DRC INTERVIEW WITH RANDY SHUMWAY, CEO, CICERO GROUP

Cicero INSTITUTE

The Chinese Government's central research arm (DRC) recently conducted an executive interview with Cicero Group CEO Randy Shumway. The interview, on how big data will transform the Chinese economy, marks the first interview with a foreigner ever published by the DRC. We are excited to make the full transcript available here in English.

With the rapid development of the Internet and information technology during the past few decades, the word "big data" has been attached to great importance. Meanwhile, an increasing number of countries and companies have recognized the significance of big data technology, and regarded it as a top priority from the strategic perspective.

In the Guiyang International Industry Summit held in May 2015, it was revealed that China is developing its national strategies and action plans by utilizing big data technology. On July 1, the General Office of the State Council issued a guidance on "Use of big data to improve public service and strengthen market supervision". The guidance calls for an expansion big data" application at the government level, which keeps up with the trend of the big data era. However, in the face of this huge opportunities in the big data industry, it is important to recognize what kind of impact big data may leave and how can we make the best use of the big data technology. For this reason, DRC interviewed Randy Shumway, the CEO of Cicero Group.

INTERVIEW TRANSCRIPT:[1]

Let's talk about "Big Data"…

DRC: What are the global trends you are seeing in big data?

Randy Shumway: Knowledge is power, and 21st century knowledge is data. For decades, businesses have overlooked their data in favor of traditional processes and expert intuition—all while the data universe expanded exponentially. During this time, a handful of businesses saw the true value and potential of data, and seized the opportunity; some of those businesses include Google, Baidu, Netflix, Alibaba, Amazon.com, Tencent, and Facebook—it has been the success of companies like these that have captured businesses' attention.

Today, the term big data is representative of the vast and ever-expanding quantities of data now accessible to businesses of all types and sizes. By 2020, data production will be 44 times greater than it was in 2009. Furthermore, in 2020, the digital universe will reach 44 zettabytes, or 44 trillion gigabytes, which is nearly as many digital bits as there are stars in the universe. With the introduction of more devices and sources (e.g., transactional data, sensors, click streams, mobile devices, etc.), greater Internet penetration rates in emerging economies, and increasing communication and technology infrastructures, big data is continuing to grow exponentially.

Estimates show that retailers using big data have the potential to increase their operating margin by as much as 60 percent

The three biggest players in the big data market in China right now are Baidu, Alibaba, and Tencent, also known as "BAT." Together, these three companies are generating an enormous amount of data, each in very specific areas—Baidu collects user search data, Alibaba collects the transactional and credit card data, and Tencent collects the social data. The potential for cross-industry application here is huge—from customer cross/upselling, to refined campaign targeting, to identifying counterfeit products, to reducing fraudulent activity. We are expecting to see some very sophisticated models and customer-centric business operations coming out of BAT, as well as the rest of China, in the coming years thanks to the power of big data.

When we look to the rest of the world, organizations are flocking to big data in hopes of uncovering golden nuggets of insight that provides a unique competitive advantage in the marketplace. This global migration toward big data is absolutely justified, especially when there are continuous big data success stories. Unsurprisingly, Cicero is very optimistic about the role big data can play in helping drive a business's competitive strategy. However, we also believe that a balanced approach to big data is necessary, as big data does have its limitations.

For instance, big data, no matter the size, only provides us with unique insights regarding patterns within the

[1] Transcript can also be read online at http://www.DRC.com.cn/DRC.Channel.Web/gylt/2015/index08.aspx

data. Such patterns can tell us how someone is likely to act, which ads are likely to be opened, which individuals are likely to succeed or fail, and other behavioral actions. However, big data does not give us the whole story, specifically the individual's story—what the experience was like, how they perceived the brand, and if they are likely to return. Big data tells us "what" but does not tell us "why." Any organization must understand both the "what" and the "why" to build the most innovative and disruptive strategies.

