European Journal of Information Systems

ISSN: 0960-085X (Print) 1476-9344 (Online) Journal homepage: https://orsociety.tandfonline.com/loi/tjis20

Utilizing big data analytics for information systems research: challenges, promises and guidelines

Oliver Müller, Iris Junglas, Jan vom Brocke & Stefan Debortoli

To cite this article: Oliver Müller, Iris Junglas, Jan vom Brocke & Stefan Debortoli (2016) Utilizing big data analytics for information systems research: challenges, promises and guidelines, European Journal of Information Systems, 25:4, 289-302, DOI: 10.1057/ejis.2016.2

To link to this article: https://doi.org/10.1057/ejis.2016.2

European Journal of Information Systems (2016) 25, 289–302

www.palgrave-journals.com/ejis/

Open

ISSUES AND OPINION

Utilizing big data analytics for information systems research: challenges, promises and guidelines

Oliver Müller[1], Iris Junglas[2], Jan vom Brocke[1] and Stefan Debortoli[1]

[1] Institute of Information Systems, University of Liechtenstein, Vaduz, Liechtenstein; [2] College of Business, Florida State University, Tallahassee, U.S.A.

Correspondence: Oliver Müller, Institute of Information Systems, University of Liechtenstein, Fürst-Franz-Josef-Strasse, Vaduz 9490, Liechtenstein
Tel: +423 265 13 00;
Email: oliver.mueller@uni.li

Abstract
This essay discusses the use of big data analytics (BDA) as a strategy of enquiry for advancing information systems (IS) research. In broad terms, we understand BDA as the statistical modelling of large, diverse, and dynamic data sets of user-generated content and digital traces. BDA, as a new paradigm for utilising big data sources and advanced analytics, has already found its way into some social science disciplines. Sociology and economics are two examples that have successfully harnessed BDA for scientific enquiry. Often, BDA draws on methodologies and tools that are unfamiliar for some IS researchers (e.g., predictive modelling, natural language processing). Following the phases of a typical research process, this article is set out to dissect BDA's challenges and promises for IS research, and illustrates them by means of an exemplary study about predicting the helpfulness of 1.3 million online customer reviews. In order to assist IS researchers in planning, executing, and interpreting their own studies, and evaluating the studies of others, we propose an initial set of guidelines for conducting rigorous BDA studies in IS.
European Journal of Information Systems (2016) **25**(4), 289–302.
doi:10.1057/ejis.2016.2; published online 9 February 2016

Keywords: big data; analytics; data source; methodology; information systems research

Why worry about big data analytics (BDA)?

The proliferation of the web, social media, mobile devices, and sensor networks, along with the falling costs for storage and computing resources, has led to a near ubiquitous and ever-increasing digital record of computer-mediated actions and communications – a record that has been termed 'big data'. Studies agree (e.g., Hilbert & López, 2011; IDC, 2011, 2014) that the volume of data being generated and stored today is growing exponentially (Kitchin, 2014) – but, big data is not just about volume (Constantiou & Kallinikos, 2015; Yoo, 2015). Accompanying the increased size of data sets is a growing variety of data sources and formats. In fact, IDC (2011) claims that more than 80% of all digital content is unstructured and that two-third is generated by individuals, rather than enterprises. The velocity of data has increased too, resulting in reduced latency between the occurrence of a real-world event and its mirroring digital footprint (vom Brocke *et al*, 2014).

As volume, variety, and velocity (Laney, 2001) increase, the veracity of data is drawn into question (Bendler *et al*, 2014). Unlike research data collected with a specific research question in mind and measured using validated instruments, big data often just 'happens'. Private and government

Received: 18 November 2014
Revised: 7 December 2015
Accepted: 1 January 2016

organisations increasingly collect big data without a concrete purpose in mind, hoping that it might provide value someday in the future (Constantiou & Kallinikos, 2015; Yoo, 2015). Marton *et al* (2013) see this 'shift to ex-post ordering' as a defining characteristic of big data:

> [Big data] can be ordered, analysed and thus made potentially informative in ways that are not predefined. The potential of big data to inform is based on ordering as much data as possible through automated computational operations after its collection; i.e., in an ex-post fashion (p. 6).

Adopting such a data collection and analysis approach for research purposes is questionable as the validity and reliability of the data cannot be assumed (Agarwal & Dhar, 2015). Furthermore, the use of big data for research raises ethical questions as individuals may not be aware that their digital footprint is being analysed for scientific purposes – or that it is recorded at all (Zimmer, 2010).

Analysing data characterised by the '4Vs' (i.e., volume, velocity, variety, and veracity) necessitates advanced analytical tools. Big data is not self-explanatory and without the application of inferential computational techniques to identify patterns in the data, it is just noise (Boyd & Crawford, 2012; Dhar, 2013; Agarwal & Dhar, 2015; Yoo, 2015). But applying standard statistical methods to big data sets or using advanced analytics for its analysis bears potential pitfalls. For example, some data mining or machine-learning techniques aim at prediction instead of explanation (Shmueli & Koppius, 2011), and are frequently labelled as 'black boxes' that hide their inner workings and make it difficult to interpret their outputs (Breiman, 2001b). For IS researchers, adopting such methods therefore can be challenging.

The first scientific disciplines to add BDA to their methodological repertoire were natural sciences (see, e.g., Hey *et al*, 2009). Slowly, the social sciences also started capitalising on the computational analysis of digital footprints (Lazer *et al*, 2009). For instance, researchers studied cultural trends, such as the adoption of new technologies, by sequencing word frequencies in digitised texts (Michel *et al*, 2011), used social media to predict daily swings of the stock market (Bollen *et al*, 2011), and tested the effect of manipulating the emotionality of social network news feeds in a field experiment with hundreds of thousands of participants (Kramer *et al*, 2014). Recently, the IS field has also published its first BDA studies (e.g., Ko & Dennis, 2011; Kumar & Telang, 2012; Oestreicher-Singer & Sundararajan, 2012; De *et al*, 2013; Goh *et al*, 2013; Stieglitz & Dang-Xuan, 2013; Zeng & Wei, 2013). Still, the novelty of BDA necessitates new norms and practices to be established on how to successfully apply BDA in IS research. Researchers who want to add BDA to their methodological toolbox must be capable of planning, executing, and interpreting their studies, and be equipped to understand and evaluate the quality of others (Agarwal & Dhar, 2015). In addition, following the arguments laid out in Sharma *et al* (2014), researchers have to acknowledge that 'insights do not emerge automatically out of mechanically applying analytical tools to data' (p. 435) and that '[i]nsights … need to be leveraged by analysts and managers into strategic and operational decisions to generate value' (p. 436). Just like the application of big data and advanced analytics in business does not automatically lead to better decisions and increased business value, researchers have to overcome a number of challenges in order to leverage big data and advanced analytics for generating new insights and translating them into theoretical contributions.

This essay is set out to shed light on the challenges that the application of BDA might bring, as well as the potentials it may hold when IS researchers choose to utilise big data and advanced analytical tools. Both are illustrated by an exemplary study using natural language processing (NLP) and machine-learning techniques to predict the helpfulness of 1.3 million online customer reviews collected from Amazon.com. On the basis of the discussion and illustration of the challenges and promises of BDA, we develop an initial set of guidelines for conducting rigorous BDA studies in IS.

The remainder of this article roughly mirrors the phases of a typical research process. We discuss issues related to appropriately framing a BDA research question, the nature of data collection, its computational analysis, and the interpretation of results. For each phase, we discuss the challenges and the promises of BDA for IS research. Subsequently, we provide an exemplary illustration of a BDA study and a set of initial guidelines for IS researchers who choose to apply BDA as part of their research endeavour.

Research question: appropriately framing a BDA study

Analytics methodologies

The literature on research methods and statistics traditionally distinguishes two broad types of approaches to modelling (see, e.g., Breiman, 2001b; Gregor, 2006). Explanatory modelling aims to statistically test theory-driven hypotheses using empirical data, while predictive models aim to make predictions about future or unknown events, using models trained on historical data. In general, the quantitative realm of IS research has a strong emphasis on developing and testing explanatory causal models – in fact, compared with predictive approaches, the '[e]xtant IS literature relies nearly exclusively on explanatory statistical modeling' (Shmueli & Koppius, 2011, p. 553). In contrast to explanatory studies, predictive studies allow for an inductive and data-driven discovery of relationships between variables in a given data set. A researcher does not have to explicitly formulate testable *a priori* hypotheses about causal relationships, but can leave it to the algorithm to uncover correlations between variables (Breiman, 2001b; Shmueli, 2010; Shmueli & Koppius, 2011; Dhar, 2013; Agarwal & Dhar, 2015). In this process, theory merely plays the role of an analytical lens providing a broad direction for the data collection and analysis process.

