Connecting big data to big insights

Rasu B. Shrestha, MD, MBA

Dr. Shrestha is the Chief Innovation Officer at University of Pittsburgh Medical Center, Pittsburgh, PA, and Executive Vice President of UPMC Enterprises. He is also Chair of the RSNA Informatics Scientific Program Committee; a Founding Member of the Executive Advisory Program, GE Healthcare; a member of the advisory boards of KLAS Research and Peer60; a member of the Board of Directors of the Society for Imaging Informatics in Medicine; a member of the boards of Pittsburgh Dataworks and Omnyx Inc., and a member of the Applied Radiology editorial board.

"In God we trust, all others bring data."
—W. Edwards Deming

Today, it seems we are data rich and information poor. A Harvard Business Review paper[1] pointed out recently that a total of 2.5 quintillion terabytes of data were generated every day in 2012 alone, and that it is estimated that as much data is now generated in two days as was created from the dawn of civilization. Furthermore, it is estimated that 90% of all the data in the world has been generated during just the last two years.[2]

But how much of this data is really meaningful, useful or actionable? As the healthcare industry marches on from analog to digital, we are seeing a massive proliferation of data sources, often siloed, often not talking to each other, and almost always created to address only a defined set of use cases. As clinicians, we often find ourselves playing the role of detective, navigating from one clinical system to another, piecing information together around our patients. While we rejoice that the digital era is upon us, we should be distraught that we really have not been able to capitalize on the sheer power of the data that we have all around us. This article is a wake-up call—a cry for us as a healthcare community to comprehend the power of data and to actively seek the insights that can be garnered from the right approaches to big data technologies.

With healthcare's focus currently on value-based care paradigms, on quality and on efficiency coupled with efficacy, data-driven decisions are of paramount importance. But there's a saying that "a frog in a well cannot conceive of the ocean." Intelligent decisions are best made with enough data to give us rich context, a fuller view of all the parameters, and all the possibilities.

However, how do we not drown in all this data we're generating? How do we stay afloat on, swim in and surf the tremendous power of this valuable resource?

A whole lot of data going on

At my institution, the University of Pittsburgh Medical Center, we currently have about 8.9 petabytes of data in real-time storage, and this amount is doubling almost every 18 months. The industry is talking not just about petabytes, but about exabytes and zettabytes of data. Data worldwide is projected to explode and reach 35 zettabytes by 2020---a 44-fold increase from 2009.[3] If you're keeping track, a zettabyte is equal to 1 sextillion bytes, which is exactly 1 million petabytes. To help put that into perspective, just 2 petabytes is equivalent to all the data stored in all U.S. academic research libraries.

This article is a wake-up call — a cry for us as a healthcare community to comprehend the power of data and to actively seek the insights that can be garnered from the right approaches to big data technologies.

That's a lot of data.

Big data technologies present a fresh opportunity in healthcare to bring previously unfathomable amounts of data to life, such that we can transform the data to valuable insights and put the data to work for us. Big data technologies will illustrate novel ways to measure improvements in quality care and patient outcomes and will drive efficiencies in clinical workflow with new insights that we did not know were possible to attain.

Big data demystified

Gartner's original definition of big data focused on the three Vs:[4] high-*volume*, high-*velocity* and high-*variety* information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

As discussed earlier, the volume of data continues to increase astronomically. Volume typically addresses "data at rest." The velocity of data adds an interesting dimension, addressing "data in motion." Data is coming to us at increasing speed, number and frequency of transactions. The velocity of big data needs to be understood, prioritized and synced up with our strategic needs both clinically and operationally. It is said that variety is the spice of life, and this is true for data, too. Variety addresses "data in many forms." We today have at our disposal an amazing array of data types, and across industries, amazing insights are being derived from text, geo-locations, log files, sensor and human generated data of various sorts. Data is now a rich organizational asset – a healthcare entity's "natural resource" that is waiting to be tapped.