For us at Cicero Group, we believe the individual's perspective should never be lost at the expense of big data engagements. As such, we complement our big data initiatives with what refer to as "small data." Small data comes in the form of surveys, focus groups, in-depth interviews, ethnographic observations, mystery shopping and the like. Together, small and big data provide a 360 view of the business objective, which allows us to provide actionable and truly accurate insights and strategies for our clients.

DRC: What industries or sectors are most benefiting from big data and how?

Randy Shumway: The benefits of big data were first realized with online companies like Google, Facebook, and Netflix. Google's search engine algorithm and targeted advertising system were primary drivers of the company's success. Similarly, Netflix used unstructured social media data and web analytics to guide the creation and acquisition of new, proprietary content that gave it an edge over other emerging online streaming providers. However, in the past several years, the benefits of big data have since grown into nearly every sector, transcending industry boundaries by optimizing business functions that are essential for every industry.

One such industry is retail. In fact, estimates show that retailers using big data have the potential to increase their operating margin by as much as 60 percent. One of the more popular big data examples in retail is Target. In 2012, Target developed an algorithm to detect changes in shopping patterns that were indicative of pregnancy. The algorithm was designed to send coupons for diapers, formula, clothes, and other related items to expecting mothers. The algorithm was so accurate and timely that consumers actually began voicing privacy concerns. As a result, Target began including greater varieties of coupons in an attempt to make their offerings appear random. It is data driven insights like this that are transforming the way the retail industry competes today.

Yet, big data is not just benefiting the retail industry. Almost all businesses today generate what we refer to as 'data exhaust,' or have the potential to. As such, we are seeing big data adoption rates increase exponentially across an array of industries, of which we believe the following stand to gain the most from big data in the coming years: healthcare, banking, telecommunications, utilities, retail, banking, insurance, energy, and travel.

DRC: What public policy issues are being positively affected by big data?

Randy Shumway: Using big data to positively inform public policy and improve social impact initiatives is an area that gets me very excited. In fact, just this past year, Cicero Group launched its own social impact department to help governments and non-profits develop measurement and evaluation capabilities that will increase transparency and improve services. Although the marriage between big data and public policy is new, we believe there is real potential for significant social good—from predicting epidemics, to decreasing air pollution, to reducing crime, to waste management, to optimizing traffic patterns, to reducing water pollution, to increasing energy efficiency, to so much more.

One of the things that have allowed big data to play a larger role in shaping public policy issues is the emergence of the Internet of Things (IoT). The IoT is a term that refers to the wide varieties of data that everyday products are now capturing and collecting. Much of this IoT data is coming from sensors, which are capable of monitoring countless data metrics—from traffic speed, to weather patterns, to air pollution, to water pressure, to machinery operations, to chemical compositions, to infinitely more. As a result, we now have consistent, accurate, and measurable data on events like never before, and it's being produced on an unprecedented scale.

Let me give just a few examples of how big data is helping to create better public policies, and in turn smarter cities. In both Stockholm and San Antonio, city transportation officials have begun leveraging big data to monitor and calibrate traffic lights to improve traffic flow and reduce congestion. The city of San Antonio estimates that since the program started they have saved over $2 billion in lost productivity by reducing the amount of time that people are caught in traffic and waiting at stop lights.

In Jerusalem the city is using sensors placed on their city water infrastructure to monitor water quality and flow as water is distributed to households. This allows them to measure water pollution or contamination that may be occurring at various points within the system. In addition, it allows rapid pinpointing of leaks or breakages so that they can be fixed before major damage occurs.

The cities of both Amsterdam and Nice, France are using sensors to track air quality data to help measure pollution sources and determine the effect of specific public transport policies aimed at improving air quality. In the United States they are using satellite image data to help monitor forest fires. Similar technology is being looked at to monitor large scale air pollution in regions around the world.

In Barcelona, the city has kicked-off over 80 different projects that incorporate big data, from public transit optimization, to efficiency around water storage and use, to energy allocation. By monitoring moisture content in soils, the city is able to dynamically manage watering of public parks, allowing the conservation of water resources and saving the city money.

