Big Data - Analytics Engine for Digital Transformation:

Where is IS?

Paulo B. Goes
Salter Professor of Information Technology and Management
Management Information Systems, University of Arizona

It is estimated that 90% of today’s existing data in the world has been created in the last two years (IBM 2015). Four Zettabytes (4 trillion GB) of data are created every year. Data come from sensors, scientific instruments, medical devices, smart phones; digital media including text, video, audio, email, blogs, postings, twitter feeds, click streams, financial and other transactions. The big promise here is that together with traditional data sources, big data offers unprecedented opportunities to study the “pulse of humanity” at granularity levels that were never available before.

For the last few years, government agencies, businesses, consultants, scientists and academics from various disciplines have been challenged with the very complex issues of how to harness big data, how to analyze big data and what to do with big data. How to capitalize on the promise of big data? The IS discipline has been challenged as well. Understanding data processing, storage, integration are integral to our field. Transforming data to information to knowledge is the essence of what we do. Supporting decision making with information and analytics has been part of our teaching and research arsenal since the concept of Decision Support Systems was introduced in the 70’s.

So what is different with big data? At one end of the spectrum the data generation, collection and storage are very different. Consider the not so futuristic scenario associated with consumers and the Internet of Things (Saleh et al, 2013). Consumers increasingly use digital devices (smartphones, tablets, their automobiles) to access, monitor, and control their connected digital products and services remotely over the Internet. Think for example heating systems, televisions, alarm systems, toys, lighting, home appliances, etc. Massive amounts of data reflecting consumers’ utilization patterns will be stored on the cloud along with data from other digital services (financial transactions, web browsing, app consumption, social media postings, temporal and spatial mobility tracking) and so on and so forth. Data from different sources will be messy, both structured and unstructured, time- and location- stamped and simply massive.

At the other end of the spectrum, with the consumer at the center, there are astonishing opportunities for innovative services and transformation brought by this new digital ecosystem centered at the consumer: insurance, health care, education, entertainment, retail, etc. Big data analytics is the engine that can power all these digital innovations.

As with many other contexts in information systems, a technology stack helps us understand the research opportunities going from big data capture to the layer of digital innovation and digital transformation. This layered depiction appears in Figure 1.

A lot has been said and written about the 4 V’s of big data: volume, variety, velocity and veracity. They reflect the zettabytes that are created every year, the diverse and highly distributed sources of data generation and the fact that there is incredible amount of noise that is inherent
with the capture of the data at the various points of generation.

**Big Data Stack**

**Infrastructure Layer**

The technical components of the infrastructure layer provide research opportunities, typically to computer scientists and engineers: data capture, streaming, storage, archiving, and parallel computing. Many of the innovations in this layer are driven by industry: Hadoop/Map Reduce, SAP’s HANA, NoSQL databases and cloud services. However, there are several topics that are germane to IS and the field should continue to do well with. One of them is data integration and reconciliation. These are critical issues in Big Data both at the semantic and syntactic levels. How to connect the information from various data sources to the same individual?  

The other side of the coin with connecting different sources of information to the same individual is privacy. The IS field is well positioned to contribute to privacy research. Some very interesting work is being generated for example by the Heinz College group at Carnegie-Mellon University, which builds on psychology, behavioral economics, and analysis of observational and experimental data (Acquisti et al. 2015). The fundamental question in privacy at the individual level is how much privacy intrusion one allows in exchange for how much more utility one gets. The big data environment magnifies the implications on both sides of this question.  

Because of the inter-relationships that exist between the technology, the people and the organizations, privacy, security and trust are essentially IS topics. Their importance grows with big data. IS security research encompasses behavioral, technical, economic and organizational approaches. In the big data realm, addressing the various security and control issues, including privacy and confidentiality are of fundamental current and future importance.
Designing and managing a platform of infrastructure services, which will often be a patchwork of several components, is also an exciting IS research area. Researchers can build on the vast existing literature that looks at what drives different infrastructure implementation choices. There are several works on optimization of cost, benefits and service levels. Provision through the cloud adds another dimension. Economic modeling based on game theoretic or principal-agent models may be useful here too, and let us not forget about contract theories, vendor management, etc. We have a solid base to build from.

Analytics Layer

Along the transformation path from big data to information to knowledge, lie a host of analytics techniques and approaches. Consider the customer-centric ecosystem mentioned above. Individuals at the center, along with their interactions, relationships, digital tracks, pose very challenging modeling issues. Variety of the data is key.

It is useful to look at the range of big data analytics through the following four categories:

- Exploration including visualization
- Explanation
- Prediction
- Prescription

The vast amounts of data, the diversity and the increasing presence of unstructured data make the analysis simply daunting. Exploration and visualization become essential parts of the analytics process. Methodologically, IS researchers have become well versed in exploratory techniques from both statistics and machine learning. The design of data visualization interfaces is a very rich area for research that involves HCI, cognitive sciences and a long tradition of IS research. How to design visualizations that are adequate for situation monitoring, alert systems and decision making?