Data pundits are also focusing on additional V's in defining big data. One is *veracity*. The argument is that big data comes in such a diverse range that quality and security become big focal points for *verification* of the data. Veracity, hence, addresses "data in doubt," the uncertainties due to data inconsistencies, ambiguities and latency. As we extract, transform and load (ETL) big data into enterprise data warehouses, verification of quality and compliance becomes an important aspect of addressing the veracity of big data.

Value is also sometimes cited as an additional V. Deriving value out of insightful analytics that could impact care processes and outcomes is an important, and often the most critical, goal of efforts related to big data. The promise of the value, especially in healthcare, is one of a tremendous wave of innovation, progress, growth and new care models,from the insights derived out of big data.

The term "big data" refers to datasets whose size is well beyond the ability of traditional database software tools to capture, store, manage and analyze.[5] The term also refers not just to the data itself, but also to an emerging set of technologies being developed to handle this massive collection of data stores. Traditionally, we have been using relational database management systems (eg, SQL) and desktop statistics and visualization packages to analyze and synthesize data. But as data has grown in every conceivable parameter, these relational database management systems have proven to be inadequate.

Dealing efficiently with big data instead requires massively parallel software running on tens, hundreds, and even thousands of servers. While relational databases perform transaction update functions very well, they struggle with the efficiency and efficacy of certain tasks key to big data management. Big data technologies (eg, NoSQL) are able to scale much better to very large sizes and are able to handle a wider array of data formats, including unstructured data for the types of searches we have now become used to using modern day search engines. Online search and social networking companies such as Google, Facebook and Amazon were the first to move away from relational databases, and they have not looked back. There has since been tremendous development in these technologies, with platforms such as the Hadoop file system, MapReduce programming language, and associated databases such as Cassandra and HBase. Healthcare has notably been slow to adopt these big data technologies, and the opportunities to leapfrog the generation of meaningful insights from the data we have are tremendous.

Cloudy with a chance of big insights

We have established that we are seeing a massive tsunami of data across the board. At the same time, however, the massive computational power required to crunch big data is now becoming increasingly accessible through cloud technologies. Hence, what we have today are massive computational capabilities, made more easily accessible, deployable and doable because of the cloud. This explosive combination offers us the potential to transform healthcare in ways we have not even begun to dream about. Big data enabled by cloud technologies could provide us with new insights –clinically, operationally and in research---even as we focus on diagnostics across complex and challenging chronic illnesses and look at populations of patients in an increasingly dynamic and cost-conscious healthcare environment that is all about accountable care and value based care. Managing, analyzing, visualizing and extracting useful information is becoming increasing sophisticated yet doable. The road ahead is an exciting one, and we will only be limited in what we do with the data by the boundaries of our imagination.

Cloud computing is a natural evolution of the widespread adoption of virtualization, service-oriented architecture (SOA) and utility computing.[6] Cloud offers high computing capabilities to process massive amounts of data for things such as 3D imaging, so we need not be tethered to an expensive, high-end advanced visualization workstation in a corner of the hospital. The cloud empowers anywhere access to 3D views and more, across literally any device, and enables a radiologist to collaborate with a vascular surgeon, leveraging capabilities to compute and batch process algorithms on large volume sets located literally anywhere across the world.

Data as an asset

Data is an invaluable resource. In many ways, the very essence of the conversation around data management has shifted with the availability of big data tools and capabilities. The debate today is less about whether we can afford to store information, and more about whether we can actually afford to throw it away. With big data technologies now at our fingertips, and the cost of data storage having dropped dramatically over time, data in most instances should not be discarded. The focus today is moving from processing volumes of data that perhaps were just not previously practical to store to dealing with massive amounts of data at a time, detecting insightful metrics and responding quickly.

The ability to run deep analytic queries on huge volumes of structured and unstructured data is a big data challenge. It requires massive parallel-processing data warehouses and purpose-built appliances for deep analytics, as well as capabilities around natural language processing (NLP) that are continuing to be perfected. Big data isn't just about data at rest – it's about data that is also in motion. Streaming data represents an entirely different big data problem -- the ability to quickly analyze and act upon data while it's still moving. There has been much progress in this area, and the possibility of correlating data elements such as hours (or months) of live waveforms from the ICU with other types of data across the healthcare enterprise is an exciting one.

