Bus Inf Syst Eng 61(3):375–380 (2019)
https://doi.org/10.1007/s12599-019-00589-0

CATCHWORD

Augmented Analytics

Nicolas Prat

Received: 15 February 2018 / Accepted: 16 July 2018 / Published online: 25 February 2019

1 Augmented Analytics: Applying Artificial Intelligence Throughout the Analytics Cycle

Business intelligence (BI) and analytics are "*the techniques, technologies, systems, practices, methodologies, and applications that analyze critical business data to help an enterprise better understand its business and market and make timely business decisions*" (Chen et al. 2012, p. 1166). Although the two terms are sometimes used jointly or interchangeably, BI often refers to reporting, OnLine Analytical Processing (OLAP), dashboards and scorecards, while analytics typically uses advanced techniques based on machine learning. A new term, "augmented analytics", coined by Gartner (2017a), is shifting the lines between BI and advanced analytics, empowering BI users with advanced machine learning techniques and artificial intelligence.

Augmented analytics brings automation to the complete analytics cycle through the application of artificial intelligence (AI), more specifically machine learning and natural language processing (NLP). Whatever term is used to designate AI-powered analytics (Gartner 2017a; Watson 2017; Henschen 2018), this is clearly a turn in the history of BI. The first generation of BI was the generation of data warehouses in the 1990s. The second generation was the one of big data analytics, with the rise of analytics in the mid-2000s, followed by the big data hype in the 2010s. It was also the generation of self-service BI (Alpar and Schulz 2016), with the emergence of powerful data-discovery tools enabling business users to explore data for insights and decision making without systematically resorting to the IT department. The third generation, starting in 2015, is that of AI-powered analytics. AI-powered analytics pushes self-service BI further: business users or analysts gain access to advanced analytics, hence the new concept of "citizen data scientist", i.e., "*a person who creates or generates models that use advanced diagnostic analytics or predictive and prescriptive capabilities, but whose primary job function is outside the field of statistics and analytics*" (Gartner 2017b). Data scientists also benefit from AI-powered analytics. Thus, one major factor explaining the interest in AI-powered analytics is the shortage of data scientists (Knight 2017). More generally, in a context of exponential flow of data (big data), augmented analytics is a promising solution to optimize the use of these data for decision making, by bringing automation to the complete analytics cycle.

To further characterize augmented analytics and detail its applications through the analytics cycle, we need to specify the phases of this cycle. Several analytics cycles or process models have been proposed, e.g., in (Erl et al. 2015; SAS 2016; Storey and Song 2017; Seddon et al. 2017). They differ in their focus and several draw on the CRISP-DM process model for data mining (Shearer 2000). The seven-phase cycle shown in Fig. 1 synthetizes and builds upon these models. It starts by identifying the business problem addressed by analytics, as well as opportunities of big data analytics for the business. Data preparation (a.k.a. wrangling) follows. It is decomposed into data profiling (quality assessment) and transformation. The data analysis phase distinguishes between data discovery (generally by analysts or business users) and model

Accepted after one revision by Prof. Dr. Weinhardt.

Prof. Dr. N. Prat (✉)
Department of Information Systems, Decision Sciences and Statistics, ESSEC Business School, 1, avenue Bernard Hirsch, CS 50105 Cergy, 95021 Cergy-Pontoise Cedex, France
e-mail: prat@essec.edu

