an ML pipeline and creates reference points for comparison when new information is uncovered. This process aligns with prior work in heuristics, implicit bias, cognition, and fake news susceptibility: deliberation—*reflection*—can correct intuitive mistakes, such as those made in data collection (Pennycook & Rand, 2019; Bago et al., 2020; Toplak et al., 2011; Saul, 2013). While it is difficult to enumerate all critical factors, this first step in our designing data process provides scaffolding, responding to our first theme (T1): *critical dimensions of data are hard to know a priori*.

As shown in Figure 2 (A), Pre-Collection Planning asks a series of questions to prompt reflexive consideration for individuals' personal biases through a series of open-text prompts, drop-down selections, and simple declarations. When the data relates to a human subject (i.e., images of faces, movement data, or voice recordings), our dashboard prompts self-reported demographic information about teams or individual users involved in developing the dataset. After, users are asked what data distributions they expect. This is an important step: prior work suggests that simply recognizing such information (Fish & Stark, 2021; Kim et al., 2017), and introducing design frictions (Gullström, 2012; Pennycook et al., 2020), may improve the quality of data work. An example of the dimensions and related distributions is shown in Figure 2. From these inputted dimensions, audits of population statistics, missing data, and undersampled subsets are presented to users. Different categories of the data, including demographics, metadata, task specifics, class representation, and intersectional categories are visualized.

II: Collection Monitoring By stating expected distributions prior to collection, auditing the data against those distributions is straightforward, allowing readjustments to be made quickly when necessary. Each view reflects the data's evolution, allowing real-time insight on where additional collection is needed. Our dashboard includes graphs highlighting distribution disparities that were perpetuated or introduced as the data was collected, as shown in Figure 2 (D). These charts benefit users by noting when the data collection process is either skewed (for example, through

convenience sampling) or when previously stated expectations did not align with reality. This step's iterative nature shortens the response time to correct fundamental errors in data, highlighting limitations that may have remained unnoticed, as described in our second and third themes: *difficulty of collection leads to compromises (T2)*, and *model failures are invisible without participation and iteration (T3).*

III: Data Familiarity After a model is trained, understanding what it has and has not learned appropriately is critical [1]. The value in this form of auditing is significant: some data may prove more difficult to learn due to their inherent complexity, or from not having enough similar examples. Despite the increased rigor of data collection as result of our earlier designing data steps, expected and actual data distributions might not match the learning needs for the model, thus requiring a stop gap such as under or oversampling, generating synthetic data, or continued data collection. We adopt density estimation (DE) to measure how familiar our model is with individual data points. While DE for OOD detection is well studied (Gawlikowski et al., 2021), our use of DE to direct data work (e.g. collection and annotation) is unique. By gaining insight on how a partially trained model perceives data, we can focus efforts on the most useful subsets, reweighting or replacing the data accordingly. In this way, even though *difficulty of collection leads to compromises (T2)*, we are able to interactively uncover remaining issues, as suggested by (*T3*).

Unfamiliar samples are "edge cases"—those that either are not represented appropriately within the dataset, are particularly challenging for the model to learn, or were erroneously collected (e.g., noisy). In early dataset development, familiarity scores act as useful checks for data with little signal—samples where we expect the model to perform well yet *may* be noisy and thus require human inspection . To measure the familiarity of the data, we incorporate density estimates of layer activations across a neural network (NN). We focus on the penultimate layer before the prediction softmax as it is the final feature representation used to make a prediction. Passing N inputs through the network produces an activation matrix $A(N) \in \mathbb{R}^{N \times M}$ for all M neurons in subset of selected layers $L' \subseteq L$. We learn a Gaussian Mixture model on these layer activations as given by the following:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i g(x \mid \mu_i, \Sigma_i) \quad (1)$$

where x is a matrix of layer activations, $w_i, i = 1, ..., M$ are the mixture weights, and $g(x|\mu_i, \Sigma_i), i = 1, ..., M$ are the component Gaussian densities [2]. For each sample in the current training set, we obtain the activations from layer l, then use PCA to reduce its dimensionality. This projection step serves two purposes: it reduces the dispersion of points that is typical in high dimensional spaces , and makes the remaining computation more tractable.

We then perform a Variational Bayesian estimation of a Gaussian mixture (Zobay, 2014) in the projected space. The fitted GMM allows us to give a familiarity score to each new sample. Given this sample, we extract the activation from the same layer l, again apply dimensionality reduction, then evaluate the log-likelihood provided by the fitted GMM—this is our familiarity score. If the sample falls into a densely populated area, its log-likelihood will be high: from the perspective of the features extracted by the current state of layer l, this sample appears as *familiar*. Conversely, if the sample falls into a scarcely populated area, its log-likelihood will be low and the sample less familiar. This measurement can be applied to new samples—those not seen by a model previously—or on training samples. Familiarity scores are presented in our dashboard through a series of graphs depicting their range and frequency, providing users direction for future data efforts; samples that are unfamiliar are reconciled with human expectations.

5. Task Selection, Data Collection, Modeling

We instantiated our designing data approach through a human activity recognition (HAR) task using *inertial measurement unit* (IMU) data— While designing data generally applies to all data collection and machine learning processes, our selection of data type and task were motivated by the unique challenges IMU data presents to building data diversity. First, IMU data inherently lacks the closeness of mapping (Blackwell et al., 2001) that image and audio data have to human models of the world, making it more difficult to audit. Recognizing when IMU data coverage is incomplete can be difficult when compared to image data as levels of abstraction often obfuscate fundamental problems in a dataset (Ramasamy Ramamurthy & Roy, 2018; Bartram et al., 2021)). Second, IMU data requires contextualization to create meaning—real-time labeling, or additional information from audio and video—and is harder to collect, unlike images and audio clips which are now ubiquitous (Ramasamy Ramamurthy & Roy, 2018). Yet IMU data still has the potential to bias ML models. Thus, we evaluate designing data's merit within a typically challenging context.

Our experimental task resembles work by Goel et al. (2013), who improved mobile text entry by categorizing different hand positions when users typed. Before collection, participants provided demographic information and metadata. This included race, ethnicity, gender, sex, age, hand length (mm), nail length (mm), hand dominance, phone version and

[1] The purpose of designing data is not to replace existing tools but rather to encourage a holistic approach that incorporates new ML techniques. For interventions *during training*, we refer to existing literature (Murphy, 2012; Japkowicz & Stephen, 2002).

[2] This mode of DE is interchangeable with other DE or OOD methods such as (Weglarczyk, 2018; Lee et al., 2018).

phone size. We collected data from 33 participants recruited over three separate periods in response to data disparities highlighted by our dashboard. In total, we collected > 3.88 million measurements from 1455 sessions. This data is used to populate dashboard, train and evaluate classifiers, then refined according to familiarity evaluations. We use 1D convolutional neural nets (CNNs) for classification [3].

6. Modeling Experiments

The ability to produce diverse subsets does not guarantee appropriate representation of sensitive attributes. Our approach to encouraging fair outcomes considers the amalgamation of both data and model. The following experiments adapt the following definition of fairness—rather than seek a high *average* accuracy across classes, we look at nuanced performance—accuracy, loss, and misclassification—between intersectional groups. To this end, we consider several questions as part of our designing data evaluation: (*Q1) Does auditing to increase data diversity improve model generalizability?* and *(Q2) Is data familiarity useful in auditing model & data?* We evaluate our interventions through a series of modeling experiments. First, we show that diverse data *does* lead to better performance. Then, we use familiarity to uncover noisy data within the dataset and describe how removing these samples impacts intersectional accuracy. Finally, we show that supplementing the dataset with unfamiliar samples improves model performance.

6.1. Diverse Data: Does auditing to increase data diversity improve model generalizability?

We sought to answer Q1 through our first set of experiments. We compare "diverse" models to "less diverse" models. We do so for two reasons: first, it may not be clear to practitioners that collecting diverse data early in development is critical to building functional tools. Second, despite best efforts to curate a list of meaningful characteristics, we did not know if these additional data dimensions had any true effect on the classification task.

Both diverse and less diverse models are trained using the same number of training samples and are evaluated on the same test data. Less diverse models are trained on data where one group (e.g. small handed) was left out. In this way, we perform leave-one-out cross-validation *and* consider the specific effects of a given demographic group. All models were trained with the same sample size. We then generate predictions from the original test set. Paying close attention to intersectional groups, we expect to see more performance stratification in less diverse models. We compare models across overall, group-specific, and intersectional accuracies. We hypothesize that some categories are more meaningful to performance than others. Performance disparity—such as lower accuracy—across categories *despite* equal sample size would support this hypothesis.

Evaluation When compared to models trained on less diverse data, we found that models with diverse data had an overall higher accuracy *and* performed better across intersectional groups, as shown in Figure 3. For example, Figure 3 B shows a significant dip in performance for the model trained without data from the **right-side-left** condition compared to diverse model (Figure 3 C). This is as we anticipated, and we see that for data where participants were typing in the **right-side-left** condition, the less-diverse model actually performs *worse* than random chance. This pattern is repeatable—across k models where we intentionally left out one group (Figure 3 A), we see a correlated diagonal of lighter color indicating lower testing accuracy, supporting our hypothesis that the extensive characteristics we collected data for *do* effect IMU performance. In contrast, diverse models show less performance variance, instead performing better across different demographics—matching what we would hope to achieve to minimize worst group generalization (Sagawa et al., 2019). We found some intersectional subgroups performed drastically different compared to their overall group performance. In typical evaluations, this nuance is often obfuscated by aggregation, yet we were able to capture it using simple interventions and visualizations.

