The term “big data” is quickly becoming part of the scientific, business, and daily lexicon. Various disciplines and enterprises have embraced the concept in an attempt to push innovation and discovery in a world where ever increasing amounts of data are available. While, in theory, there seem to be clear benefits, the ability of organizations to ingest, process, and create knowledge from vast streams of data are not trivial. What “big data” means and its implementation in other contexts may help us understand how it can be applied to the field of air sensors specifically, and environmental data in general.
Over the past several years, a new term has emerged from the esoteric information technology lexicon to become almost common parlance. “Big data” is now being discussed not only in journal literature, but in television commercials. Some of the claims and benefits may be hype, aimed at the presumably deep pockets of chief information officers (CIOs), but the reality of ever-increasing data everywhere is far from fiction. Data historically have been difficult to locate and expensive to obtain, but with the advent of computers, and ever smaller and inexpensive devices, we are now in an age of data deluge. So, what exactly is big data, and more specifically, what are the implications for environmental data from air sensors? At the 2013 EPA Air Sensors Workshop, we aimed to introduce attendees to the term and some of the aspects it touches.

**Defining Big Data**

The National Science Foundation (NSF) has embarked on a multi-million-dollar program to fund research in big data, which NSF defines as: “…advance[ing] the core scientific and technological means of managing, analyzing, visualizing, and extracting useful information from large, diverse, distributed and heterogeneous data sets so as to: accelerate the progress of scientific discovery and innovation; lead to new fields of inquiry that would not otherwise be possible….”

A more business-oriented definition might be: “…an enterprise that can mine all the data it collects right across its operations to unlock golden nuggets of business intelligence….”

Some estimates put worldwide data generation at 2.5 Exabytes per day—that is equivalent to a stack of 2 Terabyte hard drives almost 20 miles long. While we may not be interested in this all-encompassing definition, even within the confines of a discipline or enterprise, the entirety of relevant data may still be substantial and growing fast. Big data, even in the realm of science, varies with discipline and presents challenges not only with regard to the size of data sets, but also in the diversity of data being integrated.

Genomics offers a prime example of the radical change that data has made to the discipline. The first bacterium cost billions of dollars to sequence and years to analyze, but now a personal genome can be mapped in a few days for around $1,000. Sequencing costs are plummeting by 50% every five months, making enormous amounts of data available. With hard disk storage costs (MB/$) only doubling every 14 months, the result is that next-generation sequencing technology is outpacing the computational resources to store, process, and manage data. So while sequencing originally dominated the cost structure, now the costs being incurred are in data management and downstream analysis.

This has resulted in the genomics community desperately and rapidly attempting to transform the way data collection and science take place by creating a, “platform which supports an ecology of databases, interfaces, and analysis software.” The ultimate goal being: to allow users flexibility in access to tools, data, and resources, in a scalable cost-effective manner, that provides incentives to share in the process of discovery. While data streams from air sensors may not constitute the levels of data generation in genomics, challenges nonetheless still exist in other aspects such as quality, sampling protocols, and integration of disparate data sets.

**Locating the Data**

One of the first steps in such an endeavor is knowing the existence of relevant data. The data we need may exist, but if we don’t know of it or can’t find it, it is essentially useless to us. While the concept of data discovery may seem obvious and philosophical arguments of epistemology notwithstanding, the issue of how and whether we can find data is central, critical, and nontrivial to the ability to utilize them in further analysis. In addition, data and information may not reside where we like, or be available in forms we desire, or generally be in a state for us to readily ingest or process.

Still a further complication is that, in general, there are two broad categories of data: structured and unstructured. Structured data are highly organized
these standards not only make finding data easier, but also greatly ease the creation of third-party applications that utilize the data. This allows not only measurements to be democratized, but so also the applications. Given standards and an application programming interface (API), anyone can build an application to query, retrieve, and utilize data.

Standards also allow for more modern collaborative software designs such as, Service Oriented Architecture (SOA). SOA allows for pieces of self-contained code that perform certain functions to be shared and combined with other SOAs. But standards are notorious for being static, not flexible, and sometimes operationally onerous. The ideal standard needs to have a form that, if needed, can be very detailed and abstracted, but can also be implemented in a simple form when appropriate.

