MODULE ANSWER | Foundation

Include reader-centric use cases

Help readers establish *their* use cases - not just yours.

Readers of Data Cards – and any other transparency artifact, such as Model Cards, should be able to intuitively understand the utility (what the dataset can be used for) and the usability (how to implement or deploy the dataset appropriately) of the dataset.

This is especially important to help readers of the cards when they need to make critical decisions that involve multiple parties, because a clear trail of accountability in the dataset lifecycle needs to be maintained. In general, if a reader can use a Data Card to predict or explain possible outcomes of using the dataset in a specific way, then they can be empowered to use your dataset more appropriately.

There are a myriad of contexts of use for your dataset, and no dataset is unbounded in its benefits. A dataset collected for evaluation will have clear shortcomings when used to fine-tune a model without additional data augmentation methods. Datasets are often created with a specific purpose in mind, such as training a model for a particular task, such as object detection. Such a dataset will probably not contain the features necessary for face detection, even if there are faces of people present in the dataset.

The boundedness of a dataset's benefits, its motivations, and complexity are vital for readers to make decisions about your dataset.

When establishing if and how to use a dataset, readers need to grapple with many factors, especially if datasets are complex or the systems in which they use the datasets have many moving parts. For example, many ML datasets are compared on the performance of benchmark AI models on them. And some datasets are labeled using an algorithmic classifier. In both cases, the dataset inherits biases from the use of AI models, and the Data Card you would create for each will need to provide additional nuance that can easily become invisible if not surfaced in the right context.

In addition to this intrinsic complexity, readers also need to factor in your decisions when creating the dataset. For example, the motivation for creating a dataset determines what the data are and how they are collected, processed, stored, and benchmarked. Your Data Card should be objective and comprehensive enough so readers can connect the dots between how beneficial your dataset is for their context and the limitations of your dataset. The boundedness of a dataset's benefits, its motivations, and complexity are vital for readers to make decisions about your dataset.

Datasets can also become stale very quickly and are not infinitely reusable, owing to the world's dynamic technological and socio-cultural landscape. While a list of intentions, motivations, and suitable use cases can be broadly used to describe the benefits of a dataset, they do not alone guarantee appropriate or timely use. If your Data Card paints a clear picture of the limitations, thresholds for safe use, and associated risks, then readers can establish the limits of a dataset's capabilities. So readers who then choose to use your dataset in novel ways can factor these into their expectations of successes and failures they might experience.

Key Takeaways

- Datasets are bounded in their benefits. Readers will need help in understanding where benefits of datasets cease to exist so they can confidently decide what to use the dataset for (utility), and anticipate challenges when using it (usability).

- Decisions of *appropriate* use of data are shared between dataset creators, owners, publishers, and dataset users. Your Data Card should ultimately help these different parties assume accountability of how they handle and use the dataset.

- Readers will fill in the blanks with their contexts and existing mental models. Your Data Card should include sufficient explanations, justifications, and attributions so readers can understand how your dataset will work in *their* systems.

Actions For Your Team

1. **Supplement benefits with limitations.** When describing a benefit, clearly describe associated limitations, risks, and contexts which can block the benefit – so readers can make informed trade-offs.

2. **Connect the dataset to product outcomes.** Pay careful attention to possible inequities, biases, skews, and imbalances in features of your dataset. Use these to articulate plausible real-world product outcomes in your intended use cases – both positive and negative. This helps readers understand unsuitable use cases in very real-world terms.

3. **Reviews are important context clues.** Provide summaries of your dataset's performance under different conditions so readers can pick conditions that are closest to their system. These build trust and credibility when they describe the experience of other individuals and teams who have used your dataset.

4. **Documentation is a product surface.** Treat your Data Card as a product surface, and where possible, work with product teams and individuals with lived experience to inform your position on how the dataset should be used.

Considerations

→ What are the benefits of the dataset? Under what conditions do these benefits cease to exist?

→ What minor “tweaks” to the intended use or implementation of the dataset produce adverse effects?

→ If your dataset has been used by other individuals and groups, ask them to summarize what worked and what didn’t in their contexts of use and include that in your Data Card.