Data Intensive Health care

Rajendra Akerkar

Technomathematics Research Foundation
TMRF Report 03- 2012

Data intensive Healthcare

To promote innovation and increase efficiency in the healthcare sector

TMRF-report-03-2012

TMRF White Paper

By Rajendra Akerkar

March 2012

Contents

Section 1

Summary

Big data is becoming the new frontier of information management given the amount of data today's systems are generating and consuming. It has driven the need for technological infrastructure and tools that can capture, store, analyze and visualize vast amounts of disparate structured and unstructured data. In order to bring about the revolution in healthcare that modern IT promises, there are legal, technical and societal barriers that must be overcome. In this white paper we deal with the concept of Big Data as applied to healthcare, or Big Data Healthcare, and the developments it may bring. We then consider some of the current major hurdles to its acceptance in standard healthcare.

Section 2

Introduction

The amount of data in today's world has been exploding, and analysing large data sets, so-called big data, has become a strategic basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus. Businesses capture massive amount of information about their customers, suppliers, and operations. Millions of networked sensors are being embedded in the physical world in devices such as mobile phones, smart energy meters, automobiles, and industrial machines that sense, create, and communicate data in the age of the Internet of Things. Social media sites, smartphones, and other consumer devices including PCs and laptops have allowed billions of individuals around the world to contribute to the amount of big data available. And the expanding volume of multimedia content has played a vital role in the exponential growth in the amount of big data.

In near future, Big Data is expected to be deployed around existing advertising solutions and also around social networks. Big Data is also being used in various scientific research handling very large volumes of data (physics, weather, genetics, etc.) and trying to find patterns into those data sets. Moreover, Big Data is expected to offer new opportunities in all retail activities by leveraging the real time coming from smartphones, sensors and the several video formats. No doubt, Big Data should impact all sectors, including more industrial sectors, notably by fusing data from different departments of a same enterprise with external data.

Big data is more than just a matter of size; it is an opportunity to find insights in new and emerging markets based on data analysis, to make processes, businesses and public services more agile, and to answer questions that were previously considered beyond our reach. But while big data allows so many new opportunities, there are also many socio-economic, ethical, legal and privacy aspects associated to the sharing and reuse of data that must be analysed and taken into account.

Section 3

Big Data

Big Data is primarily linked with two ideas: data storage and data analysis. So the key quest is how Big Data is different from conventional data processing practices. For primary insight as to the answer to this question one need look no further than the term big data. Here, "Big" implies importance, complexity and challenge. Moreover the term "Big" also incites quantification and thus it makes difficult to provide a single definition.

The widely known definition is that included in a Meta report from 2001. Gartner proposed a 3- fold definition incorporating the "three Vs":

- Volume – large amounts of data generated;
- Velocity – frequency and speed of which data are generated, captured and shared; and
- Variety – diversity of data types and formats from various sources.

The Oracle definition is focused upon infrastructure highlighting a set of technologies like NoSQL, HDFS, Hadoop, R and relational databases. Though this definition is undoubtedly applied than others it equally lacks quantification. Intel links big data to organisations, for instance "generating a median of 300 terabytes (TB) of data weekly". Intel expresses big data through quantifying the experiences of its business partners.

Microsoft provides a concise definition: "Big data is the term increasingly used to describe the process of applying serious computing power - the latest in machine learning and artificial intelligence - to seriously massive and often highly complex sets of information".

Big Data is the collection of large volumes of varied information, used to extend our understanding of the environment, medicine, science, business and human experience.

Despite the variety and differences existing within each of the aforesaid definitions there are certain elements of similarity. Especially most definitions make at least one of the following affirmations:

1. Size: the volume of the datasets is indispensable feature.
2. Complexity: the structure, performance and mutations of the datasets are important features.
3. Technologies: the tools and techniques used to process a huge dataset are crucial features.