BDA studies can follow an explanatory or predictive approach. Yet, researchers need to be aware of the challenges and potential pitfalls that can arise from the use of big data sets or data-driven analytical methods. For example, in explanatory studies researchers have to take care when interpreting *P*-values of statistical hypothesis tests on very large samples. As Lin *et al* (2013) outline, when using traditional hypothesis tests in large samples ($n > 10{,}000$), the increase in statistical power lets *P*-values converge quickly to 0 so that 'even minuscule effects can become statistically significant' (p. 1). This may lead to findings that have high statistical, but little practical significance. Hence, when interpreting explanatory BDA studies '[t]he question is not whether differences are "significant" (they nearly always are in large samples), but whether they are interesting' (Chatfield, 1995, p. 70; as quoted in Lin *et al*, 2013).

When it comes to predictive studies, many researchers apply data mining or machine-learning algorithms. Owing to their data-driven nature, they are especially well suited to autonomously sift through data sets that are characterised by massive volume, a broad variety of attributes, and high velocity (e.g., real-time data streams). However, such a 'data dredging' approach is not quite in line with the traditional scientific method. Hence, predictive analytics has been described as unscientific and fixated on applied utility (Breiman, 2001b; Shmueli, 2010). In fact, many of the algorithms used in BDA were designed for practical applications, such as credit risk scoring or recommending products to individual customers, but 'science, unlike advertising, is not about finding patterns – although that is certainly part of the process – it is about finding explanations for those patterns' (Pigliucci, 2009; p. 534). Relying solely on correlation instead of causality might lead to the identification of patterns that not only provide little theoretical contribution, but are also vulnerable to changes and anomalies in the underlying data. The errors in flu prediction made by Google Flu Trends are a prime example of this pitfall. What used to be praised as a ground-breaking example of BDA, missed the A-H1N1 pandemic in 2009 and dramatically overestimated flu prevalence in 2012 and 2013. Google Flu Trends was over-fitting historical data, because the statistical model behind its predictions exclusively relied on inductively identified correlations between search terms and flu levels and not on biological cause-and-effect relationships; it also had only been trained on seasonal data containing no anomalies (e.g., pandemics), and it had not been updated on an ongoing basis to reflect changes in users' information seeking behaviours (Lazer *et al*, 2014). Since the *raison d'être* of the IS field is to study socio-technical systems – and not systems governed by natural laws – it therefore seems unreasonable to solely rely on inductively uncovered correlations that can change over time and contexts.

BDA for theory development

Despite the challenges, however, BDA might still be a valuable tool for IS research – especially for theory development. The massive size of big data sets allows for the detection of small effects, the investigation of complex relationships (e.g., interactions), a comparison of subsamples with one another (e.g., different geographies or time frames), the incorporation of control variables into the analysis, and the study of effects so rare that they are seldom found in small samples (e.g., fraud) (Lin *et al*, 2013). Data-driven predictive algorithms can also help to advance theory (Dhar, 2013; Agarwal & Dhar, 2015) as 'patterns [often] emerge before reasons for them become apparent' (Dhar & Chou, 2001, p. 907). In the medical field, for example, the C-Path (Computational Pathologist) system makes data-driven predictions about malignant breast tissue by examining microscopic images (Beck & Sangoi, 2011). Its predictions are not only as accurate as that of experts, but C-Path has also been able to discover new markers that are indicative of breast cancer (stromal morphologic structures) and that have been ignored by human pathologists thus far (Rimm, 2011; Martin, 2012).

IS researchers considering BDA might have to grow comfortable with the idea that research can start with data or data-driven discoveries, rather than with theory. For qualitative-interpretive researchers and those familiar with methodologies such as grounded theory, this way of thinking is not unfamiliar. Grounded theory, per definition, is concerned with the inductive generation of new theory from data, and thus less concerned about existing theory (Glaser & Strauss, 1967). It utilises a rigid methodological approach to arrive at new theory and is predominantly based upon the empirical analysis of (mostly) qualitative data, sometimes termed as 'empirically grounded induction' (Berente & Seidel, 2014, p. 4). IS researchers choosing to apply a BDA approach might want to consider some of the principles of grounded theory. Like grounded theorists, BDA researchers will spent an extraordinary amount of time on understanding the nature of the data, particularly if they have not collected them themselves. While they do not have to manually code the data, the ideas behind open, iterative, and selective coding still apply (Yu *et al*, 2011). Researchers need to be open to surprises in the data, iteratively analyse and re-analyse it, and zoom in on relevant concepts that can make a contribution. Likewise, theoretical coding, where codes and concepts are weaved into hypotheses in order to form a model, seems equally applicable for BDA researchers. As a model is said to emerge freely (and not by force), applying the principles of grounded theories can aid researchers in turning a research interest into a suitable research question. A dedicated exploratory phase where a researcher can probe the feasibility and vitality of his or her interest is not only helpful, but ensures a rigorously chosen research objective. Apart from an exploratory phase, BDA researchers, like grounded theorists, might also want to consider a phase of theoretical triangulation in which the various concepts identified are compared against extant theory. Testing and re-testing the emerging research question against existing theory can help to identify overlaps, as

well as white spots and will ensure that the emerging research question is truly unique and novel in nature.

Data collection: the nature of big data

More than size

There is more to 'big' data than the '4Vs' (Yoo, 2015). Big data tries to be exhaustive in breadth and depth and more fine-grained in resolution than traditional data, often indexing individual persons or objects instead of aggregating data at the group level (Kitchin, 2014). In economics, for example, big data collected by the Billion Prices Project (http://bpp.mit.edu/) complements traditional economic statistics by providing faster and more fine-granular data about economic activities (Einav & Levin, 2014). The project's daily price index, considering more than 5 million items from more than 300 online shops in more than 70 countries, not only correlates highly with the traditional consumer price index, but also provides pricing data for countries in which official statistics are not readily available (Cavallo, 2012).

Another characteristic that sets big data apart is related to its purpose (Marton *et al*, 2013; Constantiou & Kallinikos, 2015; Yoo, 2015). Big data is a by-product of our technological advancements and, more often than not, collected without a specific purpose in mind. Amazon's CEO Jeff Bezos, for instance, is known for saying 'we never throw away data' (Davenport & Kim, 2013). But for researchers there are big differences in access levels to those data sources. For example, Twitter grants access to its Firehose API that allows real-time access to the complete stream of tweets for selected partner businesses and a handful of research groups only (Boyd & Crawford, 2012). Others have access to only small and idiosyncratically selected sub-samples of tweets (about 10% of all public tweets). Hence, for the broader scientific community it is virtually impossible to replicate studies using Twitter data. Also, differences in accessibility could possibly create a 'data divide' (Marton *et al*, 2013) and split the research community into those that have access to big data and those that have not. Already today, the authors of many of the best known BDA studies are affiliated with big data companies, including Google (Ginsberg *et al*, 2009; Michel *et al*, 2011) and Facebook (Kramer *et al*, 2014). Finally, the use of big data for research purposes also prompts ethical questions, especially because of its fine-granular, indexical nature (Kitchin, 2014). While some individual-level data is willingly provided by users, often gladly, others are not. Hence, pursuing research objectives unequivocally – simply because the data is available and accessible – should prompt some ethical considerations from IS researchers. For instance, a study conducted by Kramer *et al* (2014) caused a public outcry when it was published. The authors conducted a field experiment on Facebook, in which they secretly manipulated the news feeds of almost 700,000 users by filtering out messages with emotional content in order to test if this had an effect on the sentiments expressed in their own posts. The results showed that when messages with positive emotions were cut out, users subsequently posted less positive and more negative messages. The opposite happened when the number of messages with negative emotions was reduced. While both the method and the results were highly interesting from a scientific standpoint, it caused numerous discussions about informed consent and real-world online experiments.

Naturally occurring and open

Do these challenges imply that IS researchers should refrain from using big data for research purposes entirely? Up to now, the IS field has almost exclusively relied on self-reports or observations collected via instruments like surveys, experiments, or case studies (Hedman *et al*, 2013). While these strategies of enquiry have many advantages (including tight controls), they also are costly and subject to biases.

