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

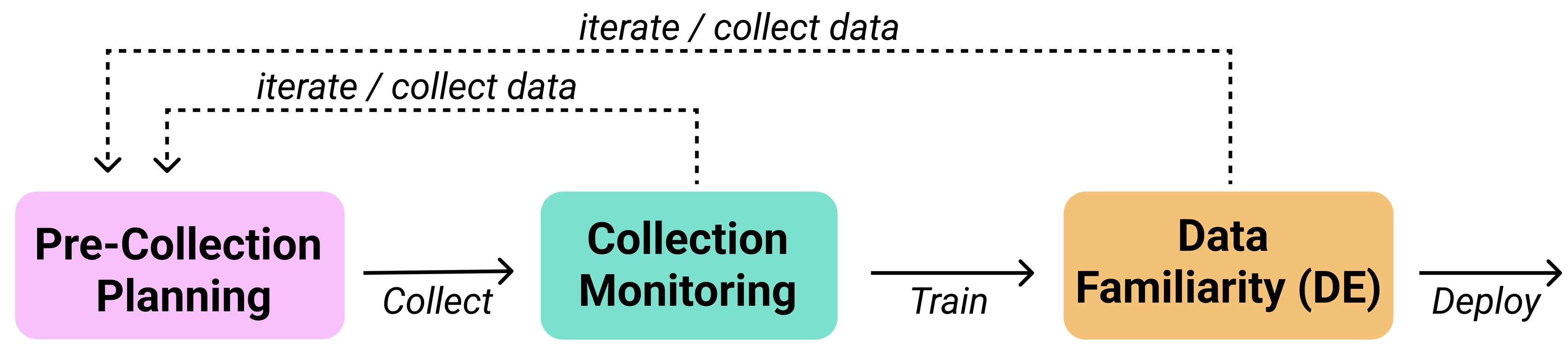
Aspen Hopkins*, Fred Hohman, Luca Zappella, Xavier Suau Cuadros, Dominik Moritz
*dataspen@mit.edu

Motivation

- Lack of diversity in data collection causes failures when deploying ML products
- Post-collection interventions are time intensive and rarely comprehensive
- Focusing only on one aspect of an ML pipeline is not enough
- New methods integrating collection, iteration, & model development to uncover “unknown unknowns” are necessary to minimize harm and maximize diversity

Contribution

Designing data is an iterative approach to data collection. It includes (1) **Pre-Collection Planning**, (2) **Collection Monitoring**, & (3) **Data Familiarity** (an application of density estimation). Each intervention complements the others, ensuring the final dataset provides as comprehensive coverage as possible.

