The elephant in the room: getting value from Big Data

Serge Abiteboul (INRIA Saclay & ENS Cachan, France),
Luna Dong (Google Inc., USA),
Oren Etzioni (Allen Institute for Artificial Intelligence, USA),
Divesh Srivastava (AT&T Labs-Research, USA),
Gerhard Weikum (Max Planck Institute for Informatics, Germany),
Julia Stoyanovich (Drexel University, USA) and
Fabian M. Suchanek (Télécom ParisTech, France)

1. INTRODUCTION

Big Data, and its 4 Vs – volume, velocity, variety, and veracity – have been at the forefront of societal, scientific and engineering discourse. Arguably the most important 5th V, value, is not talked about as much. How can we make sure that our data is not just big, but also valuable?

WebDB 2015[1] has as its theme "Freshness, Correctness, Quality of Information and Knowledge on the Web". The workshop attracted 31 submissions, of which the best 9 were selected for presentation at the workshop, and for publication in the proceedings.

To set the stage, we have interviewed several prominent members of the data management community, soliciting their opinions on how we can ensure that data is not just available in quantity, but also in quality. In this interview Serge Abiteboul, Oren Etzioni, Divesh Srivastava with Luna Dong, and Gerhard Weikum shared with us their motivation for doing research in the area of data quality, and discussed their current work and their view on the future of the field. This interview appeared as a SIGMOD Blog article.[2]

Julia Stoyanovich and Fabian M. Suchanek,
Co-Chairs of WebDB 2015.

2. SERGE ABITEBOUL

Serge Abiteboul is a Senior Researcher INRIA Saclay, and an affiliated professor at Ecole Normale Supérieure de Cachan. He obtained his Ph.D. from the University of Southern California, and a State Doctoral Thesis from the University of Paris-Sud. He was a Lecturer at the École Polytechnique and Visiting Professor at Stanford and Oxford University. He has been Chair Professor at Collège de France in 2011-12 and Francqui Chair Professor at Namur University in 2012-2013. He co-founded the company Xyleme in 2000. Serge Abiteboul has received the ACM SIGMOD Innovation Award in 1998, the EADS Award from the French Academy of sciences in 2007; the Milner Award from the Royal Society in 2013; and a European Research Council Fellowship (2008-2013). He became a member of the French Academy of Sciences in 2008, and a member the Academy of Europe in 2011. He is a member of the Conseil national du numérique. His research work focuses mainly on data, information and knowledge management, particularly on the Web.

What is your motivation for doing research on the value of Big Data?

My experience is that it is getting easier and easier to get data but if you are not careful all you get is garbage. So quality is extremely important, never over-valued and certainly relevant. For instance, with some students we crawled the French Web. If you crawl naively, it turns out that very rapidly all the URLs you try to load are wrong, meaning they do not correspond to real pages, or they return pages without real content. You need to use something such as PageRank to focus your resources on relevant pages.

So then what is your current work for finding the equivalent of "relevant pages" in Big Data?

I am working on personal information where very often, the difficulty is to get the proper knowledge and, for instance, align correctly entities from different sources. My long-term goal also working for instance with Amélie Marian is the construction of a Personal Knowledge Base that gathers all the knowledge someone can get about his/her life. For each one of us, such knowledge has enormous potential value, but for the moment it lives in different silos and we cannot get this value.

This line of work is not purely technical, but involves societal issues as well. We are living in a world where companies and governments have loads of data on us and we don't even know what they have and how they are using it. Personal Information Management is an attempt to rebalance the situation, and make personal data more easily accessible to the individuals. I have a paper on Personal Information Management Systems, talking about just that, that just appeared in CACM (with Benjamin André and Daniel Kaplan).[3].

[1] http://dbweb.enst.fr/events/webdb2015

[2] http://wp.sigmod.org/?p=1519

[3] http://cacm.acm.org/magazines/2015/5/186024-managing-your-digital-life

WebDB'15, May 31, 2015, Melbourne, VIC, Australia
Copyright 2015 ACM 978-1-4503-3627-7/15/05 ...$15.00
http://dx.doi.org/10.1145/2767109.2770014

And what is your view of the killer app of Big Data?

Relational databases was a big technical success in the 1970s-80s. Recommendation of information was a big one in the 1990s-2000s, from PageRank to social recommendation. After data, after information, the next big technical success is going to be "knowledge", say in the 2010s-20s :). It is not an easy sell because knowledge management has often been disappointing – not delivering on its promises. By knowledge management, I mean systems capable of acquiring knowledge at a large scale, reasoning with this knowledge, exchanging knowledge in a distributed manner. I mean techniques such as that used at Berkeley with Bud or at INRIA with Webdamlog. To build such systems, beyond scale and distribution, we have to solve quality issues: the knowledge is going to be imprecise, possibly missing, with inconsistencies. I see knowledge management as the next killer app!