Although the marriage between big data and public policy is new, we believe there is real potential for significant social good

There are other applications as well. Big data is being used in the public health sector. In both Canada and the United States medical monitors placed in patient homes are being used to help provide care to patients in remote areas. By daily monitoring patient data, doctors can address health issues as they arise in a patient, rather than waiting for the patient to take the time to come in for an appointment. There are also public health applications as well, looking at massive populations and health trends or epidemics.

Interestingly, some of this remote data gathering technology was used following the Sichuan earthquake, where doctors in larger metropolitan areas were able to help diagnose and treat patients in remote or isolated villages based on data coming in from remote monitoring equipment. Now that is maybe getting a little bit away from the idea of big data, in that it is also looking at very specific data about one particular patient, but it gives you a sense of what is happening in the IoT space and how we are gathering data in new and interesting ways.

There are really some amazing opportunities in public policy and other areas for taking advantage of data, data gathering, and big data to help improve society and the way we live.

DRC: Explain to me what is referred to as the "3 V's" of big data.

Randy Shumway: There is a common perception among many that big data is all about size. However, size is only part of the equation. The "3 V's" stand for volume, variety, and velocity, and are the defining characteristics of what separate big data from traditional data (i.e. relational databases). Volume refers to the scale of the data (i.e. size). Big data is inherently large, often scaled to the range of Petabytes, which can present challenges for businesses trying to store or process big data without the necessary infrastructure. Velocity refers to speed.

Big data can stream to businesses at a high rate, often requiring unique IT solutions to capture, manage, and assess the data in real time. Variety refers to the varying types of data collected. Unlike traditional data, big data comes from collections of disperse, unstructured data sources (e.g., text, numbers, audio, etc.), which requires robust processes and algorithms for proper interpretation.

DRC: What are the different sources of big data?

Randy Shumway: Simply put, big data is everywhere and its sources are constantly evolving and growing with the advent of new technologies, platforms, and processes. The individual sources of big data are truly limitless to the imagination. But, for the purposes of broad classification, we see the largest sources of big data being internal archives, documents, media, data storage, business apps, public web, social media, machine log data, and sensor data.

I'll take a moment and highlight what each of these sources entails.

- Internal Archives: These are scanned documents, statements, customer correspondence notes, etc.
- Documents: These will range from PDF, XLS, CVS, DOC, etc.
- Media: Images, videos, audio, live streams, etc.
- Data Storage: These are files within SQL, Hadoop, etc.

- Business Apps: These are apps specific to business processes, such as CRM, talent management, HR, expense management, marketing, etc.
- Public Web: This is the tracking of public databases and text scrubbing for weather, traffic, regulatory compliance, public finance, health care, etc.
- Social Media: Facebook, Weibo, LinkedIn, Twitter, Youku, QZone, RenRen, etc.
- Machine Log Data: These are inclusive of event logs, server data, clickstreams, call detail logs, mobile location, mobile app usage, etc.
- Sensor Data: Such devices may include smart electronic meters, road cameras, video games, car computers, medical tools, household appliances, etc.

DRC: What are the best analytic techniques in the field of big data?

Randy Shumway: When we take a step back and look at the scale and variety of analytic techniques being used for big data today, it is amazing. What started as a handful of analytic algorithms that were taken from varying fields like machine learning, statistics, pattern recognition, artificial intelligence, and database systems—has now evolved into far most robust and complex sets of tools that are specific to big data applications. Moreover, these analytic techniques are constantly evolving to accommodate new sources of data, as well as the unique tasks businesses are turning to big data to help solve.

It is because the use cases of big data are constantly progressing, we don't believe there are top analytic models or best practice analytic models. When it comes to big data, each type of analytic technique used has a different algorithm and thus produces a different impact or result.

At Cicero Group, the analytic technique that is used ultimately depends on the type of business problem we are trying to solve. For instance, we might use a regression model when trying to identify factors that drive profit, a classification algorithm when trying to predict which customers are likely to purchase, a clustering algorithm when identifying customer segments, and a supervised learning technique like survival analysis when predicting exactly when a customer is likely to churn.