In Big Data Healthcare the data volume is increasing and so is data velocity as continuous monitoring technology becomes ever cheaper. With so many types of tests, and the existing wide range of medical hardware and personalised monitoring devices healthcare data could not be more varied, yet data from this variety of sources must be combined for processing to reap the expected rewards. In healthcare, veracity of data is of paramount importance, requiring careful data curation and standardisation efforts but at the same time seeming to be in opposition to the enforcement of privacy rights. Extracting value out of big healthcare data for all its beneficiaries (clinicians, clinical researchers, pharmaceutical companies, healthcare policy-makers, etc.) demands significant innovations in data discovery, transparency and open-ness, explanation and provenance, summarisation and visualisation, and will constitute a major step towards the coveted democratisation of data analytics.

Section 4

Opportunities for Big Data in Healthcare

Big Data represents a new approach to analytics. It does not yet have a large or significant footprint world-wide. However, the continuing digitization of health records together with the interoperable electronic health record (EHR), presents new opportunities to investigate a myriad of clinical and administrative questions. There is potential to layer Big Data applications, in a privacy-protective manner, on top of the foundational health IT infrastructure to derive value that might not otherwise be found. What follows are some innovative ideas and solutions.

- Clinical decision support – Big Data technologies that sift through large amounts of data, understand, categorize and learn from it, and then predict outcomes or recommend alternative treatments to clinicians and patients at the point of care.
- Personalized care – Predictive data mining or analytic solutions that can leverage personalized care (e.g., genomic DNA sequence for cancer care) in real time to highlight best practice treatments to patients. These solutions may offer early detection and diagnosis before a patient develops disease symptoms.
- Public and population health – Big Data solutions that can mine web-based and social media data to predict flu outbreaks based on consumers' search, social content and query activity. Big Data solutions can also support clinicians and epidemiologists performing analyses across patient populations and care venues to help identify disease trends.
- Clinical operations – Big Data can support initiatives such as wait-time management, where it can mine large amounts of historical and unstructured data, look for patterns and model various scenarios to predict events that may affect wait times before they actually happen.
- Policy, financial and administrative – Big Data can support decision makers by integrating and analyzing data related to key performance indicators.

Section 5

Characteristics of Big Data in Healthcare

While Big Data is rather new concept, some components being leveraged by Big Data have existed for several years (e.g., data integration software that moves data, approaches to process and analyze text based data, and content management and document management for managing unstructured data). The size and complexity of Big Data makes it difficult to use traditional database management and data processing tools. This issue is being compounded by the growth in data generated by consumer, enterprise medical devices and digitized patient records where the majority of data are in different formats. Data are being created in much shorter cycles, from hours to milliseconds. There is also a trend started to create larger datasets by combining smaller datasets so that data correlations can be discovered.

Nonetheless, the arrival of Big Data in the enterprise software space has created some confusion as business leaders try to understand the differences between it and traditional data warehousing and business intelligence (DW/BI) tools. There are important distinctions and sufficient differentiating value between Big Data and DW/BI systems which make Big Data unique.

Specific to health care, the types of data anticipated to be available for use by Big Data include:

- **Genomic data** – Represents significant amounts of new gene sequencing data being made available through new investments, Big Data capabilities and business models.
- **Streamed data** – Home monitoring, tele-health, handheld and sensor-based wireless and smart devices are new data sources and types. They represent considerable amounts of real time data available for use by the health system.
- **Web and social media-based data** – Web-based data comes from Google and other search engines, consumer use of the Internet, as well as data from social networking sites.
- **Health publication and clinical reference data** – This includes text-based publications (clinical research and medical reference material) and clinical text based reference practice guidelines and health product (e.g., drug information) data.
- **Clinical data** – Eighty per cent of health data is unstructured as documents, images, clinical or transcribed notes. These semi-structured to unstructured clinical records and documents represent new data sources.

- **Business, administrative and external data** – Data which earlier has not been linked, such as commercial, scheduling, administrative, external and other non-clinical and non-health data.

Note that while there are many sources of Big Data within the health sector, it is impractical to assume that all data can be put to use for Big Data due to a range of governance, privacy, operational and technical considerations.

