What *Big Data* means to me

Philip E Bourne

Across the world, there is much talk about *Big Data* and how it is going to change the way we do business. Conferences, workshops, and funding initiatives in *Big Data* attract hundreds to thousands of stakeholders. Interestingly, no two stakeholders would be likely to define *Big Data* in exactly the same way. An obvious definition is the appropriate description, integration and sustainability of very large datasets generated by high throughput experiments. Another equally accurate and obvious definition is a large collection of small disparate, unstructured datasets which, taken together, can be analyzed to find unusual trends. I would like to offer a somewhat different viewpoint, which expands and encompasses these two definitions. For me, *Big Data* represents the emergence of the digital enterprise—the ability for an organization to take full advantage of its digital assets—which collectively can be described as large amounts of data and more. In other words, lots of data is a sign that something more is going on that has yet to be fully articulated and understood.

Pertinent to this audience is the health center, the biomedical research institution, the commercial health-related entity, or the government agency as the 'enterprise,' and digital assets that range from basic research data, to electronic health records, to courses online, to proposals submitted, to administrative data, and so on. Today, which institutions would claim to take full advantage of their digital assets? The answer, I would suggest, is few, if any, yet the need to do so to remain competitive is becoming increasingly recognized. Hence the interest in *Big Data*.

To be fair, the value and opportunities associated with digital assets take time to be appreciated in any type of enterprise. For example, until relatively recently, universities and their academic health centers were analog—patient records on paper, courses taught with slides or overheads, course notes printed, research data on shelves in notebooks, admission applications kept in endless filing cabinets, and so on. Now all of that content is, or soon will be, purely digital. The problem is that, for the most part, these data are simply electronic versions of what was maintained in hard copy. *Healthcare centers and research institutions are only now beginning to leverage the true power of the digital medium*. This time-to-adopt is not new in business; the music industry, the book and newspaper industry, the manufacturing industry, etc—and now the scholarly publishing industry—all responded slowly to the new digital reality, and when change accelerated, old business practices died and new ones emerged.

Becoming a Digital Enterprise represents a challenge for many institutions since their organizational structures make widespread data integration and analysis difficult. Research, clinical activities, hospital services, education, and administrative services are somewhat siloed, and, in many organizations, each silo maintains its own separate organizational (and sometimes duplicated) data and information infrastructure. Central services may provide computer networking and email accounts, but the needs of hospitals, clinics, schools, departments, colleges, commercial partners, government agencies, or whatever are different. Breaking down the silos takes vision, leadership, and significant resources.

The advantages of the truly digital enterprise are hard to imagine; that is just the point. Having said that, I imagine that you, like me, have said many times, "if only I could…." Here is an example of one of those situations where the stakeholders in the Digital Enterprise are Jane, an MD, PhD student; Jack a graduate student in chemistry; and Joe Smith, a neuroscience professor.

Jane scores extremely well in parts of her graduate online neurology class. Neurology professors, whose research profiles are online and well described, are automatically notified of Jane's potential based on a computer analysis of her scores against the background interests of the neuroscience professors. Consequently, Professor Smith interviews Jane and offers her a research rotation. During the rotation, she enters details of her experiments related to understanding a widespread neurodegenerative disease in an online laboratory notebook kept in a shared online research space—an institutional resource where stakeholders provide metadata, including access rights and provenance beyond that available in a commercial offering. According to Jane's preferences, the underlying computer system may automatically bring to her attention Jack, a graduate student in the chemistry department whose notebook shows that he is working on using bacteria for purposes of toxic waste cleanup. Why the connection? They reference the same gene a number of times in their notes, which is of interest to two very different disciplines—neurology and environmental sciences. In the analog academic health center, they would never have discovered each other, but thanks to the Digital Enterprise, pooled knowledge can lead to a distinct advantage. The collaboration results in the discovery of a homologous human gene product as a putative target in treating the neurodegenerative disorder. A new chemical entity is developed and patented. Accordingly, by automatically matching details of the innovation with biotech companies worldwide that might have potential interest, a licensee is found. The licensee hires Jack to continue working on the project. Jane joins Professor Smith's laboratory, and he hires another student using the revenue from the license. The research continues and leads to a federal grant award. The students are employed, further research is supported, and, in time, societal benefit arises from the technology.

A hypothetical example perhaps, but there are no technical reasons why this example cannot be realized—the discovery and analysis tools exist, and data can be sufficiently well described to make the needed connections. Missing are the organizational infrastructure, data sharing policies, tangible incentives to those who share their research products before publication, and the leaders to make this hypothetical example a reality. Continued interest in *Big Data* is the catalyst that takes us to a new and exciting place that comes to be known as the Digital Enterprise. It is our role to do our part in creating the environment in which it can become a reality.

Editor's choice
Scan to access more
free content

Correspondence to Dr Philip E Bourne, Innovation and Industry Alliances, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA; pbourne@ucsd.edu