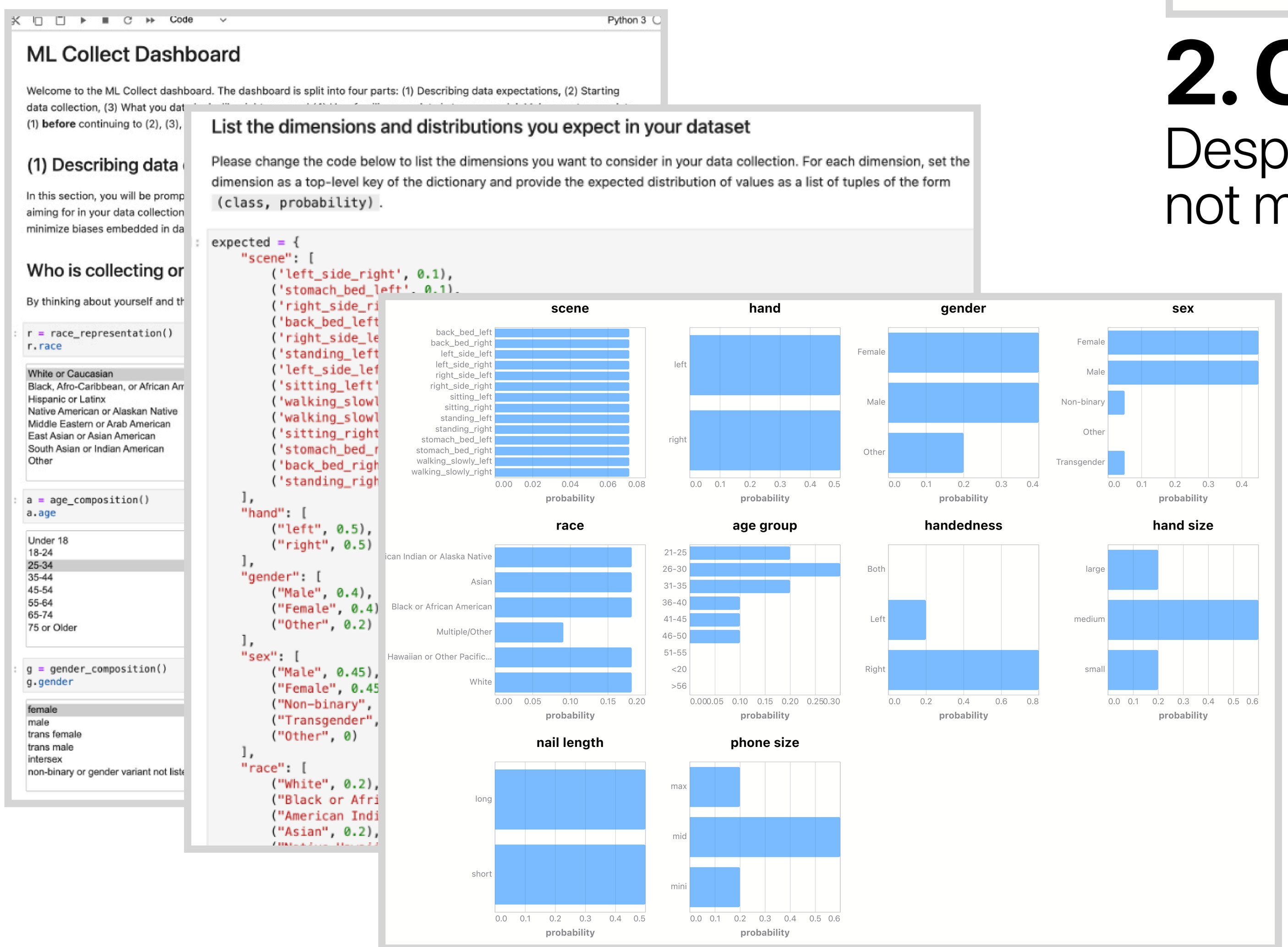


Designing Data

1. Pre-Collection Planning

Building representative datasets is an arduous, historically difficult undertaking that relies on the efficacy of human-specified data requirements.

With reflexive planning & by documenting expected distributions, collectors ensure these specifications are as compre-hensive as possible before collecting.



2. Collection Monitoring

Despite best efforts, data collected might not match expectations.

By contrasting expected distributions to real time data, we capture the dataset’s evolution, allowing users to make targeted adjustments when needed.

3. Familiarity

Despite increased rigor in collection, expected and actual data distributions may not match learning needs of model.

By understanding how a model perceives data, we can focus data collection efforts on the most useful subsets, reweighting or replacing data accordingly

To this end, we use density estimation (DE) to uncover samples that are unfamiliar to the model—those that either are not represented appropriately, are challenging to learn, or were erroneously collected.

Here, we learn a Gaussian Mixture model (GMM) on a NN’s layer activations:

