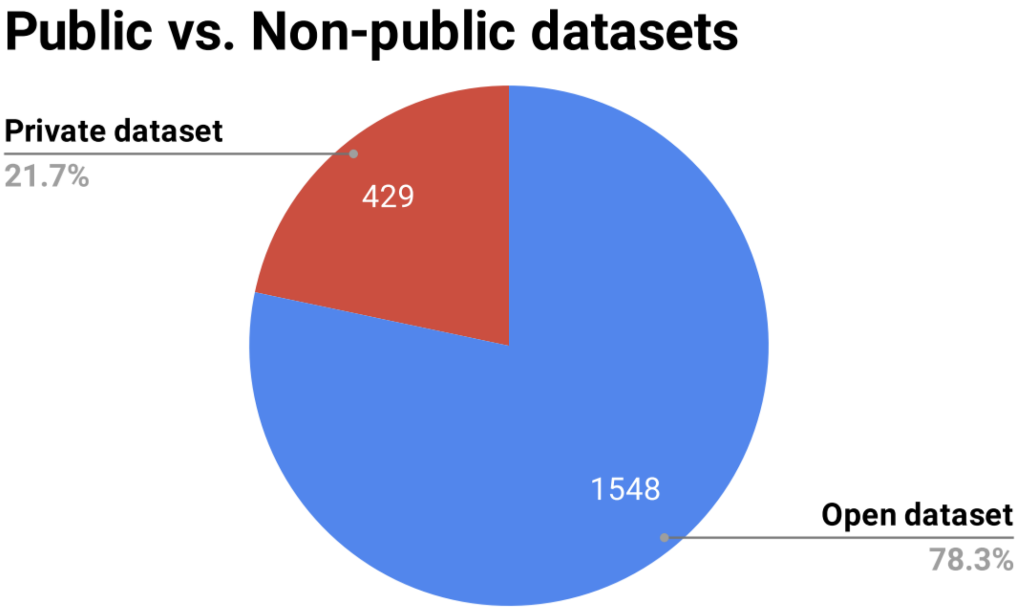
Data Engineering For Everyone

Background

During the early days of software development, many programs, tools, and architectures were closed-source (i.e., individuals could not freely peruse the codebase). The advent of open-source software to help overcome problems faced by large distributed teams of software engineers propelled software development forward massively, enhancing productivity and broadening access to resources. Today, we face a largely similar challenge with respect to data.

The Importance of Public Data

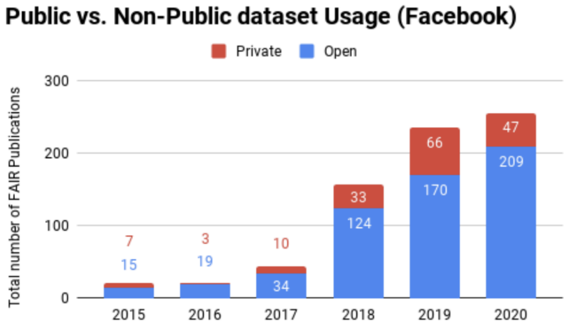
Machine learning is becoming increasingly ubiquitous, with tens of thousands of research papers now published on the subject every year. The majority of these algorithms require data to achieve their objectives, meaning that the appetite for data is concomitantly increasing. Expertise in data engineering is becoming increasingly necessary to architect, build, and manage large and high-quality datasets used for training machine learning algorithms. However, a substantial proportion of these datasets are built and managed by large corporate entities, and may or may not be open-source (approximately 1 in 5 published works by Google, Facebook, and Microsoft utilized private datasets). You can find the source of the details here.

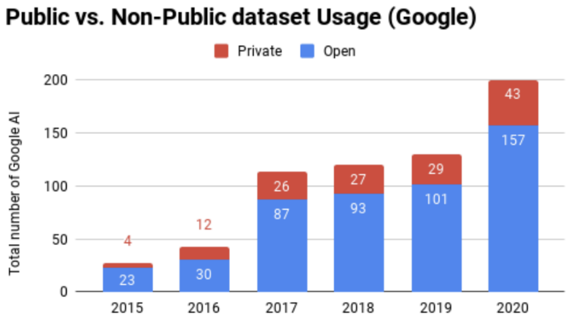

Public vs. Non-public datasets

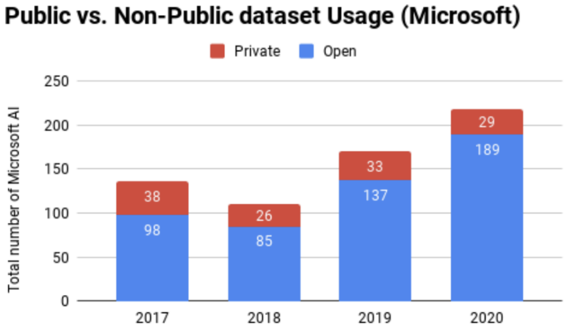
The State of Public Data

Within the machine learning community, there has been a primary focus on architectural or algorithmic problems as opposed to a focus on data. Consequently, the development of efficient architectures to crowd-source and democratize large datasets for use in machine learning algorithms has been largely sidelined. Many of the data benchmarks that exist today, such as the ImageNet and COCO datasets, were developed within academia by a small group of researchers. However, since most machine learning algorithms today require more data than can be readily produced by such groups, there is a growing need for such massive datasets to be produced through open-source contributions.

Although there has been a minor improvement in the prevalence of open-source dataset usage within academic publications over the years, this number should ideally be 100% to allow results to be readily checked and reproduced. When comparing publications by Google, Facebook, and Microsoft, the prevalence of private datasets has remained fairly consistent. This suggests that even for the giants of industry, improving the use and development of open-source datasets is not a primary goal.







Progress

Initiatives such as MLCommons have been developed to help encourage assessment using open-source benchmarks, datasets, and best practices, but there is still a long way to go before data engineering is on a similar level of open-source parity as software engineering.