6.2. Familiarity: Is data familiarity useful in auditing model & data?

There are two scenarios where familiarity is useful in directing dataset iteration: to facilitate data cleaning, and to encourage appropriate representation of diverse data for a given model. When a dataset contains noisy data, familiarity can help surface these samples for human evaluation. If a dataset has already been cleaned, then familiarity is used to highlight samples that less familiar. We evaluate familiarity's efficacy in directing dataset iteration (Q2) through a series of experiments to improve performance on the same testing data used in subsection 6.2, modifying training data but maintaining sample size parity. Modifications to the training data are completed using "self-familiarity" scores (i.e., familiarity of training data) to avoid overfitting.

6.2.1 Familiarity for Debugging Our dataset was collected "in the wild" without additional annotation from participants. While participants were given clear instructions, we anticipated that a small number of samples would show significant noise or distortion—participants might drop their phones as they type, or switch hands part way through a session. Such samples may introduce unwanted effects downstream if left uncaught. Given the size of the dataset, inspecting time series plots for each session was untenable. Instead, incorporating familiarity may *greatly* reduce the number of samples to evaluate. By using an least partially

[3] See appendix for details on data processing, modeling, and iterating with Pre-Collection Planning & Monitoring

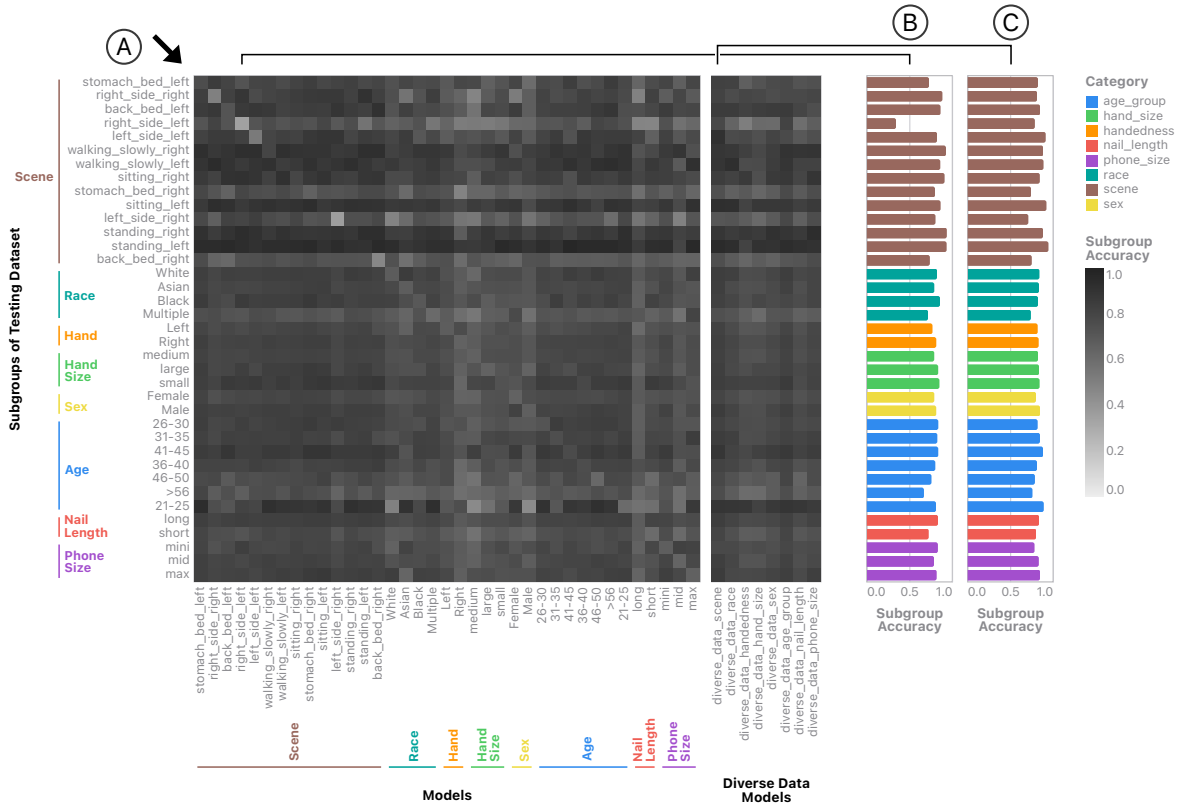

Figure 3. Performance per subgroup on testing data (y-axis) per model (x-axis), split across categories. The large square matrix (left) shows accuracy where a subgroup was left out of the training set. The rectangular matrix (right) shows performance of models trained on diverse data. (A) The arrow highlights the diagonal of the matrix: subgroups of data that perform worse than others, corresponding to the model trained without this particular subgroup. (B) E.g., taking the `left_side_right` model from the matrix, we see it performs poorly on the `left_side_right` subgroup. (C) Models trained on same sized yet more diverse dataset show no dip in accuracy.

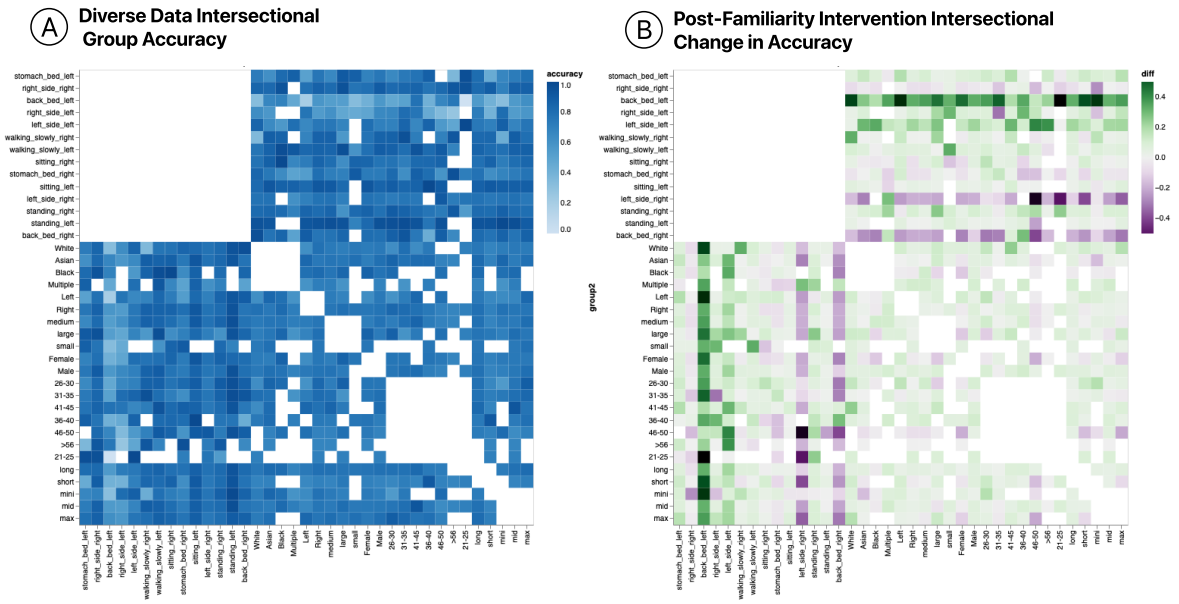

Figure 4. (A) Accuracy for intersectional subgroups in a diverse data model. (B) Difference in accuracy for subgroups before & after familiarity interventions (green is positive, purple negative).

trained model on the dataset to uncover unfamiliar samples, familiarity offers a possible solution to capture noisy data, presenting an alternative to work such as (Pleiss et al., 2020; Shen & Sanghavi, 2019), which incorporate loss as a metric for capturing noisy samples. We hypothesize that least familiar samples within a noisy dataset will include instances of noisy data posing the greatest harm to the model.

We explored how to incorporate familiarity as a tool for debugging first through an automated approach to removing data, then by incorporating human review. Our protocol is as follows: first, we train an initial model on all available training data. We apply self-familiarity to the training and testing set, selecting only 0.1% of the data corresponding to least familiar samples. This data is either removed from the dataset, or visualized and manually reviewed per sample to evaluate if they are truly noisy or simply uncommon. We removed the same number of samples with manual review of truly noisy data as with the automated removal [4]. We compared outcomes of both automated removal and removal through human review, but found that deleting a percentage of least familiar data removed both noisy data and important outliers. We evaluate this outcome by comparing the results of the following experiment to improve data diversity across each approach to data cleaning when compared to a baseline where no data is removed.

Evaluation In practice, we found that familiarity worked well as a tool for debugging. Before removing noisy data from the complete dataset, a large percentage of the least familiar samples showed significant distortion, despite our efforts to normalize the data. Because of the presence of these noisy data, running our familiarity experiments with data that was not cleaned did not show the same levels of general improvement. In this case, matching metadata to noisy samples was not the correct comparison—these samples were not exemplars of the subpopulation.