These analytic techniques are just a few of the many different types available in the big data world, others include: Association Rules Mining, Ensemble Modeling (Bagging, Boosting, Random Forest), Decision Trees (C&RT, CHAID), Text Analytics, Naive Bayes Analysis, Neural Network Modeling, Support Vector Machine Learning, Market Basket Analysis, Regression Analysis (Multivariate, Logistics).

In addition to the growing variety of analytic techniques, there are an infinite number of use cases, some of which include: lead and sales prioritization, product or service recommendations, customer acquisition scoring, risk modeling, customer value/ profitability scoring, retention and attrition prediction, content optimization, cross- and up-selling, likelihood to purchase, and customer marketing optimization.

At Cicero Group, we understand that the world of big data is constantly evolving and we must be quick to adapt—which is why we pride ourselves on staying up-to-speed on the latest industry trends, techniques, and algorithms.

To do this, we routinely send our analysts and senior management to big data conferences, seminars, and workshops to grow and refine their analytic skillsets. In addition, we pay 100% of the tuition for all of our analysts to obtain Master of Science degrees in Predictive Analytics from Northwestern University. As a result, our clients are benefiting immensely from the latest analytic methodologies.

DRC: What technologies are data scientists using to best harness the power of big data?

Randy Shumway: The biggest technology adoption that we are seeing is around the big data infrastructure. As I mentioned earlier, big data has three unique characteristics: volume, velocity, and variety, also known as the "3 V's." These three unique characteristics are also the three greatest IT challenges associated with harnessing big data. For businesses seeking to turn big data into a competitive advantage, an internal IT audit will often reveal that the necessary IT architecture is not in place for handling such challenges. The reason being, most businesses are supported by what is referred to as traditional enterprise architectures, which are designed to manage and analyze relational databases, and do not lend themselves well to unstructured data, data of large volumes, and high processing speeds. To overcome this challenge, businesses are adopting what are referred to as cluster architectures, which are systems that can cost-effectively scale to the unique demands of big data.

At a high level, cluster architectures are a set of connected computers, often low-cost commercial

computers, that work together to perform a single task. This organized cooperation between a set of connected computers enables greater processing power, especially when analyzing large, unstructured datasets. Compared to enterprise architectures, cluster architectures provide greater scalability, reliability, and processing speed, which creates an optimal platform for businesses looking to get a handle on big data. Now, the technology around cluster architectures have been around for a while, but what is new is the software platform that is lending this architecture so well to big data, and that is Hadoop.

Hadoop is an open source project under Apache, and is designed specifically for cluster architectures. What makes the marriage between Hadoop and cluster architectures so exciting is that it is enabling businesses to store as much data as they want, in whatever form they want, and in a cost-effective manner. There are two fundamental aspects that make Hadoop special, which are: (1) its data storing system and (2) its data processing system. To understand Hadoop's data storing system, one must understand the Hadoop Distributed File System (HDSF). Similarly, to understand Hadoop's data processing system, one must understand MapReduce.

Let me begin with HDSF. Imagine you had a dataset that was larger than your computer or server's storage capacity. HDSF allows you to store that same file by reducing it into smaller pieces, then distributing those pieces across multiple, connected computers. As a result, businesses are able to store datasets that would otherwise exceed the capacity of any single computer or server.

The second aspect of Hadoop that is import to understand is how Hadoop processes data (e.g., searching, reducing, enriching, etc.), which is done through MapReduce. Within traditional enterprise architectures, data are moved to where the processing task is set to occur, which has the potential to overwhelm a network, particularly when dealing with larger datasets. To overcome this challenge, MapReduce essentially moves the data processing task to each individual data location (i.e. each individual computer). In this framework, each individual computer has its own small dataset (assigned by HDSF) and is responsible for processing that data (assigned by MapReduce). MapReduce ultimately collects the output from each computer and then 'reduces' it into a single, final output. The end result is a process that is significantly quicker and more efficient than anything previously produced by traditional enterprise architectures.