However, due to the evolution of the IS field and its predominant research paradigms, the main focus of the existing data-driven IS research is explanatory. In IS we have become fairly sophisticated and competent with statistical models that are usually based on capturing linear relationships in hierarchical or structured models. But big data analytics is much more than explanation and hypothesis testing. It is about finding patterns and co-occurrences that provide understanding of the phenomena of interest and enables the design of new services. It is about using machine learning models (more predictive and less explanatory) in conjunction with explanatory models (see Shmueli and Koppius (2011) for an enlightening discussion).

One area that IS researchers need to pay attention and push forward is network modeling. Modeling explicit and implicit interactions derived from the vast amounts of data is extremely important. With big data, these interactions are now visible. A few IS researchers (Aral et al (2014), Zhang et al (2013)) have started to delve into large scale network analytics, but we need much more. The models are extremely complex, starting from sampling techniques, inference, and identifying influence. Statistical models that are built on linear relationships are no longer applicable. The networks are complex, multi-relational (links can mean different relationships), dynamic and evolve very fast.
**Prescriptive Analytics**

Companies have been spending millions of dollars in building the infrastructure for big data, hiring data scientists and building the analytics layer while the most important element of big data is to create the engine for digital innovation. The transformation power of big data lies in the design of innovative digital products and services, the digital processes that will support them and the new digital ecosystem.

To achieve the transformation, we have to create prescriptive models out of the analytics. Prescriptive models will lead to innovations in the form of policies, interventions, smart interactions, alerts, innovative services and products.

The design of digital services powered by big data analytics is the ultimate objective. Tremendous opportunities lie ahead for the design science area of IS. Not design of IT artifacts, but design of digital artifacts, which encompass technology, digital business processes and business models. Big data prescriptive analytics is what makes it possible.

**One illustrative example – Smart Campus Initiative at the University of Arizona**

Two years ago at the University of Arizona recognized the need to work on a comprehensive big data project, which would span the four layers of the architecture depicted in Figure 1. Inspired by the general theme of Smart Cities, we at the MIS department proposed to the University of Arizona administrators to work towards the development of our own Smart Campus.

We started from the bottom layer of the big data stack, and designed and implemented the infrastructure of Hadoop and MapReduce servers in our own clusters. We obtained student-related datasets from two main sources: (1) their CatCards (combination of student id, debit card and smart card) with information about what they purchased on campus, their access to facilities like their dorms, library and the gym, everything with location identification and timestamps; and (2) the logs of the students’ wifi access on campus from their laptops, tablets and smart phones. The sheer volume of the data qualified as “big”, in the Terabytes per week.

After working on solutions for both data integration and reconciliation and also de-identification to preserve the students’ privacy before the analytics were performed, we moved to data exploration and visualization. We built very insightful representations of the students’ movement on campus. This was possible because of the temporal / spatial nature of the data we had.

Subsequently, we used network models to create representations of student flow, and also student friendship maps. We were able to create networks that showed links between students if they were present at the same location. We then utilized machine learning methods to predict next location and activities of the students.

All the exploration and preliminary predictive analyses had value of their own, but the most interesting dimension of all the analyses that we did came from thinking about the prescriptive nature of the work. What could we do with all the insights we were getting?
Freshmen attrition is a huge problem faced by universities. At the University of Arizona, it is a
crucial issue. We got together with the university leadership and started to discuss ways that we
could tie our “big data” findings about student movement, student interactions and student
utilization of resources to their likelihood to drop out. By using predictive models that
incorporated these findings we were able to identify students at risk, with a much higher
accuracy than previous models that the university had that were based on traditional variables
related to demographics and performance in the classroom. Our models are not only more
accurate, but they are able to identify students at risk early enough so that prescriptive
interventions can be designed. This is where the impact of big data really is.

Now the exciting part of the work starts. The preliminary work with smart campus described
above showed us that we can move towards creating a comprehensive “student-at-the center”
digital ecosystem. Students’ interactions are already by and large digital. How they learn in a
digitally supported classroom (virtual or real), how they interact with the several university
services, how they participate in the different campus communities, their social engagement, etc.
The possibilities for an entire new layer of digital services based on big data prescriptive
analytics can be created. Improving on freshmen attrition is just one example of what can be
done.

Big data analytics allows for the creation of these “individual-at-the center” ecosystems: patients
in innovative healthcare environments, citizens in smart cities, consumers, students, members of
organizations, so on and so forth. Furthermore, these ecosystems overlap with each other. The
sky is the limit in what can be done.

IS research has to be part of all of this. We have to lead, solve the critical issues, privacy for
example, lead with the design of these digital services and systems and understand their impact
on individuals, organizations and society. The diversity of our research is our strength: the
different reference disciplines and our own different research paradigms position us well to move
forward and lead.

References

Acquisti, A., Brandimarte, L. and Loewenstein G. "Privacy and Human Behavior in the Age of
Saleh, T., Brock, J. and Yousif, N., “The Age of Digital Ecosystems: Thriving in a World of Big Data,”
July 23, 2013,
https://www.bcgperspectives.com/content/articles/information_technology_strategy_digital_economy_age
digital_ecosystems_thriving_world_big_data/
572, September 2011.
Social Network Autocorrelations for Binary Outcomes, With Applications to Technology Adoption”.