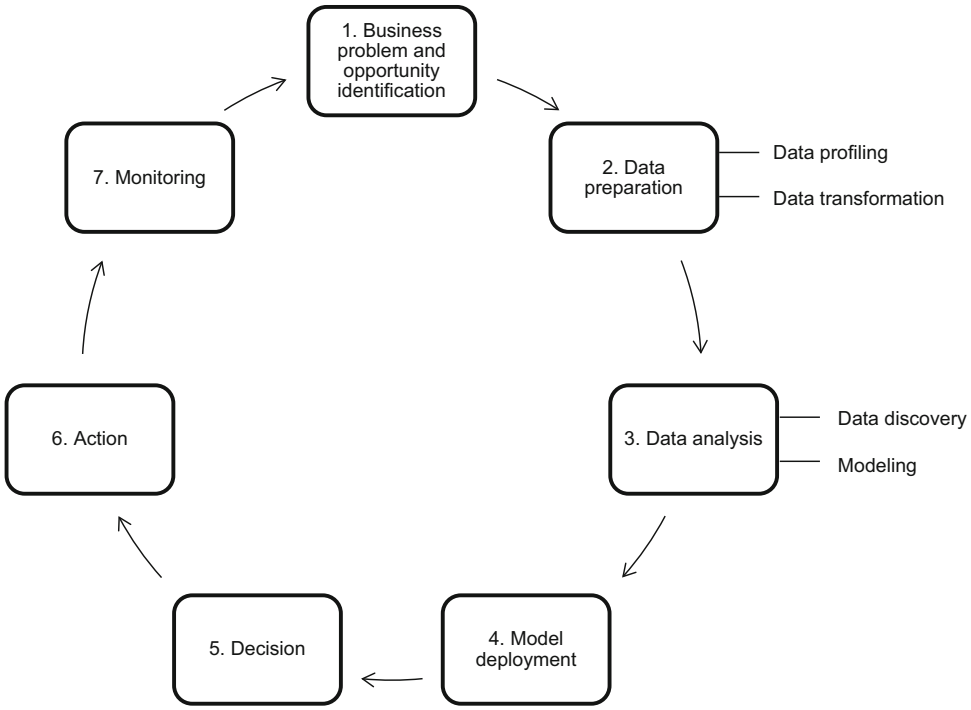

Fig. 1 Analytics cycle

building and evaluation (a task performed by data scientists). Once built and evaluated, the models are deployed in production systems. Decision making and action taking follow. Finally, monitoring reviews the action(s) taken and the performance of models, and the cycle starts again.

Note that there exist many possible instantiations of the analytics cycle, which typically implies several iterations within or between the phases. In each phase, specific methods may be used, requiring expertise in different areas. For example, knowledge of data quality assessment and improvement methods (Batini et al. 2009) and knowledge of data modeling (Storey and Song 2017) are important competencies in data preparation; knowledge of storytelling (Nussbaumer Knaflic 2015) is required in data analysis… Similarly, depending on the phase of the cycle and the organizational context, several types of stakeholders may be involved, including IT, analysts, business users, and data scientists. All phases and stakeholders of the analytics cycle may benefit from and be impacted by augmented analytics. This is what differentiates it from smart (or augmented) data discovery (Gartner 2015), which focuses on data discovery and turns domain specialists, business users, or engineers into "citizen data scientists" (Gartner 2017b; Gröger 2018).

In the following sections, we start by reviewing the current state of augmented analytics. We then delve into the limits and issues of AI in analytics. Based on these limits and issues, we identify opportunities for information systems research in augmented analytics. Even if "augmented analytics" is commonly understood as analytics augmented (i.e., powered) by AI, the term may be used differently in other contexts. More specifically, it may refer to immersive analytics in augmented reality environments (Chandler et al. 2015; Stein et al. 2018). Immersive analytics investigates how immersive environments (in virtual or augmented reality) may be used to support analytical reasoning and decision making (Chandler et al. 2015). Immersive analytics is the modern version of visual analytics. The latter combines automated analysis techniques (e.g., data mining) and interactive visualization to analyze large and complex data sets (a.k.a. big data sets) (Keim et al. 2008). While acknowledging the fact that "augmented analytics" may also evoke the domains of visual and immersive analytics, throughout this paper, we use the term to refer to AI-powered analytics, as proposed by Gartner (2017a).

2 Applications

To review the current state of augmented analytics, Table 1 shows the applications of AI through the phases of the analytics cycle. These applications differ in maturity. For example, the suggestion of visualizations for pre-selected data appeared with self-service BI, while applications of NLP are far less mature. We illustrate with examples of tools in the market that propose the applications mentioned. The list is not meant to be exhaustive. Its sole propose is to illustrate that the applications are implemented in software tools.

