"Too Little, Too Late": Big Data May Change It

Xiaoyan Zhang, Ph.D.

"Too little, too late" is a problem that many decision makers who want to use data-driven decision making processes face every day. They need actionable intelligence based on timely and reliable information now.

This is a challenge that well-trained data analysts have not been able to overcome. Following the modern statistical analysis principles, the operation of a survey based research project is so complex that months and years are needed to design, collect, and clean data. More time is then spent to process, analyze, and report the findings. The pre-designed questionnaire has another shortcoming; it can only answer the questions per its design. The result is "too little, too late" as dynamic human behavior and social, economic, and political conditions cannot be fully understood by periodic snap shots regardless of the elegant study design.

The coming of the Big Data era may provide a new way to approach this dilemma. In other words, it may change the way we solve problems and do business. Albert Einstein once said: "We can't solve problems by using the same kind of thinking we used when we created them".[1]

When discussing Big Data in relation to government decision making, Dr. Jeremy Wu, a former senior statistician at the U.S. Census Bureau, stated: "Big Data is a new and loosely defined term for large electronic datasets that may or may not be collected according to the structure and probability principles specified in the traditional statistical systems... However, some contain important information that has not been available before for decision- and policy- making, especially when they are appropriately integrated into government data sources."[2]

The characteristics of Big Data can be summarized into 4 Vs: Volume, Velocity, Variety, and Veracity.[3] In terms of Volume, "Since 2000, the amount of information the federal government captures has increased exponentially. In 2009, the U.S. Government produced 848 petabytes of data and U.S. healthcare data alone reached 150 exabytes. Five exabytes (10^18 gigabytes) of data would contain all words ever spoken by human beings on earth. At this rate, Big Data for U.S. healthcare will soon reach zetabyte (10^21 gigabytes) scale and soon yottabytes (10^24 gigabytes)."[4] In terms of Velocity, the accumulation of digital data has become a continuous stream and data accessibility a real time phenomenon. In terms of Variety, digital data is generated in different formats via many channels (Mobile devices,

Social Media, Videos, Chat, Genomics, Sensors, etc.). In terms of Veracity, the noise level of Big Data is high due to the high speed accumulation and variety of feeding channels, posting a major challenge to data analysts.

While mind-boggling Big Data makes information overload a fundamental challenge to government agencies at all levels, one thing is clear: "Big Data provides the opportunity to transform the business of government by providing greater insight at the point of impact and ultimately better serving the citizenry, society and the world."[5]

With the shrinking federal budget and increasing demand on comprehensive and timely data for decision support, the resources and time required by the traditional survey based data collection and analysis process may become un- affordable and lose its relevance. "Small evolutionary steps to tweak and tinker the edges of the current statistical systems built on knowledge and technologies grounded in the 1970s will simply not be adequate for the Big Data Revolution."[6]

The good news is that private industries have devoted significant investment in the past decade to develop Big Data processing and analysis technology, which made the harnessing of the value of Big Data feasible and affordable by government agencies today. Most relevant Big Data technologies include, Cloud-based computing (SaaS, IaaS, PaaS, which made access to large on-demand storage, computing power feasible and affordable), MapReduce (Parallel processing which made high speed and large scale data processing possible, i.e. Hadoop), SOA (Service Oriented Architecture, made seamlessly accessing separate applications across multiple operating platforms possible), Data Mash-ups (enabled integration of data from separately stored data sets in real-time over the Internet), and Data Visualization (made complex data to display interactively over the Internet in easily understood visual presentations such as maps, charts, and graphs).

The exciting new Big Data processing and presentation capability is of little value if the same old data collection, cleaning, and analytical processes remain in place. One must think "outside the box" to innovate and develop new methodologies and processes to tame Big Data opportunities. For example, with easily scalable data processing capacity at a significantly reduced cost, it is now possible to create an "analytic sandbox" outside the production data warehouse for analysts to explore different data set combinations and integration in real time (or near real time) to test hypotheses and analytic models.

"The exciting new Big Data processing and presentation capability is of little value if the same old data collection, cleaning, and analytical processes remain in place."

Only the proven solutions are included into the production environment. Further, the Enterprise Analytic Data Set (EADS) can now contain structured data, semi-structured, and unstructured data, greatly expanding the universe of answerable questions traditionally limited by individual survey questionnaire and sampling design. Finally, the improved speed and expanded scope will impact all four primary components of a model and score management decision support system: analytical data set inputs, model definitions, model validation and reporting, and model scoring output.

The challenge to the data analysts and decision makers is to embrace a new framework for processing huge volumes of unstructured and semi-structured data streams quickly with data mining models and analysis. Further, the on-line data processing results are then ready for real-time data visualization and decision support. In other words, one must learn the art of "sifting through the haystack to find the valuable needles" rather than shoot the target using a rifle equipped with a telescope. Big Data will not replace traditional survey and census. However, it will be applied where it fits and may enable us to do more with less and faster. For example, Big Data may provide approximation of the data used to be collected using traditional survey and census, therefore saves time, reduces the cost of survey, and burden of the respondents. Big Data can also create new information that has not been available before in current census or survey, thus provides additional insights for decision making.

With greater power comes a bigger responsibility to data analysts. Due to the Voracity and Volume of the data, researchers can be fooled by spurious correlations among unrelated variables. False information has grown much faster than real information in the Big Data era. The data processing algorithms, which are guided by theory and methodology, are still the foundation for generating meaningful intelligence in support of decision making. High speed data processing requires the development of intelligent algorithms to appropriately account for data uncertainty in order to reach the right conclusions faster. The improved information will enable decision makers to take prompt action to accelerate the use of successful solutions and intercede when there are problems, with the ability to observe the effect of the decisions within a few data cycles.

Is it possible to change "too little, too late" to "comprehensive and just in time"? Time will eventually tell. Big Data is not a silver bullet. However, it is definitely a new weapon. If we don't recognize the power and understand the limitations of Big Data, then we may miss out on great new opportunities.