In contrast to data provoked by researchers through interviews, surveys, or experiments, big data is predominantly user-generated, or 'naturally occurring' (Speer, 2002; Silvermann, 2006; Hedman *et al*, 2013). It is not collected with a specific research purpose in mind and, hence, may be less contrived. As a result, naturally occurring data has the potential to alleviate many of the methodological biases originating from the use of data collection instruments (Speer, 2002). Whereas traditional instruments rely on self-reports, collect data after the fact, and are drawn from a confined pool of subjects, naturally occurring data could possibly reduce instrument biases as behaviours and attitudes are gathered in an unobtrusive way and at the time and context in which they occur, from a sampling pool that possibly comprises the entire population or at least a pool that spans a wide variety of geographical, societal, and organisational borders. Simultaneously, we have to acknowledge that even social networks, such as Facebook and Twitter with millions of active users, are self-selected tools and thus prone to non-response biases. Also, not everybody uses the tool to the same extent, and some 'users' are not even humans, but bots (Boyd & Crawford, 2012). But irrespective of these shortcomings, the pool for sampling is by far greater than that targeted with traditional methods.

Naturally occurring data may or may not be open. In the natural sciences, several open research data repositories have emerged in the last years (Murray-Rust, 2008). The Global Biodiversity Information Facility (GBIF), for example, hosts more than 15,000 data sets with more than 400 million indexed records published by over 600 institutions, capturing the occurrence of organisms over time and across the planet. Since 2008, more than 900 peer-reviewed research publications identified the use of GBIF data, allowing a wide variety of researchers to scrutinise its content and draw conclusions about its validity. Having a similar platform for IS-related data sets available would not only increase the validity of data captured in the field, but

also boost the replicability of studies. Such a data library could even become part of the publication process. *The Journal of Biostatistics*, for example, encourages authors to publish their data set and the code for analysing the data alongside their accepted paper. A dedicated 'Associate Editor for Reproducibility' then runs the code on the data to check if it indeed produces the results presented in the article (Peng, 2011). The next step would be to publicly provide 'linked and executable code and data' (Peng, 2011, p. 5), for example, in the form of an interactive computational environment (e.g., IPython Notebook) that allows readers to inspect and explore the original data and reproduce all data analysis steps.

Data analysis: the computational enquiry of unstructured data

Quantifying text

Without automated statistical inference, big data turns into a burden rather than an opportunity. Yet, statistical analysis only applies to structured numeric and categorical data; and it is estimated that more than 80% of today's data is captured in an unstructured format (e.g., text, image, audio, video) (IDC, 2011), much of it expressed in ambiguous natural language. The analysis of such data would usually prompt the use of qualitative data analysis approaches, such as reading and manual coding, but the size of data sets obtained from sources like Twitter or Facebook deems any kind of manual analysis virtually impracticable. Hence, researchers have turned to NLP techniques to enable the automated analysis of large corpora of textual data (Chen *et al*, 2012).

But the statistical analysis of textual data only scratches the surface of how humans assign meaning to natural language. For example, most topic models (i.e., algorithms for discovering latent topics in a collection of documents) treat texts as simple unordered sets of words, disregarding any syntax or word order (Blei *et al*, 2003; Blei, 2012). Consequently, such techniques struggle with linguistic concepts, such as compositionality, negation, irony, or sarcasm (Lau *et al*, 2013). Further challenges arise from a lack of information about the context in which data have been generated or collected. Every unit of text is treated in an egalitarian way irrespective of who authored it or what the situation was like when the text was created (Lycett, 2013). While the current properties of NLP are in stark contrast to the assumptions and practices of traditional qualitative social science research, which emphasises features like natural settings, local groundedness, context, richness, and holism (Miles & Huberman, 1994), the application of NLP in social science research has already produced interesting findings. For example, Michel *et al* (2011) sequenced the development of cultural trends based on the content of more than 5 million digitised books, representing about 4% of all books ever printed. The texts were examined using a simple, yet insightful, technique called *n*-gram analysis that simply counts how often a word is used over time. By computing the yearly relative frequency of all words found, the authors were able to quantify cultural phenomena, such as the adoption of technology. For example, the study found that modern societies are adopting inventions at an accelerating rate. While early inventions (in the timeframe from 1800 to 1840) took over 66 years from invention to widespread adoption (measured by the shape of the frequency distribution of the corresponding word over time), the average time-to-adoption dropped to 50 years for the timeframe from 1840 to 1880, and to 27 years for the timeframe from 1880 to 1920.

Numerous empirical studies and real-world applications have shown that 'simple' NLP models fed with sufficiently large data sets can produce surprisingly accurate results (Halevy *et al*, 2009). An experimental evaluation of the Latent Semantic Analysis (LSA) topic modelling algorithm, for example, has shown that the agreement between LSA and human raters when coding clinical interview transcripts reaches 90% (Dam & Kaufmann, 2008). And a study on neural sentiment analysis found agreements between 80 and 85% for sentiment classifications of movie reviews, performed independently by human raters and machine algorithms (Socher *et al*, 2013).

Overcoming the quantitative–qualitative divide

An increase in inter-coder reliability between human raters and machines also has the potential to close the divide between quantitative and qualitative methods – a dilemma that has long captured the attention of IS researchers (e.g., Venkatesh *et al*, 2013). Implementing techniques like topic modelling or sentiment analysis into software tools for qualitative data analysis (e.g., NVivo, Atlas.ti) could reduce the time required for qualitative data analysis and permit researchers to work with much larger samples, enable the discovery of subtle patterns in texts, and allow mixed method studies. Still, for IS researchers that decide to apply a BDA approach for studying qualitative data it is important to dedicate a significant amount of time triangulating their empirical results. This additional phase not only ensures the validity of the analysis by placing it in the context of other studies, but also adds, over time, to a growing body of studies that apply a particular BDA approach. Meticulously documenting the data analysis procedure is therefore critical. As the spectrum of NLP algorithms is already extensive, continuously in flux and advancing, ensuring the traceability of the underlying algorithm is key. Likewise, IS researchers might also want to consider collaborating with non-IS researchers when conducting BDA studies, particularly with colleagues from the fields of computer science and machine learning. Apart from an advanced understanding of the phenomenon, this ensures the development of new skill sets and the advancement of the IS field through newly emerging methodological tools.

Result interpretation: opening up the 'black box' of analytics

Accuracy *vs* interpretability

To analyse big data – be it qualitative or quantitative in nature – researchers not only apply traditional inferential statistics (e.g., analysis of variance, regression analysis), but also increasingly make use of data mining and machine-learning algorithms (see, e.g., Wu *et al,* 2008; KDnuggets, 2011; Rexer, 2013 for an overview) – especially when the study's objective is prediction (Breiman, 2001b; Shmueli, 2010; Kuhn & Johnson, 2013). It is widely documented that such algorithmic methods can outperform simpler statistical models (e.g., linear regression) in terms of their predictive accuracy, because they make less statistical assumptions, can work with high-dimensional data sets, are able to capture non-linear relationships between variables, and automatically consider interaction effects between variables (Breiman, 2001b; Shmueli, 2010; Kuhn & Johnson, 2013).

Yet, this improvement in predictive accuracy comes at a cost. Some of the most accurate algorithms, such as support vector machines or random forests, are largely incomprehensible (Martens & Provost, 2014). In other words, these 'black box' algorithms are good in predicting future or unknown events, but unable to provide explanations for their predictions. When analysis results are intended to be consumed by machines, for instance, when placing online ads or ranking search results, this lack of transparency may be coped with (Lycett, 2013). Yet, for human decision makers, or when stakes are higher, the interpretability and accountability of an algorithm grows in importance (Diakopoulos, 2014). In the financial industry, for example, regulatory requirements demand that all automated decisions made by credit scoring algorithms are transparent and justifiable, for example, to avoid discrimination of individuals. As a consequence, today's credit scoring systems are based on simple linear models with a limited number of well-understood variables – even though these models could be easily outperformed by non-linear models working on high-dimensional data sets (Martens *et al*, 2007). Likewise, in the medical field it is paramount for doctors to understand the decisions made by algorithms and to be able to explain and defend them against patients, colleagues, and insurance providers (Breiman, 2001b). ' "Doctors can interpret logistic regression." There is no way they can interpret a black box containing fifty [decision] trees hooked together. In a choice between accuracy and interpretability, they'll go for interpretability' (Breiman, 2001b, p. 209). These examples illustrate that good decisions are decisions that are both of high quality and accepted by those responsible for implementing them (Sharma *et al*, 2014).

