How Big is Big Data

Big data continues to be the topic of much discussion and hype, but “big” is really a red herring. Oil companies, telecommunications companies, and other data-centric industries have had huge datasets for a long time. And as storage capacity continues to expand, today’s “big” is certainly tomorrow’s “medium” and next week’s “small.” The most meaningful definition I’ve heard: “big data” is when the size of the data itself becomes part of the problem.

We’re discussing data problems ranging from gigabytes to zeta bytes of data. At some point, traditional techniques for working with data becomes inadequate. What are we trying to do with data that’s different? We’re trying to build information platforms or dataspaces. Information platforms are similar to traditional data warehouses, but different. They are designed for capturing and understanding the data rather than for traditional need for immediate analysis and reporting. They accept variety of data formats, including the apparently incomprehensible ones, and their structures (schemas) evolve as the understanding of the data improves.

Most of the organizations that have built data platforms have found it necessary to go beyond the relational database model. Traditional relational database systems cease to be effective at the stream of such volume(scale). Managing sharing and replication across a stack of database servers is difficult, slow and costly. The need to define a schema in advance conflicts with reality of numerous, unstructured data sources, in which you are not aware what’s important until the data is analyzed.

Data capture and storage is only part of forming a data platform, though. Data is only useful if you can do something with it, and enormous datasets present computational problems. Machine learning is another essential tool for the data scientist. We now expect web and mobile applications to incorporate recommendation engines, and building a recommendation engine is an artificial intelligence problem.

Building statistical models plays yet another important role in any data analysis. Statistics is the “grammar of data science.” It is crucial to “making data speak coherently.” We’ve all heard the joke that eating pickles causes death, because everyone who dies has eaten pickles. That joke doesn’t work if you understand what correlation means. More to the point, it’s easy to notice that one advertisement for R in a Nutshell generated 2 percent more conversions than another. But it takes statistics to know whether this difference is significant, or just a random fluctuation. Data science isn’t just about the existence of data, or making speculation about what that data might mean; it’s about testing hypotheses and making sure that the conclusions drawn from the data are valid.

Thus Statistics plays a vital role in everything from traditional business intelligence to contemporary data analytics. It isn’t superseded by newer techniques from machine learning and other disciplines; it complements them. That’s where different technologies evolve in addressing the challenges in modern day data analytics including Big Data.