The Dataset Nutrition Label:
A Framework to Drive Higher Data Quality Standards

Sarah Holland[1]*, Ahmed Hosny[2]*, Sarah Newman[3], Joshua Joseph[4], and Kasia Chmielinski[1]
April 2018 | nutrition@media.mit.edu | datanutrition.media.mit.edu

[1]Assembly, MIT Media Lab and Berkman Klein Center for Internet & Society at Harvard University, [2]Dana-Farber Cancer Institute, Harvard Medical School, [3]metaLAB (at) Harvard, Berkman Klein Center for Internet & Society, Harvard University, [4]33x.ai
**Authors contributed equally*

ABSTRACT

Artificial intelligence (AI) systems matter, and so does the data on which they are modeled. One way to improve the accuracy and fairness of these models -- models that determine everything from navigation directions to mortgage approvals -- is to make it easier for data scientists to quickly assess the viability and fitness of datasets used to train them. Current methods of data analysis and assessment are not standardized; they vary greatly across industry and domains, and are also costly in time and expertise. Drawing from the fields of nutrition and privacy, we introduce a solution: the Dataset Nutrition Label, a diagnostic framework to address and mitigate some of the challenges in this process.

The Dataset Nutrition Label (the Label) provides data scientists with a distilled yet comprehensive overview of dataset 'ingredients' before AI model development. The Label is comprised of modules that indicate key dataset attributes in a standardized format. The Label also utilizes multiple statistical and probabilistic modelling backends to capture such attributes. The modules include both qualitative and quantitative information, and vary in both the access they require to the dataset as well as the point at which they are generated or appended. Different combinations of modules can be used to customize Labels for particular datasets. Modules can also be added at a later time by data scientists using and interrogating a dataset. To demonstrate and advance this concept, we generated a Label for the ProPublica *Dollars for Docs* dataset, which documents payments made to physicians by drug companies in the U.S. between 2013-2015. The live, interactive, and open source prototype displays the following sample modules: metadata, data provenance, diagnostic statistics, variable correlations, and ground truth comparisons. The prototype is available on the Dataset Nutritional Label website.

The Dataset Nutrition Label offers many benefits. It drives robust data analysis practices by providing a pre-generated 'floor' for basic data interrogation. Data scientists selecting datasets for model development can leverage the Label to quickly compare the 'health' of multiple datasets, helping them efficiently select the best dataset for their purposes, and avoiding onerous and costly analyses. Improved dataset selection provides a secondary benefit: the quality of the models trained on that data will also improve. This is a result of using a more robust dataset generally, and also as the Label enables data scientists to check for additional issues at the time of model development (e.g. surprising variable correlations, missing data, anomalous data distributions, etc). The existence of these Labels will also prompt data scientists to question a dataset and its characteristics regardless of whether every dataset contains such a Label. Lastly, by encouraging the authors and users of datasets to create Labels, we hope to build awareness of significant shortcomings in datasets (e.g. missing data, biased collection practices), which in turn will drive better data collection practices and more responsible dataset selection and use in the future.

We also explore the limitations of the Label. Considering the variety of datasets used to build models today, some challenges arise with generalizing the Label across data type, size, and composition. Some of the modules require access to the dataset's author and the data itself; this could constrain the creation of certain modules to only those who own, manage, or have access to the data. We discuss the risk of modules that rely on 'ground truth' data, which will depend on the accuracy of such ground truth for comparison. The design of the Label itself will require additional attention to determine the appropriate amount of information for presentation, comprehension, and adoption. Finally, we discuss potential ways to move forward given the limitations identified.

The Dataset Nutrition Label is a useful, timely, and necessary intervention in the development of AI models: it will encourage the collection of better and more complete data and more responsible usage of such data; it will drive accountability across various industries, and it will mitigate harm caused by algorithms built on problematic or ill-fitting data. We lay out future directions for the Dataset Nutrition Label project, including research and public policy agendas to further advance consideration of dataset labeling.

KEYWORDS

Dataset, data, quality, bias, analysis, artificial intelligence, machine learning, model development, nutrition label, computer science, data science, governance, AI accountability, algorithms