Moreover, Big Data functions are distinctive from traditional analytic methods because they:

- support an experimental type of analytics, whereas, traditional DW/BI and statistical analyses are based on answering known questions or hypotheses;
- manage open ended "how and why" kind questions, while Business Intelligence tools are designed to query peculiar "what and where";
- process unstructured data to find patterns whereas DW systems process structured, related and mostly aggregated data;
- process, generate and index large sets of data, handling the complexities of network communication, parallel programming and fault tolerance;
- process large datasets, in a distributed environment, across clusters of computers designed to scale up from single servers to thousands of machines, each having its own local computation and storage;
- deliver results faster than Business Intelligence systems, even in real time; and
- scale to petabytes of data;
- systematically mine and flag data that is relevant for other uses and for further analytics by conventional analytics tools;
- leverage analytic concepts such as geo-mapping, data mining and predictive modelling to help with disease outbreak monitoring and forecast modelling.

Section 6

The Value of Health Big Data repositories

It is true that in order to gain this growing value there is the need not only for clinicians and researchers to acquire Big Data analytics skills and services, but also to develop a framework for data repositories which adheres to international standards for the preservation of data, sets common storage protocols and metadata, protects the integrity of data, establishes rules for different levels of access and defines common rules that facilitate the combining of datasets and improve interoperability. These frameworks could provide some of today's data protection rules and procedures invalid.

Such repositories can allow testing of the prior knowledge of clinicians, who identify the data features deemed to be key for specifying a patient's treatment, versus the correlations that big data chewing may highlight, possibly leading to further knowledge discovery.

Definitely, by statistically and semantically reasoning on the data, existing pathophysiological patterns may be divulged and recorded as a first step in a fractional factorial and model driven research process supporting physicians in their iterative and interactive quest to discovering new knowledge.

The objectives are: to be able to provide model-driven patient-specific predictions and simulations and consequent optimised personalised clinical workflows, to allow for advanced similarity search among patients, such that clinicians can find "the patient like mine", and to get support through risk stratification and outcome analysis. Ultimately it is expected that certain patho-physiological patterns can be detected, refined and made available to other clinicians and researchers in the form of pattern libraries. These pattern libraries, identifying homogenous groupings among patients and model similarities, could be shared between researchers and clinicians to al-low for data intensive pathophysiological diagnoses. Similar to the above is the ability to revolutionise health communications by making it possible, on the basis of semantically advanced repositories, to use social media among patients aware of sharing highly similar conditions, empowering them to bridge the gap with the clinicians, mainly in the case of paediatric patients and their parents.

Section 7

Challenges in Data Intensive Healthcare

The healthcare domain is known for its ontological complexity, variety of medical data standards and variable data quality, there are also healthcare cultural changes that will be required to fully capitalise on Big Data Healthcare as a movement. This especially revolves around the need for medical staff to be educated using examples of its success. Clinicians will need to understand and be persuaded of the potential of Big Data Healthcare, as well as its limits. To maximise the acceptability and utility of Big Data Healthcare solutions in the clinical arena, clinicians should therefore be involved in the development of these solutions from their inception. Thus, a clinically led development ideology will ensure that technical know-how and innovation translates into clinically useful tools that fit more naturally into clinical workflows. Clinicians and engineers must work together to translate and extend their existing and advanced data analysis technology, that is the clinically trained human mind, into targeted big data analytical approaches that will achieve clinically useful outputs. Although engineers and clinicians have long collaborated successfully, development work in Big Data Healthcare will require particularly intimate reciprocal under-standing by each disciplinary culture of the other. It will re-quire further cultural development in both areas.

This should be achieved at a grass-roots level by a greater emphasis of clinical informatics in medical curriculums and similar exposure of developing bioinformatics engineers to the unique challenges that medical bioinformatics faces.

At the same time, decision support systems will need to be more than black boxes but be capable of showing clinicians why they advocate particular courses of action, providing the necessary assurance that advice is based on sound principles.