Explaining predictions

The difficulties of interpreting the results and comprehending the underlying algorithm are well-documented drawbacks of data mining and machine-learning algorithms. Researchers have long called upon fellow researchers to provide explanations for their systems' outputs (Gregor & Benbasat, 1999). Today, there is a growing body of research seeking to disclose the inner workings of 'black box' data mining and machine-learning algorithms by bolstering their highly accurate predictions with easy to comprehend explanations (Martens *et al*, 2007; Robnik-Sikonja & Kononenko, 2008; Kayande *et al*, 2009). A dedicated explanatory phase is therefore vital for ensuring the interpretability of predictive models. Such explanations can take many forms, including comparing and contrasting predictions with existing theories and related explanatory empirical studies (Shmueli & Koppius, 2011), 'replaying' predictions for extreme or typical cases (Robnik-Sikonja & Kononenko, 2008), performing a sensitivity analysis for selected model inputs and visualising the corresponding model inputs and outputs (Cortez & Embrechts, 2013), or extracting comprehendible decision rules from complex ensemble models (Martens *et al*, 2007). In this way, a researcher can first use data mining or machine-learning algorithms to identify correlated variables and then use more traditional statistics on the most relevant variables to read and interpret the identified patterns.

Illustrative BDA study

In the previous sections, we have discussed the challenges and promises of BDA along the various phases of the research process, including the definition of the research question, the nature of data collection, its computational analysis, and the interpretation of results. In this section, we illustrate how BDA can be applied in IS research by presenting an integrated, exemplary BDA study from the area of online customer reviews. The presentation of this study mirrors the structure of the previous chapters and with it the various phases of the research process.

Research question

As our research objective, we have chosen to revisit the question of 'What makes a helpful online review?' (Mudambi & Schuff, 2010). Online customer reviews, defined as 'peer-generated product evaluations posted on company or third-party websites' (Mudambi & Schuff, 2010; p. 186), are ubiquitous testaments of product experiences and are shown to influence a customer's decision-making process (BrightLocal, 2014). While most studies on review helpfulness in IS research have been explanatory (e.g., Mudambi & Schuff, 2010; Cao *et al*, 2011), our exemplary BDA study is predictive in nature. Specifically, it predicts perceived review helpfulness based on the characteristics of the review itself, that is, users' numerical product ratings and textual comments.

A predictive model for review helpfulness might be valuable for practical and theoretical reasons. It might be able to determine the proper sorting and filtering of new reviews or to pinpoint how to write more effective reviews. In addition, and because of the inductive and data-driven

nature of many machine-learning and data mining algorithms used for prediction, a predictive model for review helpfulness might also hold the potential to contribute to theory by discovering to date unknown relationships between review characteristics and perceived review helpfulness.

Data collection

Leading e-commerce platforms like Amazon, TripAdvisor, or Yelp allow users not only to write reviews, but also to rate the helpfulness of other users' reviews. To collect this type of data for analysis, it can either be crawled and scraped directly from the originating websites (which may be against the terms of use of some websites), or downloaded from open data repositories. For this example, we used the open Amazon product reviews data set curated by McAuley and colleagues (McAuley *et al*, 2015a, b) and publicly available at http://jmcauley.ucsd.edu/data/amazon/. The data set comprises 143.7 million reviews of 9.4 million products, spanning a period of 18 years from May 1996 to July 2014. In the following, we focus on the 1.3 million reviews available for the video games product category. We chose video game reviews as the unit of analysis as video games represent a type of software that customers happily discuss online; also, video games are representative of a strong market with a total revenue of more than U.S.$22 billion in 2014 (Entertainment Software Association, 2015).

While customers assign high relevance to peer-generated online reviews and even trust them as much as personal recommendations (BrightLocal, 2014), it is important to note that reviews can be subject to biases. For example, it is well-known that online reviewers tend to be extreme in their rating, leaning either to the very positive or very negative end of the spectrum (Hu *et al*, 2009). Nonetheless, the use of real reviews as a data source is common practice in the IS field (e.g., Mudambi & Schuff, 2010; Cao *et al*, 2011; Ghose & Ipeirotis, 2011), thereby acknowledging its validity and reliability for research purposes. The richness in information that online reviews provide, combined with the high volume of online reviews available on the web, make them ideal for studying customer information-seeking and decision-making behaviours.

Although McAuley and colleagues' data set has already been cleaned to some extent (i.e., duplicate reviews have been removed), we pre-processed and transformed the data even further for our illustrative example. First, an exploratory data analysis revealed that about half of the reviews did not possess a single helpfulness rating (neither positive nor negative). We excluded reviews with less than two helpfulness ratings in order to increase the reliability of our analysis. This reduced our data set to 495,358 video game reviews. A second issue we encountered was related to the distribution of the dependent variable, that is, review helpfulness. Following prior research (e.g., Mudambi & Schuff, 2010), we measured review helpfulness as the ratio of helpful votes to total votes for a given review. The resulting measure showed a *w*-shaped distribution. In order to avoid statistical concerns arising from the extreme distribution of values, we dichotomised the review helpfulness variable (i.e., reviews with a helpfulness ratio of > 0.5 were recoded as helpful, and reviews $\leqslant 0.5$ as not helpful) (Lesaffre *et al*, 2007).

Data analysis

While early research on review helpfulness mostly focused on quantitative review characteristics, especially the numerical 'star' rating, more recent studies have started to analyse their textual parts (e.g.,Mudambi & Schuff, 2010; Cao *et al*, 2011; Ghose & Ipeirotis, 2011; Pan & Zhang, 2011; Korfiatisa *et al*, 2012). Yet, these studies largely focus on easily quantifiable aspects, including syntactic or stylistic features such as review length, sentiment or readability, and mostly neglect the actual content of the review text (i.e., what topics is a reviewer writing about?). In order to capture a review's content and its impact on review helpfulness, we applied probabilistic topic modelling using the Latent Dirichlet Allocation (LDA) algorithm.

Probabilistic topic models are unsupervised machine-learning algorithms that are able to discover topics running through a collection of documents and annotate individual documents with topic labels (Blei *et al*, 2003; Blei, 2012). The underlying idea is the 'distributional hypothesis' of statistical semantics, that is, words that co-occur together in similar contexts tend to have similar meanings (Turney & Pantel, 2010). Consequently, sets of highly co-occurring words (e.g., 'work', 'doesn't', 'not', 'download', 'install', 'windows', 'run') can be interpreted as a topic (e.g., installation problems) and can be used to annotate documents with labels (Boyd-Graber *et al*, 2014). In statistical terms, each topic is defined by a probability distribution over a controlled vocabulary of words, and each document is assigned a probabilistic distribution over all the topics. The per-document topic distribution vector represents a document's content at a high level and can be used as input for further statistical analysis (e.g., predictive modelling).

Using LDA we extracted the 100 most prevalent topics from our corpus of video game reviews and annotated each review with a vector of topic probabilities. (We used the LDA implementation provided by Mine MyText.com and pre-processed all texts by filtering out standard stop words (e.g., 'and', 'the', 'I') as well as a short list of custom stop words (e.g., 'game', 'play') and lemmatizing all words of the corpus. Overall, the computation took about 6 h on a server with 24 CPU cores and 120 GB main memory. The results of the analysis are publicly available at https://app.minemytext.com/amazon-video-games-reviews.) These topic probabilities along with review rating and review length, two variables that were consistently found to be associated with review helpfulness in prior research (Trenz & Berger, 2013), were used as predictors for determining whether a new review would be perceived as helpful or not.

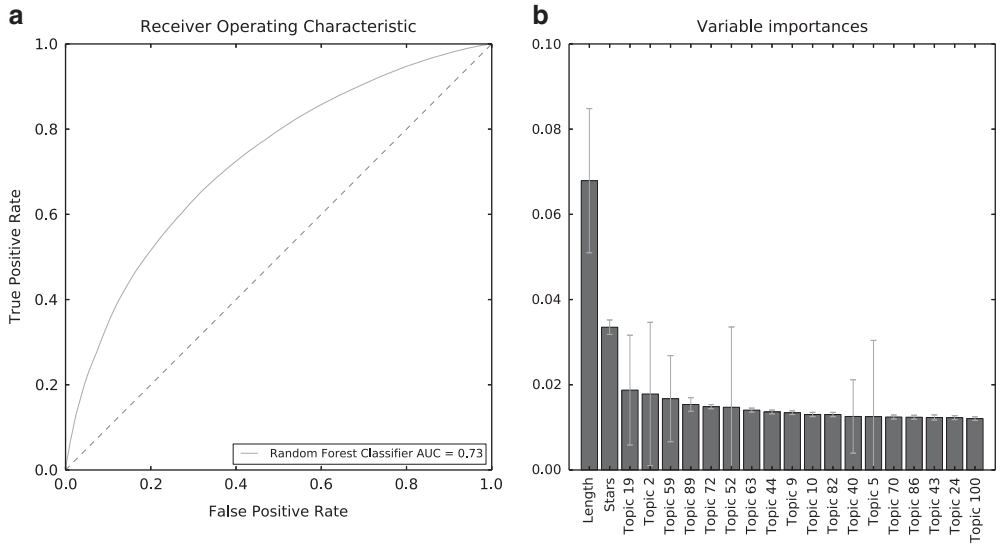

Figure 1 ROC curve and variable importance measures of random forest classifier.

