# Why Real Data Matters

## Garbage In, Garbage Out

The phrase "Garbage in, garbage out" is an old computer science adage that is particularly relevant to machine learning systems. The implication of this statement is that the data in our training set needs to match the data our system will observe during deployment as closely as possible, otherwise we risk a loss of performance and other undesirable effects. The mismatch between the data we use to train our models and that which is seen during deployment can manifest in a number of different ways.

Firstly, a scarcity of data is typically our most prevalent issue. It takes a lot of time and effort to procure and manage a large and high-quality dataset. Fortunately for us, machine learning models can often function with as little as a few hundred or thousand data samples, but this much data can still be challenging to obtain. The more complex our data distribution is, the more samples we need for our machine learning model to learn an accurate approximation of this distribution. If we are trying to build an algorithm to understand voice commands, we need to take into account different types of accents, pitches, inflections and intonations. This can be particularly problematic if you are part of a small team, and you may need to get creative and do some crowdsourcing of data (as Pete Warden did with his Speech Commands dataset!).

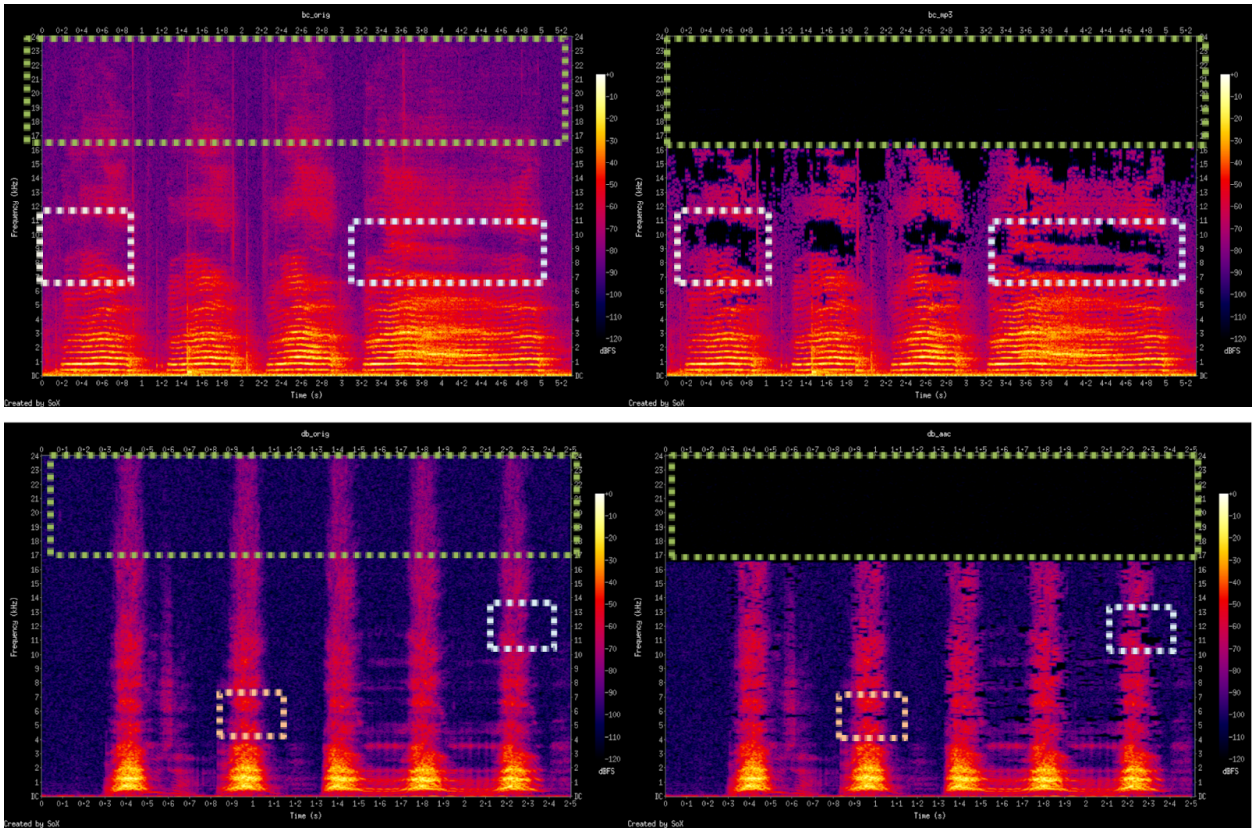
A separate but related issue is how representative our data is of the real-world use case. As we just highlighted, a model to respond to voice commands should work for a broad range of accents and voice types. If we built a model based purely on data of our own voice, it might work great when you give a voice command, but work terribly for everyone else (even if they try to imitate you!). The reason this example fails is our model is not representative of our data distribution, only a small subset of it. To improve our model's performance, we would have to obtain more data samples that have a greater diversity of accents and voice types.

## Legal Challenges

So, we have established that curating a dataset is difficult. However, you may have had a thought while reading the last section - why don't we just use datasets that already exist? This is an excellent question. In our example, can't we just go ahead and use Pete Warden's Speech Commands dataset? The answer is that it depends. Some datasets are shared under specific licenses, and if we plan to use them, we must make sure to abide by the stipulations of those licenses, especially if we plan to use the dataset in any commercial capacity. However, whether commercial models can be deployed that are trained on proprietary datasets is a bit of a gray area, highlighted nicely by Authors Guild v. Google. In this case, Google was sued for training a search algorithm on copyrighted books. However, Google won the case because it was deemed that the algorithm actually benefited the copyright holders since the search algorithm made it easier for the books to be found via a search engine. While this is an isolated case, this sets an interesting and note-worthy precedent.

## Technical Challenges

Beyond legal limitations, there are also technical limitations associated with using external datasets. Once again, they might still not be representative of our end user use case. As an example, audio data that was recorded in a recording studio might not represent the real environment, where there may be background noise such as traffic or people talking. The difference between these two environments would add bias to our data that might impact our model's performance for the end user. Even using different compression techniques can impact the performance of our model. In the examples below, we see spectrograms of different encodings for two examples, (1) a baby crying, and (2) a dog barking, before and after different audio compression techniques. These techniques leave characteristic markers in our data that are not obvious to the listener but would be obvious to a machine learning algorithm. Thus, sometimes merely transferring between data types can impact the usability of our data!




Similarly, training an algorithm on images that are always horizontal might result in a performance loss if our camera is placed at an angle during deployment, and poor lighting may cause similar problems if this was not accounted for in our training data. Naturally, it can be difficult to consider all of the possible nuances of our deployment environment, which is why curating high-quality datasets is challenging. One possible way around this is to fine-tune our model on a small set of data from our deployment environment, after broader training using a

larger dataset. This provides the benefit of generalizability, while also taking into consideration the particularities of our particular deployment environment.

Hopefully, you now have a greater appreciation for how important, but also how difficult, it is to produce a high-quality dataset to train our machine learning models. We will delve deeper into this in the following sections, but the key takeaway here is to always keep the end-use case in mind when developing your training dataset.

## Additional Resources

Estimating required sample size for model training
Data-Centric Ai Resource Hub