**Introduction**

**Big data** is a blanket term for the non-traditional strategies and technologies needed to gather, organize, process, and gather insights from large datasets. While the problem of working with data that exceeds the computing power or storage of a single computer is not new, the pervasiveness, scale, and value of this type of computing has greatly expanded in recent years.

In this article, we will talk about big data on a fundamental level and define common concepts you might come across while researching the subject. I will also take a high-level look at some of the processes and technologies currently being used in this space.

**What Is Big Data?**

An exact definition of "big data" is difficult to nail down because projects, vendors, practitioners, and business professionals use it quite differently. With that in mind, generally speaking, **big data** is:

* large datasets
* the category of computing strategies and technologies that are used to handle large datasets

In this context, "large dataset" means a dataset too large to reasonably process or store with traditional tooling or on a single computer. This means that the common scale of big datasets is constantly shifting and may vary significantly from organization to organization.

**Critiques in Big Data**

This looks to be the year that we reach peak big data hype. From wildly popular [big data conferences](http://strataconf.com/strata2013) to [columns in major newspapers](http://www.nytimes.com/2013/02/05/opinion/brooks-the-philosophy-of-data.html), the business and science worlds are focused on how large datasets can give insight on previously intractable challenges. The hype becomes problematic when it leads to what I call “data fundamentalism,” the notion that correlation always indicates causation, and that massive data sets and predictive analytics always reflect objective truth. Former Wired editor-in-chief Chris Anderson [embraced this idea](http://www.wired.com/science/discoveries/magazine/16-07/pb_theory) in his comment, “with enough data, the numbers speak for themselves.” But can big data really deliver on that promise? Can numbers actually speak for themselves?

Sadly, they can’t. Data and data sets are not objective; they are creations of human design. We give numbers their voice, draw inferences from them, and define their meaning through our interpretations. Hidden biases in both the collection and analysis stages present considerable risks, and are as important to the big-data equation as the numbers themselves.

For example, consider the Twitter data generated by Hurricane Sandy, more than 20 million tweets between October 27 and November 1. A [fascinating study](http://sm.rutgers.edu/pubs/Grinberg-SMPatterns-ICWSM2013.pdf) combining Sandy-related Twitter and Foursquare data produced some expected findings (grocery shopping peaks the night before the storm) and some surprising ones (nightlife picked up the day after — presumably when cabin fever strikes). But these data don’t represent the whole picture. The greatest number of tweets about Sandy came from Manhattan. This makes sense given the city’s high level of smartphone ownership and Twitter use, but it creates the illusion that Manhattan was the hub of the disaster. Very few messages originated from more severely affected locations, such as Breezy Point, Coney Island and Rockaway. As extended power blackouts drained batteries and limited cellular access, even fewer tweets came from the worst hit areas. In fact, there was much more going on outside the privileged, urban experience of Sandy that Twitter data failed to convey, especially in aggregate. We can think of this as a “signal problem”: Data are assumed to accurately reflect the social world, but there are significant gaps, with little or no signal coming from particular communities.

While massive datasets may feel very abstract, they are intricately linked to physical place and human culture. And places, like people, have their own individual character and grain. For example, Boston has a problem with potholes, patching approximately 20,000 every year. To help allocate its resources efficiently, the City of Boston released the excellent [StreetBump smartphone app](http://streetbump.org/), which draws on accelerometer and GPS data to help passively detect potholes, instantly reporting them to the city. While certainly a clever approach, StreetBump has a signal problem. People in lower income groups in the US are less likely to have smartphones, and this is particularly true of older residents, where smartphone penetration can be as low as 16%. For cities like Boston, this means that smartphone data sets are missing inputs from significant parts of the population — often those who have the fewest resources.

Fortunately Boston’s Office of New Urban Mechanics is aware of this problem, and works with a range of academics to take into account issues of equitable access and digital divides. But as we increasingly rely on big data’s numbers to speak for themselves, we risk misunderstanding the results and in turn misallocating important public resources. This could well have been the case had public health officials relied exclusively on Google Flu Trends, which [mistakenly estimated](http://www.nature.com/news/when-google-got-flu-wrong-1.12413) that peak flu levels reached 11% of the US public this flu season, almost double the CDC’s estimate of about 6%. While Google will not comment on the reason for the overestimation, it seems likely that it was caused by the extensive media coverage of the flu season, creating a spike in search queries. Similarly, we can imagine the substantial problems if FEMA had relied solely upon tweets about Sandy to allocate disaster relief aid.