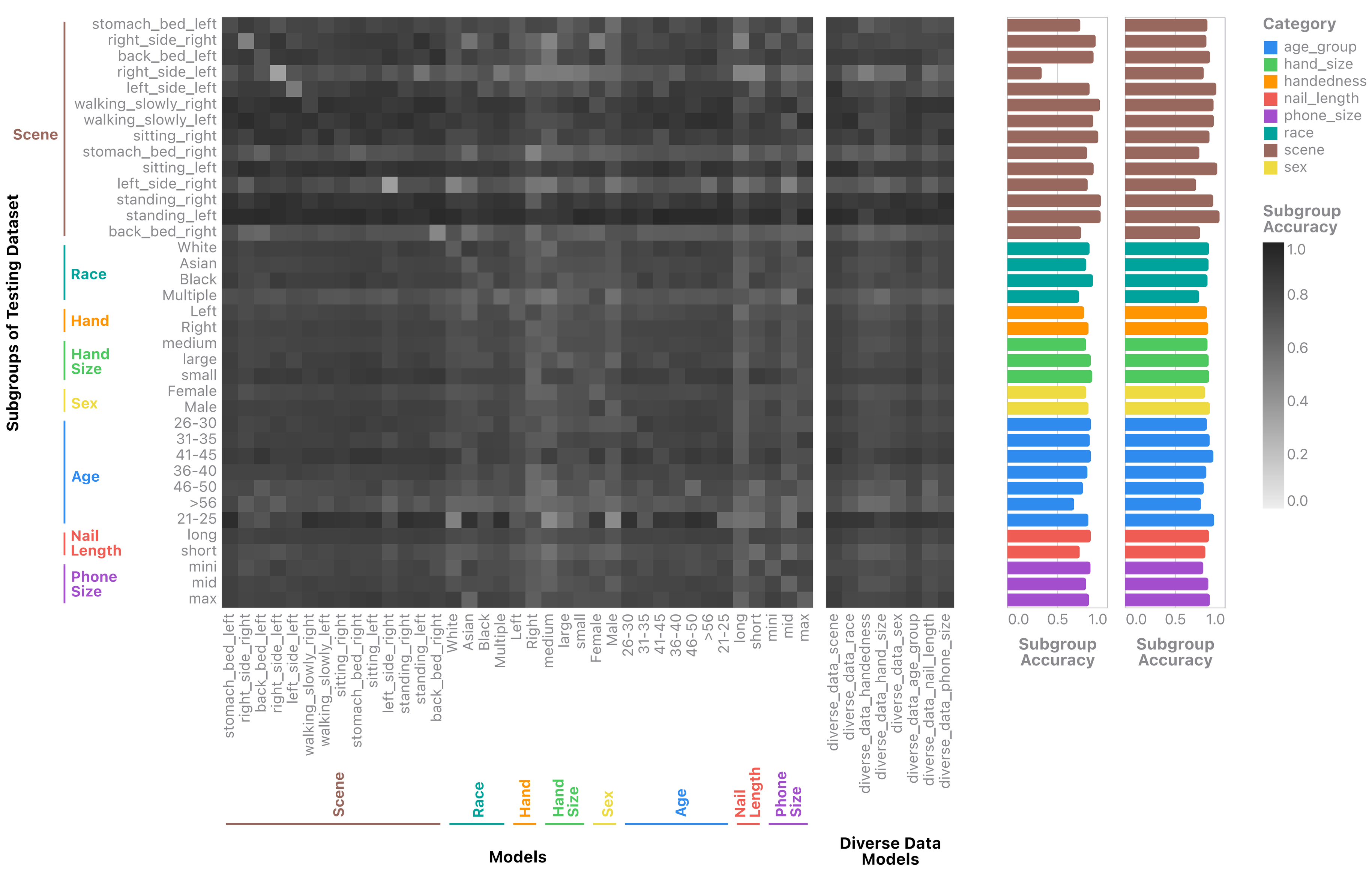
$$p(x \mid \lambda) = \sum_{i=1}^{M} w_i g(x \mid \mu_i, \Sigma_i)$$

Where x is the matrix of layer activations, $w_i, i = 1,\ldots,M$ are the mixture weights, and $g(x \mid \mu_i, \Sigma_i), i = 1,\ldots,M$ are the Gaussian densities. PCA is used to reduce the dimensionality. The resulting log-likelihood values are the **familiarity scores.**

These are used to debug a dataset early in data collection. Later, it informs data iteration, improving diversity and coverage. While DE for OOD detection is well studied, our use of DE to direct data work is unique.

Results

Does auditing to increase data diversity improve model generalizability?



Is data familiarity useful for auditing model & data?