For training the predictive model we used random forests, an ensemble supervised-learning technique that is able to process high-dimensional data sets, model non-linear relationships between variables, consider complex variable interactions, and is robust against data anomalies (e.g., non-normally distributed variables, outliers) (Kuhn & Johnson, 2013). A random forest model is constructed by generating a multitude of decision trees based on bootstrapped sub-samples such that only a random sample of the available variables at each split of the tree is considered a potential split candidate (Breiman, 2001a). Model training was performed on 80% of the overall data set. We used the implementation provided by the scikit-learn Python package and set the number of trees to 128. (The code and data are publicly available at http://www.minemytext.com/public-data/.)

Figure 1 (left) shows the Receiver Operating Characteristic (ROC) curve, illustrating the predictive performance of our binary classification on the holdout test set (20% of the overall data set) by plotting the true positive rate against the false positive rate. The area under the ROC curve (AUC) amounts to 0.73, which means that the model can distinguish between a randomly drawn helpful review and a randomly drawn unhelpful review with an accuracy of 73% (Fawcett, 2006). (Note that a random classifier has an AUC score of 0.5, whereas perfect classification is represented by a score of 1.0.)

Random forests usually have a high predictive accuracy, yet it can be difficult to interpret their output as they lack regression coefficients and *P*-values known from regression analysis. The only diagnostic parameters available to interpret a random forest model are the variable importance measures (i.e., mean decrease in Gini coefficient), which measure how each variable contributes to the homogeneity of the nodes and leaves in the random forest. Figure 1 (right) shows the 20 most influential variables for predicting the helpfulness of a video game review. The two most important variables are review length and star rating; the remaining variables comprise topics extracted through LDA.

In order to take a closer look at the topics extracted for predicting review helpfulness, Table 1 illustrates the most probable words associated with each topic and our interpretation of the respective topic derived by co-examining the per-topic word distributions and per-document topic distributions. Overall, the list reveals a broad range of topics that impact a review's helpfulness, including, for example, overall affective judgements (e.g., 19, 43, 82), economic factors (e.g., 10, 86), game features (e.g., 44, 72), technical issues (e.g., 89, 9), topics related to specific audiences (e.g., 63, 100), comparison with other games (e.g., 2, 5), and service-related aspects (e.g., 40, 24). One topic (52) represented Spanish reviews, mainly for the game Grand Theft Auto.

To ensure the measurement validity of our study, we empirically triangulated the LDA results, that is, the per-topic word distributions and the per-document topic distributions, with judgments of human coders by adapting a method proposed by Chang *et al* (2009). In a first step, we employed a word intrusion task to measure the semantic coherence of topics. Since topics are represented by words that co-occur with high probability, the idea behind the word intrusion task is to insert a randomly chosen word (intruder) into a set of words representative of a topic and ask human judges to identify the intruder. For each topic, we generated five randomly ordered sets of six words: the five most probable words for the given topic plus one randomly chosen word with low probability for the respective topic. We presented these sets to three independent human coders via the crowdsourcing platform Amazon Mechanical Turk and prompted them to identify the intruder. In total, this procedure resulted in 270

Table 1 Topics for predicting review helpfulness ordered by importance

Topic	*Most probable words ordered by probability*	*Interpretation*
19	bad buy make money graphic terrible horrible waste good before	Negative affect
2	call duty cod map multiplayer ops good black battlefield campaign	Call of Duty series
59	dont good buy thing alot didnt bad make people doesnt	Don't buy it
89	ea city drm server buy internet online simcity make connection	Digital rights management
72	great graphic amaze love awesome recommend story good gameplay buy	Gameplay and storyline
52	gta city de mission la grand auto theft car el	Spanish reviews
63	love son gift buy christmas great year grandson purchase daughter	Gift for a family member
44	good pretty graphic fun great bad nice cool thing lot	Graphics
9	work download window computer install run steam software problem instal	Installation problems
10	money buy worth waste time spend save pay good work	Not worth the money
82	fun great lot love good time recommend enjoy challenge easy	Fun and enjoyment
40	amazon customer send work product return back service support receive	Customer service
5	halo mutliplayer xbox campaign good reach great stroy map weapon	Halo series
70	great work good love buy prices product recommend awesome deal	Quality of console hardware
86	price buy great worth good pay amazon deal cheap find	Good deal
43	love awesome cool buy fun great good thing graphic make	Positive affect
24	review star give read buy people rate write good bad	Other customers' reviews
100	kid love fun year child young great adult age enjoy	Appropriateness for kids

comparisons between algorithm and human, who, when averaged over all topics, agreed in 71% of the cases (SD: 19%). In a second step, we conducted a best topic task to validate the topic assignments for each review. (The task is a variation of the topic intrusion task developed by Chang *et al* (2009). Instead of identifying an intruder among a set of highly probable topics, we chose to identify the best match of a topic. We applied the best topic task because the LDA algorithm had shown that the majority of reviews were dominated by one topic. Hence, it was impossible to create random sets of three or four highly probable topics for any given review.) Again, for each topic we randomly selected reviews and presented them to human coders on Amazon Mechanical Turk. For each review, three topics were presented: one that was most probable and two that were randomly selected low-probability topics. Each topic was described by its 10 most probable words. Again, for each of the 18 topics we generated 5 tasks and 3 independent workers were prompted to identify which of the 3 topics best matched the review presented (again, resulting in 270 comparisons overall). The agreement between algorithm and coders, averaged over all topics, amounted to 85% (SD: 17%). (For both tasks, testing for differences in mean percentage agreement for randomly selected subsamples (50% of cases) of the 270 comparisons showed no significant differences, indicating that the experiments produced reliable results.)

Result interpretation

Interpreting the results of a black-box algorithm like random forests can be challenging, as the results of model fitting are not as easy to 'read' like, for example, the outputs of a logistic regression. Variable importance measures only provide a hint about the predictive power of individual variables. Without further analysis it remains unclear whether a variable has a positive or negative influence on the probability of belonging to a certain class (i.e., helpful or unhelpful). One way to shed more light on a random forest model is to plot the values of a selected independent variable against the class probabilities predicted by the model (i.e., predictions of the dependent variable) (Friedman *et al*, 2013). Figure 2 shows such partial dependence plots of three selected topic probabilities that are charted against the predicted probability of being helpful. Topic 40 (customer service), for example, exhibits a strong negative correlation with review helpfulness. If a reviewer spends more than 10% of his or her words on customer service issues, it is highly likely that the review, on average, will be perceived as unhelpful. In contrast, the more reviewers write about problems with digital rights management (Topic 89) or the appropriateness for kids (Topic 100), the more likely a review receives positive rather than negative helpfulness ratings.