Table 1 Applications of AI through the analytics cycle

Phase of the analytics cycle	Applications of AI
1. Business problem and opportunity identification	(Possibly feeds on input from previous cycles)
2. Data preparation	
Data profiling	Automatic assessment of data quality
Data transformation	Suggestions for data cleansing, restructuring, blending and enrichment
3. Data analysis	
Data discovery	Suggestions of visualizations for pre-selected data, enhancement of visualizations with advanced analytics, guidance in data discovery, natural language data exploration, natural language generation
Modeling	Feature engineering, model tournaments
4. Model deployment	Direct model deployment and embedding into production systems
5. Decision	(Possibly automated)
6. Action	(Possibly automated)
7. Monitoring	Dynamic adjustment of models

The phase of business-driven problem and opportunity identification is often considered outside the scope of BI and analytics tools and is not easily amenable to AI-based automation. However, it may benefit from inputs from previous cycles in the analytics process. For example, insights generated with AI-powered data-discovery tools may result in the identification of new problems or new opportunities.

Data preparation takes a significant amount of time in the analytics process, sometimes as much as 80% (Lohr 2014). Therefore, automating this phase may dramatically increase the productivity of analytics and enable data scientists and analysts to allocate their time to more value-adding phases. Thanks to AI, the tools on the market, e.g., Trifacta Wrangler,[1] bring automation to the iterative cycle of data profiling and transformation. These tools consider both the syntax and semantics of data and support different formats of small and big data, with a focus on structured and semi-structured data. Data profiling is partly automated, e.g., by detecting outliers, null values, inconsistent values, or abnormal data distributions. Transformations are suggested for data cleaning (treatment of null values, standardization…), reorganization (column splitting, aggregation…), blending and enrichment (identification of join columns or suggestion of new data sets).

In data discovery, visualization tools like Tableau[2] suggest visualization types (map, scatter plot…) with predefined parameters, based on the data selected for a visualization. Visualizations may be enhanced with advanced analytics such as clustering or forecasting. Watson Analytics[3] guides data discovery by analyzing data and automatically suggesting visualizations. The list of relevant visualizations, ordered by relevancy, is updated as data discovery proceeds. With the progress of NLP, tools like Watson Analytics enable data querying in natural language (e.g., "What is the cost of courses by organization?"). The syntax for asking questions is constrained and having a real dialogue with tools is challenging, requiring them to memorize the context of previous queries. However, AI is making progress on that front (Henschen 2018). Other tools, like Narratives for Tableau,[4] automatically generate insights in natural language from visualizations, synthetizing what is important (e.g., trends, best performers, aggregates…).

AI supports the model building and evaluation phase. Tools like Driverless AI[5] automate feature engineering, which prepares the variables to be used by machine-learning algorithms. "Model tournaments" (SAS 2016) apply machine learning to automate machine learning (Knight 2017): millions of combinations of features, machine-learning algorithms and model parameters may be tested and ranked on their performance. This not only improves the productivity of modeling, but also reduces the risk of biases towards certain algorithms (Knight 2017). An example of system automating model tournaments for predictive analytics is DataRobot.[6]

The transition between modeling and model deployment in production systems often lacks fluidity. This is partly due to the change of IT environments between these two phases, as well as the change of actors, typically from data

[1] https://www.trifacta.com/products/wrangler/. Accessed 28 Feb 2018.

[2] https://www.tableau.com. Accessed 28 Feb 2018.

[3] https://www.ibm.com/watson-analytics. Accessed 1 Mar 2018.

[4] https://narrativescience.com/Partners/Business-Intelligence/Tableau. Accessed 1 Mar 2018.

[5] https://www.h2o.ai/driverless-ai/. Accessed 28 Feb 2018.

[6] https://www.datarobot.com/. Accessed 6 Mar 2018.

scientists to IT (SAS 2016). AI-powered automation facilitates the transition between the two worlds by enabling direct model deployment and embedding into production systems without requiring lengthy recoding. Alteryx Promote,[7] for example, automates the deployment of predictive models. Automation extends to model monitoring (Kobielus 2017): to optimize the predictive performance of models in production, they are automatically retrained with fresh data, and redeployed as necessary.