3. OREN ETZIONI

Oren Etzioni is Chief Executive Officer of the Allen Institute for Artificial Intelligence. He has been a Professor at the University of Washington's Computer Science department starting in 1991, receiving several awards including GeekWire's Geek of the Year (2013), the Robert Engelmore Memorial Award (2007), the IJCAI Distinguished Paper Award (2005), AAAI Fellow (2003), and a National Young Investigator Award (1993). He was also the founder or co-founder of several companies including Farecast (sold to Microsoft in 2008) and Decide (sold to eBay in 2013), and the author of over 100 technical papers that have garnered over 22,000 citations. The goal of Oren's research is to solve fundamental problems in AI, particularly the automatic learning of knowledge from text. Oren received his Ph.D. from Carnegie Mellon University in 1991, and his B.A. from Harvard in 1986.

Oren, how did you get started in your work on Big Data?

I like to say that I've been working on Big Data from the early days when it was only "small data". Our 2003 KDD paper on predictive pricing started with a data set with 12K data points. By the time Farecast was sold to Microsoft, in 2008, we were approaching a trillion labeled data points. Big price data was the essence of Farecast's predictive model, and had the elegant property that it was "self labeling". That is, if we can label the airfare on a flight from Seattle to Boston with either a "buy now" or "wait" label – all we have to do is monitor the price movement over time to determine the appropriate label. 20/20 hindsight allows us to produce labels automatically. But for Farecast, and other applications of Big Data, the labeled data points are only part of the story. Background knowledge, reasoning, and more sophisticated semantic models are necessary to take predictive accuracy to the next level.

So what is the AI2 working on to bring us to this next level?

Beginning in January 1, 2014 we launched the Allen Institute for AI[4], a research center dedicated to leveraging modern data mining, text mining, and more in order to make progress on fundamental AI questions, and to develop high-impact AI applications. One key project is Semantic Scholar, which utilizes big data over millions of academic papers to revolutionize the process of homing in on relevant papers and important citations. We are developing information extraction methods that map PDF files to key attributes including the problem in the paper, the methods used, the data sets employed, and the results reported. See this link[5] for more information, and to be notified when Semantic Scholar launches as a free service later in 2015. Beyond text, we find that figures are an important source of information in academic papers and have begun a research program to extract figures from the papers and analyze them. The first results in this research program (including open-source software for extracting figures) are available here[6].

And thinking ahead, what would be the killer application that you have in mind for Big Data?

Ideas like "background knowledge" and "common-sense reasoning" are investigated in AI whereas Big Data and data mining has developed into its own vibrant community. Over the next 10 years, I see the potential for these communities to re-engage with the goal of producing methods that are still scalable, but require less manual engineering and "human intelligence" to work. The killer application would be a Big Data application that easily adapts to a new domain, and that doesn't make egregious errors because it has "more intelligence". More concretely, we are looking at systems like Semantic Scholar, which will operate over a graph of more than 100M papers linked by citations to each other, as an application that will drive exciting research in AI and Big Data methods coming together to make literature search, which scientists and doctors do daily, more efficient than ever.

4. DIVESH SRIVASTAVA & LUNA DONG

Divesh Srivastava is the head of the Database Research Department at AT&T Labs-Research. He received his Ph.D. from the University of Wisconsin, Madison, and his B.Tech from the Indian Institute of Technology, Bombay. He is an ACM fellow, on the board of trustees of the VLDB Endowment, the managing editor of the Proceedings of the VLDB Endowment (PVLDB), and an associate editor of the ACM Transactions on Database Systems. His research interests and publications span a variety of topics in data management.

Xin Luna Dong is a senior research scientist at Google. She works on enriching and cleaning knowledge for the Google Knowledge Graph. Her research interest includes data integration, data cleaning, and knowledge management. Prior to joining Google, she worked for AT&T Labs-Research and received her Ph.D. in Computer Science and Engineering at the University of Washington. She is the co-chair for WAIM'15 and has served as an area chair for

[4] http://www.allenai.org

[5] http://allenai.org/semantic-scholar.html

[6] http://allenai.org/content/publications/clark_divvala.pdf

SIGMOD'15, ICDE'13, and CIKM'11. She won the best-demo award in SIGMOD'05.

Divesh and Luna, you have been working on several aspects of Big Data Value. What attracts you to this topic?

Value, the 5th V of big data, is arguably the promise of the big data era. The choices of what data to collect and integrate, what analyses to perform, and what data-driven decisions to make, are driven by their perceived value – to society, to organizations, and to individuals. It is worth noting that while value and quality of big data may be correlated, they are conceptually different. For example, one can have high quality data about the names of all the countries in North America, but this list of names may not have much perceived value. In contrast, even relatively incomplete data about the shopping habits of people can be quite valuable to online advertisers.