The main problem with standards, however, is the difficulty in agreeing upon just one. Standards are like opinions, everyone has one or at least one they prefer. But standards will play a key role in the rate of advancement and ability to integrate data from air sensors. The format wars of VHS versus Betamax or DVD versus Blue-ray are testaments to both the promise and pitfalls of technology adoption and market penetration.

Practical Solutions

Given the needs at hand, organizations are now putting efforts into providing practical solutions for focused problems. The IBM Smart Cities Challenge for Louisville, KY, is one such effort. Despite the envious number of desirable rankings as a great place to live and do business, Louisville also had a problem: air pollution and public health. It ranked second for most unhealthy days for asthma, ninth for annual particulate pollution, and things seemed to be trending the wrong way. The city’s topography and proximity to regional air pollution sources seemed to be adversely affecting the population and tarnishing its image. But this wasn’t just an image problem with a marketing solution. Asthma has real economic consequences from productivity to healthcare costs to mortality. The challenge for IBM was to identify and analyze

So the situation may be that we don’t know whether or where the information we seek exists, or if it does exist, we can’t readily find it. If we do find it, access and format may be too onerous. Finally, the nature of the information (unstructured text and images) may prove too difficult to deal with and process. In the context of air sensors, the ability to discover new data sources in a distributed environment and be able to integrate that information to address a given problem will be require a large shift from the current centralized air pollution monitoring paradigm.

Setting Standards

An obvious way forward from big data’s four Vs—volume, velocity, variety, and veracity—is by way of standards. Standards essentially provide some basic structure with which to base queries for information. This is especially important given that most data are unstructured. We have touched upon the volume and velocity of data, variety is just that, a diverse set of data elements from various sources that contain information potentially important to an endeavor. And as was discussed elsewhere at workshop, veracity, in some ways, may be the most important, since it represents the accuracy or trustworthiness of the data. Without some confidence on the quality of the data at hand, it is very difficult to draw any conclusions or place any confidence bounds on estimates. Again standards can help if we adhere to them.

The Open Geospatial Consortium (OGC) has developed a number of standards related to sensors, namely Sensor Model Language (SensorML), Sensor Web Enablement (SWE), and Sensor Observation Service (SOS). If conformed to,
actionable data—including big data—to provide insights around the increasing level of asthma and the burden of this disorder in the community.20

Meeting the challenge required an initiative that not only implemented big data elements such as data mining, machine learning, and predictive modeling, but also required changes in governance (public–private partnerships) and community engagement (encouraging citizens to use mobile apps to collect and upload data).21 While the outcomes are not yet clear, the potential benefit from the infusion of data that numerous air sensors could provide in determining spatio-temporal patterns of air pollution is intriguing. It also demonstrates the sort of comprehensive and systematic changes that are required to meet complex challenges.

Clearly, much careful thought and deliberation among stakeholders is required when we are confronting a future that might include the possibility a vast number of new air quality sensors of varying quality being deployed. Many of these new sensors are likely to be portable and also have GPS capabilities allowing for mobile collection and transmission of data. But questions remain regarding data quality and how these measurements can best be utilized given a variety of use cases. One suggested framework was that of a “generative platform” (i.e., one that invites contributions from anyone who cares to make them).22 The contributions start among amateurs, who participate more for fun and whimsy than for profit. Their work, previously unnoticed in the mainstream, begins to catch on, and the power of the market kicks in to regularize their innovations and deploy them in the markets for larger than the amateurs’ domain.23

One realization of the generative platform might be the development of concepts such as, “Open Air,”24 which allows for a diversity of participants, from governments to individuals, using a variety of equipment, from reference grade stationary monitors to inexpensive mobile sensors, collecting on differing time and spatial scales. But even in such a utopian vision, the ways in which to integrate disparate data sets, where the data reside, the quality of the data, and who “owns” them, still remain critical and non-trivial questions.

Given that this is a rapidly evolving area of research and product development, growing pains are to be expected. The workshop attendees were clearly engaged and enthusiastic about the prospects that air sensors might, in the near future, provide air monitoring capabilities that are cheaper, faster, and more democratized, allowing broader engagement of public in issues of environmental health. em

References
4. Assuming 2-TB hard drives each 1 inch high and 63,360 inches/mile.
23. Johnathan Zittrain, “The Future of the Internet and how to stop it”.

Disclaimer: The U.S. Environmental Protection Agency (EPA) through its Office of Research and Development funded and managed the research described here. It has been subjected to agency’s administrative review and approved for publication as an EPA document.