Final decision making and action taking are often considered outside the scope of analytics tools (see the process model of Seddon et al. (2017) for example). However, with the advent of big data, operational decisions are increasingly automated, by deploying and executing machine-learning models. This may lead to automated action immediately following decision, as in the case of high-frequency trading.

Beyond the applications of AI currently implemented to varying degrees in analytics tools and summarized in Table 1, other applications are likely to emerge. These may include applications that we cannot imagine today. However, the limits and issues of augmented analytics should be addressed, leading to research opportunities for the information systems (IS) community.

3 Limits and Research Issues

This section reviews the main limits and issues of AI-powered analytics. From there, it identifies research opportunities in augmented analytics, for the main research approaches in IS: behavioral research, design science research, and economics of IS.

3.1 Limits and Issues of Artificial Intelligence in Analytics

A major limit of AI-powered analytics is its dependence on input data (Underwood 2017). AI-enabled automation does not eliminate the need for careful data selection and human intervention in data preparation. Data quality governance is even more crucial as augmented analytics democratizes access to data selection and preparation. Beyond data quality issues, machine-learning algorithms are subject to biases, some of which may result from biases in the data used to train these algorithms (Brynjolfsson and McAfee 2017). Thus, trust and transparency are crucial in ensuring the success of augmented analytics (Henschen 2018). For some algorithms, such as those based on neural networks, providing transparency and explaining the results of models is challenging.

Some limits of augmented analytics are more specifically related to certain phases in the analytics cycle. Business problem and opportunity identification heavily relies on managers and business users. In this crucial phase, a major issue is finding the business problem addressed by analytics (e.g., "Improve the retention of high-value customers in the tablet segment", "Prevent product shrinkage in the warehouse"). Machines may be very good at solving problems, but posing problems is inherently human (Brynjolfsson and McAfee 2017). In the data preparation phase, human judgment remains essential, e.g., in the interpretation of outliers. Finally, automating decisions and subsequent actions is limited to operational decisions. Many decisions require a sense of ethics, empathy, and other capacities that, at the current stage of AI research, remain the preserve of humans.

Beyond the limits of AI-enabled automation, augmented analytics raises many issues related to technologies, people, processes, and their interactions. One issue is the redefinition of the roles of the actors in the analytics cycle, following the changes brought about by automation. For example, if model building and evaluation are increasingly automated, how should the role of data scientists evolve, what are their most added-value activities beyond modeling? One other major challenge is the orchestration of the analytics process. This orchestration is complex because it generally involves different categories of stakeholders, as well as different tools and IT environments. Democratized access to analytics thanks to AI automation makes the governance of analytics even more challenging, e.g., to ensure the quality of data and the compliance to common standards. What further complicates the orchestration and governance of the analytics process is the fact that it is not purely sequential and may be instantiated in many different ways (Seddon et al. 2017).

3.2 Research Directions

The limits and issues identified above suggest research avenues for IS. For example, IS academics should focus on ways to measure data veracity more holistically. Veracity is a multidimensional concept. For textual information, it comprises three dimensions (Lukoianova and Rubin 2014): objectivity, truthfulness, and credibility (a.k.a. believability). In big data analytics, data are often uncertain by nature (e.g., weather data, the future behavior of consumers…) (IBM 2012). Even if total veracity may not be guaranteed, the data may still be useful for decision-making, but decision makers should know their degree of veracity. A holistic measure of veracity would facilitate veracity improvement, transparency, and would likely positively

[7] https://www.alteryx.com/products/alteryx-promote. Accessed 3 Mar 2018.

affect trust in augmented analytics. Another research avenue concerns the governance issues of data analytics, e.g., to control the quality of data or orchestrate the analytics process.

To identify research directions for augmented analytics, our approach draws on Abbasi et al. (2016). These authors propose a big data research agenda in IS by considering the interplay between the characteristics of big data, the information value chain, and the main research approaches in IS (behavioral, design, and economics of IS). Here, we consider the interplay between AI (instead of big data characteristics), the analytics cycle (instead of the information value chain), and research approaches.