It should not be surprising that early efforts to extract value from big data have focused on integrating and extracting knowledge from the low-hanging fruit of "head" data – data about popular entities, in the current world, from large sources. This is true both in industry (often heavily relying on manual curation) and in academia (often as underlying assumptions of the proposed integration techniques). However, focusing exclusively on head data leaves behind a considerable volume of "tail" data, including data about less popular entities, in less popular verticals, about non-current (historical) facts, from smaller sources, in languages other than English, and so on. While each data item in the "long tail" may provide only little value, the total value present in the long tail can be substantial, possibly even exceeding the total value that can be extracted solely from head data. This is akin to shops making a big profit from a large number of specialty items, each sold in a small quantity, in addition to the profit made by selling large quantities of a few popular items.

We issue a call to arms – "leave no valuable data behind" in our quest to extract significant value from big data.

What is your recent work in this quest?

Our work in this area focuses on the acquisition, integration, and knowledge extraction from big data. More recently, we have been considering a variety of ideas, including looking at collaboratively edited databases, news stories, and "local" information, where multiple perspectives and timeliness can be even more important than guaranteeing extremely high accuracy (e.g., 99% accuracy requirement for Google's Knowledge Graph).

We started this body of work a few years ago with the Solomon project for data fusion, to make wise decisions about finding the truth when faced with conflicting information from multiple sources. We quickly realized the importance of copy detection between sources of structured data to solve this problem, and developed techniques[7] that iteratively perform copy detection, source trustworthiness evaluation, and truth discovery. The Knowledge Vault (KV) project and the Sonya project naturally extend the Solomon project to address the challenge of web-scale data. They focus on knowledge fusion, finding truthfulness of extracted knowledge from web-scale data (see here[8]), and building probabilistic knowledge bases, in the presence of source errors and extraction errors, with the latter dominating (see here[9]). The Sonya project[10] in addition measures knowledge-based trust, determining the trustworthiness of web sources based on the correctness of the facts they provide.

Big data often has a temporal dimension, reflecting the dynamic nature of the real-world, with evolving entities, relationships and stories. Over the years we have worked on many big data integration problems dealing with evolving data. For example, our work on temporal record linkage[11] addressed the challenging problem of entity resolution over time, which has to deal with evolution of entities wherein their attribute values can change over time, as well as the possibility that different entities are more likely to share similar attribute values over time. We have also looked at quality issues in collaboratively edited databases, with some recent work on automatically identifying fine-grained controversies over time in Wikipedia articles ("Scaling up copy detection" in ICDE 2015).

More recently, we have been working on the novel topic of data source management, which is of increasing interest because of the proliferation of a large number of data sources in almost every domain of interest. Our initial research on this topic involves assessing the evolving quality of data sources, and enabling the discovery of valuable sources to integrate before actually performing the integration (see here[12] and here[13]).

Finally, we make a shameless plug for our new book "Big Data Integration" that was recently published[14], which we hope will serve as a starting point for interested readers to pursue additional work on this exciting topic.

And where do you think will research head tomorrow?

In keeping with our theme of "no valuable data left behind", we think that effectively collecting, integrating, and using tail data is a challenging research direction for the big data community. There are many interesting questions that need to be answered. How should one acquire, integrate, and extract knowledge on tail entities, and for tail verticals, when there may not be many data sources providing relevant data? How can one understand the quality and value of tail data sources? How can such sources be used without compromising on value, even if the data are not of extremely high quality? How does one integrate historical data, including entities that evolve over time, and enable the exploration of the history of web data sources? In addition to freshness, what additional metrics are relevant to capturing quality over time? How does one deal with sources that provide data about future events? How

[7] http://www.vldb.org/pvldb/vol6/p97-li.pdf

[8] http://www.vldb.org/pvldb/vol7/p881-dong.pdf

[9] http://doi.acm.org/10.1145/2623330.2623623

[10] http://arxiv.org/abs/1502.03519

[11] http://www.vldb.org/pvldb/vol4/p956-li.pdf

[12] http://www.vldb.org/pvldb/vol6/p37-dong.pdf

[13] http://www.cidrdb.org/cidr2015/Papers/CIDR15_Paper21.pdf

[14] http://dx.doi.org/10.2200/S00578ED1V01Y201404DTM040

can one integrate data across multiple languages and cultures? Answering these challenging questions will keep our community busy for many years to come.

5. GERHARD WEIKUM

Gerhard Weikum is a scientific director at the Max Planck Institute for Informatics in Saarbrücken, Germany, where he is leading the department on databases and information systems. He co-authored a comprehensive textbook on transactional systems, received the VLDB 10-Year Award for his work on automatic DB tuning, and is one of the creators of the YAGO knowledge base[15]. Gerhard is an ACM Fellow, a member of several scientific academies in Germany and Europe, and a recipient of a Google Focused Research Award, an ACM SIGMOD Contributions Award, and an ERC Synergy Grant.

