Journal of Universal Computer Science, vol. 23, no. 11 (2017), 1035-1037
submitted: 15/10/17, accepted: 26/11/17, appeared: 28/11/17

Big Data in Cross-Disciplinary Research

J.UCS Focused Topic

Giangiacomo Bravo
(Linnaeus University, Växjö, Sweden
giangiacomo.bravo@lnu.se)

Mikko Laitinen
(Linnaeus University, Växjö, Sweden
and
University of Eastern Finland, Joensuu, Finland
mikko.laitinen@lnu.se; mikko.laitinen@uef.fi)

Magnus Levin
(Linnaeus University, Växjö, Sweden
magnus.levin@lnu.se)

Welf Löwe
(Linnaeus University, Växjö, Sweden
welf.lowe@lnu.se)

Göran Petersson
(Linnaeus University, Kalmar, Sweden
goran.petersson@lnu.se)

This focused topic presents four investigations from four fields in which the focus is either on utilizing big data or charting the benefits of using such evidence in basic research. When the initial steps for this publication were taken, the editors envisaged the overarching theme to be 'from hype to solid cross-disciplinary research'. At the time, there certainly was hype around big data, as is clearly seen in Figure 1 below. It visualizes the occurrences of 'big data' in big data. It is based on a very rudimentary, yet illustrative, search of this phrase in a 5.7 billion-word text corpus of newspapers in 20 English-speaking countries in the News on the Web text corpus NOW between February 2010 and autumn 2017 (see https://corpus.byu.edu/now/). The results show that the use of 'big data' peaked in public discourse around 2014–15, but it is now decreasing and is probably being replaced by other buzz terminology, such as 'deep learning', 'edge computing', and 'IoT platforms', etc.

This informal observation regarding the hype around big data can be contrasted to the so-called hype cycles that are occasionally used in the commercial world to distinguish buzz and identify long-term commercial viability. One such (randomly selected) example is the Gartner Hype Cycles (see https://preview.tinyurl.com/ycvac3ob), which provides graphic representations of the stages in which technologies and applications are being adopted. One of the various cycles is the one depicting the emerging

technologies cycle. It contains five phases (i.e. Technology Trigger, Peak of Inflated Expectations, Trough of Disillusionment, Slope of Enlightenment, and Plateau of Productivity). In the first phase, technological breakthroughs lead to first proof-of-concept stories, giving rise to considerable publicity. These first fables, one might say, often lead to inflated expectations in the second phase. In the disillusionment stage, the interest starts to decrease as the majority of the endeavors fail. In the last two phases, technologies and their benefits first become more widely understood in the enlightenment state, and in the plateau of productivity, a certain technology is adopted by mainstream. While this hype cycle clearly refers to commercial applications, analogies can also be made to how buzzwords are adopted in the academic world.

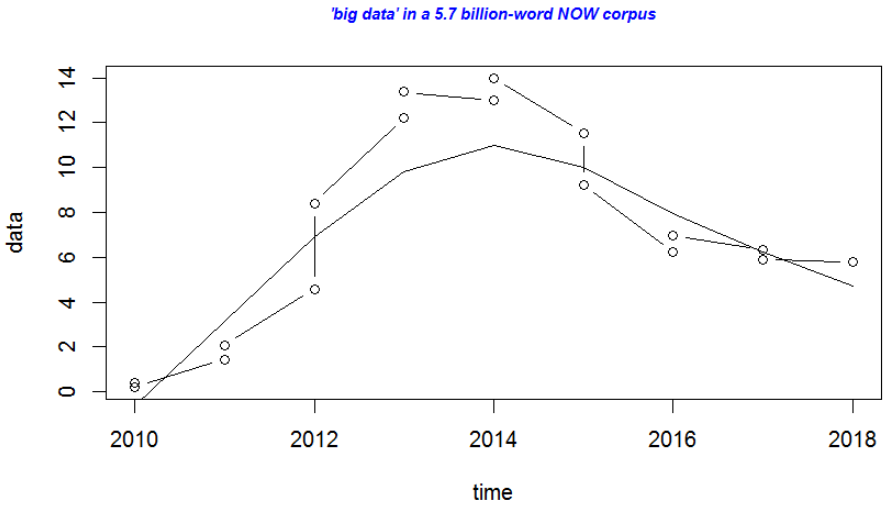

Figure 1: The use of 'big data' in public discourse in the NOW text corpus

If we use the NOW data as an estimate, it is fair to claim that with regard to big data as a buzzword in public discourse, we are now clearly past the Peak of Inflated Expectations. Most importantly, we hope that the contributions in this special issue illustrate that rather than being in the Trough of Disillusionment, we are nearing the Slope of Enlightenment and even the Plateau of Productivity.

What is, however, clear at the time of writing is that the basic tenets of using big data in cross-disciplinary research are even more obvious than they were when this special issue was first planned, and our initial idea still holds when 'big data' has become a household term in various fields. Massive amounts of data are available through various channels, and immense possibilities exist in capturing and storing data. Moreover, high-power computing allows large datasets to be mined and modelled efficiently. Data can now be effectively transferred, combined, and annotated, and powerful visualization tools can be used to analyze various types of data. These opportunities offer immense possibilities for research projects that cross disciplinary boundaries.

The general editors of this issue come from a range of fields, and space permits us to review similar work only cursorily. Yet, it is fair to say that big and rich data from nearly all areas of human life, from library and healthcare databases and from social media to IoT devices, and industry sensors have turned the world as we know it into a massive repository of information in ever-increasing numbers of areas.

For instance, English linguists are fortunate because English has left a more fully documented written history than possibly any other language. For instance, Google Books, the largest megacorpus with over two trillion words, offers access to millions of books from the Middle Ages to the present day. Such data sources alongside social media platforms, such as Twitter, have been used to provide not only new answers to old questions in (socio)linguistics, but they have also led to completely new questions in the field. In addition, big data from social media platforms have been used in sentiment analysis and predictive data mining in political sciences, finances, and business studies/marketing (an overview in Laitinen et al. this volume).

Twitter and other social media data, along with data from several other IT sources, have also become crucial for computational social sciences. For the first time in history, social sciences have access to large quantities of data reflecting actual human behavior (in contrast with the reported behavior typical of traditional surveys) and structured in large and complicated networks of interaction. This represents a crucial methodological and theoretical challenge for the social sciences, which far too often seem slow to react and risk of being marginalized by other more dynamic disciplines focusing on human behaviors from different perspectives (see Bravo and Farjam in this volume).

Industry 4.0 and digitalization are yet another hype that is about to prove its long-term commercial viability. This one enriches plants and logistic chains with sensor data. The aggregation of raw sensor data to higher-value information and actionable knowledge not only leverages automation and optimization of manufacturing processes, but it also creates new business opportunities and transforms traditional industries into IT businesses (see Heberle et al. in this volume).

Library and information sciences represent another field that has experienced major changes following the advent of the big data era. Golub and Hansson (this volume) argue for the need to find new ways of organizing documentation, information and data, and to develop instruments for research evaluation in a world where the differences between formal publications, other forms of communication and data sharing are becoming weaker.

We are only beginning to comprehend the effects and trajectories of what digitalization, technologization, and massive amounts of data mean for basic research. The prospects are immense, but only if practitioners build on solid theoretical foundations and make use of data to suggest alternatives and to construct new what-if scenarios.