The final step of the results interpretation phase is to compare and contrast the data-driven discoveries with extant theory and literature. While a full discussion is out of scope of this illustrative study, we want to highlight two interesting findings. First, as mentioned before, Topic 40 (customer service) has a strong negative impact on review helpfulness. A potential explanation for this observation might be that a considerable portion of people use Amazon as a source of product reviews but not as a purchasing platform. Consequently, judgments about Amazon's customer service, such as speed of delivery or return policies, are irrelevant (or not helpful) for these users, as they are not related to the quality of the product. Second, Topics 89 (digital rights management) and 100 (appropriateness for kids) show a strong positive relationship with review helpfulness. Both topics are connected to the concept of

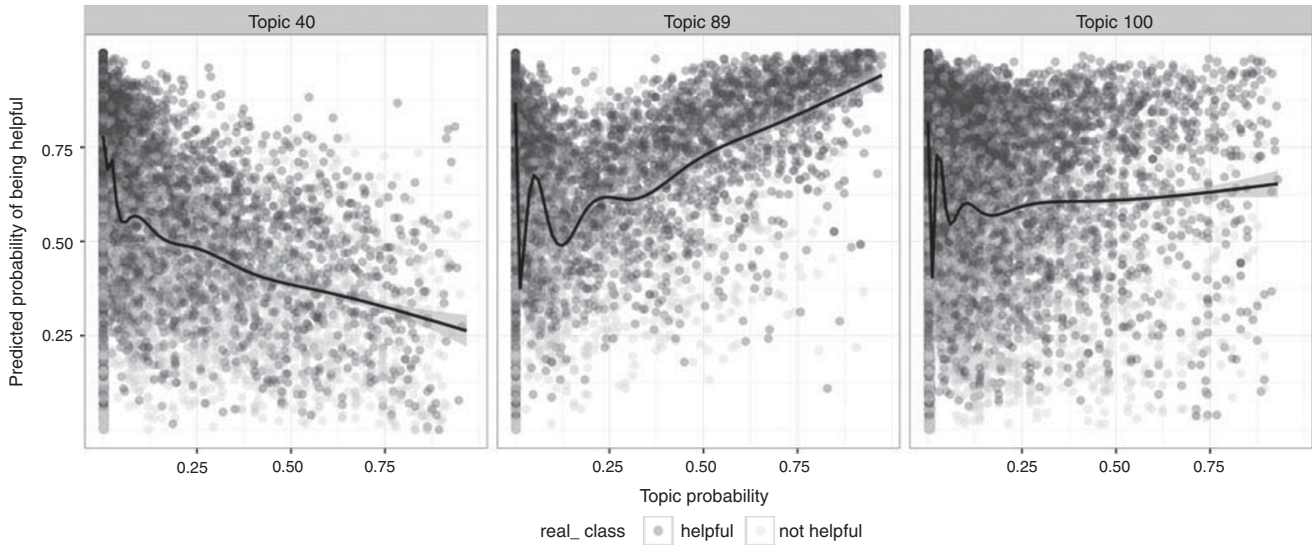

Figure 2 Topic probabilities *vs* predicted class probabilities.

experience goods – products that have qualities that are difficult for a consumer to evaluate in advance, but can be easily assessed after consumption (Nelson, 1970). In our case, customers find it helpful that other gamers or parents include warnings about digital rights management restrictions (especially the need to always stay connected to a remote server when playing) or inappropriate content (e.g., violence, sex). This reinforces the findings of prior research that has compared helpfulness ratings of experience goods *vs* search goods (e.g., Mudambi & Schuff, 2010).

Summary

When reflecting on the illustrative BDA study, a few findings should be noted. With regards to the research question and contribution, we have showcased how a BDA study that is data-driven and predictive in nature can nonetheless be the basis for theorisation. In terms of data collection, we illustrated the value of open data repositories and their remaining need for exploratory data analysis and data preparation. In the data analysis phase, we showed the potential of NLP and machine-learning techniques for discovering patterns and relationships in large and unstructured data sets, and how the results of such techniques can be triangulated with human judgements. And finally, we demonstrated the use of intuitive visualisations and theoretical triangulation for result interpretation. In the next section, we will reflect on the broader changes that BDA might bring to the IS discipline and propose a set of initial guidelines for conducting rigorous BDA studies in IS.

Initial BDA guidelines for IS researchers

Big data analytics signifies a novel and complementary approach to using large, diverse, and dynamic sets of user-generated content and digital trace data as well as new methods of analytics to augment scientific knowledge. But for BDA to take place, the emphasis we put on the various phases of the traditional research process might have to re-weighted. During the initial phase, this means that researchers should spend more time framing the research question and exploring its feasibility; during the data collection phase, it means that researchers – while able to reduce the amount of effort needed for the actual data collection – have to invest more time in understanding and preparing raw data; during the analysis phase, it means that researchers should verify their results based on new evaluation criteria (e.g., inter-coder reliability between humans and algorithms); and lastly, during the interpretation phase, it means that researchers need to spend an extraordinary amount of time making sense of the results, sometimes even necessitating an additional explanatory phase. Explicit periods of triangulation should enter the research process in more salient ways than before. As to theoretical triangulation, researchers will need to pay close attention to existing theoretical constructs in order to ensure the legitimacy of the research question as well as the interpretability of findings. Ensuring that studies are not merely built on volatile correlational relationships without enhancing theory is of vital importance. As to empirical triangulation, researchers will have to place their own empirical results in relation to others. Until sufficient evaluation criteria are developed, findings of first generation BDA studies should be contrasted to findings from studies that have applied more traditional strategies of enquiry. Table 2 summarises our initial set of guidelines for IS researchers applying BDA. Note that each of the guidelines has been addressed, to a certain extent, in the illustrative BDA study presented in the previous section. While we cannot claim that the

Table 2 A summary of guidelines for IS researchers applying BDA

Research phase	*Guidelines*
Research question	● Grow comfortable that research can start with data instead of theory ● Undergo a phase of theoretical triangulation in order to verify the vitality of a data-driven research question ● Position your study as an explanatory or predictive study and follow the corresponding methodology ● Particularly for predictive studies, do not only aim for practical utility, but plan to make a theoretical contribution ● Plan to re-adjust the time and effort spent on the various phases of the research process
Data collection	● Justify the selection of big data as the primary data source ● Discuss the nature of the data with regards to its validity and reliability ● Document the data collection process in detail; ensure transparency about the type and nature of data used ● If applicable, provide access to the data used in form of a database that accompanies the research paper
Data analysis	● Document the data pre-processing steps in detail, especially for studies applying NLP ● Algorithms evolve and are in flux at all times; rely on other disciplines, particularly computer science and machine learning, to ensure their validity ● If possible, provide access to the algorithms used for data analysis ● Apply empirical triangulation in order to ensure the statistical validity of the analysis; select appropriate evaluation criteria in order to ensure comparability with other studies
Result interpretation	● Make 'black box' algorithms transparent by adding an explicit explanatory phase to the research process ● Theoretically triangulate the results by drawing on existing theory; at the very least, discuss the results against the backdrop of existing studies ● If applicable, try to replicate findings using traditional data analysis methods

proposed set of guidelines is complete, nor that all of the presented guidelines will stand the test of time, we nevertheless contend that they represent a valid starting point for jointly and iteratively testing and revising quality criteria for BDA in IS research. Reflecting on the guidelines, we can observe that each phase of the research process requires a revised set of actions and abilities. A shift in weight and emphasis for the various phases of the research process will most probably also result in a skill set change for IS researchers. Stronger emphasis needs to be placed on developing skills for data preparation and the deployment of analytical tools and cross-instrumental evaluation criteria. IS researchers should extend their repertoire of statistical methods to also include approaches that go beyond statistical hypothesis testing. Among these are data mining and machine-learning algorithms, NLP techniques, as well as graphical methods for visualising large data sets in intuitive ways.

Conclusion

In the natural sciences, the evolution of the scientific method is often portrayed as four eras (Bell *et al*, 2009; Hey *et al*, 2009). In the early times, research was based on empirical observation; this was followed by an era of theoretical science, in which building causal models was cultivated. As models became more complex, an era of computational research using simulations emerged. Today, many natural science disciplines find themselves in an era of massive data exploration, in which data is captured in real time via ubiquitous sensors and continuously analysed through advanced statistical models. Jim Gray, the recipient of the Association of Computing Machinery's 1989 Turing Award, referred to this latter epoch as 'data-intensive scientific discovery', or simply 'the fourth paradigm' (Hey *et al*, 2009). The IS discipline is not a natural science, but rather studies socio-technical systems that are more difficult to measure and theorise. Hence, we do not argue that the advent of BDA represents an evolutionary step or paradigm shift for IS research. However, we must not ignore the potential that BDA as a novel and complementary data source and data analysis methodology can offer. In order to foster its diffusion, in this essay we discussed and illustrated its challenges and promises and proposed a set of initial guidelines intended to assist IS researchers in conducting rigorous BDA studies, and reviewers, editors, and readers in evaluating such studies. At the same time, we advise against a blind application of these guidelines, and recommend to check their applicability for each specific research project, and adapt and extend them where necessary.

About the authors

Oliver Müller is an Assistant Professor at the Institute of Information Systems at the University of Liechtenstein. He holds a Ph.D. and Master's Degree from the University of Münster, Germany. His research interests are decision

support systems, business analytics, text mining, and big data. His work has been published in internationally renowned journals (e.g., *Journal of the Association for Information Systems, Communications of the Association for Information Systems, Business & Information Systems Engineering, Computers & Education, IEEE Transactions on Engineering Management*) and presented at major international conferences.