Diffusion of medical technologies is necessarily a lengthy process. This means that these processes of education and culture shift should begin now, preparing new generations of clinicians and bioinformatics engineers for forthcoming data intensive methodologies and collaboration.

The following points will need therefore to be fleshed out:

- Management of big data
- Seamless end-to-end big data curation
- Data discovery, profiling, extraction, cleaning, integration, analysis, visualisation, summarisation, explanation

- Use of big data
- Appropriate use of big data – avoiding over-reliance
- Responsible use of automated techniques
- Communicating big data findings to patients
- Integrating data analytics into clinical workflows
- Data (clinical) scientist

One concern is that data generated in the routine care of patients may be limited in its use for analytical purposes. For example, such data may be inaccurate or incomplete. It may be transformed in ways that undermine its meaning (e.g., coding for billing priorities). It may exhibit the well-known statistical phenomenon of censoring, i.e., the first instance of disease in record may not be when it was first manifested (left censoring) or the data source may not cover a sufficiently long time interval (right censoring). Data may also incompletely adhere to well-known standards, which makes combining it from different sources more difficult. Finally, clinical data mostly only allows observational and not experimental studies, thus raising issues of cause-and-effect of findings discovered.

There are many other challenges around analytics and big data. Sometimes research questions asked of the data tend to be driven by what can be answered, as opposed to prospective hypotheses. Moreover, the data are not always as objective as we might like, and that "bigger" is not necessarily better.

Finally, there are ethical concerns over how the data of individuals is used, the means by which it is collected, and the possible divide between those who have access to data and those who do not.

Policies on the ethical use of these data types and sources may need to be established. Legal implications for patient's and clinician's use or non-use of Big Data should be assessed to understand if there are any new implications created by Big Data.
o **Ethical** – Big Data represents a new capability that allows stakeholders to answer new questions or needs for knowledge that were not anticipated when the original data collection methods were put in place. While this represents a huge opportunity for health research, ethics policies should be reviewed to assess whether they allow and enable these experimental analytics.
o **Legal** – Legal and liability implications of leveraging Big Data technologies such as genomics, natural language processing and artificial intelligence need to be considered. Big Data solutions provide access to health knowledge bases for various stakeholders (e.g., patient-consumers) that can be searched instantly and displayed back as medical advice and treatment recommendations. On the one hand, Big Data may engage stakeholders by providing instant access to health information regardless of location. On the other hand, it may fuel self-diagnosis, which may prevent consumers from seeking proper medical attention or result in misdiagnosis.

Section 8

Conclusions

Clearly there is great promise ahead for healthcare driven by data analytics. The growing quantity of clinical and research data, along with methods to analyze and put it to use, can lead to improve personal health, healthcare delivery, and biomedical research. However, there is also a continued need to improve the completeness and quality of data as well as conduct research to demonstrate how to best apply it to solve real-world problems. In addition, human expertise, including in informatics, will be required to optimally carry out such work.

Some important points:

1. Big Data Analytic platforms will examine data from multiple sources, such as clinical records, genomic data, financial systems, and administrative systems
2. Analytics is necessary to transform data to information and knowledge
3. Accountable care organizations and other new models of healthcare delivery will rely heavily on analytics to analyze financial and clinical data
4. There is a great demand for skilled data analysts in healthcare; expertise in informatics will be important for such individuals.

Finally, there is a continued need to improve the completeness and quality of data as well as conduct research to demonstrate how to best apply it to solve real-world problems.

Section 9

Bibliography

1) Davenport TH and Harris JG, Competing on Analytics: The New Science of Winning. 2007, Cambridge, MA: Harvard Business School Press.
2) Adams J and Klein J. Business Intelligence and Analytics in Health Care - A Primer. 2011, The Advisory Board Company: Washington, DC.
3) Zikopoulos P, Eaton C, deRoos D, Deutsch T, and Lapis G. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. 2011, New York, NY: McGraw-Hill.
4) Kuperman GJ. Health-information exchange: why are we doing it, and what are we doing? Journal of the American Medical Informatics Association, 2011. 18: 678-682.