In manually evaluating our unfamiliar data before removal, we uncovered cases where unfamiliar samples were not noisy but rather underrepresented intersectional groups. For instance, a person identifying as an **Asian Female** with **Small** hands and **Large** phone was as unfamiliar to the model as incredibly noisy examples. Noisy data has different implications for the model than unusual samples. For this reason, removing 0.1% of the least familiar data prior to experiments did not improve performance to the extent seen when manual review was implemented. That is, familiarity cannot distinguish between distorted noise and underrepresented or OOD data. For datasets where significant noise is present, *human review* is necessary to evaluate data quality.

6.2.2 Familiarity for Diverse Data Coverage Next, we compare familiarity scores across the different descriptive groups, as described in Section 6.1. These scores are used determine next steps for additional data collection, augmentation, or modified (over/under) sampling. Here, we sample *out* a percentage of the *most* familiar data, and *add* data matching the intersectional characteristics for the same percentage of *least* familiar data. Substituted data was held out from training in Section 6.1, thus is new to the models. We replace samples based on metadata characteristics, using combinatorial optimization to best match unfamiliar data. Using PCA, we project down to 50 dimensions, then fit 5 GMMs to last dense layer for each 1D CNN trained in Section 6.1. Scoring our training data, we save familiarity scores and model weights. We vary the range of familiarity scores to sample from, percentage of data, and two sampling in methods—top k and random selection from a least familiar data range—compared to a random baseline. We structure these experiments—varying window size and sampling percentage—to uncover a sweet spot: removing too much data may harm performance on familiar groups, and too large a window might impinge on less familiar data.

Evaluation Self-familiarity scores create a distinct curve, with unfamiliar data falling into the long-tail. We find accuracy scores are far more striated prior to familiarity interventions, showing some concepts are learned better than others. Following familiarity interventions, models do show targeted improvement. Models performed more poorly in regions with high numbers of low familiarity data (an example of which is shown in Figure 4). Of note, models did not necessarily improve overall—although this was frequently the case— instead showing improvement in areas of low performance and regression in those with high performance. In Figure 4, we show model performances on intersectional groups: (A) is a model trained on the **diverse-data-scene** dataset with no familiarity intervention. We can see lower accuracies in **back-bed-left** and **right-side-left** compared to other subgroups. In contrast, **sitting-left** and **back-bed-right** had much higher accuracy compared to other subgroups. Figure 4 (B) shows the difference between this model and one trained on the same data with familiarity interventions. Regions with relative poor performance improved dramatically, while those with higher accuracies showed some regression. Incorporating 0.1% least familiar data was optimal for our experiments. Overall, familiarity consistently captured under-represented samples.

7. Conclusion

We need processes that integrate data *and* models in systematic, transparent ways. While each step of *designing data* can be incorporated in isolation, the interventions are complimentary, compensating for the cascading effects of upstream and downstream missteps. Data is shaped by the perspective of the observer; our work highlights how systematic processes may curtail bias early in development.

[4]Experimental protocol can be found in the Appendix.

Acknowledgements

We thank our colleagues at Apple for their time and effort participating in our research. We especially thank Donghao Ren and Halden Lin for their help building the iOS application and Kayur Patel for his guidance. Aspen Hopkins was partially supported by the Siebel Fellowship.

References

Facets. *Google PAIR*, 2017. URL https://pair-code.github.io/facets/.

Aisch, G., Cox, A., and Quealy, K. You draw it: How family income predicts children's college chances. *The New York Times*, 28, 2015.

Amershi, S., Chickering, M., Drucker, S. M., Lee, B., Simard, P., and Suh, J. Modeltracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 337–346, 2015.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2019.

Anari, N., Gharan, S. O., and Rezaei, A. Monte carlo markov chain algorithms for sampling strongly rayleigh distributions and determinantal point processes. In *Conference on Learning Theory*, pp. 103–115, 2016.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. *ProPublica*, 23, 2016.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., Nair, R., Ramamurthy, K. N., Olteanu, A., Piorkowski, D., et al. Factsheets: Increasing trust in ai services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5):6–1, 2019.

Asudeh, A., Jin, Z., and Jagadish, H. Assessing and remedying coverage for a given dataset. In *2019 IEEE 35th International Conference on Data Engineering*, pp. 554–565. IEEE, 2019.

Bago, B., Rand, D. G., and Pennycook, G. Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*, 2020.

Bartram, L., Correll, M., and Tory, M. Untidy data: The unreasonable effectiveness of tables. *arXiv preprint arXiv:2106.15005*, 2021.

Ben-David, S., Hrubeš, P., Moran, S., Shpilka, A., and Yehudayoff, A. Learnability can be undecidable. *Nature Machine Intelligence*, 1(1):44, 2019.

Bender, E. M. and Friedman, B. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

Blackwell, A. F., Britton, C., Cox, A., Green, T. R., Gurr, C., Kadoda, G., Kutar, M., Loomes, M., Nehaniv, C. L., Petre, M., et al. Cognitive dimensions of notations: Design tools for cognitive technology. In *International Conference on Cognitive Technology*, pp. 325–341. Springer, 2001.

Buchanan, L., Park, H., and Pearce, A. You draw it: What got better or worse during obama's presidency. *The New York Times*, 15, 2017.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *ACM Conference on Fairness, Accountability and Transparency*, pp. 77–91, 2018.

Celis, L. E., Deshpande, A., Kathuria, T., and Vishnoi, N. K. How to be fair and diverse? *arXiv preprint arXiv:1610.07183*, 2016.

Celis, L. E., Keswani, V., Straszak, D., Deshpande, A., Kathuria, T., and Vishnoi, N. K. Fair and diverse dpp-based data summarization. *arXiv preprint arXiv:1802.04023*, 2018.

Chao, A., Chiu, C.-H., and Jost, L. Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through hill numbers. *Annual Review of Ecology, Evolution, and Systematics*, 45:297–324, 2014.

Corbett-Davies, S. and Goel, S. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Cruciani, F., Vafeiadis, A., Nugent, C., Cleland, I., McCullagh, P., Votis, K., Giakoumis, D., Tzovaras, D., Chen, L., and Hamzaoui, R. Feature learning for human activity recognition using convolutional neural networks. *CCF Transactions on Pervasive Computing and Interaction*, 2 (1):18–32, 2020.

Dodgson, J. E. Reflexivity in qualitative research. *Journal of Human Lactation*, 35(2):220–222, 2019.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pp. 2791–2801, 2018.

Drosou, M., Jagadish, H., Pitoura, E., and Stoyanovich, J. Diversity in big data: A review. *Big Data*, 5(2):73–84, 2017.

D'souza, D., Nussbaum, Z., Agarwal, C., and Hooker, S. A tale of two long tails. *arXiv preprint arXiv:2107.13098*, 2021.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012.

Fish, B. and Stark, L. Reflexive design for fairness and other human values in formal models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 89–99, 2021.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 329–338, 2019.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., and Crawford, K. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Goel, M., Jansen, A., Mandel, T., Patel, S. N., and Wobbrock, J. O. Contexttype: Using hand posture information to improve mobile touch screen text entry. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2795–2798, 2013.

Grout, J., Frederic, J., Corlay, S., and Ragan-Kelley, M. Ipywidgets jupyter widgets, 2019. URL https://github.com/jupyter-widgets/ipywidgets.

Gullström, C. Design frictions. *AI & Society*, 27(1):91–110, 2012.

Görtler, J., Hohman, F., Moritz, D., Wongsuphasawat, K., Ren, D., Nair, R., Kirchner, M., and Patel, K. Neo: Generalizing confusion matrix visualization to hierarchical and multi-output labels. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2022. doi: 10.1145/3491102.3501823.

Hao, K. and Stray, J. Can you make ai fairer than a judge? play our courtroom algorithm game. *MIT Technology Review*, 2019.

Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

Hohman, F., Kahng, M., Pienta, R., and Chau, D. H. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2018. doi: 10.1109/TVCG.2018.2843369. URL https://fredhohman.com/visual-analytics-in-deep-learning/.

Hohman, F., Conlen, M., Heer, J., and Chau, D. H. P. Communicating with interactive articles. *Distill*, 5(9):e28, 2020a.

Hohman, F., Wongsuphasawat, K., Kery, M. B., and Patel, K. Understanding and visualizing data iteration in machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2020b.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., and Wallach, H. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2019.

Hooker, S. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4):100241, 2021.

Hopkins, A. and Booth, S. Machine learning practices outside big tech: How resource constraints challenge responsible development. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pp. 134–145, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462527. URL https://doi.org/10.1145/3461702.3462527.

Inc., G. Know your data, 2021. URL https://knowyourdata.withgoogle.com/.

Japkowicz, N. and Stephen, S. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449, 2002.

Kale, A., Kay, M., and Hullman, J. Decision-making under uncertainty in research synthesis: Designing for the garden of forking paths. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2019.

Katz, J. You draw it: Just how bad is the drug overdose epidemic. *The New York Times*, 2017.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Kim, H. K., Kim, N., and Park, J. Relationship analysis between body flexion angles and smartphone tilt during smartphone use. *International Journal of Industrial Ergonomics*, 80:103034, 2020.

Kim, Y.-S., Reinecke, K., and Hullman, J. Explaining the gap: Visualizing one's predictions improves recall and comprehension of data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1375–1386, 2017.

Kim, Y.-S., Reinecke, K., and Hullman, J. Data through others' eyes: The impact of visualizing others' expectations on visualization interpretation. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):760–769, 2018. doi: 10.1109/TVCG.2017.2745240.

