Permalink: **http://www.odbms.org/blog/2012/06/big-data-for-good/**

Big Data for Good.

A distinguished panel of experts discuss how Big Data can be used to create Social Capital.

by Roberto V. Zicari, Editor.

June 5, 2012.

Every day, 2.5 quintillion bytes of data are created. This data comes from digital pictures, videos, posts to social media sites, intelligent sensors, purchase transaction records, cell phone GPS signals to name a few. This is Big Data.

There is a great interest both in the commercial and in the research communities around Big Data. It has been predicted that "analyzing Big Data will become a key basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus", according to research by MGI and McKinsey's Business Technology Office.

But very few people seem to look at how Big Data can be used for solving social problems. Most of the work in fact is not in this direction. Why this? What can be done in the international research community to make sure that some of the most brilliant ideas do have an impact also for social issues?

I have invited a panel of distinguished well known researchers and professionals to discuss this issue. The list of panelists include:

- ***Roger Barga, Microsoft Research, group lead eXtreme Computing***

Group, USA

- Laura Haas, IBM Fellow and Director Institute for Massive Data, Analytics and Modeling IBM Research, USA

- Alon Halevy, Google Research, Head of the Structured Data Group, USA

- Paul Miller, Consultant, Cloud of Data, UK

This Q&A panel focuses exactly at this question: is it possible to conduct research for a corporation and/or a research lab, and at the same time make sure that the potential output of our research has also a social impact?

We take Big Data as a key example. Big Data is clearly of interest to marketers and enterprises a like who wish to offer their customers better services and better quality products. Ultimately their goal is to sell their products/services.

This is good, but how about digging into Big Data to help people in need? Preventing / predicting natural catastrophes, helping offering services "targeting" to people and structures in social need?

Hope you`ll find this interview interesting, as well as eye-opening. RVZ

Q1. In your opinion, would it be possible to exploit some of the current and future research and developments efforts on Big Data for achieving social capital?

Alon: Yes, Big data is not just the size of an individual data set, but rather the collection of data that is available to us online (e.g., government data, NGOs, local governments, journalists, etc). By putting these data together we help tell stories about the data and make them of interest and of value to the wider public. As one simple example, a recent **Danish Journalism Award** was given to a nice visualization of data about which doctors are being sponsored by the medical industry. The ability to communicate this data with the public is certainly part of the Big Data agenda.

Laura: Absolutely. In fact, many of the efforts that we are engaged in today are exactly in this direction. Much of our "**Smarter Planet**" related research is around utilizing more intelligently the large amounts of data coming from instrumenting, observing, and capturing the information about phenomena on planet earth, both natural and man-made.

Paul: First, it's important to recognise that technological advances, new techniques, and new ways of working often deliver both tangible and intangible social benefit as a by-product of something else. **Robert Owen** and his peers in the late 18th and early 19th centuries might

have had genuine motives for the social welfare and educational programmes they delivered for workers in their factories, but it was the commercial success of the factories themselves that paid for the philanthropy. And better educated children became better integrated factory workers, so it wasn't completely altruistic.

That said, there is clearly scope for Big Data to deliver direct benefits in areas that aid society. **Google Flu Trends** is perhaps the best-known example – analysis of many millions of searches for flu-related terms (symptoms, medicines, etc) enabling Google's non-profit Foundation to provide early visibility of illness in ways that could/should assist local healthcare systems. Google's search engine isn't **about** flu, and its indices aren't **for** flu detection or prediction; this piece of societal value simply emerges from the 'data exhaust' of all those people searching a single site. Flu Trends isn't alone; Harvard researchers found that **Twitter data could be analysed to track the spread of Cholera on Haiti** in a way that proved "substantially faster" than traditional techniques. According to Mathew Ingram's write-up of the research, "What the Harvard and HealthMap study shows is that analyzing the data from large sets like the tweets around Haiti isn't just good at tracking patterns or seeing connections after an event has occurred, but **can actually be of use to researchers on the ground while those events are underway**" (my emphasis).

Roger: Absolutely, we have already seen several such examples. One such example in science is Jim Gray and Alex Szalay's collaboration to build a virtual observatory for astronomy, which leveraged relational database technology. The **SDSS Sky Server** has since supported hundreds of researchers and resulted in thousands of publications over the years. Another, more recent example, is the language translation system researchers in Microsoft Research built for the aid relief worker in Haiti after the 2010 earthquake. They leveraged the same technology we leverage in our search operations to build a statistical machine translation engine to translate Haitian Creole to English from scratch in 4 days, 17 hours, and 30 minutes and delivered to aid workers in Haiti.