What is your motivation for doing research in the area of Big Data Value?

Big Data is the New Oil! This often heard metaphor refers to the world's most precious raw asset – of this century and of the previous century. However, raw oil does not power any engines or contribute to high-tech materials. Oil needs to be cleaned, refined, and put in an application context to gain its true value. The same holds for Big Data. The raw data itself does not hold any value, unless it is processed in analytical tasks from which humans or downstream applications can derive insight. Here is where data quality comes into play, and in a crucial role.

Some applications may do well with huge amounts of inaccurate or partly erroneous data, but truly mission-critical applications would often prefer less data of higher accuracy and correctness. This Veracity dimension of the data is widely underestimated. In many applications, the workflows for Big Data analytics include major efforts on data cleaning, to eliminate or correct spurious data. Often, a substantial amount of manual data curation is unavoidable and incurs a major cost fraction.

OK, I see. Then what is your recent work in the area of oil refinery?

Much of the research in my group at the Max Planck Institute could actually be cast under this alternative – and much cooler – metaphor: Big Text Data is the New Chocolate!

We believe that many applications would enormously gain from tapping unstructured text data, like news, product reviews in social media, discussion forums, customer requests, and more. Chocolate is a lot more sophisticated and tasteful than oil – and so is natural-language text. Text is full of finesse, vagueness and ambiguities, and so could at best be seen as Uncertain Big Data. A major goal of our research is to automatically understand and enrich text data in terms of entities and relationships and this way enable its use in analytic tasks – on par with structured big data.

We have developed versatile and robust methods for discovering mentions of named entities in text documents, like news articles or posts in social media, and disambiguating them onto entities in a knowledge base or entity catalog. The AIDA software[16] is freely available as open source code. These methods allow us to group documents by entities, entity pairs or entity categories, and compute aggregates on these groups. Our STICS demonstrator[17] shows some of the capabilities for semantic search and analytics. We can further combine this with the detection and canonicalization of text phrases that denote relations between entities, and we can capture special kinds of text expressions that bear sentiments (like/dislike/support/oppose/doubt/etc.) or other important information.

Having nailed down the entities, we can obtain additional structured data from entity-indexed data and knowledge bases to further enrich our text documents and document groups. All this enables a wealth of co-occurrence-based analytics for comparisons, trends, outliers, and more. Obviously, for lifting unstructured text data to this value-added level, the Veracity of mapping names and phrases into entities and relations is decisive.

For example, when performing a political opinion analysis about the Ukrainian politician and former boxer Klitschko, one needs to be careful about not confusing him with his brother who is actively boxing. A news text like "former box champion Klitschko is now the mayor of Kiev" needs to be distinguished from "box champion Klitschko visited his brother in Kiev". Conversely, a recent text like "the mayor of Kiev met with the German chancellor" should also count towards the politician Vitali Klitschko although it does not explicitly mention his name.

Why New Chocolate?

Well, in the Aztec Empire, cocoa beans were so valuable that they were used as currency! Moreover, cocoa contains chemicals that trigger the production of the neurotransmitter Serotonin in our brains – a happiness substance! Yes, you may have to eat thousands of chocolate bars before you experience any notable kicks, but for the sake of the principle: chocolate is so much richer and creates so much more happiness than oil.

Thanks for this culinary encouragement! What do you think will be the future of the field?

Data quality is more than the quest for Veracity. Even if we could ensure that the database has only fully correct and accurate data points, there are other quality dimensions that often create major problems: incompleteness, bias and staleness are three aspects of paramount importance.

No data or knowledge base can ever be perfectly complete, but how do we know which parts we do not know? For a simple example, consider an entertainment music database, where we have captured a song and ten different cover versions of it. How can we tell that there really are only ten covers of that song? If there are more, how can we rule out that our choice of having these ten in the database is not biased in any way – for example, reflecting only Western culture versions and ignoring Asian covers? Tapping into text sources, in the New Chocolate sense, can help completing the data, but is also prone to "reporting bias". The possibility that some of the data is stale, to different degrees, makes

[15] `http://yago-knowledge.org`

[16] `http://www.mpi-inf.mpg.de/yago-naga/aida/`

[17] `https://stics.mpi-inf.mpg.de/`

the situation even more complex.

Finally, add the Variety dimension on top of all this – not a single database but many independent data and text sources with different levels of incompleteness, bias, and staleness. Assessing the overall quality that such heterogeneous and diverse data provides for a given analytic task is a grand challenge. Ideally, we would like to understand how the quality of that data affects the quality of the insight we derive from it. If we consider data cleaning measures, what costs do we need to pay to achieve which improvements in data quality and analytic-output quality? I believe these are pressing research issues; their complexity will keep the field busy for the coming years.