There is a merging of traditional and big data approaches to handling these data elements. If the traditional approach was structured and

repeatable analysis, the big data approach is one of iterative and exploratory analysis. Big data then delivers a fluid platform to enable *creative discovery*, and the user (eg, clinician, administrator or analyst) explores the facets and dimensions around the many ways intelligent insights could be asked or derived.

The opportunity at hand is to be able to scan these massive stores of data and connect them with other types of data that may be able to provide new insights and meaning. Correlating clinical data with cost, outcomes and performance data, and then tying these to evidence-based guidelines and clinical best practices, could reveal entirely new insights and opportunities to continue to push the needle forward with newer care models.

Gardeners of big data

Getting insights from data requires some level of ground work. Much like how a gardener sows his seeds, and cares for and nurtures his garden, managing data, especially at scale, requires some discipline and, arguably, a good deal of passion.

Managing data entails having disciplined methodologies around data integration, data governance, and data quality, security and information lifecycle management. Like gardeners, data stewards may need to do some weeding and pruning before data analysts and data scientists can start harvesting from data farms. The crops of data may yield insightful ingredients that cooks, professional or otherwise, may then want to conjure together into an appetizing and nutritious meal. With the right set of tools and capabilities, data analysts and data scientists can serve up the right capabilities for clinicians, administrators and technologists to glean truly meaningful insights. What's perhaps even more interesting is the movement towards self-service capabilities, where front-line users, such as clinicians, can have direct access to simple tools that can yield tasteful and actionable information visualizations with minimal effort.

What's the future of big data?

I believe the value of big data will rise exponentially, especially as ways to tame its veracity continue to be addressed alongside well-established methodologies to deal with its volume, variety and velocity.

"Big data" not only changes the tools we can use for predictive analytics, it also changes our entire way of thinking about knowledge extraction and interpretation. There's been much development in artificial intelligence, machine learning, deep learning and what is now being called cognitive computing. Machine learning is to big data as human learning is to life experience: We interpolate and extrapolate from vast past experiences to deal with specific unfamiliar situations. Machine learning with big data will duplicate this behavior, at massive scales. Big data coupled with the "pattern recognition at scale" capabilities of machine learning will allow us to program systems to automatically learn and to improve with experience, such as learning to predict which patients will respond best to which treatments, by analyzing multiple streams of data and experience at scale, often in real time. This will continue to allow us to develop algorithms that discover general conjectures and knowledge from specific data and experience, based on sound statistical and computational principles.

Indeed, the future ahead could look nothing like the past we left behind.

References

1. Shah, ND and Pathak J.Why health care may finally be ready for big data. *Harvard Business Review;* December 3, 2014. https://hbr.org/2014/12/why-health-care-may-finally-be-ready-for-big-data. Accessed Feb. 17, 2016.
2. Big Data, for better or worse: 90% of world's data generated over last two years. *Science Daily.* [Online] SINTEF, May 22, 2013. [Cited: February 10, 2016.] https://www.sciencedaily.com/releases/2013/05/130522085217.htm.
3. Gantz J, Reinsel D. The digital universe decade – are you ready? IDC, May 2010. http://www.emc.com/collateral/analyst-reports/idc-digital-universe-are-you-ready.pdf. Accessed Feb. 17, 2016.
4.Laney, D. *3-D data management: Controlling data volume,* velocity and variety. Meta Group (Gartner), 6 February 2001. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf. Accessed Feb. 17, 2016.
5. Company, McKinsey &. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, June 2011. http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation. Accessed Feb. 17, 2016.
6. Shrestha, RB. Webinar: Real world challenges and benefits of cloud technologies. *Society for Imaging Informatics in Medicine.* Jan. 28, 2016. http://siim.org/?page=web16_real_world.