Q2. If yes, what are the areas where in your opinion Big Data could have a real impact on social capital?

Alon: Bringing data that is otherwise hidden from view to the eyes of the interested public. Data activists and journalists world-wide need to be able to easily discover data sets, merge them in a sensible fashion and tell stories about them that will grab people's attention. As another example, helping people in crisis response situations has huge potential. As two examples, people have used **Google Fusion Tables** to create maps with critical information for people after the Japan earthquake in 2011 and before the hurricane in NYC later that year.

Laura: Healthcare is an obvious one, where leveraging the vast amounts of genomic information now being produced together with patient records, and the medical literature could help us provide the best known treatments to an individual patient — or discover new therapies that may be more effective than those currently in use. We have worked already on leveraging big data and machine learning to predict the best therapeutic regimens for AIDS patients, for example. Or, when it comes to natural resources, we are leveraging big data to optimize the placement of turbines in a wind farm so that we get the most power for the least environmental impact. We can also look at man-made phenomena — for example, understanding traffic patterns and using the insight to do better planning or provide incentives that can reduce traffic during crunch hours. Many other examples can be given of how Big Data is being used to improve the planet!

Paul: The opportunities must – surely – be enormous? Any of the big issues affecting society, from environmental change, to population growth, to the need for clean water, food, and healthcare; all of these affect large groups of people and all of them have aspects of policy formulation or delivery that are (or should be, if anyone collected it) data-rich. The Volume, Velocity and Variety of data in many of these areas should offer challenging research opportunities for practitioners… and tangible benefits to society when they're successful.

Roger: Top of mind is to advance scientific research, what has been referred to as **eScience** which covers both the traditional hard sciences from astronomy, oceanography, to the social sciences and economics. Our ability to acquire and analyze unprecedented amounts of data has the potential to have a profound impact on science. It is a leap from the application of computing to support scientists to 'do' science (i.e. 'computational science') to the integration of computer science and ability to analyze volumes of data to extract insights into the very fabric of science. While on the face of it, this change may seem subtle, we believe it to be fundamental to science and the way science is practiced. Indeed, we believe this development represents the foundations of a new revolution in science. We captured stories from many different scientific investigations in the book "**The Fourth Paradigm: Data-Intensive Scientific Discovery**."

Q3. What are the main challenges in such areas?

Alon: *Data discovery* is a huge challenge (how to find high-quality data from the vast collections of data that are out there on the Web). Determining the *quality* of data sets and relevance to particular issues (i.e., is the data set making some underlying assumption that renders it biased or not informative for a particular question). *Combining* multiple data sets by people who have little knowledge of database techniques is a constant challenge.

Laura: With any big data project, many of the same issues exist. I'll mention three major categories of issues: those related to the *data*, itself, those related to the *process* of deriving insight and benefit from the data, and finally, those related to *management* issues such as data privacy, security, and governance in general. In the data space, we talk about the 4 V's of data — Volume (just dealing with the sheer size of it), Variety (handling the multiplicity of types and sources and formats), Velocity (reacting to the flood of information in the time required by the application), and, last and perhaps least understood, Veracity (how can we cope with uncertainty, imprecision, missing values, and yes, occasionally, mis-statements or untruths?). The challenges with deriving insight include capturing data, aligning data from different sources (e.g., resolving when two objects are the same), transforming the data into a form suitable for analysis, modeling it, whether mathematically, or through some form of simulation, etc, and then understanding the output — visualizing and sharing the results, for example. And governance includes ensuring that data is used correctly (abiding by its intended uses and relevant laws), tracking how the data is used, transformed, derived, etc, and managing its lifecycle. There are research topics in ALL of these areas!

Paul: *Data availability* – is there data available, at all? Increasingly, there is. But coverage and comprehensiveness often remain patchy, and the rigour with which datasets are compiled may still raise concerns. A good process will, typically, make bad decisions if based upon bad data. *Data quality* – how good is the data? How broad is the coverage? How fine is the sampling resolution? How timely are the readings? How well understood are the sampling biases? What are the implications in, for example, a Tsunami that affects several Pacific Rim countries? If data is of high quality in one country, and poorer in another, does the Aid response skew 'unfairly' toward the well-surveyed country or toward the educated guesses being made for the poorly surveyed one? *Data comprehensiveness* – are there areas without coverage? What are the implications? *Personally Identifiable Information* – much of this information is about people. Can we extract enough information to help people without extracting so much as to compromise their privacy? Partly, this calls for effective industrial practices. Partly, it calls for effective oversight by Government. Partly – perhaps mostly – it requires a realistic reconsideration of what privacy really means… and an informed grown up debate about the real trade-off between aspects of privacy 'lost' and benefits gained. Rather than offering blanket privacy policies, perhaps customers, regulators and software companies should be moving closer to some form of explicit data agreement; if you give me access to X, Y, and Z about yourself, I will use it for purposes A, B, and C… and you will gain benefits/services D, E, and F. The first two parts are increasingly in place, albeit informally. The final part – the benefits – is far less well expressed. *Data*