Behavioral research – quantitative or qualitative – may investigate questions such as the following: What is the impact of different governance mechanisms (e.g., procedures and roles) on the effective use of augmented analytics? How should the role of data scientists evolve in the age of augmented analytics, in what tasks (beyond modeling) do they add most value? Should all business users take the role of "citizen data scientists", or should a specific category of business users be devoted to this role and, if so, what category? To what extent does AI affect the perceived usefulness and perceived ease of use (Davis 1989) of analytics by business users? What are the major determinants of trust and credibility in augmented analytics? To what extent does augmented analytics enable decision makers to make better decisions?

In *design science research*, conceptual modeling may help in addressing several issues, in the same way as it is relevant in big data research (Storey and Song 2017). For example, research in conceptual modeling has a long tradition in data integration, representation and exploitation of semantics, and information or data quality assessment. All these topics are especially relevant in the data profiling and transformation phase. One issue worth investigating is the assessment of believability (an important dimension of veracity) based on the provenance (a.k.a. lineage) of data (Prat and Madnick 2008). In the context of augmented analytics, data preparation tools generally capture metadata, including the tracing of data lineage along the transformation process. This facilitates the provenance-based evaluation of the different sub-dimensions of data believability and, more generally, the computation of quality scores at different levels of detail. Beyond data preparation, design science research may also contribute to other phases in the analytics cycle, e.g., data discovery. A key feature of data-discovery tools is the ability to navigate data at different aggregation levels (rollup or drill-down). Not all rollup or drill-down operations allowed by data-discovery tools make sense, and users may be guided in the aggregation process, e.g., with semantic or syntactic aggregation rules as suggested by Prat et al. (2011). The aggregation rules proposed by these authors may be extended and implemented in rule-based expert systems, which have long been a major area of AI.

For researchers in *IS economics*, an essential question is the value provided by automating analytics with AI. What are the productivity gains from AI in analytics, and, more generally, how is the value of augmented analytics (as opposed to more traditional analytics) computed? Another question is the impact of augmented analytics on the job market, as the roles of data scientists and other key actors in the analytics cycle evolve.

Finally, augmented analytics does not only raise many issues for IS academics. It is also a new tool for researchers to conduct their investigations. As stated by Agarwal and Dhar (2014, p.447), "*As a community of scholars we would be remiss not to take full advantage of the scientific possibilities created by the availability of big data, sophisticated analytical tools, and powerful computing infrastructures.*" Big data provides a wealth of material for research, and augmented analytics eases the preparation and analysis of these data by "citizen data scientists", including (IS) academics.

References

Abbasi A, Sarker S, Chiang RHL (2016) Big data research in information systems: toward an inclusive research agenda. J Assoc Inf Syst 17 (2):i-xxxii

Agarwal R, Dhar V (2014) Editorial—Big data, data science, and analytics: the opportunity and challenge for IS research. Inf Syst Res 25(3):443–448

Alpar P, Schulz M (2016) Self-service business intelligence. Bus Inf Syst Eng 58(2):151–155

Batini C, Cappiello C, Francalanci C, Maurino A (2009) Methodologies for data quality assessment and improvement. ACM Comput Surv 41(3):1–52

Brynjolfsson E, McAfee A (2017) The business of artificial intelligence: What it can – and cannot – do for your organization. Harvard Business Review Digital Articles: 3-11. https://hbr.org/cover-story/2017/07/the-business-of-artificial-intelligence

Chandler T, Cordeil M, Czauderna T, Dwyer T, Glowacki J, Goncu C, Klapperstueck M, Klein K, Marriott K, Schreiber F, Wilson E (2015) Immersive analytics. In: IEEE international symposium on big data visual analytics, Hobart, pp 1–8

Chen H, Chiang RHL, Storey VC (2012) Business intelligence and analytics: from big data to big impact. MIS Q 36(4):1165–1188