Klawonn, M., Heim, E., and Hendler, J. Exploiting class learnability in noisy data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4082–4089, 2019.

Koesten, L., Kacprzak, E., Tennison, J., and Simperl, E. Collaborative practices with structured data: Do tools support what users need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2019.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Lee, S., Lee, H., Lee, J., Ryu, H., Kim, I. Y., and Kim, J. Clip-on imu system for assessing age-related changes in hand functions. *Sensors*, 20(21):6313, 2020.

MacKenzie, I. S. and Soukoreff, R. W. Phrase sets for evaluating text entry techniques. In *CHI'03 Extended Abstracts on Human Factors in Computing Systems*, pp. 754–755, 2003.

Milani, M., Huang, Y., and Chiang, F. Diversifying anonymized data with diversity constraints. *arXiv preprint arXiv:2007.09141*, 2020.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, 2019.

Mitchell, M., Baker, D., Moorosi, N., Denton, E., Hutchinson, B., Hanna, A., Gebru, T., and Morgenstern, J. Diversity and inclusion metrics in subset selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 117–123, 2020.

Mouchet, M. A., Villéger, S., Mason, N. W., and Mouillot, D. Functional diversity measures: An overview of their redundancy and their ability to discriminate community assembly rules. *Functional Ecology*, 24(4):867–876, 2010.

Murphy, K. P. *Machine learning: A probabilistic perspective*. MIT press, 2012.

Noble, S. U. *Algorithms of oppression: How search engines reinforce racism*. NYU Press, 2018.

Noroozi, F., Kaminska, D., Corneanu, C., Sapinski, T., Escalera, S., and Anbarjafari, G. Survey on emotional body gesture recognition. *IEEE Transactions on Affective Computing*, 2018.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., and Mordvintsev, A. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.

Palanica, A., Thommandram, A., Lee, A., Li, M., and Fossat, Y. Do you understand the words that are comin outta my mouth? voice assistant comprehension of medication names. *NPJ Digital Medicine*, 2(1):1–6, 2019.

Papadopoulos, A., Kyritsis, K., Klingelhoefer, L., Bostanjopoulou, S., Chaudhuri, K. R., and Delopoulos, A. Detecting parkinsonian tremor from imu data collected in-the-wild using deep multiple-instance learning. *IEEE Journal of Biomedical and Health Informatics*, 24(9): 2559–2569, 2020. doi: 10.1109/JBHI.2019.2961748.

Pennycook, G. and Rand, D. G. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188: 39–50, 2019.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., and Rand, D. G. Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, 31(7):770–780, 2020.

Pleiss, G., Zhang, T., Elenberg, E. R., and Weinberger, K. Q. Identifying mislabeled data using the area under the margin ranking. *arXiv preprint arXiv:2001.10528*, 2020.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. The MIT Press, 2009.

Ramasamy Ramamurthy, S. and Roy, N. Recent trends in machine learning for human activity recognition—a survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1254, 2018.

Rosenfeld, E., Ravikumar, P., and Risteski, A. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Samadi, S., Tantipongpipat, U., Morgenstern, J. H., Singh, M., and Vempala, S. The price of fair pca: One extra dimension. In *Advances in Neural Information Processing Systems*, pp. 10976–10987, 2018.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. K., and Aroyo, L. M. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. 2021.

Saul, J. Scepticism and implicit bias. *Disputatio*, 5(37): 243–263, 2013.

Schapire, R. E. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pp. 5014–5026, 2018.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., and Dennison, D. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems*, pp. 2503–2511, 2015.

Shen, Y. and Sanghavi, S. Learning with bad training data via iterative trimmed loss minimization. In *International Conference on Machine Learning*, pp. 5739–5748. PMLR, 2019.

Soedirgo, J. and Glas, A. Toward active reflexivity: Positionality and practice in the production of knowledge. *PS: Political Science & Politics*, 53(3):527–531, 2020.

Stray, J. and Hao, K. Interactive visualization of fairness tradeoffs. *Computation + Journalism Symposium*, 2020.

Tayi, G. K. and Ballou, D. P. Examining data quality. *Communications of the ACM*, 41(2):54–57, 1998.

Toplak, M. E., West, R. F., and Stanovich, K. E. The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, 39(7): 1275, 2011.

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., and Steyerberg, E. W. Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):1–7, 2019.

VanderPlas, J., Granger, B., Heer, J., Moritz, D., Wongsuphasawat, K., Satyanarayan, A., Lees, E., Timofeev, I., Welsh, B., and Sievert, S. Altair: Interactive statistical visualizations for python. *Journal of Open Source Software*, 3(32):1057, 2018. doi: 10.21105/joss.01057. URL https://doi.org/10.21105/joss.01057.

Weglarczyk, S. Kernel density estimation and its application. In *ITM Web of Conferences*, volume 23, pp. 00037. EDP Sciences, 2018.

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viegas, F., and Wilson, J. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE TVCG*, 26, 2019. doi: 10.1109/tvcg.2019.2934619.

Whittaker, R. H. Evolution and measurement of species diversity. *Taxon*, 21(2-3):213–251, 1972.

Wieringa, M. What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 1–18, 2020.

Yang, K., Qinami, K., Fei-Fei, L., Deng, J., and Russakovsky, O. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 547–558, 2020.

Zheng, W., Wang, X., Fang, H., and Cheng, H. Coverage-based search result diversification. *Information Retrieval*, 15(5):433–457, 2012.

Zobay, O. Variational Bayesian inference with Gaussian-mixture approximations. *Electronic Journal of Statistics*, 8(1):355 – 389, 2014.

A. Data Collection

We built a custom iOS data collection app using the Swift programming language with public frameworks including SwiftUI, CommonCrypto, and CoreMotion. The app collects right and left-handed texting data across different contexts people may type in, such as texting while walking or laying down. Example views of the app are shown in Figure 5. All data were collected with informed consent. During collection, participants were given instructed how to hand posture when holding the phone. While this approach produces consistent data for training purposes, this limited

the inherent variability within the data. We simplified the task to binary handedness (typing with the left or right hand), but introduced a source of natural variability by prompting users to perform a series of actions in parallel while typing. This contrasts (Goel et al., 2013), who collected data while users sat in a lab environment. We introduced 6 typing scenarios (walking, sitting, standing, lying on your back, lying on your side, and lying on your front) to represent a selection of possible contexts users might normally experience typing. Participants were then asked to type an English phrase from MacKenzie and Soukoreff's phrase set (MacKenzie & Soukoreff, 2003) using their left or right hand for a total of 6 (positions) $\times$ 4 (typing sessions) $\times$ 2 (hands) $=$ 72 trials. Entered phrases were recorded as a trial. Participants were automatically redirected to the iOS *Measure* app to determine hand and nail length. The app disabled keyboard autocomplete and autocorrect. Beginning when people pressed "start", we collected data was collected at a sampling rate of 200 Hz. When the task was finished, metadata, IMU data, and trial metadata (including scenario, phrase, keyboard recording, and session time) were pushed to an encrypted database.

B. Dashboard & Questions

We built our dashboard in JupyterLab using Altair (VanderPlas et al., 2018), a declarative visualization library for Python, and Ipywidgets (Grout et al., 2019). As shown in Main Body Figure 2 (A), the dashboard asks a series of questions to prompt reflexive consideration for the team or individuals personal biases. Using drop-down selections, they are asked to describe their team's representational makeup—including race, accessibility needs, age, sex and gender identities. After filling out this information, users are told: *"The following groups and their subsequent intersections are not participating in your project development. To ensure an optimal result, take steps to consider how their experiences and views might differ from the currently represented ones."* This notice is followed by a list of demographic information not identified by the users. This is an important step: prior work has shown that simply recognizing such information (Fish & Stark, 2021; Kim et al., 2017), and introducing design frictions (Gullström, 2012; Pennycook et al., 2020), can increase the quality and consideration that goes into data work.

One limitation to this procedural approach is how scoped our questions are—they are not all-encompassing, but intended to start the process of reflection and inquiry early on. As a final step in this reflexive process, and to minimize this limitation, a series of open-ended questions that take in free-form text asks, "What's missing, in the context of your project?" followed by some examples that expand on axes of diversity teams might need to consider. Users are then asked to enter expected dimensions and distributions of data before collecting data (Main Body Figure 2 B,C). When distributions have incorrect values (i.e., do not add up to 100%), the dashboard normalizes. This active expression of expected data encourages users to acknowledge and document the specific limitations of their data, setting a precedent of conscious decision making from the beginning. It creates a simple provenance for early assumptions and a baseline to evaluate against during data collection.

From these inputted dimensions, audits of population statistics, missing data, and undersampled subsets are presented to users, reflecting *During Collection*. New dimensions can be added as needed. Different categories of the data, including demographic information and metadata (described in the participant subsection), task specifics, and class representation are then shown in visualizations Figure 2. Similarly, intersectional categories (such as age *and* hand size) are shown. This view reflects the data evolution as more data is collected, allowing real-time insight into what additional collection might be required. Following collection, the dashboard allows users to use a pre-trained model or train their own. Following training, saved states of a neural network and model architecture are loaded into our familiarity functions. Data is inputted to build out familiarity scores, the final step in our designing data process.