dogmatism – analysis of big data can offer quite remarkable insights, but we must be wary of becoming too beholden to the numbers. Domain experts – and common sense – must continue to play a role. It would be worrying, indeed, if the healthcare sector **only** responded to flu outbreaks when Google Flu Trends told them to! See, for example, a recent **blog post of mine** –

Roger: The first important step is to embrace a data-centric view. The goal is not merely to store data for a specific community but to improve data quality and to deliver as a service accurate, consistent data to operational systems. It isn't simply a matter of connecting the plumbing between many different data sources, there's a quality function that has to be applied, to clean, and reconcile all of this information. Researchers don't simply need data, they need services-based information over this data to support their work.

Q4. What are the main difficulties, barriers hindering our community to work on social capital projects?

Alon: I don't think there are particular barriers from a technical perspective. Perhaps the main barrier is ideas of how to actually take this technology and make social impact. These ideas typically don't come from the technical community, so we need more inspiration from activists.

Laura: Funding and availability of data are two big issues here. Much funding for social capital projects comes from governments — and as we know, are but a small fraction of the overall budget. Further, the market for new tools and so on that might be created in these spaces is relatively limited, so it is not always attractive to private companies to invest. While there is a lot of publicly available data today, often key pieces are missing, or privately held, or cannot be obtained for legal reasons, such as the privacy of individuals, or a country's national interests. While this is clearly an issue for most medical investigations, it crops up as well even with such apparently innocent topics as disaster management (some data about, e.g., coastal structures, may be classified as part of the national defense).

Paul: Perceived lack of easy access to data that's unencumbered by legal and privacy issues? The large-scale and long term nature of most of the problems? It's not as 'cool' as something else? A perception (whether real or otherwise) that academic funding opportunities push researchers in other directions? Honestly, I'm not sure that there are significant insurmountable difficulties or barriers, if people want to do it enough. As Tim O'Reilly said in 2009 (and many times since), developers should "Work on stuff that matters." The same is true of researchers.

Roger: The greatest barrier may be social. Such projects require community awareness to bring people to take action and often a champion to frame the technical challenges in a way that is approachable by the community. These projects will likely require close collaboration between the technical community and those familiar with the problem.

Q5. What could we do to help supporting initiatives for Big Data for Good?

Alon: Building a collection of high quality data that is widely available and can serve as the backbone for many specific data projects. For example, data sets that include boundaries of countries/counties and other administrative regions, data sets with up-to-date demographic data. It's very common that when a particular data story arises, these data sets serve to enrich it.

Laura: Increasingly, we see consortiums of institutions banding together to work on some of these problems. These Centers may provide data and platforms for data-intensive work, alleviating some of the challenges mentioned above by acquiring and managing data, setting up an environment and tools, bringing in expertise in a given topic, or in data, or in analytics, providing tools for governance, etc. My own group is creating just such a platform, with the goal of facilitating such collaborative ventures. Of course, lobbying our governments for support of such initiatives wouldn't hurt!

Paul: Match domains with a need to researchers/companies with a skill/product. Activities such as the recent Big Data Week Hackathons might be one route to follow – encourage the organisers (and companies like Kaggle, which do this every day) to run Hackathons and competitions that are explicitly targeted at a 'social' problem of some sort. Continue to encourage the **Open Data** release of key public data sets. Talk to the agencies that are working in areas of interest, and understand the problems that they face. Find ways to help them do what they already want to do, and build trust and rapport that way.

Roger: Provide tools and resources to empower the long tail of research. Today, only a fraction of scientists and engineers enjoy regular access to high performance and data-intensive computing resources to process and analyze massive amounts of data and run models and simulations quickly. The reality for most of the scientific community is that speed to discovery is often hampered as they have to either queue up for access to limited resources or pare down the scope of research to accommodate available processing power. This problem is particularly acute at the smaller research institutes which represent the long tail of the research community. Tier 1 and some tier 2 universities have sufficient funding and infrastructure to secure and

support computing resources while the smaller research programs struggle. Our funding agencies and corporations must provide resources to support researchers, in particular those who do not have access to sufficient resources.