Davis FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 13(3):319–340

Erl T, Khattak W, Buhler P (2015) Big data fundamentals: concepts, drivers & techniques. Prentice Hall, Upper Saddle River

Gartner (2015) Smart data discovery will enable a new class of citizen data scientist. https://www.gartner.com/doc/3084217/smart-data-discovery-enable-new. Accessed 02 Mar 2018

Gartner (2017a) Augmented analytics is the future of data and analytics (overview of the report in Rita Sallam's public article). https://blogs.gartner.com/rita-sallam/2017/07/31/just-buying-

into-modern-bi-and-analytics-get-ready-for-augmented-analytics-the-next-wave-of-market-disruption/. Accessed 20 Feb 2018
Gartner (2017b) Gartner says more than 40 percent of data science tasks will be automated by 2020. https://www.gartner.com/newsroom/id/3570917. Accessed 25 June 2018
Gröger C (2018) Building an industry 4.0 analytics platform. Datenbank-Spektrum 18(1):5–14. https://doi.org/10.1007/s13222-018-0273-1
Henschen D (2018) How ML and AI will transform business intelligence and analytics. http://www.zdnet.com/article/how-machine-learning-and-artificial-intelligence-will-transform-business-intelligence-and-analytics/. Accessed 1 Mar 2018
IBM (2012) Analytics: The real-world use of big data. https://www.ibm.com/services/us/gbs/thoughtleadership/ibv-big-data-at-work.html. Accessed 28 June 2018
Keim D, Andrienko G, Fekete J-D, Görg C, Kohlhammer J, Melançon G (2008) Visual analytics: definition, process, and challenges. In: Kerren A, Stasko JT, Fekete J-D, North C (eds) Information visualization: human-centered issues and perspectives. Springer, Heidelberg, pp 154–175
Knight W (2017) You could become an AI master before you know it. Here's how. https://www.technologyreview.com/s/608921/you-could-become-an-ai-master-before-you-know-it-heres-how/. Accessed 28 Feb 2018
Kobielus J (2017) Even data scientists are facing AI takeover. https://www.infoworld.com/article/3234465/data-science/even-data-scientists-are-facing-ai-automation.html. Accessed 28 Feb 2018
Lohr S (2014) For big-data scientists, 'janitor work' is key hurdle to insights. https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html. Accessed 05 Mar 2018
Lukoianova T, Rubin VL (2014) Veracity roadmap: Is big data objective, truthful and credible? Adv Classif Res Online 24(1):4–15
Nussbaumer Knaflic C (2015) Storytelling with data: a data visualization guide for business professionals. Wiley, Hoboken
Prat N, Madnick S (2008) Measuring data believability: a provenance approach. In: 41st Annual Hawaii international conference on system sciences, Big Island
Prat N, Comyn-Wattiau I, Akoka J (2011) Combining objects with rules to represent aggregation knowledge in data warehouse and OLAP systems. Data Knowl Eng 70(8):732–752
SAS (2016) Managing the analytical life cycle for decisions at scale. https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/manage-analytical-life-cycle-continuous-innovation-106179.pdf. Accessed 20 Feb 2018
Seddon PB, Constantinidis D, Tamm T, Dod H (2017) How does business analytics contribute to business value? Inf Syst J 27(3):237–269
Shearer C (2000) The CRISP-DM model: the new blueprint for data mining. J Data Warehous 5(4):13–22
Stein M, Janetzko H, Lamprecht A, Breitkreutz T, Zimmermann P, Goldlücke B, Schreck T, Andrienko G, Grossniklaus M, Keim DA (2018) Bring it to the pitch: combining video and movement data to enhance team sport analysis. IEEE Trans Vis Comput Graph 24(1):13–22
Storey VC, Song I-Y (2017) Big data technologies and management: What conceptual modeling can do. Data Knowl Eng 108:50–67
Underwood J (2017) We're entering a new era of augmented analytics. https://www.informationweek.com/big-data/were-entering-a-new-era-of-augmented-analytics/d/d-id/1329593. Accessed 19 Feb 2018
Watson H (2017) The cognitive decision-support generation. Bus Intell J 22(2):5–14