B.1. Visual Examination of Familiarity

Log-likelihood is a relative measure. Given our interest in the spread of familiarity scores across different categories within the dataset, we introduced the visualizations shown in Figure 8 and Figure 9 into the dashboard for quick insight into if there are particular subsets that are seem less familiar than others. An overview of the familiarity scores can be seen in Figure 6 For example, with a close look we see that the *least* familiar samples across genders are consistently *female*. Similarly, we see our oldest age-range (¿56) is least familiar for age. This view can be compared across iterations of data or model development for a gestalt of the changes in familiarity.

C. Case Study: Reflecting on Our Data Collection Process

It was unclear how diverse IMU data would influence our modeling experiments, or if the categories developed through our reflexive prompts would meaningfully align with variation within the data. Prior work on human gestures has shown age (Lee et al., 2020), emotion (Noroozi et al., 2018), and health (Papadopoulos et al., 2020) influence gesture presentation. The tilt of a smartphone during texting is captured by IMU data, and can be associated to back posture (Kim et al., 2020). Hand position during typing is similarly distinguishable (Goel et al., 2013), yet IMU

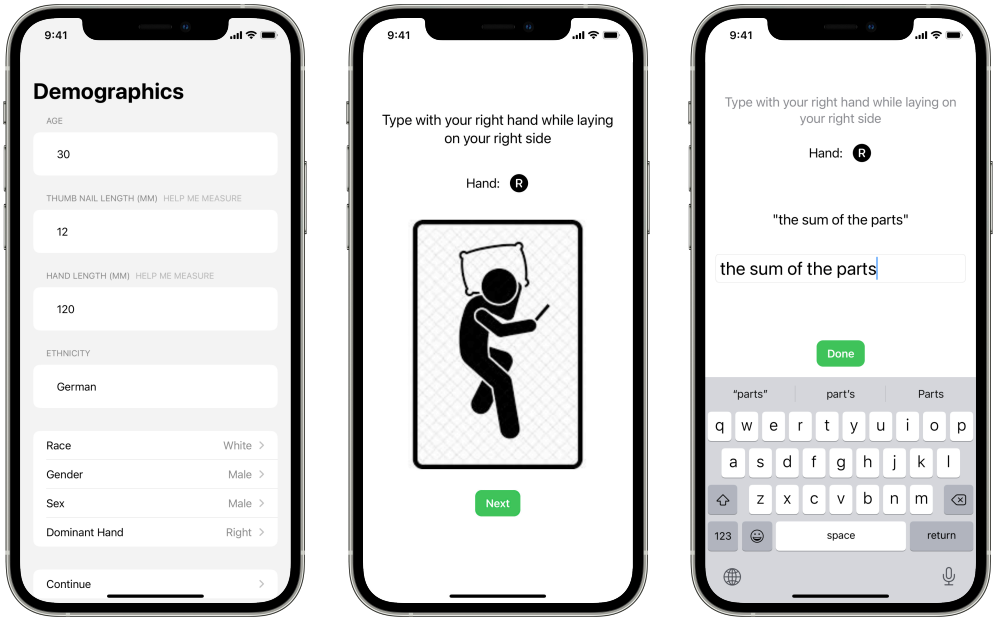

Figure 5. Our IMU Data Collection App showing the primary screens of the study. Participants entered their demographic information such as hand dominance, hand and nail length, sex and gender, and age (left). The app instructs a participant to type on their phone using either their right or left hand in a physical configuration (middle). Participants then type the presented phrase (right).

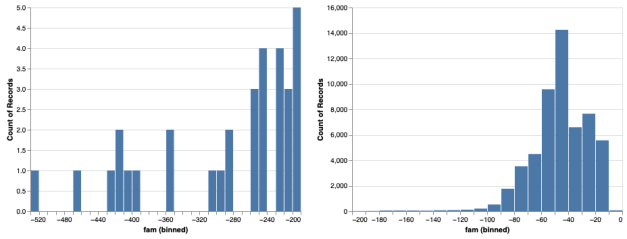

Figure 6. Overview of familiarity over the two tail ends of the distribution (most familiar on the right, least familiar on the left).

data collection rarely includes meta information about participants. Similarly, context is often not documented for image data (e.g. the proverbial question of *what's outside the frame?*).

We explored how these and other common demographic categories influenced model performance and show how consideration for diverse data should be emphasized *prior to* deployment, not just after deployment. We incorporated our dashboard's *Pre-Collection* suite of prompts in our own IMU data collection to determine what characteristics to collect for. It was through these prompts that we realized a need to measure additional information as typical demographics did not capture how people held their phones. We noted that `Phone Size`, `Hand Size`, `Handedness`, and `Nail Length` (particularly in the case of acrylic nails) may play a role in how people text, despite not being variables typically considered in such tasks. It was also through this process of reflecting on impactful features that we realized the advantages of asking participants to act out various behaviors during their typing tasks. This meaningfully shaped our task. Other measurements that were noted during this process would have added substantial complexity to our collection procedure. Most prominent of these was hand strength and dexterity. These features are impactful to how individuals type—a person with carpel tunnel or arthritis will type differently compared to someone without these conditions—but required additional tooling to accurately capture. Instead, we noted strength and dexterity for future evaluation, to be completed prior to public deployment.

The results of *Collection Monitoring* led to three instances of additional, retargeted data collection efforts based on unexpected skew. During our initial data collection, there were multiple categories which did not match our previously described distributions. We call out several in Figure 10 to demonstrate the outcomes of our data collection re-targeting. While all participants were employees from a large tech company, we did not anticipate our initial wave of data collection to be so skewed towards individuals between the ages of 21 and 40. In truth, we initially had no participants over 38. As a result, we emphasized diversity in participant age moving forward with some success—each consecutive wave improved the data coverage. Familiarity also directed

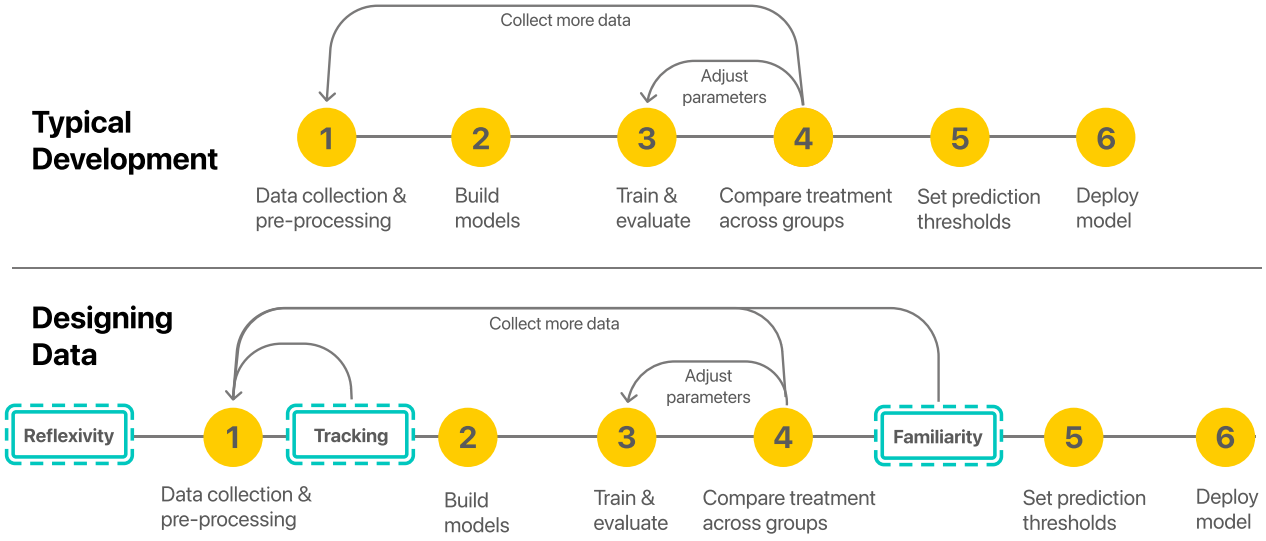

Figure 7. Designing data process compared to conventional machine learning development. *Reflexivity* ensures appropriate consideration of positionality and expectations is given before prior to collection. *Monitoring* provides insight into unexpected trends during collection. *Familiarity* facilitates debugging and highlights potentially noisy or underrepresented subpopulations to direct iteration. This figure represents a simplification of the data collection process. The results of *Familiarity* can also be incorporated into training (after cleaning the dataset) and tracking (after training an initial model).

our collection efforts, leading us to seek additional data for an intersectional subset of `Female` participants with `Large` phones and `Small` hands.

C.1. Collection Retargeting

Similarly, we noted that the majority of our participants self-described as White, but that there were no Black or Indigenous participants whatsoever. Despite our best attempts, this was only partially amended; in optimizing across intersectional groups, we were not able to perfectly match our expected or updated distributions (at least within the context of this set of iterations). In contrast, our initially collected Handedness distribution perfectly matched our expectations. This prompted a discussion of whether this distribution was reasonable—despite matching the ratio of left versus right handedness in US populations, we believed that right and left handed individuals would type in dramatically different ways, thus may require sample size parity to be appropriately learned by the model. Examples of our retargeting efforts can be seen in Figure 10.

