Against the Grain

Volume 26 | Issue 4 Article 22

2014

ATG Special Report: Big Data Takeaways

Ho Jung Yoo
UC-San Diego Library, hjsyoo@ucsd.edu

Reid Otsuji
UC-San Diego Library, rotsuji@ucsd.edu

ATG Special Report — Big Data Takeaways

by **Ho Jung Yoo** (Research Data Curation Program, UC-San Diego Library) <hjsyoo@ucsd.edu>

and **Reid Otsuji** (Research Data Curation Program, UC-San Diego Library) <rotsuji@ucsd.edu>

On March 12, 2014, **UC San Diego (UCSD) Extension** held a two-hour special event called "Big Data at Work: A Conversation with the Experts" on the **UCSD** main campus. The purpose of the event was to showcase the topics of big data and data mining, where training in the latter is offered as a certificate program at **UCSD Extension**. The evening was kicked off with an introduction by **Hugo Villar Ph.D., MBA**, Director of Science and Technology at **UCSD Extension**.

What is this Big Data hype all about? What can we do with it? What is the future of Big Data? These questions were addressed by four experts representing the areas of academia, computing infrastructure, fraud and security, and commercial big data industry. The format of the event was 10-15 minute presentations followed by a panel discussion moderated by **Natasha Balac Ph. D.**, Director of the Center for Predictive Analytics at the **San Diego Supercomputer Center (SDSC)**. The four speakers were **Larry Smarr Ph.D.**, Founding Director of **CALIT2** and Professor of Computer Science & Engineering at **UCSD**; **Mike Norman Ph.D.**, Director of **SDSC**; **Stefan Savage Ph.D.**, Professor of Computer Science & Engineering; and **Michael Zeller Ph.D.**, CEO of **Zementis**, a San Diego-based software company focused on predictive analytics for big data and real-time scoring.

Several key points arose from the presentations. One is that the tsunami metaphor we often hear applied to the rise of big data is somewhat misleading. It's more of a sea-level rise or a high-speed elevator, because unlike the tsunami, which hits a peak then recedes, big data is only going to increase in size over time. Never before has humanity had to deal with this data growth rate. The growth phenomenon is coming from the multitudes of sensors (e.g., smartphones, automobiles, personal health devices) existing around the globe, their confluence, and the means of moving all those data around the network. A second point is that we often think of the three Vs that characterize big data: Volume, Velocity, and Variety. "Volume" refers to the vast amount of data being accumulated and which has gone from being measured globally in terabytes to zettabytes. "Velocity" is the high speed at which big data needs to move; this movement is transitioning from batch movement to networked streaming data. "Variety" refers to the many different types of data out there created primarily by digital sensors. At a high level, data types can be classified as transitioning from structured data to both structured and unstructured data, which can be more challenging to deal with. But there is an additional fourth V that we need to consider when thinking of big data: adding "Value" to the other three Vs. In industry, the concern is business value, and in academia, value comes from information. To maximize the value of big data in industry, we need to move from descriptive analytics, which answers the question, "What happened?," to predictive analytics, which asks, "What will happen next?" We also need to minimize latency from data to decision. This requires development of lots of automated models running simultaneously so that big data can be used to inform decisions in real time.

The speakers gave many examples of how we are deriving value from big data. **Zeller** noted the financial industry has been ahead of the game in terms of risk-scoring, prediction modeling, and fraud detection. Additionally, his key point was that big data in industries create opportunities to develop new platforms, capabilities, and business opportunities. **Savage** spoke of the security risks associated with having centralized data, and how **UCSD** researchers are using big data to improve security. For example, one group of researchers is using big data to battle email spam by behaving as naive spam users and conducting spam analytics. By analyzing those click trajectories, they were able to discover a great deal of spam traffic could be blocked by disabling just a handful of nodes (i.e., the spam enterprise or bank) within the spamming network.

In another example, the world's largest cryptocurrency, **Bitcoin**, had been touted for its pseudonymous transactions, which make transfers opaque. But are they truly opaque? Researchers at **UCSD** developed new data mining clustering techniques that allowed tracking of Bitcoin transactions in a way that made them less anonymous. The essence of big data security problems is understanding the information environment better, faster, and more efficiently than your adversary.

In the panel discussion, several interesting points were made about the usage and future of big data. **Zeller** called for the need to merge data analytics with the software technologies and to automate these processes so that predictive models based on big data can quickly inform our decisions. **Norman** described how **Jim Fowler**, a **UCSD** Professor of Political Science, has found there is a Facebook "Like" effect in national elections. Academic research is sometimes driven by events in the social sphere, and this leads research in unpredictable directions. When asked about the future of big data, **Zeller** predicts the hype will die down, but big data will have brought us lots of new, data-driven applications. **Norman** believes big data will give rise to a new discipline, much as "supercomputing," once a buzz word, made way for the discipline of computational science. The development of educational programs in data science at universities will be driven by the public's desire for such training. **Smarr** predicts we will have intelligent personal assistants that will interact with us to inform us of our health, decisions, and interpersonal relationships. **Savage** foresees a major trend towards cloud computing. We will not be able to hold on to our own data anymore because they are being generated too quickly and are becoming too bulky and costly for transport. We will leave the data where they are, in the cloud, and send computational requests over high-speed networks to the data where they reside.

Op Ed — Embracing the Digital ...
from page 28

provide enough incentive for libraries and individuals to lease and buy eBooks on their own.

This is a healthy thing for the overall academic book marketplace. In fact, book-sharing between libraries could be viewed by presses as a sort of "freemium" approach to promoting their content: libraries and patrons that borrow virtual books offered by a publisher might be more likely to buy the book or other titles from the publisher outright.

Librarians are right to be concerned about the long-term integrity and stewardship of electronic book content. **Amazon's Orwellian** recall of *1984* is often invoked as a cautionary tale. But national or global preservation strategies by trusted players such as **Portico** or **HathiTrust** make more sense for long term preservation than relying on even the largest research libraries to create digital vaults for their eBooks.

Some readers still prefer print to eBooks, and the subject matter of some books just works best in print. Thus purchasing and managing print books will continue to remain an important, if smaller, aspect of academic libraries' missions. Academic books are now commonly released in both digital and print formats, and for now, print collections and library sharing networks can serve as a fallback when digital access falters.

As books inevitably come to be released in digital-only format (see for example **Amazon** Kindle Singles in the consumer eBook world) libraries will have no choice but to come to terms with the the digital marketplace. Those libraries who swim with the digital current will have the most success in creating robust eBook collections for their patrons.