Big data’s signal problems won’t disappear as the use of smartphones and other digital technologies increases. As the geographers Michael Crutcher and Matthew Zook [noted](http://www.sciencedirect.com/science/article/pii/S0016718509000049) after Hurricane Katrina, technologies are always differentially adopted, and “any divide in accessing digital technology is not a one-time event but a constantly moving target as new devices, software and cultural practices emerge.” As we move into an era in which personal devices are seen as proxies for public needs, we run the risk that already existing inequities will be further entrenched. Thus, with every big data set, we need to ask which people are excluded. Which places are less visible? What happens if you live in the shadow of big data sets?

This points to the next frontier: how to address these weaknesses in big data science. In the near term, data scientists should take a page from social scientists, who have a long history of asking where the data they’re working with comes from, what methods were used to gather and analyze it, and what cognitive biases they might bring to its interpretation (for more, see “[Raw Data is an Oxymoron](http://www.amazon.com/Raw-Data-Is-Oxymoron-Infrastructures/dp/0262518287#reader_0262518287)“). Longer term, we must ask how we can bring together big data approaches with small data studies — computational social science with traditional qualitative methods. We know that data insights can be found at multiple levels of granularity, and by combining methods such as ethnography with analytics, or conducting semi-structured interviews paired with information retrieval techniques, we can add depth to the data we collect. We get a much richer sense of the world when we ask people the why and the how not just the “how many”. This goes beyond merely conducting focus groups to confirm what you already want to see in a big data set. It means complementing data sources with rigorous qualitative research. Social science methodologies may make the challenge of understanding big data more complex, but they also bring context-awareness to our research to address serious signal problems. Then we can move from the focus on merely “big” data towards something more three-dimensional:

**Technologies that have been born because of Big Data**

The bubble around Big Data has certainly started to burst and the coming year awaits reasonable developments in the applications of the Big Data world. Well, most of us are now more than familiar with terms like Hadoop, Spark, NO-SQL, Hive, Cloud etc. We know there are at least 20 NO-SQL databases and a number of other Big Data solutions emerging every month. But which of these technologies see prospects going forward? Which technologies are going to fetch you big benefits?

With the year coming to an end, this is a good time to make some predictions on how the Big Data industry will shape up. In this article, I have listed the top 5 technologies to emerge/advance in Big Data in the coming year based upon how Big Data has been doing so far and the upcoming industry trend.

### ****1. Hadoop will continue to rock****

Hadoop has been widely adopted by enterprises for their data warehouse needs in the past year. The trend seems to continue and grow in the coming year as well. Companies that have not explored Hadoop so far will most likely see its advantages and applications.

In terms of the technological developments, Hadoop will come up with features that would make it more enterprise ready. Once Hadoop security projects like Sentry, Rhino etc. gain stability, Hadoop’s implementation will expand across many more sectors and companies can use the solutions without much of security concerns.

### ****2. Real-Time Solutions will expand****

All the companies by now have the data and know how to store and process Big Data. The real difference is going to be how fast they can deliver analytics solutions for better business decisions. The Focus in 2017 is going to be Speed. Processing Capabilities of Big Data solutions will certainly increase. Projects like Spark, Storm, Kafka etc. were developed with this aspect in mind. We will see companies advancing from POCs to real world applications with these technologies.

### ****3. Cloud solutions will power Big Data solutions****

With Internet of Things (IOT) taking front seat, data generation is on its increase. Applications involving IOT will require a perfect scalable solution for managing huge volumes of Data. What other than cloud services can do this better? Advantages of Hadoop on cloud has already been realized by many organizations and technologies pertaining to the coupling of Big Data technologies like Hadoop, Spark, IOT and cloud is expected to be well on rise in the coming year as well.

### ****Traditional Database world will revolutionize****

RDBMS systems have been dominating the database world for decades when structured data formed the major proportion of data in any organization. Looking at the data sources today – Social media data, IOT, sensors etc. – where each one of us is generating volumes of data on a daily basis, it’s clear that the amount of unstructured data is steadily increasing and companies have started realizing the potential insights one can gain from such data. Well, now to manage and process such data NO-SQL databases have been the best option in the last few years. Well, this trend will continue to grow. Applications on NO-SQL databases that were mostly POCs are expected to move into deployment phase. The most popular No-SQL databases like MongoDB, Cassandra will continue to be implemented by more vendors. Also graph databases like Neo4j will gain more market.

### ****5. Self-Service Big Data applications will emerge****

Applications that simplify data cleaning, data preparation and data exploration tasks is expected to increase. Tools like Tableau with Hadoop has seen increasing popularity in that last 2 years. These products will greatly minimize the effort of the end-users. Companies like Informatics have already shown innovations in this frontier. We can see more such products and more companies working towards such self-service solutions.

To summarize, Big Data is still very much on rise with more adoptions and more applications of the existing technologies and launch of newer solutions related to Big Data security, Cloud integrations, data mining etc.