D. Modeling

Using Keras, our architecture included two 1D convolutional layers (standard for time series data), max-pooling layers, a dropout layer, a fully connected dense layer with ReLu activations, and a fully connected dense layer with softmax activations. We followed the generic *Activity Recognition Chain*, which includes pre-processing, segmentation, feature extraction, and classification steps (Cruciani et al., 2020) for our approach to modeling. We used 1D convolutional neural nets (CNNs) for sequence classification. The CNN's architecture performs feature extraction through the convolution of the input signal with a kernel (or filter). To pre-process data, each session was segmented into $200ms$ windows, with $40ms$ overlap between segments. Session timing varied by how long it took participants to finish typing a given phrase. We corrected our IMU dataset to account for gravitational acceleration effects, then normalized (using a direct current blocker) and segmented IMU data in series to ensure all windows were of equal length. We discarded windows containing data from multiple sessions. For training, one sample equated to a window of the time series data. Training batch sizes were 256 (batch) $\times$ 200 (ms) $\times$ 3 (IMU), where ms is the window of time, $batch$ is the batch size, and IMU represents the three accelerometer data sources. All computations were run on NVIDIA V100 GPUs. Training all models took approximately 30 hours.

E. Experiments

E.1. Diverse Data Experimental Protocol

We save every model, their weights, training accuracy and loss curves, and training dataset, keeping track of model version. Note that each model was trained 10 x 10 times, keeping the random seed stable for r times, then repeating

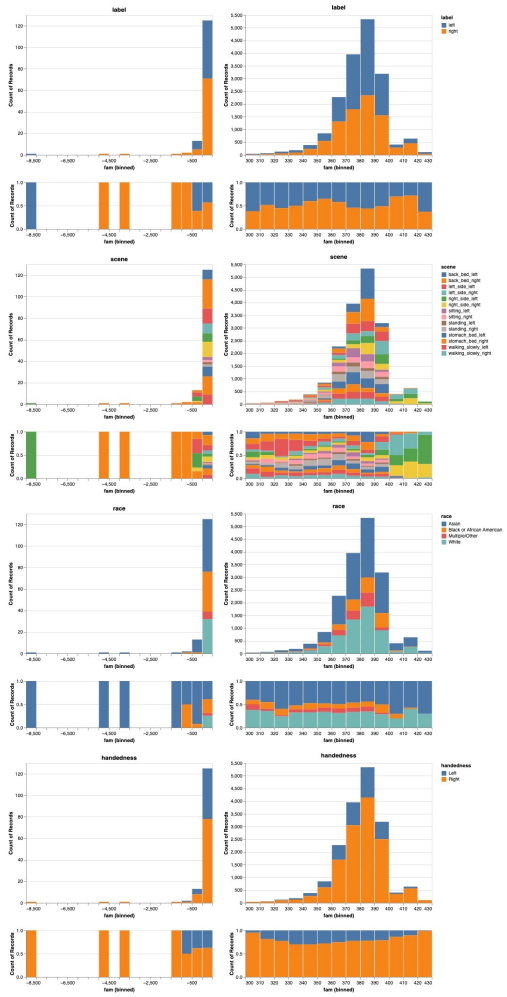

Figure 8. Visual presenting familiarity distributions across different categories (part 1).

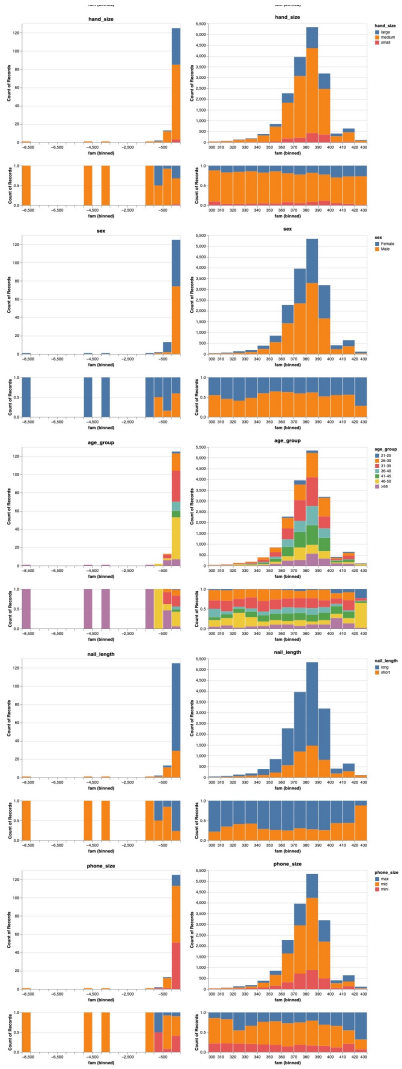

Figure 9. Visual presenting familiarity distributions across different categories (part 2).

Figure 10. Selected views of data distributions across three different waves of data collection. Each new iteration of data collection was the result of evolving understanding about the data. For Age, initially there were no participants over the age of 38. $Handedness$ initially met expected distributions, however these did not match downstream needs. In $Race$, there was substantial skew for White participants. An error in data collection led to a skew in which scenes were presented to participants.

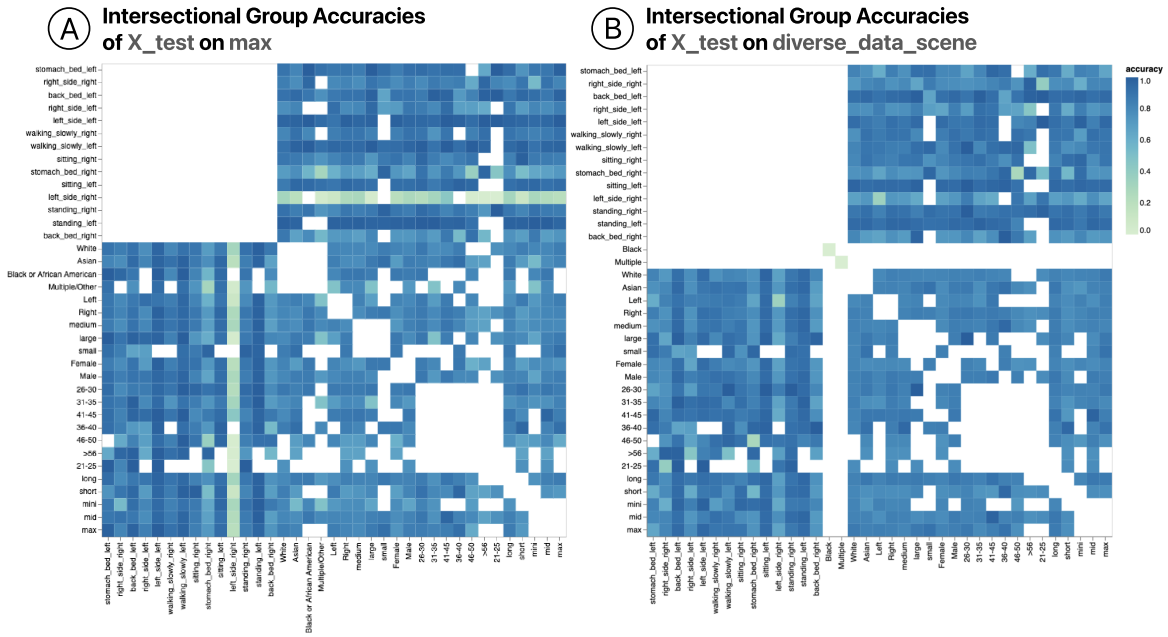

Figure 11. Comparison of intersectional groups of a less diverse model (A) to a diverse model (B). Striated accuracy across populations—as described by the metadata descriptions—performed worse when groups were left out, indicating that these characteristics were aligned with meaningful diversity in the data.

with a different seed. Weights from the models with highest accuracy across the trials were kept for later experiments.

First, shuffle then randomly select train and test datasets such that no typing trials are split across datasets. We save the test dataset to evaluate every model trained with the current train/test split. Then, we compare the full dataset and train/test distributions using a visual check and earth mover's distance (EMD). If the difference in distributions are significant, repeat the first step. We group the data by category (e.g., Sex) then by type (e.g., $Female$). Next, randomly downsample each group such that each subset is of equal size. We compare the distribution of data sampled out to the downsampled group data, repeating sampling in the case of skew, then save the sampled out data. Lastly, we repeat the prior step but do so from the complete training set: this is the "diverse" dataset for the category. For each group within a category, we append all other groups together and then leave the current group out such that new training groups were $Female \bigcup Intersex$ (for example). For each set within a category, train a new model using our previously described 1D CNN. We monitor for overfitting by setting early stopping based on loss with a patience of 4.

Finally, the influence of demographics on performance varied. In general, removing activity and age group subsets was more harmful to models than hand size, gender, or race subsets, but there were exceptions.

E.2. Familiarity Experiments: Capturing Noisy Data

Different layers of an NN capture distinct features of the input data (Olah et al., 2018). Familiarity scores can therefore be extracted from any layer. Earlier layers capture fundamental structure found in the input data, while deeper layers capture semantic content. For this reason, we focus on the final dense layer, which holds the closest semantic alignment to human perception. A visual overview of the familiarity implementation can be seen in Figure 13.

In general, any DE technique can replace our implementation. Note that the same data shown to a different model is likely to obtain a different familiarity score; each sample is tightly coupled to how a specific model perceives it. This analysis cannot be done on the dataset without the guidance of an at least partially trained model.