Q6. Are you aware of existing projects/initiatives for Big Data for Good?

Laura: Yes, many! See above for some examples. IBM Research alone has efforts in each of the areas mentioned — and many more. For example, we've been working with the city of Rio, in Brazil, to do detailed flood modeling, meter by meter; with the Toronto Children's Hospital to monitor premature babies in the neonatal ward, allowing detection of life-threatening infections up to 24 hours earlier; and with the Rizzoli Institute in Italy to find the best cancer treatments for particular groups of patients.

Roger: Yes, the United Nations Global Pulse initiative is one example. Earlier this year at the 2012 Annual Meeting in Davos, the World Economic Forum published a white paper entitled "**Big Data, Big Impact: New Possibilities for International Development**". The WEF paper lays out several of the ideas which fundamentally drive the Global Pulse initiative and presents in concrete terms the opportunity presented by the explosion of data in our world today, and how researchers and policymakers are beginning to realize the potential for leveraging Big Data to extract insights that can be used for Good, in particular for the benefit of low-income populations. What I find intriguing about this project from a technical perspective is how to extract insight from ambient data, from GPS devices, cell phones and medical devices, combined with crowd sourced data from health and aid workers in the field, then analyzed with machine learning and analytics to predict a potential social need or crisis in advance while remediation is still viable.

Q7. Anything else you wish to add?

Alon: Google Fusion Tables has been used in many cases for social good, either though journalists, crisis response or data activists making a compelling visualization that caught people's attention. This has been one of the most gratifying aspects of working on Fusion Tables and has served as a main driver for prioritizing our work: make it easy for people with passion for the data (rather than database expertise) to get their work done; make it easier for them to find relevant data and combine it with their own. We look very carefully at the workflow of these professionals and try to make it as efficient as possible.

Laura: I think our community has the ability to do a lot of good by leveraging the tools we are developing, and our expertise, to attack

some of the critical problems facing our world. We may even create economic value (not a bad thing, either!) while doing so.

Dr. Roger Barga has been with the Microsoft Corporation since 1997, first working as a researcher in the database research group of Microsoft Research, then as architect of the Technical Computing Initiative, followed by architect and engineering group lead in the eXtreme Computing Group of Microsoft Research. He currently leads a product group developing an advanced analytics service on Windows Azure. Roger holds a PhD in Computer Science (database systems), MS in Computer Science (machine learning), and a BS in Mathematics.

Dr. Alon Halevy heads the Structured Data Group at Google Research. Prior to that, he was a Professor of Computer Science at the University of Washington, where he founded the Database Research Group. From 1993 to 1997 he was a Principal Member of Technical Staff at AT&T Bell Laboratories (later AT&T Laboratories). He received his Ph.D in Computer Science from Stanford University in 1993, and his Bachelors degree in Computer Science and Mathematics from the Hebrew University in Jerusalem in 1988. Dr. Halevy was elected Fellow of the Association of Computing Machinery in 2006.

Dr. Laura Haas is an IBM Fellow, and Director of IBM Research's new Institute for Massive Data, Analytics and Modeling; she also serves as a "catalyst" for ambitious research across IBM's worldwide research labs. She was the Director of Computer Science at IBM's Almaden Research Center from 2005 to 2011. From 2001-2005, she led the Information Integration Solutions architecture and development teams in IBM's Software Group. Previously, Dr. Haas was a research staff member and manager at Almaden. She is best known for her work on the Starburst query processor, from which DB2 LUW was developed, on Garlic, a system which allowed integration of heterogeneous data sources, and on Clio, the first semi-automatic tool for heterogeneous schema mapping. She has received several IBM awards for Outstanding Innovation and Technical Achievement, an IBM Corporate Award for her work on information integration technology, and the Anita Borg Institute Technical Leadership Award. Dr. Haas was Vice President of the VLDB Endowment Board of Trustees from 2004-2009, and is a member of the National Academy of Engineering and the IBM Academy of Technology, an ACM Fellow, and Vice Chair of the board of the Computing Research Association.

Dr. Paul Miller is Founder of the Cloud of Data, a UK-based consultancy primarily concerned with Cloud Computing, Big Data, and Semantic Technologies. He works with public and private sector clients in Europe and North America, and has a Ph.D in Archaeology

(Geographic Information Systems) from the University of York

Acknowledgement: I would like to thank **Michael J. Carey** with whom I have brainstormed about this project at EDBT in Berlin. RVZ