Iris Junglas is an Associate Professor for Information Systems at Florida State University. Her research interest captures a broad spectrum of topics, most prominent are the areas of E-, M- and U-Commerce, health-care information systems, the consumerization of IT and business analytics. Her research has been published in the *European Journal of Information Systems, Information Systems Journal, Journal of the Association of Information Systems, Management Information Systems Quarterly, Journal of Strategic Information Systems*, and various others. She serves on the editorial board of the *Management Information Systems Quarterly Executive* and the *Journal of Strategic Information Systems* and is also a senior associate editor for the E*uropean Journal of Information Systems*.

Jan vom Brocke is professor for Information Systems at the University of Liechtenstein. He is the Hilti Endowed Chair of Business Process Management, Director of the Institute of Information Systems, Co-Director of the International Master Program in Information Systems, Director of the Ph.D. Program in Business Economics, and Vice-President Research and Innovation at the University of Liechtenstein. In his research he focuses on digital innovation and transformation capturing business process management, design science research, Green IS, and Neuro IS, in particular. His research has been published in *Management Information Systems Quarterly*, *Journal of Management Information Systems*, *Business & Information Systems Engineering*, *Communications of the Association for Information Systems*, *Information & Management* and others. He is author and editor of seminal books, including the *International Handbook on Business Process Management* as well as the book *BPM - Driving Innovationin a Digital World*. He has held various editorial roles and leadership positions in Information Systems research and education.

Stefan Debortoli is a research assistant and Ph.D. candidate at the Institute of Information Systems at the University of Liechtenstein. His doctoral studies focus on applying Big Data Analytics as a new strategy of inquiry in Information Systems Research. Within the field of Big Data Analytics, he focuses on the application of text mining techniques for research purposes. His work has been published in the *Business & Information Systems Engineering Journal*, the *Communications of the Association for Information Systems*, and the *European Conference on Information Systems*. Before joining the team, he has gained over 5 years of working experience in the field of software engineering and IT project management across various industries.

References

AGARWAL R and DHAR V (2015) Editorial – big data, data science, and analytics: the opportunity and challenge for IS research. *Information Systems Research* **25**(**3**), 443–448.

BECK A and SANGOI A (2011) Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science Translational Medicine* **3**(**108**), 108–113.

BELL G, HEY T and SZALAY A (2009) Beyond the data deluge. *Science* **323**(**5919**), 1297–1298.

BENDLER J, WAGNER S, BRANDT T and NEUMANN D (2014) Taming uncertainty in big data. *Business & Information Systems Engineering* **6**(**5**), 279–288.

BERENTE N and SEIDEL S (2014) Big data & inductive theory development: towards computational grounded theory? In *Proceedings of the Americas Conference on Information Systems* (TIWANA A and RAMESH B, Eds), Association for Information Systems, Savannah, USA.

BLEI D (2012) Probabilistic topic models. *Communications of the ACM* **55**(**4**), 77–84.

BLEI D, NG A and JORDAN M (2003) Latent dirichlet allocation. *Journal of Machine Learning Research* **3**(**1**), 993–1022.

BOLLEN J, MAO H and ZENG X (2011) Twitter mood predicts the stock market. *Journal of Computational Science* **2**(**1**), 1–8.

BOYD D and CRAWFORD K (2012) Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* **15**(**5**), 662–679.

BOYD-GRABER J, MIMNO D and NEWMAN D (2014) Care and feeding of topic models: problems, diagnostics, and improvements. In *Handbook of Mixed Membership Models and Their Applications* (AIROLDI EM, BLEI D, EROSHEVA EA and FIENBERG SE, Eds), pp. 3–34, CRC Press, Boca Raton.

BREIMAN L (2001a) Random forests. *Machine Learning* **45**(**1**), 5–32.

BREIMAN L (2001b) Statistical modeling: the two cultures. *Statistical Science* **16**(**3**), 199–231.

BRIGHTLOCAL (2014) Local consumer review survey 2014. [WWW document] https://www.brightlocal.com/2014/07/01/local-consumer-review-survey-2014/ (accessed 14 January 2016).

CAO Q, DUAN W and GAN Q (2011) Exploring determinants of voting for the 'helpfulness' of online user reviews: a text mining approach. *Decision Support Systems* **50**(**2**), 511–521.

CAVALLO A (2012) Scraped data and sticky prices. [WWW document] http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1711999 (accessed 6 July 2015).

CHANG J, BOYD-GRABER J, GERRISH S, WANG C and BLEI D (2009) Reading tea leaves: how humans interpret topic models. In *Proceedings of the Advances in Neural Information Processing Systems Conference* (BENGIO Y, SCHUURMANS D, LAFFERTY JD, WILLIAMS CKI and CULOTTA A, Eds), pp. 1–9, Neural Information Processing System, Vancouver, Canada.

CHATFIELD C (1995) *Problem Solving: A statistician's Guide*. Chapman & Hall, Boca Raton.

CHEN H, CHIANG R and STOREY V (2012) Business intelligence and analytics: from big data to big impact. *MIS Quarterly* **36**(**4**), 1165–1188.

CONSTANTIOU I and KALLINIKOS J (2015) New games, new rules: big data and the changing context of strategy. *Journal of Information Technology* **30**(**1**), 44–57.

CORTEZ P and EMBRECHTS M (2013) Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences* **225**(**1**), 1–17.

DAM G and KAUFMANN S (2008) Computer assessment of interview data using latent semantic analysis. *Behavior Research Methods* **40**(**1**), 8–20.

DAVENPORT T and KIM J (2013) *Keeping Up with the Quants: Your Guide to Understanding and Using Analytics*. Harvard Business School Press, Boston.

DE P, HU Y and RAHMAN M (2013) Product-oriented web technologies and product returns: an exploratory study. *Information Systems Research* **24**(**4**), 998–1010.