A comparison of noisy data to OOD and more typical samples is shown in Figure 14.

E.3. Diverse Data With Familiarity

Given our intention was to *encourage* diversity in our dataset, an ineffective sampling strategy might exacerbate edge case failures. We then select a range from the data distribution to sample the most familiar data from. Thus we explored three general approaches:

Figure 12. The performance of every model on the same test set, split across categories. For each category, a model's label indicates which subgroup was held out from its training set. To ensure fair comparison, within each category each model was trained on the same number of instances. Notice for each category that the "diverse" model (highlighted with a darker color), i.e., the model with no subgroup held out, almost exclusively performs the best, despite having the same number of data instances as the other models.

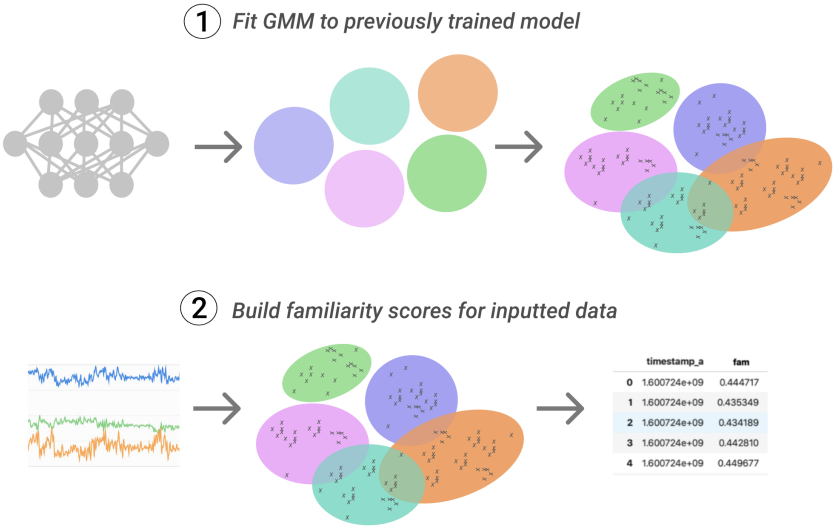

Figure 13. Overview of familiarity implementation. (1) We fit a Gaussian Mixture Model to our model of interest. (2) We present samples (e.g. our training data for "self-familiarity") to the GMM, returning log-likelihood scores.

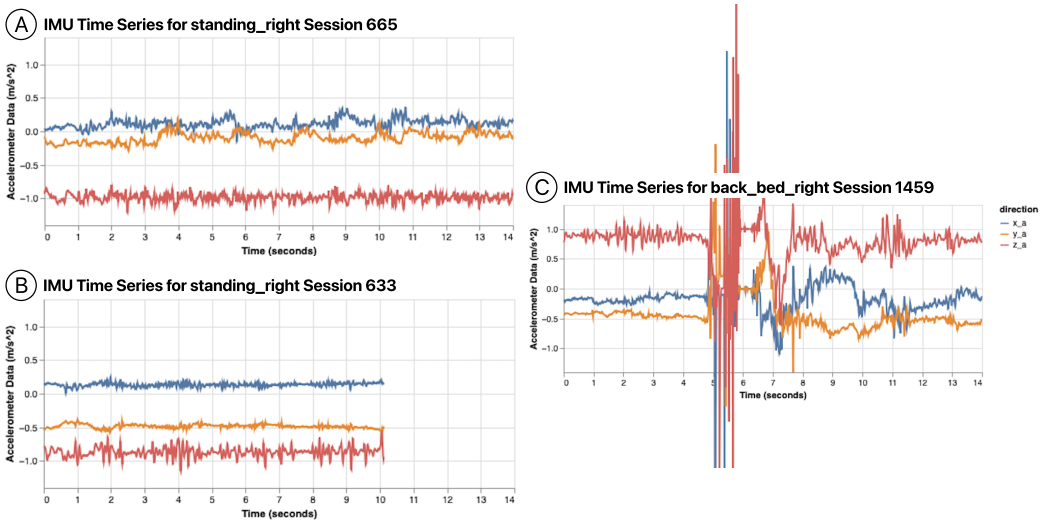

Figure 14. Distinct examples of data characterized as most "unfamiliar" to a model. (A) is an example of data we would consider out of distribution, (B) presents a case of sensor failure—the sensor stopped recording part way through the task—and (C) shows a particularly noisy sample, likely where someone dropped their phone mid-typing.

1. Replace k most familiar samples with k least familiar samples;
2. Distributed sampling across a window of $k + i$ most familiar samples with k least familiar samples, where i represents a percentage of the overall training set
3. Distributed sampling across a window of $k + i$ most familiar samples with $k + i$ least familiar samples, where i is a percentage of the training set

Each sampling mechanism was compared across multiple k and $k + i$ values to determine the relative "sweet spot" for our sampling strategy given a particular training dataset. We randomly select X percent of best and worst scores, varying percentage between 0.5%—0.01%. We train model(s) on each variation of window size and sampling percentage, repeating the previous steps k times to ensure a multifold validation, then comparing the intersectional performances of M_1 to the new model M_i trained on the familiarity-informed dataset. This is repeated per the scenario described in the full paper.

F. Limitations and Future Extensions of Familiarity

Distinguishing between rare and noisy data A weakness of familiarity is that we have no current method of distinguishing noisy samples from out of distribution data. While an unfamiliar sample might stand out to the model, in many cases, human review is necessary to evaluate its implications. For this reason, *in very noisy datasets*, familiarity may be a tool best used for debugging. Future research might seek to incorporate algorithmic methods of distinguishing sources of uncertainty, however there is currently little work on the topic. Existing research either relies on the learning rate as a proxy for discriminating types of uncertainty as aleatoric or epistemic (D'souza et al., 2021), for example, or builds on Bayesian networks (Kendall & Gal, 2017). Both face various weaknesses, and there remains great need for additional techniques and evaluations.

Familiarity for new data While we explored familiarity largely from the perspective of "self-familiarity"—that is, what the model has already been exposed to, it also introduces a mechanism by which we can understand how a model responds to data it's previously not seen. This may offer a mechanism of transparency through which future users could evaluate how a model responds to new data. In this work, we computed familiarity from a single layer. In future work, we will explore how familiarity computed at different layers can be leveraged. Given that each layer captures distinct features within the data, aggregating information across depths of the model may lead to more holistic identification of unfamiliar data *and* what features are more specifically so to the model. One way to do this is through a Product of Experts (PoE) paradigm (Hinton, 2002) where each layer is considered an "expert".

Comparisons against active learning On the surface, familiarity appears similar to active-learning (AL). AL requires practitioners to choose which data to use given a large collection. In our scenario, we must understand which data to *collect* or gather when there is no additional data readily available to run the AL algorithm on. One way to

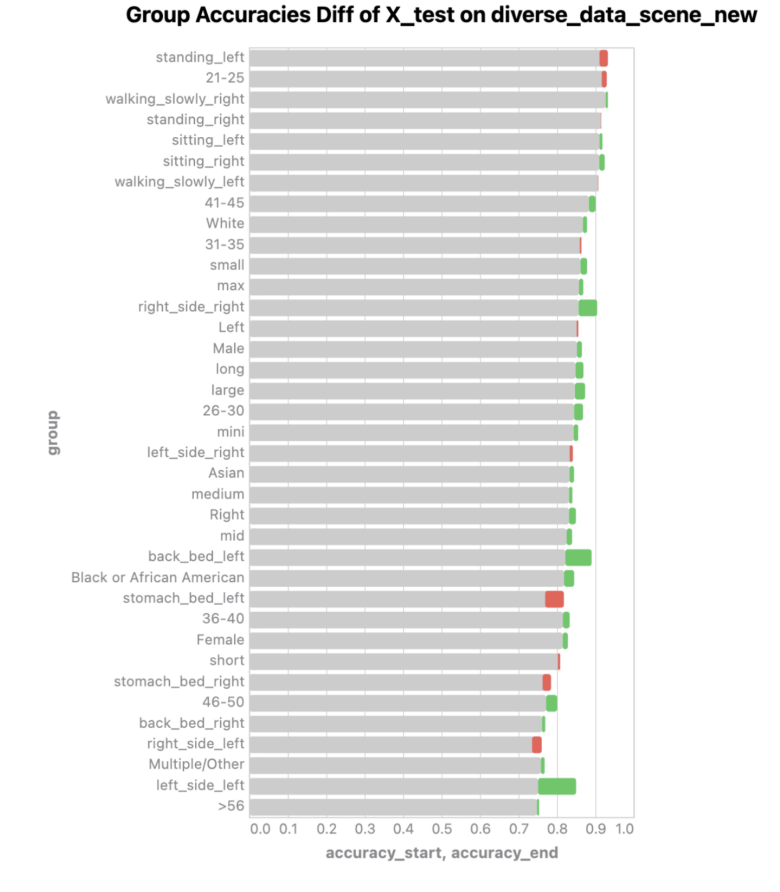

Figure 15. Comparison of accuracy on testing data before and after familiarity intervention. Minor regressions are shown, while accuracy improvements in contrast

circumvent this difference is to apply AL to the training set, and then extract statistics on the metadata that the AL algorithm indicates as most useful. For example, one could use the entropy of the logits: high entropy on a data point might be an indication that the model is still uncertain about such type of data. An issue with such approach is that it implicitly assumes models are well calibrated, which is not always the case.

