**Open Data in Context**

Datasets are the lifeblood of myriad forms of news analysis, academic research and business intelligence. Much of these data are gathered by corporations, universities, governments and nonprofit organizations to use in analysis, assessment, and exploration of various research questions, to make decisions, explore patterns of behavior, and monitor natural and man-created phenomena. While open data are freely available, modifiable, machine-readable, and shareable, other data are not - and can be distinguished from open data based on how they are collected, distributed and accessed.

**Big Data:** [**Volume, Velocity, & Variety**](https://datafloq.com/read/small-data-vs-big-data-back-to-the-basic/706)

Big Data is a catch-all phrase that encompasses the large flows of data captured by large organizations, including governments, healthcare providers, internet service providers, websites, transportation companies, retailers, financial services companies, and more. This high volume data comes in streaming, social media, and public form as a result of activities people undertake - such as reading and responding to email, clicking on website links, shopping on retail sites, conducting internet searches, and choosing transportation routes - all can be collected and mined for insights - often using analytical tools to extract elements of the data and present them in context with other information. While big data and open data overlap, not all big data is open, and not all open data is big data. In particular, some big data are proprietary, created through collection and analysis methods that publishers keep private to maintain competitive advantage or financial value. Conversely, some open data consist of small datasets that, while useful to the public and government, are not considered “big.”

**Paid and Proprietary Data**

Not all available data are open, free, and easily augmented, even though they are still accessible online. Organizations and companies can make data they have developed available for purchase, such as some of the [datasets that ProPublica has developed](https://www.propublica.org/datastore/datasets) as a result of its data journalism work. ProPublica’s data are particularly interesting because they can be an amalgam of government open data and data collected through the Freedom of Information Act. Thomson Reuters’, for example, also offers their [ESG (Environmental, Social and Governance)](https://financial.thomsonreuters.com/en/products/data-analytics/company-data/esg-research-data.html) proprietary datasets for a fee, which are used to develop measures of the potential return on investment in financial decision-making. There are thousands of examples of these types of useful and paid datasets across all industries.

Other companies offer their data publicly to varying degrees. For instance, ESRI, the world’s largest provider of geospatial data analysis software provides an [open data platform](http://opendata.arcgis.com/) on which it encourages its users to share their data, globally. In contrast, UBER released a tool, [Movement](https://movement.uber.com/cities), that provides some data on traffic patterns and rideshare behavior for its users in major cities. To date, however, the public does not have unfettered access, and city planning officials have access to limited elements of the data.

**Personally Identifiable Information**

Major organizations, such as school districts and healthcare providers, capture data to assess internal metrics. Because of the sensitive nature of this information, these data are protected by privacy laws. For instance, Kaiser Permanente uses the data it collects as a function of providing healthcare services to almost [12 million people](https://share.kaiserpermanente.org/article/fast-facts-about-kaiser-permanente/) to optimize preventive and acute care delivery to reduce overall system costs. While Kaiser has access to all of its patients’ data across its system, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) ensures that [individuals’ data cannot be released with any personally identifying information](https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html).

However, this does not mean this data cannot be used in collaborative ways. In recent years, government-facilitated efforts to help pharmaceutical companies, healthcare providers, and others in the healthcare domain share de-identfied data have led to major breakthroughs in [cancer treatment](https://www.sciencedaily.com/releases/2017/02/170201093613.htm) and [diabetes management](https://healthitanalytics.com/news/research-collaboration-to-use-big-data-for-diabetes-prevention).