DRC: In what domains in particular does Cicero Group have expertise creating value?

Randy Shumway: What is unique about Cicero Group is our ability to translate data into meaningful and actionable direction across all domains, not just a select few. One of the things I want to emphasize is that the power of data, big or small, is only realized when it is accurately translated into information that drives greater demand from the consumer. Whether its understanding market needs, developing better customer relationships to improve service and loyalty, or decreasing manufacturing costs while simultaneously improving output, the data needs to be accurately gathered, translated, and implemented to create any meaningful value.

Big data has three unique characteristics: volume, velocity, and variety, also known as the "3 V's."

It is that ability to accurately gather, translate, and implement that sets Cicero Group apart from other strategy consulting firms. When you collect the data properly and understand best industry practices, you can harness meaningful data-driven solutions that generate global impact. Furthermore, we have found that successful value creation requires a strategy that is not only informed by hard facts (i.e. data), but also continuously refined through ongoing collaborative dialogue with our clients, designed with implementation in clear focus, and structured with flexibility to take advantage of emergent opportunities.

DRC: What advice do you have for organizations throughout China that want to generate the most value out of the data that is accessible to them?

Randy Shumway: Over the past decade, numerous businesses have sought to implement a big data strategy—some of which have been very successful while others have not. Those who have been successful are the businesses that have transformed themselves into a datadriven culture, as opposed to just viewing data as a "siloe'd" initiative. Having a datadriven culture means an organization uses analytics extensively and systematically to out-think and out-execute the competition.

For Chinese businesses looking to generate the most value out of their data, they need to understand two

key points. The first is that big data is worthless in a vacuum. Meaning, its potential is unlocked only when it can be leveraged to drive real decision making. This is achieved through harnessing best practices and close collaboration between the decision makers and analysts. The second point is that data is only part of the process for generating meaningful results. Think of data-driven insights as an iceberg, the portion above the surface is the data component, but underneath is a far larger piece of the iceberg, which encompasses the firm's data strategy, data skill sets, data processes, and data systems. Without these components, and without properly tuning these components to interact with one another, business cannot effectively compete using big data.

At Cicero Group, this is what we do and it's something we are passionate about. We see incredible potential in China for organizations looking to harness the power of big data, and we are here to help transition those organizations – from government, to state owned enterprises, to private and public companies – into true analytic competitors. Throughout the transformation process, we not only assist businesses in collecting and organizing their big data, but also help translate and implement that data into the most innovative and effective solutions.

DRC: Walk me through the methodology you would apply to create a data system for an organization.

Randy Shumway: Every organization's goals and challenges are different. In addition, the type of data that is accessible to them might be unique. Therefore, virtually every data system we design is customized and individualized for the organization. And the possibilities and impact of the different data systems are absolutely endless. We build data systems for hospital networks to improve patient care while simultaneously reducing costs. We build data systems, leveraging pollution data sensors, to identify the type of pollutants both in the air and in the water, to monitor their movement, and to locate the source and impact of the pollutants in order to create solutions to decrease air and water pollution. We build data systems to improve enrollment at vocational colleges and to then design the appropriate programs and pedagogy to best improve the relevance and depth of what graduates know and are trained to do. We build data systems for retailers that help them more quickly and less expensively obtain new customers and then more effectively retain those customers for the long-term. We build data systems for banks that leverage robust machine learning algorithms and Big Data technology to detect and prevent fraud as well as to improve debt recovery. And we build data systems to help organizations acquire the right companies overseas and then to more effectively integrate the two operations. As you can see, the possibilities and impact of customized data systems are absolutely endless.

Cicero Group is a premier data-driven strategy consulting firm. Cicero integrates inductive problem solving with insightful data analytics to guide business strategy.

For more information, visit CiceroGroup.com or follow us on Twitter: @Cicero_Group.