Interactive systems for designing data Designing data advances how we account for the interplay between data and model (Hooker, 2021), considering both within the deployment cycle to compensate for missteps in either. In our case study, we use data visualizations (e.g., Figure 2) to compare practitioners' expectations against collected data distributions, then visualize familiarity to explore rare or noisy samples. The visualizations and interfaces used in this work are largely static; however, we see a great opportunity to build the designing data process into future interactive systems and tools for better data work and model evaluation. From the HCI and visualization communities, there are a number of interactive systems that have helped ML practitioners explore their data (Inc., 2021; 201, 2017; Hohman et al., 2020b) and evaluate their models (Amershi et al., 2015; Wexler et al., 2019; Görtler et al., 2022); for an in-depth survey on visual analytics for ML see (Hohman et al., 2018). Directions for future interactive systems might include tools to help practitioners reflect on their data collection practices (e.g., digging into their expectations, as discussed in Appendix C), or tools to direct familiarity analyses.

G. Extended Literature Review

G.1. Fairness, Diversity, and Heterogeneity

Fairness has become an increasingly common consideration for algorithmic contexts (Drosou et al., 2017). However, the term has been conflated with justice, bias, ethical machine learning, equity, diversity, heterogeneity, and inclusion due to cross-disciplinary use and overloaded terminology (Mitchell et al., 2020; Celis et al., 2018).

Algorithmic notions of fairness are often presented through mathematical formalizations intended to ensure anti-discrimination in the context of classification systems. They most commonly focus on ML model outcomes, but may also describe input data or how systems use model results (Mitchell et al., 2020). ML models incorporate outcome-specific fairness by adding constraints to ensure either group or individual parity across classification error metrics (Mitchell et al., 2020). Group fairness enhanced models (Friedler et al., 2019) use either anti-classification (wherein protected attributes are not included in decision making), classification parity (groups across protected attributes have similar predictive performance), or calibration (results are independent of protected attributes) methods (Corbett-Davies & Goel, 2018). In contrast to group fairness, individual fairness (Dwork et al., 2012) asks that individuals similar for a task be treated similarly throughout that task. While well-intentioned, each of these fairness enhancing approaches have incited criticism (Corbett-Davies & Goel, 2018).

Mitchell et al. described diversity, heterogeneity, and fairness as related but distinct concepts (Mitchell et al., 2020). In contrast to prior work by (Asudeh et al., 2019; Zheng et al., 2012), which considers a more encompassing definition of diversity—a measure to capture nuance of collection based on variety of constituent elements—Mitchell et al.'s diversity emphasizes attributes of social concern. They argue diversity measures that are not specific to social groups should instead be considered heterogeneity measures, though diversity has long been interchangeable with heterogeneity. Mitchell et al.'s argument is understandably motivated by concerns of specificity and impact, but reconciling what data may confer biasing effects for social groups with what does not is difficult: can human-agnostic data exist in data collected by and in reference to people?

Fields with historic interest in diversity or heterogeneity include information retrieval, ecology, biology, organization science, sociology, and chemistry, all of which have developed or employed approaches to measure diversity. These approaches largely fall into one of the seven following buckets: geometric or distance-based, combinatorial, aggregate, utility and ranking, coverage, and hybrid distance and coverage based diversity measures (Drosou et al., 2017; Zheng et al., 2012). Examples of geometric or distance-based measures of diversity include by dataset's volume (Mouchet et al., 2010; Chao et al., 2014; Anari et al., 2016; Celis et al., 2016; Whittaker, 1972) or by variance such as in principle component analysis (PCA) (Samadi et al., 2018). In general, these metrics more closely match with Mitchell et al.'s definition of heterogeneity as they do not explicitly refer to features with societal import and context (Mitchell et al., 2020). Ultimately, Mitchell et al.'s interpretation of diversity is intended to bring to light social inequalities found in ML products *specifically*, while others emphasize the measure of heterogeneity—comprehensive variety within the dataset.

Diversity metrics in ML contexts have been used to direct bias mitigation efforts (Celis et al., 2016). Instances of this can be found in subset selection, where declarations of "diversity constraints" define expected frequencies for sensitive values that the data must satisfy (Milani et al., 2020), or to measure relative coverage within a dataset (Mitchell et al., 2020). An alternative subset selection technique associates diversity scores to subsets then chooses subsets

when probabilities are proportional to this score (Celis et al., 2016; 2018). Uniformly, these methods act as stopgaps to biased and/or homogeneous data—particularly common in the context of data summarizations of big datasets (Celis et al., 2018).

The ability to produce diverse subsets using diversity metrics does not guarantee fairness across samples in the form of appropriate representation of sensitive attributes (Celis et al., 2018). Partially, this is because fairness has multiple measures (Mitchell et al., 2020). Fair treatment across social groups may also require *different things for different contexts*. For instance, consider a dataset in which each data point has a gender. One notion of group fairness, useful for ensuring that the ground truth is not distorted, is proportional representation, i.e., the distribution of sensitive characteristics in the output set should be identical to that of the input dataset. Another notion of fairness, argued to be necessary to reverse the effect of historical biases, could be equal representation where the representation of sensitive characteristics should be equal independent of the ratio in the input dataset. The plethora of metrics and evaluative tools vary and each have ongoing discussions of merit. Ultimately, datasets must be considered in relation to the use at hand, and the potential harm any failures may cause.

Qualitative research methods, statistics, and survey literature have historically managed representative data collection in a variety of ways, from expert panels to standards in population survey techniques, yet these methods face their own complications and do not necessarily translate to the needs of machine learning teams.

G.2. Bias Mitigation

While examples of algorithmic bias are often highlighted by media outlets, this frequency belies the difficulty of initial discovery—failures are hard to uncover during development, thus responsibility often falls to the public via product engagement. In the case of Google Photos, the model could not distinguish between a gorilla and a Black person and it was a member of the public that flagged the concerning labeling. Situations like racist photo identification algorithms are not uncommon, and once uncovered are responded to in a variety of ways. Google's response was to censor outputted labels such as "chimpanzee" and "gorilla" within the public Photos app search results (**?**). In their case, the model could not distinguish between a gorilla and a Black person because it was not trained to do so—there were not enough instances of Black faces in the dataset (Asudeh et al., 2019). Biased data and lack of diverse representation has broad impacts in domains beyond computer vision. Popular personalized voice agents, for example, struggle to recognize foreign accents under certain contexts such as understanding medication names when users *correctly* pronounced them (Palanica et al., 2019).

Outside of censoring outputs, bias mitigation strategies can be divided into three stages of a ML model development: pre-training (e.g., sample weighting or dataset balancing), in-training (e.g., adding specific constraints in the function that is being optimized) and post-training (e.g., by tweaking the prediction in order to ensure some fairness metric) (Donini et al., 2018). Pre-training can be further divided into collection and post-collection. Despite the fundamental nature of data collection, typical technical approaches to bias mitigation focus on post-collection efforts—e.g., modification of the dataset, reweighting and fine-tuning of hyperparameters, filtering output, or some combination therein—which largely act as stopgaps, are sensitive to underlying data (Wieringa, 2020) and ultimately may not resolve underlying issues at hand.

Additional complications arise when working with data acquired independently, possibly through a process in which the data scientist has little or no control. This "found data" (Ramasamy Ramamurthy & Roy, 2018) introduces unique challenges to ensuring data coverage for scientists and engineers. For both found data and big data contexts, post-collection approaches such as subset selection and class imbalance corrections like under- or over-sampling are introduced to counter bias and skew (Japkowicz & Stephen, 2002). Yet sampling methods can obfuscate information about appearance frequency in a dataset—if there are limited examples of X and so we oversample, then those instances of X are not unique and may cause the model to infer incorrect characteristics of a class, affecting accuracy metrics as well as production performance. These approaches are limited; for example, when improving a model without access to the original training dataset, balancing in the traditional sense is impossible. Regardless of the context, these data balancing approaches ignore data learnability: having an equal number of samples per class neglects the fact that some classes are inherently easier to learn than others (Ben-David et al., 2019; Schapire, 1990; Klawonn et al., 2019).

Work on ML generalization has looked to avoid the issue of biases in data distribution and collection altogether by focusing on the causal relationships within data. Supposedly robust to lack of coverage or variation within a dataset, Invariant Risk Minimization (IRM) relies on latent variables within the data—information not explicitly observed, but rather which are inferred from existing data—to learn concepts more closely aligned with ground truth (Arjovsky et al., 2019). IRM has met been met with scepticism, however, and work by Rosenfeld et al. found that IRM can "fail catastrophically" unless the test and training distributions are sufficiently similar (Rosenfeld et al., 2020), ultimately coming back to the intitial question of how to develop diverse datasets that more accurately reflect real world use

cases.

G.3. Contrasting Familiarity to Other Methods

Lee et al (Lee et al., 2018) is most similar to our proposed method of familiarity. However, there are several differences between our work and theirs: they use class-conditional density estimation, while we do not use labels, thus fit one model across all classes. We use PCA to reduce activation dimensionality before fitting the model (DE in high dimensional space is difficult). Finally, we use a variational Bayesian estimation of the GMM; they estimate Gaussian mean and covariance differently.