DHAR V (2013) Data science and prediction. *Communications of the ACM* **56**(**12**), 64–73.
DHAR V and CHOU D (2001) A comparison of nonlinear models for financial prediction. *IEEE Transactions on Neural Networks* **12**(**4**), 907–921.
DIAKOPOULOS N (2014) Algorithmic accountability reporting: on the investigation of black boxes. [WWW document] http://towcenter.org/research/algorithmic-accountability-on-the-investigation-of-black-boxes-2/ (accessed 3 July 2015).
EINAV L and LEVIN J (2014) Economics in the age of big data. *Science* **346**(**6210**), 715–721.
ENTERTAINMENT SOFTWARE ASSOCIATION (2015) Essential facts about the computer and video game industry – 2015 sales, demographic and usage data. [WWW document] http://www.theesa.com/wp-content/uploads/2015/04/ESA-Essential-Facts-2015.pdf (accessed 14 January 2016).
FAWCETT T (2006) An introduction to ROC analysis. *Pattern Recognition Letters* **27**(**8**), 861–874.
FRIEDMAN J, HASTIE T and TIBSHIRANI R (2013) *The Elements of Statistical Learning*. Springer, New York.
GHOSE A and IPEIROTIS P (2011) Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics. *IEEE Transactions of Knowledge and Data Engineering* **23**(**10**), 1498–1512.
GINSBERG J, MOHEBBI MH, PATEL RS, BRAMMER L, SMOLINSKI MS and BRILLIANT L (2009) Detecting influenza epidemics using search engine query data. *Nature* **457**(**7232**), 1012–1014.
GLASER B and STRAUSS A (1967) *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine Pub, Chicago.
GOH K, HENG C and LIN Z (2013) Social media brand community and consumer behavior: quantifying the relative impact of user-and marketer-generated content. *Information Systems Research* **24**(**1**), 88–107.
GREGOR S (2006) The nature of theory in information systems. *MIS Quarterly* **30**(**3**), 611–642.
GREGOR S and BENBASAT I (1999) Explanations from intelligent systems: theoretical foundations and implications for practice. *MIS Quarterly* **23**(**4**), 497–530.
HALEVY A, NORVIG P and PEREIRA F (2009) The unreasonable effectiveness of data. *IEEE Intelligent Systems* **24**(**2**), 8–12.
HEDMAN J, SRINIVASAN N and LINDGREN R (2013) Digital traces of information systems: sociomateriality made researchable. In *Proceedings of the International Conference on Information Systems* (BASKERVILLE R and CHAU M, Eds), Association for Information Systems, Milan, Italy.
HEY T, TANSLEY S and TOLLE K, Eds (2009) *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond.
HILBERT M and LÓPEZ P (2011) The world's technological capacity to store, communicate, and compute information. *Science* **332**(**60**), 60–65.
HU N, ZHANG J and PAVLOU PA (2009) Overcoming the J-shaped distribution of product reviews. *Communications of the ACM* **52**(**10**), 144.
IDC (2011) The 2011 digital universe study. [WWW document] https://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf (accessed 14 January 2016).
IDC (2014) The 2014 digital universe study. [WWW document] http://www.emc.com/leadership/digital-universe/index.htm#2014 (accessed 14 January 2016).
KAYANDE U, BRUYN DE A, LILIEN G, RANGASWAMY A and VAN BRUGGEN GH (2009) How incorporating feedback mechanisms in a DSS affects DSS evaluations. *Information Systems Research* **20**(**4**), 527–546.
KDNUGGETS (2011) Algorithms for data analysis/data mining. [WWW document] http://www.kdnuggets.com/polls/2011/algorithms-analytics-data-mining.html (accessed 14 January 2016).
KITCHIN R (2014) *The Data Revolution: Big Data, Open Data, Data Infrastructures and their Consequences*. Sage, Los Angeles.
KO D-G and DENNIS A (2011) Profiting from knowledge management: the impact of time and experience. *Information Systems Research* **22**(**1**), 134–152.
KORFIATISA K, GARCÍA-BARIOCANALB E and SÁNCHEZ-ALONSOB S (2012) Evaluating content quality and helpfulness of online product reviews: the interplay of review helpfulness *vs*. review content. *Electronic Commerce Research and Applications* **11**(**3**), 205–2017.
KRAMER A, GUILLORY J and HANCOCK J (2014) Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Science of the United States of America* **111**(**24**), 8788–8790.
KUHN M and JOHNSON K (2013) *Applied Predictive Modeling*. Springer, New York.
KUMAR A and TELANG R (2012) Does the web reduce customer service cost? Empirical evidence from a call center. *Information Systems Research* **23**(**3**), 721–737.
LANEY D (2001) 3D data management: controlling data volume, velocity, and variety. [WWW document] http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (accessed 14 January 2016).
LAU J, BALDWIN T and NEWMAN D (2013) On collocations and topic models. *ACM Transactions on Speech and Language Processing* **10**(**3**), 1–14.
LAZER D, KENNEDY R, KING G and VESPIGNANI A (2014) The parable of google flu: traps in big data analysis. *Science* **343**(**6176**), 1203–1205.
LAZER D *et al* (2009) Life in the network: the coming age of computational social science. *Science* **323**(**5915**), 721–723.
LESAFFRE E, RIZOPOULOS D and TSONAKA R (2007) The logistic transform for bounded outcome scores. *Biostatistics* **8**(**1**), 72–85.
LIN M, LUCAS HC and SHMUELI G (2013) Too big to fail: large samples and the *P*-value problem. *Information Systems Research* **24**(**4**), 906–917.
LYCETT M (2013) 'Datafication': making sense of (big) data in a complex world. *European Journal of Information Systems* **22**(**4**), 381–386.
MARTENS D, BAESENS B, VAN GESTEL T and VANTHIENEN J (2007) Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research* **183**(**13**), 1466–1476.
MARTENS D and PROVOST F (2014) Explaining data-driven document classifications. *MIS Quarterly* **38**(**1**), 73–99.
MARTIN M (2012) C-path: updating the art of pathology. *Journal of the National Cancer Institute* **104**(**16**), 1202–1204.
MARTON A, AVITAL M and JENSEN TB (2013) Reframing open big data. In *Proceedings of the European Conference on Information Systems* (BATENBURG R, VAN HILLEGERSBERG J, VAN HECK E, SPIEKERMANN S and CONNOLLY R, Eds), Association for Information Systems, Utrecht, The Netherlands.
MCAULEY J, PANDEY R and LESKOVEC J (2015a) Inferring networks of substitutable and complementary products. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (LESKOVEC J, WANG W and GHANI R, Eds), Association for Computing Machinery, Sydney.
MCAULEY J, TARGETT C, SHI Q and VAN DEN HENGEL A (2015b) Image-based recommendations on styles and substitutes. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval* (LALMAS M, MOFFAT A and RIBEIRO-NETO B, Eds), Association for Computing Machinery, Chile.
MICHEL J-B *et al* (2011) Quantitative analysis of culture using millions of digitized books. *Science* **331**(**6014**), 176–182.
MILES M and HUBERMAN A (1994) *Qualitative Data Analysis: An Expanded Sourcebook*. Sage Publications, Thousand Oaks.
MUDAMBI S and SCHUFF D (2010) What makes a helpful review? A study of customer reviews on amazon.com. *MIS Quarterly* **34**(**1**), 185–200.
MURRAY-RUST P (2008) Open data in science. *Serials Review* **34**(**1**), 52–64.
NELSON P (1970) Information and consumer behavior. *Journal of Political Economy* **78**(**2**), 311–329.
OESTREICHER-SINGER G and SUNDARARAJAN A (2012) Recommendation networks and the long tail of electronic commerce. *Management Information Systems Quarterly* **36**(**1**), 65–83.
PAN Y and ZHANG J (2011) Born unequal: a study of the helpfulness of user-generated product reviews. *Journal of Retailing* **87**(**4**), 598–612.
PENG R (2011) Reproducible research in computational science. *Science* **334**(**6060**), 1226–1227.
PIGLIUCCI M (2009) The end of theory in science? *EMBO Reports* **10**(**6**), 534.
REXER K (2013) 2013 Data miner survey – summary report. [WWW document] http://www.rexeranalytics.com/Data-Miner-Survey-Results-2013.html (accessed 14 January 2016).
RIMM D (2011) C-path: a Watson-like visit to the pathology lab. *Science Translational Medicine* **3**(**108**), 108.
ROBNIK-SIKONJA M and KONONENKO I (2008) Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering* **20**(**5**), 589–600.

SHARMA R, MITHAS S and KANKANHALLI A (2014) Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations. *European Journal of Information Systems* **23**(**4**), 433–441.

SHMUELI G (2010) To explain or to predict? *Statistical Science* **25**(**3**), 289–310.

SHMUELI G and KOPPIUS O (2011) Predictive analytics in information systems research. *MIS Quarterly* **35**(**3**), 553–572.

SILVERMANN D (2006) *Interpreting Qualitative Data*. Sage Publications, London.

SOCHER R *et al* (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (LAPATA M and NG HT, Eds), Association for Computational Linguistics, Seattle, Washington DC.

SPEER S (2002) 'Natural' and 'contrived' data: a sustainable distinction? *Discourse Studies* **4**(**4**), 511–525.

STIEGLITZ S and DANG-XUAN L (2013) Emotions and information diffusion in social media – sentiment of microblogs and sharing behavior. *Journal of Management Information Systems* **29**(**4**), 217–247.

TRENZ M and BERGER B (2013) Analyzing online customer reviews-an interdisciplinary literature review and research agenda. In *European Conference on Information Systems* (BATENBURG R, VAN HILLEGERSBERG J, VAN HECK E, SPIEKERMANN S and CONNOLLY R, Eds), Association for Information Systems, Utrecht, The Netherlands.

TURNEY P and PANTEL P (2010) From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research* **37**(**1**), 141–188.

VENKATESH V, BROWN S and BALA H (2013) Bridging the qualitative-quantitative divide: guidelines for conducting mixed methods research in information systems. *MIS Quarterly* **37**(**1**), 21–54.

VOM BROCKE J, DEBORTOLI S, MÜLLER O and REUTER N (2014) How in-memory technology can create business value: insights from the Hilti case. *Communications of the Association for Information Systems* **34**(**7**), 151–168.

WU X *et al* (2008) Top 10 algorithms in data mining. *Knowledge and Information Systems* **14**(**1**), 1–37.

YOO Y (2015) It is not about size: a further thought on big data. *Journal of Information Technology* **30**(**1**), 63–65.

YU C, JANNASCH-PENNEL A and DIGANGI S (2011) Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. *The Qualitative Report* **16**(**3**), 730–744.

ZENG X and WEI L (2013) Social ties and user content generation: evidence from Flickr. *Information Systems Research* **24**(**1**), 71–87.

ZIMMER M (2010) 'But the data is already public': on the ethics of research in facebook. *Ethics and Information Technology* **12**(**4**), 313–325.